THE ROLE OF GENETIC AND ENVIRONMENTAL FACTORS IN THE AETIOLOGY OF ESOPHAGEAL CANCER

Hannah Simba

Dissertation presented for the degree of Doctor of Philosophy in Public Health in the Faculty of Medicine and Health Sciences,
Stellenbosch University

Supervisor:
Professor Vikash Sewram
African Cancer Institute, Department of Global Health
Faculty of Medicine and Health Sciences
Stellenbosch University

Co-Supervisors:
Professor Helena Kuivaniemi
Division of Molecular Biology and Human Genetics
Department of Biomedical Sciences
Faculty of Medicine and Health Sciences
Stellenbosch University

Professor Gerard Tromp
Division of Molecular Biology and Human Genetics
Department of Biomedical Sciences
Faculty of Medicine and Health Sciences
Stellenbosch University

Dr Christian Abnet
Division of Cancer Epidemiology and Genetics
National Cancer Institute, National Institutes of Health
Bethesda, USA

Declaration of originality

By submitting this thesis electronically, I declare that the entirety of the work contained therein is my own, original work, that I am the sole author thereof (save to the extent explicitly otherwise stated), that reproduction and publication thereof by Stellenbosch University will not infringe any third party rights and that I have not previously in its entirety or in part submitted it for obtaining any qualification.

This thesis includes one published manuscript, a systematic review: Systematic Review of Genetic Factors in the Etiology of Esophageal Squamous Cell Carcinoma in African Populations. 2019 Hannah Simba, Helena Kuivaniemi, Vittoria Lutje, Gerard Tromp and Vikash Sewram. Frontiers in Genetics. 10:642; 1-20 https://doi.org/10.3389/fgene.2019.00642. I had the principal responsibility of developing and writing this manuscript and the nature and extent of contribution of co-authors is indicated accordingly.

Declaration by the candidate:

The nature and scope of my contribution regarding the published systematic review (Chapter 2), is as follows: I was the project lead, carried out literature searches, appraised the articles, summarized the results, prepared the tables and figures, drafted the manuscript and revised the manuscript when submitted for publication.

Nature of contribution Extent of contribution (%)

The following co-authors have contributed to Chapter 2:

Name	e-mail address	Nature of contribution	Extent of contribution (%)
Hannah Simba	@sun.ac.za	Project lead, carried out literature searches, appraised the articles, summarized the results, prepared the tables and figures, drafted the manuscript and revised the manuscript when submitted for publication	70
Helena Kuivaniemi	@sun.ac.za	Conceptualized the idea for the research, supervised the project, appraised the articles, summarized the results, prepared the tables and figures, drafted the manuscript	9
Vittoria Lutje	@lstmed.ac.uk	Carried out literature searches, provided specialist expertise and knowledge, and critically reviewed the manuscript.	3
Gerard Tromp	@sun.ac.za	Carried out the r^2 analyses, prepared the r^2 figure and table, and critically reviewed and revised the manuscript	9
Vikash Sewram	@sun.ac.za	Conceptualized the idea for the research, supervised the project, carried out literature searches, appraised the articles, summarized the results, prepared the tables and figures, and critically reviewed the manuscript.	9

Signature of candidate: ....

Date: ....27/05/2021..................

Declaration by co-authors:

The undersigned hereby confirm that

1. the declaration above accurately reflects the nature and extent of the contributions of the candidate and the co-authors to Chapter 2,
2. no other authors contributed to Chapter 2 besides those specified above, and
3. potential conflicts of interest have been revealed to all interested parties and that the necessary arrangements have been made to use the material in Chapter 2 of this dissertation.

Signature	**Institutional affiliation**	**Date**
	Stellenbosch University, South Africa	27.5.2021
	Cochrane Infectious Diseases Group, Liverpool, United Kingdom	02.06.2021
	Stellenbosch University, South Africa	01.06.2021
	Stellenbosch University, South Africa	06.06.2021

Abstract

Esophageal cancer (EC) is an aggressive cancer contributing an estimated 572,034 new cases and 508,585 deaths annually. Because no early detection programs exist, late presentation and high mortality are the rule. Prevalence rates are high in East Asia, Southern Europe, as well as in Eastern and Southern Africa. This peculiar distribution draws attention on the specificity of certain risk factors to particular regions. South Africa is a hotspot for EC; high prevalence has been reported in the Eastern Cape for the past five decades. Little research attention is given to EC in Africa; therefore, the epidemiology, as well as the genetic and environmental basis of EC is not well understood. The high incidence of EC, and the fatal nature of the disease, warrants a dedicated study to understand risk factors and pathobiology to facilitate strategies on prevention and screening.

The aim of this study was to assess the role of genetic and environmental factors in the development of EC, and investigate the underlying molecular pathobiology using gene expression. Genetic variants associated with esophageal squamous cell carcinoma (ESCC) in African populations were assessed in 23 studies. Altogether, 25 variants in 20 genes were reported with a statistically significant association. In addition, eight studies identified somatic alterations in 17 genes and evidence of loss of heterozygosity, copy number variation, and microsatellite instability. This was the first genetic systematic review in African populations.

A meta-analysis on 27 studies investigating environmental and lifestyle risk factors for ESCC (tobacco, alcohol use, combined tobacco and alcohol use, polycyclic aromatic hydrocarbon exposure, esophageal injury and fruit and vegetable consumption) was carried out. Adverse associations between ESCC risk and all the risk factors were found, whereas fruit and vegetable consumption showed a protective effect. The proportion of ESCC attributable to tobacco (17%), alcohol use (13%), combined tobacco and alcohol use (23%), polycyclic aromatic hydrocarbon exposure (5%), esophageal injury (17%) and fruit and vegetable consumption (-11%) were estimated using population attributable fraction analysis. This study was the most comprehensive systematic review and meta-analysis on African literature.

Genes and pathways with differential mRNA expression were identified using datasets on ESCC, esophageal adenocarcinoma (EAC) and Barrett's esophagus (BE) using the Rank Product Method, and gene set enrichment analysis (SetRank), with the Reactome Annotation Database. A total of 18 publicly available GEO mRNA expression datasets on 906 tissue samples, were analyzed. Overall, 1,107 upregulated genes and 1,537 downregulated genes were outputted for BE, EAC and ESCC. Significantly associated pathways included “Extracellular matrix organisation”, “Collagen chain trimerization”, “TP53 regulates transcription of several additional cell death genes whose specific roles in p53-dependent apoptosis remain uncertain”, and “Cyclin B2 mediated events”. Pathways not previously discussed or interpreted for EC in literature were identified, which warrant further investigation.

These results highlight the multifactorial and complex etiology of EC. Comprehensive large-scale studies on the genetic basis and pathobiology of ESCC are still lacking in Africa. Understanding EC requires an integrated approach incorporating different study designs to assess both environmental and genetic factors of EC.

Opsomming

Esofageale kanker (EK) is ‘n aggressiewe kanker wat ‘n benaderde 572034 nuwe gevalle en 508585 sterftes jaarliks bydra. Omdat geen vroeë waarnemingsprogram bestaan nie, is laat waarneming en hoë mortaliteit die reël. Die voorkomskoers is hoog in Oos-Asië, Suid-Europa en Oos- en Suidelike Afrika. Hierdie eienaardige verspreiding lig spesifieke risikofaktore in sekere gebiede uit.

Suid-Afrika is ‘n knelgebied vir EK; ‘n hoë voorkomskoers is aangemeld in die Oos-Kaap gedurende die vorige vyf dekades. Min navorsingsaandag is gevestig op EK in Afrika; daarom word die epidemiologie, en die genetiese en omgewingsbasis van EK, nie goed begryp nie. Die hoë insidensie van EK en die dodelike geaardheid daarvan regverdig geöormerkte bestudering daarvan om die risikofaktore en patologiese biologie te verstaan en stategieë vir voorkoming en sifting te bewerkstellig.

Die doel van hierdie studie was om die rol van genetiese faktore, omgewingsfaktore en die onderliggende patologiese biologie in die ontwikkel van EK te ondersoek, deur geenuitdrukkingsdata te gebruik. Genetiese variante wat geassosieer is met esofageale plaveiselepiteel karsinoom (EPEK) in Afrika-bevolkings, is in 23 studies bestudeer. Altesaam 25 variante in 20 gene is met statistiese beduidendheid gerapporteer. Daarby het agt studies ook somatiese veranderinge in 17 gene en bewyse van verlies van heterosigositeit, kopie aantal variasie en mikrosatelliet onstabiliteit gewys. Hierdie is die eerste genetiese sistematiese oorsig in Afrika-bevolkings.

‘n Meta-analise van 27 studies wat omgewings- en leefstylrisikofaktore vir EPEK ondersoek het (tabak, alkoholverbruik, tabak- en alkoholverbruik, polisikliese aromatiese hidrokoolstofblootstelling, esofageale besering en groente en vrugte verbruik) is uitgevoer. Ongewenste assosiasies tussen EPEK risiko en al die risikofaktore is gevind, terwyl groente en vrugte verbruik ‘n beskermende effek getoon het. Die breukdeel van EPEK toegeken aan tabak (17%), alkoholverbruik (13%), tabak- en alkoholverbruik (23%), polisikliese aromatiese hidrokoolstof blootstelling (5%), esofageale besering (17%) en groente en vrugte verbruik (-11%) is bepaal deur bevolkingstoekenningsfraksie analise. Hierdie studie is die mees deeglike sistematiese oorsig en meta-analise nóg van Afrika literatuur.

Gene en biologiese netwerke met differensiële bRNS uitdrukking is geïdentifiseer in EPEK, esofageale adenokarsinoom (EAK) en Barret se esofagus (BE) datastelle deur die rangorde produk metode te gebruik, tesame met geenversameling verrykingsanalise (SetRank), met die Reactome Annotation databasis. ‘n Totaal van 18 publieke beskikbare GEO bRNS geenuitdrukkingsdatastelle vir 906 weefselmonsters, is geanaliseer. In totaal is 1107 oorgereguleerde, en 1537 ondergereguleerde gene gevind vir BE, EAC en ESCC. Beduidende geassosieerde biologiese netwerke het “Ekstra-sellulêre matriks organisasie”, “Kollageenketting trimerisasie”, “TP53 reguleer transkripsie van verskeie addisionele seldood gene waarvan die spesifieke rolle in p53-afhanklike apoptose onseker bly”, en “Siklien B2-gemediëerde gebeurtenisse” ingesluit. Biologiese netwerke wat nie voorheen in EK literatuur bespreek of geïnterpreteer is nie, is gevind en verdere ondersoeke daarvan is geregverdig.

Hierdie resultate lig die veelvuldige betrokke faktore en komplekse etiologie van EK, uit. Deeglike grootskaalse studies oor die genetiese basis en patologiese biologie van EPEK is steeds beperk in Afrika. ‘n Begrip van EK benodig ‘n geïntegreerde benadering wat verskillende studie ontwerpe insluit, om beide omgewings- en genetiese faktore te ondersoek.

Acknowledgements

I would like to acknowledge everyone who have been a part of this journey with me. I thank you with all my heart.

To my family, Acub, Siwenyu, Eleazer and Precious, for their unwavering sacrifice, patience and support. I will forever be grateful.

To my supervisors and mentors, Helena, Gerard, Vikash, Christian. I appreciate your dedication to my thesis, all the knowledge you imparted on me, and for allowing me to grow and develop into the scientist I am now. I cannot thank you enough for your support throughout this PhD journey.

To Helena, thank you for being my rock, teacher, mentor throughout this PhD journey. I appreciate you for and sacrifice, care and unwavering support.

To all my friends, thank you for cheering me on, and wishing me well, and being there for me through the small wins, big wins and the tough days. Special thanks to Noxolo, Given, Leonard, and Yohane

To the PhD discussion group, thank you Susan for creating a safe space for reflection and growth as a PhD student.

To the PhD support group, Amokelani, Jimmy and Derrick, thank you all for the support.

Finally, I would like to thank God for giving me the strength and resilience I needed for this journey.

Scholarships, Awards, Conference Presentations, & Research Outputs

A. SCHOLARSHIPS

1. Collaboration for Evidence-Based Healthcare and Public Health in Africa (CEBHA+) PhD Scholarship (2019 – 2020). The scholarship supports public health research in LMICs and is awarded to PhD students working on non-communicable diseases.

2. Margaret McNamara Education Grant (2020). The scholarship is awarded to a select group of women who have demonstrated financial need, academic merit, and are committed to working for the well-being of women and children.

B. GRANTS

1. Harry Crossley Grant (year). Awarded in support of the research activities towards the PhD project and administered by the Faculty of Medicine and Health Sciences,Stellenbosch University.

2. Department of Science and Innovation/National Research Foundation Grant for Esophageal Cancer Research, which in part, aided the conduct of the systematic review of risk factors for EC in Africa.

C. AWARDS

1. L'Oréal-UNESCO for Women in Science Prize for Sub-Saharan African PhD Student (2020) Awarded to "*outstanding women researchers who have contributed to scientific progress*".

D. CONFERENCE PRESENTATIONS AND SEMINARS

International

1. **Simba H**, Kuivaniemi H, Lutje V, Tromp G, and Sewram V (2019) PhD Research Project Progress Presentation. CEBHA+ Network Meeting. Stellenbosch, South Africa. 3 March 2019. (Oral Presentation)

2. **Simba H**, Kuivaniemi H, Lutje V, Tromp G, and Sewram V (2019) Systematic review of genetic factors in the aetiology of squamous cell carcinoma of the oesophagus in African populations. African Organization for Research and Training in Cancer (AORTIC) International Cancer Conference. Maputo, Mozambique. 6-9 November 2019 (Poster Presentation)

National

1. **Simba H**, Kuivaniemi H, and Sewram V. Global Health PhD Seminar, Faculty of Medicine and Health Sciences, Stellenbosch University, May 2019. (Oral Presentation).

2. **Simba H**, Kuivaniemi H, Lutje V, Tromp G, and Sewram V (2019) Systematic review of genetic factors in the aetiology of squamous cell carcinoma of the oesophagus in African populations. Non-Communicable Disease (NCD) Research Symposium, 4 March 2020. Stellenbosch, South Africa. (Poster Presentation)

3. **Simba H**, Kuivaniemi H, Lutje V, Tromp G, and Sewram V. 2019 Systematic review of genetic factors in the aetiology of squamous cell carcinoma of the oesophagus in African populations. Annual Academic Day, Faculty of Medicine and Health Sciences, Stellenbosch University, 21 August 2019. (Oral Presentation).

4. **Simba H**, Kuivaniemi H, Sewram V, and Tromp G (2020) Identification of genes and pathways with altered mRNA expression in oesophageal cancer. South African Society for Bioinformatics (SASBi) Online Student Symposium, 4-6 August 2020. (Oral Presentation)

5. **Simba H**, Kuivaniemi H, Sewram V, and Tromp G (2020) Identification of genes and pathways with altered mRNA expression in oesophageal cancer. Annual Academic Day, Faculty of Medicine and Health Sciences, Stellenbosch University, 27 August 2020. (Poster Presentation)

6. **Simba H**, Kuivaniemi H, Sewram V, and Tromp G (2021) Identification of genes and pathways with differential mRNA expression in oesophageal cancer. Annual Academic Day, Faculty of Medicine and Health Sciences, Stellenbosch University, 10-19 August 2021. (Poster Presentation)

E. PUBLICATIONS EMANATING FROM PHD FINDINGS

Peer-reviewed published articles

1. **Simba, H.,** H. Kuivaniemi, V. Lutje, et al. 2019. Systematic Review of Genetic Factors in the Etiology of Esophageal Squamous Cell Carcinoma in African Populations. Frontiers in Genetics. 10: 642; 1-20 https://doi.org/10.3389/fgene.2019.00642 (Impact Factor 1.469)

Table of contents

List of abbreviations

1-OHPG	1-hydroxy pyrene glucuronide
3′ UTR	3 prime untranslated region
95% CI	95% Confidence interval
ADH1B	Alcohol dehydrogenase 1B (class I), beta polypeptide
AGR2	Anterior gradient 2, protein disulphide isomerase family member
ALDH	Aldehyde dehydrogenase
ALDH2	Aldehyde dehydrogenase 2 family member
ANPEP	Alanyl aminopeptidase, membrane
ASIR	Age-standardised incidence rate
ATAC	Assay for Transposase-Accessible Chromatin
BE	Barrett’s Esophagus
BMI	Body mass index
BP	Benzo[a]pyrene
C20orf54/SLC52A3	Solute carrier family 52 member 3
CASP8	Caspase 8
CCND1	Cyclin D1
CDKN2A	Cyclin-dependent kinase inhibitor 2A
cfDNA	Cell-free DNA
ChIP	Chromatin immunoprecipitation
CNV	Copy number variation
COL1A2	Collagen type I alpha 2 chain
COL11A1	Collagen type XI alpha 1 chain
COX-2	Cyclooxygenase-2
CYP1A1	Cytochrome P450 family 1 subfamily A member 1
CYP2E1	Cytochrome P450 family 2 subfamily E member 1
DEGs	Differentially expressed genes

DSC3	Desmocollin 3
DSG3	Desmoglein 3
DNA	Deoxyribonucleic acid
ECM	Extracellular matrix
EGF	Epidermal growth Factor
ELISA	Enzyme-linked immunosorbent assay
ERCC2/XPD	ERCC excision repair 2, TFIIH core complex helicase subunit
EAC	Esophageal adenocarcinoma
EMP1	Epithelial membrane protein 1
ESCC	Esophageal squamous cell carcinoma
ESD	Esophageal squamous dysplasia
EU	European Union
Fas	Fas cell surface death receptor
FC	Fold change
FFPE	Formalin fixed paraffin embedded
GATA6	GATA binding protein 6
GCMS/MS	Gas chromatography-tandem mass spectrometry
GCMS	Gas chromatography-mass spectrometry
GEO	Gene expression omnibus
GERD	Gastroesophageal reflux disease
GLOBOCAN	Global cancer statistics
GMST1	Glutathione S-transferase mu 1
GOSH	Graphic display of heterogeneity
GSEA	Gene Set Enrichment Analysis
GSTP1	Glutathione S-transferase pi 1
GSTT1	Glutathione S-transferase theta 1
GWAS	Genome-wide association study
HIV	Human immunodeficiency virus

hOGG1	8-oxoguanine DNA glycosylase
HOTAIR	HOX transcript antisense RNA
HPLC	High performance liquid chromatography
HPV	Human papilloma virus
HREC	Health Research Ethics Committee
IARC	International Agency for Research on Cancer
ICP-MS	Inductively coupled plasma-mass spectrometry
IL-18	Interleukin 18
JBI-MAStARI	Joanna briggs institute meta-analysis of statistics assessment and review instrument
LCMS/MS	Liquid chromatography-tandem mass spectrometry
LD	Linkage disequilibrium
LMIC	Low-to-Middle Income Country
lncRNA	Long non-coding RNA
MAF	Minor allele frequency
MDM2	Murine double minute 2
MIE	Minimally invasive esophagectomy
MMP1	Matrix metallopeptidase 1
MMP2	Matrix metallopeptidase 2
MNSOD	Manganese superoxide dismutase
mRNA	Messenger RNA
miRNA	Micro RNA
MSI	Microsatellite instability
MUC13	Mucin13, cell surface associated
MTHFR	Methylenetetrahydrofolate reductase
NAT2	N-acetyltransferase 2
NCBI	National Center for Biotechnology Information
NCI	National Cancer Institute

ncRNA	Non-coding RNA
NAT	Normal adjacent tissue
NGS	Next generation sequencing
NNK	4(methylnitrosamino)-1,3-pyridyl-1-butanone
NNN	N’-nitrosonornicotine
NSAIDS	Nonsteroidal anti-inflammatory drugs
NQO1 NAD(P)H	Quinone dehydrogenase 1
OR	Odds ratio
PAF	Population attributable fraction
PAH	Polycyclic aromatic hydrocarbon
PAR	Population attributable risk
pfp	Percentage of false prediction
PLCE1	Phospholipase C epsilon 1
PRS	Polygenic risk score
PPP1R3C	Protein phosphatase 1 regulatory subunit 3C
PRISMA	Preferred reporting items for systematic reviews and meta-analyses guidelines
PROSPERO	International prospective register of systematic reviews
QC	Quality control
RAMIE	Robotic-assisted minimally invasive esophagectomy
RB	Retinoblastoma-associated protein
RNA	Ribonucleic acid
RR	Relative risk
rRNA	Ribosomal RNA
SCEL	Sciellin
SEER	Surveillance Epidemiology and End Results
SLC52A3	Solute carrier family 52 member 3
SNP	Single nucleotide polymorphism

STK15/AURKA	Serine/threonine kinase 15
STREGA	STrengthening the REporting of Genetic Association studies
YGCA	The cancer genome atlas
TMB	Tumour mutational burden
TNF-α	Tumor necrosis factor alpha
TNM	Tumour, Node, Metastasis
TP53	Tumor protein p53
tRNA	Transfer RNA
WES	Whole exome sequencing
WGS	Whole genome sequencing
XRCC1	X-ray repair cross complementing 1

Note - US spelling throughout for the following words: esophagus, esophageal, etiology. This is to avoid differences of abbreviations between published works and the dissertation (thesis).

List of Figures

Chapter 1

Chapter 2

Chapter 3 Figures

Chapter 4 Figures

List of Tables

Chapter 1

Chapter 2

Chapter 3

Chapter 4

Outline of thesis

The thesis focussed on genetic and environmental risk factors associated with esophageal squamous cell carcinoma (ESCC) in African populations as well as identifying biological pathways enriched with differentially expressed genes in esophageal cancer (EC). It is divided into six parts.

Chapter 1 introduces the problem of ESCC and gives a global view of the research activities being undertaken worldwide. The epidemiology of ESCC is described. Risk factors associated with ESCC reported in the literature are also described and these include environmental, lifestyle and genetic risk factors. Tools and techniques used in the assessment of risk factors in cancer research are also described. In addition, the study rationale, overall aims and objectives of the study are outlined in this chapter.

Chapter 2 describes the genetic factors associated with ESCC risk in African populations. This review systematically screened and critically appraised relevant studies which reported germline and somatic variants. Genetic variants reported to be associated with ESCC are summarised and linkage disequilibrium for SNPs reported in the same gene is described. The need for more comprehensive large-scale genetic studies in Africa is emphasised. This systematic review has been published in the journal, *Frontiers in Genetics*.

Chapter 3 is a systematic review on the lifestyle and environmental risk factors associated with ESCC development in African population. Studies which met the selection criteria were critically appraised and numerous known and emerging risk factors were identified and described. These included tobacco smoking and alcohol consumption, socioeconomic status, diet, PAH exposure, consumption of hot food and beverages, oral health, infectious agents, esophageal injury, family history of cancer and non-acid gastro-esophageal reflux. Meta-analyses were carried out and population attributable fractions were calculated, where adequate information was provided.

Chapter 4 focusses on the identification of genes with differential mRNA expression in EC using meta-analysis of differentially expressed genes (DEGs) and Gene Set Enrichment Analysis of GEO (Gene Expression Omnibus) datasets. A total of 18 publicly available GEO mRNA expression datasets, with expression data on 906 individual tissue samples, were included in the analysis. Of the 18 datasets, three used esophageal adenocarcinoma (EAC) tissue, eleven used ESCC tissue and nine used Barrett's esophagus (BE) tissue. One dataset included EAC, ESCC and BE tissue, whilst one dataset included both EAC and squamous dysplasia tissue samples. This analysis provides novel insights into the mechanisms linked to EC development, and the differences between the different types of EC as well as the precancerous lesion, BE.

Chapter 5 is a general discussion integrating the study findings and highlighting the key findings of this thesis as well as limitations. The chapter provides a unified hypothesis for the PhD study.

Chapter 6 contains the conclusions of the PhD study. Future directions of research are also discussed.

Chapter 1: Introduction

1.1 Esophageal Cancer

Esophageal cancer (EC) is a lethal malignant tumour of the esophagus ranking as the 6th most common cause of cancer mortality and the 7th most common cancer worldwide.(1) According to the Global Cancer Statistics of 2018, a total of 572,034 new cases and 508,585 deaths were reported, indicative of the high fatality associated with EC diagnosis.(1) This translates to approximately 1 in every 20 cancer deaths being attributable to EC.(1) The 5-year survival of EC ranges between 13–18% and this is due to the lack of primary and secondary prevention methods.(2)

Malignant esophageal tumours are characterized by two major subtypes; esophageal squamous cell carcinoma (ESCC), which is the more common type and contributes 90%, and esophageal adenocarcinoma (EAC).(3, 4)).(3, 4) ESCC occurs proximal to the squamocolumnar junction and develops as a result of inflammation or carcinogenic and mutagenic factors which lead to dysplasia *in situ* and eventual malignant tumour development.(5) EAC commonly occurs in the distal and mid-esophagus, and develops from specialized intestinal epithelium (Barrett's esophagus, BE) as a result of chronic exposure to gastric acid, bile, pancreatic juice, and pepsin.(5) The two subtypes have different etiologies, pathophysiology, prognosis, treatment and different geographically defined high incidence regions. Over 80% of EC cases and deaths are reported in developing countries.(6) In Western populations; North America, Australia and Europe, EAC is more common. The squamous cell carcinoma subtype is most common in the developing countries of Africa, Asia and South America.(7) This thesis will mainly focus on the epidemiology and risk factors of ESCC. Figure 1 shows the estimated age standardised incidence rates for EC in 2018 according to GLOBOCAN.(1)

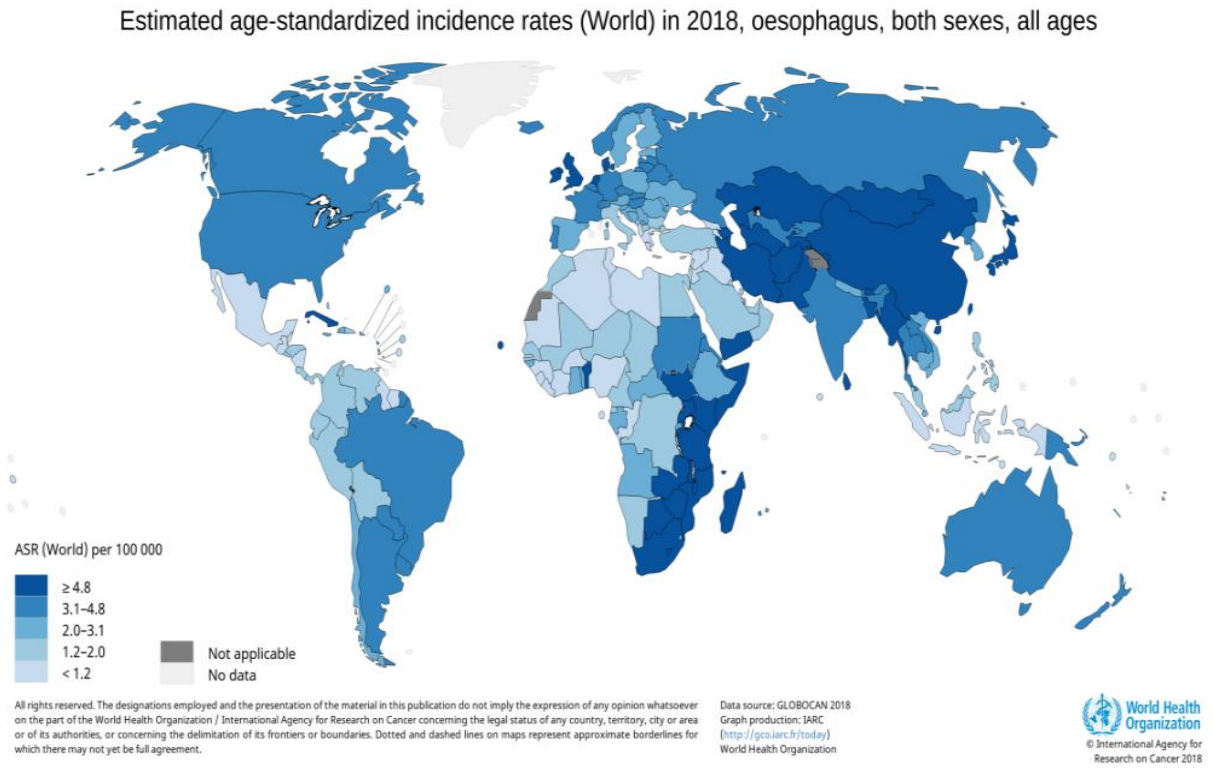

Estimated age-standardized incidence rates (World) in 2018, oesophagus, both sexes, all ages
ASR (World) per 100 000
≥ 4.8
3.1–4.8
2.0–3.1
1.2–2.0
< 1.2
Not applicable
No data
All rights reserved. The designations employed and the presentation of the material in this publication do not imply the expression of any opinion whatsoever on the part of the World Health Organization / International Agency for Research on Cancer concerning the legal status of any country, territory, city or area or of its authorities, or concerning the delimitation of its frontiers or boundaries. Dotted and dashed lines on maps represent approximate borderlines for which there may not yet be full agreement.
Data source: GLOBOCAN 2018
Graph production: IARC
(http://gco.iarc.fr/today)
World Health Organization
World Health Organization
© International Agency for Research on Cancer 2018

Figure 1.1: *Estimated age standardised incidence rates for EC in 2018 according to GLOBOCAN.* Reproduced from *https://gco.iarc.fr/today*

1.2 Epidemiology of ESCC

1.2.1 Worldwide

There is a striking variation in the ESCC incidence worldwide. Variability in the incidence between high risk and low risk areas worldwide has been reported to be up to 10-fold.(8) The highest incidence rates are recorded on two major geographical belts; north central China through the central Asian republics to northern Iran, and one from eastern to southern Africa.(1) The highest incidence rates ever recorded worldwide were reported in the 1970s in Iran with incidence rates of 165 per 100,000/year in men and 195 per 100,000/year in women.(9) Although incidence rates have declined in most regions worldwide, ESCC remains a common and fatal malignancy in these hotspot regions. According to country, Malawi has the highest incidence rates globally for both men and women with Age Standardised Incidence Rates (ASIR) of 18.7 per 100,000, followed by Mongolia and Kenya with ASIR of 18.5 and 18.4 per 100,000, respectively. According to region, Eastern Asia ranks first with

ASIR of 12.2 per 100,000.(1) Table 1 shows the top 20 countries worldwide with the highest ASIR per 100,000 and the corresponding incident cases according to GLOBOCAN 2018.(1)

Table 1.1: Top 20 countries with the highest ASIR per 100 000 worldwide in 2018 according to GLOBOCAN

Country	Incident cases	ASIR
Malawi	1,844	18.7
Mongolia	397	18.5
Kenya	4,380	18.4
Bangladesh	20,906	14.8
China	307,359	13.9
Zimbabwe	920	12.4
Tajikistan	619	11.1
Uganda	1,749	10.8
Cabo Verde	39	10.4
Burundi	520	10.2
Turkmenistan	408	9.2
Tanzania, United Republic of	2,516	8.9
Afghanistan	1,312	8.2
Kazakhstan	1,615	8.1
Madagascar	1,085	7.9
Comoros	35	7.9
South Africa	3,697	7.8
South Sudan	523	7.6
Somalia	524	7.5
Botswana	105	6.9

There is a lot of uncertainty regarding the accuracy of national incidence rates, particularly for LMICs, particularly due to lack of human capacity for registration, infrastructure and government commitment. Population based cancer incidence registries in the LMICs therefore provide with more accurate incidence rates as they provide with region specific statistics for current ESCC hotspots. However there remains uncertainties due to the paucity of comprehensive information nationally and the scarcity of resources for histological diagnosis. The Cancer Incidence in Five

Continents (CI5) series, provides high quality data from cancer registries worldwide, and is published by the International Agency for Research on Cancer (IARC) and the International Association of Cancer Registries (IACR).(10) The current volume contains data collected from 2008 to 2012. The top 20 regions worldwide with the highest incidence rates for males and females are shown in Table 1.2 and 1.3 respectively

Table 1.2: EC incidence rates for males from the Cancer in 5 continents XI publication

Population	Cases	Crude Rate	ASR (W)
China, Cixian County	1847	114.7 (2.67)	162.8 (3.99)
China, Yanting County	1670	106.3 (2.60)	101.3 (2.60)
China, Linzhou County	2575	95.3 (1.88)	91.4 (1.89)
China, Shexian County	865	83.3 (2.83)	77.6 (2.72)
China, Jianhu County	994	81.1 (2.57)	59.5 (1.93)
China, Huaiyin District, Huai'an	1565	83.2 (2.10)	58.8 (1.51)
China, Yanshi	447	48.1 (2.27)	45.7 (2.21)
China, Tongling City	573	50.3 (2.10)	45.7 (2.03)
China, Sheyang	1393	56.1 (1.50)	35.2 (0.96)
China, Xianju	379	48.9 (2.51)	34.2 (1.82)
China, Hefei	1276	37.6 (1.05)	32.8 (0.94)
China, Xiping	483	36.2 (1.65)	32.0 (1.50)
China, Guanyun	1019	37.5 (1.17)	31.5 (1.02)
India, Mizoram	546	19.9 (0.85)	27.6 (1.21)
China, Maanshan	539	33.3 (1.44)	27.2 (1.21)
India, Kamrup Urban District	508	20.9 (0.93)	24.9 (1.15)
South Africa, Eastern Cape	374	15.1 (0.78)	24.0 (1.27)
Iran, Golestan	508	14.6 (0.65)	22.3 (1.04)
Uganda, Kyadondo	281	5.3 (0.32)	21.1 (1.39)
China, Lianyungang	647	27.3 (1.07)	20.3 (0.81)

Table 1.3: EC incidence rates for females from the Cancer in 5 continents XI publication

Population	Cases	Crude Rate	ASR (W)
China, Cixian County	1320	84.0 (2.31)	101.9 (2.89)
China, Yanting County	1242	85.6 (2.43)	69.4 (1.99)
China, Linzhou County	2021	78.8 (1.75)	61.9 (1.42)
China, Shexian County	425	43.9 (2.13)	38.1 (1.90)
China, Jianhu County	667	56.0 (2.17)	38.0 (1.51)
China, Huaiyin District, Huai'an	820	46.8 (1.63)	29.0 (1.07)
China, Yanshi	325	35.8 (1.99)	26.2 (1.51)
China, Xianju	222	30.9 (2.07)	21.7 (1.54)
China, Sheyang	831	35.2 (1.22)	20.1 (0.71)
India, Kamrup Urban District	283	12.5 (0.74)	16.6 (1.02)
Iran, Golestan	357	10.4 (0.55)	16.2 (0.88)
China, Guanyun	479	19.3 (0.88)	15.9 (0.75)
China, Xiping	265	21.0 (1.29)	15.4 (1.00)
Kenya, Nairobi	261	3.4 (0.21)	15.2 (1.07)
South Africa, Eastern Cape	410	14.2 (0.70)	14.6 (0.75)
Zimbabwe, Harare: African	110	4.9 (0.47)	12.5 (1.28)
Turkey, Erzurum	136	11.7 (1.01)	11.9 (1.05)
Uganda, Kyadondo	177	3.0 (0.23)	11.7 (0.96)
China, Tongling City	131	12.0 (1.05)	10.3 (0.92)
China, Hefei	437	13.8 (0.66)	10.2 (0.51)

Only seven registries have been submitted from Africa to the CI5 XI volume, from 6 countries (South Africa, Eastern Cape; Uganda, Kyadondo; Zimbabwe, Harare: African; Kenya, Nairobi; Seychelles; Algeria, Setif; Algeria, Batna), and this covers only 1% of Africa's population.(10) The paucity of population-based cancer registries in Africa means precise figures on incidence rates are lacking. The EC incidence rates from the seven registries are shown in Tables 1.4 and 1.5

Table 1.4: EC incidence rates for males from the Cancer in 5 continents XI publication (Africa)

Population	Cases	Crude Rate	ASR (W)
South Africa, Eastern Cape	374	15.1 (0.78)	24.0 (1.27)
Uganda, Kyadondo	281	5.3 (0.32)	21.1 (1.39)
Zimbabwe, Harare: African	128	6.0 (0.53)	15.5 (1.55)
Kenya, Nairobi	322	4.0 (0.22)	14.7 (0.97)
Seychelles	12	6.8 (1.95)	7.5 (2.20)
Algeria, Setif	17	0.5 (0.13)	0.6 (0.16)
Algeria, Batna	10	0.3 (0.11)	0.5 (0.15)

Table 1.5: EC incidence rates for females from the Cancer in 5 continents XI publication (Africa)

Population	Cases	Crude Rate	ASR (W)
Kenya, Nairobi	261	3.4 (0.21)	15.2 (1.07)
South Africa, Eastern Cape	410	14.2 (0.70)	14.6 (0.75)
Zimbabwe, Harare: African	110	4.9 (0.47)	12.5 (1.28)
Uganda, Kyadondo	177	3.0 (0.23)	11.7 (0.96)
Seychelles	3	1.7 (0.99)	1.9 (1.13)
Algeria, Setif	17	0.6 (0.14)	0.7 (0.18)
Algeria, Batna	12	0.4 (0.12)	0.5 (0.16)

1.2.2 Africa

The peculiar geographical distribution of ESCC is also present in Africa, with high incidence areas lying on what is known as the African ESCC corridor or the East African Rift Valley. The African ESCC corridor runs from Kenya down to South Africa on the easterly side of Africa. This African corridor includes Malawi, Kenya, Zimbabwe, Uganda, Burundi, Tanzania, Madagascar, Comoros and South Africa with ASIR per 100,000 ranging from 18.7 to 7.8.(1, 11, 12) A comparison of data from the C15 XI with GLOBOCAN data shows higher ASIR for South Africa (14.6 vs 7.8), Zimbabwe (12.5 vs 12.4), Uganda (11.7 vs 10.8) and lower ASIR for Kenya (15.2 vs 18.4) High incidence rates from this corridor have been reported as far back as 1969.(13) ESCC cases from the African ESCC corridor are also reported to be younger than those found elsewhere in the world.(14, 15) This presents with possible unique risk factors for this region.(16) The average age of diagnosis for ESCC worldwide is 60 years whilst in sub-Saharan Africa, Malawi has a median age at diagnosis of 47.5 years and 51 years in Uganda.(15)

The young age at presentation for EC cannot be fully explained by the young age structure in Africa. There are likely other contributing factors including genetic, environmental, as well as interplay of both. Investigation is needed to elucidate the driving factors of young age at presentation for esophageal cancer and the extent to which the young age structure contributes to this incidence.

1.2.3 South Africa

South Africa currently has the 17th highest incidence of ESCC in the world and 10th highest incidence of ESCC in Africa with ASIR of 8.7 per 100,000.(1) Men have ASIR of 11.4 per 100,000, whilst women have ASIR of 5.4 per 100,000. In South Africa, ESCC is the 8th most common cancer for men with an estimated lifetime risk of 1:178 and the 9th most common cancer for women with an estimated lifetime risk of 1:326.(17) Incidence rates are disproportionately higher in the Eastern Cape Province, where it was first identified as a health issue over five decades ago, among the isiXhosa speaking people.(18) This region has continued to be a major hotspot for the disease to this day, contributing the bulk of the cases nationwide. ESCC in the Eastern Cape Province is the most common cancer for men with ASIR of 23.2 per 100,000 and contributing 30% of all the cancers.(19) It is the 2nd most common cancer for women with ASIR of 14.5 per 100,000 and contributing 20% of all the cancers.(19)

1.3 ESCC diagnosis and prognosis

1.3.1 Pathophysiology

ESCC normally starts in the cells of the mucosa (inner layer) and grows to the sub-mucosa and the muscular layers of the esophagus (Figure 2).(20) It commonly occurs in the upper two thirds of the esophagus. Squamous dysplasia of the esophagus consists of precursor lesions known to develop into malignant ESCC tumours. Development of ESCC, therefore, starts from squamous hyperplasia to squamous dysplasia, and then progresses from low-grade intraepithelial neoplasia to high grade, and subsequently carcinoma *in situ* and finally invasive cancer.(21) Dysregulation in cell cycle regulators including TP53, retinoblastoma-associated protein (RB) and cyclin-dependent kinase inhibitor 2A (CDKN2A) is evident in cancerous and precursor lesions.(21) Overall, the pathophysiology is complex and not fully understood. Several genes are likely to be dysregulated and contribute to the disease development and progression (see 6.4.1).

1.3.2 Diagnosis

The gold standard for EC diagnosis is endoscopy and biopsy with pathologic reading of the prepared tissues of the esophagus, which can detect pre-malignant and malignant changes.(21) Chromoendoscopy with Lugol's iodine is important in the diagnosis of preneoplasia of ESCC called esophageal squamous dysplasia (ESD). Use of esophagogastroduodenoscopy (EGD) with Luogl's iodine staining with biopsy leads to >95% sensitivity and specificity for ESD and is established as the gold standard for screening. Non-invasive experimental methods to screening of pre-cancerous include the EsophaCap and cytosponge -TFF3 which can detect genetic and epigenetic alterations on samples collected.(5) Grading is done using the Tumour, Node, Metastasis (TNM) system and is based on the depth of tumour invasion in the mucosa, submucosa, muscle, connective tissue, lymph nodes and into nearby structures.(21) There are five stages of diagnosis ranging from stage 0 to stage IV, with stage 0 being high grade dysplasia and the remaining four stages divided into two sub-stages each. Table 2 shows the staging for ESCC and EAC according to the American Joint Committee on Cancer (AJCC) of 2018.(22)

Table 1.6: Staging of EC according to the American Joint Committee on Cancer (AJCC) TNM system

ESCC		EAC	
Stage	**Description**	**Stage**	**Description**
0	Cancer cells only present in the epithelial cell of the esophagus and has not spread to any lymph nodes or distant organs. High-grade dysplasia. Cancer grade - does not apply.	0	Cancer cells only present in the epithelial cell of the esophagus and has not spread to any lymph nodes or distant organs. High-grade dysplasia. Cancer grade - does not apply.
IA	Cancer cells have spread and growing into the tissue under the epithelium (lamina propria or muscularis mucosa). Cancer grade - grade 1 or an unknown grade	IA	Cancer cells have spread and growing into the tissue under the epithelium (lamina propria or muscularis mucosa). Cancer grade - grade 1 or an unknown grade
IB	Cancer cells are growing into the lamina propria, muscularis mucosa, submucosa or the thick muscle layer (muscularis propria). Cancer grade - any grade or an unknown grade.	IB	Cancer cells are growing into the lamina propria, muscularis mucosa, or the submucosa. Cancer grade - grade 1 or 2 or an unknown grade.
	Not applicable	IC	Cancer cells are growing into the lamina propria, muscularis mucosa, submucosa or the thick muscle layer. Cancer grade -grade 1, 2 or 3.
IIA	Cancer cells are growing into the muscularis propria. Cancer grade - grade 2 or 3, or unknown grade. OR	IIA	The cancer is growing into the muscularis propria. Cancer grade - grade 3 or an unknown grade.

	The cancer is growing into the outer layer of the esophagus (the adventitia). Cancer grade - any grade and located in the lower esophagus, or grade 1 and located in the upper or middle esophagus.		
IIB	Cancer cells are growing into the outer layer of the esophagus (the adventitia). Cancer grade - grade 2 or 3 and located in the upper or middle of the esophagus, or unknown grade and located anywhere in the esophagus or any grade and have an unknown location in the esophagus. OR Cancer cells are growing into the lamina propria, muscularis mucosa or into the submucosa, and spread to 1 or 2 nearby lymph nodes. Cancer grade - any grade.	IIB	Cancer cells are growing into the lamina propria, muscularis mucosa, or the submucosa. It has spread to 1 or 2 nearby lymph nodes. Cancer grade - any grade. OR The cancer is growing into the outer layer of the esophagus (the adventitia). It has not spread nearby lymph nodes. Cancer grade - any grade.
IIIA	Cancer cells are growing into the lamina propria, muscularis mucosa, submucosa or the thick muscle layer. The cancer has spread to no more than 6 nearby lymph nodes. Cancer grade - any grade and located anywhere in the esophagus.	IIIA	The cancer is growing into the lamina propria, muscularis mucosa, the submucosa, or the thick muscle layer. It has spread to no more than 6 nearby lymph nodes. It has not spread to distant organs.

			Cancer grade - any grade.
IIIB	Cancer cells are growing into: i)The thick muscle layer and spread to no more than 6 nearby lymph nodes OR ii)The outer layer of the esophagus and spread to no more than 6 nearby lymph nodes OR iii)The pleura (a thin layer of tissue covering the lungs), the pericardium (a thin sac covering the heart), or the diaphragm (the muscle below the lungs) and spread to no more than 2 nearby lymph nodes. Cancer grade - any grade and located anywhere in the esophagus.	IIIB	The cancer is growing into: i)the thick muscle layer (muscularis propria) and spread to no more than 6 nearby lymph nodes OR ii)the outer layer of the esophagus and spread to no more than 6 nearby lymph nodes OR iii)the pleura (the thin layer of tissue covering the lungs), the pericardium (the thin sac surrounding the heart), or the diaphragm (the muscle below the lungs that separates the chest from the abdomen) and spread to no more than 2 nearby lymph nodes. It has not spread to distant organs. Cancer grade - any grade.
IVA	Cancer cells are growing into	IVA	The cancer is growing into: i)the pleura (the thin layer of tissue covering the lungs), the pericardium (the thin sac surrounding the heart),

	i)the pleura, the pericardium, or the diaphragm and spread to no more than 6 nearby lymph nodes OR ii) the trachea, the aorta, the spine, or other crucial structures and no more than 6 nearby lymph nodes OR iii) any layers of the esophagus and spread to 7 or more nearby lymph nodes. Cancer grade - any grade.		or the diaphragm (the muscle below the lungs that separates the chest from the abdomen) and spread to no more than 6 nearby lymph nodes OR ii)the trachea (windpipe), the aorta (the large blood vessel coming from the heart), the spine, or other crucial structures and no more than 6 nearby lymph nodes OR iii)any layers of the esophagus and spread to 7 or more nearby lymph nodes. The cancer has not spread to distant organs. Cancer grade - any grade.
IVB	Cancer cells have spread to distant lymph nodes and/or other organs, such as the liver and lungs. Cancer grade - any grade and located anywhere in the esophagus.	IVB	The cancer has spread to distant lymph nodes and/or other organs such as the liver and lungs. Cancer grade - any grade.

Cancer staging is important in finding out if the tumour has spread and serves as a guide for treatment options. Figure 2 shows Stage IIIB cancer that has spread into the thick muscle layer or the connective tissue layer of the esophagus wall and into the lymph nodes. The role of screening at-risk populations for surveillance remains a contentious topic between scientists, as cost benefits and survival benefits have not been proven as yet. There is, therefore, a need for development and validation of risk models which can identify persons and populations with a higher risk of ESCC development and eligible for screening. There have been successful ESCC screening programs developed in China. One of the main programs was the endoscopic screening and intervention program of 3,319 participants and 797 controls to assess reduction in mortality.(23) Screening for ESCC was done using endoscopy with Lugol's iodine staining. The authors reported a reduction in ESCC incidence and mortality in the intervention group compared to the control group.(23) In a recent study, short term ESCC and ESD screening efficacy was assessed on 52,729 intervention participants and 43,068 controls using endoscopy.(24) Follow up on this study is still ongoing.

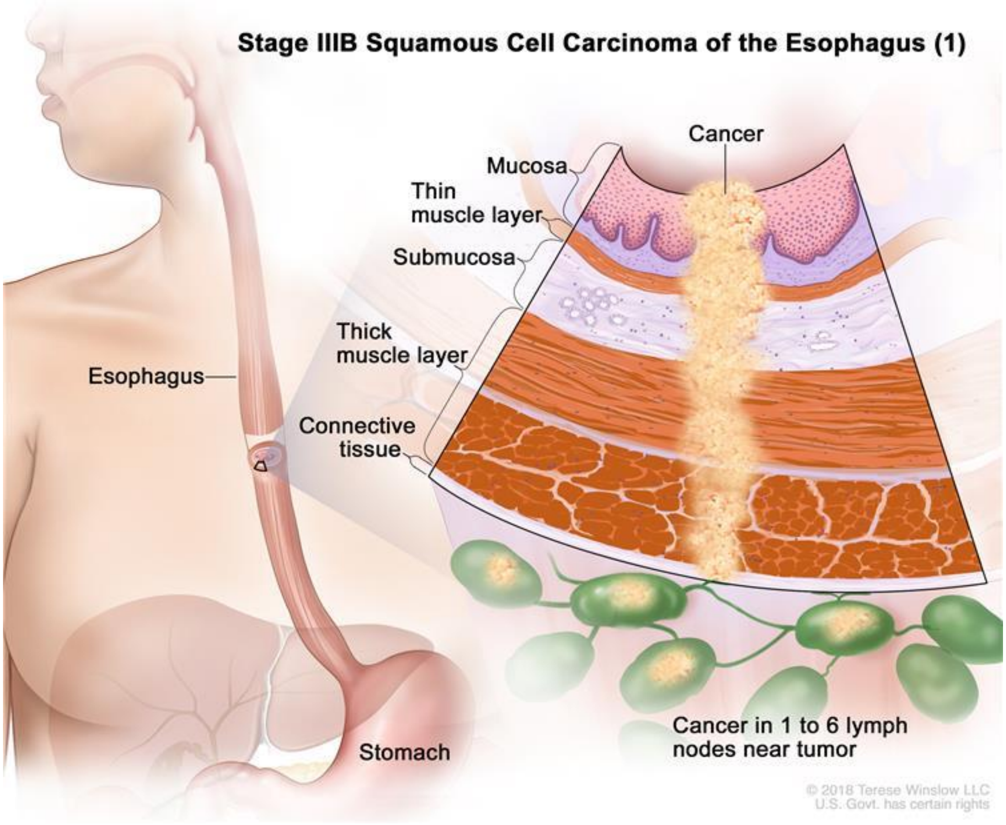

Stage IIIB Squamous Cell Carcinoma of the Esophagus (1)
Cancer
Mucosa
Thin muscle layer
Submucosa
Thick muscle layer
Esophagus
Connective tissue
Stomach
Cancer in 1 to 6 lymph nodes near tumor
© 2018 Terese Winslow LLC
U.S. Govt. has certain rights

Figure 1.2: *Stage IIIB ESCC that has spread into the thick muscle layer or the connective tissue layer of the esophagus wall and into the lymph nodes. [Reproduced with permission https://surgery.ucsf.edu/conditions--procedures/esophageal-cancer.aspx]*

1.3.3 Symptoms and treatment

ESCC has a poor prognosis as patients usually present at advanced stage and at metastatic disease. At this stage, treatment is limited to palliative chemo-radiation and stenting.(21) Common symptoms include dysphagia and unexplained weight loss.(21) If detected early, options for treatment include endoscopic mucosal resection or radiofrequency ablation for low-grade, high-grade and carcinoma *in situ*.(21) For invasive ESCC, treatment options include esophagectomy with lymphadenectomy or definitive chemoradiotherapy.(21, 25) When esophageal resection is possible, with no neoadjuvant therapy, a 5-year survival rate of 12–27% is reported, whilst surgery with neoadjuvant therapy increases the 5-year survival rate up to 57%.(21) Minimally invasive methods for surgical resection, namely, robotic-assisted minimally invasive esophagectomy (RAMIE) and minimally invasive esophagectomy (MIE) are reported to have better short-term surgical outcomes and quality of life compared to open esophagectomy.(25) There is an expected influx of new robotic platforms for surgical resection which will incorporate the use of artificial intelligence and impact robotic surgery to be less invasive and have better outcomes.(25) Immunotherapy for EC treatment is still being explored with potential targets for tyrosine-kinase receptors and epidermal growth factor receptors.(21) Despite advances in management and treatment, ESCC prognosis is still poor with a survival rate of <5% in low income countries, and 21% in China(26). The variation of survival rates is indicative of the potential benefits of primary and secondary prevention in reducing mortality.

1.4 Environmental and lifestyle risk factors of ESCC

The variability in incidence rates of ESCC according to geographical location points to multifactorial and population dependent risk factors. This is evident with the most common risk factors for ESCC, tobacco smoking and alcohol consumption. In a study done in the USA(27), cigarette smoking, alcohol consumption, and low consumption of fruits and vegetables were found to have a population attributable risk (PAR) of 89%, whilst in a large study done in China(28) both tobacco smoking and alcohol consumption were reported to not increase risk of ESCC. It is still unclear if exposure rates play a role in these differences in effects, however this highlights the importance of population specific estimates for ESCC risk factors.

A number of environmental and lifestyle risk factors have been reported to be associated with ESCC development..(9) Figure 3 summarizes the risk factors of ESCC worldwide.

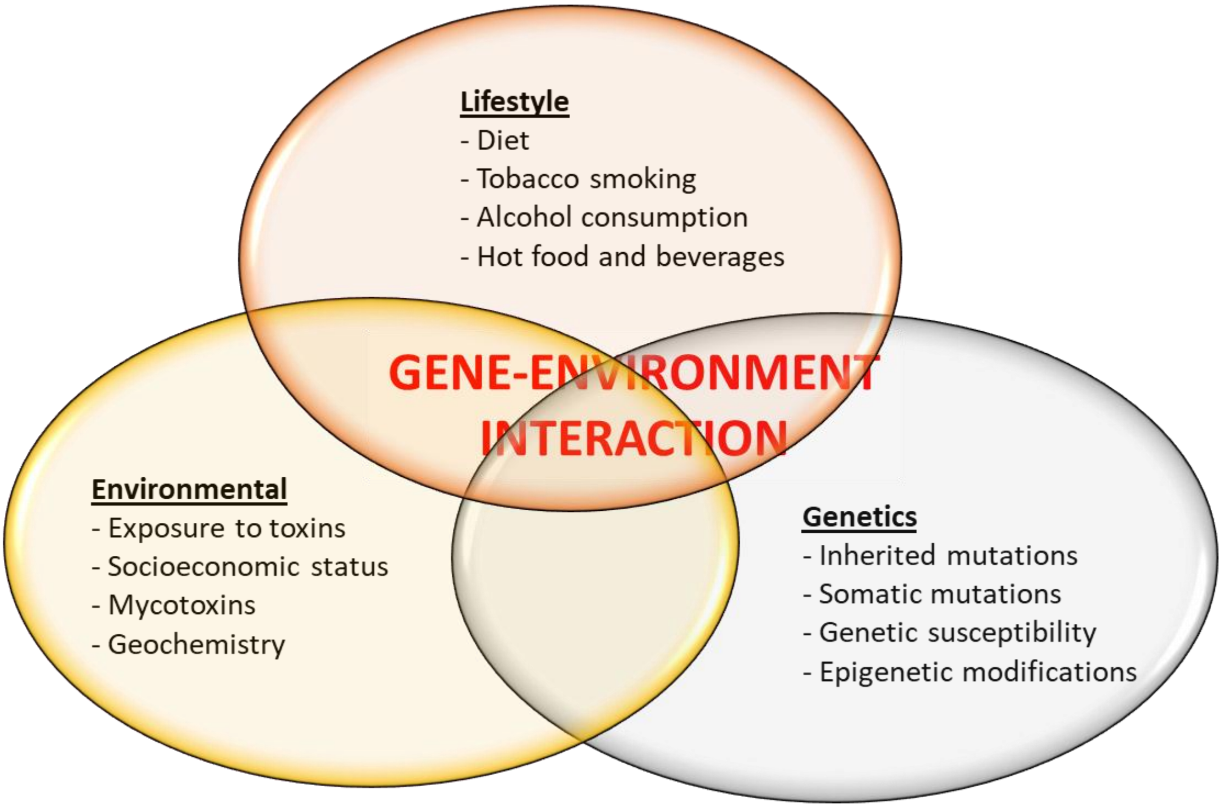


Figure 1.3: *Summary of ESCC risk factors worldwide*

Reports from 2005 projected that the number of ESCC cases will dramatically increase by 140% by the year 2025.(29, 30) Therefore, research agenda needs to focus on determining risk factors associated with the development and progression of ESCC and strategies for its prevention.

1.4.1 Tobacco smoking and chewing

Tobacco is classified as a class 1 carcinogen by the International Agency for Research on Cancer (IARC).(15) Smoking commercial cigarettes causes exposure to a number of chemicals including polycyclic aromatic hydrocarbons (PAH), acetaldehyde, and nitrosamines such as N’-nitrosonornicotine (NNN), and 4(methylnitrosamino)-1,3-pyridyl-1-butanone (NNK), reactive oxygen species, and nitric oxide which have been reported to have carcinogenic, mutagenic and toxic componets.(15) Tobacco smoking also inhibits the aldehyde dehydrogenase (ALDH) enzyme which metabolizes acetaldehyde, resulting in higher levels of acetealdehyde.(15) It has been shown that the risk of ESCC increases with high tobacco exposure intensity and increased duration of tobacco exposure.(31, 32) Second-hand smoke has also been shown to increase the risk of ESCC in a study done on a high-risk population of India.(33) Smoking tobacco after an ESCC diagnosis is reported to reduce survival times.(34) Chewing and smoking tobacco has been reported to increase the risk of ESCC in African, Asian and Western populations.(15, 35, 36) Other forms of tobacco use include pipes, hookah, cigars and snuff. In a recent systematic review on the risk factors of ESCC in Africa by Asombang et al(35) tobacco smoking was found to be a significant risk factor in African populations, associated with a 3-fold increased risk.

1.4.2 Opium use

Opium use has been reported to increase the risk of ESCC in a number of studies.(37) It is suspected that opium smoking, similar to tobacco smoking, may result in exposure to PAHs due to increases in the urinary levels of 1-OHPG (1-hydroxy pyrene glucuronide).(15) Other studies have indicated that opium can be contaminated with lead, hence its association with cancer development.(38) Studies on the role of opium use are prone to discrepancies due to the differences in doses of opioids, route of administration or the time and method of exposure to opioids (39)

In one of the largest EC studies, the Golestan Cohort study in Iran, which collected data from 50, 000 participants (2004-2008) and followed up for 10 years, opium use emerged as a major risk factor in the population.(40)

1.4.3 Alcohol consumption

IARC has classified alcohol as carcinogenic to humans and stated that alcohol is associated with EC development.(41) Alcohol consumption causes exposure to acetaldehyde, which is a class 1 carcinogen associated with ESCC development and a primary metabolite for ethanol metabolism.(4) Heavy to moderate alcohol consumption has been reported to increase the risk of ESCC in high risk populations.(41) In Kenyan, Zambian and South African populations, the consumption of homemade/traditional beer has been reported to increase ESCC risk.(35, 42) Furthermore, preparation of the beer is often done in containers which may contain carcinogens e.g. oil drums may be mixed with potentially carcinogenic compounds.(15) Homemade and traditional beer forms part of the unrecorded alcohol consumption, and it is estimated that low income countries consume as much as 47.9% of unrecorded alcohol, with middle-income, upper middle-income, and high-income countries consuming 38.9%, 30.5%, and 11.2% respectively.(43) The consumption of unrecorded alcohol is important as it impacts health and is an major risk factor for ESCC. In a systematic review done on the prevalence of unrecorded alcohol consumption, data from sub-Saharan Africa showed that traditional beverages were less expensive than commercial beers, and were predominantly consumed in low socioeconomic settings.(43) Similar to tobacco use, alcohol consumption after ESCC diagnosis impacts survival negatively.(15)

Two African studies, on Nigerian and Ugandan populations, assessed the population attributable fractions (PAF) for alcohol use.(44, 45) In Nigeria, 42.1% of ESCC cases were attributable to alcohol.(44) However in Uganda, the PAF for alcohol use was lower (10.17%), and the combined PAF for smoking and alcohol was 12.66%.(45) This indicates that smoking and alcohol use are not major risk factors in Uganda, or that ESCC development in this population is through multifactorial interactions of risk factors.(45)

Tobacco smoking and alcohol use can interact resulting in increased risk of ESCC. A combination of tobacco smoking and alcohol consumption is reported to have a synergistic effect on ESCC risk.(15) A number of tumorigenesis mechanisms for combined tobacco smoking and alcohol use have been hypothesised which may happen concurrently. They include i) cellular DNA damage leading to a reduction in metabolic activity, detoxification capability and increased oxidation, ii) alcohol may cause the solvent effect on the esophagus resulting in increased permeability of the esophageal lumen to potential carcinogens, iii) smoking may instigate changes in the oral microbiome causing increased levels of acetaldehyde in saliva following alcohol consumption, and iv) prolonged alcohol consumption may induce cytochrome P450 enzymes resulting in an increase of carcinogenic ethanol metabolites.(15)

1.4.4 Diet

Observational studies and systematic reviews have shown that intake of fruits and vegetables decreases ESCC risk.(4, 46, 47) Among smokers, an increase in the consumption of vegetables and fruits combined and fruit consumption independently was associated with a 12% and 24% reduction of ESCC risk, respectively, in a large European study.(48) The Golestan Cohort Study in Iran reported that low intake of fruit and vegetables increased ESCC risk.(40) The proposed protective mechanisms of fruits and vegetables include; DNA methylation modulation; DNA damage protection and repair; detoxifying phase-II enzymes induction and promotion of apoptosis.(47) The role of micronutrients in ESCC development has been described in the literature. These micronutrients, vitamin A, pro-vitamin A carotenoids, selenium, retinol, thiamine, riboflavin, β-cryptoxathin, zinc, and iron are reported to reduce ESCC risk.(4, 49, 50) Whilst high levels of selenium have been reported to reduce ESCC risk, excess iron levels have been reported to increase ESCC risk. (51, 52) It is important to state that no action or intervention for iron risk reduction is recommended if ferritin and transferrin levels are within normal reference ranges recommended by the WHO (WHO, 2020) selenium deficiency has been reported in Malawi, which has the highest ESCC incidence rate in Africa. It is theorised that this is due to low levels of selenium in the staple food, maize.(53) Schaafsma et al. (54) performed a large ecological study assessing the role of seven micronutrients (calcium, iron, copper, iodine, magnesium, selenium, and zinc) in the development of ESCC in 32 African countries. Iron, zinc and selenium were described to have a protective effect in males and females, whilst magnesium was reported to be protective in females. A randomized trial was conducted in China, which constituted of six years of multivitamin supplementation and 20 years of post-intervention follow-up in 29 584 participants from 1986-1991.(55) Supplementation was done in combinations of multivitamins: A (retinol/zinc), B (riboflavin/niacin), C (vitamin C/molybdenum), and/or D (selenium/vitamin E/beta-carotene). Daily supplementation of factors A, B and C did not have an effect of total mortality, whilst factor D (selenium/vitamin E/beta-carotene) showed a protective effect against ESCC.(55)

This section contains a number of epidemiological study designs which have been utilized to explore the role of diet on ESCC development, these includes case control, cohort, ecological, and randomized case control trial studies. The quality and efficacy of the evidence from these studies differs due to biases linked to the study designs. Generally, researching the effect of diet on health is difficult, this is because consumption of a healthy diet is usually linked to other lifestyle factors which may be protective of cancer such as exercise, lack of smoking and lack of alcohol consumption.(56) Case control studies were used to retrospectively assess the dietary factors associated with ESCC. Case controls studies are a good design for chronic long latency diseases with multiple exposures like ESCC. However, particularly for dietary intake assessment, this study design is riddled with recall bias for past exposures, selection bias, as well as reverse causation where there is behavioural change from early ESCC symptoms or post diagnosis.(57) Cohort studies assess the effect of a specific dietary intake on ESCC for exposed and non-exposed participants prospectively or retrospectively. Cohort studies are less likely to be subject to recall bias as in case control studies. However, cohort studies still need to control for confounders and additionally, assessment of diet as a risk factor in cancer in cohort studies has not produced conclusive evidence. Other biases linked to cohort studies include sampling, observer biases, as well as non-response and loss to follow up.(57) In ecological studies associations between exposures and outcomes are assessed for entire populations or groups of people in different geographical regions at a single point in time. Evidence from ecological studies is confounded by ecological fallacy, where associations determined at the group levels are assumed to hold true at the individual level.(57) The use of an ecological study design for the assessment of dietary factors is therefore questionable, especially because of the heterogeneity of dietary patterns within populations. Randomized control trials (RCT)are considered the gold standard of study designs, if designed and implemented well they are the least subject to biases including confounding and selection biases. Limitations of RCTs include non-compliance, dropouts and randomization issues.(57) Particularly when assessing for the effect of diet and cancer, there are limitations as to which exposures are assignable to one group and not the other. Meta-analysis of dietary data also has limitations as adjustment of confounders can only be done for the individual studies. It is therefore important to take all these biases and shortcomings into considering when appraising studies which assessed dietary intake for ESCC.”

1.4.5 Socioeconomic status

ESCC development is highly depended on socioeconomic status.(4) Lower socioeconomic status has been linked with increased risk of ESCC.(4) Although socioeconomic status cannot influence cancer development or progression biologically, it can be associated with certain lifestyle, environmental and dietary factors which increase the risk of ESCC development. This disparity has been consistent in studies done in developed and developing countries.(4) Studies carried out in Sweden and the USA reported that risk was increased for populations with a lower income or not having a high school education.(58, 59) A study done in India reported an association between low socioeconomic status and risk of ESCC development.(60) Lower income, and lack of education and housing facilities were reported to increase ESCC risk in the population.(60) In a systematic review on the risk factors of ESCC in African populations, low socioeconomic status was reported as an important risk factor.(35)

1.4.6 Infectious agents

The infectious agents, human papilloma virus (HPV) and *Helicobacter pylori* have been implicated in the development of ESCC.(61) However, studies on both HPV and *H. pylori* have been inconsistent, hence the evidence for infectious aetiology for ESCC has not been conclusive.(61, 62) The discrepancies in the case of HPV may be attributed to the different methods employed in the detection of HPV DNA in different studies.(63) Additionally, sample collection and storage may also play a role. The role of the bacterial pathogens in esophageal carcinogenesis has been described in the literature. The human microbiome has been reported to induce carcinogenesis in several types of cancer through DNA damage (oxidative stress, and production of mutagens) and activation of metabolic pathways.(64-67) The microbiome is an environmental factor which we are constantly being exposed to.(67) Studies looking at the human microbiome have reported its association to ESCC development. The microbiome is acquired over the first few years of life, but is greatly influenced by diet, environmental exposure and oral health throughout the life.(68) Genetic factors of the host have also been reported to influence the human microbiome.(68) The majority of the studies have investigated the role of gastric and esophageal microbiota in ESCC, and have established an association.(64, 69, 70) Recently, a few studies have begun to analyse the possibility of the oral microbiome playing a role in esophageal carcinogenesis using saliva/mouthwash samples.(71) A study done by May et al. showed that the esophageal microbiome has a similar composition to the oral microbiome giving the possibility of a non-invasive approach in analysing the human microbiome.(72) One of these studies, done by Peters *et al.* (2017), showed that *Tannerella forsythia* and *Porphyromonas gingivalis* were associated with a high risk of ESCC, whilst a reduction in the genus *Neisseria* and the species *Streptococcus pneumoniae* resulted in a lower risk.(73) *F. nucleatum,* a microbe which is found in the oral cavity, has been analyzed from ESCC tissue samples and found to be associated with shorter survival of ESCC patients.(66)

1.4.7 Polycyclic aromatic hydrocarbons (PAHs)

Polycyclic aromatic hydrocarbons (PAH) are a group of compounds formed during the incomplete combustion of coal, oil, gas, wood, garbage, or other organic substances, such as tobacco and charbroiled meat.(15) They are present in the air, water, soil and food and therefore exposure occurs through inhalation, ingestion or percutaneous penetration.(9) PAHs are classified as class 1 carcinogens.(74) Once in the body, PAHs are metabolized within cells into active diol-epoxides via hydroxylation of the methylene carbon by cytochrome P450 enzymes, and elicit their effect by binding to macromolecules like DNA and initiating mutagenesis.(9) This is further elaborated in section 5.3.1: DNA adducts and cancer. Exposure to PAHs has been reported to contribute considerably to carcinogenesis of ESCC.(9) An excess risk of ESCC has been reported in individuals with direct exposure to ESCC in several countries.(9, 75, 76) The major pathways of PAH exposure are through:

i) Dietary exposure

More than 96% of the daily intake of PAHs is attributable to dietary intake.(9) Exposure to PAH can occur through ingestion of food items which have naturally high PAH content. An example is the yerba mate (*Ilex paraguayensis*) herb, commonly infused with water and drank in parts of South America.(9, 77) High levels of the benzo-a-pyrene, a PAH metabolite and other PAHs have been found in the leaves, and according to a study done by Kamangar et al. *"drinking a gourd of mate in the traditional manner can expose an individual to as much benzo-a-pyrene as smoking a typical pack of 20 cigarettes".*(9, 78) Addition of certain compounds to food can also lead to PAH exposure e.g. the mursik beverage consumed by the Kenyan population.(79) Mursik is a popular drink made by fermenting milk and lacing it with charcoal powder from a local herb plant to flavour the drink. The charcoal is likely to contain PAHs which increase the risk of ESCC.(79) Additionally food can be contaminated by PAHs during processing, preparation or accidental atmospheric contamination. Food preparation methods reported to increase PAH levels include smoking, grilling, barbequing (mainly because exposure to smoke), charbroiling and deep frying (mainly because of exposure to oil).(9)

ii) Tobacco smoking

Tobacco smoking is positively associated with carcinogenic urinary PAH biomarker, 1-OHPG.(78) It is reported that 99% of PAH urinary excretion is attributable to active smoking and consumption of food containing PAHs, in individuals who are not occupationally exposed to PAHs.(80) Whilst smokers are reported to have up to ten times the amount of 1-OHPG compared to non-smokers, second hand smoking is also associated with high levels of 1-OHPG.(9)

iii) Opium use

Opium consumption regardless of route of use, is positively associated with three PAHs, 1-hydroxynaphthalene, 3-hydroxyfluorene, and 1-hydroxypyrene.(81)

iv) Occupational exposure

Occupational exposure to carcinogens is an important risk factor of ESCC. The main carcinogen implicated in occupational exposures associated with ESCC is PAH. Exposure levels vary between occupations, with high levels of PAH exposure reported for the following occupations: foundry workers, chimney sweeps, blast furnace and coke-oven workers, vendors of broiled food, and steel plant and waste incineration workers.(9) Route of exposure is mainly inhalation and dermal.

v) Indoor air pollution

A major source of indoor PAH exposure is through the combustion of solid fuels for heating or cooking. In Chinese homes where coal is used for cooking and heating, high levels of PAH were found in indoor air.(9) Studies from South Africa, Kenya and Zambia have reported an association between indoor air pollution from heating and cooking fuel and ESCC development.(79, 82-85). The Golestan Cohort study also reported increased risk of ESCC from PAH exposure through indoor air pollution.(40) In a systematic review on the role of biomass fuel (wood, charcoal, coal, dung, and crop residues) in ESCC development, the authors reported that use of biomass fuel for heating and cooking was associated with ESCC development, due to smoke exposure.(86) Africa and Asia were reported to have the highest risk.

vi) Environmental air pollution

The main sources of atmospheric PAHs are from incomplete combustion of solid fuels from natural, industrial, commercial, vehicular and residential sources.(9, 87) Atmospheric PAH can travel for long distances and eventually deposits onto vegetation, soil and water bodies.(87)

1.4.8 Esophageal injury

1.4.8.1 Hot food and beverages

Esophageal thermal injury due to consumption of hot food and beverages has been associated with increased risk of ESCC. This includes hot teas and hot porridge. Hot beverages have been classified as probably carcinogenic to humans (Group 2A).(15) Hot beverage temperatures of 65^{0}C or higher have a positive association with ESCC development.(15, 88) High tea temperatures have been reported to be associated with ESCC in a number of populations including Kenyan, Tanzanian, Chinese, Iranian and

South American populations.(15, 40, 77, 88-91) There has been an increase in studies assessing the role of hot tea as a risk factor for ESCC in African populations.(35) ESCC development due to hot substance consumption can occur in a number of biological pathways. Firstly, the thermal injury from hot beverages can result in activation of the immune system, i.e., heat shock proteins, cytokines, and chemokines. This may result in chronic inflammation which will affect the intracellular signaling pathways and thereby facilitating the occurrence and survival of mutant cells and subsequently tumorigenesis.(15, 88) Secondly, esophageal thermal injury may result in the formation of endogenous nitrogen species, followed by formation of nitrosamines which are classified as Group 1 human carcinogens and mutagenic.(15) Another pathway is that thermal injury may allow for carcinogens in the beverage or food to permeate the esophagus.(15)

It has been established that high levels of PAHs are present in the mate beverage of South America. The anti-carcinogenic effects of tea leaves constituents have been established in a number of animal and some human studies, it however remains uncertain whether the consumption of tea or coffee at lower temperatures is protective of cancer development.(15, 92) Certain brands of black tea have been reported to contain PAHs, whilst certain teas, due to method of processing or preparation, can contain heavy metals, dioxins, fluoride, pesticides, and mycotoxins.(15) There is however, a lack of epidemiological evidence to support the theory of intraluminal carcinogens due to thermal injury as a risk factor for ESCC. Finally, thermal injury can also stimulate the proliferation of esophageal epithelial basal cells, and this has been confirmed by the increase of Ki-67 and CK-14 and decrease in CK5 positive cells on immunohistochemical analyses following thermal injury.(15) A major limitation in this research area is that most studies rely on self-reporting, which may not be accurate. Overall, more research is needed in this area.

1.4.8.2 Self-induced vomiting

In Africa, the role of self-induced vomiting and ESCC risk has almost exclusively been studied in two South African studies.(93, 94) Induced vomiting is done as a part of the indigenous traditional rituals with various methods to induce vomiting which include use of salt water, traditional medicine, warm water, holy water, and vinegar water.(93)

The two studies show conflicting results with one study reporting an association between induced vomiting and the other not.(93, 94) Self-induced vomiting is also a common practice in people with bulimia nervosa, and some studies have indicated an increased risk for ESCC in these patients who have no history of other major ESCC exposures. Self-induced vomiting may result in esophageal injury that can cause chronic inflammation of the esophagus and subsequently tumourigenesis.(93) The weakening of the esophageal epithelium may also allow carcinogens which may be present in the traditional emetics to permeate.

1.4.9 Oral Health

Poor oral health is an independent risk factor for ESCC, reported in a number of studies.(15) Both poor oral health and poor dentition have been associated with esophageal squamous cell dysplasia(95) and ESCC development in a number of countries including India(96), Kenya(97), China(98), Iran(40, 99), Japan.(100) A multicentric study including countries in Latin America (Argentina, Brazil, and Cuba), and Europe (Russia, Romania, and Poland) reported that periodontal disease increased the risk of ESCC development(101). However, in a cohort study from the USA consisting of 51,529 men followed up for 18 years (1986-2000), no association was reported between periodontal disease or tooth loss and EC.(102) In the Linxian Dysplasia Nutrition Intervention Trial Cohort done in China, which included a 30-year follow-up, moderate tooth loss was reported to be associated with an increase in ESCC mortality.(103)One African study which comprehensively looked at the role of oral hygiene as a risk factor for ESCC was done in Kenya by Menya at al.(97), and found significant associations between tooth loss and decayed teeth and ESCC development.(97) Another Kenyan study reported an association between tooth loss and ESCC. (79) It is yet unclear whether good oral hygiene induces protective or preventative effect on ESCC, whether it alters the oral microbiome or if it removes possible carcinogens from teeth.(15)

The oral microbiome, which colonizes the oral cavity, comprises 700 bacterial species.(2) It has been reported that a balance in the oral microbiome is crucial in maintaining oral health, hence poor oral health is linked to an imbalance in the oral microbiome.(2) The oral microbiome plays a role in tooth decay and periodontal disease.(2) Studies have shown that people diagnosed with ESCC have lower salivary microbial diversity compared to healthy controls.(2) It is also hypothesized that individuals with low oral microbial richness are at higher risk of esophageal squamous dysplasia.(2) The hypothesised biological pathways for the oral microbiome are that the oral microbial species: 1) Facilitate the production of a carcinogen, e.g., nitrite to nitrosamines or ethanol to acetaldehyde, or 2) Induce systemic inflammatory processes.(15)

1.4.10 Mycotoxins

Mycotoxins are secondary toxic metabolites produced by fungi, and encompass a class of toxins including fumonisins, aflatoxins, ochratoxin A, zearalenone, deoxynivalenol, diacetoxyscirpenol, nivalenol and trichothecenes.(15, 104) They mainly affect agricultural commodities and therefore can be inhaled, ingested or absorbed through skin. They are implicated in a number of diseases in both humans and animals. In humans, mycotoxin exposure has been reported to be associated with allergies, immunosuppression, hepatocellular carcinoma and renal cell carcinoma.(15) Their role on ESCC development is contested among scientists. High levels of fumonisin contamination have been reported in the ESCC high risk areas of South Africa, Iran and China.(105-111) Additionally, the recorded farming and dietary pattern change of maize to sorghum in the South African population and in the Italian population (after World War II), is reported to be associated with increased rates of OC in these countries.(15) This is because sorghum does not facilitate the growth of *fusarium* fungi, whereas maize does.(15) Sorghum is the staple diet in Nigeria, and ESCC is rare in the population.(15) In South Africa contaminated maize had also been reported to be used to make traditional or homemade beer, hence mycotoxins have been found in these beverages.(112, 113) Consumption of maize beer made with contaminated maize has been reported to be associated with ESCC.(112) This provides with a potential additive pathway in the risk to ESCC. However, no studies have been done to assess if there are additional carcinogens present in the homemade beers responsible for the increased risk.

Mainly fumonisins and aflatoxins have been associated with ESCC.(104) The mechanism of action for fumonisins is still unclear, but is postulated to be nongenotoxic.(104) Other pathways include oxidative damage, disruption of lipid metabolism, apoptosis and inhibition of sphingolipid metabolism.(104) Fumonisin B_1 is classified by IARC as possibly carcinogenic to humans (Group 2B), as there is compelling evidence from animal studies, but a lack of epidemiological evidence. In a prospective nested case-control study done in China, no significant associations were found between sphingolipid levels, which are biomarkers of fumonisin exposure, and ESCC incidence.(114)

1.4.11 Sex hormones

The current evidence on the role of sex hormones in ESCC development is inadequate and unclear. In a study done on the Survival Epidemiology and End Results (SEER9) dataset from 1977–2004, patients who developed prostate cancer had a reduced risk of EAC and ESCC.(115) It is unclear whether the androgen deprivation therapy for prostate cancer patients, contributed to this low risk, or if changes to lifestyle following a prostate cancer diagnosis contributed to this reduced risk.(115) Other confounders may also be responsible such as socioeconomic status or study biases.(115) In another study mutations in the androgen receptor genes were reported to be associated with ESCC development.(116) In animal studies, oestrogen is reported to have an inhibitory effect on esophageal tumour growth mediated by the oestrogen receptors.(117) In a case-control study done in Iran, female hormonal factors were reported to have a protective effect of ESCC risk.(118) However the evidence has been inconsistent and confounded by other factors. There has also been a lack of human and epidemiological studies to confirm these hypotheses. The reason for the difference of ESCC incidence rates between men and women in certain regions of the world, therefore, remains to be elucidated.

1.4.12 Environmental geochemistry

The high incidence of ESCC has been reported to have a geochemical contribution.(53) The peculiar distribution of ESCC high incidence along the African Cancer Corridor or the East African Rift Valley has not been adequately explained. Incidence rates in this corridor have been reported to be as high as 20-fold compared to West Africa, and the argument is that this difference cannot be entirely attributable to the established risk factors such as tobacco and alcohol.(53) The "Rift Valley Hypothesis" has been brought forward by McCormack et al. to explain the high incidence rates in this region.(119) Due to its volcanic pact, this region has distinctive physiography and residential altitudes are often above 2,000 meters and a near surface lithology of volcanic rock-types and subsequently a unique geochemistry.(53) This physiography is reported to yield a severe enrichment and depletion of both essential and possibly toxic elements in soil, groundwater and surface water.(53) The disease profile in the region is also distinct, with nonfilarial elephantiasis (podoconiosis), upper respiratory tract infections, Rift Valley fever, dental, skeletal fluorosis. Iodine deficiency as well as other trace element deficiencies are also common in this region. One hypothesis that has been put forward is micronutrient deficiencies, which has been described in section 4.4: Diet. More inter-disciplinary studies are needed to elucidate the role of environmental geochemistry in the aetiology of ESCC and confirm the current hypotheses and observations.

1.5 Genetic risk factors for ESCC

Genetic factors contribute to the increased risk of EC development. Genetic basis and susceptibility to ESCC has been studied, with reports of genomic alterations resulting in tumour development. (120, 121) It is important to note that the majority of the gene variants including single nucleotide polymorphisms (SNPs), nucleotide deletions and insertions are considered non-pathogenic with many having unknown effects and others contributing to differences between humans. Only a few variants have been directly associated with the development and progression of ESCC. Genetic variants reported in various types of cancers can be classified according to the new 2017 guidelines by the Association for Molecular Pathology, American Society of Clinical Oncology, and College of American Pathologists using a four-tiered system.(122) The tiers are as follows: tier I, variants of strong clinical significance; tier II, variants of potential clinical significance; tier III, variants of unknown significance; and tier IV, variants of known insignificance (i.e., likely benign or benign). Further elaboration of classification is recommended for somatic variants according to level of evidence and clinical significance in diagnosis, prognosis and therapeutics as shown in Table 2.

Table 1.7: Categorization of somatic variants into four tiers according to the Association for Molecular Pathology, American Society of Clinical Oncology, and College of American Pathologists 2017 guidelines

Tier	I Variants of strong clinical significance	II Variants of potential clinical significance	III Variants of unknown clinical significance	IV Benign or likely benign variants
Level of significance	Therapeutic, prognostic and diagnostics	Therapeutic, prognostic and diagnostics		
Level of evidence	Level A FDA approved therapies and included in professional guidelines	Level C FDA approved therapies, or investigational therapies, and multiple small published studies	Absence of convincing published data	Absence of published data
	Level B Well powered studies with expert consensus	Level D Preclinical trials or few case reports		

The reporting and classification of germline variants can be done according to a Joint Consensus Recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology .(123) The recommendation states that Mendelian disorders be classified in five categories as :‘pathogenic’, ‘likely pathogenic’, ‘uncertain significance’, ‘likely benign’, and ‘benign’, taking into consideration types of variant evidence i.e., population data, computational data, functional data, segregation data e.t.c.(123)

1.5.1 Genetics

ESCC has both an inherited and cellular genetic basis. (4, 121) Familial syndromes which have been associated with increased risk of malignancy include Tylosis and Fanconi anaemia.(4) Somatic mutations have been identified using DNA sequence analysis for targeted genes. Whole genome sequencing (WGS) and whole exome sequencing (WES) techniques have been reported to provide more comprehensive results in terms of mutation signatures.(124) Associations between genetic variants and risk of developing ESCC have been determined using genome wide association studies (GWAS) in Chinese and European studies.(4) GWAS involves high throughput genotyping which can identify genetic associations with a disease.(4) GWAS uses SNPs, to find which allele is more common in people with a specific disease than not. Overall, GWAS studies on African populations are scarce. The National Human Genome Research Institute–European Bioinformatics Institute (NHGRI-EBI) GWAS Catalogue, which contains information on over 4,346 published GWAS studies across more than 4,933 diseases and traits, shows limited geographic and demographic diversity.(125, 126) For cancer research the majority are European-ancestry groups which make up 86.15% of all GWAS participants, in contrast with 0.09% African, 13.29% Asian, 0.05% African-American and Afro-Caribbean, and 0.42% Hispanic or Latin American ancestries.(125) Figure 4 shows the total GWAS participants diversity by country according the GWAS Diversity Monitor.(126) WGS and WES studies have also improved knowledge on molecular signatures responsible for diseases, mostly in western and Asian populations.

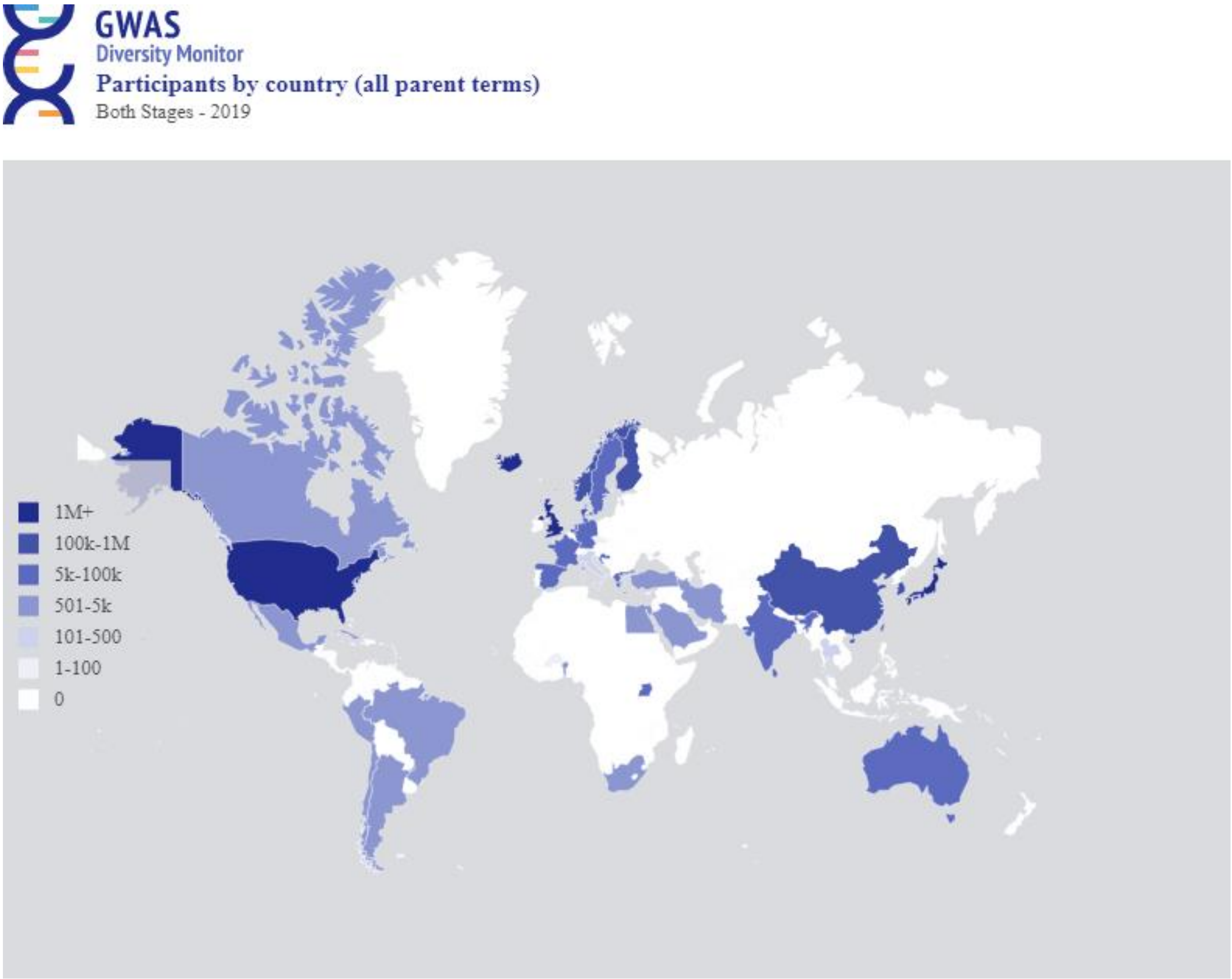


Figure 1.4: *Total GWAS participants diversity according to the GWAS Diversity Monitor for 4,933 diseases and traits. Reproduced from* [*https://gwasdiversitymonitor.com/*](https://gwasdiversitymonitor.com/) *Accessed 11/05/2020*

Genes which have been implicated in the development of ESCC include phospholipase c epsilon 1 gene *(PLCE1)*, caspase 8 gene *(CAP8)*, runt-related transcription factor 1 *(RUNX1)*, checkpoint kinase 2 *(CHEK2)*, phosphodiesterase 4D *(PDE4D)*, trans-membrane protein 173 *(TMEM173*), tumour protein 53 *(TP53)*, and human leukocyte antigen *(HLA)*.(4) A WGS study done by Liu et al. (2016) revealed mutations associated with ESCC in the following genes*: TP53, RB1, CDKN2A, PIK3CA, NOTCH1,* and *NFE2L2.(127)* In South Africa, a genetic association study by Bye *et al.* (2012) detected variants in *PLCE1* gene associated with ESCC development.(128) Other studies done in South Africa have shown associations with variants in *GSTP1, CYP2E1, SULT1A1* and *CYP3A5.*(129) A few studies in Africa have reassessed SNPs which were identified to be associated with ESCC in Chinese

GWAS studies, and only two studies carried out WES.(127, 130, 131) A number of systematic and meta-analyses have confirmed several genetic association to ESCC. However, these studies have been exclusively reported in the Western and Asian populations. Many early studies reported significant associations between various genes and risk of ESCC(132-139) but few of these findings replicated in GWAS, the gold standard for delineating genes associated with human diseases. At this time, no GWAS of oesophageal cancer in Africans have been published and validation of the current hypotheses awaits this more definitive study design.

In a recent study, 552 ESCC genomes from eight countries were assessed with the aim of elucidating the driving factors of the global variation in ESCC.(140) The authors combined whole genome sequencing and mutational signature analysis with epidemiological questionnaire data to assess if a specific environmental mutagen is driving the difference in ESCC global incidence.(140) The countries studies comprised of : Brazil, China, Iran, Japan, Kenya, Malawi, Tanzania and United Kingdom, and the risk factors assessed were tobacco smoking and alcohol consumption, hot food and drink consumption, indoor air pollution and opioid use. The authors found no evidence of a mutational signature linked to an environmental exposure which could explain the global difference in ESCC incidence.(140)

1.5.2 Epigenetics

Epigenetic modifications resulting from DNA methylation and histone modifications are associated with gene silencing, and other mechanisms that control gene expression.(141) Whilst these modifications occur naturally to regulate gene expression in the body, disruptions can occur due to changes in the internal body environment, genetic predisposition, exposure to environmental carcinogens, and lifestyle factors resulting in tumour development. Aberrant epigenetic modifications, particularly hypermethylation, have been reported to be associated with ESCC development.(3, 142) Hypermethylation of the *CDKN2A (p16INK4a)* tumour suppressor gene was reported to be associated with ESCC development.(3) This hypermethylation is associated with loss of expression and high grade tumours.(3) Another common epigenetic modification associated with ESCC development is hypermethylation of the *MGMT* gene.(3) *MGMT* hypermethylation results in a reduced level of MGMT protein, which plays a role in repairing DNA damage caused by the environmental carcinogens, nitrosames.(3) A summary of genes reported to be hypermethylated in ESCC, according to a review by Kaz et al.(3) is shown in Table 4. The table also includes percentage of hypermethylation which occurs in specific genes of ESCC patients.

Table 1.8: *Frequency of hypermethylation of genes studied in ESCC cases. [Table adapted from Kaz et al.]*

Official Gene Symbol (NCBI)	Full name	Frequency of hypermethylation (%)[1]	Reference
CDKN2A	Cyclin-dependent kinase	40 – 62	Guo et al 2006(143)
MGMT	O-6-Methylguanine-DNA	33 – 39	Guo et al 2006(143)
APC	Adenomatous polyposis	50	Kawakami 2000(149)
p14ARF	Protein product of CDKN2A	15	Xing et al 1999(147)
p15INK4b	Protein product of CDKN2A	12	Xing et al 1999(147)
DAB2	Disabled 2	20	Anupam et al 2006(150)
HIN-1[2]	High in normal 1	50	Guo et al 2008(151)
MLH1	MutL Homolog 1	23	Guo et al 2006(143)
RARB	Retinoic acid receptor β2	36 – 70	Guo et al 2006(143)
CDH1	Cadherin-1	34	Guo et al 2006(143)
DAPK1	Death-associated protein	26	Guo et al 2006(143)
ECRG4	Esophageal Cancer	60	Yue et al 2003(154)
FHIT	Fragile histidine triad	45 – 69	Noguchi 2003(155)
GNG7	G protein gamma 7	41	Ohta et al 2008(156)
TMEFF2	Transmembrane protein	54	Zhao et al 2008(157)
VHL	von Hippel-Lindau tumour	13	Kuroki et al 2003(153)
RASSF1	Ras association domain	51	Kuroki et al 2003(153)
CADM1	Cell adhesion molecule 1	50	Ito et al 2003(158)
UCHL1	Protein gene product 9.5	42	Mandelker et al 2005(159)
RPRM	Reprimo, TP53 dependent	13	Hamilton 2006(160)
SST	Somatostatin	54	Jin et al 2008(161)
CDH13	Cadherin-13	19	Jin et al 2008(162)
TAC1[2]	Tachykinin-1	50	Jin et al 2007(163)
NELL1	Nel-like 1	12	Jin et al 2007(164)

[1]Frequency of hypermethylation; percentage of cases demonstrating methylation of given gene
[2]NCBI gene symbol not found for the gene

1.5.3 Gene and protein expression

Gene expression is the process of transcription which results in phenotypic manifestation of genes. Gene expression also encompasses the regulation of these processes. This directional flow of information from DNA to RNA (transcription) and from RNA to protein (translation) is called the Central Dogma of molecular biology.(165) Whilst proteins dictate cell function, the amount and types of RNA in cell also reflect the function of the cell, making transcription the key control point for gene expression. Gene expression analysis determines the pattern of genes expressed at the transcription level, and identifies key genes and biological pathways associated with disease development, progression, and survival.(166) Gene expression can be analysed by microarray based or RNA sequencing approaches. Useful methods for studying transcriptional control and protein expression promoter analysis, protein expression profiling, and post-translational modification analysis.

Few studies have analysed gene expression in EC. Differential mRNA expression analysis has identified a number of genes and pathways associated with EC, and these include *KRAS, SPARC, SPP1, FOXM1, WDR66, PTGS2*, V-ATPase genes, tumour suppressor genes, and PI3K signalling pathway.(167-170) Non coding RNAs have been implicated in EC development, particularly long non-coding RNAs (lncRNAs) and microRNAs (miRNAs), which can either be tumour promoters or suppressors.(171) A number of miRNA have been reported to dysregulated in EC(172), with miR-145, miR-30a-3p, miR133a and miR-133b reported to be tumour promoters.(173, 174)

1.5.4 Gene–environment interactions

Gene-environment interactions and combined effects occur when a person’s genotype is susceptible to environmental exposures, which may lead to disease formation.(175) Understanding these interactions brings clarity to the aetiology of diseases, particularly ESCC, whose risk factors include both genetic and environmental factors.(175) Knowledge of gene-environment interactions is important in identifying populations and individuals at a higher risk, in order to make informed decisions on prevention interventions and treatment strategies.(176, 177) The diverse geographical pattern apparent in ESCC incidence suggests gene-environment interactions play a role in the aetiology and disease progression. This means that individuals with genetic variations associated with ESCC may be at an increased risk for developing the malignancy if a particular environmental factor is present or absent. In a study done by Bhat et al. (2017), variants in the *CYP2D6* gene were reported to be associated with tobacco smoking in a population in India.(178) CYP2D6 belong to the phase I xenobiotic metabolizing enzymes.(178) However, there have been few studies which have analysed the role of CYP2D6 in ESCC development in the presence of tobacco smoking. More studies are needed to ascertain if testing for CYP2D6 variants in ESCC patients exposed to tobacco can be recommended. The *CYP2C19*2* allele has combined effects with drinking hot tea and eating pickled vegetables in a Chinese population.(179) The *CYP2C19*2* allele is additionally associated with reduced activity of the enzyme cytochrome p450 2C19 (*CYP2C19*) which is involved in drug metabolism.(179)

1.5.4.1 DNA adducts and cancer

The genotoxicity of chemical carcinogens occurs through various pathways including formation of DNA adducts, DNA strand breaks, DNA-protein crosslinks, sister chromatid exchanges, chromosomal aberrations, and formation of micronuclei.(180-182) Chemical carcinogens have the ability to form covalent bonds with DNA. This results in DNA adducts which reflect exposure to carcinogens.(183) PAH-DNA adducts have been reported to cause genetic changes and mutations in proto-oncogenes and tumour-suppressor genes that can initiate carcinogenesis.(9, 184) Another mechanism of action of PAH-DNA adducts is through the generation of reactive intermediates through a one-electron oxidation process which may result in chemical alkylation of DNA and potentially mutagenic depurination.(9) DNA adducts are used as biomarkers of chemical/toxic exposure and for human risk assessment. Their quantification and analysis sheds more light on their association with cancer risk and carcinogenesis.(184) In a study done in 2010 by Marjani *et al.*, on an Iranian population, the number of PAH-DNA adducts found in ESCC tissue biopsies was significantly higher than that in esophageal tissue from controls.(75) The study established that PAH-DNA adducts have the potential to be used as biomarkers for PAH exposure and ESCC risk.(75)

1.6 Tools and techniques for assessment of risk factors in cancer research

Numerous tools and techniques are used in assessing risk factors in cancer research. These include animal, biomonitoring and human observational studies (Figure 5). Research on cancer risk factor aids in the understanding of carcinogenesis and in the identification of new approaches to disease prevention and treatment. Additionally, the data generated from cancer research focussed on risk factors can be used to lobby regulatory bodies and other stakeholders to set safety standards that reduce exposure to carcinogens and mutagens associated to cancer development.

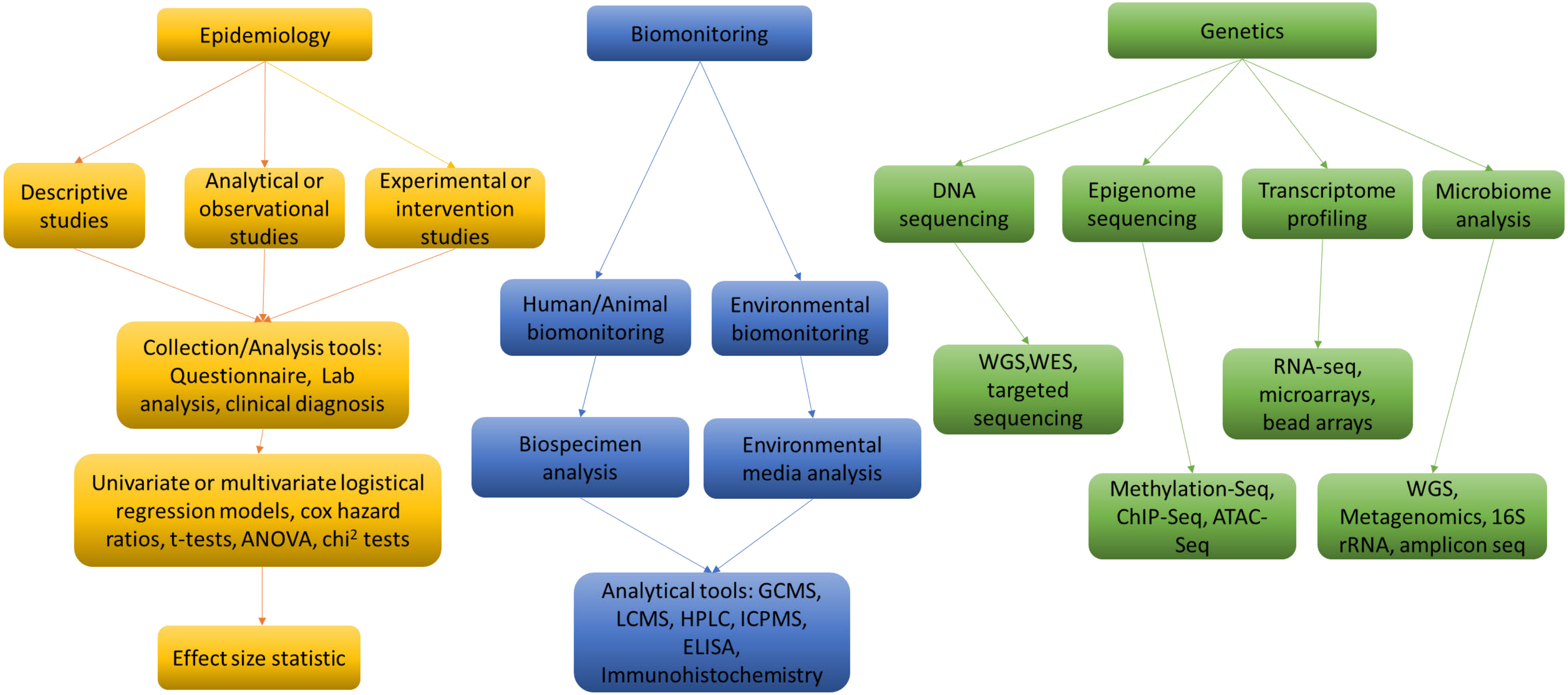

Epidemiology
Descriptive studies
Analytical or observational studies
Experimental or intervention studies
Collection/Analysis tools: Questionnaire, Lab analysis, clinical diagnosis
Univariate or multivariate logistical regression models, cox hazard ratios, t-tests, ANOVA, chi² tests
Effect size statistic
Biomonitoring
Human/Animal biomonitoring
Environmental biomonitoring
Biospecimen analysis
Environmental media analysis
Analytical tools: GCMS, LCMS, HPLC, ICPMS, ELISA, Immunohistochemistry
Genetics
DNA sequencing
Epigenome sequencing
Transcriptome profiling
Microbiome analysis
WGS,WES, targeted sequencing
RNA-seq, microarrays, bead arrays
Methylation-Seq, ChIP-Seq, ATAC-Seq
WGS, Metagenomics, 16S rRNA, amplicon seq

***Figure 1.5**: A summary of tools and techniques used in the assessment of risk factors in cancer research*

1.6.1 Epidemiological study designs

Epidemiology is essential in the assessment of risk factors for cancer. The basic elements of epidemiological studies include(185):

I. Formulation of the research question or hypothesis
II. Selection of study populations and study samples
III. Selection of indicators of exposure
IV. Measurement of exposure and disease
V. Analysis of the relationship between exposure and outcome
VI. Evaluation of the role of biases
VII. Evaluation of the role of chance

The main types of epidemiologic study designs include descriptive studies, analytical or observational studies, and experimental or intervention studies.(185) These study designs can be used in the assessment of risk factors associated with cancer.

Descriptive studies examine the patterns of occurrence of disease according to person, place, and time.(185) Analytical studies test a specific hypothesis and can be divided into four groups; ecological, cross sectional, case-control and cohort studies.(185) In ecological studies associations between exposures and outcomes are assessed for entire populations or groups of people in different geographical regions at a single point in time. Cross-sectional studies assess exposure outcomes from a subset of a population at a specific point in time. Cohort studies investigate risk of specific health outcomes for individuals over a period of time. Case-control studies assess outcomes due to exposure by comparing individuals who already have the outcome (cases) versus individuals from the same population who do not have the outcome (controls). Experimental studies include the introduction of an intervention into the study to assess the cause and effect of an exposure. They also include randomisation of the study cohort into two groups, the exposure group and the control group.

Questionnaires are commonly used for collection of data used in exposure assessment for epidemiological studies. This is because in most cases they provide the most efficient data collection method. Questionnaires can either be self-administered or administered by an interviewer over several platforms including face-to-face, handed out, mailed, over-the-phone interview, online form, questionnaire diaries and obtaining information from a proxy responded in the case of death.(186) Statistical tests commonly used to analyse data from the questionnaire include univariate or multivariate logistic regression models, or cox proportional hazard models. These statistical analyses provide with quantifiable risk/odds of developing the disease or risk/odds of dying from the disease due to a specific risk factor. Data from these studies can be critically appraised and synthesized qualitatively or quantitatively using systematic reviews and meta-analysis. Systematics reviews and meta-analysis can validate the role of reported risk factors and provide aggregated effect sizes.

Table 1.9: Advantages and disadvantages of epidemiological study designs

Method	Summary	Advantages	Disadvantages
Case-control studies	Measurement of exposures retrospectively in diseases and non-diseased individuals	Assessment of multiple exposures Suitable for rare diseases with long latency periods Relatively cost effective, efficient often less time consuming	Recall bias Sampling bias Multiple outcomes cannot be assessed Reverse causation Calculation of prevalence, incidence, population relative risk or attributable risk not possible Confounding factors

Cross-sectional studies	Assessment of existing disease and current exposure levels at one point in time	Quick and inexpensive Good for hypothesis testing Assessment of multiple exposures and outcomes Estimations of prevalence	Cannot provide cause and effect Recall bias May miss latent disease Confounding factors
Cohort studies	Assessment of outcomes for exposed and non exposed participants prospectively or retrospectively	Low recall bias Multiple outcomes can be assessed Can measure population based incidence rates Can assess rare exposures	More prone to selection bias Can be expensive and time consuming Sampling and observer bias Study non-response Loss-to-follow-up Confounding factors
Ecological studies	Assessment of population-level effect	Assesses health status and needs of community	Ecological fallacy

	of exposures on a disease	Generation of a hypothesis	
Randomised control studies	Experimental studies where exposure/treatment is assigned	Has the least biases of all the study designs	Expensive Loss-to-follow-up Has to take intention to treat into considerations Not all exposures/treatments are assignable Randomization issues due to crossovers and non compliance

Meta-analysis	Aggregated data analysis from multiple individual studies	Increases external validity Method easily replicable Analysis done according to rigorous rules Increases statistical power Improves effect size estimates	Adjustment of confounders can only be done for the individual studies Publication bias

1.6.2 Biomonitoring study designs

Human biomonitoring studies

Human biomonitoring is the measurement of exposure to chemicals or their metabolites in human bodily fluids (blood, urine, saliva and breast milk) and tissues (hair, nails, fat and bone).(187) This allows for the quantification of potentially toxic chemicals to humans, extent of exposure and how these exposures may be changing over time and most importantly, the health risks associated.(187) The data from human biomonitoring studies can also be used as evidence to support policy making and health impact assessments with the overall agenda of reducing exposure to toxic and carcinogenic chemicals.

According to the European Union's (EU) 7th Environmental Action Programme of 2013, human biomonitoring is an important tool that can provide "authorities with a more comprehensive view of actual exposure of the population to pollutants, especially sensitive groups such as children, and can provide better evidence from guiding appropriate responses".(188)

The analytical methodologies used in biomonitoring studies are mostly mass spectrometry based. A substance or biomarker/s of interest is identified which will guide the type of bio-specimen used and the assay used. The assays mainly used include inductively coupled plasma mass spectrometry (ICP-MS) for trace elements, gas chromatography mass spectrometry (GC/MS) or gas chromatography-tandem mass spectrometry (GC-MS/MS) for volatile organic compounds and their metabolites, and liquid chromatography-tandem mass spectrometry (LCMS/MS) for semi- and non-volatile compounds and their metabolites.(189, 190) DNA and protein adducts are assessed using High Performance Liquid Chromatography (HPLC), Enzyme-Linked Immunosorbent Assay (ELISA), and immunohistochemical techniques.(191)

Environmental biomonitoring

Environmental monitoring is the sampling and analysis of toxic chemicals, and bacteria in environmental media (soil, plants, air and water). It is an important aspect in the maintenance of both environmental and human health.(192) The assays used to detect and quantify contaminants in environmental media are also dependant on the type of contaminant and type of media. Assays similar to those mentioned under the human biomonitoring section are also used in environmental biomonitoring. Mutagenicity of environmental chemicals can also be investigated using laboratory organism with the Ames and Comet assays.(193) Ionizing radiation in the environment is measured using a Geiger counter.

1.6.3 Animal studies in Esophageal Cancer

Animal models play a significant role in understanding the pathophysiology of EC. This includes tumour-host interaction, environmental exposure assessment, mechanisms of metastasis, and the development of therapeutic interventions.(194-198) Laboratory animal models used in the EC research include rat (*Rattus norvegicus*), mouse (*Mus musculus*), rabbit (*Oryctolagus cuniculus*), guinea pig (*Cavia porcellus*) and hamster models (*Mesocricetus auratus*).(194)

Three main rat models are used in EC studies. This includes the rat reflux model that is suitable for studying EAC, very few studies have used this model to study ESCC.(194) The model is appropriate to use in EAC due to the rat being larger in size compared to the mouse, additionally, the pathophysiology of EAC in the rat is similar to that in humans.(194) (199) Establishment of the model is normally done through esophagoduodenal anastomosis with total gastrectomy surgical procedure.(194) Development of the model usually takes 40 weeks. Second is the orthotopic xenograft implantation rat model which is used in ESCC studies. Human ESCC cells are injected subcutaneously into the cervical esophagus of athymic nude rats.(194, 200) The model develops in six weeks. In the chemical induction model, rats are subcutaneously injected with N-nitrosomethylbenzylamine (NMBA) for five weeks.(194) NMBA is a mutagen known to induce ESCC development in rats.(194)

There are five mouse models commonly used in EC studies. The subcutaneous xenograft model is the most common and mature mouse model used in EC studies. Human esophageal cancer cell lines are used to induce both EAC and ESCC but being injected into the dorsolateral flanks of immunodeficient mice subcutaneaosly.(194) The transgenic model has mice that are genetically modified and therefore the development of the model takes a shorter time. Induction of ESCC is mostly done by insertion of the Epstein–Barr virus ED-L2.(199) This is an early lytic cycle promoter that targets the cyclin D1 in the transgenic mouse model leading to dysplastic conditions in the esophagus.(194) Induction of EAC in the transgenic model is done through insertion of interleukin-1β cDNA into the Epstein–Barr virus ED-L2.(194, 199) This results in metaplasia of the esophagus similar to BE. The orthotopic xenograft model can be used to analyse both tumour development and metastasis, unlike the previous two models which are used to analyse tumour development. The model is developed by surgical transplantation of histologically intact human cancer fragments through a series of steps.(194, 199) The patient-derived xenograft model is similar to subcutaneous xenograft model apart from the fact that the immunodeficient mice in this model are injected with tumour biopsy cells from EC patients.(194, 199) The tumour takes three months to develop. Mice are the most commonly used animals for tumour xenograft models due to the fact that they have a comparable genome size with humans, a short reproductive cycle, large litter size, low maintenance expenses and are easy to manipulate.(199) Lastly, the chemical induction in transgenic model is one of the best mouse models (together with the transgenic model) to study ESCC.(194) The mutagens 4-nitroquinoline 1-oxide (4-NQO), NMBA, deoxycholate, and N-methyl-N-nitrosourea are used to induce ESCC.(194) It takes about 24 weeks for the model to develop post induction.

Two rabbit models are commonly used in EC studies. The induction of tumour cells by surgical technique model is used in ESCC studies. ESCC is induced by surgically introducing the VX2 tumour cell suspension into the submucosal layer or muscular layer of the cervical esophagus, from the esophageal tunica adventitia.(194) The second model, endoscopic implantation model, has a similar induction method to the induction of tumour cells by surgical technique model, but the induction is done using an endoscope.(194) The endoscope introduces the tumour cells into the submucosa of the esophagus. VX2 tumour cells lines can mimic human ESCC and are able to produce a variety of pathological characteristics.(194)

Gastresophageal reflux model is used to study EAC in guinea pigs.(194) Development of the reflux model is done through perfusion of esophagus with HCl (containing 1 g/L pepsin) for 20 minutes/day.(194) Inflammation is induced due the acid, which develops into Barrett's esophagus and subsequently EAC.(194) Whilst hamsters are commonly used for oral cancer, few studies have included them in EC. The benzo[a]pyrene (BP) induced model uses instillation of BP to induce ESCC in hamsters.(194)

Animal models play a critical role in our understanding on EC, and in the development of effective drugs. They form the bridge from basic to clinical research and can complement in vitro studies.

1.6.4 Genetic study designs

In the past 15 years, the emergence of high throughput DNA sequencing technologies have dramatically transformed human genetics research and clinical diagnostics.(201) These technologies are collectively termed next generation sequencing (NGS). NGS includes WGS, targeted sequencing, metagenomics, transcriptome quantitation and sequencing, DNA methylation analysis, ribosome profiling and ChIP sequencing.(202) A summary of advantages and limitations of high throughput methods used in genome analysis are shown in Table 5.(203-208) Biospecimens used for analysis include blood, saliva, urine, stool and tumour samples. There are a number of NGS approaches including Illumina sequencing-by-synthesis (Solexa technology), Roche 454 pyrosequencing, AB SOLID colour-based sequencing by ligation, Ion Torrent semiconductor sequencing and Single-molecule sequencing (PacBio, MinION, etc).(209) Types of variants detectable using NGS include large amplifications, large deletions, point mutations (SNPs), insertions/deletions, rearrangements, copy number variations (CNVs), and fusions/splice variants.

Advantages of NGS compared to previous sequencing technologies i.e., Sanger sequencing, which was laborious, time consuming and expensive (dependant on number of genes analysed, fewer genes are less costly) were that(210):

I. They do not require electrophoretic separation of sequencing products
II. They are relatively cheaper, and prices continue to decrease
III. They are high throughput, whilst sanger sequencing allows for only one read (1 kb), NGS allows for up to thousands of GB of DNA to be read on a single run on a single chip.
IV. They have higher accuracy, due to the generation of multiple data points on a single nucleotide locus
V. A single molecule of DNA can generate a nucleotide sequence, whilst in Sanger sequencing several thousands of copies of DNA are needed as input.

One of the major drawbacks of NGS is that is produces short reads, this results in heavily fragmented genomes in the form of contigs.(210) Short reads have a restricted capacity to link independent variations present on the same DNA molecule.(210) Subsequently, NGS methods are not well suited to differentiate and phase alleles to their corresponding parental homologs, an important aspect in human genetics.(210) Additionally, detection and characterization of larger structural variants (SVs) is difficult with NGS. This is particularly important considering that SVs have been associated with the development of many diseases.(210) Finally NGS techniques are dependent on PCR which causes difficulties in amplifying regions of high GC content.(210)

Following the introduction of NGS, Third-Generation Sequencing (TGS)/Long-Read Sequencing emerged. One of the main distinguishing features of TGS is the single-molecule sequencing (SMS) and sequencing in real-time which is not present in NGS.(210) The first viable TGS technology was produced by PacBio, called ‘single-molecule real-time’ (SMRT) sequencing. A more recent TGS technology was produced by Oxford Nanopore Technologies (ONT) called Nanopore sequencing. A major feature of both SMRT sequencing and Nanopore sequencing is the production if long reads.(210) Other advantages of TGS methods, compared to NGS include(210):

I. Generation of reads of up to tens of kilobase up to 1 Mb, compared to NGS which generates hundreds of base pairs

II. Lack of PCR amplification which results in less bias and increased homogeneity in genome coverage

III. Enhanced performance in the analysis of repeated regions and SVs, haplotype phasing, and transcriptome analysis

Disadvantages of TGS technologies, particularly Nanopore sequencing is the high error rate. Overall, TGS allows for the analysis of genomes, transcriptomes, and metagenomes at an unprecedented resolution. The third revolution of sequencing is only at the beginning and new technologies and innovations will continue being introduced.

Table 1.10: Summary of high throughput methods used in genetic analysis of human diseases

Method	Summary	Advantages	Disadvantages
Whole genome sequencing (WGS)	Sequencing of the entire genomic DNA sequence of an organism. Comprehensive characterisation of the genome, which is a starting point for the elucidation of function	Provides complete coverage of the coding and non-coding regions of the genome. Has better determination of structural variants due to having reads longer than 2×100 paired Easier fine mapping	May be costly for developing countries Challenges in interpretation of data Storage and analysis issues due to large amount of data produced Relatively less accurate compared to SNP arrays
Whole exome sequencing (WES)	Entails targeting exonic regions of the genome.	Ensures higher depth of coverage Provides a platform for creating custom panels Cheaper than WGS	It only covers the protein coding regions of the genome Representation of genomic SV is limited

Targeted sequencing	Detection of known and novel variants in selected sets of genes or genomic regions	Rapid and cost effective	Limited to only selected genes
SNP arrays	A type of DNA array used to detect SNPs in a DNA sample. It contains designed probes which can determine specific alleles in a given sample	Relatively less costly compared to WGS Reliable and highly accurate Well established pipelines for analysis	Genomic coverage is mostly restricted towards high frequency variants, and biased towards variants present in highly sequenced populations
Methylation-Sequencing (methyl-seq or bisulfite sequencing)	Profiles the methylation status of genomic regions of interest with single nucleotide resolution	Generates resolution at DNA level. High coverage of sparse CpG dinucleotides Effective in providing evidence of cytosine methylation	Difficulty in distinguishing between methylated and hemi-methylated cytosine Low coverage of CGIs

ChIP-Sequencing	In-situ genome- wide profiling of DNA-binding proteins and histone and nucleosome modifications.	Fast and well-studied. Compatible with both array-and sequencing based analysis, therefore possible to perform genome-wide analysis Unlimited dynamic range Multiplexing possible High throughput analysis of large numbers of single cells	Relies on antibody specificity Requires a lot of tissue as input samples Low coverage of sequencing reads
ATAC-Sequencing	Assay for Transposase-Accessible Chromatin (ATAC-seq) sequencing assesses genome- wide chromatin accessibility. Analysis uncovers how chromatic packaging and other factors are associated with gene expression.	Requires no prior knowledge of regulatory elements Low number of cells as unput material Shorter experimental time compared to other methods Uses paired-end sequencing technology to map nucleosome positioning and	ATAC-generated data contains mitochondrial DNA DNA fragments are joined at random by adapter, resulting in a 1 in 2 chances that the adapters at both ends of the adapter are the same. This produces unusable fragments. Low coverage of sequencing reads

		occupancy, resulting in more accurate mapping.	
Metagenomics	Sequencing and identification of genetic material from multiple taxa	Ability to identify genetic material from different kingdoms of organisms simultaneously Analysis is hypothesis free.	Difficulties in the interpretation of results due to contamination and colonization Difficulties in the selection and validation of databases used for analysis
16S rRNA sequencing	The use of 16S rRNA sequencing as the gold standard for identification and classification of bacterial species is because it is present in almost all bacteria. 16S rRNA gene function has not changed over time The 16S rRNA gene is 1,500 bp, and therefore easy to process with basic sequencing methods and informatics.	It is highly conserved between different species of bacteria and archaea therefore universal primers can be used. Data analysis can be done on pre-existing pipelines	Sequences only a single region of the bacterial genome, unlike WGS which has more bacterial species per read

Ampliconsequencing	High throughput sequencing of a PCR product (amplicon) that targets a specific region of the genome	It is ideal for the detection of rare variants. phase variation assessment, haplotype description of complex immune regions, detection of somatic variants in complex samples, taxonomic classification of microorganisms, validation and follow-up of WES, and genome editing Comprehensive reference databases available	Functional annotations are inferential Low confidence in the characterisation of taxa at the species level Overestimation of richness

1.6.4.1 Global genomics project on cancer

An interdisciplinary team of scientists from four continents recently embarked on a global genomics project which aimed to understand the full genetic complexity of cancer. The Pan-Cancer Analysis of Whole Genomes (PCAWG) Consortium performed the most comprehensive meta-analysis of cancer genomes to date, resulting in studies of 2658 tumours and 38 tissues.(211) Of these, 98 were EAC tumours. A total of six main papers were produced from the study with each paper assessing a key aspect of cancer genomics i.e. cancer drivers(212), non-coding changes(213), mutational signatures, structural variants, cancer evolution, and RNA alternations.

In the first study, which focussed on the assessment of whole genomes, authors reported that 91% of the tumours had at least one identified driver mutation.(212) On average, cancer genomes contained four to five driver mutations, with coding and non-coding genomic elements combined.(212) In 5% of the cases no cancer drivers were identified. Many of the tumours exhibited patterns of clustered mutations including chromoplexy (17.8% of samples), kataegis (65% of cancers), and chromothripsis (22.3% of samples).(212) Chromoplexy is the re-arrangement of co-occurring double stranded DNA breaks, normally on different chromosomes. Kataegis is a process where focal hypermutation occurs resulting in locally clustered nucleotide substitutions, with bias towards a single strand. Chromothripsis is a mutational process in which simultaneous DNA breaks occur (tens to hundreds) on one or few chromosomes with subsequent near-random stitching of fragments.

In the second study, which explored the non-coding somatic drivers in the whole genomes, the authors used rigorous statistical models to reliably identify non-coding drivers.(213) Overall, the study showed that compared to protein coding driver, non-protein coding drivers are rare. This may be due to low discovery power or low sequence coverage. The most frequently mutated non-coding driver was the *TERT* promoter, with mutations being associated with high *TERT* overexpression.(213) The TERT enzyme facilitates the uncontrollable division of tumour cells.(211) The authors reported recurrent somatic events in the 3′ UTRs of *TOB1* (carcinoma and pan-cancer meta-cohorts), *NFKBIZ* (lymphomas) and *ALB* (liver cancer).(213) Significant recurrence of mutations in the non-coding region of the tumour suppressor gene *TP53* were also reported.(213) The mutations were coupled with loss of heterozygosity and reduced mRNA expression levels.(213) Interestingly, previously reported non-coding drivers including the ncRNAs, *NEAT1* and *MALAT1* were found to not be genuine drivers.(213)

Li *et a*l(214) and Alexandrov *et al*(215) explored the mutational signatures in cancer together identified 97 signatures. In the study by Li *et al*(214), the authors developed methods for grouping, classifying and describing somatic structural variants. Structural variation occurs in the form in genomic segments rearrangements i.e. deletions, amplification and reordering.(214) The authors identified 16 signatures of somatic variants, and the roles of these structural variants were also ascertained.(214) Structural variants facilitate tumour development by affecting cancer genes through copy number alteration, tumour suppressor gene suppression, creation of fusion genes, and the juxtaposition of one gene's DNA and the regulatory apparatus of another.(214) This publication represents one of the first studies to discover reproducible structural variant signatures.(214) The study by Alexandrov *et al*(215) identified 79 structural somatic variants. These included 49 single-base-substitution, eleven doublet-base-substitution, four clustered-base-substitution and 17 small insertion-and-deletion signatures. New signatures were discovered with included, SBS31, SBS32, SBS35, SBS36, SBS42 and SBS44.

In the 5th paper, Gerstung *et al (216)* characterised the evolutionary history of the 2568 tumour samples and 38 cancer types. This was done by inferring timing and patterns of chromosomal evolution in each tumour type and learning the characteristic sequences of mutations.(216) This was the first large-scale genome-wide study to reconstruct the evolutionary patterns of cancer, by reconstructing precancer and later stages of 38 tumour types.(216) Progression of most precancerous lesions to tumours normally occurs over years to decades. This study corroborated this and further extended these timeframes to tumours without detectable precursor lesions.(216)

The final paper, presented the most comprehensive catalogue of cancer-associated gene alterations to date, which was done by characterisation of tumour transcriptomes from 1,188 donors in the PCAWG dataset.(217) This is, to date, the largest resource of RNA phenotype and their underlying genetic mechanisms in cancer. Calabrese *et al* (217) used WGS data to explore associations between altered RNA expression and germline and somatic DNA alterations. Hundreds of single nucleotide variants were identified and were associated with expression of nearby genes. Somatic copy number alterations were the major mutational drivers of total gene and allele-specific expression.(217) The study revealed that 731 genes were recurrently altered by several mechanisms, including TP53 and GAS7.(217) Overall the study demonstrated that RNA analyses reveals cancer-associated pathway alterations that have not been discovered by DNA methods.

1.6.5 Gene expression study designs

There are an array of techniques used to analyse and quantify gene expression and its regulation. Older or low to mid-plex techniques include reporter gene assays (DNA regulation), northern blotting (RNA expression), western blotting (protein expression), fluorescent in situ hybridization (FISH) (identification and location of gene sequences), Reverse transcription polymerase chain reaction (RT-PCR) (detecting and quantifying mRNA). Gene expression data can also be determined using NGS techniques.(218, 219) These include microarray and RNA sequencing. Bioinformatics tools are used to annotate, filter and analyse variants associated with diseases and determine expression levels of mRNAs, miRNAs and lnc RNAs.

Table 1.11: Summary of high throughput methods used in gene and protein expression analysis of human diseases

Method	Summary	Advantages	Disadvantages
RNA sequencing	Profiles the transcriptome (mRNA, rRNA and tRNA etc). RNA- sequencing provides information on which genes are turned on or off in a cell, their level of expression, and what times they are activated or shut off.	Direct, quantitative and high throughput method. Does not require a prior knowledge on genomic features. Suitable for gene, transcripts (including alternative gene spliced transcripts) or allele-specific expression detection Does not rely on previous sequence information High dynamic range with no saturation Direct sequence alignment, hybridisation not required,	Has high sequence resemblance between alternatively spliced isoforms Relatively expensive compared to microarrays High powered computing facilities required Splice variants analysis complex

Microarray	Extensively studied method with well defined pipelines for analysis and well defined protocols for hybridisation High throughput and quantitative method Based on fluorescence therefore no need of radioactive probes Low cost compared to RNA sequencing	Analysis is only for predefined sequences High variance detected for low expressed genes Reliant on hybridisations which can be non-specific Generally does not specify splice variants Limited dynamic range Complex data analysis

1.7 Rationale for the Study

The geographical distribution of ESCC worldwide points to population-specific risk factors. Despite there being more data coming out from Western and Asian research, there is still a dearth of information regarding the etiology of ESCC, particularly on the role of genetic risk factors and the pathobiology of ESCC. There still remains a high level of uncertainty regarding the role of genetic, environmental and lifestyle risk factors. The questions which are currently unanswered for the African populations are:

I. What are the genetic variants that are associated with ESCC development?
II. What is the underlying molecular pathobiology driving EC development, progression?
III. What is the role for gene-environment interaction in tumour development?
IV. Which environmental and lifestyle factors increase susceptibility to ESCC?

Answering these questions will not only bring in new and relevant knowledge on the disease but will also inform prevention, early screening and diagnostic strategies for ESCC. It will provide a platform to bring in solutions which are relevant to the African population.

Scientific evidence has shown that ESCC etiology includes both genetic and environmental factors. Genetic variants that increase person's susceptibility to develop tumours and mutations which induce tumorigenesis have been described in literature, more-so for the Western and Asian populations. Studies in Africa and South Africa are still lacking, and the contribution of genetic variants has not been fully resolved. Recently, in an African first, a study was done on a Malawian population (2016) looking at the genetic basis of ESCC using WES and RNA-sequencing (127). Some studies in South Africa have reassessed the polymorphisms found in the Asian studies on the South African population.(128, 129, 220). However, the studies highlighted the complexities of translating and analysing findings from one population to another. Whilst it is clear that more studies need to be done, the major hurdles that remain include having studies with small sample sizes and patient groups that were not well characterized. A number of environmental factors have been described as well (4, 11, 36, 46, 49). Exposure to PAH and acetaldehyde has been reported to influence carcinogenesis. Such evidence in the African population has been indirect as no studies have looked at the direct measure of these exposures in individuals. Environmental factors, including PAH, acetaldehyde have been reported to influence carcinogenesis through host DNA damage and activation of oncogenic pathways (65, 66, 175, 182, 221). PAH in particular has been reported to form DNA adducts. Carcinogen-DNA adducts are biomarkers of exposure, which are reported to cause genomic alterations which result in carcinogenesis.

It is clear that gene-environment interactions and combined effects underlie EC aetiology. The diverse geographical pattern apparent in EC incidence suggests gene-environmental interactions play a role in susceptibility, disease progression and survival. These interactions are important in identifying populations and individuals which may be at a higher risk, to make informed and personalized decisions on screening, prevention interventions and treatment strategies. They also shed light on disease aetiology. There have been very few studies looking at gene-environment interactions and combined effects for EC in Africa.

Understanding the genetic basis and pathobiology of EC, and the link of the major environmental exposures to genomic alterations is what makes this study of interest to us. The high incidence of ESCC in South Africa, and the fatal nature of the disease, warrants a dedicated study on the genetic and environmental basis of disease, whose epidemiology differs from that of the West and Asia.

1.8 Aims and objectives

The aim of this study is to assess and characterise the role of genetic and environmental factors in the development of EC, and investigate the underlying molecular pathobiology using gene expression. This will be achieved through the following objectives:

1. **Chapter 2:** To assess all genetic risk factors for ESCC reported in relevant African literature using a systematic review

2. **Chapter 3:** To assess all ESCC risk factors reported in relevant African literature using a systematic review and meta-analysis

3. **Chapter 4:** To identify genes and pathways with differential mRNA expression in EC (EAC, ESCC and BE) using meta-analysis of DEGs and Gene Set Enrichment Analysis of GEO (Gene Expression Omnibus) datasets.

1.8.1 Why is this PhD study of importance?

This Introduction chapter gave an analysis of EC epidemiology as well as the current state of research activities being done worldwide. This global assessment highlighted that there is a plethora of risk factors associated with EC, some of which are globally significant whilst others are only locally/regionally relevant. The strength of the evidence for various risk factors are also inconsistent between studies. Therefore, a systematic review to assess the strength of the evidence across various epidemiological studies, and reporting of pooled estimates for various risk factors, was considered a necessity. Chapters 2 and 3 which are systematic reviews assessed the role of genetic factors, and environmental and lifestyle factors, respectively, giving a more in-depth analysis on the association of these risk factors on ESCC development, with a focus on African populations. The aim was to critically appraise the studies in order understand the merits, strengths and weaknesses of these studies included in the analysis, and to compute and analyse pooled risk estimates in order to determine the gravity of the evidence for ESCC risk factors in African populations. Assessing the genetic risk factors was particularly important in highlighting the research gaps which exists, considering the dearth of information on genetic studies from African countries. Gene-environment interactions were also described in the systematic review focussed on the role of genetic variants. Assessing the pathobiology of EC in chapter 4 provided an opportunity for further in-depth analysis into the genomics of EC through the analysis and identification of genes and biological pathways involved in its pathogenesis. This was achieved through leveraging available microarray data from

GEO datasets platform and performing a comprehensive meta-analysis of DEGs from multiple datasets. Insights on plausible environmental factors associated with the dysregulated pathways identified in the study were also discussed. The lack of data from African populations in this chapter highlights the lack of genetic studies in Africa, and the need for more research to be done.

1.9 References

1. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: a cancer journal for clinicians* (2018) 68(6):394-424.

2. Wang Q, Rao Y, Guo X, Liu N, Liu S, Wen P, et al. Oral Microbiome in Patients with Oesophageal Squamous Cell Carcinoma. *Sci Rep* (2019) 9(1):19055. Epub 2019/12/15. doi: 10.1038/s41598-019-55667-w. PubMed PMID: 31836795; PubMed Central PMCID: PMCPMC6910992.

3. Kaz AM, Grady WM. Epigenetic biomarkers in esophageal cancer. *Cancer Letters* (2014) 342(2):193-9. doi: http://dx.doi.org/10.1016/j.canlet.2012.02.036.

4. Abnet CC, Arnold M, Wei W-Q. Epidemiology of esophageal squamous cell carcinoma. *Gastroenterology* (2017).

5. Cummings D, Wong J, Palm R, Hoffe S, Almhanna K, Vignesh S. Epidemiology, Diagnosis, Staging and Multimodal Therapy of Esophageal and Gastric Tumors. *Cancers (Basel)* (2021) 13(3). Epub 2021/02/06. doi: 10.3390/cancers13030582. PubMed PMID: 33540736; PubMed Central PMCID: PMCPMC7867245.

6. Van Loon K, Mwachiro MM, Abnet CC, Akoko L, Assefa M, Burgert SL, et al. The African Esophageal cancer Consortium: A call to action. *Journal of Global Oncology* (2018) (pagination). doi: http://dx.doi.org/10.1200/JGO.17.00163. PubMed PMID: 624435622.

7. Thrift AP. The epidemic of oesophageal carcinoma: Where are we now? *Cancer epidemiology* (2016) 41:88-95. Epub 2016/02/07. doi: 10.1016/j.canep.2016.01.013. PubMed PMID: 26851752.

8. Abnet CC, Arnold M, Wei WQ. Epidemiology of Esophageal Squamous Cell Carcinoma. *Gastroenterology* (2017) 154(2):360-73. Epub 2017/08/22. doi: doi:10.1053/j.gastro.2017.08.023. PubMed PMID: 28823862; PubMed Central PMCID: PMCPMC5836473.

9. Roshandel G, Semnani S, Malekzadeh R, Dawsey SM. Polycyclic aromatic hydrocarbons and esophageal squamous cell carcinoma. *Arch Iran Med* (2012)

15(11):713-22. Epub 2012/10/30. doi: 0121511/aim.0013. PubMed PMID: 23102250; PubMed Central PMCID: PMCPMC5757504.

10. Bray F, Colombet M, Mery L, Piñeros M, Znaor A, R Z, et al. Cancer Incidence in Five Continents, Vol. XI. France: IARC, (2017).

11. Munishi MO, Hanisch R, Mapunda O, Ndyetabura T, Ndaro A, Schüz J, et al. Africa's oesophageal cancer corridor - do hot beverages contribute? *Cancer causes & control : CCC* (2015) 26(10):1477-86. doi: 10.1007/s10552-015-0646-9. PubMed PMID: PMC4838015.

12. Parkin DM, Bray F, Ferlay J, Jemal A. Cancer in Africa 2012. *Cancer epidemiology, biomarkers & prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology* (2014) 23(6):953-66. Epub 2014/04/05. doi: 10.1158/1055-9965.epi-14-0281. PubMed PMID: 24700176.

13. McGlashan ND. Oesophageal cancer and alcoholic spirits in central Africa. *Gut* (1969) 10(8):643-50. Epub 1969/08/01. PubMed PMID: 5810975; PubMed Central PMCID: PMCPMC1552917.

14. Odera JO, Odera E, Githang'a J, Walong EO, Li F, Xiong Z, et al. Esophageal cancer in Kenya. *American journal of digestive disease* (2017) 4(3):23.

15. Chetwood JD, Garg P, Finch P, Gordon M. Systematic review: the etiology of esophageal squamous cell carcinoma in low-income settings. *Expert Rev Gastroenterol Hepatol* (2019) 13(1):71-88. Epub 2019/02/23. doi: 10.1080/17474124.2019.1543024. PubMed PMID: 30791842.

16. Van Loon K, Mwachiro MM, Abnet CC, Akoko L, Assefa M, Burgert SL, et al. The African Esophageal Cancer Consortium: A Call to Action. *Journal of global oncology* (2018) (4):1-9. Epub 2018/09/23. doi: 10.1200/jgo.17.00163. PubMed PMID: 30241229; PubMed Central PMCID: PMCPMC6223465.

17. National Cancer Registry. Cancer in South Africa South Africa(2017) [cited 2021 16 June 2021]. Available from: https://www.nicd.ac.za/centres/national-cancer-registry/.

18. Burrell R. Oesophageal cancer in the Bantu. *South African Medical Journal* (1957) 31(17):401-9.

19. Somdyala NI, Parkin DM, Sithole N, Bradshaw D. Trends in cancer incidence in rural Eastern Cape Province; South Africa, 1998-2012. *Int J Cancer* (2015) 136(5):E470-4. Epub 2014/09/23. doi: 10.1002/ijc.29224. PubMed PMID: 25236502.

20. Society AC. Esophageal Cancer USA: American Cancer Society (2018) [updated January 2018; cited 2018 03 Februaty 2018]. Available from: https://www.cancer.org/cancer/esophagus-cancer/about/what-is-cancer-of-the-esophagus.html.

21. Auld M, Srinath H, Jeyarajan E. Oesophageal Squamous Dysplasia. *Journal of Gastrointestinal Cancer* (2018) 49(3):385-8. doi: 10.1007/s12029-018-0122-3.

22. Oweira H, Schmidt J, Mehrabi A, Kulaksiz H, Schneider P, Schöb O, et al. Validation of the eighth clinical American Joint Committee on Cancer stage grouping for esophageal cancer. *Future oncology* (2018) 14(1):65-75.

23. Wei WQ, Chen ZF, He YT, Feng H, Hou J, Lin DM, et al. Long-Term Follow-Up of a Community Assignment, One-Time Endoscopic Screening Study of Esophageal Cancer in China. *J Clin Oncol* (2015) 33(17):1951-7. Epub 2015/05/06. doi: 10.1200/jco.2014.58.0423. PubMed PMID: 25940715; PubMed Central PMCID: PMCPMC4881309 online at www.jco.org. Author contributions are found at the end of this article.

24. Guan CT, Song GH, Li BY, Gong YW, Hao CQ, Xue LY, et al. Endoscopy screening effect on stage distributions of esophageal cancer: A cluster randomized cohort study in China. *Cancer Sci* (2018) 109(6):1995-2002. Epub 2018/04/11. doi: 10.1111/cas.13606. PubMed PMID: 29635717; PubMed Central PMCID: PMCPMC5989864.

25. van Boxel GI, Kingma BF, Voskens FJ, Ruurda JP, van Hillegersberg R. Robotic-assisted minimally invasive esophagectomy: past, present and future. *J Thorac Dis* (2020) 12(2):54-62. Epub 2020/03/20. doi: 10.21037/jtd.2019.06.75. PubMed PMID: 32190354; PubMed Central PMCID: PMCPMC7061186.

26. Murphy G, McCormack V, Abedi-Ardekani B, Arnold M, Camargo MC, Dar NA, et al. International cancer seminars: a focus on esophageal squamous cell carcinoma. *Ann Oncol* (2017) 28(9):2086-93. Epub 2017/09/16. doi: 10.1093/annonc/mdx279. PubMed PMID: 28911061.

27. Engel LS, Chow WH, Vaughan TL, Gammon MD, Risch HA, Stanford JL, et al. Population attributable risks of esophageal and gastric cancers. *J Natl Cancer Inst* (2003) 95(18):1404-13. Epub 2003/09/18. doi: 10.1093/jnci/djg047. PubMed PMID: 13130116.

28. Tran GD, Sun XD, Abnet CC, Fan JH, Dawsey SM, Dong ZW, et al. Prospective study of risk factors for esophageal and gastric cancers in the Linxian general population trial cohort in China. *Int J Cancer* (2005) 113(3):456-63. Epub 2004/09/30. doi: 10.1002/ijc.20616. PubMed PMID: 15455378.

29. Demeester SR. Epidemiology and biology of esophageal cancer. *Gastrointestinal cancer research : GCR* (2009) 3(2 Suppl):S2-S5. PubMed PMID: 19461918.

30. Lambert R, Hainaut P. Epidemiology of oesophagogastric cancer. *Best Practice & Research Clinical Gastroenterology* (2007) 21(6):921-45. doi: https://doi.org/10.1016/j.bpg.2007.10.001.

31. Lee KD, Wang TY, Lu CH, Huang CE, Chen MC. The bidirectional association between oral cancer and esophageal cancer: A population-based study in Taiwan over a 28-year period. *Oncotarget* (2017) 8(27):44567-78. Epub 2017/06/01. doi:

10.18632/oncotarget.17818. PubMed PMID: 28562351; PubMed Central PMCID: PMCPMC5546502.

32. Meyers TJ, Chang SC, Chang PY, Morgenstern H, Tashkin DP, Rao JY, et al. Case-control study of cumulative cigarette tar exposure and lung and upper aerodigestive tract cancers. *Int J Cancer* (2017) 140(9):2040-50. Epub 2017/02/07. doi: 10.1002/ijc.30632. PubMed PMID: 28164274; PubMed Central PMCID: PMCPMC5552057.

33. Rafiq R, Shah IA, Bhat GA, Lone MM, Islami F, Boffetta P, et al. Secondhand Smoking and the Risk of Esophageal Squamous Cell Carcinoma in a High Incidence Region, Kashmir, India: A Case-control-observational Study. *Medicine (Baltimore)* (2016) 95(1):e2340. Epub 2016/01/07. doi: 10.1097/md.0000000000002340. PubMed PMID: 26735535; PubMed Central PMCID: PMCPMC4706255.

34. Wang N, Tan B, Cao F, Song Q, Wang J, Jia Y, et al. Prognostic influence of smoking on esophageal squamous cell carcinoma. *Int J Clin Exp Med* (2015) 8(10):18867-72. Epub 2016/01/16. PubMed PMID: 26770509; PubMed Central PMCID: PMCPMC4694409.

35. Asombang AW, Chishinga N, Nkhoma A, Chipaila J, Nsokolo B, Manda-Mapalo M, et al. Systematic review and meta-analysis of esophageal cancer in Africa: Epidemiology, risk factors, management and outcomes. *World J Gastroenterol* (2019) 25(31):4512-33. Epub 2019/09/10. doi: 10.3748/wjg.v25.i31.4512. PubMed PMID: 31496629; PubMed Central PMCID: PMCPMC6710188.

36. Sewram V, Sitas F, O'Connell D, Myers J. Tobacco and alcohol as risk factors for oesophageal cancer in a high incidence area in South Africa. *Cancer epidemiology* (2016) 41:113-21.

37. Kamangar F, Shakeri R, Malekzadeh R, Islami F. Opium use: an emerging risk factor for cancer? *The Lancet Oncology* (2014) 15(2):e69-77. Epub 2014/02/01. doi: 10.1016/s1470-2045(13)70550-3. PubMed PMID: 24480557.

38. Mahmoodpoor A, Golzari SE. Epigenetics, opium, and cancer. *The Lancet Oncology* (2014) 15(4):e153. Epub 2014/04/04. doi: 10.1016/s1470-2045(14)70077-4. PubMed PMID: 24694638.

39. Szczepaniak A, Fichna J, Zielinska M. Opioids in Cancer Development, Progression and Metastasis: Focus on Colorectal Cancer. *Current treatment options in oncology* (2020) 21(1):6. Epub 2020/01/24. doi: 10.1007/s11864-019-0699-1. PubMed PMID: 31970561; PubMed Central PMCID: PMCPMC6976545.

40. Sheikh M, Poustchi H, Pourshams A, Etemadi A, Islami F, Khoshnia M, et al. Individual and Combined Effects of Environmental Risk Factors for Esophageal Cancer Based on Results From the Golestan Cohort Study. *Gastroenterology* (2019) 156(5):1416-27. Epub 2019/01/07. doi: 10.1053/j.gastro.2018.12.024. PubMed PMID: 30611753; PubMed Central PMCID: PMCPMC7507680.

41. Islami F, Fedirko V, Tramacere I, Bagnardi V, Jenab M, Scotti L, et al. Alcohol drinking and esophageal squamous cell carcinoma with focus on light-drinkers and never-smokers: a systematic review and meta-analysis. *Int J Cancer* (2011) 129(10):2473-84. Epub 2010/12/31. doi: 10.1002/ijc.25885. PubMed PMID: 21190191.

42. Menya D, Kigen N, Oduor M, Maina SK, Some F, Chumba D, et al. Traditional and commercial alcohols and esophageal cancer risk in Kenya. *International journal of cancer* (2019) 144(3):459-69. doi: http://dx.doi.org/10.1002/ijc.31804. PubMed PMID: 624876329.

43. Rehm J, Kailasapillai S, Larsen E, Rehm MX, Samokhvalov AV, Shield KD, et al. A systematic review of the epidemiology of unrecorded alcohol consumption and the chemical composition of unrecorded alcohol. *Addiction* (2014) 109(6):880-93. doi: 10.1111/add.12498.

44. Odutola MK, Jedy-Agba EE, Dareng EO, Adebamowo SN, Oga EA, Igbinoba F, et al. Cancers attributable to alcohol consumption in Nigeria: 2012-2014. *Frontiers in Oncology* (2017) 7(AUG):183. doi: http://dx.doi.org/10.3389/fonc.2017.00183.

45. Okello S, Churchill C, Owori R, Nasasira B, Tumuhimbise C, Abonga CL, et al. Population attributable fraction of Esophageal squamous cell carcinoma due to smoking and alcohol in Uganda. *BMC Cancer* (2016) 16:446. Epub 2016/07/13. doi: 10.1186/s12885-016-2492-x. PubMed PMID: 27400987; PubMed Central PMCID: PMCPMC4940693.

46. Sewram V, Sitas F, O'Connell D, Myers J. Diet and esophageal cancer risk in the Eastern Cape Province of South Africa. *Nutrition and cancer* (2014) 66(5):791-9.

47. Liu J, Wang J, Leng Y, Lv C. Intake of fruit and vegetables and risk of esophageal squamous cell carcinoma: a meta-analysis of observational studies. *Int J Cancer* (2013) 133(2):473-85. Epub 2013/01/16. doi: 10.1002/ijc.28024. PubMed PMID: 23319052.

48. Jeurnink SM, Buchner FL, Bueno-de-Mesquita HB, Siersema PD, Boshuizen HC, Numans ME, et al. Variety in vegetable and fruit consumption and the risk of gastric and esophageal cancer in the European Prospective Investigation into Cancer and Nutrition. *Int J Cancer* (2012) 131(6):E963-73. Epub 2012/03/07. doi: 10.1002/ijc.27517. PubMed PMID: 22392502.

49. Schaafsma T, Wakefield J, Hanisch R, Bray F, Schüz J, Joy EJM, et al. Africa's Oesophageal Cancer Corridor: Geographic Variations in Incidence Correlate with Certain Micronutrient Deficiencies. *PLoS ONE* (2015) 10(10):e0140107. doi: 10.1371/journal.pone.0140107. PubMed PMID: PMC4598094.

50. Choi S, Cui C, Luo Y, Kim SH, Ko JK, Huo X, et al. Selective inhibitory effects of zinc on cell proliferation in esophageal squamous cell carcinoma through Orai1. *Faseb j* (2018) 32(1):404-16. Epub 2017/09/21. doi: 10.1096/fj.201700227RRR. PubMed PMID: 28928244.

51. Jaskiewicz K, Marasas W, Rossouw J, Van Niekerk F, Tech E. Selenium and other mineral elements in populations at risk for esophageal cancer. *Cancer* (1988) 62(12):2635-9.

52. Strickland NJ, Matsha T, Erasmus RT, Zaahl MG. Molecular analysis of Ceruloplasmin in a South African cohort presenting with oesophageal cancer. *International journal of cancer* (2012) 131(3):623-32.

53. Middleton DR, McCormack VA, Watts MJ, Schüz J. Environmental geochemistry and cancer: a pertinent global health problem requiring interdisciplinary collaboration. *Environmental geochemistry and health* (2019):1-10.

54. Schaafsma T, Wakefield J, Hanisch R, Bray F, Schuz J, Joy EJ, et al. Africa's Oesophageal Cancer Corridor: Geographic Variations in Incidence Correlate with Certain Micronutrient Deficiencies. *PloS one* (2015) 10(10):e0140107. Epub 2015/10/09. doi: 10.1371/journal.pone.0140107. PubMed PMID: 26448405; PubMed Central PMCID: PMCPMC4598094.

55. Wang SM, Taylor PR, Fan JH, Pfeiffer RM, Gail MH, Liang H, et al. Effects of Nutrition Intervention on Total and Cancer Mortality: 25-Year Post-trial Follow-up of the 5.25-Year Linxian Nutrition Intervention Trial. *J Natl Cancer Inst* (2018) 110(11):1229-38. Epub 2018/04/05. doi: 10.1093/jnci/djy043. PubMed PMID: 29617851; PubMed Central PMCID: PMCPMC6235687.

56. Cantwell MM. The Role of Diet in Cancer Development and Prevention. *Current Nutrition Reports* (2012) 1(1):1-7. doi: 10.1007/s13668-011-0002-y.

57. Munnangi S, Boktor SW. Epidemiology Of Study Design. *StatPearls*. Treasure Island (FL): StatPearls Publishing

Copyright © 2021, StatPearls Publishing LLC. (2021).

58. Jansson C, Johansson AL, Nyrén O, Lagergren J. Socioeconomic factors and risk of esophageal adenocarcinoma: a nationwide Swedish case-control study. *Cancer Epidemiology and Prevention Biomarkers* (2005) 14(7):1754-61.

59. Gammon MD, Ahsan H, Schoenberg JB, West AB, Rotterdam H, Niwa S, et al. Tobacco, alcohol, and socioeconomic status and adenocarcinomas of the esophagus and gastric cardia. *Journal of the National Cancer Institute* (1997) 89(17):1277-84.

60. Dar NA, Shah IA, Bhat GA, Makhdoomi MA, Iqbal B, Rafiq R, et al. Socioeconomic status and esophageal squamous cell carcinoma risk in Kashmir, India. *Cancer science* (2013) 104(9):1231-6.

61. Liyanage SS, Segelov E, Garland SM, Tabrizi SN, Seale H, Crowe PJ, et al. Role of human papillomaviruses in esophageal squamous cell carcinoma. *Asia-Pacific Journal of Clinical Oncology* (2013) 9(1):12-28.

62. Xie F-J, Zhang Y-P, Zheng Q-Q, Jin H-C, Wang F-L, Chen M, et al. Helicobacter pylori infection and esophageal cancer risk: An updated meta-analysis.

World Journal of Gastroenterology : WJG (2013) 19(36):6098-107. doi: 10.3748/wjg.v19.i36.6098. PubMed PMID: PMC3785633.

63. Pastrez PRA, Mariano VS, da Costa AM, Silva EM, Scapulatempo-Neto C, Guimarães DP, et al. The Relation of HPV Infection and Expression of p53 and p16 Proteins in Esophageal Squamous Cells Carcinoma. *J Cancer* (2017) 8(6):1062-70. doi: 10.7150/jca.17080. PubMed PMID: 28529620.

64. Zhang C, Cleveland K, Schnoll-Sussman F, McClure B, Bigg M, Thakkar P, et al. Identification of low abundance microbiome in clinical samples using whole genome sequencing. *Genome biology* (2015) 16(1):265.

65. Baba Y, Yamamura K, Nakagawa S, Mima K, Ishimoto T, Iwatsuki M, et al. *Genetic and epigenetic characteristics of esophageal cancer tissues with microbiome fusobacterium nucleatum*. AACR (2017).

66. Yamamura K, Baba Y, Nakagawa S, Mima K, Miyake K, Nakamura K, et al. Human microbiome Fusobacterium nucleatum in esophageal cancer tissue is associated with prognosis. *Clinical Cancer Research* (2016) 22(22):5574-81.

67. Bultman SJ. Emerging roles of the microbiome in cancer. *Carcinogenesis* (2013) 35(2):249-55.

68. Goodrich JK, Davenport ER, Clark AG, Ley RE. The Relationship Between the Human Genome and Microbiome Comes into View. *Annu Rev Genet* (2017) 51:413-33. Epub 2017/09/22. doi: 10.1146/annurev-genet-110711-155532. PubMed PMID: 28934590; PubMed Central PMCID: PMCPmc5744868.

69. Nasrollahzadeh D, Malekzadeh R, Ploner A, Shakeri R, Sotoudeh M, Fahimi S, et al. Variations of gastric corpus microbiota are associated with early esophageal squamous cell carcinoma and squamous dysplasia. *Scientific reports* (2015) 5:8820.

70. Hall AB, Tolonen AC, Xavier RJ. Human genetic variation and the gut microbiome in disease. *Nat Rev Genet* (2017) 18(11):690-9. Epub 2017/08/22. doi: 10.1038/nrg.2017.63. PubMed PMID: 28824167.

71. Chen X, Winckler B, Lu M, Cheng H, Yuan Z, Yang Y, et al. Oral microbiota and risk for esophageal squamous cell carcinoma in a high-risk area of China. *PLoS One* (2015) 10(12):e0143603.

72. May M, Abrams JA. Emerging Insights into the Esophageal Microbiome. *Curr Treat Options Gastroenterol* (2018):1-14.

73. Peters BA, Wu J, Pei Z, Yang L, Purdue MP, Freedman ND, et al. Oral Microbiome Composition Reflects Prospective Risk for Esophageal Cancers. *Cancer research* (2017) 77(23):6777-87.

74. IARC. *IARC Working Group on the Evaluation of Carcinogenic Risks to Humans: Some Non-heterocyclic Polycyclic Aromatic Hydrocarbons and Some Related Exposures*: World Health Organization (2010).

75. Marjani H, Biramijamal F, Rakhshani N, Hossein-Nezhad A, Malekzadeh R. Investigation of NQO1 genetic polymorphism, NQO1 gene expression and PAH-DNA adducts in ESCC. A case-control study from Iran. *Genetics and Molecular Research* (2010) 9(1):239-49.

76. Pratt MM, John K, MacLean AB, Afework S, Phillips DH, Poirier MC. Polycyclic aromatic hydrocarbon (PAH) exposure and DNA adduct semi-quantitation in archived human tissues. *International journal of environmental research and public health* (2011) 8(7):2675-91.

77. Sewram V, De Stefani E, Brennan P, Boffetta P. Mate consumption and the risk of squamous cell esophageal cancer in uruguay. *Cancer Epidemiol Biomarkers Prev* (2003) 12(6):508-13. Epub 2003/06/20. PubMed PMID: 12814995.

78. Kamangar F, Schantz MM, Abnet CC, Fagundes RB, Dawsey SM. High levels of carcinogenic polycyclic aromatic hydrocarbons in mate drinks. *Cancer Epidemiol Biomarkers Prev* (2008) 17(5):1262-8. Epub 2008/05/17. doi: 10.1158/1055-9965.epi-08-0025. PubMed PMID: 18483349.

79. Patel K, Wakhisi J, Mining S, Mwangi A, Patel R. Esophageal Cancer, the Topmost Cancer at MTRH in the Rift Valley, Kenya, and Its Potential Risk Factors. *ISRN oncology* (2013) 2013:503249. Epub 2014/02/04. doi: 10.1155/2013/503249. PubMed PMID: 24490085; PubMed Central PMCID: PMCPMC3893746.

80. Van Rooij JG, Veeger MM, Bodelier-Bade MM, Scheepers PT, Jongeneelen FJ. Smoking and dietary intake of polycyclic aromatic hydrocarbons as sources of interindividual variability in the baseline excretion of 1-hydroxypyrene in urine. *Int Arch Occup Environ Health* (1994) 66(1):55-65. Epub 1994/01/01. doi: 10.1007/bf00386580. PubMed PMID: 7927844.

81. Etemadi A, Poustchi H, Calafat AM, Blount BC, De Jesus VR, Wang L, et al. Opiate and Tobacco Use and Exposure to Carcinogens and Toxicants in the Golestan Cohort Study. *Cancer Epidemiol Biomarkers Prev* (2020) 29(3):650-8. Epub 2020/01/10. doi: 10.1158/1055-9965.epi-19-1212. PubMed PMID: 31915141.

82. Kayamba V, Bateman AC, Asombang AW, Shibemba A, Zyambo K, Banda T, et al. HIV infection and domestic smoke exposure, but not human papillomavirus, are risk factors for esophageal squamous cell carcinoma in Zambia: a case-control study. *Cancer medicine* (2015) 4(4):588-95. Epub 2015/02/03. doi: 10.1002/cam4.434. PubMed PMID: 25641622; PubMed Central PMCID: PMCPMC4402073.

83. Dandara C, Ballo R, Parker MI. CYP3A5 genotypes and risk of oesophageal cancer in two South African populations. *Cancer Lett* (2005) 225(2):275-82. Epub 2005/06/28. doi: 10.1016/j.canlet.2004.11.004. PubMed PMID: 15978331.

84. Dandara C, Li D-P, Walther G, Parker MI. Gene-environment interaction: the role of SULT1A1 and CYP3A5 polymorphisms as risk modifiers for squamous cell carcinoma of the oesophagus. *Carcinogenesis* (2006) 27(4):791-7.

85. Pacella-Norman R, Urban MI, Sitas F, Carrara H, Sur R, Hale M, et al. Risk factors for oesophageal, lung, oral and laryngeal cancers in black South Africans. *British journal of cancer* (2002) 86(11):1751-6. Epub 2002/06/28. doi: 10.1038/sj.bjc.6600338. PubMed PMID: 12087462; PubMed Central PMCID: PMCPMC2375408.

86. Okello S, Akello SJ, Dwomoh E, Byaruhanga E, Opio CK, Zhang R, et al. Biomass fuel as a risk factor for esophageal squamous cell carcinoma: a systematic review and meta-analysis. *Environmental Health: A Global Access Science Source* (2019) 18(1):60-. doi: 10.1186/s12940-019-0496-0. PubMed PMID: 31262333.

87. Kim KH, Jahan SA, Kabir E, Brown RJ. A review of airborne polycyclic aromatic hydrocarbons (PAHs) and their human health effects. *Environ Int* (2013) 60:71-80. Epub 2013/09/10. doi: 10.1016/j.envint.2013.07.019. PubMed PMID: 24013021.

88. Maghsudlu M, Farashahi Yazd E. Heat-induced inflammation and its role in esophageal cancer. *J Dig Dis* (2017) 18(8):431-44. Epub 2017/07/28. doi: 10.1111/1751-2980.12511. PubMed PMID: 28749599.

89. Middleton DR, Menya D, Kigen N, Oduor M, Maina SK, Some F, et al. Hot beverages and oesophageal cancer risk in western Kenya: Findings from the ESCCAPE case-control study. *International journal of cancer* (2019) 144(11):2669-76. doi: https://dx.doi.org/10.1002/ijc.32032.

90. Munishi MO, Hanisch R, Mapunda O, Ndyetabura T, Ndaro A, Schuz J, et al. Africa's oesophageal cancer corridor: Do hot beverages contribute? *Cancer Causes Control* (2015) 26(10):1477-86. Epub 2015/08/08. doi: 10.1007/s10552-015-0646-9. PubMed PMID: 26245249; PubMed Central PMCID: PMCPMC4838015.

91. Okaru AO, Rullmann A, Farah A, Gonzalez de Mejia E, Stern MC, Lachenmeier DW. Comparative oesophageal cancer risk assessment of hot beverage consumption (coffee, mate and tea): the margin of exposure of PAH vs very hot temperatures. *BMC Cancer* (2018) 18:1-. doi: 10.1186/s12885-018-4060-z. PubMed PMID: 128244410. Language: English. Entry Date: In Process. Revision Date: 20180927. Publication Type: journal article.

92. Yang CS, Lambert JD, Ju J, Lu G, Sang S. Tea and cancer prevention: Molecular mechanisms and human relevance. *Toxicology and Applied Pharmacology* (2007) 224(3):265-73. doi: https://doi.org/10.1016/j.taap.2006.11.024.

93. Matsha T, Stepien A, Blanco-Blanco E, Brink LT, Lombard CJ, Van Rensburg S, et al. Self-induced vomiting -- risk for oesophageal cancer? *S Afr Med J* (2006) 96(3):209-12. Epub 2006/04/12. PubMed PMID: 16607430.

94. Sammon AM. A case-control study of diet and social factors in cancer of the esophagus in Transkei. *Cancer* (1992) 69(4):860-5. Epub 1992/02/15. PubMed PMID: 1735077.

95. Sun Y, Zhang T, Wu W, Zhao D, Zhang N, Cui Y, et al. Risk Factors Associated with Precancerous Lesions of Esophageal Squamous Cell Carcinoma: a

Screening Study in a High Risk Chinese Population. *J Cancer* (2019) 10(14):3284-90. Epub 2019/07/11. doi: 10.7150/jca.29979. PubMed PMID: 31289600; PubMed Central PMCID: PMCPMC6603371.

96. Dar NA, Islami F, Bhat GA, Shah IA, Makhdoomi MA, Iqbal B, et al. Poor oral hygiene and risk of esophageal squamous cell carcinoma in Kashmir. *Br J Cancer* (2013) 109(5):1367-72. Epub 2013/08/01. doi: 10.1038/bjc.2013.437. PubMed PMID: 23900216; PubMed Central PMCID: PMCPMC3778293.

97. Menya D, Maina SK, Kibosia C, Kigen N, Oduor M, Some F, et al. Dental fluorosis and oral health in the African Esophageal Cancer Corridor: Findings from the Kenya ESCCAPE case-control study and a pan-African perspective. *International Journal of Cancer* (2019). doi: http://dx.doi.org/10.1002/ijc.32086. PubMed PMID: 625915031.

98. Chen X, Yuan Z, Lu M, Zhang Y, Jin L, Ye W. Poor oral health is associated with an increased risk of esophageal squamous cell carcinoma - a population-based case-control study in China. *Int J Cancer* (2017) 140(3):626-35. Epub 2016/10/26. doi: 10.1002/ijc.30484. PubMed PMID: 27778330.

99. Abnet CC, Kamangar F, Islami F, Nasrollahzadeh D, Brennan P, Aghcheli K, et al. Tooth loss and lack of regular oral hygiene are associated with higher risk of esophageal squamous cell carcinoma. *Cancer Epidemiol Biomarkers Prev* (2008) 17(11):3062-8. Epub 2008/11/08. doi: 10.1158/1055-9965.epi-08-0558. PubMed PMID: 18990747; PubMed Central PMCID: PMCPMC2586052.

100. Hiraki A, Matsuo K, Suzuki T, Kawase T, Tajima K. Teeth loss and risk of cancer at 14 common sites in Japanese. *Cancer Epidemiol Biomarkers Prev* (2008) 17(5):1222-7. Epub 2008/05/17. doi: 10.1158/1055-9965.epi-07-2761. PubMed PMID: 18483345.

101. Guha N, Boffetta P, Wunsch Filho V, Eluf Neto J, Shangina O, Zaridze D, et al. Oral health and risk of squamous cell carcinoma of the head and neck and esophagus: results of two multicentric case-control studies. *Am J Epidemiol* (2007) 166(10):1159-73. Epub 2007/09/01. doi: 10.1093/aje/kwm193. PubMed PMID: 17761691.

102. Michaud DS, Liu Y, Meyer M, Giovannucci E, Joshipura K. Periodontal disease, tooth loss, and cancer risk in male health professionals: a prospective cohort study. *Lancet Oncol* (2008) 9(6):550-8. Epub 2008/05/09. doi: 10.1016/s1470-2045(08)70106-2. PubMed PMID: 18462995; PubMed Central PMCID: PMCPMC2601530.

103. Zhang S, Yu P, Wang JB, Fan JH, Qiao YL, Taylor PR. Association between tooth loss and upper gastrointestinal cancer: A 30-year follow-up of the Linxian Dysplasia Nutrition Intervention Trial Cohort. *Thorac Cancer* (2019) 10(4):966-74. Epub 2019/03/19. doi: 10.1111/1759-7714.13037. PubMed PMID: 30883021; PubMed Central PMCID: PMCPMC6449253.

104. Kigen G, Busakhala N, Kamuren Z, Rono H, Kimalat W, Njiru E. Factors associated with the high prevalence of oesophageal cancer in Western Kenya: a

review. *Infect Agent Cancer* (2017) 12:59. Epub 2017/11/17. doi: 10.1186/s13027-017-0169-y. PubMed PMID: 29142587; PubMed Central PMCID: PMCPMC5670732.

105. Alizadeh AM, Rohandel G, Roudbarmohammadi S, Roudbary M, Sohanaki H, Ghiasian SA, et al. Fumonisin B1 contamination of cereals and risk of esophageal cancer in a high risk area in northeastern Iran. *Asian Pac J Cancer Prev* (2012) 13(6):2625-8. Epub 2012/09/04. doi: 10.7314/apjcp.2012.13.6.2625. PubMed PMID: 22938431.

106. Ghasemi-Kebria F, Joshaghani H, Taheri NS, Semnani S, Aarabi M, Salamat F, et al. Aflatoxin contamination of wheat flour and the risk of esophageal cancer in a high risk area in Iran. *Cancer epidemiology* (2013) 37(3):290-3. Epub 2013/02/26. doi: 10.1016/j.canep.2013.01.010. PubMed PMID: 23434312.

107. Kamangar F, Chow WH, Abnet CC, Dawsey SM. Environmental causes of esophageal cancer. *Gastroenterol Clin North Am* (2009) 38(1):27-57, vii. Epub 2009/03/31. doi: 10.1016/j.gtc.2009.01.004. PubMed PMID: 19327566; PubMed Central PMCID: PMCPMC2685172.

108. Marasas WF, van Rensburg SJ, Mirocha CJ. Incidence of Fusarium species and the mycotoxins, deoxynivalenol and zearalenone, in corn produced in esophageal cancer areas in Transkei. *J Agric Food Chem* (1979) 27(5):1108-12. Epub 1979/09/01. doi: 10.1021/jf60225a013. PubMed PMID: 161914.

109. van Rensburg SJ, Benade AS, Rose EF, du Plessis JP. Nutritional status of African populations predisposed to esophageal cancer. *Nutrition and cancer* (1983) 4(3):206-16. Epub 1983/01/01. doi: 10.1080/01635588209513759. PubMed PMID: 6844145.

110. Xue KS, Tang L, Sun G, Wang S, Hu X, Wang JS. Mycotoxin exposure is associated with increased risk of esophageal squamous cell carcinoma in Huaian area, China. *BMC Cancer* (2019) 19(1):1218. Epub 2019/12/18. doi: 10.1186/s12885-019-6439-x. PubMed PMID: 31842816; PubMed Central PMCID: PMCPMC6916103.

111. Yoshizawa T, Yamashita A, Luo Y. Fumonisin occurrence in corn from high- and low-risk areas for human esophageal cancer in China. *Appl Environ Microbiol* (1994) 60(5):1626-9. Epub 1994/05/01. PubMed PMID: 8017941; PubMed Central PMCID: PMCPMC201527.

112. Dlamini Z, Bhoola K. Esophageal cancer in African blacks of Kwazulu Natal, South Africa: an epidemiological brief. *Ethn Dis* (2005) 15(4):786-9. Epub 2005/11/02. PubMed PMID: 16259509.

113. Dlamini Z, Mbita Z, Skhosana L. Maize beer carcinogenesis: Molecular implications of fumonisins, aflatoxins and prostaglandins. *Beer in Health and Disease Prevention*. (2008). p. 651-6.

114. Abnet CC, Borkowf CB, Qiao YL, Albert PS, Wang E, Merrill AH, Jr., et al. Sphingolipids as biomarkers of fumonisin exposure and risk of esophageal

squamous cell carcinoma in china. *Cancer Causes Control* (2001) 12(9):821-8. Epub 2001/11/21. doi: 10.1023/a:1012228000014. PubMed PMID: 11714110.

115. Cooper S, Trudgill N. Subjects with prostate cancer are less likely to develop esophageal cancer: analysis of SEER 9 registries database. *Cancer Causes & Control* (2012) 23(6):819-25.

116. Dietzsch E, Laubscher R, Parker MI. Esophageal cancer risk in relation to GGC and CAG trinucleotide repeat lengths in the androgen receptor gene. *International journal of cancer* (2003) 107(1):38-45.

117. Utsumi Y, Nakamura T, Nagasue N, Kubota H, Morikawa S. Role of estrogen receptors in the growth of human esophageal carcinoma. *Cancer* (1989) 64(1):88-93.

118. Islami F, Cao Y, Kamangar F, Nasrollahzadeh D, Marjani H-A, Shakeri R, et al. Reproductive factors and risk of esophageal squamous cell carcinoma in northern Iran-A case-control study in a high risk area and literature review. *European journal of cancer prevention: the official journal of the European Cancer Prevention Organisation (ECP)* (2013) 22(5):461.

119. McCormack VA, Menya D, Munishi MO, Dzamalala C, Gasmelseed N, Leon Roux M, et al. Informing etiologic research priorities for squamous cell esophageal cancer in Africa: A review of setting-specific exposures to known and putative risk factors. *International Journal of Cancer* (2017) 140(2):259-71. doi: 10.1002/ijc.30292.

120. Huang F-L, Yu S-J. Esophageal cancer: Risk factors, genetic association, and treatment. *Asian Journal of Surgery* (2016).

121. Coleman HG, Xie S-H, Lagergren J. The Epidemiology of Esophageal Adenocarcinoma. *Gastroenterology* (2017).

122. Li MM, Datto M, Duncavage EJ, Kulkarni S, Lindeman NI, Roy S, et al. Standards and guidelines for the interpretation and reporting of sequence variants in cancer: a joint consensus recommendation of the Association for Molecular Pathology, American Society of Clinical Oncology, and College of American Pathologists. *The Journal of molecular diagnostics* (2017) 19(1):4-23.

123. Richards S, Aziz N, Bale S, Bick D, Das S, Gastier-Foster J, et al. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet Med* (2015) 17(5):405-24. Epub 2015/03/06. doi: 10.1038/gim.2015.30. PubMed PMID: 25741868; PubMed Central PMCID: PMCPMC4544753.

124. Chang J, Tan W, Ling Z, Xi R, Shao M, Chen M, et al. Genomic analysis of oesophageal squamous-cell carcinoma identifies alcohol drinking-related mutation signature and genomic alterations. *Nature communications* (2017) 8:15290.

125. National Human Genome Research Institute–European Bioinformatics Institute (NHGRI-EBI). GWAS Diversity Monitor (2020) [updated 2020/03/01; cited 2020 14 September]. Available from: https://gwasdiversitymonitor.com/.

126. Mills MC, Rahal C. The GWAS Diversity Monitor tracks diversity by disease in real time. *Nature Genetics* (2020) 52(3):242-3. doi: 10.1038/s41588-020-0580-y.

127. Liu W, Snell JM, Jeck WR, Hoadley KA, Wilkerson MD, Parker JS, et al. Subtyping sub-Saharan esophageal squamous cell carcinoma by comprehensive molecular analysis. *JCI insight* (2016) 1(16):e88755. doi: 10.1172/jci.insight.88755. PubMed PMID: PMC5053149.

128. Bye H, Prescott NJ, Lewis CM, Matejcic M, Moodley L, Robertson B, et al. Distinct genetic association at the PLCE1 locus with oesophageal squamous cell carcinoma in the South African population. *Carcinogenesis* (2012) 33(11):2155-61.

129. Bye H, Prescott NJ, Matejcic M, Rose E, Lewis CM, Parker MI, et al. Population-specific genetic associations with oesophageal squamous cell carcinoma in South Africa. *Carcinogenesis* (2011) 32(12):1855-61.

130. Chen WC, Bye H, Matejcic M, Kerr R, Singh E, Prescott NJ, et al. *Abstract A34: The genetic etiology of esophageal cancer in South African Black populations*. AACR (2017).

131. Mathew CG. Abstract IA8: The genetics and genomics of African esophageal cancer. *Cancer Research* (2017) 77(22 Supplement):IA8-IA. doi: 10.1158/1538-7445.newfront17-ia8.

132. Mao N, Nie S, Hong B, Li C, Shen X, Xiong T. Association between alcohol dehydrogenase-2 gene polymorphism and esophageal cancer risk: a meta-analysis. *World J Surg Oncol* (2016) 14(1):191. Epub 2016/07/28. doi: 10.1186/s12957-016-0937-y. PubMed PMID: 27450204; PubMed Central PMCID: PMCPMC4957421.

133. Gong FF, Lu SS, Hu CY, Qian ZZ, Feng F, Wu YL, et al. Cytochrome P450 1A1 (CYP1A1) polymorphism and susceptibility to esophageal cancer: an updated meta-analysis of 27 studies. *Tumour Biol* (2014) 35(10):10351-61. Epub 2014/07/23. doi: 10.1007/s13277-014-2341-y. PubMed PMID: 25048966.

134. Ding DP, Ma WL, He XF, Zhang Y. XPD Lys751Gln polymorphism and esophageal cancer susceptibility: a meta-analysis of case-control studies. *Molecular biology reports* (2012) 39(3):2533-40. Epub 2011/06/15. doi: 10.1007/s11033-011-1005-x. PubMed PMID: 21667112.

135. Zhu ML, He J, Wang M, Sun MH, Jin L, Wang X, et al. Potentially functional polymorphisms in the ERCC2 gene and risk of esophageal squamous cell carcinoma in Chinese populations. *Sci Rep* (2014) 4:6281. Epub 2014/09/12. doi: 10.1038/srep06281. PubMed PMID: 25209371; PubMed Central PMCID: PMCPMC4160711.

136. Ge Y, Jiang R, Zhang M, Wang H, Zhang L, Tang J, et al. Analyzing 37,900 samples shows significant association between HOTAIR polymorphisms and cancer

susceptibility: a meta-analysis. *Int J Biol Markers* (2017) 32(2):e231-e42. Epub 2016/10/30. doi: 10.5301/jbm.5000235. PubMed PMID: 27791260.

137. Li X, Ren D, Li Y, Xu J, Liu C, Zhao Y. Increased cancer risk associated with the -607C/A polymorphism in interleukin-18 gene promoter: an updated meta-analysis including 12,502 subjects. *J buon* (2015) 20(3):902-17. Epub 2015/07/28. PubMed PMID: 26214646.

138. Li X, Qu L, Zhong Y, Zhao Y, Chen H, Daru L. Association between promoters polymorphisms of matrix metalloproteinases and risk of digestive cancers: a meta-analysis. *J Cancer Res Clin Oncol* (2013) 139(9):1433-47. Epub 2013/05/07. doi: 10.1007/s00432-013-1446-9. PubMed PMID: 23644699.

139. Sun GG, Wang YD, Lu YF, Hu WN. Different association of manganese superoxide dismutase gene polymorphisms with risk of prostate, esophageal, and lung cancers: evidence from a meta-analysis of 20,025 subjects. *Asian Pac J Cancer Prev* (2013) 14(3):1937-43. Epub 2013/05/18. doi: 10.7314/apjcp.2013.14.3.1937. PubMed PMID: 23679296.

140. Moody S, Senkin S, Islam SA, Wang J, Nasrollahzadeh D, Penha RCC, et al. Mutational signatures in esophageal squamous cell carcinoma from eight countries of varying incidence. *medRxiv* (2021).

141. Esteller M. Epigenetics in cancer. *New England Journal of Medicine* (2008) 358(11):1148-59.

142. Kit OI, Vodolazhskiy DI, Kolesnikov EN, Timoshkina NN. Epigenetic markers of esophageal cancer: DNA methylation. *Biochemistry (Moscow), Supplement Series B: Biomedical Chemistry* (2017) 11(1):55-61. doi: 10.1134/s1990750817010048.

143. Guo M, Ren J, House MG, Qi Y, Brock MV, Herman JG. Accumulation of promoter methylation suggests epigenetic progression in squamous cell carcinoma of the esophagus. *Clin Cancer Res* (2006) 12(15):4515-22. Epub 2006/08/11. doi: 10.1158/1078-0432.ccr-05-2858. PubMed PMID: 16899597.

144. Salam I, Hussain S, Mir MM, Dar NA, Abdullah S, Siddiqi MA, et al. Aberrant promoter methylation and reduced expression of p16 gene in esophageal squamous cell carcinoma from Kashmir valley: a high-risk area. *Mol Cell Biochem* (2009) 332(1-2):51-8. Epub 2009/06/11. doi: 10.1007/s11010-009-0173-7. PubMed PMID: 19513816.

145. Taghavi N, Biramijamal F, Sotoudeh M, Khademi H, Malekzadeh R, Moaven O, et al. p16INK4a hypermethylation and p53, p16 and MDM2 protein expression in esophageal squamous cell carcinoma. *BMC Cancer* (2010) 10:138. Epub 2010/04/15. doi: 10.1186/1471-2407-10-138. PubMed PMID: 20388212; PubMed Central PMCID: PMCPMC2868052.

146. Hibi K, Taguchi M, Nakayama H, Takase T, Kasai Y, Ito K, et al. Molecular detection of p16 promoter methylation in the serum of patients with esophageal squamous cell carcinoma. *Clin Cancer Res* (2001) 7(10):3135-8. Epub 2001/10/12. PubMed PMID: 11595706.

147. Xing EP, Nie Y, Song Y, Yang GY, Cai YC, Wang LD, et al. Mechanisms of inactivation of p14ARF, p15INK4b, and p16INK4a genes in human esophageal squamous cell carcinoma. *Clinical cancer research : an official journal of the American Association for Cancer Research* (1999) 5(10):2704-13. Epub 1999/10/28. PubMed PMID: 10537333.

148. Zhang L, Lu W, Miao X, Xing D, Tan W, Lin D. Inactivation of DNA repair gene O6-methylguanine-DNA methyltransferase by promoter hypermethylation and its relation to p53 mutations in esophageal squamous cell carcinoma. *Carcinogenesis* (2003) 24(6):1039-44. Epub 2003/06/17. doi: 10.1093/carcin/bgg062. PubMed PMID: 12807758.

149. Kawakami K, Brabender J, Lord RV, Groshen S, Greenwald BD, Krasna MJ, et al. Hypermethylated APC DNA in plasma and prognosis of patients with esophageal adenocarcinoma. *J Natl Cancer Inst* (2000) 92(22):1805-11. Epub 2000/11/18. doi: 10.1093/jnci/92.22.1805. PubMed PMID: 11078757.

150. Anupam K, Tusharkant C, Gupta SD, Ranju R. Loss of disabled-2 expression is an early event in esophageal squamous tumorigenesis. *World J Gastroenterol* (2006) 12(37):6041-5. Epub 2006/09/30. doi: 10.3748/wjg.v12.i37.6041. PubMed PMID: 17009406; PubMed Central PMCID: PMCPMC4124415.

151. Guo M, Ren J, Brock MV, Herman JG, Carraway HE. Promoter methylation of HIN-1 in the progression to esophageal squamous cancer. *Epigenetics* (2008) 3(6):336-41. Epub 2008/12/23. doi: 10.4161/epi.3.6.7158. PubMed PMID: 19098448.

152. Wang Y, Fang MZ, Liao J, Yang GY, Nie Y, Song Y, et al. Hypermethylation-associated inactivation of retinoic acid receptor beta in human esophageal squamous cell carcinoma. *Clin Cancer Res* (2003) 9(14):5257-63. Epub 2003/11/14. PubMed PMID: 14614007.

153. Kuroki T, Trapasso F, Yendamuri S, Matsuyama A, Alder H, Mori M, et al. Allele loss and promoter hypermethylation of VHL, RAR-beta, RASSF1A, and FHIT tumor suppressor genes on chromosome 3p in esophageal squamous cell carcinoma. *Cancer research* (2003) 63(13):3724-8. Epub 2003/07/04. PubMed PMID: 12839965.

154. Yue CM, Deng DJ, Bi MX, Guo LP, Lu SH. Expression of ECRG4, a novel esophageal cancer-related gene, downregulated by CpG island hypermethylation in human esophageal squamous cell carcinoma. *World J Gastroenterol* (2003) 9(6):1174-8. Epub 2003/06/12. doi: 10.3748/wjg.v9.i6.1174. PubMed PMID: 12800218; PubMed Central PMCID: PMCPMC4611778.

155. Noguchi T, Takeno S, Kimura Y, Uchida Y, Daa T, Yokoyama S, et al. FHIT expression and hypermethylation in esophageal squamous cell carcinoma. *Int J Mol Med* (2003) 11(4):441-7. Epub 2003/03/13. PubMed PMID: 12632095.

156. Ohta M, Mimori K, Fukuyoshi Y, Kita Y, Motoyama K, Yamashita K, et al. Clinical significance of the reduced expression of G protein gamma 7 (GNG7) in oesophageal cancer. *Br J Cancer* (2008) 98(2):410-7. Epub 2008/01/26. doi:

10.1038/sj.bjc.6604124. PubMed PMID: 18219292; PubMed Central PMCID: PMCPMC2361448.

157. Zhao BJ, Tan SN, Cui Y, Sun DG, Ma X. Aberrant promoter methylation of the TPEF gene in esophageal squamous cell carcinoma. *Dis Esophagus* (2008) 21(7):582-8. Epub 2008/12/02. doi: 10.1111/j.1442-2050.2007.00808.x. PubMed PMID: 19040536.

158. Ito T, Shimada Y, Hashimoto Y, Kaganoi J, Kan T, Watanabe G, et al. Involvement of TSLC1 in progression of esophageal squamous cell carcinoma. *Cancer Res* (2003) 63(19):6320-6. Epub 2003/10/16. PubMed PMID: 14559819.

159. Mandelker DL, Yamashita K, Tokumaru Y, Mimori K, Howard DL, Tanaka Y, et al. PGP9.5 promoter methylation is an independent prognostic factor for esophageal squamous cell carcinoma. *Cancer Res* (2005) 65(11):4963-8. Epub 2005/06/03. doi: 10.1158/0008-5472.can-04-3923. PubMed PMID: 15930319.

160. Hamilton JP, Sato F, Jin Z, Greenwald BD, Ito T, Mori Y, et al. Reprimo methylation is a potential biomarker of Barrett's-Associated esophageal neoplastic progression. *Clin Cancer Res* (2006) 12(22):6637-42. Epub 2006/11/24. doi: 10.1158/1078-0432.ccr-06-1781. PubMed PMID: 17121882.

161. Jin Z, Mori Y, Hamilton JP, Olaru A, Sato F, Yang J, et al. Hypermethylation of the somatostatin promoter is a common, early event in human esophageal carcinogenesis. *Cancer* (2008) 112(1):43-9. Epub 2007/11/14. doi: 10.1002/cncr.23135. PubMed PMID: 17999418.

162. Jin Z, Cheng Y, Olaru A, Kan T, Yang J, Paun B, et al. Promoter hypermethylation of CDH13 is a common, early event in human esophageal adenocarcinogenesis and correlates with clinical risk factors. *Int J Cancer* (2008) 123(10):2331-6. Epub 2008/08/30. doi: 10.1002/ijc.23804. PubMed PMID: 18729198.

163. Jin Z, Olaru A, Yang J, Sato F, Cheng Y, Kan T, et al. Hypermethylation of tachykinin-1 is a potential biomarker in human esophageal cancer. *Clin Cancer Res* (2007) 13(21):6293-300. Epub 2007/11/03. doi: 10.1158/1078-0432.ccr-07-0818. PubMed PMID: 17975140.

164. Jin Z, Mori Y, Yang J, Sato F, Ito T, Cheng Y, et al. Hypermethylation of the nel-like 1 gene is a common and early event and is associated with poor prognosis in early-stage esophageal adenocarcinoma. *Oncogene* (2007) 26(43):6332-40. Epub 2007/04/25. doi: 10.1038/sj.onc.1210461. PubMed PMID: 17452981.

165. Crick F. Central dogma of molecular biology. *Nature* (1970) 227(5258):561-3.

166. Zhang C, Sun Q. Weighted gene co-expression network analysis of gene modules for the prognosis of esophageal cancer. *Journal of Huazhong University of Science and Technology [Medical Sciences]* (2017) 37(3):319-25. doi: 10.1007/s11596-017-1734-8.

167. Kim SM, Park YY, Park ES, Cho JY, Izzo JG, Zhang D, et al. Prognostic biomarkers for esophageal adenocarcinoma identified by analysis of tumor transcriptome. *PLoS One* (2010) 5(11):e15074. Epub 2010/12/15. doi: 10.1371/journal.pone.0015074. PubMed PMID: 21152079; PubMed Central PMCID: PMCPMC2994829.

168. Lee JJ, Natsuizaka M, Ohashi S, Wong GS, Takaoka M, Michaylira CZ, et al. Hypoxia activates the cyclooxygenase-2-prostaglandin E synthase axis. *Carcinogenesis* (2010) 31(3):427-34. Epub 2010/01/01. doi: 10.1093/carcin/bgp326. PubMed PMID: 20042640; PubMed Central PMCID: PMCPMC2832548.

169. Nicolau-Neto P, Da Costa NM, de Souza Santos PT, Gonzaga IM, Ferreira MA, Guaraldi S, et al. Esophageal squamous cell carcinoma transcriptome reveals the effect of FOXM1 on patient outcome through novel PIK3R3 mediated activation of PI3K signaling pathway. *Oncotarget* (2018) 9(24):16634-47. Epub 2018/04/24. doi: 10.18632/oncotarget.24621. PubMed PMID: 29682174; PubMed Central PMCID: PMCPMC5908275.

170. Peng D, Guo Y, Chen H, Zhao S, Washington K, Hu T, et al. Integrated molecular analysis reveals complex interactions between genomic and epigenomic alterations in esophageal adenocarcinomas. *Sci Rep* (2017) 7:40729. Epub 2017/01/20. doi: 10.1038/srep40729. PubMed PMID: 28102292; PubMed Central PMCID: PMCPMC5244375.

171. Hou X, Wen J, Ren Z, Zhang G. Non-coding RNAs: new biomarkers and therapeutic targets for esophageal cancer. *Oncotarget* (2017) 8(26):43571-8. Epub 2017/04/08. doi: 10.18632/oncotarget.16721. PubMed PMID: 28388588; PubMed Central PMCID: PMCPMC5522170.

172. Liu R, Gu J, Jiang P, Zheng Y, Liu X, Jiang X, et al. DNMT1-microRNA126 epigenetic circuit contributes to esophageal squamous cell carcinoma growth via ADAM9-EGFR-AKT signaling. *Clinical cancer research : an official journal of the American Association for Cancer Research* (2015) 21(4):854-63. Epub 2014/12/17. doi: 10.1158/1078-0432.Ccr-14-1740. PubMed PMID: 25512445.

173. Suzuki H, Maruyama R, Yamamoto E, Kai M. DNA methylation and microRNA dysregulation in cancer. *Molecular oncology* (2012) 6(6):567-78. Epub 2012/08/10. doi: 10.1016/j.molonc.2012.07.007. PubMed PMID: 22902148.

174. Alaouna M, Hull R, Penny C, Dlamini Z. Esophageal cancer genetics in South Africa. *Clinical and experimental gastroenterology* (2019) 12:157-77. doi: 10.2147/CEG.S182000. PubMed PMID: 31114287.

175. Matejcic M, Iqbal Parker M. Gene–environment interactions in esophageal cancer. *Critical Reviews in Clinical Laboratory Sciences* (2015) 52(5):211-31. doi: 10.3109/10408363.2015.1020358.

176. Fletcher O, Dudbridge F. Candidate gene-environment interactions in breast cancer. *BMC medicine* (2014) 12(1):195.

177. Zhang L, Jiang Y, Wu Q, Li Q, Chen D, Xu L, et al. Gene–environment interactions on the risk of esophageal cancer among Asian populations with the G48A polymorphism in the alcohol dehydrogenase-2 gene: a meta-analysis. *Tumor Biology* (2014) 35(5):4705-17.

178. Bhat GA, Bhat AB, Lone MM, Dar NA. Association of genetic variants of CYP2C19 and CYP2D6 with esophageal squamous cell carcinoma risk in Northern India, Kashmir. *Nutrition and cancer* (2017) 69(4):585-92.

179. Peng X-E, Chen H-F, Hu Z-J, Shi X-S. Independent and combined effects of environmental factors and CYP2C19 polymorphisms on the risk of esophageal squamous cell carcinoma in Fujian Province of China. *BMC medical genetics* (2015) 16(1):15.

180. Nieminen MT, Salaspuro M. Local Acetaldehyde—An Essential Role in Alcohol-Related Upper Gastrointestinal Tract Carcinogenesis. *Cancers* (2018) 10(1):11.

181. Salaspuro M. Key role of local acetaldehyde in upper GI tract carcinogenesis. *Best Practice & Research Clinical Gastroenterology* (2017).

182. Brooks PJ, Zakhari S. Acetaldehyde and the genome: beyond nuclear DNA adducts and carcinogenesis. *Environmental and molecular mutagenesis* (2014) 55(2):77-91.

183. Ceppi M, Munnia A, Cellai F, Bruzzone M, Peluso ME. Linking the generation of DNA adducts to lung cancer. *Toxicology* (2017) 390:160-6.

184. Yun BH, Xiao S, Yao L, Krishnamachari S, Rosenquist TA, Dickman KG, et al. A Rapid Throughput Method To Extract DNA from Formalin-Fixed Paraffin-Embedded Tissues for Biomonitoring Carcinogenic DNA Adducts. *Chemical research in toxicology* (2017) 30(12):2130-9.

185. Blumenthal U, Fleisher J, Esrey SA, Peasey A. Epidemiology: a tool for the assessment of risk. *WHO Water Quality Guidelines, Standards and Health: Assessment of risk and risk management for water-related infectious disease* (2001):135.

186. Nieuwenhuijsen MJ. Design of exposure questionnaires for epidemiological studies. *Occupational and Environmental Medicine* (2005) 62(4):272-80. doi: 10.1136/oem.2004.015206.

187. Ganzleben C, Antignac JP, Barouki R, Castano A, Fiddicke U, Klanova J, et al. Human biomonitoring as a tool to support chemicals regulation in the European Union. *Int J Hyg Environ Health* (2017) 220(2 Pt A):94-7. Epub 2017/03/13. doi: 10.1016/j.ijheh.2017.01.007. PubMed PMID: 28284775.

188. EU. 1386/2013/EU of the European Parliament and of the Council of 20 November 2013 on a General Union Environment Action Programme to 2020 'Living well, within the limits of our planet'. (2013) Contract No.: 171.

189. APHL. Guidance for laboratory biomonitoring programs. (2012).

190. Crinnion WJ. The CDC Fourth National Report on Human Exposure to Environmental Chemicals: What it Tells Us About our Toxic Burden and How it Assists Environmental Medicine Physicians. *Alternative medicine review* (2010) 15(2).

191. Ewa B, Danuta M-Š. Polycyclic aromatic hydrocarbons and PAH-related DNA adducts. *Journal of applied genetics* (2017) 58(3):321-30.

192. Artiola JF, Warrick AW. Sampling and data quality objectives for environmental monitoring. *Environmental Monitoring and Characterization*. Elsevier Inc. (2004). p. 11-27.

193. Beiras R. Chapter 16 - Biological Tools for Monitoring: Biomarkers and Bioassays. In: Beiras R, editor. *Marine Pollution*. Elsevier (2018). p. 265-91.

194. Nair DV, Reddy AG. Laboratory animal models for esophageal cancer. *Vet World* (2016) 9(11):1229-32. Epub 2016/12/14. doi: 10.14202/vetworld.2016.1229-1232. PubMed PMID: 27956773; PubMed Central PMCID: PMCPMC5146302.

195. Tajaldini M, Samadi F, Khosravi A, Ghasemnejad A, Asadi J. Protective and anticancer effects of orange peel extract and naringin in doxorubicin treated esophageal cancer stem cell xenograft tumor mouse model. *Biomed Pharmacother* (2020) 121:109594. Epub 2019/11/11. doi: 10.1016/j.biopha.2019.109594. PubMed PMID: 31707344.

196. Lu S, Wang TD. In vivo cancer biomarkers of esophageal neoplasia. *Cancer Biomark* (2008) 4(6):341-50. Epub 2009/01/08. doi: 10.3233/cbm-2008-4606. PubMed PMID: 19126962; PubMed Central PMCID: PMCPMC3226753.

197. Peterson MK, Mohar I, Lam T, Cook TJ, Engel AM, Lynch H. Critical review of the evidence for a causal association between exposure to asbestos and esophageal cancer. *Crit Rev Toxicol* (2019) 49(7):597-613. Epub 2020/01/23. doi: 10.1080/10408444.2019.1692190. PubMed PMID: 31965908.

198. Hao J, Liu B, Yang CS, Chen X. Gastroesophageal reflux leads to esophageal cancer in a surgical model with mice. *BMC Gastroenterol* (2009) 9:59. Epub 2009/07/25. doi: 10.1186/1471-230x-9-59. PubMed PMID: 19627616; PubMed Central PMCID: PMCPMC2723127.

199. Lee NP, Chan CM, Tung LN, Wang HK, Law S. Tumor xenograft animal models for esophageal squamous cell carcinoma. *J Biomed Sci* (2018) 25(1):66. Epub 2018/08/31. doi: 10.1186/s12929-018-0468-7. PubMed PMID: 30157855; PubMed Central PMCID: PMCPMC6116446.

200. Chen MF, Chen PT, Lu MS, Chen WC. Role of ALDH1 in the prognosis of esophageal cancer and its relationship with tumor microenvironment. *Molecular carcinogenesis* (2018) 57(1):78-88. Epub 2017/09/10. doi: 10.1002/mc.22733. PubMed PMID: 28888039.

201. Hoffjan S. SI: Next-generation sequencing in human molecular genetic diagnostics. *Mol Cell Probes* (2019) 45:69. Epub 2019/05/19. doi: 10.1016/j.mcp.2019.05.005. PubMed PMID: 31102645.

202. Mehta NAL, Battram AM. DNA sequencing technologies and emerging applications in drug discovery [Article]. (2011) [updated 2020; cited 2020 6 May]. Available from: https://www.europeanpharmaceuticalreview.com/article/10409/dna-sequencing-technologies-and-emerging-applications-in-drug-discovery/.

203. Gasperskaja E, Kučinskas V. The most common technologies and tools for functional genome analysis. *Acta medica Lituanica* (2017) 24(1):1-11. doi: 10.6001/actamedica.v24i1.3457. PubMed PMID: 28630587.

204. Ranjan R, Rani A, Metwally A, McGee HS, Perkins DL. Analysis of the microbiome: Advantages of whole genome shotgun versus 16S amplicon sequencing. *Biochem Biophys Res Commun* (2016) 469(4):967-77. Epub 2016/01/01. doi: 10.1016/j.bbrc.2015.12.083. PubMed PMID: 26718401; PubMed Central PMCID: PMCPMC4830092.

205. Lee R. Metagenomic Next Generation Sequencing: How Does It Work and Is It Coming to Your Clinical Microbiology Lab? US(2019) [updated 2019; cited 2020 3 October 2020]. Available from: https://asm.org/Articles/2019/November/Metagenomic-Next-Generation-Sequencing-How-Does-It#:~:text=Metagenomic%20NGS%20(mNGS)%20is%20simply,present%20and%20in%20what%20proportions.

206. Tam V, Patel N, Turcotte M, Bossé Y, Paré G, Meyre D. Benefits and limitations of genome-wide association studies. *Nature Reviews Genetics* (2019) 20(8):467-84. doi: 10.1038/s41576-019-0127-1.

207. Sun Y, Miao N, Sun T. Detect accessible chromatin using ATAC-sequencing, from principle to applications. *Hereditas* (2019) 156:29-. doi: 10.1186/s41065-019-0105-9. PubMed PMID: 31427911.

208. Lo P-K, Zhou Q. Emerging techniques in single-cell epigenomics and their applications to cancer research. *Journal of clinical genomics* (2018) 1(1):10.4172/JCG.1000103. Epub 2018/03/05. doi: 10.4172/JCG.1000103. PubMed PMID: 30079405.

209. Buermans HPJ, den Dunnen JT. Next generation sequencing technology: Advances and applications. *Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease* (2014) 1842(10):1932-41. doi: https://doi.org/10.1016/j.bbadis.2014.06.015.

210. van Dijk EL, Jaszczyszyn Y, Naquin D, Thermes C. The Third Revolution in Sequencing Technology. *Trends in Genetics* (2018) 34(9):666-81. doi: https://doi.org/10.1016/j.tig.2018.05.008.

211. Cieslik M, Chinnaiyan AM. Global genomics project unravels cancer's complexity at unprecedented scale. *Nature* (2020) 578(7793):39-40. Epub 2020/02/07. doi: 10.1038/d41586-020-00213-2. PubMed PMID: 32025004.

212. Campbell PJ, Getz G, Korbel JO, Stuart JM, Jennings JL, Stein LD, et al. Pan-cancer analysis of whole genomes. *Nature* (2020) 578(7793):82-93. doi: 10.1038/s41586-020-1969-6.

213. Rheinbay E, Nielsen MM, Abascal F, Wala JA, Shapira O, Tiao G, et al. Analyses of non-coding somatic drivers in 2,658 cancer whole genomes. *Nature* (2020) 578(7793):102-11. doi: 10.1038/s41586-020-1965-x.

214. Li Y, Roberts ND, Wala JA, Shapira O, Schumacher SE, Kumar K, et al. Patterns of somatic structural variation in human cancer genomes. *Nature* (2020) 578(7793):112-21. doi: 10.1038/s41586-019-1913-9.

215. Alexandrov LB, Kim J, Haradhvala NJ, Huang MN, Tian Ng AW, Wu Y, et al. The repertoire of mutational signatures in human cancer. *Nature* (2020) 578(7793):94-101. doi: 10.1038/s41586-020-1943-3.

216. Gerstung M, Jolly C, Leshchiner I, Dentro SC, Gonzalez S, Rosebrock D, et al. The evolutionary history of 2,658 cancers. *Nature* (2020) 578(7793):122-8. doi: 10.1038/s41586-019-1907-7.

217. Calabrese C, Davidson NR, Demircioğlu D, Fonseca NA, He Y, Kahles A, et al. Genomic basis for RNA alterations in cancer. *Nature* (2020) 578(7793):129-36. doi: 10.1038/s41586-020-1970-0.

218. Mu W, Li B, Wu S, Chen J, Sain D, Xu D, et al. Detection of structural variation using target captured next-generation sequencing data for genetic diagnostic testing. *Genetics in Medicine* (2019) 21(7):1603-10. doi: 10.1038/s41436-018-0397-6.

219. Teng S. NGS for Sequence Variants. *Adv Exp Med Biol* (2016) 939:1-20. Epub 2016/11/04. doi: 10.1007/978-981-10-1503-8_1. PubMed PMID: 27807741.

220. Chen WC, Bye H, Matejcic M, Kerr R, Singh E, Prescott NJ, et al. Abstract A34: The genetic etiology of esophageal cancer in South African Black populations. *Cancer Research* (2017) 77(22 Supplement):A34-A. doi: 10.1158/1538-7445.newfront17-a34.

221. Peters BA, Wu J, Pei Z, Yang L, Purdue MP, Freedman ND, et al. *The oral microbiome and prospective risk for esophageal cancer: A population-based nested case-control study*. AACR (2017).

Chapter 2: Systematic Review of Genetic Factors in the Etiology of Esophageal Squamous Cell Carcinoma in African Populations

Publication: Systematic Review of Genetic Factors in the Etiology of Esophageal Squamous Cell Carcinoma in African Populations. 2019 Hannah Simba, Helena Kuivaniemi, Vittoria Lutje, Gerard Tromp and Vikash Sewram. Frontiers in Genetics. 10:642; 1-20 https://doi.org/10.3389/fgene.2019.00642

Table of Contents

2.1 Abstract

Background: Esophageal squamous cell carcinoma (ESCC), one of the most aggressive cancers, is endemic in Sub-Saharan Africa, constituting a major health burden. It has the most divergence in cancer incidence globally, with high prevalence reported in East Asia, Southern Europe, and in East and Southern Africa. Its etiology is multifactorial, with lifestyle, environmental and genetic risk factors. Very little is known about the role of genetic factors in ESCC development and progression among African populations. The study aimed to systematically assess the evidence on genetic variants associated with ESCC in African populations.

Methods: We carried out a comprehensive search of all African published studies up to April 2019, using PubMed, Embase, Scopus and African Index Medicus databases. Quality assessment and data extraction were carried out by two investigators. The strength of the associations was measured by odds ratios and 95% confidence intervals.

Results: Twenty-three genetic studies on ESCC in African populations were included in the systematic review. They were carried out on Black and Admixed South African populations, as well as on Malawian, Sudanese and Kenyan populations. Most studies were candidate gene studies and included DNA sequence variants in 58 different genes. Only one study carried out whole exome sequencing of 59 ESCC patients. Sample sizes varied from 18 to 880 cases and 88 to 939 controls. Altogether over 100 variants in 37 genes were part of 17 case-control genetic association studies to identify susceptibility loci for ESCC. In these studies 25 variants in 20 genes were reported to have a statistically significant association. In addition, eight studies investigated changes in cancer tissues and identified somatic alterations in 17 genes and evidence of loss-off-heterozygosity, copy number variation and microsatellite instability. Two genes were assessed for both genetic association and somatic mutation.

Conclusions: Comprehensive large-scale studies on the genetic basis of ESCC are still lacking in Africa. Sample sizes in existing studies are too small to draw definitive conclusions about ESCC etiology. Only a small number of African populations have been analysed, and replication and validation studies are missing. The genetic etiology of ESCC in Africa is, therefore, still poorly defined.

2.2 Introduction

Esophageal cancer is an aggressive and fatal cancer of the digestive tract. It accounts for an estimated 455,800 new cases and 400,200 deaths per year globally, making it the 8th most common cancer in the world (1). The malignant tumours are characterized by two major subtypes; esophageal squamous cell carcinoma (ESCC), which is the more common type and contributes 90%, and esophageal adenocarcinoma (EAC) (2, 3). ESCC presents with poor prognosis and low survival rate (<5%) in low resource settings (1, 4). The asymptomatic development of ESCC results in diagnosis at late stage for patients and is characterized by dysphagia. At this stage, treatment is limited to palliative care.

ESCC is endemic in specific geographic locations worldwide and has the most divergence in cancer incidence globally, with high prevalence reported in East Asia, Southern Europe, as well as in Eastern and Southern Africa (3). This peculiar distribution draws questions on the specificity of certain risk factors to particular populations. The African ESCC corridor, which includes Ethiopia, Rwanda, Burundi, Malawi, Kenya, Uganda, Tanzania and South Africa, is an ESCC hotspot region (5, 6). It has also been reported that in Sub-Saharan Africa ESCC develops in younger patients than in other regions (7).

The etiology of esophageal carcinoma is multifactorial. The risk factors reported worldwide comprise several lifestyle, environmental and genetic factors (8-12). Growing evidence supports the hypothesis that genomic alterations and epigenetic modifications contribute to tumour development (13). ESCC has both an inherited and cellular genetic basis (3, 14). Familial syndromes associated with increased risk of malignancy include tylosis and Fanconi anemia (3). The majority of genetic studies on ESCC have been case-control association studies analysing single-nucleotide polymorphisms (SNPs) in various candidate genes. However, the reproducibility of these studies has been low. Some of the more common SNPs associated with ESCC have been identified in the aldehyde dehydrogenase 2 family gene (*ALDH2)* and an acetaldehyde dehydrogenase gene *(ADH1B)* (3). Variants in these genes have been shown to increase susceptibility to ESCC development, and they are also associated with alcohol consumption (3). Two meta-analyses published in 2018 reported associations between the genes *MTHFR* and *GSTT1* and esophageal cancer development (15, 16). However, the meta-analyses were done on predominantly Asian and Western populations. In recent years, the focus of ESCC research in the Western and Asian countries has shifted from candidate gene studies to genome-wide association studies (GWAS) and whole exome sequencing (WES) to identify variants associated with ESCC. Combined analysis of different study designs has provided a better understanding of ESCC etiology in Asian populations (3, 17). Genes with variants implicated in the development of ESCC in these populations include phospholipase c epsilon 1 *(PLCE1)*, caspase 8 *(CAP8)*, tumour protein 53 *(TP53)*, and human leukocyte antigen *(HLA)* (3).

The genetic etiology of ESCC in Africa is not well understood, since there have been very few studies on ESCC in African populations. This is in part due to the unavailability of adequate research infrastructure. A lack of comprehensive assessment and validation of existing evidence through systematic reviews has also contributed to this knowledge gap. A number of small studies on African populations have yielded varied associations between genetic variants and ESCC. There is, therefore, a need to systematically assess the current evidence in order to map out the contribution of genetic factors in the development of ESCC in African populations using critically appraised data.

The aim of the current systematic review was to assess all genetic (cross-sectional, case-control, and cohort) studies reporting on germline and somatic variants where risk factor estimates were calculated. This was achieved through the following: 1) Critical appraisal of African literature on association of genetic factors to ESCC development; 2) Comprehensive analysis of genetic (germline and somatic) variants in the reported studies; 3) Data synthesis through pooled analysis, if feasible; and 4) Comparison of genetic variants identified in African populations to those reported in other geographic regions.

2.3 Materials and methods

We followed the Preferred Reporting Items for Systematic Reviews and Meta-Analyses guidelines (PRISMA)(18). However, because PRISMA is not a quality assessment tool, other instruments were used to assess quality control.

2.3.1 Data Sources and Search Strategy

We carried out a literature search on all published African ESCC studies up to April 2019. We developed a comprehensive set of search terms subjectively and iteratively. We searched the following electronic bibliographic databases without time or language limits: Medline (PubMed), Embase (OViD), Scopus, African index medicus, and Africa-wide information (EbsCOHost). We also checked the reference lists of potentially relevant articles for additional citations, and used the "related citations" search key in PubMed to identify similar papers.

We checked Medline (PubMed) to identify controlled vocabulary (MeSH) terms related to esophageal cancer, and also identified text keywords based on our knowledge of the field (Table 1). Medline search terms were modified for other electronic databases to conform to their search functions.

Table 2.1. Medline (Pubmed) search strategy to identify published African ESCC literature.

Search Number	Search terms
#1	Search cancer or carcinoma or neoplasm* Field: Title/Abstract
#2	Search Esophageal or oesophageal Field: Title/Abstract
#3	#1 and #2
#4	Search "Esophageal cancer " Field: Title/Abstract
#5	Search "oesophageal cancer " or “oesophageal neoplasm*” Field: Title/Abstract
#6	Search "Esophageal Neoplasms"[Mesh]
#7	Search "Esophageal Neoplasms" Field: Title/Abstract
#8	Search “Esophageal squamous cell carcinoma” or “oesophageal squamous cell carcinoma” or ESCC Field: Title/Abstract
#9	Search ((((#3) OR #4) OR #5) OR #6) OR #7 OR #8
#10	Search "Africa"[Mesh]
#11	Search algeria OR angola OR benin OR botswana OR burkina faso OR burundi OR cameroon OR cape verde OR central african republic OR chad OR comoros OR congo OR "Democratic Republic of Congo" OR DRC OR djibouti OR equatorial guinea OR egypt OR eritrea OR ethiopia OR gabon OR gambia OR ghana OR guinea OR bissau OR ivory coast OR (Côte d' Ivoire) OR jamahiriya OR kenya OR lesotho OR liberia OR Libya OR madagascar OR malawi OR mali OR mauritania OR mauritius OR mayotte OR morocco OR mozambique OR namibia OR niger OR nigeria OR principe OR reunion OR rwanda OR "Sao Tome" OR senegal OR seychelles OR "Sierra Leone" OR somalia OR "South Africa" OR st helena OR sudan OR swaziland OR tanzania OR togo OR tunisia OR uganda OR zaire OR zambia OR zimbabwe OR "Central Africa" OR "West Africa" OR "East Africa" OR "Southern Africa" OR "South Africa" Field: Title/Abstract
#12	Search (#10) or #11
#13	Search (#9) AND #12

Screening for eligible studies was carried out by two authors (HS and HK). First, the two authors read the titles and abstracts independently and then met to finalise an initial list. Full articles of the studies were selected based on the initial screening, read and assessed for inclusion to the systematic review. Figure 1 shows the outline for selection of eligible studies.

2.3.2 Quality Control and Data Extraction

Quality of the methodology used in the published studies was assessed using a quality assessment tool adapted from the STrengthening the REporting of Genetic Association studies (STREGA) statement (18). The quality assessment for genetic association studies to identify ESCC susceptibility loci included reporting on power calculations, detailed population characteristics for cases, description of ESCC diagnosis, screening of cases and controls, reporting a measure of association using odds ratios, adjustment of population stratification, assessment of genotyping error, reporting the Hardy–Weinberg equilibrium, correction for multiple testing, and reporting of NCBI rs numbers for variants (Table S1).

For somatic mutation studies quality assessment included the following: description of ESCC diagnosis, reporting of tissues used [cancerous (Ca) and normal neighboring tissue (NET)], detailed population characteristics, variant classification and type, confirmation of variants identified, reporting of amino acid change and use of pathogenicity scoring (Table 2A2).

Data extraction was carried out by two authors (HS and HK) using data extraction forms. Two separate extraction forms were prepared for the germline (genetic susceptibility) and somatic mutation studies. The data extraction form for the genetic susceptibility studies included the following: description of the population (age, sex, sample size, smoking and alcohol use for cases and controls separately), genotyping method, statistical analysis test, minor allele frequency (MAF), genotype frequency, haplotype frequency, and environmental association frequency. The somatic mutation study extraction form had the same variables excluding gene-environment interaction frequency and haplotype frequency.

The South African Admixed Population is reported as Mixed Ancestry in the tables according to how it was reported in the articles.

2.3.3 Data Analysis

A meta-analysis could not be performed as there were only two SNPs analyzed in more than one study and even those were analysed in only two independent studies. For a meta-analysis to be carried out, SNPs have to be assessed in at least three separate case-control studies. *TP53* in the somatic variant studies was analysed in four separate studies, but two of the studies were had cases only with no controls, and the remaining two assessed different parts of the gene. The results of this systematic review will, therefore, be reported in a descriptive manner.

We were able to find rs-numbers for most of the variants even if the authors of the original studies did not report them, and have included them in the tables of this systematic review. We used the canonical SNP identifier (rs-number) and dbSNP (version 152; April 2019) database at NCBI (https://www.ncbi.nlm.nih.gov/snp/)for this. We also determined the locus positions of the microsatellite markers reported in a study by Naidoo et al. 2005 (19) using the primer-BLAST database at NCBI (https://www-ncbi-nlm-nih-gov.ez.sun.ac.za/tools/primer-blast).

To determine the linkage disequilibrium (LD) measures between the SNPs reported in the same genes, we obtained the imputed data set from the Thousand Genomes project (1000 Genomes Release Phase 3 2013-05-02; ref 1kG), and used bcftools to extract all individuals from African populations not including African Americans, and the 77 SNPs discussed here using all synonyms (alternative rs IDs) for SNPs (20). We obtained a dataset of 504 individuals and 67 SNPs. We computed all pair-wise r^2 using PLINK (v1.09) (21, 22).

2.4 Results

2.4.1 Systematic Review Outline

The selection process for all the included studies is shown in Figure 1. The initial database search identified 2,235 articles. Titles and abstracts of these articles were reviewed and 2,168 studies were removed for not being original genetic studies. The 67 articles that remained were selected for full-text eligibility assessment. This process resulted in the removal of 40 articles: 15 review articles, 18 chromosomal, gene or protein expression studies, four blood group studies, one duplicate and two abstracts. A total of 27 full articles were then assessed for eligibility, and four articles were removed for not meeting the criteria, as follows: one study had no cancer patients/cases (23), one focused on the Chinese population (Li et al., 2016), whilst one focused on protein expression (9, 24) and the other was a mathematical model study (25). In the end, 23 studies were included and analysed in the systematic review.

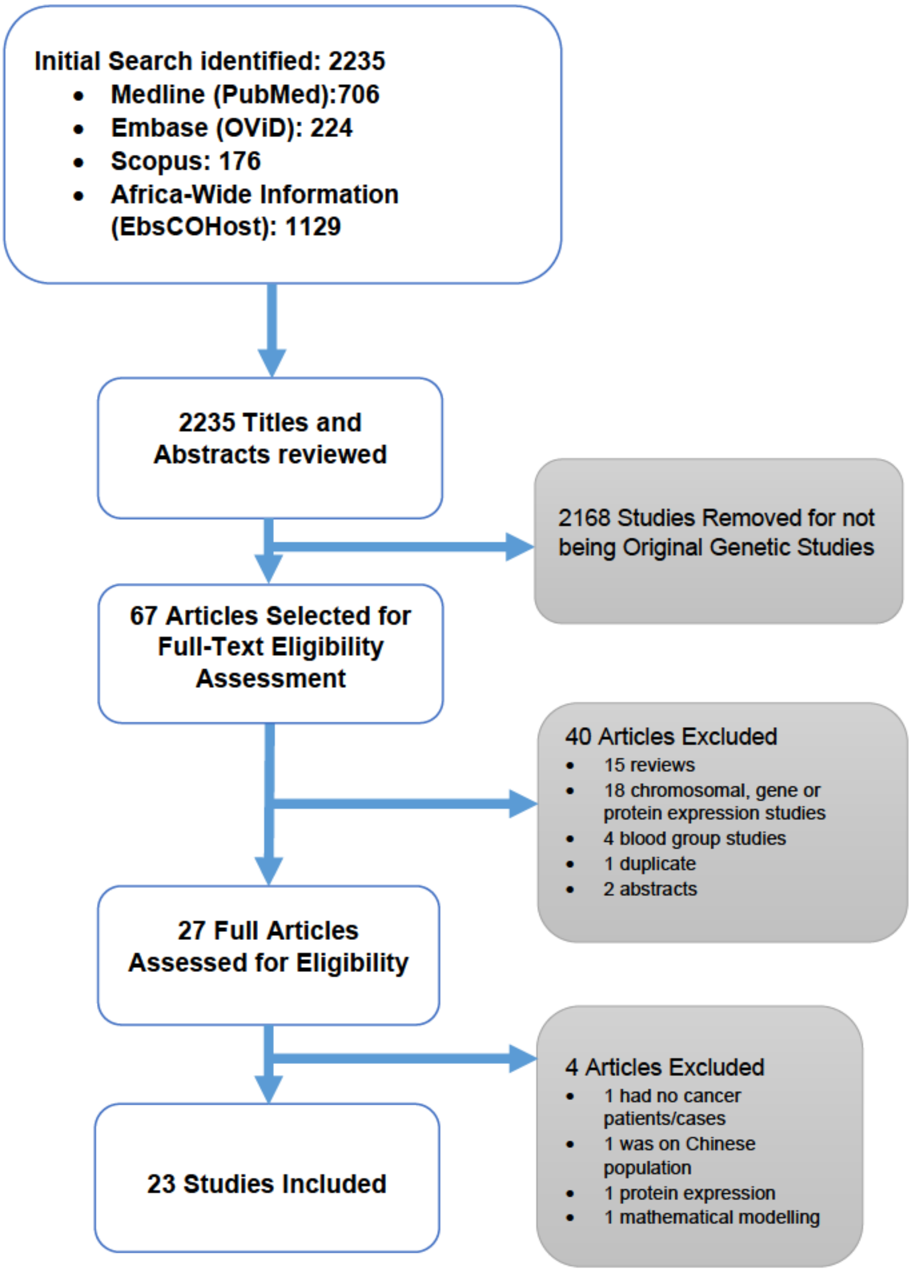

Initial Search identified: 2235
Medline (PubMed):706
Embase (OViD): 224
Scopus: 176
Africa-Wide Information (EbsCOHost): 1129
2235 Titles and Abstracts reviewed
2168 Studies Removed for not being Original Genetic Studies
67 Articles Selected for Full-Text Eligibility Assessment
40 Articles Excluded
15 reviews
18 chromosomal, gene or protein expression studies
4 blood group studies
1 duplicate
2 abstracts
27 Full Articles Assessed for Eligibility
4 Articles Excluded
1 had no cancer patients/cases
1 was on Chinese population
1 protein expression
1 mathematical modelling
23 Studies Included

Figure 1: Systematic Review outline

2.4.2 Study Characteristics

The characteristics of all the genetic susceptibility and somatic variant studies included are shown in Table 2 and 3, respectively. The 23 studies included in the study were published between 1990 and 2019. There were 17 genetic susceptibility and eight somatic variant studies. Two studies reported on both genetic susceptibility and somatic variants.

Table 2.2: Characteristics of genetic susceptibility studies for ESCC in African populations.

Study (PMID)	**Location**	**Year**	**Population**	**Age, y (SD)**		**Sample size**		**Sex, cases n (%)**		**Sex, ctrl n (%)**		**Clinical assessment**		**Analysis method**	**Smoking n (%)**		**Alcohol n (%)**	
				Cases	**Ctrl**	**Cases**	**Ctrl**	**Male**	**Female**	**Male**	**Female**	**Cases**	**Ctrl**		**Cases**	**Ctrl**	**Cases**	**Ctrl**
Bye et al (21926110)	South Africa	2011	Black	59.8 (11.3)	-	358	477	182 (50.8)	176 (49.2)	-	-	Histology	-	Taqman Assay	228 (63.7)	-	228 (63.7)	-
			Mixed Ancestry	60.5 (10.6)	-	201	427	131 (65.2)	70 (34.8)	-	-	Histology	-	Taqman Assay	189 (94.1)	-	163 (81.1)	-
Bye et al (22865593)	South Africa	2012	Black	59.8 (11.3)	48.8 (16.7)	407	849	199 (48.9)	208 (51.1)	335 (39.5)	511 (60.2)	Histology	-	Taqman Assay and KASP	242 (59.5)	333 (39.2)	253 (62.2)	452 (53.2)
			Mixed Ancestry	60.6 (10.6)	46.7 (16.8)	257	860	165 (64.2)	91 (35.4)	309 (35.9)	551 (64.1)	Histology	-	Taqman Assay and KASP	240 (93.4)	597 (69.4)	212 (82.5)	419 (48.7)
Chelule et al (17264406)	South Africa	2006	Black	18-74[1]	18-74	70	261	-	-	-	-	Histology	-	PCR-RFLP	-	-	-	-
Chen et al (30753320)	South Africa	2019	Black[7]	60.2 (11.3)	48.9 (16.8)	591	852	284 (48.1)	307 (51.9)	342 (40.1)	507 (59.5)	Histology	-	Taqman Assay	364 (61.6)	338 (39.7)	370 (62.6)	458 (53.7)

	South Africa		Black[8]	58.2 (10.2)	50.0 (15.5)	880	939	545 (61.9)	332 (37.7)	240 (25.6)	698 (74.3)	Histology		iPLEX and Taqman assays	598 (68.0)	333 (35.5)	473 (53.8)	633 (67.4)
Dandara et al (15978331)	South Africa	2005	Black	-	-	142	178	-	-	-	-	Histology	-	PCR-RFLP	179	162	171	160
	South Africa	2005	Mixed Ancestry			99	94					Histology		PCR-RFLP				
Dandara et al (16272171)	South Africa	2006	Black	61.23	61.85	145	194	85 (59)	60 (41)	111 (57)	83 (43)	Histology	-	PCR-RFLP	95 (65)	123 (63)	98 (68)	127 (65)
			Mixed Ancestry	61.49	69.53	100	94	78 (78)	22 (22)	45 (48)	49 (52)	Histology	-	PCR-RFLP	93 (93)	74 (79)	73 (73)	45 (48)
Dietzsch et al (12925954)	South Africa	2003	Black & Mixed Ancestry	59.6	58.7	58[2]	226	44	14	167	59	-	-	PCR and PAGE				
Eltahir et al (23053979)	Sudan	2012				18	235					Histology		PCR-RFLP				
Li et al (15899651)	South Africa	2005	Black & Mixed Ancestry	61.1 (10.5)	65.7(10.2)	189	198	-	-	-	-	Histology	-	PCR-SSCP and DNA sequencing	144 (76)	122 (62)	133 (70)	114 (58)

Li et al (18254707)	South Africa	2008	Black[3]	-	-	142	178	-	-	-	-	Histology	-	PCR-RLFP	179	162	71	160
	South Africa	2008	Mixed[3] Ancestry			101	100					Histology		PCR-RFLP				
Li et al (20540773)	South Africa	2010	Black[3]	61.23	61.85	145	194	85 (59)	60 (41)	111 (57)	83 (43)	Histology	-	PCR-RFLP	95 (65)	123 (63)	98 (68)	127 (65)
			Mixed[3] Ancestry	61.49	69.53	100	94	78 (78)	22 (22)	45 (48)	49 (52)	Histology	-	PCR-RFLP	93 (93)	74 (79)	73 (73)	45 (48)
Matejcic et al (22216261)	South Africa	2011	Black	-	-	330	479	-	-	-	-	Histology	-	Taqman Assay and Gel Electrophoresis	210	-	204	-
			Mixed Ancestry	-	-	232	428	-	-	-	-	Histology	-	Taqman Assay and Gel Electrophoresis	216	-	189	-
Matejcic et al (26447020)	South Africa	2015	Black	59.6 (10.7)	56.7 (15.0)	463	480	229 (49)	234 (51)	235 (49)	245 (51)	Histology	-	Taqman Assay	280 (60)	222 (46)	286 (62)	278 (58)
			Mixed Ancestry	60.7 (10.3)	57.7 (14.3)	269	288	177 (66)	92 (34)	178 (62)	110 (38)	Histology	-	Taqman Assay	250 (93)	226 (78)	215 (80)	172 (60)

Strickland et al (21901748)	South Africa	2012	Black	59/66[4]	-	96	88	48	48	-	-	Histology	Brush biopsy	HEX SSCP and DNA sequencing	58	-	58	-
Vogelsang et al (22623965)	South Africa	2012	Black	59.8 (11.3)	56.1 (16.2)	345[5]	344	166 (48.1)	179 (51.9)	120 (34.9)	224 (65.1)	Histology	-	Allele-specific quantitative PCR	209 (60.6)	117 (34.0)	160 (46.4)	92 (26.7)
			Mixed Ancestry	60.7 (10.2)	56.8 (16.5)	205[6]	266	136 (66.3)	69 (33.7)	82 (30.8)	184 (69.2)	Histology	-	Allele-specific quantitative PCR	189 (92.2)	162 (60.9)	118 (57.6)	38 (14.3)
Vos et al (12550754)	South Africa	2003	Black	57 (11)	57 (11)	74	118	-	-	-	-	Histology	-	SSCP and DNA sequencing	-	-	-	-
Zaahl et al (15860357)	South Africa	2005	Mixed Ancestry	-	-	105	110	82	23	43	67	Histology	-	SSCP and DNA sequencing	-	-	-	-

[1] Only range of age was reported for the combined group of cases and controls.
[2] 57 had ESCC.
[3] Same population as in Dandara et al. 2005 study.
[4] 59+/-13 for male (n=48) and 66+/- (n=48) for female patients.
[5] 326 had ESCC.
[6] 182 had ESCC.
[7] Western and Eastern Cape Province Black Population
[8] Gauteng Province Black Population

Ctrl, controls; ESCC, Esophageal squamous cell carcinoma; HEX, heteroduplex; KASP, competitive allele specific PCR; PAGE, polyacrylamide gel electrophoresis; PCR, polymerase chain reaction; RFLP, restriction fragment length polymorphism; SD, Standard deviation; SSCP, single-strand conformation polymorphis

Table 2.3: Characteristics of studies on somatic changes in ESCC in African populations.

Study (PMID)	**Country**	**Year**	**Population**	**Sample size**			**Age, y (SD)**	**Sex n (%)**		**Clinical assessment**		**Analysis method**	**Smoking n (%)**	**Alcohol n (%)**
				Ca	NET	Blood	Cases	Male	Female	Ca	NET			
Dietzsch et al (12435113)	South Africa	2002	Black	33	33	-	57.4	23 (70)	10 (30)	Histology	-	PCR and DNA sequencing analysis	-	-
Dietzsch et al (12925954)	South Africa	2003	Black & Mixed ancestry	58[1]	58	-	59.6	29 (67)	14 (33)	-	-	PCR and PAGE	-	-
Gamieldien et al (9808520)	South Africa	1998	Black	76	9	50	57 (11)	49 (65)	27 (35)	Histology	Histology	PCR and HEX-SSCP	-	-
Liu et al (29148985)	Malawi	2016	Malawian	59	-	59	56	27 (45.8)	31 (52.5)	Histology	-	WES	24 (40.7)	14 (23.7)
Naidoo et al (15735161)	South Africa	2005	South African	100	100	-	56	53 (54)	45 (46)	Histology	Histology	PCR	-	-
Patel et al (22040862)	Kenya	2011	Kenyan	28	-	-	56.03 (12.30)	13 (46)	15 (54)	-	-	PCR and DNA sequencing	6 (21)	10 (36)
Victor et al (21901748)	South Africa	1990	Black & Mixed ancestry	27	-	-	-	-	-	-	-	PCR and Dot Blot Hybridisation	-	-
Vos et al (12550754)	South Africa	2003	South African	74	-	37	-	-	-	Histology	-	SSCP and DNA sequencing	-	-

Ca, cancer tissue; HEX-SSCP, heteroduplex single-strand conformation polymorphism; NET, neighboring tissue; PAGE, polyacrylamide gel electrophoresis; PCR, polymerase chain reaction; WES, whole exome sequencing.

[1] 57 had ESCC and 1 had adenocarcinoma

2.4.3 Genetic Susceptibility Studies

The 17 genetic susceptibility studies (Table 2) were all case-control studies (26-42) published between 2003 and 2019. Sixteen articles reported on the South African population and one article on the Sudanese population. The majority (13/16; 81%) of the studies reported on the main subject characteristics (ethnicity, sex, age, and type of clinical assessment). Sample sizes for ESCC patients ranged from 18 to 880 with six of the studies having over 200 patient samples. Sample sizes for controls ranged from 88 to 939 with nine of the studies having over 200 control samples. It is difficult to estimate the total number of patients analysed in these 17 studies, since it appears that the same authors used the same sample set for different SNPs in different publications. Our assessment showed that Bye et al. 2011 (27) and Bye et al. 2012 (26) used the same participants. In addition studies by Li et al. 2005 (40) and Li et al. 2008 (33) used the same participants as Dandara et al. 2005 (29). The remaining 12 studies do not seem to have any obvious sample overlap.

Altogether 16 out of 17 studies clinically assessed for ESCC through histology. None of the studies clinically assessed controls for ESCC with the exception of one study (35) which assessed controls using a brush biopsy. Nine studies reported on smoking and alcohol consumption status for all participants (26, 29, 30, 34, 36, 39-41), whilst three (27, 35, 42) reported those risk factors for only the ESCC patients.

The Hardy Weinberg equilibrium deviation was assessed in eleven (65%) studies, however only six (35%) of the studies reported power calculations and three (18%) studies reported the evaluation of a genotyping error. Detailed characteristics of the study population were reported in twelve of the studies for cases and ten for controls. Correction for multiple testing was reported in only seven (41%) studies. NCBI rs numbers were reported in eight (47%) studies. Our quality assessment scoring had 11 items (Table S1), and each item had a weight of 1 point, therefore total maximum quality score was 11. Overall, only seven of the 17 (41%) studies scored half or above half (5.5). The highest score was 9 (36, 39) and the lowest score was 1 (37, 38).

2.4.3 Somatic Variant Studies

Somatic variant studies (Table 3) constituted of eight studies published between 1990 and 2016 (19, 31, 37, 43-47). A total of 455 patients were assessed, with the control group comprising 200 NET and 146 blood samples. Of the 455 patient samples, one was reported to be an adenocarcinoma from one study; therefore the exact ESCC patient population was 454. The study populations were from South Africa, Kenya, and Malawi.

Clinical diagnosis of ESCC was determined by histology in five (75%) studies, and the remaining three did not report on how clinical assessment was done. Four (50%) studies reported using both cancer tissue and NET for assessment. Three of these studies had an equal number of cancer tissue and NET samples. Two (25%) studies did not have any control samples and the remaining two (25%) studies collected blood samples only as controls. Only two studies reported on smoking and alcohol consumption status. On patient characteristics, age and sex were reported in six (75%) of the studies. Variant classification and type was reported in all of the studies, but confirmation of results was reported in only two studies. No studies used pathogenicity scoring. Amino acid change was also reported in only two of the studies. Our quality assessment score had seven items (Table 2A2), and each item had a weight of 1 point, therefore total maximum score for the quality assessment was 7. Overall, six of the eight (75%) studies scored half or above half (3.5). The highest score was 6 (44) and the lowest score was zero (47).

2.4.4 Description of Genes Studied

A total of 58 genes were investigated in the 23 studies which were selected for the systematic review, with 37 genes studied in the genetic susceptibility studies and 23 in the somatic variant studies. Two genes were investigated in both studies. In addition, the somatic studies investigated six genetic loci without specific gene names. A summary of SNPs analyzed in the genetic susceptibility studies is shown in Table 4. Over 100 SNPs were analysed and 25 SNPs were reported to be associated with ESCC (four SNPs using p values only, and 21 SNPs using p values and odds ratios). The 25 SNPs were in 20 genes; *ADH1B, ADH3, ALDH2, AR, CASP8, CHEK2, CP, CYP2E1, CYP3A5, GSTT2B, MGMT, MLH3, MSH3, NAT2, PTGS2 (also known as COX-2), PLCE1, PMS1, RUNX1, SLC11A1, and TP53*. The associations with all 25 SNPs were identified in South African populations, whilst none were found in the Sudanese population.

Table 2.4: *Summary of studies investigating genetic susceptibility of ESCC in African populations.*

Gene	**Variant (rs number)**	**Study**	**Population**	**ESCC**		**Controls**		**Effect Allele**^{2}	**Findings and Comments**
				n	**MAF**	**n**	**MAF**		
ADH1B	rs1229984 (Arg48His)	Bye et al 2011 (21926110)	Black South African	358	0	477	0		Not informative
		Bye et al 2011 (21926110)	Mixed ancestry South African	201	0.054	427	0.098	A	OR = 0.52 (0.32–0.86) p=0.009
ADH2	ADH2*1/*2/*3	Li et al 2008 (18254707)	Black South African	142	0.01	174	0.01		Not informative
		Li et al 2008 (18254707)	Mixed ancestry South African	96	0.03	94	0.03		Not informative
ADH3	ADH3*1/*2	Li et al 2008 (18254707)	Black South African	141	0.46	174	0.32		NS
		Li et al 2008 (18254707)	Mixed ancestry South African	96	0.38	94	0.31	*2	OR = 1.80; p=0.0004
ADH7	rs1573496 (Gly92Ala)	Bye et al 2011 (21926110)	Black South African	358	0	477	0.001		Not informative
		Bye et al 2011 (21926110)	Mixed ancestry South African	201	0.014	427	0.02		NS
ALDH2	rs671 (Glu504Lys)	Bye et al 2011 (21926110)	Black South African	358	0	477	0		Not informative
		Bye et al 2011 (21926110)	Mixed Ancestry South African	201	0	427	0		Not informative
	rs441 (-261 C/T)	Bye et al 2011 (21926110)	Black South African	358	0.154	477	0.145		NS
		Bye et al 2011	Mixed Ancestry South African	201	0.18	427	0.194		NS

	rs886205 (+82 A/G)	Bye et al 2011 (21926110)	Black South African	358	0.247	477	0.252		NS
		Bye et al 2011 (21926110)	Mixed Ancestry South African	201	0.402	427	0.489	G	OR = 0.70 (0.55–0.89); p=0.004
	ALDH2*1/*2	Li et al 2008 (18254707)	Black South African	142	0.10	174	0.04	*2	OR = 2.35; p=0.008
		Li et al 2008 (18254707)	Mixed Ancestry South African	101	0.03	1004	0.04		Not informative
	rs4767364 (A/G)	Chen et al 2019 (30753320)	Black South African[5]	880	0.12	939	0.11		NS
ALS2CR12	rs13016963 (G/A)	Chen et al 2019 (30753320)	Black South African[4]	591	0.35	852	0.35		NS
		Chen et al 2019 (30753320)	Black South African[5]	880	0.39	939	0.38		NS
	rs10201587 (A/G)	Chen et al 2019 (30753320)	Black South African[5]	880	0.38	939	0.39		NS
AR	CAG-repeat in exon 1	Dietzsch et al 2003 (12925954)	Black South African males	29		109			NS
		Dietzsch et al 2003 (12925954)	Mixed Ancestry South African males	15		58			NS
	GGC-repeat in exon 1	Dietzsch et al 2003 (12925954)	Black South African males	29		109		$(GGC)_{\leq 16}$	OR = 2.7 (1.14-6.36); p=0.018
		Dietzsch et al 2003 (12925954)	Mixed Ancestry South African males	15		58			NS
ATP1B2/ TP53	rs1642764 (C/T)	Chen et al 2019 (30753320)	Black South African[4]	591	0.21	852	0.20		NS
		Chen et al 2019 (30753320)	Black South African[5]	880	0.18	939	0.18		NS

	rs1641511 (A/G)	Chen et al 2019 (30753320)	Black South African[5]	880	0.39	939	0.42		NS
C20orf54	rs13042395	Bye et al 2012 (22865593)	Black South African	407	0.002	849	0.005		Not informative
		Bye et al 2012 (22865593)	Mixed Ancestry South African	257	0.067	860	0.068		NS
CASP8	rs1045485 (Asp302His)	Bye et al 2011 (21926110)	Black South African	358	0.154	477	0.152		NS
		Bye et al 2011 (21926110)	Mixed Ancestry South African	201	0.169	427	0.126	C	OR = 1.42 (1.01–1.98); p=0.040
	rs3834129 (-652 6N ins/del)	Bye et al 2011 (21926110)	Black South African	358	0.518	477	0.502		NS
		Bye et al 2011 (21926110)	Mixed Ancestry South African	201	0.385	427	0.386		NS
	rs10931936 (C/T)	Chen et al 2019 (30753320)	Black South African[4]	591	0.19	852	0.20		NS
		Chen et al 2019 (30753320)	Black South African[5]	880	0.22	939	0.20		NS
CHEK2	rs4822983 (C/T)	Chen et al 2019 (30753320)	Black South African[4]	591	0.46	852	0.39	T	OR=1.32 (1.12-1.56); p=0.001
		Chen et al 2019 (30753320)	Black South African[5]	880	0.43	939	0.42		NS
	rs1033667 (C/T)	Chen et al 2019 (30753320)	Black South African[4]	591	0.44	852	0.38	T	OR=1.30 (1.10-1.53) P=0.002
		Chen et al 2019 (30753320)	Black South African[5]	880	0.42	939	0.39		NS
CP	rs34053109 (C/G)	Strickland et al. 2012 (21901748)	Black South African	84	0	85	0.01		Not informative

rs17838834 (T/C)	Strickland et al. 2012 (21901748)	Black South African	90	0.33	85	0.23		NS
rs701749 (C/T)	Strickland et al. 2012 (21901748)	Black South African	79	0.01	78	0.02		Not informative
rs17838833 (delT)	Strickland et al. 2012 (21901748)	Black South African	79	0.01	78	0		Not informative
rs17838832 (T/C)	Strickland et al. 2012 (21901748)	Black South African	80	0.33	78	0.3		NS
rs34334174 (C/T)	Strickland et al. 2012 (21901748)	Black South African	80	0.14	78	0.08		NS
5'UTR-308G/A	Strickland et al. 2012 (21901748)	Black South African	52	0.05	64	0	A	p=0.012; sample size very small
rs17838831 (A/G)	Strickland et al. 2012 (21901748)	Black South African	53	0.21	64	0.22		NS
rs138512757 (Thr83)	Strickland et al. 2012 (21901748)	Black South African	92	0.02	84	0.01		Not informative
rs35438054 (Val223)	Strickland et al. 2012 (21901748)	Black South African	95	0.01	85	0.01		Not informative
rs797045480 (Val246Ala)	Strickland et al. 2012 (21901748)	Black South African	95	0.01	85	0		Not informative
rs34067682 (IVS4-14C/T)	Strickland et al. 2012 (21901748)	Black South African	84	0.12	83	0.12		NS
rs34624984 (Arg367Cys)	Strickland et al. 2012 (21901748)	Black South African	94	0.02	86	0.01		Not informative
rs34237139 (Tyr425)	Strickland et al. 2012 (21901748)	Black South African	91	0.01	87	0		Not informative
rs35272481 (IVS7+9T/C)	Strickland et al. 2012 (21901748)	Black South African	91	0.01	87	0		Not informative

	rs701753 (D544E)	Strickland et al. 2012 (21901748)	Black South African	95	0.23	81	0.27		NS
	rs147192657 (Gly633 T/C)	Strickland et al. 2012 (21901748)	Black South African	88	0.07	84	0	C	p=0.0004
	rs16861582 (IVS15-12T/C)	Strickland et al. 2012 (21901748)	Black South African	93	0,44	88	0.41		NS
CYP2E1	CYP2E1*1 (c1) / CYP2E1*5 (c2)	Chelule et al 2006 (17264406)	Black South African	30	0.04	331	0.06		Limited power
	-1053C/T	Li et al 2005 (15899651)	Black & Mixed Ancestry South African	189	0.01	198	0.02		NS
	-1293G/A	Li et al 2005 (15899651)	Black & Mixed Ancestry South African	189	0.01	198	0.03		NS
	7632T/A	Li et al 2005 (15899651)	Black & Mixed Ancestry South African	189	0.18	198	0.07	A	OR = 5.90 (3.25-10.7); p=0.001 for genotype distribution
CYP3A5	CYP3A5*1	Dandara et al 2005 (15978331)	Black South African	142	0.627	178	0.638		NS
		Dandara et al 2005 (15978331)	Mixed Ancestry South African	99	0.384	94	0.287		NS
	CYP3A5*3 (6986A/G)	Dandara et al 2005 (15978331)	Black South African	142	0.155	178	0.138		NS
		Dandara et al 2005 (15978331)	Mixed Ancestry South African	99	0.475	94	0.590	G	OR = 0.60 (0.39–0.94); p=0.025
	CYP3A5*6 (1490G/A)	Dandara et al 2005 (15978331)	Black South African	142	0.190	178	0.213		NS
		Dandara et al 2005 (15978331)	Mixed Ancestry South African	99	0.136	94	0.122		NS

	CYP3A5*7 (27131-32insT; frameshift)	Dandara et al 2005 (15978331)	Black South African	142	0.028	178	0.011	NS
		Dandara et al 2005 (15978331)	Mixed Ancestry South African	99	0.005	94	0	Not informative
	CYP3A5 all variants	Dandara et al 2005 (15978331)	Black South African	142	0.373	178	0.441	NS
		Dandara et al 2005 (15978331)	Mixed Ancestry South African	99	0.616	94	0.713	OR = 0.65 (0.42–0.99); p=0.045
FAS	rs1800682 (-670 G>A)	Bye et al 2011 (21926110)	Black South African	358	0.219	477	0.225	NS
		Bye et al 2011 (21926110)	Mixed Ancestry South African	201	0.356	427	0.406	NS
	rs2234767 (-1377 G>A)	Bye et al 2011 (21926110)	Black South African	358	0.096	477	0.072	NS
		Bye et al 2011 (21926110)	Mixed Ancestry South African	201	0.139	427	0.183	NS
FASL	rs763110 (-844 T>C)	Bye et al 2011 (21926110)	Black South African	358	0.192	477	0.189	NS
		Bye et al 2011 (21926110)	Mixed Ancestry South African	201	0.416	427	0.386	NS
GSTP1	rs1695 (Ile105Val)	Matejcic et al. 2011	Black South African	325	0.518	474	0.534	NS
	rs1695 (Ile105Val)	Matejcic et al. 2011	Mixed Ancestry South African	229	0.454	428	0.438	NS
	rs1695 (Ile105Val)	Li et al 2010 (20540773)	Black South African		0.39		0.37	NS
	rs1695 (Ile105Val)	Li et al 2010 (20540773)	Mixed Ancestry South African		0.38		0.41	NS

	rs1138272 (Ala114Val)	Li et al 2010 (20540773)	Black South African		0.22		0.07		NS
	rs1138272 (Ala114Val)	Li et al 2010 (20540773)	Mixed Ancestry South African		0.19		0.03		NS
GSTT1	Deletion allele	Matejcic et al. 2011 (22216261)	Black South African	311	0.574	462	0.554		NS
		Matejcic et al. 2011 (22216261)	Mixed Ancestry South African	217	0.493	414	0.495		NS
GSTT2B	Deletion allele	Matejcic et al. 2011 (22216261)	Black South African	320	0.336	461	0.371		NS
		Matejcic et al. 2011 (22216261)	Mixed Ancestry South African	226	0.418	425	0.501		OR = 0.71 (0.57-0.90); p=0.004
MGMT	rs12917 (Leu84Phe)	Bye et al 2011 (21926110)	Black South African	358	0.189	477	0.195		NS
		Bye et al 2011 (21926110)	Mixed Ancestry South African	201	0.222	427	0.168		OR = 1.41 (1.05–1.91); p=0.023
MLH1	rs13320360 (c.546-191T/C)	Volgesang et al. 2012 (22623965)	Black South African	343	0.15	340	0.17		NS
		Volgesang et al.2012 (22623965)	Mixed Ancestry South African	203	0.07	264	0.06		NS
MLH3	rs28756991 (Arg797His)	Volgesang et al. 2012 (22623965)	Black South African	345	0.11	342	0.12		NS
		Volgesang et al. 2012 (22623965)	Mixed Ancestry South African	205	0.09	264	0.4	G	OR = 2.07 (1.04-4.12); p=0.038
MSH2	rs17217772 (Asn127Ser)	Volgesang et al. 2012 (22623965)	Black South African	341	0.06	343	0.06		NS
		Volgesang et al. 2012 (22623965)	Mixed Ancestry South African	204	0.03	264	0.03		NS

	rs10188090 (c.2635-765G/A)	Volgesang et al. 2012 (22623965)	Black South African	343	0.09	342	0.10		NS
		Volgesang et al. 2012 (22623965)	Mixed Ancestry South African	205	0.31	265	0.33		NS
	rs3771280 (c.1510+118T/C)	Volgesang et al. 2012 (22623965)	Black South African	344	0.11	339	0.12		NS
		Volgesang et al. 2012 (22623965)	Mixed Ancestry South African	202	0.35	266	0.37		NS
MSH3	rs26279 (Ala1045Thr)	Volgesang et al. 2012 (22623965)	Black South African	341	0.40	344	0.43		NS
		Volgesang et al. 2012 (22623965)	Mixed Ancestry South African	204	0.38	263	0.32	A	OR = 2.71 (1.34-5.50); p=5.71x10^{-3}
	rs1428030 (c.1341-12568A/G)	Volgesang et al. 2012 (22623965)	Black South African	342	0.29	342	0.27		NS
		Volgesang et al. 2012 (22623965)	Mixed Ancestry South African	201	0.23	264	0.20		NS
	rs1805355 (Pro231Pro)	Volgesang et al. 2012 (22623965)	Black South African	343	0.28	339	0.29		NS
		Volgesang et al. 2012 (22623965)	Mixed Ancestry South African	203	0.24	265	0.22		NS
NAT1	rs1057126 (1088T>A NAT1*10)	Matejcic et al. 2015 (26447020)	Black South African	463	54.8	480	57.7		NS
		Matejcic et al. 2015 (26447020)	Mixed Ancestry South African	269	43.4	288	40.1		NS
	rs15561 (1095C>A NAT1*10, NAT1*3)	Matejcic et al. 2015 (26447020)	Black South African	463	55.7	480	57.7		NS
		Matejcic et al. 2015 (26447020)	Mixed Ancestry South African	269	46.5	288	43		NS

NAT2	rs1799930 (590G/A NAT2*6)	Matejcic et al. 2015 (26447020)	Black South African	463	24.7	480	21.4		NS
		Matejcic et al. 2015 (26447020)	Mixed Ancestry South African	269	22.4	288	22		NS
	rs1801280 (341T/C NAT2*5)	Matejcic et al. 2015 (26447020)	Black South African	463	27.1	480	29		NS
		Matejcic et al. 2015 (26447020)	Mixed Ancestry South African	269	25.2	288	33.2	C	0R = 0.57 (0.38-0.87) p=0.01
	rs1799931 (857G/A NAT2*7)	Matejcic et al. 2015 (26447020)	Black South African	463	0.01	480	0.05		Not informative
		Matejcic et al. 2015 (26447020)	Mixed Ancestry South African	269	0.05	288	0.04		NS
	rs1801279 (191G/A NAT2*14)	Matejcic et al. 2015 (26447020)	Black South African	463	0.053	480	0.063		NS
		Matejcic et al. 2015 (26447020)	Mixed Ancestry South African	269	0.038	288	0.023		NS
UNC5CL	rs10484761 (G/A)	Bye et al 2012 (22865593)	Black South African	407	0.467	849	0.477		NS
		Bye et al 2012 (22865593)	Mixed Ancestry South African	257	0.354	860	0.314		NS
PTGS2	rs20417 (-765 G/C)	Bye et al 2011 (21926110)	Black South African	358	0.471	477	0.513		NS
		Bye et al 2011 (21926110)	Mixed Ancestry South African	201	0.376	427	0.321		NS
	rs689466 (-1195 A/G)	Bye et al 2011 (21926110)	Black South African	358	0.064	477	0.053		NS
		Bye et al 2011 (21926110)	Mixed Ancestry South African	201	0.103	427	0.155	G	OR = 0.63 (0.43–0.91); p=0.014

PDE4D	rs10052657 (C/A)	Bye et al 2012 (22865593)	Black South African	407	0.137	849	0.128		NS
		Bye et al 2012 (22865593)	Mixed Ancestry South African	257	0.175	860	0.155		NS
PLCE1	rs2274223 (His1927Arg)	Bye et al 2012 (22865593)	Black South African	407	0.416	849	0.403		NS
		Bye et al 2012 (22865593)	Mixed Ancestry South African	257	0.437	860	0.40		NS
	rs17417407 (Arg548Leu)	Bye et al 2012 (22865593)	Black South African	407	0.166	849	0.211	T	OR = 0.74 (0.60-0.93); p=0.008
		Bye et al 2012 (22865593)	Mixed Ancestry South African	257	0.174	860	0.18		NS
	rs1438095332 (5’UTR 14 bp indel)	Bye et al 2012 (22865593)	Black South African	321	0.234	456	0.242		NS
	rs199781223 (Gly1199Ser)	Bye et al 2012 (22865593)	Black South African	321	0.053	449	0.045		NS
	rs3765525[3] (Ile1777Thr)	Bye et al 2012 (22865593)	Black South African	316	0.472	452	0.463		NS
	rs58539480 (Pro1890Leu)	Bye et al 2012 (22865593)	Black South African	307	0.073	429	0.064		NS
	rs17417407 (G/T)	Chen et al 2019 (30753320)	Black South African[4]	591	0.17	852	0.21	T	OR = 0.76 (0.60-0.95); p=0.014
		Chen et al 2019 (30753320)	Black South African[5]	880	0.19	939	0.19		NS
	rs7084339 (G/A)	Chen et al 2019 (30753320)	Black South African[5]	880	0.48	939	0.46		NS
	rs3765524 (T/C)	Chen et al 2019 (30753320)	Black South African[4]	591	0.47	852	0.47		NS

		Chen et al 2019 (30753320)	Black South African[5]	880	0.48	939	0.46		NS
	rs2274223 (A/G)	Chen et al 2019 (30753320)	Black South African[4]	591	0.42	852	0.40		NS
		Chen et al 2019 (30753320)	Black South African[5]	880	0.41	939	0.43		NS
	rs11187850 (A/G)	Chen et al 2019 (30753320)	Black South African[5]	880	0.21	939	0.19		NS
PMS1	rs5742938 (c.-21+639G/A)	Volgesang et al. 2012 (22623965)	Black South African	345	0.18	344	0.15		NS
		Volgesang et al. 2012 (22623965)	Mixed Ancestry South African	203	0.43	266	0.48	A	OR = 1.73 (1.07-2.79); p=0.027
	rs13404927 (c.699+3331G/A)	Volgesang et al. 2012 (22623965)	Black South African	342	0.18	339	0.19		NS
		Volgesang et al. 2012 (22623965)	Mixed Ancestry South African	204	0.14	264	0.12		NS
RUNX1	rs2014300 (A/G)	Bye et al 2012 (22865593)	Black South African	407	0.378	849	0.403		NS
		Bye et al 2012 (22865593)	Mixed Ancestry South African	257	0.438	860	0.370	G	OR = 1.33 (1.09–1.63); p=0.0055
	rs2014300 (A/G)	Chen et al 2019 (30753320)	Black South African[4]	591	0.38	852	0.40		NS
		Chen et al 2019 (30753320)	Black South African[5]	880	0.36	939	0.36		NS
	rs2834718 (T/A)	Chen et al 2019 (30753320)	Black South African[5]	880	0.33	939	0.33		NS
SLC11A1	-237C/T	Zaahl et al. 2005 (15860357)	Mixed Ancestry South African	105	0.029	110	0.1		p<0.004

	-8G/A	Zaahl et al. 2005 (15860357)	Mixed Ancestry South African	105	0.004	110	0.009		NS
	IVSI-28C/T	Zaahl et al. 2005 (15860357)	Mixed Ancestry South African	105	0.028	110	0.0004		p<0.05
	GT-repeat	Zaahl et al. 2005 (15860357)	Mixed Ancestry South African		0.171		0.191		NS
SULT1A1	638G/A in Exon 7	Dandara et al 2006 (16272171)	Black South African	145	0.42	194	0.37		NS[1]
		Dandara et al 2006 (16272171)	Mixed Ancestry South African	100	0.40	94	0.29		NS
TMEM173	rs13181561 (A/G)	Chen et al 2019 (30753320)	Black South African[5]	880	0.48	939	0.49		NS
	rs13153461 (G/A)	Chen et al 2019 (30753320)	Black South African[4]	591	0.04	852	0.05		NS
TP53	16-bp insertion in intron 3	Vos et al. 2003 (12550754)	Black South African	74	0.108	118	0.364		
	rs200073907 (Exon 4 codon 34)	Vos et al. 2003 (12550754)	Black South African	74	0.115	118	0.102		NS
	rs750578863 (Exon 4 codon 36)	Vos et al. 2003 (12550754)	Black South African	73	0.089	115	0.143		NS
	Arg72Pro	Vos et al. 2003 (12550754)	Black South African	73	0.356	115	0.409		p<0.05
	Arg72Pro	Eltahir et al 2012 (23053979)	Sudanese	25	0.49	235	0.51		NS
	rs1800371 (G/A)	Chen et al 2019 (30753320)	Black South African[4]	591	0.02	852	0.03		NS
		Chen et al 2019 (30753320)	Black South African[5]	880	0.03	939	0.02		NS

XBP1	rs2239815 (C/T)	Chen et al 2019 (30753320)	Black South African[4]	591	0.21	852	0.16	T	OR=1.41 (1.15-1.74) p=0.001
		Chen et al 2019 (30753320)	Black South African[5]	880		939			NS

[1] Increased risk among smokers with *SULT1A1*2/*2* genotype, but sample size was small.

[2] When OR>1, effect allele =increased risk ; When OR<1, effect allele = protective effect.

[3] rs3765525 has been merged into rs959421.

[4] Western and Eastern Cape Province Black Population.

[5] Gauteng Province Black Population.

Table 2.5 shows a summary of the pathways for the 20 genes. All the genes encode for proteins. Four of the genes *ADH1B, ADH3, ALDH2* are involved in alcohol metabolism (27, 33). Three mismatch repair genes, *MLH3, MSH3,* and *PMS1* play a role in genomic integrity (36). They are reported to also play a role in carcinogenesis. MGMT is involved in cell defence against mutagens, and mutations in the gene are reported to be associated with cancer formation (27). *NAT2* and *GSTT2B* play a role in the activation and deactivation of drugs and carcinogens, with reports of mutations being associated with carcinogenesis (34). Genes regulating cell apoptosis are *TP5, CHEK2 and CASP8* (27, 32, 37, 39). *TP53* and *CHEK2* is also involved in gene expression, and DNA repair. Regulation of gene expression is facilitated by *PLCE1* and *SLC11A1* (26, 38). The *AR* gene regulates the sex hormones, androgens (31), whilst *CYP2E1* and *CYP3A5* is involved in steroid, cholesterol and lipid synthesis (28, 29, 40). *CYP2E1* also metabolizes drugs and has been implicated in carcinogenesis. *CP* facilitates transportation of iron from organs into the blood cells*, RUNX1* plays a role in hematopoiesis and *PTGS2* in inflammation and mitogenesis (26, 27, 35).

Table 2.5: Biological pathways for genetic susceptibility studies showing putative association with ESCC in African populations

Gene	Full Name	Pathway
ADH1B	Alcohol dehydrogenase 1B (class I), beta polypeptide	Ethanol metabolism
ADH3	Alcohol dehydrogenase ADH3	Metabolises ethanol into acetaldehyde
ALDH2	Aldehyde dehydrogenase 2 family member	Alcohol metabolism. Implicated in increased susceptibility for cancer
AR	Androgen Receptor	Regulates binding of androgens on androgen receptor
CASP8	Caspase 8	Cell apoptosis
CHEK2	Checkpoint kinase 2	Tumour suppressor gene. Mutations associated with predisposition to carcinogenesis
CP	Ceruloplasm	Peroxidation of iron through its transportation from organs and tissue into blood
CYP2E1	Cytochrome P450 family 2 subfamily E member 1	Drug metabolism and catalysis and synthesis of cholesterol, steroids and other lipids. Implicated in cancer development
CYP3A5	cytochrome P450 family 3 subfamily A member 5	Involved in drug metabolism and in the synthesis of cholesterol, steroids and other lipids
GSTT2B	Glutathione S-transferase theta 2B (gene/pseudogene)	Conjugation of glutathione to electrophilic and hydrophobic compounds. Plays a role in carcinogenesis
MGMT	O-6-methylguanine-DNA methyltransferase	DNA repair and defence from alkylating agents which cause mutagenesis and toxicity. Implicated in several cancers.
MLH3	MutL homolog 3	Maintenance of genomic integrity following cell division and DNA replication. Germline mutations implicated in cancer and somatic mutations implicated in microsatellite instability
MSH3	MutS homolog 3	Forms heterodimers with MSH2. Involved in mismatch repair and implicated in cancer development.

NAT2	N-acetyltransferase 2	Activation and deactivation of arylamine and hydrazine drugs and carcinogens. Implicated in high cancer incidence and drug toxicity.
PTGS2	Prostaglandin-endoperoxide synthase 2	A dioxygenase and a peroxidase involved in both inflammation and mitogenesis
PLCE1	Phospholipase C epsilon 1	Regulation of cell growth, differentiation, and gene expression.
PMS1	PMS1 homolog 1, mismatch repair system component	Mismatch repair gene. Mutations implicated in cancer development.
RUNX1	Runt related transcription factor 1	Development of hematopoiesis
SLC11A1	Solute carrier family 11 (proton-coupled divalent metal ion transporter), member 1	Regulation of gene expression.
TMEM173	Transmembrane protein 173	Regulation of the innate immune response to viral and bacterial infections. Role in tumourigenesis still inadequate
TP53	Tumour protein 53	Regulation of gene expression, cell cycle, apoptosis, and DNA repair.
XBP1	X-box binding protein 1	Regulation of genes involved in endoplasmic reticulum protein synthesis, folding, glycosylation, redox metabolism, autophagy, lipid biogenesis and vesicular trafficking. Associated with development of cancer.

Nine of the 25 associated SNPs were from small studies with fewer than 150 cases and controls. These SNPs are in the following six genes: *ADH3, AR, CP, CYP3A5, SLC11A1, and TP53*. Because of the small sample size, the reliability and replicability of these results is uncertain. Sixteen of the SNPs came from studies with at least 150 cases and controls, and one study with 142 cases. These sample sizes could potentially give reliable and replicable results. The 16 SNPs were from the following genes: *ADH1B, ALDH2, CASP8, CHEK2, CYP2E1, GSTT2B, MGMT, MLH3, MSH3, NAT2, PLCE1, PMS1, PTGS2, and RUNX1*.

Two of the 16 SNPs in the *ALDH2* gene and were analysed in two different studies. However, it is not clear whether these two SNPs are the same because whilst one study reported the NCBI rs number (rs886205) (27) the other study did not (33).The two SNPs reported very different MAF, and opposite odds ratios of 2.35 and 0.70 demonstrating increased risk and a protective effect, respectively.

Six of the 16 SNPs were reported to reduce the risk of ESCC, and they are: *ADH1B* (Arg48His; rs1229984), *ALDH2* (+82 A>G; rs886205), *GSTT2B* (deletion allele), *NAT2* (341T>C; rs1801280), *PTGS2* (-1195 A>G; rs689466) and *PLCE1* (Arg548Leu; rs17417407). The remaining ten SNPs were reported to increase the risk of ESCC: *ALDH2* (ALDH2*1/*2), *CASP8* (Asp302His; rs1045485), *CHEK2* (rs4822983 C>T, and rs1033667, C>T), *CYP2E1* (7632T>A), *MGMT* (Leu84Phe; rs12917), *MLH3* (Arg797His; rs28756991), *MSH3* (Ala1045Thr; rs26279), *PMS1* (c.-21+639G>A; rs5742938), and *RUNX1* (rs2014300). The effect alleles associated with these SNPs are indicated in Table 2.4. Eleven of the 16 SNPs showed association in the South African Admixed population, whilst only 4 showed association in the Black South African population and one in a combined South African population. All the studies used PCR-based methods for genotyping.Using the 1000 Genomes Database, r^2 analysis was carried out on SNPs reported in the same gene, to assess the LD between the SNPs. Thirteen pairs of SNPs in *MHS2, CP, MSH3, PLCE1,CHEK2,* and *NAT1* genes had $r^2 > 0.45$, shown in Figure 2 and Table 2A3.

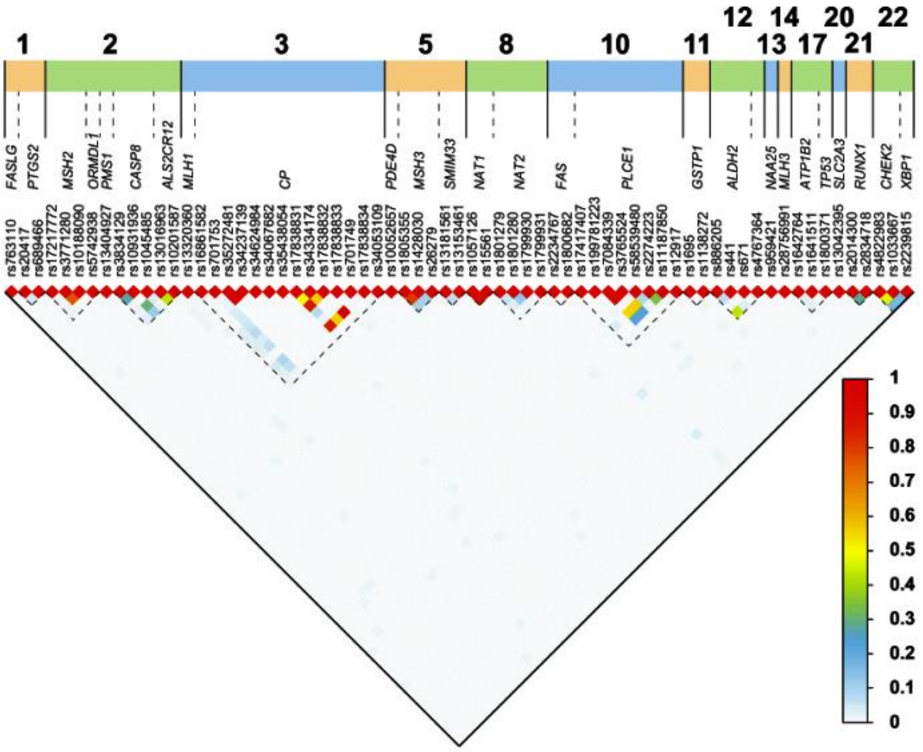


***Figure 2.2:** Linkage Disequilibrium Plot for Paired SNPs. We obtained the rs-numbers of the variants from dbSNP (version 152; April 2019; (https://www.ncbi.nlm.nih.gov/snp/) and used the canonical SNP identifier. To determine the linkage disequilibrium (LD) between the SNPs, we obtained the imputed data set from the Thousand Genomes project (1000 Genomes Release Phase 3 2013-05-02; ref 1kG), and used bcftools to extract all individuals from African populations not including African Americans, and the 77 SNPs discussed here using all synonyms (alternative rs IDs) for SNPs (Auton et al., 2015). We obtained a dataset of 504 individuals and 67 SNPs. We computed all pair-wise r2 using PLINK (v1.09) (Danecek et al., 2011;Chang et al., 2015)*

Altogether 44 somatic changes were reported in the following 22 genes: AR, *CCND1, CDKN2A, COL1A2, EFGR, EP300, FAT1, FAT2, FAT3, FAT4, FBXW7, JAG1, KMT2C (MLL3), KMT2D (MLL2), MUC2, NFE2L2, NOTCH1, NOTCH3, PIK3CA, SERPINB4, TP53, and TP63,* and six genetic loci without specific gene names (Table 6). The specific locus positions with the corresponding microsatellite markers are as follows: 2p (D2S123), 3p13 (D3S659), 3p24.2-25 (D3S1255), 4q12 (Bat 25), 2p21-p16.3 (Bat 26), and 1p12-13.3 (Bat 40). These variants were reported in the South African (20 variants), Kenyan (three variants), and Malawian (21 variants) populations.

Whilst the majority of the studies used PCR-based methods, a more recent study used WES as the analysis method (45). A total of 18 of the 22 genes with somatic variants in cancer tissue were discovered using WES. Statistical significance was not reported for any of the 44 variants. The most common type of somatic variants was missense mutations, reported in 14 of the 22 genes (64%) (45, 46). Other somatic changes included copy number gains (14%), copy number losses (5%), deletions (14%), insertions (14%), and frameshift mutations (14%). In three studies (19, 31, 43) microsatellite instability and loss of heterozygosity (LOH) was reported (14%).

Table 2.6: Summary of studies investigating somatic changes linked to ESCC in African patients

Gene	Study (PMID)	Population	Findings
AR	Dietzsch et al 2003 (12925954)	Black and Mixed Ancestry South African	LOH at CAG locus
CCND1	Liu et al 2016 (29148985)	Malawian	Enriched copy number gains
CDKN2A	Gamieldien et al 1998 (9808520)	Black South African	Insertions Deletions Frameshift mutations
	Liu et al 2016 (29148985)	Malawian	Copy number losses
COL1A2	Dietzsch et al 2002 (12435113)	Black South African	LOH (promoter and 1st intron) No evidence of MSI or allelic amplification
EFGR	Liu et al 2016 (29148985)	Malawian	Copy number gains
EP300	Liu et al 2016 (29148985)	Malawian	Missense mutations
FAT1	Liu et al 2016 (29148985)	Malawian	Nonsense mutations
FAT2	Liu et al 2016 (29148985)	Malawian	Missense mutations
FAT3	Liu et al 2016 (29148985)	Malawian	Missense mutations
FAT4	Liu et al 2016 (29148985)	Malawian	Missense mutations
FBXW7	Liu et al 2016 (29148985)	Malawian	Frameshift mutations
JAG1	Liu et al 2016 (29148985)	Malawian	Missense mutations
KMT2C (MLL3)	Liu et al 2016 (29148985)	Malawian	Missense mutations
KMT2D(MLL2)	Liu et al 2016 (29148985)	Malawian	Nonsense mutations
Mismatch repair genes	Naidoo et al 2005 (15735161)	South African	LOH and MSI at: • D2S123 (2p) • D3S659 (3p13) • D3S1255 (3p3p24.2-25) • Bat 25 (4q12)

			• Bat 26 (2p2p21-p16.3) • Bat 40 (1p12-13.3)
MUC2	Liu et al 2016 (29148985)	Malawian	Missense mutations
NFE2L2	Liu et al 2016 (29148985)	Malawian	Missense mutations
NOTCH1	Liu et al 2016 (29148985)	Malawian	Missense mutations
NOTCH3	Liu et al 2016 (29148985)	Malawian	Missense mutations
PIK3CA	Liu et al 2016 (29148985)	Malawian	Missense mutations
Ras genes	Victor et al 1990 (21901748)	South African	No mutations found in codon 12, 13 or 61
SERPINB4	Liu et al 2016 (29148985)	Malawian	Missense mutations
TP53	Liu et al 2016 (29148985)	Malawian	Missense and nonsense mutations
	Gamieldien et al 1998 (9808520)	Black South African	Exon 5-8 frameshift mutations: point mutations, deletions and insertions
	Patel et al 2011 (22040862)	Kenyan	Exon 5-8 mutations: missense, nonsense and deletions
	Vos et al 2003 (12550754)	South African	16-bp insertion in intron 3
	Vos et al 2003 (12550754)	South African	Exon 4 polymorphism in codons 34, 36 and 72 LOH (16-bp repeat locus)
TP63	Liu et al 2016 (29148985)	Malawian	Copy number gains

LOH, loss of heterozygosity; MSI, microsatellite instability.

Table 7 shows a summary of the pathways in the 22 genes reporting somatic changes. Five genes*; AR, EP300, KMT2D, KMT2C,* and *TP53*, play a role in the regulation of transcription (31, 37, 44-46). The encoded protein for the *AR* gene functions as a steroid hormone activated transcription factor, whilst KMT2D has a role in methylation. Both *TP53* and *EP300* have been implicated in a number of cancers (37, 44-46). *TP53* additionally functions in DNA repair, gene expression and apoptosis. The mismatch repair genes also facilitate DNA repair (19). *CCND1, CDKN2A, FAT1/2/3/4,* and *Ras* genes are all reported to be involved in cell cycle pathways including regulation of mitotic events, cell proliferation, cell growth and death (44, 45, 47). *NOTCH1* and *NOTCH3* both facilitate cell and tissue development (45). *JAG1* plays a role in hematopoiesis whilst *NFE2L2* is involved in response to inflammation including production of free radicals (45). *PIK3CA* is an oncogene implicated in tumour development whilst *SERPINB4* modulates response against tumour cells (45). *EGFR* and *COL1A2* genes encode for epidermal growth factor and type 1 collagen, respectively (43, 45). *FBXW7* is a tumour suppressor involved in ubiquitin degradation (45). *MUC2* facilitates the formation of a mucous barrier that protects the gut lumen (45). *TP63* gene is involved in tissue and organ development including skin and heart, and in adult stem cell regulation (45).

Table 2.7: Biological pathways for somatic changes studies showing putative association with ESCC in African populations.

Gene	Full Name	Pathway
AR	Androgen receptor gene	Regulation of gene expression and the protein functions as a steroid-hormone activated transcription factor.
CCND1	Cyclin D1	Regulators of CDK kinases and mitotic events. Mutations and overexpression of the gene has been associated with cancer development.
CDKN2A	Cyclin dependent kinase inhibitor 2A	A tumour suppressor gene which regulates the cell cycle. Commonly inactivated in a variety of tumours.
CHEK2		
COL1A2	Collagen type I, alpha 2 chain	Encodes for type I collagen, which is an abundant connective tissue protein and part of extracellular matrix.
EFGR	Epidermal growth factor receptor	Encodes for the growth factor epidermal growth factor receptor.
EP300	E1A binding protein p300	Encodes the adenovirus E1A-associated cellular p300 transcriptional co-activator protein which functions in transcription regulation. Mutations have been implicated in tumorigenesis.
FAT1/2/3/4	FAT atypical cadherin 1/2/3/4	Human homologues of the Drosophila FAT genes. Putative tumour suppressor involved in cell proliferation during Drosophila development.
FBXW7	F-box and WD repeat domain containing 7	Encodes an F-Box protein which binds directly to cyclin E and potentially targets cyclin E for ubiquitin-mediated degradation.
JAG1	Jagged 1	Encodes for the human homolog of the Drosophila jagged 1 protein which is involved in hematopoiesis.
KMT2C (MLL3)	Lysine methyltransferase 2C	The gene is member of the myeloid/lymphoid or mixed-lineage leukemia (MLL) family. It encodes a nuclear protein involved in transcriptional regulation.
KMT2D (MLL2)	Lysine methyltransferase 2D	Methylation of histones and transcriptional regulation.

Mismatch repair genes	Mismatch repair genes	DNA repair. Mutations have been implicated in cancer.
MUC2	Mucin 2, oligomeric mucus/gel-forming	Formation of insoluble mucous barrier that protects the gut lumen.
NFE2L2	nuclear factor, erythroid 2 like 2	Encodes for proteins involved in response to inflammation including free radical production.
NOTCH1	NOTCH1	Development of cell and tissue. Mutations have been reported to be linked with tumorigenesis.
NOTCH3	NOTCH3	The third discovered human homologue of the Drosophilia melanogaster type I membrane protein notch. Involved in intercellular signalling pathways in neural development.
PIK3CA	Phosphatidylinositol-4,5-bisphosphate 3-kinase catalytic subunit alpha	Oncogenic and implicated in cancer development.
Ras genes	Rat sarcoma	Regulation of cell signalling pathways, cell growth and death.
SERPINB4	Serpin family B member 4	Inactivation of granzyme M, an enzyme that kills tumour cells. Highly expressed in tumour cells.
TP53	Tumor protein p53	Regulates transcription, expression of target genes, thereby inducing cell cycle arrest, apoptosis, senescence, DNA repair, or changes in metabolism. Implicated in a number of cancers.
TP63	Tumor protein p63	Involved in the following processes in skin development and maintenance, adult stem/progenitor cell regulation, heart development and premature aging.

2.4.5 Interaction Studies

Combinations of specific genotypes with environmental factors were also reported to be associated with ESCC in a number of studies (Table 2). The main two environmental factors studied were smoking and alcohol consumption. The interaction between smoking and alcohol status and specific genotypes was measured and reported as frequency (percentage) and assessed using p values and odds ratios in eight genetic susceptibility studies (27, 29, 30, 33, 34, 36, 40-42). Four studies showed statistically significant associations between both alcohol and smoking status and variants in the *CYP3A5, CYP2E1, GST* and *NAT2* genes (29, 34, 40). *SULT1A1* variants were associated with smoking status only (30). Other interaction studies included wood/charcoal use and mutations in the *GST* genes (41), as well as red and white meat intake and SNPs in *NAT1/2* genes (34).

2.5 Discussion

2.5.1 General Systematic Review Findings

In this study, we systematically evaluated the genetic variants reported to be associated with ESCC in African populations providing the first systematic review on genetic factors of ESCC in this region. Of all studies that have been published on genetic association to ESCC in the African populations only 23 fit our selection criteria. It was clear from the beginning that there is a dearth of information on this topic. Our analysis showed that 25 germline SNPs were reported to be associated with ESCC in the South African population. However, none of these SNPs were repeated in three or more independent studies, hence a meta-analysis was not possible. Additionally, only three (*ALDH2 and CYP2E1*) of the 23 genes were analyzed in two independent studies, but testing for different SNPs. We determined that it was unlikely that the two *ALDH2* SNPs analysed were the same SNPs. This is because the MAFs were significantly different and whilst one SNP had a protective effect (reduced risk), the other increased risk. The lack of studies re-assessing the same genetic variants poses a major hurdle in validating existing evidence on the association between genetic variants and ESCC development. This makes resolving the genetic etiology of ESCC in African populations difficult.

2.5.2 Genetic Susceptibility to ESCC

Of the 25 SNPs from the genetic susceptibility studies which showed an association to ESCC, we concluded that results on 16 SNPs had the potential to be reliable and reproducible due to the larger sample sizes. Ten of the SNPs were reported to increase the risk of ESCC, whilst six were reported to reduce the risk. However, it was noted that the majority (11) of these SNPs showed association in the South African Admixed population and the studies did not report controlling for population stratification. This is a highly admixed population (48), in which the predominant ancestral lines are Khoesan (32-43%), Bantu speaking Africans (20-36%), European

(21-28%) and Asian (9-11%) (49). This diverse population is a result of South Africa's colonial and trade history, and constitutes 9% of the total South African population (49). Genetic variability can also be seen in the Black South African population (48). Without controlling for population stratification, the reproducibility of these results is questionable. It is, however, important to note that the majority of these studies were carried out several years ago and information on population stratification and methods to detect it may not have been available as yet.

Re-examination of common SNPs from the Chinese population was done in three of the studies (26, 27, 39), but the findings were not conclusive. It is possible that there may be population-specific differences influencing the genetic etiology of ESCC in the African populations. This may also point to the role of environmental factors contributing to the genetic susceptibility to ESCC through gene-environment interactions.

2.5.3 Somatic Changes in ESCC

Forty-four somatic variants were reported but only two were significantly associated with ESCC. The paucity of information was also evident in the somatic variant studies. There were significantly fewer studies (eight) on somatic variants than on genetic susceptibility (16). The molecular profiling of tumours is of great importance as it is relevant in the development of targeted cellular therapeutics. One gene (*CDKN2A*) was analyzed in two studies, but these studies focussed on a different variant. Another gene, TP53 was analyzed in four studies, but two studies analysed different parts of the gene and two had no control data. It was evident, however, that the WES study provided with a wider variety of genetic variants associated with ESCC (45). The WES study overall had the largest number of genetic variants of all the 22 studies, and was able to identify variants in an unbiased manner.

2.5.4 Common Limitations among the African Studies

There were no GWAS studies among the studies we analysed, but reports from the Chinese and European studies demonstrated that GWAS studies are able to successfully identify common genetic variants associated with ESCC (3). To date, GWAS has successfully identified more than 700 loci for cancer risk. However these studies have been predominantly done in populations of European ancestry (80%), with African and Latin American populations contributing less than 1% (Van Loon et al., 2018). A shift to WES and GWAS studies on the African populations might, therefore, yield better results in identifying variants that play a role in ESCC development. The African Esophageal Cancer Consortium, which was initiated in 2016 by African investigators and International partners, released a call to action to, among other priority activities, increase molecular research on esophageal cancer in Africa, particularly GWAS and genomic profiling (Van Loon et al., 2018).

One of the main deficiencies in the studies was that the majority of the genetic susceptibility studies did not report a power calculation, or a genotyping error and this may have resulted in studies being underpowered and with increased type II error. Few studies reported correction for multiple testing, however many of the studies were not analysing multiple variants at the same time. The lack of correction for multiple testing, therefore, is not a reflection on the methodological quality. Very few studies reported NCBI rs numbers. In most studies, the diagnosis of ESCC in patients was adequately defined with no ambiguity on the number of patients with ESCC. There were, however, three studies that combined samples from patients with squamous cell and adenocarcinoma into one case group, which could introduce bias (31, 32, 36).

It is important to note that rs numbers were poorly documented in the majority of the studies assessed in this systematic review. Additionally, in many of these studies, the positions of the SNPs using genome coordinates were not reported, hence making it difficult to locate the SNPs. In the absence of an rs number, we recommend that authors report the position using genome coordinates and the version of the genome used as a reference.

The somatic variant studies also had adequately defined ESCC diagnosis for the majority of the studies. Whilst the variant classification and type was reported by most studies, there was no confirmation of the results (except for two studies). Overall, for both the germline and somatic variant studies, the quality of reporting for the majority of the studies was not adequate. Other important limitations and biases are the lack of controlling for population stratification and small sample sizes in the study populations, which may have led to unreliable results.

2.5.5 Limitations of the Systematic Review

Whilst we did a comprehensive search in four of the main literature databases, it is possible that we could have missed some non-English studies on African populations.

Because of the lack of replication and validation studies, we could not carry out a meta-analysis in the current study. Furthermore, we did not re-analyse the data, and relied on reported p values and odds ratios for descriptive analysis.

2.6 Conclusions

Whilst this review has highlighted a number of genes that may be potentially associated with ESCC in the African populations, limitations such as lack of reproducibility, quality of reporting and quality of assessment remain a major concern. The implications of having these inconsistencies, and lack of reproducibility is that the genetic etiology of ESCC in Africa will continue to be unclear. The region lags behind in contributing to genetic knowledge and literature on ESCC. Importantly any preventative, diagnostic or therapeutic interventions cannot be effectively identified or applied in these populations.

The identification of genetic markers of esophageal cancer susceptibility has clear translational benefits to African populations in understanding the underlying disease risk and heritability. Benefits include the utilization of genetic information to improve risk prediction, which can be translated into prevention and screening programs relevant and specific to the African population. These studies also play a role in identifying and quantifying the interactions of modifiable environmental risk factors which interact with these genetic variants, and hence provide a platform for better targeted interventions. The ability to sufficiently translate genetic research on the African population is dependent on more genetic studies done on the population.

Our recommendations are that more and larger genetic studies be done on the African populations, particularly focussing on WES and GWAS approaches. This will require multinational collaborations between the African countries.

2.7 Authors' contributions

VL, VS and HS carried out literature searches. HS and HK appraised the articles, summarized results, prepared the tables and figures and drafted the manuscript. VS and VL reviewed the articles and edited the manuscript. VS and HK conceptualized the idea for the research, obtained funding, supervised the project and wrote sections of the manuscript. VL provided specialist expertise and knowledge, and critically reviewed the manuscript. GT carried out the r^2 analyses, prepared the r^2 figure and table and critically reviewed and revised the manuscript. All authors approved the final version of the manuscript.

2.8 Funding

This work was supported by the African Cancer Institute. HS acknowledges the Beit Trust hardship fund for providing a Doctoral Scholarship and the Collaboration for Evidence-based Healthcare and Public Health in Africa (CEBHA+), as part of the Research Networks for Health Innovation in Sub-Saharan Africa Funding Initiative of the German Federal Ministry of Education and Research. HK was supported by the Faculty of Medicine and Health Sciences, Stellenbosch University, South Africa. GT was supported by the South African Tuberculosis Bioinformatics Initiative (SATBBI), a Strategic Health Innovation Partnership grant from the South African Medical Research Council and South African Department of Science and Technology.

2.9 Supplementary Material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org.

2.10 Conflict of Interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

2.11 References

1. Murphy G, McCormack V, Abedi-Ardekani B, Arnold M, Camargo MC, Dar NA, et al. International cancer seminars: a focus on esophageal squamous cell carcinoma. *Ann Oncol* (2017) 28(9):2086-93. Epub 2017/09/16. doi: 10.1093/annonc/mdx279. PubMed PMID: 28911061.

2. Kaz AM, Grady WM. Epigenetic biomarkers in esophageal cancer. *Cancer letters* (2014) 342(2):193-9. Epub 2012/03/13. doi: 10.1016/j.canlet.2012.02.036. PubMed PMID: 22406828; PubMed Central PMCID: PMCPMC3395756.

3. Abnet CC, Arnold M, Wei WQ. Epidemiology of Esophageal Squamous Cell Carcinoma. *Gastroenterology* (2017) 154(2):360-73. Epub 2017/08/22. doi: doi:10.1053/j.gastro.2017.08.023. PubMed PMID: 28823862; PubMed Central PMCID: PMCPMC5836473.

4. Yazbeck R, Jaenisch SE, Watson DI. From blood to breath: New horizons for esophageal cancer biomarkers. *World Journal of Gastroenterology* (2016) 22(46):10077-83. doi: 10.3748/wjg.v22.i46.10077. PubMed PMID: PMC5155166.

5. Munishi MO, Hanisch R, Mapunda O, Ndyetabura T, Ndaro A, Schüz J, et al. Africa's oesophageal cancer corridor: Do hot beverages contribute? *Cancer Causes Control* (2015) 26(10):1477-86. doi: 10.1007/s10552-015-0646-9. PubMed PMID: 26245249.

6. Schaafsma T, Wakefield J, Hanisch R, Bray F, Schüz J, Joy EJM, et al. Africa's Oesophageal Cancer Corridor: Geographic Variations in Incidence Correlate with Certain Micronutrient Deficiencies. *PloS one* (2015) 10(10):e0140107-e. doi: 10.1371/journal.pone.0140107. PubMed PMID: 26448405.

7. Kayamba V, Bateman AC, Asombang AW, Shibemba A, Zyambo K, Banda T, et al. HIV infection and domestic smoke exposure, but not human papillomavirus, are risk factors for esophageal squamous cell carcinoma in Zambia: a case-control study. *Cancer medicine* (2015) 4(4):588-95. Epub 2015/02/03. doi: 10.1002/cam4.434. PubMed PMID: 25641622; PubMed Central PMCID: PMCPMC4402073.

8. Chen X, Winckler B, Lu M, Cheng H, Yuan Z, Yang Y, et al. Oral Microbiota and Risk for Esophageal Squamous Cell Carcinoma in a High-Risk Area of China. *PloS one* (2015) 10(12):e0143603. Epub 2015/12/08. doi: 10.1371/journal.pone.0143603. PubMed PMID: 26641451; PubMed Central PMCID: PMCPMC4671675.

9. Huang FL, Yu SJ. Esophageal cancer: Risk factors, genetic association, and treatment. *Asian J Surg* (2018) 41(3):210-5. Epub 2016/12/18. doi: 10.1016/j.asjsur.2016.10.005. PubMed PMID: 27986415.

10. Pink RC, Bailey TA, Iputo JE, Sammon AM, Woodman AC, Carter DR. Molecular basis for maize as a risk factor for esophageal cancer in a South African population via a prostaglandin E2 positive feedback mechanism. *Nutrition and cancer* (2011) 63(5):714-21. Epub 2011/06/15. doi: 10.1080/01635581.2011.570893. PubMed PMID: 21667399.

11. Sewram V, Sitas F, O'Connell D, Myers J. Tobacco and alcohol as risk factors for oesophageal cancer in a high incidence area in South Africa. *Cancer epidemiology* (2016) 41:113-21. Epub 2016/02/24. doi: 10.1016/j.canep.2016.02.001. PubMed PMID: 26900781.

12. Sewram V, Sitas F, O'Connell D, Myers J. Diet and esophageal cancer risk in the Eastern Cape Province of South Africa. *Nutrition and cancer* (2014) 66(5):791-9. Epub 2014/06/01. doi: 10.1080/01635581.2014.916321. PubMed PMID: 24877989.

13. Baba Y, Yamamura K, Nakagawa S, Mima K, Ishimoto T, Iwatsuki M, et al. Abstract 4930: Genetic and epigenetic characteristics of esophageal cancer tissues with microbiome fusobacterium nucleatum. *Cancer Research* (2017) 77(13 Supplement):4930-. doi: doi:10.1158/1538-7445.am2017-4930.

14. Coleman HG, Xie SH, Lagergren J. The Epidemiology of Esophageal Adenocarcinoma. *Gastroenterology* (2018) 154(2):390-405. Epub 2017/08/07. doi: 10.1053/j.gastro.2017.07.046. PubMed PMID: 28780073.

15. He F, Liu C, Zhang R, Hao Z, Li Y, Zhang N, et al. Association between the Glutathione-S-transferase T1 null genotype and esophageal cancer susceptibility: A meta-analysis involving 11,163 subjects. *Oncotarget* (2018) 9(19):15111-21. doi: 10.18632/oncotarget.24534.

16. Kumar P, Rai V. MTHFR C677T polymorphism and risk of esophageal cancer: An updated meta-analysis. *Egyptian Journal of Medical Human Genetics* (2018) 19(4):273-84. doi: 10.1016/j.ejmhg.2018.04.003.

17. Lesseur C, Ferreiro-Iglesias A, McKay JD, Bossé Y, Johansson M, Gaborieau V, et al. Genome-wide association meta-analysis identifies pleiotropic risk loci for aerodigestive squamous cell cancers. *PLoS Genet* (2021) 17(3):e1009254. Epub 2021/03/06. doi: 10.1371/journal.pgen.1009254. PubMed PMID: 33667223; PubMed Central PMCID: PMCPMC7968735

18. Little J, Higgins JP, Ioannidis JP, Moher D, Gagnon F, von Elm E, et al. STrengthening the REporting of Genetic Association Studies (STREGA)--an extension of the STROBE statement. *Genetic epidemiology* (2009) 33(7):581-98. Epub 2009/03/12. doi: 10.1002/gepi.20410. PubMed PMID: 19278015.

19. Naidoo R, Ramburan A, Reddi A, Chetty R. Aberrations in the mismatch repair genes and the clinical impact on oesophageal squamous carcinomas from a high incidence area in South Africa. *Journal of clinical pathology* (2005) 58(3):281-4. Epub 2005/03/01. doi: 10.1136/jcp.2003.014290. PubMed PMID: 15735161; PubMed Central PMCID: PMCPMC1770598.

20. Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, Korbel JO, et al. A global reference for human genetic variation. *Nature* (2015) 526(7571):68-74. Epub 2015/10/04. doi: 10.1038/nature15393. PubMed PMID: 26432245; PubMed Central PMCID: PMCPMC4750478.

21. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience* (2015) 4:7. Epub 2015/02/28. doi: 10.1186/s13742-015-0047-8. PubMed PMID: 25722852; PubMed Central PMCID: PMCPMC4342193.

22. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, et al. The variant call format and VCFtools. *Bioinformatics (Oxford, England)* (2011) 27(15):2156-8. Epub 2011/06/10. doi: 10.1093/bioinformatics/btr330. PubMed PMID: 21653522; PubMed Central PMCID: PMCPMC3137218.

23. Adams CH, Werely CJ, Victor TC, Hoal EG, Rossouw G, van Helden PD. Allele frequencies for glutathione S-transferase and N-acetyltransferase 2 differ in African population groups and may be associated with oesophageal cancer or tuberculosis incidence. *Clinical chemistry and laboratory medicine* (2003) 41(4):600-5. Epub 2003/05/16. doi: doi:10.1515/cclm.2003.090. PubMed PMID: 12747608.

24. Jaskiewicz K, De Groot KM. p53 gene mutants expression, cellular proliferation and differentiation in oesophageal carcinoma and non-cancerous epithelium. *Anticancer research* (1994) 14(1A):137-40.

25. Uys P, van Helden PD. On the nature of genetic changes required for the development of esophageal cancer. *Molecular carcinogenesis* (2003) 36(2):82-9. Epub 2003/01/31. doi: 10.1002/mc.10100. PubMed PMID: 12557264.

26. Bye H, Prescott NJ, Lewis CM, Matejcic M, Moodley L, Robertson B, et al. Distinct genetic association at the PLCE1 locus with oesophageal squamous cell carcinoma in the South African population. *Carcinogenesis* (2012) 33(11):2155-61. Epub 2012/08/07. doi: 10.1093/carcin/bgs262. PubMed PMID: 22865593; PubMed Central PMCID: PMCPMC3584964.

27. Bye H, Prescott NJ, Matejcic M, Rose E, Lewis CM, Parker MI, et al. Population-specific genetic associations with oesophageal squamous cell carcinoma in South Africa. *Carcinogenesis* (2011) 32(12):1855-61. Epub 2011/09/20. doi: 10.1093/carcin/bgr211. PubMed PMID: 21926110; PubMed Central PMCID: PMCPMC3220606.

28. Chelule PK, Pegoraro RJ, Gqaleni N, Dutton MF. The frequency of cytochrome P450 2E1 polymorphisms in Black South Africans. *Disease markers* (2006) 22(5-6):351-4. Epub 2007/02/01. PubMed PMID: 17264406; PubMed Central PMCID: PMCPMC3851364.

29. Dandara C, Ballo R, Parker MI. CYP3A5 genotypes and risk of oesophageal cancer in two South African populations. *Cancer letters* (2005) 225(2):275-82. Epub 2005/06/28. doi: 10.1016/j.canlet.2004.11.004. PubMed PMID: 15978331.

30. Dandara C, Li DP, Walther G, Parker MI. Gene-environment interaction: the role of SULT1A1 and CYP3A5 polymorphisms as risk modifiers for squamous cell carcinoma of the oesophagus. *Carcinogenesis* (2006) 27(4):791-7. Epub 2005/11/08. doi: 10.1093/carcin/bgi257. PubMed PMID: 16272171.

31. Dietzsch E, Laubscher R, Parker MI. Esophageal cancer risk in relation to GGC and CAG trinucleotide repeat lengths in the androgen receptor gene. *International journal of cancer* (2003) 107(1):38-45. Epub 2003/08/20. doi: 10.1002/ijc.11314. PubMed PMID: 12925954.

32. Eltahir HA, Adam AA, Yahia ZA, Ali NF, Mursi DM, Higazi AM, et al. p53 Codon 72 arginine/proline polymorphism and cancer in Sudan. *Molecular biology reports* (2012) 39(12):10833-6. Epub 2012/10/12. doi: 10.1007/s11033-012-1978-0. PubMed PMID: 23053979.

33. Li D-P, Dandara C, Walther G, Parker MI. Genetic polymorphisms of alcohol metabolising enzymes: their role in susceptibility to oesophageal cancer. *Clinical chemistry and laboratory medicine* (2008) 46(3):323-8.

34. Matejcic M, Vogelsang M, Wang Y, Parker MI. Erratum to: NAT1 and NAT2 genetic polymorphisms and environmental exposure as risk factors for oesophageal squamous cell carcinoma: a case-control study. *BMC cancer* (2015) 15:658. Epub 2015/10/09. doi: 10.1186/s12885-015-1681-3. PubMed PMID: 26447020; PubMed Central PMCID: PMCPMC4597372.

35. Strickland NJ, Matsha T, Erasmus RT, Zaahl MG. Molecular analysis of ceruloplasmin in a South African cohort presenting with oesophageal cancer. *International journal of cancer* (2012) 131(3):623-32. Epub 2011/09/09. doi: 10.1002/ijc.26418. PubMed PMID: 21901748.

36. Vogelsang M, Wang Y, Veber N, Mwapagha LM, Parker MI. The cumulative effects of polymorphisms in the DNA mismatch repair genes and tobacco smoking in oesophageal cancer risk. *PloS one* (2012) 7(5):e36962. Epub 2012/05/25. doi: 10.1371/journal.pone.0036962. PubMed PMID: 22623965; PubMed Central PMCID: PMCPMC3356375.

37. Vos M, Adams CH, Victor TC, van Helden PD. Polymorphisms and mutations found in the regions flanking exons 5 to 8 of the TP53 gene in a population at high risk for esophageal cancer in South Africa. *Cancer genetics and cytogenetics* (2003) 140(1):23-30. Epub 2003/01/29. PubMed PMID: 12550754.

38. Zaahl MG, Warnich L, Victor TC, Kotze MJ. Association of functional polymorphisms of SLC11A1 with risk of esophageal cancer in the South African Colored population. *Cancer genetics and cytogenetics* (2005) 159(1):48-52. Epub 2005/04/30. doi: 10.1016/j.cancergencyto.2004.09.017. PubMed PMID: 15860357.

39. Chen WC, Bye H, Matejcic M, Amar A, Govender D, Khew YW, et al. Association of genetic variants in CHEK2 with oesophageal squamous cell carcinoma in the South African Black population. *Carcinogenesis* (2019). doi: https://dx.doi.org/10.1093/carcin/bgz026.

40. Li D, Dandara C, Parker MI. Association of cytochrome P450 2E1 genetic polymorphisms with squamous cell carcinoma of the oesophagus. *Clinical chemistry and laboratory medicine* (2005) 43(4):370-5. Epub 2005/05/19. doi: 10.1515/cclm.2005.067. PubMed PMID: 15899651.

41. Li D, Dandara C, Parker MI. The 341C/T polymorphism in the GSTP1 gene is associated with increased risk of oesophageal cancer. *BMC genetics* (2010) 11:47. Epub 2010/06/15. doi: 10.1186/1471-2156-11-47. PubMed PMID: 20540773; PubMed Central PMCID: PMCPMC2891604.

42. Matejcic M, Li D, Prescott NJ, Lewis CM, Mathew CG, Parker MI. Association of a deletion of GSTT2B with an altered risk of oesophageal squamous cell carcinoma in a South African population: a case-control study. *PloS one* (2011) 6(12):e29366. Epub 2012/01/05. doi: 10.1371/journal.pone.0029366. PubMed PMID: 22216261; PubMed Central PMCID: PMCPMC3246501.

43. Dietzsch E, Parker M. *Infrequent Somatic Deletion of the 5' Region of the COL1A2 Gene in Oesophageal Squamous Cell Cancer Patients*(2002). 941-5 p.

44. Gamieldien W, Victor TC, Mugwanya D, Stepien A, Gelderblom WC, Marasas WF, et al. p53 and p16/CDKN2 gene mutations in esophageal tumors from a high-incidence area in South Africa. *International journal of cancer* (1998) 78(5):544-9. Epub 1998/11/10. PubMed PMID: 9808520.

45. Liu W, Snell JM, Jeck WR, Hoadley KA, Wilkerson MD, Parker JS, et al. Subtyping sub-Saharan esophageal squamous cell carcinoma by comprehensive molecular analysis. *JCI insight* (2016) 1(16):e88755. Epub 2016/10/14. doi: 10.1172/jci.insight.88755. PubMed PMID: 27734031; PubMed Central PMCID: PMCPMC5053149.

46. Patel K, Mining S, Wakhisi J, Gheit T, Tommasino M, Martel-Planche G, et al. TP53 mutations, human papilloma virus DNA and inflammation markers in esophageal squamous cell carcinoma from the Rift Valley, a high-incidence area in Kenya. *BMC research notes* (2011) 4:469. Epub 2011/11/02. doi: 10.1186/1756-0500-4-469. PubMed PMID: 22040862; PubMed Central PMCID: PMCPMC3216406.

47. Victor T, Du Toit R, Jordaan AM, Bester AJ, van Helden PD. No evidence for point mutations in codons 12, 13, and 61 of the ras gene in a high-incidence area for esophageal and gastric cancers. *Cancer research* (1990) 50(16):4911-4. Epub 1990/08/15. PubMed PMID: 2199031.

48. Chimusa ER, Daya M, Moller M, Ramesar R, Henn BM, van Helden PD, et al. Determining ancestry proportions in complex admixture scenarios in South Africa using a novel proxy ancestry selection method. *PloS one* (2013) 8(9):e73971. Epub 2013/09/26. doi: 10.1371/journal.pone.0073971. PubMed PMID: 24066090; PubMed Central PMCID: PMCPMC3774743.

49. de Wit E, Delport W, Rugamika CE, Meintjes A, Moller M, van Helden PD, et al. Genome-wide analysis of the structure of the South African Coloured Population in the Western Cape. *Human genetics* (2010) 128(2):145-53. Epub 2010/05/22. doi: 10.1007/s00439-010-0836-1. PubMed PMID: 20490549.

Chapter 3: Environmental and life-style risk factors for esophageal squamous cell carcinoma in Africa: A systematic review

Table of Contents

3.1 Abstract

The African esophageal squamous cell carcinoma (ESCC) corridor, which runs from Ethiopia down to South Africa, is an esophageal cancer (EC) hotspot. High incidence and mortality rates of EC have been reported from this region. We systematically assessed the evidence on environmental and life-style related risk factors associated with ESCC in African populations. We followed the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines. We carried out a comprehensive search of all African published studies up to January 2020 using PubMed, Embase, Scopus and African Index Medicus databases. We identified 31 studies with measures of association [odds ratio (OR), relative risk (RR), and 95% confidence intervals (95%CI)], which reported smoking and alcohol consumption, socioeconomic status, diet, polycyclic aromatic hydrocarbon exposure, consumption of hot food and beverages, oral health, infectious agents, esophageal inflammation, family history of cancer and non-acid gastro-esophageal reflux as risk factors for ESCC in Africa. We performed a meta-analysis on 27 studies investigating tobacco, alcohol use, combined tobacco and alcohol use, polycyclic aromatic hydrocarbon exposure, esophageal injury and fruit and vegetable consumption. We found adverse associations between ESCC risk and all the risk factors. Analysis of fruit and vegetable consumption showed a protective effect. Using population attributable fraction (PAF) analysis, we calculated the proportion of ESCC attributable to tobacco (17%), alcohol use (13%), combined tobacco and alcohol use (23%), polycyclic aromatic hydrocarbon exposure (5%), esophageal injury (17%) and fruit and vegetable consumption (-11%). In conclusion, tobacco smoking and alcohol consumption were the most studied risk factors overall, whilst esophageal injury emerged as an important risk factor for ESCC in our meta-analysis and population attributable fraction analysis, although understudied. Our results point to a multifactorial etiology of ESCC in African populations.

3.2 Introduction

Esophageal cancer (EC) is a lethal malignancy ranking as the 6th most common cause of cancer mortality and the 7th most common cancer worldwide.(1) In 2018, 572,034 new cases and 508,585 deaths were reported, indicative of the high fatality associated with EC diagnosis.(1) About 80% of EC cases and deaths occur in economically developing countries, where esophageal squamous cell carcinoma (ESCC) is more prevalent.(2, 3) ESCC is the major subtype contributing approximately 90% of all ECs, in contrast to esophageal adenocarcinoma, which is more prevalent in the Western countries.(3) The high mortality rate is attributable to late diagnosis with patients presenting at advanced stage due to a lack of early symptoms. ESCC has a peculiar geographical distribution with high incidence rates reported from China to Iran, parts of South America and from Eastern to Southern Africa.(4, 5) The variability in incidence between high risk and low risk areas across the globe has been reported to be up to 10-fold.(6) Variations within regions and countries have also been noted. This peculiar distribution draws questions on the specificity of certain risk factors to particular regions.

The African EC corridor, which runs from Kenya down to South Africa on the easterly side of Africa, is an ESCC hotspot region. This African corridor includes Ethiopia, Burundi, Malawi, Kenya, Uganda, Tanzania, Zimbabwe, Madagascar and South Africa.(7, 8) High incidence rates from this corridor have been reported as far back as 1969.(9) ESCC cases from the African cancer corridor are also reported to be younger than those found elsewhere in the world.(10) This presents with possible unique risk factors for this region.(2)

EC has a multi-factorial etiology. The risk factors reported worldwide comprise lifestyle, environmental and genetic factors. The lifestyle and environmental factors include smoking, alcohol consumption, poor diet, micronutrient deficiency, exposure to polycyclic aromatic hydrocarbons (PAHs) through cooking and heating methods, esophageal thermal injury, obesity, infectious agents, low socioeconomic status, and exposure to contaminants which have carcinogenic effects.(6, 11-13) Genetic basis and susceptibility to esophageal carcinoma has also been studied, with reports of single nucleotide polymorphisms (SNPs), genomic alterations and epigenetic modifications contributing to tumour development.(6, 14, 15) Familial syndromes have been reported to be associated with increased risk of malignancy, including tylosis and Fanconi anemia.(16)

Despite advances in management and treatment, ESCC prognosis is still poor with a survival rate of <5% in economically developing countries.(4) The etiology of ESCC and the reasons for the high EC burden in Africa are not well understood. The rapid fatality of the cancer, poor prognosis, and contribution of reported modifiable risk factors make ESCC research important in Africa and worldwide. A number of studies have been done in Africa looking at the association between risk factors and ESCC. This body of evidence, when systematically assessed and analysed, will shed light on the epidemiology of ESCC in the African populations. It will also substantiate the role of reported risk factors on esophageal carcinogenesis, shed light on emerging risk factors, and provide knowledge on the pathobiology of EC. Improved understanding of EC is required to design better prevention and treatment modalities.

The aim of this systematic review was to provide an in-depth analysis of key environmental, and lifestyle related factors associated with ESCC development in African populations and perform a meta-analysis and population attributable fraction (PAF) analysis. Whilst the meta-analysis provides information on the overall risk of a specific risk factor to disease, PAF is the proportional reduction in disease that would occur in a population if exposure to a risk factor were modified or removed. Genetic factors were not included in the current study, since they have been reported recently in a separate study.(17) The aim of this study was achieved through: 1) critical appraisal of reported African studies on known and emerging risk factors; 2) data synthesis through pooled analysis of each risk factor using meta-analysis; 3) quantifying contribution of risk factors to disease burden using PAF analysis; and 4) comparison of risk factors reported in other ESCC high-risk regions.

3.3 Materials and methods

3.3.1 Study design

The study assessed all environmental and lifestyle risk factors reported in relevant African literature (cross sectional, case-control, and cohort studies) and tested for an association with ESCC development or progression. We followed the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines.(18) To assess the quality of methods and reporting of the published studies, the Joanna Briggs Institute Meta Analysis of Statistics Assessment and Review Instrument (JBI-MAStARI) for Cohort and Case-Control Studies was used.(19)

3.3.2 Data sources, search strategy and extraction

We carried out a literature search on all published African ESCC studies up to January 2020. We developed a comprehensive set of search terms subjectively and iteratively. We searched the following electronic bibliographic databases without time or language limits: Medline (PubMed), Embase (OViD), Scopus, African index medicus, and Africa-wide information (EbsCOHost). We also checked the reference lists of potentially relevant articles for additional citations and used the "related citations" search key in PubMed to identify similar papers.

We checked Medline (PubMed) to identify controlled vocabulary (MeSH) terms related to EC, and identified text keywords based on our knowledge of the field (Table 1). Medline search terms were modified for other electronic databases to conform to their search functions.

Two authors (HS and VS) carried out screening for eligible studies. First, the two authors read the titles and abstracts independently and then met to finalise an initial list. Full articles of the studies selected, based on the initial screening, were read and assessed for inclusion to the systematic review. Figure 3.1 shows the outline for selection of eligible studies. Data extraction was carried out by two authors (HS and VS) using data extraction forms in Microsoft Excel software.

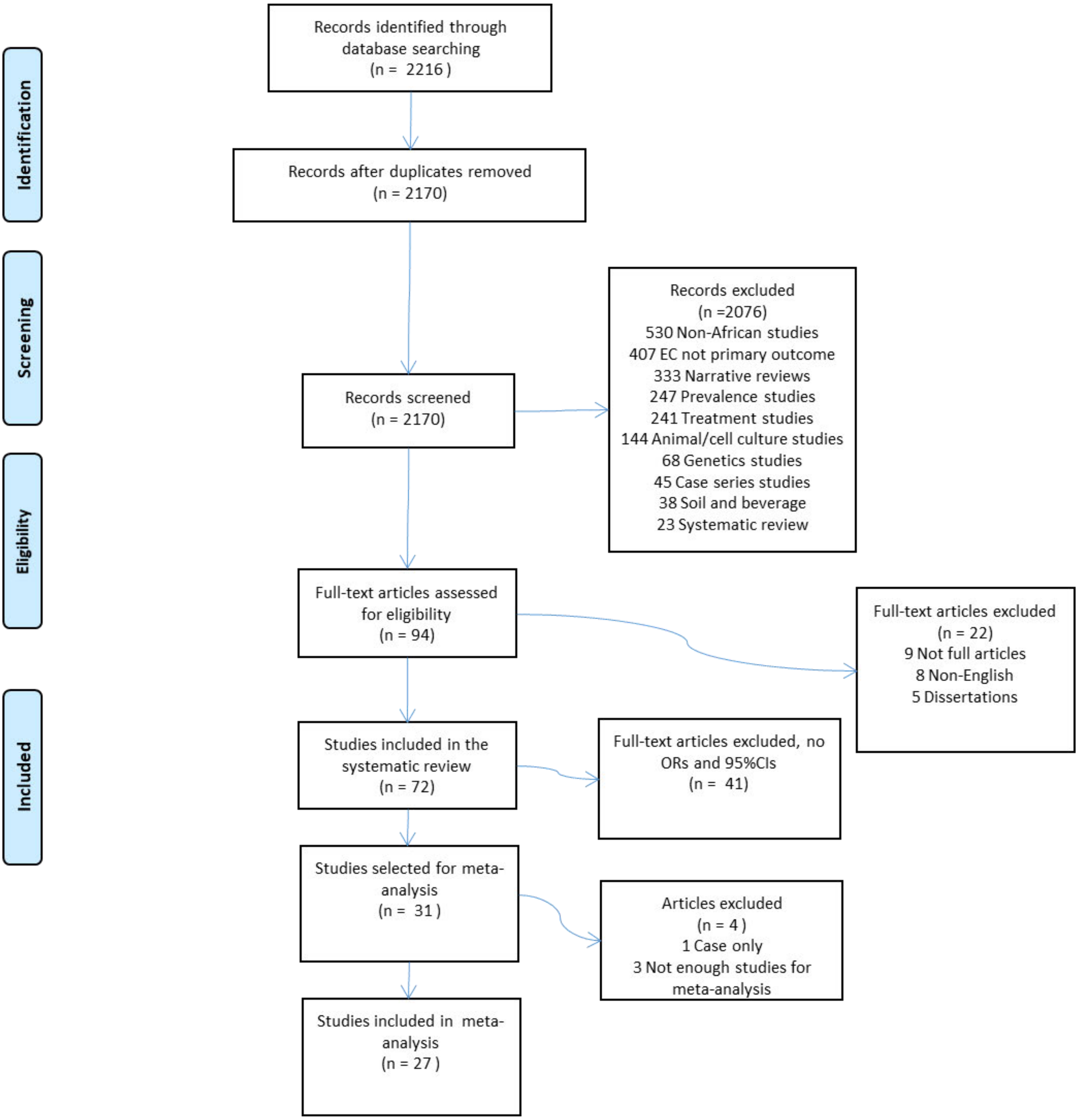

Identification
Screening
Eligibility
Included
Records identified through database searching (n = 2216)
Records after duplicates removed (n = 2170)
Records screened (n = 2170)
Records excluded (n =2076)
530 Non-African studies
407 EC not primary outcome
333 Narrative reviews
247 Prevalence studies
241 Treatment studies
144 Animal/cell culture studies
68 Genetics studies
45 Case series studies
38 Soil and beverage
23 Systematic review
Full-text articles assessed for eligibility (n = 94)
Full-text articles excluded (n = 22)
9 Not full articles
8 Non-English
5 Dissertations
Studies included in the systematic review (n = 72)
Full-text articles excluded, no ORs and 95%CIs (n = 41)
Studies selected for meta-analysis (n = 31)
Articles excluded (n = 4)
1 Case only
3 Not enough studies for meta-analysis
Studies included in meta-analysis (n = 27)

Figure 3.1*: Outline of the study using the PRISMA diagram*

3.3.4 Quality of methodology and reporting assessment, and data extraction

Quality of the methods and reporting used in the published studies was assessed using a quality assessment tool adapted from the JBI-MAStARI.(19) The selection, stage of EC in patients, confounding factors, assessment of outcomes in cases and controls, reliability of assessment of outcomes methods, and statistical analysis used. External validity and representativeness of sample to the population was confirmed if the study had at least 150 cases and/or controls. Statistical analysis was measured by determining if the correct analysis was used as well as if enough information was reported regarding the analysis used. Studies were classified as low quality (score of 1–3), moderate quality (score of 4–6) or high quality (score of 7+). Only studies, which reported on measures of association [odds ratio (OR), relative risk (RR), and 95% confidence intervals (95%CI)], were assessed for quality of methodology and reporting.

3.3.5 Data analysis

Where high quality data were available from a minimum of three studies, pooled statistical analysis was carried out using meta-analysis and attributable risk analysis. Where statistical pooling was not possible, the results were presented in a narrative form.

Meta-analysis was performed using R statistical software.(20) The metagen R package was the main package used for the analysis. A random effects model was used in the analysis, using the Sidik-Jonkman estimator. To assess for heterogeneity in the meta-analysis, a test for heterogeneity was done. A test for heterogeneity and between study variance was done as part of the meta-analysis using the Chi-squared test. Outlier detection method was used to remove studies with extreme effect sizes from the meta-analysis.(21), shown in the results section. If a study's confidence interval did not overlap with the confidence interval of the pooled effect from the initial meta-analysis it was considered an outlier. Influence analysis was further performed to detect studies in the meta-analysis exerted high influence on the overall results. This was done by repeatedly recalculating the results of the meta-analysis, and each time leaving out one study.(21) This allowed for better assessment of studies that influenced or distorted the overall pooled effect. To further explore the robustness of the meta-analyses, Graphic Display of Heterogeneity (GOSH) plots were generated to identify the patterns of effect sizes and heterogeneity in the data.(22) This is a more vigorous and computationally intensive method. A second meta-analysis was done after the removal of outliers.

Publication bias analysis was done through funnel plots to determine and visualize whether small studies with small effect sizes are missing from the meta-analysis, and this was visualised through funnel plots. Egger's test of the intercept was done to test for funnel plot asymmetry.

Attributable risk is used to determine how much of an outcome is attributable to a particular risk factor, and hence provides with estimates (proportions or percentages) of how an outcome can be influenced with the removal or reduction of that risk factor. The PAF was computed using the formula (23):

$$\sum \frac{p * (RR - 1)}{P * (RR - 1) + 1}$$

where p is the proportion of people in the population exposed to the risk factor, and RR is the relative risk. Where only ORs were reported, we converted OR to RR using data provided in the studies.(24) The formula was executed using a function that we generated in R statistical software. The overall PAF value for each risk factor was computed using weighted PAF values. The final PAF values where therefore calculated using the following equation, incorporating weighted PAF values.

$$\sum \frac{p}{\sum p} * PAF$$

Where p is the proportion of people in the population exposed to the risk factor and ∑p is the sum of p.

We calculated the combined PAF from exposure to tobacco smoking, alcohol consumption, esophageal injury, and PAH using the following equation (25).

$$PAF = 1 - (1 - PAF_1) * (1 - PAF_2) * (1 - PAF_3) * (1 - PAF_4)$$

where PAF_1 is the PAF for tobacco smoking, PAF_2 is the PAF for alcohol consumption, PAF_3 is the PAF for esophageal injury, and PAF_4 is the PAF for PAH exposure. It is important to note that this equation assumes independence of exposure from the four sources.

3.4 Results

3.4.1 Review outline

A summary of the search results, as well as the screening and inclusion process for the review are presented as a PRISMA diagram in Figure 3.1. The initial search produced 2,216 articles, which were screened for duplicates and 46 duplicates were removed. The remaining 2,170 articles were screened using titles and abstracts for eligibility. A total of 2,076 articles were removed after the screening, since they were not original observational studies reporting associations between environmental and lifestyle risk factors and EC in Africa. Full text assessment was done on the remaining 94 articles. Twenty-two articles were removed for the following reasons: nine had no full text, eight were non-English articles and five were dissertations. Finally, 72 studies we included in the study for appraisal and analysis. Studies that did not report on ORs or RR and 95%CIs were not included in the meta-analysis or the PAF analysis.

3.4.2 Study selection and characteristics

Risk factors reported in the 72 included studies were smoking and alcohol consumption, low socioeconomic status, poor diet, PAH exposure, consumption of hot food and beverages, poor oral health, infectious agents, esophageal inflammation, family history of cancer and non-acid gastro-esophageal reflux. The studies were published between 1972 and 2019. The diagnostic methods for ESCC used included histopathology, barium swallow, and brush. Of the 72 studies, only 31 (43%) reported association of a risk factor to ESCC using ORs or RRs and 95%CIs. Some of these studies (n=8) reported on more than one environmental and lifestyle risk factor. The least reported characteristics were the EC stage (n=1) of the patients and the response rates of participants (n=2).

Quality of methods and reporting assessment was done on the 31 articles that reported ORs or RRs and 95%CIs and hence qualified for quantitative assessment. The quality scores are reported in Appendix Table 3A1. The majority of the articles (77.5%) were of moderate quality (score of 4–6). Five studies (16.5%) were of low quality (score of 1-3). Only two studies (6%) had high quality reporting (score of 7-9). The least reported characteristics were the EC stage of the patients and the response rates of participants.

Table 3.1 shows a list of African countries (majority from the African Esophageal Cancer Corridor) according to incidence numbers and Age Standardized Incidence Rater (ASIR) per 100,000, which had studies included in our systematic review and meta-analysis. The risk factors analyzed in each country are presented in the table. This information is further depicted in Figure 3.2 and 3.3 as pie charts.

Table 3.1: List of African countries according to incidence numbers and ASIR, which had studies included in the systematic review and meta-analysis

Population	**Incidence number**	**ASIR (World)**	**Rank in Africa according to ASIR**	**Number of studies eligible for meta-analysis**	**Number of studies included in the systematic review**	**Risk factors investigated in case control studies**
Malawi	1844	18.70	1	0	3	Alcohol, Tobacco, Infectious agents
Kenya	4380	18.40	2	4	9	Alcohol, Tobacco, Diet, PAH, Hot food and beverages, Socioeconomic status, Esophageal Injury, Family history of ESCC, Oral health, Water source
Zimbabwe	920	12.40	3	2	4	Alcohol, Tobacco, Socioeconomic status
Uganda	1749	10.80	4	2	2	Alcohol, Tobacco

United Republic of Tanzania	2516	8.90	7	0	2	Alcohol, Tobacco, Hot food and beverages
South Africa	3697	7.80	10	17	43	Alcohol, Tobacco, Infectious agents, Diet, PAH, Socioeconomic status, Esophageal Injury, Non-acid gastro esophageal reflux
Somalia	524	7.50	12	0	1	Infectious agents
Zambia	389	5.30	17	2	2	Alcohol, Tobacco, Infectious agents, Diet, PAH
Ethiopia	1752	3.00	23	3	4	Alcohol, Tobacco, Infectious agents, Diet, PAH, Hot food and beverages
Mali	190	1.90	36	0	1	Diet
Egypt	1034	1.30	38	0	1	Alcohol, Tobacco

ASIR, age standardised incidence rates

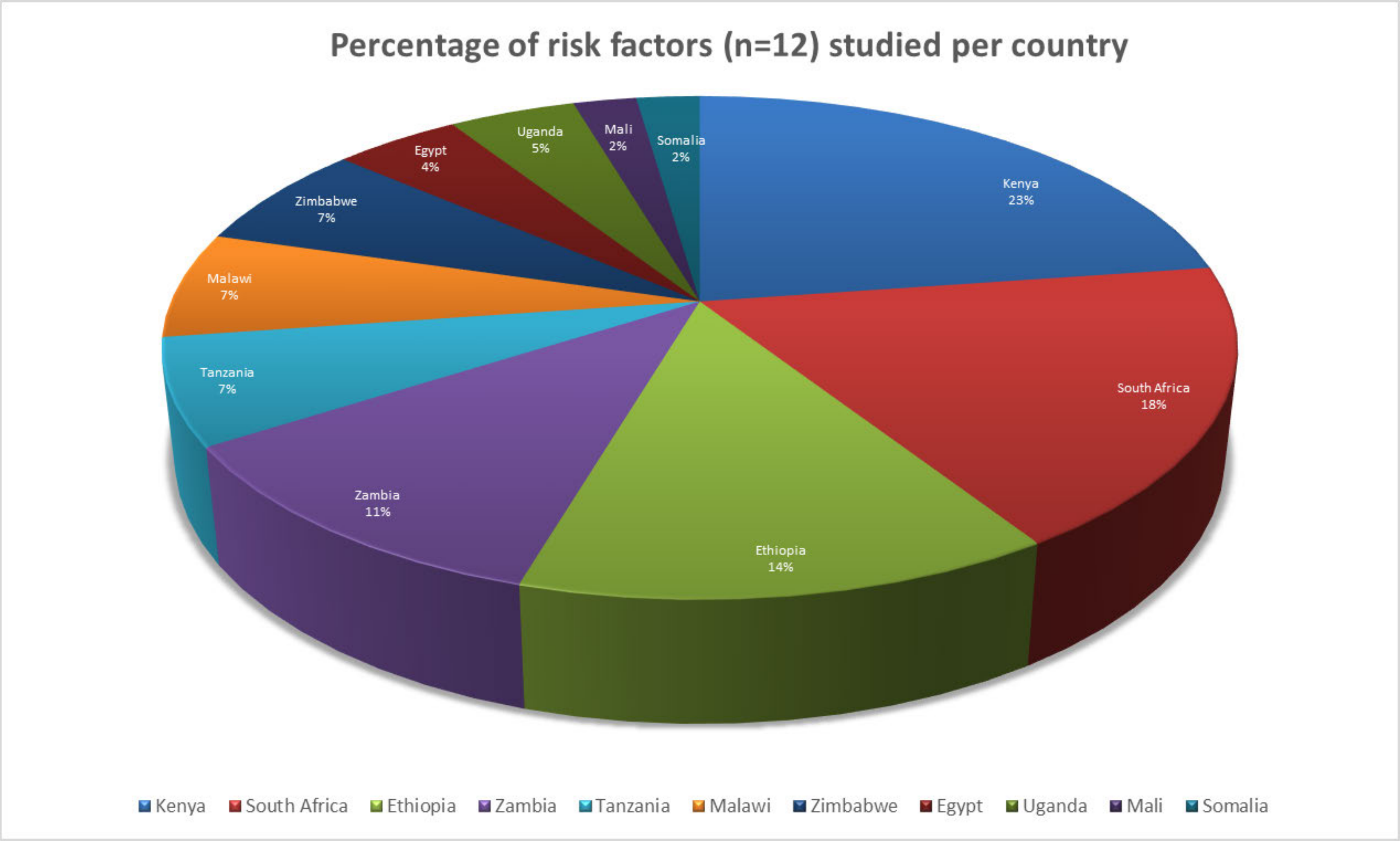


Figure 3.2: *Risk factors investigated in the 11 African countries which reported on case control studies*

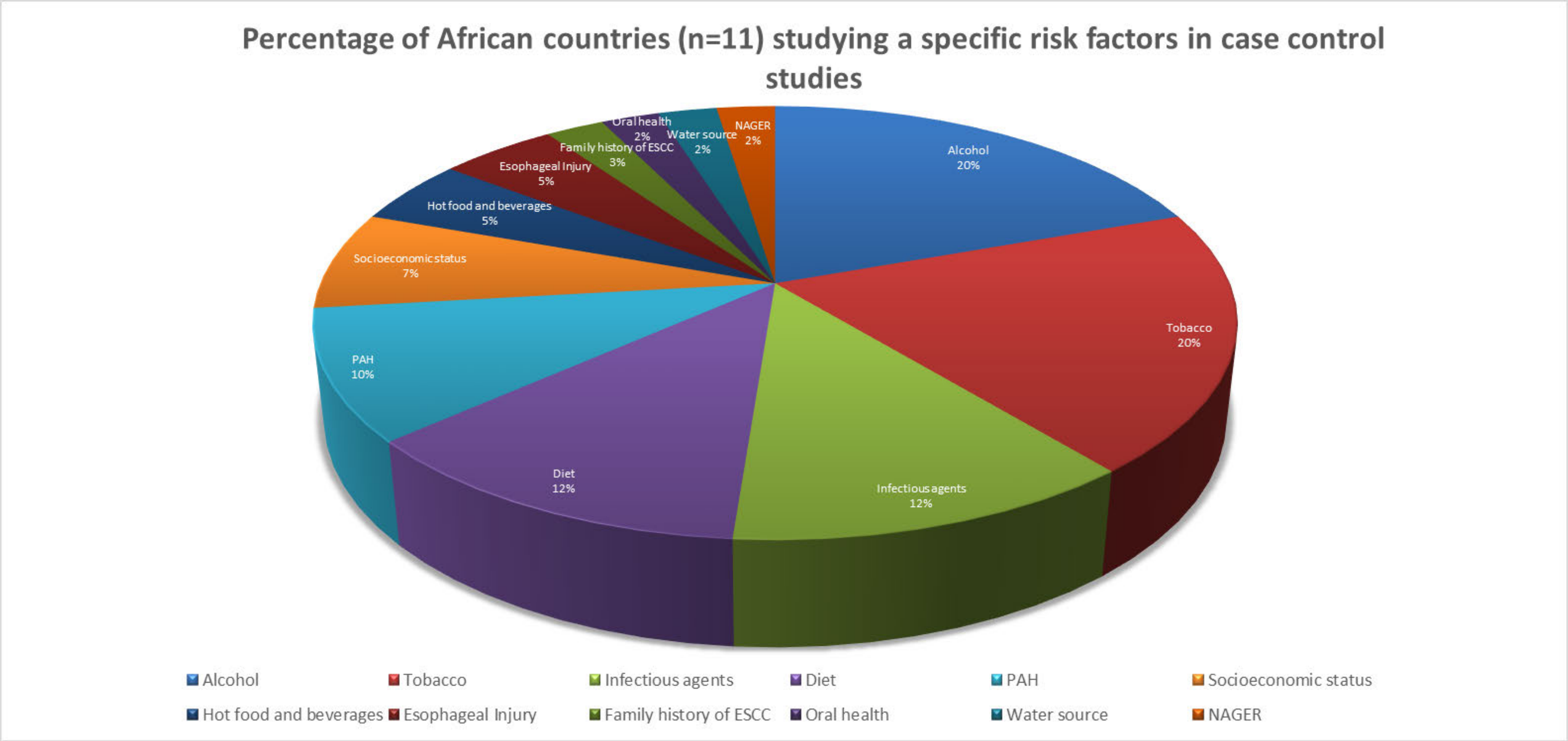


Figure 3.3: *African countries (n=11) studying a specific risk factor. NAGER; non-acid gastroesophageal reflux*Using data from GLOBOCAN, Table 3.2 shows environmental and lifestyle risk factors investigated in other ESCC high risk regions which are reported in the systematic review.

Table 3.2: Risk factors investigated in other ESCC high-risk regions and reported in this systematic review

Risk factor for ESCC	High-risk non-African country and region reporting risk factors
Alcohol	Argentina, Brazil, Paraguay, Uruguay, China, Thailand, Japan (26, 27)
Tobacco	Argentina, Brazil, Paraguay, Uruguay, China, Thailand, Japan (26, 27)
Helicobacter pylori	China, India, Japan, Korea, Iran (28)
Diet (fruits and vegetables)	China, Japan, Uruguay (29)
PAH	Argentina, Brazil, Paraguay, and Uruguay, India, China, Iran (30-32)
Hot food and beverages, and esophageal inflammation	China, India, Iran, Argentina, Brazil, Uruguay, Paraguay (33-35)
Socioeconomic status	India, China (36, 37)
Family history of cancer	China, India (38, 39)
Non-acid gastro esophageal reflux	Japan (40)
Oral health	China, Iran (37, 41)
Water source	Iran (42)

3.4.3 Tobacco smoking and chewing

Tobacco smoking was the most commonly investigated risk factor, with 21 (68%) of the 31 reporting quantifiable associations between smoking and EC (Figure 2). Twenty-one studies were case-control studies and included 6,984 cases and 15,322 controls. The majority (n=15) of the studies were from South Africa, with two from Zambia, two from Zimbabwe, two from Kenya, and two from Uganda. The studies were published between 1988 and 2019. Of the 21 studies, 15 (71%) reported significant associations between tobacco use and the ESCC development.

Of the 15 studies that reported a significant association, all indicated an increased risk of ESCC in people who smoke or chew tobacco, with ORs ranging from 1.05 to 11.24.(43, 44) The highest risk was reported in studies done on a Zambian population and a South African female population, with ORs of 11.24 (1.34-92.4 95%CI) and 11.1 (4.5-27 95%CI), respectively.(44, 45) This was followed by studies done on a Zambian population and two Zimbabwean male populations which reported ORs of 8 (2.8-22.7

95%CI), 5.7(3.7-8.5 95%CI) and 5.6 (3.8-8.4 95%CI), respectively.(46-48) The rest of the studies stated increased risk of ESCC with ORs ≤ 4.11. In studies that assessed the tobacco as a risk factor separately for men and women, men had a slightly higher risk than women.(49, 50)

3.4.4 Alcohol consumption

Alcohol consumption was investigated as a risk factor in 21 of the 31 (66%) studies with quantifiable associations to ESCC. All the studies were case-control studies with the majority (n=15) of the studies coming from South Africa, and two from Zambia, two from Zimbabwe, two from Kenya, and two from Uganda. They included 6,773 cases and 15,111 controls. There was a significant overlap with the studies reporting the effects of tobacco smoking. The studies were published between 1988 and 2019.

Only eight (38%) of the 21 studies reported statistically significant associations and increased risk to ESCC. The highest risk was reported in a study done on a South African population with OR of 15.4 (5.5-43.4 95%CI) for men and 9.9 (3.4-28.6 95%CI) for women consuming commercial beer.(45) This was followed by a South African study which reported an OR of 5.09 (3.4-7.6 95%CI) and 3.86 (2.7-5.5 95%CI) for traditional beer and commercial spirits, respectively.(51) The remaining eight studies documented increased risk with ORs ranging from 2.2 to 3.48 and 95%CI of 1.23 to 6.07.(50) One of the studies by Patel et al (52), after adjusting for gender, age, smoking, snuff use, and cooking and sleeping in the same room, ESCC patients were 45% more likely to have EC due to alcohol consumption as compared to controls.

3.4.5 Tobacco and alcohol

A combination of smoking and alcohol as a risk factor for ESCC was investigated in 11 of the 31 studies, with 10 reporting statistically significant interactions between the two factors. They included 4,052 cases and 7,007 controls. Both alcohol and tobacco independent risk factors for ESCC The combination of tobacco and alcohol was reported to increase the risk in all the studies with ORs ranging from 1.95 to 19.06 and 95%CIs ranging from 1.4 to 41.7.(45, 53-55) A South African study described an increased risk of OR 18.2 (8.1-41.7 95%CI) for women, which was significantly higher than that for men 3.5 (1.5-8.4 95%CI).(45) In another study which assessed the risk for men and women separately, the risk for women was slightly higher, with an OR of 4.8 (3-7.8 95%CI), than for men, 4.7 (2.8-7.9 95%CI).(49) It is important to determine whether the joint effect of adverse risk factors has an additive or multiplicative effect.

3.5.6 Diet

Nine studies investigated the effect of poor diet on ESCC. This included food groups, food items, vitamins and trace elements. A total of 1,509 cases and 2,188 controls were assessed. Seven of the studies reported statistically significant associations to ESCC development; six were case-control studies whilst one was an ecological study. Three South African studies reported increased risk of ESCC in participants who consume wild vegetables.(56-58) The wild vegetables comprised imifino, Uthyuthu (*Amaranthus thunbergii)*, imbikicane (*Chenopodium album*), and umsobo (*Sofanum nigrum*). One of the studies on South African women reported an increased risk of ESCC in consumers of wild imifino vegetables with OR of 1.84 (1.04-3.27 95%CI).(58) The highest risk was observed by Sammon et al (56) with a RR of 2.86 (1.16-8 95%CI).

Consumption of fruits, vegetables and green legumes was individually associated with a protective effect to ESCC development in a South African study by Sewram et al (58). Eating fruits 5-7 times a week was associated with a protective effect of OR 0.51 and 0.42 for men and women, respectively. Consumption of vegetables 5-7 times a week, also had a protective effect of OR 0.62 and 0.5 for men and women, respectively. A study done by Leon et al (59) reported that not eating vegetables at least once a week, or not eating green vegetables at all significantly increased the risk of ESCC with OR of 12.68 (1.99-80.96 95%CI) and 400 (12-3,345 95%CI), respectively.(59) Additionally, consumption of meat 5-7 days per week was reported to have a protective effect.(58)

Other food items that were reported to increase the risk of ESCC development, were: purchased maize, pumpkin, beans, sorghum and porridge reported in three South African and one Ethiopian study. (57, 58, 60, 61) The Ethiopian study by Leon et al (59), also indicated that saltiness in food increased the risk of ESCC with OR of 7.79 (1.21-50.3 95%CI). In another South African study, daily and weekly consumption of margarine was reported to have a protective effect with OR of 0.51 and 0.71, respectively.(61) Schaafsma et al (62) performed an ecological study assessing the

ESCC development and seven micronutrients (calcium, copper, iodine, magnesium, selenium, zinc) in 32 African countries. Iron, zinc and selenium were described to have a protective effect in males and females, whilst magnesium was reported to be protective in females only.

3.5.7 Socio-economic status

Low socio-economic status was assessed as a risk factor for ESCC development in four case-control studies. They included 1,268 cases and 3,723 controls. One of the studies was from South Africa, one from Zimbabwe and two were from Kenya. Socio-economic status was measured using salaries, occupational status and type of housing. The South African study reported an increased risk of ESCC associated with lower salaries.(51) The study found RR ranging from 1.23 to 74.94 for various low-salary levels.(51) The Zimbabwean study evaluated the effect of occupational status in men.(48) Low occupational status and mining as an occupation were found to increase the risk for ESCC when compared to high occupational status in men with OR of 1.5 and 2.5, respectively.(48) One Kenyan study showed that a monthly salary of over 100 dollars reduced the risk of ESCC with OR of 0.59 (0.46-0.77 95%CI).(52) The second Kenyan study showed that poor housing increased the risk of ESCC with OR of 1.98 (1.11-3.53 95%CI).(43)

3.5.8 PAH exposure

Six case control studies reported PAH exposure as a risk factor for ESCC. They included 1,176 cases and 3,076 controls. Indoor air pollution through: smokiness in the home, heating and cooking fuel, and mursik (a fermented milk beverage which may contain charcoal) was classified as PAH exposures. Five studies from South Africa, Kenya and Zambia reported significant associations between PAH exposure and ESCC. Pacella-Norman et al (49) assessed the effect of heating fuel (paraffin) in South African men and women and found an increased risk of ESCC in women OR 3.5 (1.1-11.3 95%CI). In another South African study, the use of wood and charcoal for heating and cooking were reported to increase ESCC risk with OR 15.2 (8.15-28.2 95%CI).(63) The use of charcoal and wood as cooking fuel was reported to increase risk in a Zambian population with OR 3.0 (1.2-7.4 95%CI).(46) Charcoal and wood use for cooking were also assessed in a Kenyan study, which reported an increased risk with OR 2.32 (1.41-3.84 95%CI) and in a South African study with OR 7.1 (4.6-11 95% CI).(52, 55) The same Kenyan study reported on the use of mursik, the consumption of which increased the risk of ESCC with OR 3.72 (1.96-7.14 95%CI).(52)

3.5.9 Hot food and beverages

Two Kenyan studies assessed the consumption of hot food and beverages and their association to ESCC. They included 589 cases and 599 controls. Both studies were case-control studies, and reported that drinking hot, and very hot beverages increased the risk of ESCC with OR of 12.78 (6.98-23.6 95%CI) and 3.66 (2.1-6.5 95%CI), respectively.(52, 64) A few studies from Kenya, Ethiopia, Tanzania and Malawi have also shown that the consumption of hot tea, hot food and hot chai are important risk factors for ESCC.(65-67) However, these studies did not report risk estimates.

3.5.10 Oral Health

Two Kenyan and two Tanzanian case-control studies explored poor oral health as a risk factor for ESCC. They included 1,370 cases and 1,380 controls. A study by Patel et al (52) showed that tooth loss was associated with an increased risk of ESCC with OR 5.28 (2.98-9.41 95%CI). Tooth loss was also associated with an increased risk of ESCC in a study by Menya et al (68). In this study, other components of oral health were assessed which showed an increased risk, these include: decayed teeth (≥3) OR 4.4 (2.3-8.3 95%CI), brushing teeth only once per week OR 2.3 (1.0-5.5 95%CI), never having brushed teeth OR 2.5 (1.0-6.0 95%CI), oral leukoplakia OR 3.1 (1.8-5.3 95%CI), and the sum of number of decayed + missing + filled teeth ≥8 OR 3.0 (1.5-6.1 95%CI).(68)

3.5.11 Infectious agents

Human papillomavirus (HPV) and HIV infection were assessed as risk factors in three studies (two Zambian and one South African). They included 200 cases and 720 controls. Two studies reported statistically significant associations between HPV, HIV and ESCC development. In a South African study HPV16 was associated with an increased risk of OR 1.59 (1.19-2.13 95%CI).(69) One of the Zambian studies did not find statistically significant associations between HPV and ESCC; however HIV infection in patients over 60 years was associated with increased risk of ESCC with

OR 5.5 (1-27.7 95%CI).(70) The second Zambian study did not report significant association between HIV infection and ESCC development.(71) The discrepancies in HPV detection may be attributable to the different methods employed in the detection of HPV DNA in different studies as well conditions surrounding sample collection and storage

3.5.12 Esophageal injury

Esophageal inflammation due to self-induced vomiting and caustic ingestion was reported as a risk factor in two South African studies and one Kenyan study. The case-control studies had a total of 661 cases and 266 controls. In the South African study, induced vomiting was associated with ESCC, reporting OR of 1.83 (1.13-2.96 95%CI).(72) The study reported on various methods used by the participants to induce vomiting which include use of salt water, traditional medicine, warm water, holy water, and vinegar water. The South African case-control study did not show a statistically significant association between induced vomiting or use of traditional emetics and ESCC development.(56) The Kenyan study reported that caustic ingestion was associated with an increased risk of ESCC with OR 11.35 (3.04-42.46 95%CI).(43) The use of traditional medicines, which can be used as emetics, was reported in a South African case-control study by Sammon et al. (56) However, the association between traditional medicines and EC development was not statistically significant.

3.5.13 Water sources

Water source was assessed as a risk factor for ESCC development in a Kenyan case control study of 430 cases and 440 controls.(68) Use of spring/river water compared to piped/rain water was reported to be associated with ESCC development with OR 3.1 (1.5-6.5 95%CI). Use of borehole and piped water did not show statistically significant associations with ESCC.(68)

3.5.14 Family history of cancer

Family history of cancer was reported to increase the risk of ESCC in a Kenyan case-control study with OR of 3.50 (1.29-9.49 95%CI).(43) Sample size was small with only 83 cases and 166 controls. Ten (12%) of the cases and eight (4%) of the controls had a positive family history of cancer.

3.5.15 Non-acid gastroesophageal reflux

Non-acid gastroesophageal reflux was reported to increase the risk of ESCC in a South African case-control study with OR of 8.8 (3.2-24.5 95%CI).(73) The authors measured non-acid gastroesophageal reflux using a digi-trapper high-definition multichannel impedance and pH medical measurement system, which involved placing a test catheter near the esophagogastric junction for 24 hours. Sample size was very small with only 32 cases and 49 controls. Non-acid gastroesophageal reflux was reported in 23 (73%) of the cases and in 11 (22%) of the controls.

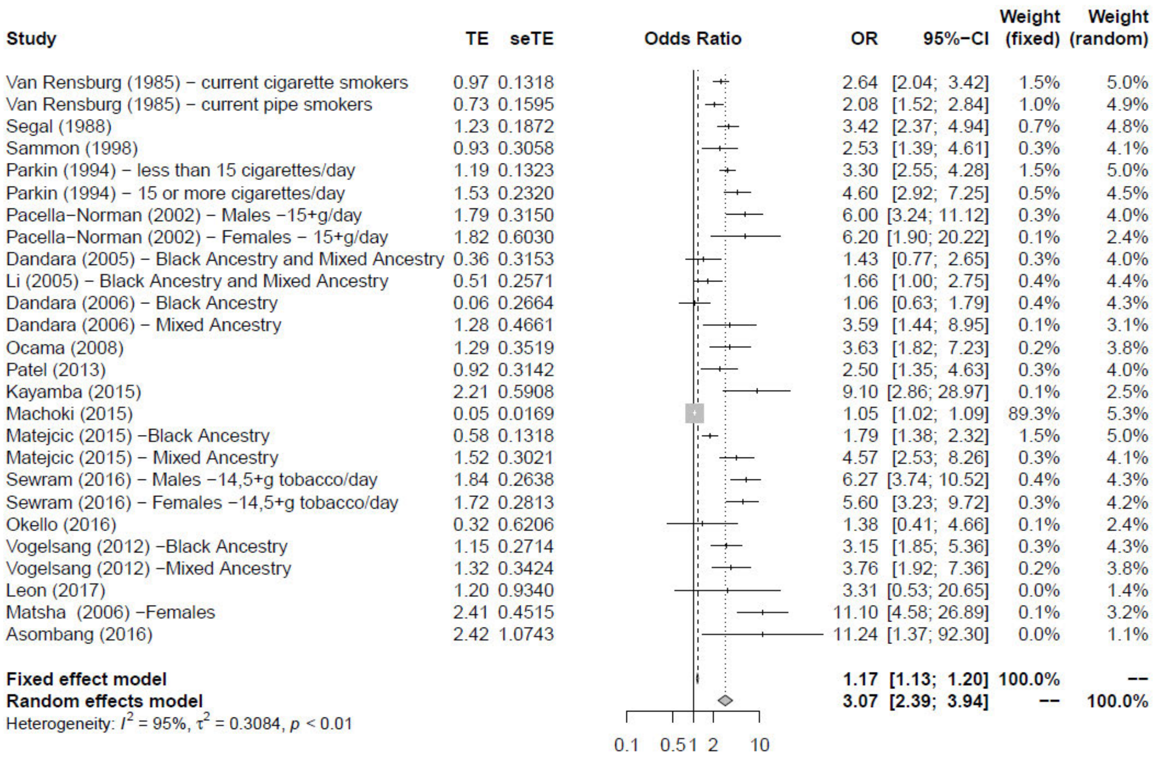
3.5.16 Synthesized findings

3.5.16.1 Meta-analysis

We performed a meta-analysis for six of the risk factors: tobacco smoking, alcohol consumption, combined tobacco and alcohol use, esophageal injury, fruit and vegetable consumption and PAH exposure. We first included all studies in the meta-analysis and then using an outlier detection method, removed the studies with extreme effect sizes from the final meta-analysis. The outliers are still displayed in the second meta-analysis forest plots, however their weight is set to 0%, indicating that we did not include them in the pooled analysis. Influence analysis was also done to detect studies which were distorting the overall effect size the most as well as to corroborate the results from the outlier detection methods. Three studies, van Rensburg et al 1985(61), Segal et al 1988(51), and Sammon et al 1992(56), reported their effect sizes as RR, therefore ORs were calculated from the exposed vs non-exposed data provided in the respective publications and used in our meta-analysis.

The pooled analysis for tobacco smoking showed an effect size of OR of 3.07 (2.39-3.94 95%CI) (Figure 3.4). Heterogeneity (I^2) of 95% with $p < 0.01$ was recorded. Using the outliers detection method, influence analysis and GOSH method, 12 studies were identified as outliers, distorting the overall effect size, and contributing to the high heterogeneity and cluster imbalance. A second meta-analysis following the removal of these seven studies, resulted in a pooled effect size of OR of 4.14 (3.26-5.26 95%CI) and I^2 of 36% ($p = 0.09$) (Figure 3.5). The forest plot is shown in Figure 3.6. Funnel plot assessing for publication bias showed significant asymmetry, eggers test had a $p < 2.74 \times 10^{-8}$ confirming significant asymmetry and possible publication bias

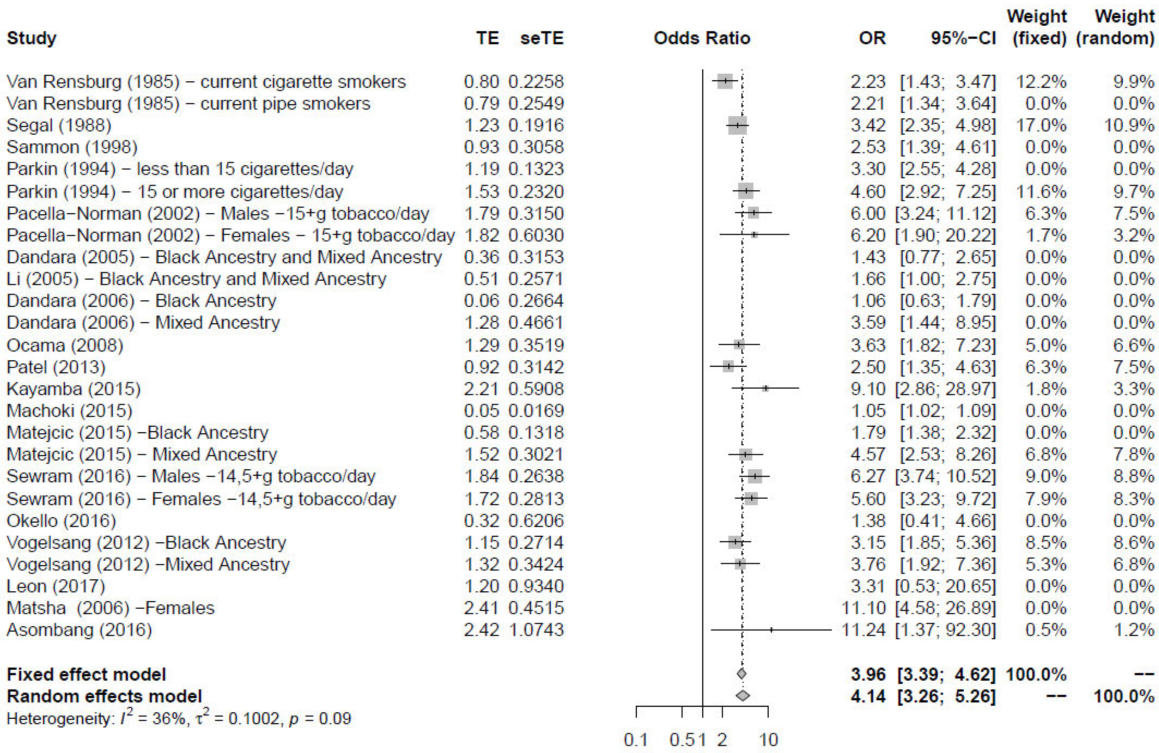
(Figure 3.9). One of the studies included in this analysis, Machoki et al (74) is not indexed on PubMed®.



Study	TE	seTE	OR	95%−CI	Weight (fixed)	Weight (random)
Van Rensburg (1985) − current cigarette smokers	0.97	0.1318	2.64	[2.04; 3.42]	1.5%	5.0%
Van Rensburg (1985) − current pipe smokers	0.73	0.1595	2.08	[1.52; 2.84]	1.0%	4.9%
Segal (1988)	1.23	0.1872	3.42	[2.37; 4.94]	0.7%	4.8%
Sammon (1998)	0.93	0.3058	2.53	[1.39; 4.61]	0.3%	4.1%
Parkin (1994) − less than 15 cigarettes/day	1.19	0.1323	3.30	[2.55; 4.28]	1.5%	5.0%
Parkin (1994) − 15 or more cigarettes/day	1.53	0.2320	4.60	[2.92; 7.25]	0.5%	4.5%
Pacella−Norman (2002) − Males −15+g/day	1.79	0.3150	6.00	[3.24; 11.12]	0.3%	4.0%
Pacella−Norman (2002) − Females − 15+g/day	1.82	0.6030	6.20	[1.90; 20.22]	0.1%	2.4%
Dandara (2005) − Black Ancestry and Mixed Ancestry	0.36	0.3153	1.43	[0.77; 2.65]	0.3%	4.0%
Li (2005) − Black Ancestry and Mixed Ancestry	0.51	0.2571	1.66	[1.00; 2.75]	0.4%	4.4%
Dandara (2006) − Black Ancestry	0.06	0.2664	1.06	[0.63; 1.79]	0.4%	4.3%
Dandara (2006) − Mixed Ancestry	1.28	0.4661	3.59	[1.44; 8.95]	0.1%	3.1%
Ocama (2008)	1.29	0.3519	3.63	[1.82; 7.23]	0.2%	3.8%
Patel (2013)	0.92	0.3142	2.50	[1.35; 4.63]	0.3%	4.0%
Kayamba (2015)	2.21	0.5908	9.10	[2.86; 28.97]	0.1%	2.5%
Machoki (2015)	0.05	0.0169	1.05	[1.02; 1.09]	89.3%	5.3%
Matejcic (2015) −Black Ancestry	0.58	0.1318	1.79	[1.38; 2.32]	1.5%	5.0%
Matejcic (2015) − Mixed Ancestry	1.52	0.3021	4.57	[2.53; 8.26]	0.3%	4.1%
Sewram (2016) − Males −14,5+g tobacco/day	1.84	0.2638	6.27	[3.74; 10.52]	0.4%	4.3%
Sewram (2016) − Females −14,5+g tobacco/day	1.72	0.2813	5.60	[3.23; 9.72]	0.3%	4.2%
Okello (2016)	0.32	0.6206	1.38	[0.41; 4.66]	0.1%	2.4%
Vogelsang (2012) −Black Ancestry	1.15	0.2714	3.15	[1.85; 5.36]	0.3%	4.3%
Vogelsang (2012) −Mixed Ancestry	1.32	0.3424	3.76	[1.92; 7.36]	0.2%	3.8%
Leon (2017)	1.20	0.9340	3.31	[0.53; 20.65]	0.0%	1.4%
Matsha (2006) −Females	2.41	0.4515	11.10	[4.58; 26.89]	0.1%	3.2%
Asombang (2016)	2.42	1.0743	11.24	[1.37; 92.30]	0.0%	1.1%
Fixed effect model			**1.17**	**[1.13; 1.20]**	**100.0%**	**--**
Random effects model			**3.07**	**[2.39; 3.94]**	**--**	**100.0%**

Heterogeneity: $I^2 = 95\%$, $\tau^2 = 0.3084$, $p < 0.01$

Figure 3.4: Effect of tobacco smoking on esophageal cancer in Africa. *Forest plot showing the pooled effect of tobacco smoking on esophageal cancer development in Africa. The figure was generated using R software. Study ID gives the first author and the year of the publication. OR, odds ratio: TE, logged effect size: seTE, standard error of effect size: CI, confidence interval*



Study	TE	seTE	OR	95%−CI	Weight (fixed)	Weight (random)
Van Rensburg (1985) − current cigarette smokers	0.80	0.2258	2.23	[1.43; 3.47]	12.2%	9.9%
Van Rensburg (1985) − current pipe smokers	0.79	0.2549	2.21	[1.34; 3.64]	0.0%	0.0%
Segal (1988)	1.23	0.1916	3.42	[2.35; 4.98]	17.0%	10.9%
Sammon (1998)	0.93	0.3058	2.53	[1.39; 4.61]	0.0%	0.0%
Parkin (1994) − less than 15 cigarettes/day	1.19	0.1323	3.30	[2.55; 4.28]	0.0%	0.0%
Parkin (1994) − 15 or more cigarettes/day	1.53	0.2320	4.60	[2.92; 7.25]	11.6%	9.7%
Pacella−Norman (2002) − Males −15+g tobacco/day	1.79	0.3150	6.00	[3.24; 11.12]	6.3%	7.5%
Pacella−Norman (2002) − Females − 15+g tobacco/day	1.82	0.6030	6.20	[1.90; 20.22]	1.7%	3.2%
Dandara (2005) − Black Ancestry and Mixed Ancestry	0.36	0.3153	1.43	[0.77; 2.65]	0.0%	0.0%
Li (2005) − Black Ancestry and Mixed Ancestry	0.51	0.2571	1.66	[1.00; 2.75]	0.0%	0.0%
Dandara (2006) − Black Ancestry	0.06	0.2664	1.06	[0.63; 1.79]	0.0%	0.0%
Dandara (2006) − Mixed Ancestry	1.28	0.4661	3.59	[1.44; 8.95]	0.0%	0.0%
Ocama (2008)	1.29	0.3519	3.63	[1.82; 7.23]	5.0%	6.6%
Patel (2013)	0.92	0.3142	2.50	[1.35; 4.63]	6.3%	7.5%
Kayamba (2015)	2.21	0.5908	9.10	[2.86; 28.97]	1.8%	3.3%
Machoki (2015)	0.05	0.0169	1.05	[1.02; 1.09]	0.0%	0.0%
Matejcic (2015) −Black Ancestry	0.58	0.1318	1.79	[1.38; 2.32]	0.0%	0.0%
Matejcic (2015) − Mixed Ancestry	1.52	0.3021	4.57	[2.53; 8.26]	6.8%	7.8%
Sewram (2016) − Males −14,5+g tobacco/day	1.84	0.2638	6.27	[3.74; 10.52]	9.0%	8.8%
Sewram (2016) − Females −14,5+g tobacco/day	1.72	0.2813	5.60	[3.23; 9.72]	7.9%	8.3%
Okello (2016)	0.32	0.6206	1.38	[0.41; 4.66]	0.0%	0.0%
Vogelsang (2012) −Black Ancestry	1.15	0.2714	3.15	[1.85; 5.36]	8.5%	8.6%
Vogelsang (2012) −Mixed Ancestry	1.32	0.3424	3.76	[1.92; 7.36]	5.3%	6.8%
Leon (2017)	1.20	0.9340	3.31	[0.53; 20.65]	0.0%	0.0%
Matsha (2006) −Females	2.41	0.4515	11.10	[4.58; 26.89]	0.0%	0.0%
Asombang (2016)	2.42	1.0743	11.24	[1.37; 92.30]	0.5%	1.2%
Fixed effect model			**3.96**	**[3.39; 4.62]**	**100.0%**	**--**
Random effects model			**4.14**	**[3.26; 5.26]**	**--**	**100.0%**

Heterogeneity: $I^2 = 36\%$, $\tau^2 = 0.1002$, $p = 0.09$

Figure 3.5: Final estimates of the of tobacco smoking on esophageal cancer in Africa. *Forest plot showing the pooled effect of tobacco smoking on esophageal cancer development in Africa, after excluding outliers and studies which account for distorted effect size, high-heterogeneity and clustering. The figure was generated using R software. Study ID gives the first author and the year of the publication. OR, odds ratio: TE, logged effect size: seTE, standard error of effect size: CI, confidence interval*

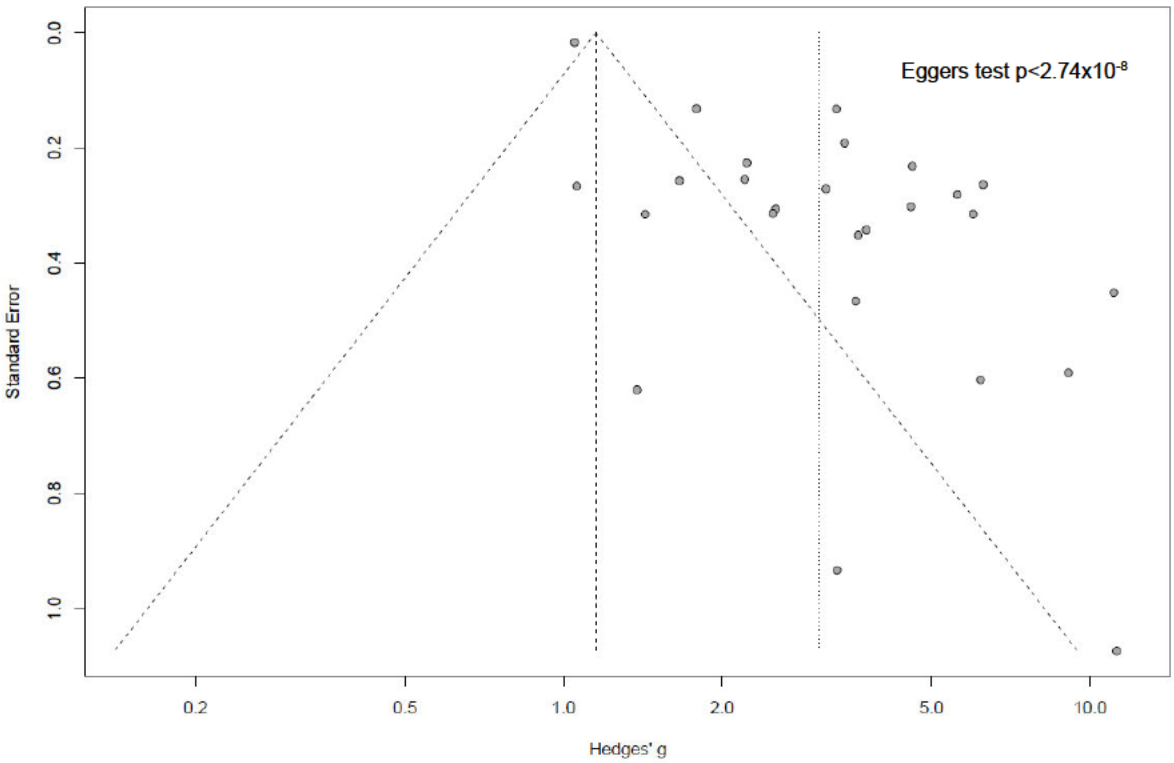
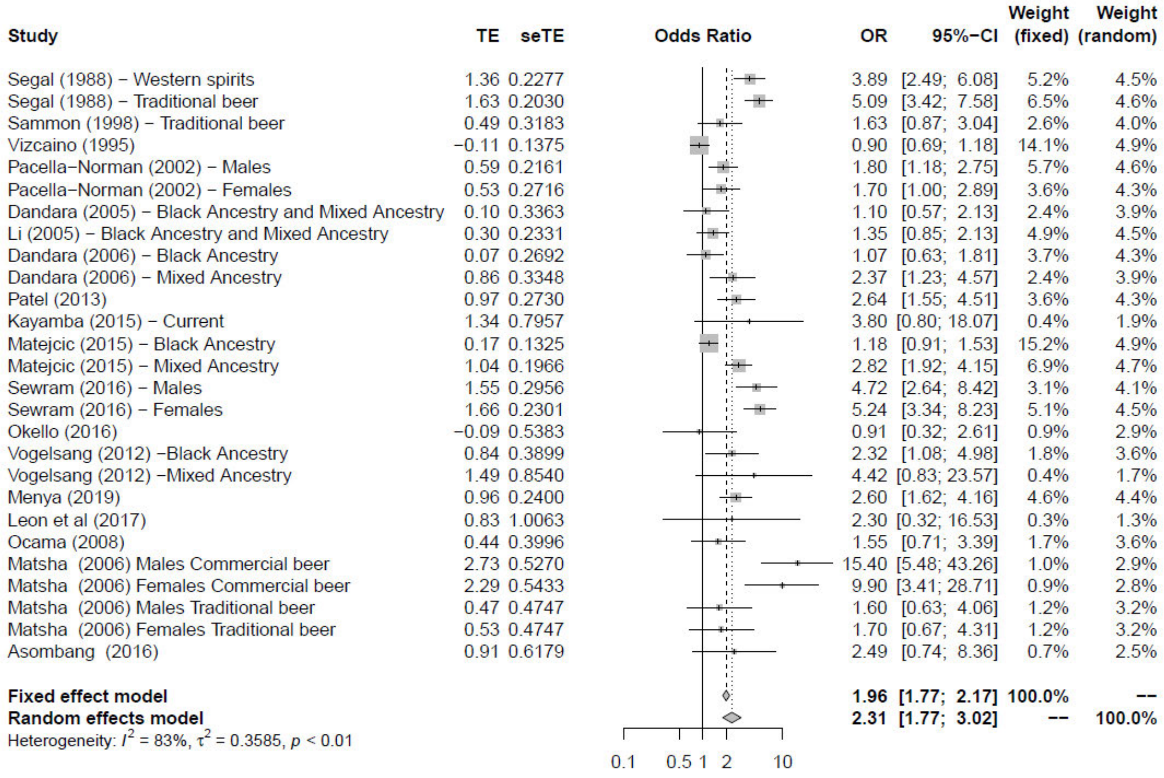


***Figure 3.6: Funnel plot for assessing publication bias for tobacco smoking.** The figure was generated using R software.*

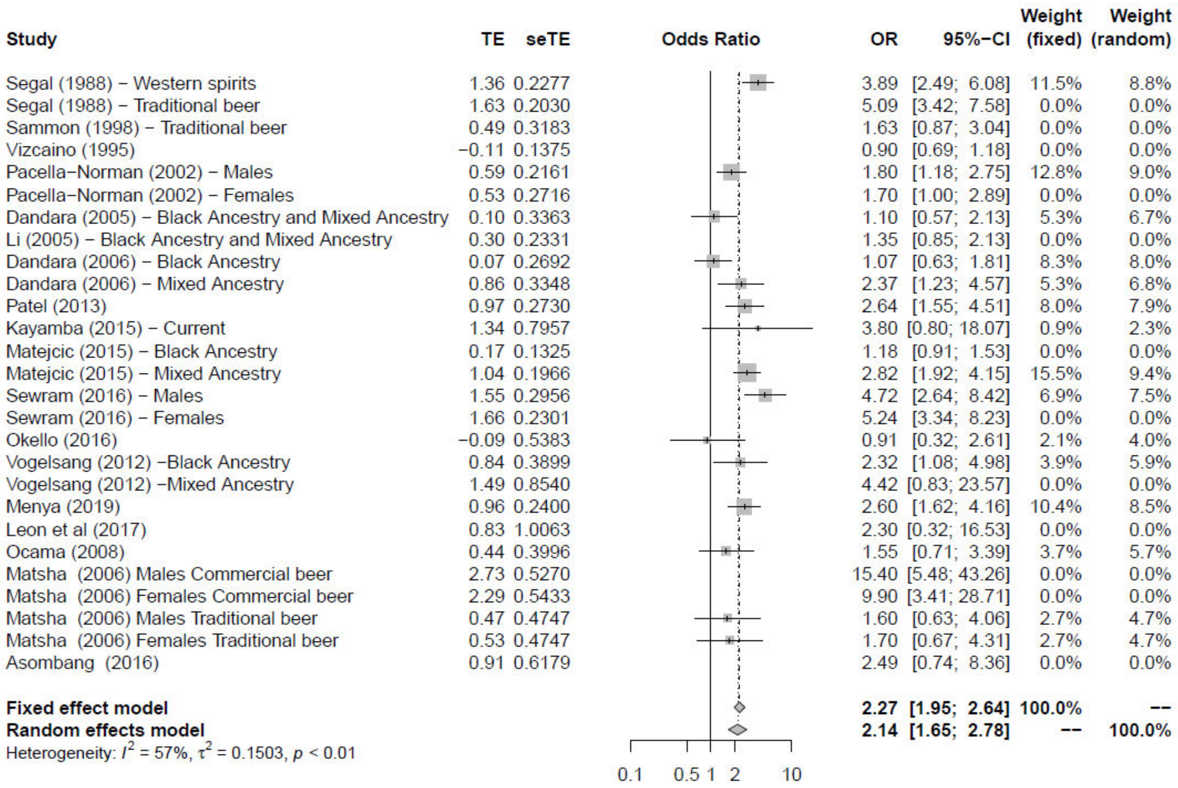
The pooled analysis for alcohol consumption demonstrated an effect size of OR 2.31 (1.77-3.02 95%CI) (Figure 3.7), indicating that alcohol users are twice as likely to develop ESCC compared to non-alcohol users. Test for heterogeneity showed I^2 of 83% with a $p < 0.01$. A second meta-analysis excluding the outliers has OR of 2.14 (1.65-2.78 95%CI) and I^2 of 57% ($p < 0.01$) (Figure 3.8). The funnel plot (Figure 3.9) did not show asymmetry with the Egger's test indicating a $p < 0.07$.



Study	TE	seTE	OR	95%−CI	Weight (fixed)	Weight (random)
Segal (1988) − Western spirits	1.36	0.2277	3.89	[2.49; 6.08]	5.2%	4.5%
Segal (1988) − Traditional beer	1.63	0.2030	5.09	[3.42; 7.58]	6.5%	4.6%
Sammon (1998) − Traditional beer	0.49	0.3183	1.63	[0.87; 3.04]	2.6%	4.0%
Vizcaino (1995)	−0.11	0.1375	0.90	[0.69; 1.18]	14.1%	4.9%
Pacella−Norman (2002) − Males	0.59	0.2161	1.80	[1.18; 2.75]	5.7%	4.6%
Pacella−Norman (2002) − Females	0.53	0.2716	1.70	[1.00; 2.89]	3.6%	4.3%
Dandara (2005) − Black Ancestry and Mixed Ancestry	0.10	0.3363	1.10	[0.57; 2.13]	2.4%	3.9%
Li (2005) − Black Ancestry and Mixed Ancestry	0.30	0.2331	1.35	[0.85; 2.13]	4.9%	4.5%
Dandara (2006) − Black Ancestry	0.07	0.2692	1.07	[0.63; 1.81]	3.7%	4.3%
Dandara (2006) − Mixed Ancestry	0.86	0.3348	2.37	[1.23; 4.57]	2.4%	3.9%
Patel (2013)	0.97	0.2730	2.64	[1.55; 4.51]	3.6%	4.3%
Kayamba (2015) − Current	1.34	0.7957	3.80	[0.80; 18.07]	0.4%	1.9%
Matejcic (2015) − Black Ancestry	0.17	0.1325	1.18	[0.91; 1.53]	15.2%	4.9%
Matejcic (2015) − Mixed Ancestry	1.04	0.1966	2.82	[1.92; 4.15]	6.9%	4.7%
Sewram (2016) − Males	1.55	0.2956	4.72	[2.64; 8.42]	3.1%	4.1%
Sewram (2016) − Females	1.66	0.2301	5.24	[3.34; 8.23]	5.1%	4.5%
Okello (2016)	−0.09	0.5383	0.91	[0.32; 2.61]	0.9%	2.9%
Vogelsang (2012) −Black Ancestry	0.84	0.3899	2.32	[1.08; 4.98]	1.8%	3.6%
Vogelsang (2012) −Mixed Ancestry	1.49	0.8540	4.42	[0.83; 23.57]	0.4%	1.7%
Menya (2019)	0.96	0.2400	2.60	[1.62; 4.16]	4.6%	4.4%
Leon et al (2017)	0.83	1.0063	2.30	[0.32; 16.53]	0.3%	1.3%
Ocama (2008)	0.44	0.3996	1.55	[0.71; 3.39]	1.7%	3.6%
Matsha (2006) Males Commercial beer	2.73	0.5270	15.40	[5.48; 43.26]	1.0%	2.9%
Matsha (2006) Females Commercial beer	2.29	0.5433	9.90	[3.41; 28.71]	0.9%	2.8%
Matsha (2006) Males Traditional beer	0.47	0.4747	1.60	[0.63; 4.06]	1.2%	3.2%
Matsha (2006) Females Traditional beer	0.53	0.4747	1.70	[0.67; 4.31]	1.2%	3.2%
Asombang (2016)	0.91	0.6179	2.49	[0.74; 8.36]	0.7%	2.5%
Fixed effect model			**1.96**	**[1.77; 2.17]**	**100.0%**	**--**
Random effects model			**2.31**	**[1.77; 3.02]**	**--**	**100.0%**

Heterogeneity: I^2 = 83%, τ^2 = 0.3585, p < 0.01

Figure 3.7: Effect of alcohol consumption on ESCC in Africa. *Forest plot showing the pooled effect of alcohol consumption on ESCC development in Africa. The figure was generated using R software. Study ID gives the first author and the year of the publication. OR, odds ratio: TE, logged effect size: seTE, standard error of effect size: CI, confidence interval*



Study	TE	seTE	OR	95%−CI	Weight (fixed)	Weight (random)
Segal (1988) − Western spirits	1.36	0.2277	3.89	[2.49; 6.08]	11.5%	8.8%
Segal (1988) − Traditional beer	1.63	0.2030	5.09	[3.42; 7.58]	0.0%	0.0%
Sammon (1998) − Traditional beer	0.49	0.3183	1.63	[0.87; 3.04]	0.0%	0.0%
Vizcaino (1995)	−0.11	0.1375	0.90	[0.69; 1.18]	0.0%	0.0%
Pacella−Norman (2002) − Males	0.59	0.2161	1.80	[1.18; 2.75]	12.8%	9.0%
Pacella−Norman (2002) − Females	0.53	0.2716	1.70	[1.00; 2.89]	0.0%	0.0%
Dandara (2005) − Black Ancestry and Mixed Ancestry	0.10	0.3363	1.10	[0.57; 2.13]	5.3%	6.7%
Li (2005) − Black Ancestry and Mixed Ancestry	0.30	0.2331	1.35	[0.85; 2.13]	0.0%	0.0%
Dandara (2006) − Black Ancestry	0.07	0.2692	1.07	[0.63; 1.81]	8.3%	8.0%
Dandara (2006) − Mixed Ancestry	0.86	0.3348	2.37	[1.23; 4.57]	5.3%	6.8%
Patel (2013)	0.97	0.2730	2.64	[1.55; 4.51]	8.0%	7.9%
Kayamba (2015) − Current	1.34	0.7957	3.80	[0.80; 18.07]	0.9%	2.3%
Matejcic (2015) − Black Ancestry	0.17	0.1325	1.18	[0.91; 1.53]	0.0%	0.0%
Matejcic (2015) − Mixed Ancestry	1.04	0.1966	2.82	[1.92; 4.15]	15.5%	9.4%
Sewram (2016) − Males	1.55	0.2956	4.72	[2.64; 8.42]	6.9%	7.5%
Sewram (2016) − Females	1.66	0.2301	5.24	[3.34; 8.23]	0.0%	0.0%
Okello (2016)	−0.09	0.5383	0.91	[0.32; 2.61]	2.1%	4.0%
Vogelsang (2012) −Black Ancestry	0.84	0.3899	2.32	[1.08; 4.98]	3.9%	5.9%
Vogelsang (2012) −Mixed Ancestry	1.49	0.8540	4.42	[0.83; 23.57]	0.0%	0.0%
Menya (2019)	0.96	0.2400	2.60	[1.62; 4.16]	10.4%	8.5%
Leon et al (2017)	0.83	1.0063	2.30	[0.32; 16.53]	0.0%	0.0%
Ocama (2008)	0.44	0.3996	1.55	[0.71; 3.39]	3.7%	5.7%
Matsha (2006) Males Commercial beer	2.73	0.5270	15.40	[5.48; 43.26]	0.0%	0.0%
Matsha (2006) Females Commercial beer	2.29	0.5433	9.90	[3.41; 28.71]	0.0%	0.0%
Matsha (2006) Males Traditional beer	0.47	0.4747	1.60	[0.63; 4.06]	2.7%	4.7%
Matsha (2006) Females Traditional beer	0.53	0.4747	1.70	[0.67; 4.31]	2.7%	4.7%
Asombang (2016)	0.91	0.6179	2.49	[0.74; 8.36]	0.0%	0.0%
Fixed effect model			**2.27**	**[1.95; 2.64]**	**100.0%**	**--**
Random effects model			**2.14**	**[1.65; 2.78]**	**--**	**100.0%**

Heterogeneity: $I^2 = 57\%$, $\tau^2 = 0.1503$, $p < 0.01$

***Figure 3.8: Final estimates of the effect of alcohol consumption on ESCC in Africa**. Forest plot showing the pooled effect of alcohol consumption on ESCC development in Africa, after excluding outliers and studies which account for distorted ES, high-heterogeneity and clustering. The figure was generated using R software. Study ID gives the first author and the year of the publication. OR, odds ratio: TE, logged effect size: seTE, standard error of effect size: CI, confidence interval*

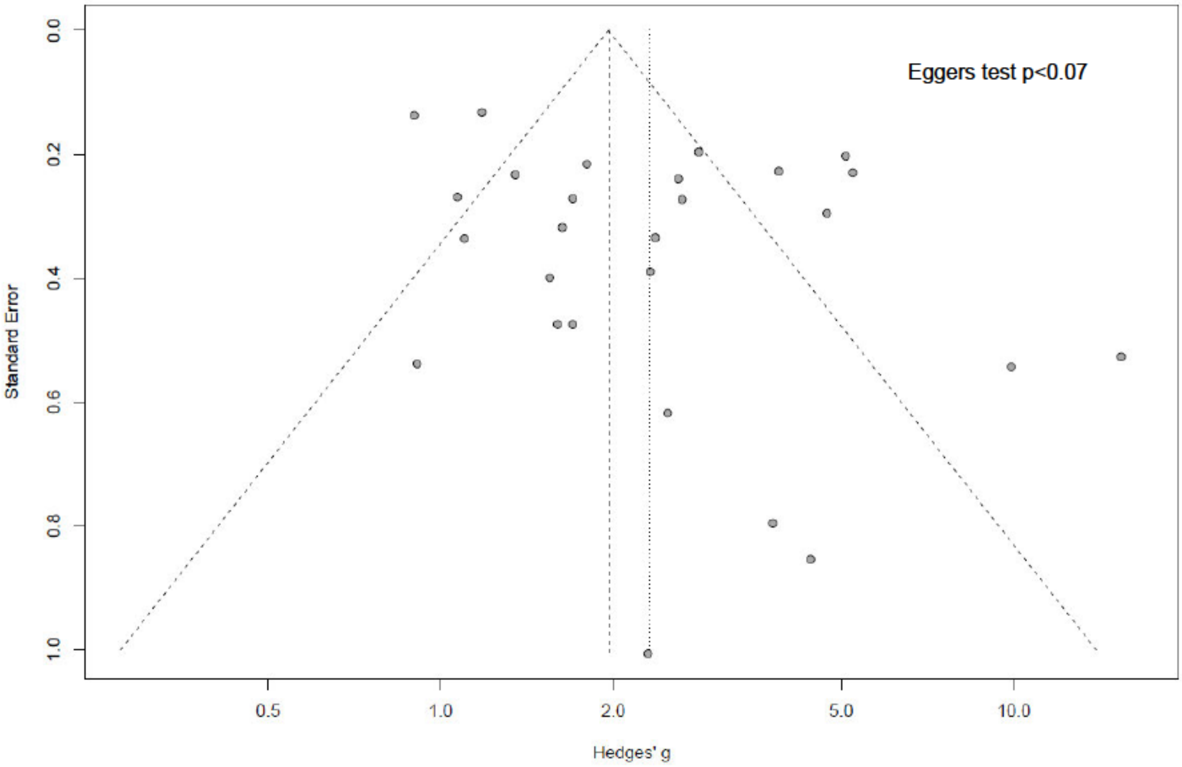


Figure 3.9: Funnel plot for assessing publication bias for alcohol consumption. *The figure was generated using R software.*

The pooled analysis of combined alcohol and tobacco, had an effect size of OR 3.95 (2.53-6.17 95%CI) (Figure 3.10), indicating that alcohol users are four times more likely to develop ESCC compared to non-alcohol users. Test for heterogeneity showed I^2 of 88% ($p < 0.01$). After removing outliers, a second pooled analysis showed OR 4.48 (3.37-5.95 95%CI), with I^2 of 25% ($p = 0.25$) (Figure 3.11). The funnel plot showed some symmetry with a few studies which were outliers (Figure 3.12). Egger's test reported $p < 0.22$.

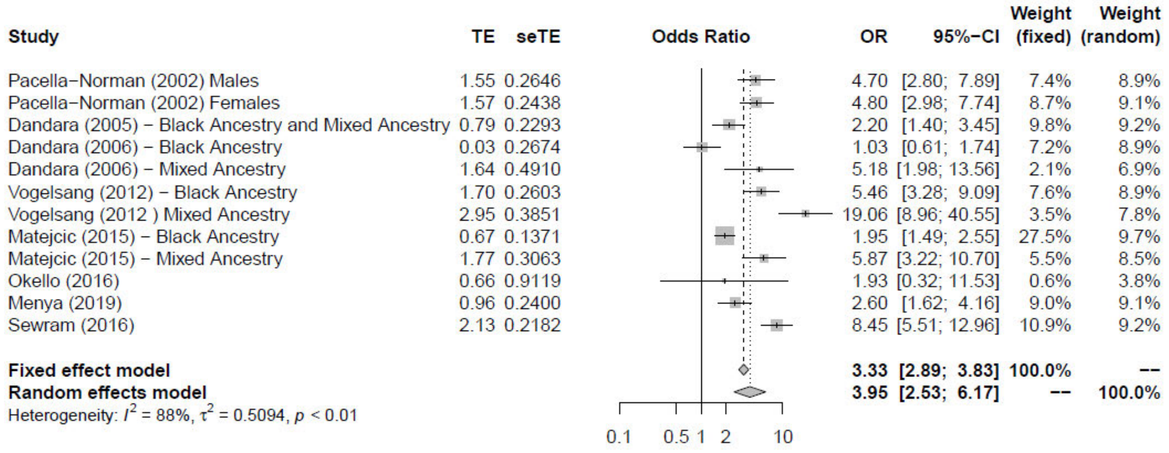


Figure 3.10: Effect of combined tobacco and alcohol use on ESCC in Africa. *Forest plot showing the pooled effect of combined tobacco and alcohol use on ESCC development in Africa. The figure was generated using R software. Study ID gives the first author and the year of the publication. OR, odds ratio: TE, logged effect size: seTE, standard error of effect size: CI, confidence interval*

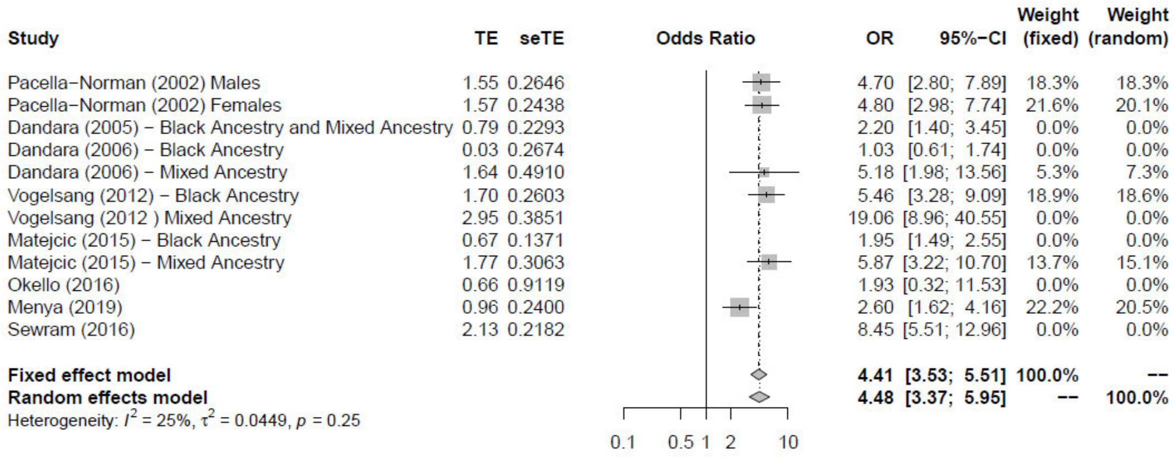


Figure 3.11: Final estimates of the effect of combined tobacco and alcohol consumption on ESCC in Africa*. Forest plot showing the pooled effect of combined tobacco and alcohol use on ESCC development in Africa, after excluding outliers and studies which account for distorted ES, high-heterogeneity and clustering. The figure was generated using R software. Study ID gives the first author and the year of the publication. OR, odds ratio: TE, logged effect size: seTE, standard error of effect size: CI, confidence interval*

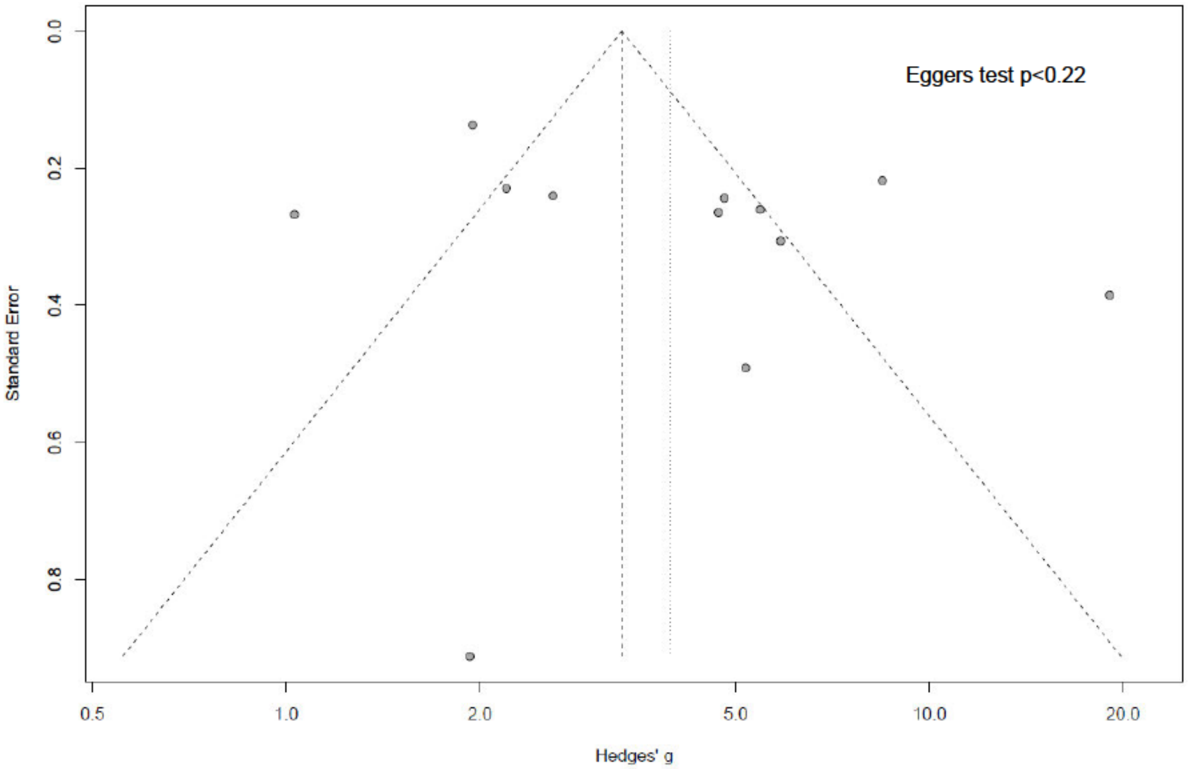
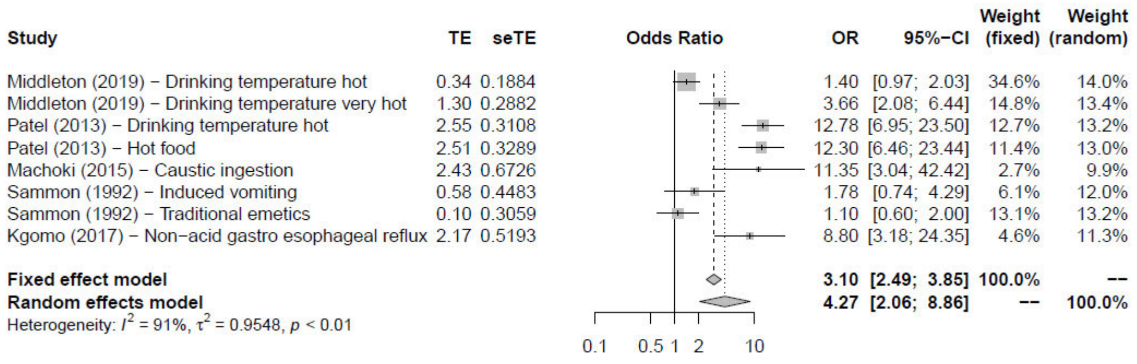


Figure 3.12: Funnel plot for assessing publication bias for combined tobacco and alcohol. *The figure was generated using R software.*

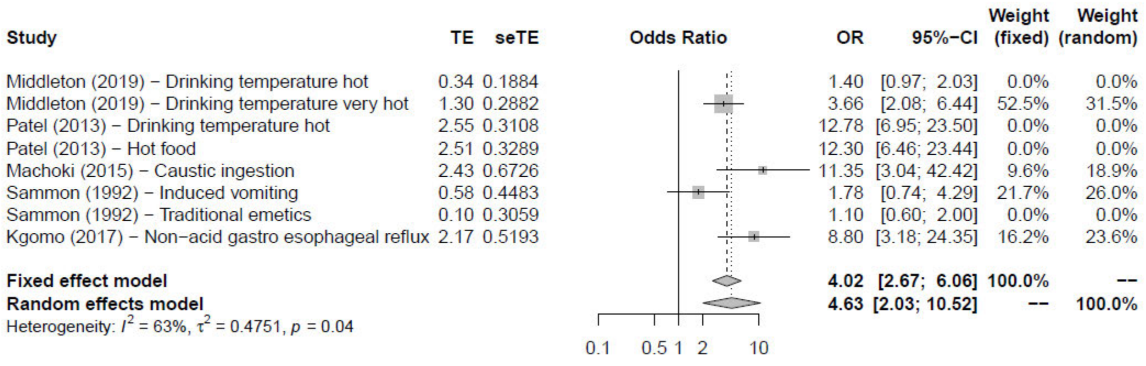
An overall effect estimate of OR 4.27 (2.06-8.86 95%CI) and I^2 of 91% ($p < 0.01$) was reported for the pooled analysis of esophageal injury exposure, showing that esophageal injury doubles the risk of developing EC (Figure 3.13). A second pooled analysis excluding outliers showed OR 4.63 (2.03-10.52 95%CI) and I^2 of 63% (p = 0.04) (Figure 3.14). The funnel plot showed asymmetry (Figure 3.15). Egger's test could not be performed due to the small number of studies. One of the studies included in this analysis, Machoki et al (74) is not indexed on PubMed®, and contributes to the final effect size.



Study	TE	seTE	OR	95%−CI	Weight (fixed)	Weight (random)
Middleton (2019) − Drinking temperature hot	0.34	0.1884	1.40	[0.97; 2.03]	34.6%	14.0%
Middleton (2019) − Drinking temperature very hot	1.30	0.2882	3.66	[2.08; 6.44]	14.8%	13.4%
Patel (2013) − Drinking temperature hot	2.55	0.3108	12.78	[6.95; 23.50]	12.7%	13.2%
Patel (2013) − Hot food	2.51	0.3289	12.30	[6.46; 23.44]	11.4%	13.0%
Machoki (2015) − Caustic ingestion	2.43	0.6726	11.35	[3.04; 42.42]	2.7%	9.9%
Sammon (1992) − Induced vomiting	0.58	0.4483	1.78	[0.74; 4.29]	6.1%	12.0%
Sammon (1992) − Traditional emetics	0.10	0.3059	1.10	[0.60; 2.00]	13.1%	13.2%
Kgomo (2017) − Non−acid gastro esophageal reflux	2.17	0.5193	8.80	[3.18; 24.35]	4.6%	11.3%
Fixed effect model			**3.10**	**[2.49; 3.85]**	**100.0%**	**--**
Random effects model			**4.27**	**[2.06; 8.86]**	**--**	**100.0%**

Heterogeneity: $I^2 = 91\%$, $\tau^2 = 0.9548$, $p < 0.01$

Figure 3.13: Effect of esophageal injury on ESCC in Africa. *Forest plot showing the pooled effect of esophageal injury on ESCC development in Africa. The figure was generated using R software. Study ID gives the first author and the year of the publication. OR, odds ratio: TE, logged effect size: seTE, standard error of effect size: CI, confidence interval*



Study	TE	seTE	OR	95%−CI	Weight (fixed)	Weight (random)
Middleton (2019) − Drinking temperature hot	0.34	0.1884	1.40	[0.97; 2.03]	0.0%	0.0%
Middleton (2019) − Drinking temperature very hot	1.30	0.2882	3.66	[2.08; 6.44]	52.5%	31.5%
Patel (2013) − Drinking temperature hot	2.55	0.3108	12.78	[6.95; 23.50]	0.0%	0.0%
Patel (2013) − Hot food	2.51	0.3289	12.30	[6.46; 23.44]	0.0%	0.0%
Machoki (2015) − Caustic ingestion	2.43	0.6726	11.35	[3.04; 42.42]	9.6%	18.9%
Sammon (1992) − Induced vomiting	0.58	0.4483	1.78	[0.74; 4.29]	21.7%	26.0%
Sammon (1992) − Traditional emetics	0.10	0.3059	1.10	[0.60; 2.00]	0.0%	0.0%
Kgomo (2017) − Non−acid gastro esophageal reflux	2.17	0.5193	8.80	[3.18; 24.35]	16.2%	23.6%
Fixed effect model			**4.02**	**[2.67; 6.06]**	**100.0%**	**--**
Random effects model			**4.63**	**[2.03; 10.52]**	**--**	**100.0%**

Heterogeneity: $I^2 = 63\%$, $\tau^2 = 0.4751$, $p = 0.04$

Figure 3.14: Final estimates of effect of esophageal injury on ESCC in Africa. *Forest plot showing the pooled effect of esophageal injury on ESCC development in Africa, after excluding outliers and studies which account for distorted ES, high-heterogeneity and clustering. The figure was generated using R software. Study ID gives the first author and the year of the publication. OR, odds ratio: TE, logged effect size: seTE, standard error of effect size: CI, confidence interval*

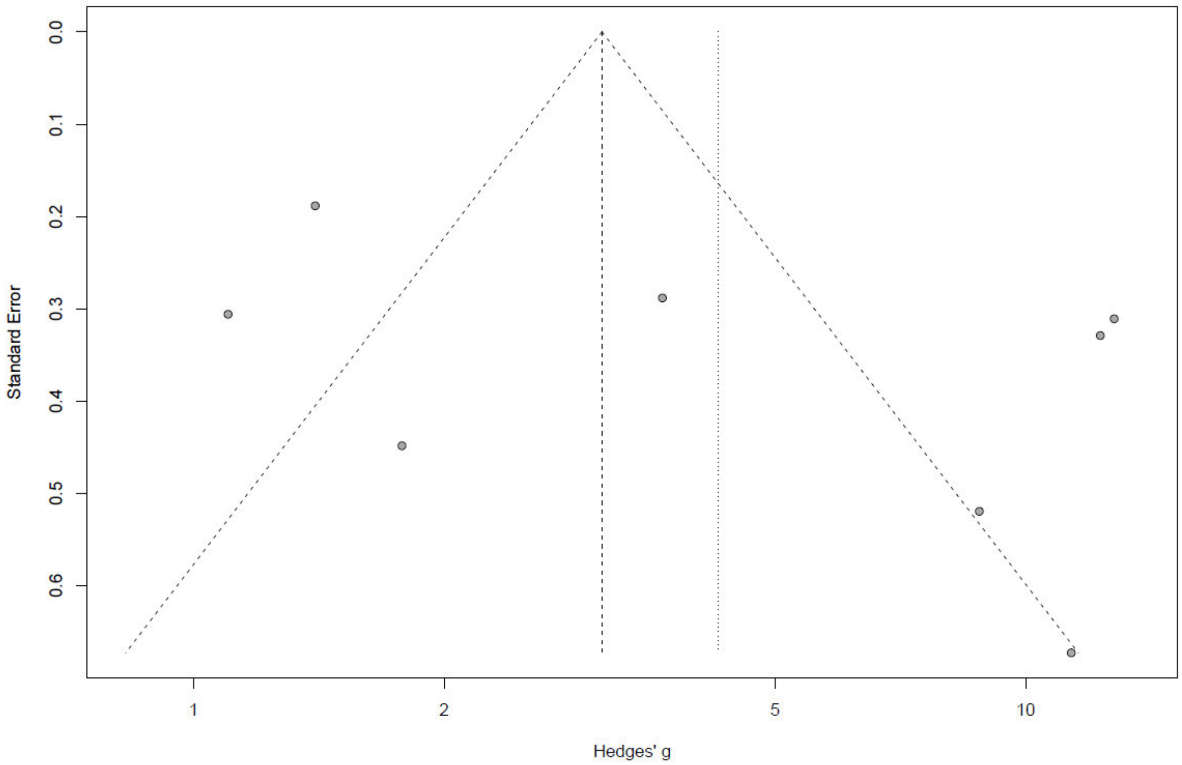
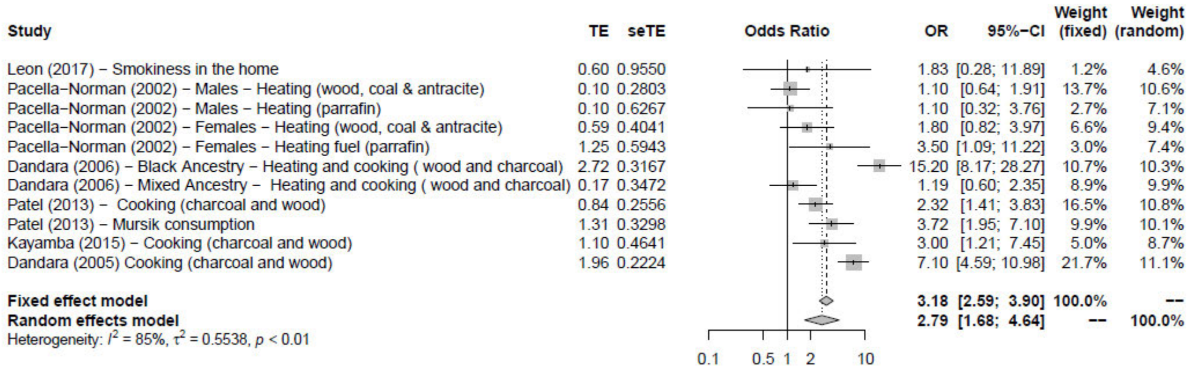


Figure 3.15: Funnel plot for assessing publication bias for esophageal injury. *The figure was generated using R software. Egger's test could not be performed.*

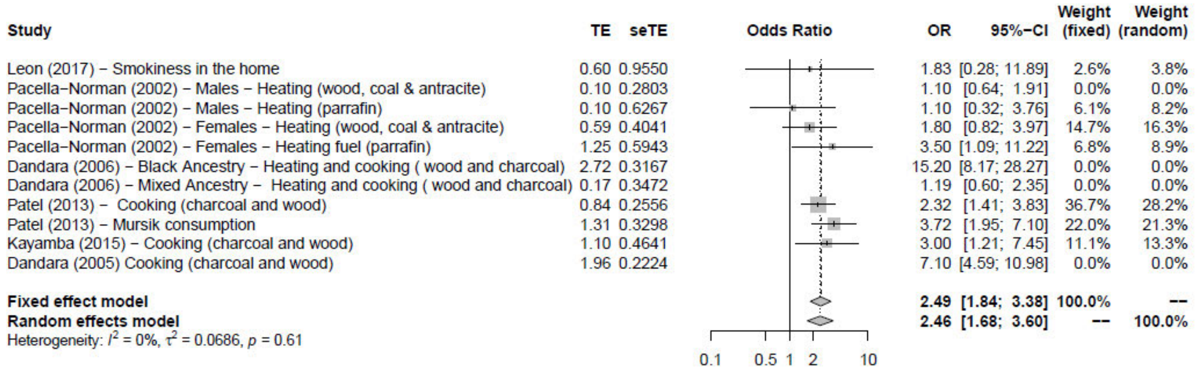
A forest plot of PAH exposure showed an effect estimate of OR of 2.79 (1.68-4.64 95%CI) (Figure 3.16). This indicates that exposure to PAHs increases the risk of developing EC by a factor of 2. Test for heterogeneity showed an I^2 of 85% ($p < 0.01$). The second pooled analysis resulted in OR 2.46 (1.68-3.60 95%CI) and I^2 of 0% (Figure 3.17). The funnel plot shows moderate symmetry and some outliers (Figure 3.18). Egger's test gave $p < 0.43$.



Study	TE	seTE	OR	95%−CI	Weight (fixed)	Weight (random)
Leon (2017) − Smokiness in the home	0.60	0.9550	1.83	[0.28; 11.89]	1.2%	4.6%
Pacella−Norman (2002) − Males − Heating (wood, coal & antracite)	0.10	0.2803	1.10	[0.64; 1.91]	13.7%	10.6%
Pacella−Norman (2002) − Males − Heating (parrafin)	0.10	0.6267	1.10	[0.32; 3.76]	2.7%	7.1%
Pacella−Norman (2002) − Females − Heating (wood, coal & antracite)	0.59	0.4041	1.80	[0.82; 3.97]	6.6%	9.4%
Pacella−Norman (2002) − Females − Heating fuel (parrafin)	1.25	0.5943	3.50	[1.09; 11.22]	3.0%	7.4%
Dandara (2006) − Black Ancestry − Heating and cooking (wood and charcoal)	2.72	0.3167	15.20	[8.17; 28.27]	10.7%	10.3%
Dandara (2006) − Mixed Ancestry − Heating and cooking (wood and charcoal)	0.17	0.3472	1.19	[0.60; 2.35]	8.9%	9.9%
Patel (2013) − Cooking (charcoal and wood)	0.84	0.2556	2.32	[1.41; 3.83]	16.5%	10.8%
Patel (2013) − Mursik consumption	1.31	0.3298	3.72	[1.95; 7.10]	9.9%	10.1%
Kayamba (2015) − Cooking (charcoal and wood)	1.10	0.4641	3.00	[1.21; 7.45]	5.0%	8.7%
Dandara (2005) Cooking (charcoal and wood)	1.96	0.2224	7.10	[4.59; 10.98]	21.7%	11.1%
Fixed effect model			**3.18**	**[2.59; 3.90]**	**100.0%**	**--**
Random effects model			**2.79**	**[1.68; 4.64]**	**--**	**100.0%**

Heterogeneity: $I^2 = 85\%$, $\tau^2 = 0.5538$, $p < 0.01$

Figure 3.16: Effect of PAH on ESCC in Africa. *Forest plot showing the pooled effect of PAH on ESCC development in Africa. The figure was generated using R software. Study ID gives the first author and the year of the publication. OR, odds ratio: TE, logged effect size: seTE, standard error of effect size: CI, confidence interval*



Study	TE	seTE	OR	95%−CI	Weight (fixed)	Weight (random)
Leon (2017) − Smokiness in the home	0.60	0.9550	1.83	[0.28; 11.89]	2.6%	3.8%
Pacella−Norman (2002) − Males − Heating (wood, coal & antracite)	0.10	0.2803	1.10	[0.64; 1.91]	0.0%	0.0%
Pacella−Norman (2002) − Males − Heating (parrafin)	0.10	0.6267	1.10	[0.32; 3.76]	6.1%	8.2%
Pacella−Norman (2002) − Females − Heating (wood, coal & antracite)	0.59	0.4041	1.80	[0.82; 3.97]	14.7%	16.3%
Pacella−Norman (2002) − Females − Heating fuel (parrafin)	1.25	0.5943	3.50	[1.09; 11.22]	6.8%	8.9%
Dandara (2006) − Black Ancestry − Heating and cooking (wood and charcoal)	2.72	0.3167	15.20	[8.17; 28.27]	0.0%	0.0%
Dandara (2006) − Mixed Ancestry − Heating and cooking (wood and charcoal)	0.17	0.3472	1.19	[0.60; 2.35]	0.0%	0.0%
Patel (2013) − Cooking (charcoal and wood)	0.84	0.2556	2.32	[1.41; 3.83]	36.7%	28.2%
Patel (2013) − Mursik consumption	1.31	0.3298	3.72	[1.95; 7.10]	22.0%	21.3%
Kayamba (2015) − Cooking (charcoal and wood)	1.10	0.4641	3.00	[1.21; 7.45]	11.1%	13.3%
Dandara (2005) Cooking (charcoal and wood)	1.96	0.2224	7.10	[4.59; 10.98]	0.0%	0.0%
Fixed effect model			**2.49**	**[1.84; 3.38]**	**100.0%**	**--**
Random effects model			**2.46**	**[1.68; 3.60]**	**--**	**100.0%**

Heterogeneity: $I^2 = 0\%$, $\tau^2 = 0.0686$, $p = 0.61$

Figure 3.17: Final estimates of effect of PAH on ESCC in Africa. *Forest plot showing the pooled effect of PAH on ESCC development in Africa, after excluding outliers and studies which account for distorted ES, high-heterogeneity and clustering. The figure was generated using R software. Study ID gives the first author and the year of the publication. OR, odds ratio: TE, logged effect size: seTE, standard error of effect size: CI, confidence interval*

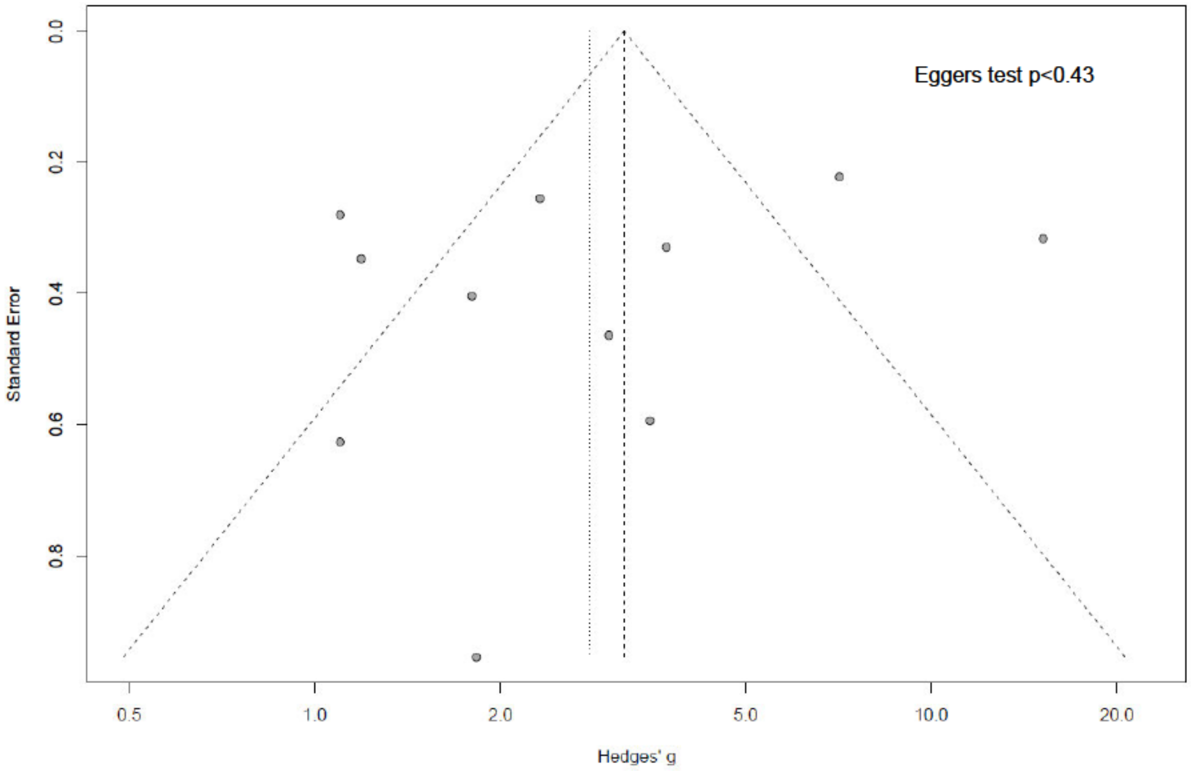


***Figure 3.18: Funnel plot for assessing publication bias in PAH studies.** The figure was generated using R software.*

Pooled analysis for fruit and vegetables resulted in an overall OR of 0.64 (0.47-0.87 95%CI) and I^2 of 70% (p < 0.01) (Figure 3.19). The second pooled analysis, after removal of outliers gave an OR 0.60 (0.43-0.83 95%CI) and I^2 of 45% (p = 0.08) (Figure 3.20). Funnel plot showed moderate symmetry (Figure 3.21), however an Egger's test could not be performed due to the low number of studies. Egger's test is accurate when done on ten or more studies.

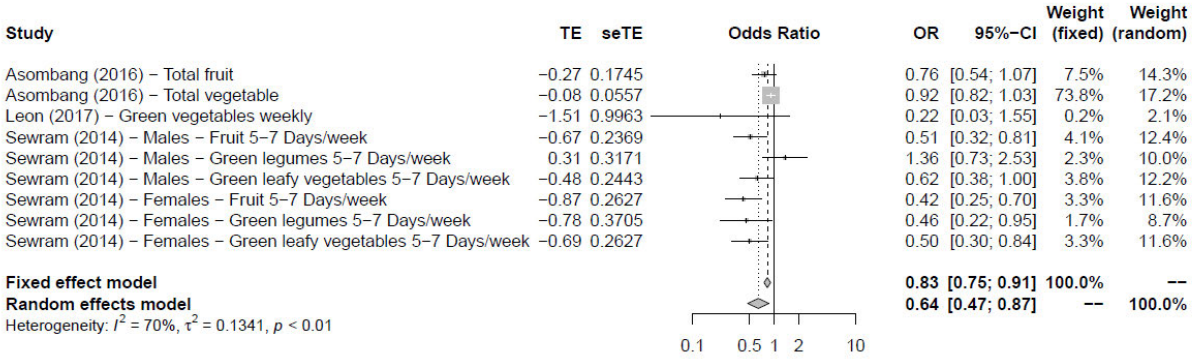


Figure 3.19: Effect of fruit and vegetables consumption on ESCC in Africa. *Forest plot showing the pooled effect of fruit and vegetables consumption on ESCC development in Africa. The figure was generated using R software. Study ID gives the first author and the year of the publication. OR, odds ratio: TE, logged effect size: seTE, standard error of effect size: CI, confidence interval*

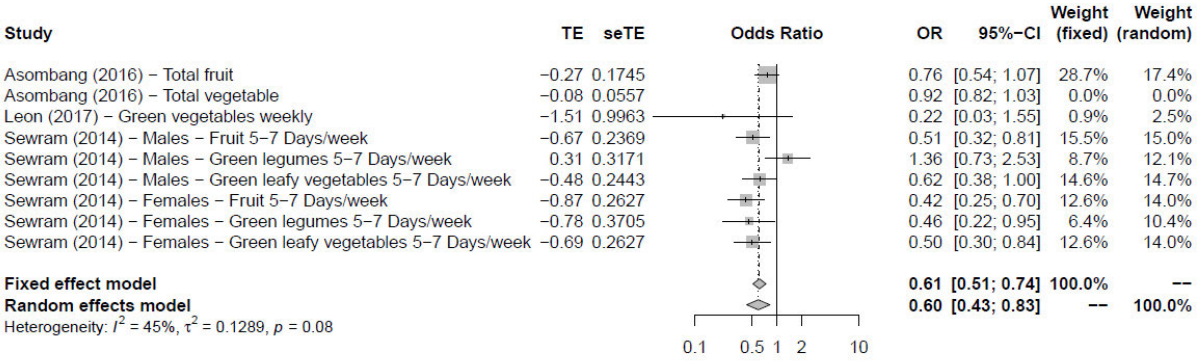


Figure 3.20: Final estimate of effect of fruit and vegetables consumption on ESCC in Africa. *Forest plot showing the pooled effect of fruit and vegetables consumption on ESCC development in Africa, after excluding outliers and studies which account for distorted ES, high-heterogeneity and clustering. The figure was generated using R software. Study ID gives the first author and the year of the publication. OR, odds ratio: TE, logged effect size: seTE, standard error of effect size: CI, confidence interval*

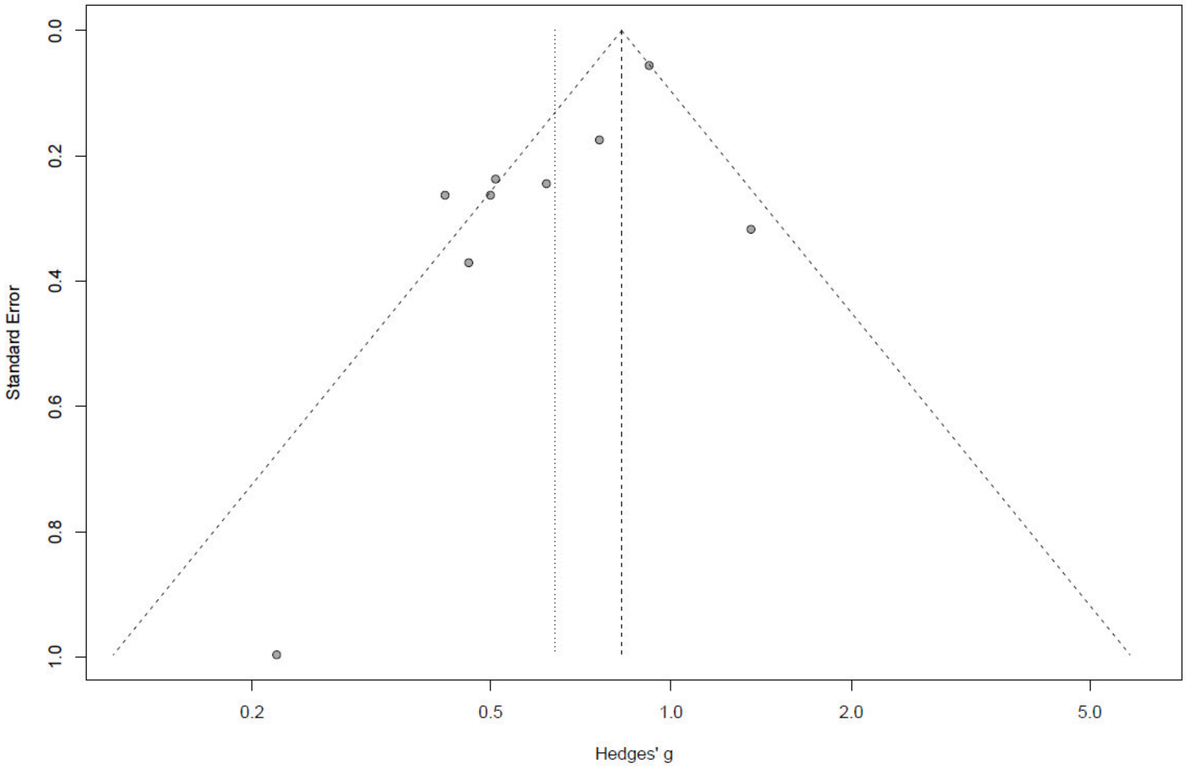


***Figure 3.21: Funnel plot for assessing publication bias for fruit and vegetable consumption.** The figure was generated using R software. Egger's test could not be performed.*

3.5.16.2 Population attributable fraction

PAF calculations were done for six risk factors: tobacco smoking, alcohol consumption, combined tobacco and alcohol use, esophageal injury, fruit and vegetable consumption and PAH exposure (Table 3.3). The data were taken from the studies selected for the meta-analysis. Five studies were excluded from the analysis due to not having enough information on exposure. The PAF attributable to tobacco smoking was 17%, whilst for alcohol consumption it was 13%. According to our analysis, the combined exposure of tobacco and alcohol attributed 23% of the esophageal cancers. Esophageal injury was responsible for 17% of ESCC cases. Exposure to PAHs contributed 5% of ESCC cases. Fruit and vegetable consumption, due to its protective effect, showed a negative PAF of -11%. Our estimates show that

43% of ESCC cases are attributable to the combined effects of tobacco smoking, alcohol consumption, esophageal injury and PAH exposure.

Table 3.3: PAF values for risk factors associated with ESCC in African populations

Risk factor for ESCC	**PAF value (%)**	**Studies from which the PAF was calculated**
Alcohol	0.17 (17)	(11, 46, 48, 49, 51-53, 55, 57, 71, 75-82)
Tobacco	0.13 (13)	(11, 46, 52, 53, 61, 71, 74, 78, 80, 81, 83)
Tobacco and alcohol	0.23 (23)	(11, 49, 53, 55, 77, 79, 81)
Esophageal injury	0.17 (17)	(52, 56, 64, 84)
PAH	0.05 (5)	(46, 49, 52, 75, 78)
Fruits and vegetables	-0.11	(12, 78)
Total*	0.43 (43)	

*Combined PAF for Alcohol, tobacco, esophageal injury and PAH only

3.6 Discussion

EC constitutes a major health burden in specific geographic regions of the world, i.e., China to Iran, parts of South America and from Eastern to Southern Africa.(6) Our systematic review identified 31 studies which reported smoking and alcohol consumption, low socioeconomic status, poor diet, PAH exposure, consumption of hot food and beverages, poor oral health, infectious agents, esophageal injury, family history of cancer, water source, and non-acid gastro-esophageal reflux as environmental and life-style risk factors for ESCC in Africa using risk estimates. This points to a multifactorial etiology of ESCC which was also reported in two other recent systematic reviews, one published while the current study was in progress.(33, 85) Most of the studies in our systematic review were reported from the African oesophageal cancer corridor. We performed a meta-analysis of eligible studies. Our study also aggregated data from additional sources other than PubMed and performed PAF analysis, which has not been done before.

Meta-analysis was done for the following risk factors, which had enough studies for a pooled analysis: tobacco smoking, alcohol consumption, combined smoking and tobacco exposure, PAH exposure, esophageal injury, as well as fruit and vegetable consumption. Our systematic review differed from the already published systematic review(85) in that we performed meta-analysis on an additional four risk factors (combined smoking and tobacco exposure, PAH exposure, esophageal injury, and fruit and vegetable consumption) and performed PAF analysis.

We performed two meta-analysis for each risk factors. The initial meta-analyses for all risk factors showed significant heterogeneity, and therefore for each pooled analysis a second meta-analysis was done after removal of outliers and studies which accounted for the distorted effect size, high-heterogeneity and clustering.

Tobacco smoking was the most studied risk factor and emerged as a plausible contributing agent for ESCC in South Africa, Uganda, Zambia, and Zimbabwe. Alcohol consumption was the second most reported risk factor, however only 38% of the studies reported statistically significant associations. Our pooled analysis showed that tobacco smoking increased ESCC risk by 3.07 (2.39-3.94 95%CI), but the heterogeneity of the studies was high (I^2 = 95%). The second meta-analysis after the removal of outliers and studies which account for distorted effect size, high-heterogeneity and clustering showed an OR of 4.14 (3.26-5.26 95%CI) and I^2 of 36% ($p = 0.09$). The heterogeneity was reduced from 95% to 36% and the effect size increased. The funnel plot analysis showed potential publication bias, possibly due to meta-analysis of non-standardized exposures and intensities.

Home-brewed beer consumption was a major risk factor for ESCC in South African, Zambian and Kenyan populations. Preparation of this beer is often done in oil drums which may contain iron and other carcinogens.(86) Our meta-analysis showed that alcohol consumption increased ESCC risk of OR 2.31 (1.77-3.02 95%CI) and high heterogeneity (I^2 = 83%). The second meta-analysis after the removal of outliers showed OR of 2.14 (1.65-2.78 95%CI) and I^2 of 57% ($p < 0.01$), also showing a significant reduction in heterogeneity. Alcohol consumption therefore increased the risk of developing ESCC by a factor of two. The funnel plot showed moderate symmetry (Figure 9) and the Egger's test had a $p < 0.07$. This means that publication bias in this pooled analysis was unlikely, although a few outliers are present.

Whilst data from high income countries have conclusively implicated tobacco smoking and alcohol consumption as the main risk factors for ESCC(16), more evidence is needed to assess if they are major causative agents of ESCC in the African EC corridor. Some studies reported the interaction between tobacco and alcohol, with high risk estimates, indicative of synergistic effects of combined exposure to tobacco and alcohol. However, because none of these studies were originally designed as interaction studies, they may not have had enough statistical power to detect interactions. In a systematic review and meta-analyses done by Prabhu et al (87) pooled analysis of five studies from Asian populations showed a positive synergistic effect of tobacco and alcohol exposure. The synergy factor was reported as an OR of 3.28 (2.11- 508 95%CI), Cochrane's Q P value = 0.05 and I^2 = 55.3 %.(87) The authors reported that the combined effect of tobacco and alcohol exposure was approximately twice that of each risk factor alone.(87) In our study, pooled analysis of combined tobacco and alcohol use showed similar findings with OR 3.95 (2.53-6.17 95%CI) and an I^2 of 88%. The second pooled analysis following the removal of outliers showed OR 4.48 (3.37-5.95 95%CI), with I^2 of 25% ($p = 0.25$). The results support the synergistic effects of combined alcohol and tobacco exposure reported in literature, with a two-fold increase in risk compared to alcohol consumption alone. Regarding tobacco smoking alone, the combined effect of tobacco smoking and alcohol consumption only slightly increases. A significant reduction in heterogeneity was reported, from 88% to 25%. The funnel plot showed moderate symmetry with some outliers, indicating that publication bias was unlikely.

Esophageal injury emerged as an important risk factor for ESCC in our meta-analysis, however, only three studies have assessed this risk factor in African populations. Esophageal injury can occur due to consumption of hot food and beverages, self-induced vomiting, use of traditional emetics, and non-acid gastroesophageal reflux. In our study, the role of self-induced vomiting and traditional emetics in ESCC development was conflicting. Over the past few years, there has been an increase in the number of African studies assessing the role of hot food and beverages in ESCC pathogenesis. The studies reported an association, corroborated with other studies in China, Iran and South America which demonstrated the same effect.(33) Additionally, results from our meta-analysis were consistent with these findings, showing that esophageal injury was associated with increased risk of ESCC, with OR 4.27 (2.06-8.86 95%CI) and I^2 of 91% ($p < 0.01$). A second pooled analysis showed OR 4.63 (2.03-10.52 95%CI) and I^2 of 63% ($p = 0.04$). Esophageal injury therefore increases the risk of developing ESCC by a factor on four.

Similar to our study, low socioeconomic status was reported to be associated with increased ESCC risk in Indian, American and Swedish studies.(88-90) Our study assessed PAHs from different sources (heating and cooking fuel, and consumption of charcoal powder when drinking mursik) and was found to increase ESCC risk. In a systematic review on the role of biomass fuel (wood, charcoal, coal, dung, and crop residues) in ESCC development, the use of biomass fuel for heating and cooking was associated with ESCC development, due to smoke exposure.(30) The highest risk was reported in Africa and Asia. These results were corroborated in our study, as pooled analysis demonstrated that PAH exposure was associated with increased risk of developing ESCC with OR of 2.79 (1.68-4.64 95%CI) and I^2 of 85% ($p < 0.01$). The second pooled analysis gave an OR 2.46 (1.68-3.60 95%CI) and I^2 of 0%. PAH exposure therefore increases the risk of developing ESCC by a factor of two. After removal of outliers, heterogeneity was reduced to 0%, representing the biggest reduction in heterogeneity in our meta-analysis. The funnel plot shows moderate symmetry and some outliers, with eggers test $p < 0.43$.

Consumption of fruits and vegetables reduced the risk of ESCC in our study. This evidence is supported by a previous meta-analysis done by Liu et al (29) on 32 studies on Asian, European, North and South American populations. The study reported that consumption of vegetables and fruits significantly reduced the risk of ESCC with summary relative risks of 0.56 (0.45–0.69 95%CI) and 0.53 (0.44–0.64 95%CI), respectively.(29) Our pooled analysis for fruit and vegetable consumption among African populations showed OR 0.64 (0.47-0.87 95%CI) and I^2 of 70% ($p < 0.01$). A second pooled analysis after removal of outliers gave an OR 0.60 (0.43-0.83 95%CI) and I^2 of 45% ($p = 0.08$). Fruit and vegetable consumption had a protective effect on ESCC, meaning it reduced the risk of developing ESCC by 40%. Due to the small number of studies, fruit and vegetables consumption could not be analysed separately. The role of micronutrient deficiencies in the etiology of ESCC is contested in the literature. Our study found that micronutrient deficiencies of iron, zinc, selenium and magnesium increase the risk of ESCC. In one of the biggest micronutrient studies done in 32 African countries, iron, zinc and selenium were described to have a protective effect in males and females, whilst magnesium was reported to be protective in females only.(91)The study was an ecological study, which is an observational study where data are analysed for entire populations in different geographical regions at a single point in time. In a systematic review on micronutrients and EC by Velenzuela et al(92) increased dietary intake of total iron and zinc, and reduced heme iron intake was reported to be protective against.

Our meta-analysis showed that tobacco smoking, alcohol use, combined tobacco and alcohol use, PAH exposure and esophageal injury all enhanced the risk of ESCC. However, these results also showed high heterogeneity, and hence should be interpreted with caution. It is inevitable that studies pooled together in a meta-analysis will have some level of heterogeneity. Based on the recommendations of the Cochrane Handbook for Systematic Reviews of Interventions version 6.0 (93), five (tobacco exposure, alcohol consumption, combined tobacco and alcohol exposure, PAH, and esophageal injury) of our initial meta-analysis results fall within the 75-90% heterogeneity, indicating considerable heterogeneity. Fruit and vegetable consumption fell within the 50% to 90%, representing substantial heterogeneity. The second set of meta-analyses after removal of outliers resulted in three (tobacco exposure, alcohol consumption, fruit and vegetables exposure) meta-analyses falling within 30% to 60%, representing moderate heterogeneity. The analysis on PAH fell within 0% to 40% of heterogeneity whilst the analysis on esophageal injury had heterogeneity within 50% to 90%. It is important to note that our meta-analysis was performed using the random effects model, which incorporates heterogeneity among studies.

Our study provides evidence-based assessment of the proportion of ESCC cases attributable to certain risk factors. To quantify the contribution of a risk factor to the development of ESCC, PAF analysis was performed. An estimation of PAF is imperative as it plays an important role in cancer control and prevention. Our analysis showed that tobacco smoking and alcohol consumption contributed to 17% and 13% of the ESCC cases, respectively, whilst their combined exposure contributed 23% of the ESCC cases. The combined PAF corroborates our results from the meta-analysis which also showed synergistic effects of combined and tobacco exposure. In a Chinese study, the contribution of alcohol consumption in EC cases was reported to be PAF of 10.9% (15.2% for men and 1.3% for women) (94), which was lower than our overall estimate of 17%. Another Chinese study reported the combined contribution of tobacco smoking and alcohol consumption in ESCC as 40.9%.(95) A study done on a Ugandan population reported PAF values of 15.62 for tobacco smoking, 10.17% for alcohol consumption, and 13% for combined tobacco smoking and alcohol use.(81). A recent study on a Kenyan population reported that PAF of ESCC for alcohol consumption was 48% (59% for men and 27% for women).(79) A Lebanese study reported higher PAF values of tobacco smoking; (43%) and alcohol (31%) for men and smoking (33%), and alcohol (6%) for women.(96) Combined attributable risk for tobacco smoking and alcohol consumption in men was 76% and 37% in women.(96) In a meta-analysis of large-scale population-based cohort studies in Japan, the authors reported PAF values of 55.4%, 61.2%, and 81.4% for smoking, alcohol consumption, and combined smoking and alcohol consumption respectively.(97) An Australian study in 2013 reported that PAF for ESCC due to tobacco smoking was 49% and for alcohol consumption it was 32%.(98) Attributable fraction for ESCC was reported in a Parkistani population with the following PAF values; chewing areca nut (10.8%), chewing betel quid with tobacco (47.6%), oral snuff (10.1), and cigarette smoking (22.3).(99) Tobacco chewing and snuff use is a common practice in African populations, but understudied.

Whilst our findings from the PAF analysis point to tobacco use being one of the more important risk factors, interestingly, esophageal injury also showed a similar attributable fraction of 17%. Esophageal injury therefore emerged as an important risk factor. In a 2003 study done in Paraguay, maté consumption, which is normally consumed at high temperatures, had a population-attributable risk of 53%.(100) However the authors mention that two competitive mechanisms could explain this high attributable risk; the high temperature of the mate causing thermal injury of the esophagus, and the carcinogenic effects of the herbs used in preparation of the drink. Exposure to PAH contributed 5% of ESCC cases in our analysis. Whilst our study did not assess for low intake of fruits and vegetables, this has been shown in an Australian study to attribute to 9% of ESCC cases (98), and 2% in a Canadian study.(101)

In another study which used data from Global Burden of Diseases, Injuries, and Risk Factors Study (GBD) 2017, the authors identified smoking, alcohol use, high body mass index (BMI), a diet low in fruits, and chewing tobacco as contributing significantly to the proportion of oesophageal cancer disability-adjusted life-years (DALYs).(102) The data originated from 195 countries and territories between 1990 and 2017. The combined attributable risk of four of the risk factors (tobacco smoking, alcohol consumption, esophageal injury and PAH exposure) was reported to be 43%. This suggests that 43% of ESCC cases would be prevented if these exposures were removed. Overall our results show that certain risk factors are population- and region-specific, and point to a multifactorial etiology of ESCC.

Overall, our study showed that there is a relatively large body of evidence for smoking and alcohol being associated with ESCC, compared to other risk factors. Areas where there is an emerging body of evidence include hot food and beverages, and oral health. At the same time new avenues of research are also emerging in PAH exposure, and diet as risk factors. A number of South African studies reported on risk according to specific populations, i.e. black population and mixed ancestry. This South African mixed ancestry population is a highly admixed population (103), in which the predominant ancestral lines are Khoesan (32-43%), Bantu speaking Africans (20-36%), European (21-28%) and Asian (9-11%).(104)

Strengths

The strength of this study is that it provides the most comprehensive meta-analysis on environmental risk factors associated with ESCC in African populations. It is also the first study to perform an aggregated PAF analysis of multiple risk factors which contribute to ESCC cases in African populations.

Limitations

One of the main limitations was the heterogeneity of the original studies. Studies assessing the same risk factor often had a different study design, geographical location, exposure measurement, exposure assessment category, exposure intensity, and confounding factors adjusted for, variability in sample size (the majority of the studies were small case-control studies), therefore caution is needed when interpreting the magnitude of the risk estimates. None of the studies directly measured carcinogen levels in tobacco, alcohol or indoor smoke. Tobacco exposure had multiple routes of exposure, including smoking, chewing, and snuff, therefore the carcinogen levels may have differed. Acetaldehyde levels of alcohol also may have differed, especially in home-brewed beer.

A number of studies did not report risk estimates, and hence could not be included in the meta-analyses. Most of these studies were published before 1990. There were some deficiencies in the quality of reporting and methods in some of the studies. About one third of the studies did not report on adjusting for confounding factors. Unmeasured and unadjusted confounders can result in a false association. Response rates were also reported in only two of the studies.

There is a lack of interaction studies to investigate the multifactorial etiology of ESCC. There are a number of gene-environment interaction studies which have been published (105) assessing the interaction between genetic variants and family history of cancer, age, gender, food hygiene, eating habits,(106) tobacco smoking (107, 108) and alcohol consumption.(107-109) However, the majority of the studies have been published on Asian populations.

Whilst we reported on PAFs, there are potential biases in our estimates due to the uncertainty of the magnitude (RR) of the effect given the sparse literature and the lack of population-based estimates of exposure. Additionally, for risk factors such as alcohol consumption and tobacco smoking, sex aggregated data would have given a clearer picture of the attributable risk as prevalence of exposure may differ according to gender. Both the meta-analysis and PAF analysis combined fruits and vegetables as one factor, however, the nutritional value of fruits and vegetables differ.(110) The combination of these factors was done due to limited number of studies for analysis.

Whilst we identified these limitations in the included studies, we also acknowledge the challenges that come with doing research in low-to-middle-income countries. These include the availability of resources and infrastructure to perform research, including time and expenses. There is also a lack of suitable methods and technologies, which results in the use of non-standardized assessment tools. Most data collection tools are based on self-reporting of lifestyle behaviors and environmental exposures via questionnaires which have a low precision of accuracy and can result in recall and misclassification bias.(111) It is important to highlight the lack of prospective cohort studies, which have the capability of significantly reducing some of the biases common in case-control studies.

3.7 Conclusions

Studies investigating the etiology of ESCC in Africa are very limited, therefore more research needs to be done to understand the high prevalence seen in the African EC corridor. A standardized way of measuring risk factors will allow for future systematic reviews to report with certainty pooled estimates which can be generalized to the region. The results of our study point to a multifactorial etiology, which includes genetic predisposition and multiple environmental and life-style risk factors, playing a role in ESCC risk. More interaction studies, are, therefore needed to elucidate the etiology of ESCC in the African populations. Whilst the majority of the risk factors here can be generalized to most African countries, more studies are needed to investigated if there are risk factors which are specific to the African ESCC corridor, which may explain the high incidence in this region. In particular, the role of gene x environment interactions needs to be further investigate as well as the role of geochemistry.

3.8 References

1. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: a cancer journal for clinicians* (2018) 68(6):394-424. Epub 2018/09/13. doi: DOI10.3322/caac.21492. PubMed PMID: 30207593.

2. Van Loon K, Mwachiro MM, Abnet CC, Akoko L, Assefa M, Burgert SL, et al. The African Esophageal Cancer Consortium: A Call to Action. *Journal of global oncology* (2018) (4):1-9. Epub 2018/09/23. doi: 10.1200/jgo.17.00163. PubMed PMID: 30241229; PubMed Central PMCID: PMCPMC6223465.

3. Arnold M, Laversanne M, Brown LM, Devesa SS, Bray F. Predicting the Future Burden of Esophageal Cancer by Histological Subtype: International Trends in Incidence up to 2030. *The American Journal Of Gastroenterology* (2017) 112:1247. doi: 10.1038/ajg.2017.155

4. Murphy G, McCormack V, Abedi-Ardekani B, Arnold M, Camargo MC, Dar NA, et al. International cancer seminars: a focus on esophageal squamous cell carcinoma. *Ann Oncol* (2017) 28(9):2086-93. Epub 2017/09/16. doi: 10.1093/annonc/mdx279. PubMed PMID: 28911061.

5. Codipilly DC, Qin Y, Dawsey SM, Kisiel J, Topazian M, Ahlquist D, et al. Screening for esophageal squamous cell carcinoma: recent advances. *Gastrointestinal endoscopy* (2018) 88(3):413-26. Epub 2018/05/02. doi: doi:10.1016/j.gie.2018.04.2352. PubMed PMID: 29709526.

6. Abnet CC, Arnold M, Wei WQ. Epidemiology of Esophageal Squamous Cell Carcinoma. *Gastroenterology* (2017) 154(2):360-73. Epub 2017/08/22. doi: doi:10.1053/j.gastro.2017.08.023. PubMed PMID: 28823862; PubMed Central PMCID: PMCPMC5836473.

7. Munishi MO, Hanisch R, Mapunda O, Ndyetabura T, Ndaro A, Schüz J, et al. Africa's oesophageal cancer corridor - do hot beverages contribute? *Cancer causes & control : CCC* (2015) 26(10):1477-86. doi: 10.1007/s10552-015-0646-9. PubMed PMID: PMC4838015.

8. Parkin DM, Bray F, Ferlay J, Jemal A. Cancer in Africa 2012. *Cancer epidemiology, biomarkers & prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology* (2014) 23(6):953-66. Epub 2014/04/05. doi: 10.1158/1055-9965.epi-14-0281. PubMed PMID: 24700176.

9. McGlashan ND. Oesophageal cancer and alcoholic spirits in central Africa. *Gut* (1969) 10(8):643-50. Epub 1969/08/01. PubMed PMID: 5810975; PubMed Central PMCID: PMCPMC1552917.

10. Odera JO, Odera E, Githang'a J, Walong EO, Li F, Xiong Z, et al. Esophageal cancer in Kenya. *American journal of digestive disease* (2017) 4(3):23.

11. Sewram V, Sitas F, O'Connell D, Myers J. Tobacco and alcohol as risk factors for oesophageal cancer in a high incidence area in South Africa. *Cancer epidemiology* (2016) 41:113-21.

12. Sewram V, Sitas F, O'Connell D, Myers J. Diet and esophageal cancer risk in the Eastern Cape Province of South Africa. *Nutrition and cancer* (2014) 66(5):791-9.

13. Menya D, Kigen N, Oduor M, Maina SK, Some F, Chumba D, et al. Traditional and commercial alcohols and esophageal cancer risk in Kenya. *Int J Cancer* (2019) 144(3):459-69. Epub 2018/08/18. doi: 10.1002/ijc.31804. PubMed PMID: 30117158; PubMed Central PMCID: PMCPMC6294681.

14. Simba H, Kuivaniemi H, Lutje V, Tromp G, Sewram V. Systematic Review of Genetic Factors in the Etiology of Esophageal Squamous Cell Carcinoma in African Populations. *Frontiers in genetics* (2019) 10:642.

15. Alaouna M, Hull R, Penny C, Dlamini Z. Esophageal cancer genetics in South Africa. *Clinical and experimental gastroenterology* (2019) 12:157-77. doi: 10.2147/CEG.S182000. PubMed PMID: 31114287.

16. Abnet CC, Arnold M, Wei W-Q. Epidemiology of esophageal squamous cell carcinoma. *Gastroenterology* (2017).

17. Simba H, Kuivaniemi H, Lutje V, Tromp G, Sewram V. Systematic Review of Genetic Factors in the Etiology of Esophageal Squamous Cell Carcinoma in African Populations. *Front Genet* (2019) 10:642. Epub 2019/08/21. doi: 10.3389/fgene.2019.00642. PubMed PMID: 31428123; PubMed Central PMCID: PMCPMC6687768.

18. Moher D, Liberati A, Tetzlaff J, Altman DG. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *Annals of internal medicine* (2009) 151(4):264-9.

19. Schultz T, Florence Z. *Joanna Briggs institute meta-analysis of statistics assessment and review instrument*. Adelaide: Joanna Briggs Institute (2007).

20. Team RC. *R: A language and environment for statistical computing*. Vienna, Austria (2013).

21. Viechtbauer W, Cheung MW-L. Outlier and influence diagnostics for meta-analysis. *Research Synthesis Methods* (2010) 1(2):112-25. doi: https://doi.org/10.1002/jrsm.11.

22. Olkin I, Dahabreh IJ, Trikalinos TA. GOSH – a graphical display of study heterogeneity. *Research Synthesis Methods* (2012) 3(3):214-23. doi: https://doi.org/10.1002/jrsm.1053.

23. Levin ML. The occurrence of lung cancer in man. *Acta Unio Int Contra Cancrum* (1953) 9(3):531-41. Epub 1953/01/01. PubMed PMID: 13124110.

24. Grant RL. Converting an odds ratio to a range of plausible relative risks for better communication of research findings. *BMJ : British Medical Journal* (2014) 348:f7450. doi: 10.1136/bmj.f7450.

25. Ezzati M, Hoorn SV, Rodgers A, Lopez AD, Mathers CD, Murray CJ. Estimates of global and regional potential health gains from reducing multiple major risk factors. *Lancet* (2003) 362(9380):271-80. Epub 2003/08/02. doi: 10.1016/s0140-6736(03)13968-2. PubMed PMID: 12892956.

26. Castellsagué X, Muñoz N, De Stefani E, Victora CG, Castelletto R, Rolón PA, et al. Independent and joint effects of tobacco smoking and alcohol drinking on the risk of esophageal cancer in men and women. *International Journal of Cancer* (1999) 82(5):657-64. doi: https://doi.org/10.1002/(SICI)1097-0215(19990827)82:5<657::AID-IJC7>3.0.CO;2-C.

27. Prabhu A, Obi KO, Rubenstein JH. The Synergistic Effects of Alcohol and Tobacco Consumption on the Risk of Esophageal Squamous Cell Carcinoma: A

Meta-Analysis. *Official journal of the American College of Gastroenterology | ACG* (2014) 109(6):822-7. doi: 10.1038/ajg.2014.71. PubMed PMID: 00000434-201406000-00010.

28. Hošnjak L, Poljak M. A systematic literature review of studies reporting human papillomavirus (HPV) prevalence in esophageal carcinoma over 36 years (1982-2017). *Acta Dermatovenerol Alp Pannonica Adriat* (2018) 27(3):127-36. Epub 2018/09/24. PubMed PMID: 30244262.

29. Liu J, Wang J, Leng Y, Lv C. Intake of fruit and vegetables and risk of esophageal squamous cell carcinoma: a meta-analysis of observational studies. *Int J Cancer* (2013) 133(2):473-85. Epub 2013/01/16. doi: 10.1002/ijc.28024. PubMed PMID: 23319052.

30. Okello S, Akello SJ, Dwomoh E, Byaruhanga E, Opio CK, Zhang R, et al. Biomass fuel as a risk factor for esophageal squamous cell carcinoma: a systematic review and meta-analysis. *Environmental Health: A Global Access Science Source* (2019) 18(1):60-. doi: 10.1186/s12940-019-0496-0. PubMed PMID: 31262333.

31. Zhang M, Wang J, Zhao Q, Mishra V, Fan J, Sun Y. Polycyclic aromatic hydrocarbons (PAHs) and esophageal carcinoma in Handan-Xingtai district, North China: a preliminary study based on cancer risk assessment. *Environmental Monitoring and Assessment* (2020) 192(9):596. doi: 10.1007/s10661-020-08499-5.

32. Roshandel G, Semnani S, Malekzadeh R, Dawsey SM. Polycyclic aromatic hydrocarbons and esophageal squamous cell carcinoma. *Arch Iran Med* (2012) 15(11):713-22. Epub 2012/10/30. doi: 0121511/aim.0013. PubMed PMID: 23102250; PubMed Central PMCID: PMCPMC5757504.

33. Chetwood JD, Garg P, Finch P, Gordon M. Systematic review: the etiology of esophageal squamous cell carcinoma in low-income settings. *Expert Rev Gastroenterol Hepatol* (2019) 13(1):71-88. Epub 2019/02/23. doi: 10.1080/17474124.2019.1543024. PubMed PMID: 30791842.

34. Maghsudlu M, Farashahi Yazd E. Heat-induced inflammation and its role in esophageal cancer. *J Dig Dis* (2017) 18(8):431-44. Epub 2017/07/28. doi: 10.1111/1751-2980.12511. PubMed PMID: 28749599.

35. Islami F, Boffetta P, Ren J-S, Pedoeim L, Khatib D, Kamangar F. High-temperature beverages and foods and esophageal cancer risk—A systematic review. *International Journal of Cancer* (2009) 125(3):491-524. doi: https://doi.org/10.1002/ijc.24445.

36. Dar NA, Islami F, Bhat GA, Shah IA, Makhdoomi MA, Iqbal B, et al. Poor oral hygiene and risk of esophageal squamous cell carcinoma in Kashmir. *Br J Cancer* (2013) 109(5):1367-72. Epub 2013/08/01. doi: 10.1038/bjc.2013.437. PubMed PMID: 23900216; PubMed Central PMCID: PMCPMC3778293.

37. Wei WQ, Abnet CC, Lu N, Roth MJ, Wang GQ, Dye BA, et al. Risk factors for oesophageal squamous dysplasia in adult inhabitants of a high risk region of China. *Gut* (2005) 54(6):759-63. doi: 10.1136/gut.2004.062331. PubMed PMID: 15888779.

38. Chen T, Cheng H, Chen X, Yuan Z, Yang X, Zhuang M, et al. Family history of esophageal cancer increases the risk of esophageal squamous cell carcinoma. *Scientific Reports* (2015) 5(1):16038. doi: 10.1038/srep16038.

39. Bhat GA, Shah IA, Rafiq R, Nabi S, Iqbal B, Lone MM, et al. Family history of cancer and the risk of squamous cell carcinoma of oesophagus: a case-control study in Kashmir, India. *Br J Cancer* (2015) 113(3):524-32. Epub 2015/07/01. doi: 10.1038/bjc.2015.218. PubMed PMID: 26125444; PubMed Central PMCID: PMCPMC4522628.

40. Uno K, Iijima K, Hatta W, Koike T, Abe Y, Asano N, et al. Direct measurement of gastroesophageal reflux episodes in patients with squamous cell carcinoma by 24-h pH-impedance monitoring. *Am J Gastroenterol* (2011) 106(11):1923-9. Epub 2011/09/21. doi: 10.1038/ajg.2011.282. PubMed PMID: 21931379.

41. Abnet CC, Kamangar F, Islami F, Nasrollahzadeh D, Brennan P, Aghcheli K, et al. Tooth loss and lack of regular oral hygiene are associated with higher risk of esophageal squamous cell carcinoma. *Cancer Epidemiology and Prevention Biomarkers* (2008) 17(11):3062-8.

42. Golozar A, Etemadi A, Kamangar F, Fazeltabar Malekshah A, Islami F, Nasrollahzadeh D, et al. Food preparation methods, drinking water source, and esophageal squamous cell carcinoma in the high-risk area of Golestan, Northeast Iran. *European Journal of Cancer Prevention* (2016) 25(2):123-9. doi: 10.1097/cej.0000000000000156. PubMed PMID: 00008469-201603000-00005.

43. Machoki MS, Saidi H, Raja A, Ndonga A, Njue A, Biomdo I, et al. Risk Factors for Esophageal Squamous Cell Carcinoma in a Kenyan Population. *Annals of African Surgery* (2015) 12(1).

44. AsombangAkwi W, Kayamba V, Lisulo MM, Trinkaus K, Mudenda V, Sinkala E, et al. Esophageal squamous cell cancer in a highly endemic region. *World journal of gastroenterology* (2016) 22(9):2811-7.

45. Matsha T, Brink L, van Rensburg S, Hon D, Lombard C, Erasmus R. Traditional home-brewed beer consumption and iron status in patients with esophageal cancer and healthy control subjects from Transkei, South Africa. *Nutrition and cancer* (2006) 56(1):67-73. Epub 2006/12/21. doi: 10.1207/s15327914nc5601_9. PubMed PMID: 17176219.

46. Kayamba V, Bateman AC, Asombang AW, Shibemba A, Zyambo K, Banda T, et al. HIV infection and domestic smoke exposure, but not human papillomavirus, are risk factors for esophageal squamous cell carcinoma in Zambia: a case-control study. *Cancer medicine* (2015) 4(4):588-95. Epub 2015/02/03. doi: 10.1002/cam4.434. PubMed PMID: 25641622; PubMed Central PMCID: PMCPMC4402073.

47. Parkin DM, Vizcaino AP, Skinner ME, Ndhlovu A. Cancer patterns and risk factors in the African population of southwestern Zimbabwe, 1963-1977. *Cancer epidemiology, biomarkers & prevention : a publication of the American Association*

for Cancer Research, cosponsored by the American Society of Preventive Oncology (1994) 3(7):537-47. Epub 1994/10/01. PubMed PMID: 7827583.

48. Vizcaino AP, Parkin DM, Skinner ME. Risk factors associated with oesophageal cancer in Bulawayo, Zimbabwe. *British journal of cancer* (1995) 72(3):769-73. Epub 1995/09/01. PubMed PMID: 7669592; PubMed Central PMCID: PMCPMC2033891.

49. Pacella-Norman R, Urban MI, Sitas F, Carrara H, Sur R, Hale M, et al. Risk factors for oesophageal, lung, oral and laryngeal cancers in black South Africans. *British journal of cancer* (2002) 86(11):1751-6. Epub 2002/06/28. doi: 10.1038/sj.bjc.6600338. PubMed PMID: 12087462; PubMed Central PMCID: PMCPMC2375408.

50. Sewram V, Sitas F, O'Connell D, Myers J. Tobacco and alcohol as risk factors for oesophageal cancer in a high incidence area in South Africa. *Cancer epidemiology* (2016) 41:113-21. Epub 2016/02/24. doi: 10.1016/j.canep.2016.02.001. PubMed PMID: 26900781.

51. Segal I, Reinach SG, de Beer M. Factors associated with oesophageal cancer in Soweto, South Africa. *British journal of cancer* (1988) 58(5):681-6. Epub 1988/11/01. PubMed PMID: 3219281; PubMed Central PMCID: PMCPMC2246810.

52. Patel K, Wakhisi J, Mining S, Mwangi A, Patel R. Esophageal Cancer, the Topmost Cancer at MTRH in the Rift Valley, Kenya, and Its Potential Risk Factors. *ISRN oncology* (2013) 2013:503249. Epub 2014/02/04. doi: 10.1155/2013/503249. PubMed PMID: 24490085; PubMed Central PMCID: PMCPMC3893746.

53. Vogelsang M, Wang Y, Veber N, Mwapagha LM, Parker MI. The cumulative effects of polymorphisms in the DNA mismatch repair genes and tobacco smoking in oesophageal cancer risk. *PloS one* (2012) 7(5):e36962. Epub 2012/05/25. doi: 10.1371/journal.pone.0036962. PubMed PMID: 22623965; PubMed Central PMCID: PMCPMC3356375.

54. Matejcic M, Vogelsang M, Wang Y, Iqbal Parker M, Parker IM. NAT1 and NAT2 genetic polymorphisms and environmental exposure as risk factors for oesophageal squamous cell carcinoma: a case-control study. *BMC cancer* (2015) 15:150-.

55. Dandara C, Ballo R, Parker MI. CYP3A5 genotypes and risk of oesophageal cancer in two South African populations. *Cancer letters* (2005) 225(2):275-82. Epub 2005/06/28. doi: 10.1016/j.canlet.2004.11.004. PubMed PMID: 15978331.

56. Sammon AM. A case-control study of diet and social factors in cancer of the esophagus in Transkei. *Cancer* (1992) 69(4):860-5. Epub 1992/02/15. PubMed PMID: 1735077.

57. Sammon AM. Protease inhibitors and carcinoma of the esophagus. *Cancer* (1998) 83(3):405-8. Epub 1998/08/05. PubMed PMID: 9690530.

58. Sewram V, Sitas F, O'Connell D, Myers J. Diet and esophageal cancer risk in the Eastern Cape Province of South Africa. *Nutrition and cancer* (2014) 66(5):791-9. Epub 2014/06/01. doi: 10.1080/01635581.2014.916321. PubMed PMID: 24877989.

59. Leon ME, Assefa M, Kassa E, Bane A, Gemechu T, Tilahun Y, et al. Qat use and esophageal cancer in Ethiopia: A pilot case-control study. *PloS one* (2017) 12(6):e0178911. Epub 2017/06/09. doi: 10.1371/journal.pone.0178911. PubMed PMID: 28594883; PubMed Central PMCID: PMCPMC5464578.

60. Shewaye AB, Sime A. Risk factors associated with esophageal cancer among ethiopian patients. *South African Gastroenterology Review* (2015) 13(2):53.

61. Van Rensburg SJ, Bradshaw ES, Bradshaw D, Rose EF. Oesophageal cancer in Zulu men, South Africa: a case-control study. *British journal of cancer* (1985) 51(3):399-405. Epub 1985/03/01. PubMed PMID: 3970816; PubMed Central PMCID: PMCPMC1976950.

62. Schaafsma T, Wakefield J, Hanisch R, Bray F, Schuz J, Joy EJ, et al. Africa's Oesophageal Cancer Corridor: Geographic Variations in Incidence Correlate with Certain Micronutrient Deficiencies. *PloS one* (2015) 10(10):e0140107. Epub 2015/10/09. doi: 10.1371/journal.pone.0140107. PubMed PMID: 26448405; PubMed Central PMCID: PMCPMC4598094.

63. Dandara C, Li DP, Walther G, Parker MI. Gene-environment interaction: the role of SULT1A1 and CYP3A5 polymorphisms as risk modifiers for squamous cell carcinoma of the oesophagus. *Carcinogenesis* (2006) 27(4):791-7. Epub 2005/11/08. doi: 10.1093/carcin/bgi257. PubMed PMID: 16272171.

64. Middleton DR, Menya D, Kigen N, Oduor M, Maina SK, Some F, et al. Hot beverages and oesophageal cancer risk in western Kenya: Findings from the ESCCAPE case-control study. *International journal of cancer* (2019) 144(11):2669-76. doi: https://dx.doi.org/10.1002/ijc.32032.

65. Granger S. A case series of 15 patients with oesophageal cancer in Blantyre, Malawi. Risk factors for oesophageal squamous cell carcinoma in Sub-Saharan Africa. *BMC Proceedings* (2012) 6(SUPPL. 4).

66. Mwachiro MM, Parker RK, Pritchett NR, Lando JO, Ranketi S, Murphy G, et al. Investigating tea temperature and content as risk factors for esophageal cancer in an endemic region of Western Kenya: Validation of a questionnaire and analysis of polycyclic aromatic hydrocarbon content. *Cancer epidemiology* (2019) 60:60-6. doi: http://dx.doi.org/10.1016/j.canep.2019.03.010. PubMed PMID: 2001731031.

67. Munishi MO, Hanisch R, Mapunda O, Ndyetabura T, Ndaro A, Schuz J, et al. Africa's oesophageal cancer corridor: Do hot beverages contribute? *Cancer causes & control : CCC* (2015) 26(10):1477-86. Epub 2015/08/08. doi: 10.1007/s10552-015-0646-9. PubMed PMID: 26245249; PubMed Central PMCID: PMCPMC4838015.

68. Menya D, Maina SK, Kibosia C, Kigen N, Oduor M, Some F, et al. Dental fluorosis and oral health in the African Esophageal Cancer Corridor: Findings from the Kenya ESCCAPE case-control study and a pan-African perspective. *International*

Journal of Cancer (2019). doi: http://dx.doi.org/10.1002/ijc.32086. PubMed PMID: 625915031.

69. Sitas F, Urban M, Stein L, Beral V, Ruff P, Hale M, et al. The relationship between anti-HPV-16 IgG seropositivity and cancer of the cervix, anogenital organs, oral cavity and pharynx, oesophagus and prostate in a black South African population. *Infectious agents and cancer* (2007) 2:6-.

70. Kayamba V, Bateman AC, Asombang AW, Shibemba A, Zyambo K, Banda T, et al. HIV infection and domestic smoke exposure, but not human papillomavirus, are risk factors for esophageal squamous cell carcinoma in Zambia: a case-control study. *Cancer medicine* (2015) 4(4):588-95. Epub 2015/01/30. doi: 10.1002/cam4.434. PubMed PMID: 25641622.

71. Asombang AW, Kayamba V, Lisulo MM, Trinkaus K, Mudenda V, Sinkala E, et al. Esophageal squamous cell cancer in a highly endemic region. *World journal of gastroenterology* (2016) 22(9):2811-7. Epub 2016/03/15. doi: 10.3748/wjg.v22.i9.2811. PubMed PMID: 26973419; PubMed Central PMCID: PMCPMC4778003.

72. Matsha T, Stepien A, Blanco-Blanco E, Brink LT, Lombard CJ, Van Rensburg S, et al. Self-induced vomiting -- risk for oesophageal cancer? *South African medical journal = Suid-Afrikaanse tydskrif vir geneeskunde* (2006) 96(3):209-12. Epub 2006/04/12. PubMed PMID: 16607430.

73. Kgomo M, Mokoena TR, Ker JA. Non-acid gastro-oesophageal reflux is associated with squamous cell carcinoma of the oesophagus. *BMJ open gastroenterology* (2017) 4(1):e000180. Epub 2017/11/28. doi: 10.1136/bmjgast-2017-000180. PubMed PMID: 29177066; PubMed Central PMCID: PMCPMC5687548.

74. Machoki M, Saidi H, Raja A, Ndonga A, Njue A, Biomdo I, et al. Risk factors for esophageal squamous cell carcinoma in a Kenyan population. *Annals of African Surgery* (2015) 12(1).

75. Dandara C, Li D-P, Walther G, Parker MI. Gene-environment interaction: the role of SULT1A1 and CYP3A5 polymorphisms as risk modifiers for squamous cell carcinoma of the oesophagus. *Carcinogenesis* (2006) 27(4):791-7.

76. Li D, Dandara C, Parker MI. Association of cytochrome P450 2E1 genetic polymorphisms with squamous cell carcinoma of the oesophagus. *Clinical chemistry and laboratory medicine* (2005) 43(4):370-5. Epub 2005/05/19. doi: 10.1515/cclm.2005.067. PubMed PMID: 15899651.

77. Matejcic M, Vogelsang M, Wang Y, Iqbal Parker M. NAT1 and NAT2 genetic polymorphisms and environmental exposure as risk factors for oesophageal squamous cell carcinoma: a case-control study. *BMC cancer* (2015) 15:150. Epub 2015/04/18. doi: 10.1186/s12885-015-1105-4. PubMed PMID: 25886288; PubMed Central PMCID: PMCPMC4379954.

78. Leon ME, Assefa M, Kassa E, Bane A, Gemechu T, Tilahun Y, et al. Qat use and esophageal cancer in Ethiopia: A pilot case-control study. *PloS one* (2017) 12(6):e0178911.

79. Menya D, Kigen N, Oduor M, Maina SK, Some F, Chumba D, et al. Traditional and commercial alcohols and esophageal cancer risk in Kenya. *International Journal of Cancer* (2019) 144(3):459-69. doi: http://dx.doi.org/10.1002/ijc.31804. PubMed PMID: 624876329.

80. Ocama P, Kagimu MM, Odida M, Wabinga H, Opio CK, Colebunders B, et al. Factors associated with carcinoma of the oesophagus at Mulago Hospital, Uganda. *African health sciences* (2008) 8(2):80-4. Epub 2009/04/10. PubMed PMID: 19357755; PubMed Central PMCID: PMCPMC2584326.

81. Okello S, Churchill C, Owori R, Nasasira B, Tumuhimbise C, Abonga CL, et al. Population attributable fraction of Esophageal squamous cell carcinoma due to smoking and alcohol in Uganda. *BMC cancer* (2016) 16:446. Epub 2016/07/13. doi: 10.1186/s12885-016-2492-x. PubMed PMID: 27400987; PubMed Central PMCID: PMCPMC4940693.

82. Matsha T, Brink L, van Rensburg S, Hon D, Lombard C, Erasmus R. Traditional Home-Brewed Beer Consumption and Iron Status in Patients With Esophageal Cancer and Healthy Control Subjects From Transkei, South Africa. *Nutrition and cancer* (2006) 56(1):67-73. doi: 10.1207/s15327914nc5601_9.

83. Nieminen MT, Salaspuro M. Local Acetaldehyde—An Essential Role in Alcohol-Related Upper Gastrointestinal Tract Carcinogenesis. *Cancers* (2018) 10(1):11.

84. Kgomo M, Mokoena T, Ker J. Non-acid gastro-oesophageal reflux is associated with squamous cell carcinoma of the oesophagus. *BMJ Open Gastroenterology* (2017) 4.

85. Asombang AW, Chishinga N, Nkhoma A, Chipaila J, Nsokolo B, Manda-Mapalo M, et al. Systematic review and meta-analysis of esophageal cancer in Africa: Epidemiology, risk factors, management and outcomes. *World J Gastroenterol* (2019) 25(31):4512-33. Epub 2019/09/10. doi: 10.3748/wjg.v25.i31.4512. PubMed PMID: 31496629; PubMed Central PMCID: PMCPMC6710188.

86. Isaacson C, Bothwell TH, MacPhail AP, Simon M. The iron status of urban black subjects with carcinoma of the oesophagus. *South African medical journal = Suid-Afrikaanse tydskrif vir geneeskunde* (1985) 67(15):591-3. PubMed PMID: 3983740.

87. Prabhu A, Obi KO, Rubenstein JH. The synergistic effects of alcohol and tobacco consumption on the risk of esophageal squamous cell carcinoma: a meta-analysis. *American Journal of Gastroenterology* (2014) 109(6):822-7.

88. Jansson C, Johansson AL, Nyrén O, Lagergren J. Socioeconomic factors and risk of esophageal adenocarcinoma: a nationwide Swedish case-control study. *Cancer Epidemiology and Prevention Biomarkers* (2005) 14(7):1754-61.

89. Gammon MD, Ahsan H, Schoenberg JB, West AB, Rotterdam H, Niwa S, et al. Tobacco, alcohol, and socioeconomic status and adenocarcinomas of the esophagus and gastric cardia. *Journal of the National Cancer Institute* (1997) 89(17):1277-84.

90. Dar NA, Shah IA, Bhat GA, Makhdoomi MA, Iqbal B, Rafiq R, et al. Socioeconomic status and esophageal squamous cell carcinoma risk in Kashmir, India. *Cancer science* (2013) 104(9):1231-6.

91. Schaafsma T, Wakefield J, Hanisch R, Bray F, Schüz J, Joy EJM, et al. Africa’s Oesophageal Cancer Corridor: Geographic Variations in Incidence Correlate with Certain Micronutrient Deficiencies. *PLoS ONE* (2015) 10(10):e0140107. doi: 10.1371/journal.pone.0140107. PubMed PMID: PMC4598094.

92. Ma J, Li Q, Fang X, Chen L, Qiang Y, Wang J, et al. Increased total iron and zinc intake and lower heme iron intake reduce the risk of esophageal cancer: A dose-response meta-analysis. *Nutrition Research* (2018) 59:16-28. doi: https://doi.org/10.1016/j.nutres.2018.07.007.

93. Higgins JP, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, et al. *Cochrane handbook for systematic reviews of interventions*: John Wiley & Sons (2019).

94. Wang J-B, Fan J-H, Liang H, Li J, Xiao H-J, Wei W-Q, et al. Attributable causes of esophageal cancer incidence and mortality in China. *PloS one* (2012) 7(8):e42281-e. Epub 2012/08/02. doi: 10.1371/journal.pone.0042281. PubMed PMID: 22876312.

95. Wu M, Van't Veer P, Zhang ZF, Wang XS, Gu XP, Han RQ, et al. A large proportion of esophageal cancer cases and the incidence difference between regions are attributable to lifestyle risk factors in China. *Cancer Lett* (2011) 308(2):189-96. Epub 2011/06/15. doi: 10.1016/j.canlet.2011.05.003. PubMed PMID: 21665362.

96. Charafeddine MA, Olson SH, Mukherji D, Temraz SN, Abou-Alfa GK, Shamseddine AI. Proportion of cancer in a Middle eastern country attributable to established risk factors. *BMC cancer* (2017) 17(1):337-. doi: 10.1186/s12885-017-3304-7. PubMed PMID: 28521815.

97. Oze I, Charvat H, Matsuo K, Ito H, Tamakoshi A, Nagata C, et al. Revisit of an unanswered question by pooled analysis of eight cohort studies in Japan: Does cigarette smoking and alcohol drinking have interaction for the risk of esophageal cancer? *Cancer Med* (2019) 8(14):6414-25. Epub 2019/09/03. doi: 10.1002/cam4.2514. PubMed PMID: 31475462; PubMed Central PMCID: PMCPMC6797581.

98. Pandeya N, Olsen CM, Whiteman DC. Sex differences in the proportion of esophageal squamous cell carcinoma cases attributable to tobacco smoking and alcohol consumption. *Cancer epidemiology* (2013) 37(5):579-84. Epub 2013/07/09. doi: 10.1016/j.canep.2013.05.011. PubMed PMID: 23830137.

99. Akhtar S, Sheikh AA, Qureshi HU. Chewing areca nut, betel quid, oral snuff, cigarette smoking and the risk of oesophageal squamous-cell carcinoma in South Asians: a multicentre case-control study. *Eur J Cancer* (2012) 48(5):655-61. Epub 2011/07/08. doi: 10.1016/j.ejca.2011.06.008. PubMed PMID: 21733677.

100. Sewram V, De Stefani E, Brennan P, Boffetta P. Mate consumption and the risk of squamous cell esophageal cancer in uruguay. *Cancer Epidemiol Biomarkers Prev* (2003) 12(6):508-13. Epub 2003/06/20. PubMed PMID: 12814995.

101. Grundy A, Poirier AE, Khandwala F, McFadden A, Friedenreich CM, Brenner DR. Cancer incidence attributable to insufficient fruit and vegetable consumption in Alberta in 2012. *CMAJ Open* (2016) 4(4):E760-e7. Epub 2016/12/27. doi: 10.9778/cmajo.20160037. PubMed PMID: 28018892; PubMed Central PMCID: PMCPMC5173484.

102. GBD 2017 Oesophageal Cancer Collaborators. The global, regional, and national burden of oesophageal cancer and its attributable risk factors in 195 countries and territories, 1990-2017: a systematic analysis for the Global Burden of Disease Study 2017. *Lancet Gastroenterol Hepatol* (2020) 5(6):582-97. Epub 2020/04/05. doi: 10.1016/s2468-1253(20)30007-8. PubMed PMID: 32246941; PubMed Central PMCID: PMCPMC7232026.

103. Chimusa ER, Daya M, Moller M, Ramesar R, Henn BM, van Helden PD, et al. Determining ancestry proportions in complex admixture scenarios in South Africa using a novel proxy ancestry selection method. *PloS one* (2013) 8(9):e73971. Epub 2013/09/26. doi: 10.1371/journal.pone.0073971. PubMed PMID: 24066090; PubMed Central PMCID: PMCPMC3774743.

104. de Wit E, Delport W, Rugamika CE, Meintjes A, Moller M, van Helden PD, et al. Genome-wide analysis of the structure of the South African Coloured Population in the Western Cape. *Human genetics* (2010) 128(2):145-53. Epub 2010/05/22. doi: 10.1007/s00439-010-0836-1. PubMed PMID: 20490549.

105. Matejcic M, Iqbal Parker M. Gene-environment interactions in esophageal cancer. *Crit Rev Clin Lab Sci* (2015) 52(5):211-31. Epub 2015/07/30. doi: 10.3109/10408363.2015.1020358. PubMed PMID: 26220475.

106. Guo LY, Zhang S, Suo Z, Yang CS, Zhao X, Zhang GA, et al. PLCE1 gene in esophageal cancer and interaction with environmental factors. *Asian Pac J Cancer Prev* (2015) 16(7):2745-9. Epub 2015/04/10. doi: 10.7314/apjcp.2015.16.7.2745. PubMed PMID: 25854357.

107. Matejcic M, Mathew CG, Parker MI. The Relationship Between Environmental Exposure and Genetic Architecture of the 2q33 Locus With Esophageal Cancer in South Africa. *Frontiers in Genetics* (2019) 10(406). doi: 10.3389/fgene.2019.00406.

108. Zhang L, Jiang Y, Wu Q, Li Q, Chen D, Xu L, et al. Gene-environment interactions on the risk of esophageal cancer among Asian populations with the G48A polymorphism in the alcohol dehydrogenase-2 gene: a meta-analysis. *Tumour Biol* (2014) 35(5):4705-17. Epub 2014/01/22. doi: 10.1007/s13277-014-1616-7. PubMed PMID: 24446180.

109. Matsuo K, Hamajima N, Shinoda M, Hatooka S, Inoue M, Takezaki T, et al. Gene-environment interaction between an aldehyde dehydrogenase-2 (ALDH2) polymorphism and alcohol consumption for the risk of esophageal cancer. *Carcinogenesis* (2001) 22(6):913-6. Epub 2001/05/29. doi: 10.1093/carcin/22.6.913. PubMed PMID: 11375898.

110. Marmot M, Atinmo T, Byers T, Chen J, Hirohata T, Jackson A, et al. Food, nutrition, physical activity, and the prevention of cancer: a global perspective. (2007).

111. Delgado-Rodríguez M, Llorca J. Bias. *Journal of Epidemiology and Community Health* (2004) 58(8):635. doi: 10.1136/jech.2003.008466.

Chapter 4: Identification of Genes and Pathways with Differential mRNA Expression in Esophageal Cancer

Table of Contents

4.1 Abstract

Esophageal cancer (EC) is one of the most aggressive malignancies and a leading cause of cancer death globally. It is characterised by poor prognosis as patients present at an advanced stage, and consequently EC has a low survival rate. The pathobiology of EC is not well understood. The aim of this study was to identify biological pathways involved in EC development with genes demonstrating differential mRNA expression in EC. **Method:** We performed a comprehensive search on the GEO Database (NCBI) to identify datasets on esophageal squamous cell carcinoma (ESCC), squamous dysplasia, esophageal adenocarcinoma (EAC) and Barrett's esophagus (BE). Identification and meta-analysis of differentially expressed genes (DEGs) was done using the Rank Product Method. An advanced gene set enrichment analysis, SetRank, was used to identify enriched biological pathways using the Reactome Annotation Database. Pathways were visualised using Cytoscape. **Results:** A total of 18 publicly available GEO mRNA expression datasets, with expression data on 906 individual tissue samples, were included in the analysis. Overall, 1,107 upregulated genes and 1,537 downregulated genes were outputted for BE, EAC and ESCC. The majority of DEGs were significantly associated with the pathways involved in the extracellular matrix, including *"Extracellular matrix organisation", "Assembly of collagen fibrils and other multimeric structures",* and *"Collagen chain trimerization*". Pathways involved in cell cycle regulation were also identified, including *"TP53 regulates transcription of several additional cell death genes whose specific roles in p53-dependent apoptosis remain uncertain"*, and *"Cyclin B2 mediated events"*. **Conclusions:** We identified key pathways not previously discussed or interpreted in literature in relation to EC, which warrant further investigation. The combined bioinformatic analysis of existing GEO mRNA expression datasets on EC provided novel insights into the pathobiology of EC.

4.2 Background

Esophageal cancer (EC) is a complex disease and the 6th most common cause of cancer deaths and the 7th most common cancer worldwide.(1) It is characterised by two main histological subtypes, esophageal squamous cell carcinoma (ESCC) which constitutes about 90% of all EC cases, and esophageal adenocarcinoma (EAC).(1)

ESCC develops from precursor lesions in the mucosa of the esophagus, consisting of squamous dysplasia known to develop into malignant ESCC tumours.(2) EAC develops from intestinal metaplasia of the esophageal epithelium (Barrett's esophagus, BE) that develops in response to chronic gastroesophageal reflux.(2, 3) Whilst the origin of squamous dysplasia is definitive, the origin of BE is still contested in literature, with several hypotheses currently being brought forward (Figure 4.1).(4)

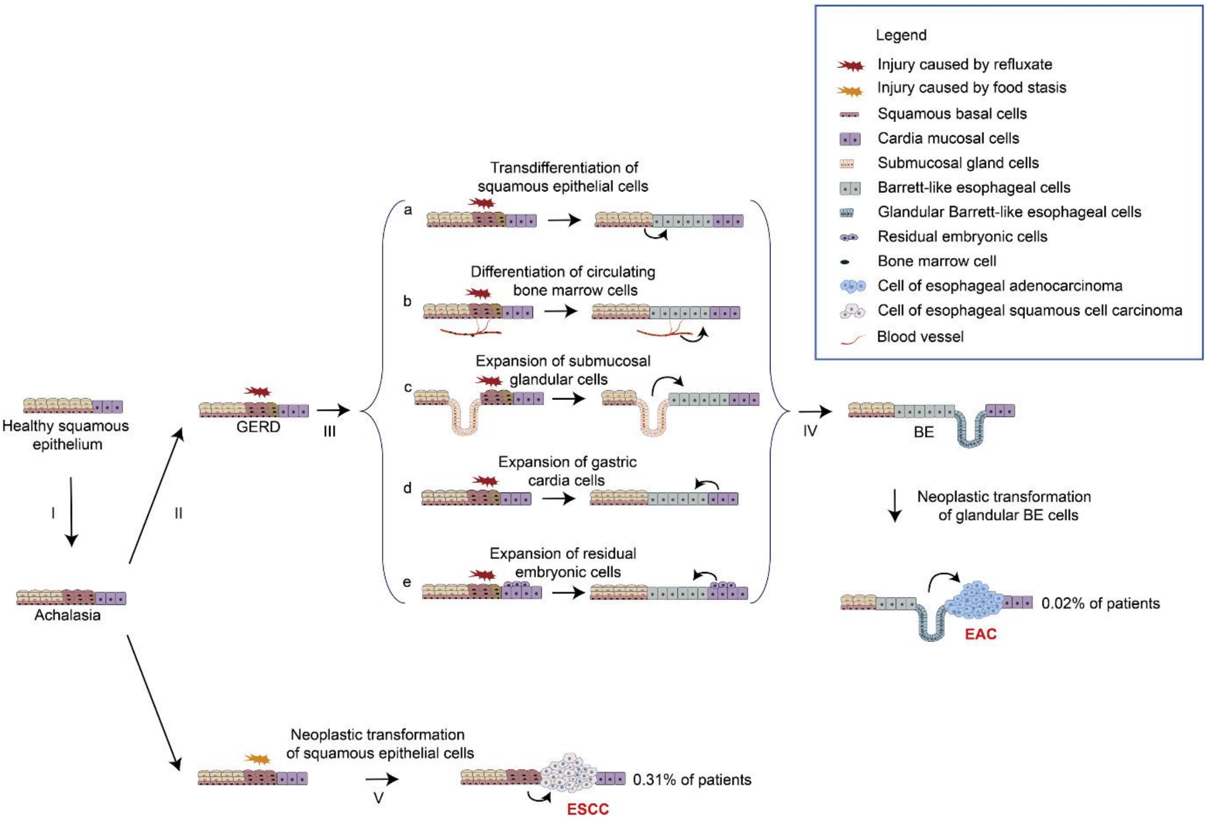

Legend
Injury caused by refluxate
Injury caused by food stasis
Squamous basal cells
Cardia mucosal cells
Submucosal gland cells
Barrett-like esophageal cells
Glandular Barrett-like esophageal cells
Residual embryonic cells
Bone marrow cell
Cell of esophageal adenocarcinoma
Cell of esophageal squamous cell carcinoma
Blood vessel
Transdifferentiation of squamous epithelial cells
a
Differentiation of circulating bone marrow cells
b
Expansion of submucosal glandular cells
c
Expansion of gastric cardia cells
d
Expansion of residual embryonic cells
e
Healthy squamous epithelium
GERD
III
IV
BE
I
II
Achalasia
Neoplastic transformation of glandular BE cells
0.02% of patients
EAC
Neoplastic transformation of squamous epithelial cells
V
0.31% of patients
ESCC

Figure 4.1: *Proposed hypothesis on the cell origins of Barrett's Esophagus and esophageal squamous dysplasia. Figure by Nesteruk et al (4). Creative Commons licence: Rightslink® by Copyright Clearance Center. BE, Barrett's esophagus; ESCC, esophageal squamous cell carcinoma; EAC, esophageal adenocarcinoma; GERD, gastroesophageal reflux disease.*

EC has a multifactorial etiology, with environmental, lifestyle and genetic risk factors reported to be associated with EC development and progression.(3) Whilst multiple genomic alterations have been reported to be associated with EC development and prognosis (5-8), the genetic basis and pathobiology of EC are still poorly understood. Investigating differential mRNA expression allows for elucidation of oncogenic pathways which drive EC development, progression, and survival. Research has shown that biological differences reflected in gene expression profiles of tumours may be associated with cancer prognosis, and could be used to select more targeted therapies.(9) High throughput technologies such as microarrays and RNA-sequencing are some of the most effective approaches used to identify key molecular events involved in tumorigenesis and have been used in the identification of novel pathways associated with cancer development.(9, 10)

Microarray analysis has been useful in revealing genetic networks and pathways associated with EC, through comprehensive screening of differentially expressed genes (DEGs). The development, progression, metastasis, response to therapy, and survival of EC is linked to changes in patterns of gene expression.(11) Microarray analysis involves simultaneous expression analysis of thousands of genes from tumour samples and is, therefore, a powerful platform for assessing complex changes in gene expression.(11) Studies have shown that one of the most commonly identified genomic alterations in ESCC is *NRF2* hyperactivation, which has been associated with a poor prognosis. (12) Several receptor tyrosine kinases (RTKs) have also been reported to have differential expression in OSCC and these include; *EFGR, IGF1R, MET, FLT1*, and *PTK7*.(13) The Cancer Genome Atlas (TGCA) project reported that upregulation of Wnt, syndecan, and p63 pathways was characteristic of ESCC.(14) Increased E-cadherin expression and upregulation of pathways involved in the regulation of E-cadherin were reported to be associated with EAC (14), whilst p53 overexpression has been linked with BE (15). In a systematic review assessing prognostic gene expression profiles using microarray data, the authors reported eight genes; *ALDH1A3, BIN1, CSPG2, DOK1, IFIT1, IFIT3, PHB,* and *SPP1*, which were differentially expressed in the survival of EAC.(16) Three genes; *ATR, MAL, PCP4* were associated with lymph node metastasis in ESCC.(16) These differences in gene expression profiles point to different genomic alterations or transcriptional regulations driving EAC, ESCC, and BE development and progression.

The development and application of gene chips over the past decades have resulted in more data being generated and stored in public databases. Cancer genomics researchers have over the past years generated a large amount of genomic data, used to address specific research questions or only in the analysis of a subset of the data.(17) A growing trend is the depositing of this data in public repositories for further use.(18) This has created opportunities for researchers to repurpose such data for answering new research questions, analysing part of the data not previously analysed, and/or performing integrated analyses of multiple datasets for more comprehensive analyses. Researchers who would have otherwise not been able to generate this type of data can also have access to it and contribute to scientific progress and new insights. Public repositories of genomic data have consequently become valuable sources of data and springboards for new research. Repositories also allow for transparency and validation of research. Several repositories exist, including the National Cancer Institute (NCI) Genomic Data Commons (GDC) which contains genomic and associated clinical data from over 60 NCI-funded and other research projects.(19) Gene expression omnibus (GEO) database is a data repository of curated gene expression datasets, containing one of the most comprehensive repositories of microarray data.(20)

The high fatality and poor prognosis of EC calls for more research focussed on elucidating the pathobiology of EC, which is not well understood. Over the past several of years, bioinformatic tools have been used to effectively explore and identify genes, proteins and pathways involved in cancer development and prognosis. The aim of this study is to identify biological pathways involved in EC development with genes demonstrating differential mRNA expression in EC. Very few studies have investigated the transcriptome in EC. Although several studies have described individual gene signatures associated with EC development, few have investigated the key modules, hub genes and functional networks involved.(9, 21) In this study, gene expression profiling of the two main histological EC subtypes, ESCC, EAC and the pre-cancerous lesions, BE using data from microarray analysis was used to analyse genes and pathways associated with EC. We included datasets that assessed ESCC, EAC, squamous dysplasia, and BE. This was done to provide a comprehensive analysis of the pre-cancer (BE and squamous dysplasia) and cancerous tissue (ESCC and EAC) and to also analyse the progression of the precursor lesions to malignant tumours for both EAC and ESCC.

This was achieved through the following objectives: i) data compilation of raw genome-wide mRNA expression data on EC from the GEO database repository, ii) meta-analysis of DEGs from the combined datasets using the Rank Product method, iii) gene set enrichment analysis using SetRank analysis, and iv) functional annotation of DEGs and gene sets using the Reactome annotation database. The study provided novel insights into the mechanisms linked to EC development. It is important to mention that only a minority of the precancerous lesions in EC develop to cancer (4), and that BE and squamous dysplasia, are not tumours.

4.3 Methods

4.3.1 Raw data acquisition

We performed a comprehensive search on the GEO Database (NCBI) to identify genome-wide mRNA expression datasets on EC. Table 4.1 shows the search strategy used. Additional manual screening was performed to remove datasets using cell lines, not tissue samples, and studies using printed microarray designs. Raw data were downloaded from the GEO database onto the Stellenbosch University aither.mb.sun.ac.za server, from which all the analyses were performed.

Table 4.1: GEO Datasets Search Strategy

Search type	Search criteria
Base search	((esophageal cancer) AND "rna"[Sample Type]) AND "gse"[Entry Type] AND homo sapiens[ORGN]
Extended search 1	((((esophageal cancer) AND "rna"[Sample Type]) AND "gse"[Entry Type] AND homo sapiens[ORGN])) AND Squamous
Extended search 2	((esophageal cancer) AND "rna"[Sample Type]) AND "gse"[Entry Type] AND homo sapiens[ORGN] AND adenocarcinoma
Extended search 3	((((esophageal cancer) AND "rna"[Sample Type]) AND "gse"[Entry Type] AND homo sapiens[ORGN])) AND barrett's

4.3.2 Summary of datasets

A total of 32 potential publicly available mRNA expression datasets were identified on GEO datasets (NCBI). After screening, 21 datasets fitting the selection criteria were identified. The 21 datasets were from three platforms: Affymetrix, Agilent and Illumina. They included Affymetrix Human Gene 1.0 ST Arrays, Affymetrix Human Exon 1.0 ST Array, Affymetrix Human Genome U133 Plus 2.0 Array, Affymetrix Human Genome U133A Array, Affymetrix Human Genome U133B Array, Agilent-026652 Whole Human Genome Microarray 4x44K v2 and Illumina Human-6 v2.0 Expression BeadChip.(22-24) A summary of the platforms is shown in Table 4.2. Of the 21 datasets included in the study, 18 were analysed using the Affymetrix platforms, whilst two datasets used the Illumina platforms, and one dataset used the Agilent platform. Differential expression analysis was done only on the 18 datasets that used the Affymetrix platforms.

Table 4.2: Summary of platforms in included datasets

Platform	Probes	Probe sets	Genes	Probes per gene	Exons	Transcripts and variants	Oligonucleotide probe size
Affymetrix Human Gene 1.0 ST Array	764,885	33,252	28,869	26		36,079	25-mer
Affymetrix Human Exon 1.0 ST Array	1,073,146	1,400,000		40	325,353	35,685	25-mer
Affymetrix Human Genome U133 Plus 2.0 Array	604,258	54,675	38,500	16		47,000	25-mer
Affymetrix Human Genome U133A Array		22,000	14,500	11		18,400	25-mer
Affymetrix Human Genome U133B Array		22,000	18,500	11		20,600	25-mer

Agilent-026652 Whole Human Genome Microarray 4x44K v2		34,184	
Illumina human-6 v2.0 expression beadchip	48,702	14,507	50 bp

The Human Gene 1.0 ST Array utilizes a subset of probes selected from the Exon 1.0 ST Array and these two arrays were therefore considered as one platform. Affymetrix Human Genome U133 Plus 2.0 Array's housekeeping/control genes are *GAPDH*, *ACTB*, *ISGF-3* (*STAT1*). The Human Gene 1.0 ST Array, Human Genome U133 Plus 2.0 and the Human Exon 1.0 ST Arrays were assessed and reported to have comparable gene-level performance and strong concordance.(24, 25) The analysis also showed a marginally improved reproducibility for the Human Gene 1.0 ST array with comparable detection thresholds.(24) The three Affymetrix platforms had a 65% overlap in the top 2,000 DEGs in an analysis done by Robinson *et al* (26) in brain and heart samples.

4.3.3 Summary of published studies

A total of 81 studies were published based on the 21 datasets. Wang *et al* (27) analysed the most datasets, nine, followed by Zhang et al (28) and He *et al* (29) which analysed seven and six datasets, respectively. The rest of the studies analysed five or fewer datasets. The 81 studies were published between 2010 and 2020. The previously published studies highlighted several genes and pathways in these datasets: *KRAS, SPARC, SPP1, FOXM1, WDR66, PTGS2*, V-ATPase genes, tumour suppressor genes, and PI3K signalling pathway. These are shown in Table 4.3. A summary of the bioinformatics and biostatistics tools used in some of these studies is shown in Table 4.4. The table only includes studies that assessed four or more GEO database datasets. Pathway and network analyses were performed using Ingenuity Pathway Analysis, Gene Ontology and KEGG analysis. Functional analysis was carried out using Gene Set Enrichment Analysis (GSEA). The false discovery rate (FDR) method was used to adjust p-values for multiple testing.

4.3.4 Studies included for analysis

Altogether 18 datasets that used the Affymetrix platform were included in the meta-analysis of DEGs. Of the 18 datasets, three used EAC tissue, 11 used ESCC tissue and nine used BE tissue. One dataset included EAC, ESCC and BE tissue, whilst one dataset included both EAC and squamous dysplasia tissue samples. GSE33426 and GSE29001 had 18 similar samples, these duplicate samples were removed from GSE33426 before analysis. In total there were 906 genome-wide mRNA expression datasets on 280 ESCC, 45 EAC, 140 BE, 2 squamous dysplasia and 439 normal tissue samples. The distribution of the sample types according to a dataset is shown in Figure 4.2. It is important to note that one dataset, GSE23400, used two Affymetrix platforms: Affymetrix Human Genome U133 Array and Affymetrix Human Genome U133B Array, therefore the dataset was analysed as two datasets, GSE2300A and GSE23400B.

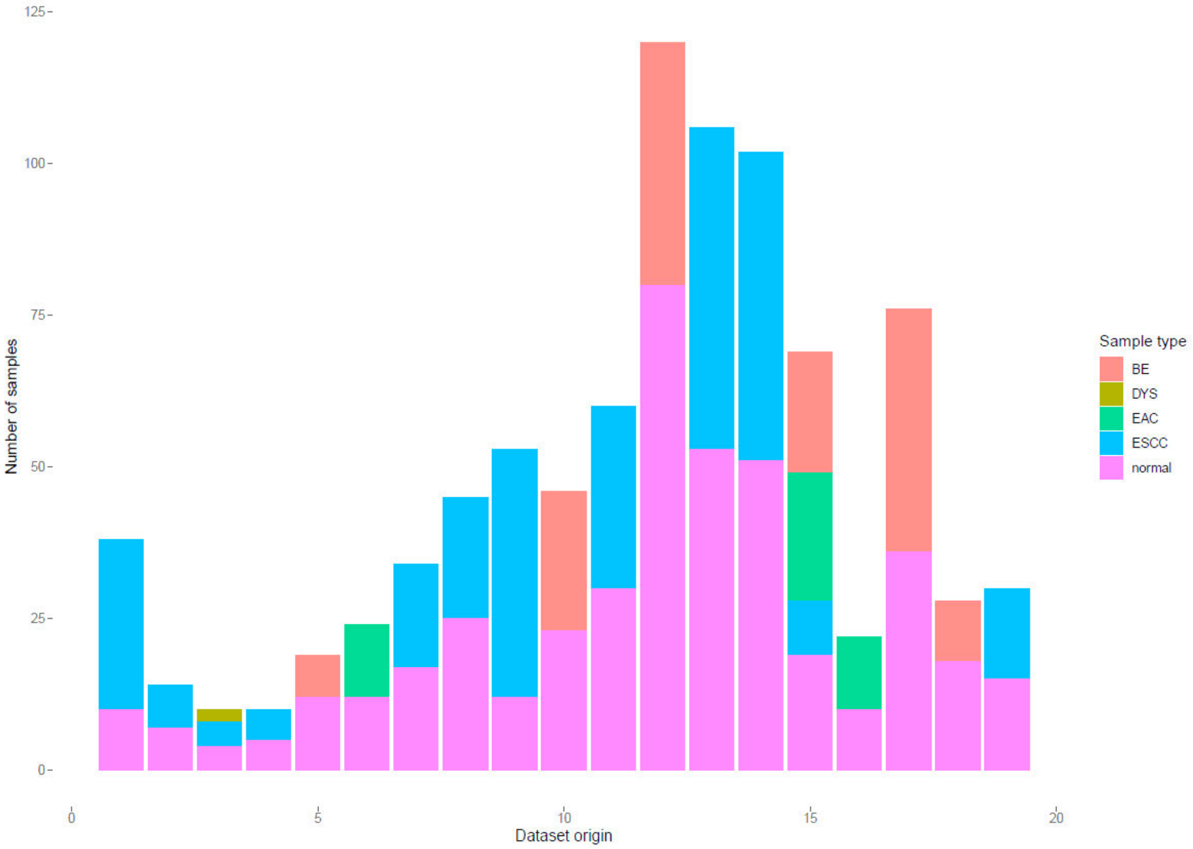


Figure 4.2: *Summary of the distribution of sample types according to each of the 18 Affymetrix datasets (The GSE23400 dataset had 2 Affymetrix platforms and was analysed as 2 datasets). BE; Barrett's esophagus, DYS; squamous dysplasia, EAC;*

esophageal adenocarcinoma, ESCC; esophageal squamous cell carcinoma, normal; non-tumour tissue samples.

Six of the 18 datasets had samples originating from China, and another six datasets had samples originating from the USA. One dataset had samples from both the USA and Japan. The remaining five datasets had samples originating from Germany, Hong Kong, Poland and the UK each. The novelty of the current study is that we analysed all available data together and were able to compare the different EC types and stages in this analysis. An overview of the microarray data analysis steps employed in the current study is shown in Figure 4.3. A summary of the characteristics of the datasets selected for analysis is shown in table 4.5.

Table 4.3: Genes and pathways highlighted in previous publications

GEO ID	GEO Series Accession	Highlighted genes/pathways	References (PMID)
200092396	GSE92396	*KRAS* associated gene signature	Peng et al 2017 (28102292)(30)
200013898	GSE13898	*SPARC* and *SPP1*	Kim et al 2010 (21152079)(31)
200001420	GSE1420	Aldo-keto reductases, aldehyde dehydrogenases , dual-specificity phosphatases, annexins, chloride channels, keratins and genes involved in the formation of desmosomes, and the cornified envelope of squamous epithelium	Kimchi et al 2005 (15833844)(10)
200028302	GSE28302	DPP4, ATP2A3, AGR2, collagens, IGFBP7, PLAU, MUC6, CA2, TFF1, AKR1C2, AKR1B10	Nancarrow et al 2011 (21829465)(32)
200013083	GSE13083	Cdx1 homeodomain transcription factor and the c-myc pathway	Stairs et al 2008 (18953412)(33)
200034619	GSE34619	HOXB genes	di Pietro et al 2012 (22603795)(34)
200039491	GSE39491	ABP1, ATP2C2, CALML4, HOXB7 KRT7, MSLN, and TFF3	Hyland et al 2014 (24714516)(35)
200036223	GSE36223	Defence and repair responses of metaplastic mucosa	Ostrowski et al 2007 (17415542)(36)
200100843	GSE100843	PTGS2 pathway of PGE2	Cummings et al 2017 (28922414)(37)
200026886	GSE26886	WDR66	Wang et al 2013 (23514407)(38) Dai et al 2018 (29223109)(39)
200020347	GSE20347	Loss of heterozygosity (LOH) and copy number (CN) change	Hu et al 2010 (20955586)

200075241	GSE75241	*FOXM1,* PI3K Signalling Pathway and V-ATPase genes	Nicolau-Neto et al 2018 (29682174),(40) Couto-Vieira et al 2020 (31901859)(41)
200100942	GSE100942	RHCG (Tumour suppressor genes)	Ming at al 2018 (29290801)(42)
200017351	GSE17351	Cyclooxygenase-2-prostaglandin E synthase axis	Lee et al 2010 (20042640)(43)
200023400	GSE23400	PLCE1	Su et al 2011 (21385931)(44) Li et al 2014 (24867265)(45) Hyland et al 2016 (26635288)(46)
200033426	GSE33426	ODC1, POSTN, ASPA and IGF2BP3	Yan et al 2012 (22280838)(47) Yan et al 2013 (23219752)(48)
200029001	GSE29001	ODC1, POSTN, ASPA	Yan et al 2012 (22280838)(47)
200045168	GSE45168	NEK6	
200045670	GSE45670	LIMCH1, SDPR, C1orf226, SLC9A9, GSTM3, IGSF10, MMP1, MMP9, MMP12 and OASL	Wen et al 2014 (24907633)(49)
200077861	GSE77861	KRT17, PRDCSH, TNFRSF6B, SELK, RAB5B, ALD, RAF	Erkizan et al 2019 (28629367)(50)
200038129	GSE38129	FOXP1, CSMD1, CDKN2A/2B, FHIT, DLEC1, and RARB	

Table 4.4: Published studies

Study (PMID)	Microarray data normalisation	Statistics[1]	Bioinformatics	Data sets (n)
Wang et al 2019 (31686859)(27)	Affy package in R	LIMMA and independent sample t test. Pearson's correlation matrix.	WGCNA, GO and KEGG	9
Zhang et al 2018 (29600044)(28)	Quantile normalisation	LIMMA and moderate sample t test. Pheatmap package in R	Differentially expressed genes (DEGs) analysis, GO, KEGG and Cytoscape	7
He et al 2018 (30505479)(29)	Robust Multichip Averaging	Student's t-test	Gene set enrichment analysis, KEGG, Blast2go	6
Wu et al 2013 (24039884)(51)		k-clique method in Subpathway Miner R packages	Fold-change analysis[3]	5
Liu et al 2015 (26489668)(52)	Least variant set (LVS) method	SAM package in R	Search Tool for the Retrieval of Interacting Genes/Proteins (STRING) and MCODE. GO and Cytoscape	5
Lv et al 2018 (29417867)(53)		LIMMA using R. SPSS, Graphpad Prism, Med-calc software	DAVID Bioinformatics Resources, GO, KEGG and Cytoscape	5
Lian at al 2019 (31164411)(54)		Multiple linear regression analysis using Matlab	'Shortest Path' module of Pathway Studio	5
McKenzie et al 2016 (27895316)(55)			Oncomine™ analysis	4
Li et al 2017 (28937628)(56)		SAM package in R and Kaplan-Meier analysis	Subpathway-GM and Cytoscape	4

Chen et al 2019 (31114367)(57)		LIMMA and Kaplan Meier analysis in R. Graphpad Prism and students t test.	TF-miRNA-mRNA network, GO, KEGG and PPI	4
Wang et al 2019 (31383552)(58)		LIMMA and Pheatmap in R. SPSS	Venn online analysis tool	4
Li et al 2020 (32208405)(59)	RMA algorithm	Volcano plot in R (ggplot2). T-test and Graphpad prism	GEO2R and Jvenn. GSEA, GEPIA, GO and KEGG	4
Xia et l 2020 (32494154)(60)		Student's t-test and Graphpad Prism. Kaplan-Meier analysis	GEPIA	4

[1] Statistical analysis and tools used to analyse/compare gene expression levels and in prognostic analysis

[2] Bioinformatics tools used for pathway, network and functional analysis

[3] Intersections of predicted targets from miRecords and ESCC DEGs from ESCC mRNA profile were computed for subsequent subpathway analysis for each differentially expressed miRNA, respectively.

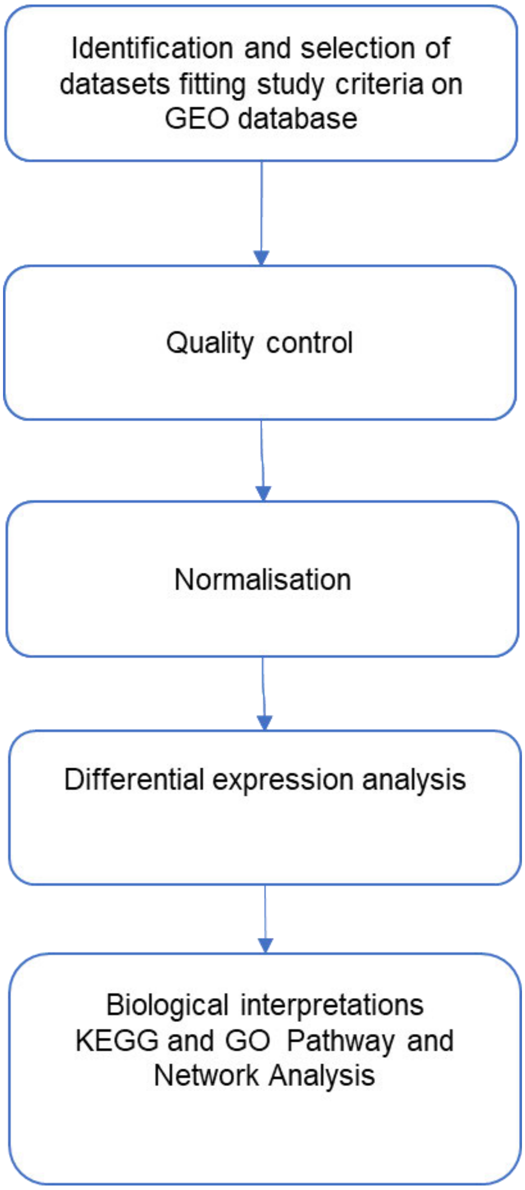

Identification and selection of datasets fitting study criteria on GEO database
Quality control
Normalisation
Differential expression analysis
Biological interpretations KEGG and GO Pathway and Network Analysis

Figure 4.3: *An overview of the microarray data analysis steps used for the study*

Table 4.5: Summary of GEO datasets included in the current study

							Sample description						
GEO ID	**GEO Series Accession**	**EC Type**	**Platform**	**Design Comment**	**Platform ID**	**Sample (n)**	**EAC**	**ESCC**	**BE**	**Normal**	**Design**	**Validation**	**Study (PMID)**
200092396	GSE92396	EAC	Affymetrix Human Gene 1.0 ST arrays	HE	GPL6244	21	12			9	Cross-sectional	None	Peng et al 2017 (28102292)(30)
200013898	GSE13898	EAC & BE	Illumina human-6 v2.0 expression beadchip	Frozen	GPL6102	118	75		15	28	Cross-sectional	qRT-PCR	Kim et al 2010 (21152079)(31)
200001420	GSE1420	EAC	Affymetrix Human Genome U133A Array	Frozen	GPL96	24	12			12		qRT-PCR	Kimchi et al 2005 (15833844)(10)
200028302	GSE28302	EAC & BE	Sentrix Human-6 Expression BeadChip	Frozen	GPL2507	54	23		22	9			Nancarrow et al 2011 (21829465)(32)
200013083	GSE13083	BE	Affymetrix Human Genome U133A Array	FF	GPL96	19			7	12	Paired	IHC	Stairs et al 2008 (18953412)(33)
200034619	GSE34619	BE	Affymetrix Human Gene 1.0 ST Array	Frozen	GPL6244	28			10	18	Paired	qRT-PCR	di Pietro et al 2012 (22603795)(34)

200039491	GSE39491	BE	Affymetrix Human Genome U133A 2.0 Array	Frozen	GPL571	120			40	80	Paired	qRT-PCR	Hyland et al 2014 (24714516)(35)
200036223	GSE36223	BE	Affymetrix Human Genome U133A 2.0 Array	Frozen	GPL571	46			23	23	Paired	qRT-PCR	Ostrowski et al 2007 (17415542)(36)
200100843	GSE100843	BE	Affymetrix Human Gene 1.0 ST Array	FFPE and HE	GPL6244	76			40	36	Longitudinal	None	Cummings et al 2017 (28922414)(37)
200026886	GSE26886	EAC, ESCC & BE	Affymetrix Human Genome U133 Plus 2.0 Array	Frozen	GPL570	69	21	9	20	19	Cross-sectional	qRT-PCR	Wang et al 2013 (23514407)(38) Dai et al 2018 (29223109)(39)
200020347	GSE20347	ESCC	Affymetrix Human Genome U133A 2.0 Array	Frozen	GPL571	34		17		17	Paired		Hu et al 2010 (20955586)
200075241	GSE75241	ESCC	Affymetrix Human Exon 1.0 ST Array	Frozen	GPL5175	30		15		15	Paired	qRT-PCR	Nicolau-Neto et al 2018 (29682174),(40) Couto-Vieira et al

											2020 (31901859)(41)
200100942	GSE100942	ESCC	Affymetrix Human Genome U133 Plus 2.0 Array	HE	GPL570	10	5	5	Paired	qRT-PCR	Ming at al 2018 (29290801)(42)
200017351	GSE17351	ESCC	Affymetrix Human Genome U133 Plus 2.0 Array	Paraffin blocks	GPL570	10	5	5	Paired	qRT-PCR	Lee et al 2010 (20042640)(43)
200023400	GSE23400	ESCC	Affymetrix Human Genome U133 Array and Affymetrix Human Genome U133B Array	Frozen	GPL96 andGPL97	106	53	53	Paired	qRT-PCR	Su et al 2011 (21385931)(44) Li et al 2014 (24867265)(45) Hyland et al 2016 (26635288)(46)
200033426	GSE33426	ESCC	Affymetrix Human Genome U133A 2.0 Array	Frozen and Hematoxylin and eosin stained	GPL571	71	59	12		IHC	Yan et al 2012 (22280838)(47) Yan et al 2013 (23219752)(48)

200029001	GSE29001	ESCC	Affymetrix Human Genome U133A 2.0 Array	Frozen	GPL571	45	21	24		IHC	Yan et al 2012 (22280838)(47)
200045168	GSE45168	ESCC	Agilent-026652 Whole Human Genome Microarray 4x44K v2		GPL13497	10	5	5	Paired	qRT-PCR	
200045670	GSE45670	ESCC	Affymetrix Human Genome U133 Plus 2.0 Array		GPL570	38	28	10		qRT-PCR	Wen et al 2014 (24907633)(49)
200077861	GSE77861	ESCC	Affymetrix Human Genome U133 Plus 2.0 Array	Frozen	GPL570	14	7	7	Paired	qRT-PCR	Erkizan et al 2019 (28629367)(50)
200038129	GSE38129	ESCC	Affymetrix Human Genome U133A 2.0 Array	Frozen	GPL571	60	30	30	Paired		

BE; Barrett’s esophagus, EAC; esophageal adenocarcinoma, ESCC; esophageal squamous cell carcinoma; FFPE, formalin-fixed paraffin-embedded, HE; hematoxylin and eosin, IHC; immunohistochemistry, qRT-PCT; quantitative reverse transcription polymerase chain reaction.

4.3.5 Data pre-processing: QC, background correction, normalisation, and annotation

All the data pre-processing was done using R statistical software, version 4.0.3. Quality Control (QC) reports were generated using the AffyQC package (61) for all the 14 Affymetrix U133 arrays, and the Oligo QC method (62) was used for the four Affymetrix human ST arrays. QC reports are provided in the Appendix file 4A1. The input data were raw CEL files from microarray analysis, and normalisation resulted in $\log_2$ transformed intensities. The packages allow for the assessment of the quality of arrays in an Affybatch object. Components assessed in the QC report include overall signal quality for the arrays, 3':5' ratio for spiked-in and control genes specific to the array type, clustering of negative and positive elements, position of the centre of intensity on the grid, and assessment of the heat map of the array-array Spearman rank correlation coefficients. QC reports from the datasets are found in Appendix file 4A1. Normalization of all datasets was done using the RMA function in the Affy package.(63) The Affy package is used for data analysis of Affymetrix oligonucleotide array probe level data. The RMA function computes expression in the following way:

i) Probe-specific background correction of the array data using a model where observed intensity is the sum of signal and noise. Spatial variation within individual arrays is corrected for.

ii) Log transformation of base 2 for each probe. This transformation corrects for skewness and generates an equal spread of ratios between up and downregulated genes.

iii) Normalization of perfect match (PM) probes using quantile normalization. This equalizes the data distribution of the arrays, by correcting for biases from non-biological sources.

iv) Expression measure calculation from probe level data using median polish. This probe normalisation is done to correct for variability within probe sets, which will now be equalized and combined into one value for a complete probe set.

The effect of normalization was assessed by comparing raw and normalized data for comparable median expression levels using box plots. Annotation of the datasets from

Probe IDs to Entrez gene IDs was done according to the microarray platform using the following R packages and respective datasets: hgu133plus2.db (GSE13083, GSE1420, GSE45670, GSE77861, GSE100942, GSE17351, GSE20347, GSE29001, GSE33426, GSE36223, GSE38129, GSE39491), hgu133a.db (GSE23400, GSE26886), hugene10sttranscriptcluster.db (GSE92396, GSE100842, GSE34619), huex10sttranscriptcluster.db (GSE75241). During annotation of duplicate Entrez gene IDs, probes that mapped to multiple genes were removed by a selection of those with higher variance. This was done to prevent misinterpretation of results and to increase the specificity of the analysis. Entrez gene IDs detected as NAs or null values were also filtered out. The processed microarray datasets were then used in the meta-analysis of DEGs.

4.3.6 Meta-analysis of DEGs

To generate one combined dataset, expression data frames for the 18 datasets were merged by common Entrez gene IDs. This merged data frame was used for downstream analysis. The R Bioconductor package RankProd (64) was used for the microarray DEG meta-analyses of the dataset, according to the manual. RankProd uses a non-parametric rank product (RP) method to detect genes and other variables that are consistently up or downregulated in repeat experiments. Combining multiple datasets for RP analysis increases the power of the statistical test and results in more genes being selected. The RankProduct method is based on relatively weak assumptions compared to other methods, and these include:

i) A minority of all the features measured are up or downregulated.

ii) Independence of measurements between replicate experiments.

iii) A majority of the changes are independent of each other.

iv) Equal measurement variance for all measurements. Variance stabilisation during the normalization process addressed this assumption in our analysis.

An RP statistic, which is a product of the RP method, for a specific gene is defined as the geometric mean of all the ranks of the gene obtained from each experiment. The equation used in the RP method to generate an RP statistic is as follows:

$$RP_i = \left(\prod_{j=1}^{K} r_{i,j} \right)^{1/K}$$

Where i is the i^{th} gene and j is the j^{th} replicate experiment and $r_{i,j}$ is the position of the i^{th} gene in the j^{th} replicate experiment in a list of up or downregulated genes ordered according to fold changes. Where K is the number of replicated experiments. Pairs of datasets were selected for the RP method analysis (tumour vs normal tissue). The input data for the analysis included the normalised expression datasets with expression levels and probe IDs from the microarray analysis, a vector containing the class labels of all the samples in the dataset, i.e., case and controls, and a vector containing the platform origins labels (platforms labelled 1 to 19) of datasets. The Rank Product identifies DEGs based on a percentage of false prediction (pfp <0.05) and fold change; log2 fold change of >1 (downregulated genes) or < 1 (upregulated genes). The fold change was additionally presented in the results section using the following calculation: [-1*(1/FC)].

4.3.7 Gene Set Enrichment Analysis (GSEA)

GSEA is a technique used to identify general trends in the lists of genes or proteins produced by functional genomics techniques and bioinformatics analyses. GSEA methods detect groups of genes that occur significantly more than expected by chance in a given list of genes.(65) There are several limitations of GSEA methods which include:

i) When defining a list of significant genes, the output is highly dependent on arbitrary parameters such as the p-value cut-off.

ii) Results frequently contain many false positive entries as a result of the overlap between many pathways and of bias introduced by the sample source. Gene sets overlap occurs due to the same genes playing a role in different pathways and processes.

iii) Correcting for testing multiple gene sets is not feasible using conventional correction methods due to the many overlaps between gene sets.

SetRank is a novel GSEA algorithm that was developed specifically to address some of these limitations. The key principle of the algorithm is that it removes gene sets that are identified as significant if their significance is only due to the overlap with another gene set.(65) SetRank analysis was performed using the SetRank package on the R statistical software. The input files for analysis included a list of genes from the Rank Product analysis, ranked by p-value, and a gene set collection compiled from the Reactome annotation database(66). The output from the SetRank analysis is a network representation of enriched pathways, based on a p-value cutoff of 0.01 and a false discovery rate cut-off of 0.05. SetRank uses a sophisticated method of sorting results using three p values. These include:

- SetRank p-value, the p-value linked to the SetRank value of the gene set
- Corrected p-value, corrected p-values of the gene set and adjusted for overlap
- Adjusted p-value, p-value adjusted for multiple testing using the Holm procedure and assigned back to the gene sets within each component

The SetRank results presented in this chapter are from the corrected p-values. The pathways were visualized and integrated using the Cytoscape platform.(67)

4.3.8 Mapping of enriched pathways to DEGs

The enriched pathways from the SetRank analysis were also mapped to the DEGs and the corresponding fold changes reported in the RankProd analysis. This was done to determine which genes were involved in pathways as well as whether they were upregulated or downregulated for each particular pathway reported.

4.3.9 Functional annotation of DEGs

The Reactome annotation database(66) was used to explore biological themes and pathways of genes in the SetRank analysis. The pathways in the Reactome database are of high quality and human/expert-curated. What makes it distinctive from other pathway databases is its focus is on the annotation of only one species, *Homo sapiens,* and the application of a single and consistent data model across the platform.(66) It contains entries for 10,867 human protein-coding genes, which is 53%

of all predicted human protein-coding genes.(66) It is the most comprehensive open-source open data pathway database, providing high coverage of the genome.

The pathways from the SetRank analysis were visualized using Cytoscape. An example and interpretation of the gene set networks which are produced in Cytoscape are shown in Figure 4.4. In summary, the diagram shows three Venn diagrams of hypothetical gene set interactions. The red dots characterize the significant genes, and the white dots represent the non-significant genes. The gene sets are represented by the nodes, and the intersections between the gene sets are represented by the edges. There are three types of edges, the first edge shows normal overlap, showing the intersection between the two gene sets. The arrow points from the less significant gene set to the more significant gene set, after subtracting the intersection between the two gene sets. The second edge occurs when the significance of each gene set is only due to the interaction between the two gene sets. The third edge occurs when one gene set is a proper subset of the second gene set.

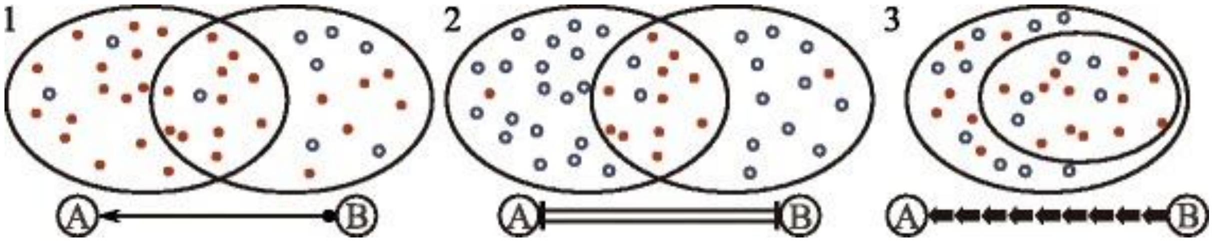


Figure 4.4: *Venn diagrams of hypothetical intersecting gene sets A and B. 1. Straight line arrow, normal overlap. Both gene sets are significant, but the significance of gene set B is partly due to its overlap with gene set A due to gene set A having more significant genes. 2. Two straight lines, intersection only. Significance of both gene set A and B is due to the intersection between both. 3. Multiple arrows, subset. Gene set B is a subset of gene set A. Figure by Simmilon et al (2017)(65) Creative commons licence http://creativecommons.org/licenses/by/4.0/*

4.3.10 Functional interaction networks using ReactomeFIViz application

Additionally, the ReactomeFIViz application (68) was used to investigate functional relationships among genes in some of the pathways identified by the SetRank analysis on the Cytoscape platform. ReactomeFIViz is used for pathway and network-based data analysis using the Reactome database.(68) The application can access Reactome pathways and convert them into Reactome Functional Interaction (FI) networks using a method described by Wu et al (69). For this study, the app was used to show network and pathway patterns from the pathways uncovered from the SetRank analysis.

4.3.11 Substudy: Reliability of subsampling

One of the limitations we had in our analysis was not being able to perform the Rank Product analysis on the entire combined dataset due to the package not being able to analyse a large number of samples at once - exceeding computational resources. We addressed this issue through performing the analysis using different pairwise contrasts of the datasets. Another approach that we used to address this limitation was subsampling. Subsampling refers to reducing the sample size through the selection of a subset of the original dataset.(70) The subsampling was done to reach the sample size which the RankProd package could analyse. We performed a substudy to ascertain the reliability of the subsampling that was performed for some of the pairwise contrasts and to provide more stable estimates of the effects. The substudy of repeated subsampling was important to ensure that we did not add enriched pathways that were arbitrarily significant, and to even out outliers. The subsampling substudy was done in one of the BE pairwise contrasts, which had the following datasets GSE36223, GSE26886, GSE100843, and GSE34619. The Rank Product method was repeated using the set seed function in R to randomise the samples (GSE100843) from where the subsampling was done. Nine repeats of the RankProd were done subsequently followed by nine repeats of the SetRank analysis. The p-value for each enriched pathway produced in the SetRank analysis in each of the repeats was recorded. These p-values were then transformed to z-values and averaged. The z-value averages were transformed back to a mean p-value. Variance for the z-values across the nine repeats was measured and recorded.

4.4 Results

4.4.1 Identification of DEGs using the Rank Product method

We performed a meta-analysis of microarray data using the Rank Product method to identify DEGs in BE, EAC, and ESCC. Merged datasets for BE, EAC, and ESCC were inputted into the RankProd package for identification and meta-analysis of DEGs. One of the major limitations of the RankProd Package is that it cannot analyse large datasets at once, due to a large number of pairwise computations needed to be performed for each analysis. This becomes computationally intensive and requires large amounts of memory. We, therefore, performed the analysis using different pairwise contrasts of datasets up to the sample size that the analysis allowed (range from 180 to 200), BE, EAC, and ESCC. For BE, there were eight pairwise contrasts, including one contrast which included sub-sampling. For ESCC, there were six pairwise contrasts, including one sub-sampling contrast. All the EAC datasets were analysed at once. Model contrasts were selected for BE and ESCC based on the number of datasets and the sample size. The pairwise contrasts are shown in Appendix Table 4A2.

4.4.2 Identification of DEGs between BE samples and normal tissue samples

The model pairwise contrasts for BE included GSE100843, GSE34619, GSE13083, GSE26886 datasets, with 77 tumour samples and 85 normal tissue samples. There were 1,181 downregulated genes and 767 upregulated genes identified in the RankProd Analysis. Figure 4.5 shows the up- and downregulated genes in diagrammatic format. In the diagram, the estimated percentage of false prediction (pfp) vs. the number of identified genes using the output from RankProd is shown. The pfp cut-off is specified as 0.05, therefore the identified genes are shown in red. Table 4.6 and 4.7 show the top 30 most up and downregulated genes, respectively. RankProd analysis identified *sciellin (SCEL), small proline rich protein 3 (SPRR3), desmocollin 3 (DSC3), and desmoglein 3 (DSG3)* as the most downregulated genes with fold-change ranging from 11.2(-0.09) to 7.2 (-0.14). The most upregulated genes were *anterior gradient 2, protein disulphide isomerase family member (AGR2), sulfotransferase family 1C member 2 (SULT1C2), mucin 13, cell surface associated (MUC13), and* *alanyl aminopeptidase, membrane (ANPEP)* with fold change ranging from 0.03(-33.3) to 0.09(-11.1).

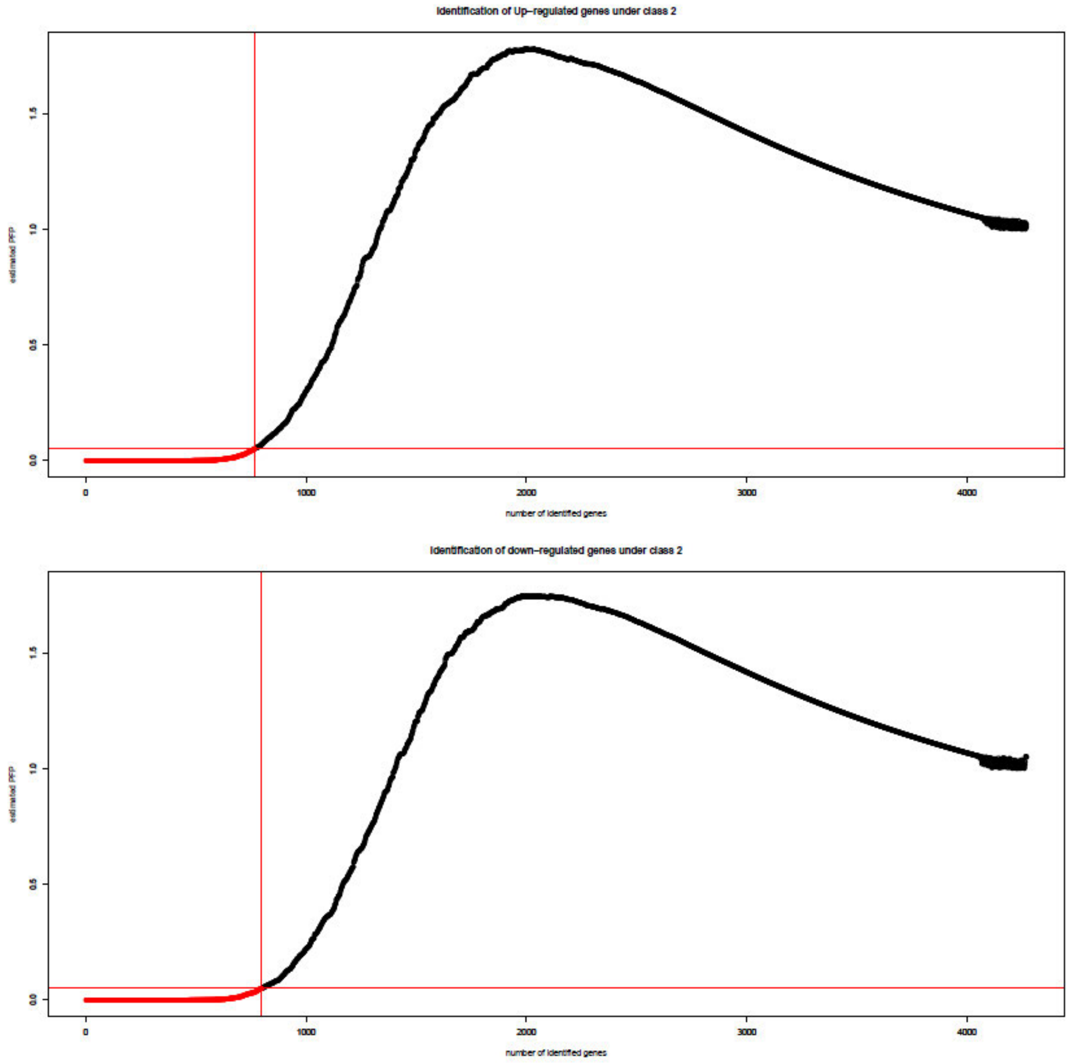


***Figure 4.5:** Graphical display of the estimated percentage of false prediction (pfp) vs. number of identified genes using the output from RankProd for BE (model contrast) analysis.*

Table 4.6: The top 30 upregulated genes in BE

Gene name	Symbol	Fold Change	pfp[1]
Anterior gradient 2, protein disulphide isomerase family member	*AGR2*	0.03	0
Sulfotransferase family 1C member 2	*SULT1C2*	0.06	2.70E-136
Mucin 13, cell surface associated	*MUC13*	0.06	4.90E-122
Alanyl aminopeptidase, membrane	*ANPEP*	0.09	8.90E-109
Claudin 18	*CLDN18*	0.08	8.71E-107
Polymeric immunoglobulin receptor	*PIGR*	0.12	2.55E-104
Transmembrane channel like 5	*TMC5*	0.08	2.47E-102
3-hydroxy-3-methylglutaryl-coa synthase 2	*HMGCS2*	0.15	1.09E-84
IQ motif containing gtpase activating protein 2	*IQGAP2*	0.17	7.28E-82
Endosome-lysosome associated apoptosis and autophagy regulator 1	*ELAPOR 1*	0.18	6.84E-81
Serpin family A member 1	*SERPINA 1*	0.15	3.98E-77
SEL1L family member 3	*SEL1L3*	0.16	1.58E-73
Fatty acid binding protein 1	*FABP1*	0.15	4.52E-72

Table 4.7: The top 30 downregulated genes in BE

Gene name	Symbol	Fold Change	pfp[1]
Sciellin	*SCEL*	11.23	1.46E-118
Small proline rich protein 3	*SPRR3*	9.10	2.11E-103
Desmocollin 3	*DSC3*	8.16	1.71E-102
Desmoglein 3	*DSG3*	7.21	1.98E-101
Epithelial membrane protein 1	*EMP1*	5.43	3.04E-91
Heat shock protein family B (small) member 8	*HSPB8*	5.84	1.62E-82
Protein phosphatase 1 regulatory subunit 3C	*PPP1R3 C*	5.41	9.13E-79
Transmembrane protein 40	*TMEM4 0*	4.10	5.38E-72
Protein tyrosine phosphatase non-receptor type 13	*PTPN13*	4.09	1.27E-64
EH domain containing 3	*EHD3*	4.03	1.50E-64
Annexin A1	*ANXA1*	3.15	2.18E-63
Paired box 9	*PAX9*	3.22	5.03E-62
Peptidyl arginine deiminase 1	*PADI1*	3.18	6.50E-62

Gasdermin B	*GSDMB*	0.21	5.66E-71	SAM and SH3 domain containing 1	*SASH1*	3.51	8.07E-62
Aldolase, fructose-bisphosphate B	*ALDOB*	0.24	2.88E-69	Endoplasmic reticulum oxidoreductase 1 alpha	*ERO1A*	3.03	2.22E-60
Villin 1	*VIL1*	0.19	1.19E-67	MAX dimerization protein 1	*MXD1*	3.20	1.56E-58
Polypeptide N-acetylgalactosaminyltransferase 6	GALNT6	0.20	3.58E-64	Pleckstrin homology like domain family A member 1	PHLDA1	3.86	6.85E-57
GATA binding protein 6	GATA6	0.16	4.73E-63	Myelin protein zero like 2	MPZL2	3.03	1.44E-56
Synaptotagmin like 2	SYTL2	0.28	3.24E-62	Sterile alpha motif domain containing 9	SAMD9	2.79	4.33E-53
CF transmembrane conductance regulator	CFTR	0.19	7.12E-60	Chromosome 18 open reading frame 25	C18orf25	3.45	1.18E-52
Tropomyosin 1	TPM1	0.27	2.83E-59	RAN binding protein 9	RANBP9	3.15	1.20E-51
5'-nucleotidase ecto	NT5E	0.21	7.05E-59	RAR related orphan receptor A	RORA	3.57	4.23E-51
ATP binding cassette subfamily C member 3	ABCC3	0.27	8.31E-57	Nicotinamide phosphoribosyltransferase	NAMPT	2.64	5.19E-51
Endoplasmic reticulum to nucleus signaling 2	ERN2	0.31	3.20E-55	Iodothyronine deiodinase 2	DIO2	3.09	5.98E-51
Acyl-coa synthetase long chain family member 5	ACSL5	0.25	4.59E-55	Dedicator of cytokinesis 9	DOCK9	3.15	1.11E-50
Golgi integral membrane protein 4	GOLIM4	0.29	4.41E-54	Tumor associated calcium signal transducer 2	TACSTD2	2.45	2.55E-49
Atpase sarcoplasmic/endoplasmic reticulum Ca2+ transporting 3	ATP2A3	0.28	7.50E-52	Solute carrier family 16 member 6	SLC16A6	3.25	5.61E-49

Glucosaminyl (N-acetyl) transferase 1	GCNT1	0.31	8.46E-52	Carbonic anhydrase 12	CA12	2.74	1.55E-47
Solute carrier family 12 member 2	SLC12A2	0.09	0	Kallikrein related peptidase 12	KLK12	3.41	1.50E-47
Solute carrier family 6 member 20	SLC6A20	0.09	2.70E-136	GM2 ganglioside activator	GM2A	2.92	1.64E-47

[1] pfp; percentage of false prediction

4.4.3 Identification of DEGs between EAC tumour samples and normal tissue samples

For EAC analysis, three datasets were included: GSE26886, GSE92396, and GSE1420. The three datasets included 45 tumour samples and 41 control samples. The merged dataset of 86 samples was analysed using Rank Product Method. A total of 1,130 genes were found to be differentially expressed. RankProd analysis identified 557 downregulated genes and 573 upregulated genes. A diagrammatic representation of the up- and downregulated genes is shown in Figure 4.6. The top 30 upregulated and downregulated genes are shown in Tables 4.8 and 4.9. The most downregulated genes were *sciellin (SCEL)*, *small proline rich protein 3 (SPRR3), desmoglein 3 (DSG3),* and *desmocollin 3 (DSC3)*, with fold change ranging from 18.9(-0.05) to 9.8(-0.10). The most upregulated genes were *anterior gradient 2, protein disulphide isomerase family member (AGR2), transmembrane channel like 5 (TMC5), and GATA binding protein 6 (GATA6),* and *mucin 13, cell surface associated (MUC13)* with fold-change ranging from 0.09(-11.1) to 0.16(-6.25)*.*

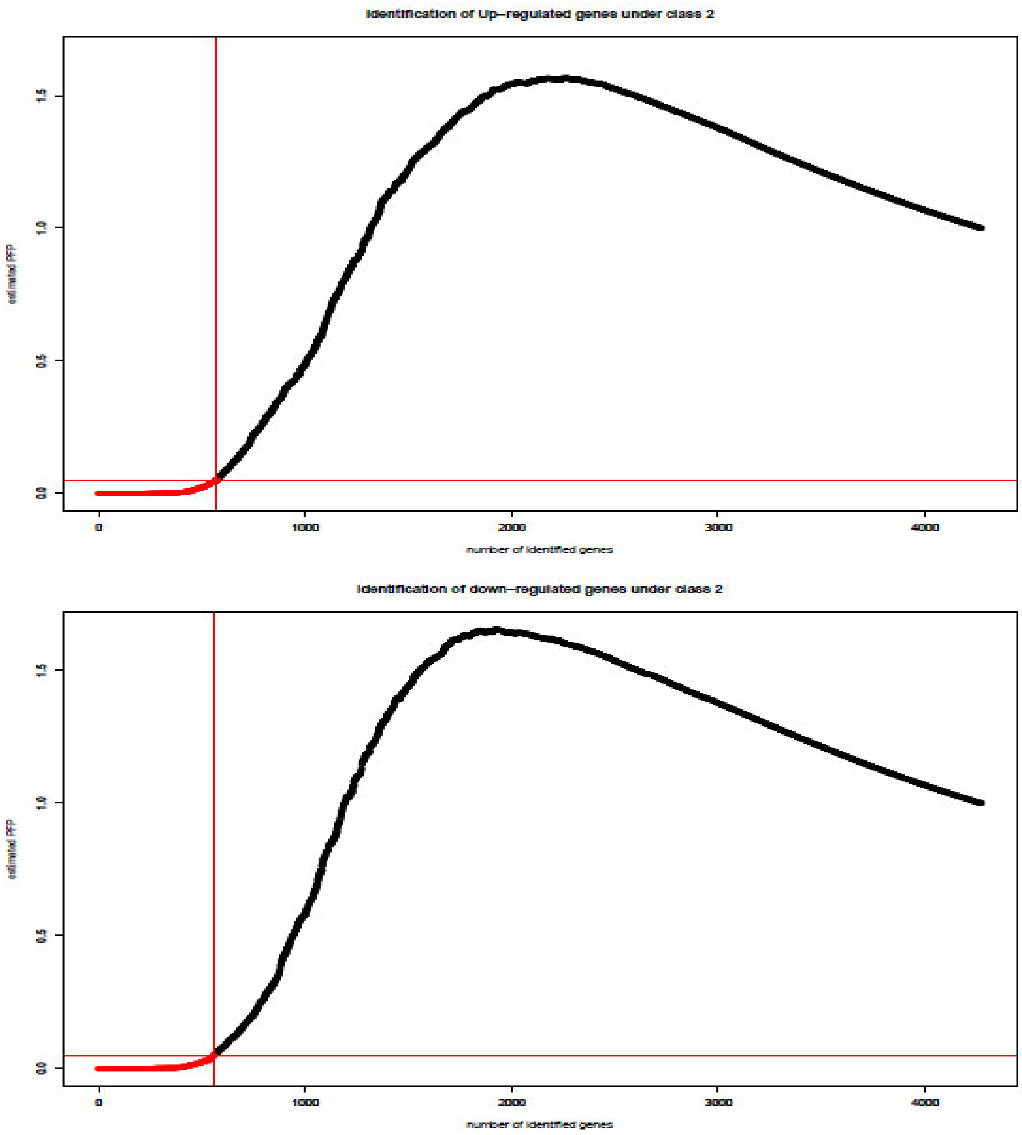


Figure 4.6: *Graphical display of the estimated percentage of false prediction (pfp) vs. number of identified genes using the output from RankProd for EAC analysis.*

Table 4.8: The top 30 upregulated genes in EAC

Gene name	Symbol	Fold Change	pfp[1]
Anterior gradient 2, protein disulphide isomerase family member	*AGR2*	0.09	1.11E-54
Mucin 13, cell surface associated	*MUC13*	0.09	2.79E-48
Transmembrane channel like 5	*TMC5*	0.12	2.14E-47
Sulfotransferase family 1c member 2	*SULT1C2*	0.16	5.81E-38
Gata binding protein 6	*GATA6*	0.19	3.16E-37
Claudin 18	*CLDN18*	0.19	8.04E-34
Agmatinase	*AGMAT*	0.21	2.35E-31
Coagulation factor v	*F5*	0.22	6.75E-29
Villin 1	*VIL1*	0.21	6.04E-29
Sel1l family member 3	*SEL1L3*	0.24	4.25E-28
Transmembrane serine protease 3	*TMPRSS3*	0.25	9.33E-27

Table 4.9: The top 30 downregulated genes in EAC

Gene name	Symbol	Fold Change	pfp[1]
Small proline rich protein 3	*SPRR3*	18.90	8.11E-64
Sciellin	*SCEL*	13.53	2.38E-59
Desmoglein 3	*DSG3*	10.88	6.24E-52
Desmocollin 3	*DSC3*	9.77	4.22E-48
Protein phosphatase 1 regulatory subunit 3C	*PPP1R3C*	6.90	4.39E-42
Epithelial membrane protein 1	*EMP1*	6.51	5.38E-41
Kallikrein related peptidase 12	*KLK12*	5.92	2.19E-37
Heat shock protein family B (small) member 8	*HSPB8*	5.71	4.74E-36
Annexin A1	*ANXA1*	4.36	5.75E-32
EH domain containing 3	*EHD3*	4.18	5.64E-29
Acid phosphatase 3	*ACP3*	4.32	8.01E-29

Serpin family a member 1	*SERPIN A1*	0.28	7.76E-26	Protein tyrosine phosphatase non-receptor type 13	*PTPN13*	4.32	5.70E-27
Collagen type i alpha 2 chain	*COL1A2*	0.28	7.53E-26	Myelin protein zero like 2	*MPZL2*	3.59	3.45E-26
Growth differentiation factor 15	*GDF15*	0.28	3.15E-25	Transmembrane protein 40	*TMEM40*	4.53	5.93E-26
Alanyl aminopeptidase, membrane	*ANPEP*	0.28	9.91E-23	MAX dimerization protein 1	*MXD1*	3.20	3.54E-25
Lumican	*LUM*	0.29	3.89E-22	Desmocollin 2	*DSC2*	3.32	2.02E-24
Gasdermin b	*GSDMB*	0.31	4.71E-21	Glycolipid transfer protein	*GLTP*	3.39	1.38E-23
5'-nucleotidase ecto	NT5E	0.29	7.77E-21	Dehydrogenase/reductase 9	DHRS9	3.49	1.03E-22
Solute carrier family 12 member 2	SLC12A2	0.30	8.04E-21	SRY-box transcription factor 2	SOX2	3.84	1.76E-22
Microtubule associated monooxygenase, calponin and lim domain containing 2	MICAL2	0.33	1.52E-20	Paired box 9	PAX9	3.43	3.05E-22
Cadherin 11	CDH11	0.32	1.72E-20	Palmdelphin	PALMD	3.28	2.55E-21
Fatty acyl-coa reductase 2	FAR2	0.31	2.36E-20	Solute carrier family 16 member 6	SLC16A6	3.26	2.95E-21
Tropomyosin 1	TPM1	0.33	4.43E-20	Kallikrein related peptidase 8	KLK8	2.87	3.21E-21
Acyl-coa synthetase long chain family member 5	ACSL5	0.31	5.29E-20	Chromosome 18 open reading frame 25	C18orf25	3.19	3.71E-21
Phospholipase a and acyltransferase 3	PLAAT3	0.33	7.00E-19	Cyclin G2	CCNG2	2.77	1.27E-20
Cf transmembrane conductance regulator	CFTR	0.31	8.47E-19	Carbonic anhydrase 12	CA12	3.07	1.44E-20
Solute carrier family 6 member 20	SLC6A20	0.29	1.07E-18	Kallikrein related peptidase 7	KLK7	2.70	2.03E-20
Solute carrier family 3 member 1	SLC3A1	0.34	1.32E-18	Iodothyronine deiodinase 2	DIO2	2.98	3.14E-20

Anterior gradient 2, protein disulphide isomerase family member	NNMT	0.09	2.05E-18	PDZ domain containing 2	PDZD2	3.05	3.36E-20
Mucin 13, cell surface associated	PLEKHB1	0.09	3.53E-18	Tumor protein p53 regulated apoptosis inducing protein 1	TP53AIP1	3.25	1.01E-19

[1]pfp; percentage of false prediction

4.4.4 Identification of DEGs between ESCC tumour samples and normal tissue samples

Four datasets were included in the ESCC model pairwise contrast, these were GSE20347, GSE33426, GSE29001, and GSE38129. The analyses comprised 109 cases and 83 controls. The RankProd analyses identified 1,832 differentially expressed genes, of these, 826 were downregulated and 1,006 were upregulated (Figure 4.7). The top 30 up and downregulated genes are shown in Tables 4.10 and 4.11. The most downregulated genes outputted by the analysis were *sciellin (SCEL), epithelial membrane protein 1 (EMP1), small proline rich protein 3 (SPRR3), and protein phosphatase 1 regulatory subunit 3C (PPP1R3C)*, with fold-change ranging from 22.5(-0.04) to 8.2(-0.12). The most upregulated genes were *collagen type XI alpha 1 chain (COL11A1), collagen type I alpha 2 chain (COL1A2), epithelial cell transforming 2 (ECT2), and inhibin subunit beta A (INHBA),* with fold-change ranging from 0.11(-9.09) to 0.14(-7.14).

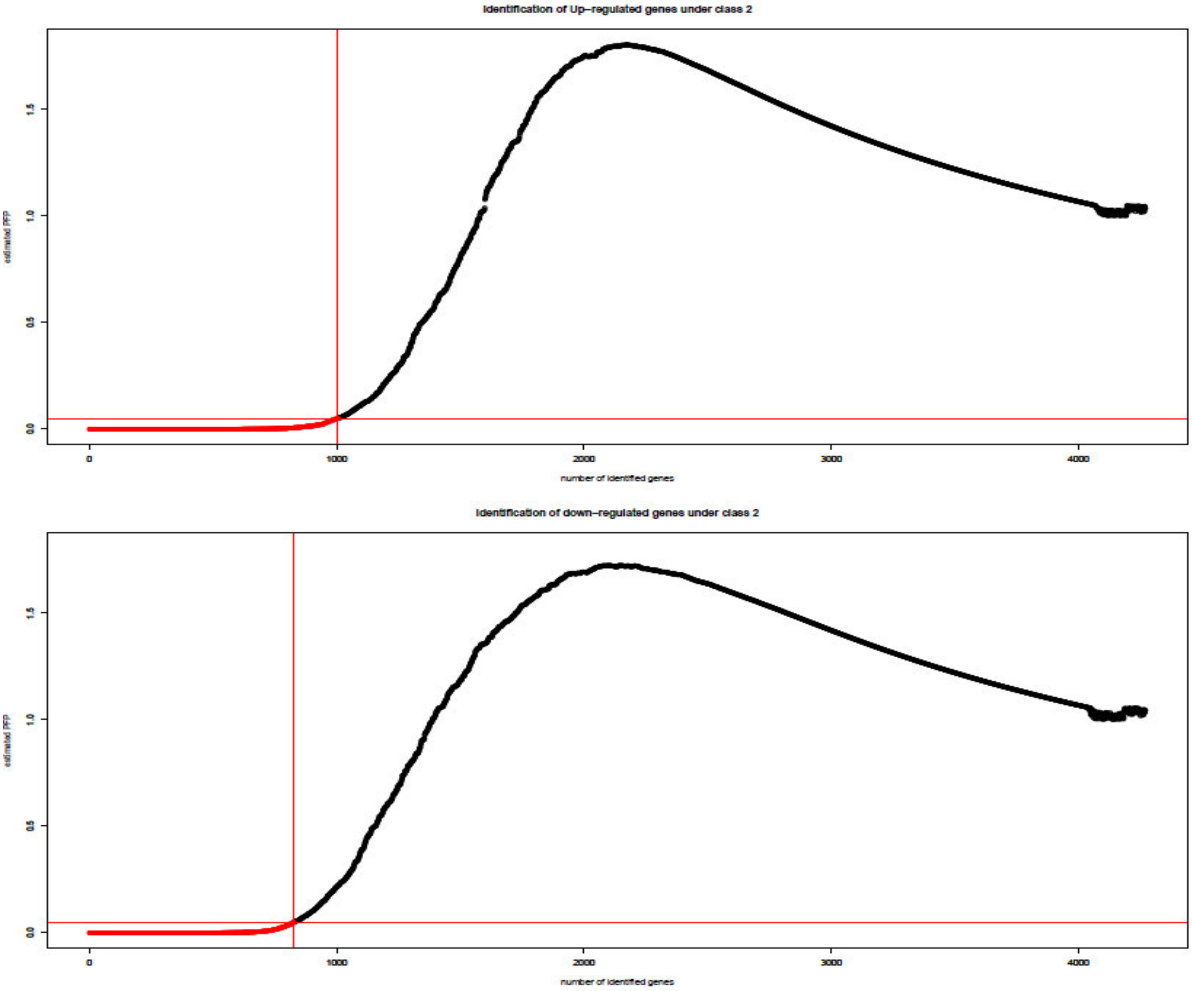


Figure 4.7: *Graphical display of the estimated percentage of false prediction (pfp) vs. number of identified genes using the output from RankProd for ESCC (model contrast) analysis.*

Table 4.10: The top 30 upregulated genes in ESCC

Gene name	Symbol	Fold Change	pfp[1]
collagen type XI alpha 1 chain	*COL11A1*	0.11	1.53E-97
collagen type I alpha 2 chain	*COL1A2*	0.14	1.49E-81
epithelial cell transforming 2	*ECT2*	0.14	3.11E-78
inhibin subunit beta A	*INHBA*	0.14	8.78E-77
Zic family member 1	*ZIC1*	0.14	2.07E-75
mitochondrial assembly of ribosomal large subunit 1	*MALSU1*	0.17	1.86E-70
DNA topoisomerase II alpha	*TOP2A*	0.20	2.49E-60
cadherin 11	*CDH11*	0.23	1.57E-54
fibronectin type III domain containing 3B	*FNDC3B*	0.21	5.44E-52
collagen type III alpha 1 chain	*COL3A1*	0.26	9.85E-51
malignant fibrous histiocytoma amplified sequence 1	*MFHAS1*	0.22	2.71E-48
mitotic arrest deficient 2 like 1	*MAD2L1*	0.25	4.87E-46

Table 4.11: The top 30 downregulated genes in ESCC

Gene name	Symbol	Fold Change	pfp[1]
sciellin	*SCEL*	22.49	8.17E-133
epithelial membrane protein 1	*EMP1*	11.91	9.01E-97
small proline rich protein 3	*SPRR3*	17.62	4.32E-95
protein phosphatase 1 regulatory subunit 3C	*PPP1R3C*	8.24	2.56E-87
serpin family B member 1	*SERPINB1*	8.01	4.25E-80
chromosome 1 open reading frame 116	*C1orf116*	8.19	2.39E-77
acid phosphatase 3	*ACP3*	7.09	3.13E-77
transmembrane serine protease 2	*TMPRSS2*	7.77	1.97E-73
monoglyceride lipase	*MGLL*	6.11	2.50E-70
kallikrein related peptidase 12	*KLK12*	6.31	5.01E-67
Cbp/p300 interacting transactivator with Glu/Asp rich carboxy-terminal domain 2	*CITED2*	6.07	6.91E-67
MAX dimerization protein 1	*MXD1*	6.65	1.65E-65

kinesin family member 23	*KIF23*	0.27	7.32E-44	synaptopodin 2 like	*SYNPO2L*	7.29	6.65E-65
cyclin dependent kinase 1	*CDK1*	0.27	1.01E-43	cytochrome P450 family 3 subfamily A member 5	*CYP3A5*	5.48	5.18E-64
transferrin receptor	*TFRC*	0.26	1.31E-43	nucleobindin 2	*NUCB2*	5.68	4.79E-63
ENAH actin regulator	*ENAH*	0.24	1.55E-43	heat shock protein family B (small) member 8	*HSPB8*	5.41	3.01E-62
denticleless E3 ubiquitin protein ligase homolog	*DTL*	0.27	1.82E-43	alcohol dehydrogenase 1B (class I). beta polypeptide	*ADH1B*	4.72	4.71E-60
BH3 interacting domain death agonist	*BID*	0.26	1.73E-43	iodothyronine deiodinase 2	*DIO2*	4.92	6.59E-58
ATPase family AAA domain containing 2	*ATAD2*	0.28	6.59E-43	polypeptide N-acetylgalactosaminyltransferase 12	*GALNT12*	4.36	1.34E-53
kinesin family member 14	*KIF14*	0.28	6.46E-43	chromosome 18 open reading frame 25	*C18orf25*	4.89	7.34E-53
FA complementation group I	*FANCI*	0.27	1.28E-42	inhibitor of DNA binding 4. HLH protein	*ID4*	4.17	1.64E-52
atypical chemokine receptor 3	*ACKR3*	0.29	3.72E-41	adhesion G protein-coupled receptor F1	*ADGRF1*	5.98	1.95E-51
asporin	*ASPN*	0.27	4.44E-41	dehydrogenase/reductase 9	*DHRS9*	4.86	3.90E-51
insulin like growth factor 2 mRNA binding protein 2	*IGF2BP2*	0.30	7.05E-41	myelin protein zero like 2	*MPZL2*	4.04	4.26E-49
neuropilin and tolloid like 2	*NETO2*	0.28	3.25E-40	endoplasmic reticulum oxidoreductase 1 alpha	*ERO1A*	4.08	6.11E-48

integrin subunit alpha 6	*ITGA6*	0.29	1.39E-39	sorbin and SH3 domain containing 2	*SORBS2*	3.86	2.13E-47
matrix metallopeptidase 11	*MMP11*	0.31	1.41E-38	cystatin E/M	*CST6*	4.05	6.50E-47
lumican	*LUM*	0.31	1.74E-38	PDZ and LIM domain 2	*PDLIM2*	4.33	1.32E-46
minichromosome maintenance 10 replication initiation factor	*MCM10*	0.30	6.42E-38	SAM and SH3 domain containing 1	*SASH1*	3.88	1.68E-46
centromere protein N	*CENPN*	0.31	1.48E-36	RIO kinase 3	*RIOK3*	3.98	2.13E-46

[1] pfp; percentage of false prediction

4.4.5 Identification of DEGs between squamous dysplasia samples and normal tissue samples

We identified only one dataset which assessed squamous dysplasia, with a sample size of six (two dysplastic tissue and four normal tissue samples). The sample size was therefore not enough for a meta-analysis and subsequent analysis.

4.4.6 Comparison of results for BE, EAC and ESCC

Overall, the RankProd analysis outputted a total of 1,107 upregulated genes for BE, EAC and ESCC with 216 overlapping genes. A total of 1,537 genes were downregulated for BE, EAC and ESCC with 341 overlapping genes. Venn diagrams of the up and downregulated genes, as well as overlapping genes for BE, EAC and, ESCC, are shown in Figure 4.8. Of the top 15 most significantly upregulated genes, eight overlapped between BE and EAC were *AGR2, SULT1C2, MUC13, ANPEP, CLDN18, TMC5, SERPINA1*, and *SEL1L3*. One gene, *COL1A2*, overlapped between EAC and ESCC. In the top 15 most significantly downregulated genes, *SCEL, SPRR3, PPP1R3C, EMP1* showed overlap among BE, EAC and ESCC, whilst seven genes, *DSC3, DSG3, EMP1, HSPB8, TMEM40, PTPN13, EHD3,* and *ANXA1* showed overlap between BE and EAC. Three genes, *ACP3, KLK12*, and *MXD1* showed overlap between EAC and ESCC.

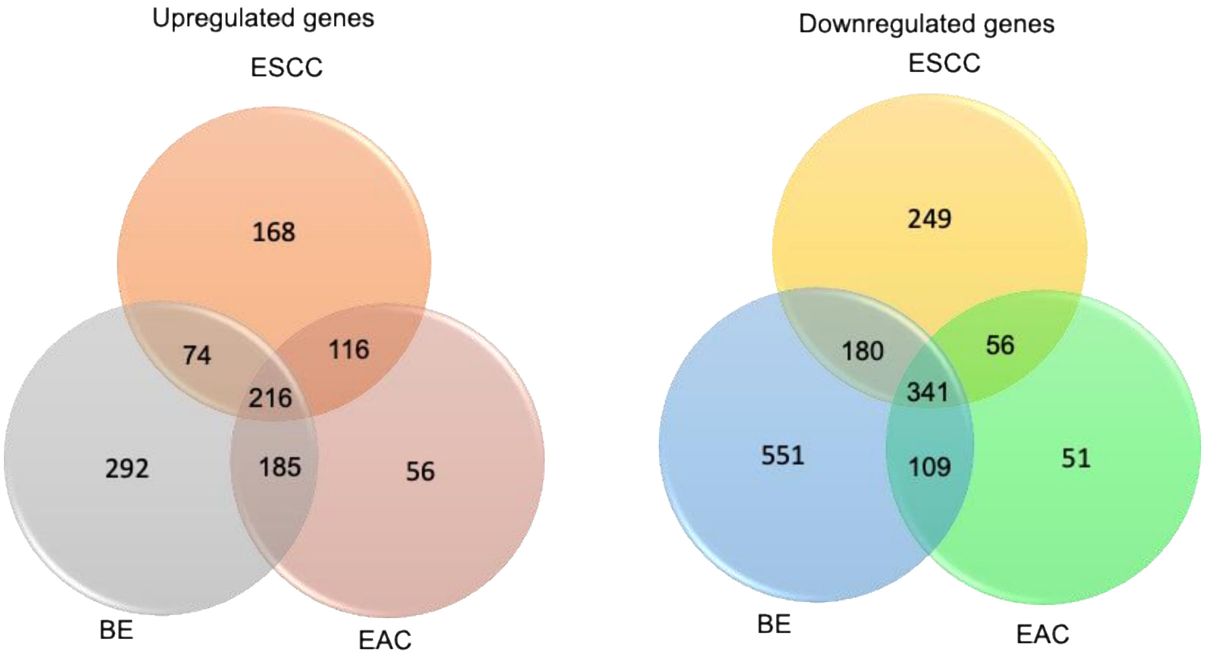


Figure 4.8: *Comparison of results for up- and downregulated genes. Venn diagrams showing the number of upregulated, downregulated, and overlapping genes for BE, EAC and ESCC obtained from the RankProd analysis*

4.4.7 Advanced gene set enrichment analysis using SetRank

Output from the RankProd meta-analysis was used in the SetRank analysis to identify gene sets and networks of enriched biological pathways. This included output from the eight pairwise contrasts for BE, six pairwise contrasts for ESCC and the output from the single EAC analysis. A total of 97 pathways were identified by the SetRank analyses for BE, EAC and ESCC, with varying p-values (Appendix Table 4A2). The most common enriched pathways which were present in all three tumour types were the formation of the cornified envelope, smooth muscle contraction, and glycogen storage diseases.

4.4.8 Advanced gene set enrichment analysis for BE

SetRank analysis for BE identified 45 enriched pathways in all the eight BE pairwise contrasts using the Reactome Database. In the model pairwise contrast, 18 enriched pathways were identified. Table 4.12 shows these pathways as well as the number of genes involved in each pathway, and the corrected and adjusted p-values. The top enriched pathways according to the corrected p-values included formation of the cornified envelope (p=1.30E-05), transport of inorganic cations/anions and amino acids/oligopeptides (p=7.49E-04), glycosaminoglycan metabolism (p=2.77E-03). The pathways neutrophil granulation had the highest number of genes, 114.

Table 4.12: Biological pathways identified by the Advanced Gene Set Enrichment Analysis. SetRank for BE

Pathway Name	**Number of genes**	**Corrected P-value[1]**	**Adjusted P-value[2]**
BMAL1: CLOCK. NPAS2 activates circadian gene expression	20	6.50E-03	4.38E-02
Cytosolic sulfonation of small molecules	5	1.77E-03	2.48E-02
Glycosaminoglycan metabolism	35	2.77E-04	4.15E-03
Hormone-sensitive lipase (HSL)-mediated triacylglycerol hydrolysis	6	3.38E-03	4.27E-02
Nicotinate metabolism	6	6.26E-03	4.38E-02
Assembly of collagen fibrils and other multimeric structures	14	3.28E-03	4.27E-02
Metabolism of Angiotensinogen to Angiotensins	3	6.40E-03	4.38E-02
Tight junction interactions	7	7.68E-03	4.38E-02
Transport of inorganic cations/anions and amino acids/oligopeptides	29	7.49E-05	1.20E-03
Smooth Muscle Contraction	14	6.55E-03	4.38E-02
Defective GALNT3 causes familial hyperphosphatemic tumoral calcinosis (HFTC)	2	3.84E-03	4.27E-02
Neutrophil degranulation	114	4.18E-03	4.27E-02
TP53 regulates transcription of several additional cell death genes whose specific roles in p53-dependent apoptosis remain uncertain	5	9.38E-03	4.38E-02
Formation of the cornified envelope	12	1.30E-06	2.29E-05
Glucose metabolism	19	4.17E-03	4.27E-02
Ketone body metabolism	3	8.33E-03	4.38E-02
RAB geranylgeranylation	21	5.06E-03	4.27E-02
O-linked glycosylation of mucins	15	3.82E-03	4.27E-02

[1] Holm correction for multiple testing for the gene set; [2] Holm correction second round. correcting for dependence among pathways. Number of pathways = 18

The pathways were visualised using the Cytoscape platform. Using a Cytoscape function, all the BE pathways from the eight pairwise contrasts were merged. The merged pathways are shown in Figure 4.9. The diagram shows the intersections and interconnectedness of 13 pathways, due to the overlap of genes in the gene sets.

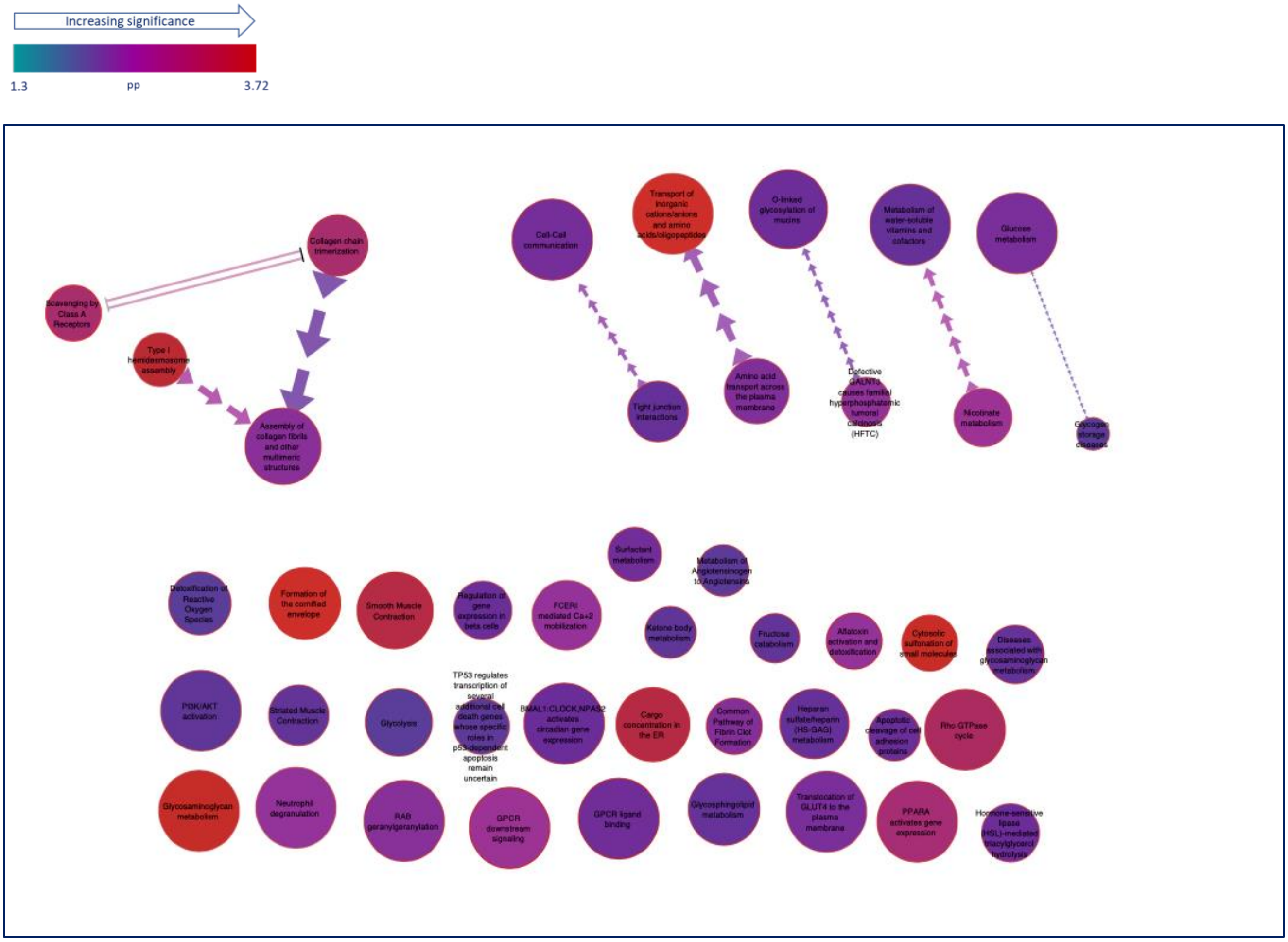

Increasing significance
1.3
pp
3.72
Collagen chain trimerization
Scavenging by Class A Receptors
Type I hemidesmosome assembly
Assembly of collagen fibrils and other multimeric structures
Cell-Cell communication
Tight junction interactions
Transport of inorganic cations/anions and amino acids/oligopeptides
Amino acid transport across the plasma membrane
O-linked glycosylation of mucins
Defective GALNT3 causes familial hyperphosphatemic tumoral calcinosis (HFTC)
Metabolism of water-soluble vitamins and cofactors
Nicotinate metabolism
Glucose metabolism
Glycogen storage diseases
Surfactant metabolism
Metabolism of Angiotensinogen to Angiotensins
Detoxification of Reactive Oxygen Species
Formation of the cornified envelope
Smooth Muscle Contraction
Regulation of gene expression in beta cells
FCERI mediated Ca+2 mobilization
Ketone body metabolism
Fructose catabolism
Aflatoxin activation and detoxification
Cytosolic sulfonation of small molecules
Diseases associated with glycosaminoglycan metabolism
PI3K/AKT activation
Striated Muscle Contraction
Glycolysis
TP53 regulates transcription of several additional cell death genes whose specific roles in p53-dependent apoptosis remain uncertain
BMAL1:CLOCK,NPAS2 activates circadian gene expression
Cargo concentration in the ER
Common Pathway of Fibrin Clot Formation
Heparan sulfate/heparin (HS-GAG) metabolism
Apoptotic cleavage of cell adhesion proteins
Rho GTPase cycle
Glycosaminoglycan metabolism
Neutrophil degranulation
RAB geranylgeranylation
GPCR downstream signaling
GPCR ligand binding
Glycosphingolipid metabolism
Translocation of GLUT4 to the plasma membrane
PPARA activates gene expression
Hormone-sensitive lipase (HSL)-mediated triacylglycerol hydrolysis

Figure 4.9: *Merged gene set network of Reactome pathways identified by the advanced GSEA analysis (SetRank) for BE. Node fill colour indicates level of significance (blue to red – increasing significance). Node size denotes number of genes in the gene set. Edge thickness denotes size of the intersection. pp – negative logarithm of the p-value.*

4.4.9 Advanced gene set enrichment analysis for EAC

Analysis for EAC identified twelve enriched pathways using the Reactome Database (Table 4.13). The top enriched pathways according to the corrected p-values, formation of the cornified envelope (p=2.23E-07), transport of inorganic cations/anions and amino acids/oligopeptides (p=5.27E-03), smooth muscle contraction (p=9.00E-03). The gene set with the most genes (78) was *"Extracellular matrix organisation".* Cytoscape visualisation showed no intersections between the gene sets (Figure 4.10).

Table 4.13: Biological pathways identified by the Advanced Gene Set Enrichment Analysis. SetRank for EAC

Pathway Name	**Number of genes**	**Corrected P-value**	**Adjusted P-value**
Extracellular matrix organization	78	4.51E-03	3.16E-02
Nicotinate metabolism	6	5.06E-03	3.16E-02
Peptide hormone metabolism	21	8.10E-03	3.56E-02
Glycogen storage diseases	1	7.11E-03	3.56E-02
Apoptotic cleavage of cell adhesion proteins	3	3.56E-03	2.84E-02
Transport of inorganic cations/anions and amino acids/oligopeptides	29	5.27E-04	5.80E-03
Smooth Muscle Contraction	14	9.00E-04	9.00E-03
Defective GALNT3 causes familial hyperphosphatemic tumoral calcinosis (HFTC)	2	8.87E-03	3.56E-02
RHO GTPases activate IQGAPs	5	8.02E-03	3.56E-02
Cargo concentration in the ER	13	7.28E-03	3.56E-02
Interleukin-4 and 13 signalling	24	9.26E-04	9.00E-03
Formation of the cornified envelope	12	2.23E-07	3.00E-06

[1]Holm correction for multiple testing for the gene set; [2]Holm correction second round. correcting for dependence among pathways. Number of pathways = 12

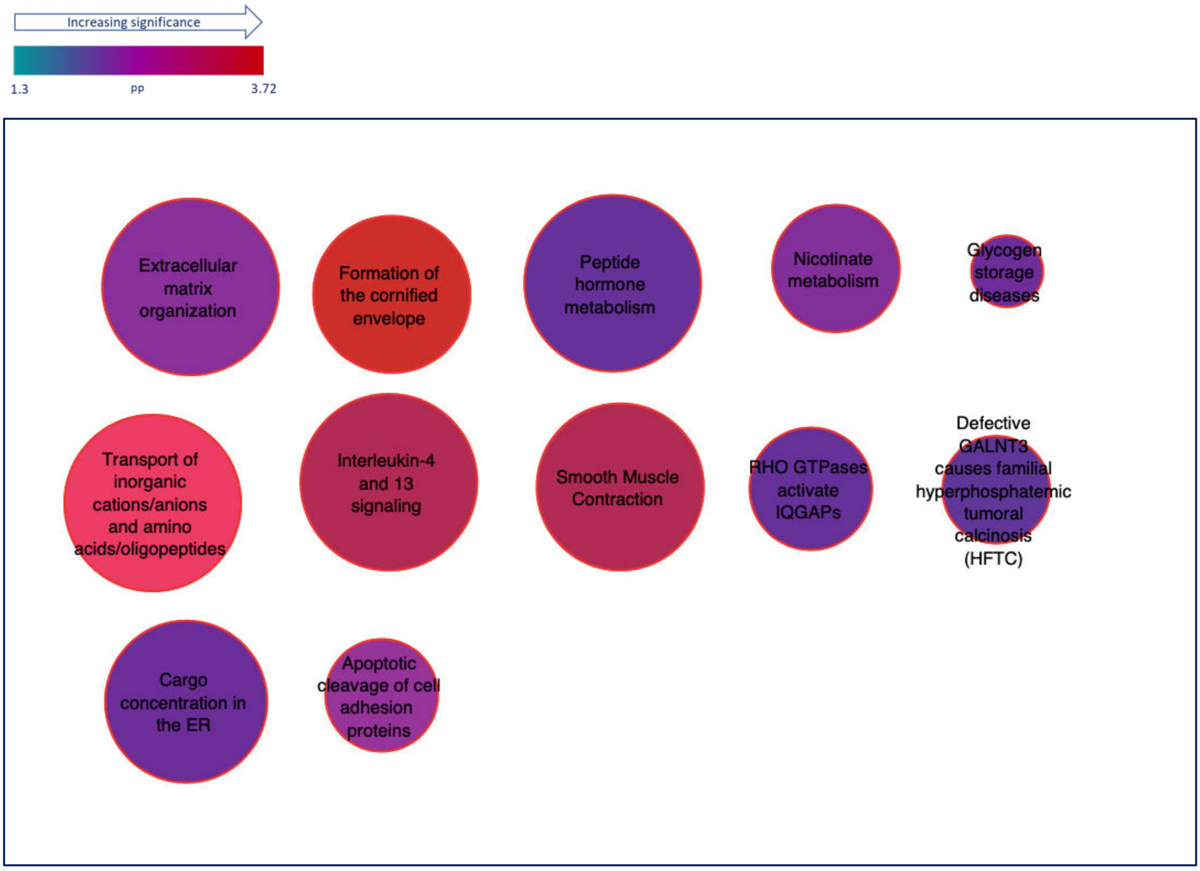

Increasing significance
1.3
pp
3.72
Extracellular matrix organization
Formation of the cornified envelope
Peptide hormone metabolism
Nicotinate metabolism
Glycogen storage diseases
Transport of inorganic cations/anions and amino acids/oligopeptides
Interleukin-4 and 13 signaling
Smooth Muscle Contraction
RHO GTPases activate IQGAPs
Defective GALNT3 causes familial hyperphosphatemic tumoral calcinosis (HFTC)
Cargo concentration in the ER
Apoptotic cleavage of cell adhesion proteins

Figure 4.10: *Merged gene set network of Reactome pathways identified by the advanced GSEA analysis (SetRank) for EAC. Node fill colour indicates level of significance (blue to red – increasing significance). Node size denotes number of genes in the gene set. Edge thickness denotes size of the intersection. pp – negative logarithm of the p-value.*

4.4.10 Advanced gene set enrichment analysis for ESCC

The SetRank analysis identified a total of 62 Reactome pathways in all the ESCC model pairwise contrasts. The model pairwise contrast had 21 pathways, with the top pathways according to the corrected p values being collagen chain trimerization (p=5.40E-04), G2/M Checkpoints (p=1.15E-03), integrin cell surface interactions (p=7.40E-03). The pathways are shown in Table 4.14 with the corresponding corrected and adjusted p values. Metabolism had the most genes in all the gene sets (545).

Table 4.14: Biological pathways identified by the Advanced Gene Set Enrichment Analysis. SetRank for ESCC

Pathway Name	Number of genes	Corrected P-value	Adjusted P-value
E2F-enabled inhibition of pre-replication complex formation	3	7.41E-03	1.38E-03
Metabolism	545	2.98E-03	1.90E-02
Cyclin B2 mediated events	2	2.78E-03	1.38E-03
Glycosaminoglycan metabolism	35	3.28E-03	1.90E-02
Metabolism of water-soluble vitamins and cofactors	31	7.61E-03	1.90E-02
Glycoprotein hormones	2	6.07E-03	3.04E-02
Phase 1 - Functionalization of compounds	19	4.49E-03	1.90E-02
Integrin cell surface interactions	21	7.40E-04	8.14E-03
Metabolism of ingested SeMet. Sec. MeSec into H2Se	3	2.11E-03	1.90E-02
Antagonism of Activin by Follistatin	2	6.07E-03	3.04E-02
Glycogen storage diseases	1	2.76E-03	2.21E-02
Arachidonate production from DAG	1	8.28E-03	3.31E-02
Smooth Muscle Contraction	14	3.95E-03	2.37E-02
Homologous DNA Pairing and Strand Exchange	10	9.64E-03	1.38E-03
Signaling by MET	26	3.00E-03	2.21E-02
Formation of the cornified envelope	12	1.15E-03	1.15E-02
G2/M Checkpoints	34	1.15E-04	1.38E-03
Amino acid synthesis and interconversion (transamination)	8	2.17E-03	1.90E-02
Metabolism of amino acids and derivatives	93	6.46E-03	1.90E-02
TFAP2 (AP-2) family regulates transcription of other transcription factors	1	9.39E-03	3.31E-02
Collagen chain trimerization	7	5.40E-05	7.00E-04
Interferon alpha/beta signaling	20	9.81E-03	3.31E-02
O-linked glycosylation of mucins	15	8.77E-03	3.31E-02

[1] Holm correction for multiple testing for the gene set; [2] Holm correction second round. correcting for dependence among pathways. Number of pathways =23

The pathways from all the ESCC pairwise contrasts were merged and visualised in Cytoscape (Figure 4.11). A total of 41 gene sets showed intersections. The pathways which were interconnected to at least three other pathways included "*Metabolism"* (8), "*Extracellular matrix organisation"* (6), *"Resolution of sister chromatic cohesion"* (3), and *"Signal transduction"* (3).

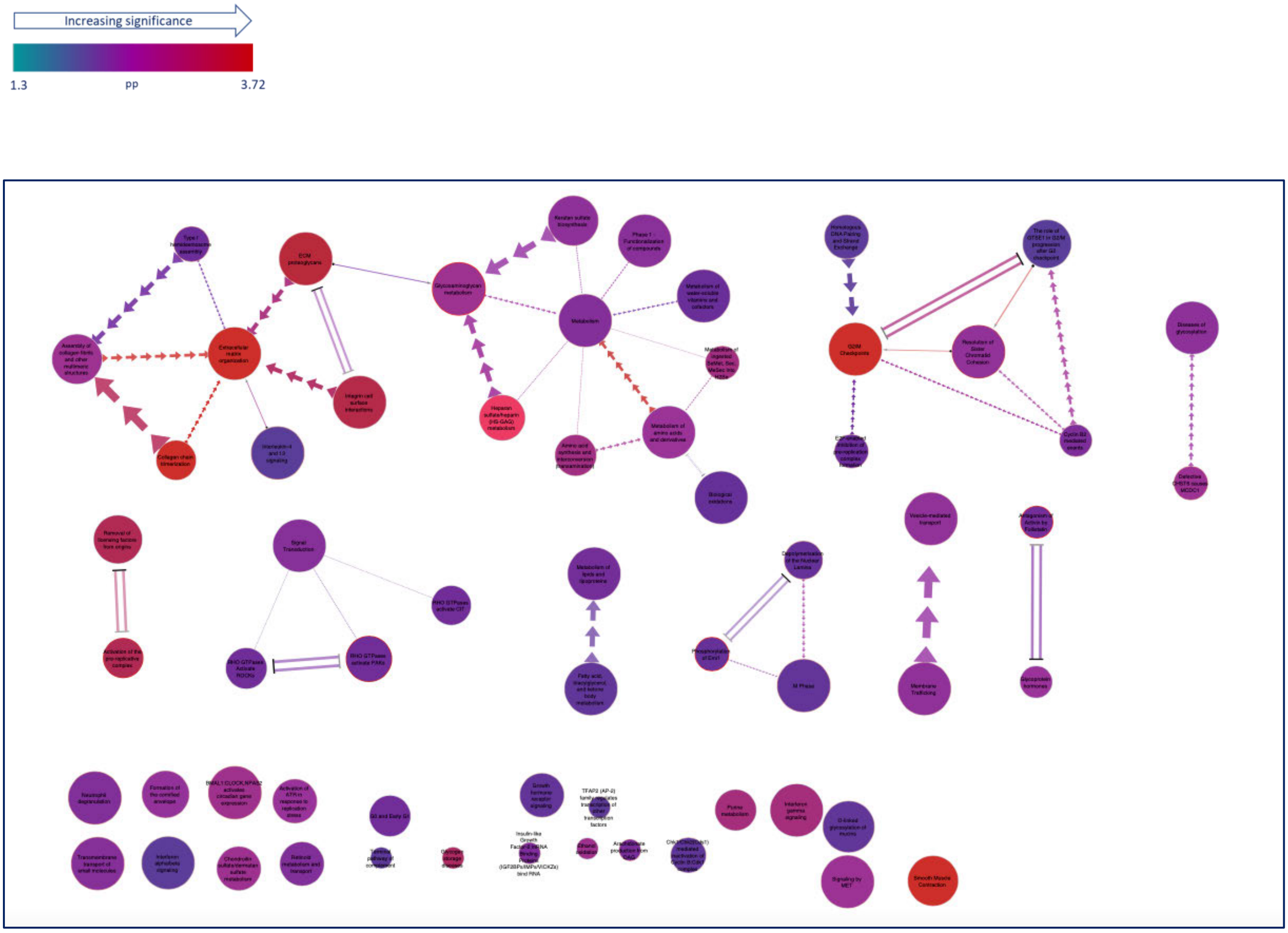

Increasing significance
1.3
pp
3.72
Type I hemidesmosome assembly
ECM proteoglycans
Glycosaminoglycan metabolism
Keratan sulfate biosynthesis
Phase I - Functionalization of compounds
Metabolism
Metabolism of water-soluble vitamins and cofactors
Assembly of collagen fibrils and other multimeric structures
Extracellular matrix organization
Integrin cell surface interactions
Heparan sulfate/heparin (HS-GAG) metabolism
Metabolism of ingested SeMet, Sec, MeSec into H2Se
Collagen chain trimerization
Interleukin-4 and 13 signaling
Amino acid synthesis and interconversion (transamination)
Metabolism of amino acids and derivatives
Biological oxidations
Homologous DNA Pairing and Strand Exchange
The role of GTSE1 in G2/M progression after G2 checkpoint
G2/M Checkpoints
Resolution of Sister Chromatid Cohesion
E2F-enabled inhibition of pre-replication complex formation
Cyclin B2 mediated events
Diseases of glycosylation
Defective CHST6 causes MCDC1
Removal of licensing factors from origins
Activation of the pre-replicative complex
Signal Transduction
RHO GTPases activate CIT
RHO GTPases Activate ROCKs
RHO GTPases activate PAKs
Metabolism of lipids and lipoproteins
Fatty acid, triacylglycerol, and ketone body metabolism
Depolymerisation of the Nuclear Lamina
Phosphorylation of Emi1
M Phase
Vesicle-mediated transport
Membrane Trafficking
Antagonism of Activin by Follistatin
Glycoprotein hormones
Neutrophil degranulation
Formation of the cornified envelope
BMAL1:CLOCK,NPAS2 activates circadian gene expression
Activation of ATR in response to replication stress
G0 and Early G1
Growth hormone receptor signaling
TFAP2 (AP-2) family regulates transcription of other transcription factors
Purine metabolism
Interferon gamma signaling
O-linked glycosylation of mucins
Transmembrane transport of small molecules
Interferon alpha/beta signaling
Chondroitin sulfate/dermatan sulfate metabolism
Retinoid metabolism and transport
Terminal pathway of complement
Glycogen storage diseases
Insulin-like Growth Factor-2 mRNA Binding Proteins (IGF2BPs/IMPs/VICKZs) bind RNA
Ethanol oxidation
Arachidonate production from DAG
Chk1/Chk2(Cds1) mediated inactivation of Cyclin B:Cdk1 complex
Signaling by MET
Smooth Muscle Contraction

Figure 4.11: *Merged gene set network of Reactome pathways identified by the advanced GSEA analysis (SetRank) for ESCC. Node fill colour indicates level of significance (blue to red – increasing significance). Node size denotes number of genes in the gene set. Edge thickness denotes size of the intersection. pp – negative logarithm of the p-value.*

4.4.11 Mapping the DEGs to the enriched pathways

Output from the SetRank analysis showing the enriched pathways were mapped to the DEGs and the corresponding fold change reported in the RankProd analysis.

In the analysis for BE, the formation of the cornified envelope showed ten downregulated genes and one upregulated. *"Transport of inorganic cations/anions and amino acids/oligopeptides"* pathway had eight upregulated and eight downregulated genes. The pathway *"Glycosaminoglycan metabolism"* had seven downregulated genes and eight upregulated genes. The top five pathways according to the corrected p-value as well the corresponding up and downregulated genes are shown in Table 4.15.

Table 4.15: Differentially expressed genes (DEGs) mapped to the enriched pathways for BE vs normal tissue

Enriched pathway	Upregulated genes	Downregulated genes
Formation of the cornified envelope	*CAPNS1. CAPN1. PERP. DSC2. KAZN. KLK8. KLK12. DSG3. DSC3. SPRR3*	*PKP4*
Transport of inorganic cations/anions and amino acids/oligopeptides	*SLC1A4, SLC15A2, SLC6A15, CALM1, SLC7A2, SLC38A2, SLC12A6, SLC26A2*	*SLC3A1, SLC12A2, SLC6A20, SLC9A1, SLC20A1, SLC26A6, SLC17A5, SLC38A1*
Glycosaminoglycan metabolism	*CD44, HS3ST3B1, B4GALT1, IDS, ST3GAL1, HPSE, SLC26A2*	*LUM, SLC9A1, CHST6, EXT1, SLC35D2, HEXA, DCN, GUSB*
Cytosolic sulfonation of small molecules	*SLC26A2*	*BPNT1, SULT1C2, SULT1E1, SULT1A1*
Assembly of collagen fibrils and other multimeric structures	*CTSB, DST*	*ITGA6, CTSS, LAMA3, ITGB4, COL3A1, COL1A2, LOXL2, COL4A5*

For EAC, ten genes involved in the pathway *"Formation of the cornified envelope*" were all downregulated. The pathway "*transport of inorganic cations/anions and amino acids/oligopeptides*" had six downregulated and six upregulated genes. The pathway "*smooth muscle contraction*" had four downregulated genes and seven upregulated genes. The top five pathways according to the corrected p-value are shown in Table 4.16 with a list of the upregulated and downregulated genes in the pathway.

Table 4.16 Differentially expressed genes (DEGs) mapped to the enriched pathways for EAC vs normal tissue

Pathway	Upregulated genes	Downregulated genes
Formation of the cornified envelope	*DSG3, PERP, KLK12, KAZN, KLK8, SPRR3, DSC2, DSC3, CAPN1, CAPNS1*	
Transport of inorganic cations/anions and amino acids/oligopeptides	*SLC38A2, SLC15A2, SLC6A15, CALM1, SLC12A6, SLC26A2*	*SLC6A20, SLC12A2, SLC3A1, SLC20A1, SLC26A6, SLC9A1*
Smooth muscle contraction	*MYH11, ACTG2, CALM1, ANXA1*	*TPM1, TPM2, ACTA2, CALD1, MYLK, MYL9, SORBS1*
Interleukin-4 and 13 signaling	*BCL6, STAT3, HSPA8, MCL1, RORA, SOX2, ANXA1*	*COL1A2, STAT1, ZEB1*
Apoptotic cleavage of cell adhesion proteins	*DSG3*	

"Collagen chain trimerization", one of the pathways by the ESCC SetRank analysis had four upregulated genes. "*G2/M checkpoints*" pathway had one downregulated gene and 18 upregulated genes, whilst *"Integrin cell surface interactions*" had three downregulated and six upregulated genes. A list of the five most significant pathway with the corresponding up and downregulated genes are shown in Table 4.17.

The full maps for BE (model contrast), EAC, and ESCC (model contrast) are shown in the Appendix tables 4A3, 4A4 and 4A5.

Table 4.17: Differentially expressed genes (DEGs) mapped to the enriched pathways for ESCC vs normal tissue

Enriched pathway	Upregulated genes	Downregulated genes
Collagen chain trimerization		*COL11A1, COL4A5, COL1A2, COL3A1*
G2/M Checkpoints	*H2BC4*	*MCM10, GTSE1, CDK1, CHEK1, CCNB2, CCNB1, MCM6, MCM4, ATR, PSMB2, RFC3, MRE11, NSD2, RAD1, NBN, PSMB4, RPA1, YWHAH*
Integrin cell surface interactions	*JAM3, ITGA8, JAM2*	*LUM, TNC, ITGA6, CD44, THBS1, FBN1*
Formation of the cornified envelope	*CAPN1, KLK8, PCSK6, CAPNS1, DSG3, KAZN, DSC2, KLK12, SPRR3*	*DSC3*
Metabolism of ingested SeMet. Sec. MeSec into H2Se	*NNMT, CBS*	*HNMT*

4.4.12 Functional interaction networks using ReactomeFIViz application

The ReactomeFIViz application was able to identify functional interaction networks among genes in the pathways described in this study. Four of the most common pathways identified in the SetRank analysis were selected further investigation using the ReactomeFIViz app. These were *"Formation of the cornified envelope"*, *"Smooth muscle contraction"*, *"BMAL1:CLOCK,NPAS2 activates circadian gene expression",* and *"Assembly of collagen fibrils and other multimeric structures".* The gene interaction networks as well as the corresponding Reactome pathway diagrams are shown in Figures 4.12, 4.13, 4.14 and 4.15.

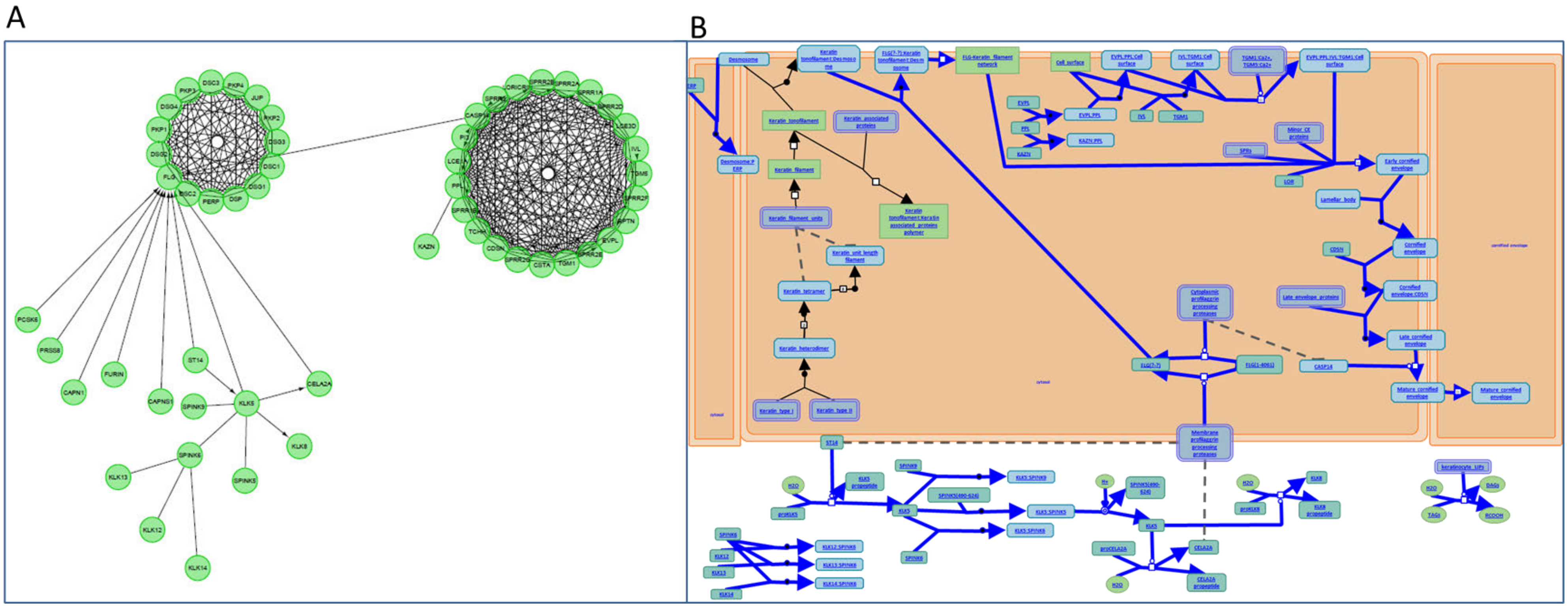


***Figure 4.12**: A. Functional interaction network among genes in the enriched pathway “Formation of the cornified envelope”. developed using the ReactomeFIViz app. B. Reactome pathway overview for “Formation of the cornified envelope.*

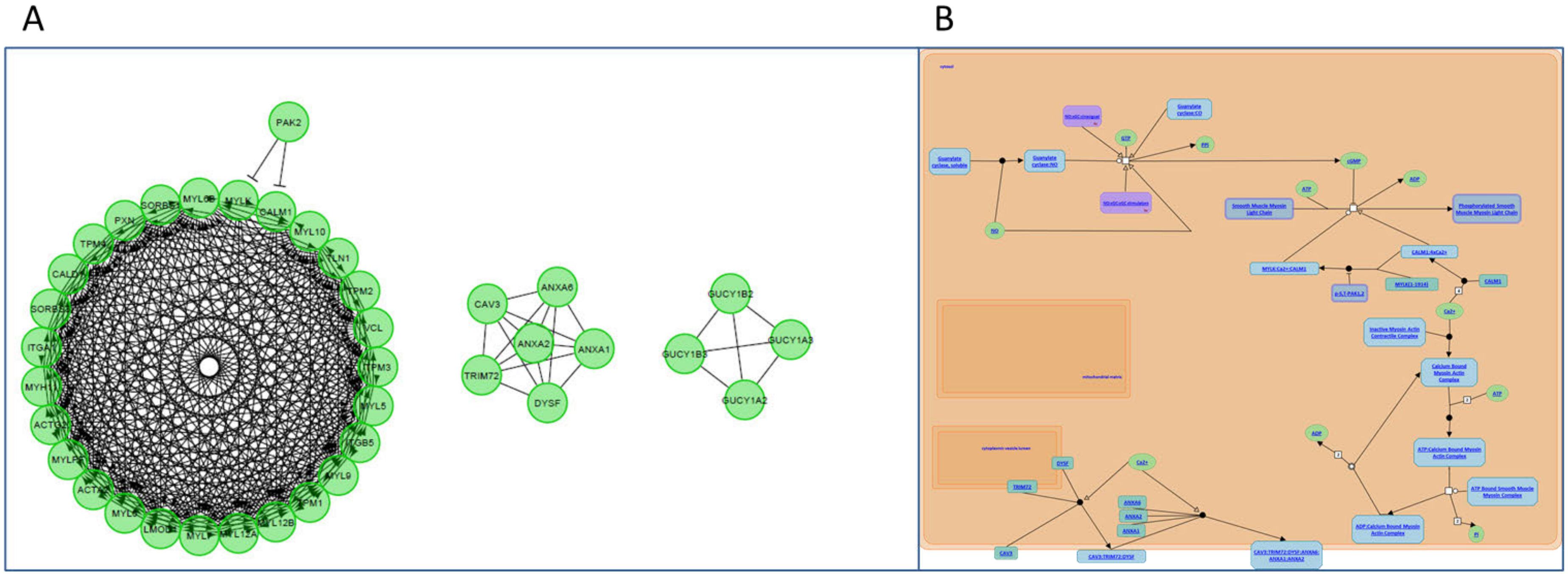


***Figure 4.13**: A. Functional interaction network among genes in the enriched pathway "Smooth muscle contraction". developed using the ReactomeFIViz app. B. Reactome pathway overview for "Smooth muscle contraction".*

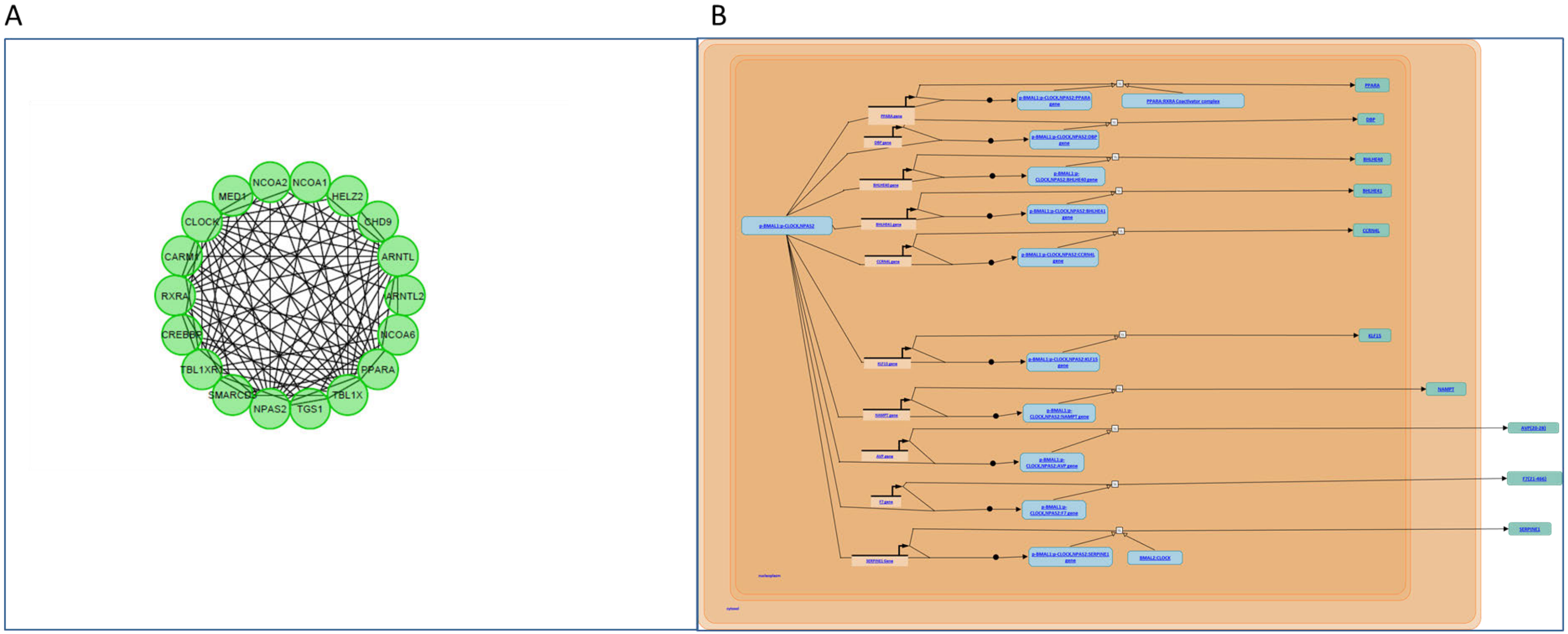

A
NCOA2
NCOA1
MED1
HELZ2
CLOCK
CHD9
CARM1
ARNTL
RXRA
ARNTL2
CREBBP
NCOA6
TBL1XR1
PPARA
SMARCD3
TBL1X
NPAS2
TGS1
B
p-BMAL1:p-CLOCK,NPAS2
PPARA gene
p-BMAL1:p-CLOCK,NPAS2:PPARA gene
PPARA:RXRA Coactivator complex
PPARA
DBP gene
p-BMAL1:p-CLOCK,NPAS2:DBP gene
DBP
BHLHE40 gene
p-BMAL1:p-CLOCK,NPAS2:BHLHE40 gene
BHLHE40
BHLHE41 gene
p-BMAL1:p-CLOCK,NPAS2:BHLHE41 gene
BHLHE41
CCRN4L gene
p-BMAL1:p-CLOCK,NPAS2:CCRN4L gene
CCRN4L
KLF15 gene
p-BMAL1:p-CLOCK,NPAS2:KLF15 gene
KLF15
NAMPT gene
p-BMAL1:p-CLOCK,NPAS2:NAMPT gene
NAMPT
AVP gene
p-BMAL1:p-CLOCK,NPAS2:AVP gene
AVP(20-28)
F7 gene
p-BMAL1:p-CLOCK,NPAS2:F7 gene
F7(21-466)
SERPINE1 Gene
p-BMAL1:p-CLOCK,NPAS2:SERPINE1 gene
BMAL2:CLOCK
SERPINE1
nucleoplasm
cytosol

***Figure 4.14**: A. Functional interaction network among genes in the enriched pathway “BMAL1:CLOCK.NPAS2 activates circadian gene expression”. developed using the ReactomeFIViz app. B. Reactome pathway overview for “BMAL1:CLOCK.NPAS2 activates*

circadian gene expression".

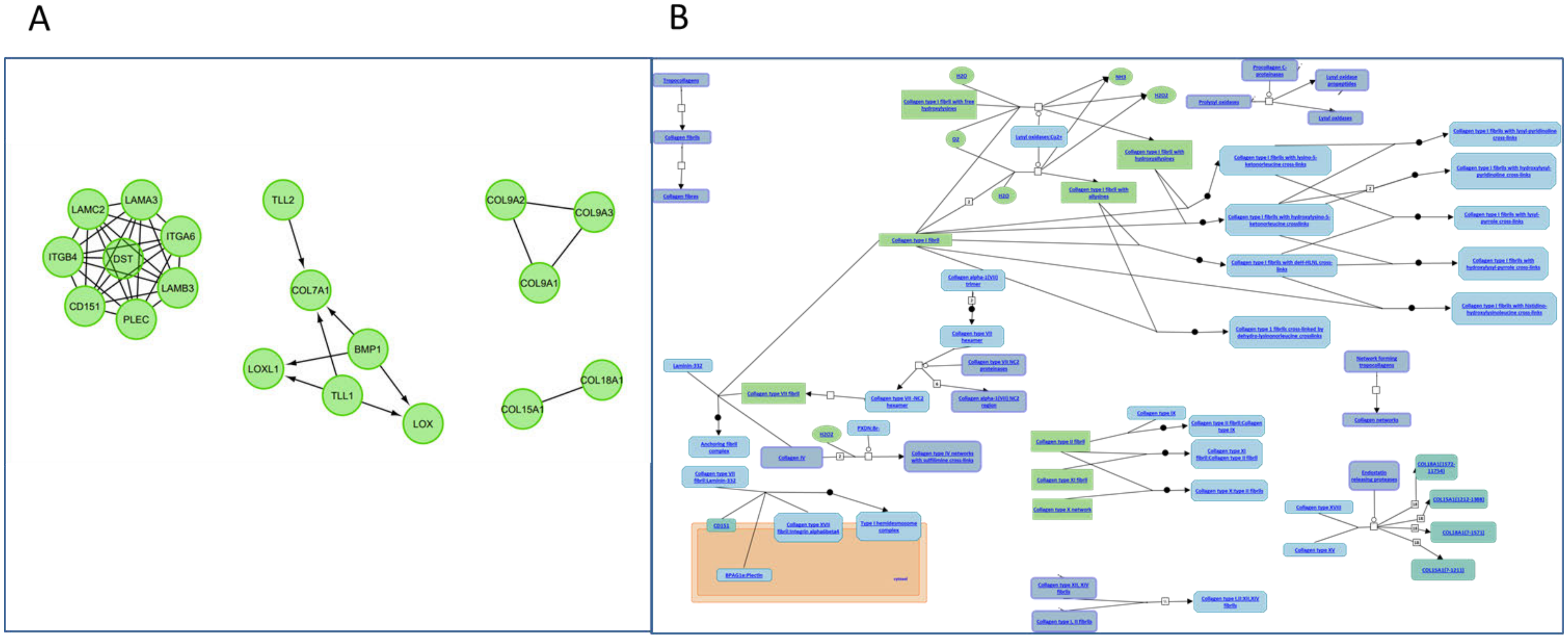

A
B
LAMC2
LAMA3
ITGA6
ITGB4
DST
LAMB3
CD151
PLEC
TLL2
COL7A1
BMP1
LOXL1
TLL1
LOX
COL9A2
COL9A3
COL9A1
COL18A1
COL15A1
Tropocollagens
Collagen fibrils
Collagen fibres
H2O
Collagen type I fibril with free hydroxylysines
O2
Lysyl oxidases:Cu2+
H2O
NH3
H2O2
Collagen type I fibril with hydroxyallysines
Collagen type I fibril with allysines
Procollagen C-proteinases
Prolysyl oxidases
Lysyl oxidase propeptides
Lysyl oxidases
Collagen type I fibril
Collagen alpha-1(VII) trimer
Collagen type VII hexamer
Collagen type VII NC2 proteinases
Collagen type VII -NC2 hexamer
Collagen alpha-1(VII) NC2 region
Collagen type I fibrils with lysino-5-ketonorleucine cross-links
Collagen type I fibrils with hydroxylysino-5-ketonorleucine crosslinks
Collagen type I fibrils with deH-HLNL cross-links
Collagen type 1 fibrils cross-linked by dehydro-lysinonorleucine crosslinks
Collagen type I fibrils with lysyl-pyridinoline cross-links
Collagen type I fibrils with hydroxylysyl-pyridinoline cross-links
Collagen type I fibrils with lysyl-pyrrole cross-links
Collagen type I fibrils with hydroxylysyl-pyrrole cross-links
Collagen type I fibrils with histidino-hydroxylysinoleucine cross-links
Laminin-332
Collagen type VII fibril
Anchoring fibril complex
Collagen type VII fibril:Laminin-332
H2O2
PXDN:Br-
Collagen IV
Collagen type IV networks with sulfilimine cross-links
CD151
Collagen type XVII fibril:integrin alpha6beta4
Type I hemidesmosome complex
BPAG1e:Plectin
cytosol
Collagen type IX
Collagen type II fibril
Collagen type II fibril:Collagen type IX
Collagen type XI fibril:Collagen type II fibril
Collagen type XI fibril
Collagen type X network
Collagen type X:type II fibrils
Collagen type XII, XIV fibrils
Collagen type I, II fibrils
Collagen type I,II:XII,XIV fibrils
Network forming tropocollagens
Collagen networks
Endostatin releasing proteases
Collagen type XVIII
Collagen type XV
COL18A1(1572-1754)
COL15A1(1212-1388)
COL18A1(?-1571)
COL15A1(?-1211)

***Figure 4.15**: A. Functional interaction network among genes in the enriched pathway "Assembly of collagen fibrils and other multimeric structures". developed using the ReactomeFIViz app. B. Reactome pathway overview for "Assembly of collagen fibrils and other multimeric structures"*

4.4.13 Substudy: Reliability of subsampling

Overall, most of the pathways from the nine repeats showed consistency in the p values. The top 21 pathways showed consistently stable p and z values with low variance between the nine repeats. The pathways are ordered according to statistical significance using the mean p value. These pathways are shown in Table 4.18, with the corresponding z and p mean values, as well as the variance. The three topmost significant pathways included "*Formation of the cornified envelope*" (p=5.33E-07), *"Smooth Muscle Contraction"* (p=1.00E-03)*, and "Type I hemidesmosome assembly"* (p=6.00E-03)*.* The variance in the top 21 enriched pathways ranged from 0.00059 to 0.14.

Table 4.18: Subsampling sub study P and Z estimates from nine repeats

Description	Mean Z value	Mean P value	Z Variance	Consistency (Mean Z/Z variance)
Formation of the cornified envelope	-4.88	5.34E-07	2.45E-02	-199
Smooth Muscle Contraction	-3.70	1.09E-04	8.55E-03	-433
Type I hemidesmosome assembly	-3.24	5.92E-04	3.60E-03	-900
Cytosolic sulfonation of small molecules	-3.12	8.90E-04	2.79E-02	-112
Assembly of collagen fibrils and other multimeric structures	-2.97	1.50E-03	5.52E-03	-537
TP53 regulates transcription of several additional cell death genes whose specific roles in p53-dependent apoptosis remain uncertain	-2.76	2.88E-03	1.33E-03	-2078
Neutrophil degranulation	-2.68	3.64E-03	4.73E-02	-57
O-linked glycosylation of mucins	-2.60	4.65E-03	1.47E-01	-18
BMAL1:CLOCK.NPAS2 activates circadian gene expression	-2.46	6.97E-03	8.90E-02	-28

Retrograde transport at the Trans-Golgi-Network	-2.41	7.91E-03	8.72E-03	-277
Aflatoxin activation and detoxification	-2.40	8.28E-03	6.71E-03	-357
Tight junction interactions	-2.36	9.09E-03	6.90E-03	-342
Interleukin-4 and 13 signalling	-2.32	1.02E-02	1.04E-02	-224
Surfactant metabolism	-2.29	1.10E-02	1.82E-02	-126
Ketone body metabolism	-2.20	1.38E-02	1.58E-03	-1395
Interaction between PHLDA1 and AURKA	-2.11	1.75E-02	5.87E-04	-3591
RHO GTPases regulate CFTR trafficking	-2.07	1.94E-02	6.71E-04	-3081
Antigen activates B Cell Receptor (BCR) leading to generation of second messengers	-2.06	1.96E-02	6.32E-04	-3264
Factors involved in megakaryocyte development and platelet production	-2.03	2.13E-02	2.28E-02	-89
Hormone-sensitive lipase (HSL)-mediated triacylglycerol hydrolysis	-1.98	2.38E-02	4.93E-03	-402

Comparisons between the pathways identified in the initial BE subsampling analysis and the BE subsampling repeats were done. The initial BE subsampling identified 16 pathways, of which the top 9 enriched pathways were also identified as the nine topmost enriched pathways in the subsampling sub study. The remaining seven pathways from the initial subsampling analysis were present in the top 45 pathways identified in the subsampling repeats. The pathways in the initial analysis and the subsampling repeats were ordered according to p-values, smallest to largest. The p-values in both analyses were comparable (Table 4.19). The differences in p-values between the two analyses ranged from 3.58E-07 to 1.99E-02. The complete data on the subsampling sub-study is shown in Appendix Table 4A6.

Table 4.19: Comparison of p values from the initial subsampling analysis and the subsampling repeats

Description	P-value[1]	Mean P-value[2]
Formation of the cornified envelope	1.76E-07	5.34E-07
Transport of inorganic cations/anions and amino acids/oligopeptides	2.20E-05	2.51E-02
Smooth Muscle Contraction	6.30E-05	1.09E-04
Cytosolic sulfonation of small molecules	6.11E-04	8.90E-04
Type I hemidesmosome assembly	9.90E-04	5.92E-04
Assembly of collagen fibrils and other multimeric structures	1.19E-03	1.50E-03
Neutrophil degranulation	1.23E-03	3.64E-03
BMAL1:CLOCK.NPAS2 activates circadian gene expression	3.70E-03	6.97E-03
TP53 regulates transcription of several additional cell death genes whose specific roles in p53-dependent apoptosis remain uncertain	3.85E-03	2.88E-03
Nicotinate metabolism	6.44E-03	4.36E-02
Tight junction interactions	6.97E-03	9.09E-03
Defective GALNT3 causes familial hyperphosphatemic tumoral calcinosis (HFTC)	6.98E-03	2.69E-02
PPARA activates gene expression	7.11E-03	2.60E-02
Metabolism of Angiotensinogen to Angiotensins	8.71E-03	2.54E-02
Glucose metabolism	8.77E-03	7.14E-02
O-linked glycosylation of mucins	9.63E-03	4.65E-03

[1] P value from the initial subsampling analysis [2] Mean p value computed from the nine subsampling repeats

4.5 Discussion

4.5.1 Overview

In this study, we analysed altered mRNA expression in EC using 18 datasets from the public repository, GEO datasets. We identified 1,107 upregulated genes and 1,537 downregulated genes for ESCC and EAC, and the precancerous tumour BE. A total of 97 enriched biological pathways were identified for ESCC, EAC and BE, using the novel GSEA, SetRank. The pathway "*Formation of the cornified envelope*" emerged as a key and common pathway explaining the pathobiology of ESCC, EAC and BE. Pathways involved in the ECM were the most common, followed by pathways involved in cell membrane regulation, cell cycle regulation and detoxification. We also identified key and interesting pathways not previously interpreted in literature.

4.5.2 Summary of results: DEGs

We performed a meta-analysis of microarray data (ESCC, EAC, BE) to identify DEGs. Analysis of squamous dysplasia could not be done as there were not enough samples for analysis. The meta-analysis, compared to individual analysis, provided greater power to detect high confidence DEGs and subsequently gene sets. It also reduced the number of false positives normally identified in individual studies. The total numbers of DEGs identified for BE (model pairwise contrast) were 1,945, for EAC were 1,130, and for ESCC (model pairwise contrast) were 1,832. The top DEGs for BE included the downregulated, *SCEL* (precursor of the cornified envelope, linked cancer development), and *SPRR3* (differentiation and cornification of keratinocytes), and the upregulated genes *AGR2* (tp53 inhibitor, cell metastasis, and linked to cancer progression), and *SULT1C2* (catalysis of endogenous and xenobiotic compounds). The top DEGs for EAC comprised of the downregulated genes, *DSG3* (cell-cell adhesion, expressed in squamous epithelium), *DSC3* (cell-cell adhesion, cancer biomarker), and the upregulated genes, *MUCIN13* (cell surface glycoproteins), and *TMC5* (cell cycle regulation). The top DEGs for ESCC included the downregulated genes EMP1 (cell migration and proliferation and implicated in cancer), *PPP1R3C* (protein phosphorylation), and the upregulated genes *COL11A1* and *COL1A2* (encodes for collagen), and *ECT2* (oncoprotein associated with cell division and growth and associated with cancer).

The top upregulated gene which overlapped between EAC and ESCC, *COL1A2,* is involved in type I collagen synthesis. In a study by Fang et al (71), on ESCC cell lines the authors reported that type I collagen expression was significantly associated with survival and cancer cells differentiation. In our study, *COL1A2* was also upregulated in BE and EAC and ESCC, pointing to similar collagen-mediated effects on tumorigenesis in the three tissue types. Hypermethylation of *COL1A2*, which has been associated with upregulation of gene expression (72), has been reported in some cancers including colorectal, bladder, breast cancers, as well as melanoma, neuroblastoma, medulloblastoma, and hepatoma.(73) In another study that re-analysed microarray datasets, *COL1A2* was reported to have expression significantly higher than normal tissue.(74) Three of the most upregulated genes which overlapped between BE and EAC have the following functions; *AGR2* (p53 inhibitor and involved in cell migration, differentiation and metastasis), *SULT1C2 (*catalysis of endogenous and xenobiotic compounds)*,* and *MUC13 (*cell surface glycoproteins). The gene *AGR2* has been implicated in promoting tumour growth in EAC (32, 75, 76) and the development of BE (32). *MUC13* is widely reported to be overexpressed in epithelial tissue tumours, as well as in ESCC (77), but there is currently no evidence in the literature regarding its association to EAC and BE.

The top downregulated genes that showed overlap among BE, EAC, and ESCC included *SCEL, SPRR3*, and *EMP1*. There is not much in the literature regarding the role of SCEL in esophageal tumorigenesis. Gene expression studies have shown overexpression of *SCEL* in ESCC.(78) *SPRR3* has been reported to be downregulated in EAC and ESCC in several studies and is strongly associated with carcinogenesis.(79, 80) Detection of *EMP1* expression using RT-PCR in EC cells showed downregulation in cancer cells compared to normal tissue.(81) *ACP3* and *MXD1* were the two most downregulated genes which overlapped between EAC and ESCC, and there is little information in the literature regarding their role in EC carcinogenesis. The most downregulated genes between BE and EAC were *DSC3, DSG3*, and *HSPB8*. The role and expression of *DSC3* and *DSG3* in cancer is still contested in literature, and there is little evidence. *DSC3* was reported to be downregulated in BE and EAC cell lines and tissue samples using qRT-PCR.(82) There is a lack of information on the role of *DSG3* in BE and EAC. *DSG3* is reported to be downregulated in oral squamous cell carcinoma and overexpressed in the skin and head and neck cancers.(83) In our study, DSG3 was downregulated in ESCC, however, in a study done by Fang et al (83) DSG3 was overexpressed in 74% of the ESCC tumours compared to normal tissue, whilst 25% showed downregulation. Additionally, in one of the few studies performed on an African population (Malawi), DSG3 and DSC3 were upregulated in ESCC tissue compared to normal tissue.(84) More studies are needed to elucidate the role of DSG3 expression in esophageal tumorigenesis. Studies have shown that desmosome deficiency plays a significant role in tumorigenesis and the progression of various epithelial cancers.(85) *HSPB8* has been reported to be downregulated in a study by Nancarrow et al (32) assessing BE and EAC tissue, as well as in a study by Yang et al in 2019 (86) which re-analysed ESCC GEO datasets. In our study, *HSPB8* was downregulated. Overall, the differential expression shown in our analysis corroborated the evidence in the literature. There are some significantly DEGs, identified in our analysis, which require further investigation regarding their role in BE, EAC and, ESCC development and progression. These include *SCEL, DSG3, ACP3, MXD1,* and *MUC13.*

Some additional genes have been reported to be associated with EC in the literature. In a previous study that we undertook, a systematic review on the ESCC genetic variants reported in African populations, we identified several genes associated with ESCC development.(8) We assumed that some of these genes would be differentially expressed in our present study. We checked for these genes in our dataset to ascertain if they were differentially expressed. The genes which showed differential expression in our gene expression dataset, which were also present in our systematic review include *ALDH2*, *ADH1B*, *TP53*, *CASP8*, *RUNX1*, *CYP3A5*, *MTHFR, AR, XBP1, GSTT1, FBXW7, JAG1, PIK3CA.(8) ALDH2, ADH1B, TP53, CASP8, RUNX1, CYP3A5, AR, GSTT1, MTHFR, XBP1* had germline variation, whilst AR, FBXW7, JAG1, PIK3CA, and TP53 genes has somatic variation. *ADH1B* (ethanol metabolism), in our study, was downregulated in ESCC and EAC and upregulated in BE. *ADH1B* is widely reported to be associated with ESCC due to its role in modulating alcohol oxidising capabilities. This links to alcohol consumption, which is one of the main risk factors of ESCC. *TP53AIP1* (mediating p53-dependent apoptosis) gene was downregulated in BE, EAC and ESCC. The gene *TP53* is one of the major tumour suppressor genes involved in the inhibition of tumour growth. EC is reported to have a *TP53* mutation rate of over 50%, one of the highest among other cancer types, which subsequently results in decreased mRNA expression.(87) It is important to note that *TP53* is downregulated in non-truncating cancers.(87) The downregulation of *TP53AIP1* in BE may indicate that the patients whose BE samples were analysed here are at a high risk of progressing into EAC. The downregulation of *TP53AIP1* in our analysis, therefore, corroborates the evidence in the literature. Another tumour suppressor gene downregulated in our analysis in BE, EAC and ESCC was *FBXW7.*

The p53 protein regulates the expression of other genes, repressing them transcriptionally, present in our study and upregulated (EAC and ESCC) is *CDK1* (cell cycle regulation). *CDK1* dysregulation has been linked to disorders in cell differentiation and cell cycle, resulting in abnormal cell differentiation and malignant tumour development.(88) It is overexpressed in other tumours including breast, cervical, gastric and oral cancers.(88) It is also reported to be upregulated in both EAC and ESCC.(88, 89) *CASP8 and FADD like apoptosis regulator* also known as *CFLAR* (cell apoptosis, similar to CASP8) was downregulated in BE, EAC and ESCC. *RUNX1*(development of hematopoiesis) was upregulated in BE and EAC and downregulated in ESCC. *CYP3A5* (drug metabolism synthesis of cholesterol, and steroids), was upregulated in BE and downregulated in EAC and ESCC. The gene *AR* (regulation of androgen binding on androgen receptor) and was downregulated in only ESCC and was not differentially expressed in BE and EAC. *XBP1* (regulation of genes) was upregulated in BE and EAC only. There is little evidence in the literature on *XBP1* expression and BE and EAC, most of the reports are on ESCC where it is overexpressed. *GSTT1* (conjugation of reduced glutathione to hydrophobic electrophiles) was downregulated in BE, EAC, and ESCC. One of the major drivers of ESCC is exposure to carcinogenic and mutagenic factors, therefore the downregulation of *GSTT1*, which plays a role in the detoxification of carcinogens, results in increased susceptibility to developing ESCC. The gene *JAG1* (involved in hematopoiesis) was upregulated in ESCC and downregulated in BE and EAC. *PIK3CA* (involved in multiple signalling pathways and oncogenic) was present in our dataset, upregulated in ESCC. *PIK3CA* expression and mutational signature were assessed by Wang et al (90) using immunohistochemistry and PCR, respectively, on ESCC tumour samples. The study reported *PIK3CA* overexpression and identified somatic point mutations associated with ESCC. *PIK3CA* is one of the most significantly mutated genes associated with ESCC, with a mutation frequency between 2.2% and 21%, and is linked to local recurrence and poor survival.(90)

We could not find any datasets on African populations in our screening and selection process. This points to the lack of genomic analysis being on the African continent. Microarray analysis is a powerful tool and also cheaper than other genome-wide technologies, and is, therefore, a good option in low-resource settings. Considering the impact that gene expression studies have in understanding the pathobiology of EC, including its diagnosis, progression, therapy, treatments, and the high burden of EC in the African EC corridor, more microarray studies are needed on African populations. Our initial plan for the PhD included a comprehensive study of South African ESCC patients residing in Eastern Cape Province, but due to the COVID-19 pandemic, participant recruitment was not possible, and the project had to be postponed. The study was designed to include collection of tissue samples from the esophagus which would have been used in genetic analysis.

4.5.3 Summary of results: Enriched pathways

Overall, we identified 97 pathways that were enriched biological pathways. The majority of the pathways are involved in ECM. These included *"Formation of the cornified envelope", "Smooth muscle contraction", "Assembly of collagen fibrils and other multimeric structures", "Type I hemidesmosome assembly", "Collagen chain trimerization", "Extracellular matrix organization", "Integrin cell surface interactions", "ECM proteoglycans",* and *"Tight junction interactions".* The ECM is a non-cellular component of tissues and organs involved in the physical scaffolding of the cells and also involved in biological processes such as tissue morphogenesis, cell differentiation and homeostasis.(91) Its involvement in crucial biochemical and biomechanical processes is what makes abnormalities in the ECM result in the development and progression of several diseases, such as cancer.(91) The ECM can modulate key processes which drive carcinogenesis, which include: cell survival and proliferation, apoptosis, angiogenesis, and migration, among others.(91) Additionally, the tumor and its microenvironment establish a feedback loop mechanism that facilitated malignant behavior in cells due to its effect on ECM stiffness, adhesion, remodeling, and other biological aspects.(92) The proteins which make up the ECM include proteoglycans and fibrous proteins (collagens, elastins, fibronectins and laminins).(93) The ECM is dynamic and constantly being reconstructed, and modulated, mainly by matrix metalloproteinases (MMP) and growth factors. In the past couple of years, the ECM has been reported to play a role in EC development and progression.(91, 92) The role of the ECM in EC carcinogenesis through activation of signaling pathways in the feedback loop mechanism is shown in Figure 4.16. ECM proteins like type I collagen are reported to be upregulated in EC (91), and this is corroborated by our results where *COL11A1, COL1A2, FNDC3B, COL3A1* and *MMP11* were upregulated. These genes were dysregulated in the *"Extracellular matrix organization"* pathway in our study. This pathway has been reported in several studies as contributing to EC development.(94, 95)

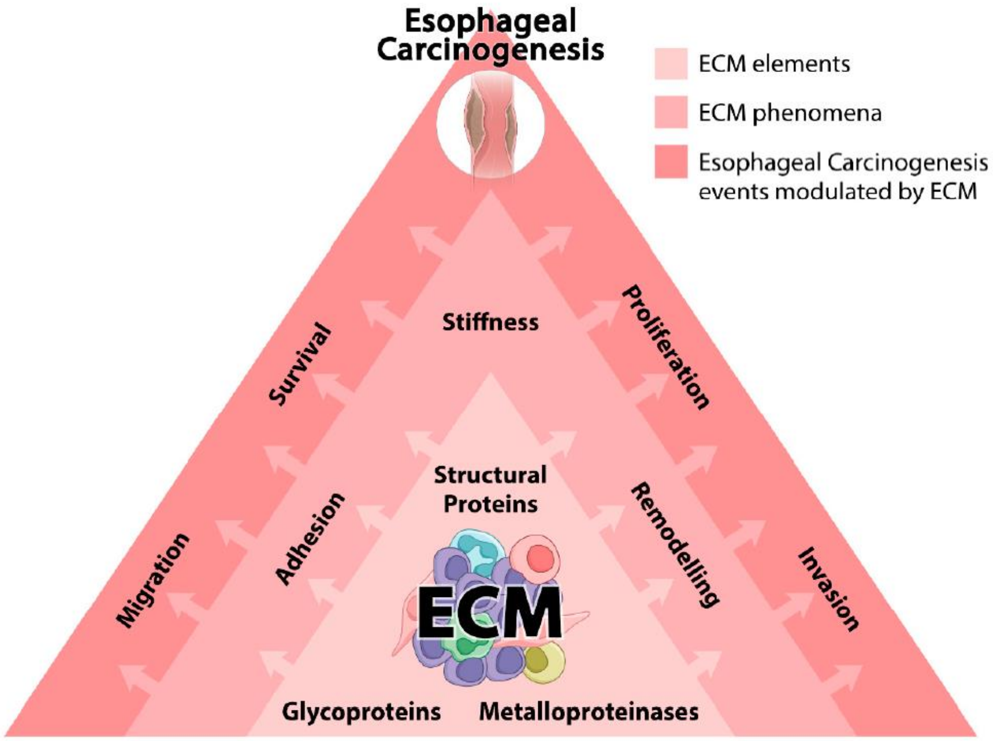


Figure 4.16: *The role of the ECM in EC carcinogenesis. The inner triangle shows the ECM proteins involved in ECM carcinogenesis. The middle triangle shows the effect of ECM protein alteration. The outer triangle shows the major cellular events facilitated by the mechanical and biochemical alterations in the ECM. which may play a role in EC carcinogenesis. Figure adapted from Palumbo* et al 2020 (92) using the MDPI open access licence MDPI | Open Access Information

The most common pathway in our study was *"Formation of the cornified envelope",* which was reported in EAC, all the BE pairwise contrasts, and all but one of the ESCC pairwise contrasts. For the EAC and the BE model pairwise contrasts, it was the most significantly enriched pathway. The majority of the genes in this pathway were downregulated. There is a lack of information in the literature regarding the role of the cornified envelope in EC carcinogenesis. *SPRR3*, a precursor protein of the cornified envelope was reported to be downregulated using Immunohistochemistry (IHC) and tissue microarray in a study by Zhang et al (96). The authors also determined that *SPRR3* was involved in anti-tumor activity. SPRR proteins have been reported to have anti-oxidative properties in the cornified envelope.(97) The role of antioxidant enzymes and SPRR proteins in the cornified envelope and EC warrants investigation. In our study, *SPRR3* was downregulated in both EAC and ESCC. The formation of the cornified envelope is the final step in the keratinization process of the epidermis. Whilst this process is commonly reported in skin and hair follicles, it also occurs in the esophagus and serves as a barrier from environmental exposures. Our study, therefore, suggests that the dysregulation of genes associated with the formation of cornified envelope impact the development of BE, EAC and ESCC, by modulating the proteins involved in the keratinization process. Other studies have listed pathways related to keratinization in ESCC analysis, which include *"Epidermis development*", *"Keratinocyte differentiation*", and *"Keratinization*".(95) However, these pathways are not interpreted in literature. More studies are needed to elucidate the role of the cornified envelope in EC development and progression, and how dysregulation of its components drive carcinogenesis.

Another enriched pathway that was common in BE, EAC and ESCC was *"Smooth muscle contraction*". Peristalsis is dependent on smooth muscle contraction of the esophagus. The genes dysregulated in this pathway in our study are involved in actin (*ACTA2, ACTG2*)(98), myosin (*MYL9, TPM2, TPM1*)(99) and calcium (*CALM1*)(100) control. Aberrant expression of Actin has been reported to be an early biomarker of cancer, through supporting oncogenic process such as cell proliferation.(98)The calcium cation, which is involved in many physiological processes in the body, is also involved in smooth muscle contraction. Interpretation of how smooth muscle contraction plays a role in EC is scarce in the literature. In a study that assessed the calcium regulation and its role in smooth muscle contraction of the esophagus in mice, reduced calcium levels resulted in reduced smooth muscle contractions.(101) The reduced smooth muscle contractions can have result in gastro-esophageal reflux, which is a risk factor for BE and subsequently EAC. Achalasia, which is also a risk factor for EC, occurs when smooth muscle fibers do not relax, is also regulated by calcium.(101) Smooth muscle disorders therefore can result in EC. Whilst it is clear that smooth muscle disorders can result in esophageal diseases, it is unclear if increasing intracellular calcium will reduce the prevalence of reflux, BE, and EAC. In a meta-analysis done on the effect of calcium intake and EC, dietary calcium intake was reported to have a protective effect on ESCC risk, in a Chinese population.(102)

The second most common grouping of the pathways was that of cell membrane regulation. These included: *"Cytosolic sulfonation of small molecules", "Transport of* *inorganic cations/anions and amino acids/oligopeptides", "Neutrophil degranulation", "Metabolism of Angiotensinogen to Angiotensins", "Tight junction interactions",* and *"RAB geranylgeranylation".* The pathway "*Cytosolic sulfonation of small molecules"* involves sulfonation of proteoglycans and other molecules making them soluble. The genes dysregulated in this pathway in our study include sulfotransferases which are involved in the conjugation of hormones, neurotransmitters, drugs, and xenobiotic compounds which may be toxic to the body. This pathway was identified in the BE analysis. *"Transport of inorganic cations/anions and amino acids/oligopeptides"* was one of the most significantly enriched pathways reported in BE (p = 7.50E-05) and EAC (p = 5.30E-04) analysis. This suggests that the dysregulation of transport of proteins and oligopeptides plays a role in the development of BE and EAC. The pathway "*Neutrophil degranulation*" was also reported in a study that re-assessed GEO datasets of BE.(103) The pathway *"Tight junction interactions"* plays a critical role in acting as a membrane barrier and regulating cell proliferation and morphogenesis, although its role in EC carcinogenesis is unclear and warrants further investigation.

The *"RAB geranylgeranylation"* pathway was identified in the BE analysis in our study (p = 5.00E-03). It is involved in the post-translational modification of proteins and lipids (an important cellular regulatory process), which regulate exocytic and endocytic pathways.(104) The genes dysregulated in this pathway in our BE analysis are *RAB* genes (*RAB27A, RAB7A, RAB20, RAB8B, RAB21, RAB29, RAB11A, RAB27B, RAB3B, RAB5A),* which are a part of the RAS oncogene family. (105) The RAB genes regulate membrane trafficking as well as cell growth, signalling and survival and are reported to be dysregulated in cancer. (105) However, this pathway has not been described for EC or BE in the literature. Two RAB genes, *RAB25* (106) and *RAB23* (107) (also dysregulated in our ESCC analysis) have been reported to be associated with ESCC development in the literature. Overall, the role of the *RAB* genes in cancer is still understudied. We hypothesize that the biological processes associated with this pathway drive formation of malignant tumours from BE. Again, this pathway warrants further investigation.

Several pathways in or analysis are involved in cell cycle regulation, and they include, "*TP53 regulates transcription of several additional cell death genes whose specific roles in p53-dependent apoptosis remain uncertain", "Cyclin B2 mediated events*", *"PI3K/AKT activation",* and, "*BMAL1:CLOCK,NPAS2 activates circadian gene expression*"*.* The pathway *"TP53 regulates transcription of several additional cell death genes whose specific roles in p53-dependent apoptosis remain uncertain"* was identified in our BE analysis. TP53 facilitates tumour suppression through regulation of transcription in several genes mainly involved in cell apoptosis, as well as others that inhibit apoptosis, leaving the cells an opportunity to repair the damage. Some of the genes involved in this pathway reported in our study include *PERP* (effector of the TP53 apoptotic pathway), *BCL6* (transcriptional repressor and regulator of germinal centers), and *BCL2L14* (apoptotic facilitation). The precise mechanisms of these pro-apoptotic genes, particularly in EC, remain unclear. Whilst P53 pathways have been described in other studies (108), this specific pathway has not been described in the literature, in relation to EC. Our study posits that the dysregulation of TP53 mediated cell death genes drives oncogenesis from BE and EAC. "*Cyclin B2 mediated events"* is a pathway involved in cell cycle regulation and was present in our ESCC analysis. The presence of this pathway in ESCC suggests that dysregulation of the pathway "*Cyclin B2 mediated events*" pathway drives carcinogenesis in ESCC. The genes dysregulated the pathway "*Antagonism of activin by follistatin*" were upregulated and they included *INHBA, FST*. This pathway was dysregulated in all of the ESCC pairwise contrasts, and the model contrast had a p value of 0.006. The response of normal and tumour cells to activin is mixed, but overall, it involved in a number of oncogenic processes including cell growth, death and migration, angiogenesis, inflammation, drug resistance, and bone loss.(109)

The pathway "*BMAL1:CLOCK,NPAS2 activates circadian gene expression*" was identified for ESCC (p=3.00E-02) and BE (p=6.50E-05). Circadian disruption is known as a cancer risk factor and classified as a carcinogen by the Word Health Organisation.(110) Circadian control is regulated by the suprachiasmatic nucleus in the hypothalamus of the brain.(111) The circadian clock genes which regulate the sleep and wake cycles also regulate normal cells and cancer cells' division, and proliferation.(110). Specifically, the circadian clock influences cancer development and progression through cell-cycle control, apoptosis, metabolism, and DNA damage response Disruptions of normal circadian rhythms associated with dysregulation in the clock genes may lead to cancer development.(110) Cancers reported to be associated with the disrupted circadian rhythm in epidemiological studies include breast, endometrial, colorectal, and ovarian cancer.(110, 112) Additionally, perturbation of the circadian clock is also reported to influence cancer therapy and survival. Interestingly, overall cancer incidence is reported to be low in in visually impaired individuals who are not sensitive to ambient light changes, and hence their daily circadian cycles are facilitated by endogenous circadian clocks which coordinate daily physiology.(110) This pathway has not been interpreted for EC in literature, and is a novel risk factor for EAC and ESCC

We expected to find pathways involved in detoxification, which may play roles in the stress and toxic response, as well as xenobiotic metabolism. The pathways we identified included *"Ethanol oxidation", "Detoxification of Reactive Oxygen Species", "Aflatoxin activation and detoxification", "Cytosolic sulfonation of small molecules", "Fatty acid, triacylglycerol, and ketone body metabolism"* and *"Nicotinate metabolism".* The pathway *"Ethanol oxidation"* (p=3.00E-02) is involved in alcohol metabolism and was identified in our ESCC analysis*.* The gene dysregulated in this pathway is *ADHB1*, which is involved in the oxidation of ethanol into acetaldehyde in the liver.(113) Polymorphisms of *ADHB* genes have consistently been shown to be associated with EC, together with *ALDH* genes, which is involved in the metabolisms of acetaldehyde to acetate.(113, 114) Acetaldehyde is a carcinogen and associated with EC development.(113) The pathway *"Cytosolic sulfonation of small molecules"* (p=1.80E-02) was identified in all the BE pairwise contrasts. The genes dysregulated in this pathway are sulfotransferases which are phase II drug metabolizing enzymes involved in the biotransformation of xenobiotic and endogenous compounds.(115)

The identification of pathways involved in detoxification is noteworthy due to their involvement in the detoxification of environmental carcinogens. Environmental exposure to carcinogens such as alcohol, tobacco and PAHs is one of the biggest drivers ESCC development worldwide. The role of environmental carcinogens is also stated in our systematic review. The identification of these pathways therefore gives biological plausibility of environmental carcinogens and EC.

An interesting pathway that was identified in our study (ESCC analysis) is *"Metabolism of ingested SeMet, Sec, MeSec into H2Se*" which involves the transformation of dietary selenium to its metabolites. Selenium is a trace element found in food, and its presence in food is dependent on selenium content in the soil from which the food is grown.(116) There is limited evidence on the role of selenium in EC, however, a few studies have shown that selenium deficiency is a risk factor for ESCC.(116-119) Selenium has anti-tumour properties which include apoptosis induction in cancer cells, angiogenesis inhibition, and anti-oxidant characteristics.(116) The genes involved in this pathway and identified in our study include *NNMT* (metabolism of drugs and xenobiotic compounds), CBS (regulation of homocysteine metabolism and involved in cellular redox status), and *HNMT* (metabolism of histamine). This pathway has not been discussed in the literature. However, our results may suggest that dysregulation of selenium metabolism could result in low levels of selenium metabolites facilitating carcinogenesis. Further work is needed to understand the role of this pathway in ESCC development.

Another interesting pathway present and significantly enriched in our BE and EAC analysis which we consider important for further inquiry is "*Defective GALNT3 causes familial hyperphosphatemic tumoral calcinosis (HFTC)",* due to the fact the *GALNT3* has been reported to be associated with some cancers including lung cancer, gastric cancer, breast cancer, colorectal cancer, hepatocellular carcinoma, and oral carcinoma.(120) However, its role in EC has not been described in the literature. *GALNT3* is involved in O-glycosylation, a process involved in the induction of tumour invasion, metastasis, and recurrence.(120) The GALNT family is linked to mucins proteins. The dysregulated genes in the pathway *"Defective GALNT3 causes familial hyperphosphatemic tumoral calcinosis (HFTC)"* in our study were *MUC4* and *MUC13*. It is important to note that the pathway "O-linked glycosylation of mucins" was enriched in our BE analysis (p=3.80E-03), and ESCC analysis (p=8.80E-03), and the dysregulated genes included the GALNT family of genes and mucin genes. We suggest that further studies look into associations between HFTC and EC, to elucidate whether the "*Defective GALNT3 causes familial hyperphosphatemic tumoral calcinosis (HFTC)"* pathway is associated with EC solely due to its link with defective *GALNT3* and the *"O-linked glycosylation of mucins",* or if there are possibly any associations between HFTC and EC.

We also identified the pathway *"Arachidonate production from DAG"* (p=6.00E-02) to be dysregulated in our ESCC analysis. This pathway is involved in the formation of prostaglandins, from arachidonate acid. Prostaglandins play an important role in inflammation, this includes redness (rubor), heat (calor), pain (dolor), and swelling (tumor).(121, 122) One of the genes dysregulated in this pathway, MGLL, has been reported to regulate tumour progression in cancer cell lines.(123) Chronic inflammation has been implicated as a risk factor for cancer.(124) Inflammatory processes are linked with hypermethylation of the promoter regions of tumour suppressor genes.(124) The is no evidence in the literature regarding the role of the pathway *"Arachidonate production from DAG"* as well as the MGLL gene. Due to the pathway's role in inflammation, we hypothesize that it may have a role to play in esophageal injury and inflammation as a risk factor for ESCC development. Esophageal injury leading to inflammation can occur through consumption of hot food and beverages, reflux, caustic ingestion, and induced vomiting. We recommend that further studies be done to elucidate the role of the pathway "*Arachidonate production from DAG*" and its dysregulated genes in ESCC development.

Overall, our results are consistent with previous studies in terms of pathways identified, our study identified pathways that need further investigation and interpretation. Importantly, we also identified pathways that have not been described in the literature in relation to EC. Lastly, our subsampling sub study confirmed and validated the reliability of the subsampling that we performed.

Strengths and limitations of the study

Our study assessed GEO datasets of microarray data. There are several strengths and limitations associated with the use of microarray data. Microarrays are an older technology, therefore there is a lot of information available on how to analyse them, and they are widely used in studies and relatively cheap on a per sample basis. There is also an ease of handling data for microarrays, as data summary is produced for each probe on the array, rather than for each DNA or RNA fragment which is characteristic of sequencing. However, microarrays require known sequences as probes and are unable to detect unknown sequences, unlike sequencing technologies. They also depend on existing knowledge of the genome sequence; this means that they can only provide information on genes that are part of the array. There can also be high background noise due to cross hybridisation.

One limitation of our study is that we did not assess clinical parameters or environmental exposures in the study sample due to limitations of data availability in the data submitted in the GEO datasets. Another limitation was the datasets were biased towards European, American and Asian populations, as there were no datasets from African populations. The majority of the ESCC cases were from Asian populations, where, similar to the African region, ESCC is prevalent. It is however not clear if these results can be generalized to African populations as the risk factors driving development may be different. Additionally, our analysis was limited to studies on the GEO datasets, other repositories exist, which we did not include as part of our inclusion criteria. There were a few squamous dysplasia samples, therefore we could not perform analysis on squamous dysplasia vs control samples. There were also limitations with the meta-analysis package that we used, RankProd, as it could not analyse all the samples at once due to computational load of a large number of samples. We resorted to breaking up the analysis into multiple pairwise contrasts, up to a sample size the package allowed.

The main strength of this study is that we combined data from GEO datasets and performed combined analysis using meta-analysis. The combined analysis increased the power of the study as well as allowed for a more comprehensive analysis. The use of a novel GSEA algorithm, SetRank, ensured that our analysis addressed issues of overlap, multiple testing, and remove false positives. We also used and validated the reliability of the subsampling that we performed on some of the datasets. The subsampling approach is an important technique in genomics, considering that sample sizes keep getting larger. Whilst development of methods and platforms that can take the computational load of large datasets is important, use and validation of techniques involved in data reduction such as subsampling is also important.

This study highlights the importance of public repositories and data sharing, without which this analysis would not have been possible. Data sharing through public repositories is important as it lowers barriers for genomic data analysis and interpretation of results for the larger scientific community. Gene expression profiles play an important role in elucidating cell functions, as well as biological and regulatory pathways. They also can significantly add to conventional clinical risk factors in the prediction of patient outcomes. Differential gene expression analysis aids in understanding disease mechanism of function, identification of new therapeutic points of intervention, and prognosis. Specific to cancer this includes identification of markers for diagnosis, clinical response to chemotherapy, recurrence, metastasis, and survival.(125)

4.6 Conclusions

Despite the significant strides made in understanding the epidemiology, risk factors and pathobiology of EC, the comprehensive and complex mechanisms involved in its pathogenesis are still unclear. The combined bioinformatic analysis of existing GEO mRNA expression datasets on EC corroborated the existing evidence in the literature and importantly, provided novel insights into the pathobiology of EC.

There is an apparent lack of gene expression studies in African populations, therefore the pathobiology of ESCC remains unclear. Gene expression studies have the capability to elucidate the pathobiology of ESCC, and give vital insights into the development, progression, metastasis, response to therapy, and survival of ESCC. Considering the high burden of ESCC in the African esophageal cancer corridor, and a lack of understanding of its genetic architecture, we recommend that more gene expression studies be prioritized in Africa.

4.7 References

1. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: a cancer journal for clinicians* (2018) 68(6):394-424.

2. Auld M, Srinath H, Jeyarajan E. Oesophageal Squamous Dysplasia. *Journal of Gastrointestinal Cancer* (2018) 49(3):385-8. doi: 10.1007/s12029-018-0122-3.

3. Huang FL, Yu SJ. Esophageal cancer: Risk factors, genetic association, and treatment. *Asian J Surg* (2018) 41(3):210-5. Epub 2016/12/18. doi: 10.1016/j.asjsur.2016.10.005. PubMed PMID: 27986415.

4. Nesteruk K, Spaander MCW, Leeuwenburgh I, Peppelenbosch MP, Fuhler GM. Achalasia and associated esophageal cancer risk: What lessons can we learn from the molecular analysis of Barrett's-associated adenocarcinoma? *Biochim Biophys Acta Rev Cancer* (2019) 1872(2):188291. Epub 2019/05/07. doi: 10.1016/j.bbcan.2019.04.007. PubMed PMID: 31059738.

5. Alaouna M, Hull R, Penny C, Dlamini Z. Esophageal cancer genetics in South Africa. *Clinical and experimental gastroenterology* (2019) 12:157-77. doi: 10.2147/CEG.S182000. PubMed PMID: 31114287.

6. He F, Liu C, Zhang R, Hao Z, Li Y, Zhang N, et al. Association between the Glutathione-S-transferase T1 null genotype and esophageal cancer susceptibility: A meta-analysis involving 11,163 subjects. *Oncotarget* (2018) 9(19):15111-21. doi: 10.18632/oncotarget.24534.

7. Kumar P, Rai V. *MTHFR C677T polymorphism and risk of esophageal cancer: An updated meta-analysis*. (2018). p. 273-84.

8. Simba H, Kuivaniemi H, Lutje V, Tromp G, Sewram V. Systematic Review of Genetic Factors in the Etiology of Esophageal Squamous Cell Carcinoma in African Populations. *Front Genet* (2019) 10:642. Epub 2019/08/21. doi: 10.3389/fgene.2019.00642. PubMed PMID: 31428123; PubMed Central PMCID: PMCPMC6687768.

9. Zhang C, Sun Q. Weighted gene co-expression network analysis of gene modules for the prognosis of esophageal cancer. *Journal of Huazhong University of Science and Technology [Medical Sciences]* (2017) 37(3):319-25. doi: 10.1007/s11596-017-1734-8.

10. Kimchi ET, Posner MC, Park JO, Darga TE, Kocherginsky M, Karrison T, et al. Progression of Barrett's metaplasia to adenocarcinoma is associated with the suppression of the transcriptional programs of epidermal differentiation. *Cancer Res* (2005) 65(8):3146-54. Epub 2005/04/19. doi: 10.1158/0008-5472.Can-04-2490. PubMed PMID: 15833844.

11. Shimada Y, Sato F, Shimizu K, Tsujimoto G, Tsukada K. cDNA microarray analysis of esophageal cancer: discoveries and prospects. *General Thoracic and Cardiovascular Surgery* (2009) 57(7):347-56. doi: 10.1007/s11748-008-0406-9.

12. Ma S, Paiboonrungruan C, Yan T, Williams KP, Major MB, Chen XL. Targeted therapy of esophageal squamous cell carcinoma: the NRF2 signaling pathway as target. *Ann N Y Acad Sci* (2018) 1434(1):164-72. Epub 2018/05/13. doi: 10.1111/nyas.13681. PubMed PMID: 29752726; PubMed Central PMCID: PMCPMC6230513.

13. Kashyap MK, Abdel-Rahman O. Expression, regulation and targeting of receptor tyrosine kinases in esophageal squamous cell carcinoma. *Mol Cancer* (2018) 17(1):54. Epub 2018/02/20. doi: 10.1186/s12943-018-0790-4. PubMed PMID: 29455652; PubMed Central PMCID: PMCPMC5817798.

14. Pennathur A, Godfrey TE, Luketich JD. The Molecular Biologic Basis of Esophageal and Gastric Cancers. *Surg Clin North Am* (2019) 99(3):403-18. Epub 2019/05/03. doi: 10.1016/j.suc.2019.02.010. PubMed PMID: 31047032.

15. Rajendra S, Sharma P. Barrett Esophagus and Intramucosal Esophageal Adenocarcinoma. *Hematol Oncol Clin North Am* (2017) 31(3):409-26. Epub 2017/05/16. doi: 10.1016/j.hoc.2017.01.003. PubMed PMID: 28501084.

16. Visser E, Franken IA, Brosens LA, Ruurda JP, van Hillegersberg R. Prognostic gene expression profiling in esophageal cancer: a systematic review. *Oncotarget* (2017) 8(3):5566-77. Epub 2016/11/17. doi: 10.18632/oncotarget.13328. PubMed PMID: 27852047; PubMed Central PMCID: PMCPMC5354930.

17. Torcivia-Rodriguez J, Dingerdissen H, Chang TC, Mazumder R. A Primer for Access to Repositories of Cancer-Related Genomic Big Data. *Methods Mol Biol* (2019) 1878:1-37. Epub 2018/11/01. doi: 10.1007/978-1-4939-8868-6_1. PubMed PMID: 30378067.

18. Patra BG, Roberts K, Wu H. A content-based dataset recommendation system for researchers-a case study on Gene Expression Omnibus (GEO) repository. *Database : the journal of biological databases and curation* (2020) 2020:1-. doi: 10.1093/database/baaa064. PubMed PMID: 33002137.

19. Heath AP, Ferretti V, Agrawal S, An M, Angelakos JC, Arya R, et al. The NCI Genomic Data Commons. *Nat Genet* (2021) 53(3):257-62. Epub 2021/02/24. doi: 10.1038/s41588-021-00791-5. PubMed PMID: 33619384.

20. Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, et al. NCBI GEO: archive for functional genomics data sets—update. *Nucleic Acids Research* (2012) 41(D1):D991-D5. doi: 10.1093/nar/gks1193.

21. Shao M, Li W, Wang S, Liu Z. Identification of key genes and pathways associated with esophageal squamous cell carcinoma development based on weighted gene correlation network analysis. *J Cancer* (2020) 11(6):1393-402. doi: 10.7150/jca.30699.

22. Zoubarev A, Hamer KM, Keshav KD, McCarthy EL, Santos JRC, Van Rossum T, et al. Gemma: a resource for the reuse, sharing and meta-analysis of expression profiling data. *Bioinformatics* (2012) 28(17):2272-3. doi: 10.1093/bioinformatics/bts430.

23. Affymetrix and Thermo Fisher Scientific. Microarray Technical Documentation (2017) [cited 2020 5 July]. Available from: http://www.affymetrix.com/support/technical/byproduct.affx?product=expressionconsole.

24. Pradervand S, Paillus A, Thomas J, Weber J, Wirapati P, Hagenbüchle O, et al. Affymetrix Whole-Transcript Human Gene 1.0 ST array is highly concordant with standard 3 ' expression arrays. *BioTechniques* (2008) 44:759-62. doi: 10.2144/000112751.

25. Thermofischer Scientific. Affymetrix Genechip Gene and Exon Array White Paper Collection USA(2007) [cited 2020 3 July]. Available from: https://assets.thermofisher.com/TFS-Assets/LSG/brochures/hugene_perf_whitepaper.pdf.

26. Robinson MD, Speed TP. A comparison of Affymetrix gene expression arrays. *BMC Bioinformatics* (2007) 8(1):449. doi: 10.1186/1471-2105-8-449.

27. Wang W, Fu S, Lin X, Zheng J, Pu J, Gu Y, et al. miR-92b-3p Functions As A Key Gene In Esophageal Squamous Cell Cancer As Determined By Co-Expression Analysis. *Onco Targets Ther* (2019) 12:8339-53. Epub 2019/11/07. doi: 10.2147/ott.S220823. PubMed PMID: 31686859; PubMed Central PMCID: PMCPMC6799829.

28. Zhang Y, Xu Y, Li Z, Zhu Y, Wen S, Wang M, et al. Identification of the key transcription factors in esophageal squamous cell carcinoma. *J Thorac Dis* (2018) 10(1):148-61. Epub 2018/03/31. doi: 10.21037/jtd.2017.12.27. PubMed PMID: 29600044; PubMed Central PMCID: PMCPMC5863113.

29. He W, Chen L, Yuan K, Zhou Q, Peng L, Han Y. Gene set enrichment analysis and meta-analysis to identify six key genes regulating and controlling the prognosis of esophageal squamous cell carcinoma. *J Thorac Dis* (2018) 10(10):5714-26. Epub 2018/12/07. doi: 10.21037/jtd.2018.09.55. PubMed PMID: 30505479; PubMed Central PMCID: PMCPMC6236198.

30. Peng D, Guo Y, Chen H, Zhao S, Washington K, Hu T, et al. Integrated molecular analysis reveals complex interactions between genomic and epigenomic alterations in esophageal adenocarcinomas. *Sci Rep* (2017) 7:40729. Epub 2017/01/20. doi: 10.1038/srep40729. PubMed PMID: 28102292; PubMed Central PMCID: PMCPMC5244375.

31. Kim SM, Park YY, Park ES, Cho JY, Izzo JG, Zhang D, et al. Prognostic biomarkers for esophageal adenocarcinoma identified by analysis of tumor transcriptome. *PLoS One* (2010) 5(11):e15074. Epub 2010/12/15. doi: 10.1371/journal.pone.0015074. PubMed PMID: 21152079; PubMed Central PMCID: PMCPMC2994829.

32. Nancarrow DJ, Clouston AD, Smithers BM, Gotley DC, Drew PA, Watson DI, et al. Whole genome expression array profiling highlights differences in mucosal defense genes in Barrett's esophagus and esophageal adenocarcinoma. *PLoS One* (2011) 6(7):e22513. Epub 2011/08/11. doi: 10.1371/journal.pone.0022513. PubMed PMID: 21829465; PubMed Central PMCID: PMCPMC3145652.

33. Stairs DB, Nakagawa H, Klein-Szanto A, Mitchell SD, Silberg DG, Tobias JW, et al. Cdx1 and c-Myc foster the initiation of transdifferentiation of the normal esophageal squamous epithelium toward Barrett's esophagus. *PLoS One* (2008) 3(10):e3534. Epub 2008/10/28. doi: 10.1371/journal.pone.0003534. PubMed PMID: 18953412; PubMed Central PMCID: PMCPMC2568822.

34. di Pietro M, Lao-Sirieix P, Boyle S, Cassidy A, Castillo D, Saadi A, et al. Evidence for a functional role of epigenetically regulated midcluster HOXB genes in the development of Barrett esophagus. *Proc Natl Acad Sci U S A* (2012) 109(23):9077-82. Epub 2012/05/19. doi: 10.1073/pnas.1116933109. PubMed PMID: 22603795; PubMed Central PMCID: PMCPMC3384195.

35. Hyland PL, Hu N, Rotunno M, Su H, Wang C, Wang L, et al. Global changes in gene expression of Barrett's esophagus compared to normal squamous esophagus and gastric cardia tissues. *PLoS One* (2014) 9(4):e93219. Epub 2014/04/10. doi: 10.1371/journal.pone.0093219. PubMed PMID: 24714516; PubMed Central PMCID: PMCPMC3979678

36. Ostrowski J, Mikula M, Karczmarski J, Rubel T, Wyrwicz LS, Bragoszewski P, et al. Molecular defense mechanisms of Barrett's metaplasia estimated by an integrative genomics. *J Mol Med (Berl)* (2007) 85(7):733-43. Epub 2007/04/07. doi: 10.1007/s00109-007-0176-3. PubMed PMID: 17415542.

37. Cummings LC, Thota PN, Willis JE, Chen Y, Cooper GS, Furey N, et al. A nonrandomized trial of vitamin D supplementation for Barrett's esophagus. *PLoS One* (2017) 12(9):e0184928. Epub 2017/09/19. doi: 10.1371/journal.pone.0184928. PubMed PMID: 28922414; PubMed Central PMCID: PMCPMC5602627.

38. Wang Q, Ma C, Kemmner W. Wdr66 is a novel marker for risk stratification and involved in epithelial-mesenchymal transition of esophageal squamous cell carcinoma. *BMC Cancer* (2013) 13:137. Epub 2013/03/22. doi: 10.1186/1471-2407-13-137. PubMed PMID: 23514407; PubMed Central PMCID: PMCPMC3610187.

39. Dai Y, Wang Q, Gonzalez Lopez A, Anders M, Malfertheiner P, Vieth M, et al. Genome-Wide Analysis of Barrett's Adenocarcinoma. A First Step Towards Identifying Patients at Risk and Developing Therapeutic Paths. *Transl Oncol* (2018) 11(1):116-24. Epub 2017/12/10. doi: 10.1016/j.tranon.2017.10.003. PubMed PMID: 29223109; PubMed Central PMCID: PMCPMC6002392.

40. Nicolau-Neto P, Da Costa NM, de Souza Santos PT, Gonzaga IM, Ferreira MA, Guaraldi S, et al. Esophageal squamous cell carcinoma transcriptome reveals the effect of FOXM1 on patient outcome through novel PIK3R3 mediated activation of PI3K signaling pathway. *Oncotarget* (2018) 9(24):16634-47. Epub 2018/04/24. doi: 10.18632/oncotarget.24621. PubMed PMID: 29682174; PubMed Central PMCID: PMCPMC5908275.

41. Couto-Vieira J, Nicolau-Neto P, Costa EP, Figueira FF, Simão TA, Okorokova-Façanha AL, et al. Multi-cancer V-ATPase molecular signatures: A distinctive balance of subunit C isoforms in esophageal carcinoma. *EBioMedicine* (2020) 51:102581. Epub 2020/01/07. doi: 10.1016/j.ebiom.2019.11.042. PubMed PMID: 31901859; PubMed Central PMCID: PMCPMC6948166.

42. Ming XY, Zhang X, Cao TT, Zhang LY, Qi JL, Kam NW, et al. RHCG Suppresses Tumorigenicity and Metastasis in Esophageal Squamous Cell Carcinoma via Inhibiting NF-κB Signaling and MMP1 Expression. *Theranostics* (2018) 8(1):185-98. Epub 2018/01/02. doi: 10.7150/thno.21383. PubMed PMID: 29290801; PubMed Central PMCID: PMCPMC5743468.

43. Lee JJ, Natsuizaka M, Ohashi S, Wong GS, Takaoka M, Michaylira CZ, et al. Hypoxia activates the cyclooxygenase-2-prostaglandin E synthase axis. *Carcinogenesis* (2010) 31(3):427-34. Epub 2010/01/01. doi: 10.1093/carcin/bgp326. PubMed PMID: 20042640; PubMed Central PMCID: PMCPMC2832548.

44. Su H, Hu N, Yang HH, Wang C, Takikita M, Wang Q-H, et al. Global gene expression profiling and validation in esophageal squamous cell carcinoma and its association with clinical phenotypes. *Clinical cancer research : an official journal of the American Association for Cancer Research* (2011) 17(9):2955-66. doi: 10.1158/1078-0432.ccr-10-2724. PubMed PMID: 21385931.

45. Li WQ, Hu N, Burton VH, Yang HH, Su H, Conway CM, et al. PLCE1 mRNA and protein expression and survival of patients with esophageal squamous cell carcinoma and gastric adenocarcinoma. *Cancer Epidemiol Biomarkers Prev* (2014) 23(8):1579-88. Epub 2014/05/29. doi: 10.1158/1055-9965.Epi-13-1329. PubMed PMID: 24867265; PubMed Central PMCID: PMCPMC4207376.

46. Hyland PL, Zhang H, Yang Q, Yang HH, Hu N, Lin SW, et al. Pathway, in silico and tissue-specific expression quantitative analyses of oesophageal squamous cell carcinoma genome-wide association studies data. *Int J Epidemiol* (2016) 45(1):206-20. Epub 2015/12/05. doi: 10.1093/ije/dyv294. PubMed PMID: 26635288; PubMed Central PMCID: PMCPMC4881832.

47. Yan W, Shih JH, Rodriguez-Canales J, Tangrea MA, Ylaya K, Hipp J, et al. Identification of unique expression signatures and therapeutic targets in esophageal squamous cell carcinoma. *BMC Res Notes* (2012) 5:73. Epub 2012/01/28. doi: 10.1186/1756-0500-5-73. PubMed PMID: 22280838; PubMed Central PMCID: PMCPMC3283499.

48. Yan W, Shih J, Rodriguez-Canales J, Tangrea MA, Player A, Diao L, et al. Three-dimensional mRNA measurements reveal minimal regional heterogeneity in esophageal squamous cell carcinoma. *The American journal of pathology* (2013) 182(2):529-39. doi: 10.1016/j.ajpath.2012.10.028. PubMed PMID: 23219752.

49. Wen J, Yang H, Liu MZ, Luo KJ, Liu H, Hu Y, et al. Gene expression analysis of pretreatment biopsies predicts the pathological response of esophageal squamous cell carcinomas to neo-chemoradiotherapy. *Ann Oncol* (2014) 25(9):1769-74. Epub 2014/06/08. doi: 10.1093/annonc/mdu201. PubMed PMID: 24907633.

50. Erkizan HV, Johnson K, Ghimbovschi S, Karkera D, Trachiotis G, Adib H, et al. African-American esophageal squamous cell carcinoma expression profile reveals dysregulation of stress response and detox networks. *BMC cancer* [Internet]. (2017 2017/06//); 17(1):[426 p.]. Available from: http://europepmc.org/abstract/MED/28629367

51. Wu B, Li C, Zhang P, Yao Q, Wu J, Han J, et al. Dissection of miRNA-miRNA interaction in esophageal squamous cell carcinoma. *PLoS One* (2013) 8(9):e73191. Epub 2013/09/17. doi: 10.1371/journal.pone.0073191. PubMed PMID: 24039884; PubMed Central PMCID: PMCPMC3764179.

52. Liu N, Li C, Huang Y, Yi Y, Bo W, Li C, et al. A functional module-based exploration between inflammation and cancer in esophagus. *Sci Rep* (2015) 5:15340. Epub 2015/10/23. doi: 10.1038/srep15340. PubMed PMID: 26489668; PubMed Central PMCID: PMCPMC4614801.

53. Lv J, Liu J, Guo L, Zhang J, Cheng Y, Chen C, et al. Bioinformatic analyses of microRNA-targeted genes and microarray-identified genes correlated with Barrett's esophagus. *Cell Cycle* (2018) 17(6):792-800. Epub 2018/02/09. doi: 10.1080/15384101.2018.1431597. PubMed PMID: 29417867; PubMed Central PMCID: PMCPMC5969547.

54. Lian X, Baranova A, Ngo J, Yu G, Cao H. UGT2B17 and miR-224 contribute to hormone dependency trends in adenocarcinoma and squamous cell carcinoma of esophagus. *Biosci Rep* (2019) 39(7). Epub 2019/06/06. doi: 10.1042/bsr20190472. PubMed PMID: 31164411; PubMed Central PMCID: PMCPMC6609598.

55. McKenzie C, D'Avino PP. Investigating cytokinesis failure as a strategy in cancer therapy. *Oncotarget* (2016) 7(52):87323-41. Epub 2016/11/30. doi: 10.18632/oncotarget.13556. PubMed PMID: 27895316; PubMed Central PMCID: PMCPMC5349991.

56. Li C, Wang Q, Ma J, Shi S, Chen X, Yang H, et al. Integrative Pathway Analysis of Genes and Metabolites Reveals Metabolism Abnormal Subpathway Regions and Modules in Esophageal Squamous Cell Carcinoma. *Molecules* (2017) 22(10). Epub 2017/09/25. doi: 10.3390/molecules22101599. PubMed PMID: 28937628; PubMed Central PMCID: PMCPMC6151487.

57. Chen D, Lu T, Tan J, Zhao K, Li Y, Zhao W, et al. Identification of a transcription factor-microRNA network in esophageal adenocarcinoma through bioinformatics analysis and validation through qRT-PCR. *Cancer Manag Res* (2019) 11:3315-26. Epub 2019/05/23. doi: 10.2147/cmar.S201274. PubMed PMID: 31114367; PubMed Central PMCID: PMCPMC6489589.

58. Wang WW, Zhao ZH, Wang L, Li P, Chen KS, Zhang JY, et al. MicroRNA-134 prevents the progression of esophageal squamous cell carcinoma via the PLXNA1-mediated MAPK signalling pathway. *EBioMedicine* (2019) 46:66-78. Epub 2019/08/07. doi: 10.1016/j.ebiom.2019.07.050. PubMed PMID: 31383552; PubMed Central PMCID: PMCPMC6711887.

59. Li M, Wang K, Pang Y, Zhang H, Peng H, Shi Q, et al. Secreted Phosphoprotein 1 (SPP1) and Fibronectin 1 (FN1) Are Associated with Progression and Prognosis of Esophageal Cancer as Identified by Integrated Expression Profiles Analysis. *Med Sci Monit* (2020) 26:e920355. Epub 2020/03/26. doi: 10.12659/msm.920355. PubMed PMID: 32208405; PubMed Central PMCID: PMCPMC7111131.

60. Xia C, Chen X, Li J, Chen P. SLC39A4 as a Novel Prognosis Marker Promotes Tumor Progression in Esophageal Squamous Cell Carcinoma. *Onco Targets Ther* (2020) 13:3999-4008. Epub 2020/06/05. doi: 10.2147/ott.S245094. PubMed PMID: 32494154; PubMed Central PMCID: PMCPMC7227820.

61. Parman C, Halling C. *affyQCReport: a package to generate QC reports for Affymetrix array data*. (2006).

62. Carvalho BS, Irizarry RA. A framework for oligonucleotide microarray preprocessing. *Bioinformatics* (2010) 26(19):2363-7.

63. Gautier L, Cope L, Bolstad BM, Irizarry RA. affy—analysis of Affymetrix GeneChip data at the probe level. *Bioinformatics* (2004) 20(3):307-15.

64. Hong F, Breitling R, McEntee CW, Wittner BS, Nemhauser JL, Chory J. RankProd: a bioconductor package for detecting differentially expressed genes in meta-analysis. *Bioinformatics* (2006) 22(22):2825-7.

65. Simillion C, Liechti R, Lischer HEL, Ioannidis V, Bruggmann R. Avoiding the pitfalls of gene set enrichment analysis with SetRank. *BMC Bioinformatics* (2017) 18(1):151. doi: 10.1186/s12859-017-1571-6.

66. Jassal B, Matthews L, Viteri G, Gong C, Lorente P, Fabregat A, et al. The reactome pathway knowledgebase. *Nucleic Acids Res* (2020) 48(D1):D498-d503. Epub 2019/11/07. doi: 10.1093/nar/gkz1031. PubMed PMID: 31691815; PubMed Central PMCID: PMCPMC7145712.

67. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* (2003) 13(11):2498-504. Epub 2003/11/05. doi: 10.1101/gr.1239303. PubMed PMID: 14597658; PubMed Central PMCID: PMCPMC403769.

68. Wu G, Dawson E, Duong A, Haw R, Stein L. ReactomeFIViz: a Cytoscape app for pathway and network-based data analysis. *F1000Res* (2014) 3:146. Epub 2014/10/14. doi: 10.12688/f1000research.4431.2. PubMed PMID: 25309732; PubMed Central PMCID: PMCPMC4184317.

69. Wu G, Feng X, Stein L. A human functional protein interaction network and its application to cancer data analysis. *Genome Biology* (2010) 11(5):R53. doi: 10.1186/gb-2010-11-5-r53.

70. Schroeder WJ, Martin KM. Overview of Visualization. In: Hansen CD, Johnson CR, editors. *The Visualization Toolkit: An Object-Oriented Approach to 3D Graphics*. 3rd ed. Burlington: Butterworth-Heinemann (2005). p. 3-35.

71. Fang S, Dai Y, Mei Y, Yang M, Hu L, Yang H, et al. Clinical significance and biological role of cancer-derived Type I collagen in lung and esophageal cancers. *Thorac Cancer* (2019) 10(2):277-88. Epub 2019/01/04. doi: 10.1111/1759-7714.12947. PubMed PMID: 30604926; PubMed Central PMCID: PMCPMC6360244.

72. Rauluseviciute I, Drabløs F, Rye MB. DNA hypermethylation associated with upregulated gene expression in prostate cancer demonstrates the diversity of epigenetic regulation. *BMC Medical Genomics* (2020) 13(1):6. doi: 10.1186/s12920-020-0657-6.

73. Misawa K, Kanazawa T, Misawa Y, Imai A, Endo S, Hakamada K, et al. Hypermethylation of collagen α2 (I) gene (COL1A2) is an independent predictor of survival in head and neck cancer. *Cancer Biomark* (2011) 10(3-4):135-44. Epub 2011/01/01. doi: 10.3233/cbm-2012-0242. PubMed PMID: 22674299.

74. Li G, Jiang W, Kang Y, Yu X, Zhang C, Feng Y. High expression of collagen 1A2 promotes the proliferation and metastasis of esophageal cancer cells. *Ann Transl Med* (2020) 8(24):1672. Epub 2021/01/26. doi: 10.21037/atm-20-7867. PubMed PMID: 33490184; PubMed Central PMCID: PMCPMC7812173.

75. Worfolk JC, Bell S, Simpson LD, Carne NA, Francis SL, Engelbertsen V, et al. Elucidation of the AGR2 Interactome in Esophageal Adenocarcinoma Cells Identifies a Redox-Sensitive Chaperone Hub for the Quality Control of MUC-5AC. *Antioxid Redox Signal* (2019) 31(15):1117-32. Epub 2019/08/23. doi: 10.1089/ars.2018.7647. PubMed PMID: 31436131.

76. Wang Z, Hao Y, Lowe AW. The adenocarcinoma-associated antigen, AGR2, promotes tumor growth, cell migration, and cellular transformation. *Cancer Res* (2008) 68(2):492-7. Epub 2008/01/18. doi: 10.1158/0008-5472.Can-07-2930. PubMed PMID: 18199544.

77. Wang H, Shen L, Lin Y, Shi Q, Yang Y, Chen K. The expression and prognostic significance of Mucin 13 and Mucin 20 in esophageal squamous cell carcinoma. *J Cancer Res Ther* (2015) 11 Suppl 1:C74-9. Epub 2015/09/02. doi: 10.4103/0973-1482.163846. PubMed PMID: 26323930.

78. Uchikado Y, Inoue H, Haraguchi N, Mimori K, Natsugoe S, Okumura H, et al. Gene expression profiling of lymph node metastasis by oligomicroarray analysis using laser microdissection in esophageal squamous cell carcinoma. *International journal of oncology* (2006) 29(6):1337-47.

79. de A Simão T, Souza-Santos PT, de Oliveira DSL, Bernardo V, Lima SCS, Rapozo DCM, et al. Quantitative evaluation of SPRR3 expression in esophageal squamous cell carcinoma by qPCR and its potential use as a biomarker. *Experimental and Molecular Pathology* (2011) 91(2):584-9. doi: https://doi.org/10.1016/j.yexmp.2011.06.006.

80. Luo A, Chen H, Ding F, Zhang Y, Wang M, Xiao Z, et al. Small proline-rich repeat protein 3 enhances the sensitivity of esophageal cancer cells in response to DNA damage-induced apoptosis. *Molecular oncology* (2013) 7(5):955-67.

81. Wang H-T, Kong J-P, Ding F, Wang X-Q, Wang M-R, Liu L-X, et al. Analysis of gene expression profile induced by EMP-1 in esophageal cancer cells using cDNA Microarray. *World journal of gastroenterology* (2003) 9(3):392.

82. Wang Q, Peng D, Zhu S, Chen Z, Hu T, Soutto M, et al. Regulation of Desmocollin3 Expression by Promoter Hypermethylation is Associated with Advanced Esophageal Adenocarcinomas. *J Cancer* (2014) 5(6):457-64. doi: 10.7150/jca.9145. PubMed PMID: 24847386.

83. Fang W-K, Chen B, Xu X-E, Liao L-D, Wu Z-Y, Wu J-Y, et al. Altered expression and localization of desmoglein 3 in esophageal squamous cell carcinoma. *Acta histochemica* (2014) 116(5):803-9.

84. Liu W, Snell JM, Jeck WR, Hoadley KA, Wilkerson MD, Parker JS, et al. Subtyping sub-Saharan esophageal squamous cell carcinoma by comprehensive molecular analysis. *JCI insight* (2016) 1(16):e88755. doi: 10.1172/jci.insight.88755. PubMed PMID: PMC5053149.

85. Dusek RL, Attardi LD. Desmosomes: new perpetrators in tumour suppression. *Nat Rev Cancer* (2011) 11(5):317-23. doi: 10.1038/nrc3051. PubMed PMID: 21508970.

86. Yang W, Zhao X, Han Y, Duan L, Lu X, Wang X, et al. Identification of hub genes and therapeutic drugs in esophageal squamous cell carcinoma based on integrated bioinformatics strategy. *Cancer Cell International* (2019) 19(1):142. doi: 10.1186/s12935-019-0854-6.

87. Wang X, Sun Q. TP53 mutations, expression and interaction networks in human cancers. *Oncotarget* (2017) 8(1):624-43. Epub 2016/11/24. doi: 10.18632/oncotarget.13483. PubMed PMID: 27880943; PubMed Central PMCID: PMCPMC5352183.

88. Han B-A, Yang X-P, Hosseini DK, Zhang P, Zhang Y, Yu J-T, et al. Identification of candidate aberrantly methylated and differentially expressed genes in Esophageal squamous cell carcinoma. *Scientific Reports* (2020) 10(1):9735. doi: 10.1038/s41598-020-66847-4.

89. Hansel DE, Dhara S, Huang RC, Ashfaq R, Deasel M, Shimada Y, et al. CDC2/CDK1 Expression in Esophageal Adenocarcinoma and Precursor Lesions Serves as a Diagnostic and Cancer Progression Marker and Potential Novel Drug Target. *The American Journal of Surgical Pathology* (2005) 29(3).

90. Wang L, Shan L, Zhang S, Ying J, Xue L, Yuan Y, et al. PIK3CA gene mutations and overexpression: implications for prognostic biomarker and therapeutic target in Chinese esophageal squamous cell carcinoma. *PLoS One* (2014) 9(7):e103021. Epub 2014/07/24. doi: 10.1371/journal.pone.0103021. PubMed PMID: 25054828; PubMed Central PMCID: PMCPMC4108430.

91. Palumbo A, Jr., Meireles Da Costa N, Pontes B, Leite de Oliveira F, Lohan Codeço M, Ribeiro Pinto LF, et al. Esophageal Cancer Development: Crucial Clues Arising from the Extracellular Matrix. *Cells* (2020) 9(2). Epub 2020/02/23. doi: 10.3390/cells9020455. PubMed PMID: 32079295; PubMed Central PMCID: PMCPMC7072790.

92. Lin EW, Karakasheva TA, Hicks PD, Bass AJ, Rustgi AK. The tumor microenvironment in esophageal cancer. *Oncogene* (2016) 35(41):5337-49. doi: 10.1038/onc.2016.34.

93. Frantz C, Stewart KM, Weaver VM. The extracellular matrix at a glance. *Journal of Cell Science* (2010) 123(24):4195-200. doi: 10.1242/jcs.023820.

94. Dong Z, Wang J, Zhang H, Zhan T, Chen Y, Xu S. Identification of potential key genes in esophageal adenocarcinoma using bioinformatics. *Exp Ther Med* (2019) 18(5):3291-8. Epub 2019/10/17. doi: 10.3892/etm.2019.7973. PubMed PMID: 31616504; PubMed Central PMCID: PMCPMC6781836.

95. Hu J, Li R, Miao H, Wen Z. Identification of key genes for esophageal squamous cell carcinoma via integrated bioinformatics analysis and experimental confirmation. *J Thorac Dis* (2020) 12(6):3188-99. Epub 2020/07/10. doi: 10.21037/jtd.2020.01.33. PubMed PMID: 32642240; PubMed Central PMCID: PMCPMC7330802.

96. Zhang Y, Feng Y-B, Shen X-M, Chen B-S, Du X-L, Luo M-L, et al. Exogenous expression of Esophagin/SPRR3 attenuates the tumorigenicity of esophageal squamous cell carcinoma cells via promoting apoptosis. *International Journal of Cancer* (2008) 122(2):260-6. doi: https://doi.org/10.1002/ijc.23104.

97. Rinnerthaler M, Bischof J, Streubel MK, Trost A, Richter K. Oxidative stress in aging human skin. *Biomolecules* (2015) 5(2):545-89. Epub 2015/04/24. doi: 10.3390/biom5020545. PubMed PMID: 25906193; PubMed Central PMCID: PMCPMC4496685.

98. Suresh R, Diaz RJ. The remodelling of actin composition as a hallmark of cancer. *Translational Oncology* (2021) 14(6):101051. doi: https://doi.org/10.1016/j.tranon.2021.101051.

99. Strand J, Nili M, Homsher E, Tobacman LS. Modulation of myosin function by isoform-specific properties of Saccharomyces cerevisiae and muscle tropomyosins. *J Biol Chem* (2001) 276(37):34832-9. Epub 2001/07/18. doi: 10.1074/jbc.M104750200. PubMed PMID: 11457840.

100. Liu T, Han X, Zheng S, Liu Q, Tuerxun A, Zhang Q, et al. CALM1 promotes progression and dampens chemosensitivity to EGFR inhibitor in esophageal squamous cell carcinoma. *Cancer Cell International* (2021) 21(1):121. doi: 10.1186/s12935-021-01801-6.

101. Kim K, Lee D, Ahn C, Kang HY, An BS, Seong YH, et al. Effects of estrogen on esophageal function through regulation of Ca(2+)-related proteins. *J*

Gastroenterol (2017) 52(8):929-39. Epub 2017/01/13. doi: 10.1007/s00535-016-1305-y. PubMed PMID: 28078471.

102. Li Q, Cui L, Tian Y, Cui H, Li L, Dou W, et al. Protective Effect of Dietary Calcium Intake on Esophageal Cancer Risk: A Meta-Analysis of Observational Studies. *Nutrients* (2017) 9(5). Epub 2017/05/20. doi: 10.3390/nu9050510. PubMed PMID: 28524093; PubMed Central PMCID: PMCPMC5452240.

103. Wang J, Di J, Wang G. ENPP4 overexpression is associated with no recovery from Barrett's esophagus. *Int J Clin Exp Pathol* (2020) 13(12):2927-36. Epub 2021/01/12. PubMed PMID: 33425094; PubMed Central PMCID: PMCPMC7791367.

104. Seabra MC. Nucleotide Dependence of Rab Geranylgeranylation: Rab ESCORT PROTEIN INTERACTS PREFERENTIALLY WITH GDP-BOUND Rab *. *Journal of Biological Chemistry* (1996) 271(24):14398-404. doi: 10.1074/jbc.271.24.14398.

105. Qin X, Wang J, Wang X, Liu F, Jiang B, Zhang Y. Targeting Rabs as a novel therapeutic strategy for cancer therapy. *Drug Discovery Today* (2017) 22(8):1139-47. doi: https://doi.org/10.1016/j.drudis.2017.03.012.

106. Tong M, Chan KW, Bao JY, Wong KY, Chen J-N, Kwan PS, et al. Rab25 is a tumor suppressor gene with antiangiogenic and anti-invasive activities in esophageal squamous cell carcinoma. *Cancer research* (2012) 72(22):6024-35.

107. Cheng L, Yang F, Zhou B, Yang H, Yuan Y, Li X, et al. RAB23, regulated by miR-92b, promotes the progression of esophageal squamous cell carcinoma. *Gene* (2016) 595(1):31-8.

108. Wang J, Xie X, Sun Y. Time series expression pattern of key genes reveals the molecular process of esophageal cancer. *Bioscience reports* (2020) 40(2).

109. Ries A, Schelch K, Falch D, Pany L, Hoda MA, Grusch M. Activin A: an emerging target for improving cancer treatment? *Expert Opinion on Therapeutic Targets* (2020) 24(10):985-96. doi: 10.1080/14728222.2020.1799350.

110. Shafi AA, Knudsen KE. Cancer and the Circadian Clock. *Cancer Research* (2019) 79(15):3806-14. doi: 10.1158/0008-5472.Can-19-0566.

111. Zmrzljak UP, Rozman D. Circadian Regulation of the Hepatic Endobiotic and Xenobitoic Detoxification Pathways: The Time Matters. *Chemical Research in Toxicology* (2012) 25(4):811-24. doi: 10.1021/tx200538r.

112. Wendeu-Foyet MG, Menegaux F. Circadian Disruption and Prostate Cancer Risk: An Updated Review of Epidemiological Evidences. *Cancer Epidemiology Biomarkers &amp;amp; Prevention* (2017) 26(7):985. doi: 10.1158/1055-9965.EPI-16-1030.

113. Matejcic M, Gunter MJ, Ferrari P. Alcohol metabolism and oesophageal cancer: a systematic review of the evidence. *Carcinogenesis* (2017) 38(9):859-72. Epub 2017/06/25. doi: 10.1093/carcin/bgx067. PubMed PMID: 28645180.

114. Zhang GH, Mai RQ, Huang B. Meta-analysis of ADH1B and ALDH2 polymorphisms and esophageal cancer risk in China. *World J Gastroenterol* (2010) 16(47):6020-5. Epub 2010/12/16. doi: 10.3748/wjg.v16.i47.6020. PubMed PMID: 21157980; PubMed Central PMCID: PMCPMC3007115.

115. Jancova P, Anzenbacher P, Anzenbacherova E. Phase II drug metabolizing enzymes. *Biomed Pap Med Fac Univ Palacky Olomouc Czech Repub* (2010) 154(2):103-16. Epub 2010/07/30. doi: 10.5507/bp.2010.017. PubMed PMID: 20668491.

116. Steevens J, van den Brandt PA, Goldbohm RA, Schouten LJ. Selenium status and the risk of esophageal and gastric cancer subtypes: the Netherlands cohort study. *Gastroenterology* (2010) 138(5):1704-13. Epub 2009/12/17. doi: 10.1053/j.gastro.2009.12.004. PubMed PMID: 20006613.

117. Mark SD, Qiao Y-L, Dawsey SM, Wu Y-P, Katki H, Gunter EW, et al. Prospective study of serum selenium levels and incident esophageal and gastric cancers. *Journal of the National Cancer Institute* (2000) 92(21):1753-63.

118. Middleton DR, McCormack VA, Watts MJ, Schüz J. Environmental geochemistry and cancer: a pertinent global health problem requiring interdisciplinary collaboration. *Environmental geochemistry and health* (2019):1-10.

119. Pritchett NR, Burgert SL, Murphy GA, Brockman JD, White RE, Lando J, et al. Cross sectional study of serum selenium concentration and esophageal squamous dysplasia in western Kenya. *BMC cancer* (2017) 17(1):835. Epub 2017/12/09. doi: 10.1186/s12885-017-3837-9. PubMed PMID: 29216866; PubMed Central PMCID: PMCPMC5721656.

120. Liu B, Pan S, Xiao Y, Liu Q, Xu J, Jia L. LINC01296/miR-26a/GALNT3 axis contributes to colorectal cancer progression by regulating O-glycosylated MUC1 via PI3K/AKT pathway. *J Exp Clin Cancer Res* (2018) 37(1):316. Epub 2018/12/15. doi: 10.1186/s13046-018-0994-x. PubMed PMID: 30547804; PubMed Central PMCID: PMCPMC6295061.

121. Yao C, Narumiya S. Prostaglandin-cytokine crosstalk in chronic inflammation. *Br J Pharmacol* (2019) 176(3):337-54. Epub 2018/11/02. doi: 10.1111/bph.14530. PubMed PMID: 30381825; PubMed Central PMCID: PMCPMC6329627.

122. Ciaccia L. Fundamentals of Inflammation. *Yale J Biol Med* (2011) 84(1):64-5. Epub 2011/03/. PubMed PMID: PMC3064252.

123. Xiang W, Shi R, Kang X, Zhang X, Chen P, Zhang L, et al. Monoacylglycerol lipase regulates cannabinoid receptor 2-dependent macrophage activation and cancer progression. *Nat Commun* (2018) 9(1):2574. Epub 2018/07/04. doi: 10.1038/s41467-018-04999-8. PubMed PMID: 29968710; PubMed Central PMCID: PMCPMC6030061.

124. Song W, Jiang R, Zhao C. Regulation of arachidonic acid in esophageal adenocarcinoma cells and tumor-infiltrating lymphocytes. *Oncol Lett* (2013)

5(6):1897-902. Epub 2013/03/20. doi: 10.3892/ol.2013.1267. PubMed PMID: 23833663.

125. Russo G, Zegar C, Giordano A. Advantages and limitations of microarray technology in human cancer. *Oncogene* (2003) 22(42):6497-507. doi: 10.1038/sj.onc.1206865.

Chapter 5: Discussion

Table of Contents

5.1 Summary of the PhD topic

Cancer remains a leading cause of mortality globally, with an estimated 18.1 million cases and 9.6 million deaths occurring in 2018.(1) Whilst cancer contributes immensely to loss of life, its disease burden also contributes an immense burden on healthcare systems, healthcare costs, and quality of life of those affected. EC is a deadly malignancy with high incidence and mortality rates (572,034 new cases and 508,585 deaths annually) worldwide. The environmental and genetic etiology of EC is still poorly understood, with research gaps in African populations. EC has a poor prognosis and a low survival rate.(1) In this study, we assessed genetic, environmental, and lifestyle factors associated with the development of EC as well as its pathobiology. To achieve this overarching aim, objectives were presented in chapter format as follows:

5.1.1 Chapter 2: Genetic risk factors for ESCC reported in African literature: a systematic review.

This study was a systematic review that investigated the genetic factors (germline and somatic variants) associated with ESCC development in African populations. The study was published in Frontiers in Genetics with the title "Systematic Review of Genetic Factors in the Etiology of Esophageal Squamous Cell Carcinoma in African Populations"(2) (doi:10.3389/fgene.2019.00642). The review systematically screened and critically appraised studies reported in African populations, specifically looking at germline variants (inherited at birth) and somatic variants (acquired throughout life and present in the tumour). The study represented the first systematic review on ESCC genetics in African populations.

Genetic variants reported to be associated with ESCC in African populations were summarized. Over 100 SNPs were analysed in 17 case-control genetic association studies. Of these, 25 SNPs in 20 genes [*ADH1B, ADH3, ALDH2, AR, CASP8, CHEK2, CP, CYP2E1, CYP3A5, GSTT2B, MGMT, MLH3, MSH3, NAT2, PTGS2 (*also known as *COX-2), PLCE1, PMS1, RUNX1, SLC11A1,* and *TP53]* were reported to be associated with ESCC. The results from the somatic studies showed 44 somatic changes in the following 22 genes: *AR*, *CCND1, CDKN2A, COL1A2, EFGR, EP300,* *FAT1, FAT2, FAT3, FAT4, FBXW7, JAG1, KMT2C (MLL3), KMT2D (MLL2), MUC2, NFE2L2, NOTCH1, NOTCH3, PIK3CA, SERPINB4, TP53, and TP63,* and six genetic loci without specific gene names. In summary, these genes are involved in alcohol metabolism, cell cycle regulation, DNA repair, tumour suppression, xenobiotic metabolism, and extracellular matrix (ECM) formation.(2)

We performed all pairwise r^2 analyses to determine the linkage disequilibrium (LD) for 67 SNPs reported in our study. The results showed overall low LD across all the SNPs. Thirteen pairs of SNPs in *MHS2, CP, MSH3, PLCE1, CHEK2, NAT1* genes had $r^2 > 0.45$. LD is the non-random association of alleles positioned at different locations in a population.(3) It has multiple applications, including identifying regions of a gene associated with the disease.(3)

Whilst we identified important genetic variants associated with ESCC in African populations, we also identified several limitations. These include the overall lack of genetic studies, small studies from which reliability and replicability of results are uncertain, and inadequate quality of reporting. We recommend comprehensive large-scale genetic studies in Africa with a particular focus on WGS and GWAS.

5.1.2 Chapter 3: Lifestyle and environmental risk factors reported in African literature: a systematic review and meta-analysis.

This chapter is a systematic review of the lifestyle and environmental risk factors for ESCC in African populations. Studies that met the selection criteria were critically appraised and several known and emerging risk factors were identified and described. These included smoking, alcohol consumption, low socioeconomic status, poor diet, PAH exposure, consumption of hot food and beverages, poor oral health, infectious agents, esophageal injury, family history of cancer, and non-acid gastro-esophageal reflux. Meta-analyses were carried out and population attributable fractions calculated on risk factors that had adequate information (tobacco, alcohol use, combined tobacco and alcohol use, PAH exposure, esophageal injury, and fruit and vegetable consumption). This study represents one of the few systematic reviews on environmental factors done on African populations. It is also the first study to perform an integrated PAF analysis from multiple African studies. PAF analysis was done on tobacco, alcohol use, combined tobacco and alcohol use, PAH exposure, esophageal injury, and fruit and vegetable consumption.

During the writing of this systematic review, another systematic review on the same topic was published.(4) Neither we nor the authors of the other systematic review had registered our protocols on the International Prospective Register of Systematic Reviews (PROSPERO), which is a registry for systematic review protocols.(5) PROSPERO allows for authors to check if a similar review already exists or is in progress. This avoids unnecessary duplication of work. Our systematic review differed from the already published systematic review(4) in that we performed a meta-analysis on an additional four risk factors (combined smoking and tobacco exposure, PAH exposure, esophageal injury, and fruit and vegetable consumption) and performed PAF analysis. For similar risk factors, our study had a more comprehensive list of studies included in the meta-analysis. Our meta-analysis methods were also more extensive. We performed tests for heterogeneity and between study variance as part of the meta-analysis. To assess the robustness of the meta-analyses, Graphic Display of Heterogeneity (GOSH) plots were generated to identify the patterns of effect sizes and heterogeneity in the data. Additionally, a second meta-analysis for each risk factor was done after the removal of outliers. We also performed PAF analysis as part of the systematic review, to ascertain the proportion of ESCC attributable to each risk factor.

Tobacco smoking and alcohol exposure are some of the main environmental risk factors for ESCC, and we expected to find them to significantly increase ESCC development in our analysis. It is important to state that alcohol is an independent risk factor for ESCC, in the absence of tobacco smoking. It has a clear dose response relationship for ESCC. In the meta-analysis, both tobacco smoking and alcohol consumption showed an increased risk of ESCC, with OR 4.14 (3.26-5.26) and 2.14 (1.65-2.78), respectively. Both risk factors also had relatively high PAF, 17% for tobacco smoking and 13% for alcohol consumption. A positive synergistic effect for combined alcohol and tobacco use was also identified in our meta-analysis, similar to the evidence presented in the literature.(6) The interactive nature of tobacco and alcohol has been substantiated by biological evidence in the literature. Molecular studies have shown that tobacco extracts and ethanol interact resulting in altered gene expression profiles, cellular morphology, and cell growth of endothelial cells and fibroblasts.(7)

We expected PAH exposure to be one of the main risk factors, as approximately 77% of the population in Africa still relies on traditional biomass fuel (wood, charcoal, coal, dung, and crop residues) for cooking and heating.(8) These fuels are a source of PAHs. In our study, exposure to PAHs from wood, and charcoal increased the risk of ESCC twofold with a PAF of 5%. Our meta-analysis results corroborated evidence from one other systematic review which investigated the role of biomass fuel exposure and ESCC.(9) Household PAH exposure mostly occurs through the burning of solid fuels for cooking and heating, which is common in African households where wood, charcoal, dung, and maize cobs are the primary fuel source.(10) PAH exposure through household air pollution may explain the ESCC burden in women and younger patients in African populations, as women are most likely to be tending to fires for cooking in small kitchen rooms with inadequate ventilation, at times in the company of their young children. Household air pollution studies using measured continuous monitoring have shown mean daily concentrations of particulate matter of 2800 to 5000 mg/m^3 in young and adult women in Kenya, 2.5 to 5 times higher than that of their male counterparts.(11) Women are reported to have the highest absolute exposure to particulate matter, with results from Ghana and Ethiopia corroborating this evidence.(12) Surprisingly, most of the evidence in household air pollution has been dominated by countries outside Africa, despite large proportions of African households being dependant on biomass fuel. More studies are needed to elucidate the role of PAHs in ESCC development in high-risk regions.

We expected consumption of fruits and vegetables to have a protective effect from ESCC, similar to other systematic reviews investigating the role of fruits and vegetable consumption (13), and citrus fruit consumption (14). Our study showed that consumption of fruits and vegetables reduced the risk of ESCC by 40%. The protective effect of fruits and vegetables may be through their antioxidant properties, alleviating the effect of oxidative stress (15). Investigating the role of dietary practices and ESCC development can be difficult due to the heterogeneity of dietary practices, therefore analysis of micronutrient deficiencies may be more effective. It is reported that micronutrient deficiencies make the esophageal epithelium more susceptible to inflammation, promoting epigenetic changes that result in tumorigenesis. Interactions between diet and genomics is an area that may be pivotal in understanding diet as a risk factor for ESCC.

Our study identified emerging risk factors in African populations which warrant further investigations, and they include consumption of hot food and beverages and oral health. Esophageal injury emerged as an important risk factor for ESCC in African populations and was comprised of consumption of hot food and beverages, induced vomiting, and caustic ingestion. Esophageal injury increased the risk of ESCC by over four-fold with OR 4.63 (2.03-10.52 95%CI) and a PAF of 17%. We believe that the role of esophageal injury in ESCC requires further investigation, particularly the consumption of hot food and beverages. We also recommend investigating the role of esophageal injury on the development of squamous dysplasia.

Overall, according to our study, 43% of ESCC is attributable to the five risk factors: tobacco, alcohol use, combined tobacco and alcohol use, PAH exposure, and esophageal injury. The PAF estimates presented here should be interpreted with caution considering the study heterogeneity, heterogeneity in exposure assessment of the original studies and potential misclassification of exposures.All these risk factors, including consumption of fruit and vegetables, are particularly appealing because they are modifiable and can make a big contribution in the primary prevention of ESCC in Africa.

5.1.3 Chapter 4: Genes and pathways with differential mRNA expression in esophageal cancer: meta-analysis of differentially expressed genes and Gene Set Enrichment Analysis of GEO (Gene Expression Omnibus) datasets

This chapter focussed on the identification of genes and pathways with differential mRNA expression in EC using meta-analysis of DEGs and Gene Set Enrichment Analysis of GEO (Gene Expression Omnibus) datasets. A total of 18 publicly available GEO mRNA expression datasets, with expression data on 906 individual tissue samples, were included in the analysis. Of the 18 datasets, three used EAC tissue, eleven used ESCC tissue, and nine used Barrett's esophagus (BE) tissue. One dataset included EAC, ESCC, and BE tissue, whilst one dataset included both EAC and squamous dysplasia tissue samples. This analysis provides novel insights into the mechanisms linked to EC development, and the differences between the different types of EC and BE.

In this chapter, shared and divergent molecular features of ESCC and EAC, which are histologically different tumours(16), as well as the precancerous lesion for EAC, BE, was analysed. Squamous dysplasia could not be analysed due to the limited number of samples that were available. Overall, 1,579 upregulated genes and 1,383 downregulated genes were outputted for EAC and ESCC, whilst 767 upregulated and 1,181 downregulated genes were outputted for BE. The majority of DEGs were significantly associated with the pathways involved in the ECM, including "*Extracellular matrix organisation", "Assembly of collagen fibrils and other multimeric structures", "Smooth muscle contraction", and "Collagen chain trimerization".* Interpretation on how the *"smooth muscle contraction"* and *"Formation of the cornified envelope"* pathways is lacking in the literature and needs further investigation. The second most common group of pathways belonged to cell membrane regulation and included "*Cytosolic sulfonation of small molecules", "Transport of inorganic cations/anions and amino acids/oligopeptides", "Neutrophil degranulation", "Metabolism of Angiotensinogen to Angiotensins", "Tight junction interactions",* and *"RAB geranylgeranylation*". We identified the pathway "*RAB geranylgeranylation*" as novel and a pathway of interest warranting further investigation. We hypothesised that this pathway, which is involved in posttranslational modification of proteins, cell growth, and signalling, may be involved in the progression of BE to EAC.

Pathways involved in cell cycle regulation were also identified, including *"TP53 regulates transcription of several additional cell death genes whose specific roles in p53-dependent apoptosis remain uncertain", "Cyclin B2 mediated events*", and "*BMAL1:CLOCK,NPAS2 activates circadian gene expression*". We hypothesised that genes dysregulated in the pathway *"TP53 regulates transcription of several additional cell death genes whose specific roles in p53-dependent apoptosis remain uncertain"* may be driving tumorigenesis from BE, whilst genes dysregulated in *"Cyclin B2 mediated events"* may be driving tumour development in ESCC. The pathway "*BMAL1:CLOCK,NPAS2 activates circadian gene expression*" has not been described for EC previously. The circadian clock genes which regulate the sleep and wake cycles also regulate normal cells and cancer cells' division, and proliferation. Disruptions of normal circadian rhythms associated with dysregulation in the clock genes which may lead to cancer development.(17) Cancers reported to be associated with disrupted circadian rhythm include breast, endometrial, colorectal, and ovarian cancer.(17) This pathway has not been interpreted for EC in literature, and provides a novel mechanistic aspect for EAC and ESCC

We expected to find pathways involved in detoxification, which may play roles in the stress and toxic response, as well as xenobiotic metabolism. The pathways we identified included *"Detoxification of Reactive Oxygen Species", "Aflatoxin activation and detoxification", "Cytosolic sulfonation of small molecules", "Fatty acid, triacylglycerol, and ketone body metabolism"* and *"Nicotinate metabolism".* We also identified a pathway involved in alcohol metabolism, *"Ethanol oxidation".* The identification of these pathways is significant due to their involvement in the detoxification of environmental carcinogens including alcohol, tobacco, and PAHs. These are some of the major risk factors reported for ESCC worldwide as well as in our systematic review (Chapter 3), which therefore gives biological plausibility of environmental carcinogens and EC.

Particularly noteworthy are the pathways we identified which have not been interpreted in literature for EC and warrant further investigation. These include *"Metabolism of ingested SeMet, Sec, MeSec into H2Se*" which involves the transformation of dietary selenium to its metabolites. Selenium deficiency is reported to be associated with ESCC development, through the promotion of oxidative stress, and DNA damage(15), and hence dysregulation of this pathway which is involved in the metabolism of selenium may result in low levels of selenium being available in the body. We noted that this pathway has not been previously described for EC in literature and hence should be considered for further investigation. We also identified the pathway "*Defective GALNT3 causes familial hyperphosphatemic tumoral calcinosis (HFTC)"* to be significantly enriched in our analysis. This pathway has also not been described and interpreted for EC in literature and it is unclear whether its association to EC is through the dysregulated *GALNT3* or HFTC. We recommend that further studies be done to elucidate the role that this pathway plays in EC tumorigenesis.

It is important to reiterate that only a minority of the precancerous lesions in EC develop to cancer, and that BE, which we assessed in our study, is not a cancer. There were unfortunately not enough samples for squamous dysplasia to be included in the analysis. Whilst the development of BE significantly increases the risk of EAC development (30-60 fold), the number of BE cases that progress to EAC is low (0.33% for non-dysplastic BE and 10% high grade dysplasia annually).(18) The inclusions of the precancerous lesions into our study were important because understanding the pathobiology of precancerous lesions provides the platform for early diagnosis and prevention. Most importantly, to identify biological pathways driving the progression of premalignant lesions to tumours. Compared to BE, there has been far less research done to understand the genetics and pathobiology of squamous dysplasia. Over the past decade, about five GWAS studies have been done on BE identifying germline variants associated with BE, whilst we could not find any done on squamous dysplasia.(18) Additionally, the risk of progression from squamous dysplasia to ESCC has not been adequately studied. The few studies available have mainly been done on the Chinese population.(19) Early detection of the precancerous lesions has the capacity to reduce the morbidity and mortality rates of EC.(20) A whole genome analysis of both ESCC and squamous dysplasia found similar variants in the two tissue types from the following genes: *TP53, CDKN2A, CCND1, SOX2, NFE2L2*, and *CDKN2A*.(21) There continues to be a push towards surveillance and screening for BE and squamous dysplasia in high incidence areas, with the intent of preventing the development of EAC and ESCC. However, the cost effectiveness and accuracy of this intervention remains a point of contention among researchers. There is therefore a need for accurate identification of BE patients who end up progressing to EAC. This requires knowledge of the genetic basis of BE and EAC, and the mechanisms which drive tumorigenesis. The importance of studying BE and squamous dysplasia can, therefore, not be understated. Additionally, studies detailing follow up after treatment of precancerous lesions are needed.

There was an apparent lack of gene expression studies on African populations, making the generalization of the results from the ESCC analysis difficult. Similar to the African Esophageal Cancer Corridor, China has a high prevalence of ESCC, however, evidence from our systematic review on genetic factors shows that associations reported in the Asian populations at times do not exist in some of the African populations (Chapter 2). It remains to be seen if this is the same regarding gene expression and the biological pathways associated with EC pathobiology.

Gene expression analysis plays an important role in understanding the pathobiology of EC, including development, prognosis, progression, metastasis, and therapy. Investing in gene expression analysis for EC in African populations will allow for the production of a wealth of information that can be used to understand the disease better and identify biological pathways driving its etiology. This will have major implications in prevention, diagnosis, and treatment. Microarrays are one of the high throughput technologies which can identify key and novel pathways associated with the pathobiology of cancer. Although they are an older technology, they are cost-effective and would be a good resource to explore gene expression in LMICs.

We encountered a limitation in our study, in that the sample size we planned to analyse exceeded the computational resources of the meta-analysis package, RankProd. The issue of large sample sizes is not uncommon in genomic studies. We addressed this limitation in two ways. Firstly, we computed multiple pairwise contrasts, which allowed us to assess different combinations of the datasets together. This also allowed us to assess all the datasets included in our study. Secondly, we explored sample size reduction using subsampling, which allowed us to assess the same datasets with reduced sample sizes. We checked and validated the reliability of the subsampling using multiple repeats of the analysis on the same datasets to get stable estimates of the effects.

5.2 A unified hypothesis

In the chapters presented here, the thesis identified environmental, genetic, and biological mechanisms underscoring the development, progression, and pathobiology of EC. The findings corroborated the evidence in the literature, identified emerging environmental risk factors in African populations, and provided evidence on germline and somatic variants driving ESCC development in African populations, and identified key and novel biological pathways involved in ESCC, EAC and BE pathobiology in

Asian, American, and European populations. The studies presented here also highlight the gaps in EC research, particularly in Africa. Overall, the PhD studies explored and highlighted the multifactorial etiology of ESCC.

Investigation and interpretation of interactions between genomics and the environment is pivotal in understanding EC. The findings in our systematic review on genetic risk factors and the differential mRNA expression study identified genetic variants, dysregulated genes, and pathways which provide additional evidence for the effect of some of the environmental risk factors in our systematic review. These risk factors included

I. Alcohol consumption

Alcohol was one of the major risk factors identified for ESCC in our environmental and lifestyle risk factor systematic review. It is an independent risk factor for ESCC. The alcohol metabolizing genes (*ALDH2, ADH1B)(22)* were reported to have variants associated with ESCC in our genetic systematic review (Chapter 2), were dysregulated in our expression profiling study, and were involved in one of the enriched pathways identified, "*Ethanol oxidation*" (Chapter 4). Alcohol consumption emerged as a consistent risk for ESCC throughout the thesis.

II. Tobacco smoking, PAHs and other environmental carcinogens

Tobacco smoke and exposure to PAHs were some of the ESCC risk factors in African populations we identified in our environmental and lifestyle risk factors systematic review (Chapter 3). We identified variants in genes involved in detoxification and xenobiotic metabolism in our systematic review which included Phase I enzyme *CYP3A5* and Phase II enzyme *GSTT1* (Chapter 2). *CYP3A5* and *GSTT1* have been reported to be involved in the metabolism and detoxification of carcinogens including alcohol, tobacco smoke, and PAHs. These genes were also dysregulated in our expression profiling study (Chapter 4). Several pathways involved in toxic response, as well as xenobiotic metabolism, were significantly enriched in our study, which are likely involved in the detoxification of environmental carcinogens. The pathways included *"Detoxification of Reactive Oxygen Species", "Aflatoxin activation and* *detoxification", "Fatty acid, triacylglycerol, and ketone body metabolism"* and *"Nicotinate metabolism"* (Chapter 4)*.*

III. Common genes in the ESCC genetic systematic review and gene expression study

We identified 13 dysregulated genes in our gene expression study that also had variants associated with ESCC in our genetic systematic review (Chapter 2 and 4). The genes included *ALDH2*, *ADH1B*, *TP53*, *CASP8*, *RUNX1*, *CYP3A5*, *MTHFR, AR, XBP1, GSTT1, FBXW7, JAG1, PIK3CA.* The identification of these genes confirmed that these genes are associated with ESCC. However, the gene expression study did not include data from any African populations, whilst for the genetic systematic review we included exclusively African populations. Some of the 13 genes are linked to stress response and detoxification of environmental carcinogens as indicated in points I and II above.

IV. We also identified the pathway *"Arachidonate production from DAG"* to be dysregulated in our gene expression pathway (Chapter 4). This pathway is involved in the formation of prostaglandins, which play an important role in inflammation.(23) This includes redness (rubor), heat (calor), pain (dolor), and swelling (tumor). We suspect that this pathway may have a role to play in esophageal injury and inflammation as a risk factor for ESCC. Esophageal injury emerged as one of the important risk factors for ESCC in our systematic review (Chapter 3). We recommend that further studies be done to elucidate the role the pathway "*Arachidonate production from DAG*" and its dysregulated genes in ESCC development.

The one theme that connects the three chapters in this PhD dissertation is an integrated analysis of multiple datasets, or data from multiple sources to elucidate the etiology and pathobiology of EC. This was achieved through systematic review and linkage disequilibrium analysis of genetic variants from 23 African studies(2), systematic review and meta-analysis of environmental and lifestyle risk factors from African 32 studies, as well as meta-analysis and of differentially expressed genes from 18 GEO datasets collected worldwide from non-African countries. This was made possible through the availability of data in the literature for the systematic reviews as well as the availability of data through public repositories for the gene expression study.

Whilst the main focus of the thesis was on ESCC, we saw an opportunity with the gene expression study to explore the pathobiology of EAC and the BE, the precancerous lesions that may develop to EAC. This study highlighted the importance of studying precancerous lesions which may develop to EC, which are BE, and squamous dysplasia. We identified genes and pathways dysregulated in BE, some of which may drive EAC development, adding to the knowledge on BE pathobiology. Further studies need to elucidate both the environmental and genetic factors resulting in the development of the precancerous lesions, to identify high-risk individuals for which endoscopy surveillance may be instituted. Similar to germline variants in the *APC* gene which increases the risk of colon cancer, and therefore allows for screening and intervention as early as 10 years old,(18) germline variants identified in BE and squamous dysplasia, have the potential to be used in early interventions. This has implications not only for early interventions for the precancerous lesions but for the tumours, EAC, and ESCC, as well.

SNPs from GWAS studies analysing germline variants can be used to calculate polygenic risk scores (PRS), which predict the susceptibility of an individual to disease. Only three studies so far have reported on PRSs for BE and EAC(24-26), however, none to our knowledge have been imputed for squamous dysplasia or ESCC. Imputation of PRS in EC is the first step in the development of a risk prediction model, which incorporates genetic markers, clinical characteristics, demographics, and other identified risk factors. In the study by Dong et al (2019), the authors imputed PRS for EAC and together with known risk factors for EAC [GERD, tobacco smoking, body mass index (BMI), and use of nonsteroidal anti-inflammatory drugs (NSAIDs)], developed a risk prediction model.(25) The PRS was strongly associated with the risk of BE and EAC, however, the risk prediction model's discriminatory ability between cases and controls was poor to moderate.(25) Comprehensive risk prediction models can incorporate genetic risk factors, environmental risk factors, as well as clinical features, and other risk factors and predict susceptibility to EC. Combining these risk factors into an ultimate risk score(18) has the potential to revolutionize screening for

EC and the associated precancerous lesions, address the issue of high costs screening as well as accuracy issues.

Overall, EC is understudied in African populations, particularly in the field of genomics. The lack of genetic studies is common in LMICs and occurs due to a lack of infrastructure, resources, and capacity. The limited studies mean that the genetic architecture of ESCC in African populations remains unclear. We recommend more genomic and transcriptomic studies on the African populations. As shown in this thesis, understanding EC requires an integrated approach, which incorporates the environmental, lifestyle, and genetic factors.

5.3 Overall strengths and limitations of the PhD study

One of the main strengths of the PhD study is that we used systematic reviews and meta-analysis of data from multiple sources, and these provide the highest quality evidence with high statistical power compared to other studies. The PhD study also gave a comprehensive view of ESCC etiology, from lifestyle and environmental factors, genetic variants as well as biological pathways associated with ESCC. Another strength of the study was that we were able to analyse EAC and ESCC as well as BE, the precancerous lesion, and compare results in the gene expression study. This allowed us to determine and compare factors driving tumorigenesis in the two histologically different tumour types. It also allowed us to assess the genetic architecture of BE and processes which may be driving its progression to EAC. For all the included studies/datasets we applied strict QC standards, and the sample sizes in our analysis were sufficient to draw meaningful conclusions Finally, we were able to analyse data from a public repository, GEO datasets, for our gene expression study. Whilst this highlighted the importance of public repositories and data sharing it was also the most comprehensive bioinformatic analysis of existing GEO mRNA expression datasets on EC.

A major limitation of this PhD study was that we were not able to collect primary data for the gene expression analysis. This meant that we could not set the parameters for data collection, particularly clinical and environmental exposures. These additional parameters in combination with the gene expression data would have aided in the interpretation of the enriched biological pathways. We were also not able to control the sample quality during collection and storage. Maintaining the quality of the RNA and making sure it does not degrade is important for accurate results. RNA integrity is important in gene expression studies. Another limitation was that there was a lack of African data in the gene expression analysis. This makes the generalization of our results to the African populations challenging.

5.4 How COVID-19 impacted the PhD studies

The COVID-19 pandemic has caused unforeseen devastation worldwide, infecting and killing millions of people.(27) It has also caused a lot if disruptions in the academic space, increasing pressures on students and researchers alike. The movement restrictions and lockdowns instituted to slow down the spread of the virus in many countries, including South Africa, meant that a lot of research slowed down or came to a standstill. Similar to students worldwide, the COVID-19 pandemic affected my studies. The original plan for my PhD primary study was to perform a hospital-based case-control study assessing the environmental and genetic basis of ESCC in South Africa. The plan was to collect epidemiological data and biospecimens in the Eastern Cape Province. This was going to be achieved through the collection of primary data (environmental, lifestyle factors, diet) through a questionnaire as well as the collection of biological samples for genetic analysis from patients attending Frere hospital in the Eastern Cape Province of South Africa. This region has the highest incidence of ESCC in South Africa and is one of the major “hotspots” of ESCC worldwide. I had received an approval from the Heath Research Ethics Committee (HREC number: S18/10/250) of Stellenbosch University for the study before the Covid-19 pandemic started. I was busy ordering supplies and getting ready to start the study participant recruitment. My plans to data collection process involved traveling to Frere hospital to facilitate the data and sample collection, storage, and subsequent transportation to Stellenbosch University for sample preparation and analysis. I intended to officially start data and sample collection in March of 2020, which coincided with several changes which were happening in the country and at Stellenbosch University regarding COVID-19.

The first case of COVID-19 in South Africa was recorded on the 6th of March 2020, and immediately after (15 March 2020) COVID-19 was declared a national disaster. On the 27th of March, the country was officially put on lockdown, with lockdown regulations and travel (international and in-country) restrictions being instituted. This meant that travel to the Eastern Cape Province for my data collection became impossible. During that time, Stellenbosch University also held a contingency meeting (25 March 2020) to discuss university regulations. The Contingency Committee for Research at Stellenbosch University issued a statement to all researchers which read;

"There will be a postponement of all research activities at Stellenbosch University, apart from research that can be conducted remotely/online and requires no human contact, and research in those areas specifically acknowledged as essential services by the South African government under the presidential regulations related to Covid-19 (e.g., clinical studies)."

This meant I could no longer carry out the data and sample collection since my research required patient contact and interaction. It would have also required travel to a different province, from the Western Cape Province to the Eastern Cape Province, which was not allowed during the initial lockdown period. At the time of these new developments nationally and at the University I had completed one systematic review (Chapter 2) on oesophageal cancer on the genetics of EC in Africa with the title "Systematic Review of Genetic Factors in the Etiology of Esophageal Squamous Cell Carcinoma in African Populations" and published in Frontiers in Genetics(2) (DOI: 10.3389/fgene.2019.00642). We were still finalizing the second systematic review (Chapter 3) on the environmental and lifestyle risk factors of EC in Africa, with the title "Environmental and life-style risk factors for esophageal squamous cell carcinoma in Africa: A systematic review".

I had also made good progress with the primary case-control research study by finalising the protocol, getting ethics approval from the University as well as the Eastern Cape Provincial Department of Health. We planned to administer the questionnaire using an application, the Mobenzi app, and had purchased the devices that we were going to use for the study. We had ordered the supplies needed for sample collection and recruited two nurses who would assist. I had also travelled to Frere hospital and met with the gastroenterologist who was to assist on the study. Unfortunately, the plans came to a halt following the implementation of the new research regulations at the university. We did not know how long the regulations would last. I, therefore, decided to explore ways in which to continue and finish my PhD thesis during the pandemic.

The decision to revise objectives was not easy, as I had made significant progress in preparing for data collection. Additionally, being an international student facing a lot of uncertainty regarding the future of my studies was difficult. I had meetings with my supervisors (Prof. Vikash Sewram, Prof. Helena Kuivaniemi, and Dr. Christian Abnet), the Director of the Doctoral Office at Stellenbosch University (Dr. J Chabilall), and my funders (CEBHA+ and Margaret McNamara personnel) who guided and supported me on pursuing an alternative set of objectives, as well as reaching out to potential collaborators for the revised objectives. I had meetings with several potential collaborators to discuss collaborations on the PhD study using existing data and developed a good network of mentors and colleagues in the process. The revised objectives included extensive bioinformatic analyses on existing GEO datasets, which then became the primary study.

The gene expression study proposal had to be approved by the Faculty Postgraduate Committee, the Health Research Ethics Committee as well as the Faculty Senate. This process took a few months, during which time, I went through extensive training in bioinformatics analyses, the R language, essential Linux usage and bash scripting, using high performance computers and how to write scripts for computational analyses. I attended the Bioinformatics Summer School training as well as the Human Genomic Epidemiology in African populations course offered by the Wellcome Genome Campus. I also attended webinars on bioinformatics offered by Bioconductor and other institutions. Throughout this process, I have continually received bioinformatics training and mentorship from Prof Gerard Tromp, who became an additional supervisor for my PhD thesis. I also learned approaches to dealing with situations where the size of the dataset exceeds the computational resources: in my research I dealt with that by resampling to provide more stable estimates of the effects.

I also attended the Cancer Epidemiology module offered as part of the MPhil (Cancer Science) Programme and received statistical software training from Prof Birhanu Ayele, who unfortunately passed away from COVID-19-related complications in March 2021.

I embarked on a passion project with a colleague during this time. We wrote a perspective manuscript on COVID-19 and women's health. This side project allowed me to take a mental break from the stress that came from waiting for approval of the revised objectives and to focus on something different. The article was published in Frontiers in Global Women's Health(28) (https://doi.org/10.3389/fgwh.2020.570666). The article was also picked up by the local news site, News24, and published(29) (https://www.news24.com/news24/columnists/guestcolumn/opinion-covid-19-and-the-impact-on-women-20201202).

Not being able to continue with the study plan we had planned for two years was stressful. What made the whole process better was the support from my supervisors. Their support and guidance made the continuation of the PhD work during a pandemic feasible. Doing a Ph.D. during a pandemic is also stressful and comes with its unique pressures. Throughout this process, I believe I gained more than I lost.

I. I made a new network of colleagues and potential future collaborators through the meetings I had as I planned to revise objectives.
II. I was able to learn and gain new skills in cancer epidemiology, bioinformatics and statistics that I would have otherwise not gained.
III. I was able to perform analysis on EAC, and the precancerous lesions, BE, and interpret the biological mechanisms driving tumorigenesis in EAC, on top of the ESCC analysis I had originally planned.
IV. I learned the importance of being flexible in research. There are unforeseen events that can happen, and it is important to be able to be flexible enough for change.

Although the objectives were revised, the main theme of the study remained the same - the genetic and environmental factors associated with EC. I am grateful for the support of my supervisors in mentoring and guiding me as I finish my PhD during a pandemic.

5.5 References

1. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: a cancer journal for clinicians* (2018) 68(6):394-424.

2. Simba H, Kuivaniemi H, Lutje V, Tromp G, Sewram V. Systematic Review of Genetic Factors in the Etiology of Esophageal Squamous Cell Carcinoma in African Populations. *Front Genet* (2019) 10:642. Epub 2019/08/21. doi: 10.3389/fgene.2019.00642. PubMed PMID: 31428123; PubMed Central PMCID: PMCPMC6687768.

3. Bilton TP, McEwan JC, Clarke SM, Brauning R, van Stijn TC, Rowe SJ, et al. Linkage Disequilibrium Estimation in Low Coverage High-Throughput Sequencing Data. *Genetics* (2018) 209(2):389. doi: 10.1534/genetics.118.300831.

4. Asombang AW, Chishinga N, Nkhoma A, Chipaila J, Nsokolo B, Manda-Mapalo M, et al. Systematic review and meta-analysis of esophageal cancer in Africa: Epidemiology, risk factors, management and outcomes. *World J Gastroenterol* (2019) 25(31):4512-33. Epub 2019/09/10. doi: 10.3748/wjg.v25.i31.4512. PubMed PMID: 31496629; PubMed Central PMCID: PMCPMC6710188.

5. Page MJ, Shamseer L, Tricco AC. Registration of systematic reviews in PROSPERO: 30,000 records and counting. *Syst Rev* (2018) 7(1):32. Epub 2018/02/22. doi: 10.1186/s13643-018-0699-4. PubMed PMID: 29463298; PubMed Central PMCID: PMCPMC5819709.

6. Prabhu A, Obi KO, Rubenstein JH. The Synergistic Effects of Alcohol and Tobacco Consumption on the Risk of Esophageal Squamous Cell Carcinoma: A Meta-Analysis. *Official journal of the American College of Gastroenterology | ACG* (2014) 109(6):822-7. doi: 10.1038/ajg.2014.71. PubMed PMID: 00000434-201406000-00010.

7. Laytragoon-Lewin N, Bahram F, Rutqvist LE, Turesson I, Lewin F. Direct effects of pure nicotine, cigarette smoke extract, Swedish-type smokeless tobacco (Snus) extract and ethanol on human normal endothelial cells and fibroblasts. *Anticancer Res* (2011) 31(5):1527-34. Epub 2011/05/28. PubMed PMID: 21617206.

8. Kayamba V, Heimburger DC, Morgan DR, Atadzhanov M, Kelly P. Exposure to biomass smoke as a risk factor for oesophageal and gastric cancer in low-income populations: A systematic review. *Malawi Med J* (2017) 29(2):212-7. Epub 2017/09/29. doi: 10.4314/mmj.v29i2.25. PubMed PMID: 28955435; PubMed Central PMCID: PMCPMC5610298.

9. Okello S, Akello SJ, Dwomoh E, Byaruhanga E, Opio CK, Zhang R, et al. Biomass fuel as a risk factor for esophageal squamous cell carcinoma: a systematic review and meta-analysis. *Environmental Health: A Global Access Science Source* (2019) 18(1):60-. doi: 10.1186/s12940-019-0496-0. PubMed PMID: 31262333.

10. Bruce N, Rehfuess E, Mehta S. *Indoor Air Pollution.* 2nd ed. Jamison DT, Breman JG, Measham AR, editors. Washington (DC): The International Bank for Reconstruction and Development / The World Bank (2006).

11. Ezzati M, Saleh H, Kammen DM. The contributions of emissions and spatial microenvironments to exposure to indoor air pollution from biomass combustion in Kenya. *Environ Health Perspect* (2000) 108(9):833-9. Epub 2000/10/06. doi: 10.1289/ehp.00108833. PubMed PMID: 11017887; PubMed Central PMCID: PMCPMC2556923.

12. Okello G, Devereux G, Semple S. Women and girls in resource poor countries experience much greater exposure to household air pollutants than men: Results from Uganda and Ethiopia. *Environment international* (2018) 119:429-37. doi: https://doi.org/10.1016/j.envint.2018.07.002.

13. Liu J, Wang J, Leng Y, Lv C. Intake of fruit and vegetables and risk of esophageal squamous cell carcinoma: a meta-analysis of observational studies. *Int J Cancer* (2013) 133(2):473-85. Epub 2013/01/16. doi: 10.1002/ijc.28024. PubMed PMID: 23319052.

14. Zhao W, Liu L, Xu S. Intakes of citrus fruit and risk of esophageal cancer: A meta-analysis. *Medicine (Baltimore)* (2018) 97(13):e0018. Epub 2018/03/30. doi: 10.1097/md.0000000000010018. PubMed PMID: 29595629; PubMed Central PMCID: PMCPMC5895383.

15. Chetwood JD, Garg P, Finch P, Gordon M. Systematic review: the etiology of esophageal squamous cell carcinoma in low-income settings. *Expert Rev Gastroenterol Hepatol* (2019) 13(1):71-88. Epub 2019/02/23. doi: 10.1080/17474124.2019.1543024. PubMed PMID: 30791842.

16. Smyth EC, Lagergren J, Fitzgerald RC, Lordick F, Shah MA, Lagergren P, et al. Oesophageal cancer. *Nature reviews Disease primers* (2017) 3:17048. Epub 2017/07/28. doi: 10.1038/nrdp.2017.48. PubMed PMID: 28748917.

17. Shafi AA, Knudsen KE. Cancer and the Circadian Clock. *Cancer Research* (2019) 79(15):3806-14. doi: 10.1158/0008-5472.Can-19-0566.

18. Callahan ZM, Shi Z, Su B, Xu J, Ujiki M. Genetic variants in Barrett's esophagus and esophageal adenocarcinoma: a literature review. *Diseases of the Esophagus* (2019) 32(8). doi: 10.1093/dote/doz017.

19. Taylor PR, Abnet CC, Dawsey SM. Squamous dysplasia--the precursor lesion for esophageal squamous cell carcinoma. *Cancer Epidemiol Biomarkers Prev* (2013) 22(4):540-52. Epub 2013/04/04. doi: 10.1158/1055-9965.Epi-12-1347. PubMed PMID: 23549398; PubMed Central PMCID: PMCPMC3681095.

20. Chung CS, Lee YC, Wu MS. Prevention strategies for esophageal cancer: Perspectives of the East vs. West. *Best Pract Res Clin Gastroenterol* (2015) 29(6):869-83. Epub 2015/12/15. doi: 10.1016/j.bpg.2015.09.010. PubMed PMID: 26651249.

21. Liu X, Zhang M, Ying S, Zhang C, Lin R, Zheng J, et al. Genetic Alterations in Esophageal Tissues From Squamous Dysplasia to Carcinoma. *Gastroenterology* (2017) 153(1):166-77. Epub 2017/04/04. doi: 10.1053/j.gastro.2017.03.033. PubMed PMID: 28365443.

22. Matejcic M, Gunter MJ, Ferrari P. Alcohol metabolism and oesophageal cancer: a systematic review of the evidence. *Carcinogenesis* (2017) 38(9):859-72. Epub 2017/06/25. doi: 10.1093/carcin/bgx067. PubMed PMID: 28645180.

23. Yao C, Narumiya S. Prostaglandin-cytokine crosstalk in chronic inflammation. *Br J Pharmacol* (2019) 176(3):337-54. Epub 2018/11/02. doi: 10.1111/bph.14530. PubMed PMID: 30381825; PubMed Central PMCID: PMCPMC6329627.

24. Choi J, Jia G, Wen W, Long J, Zheng W. Evaluating polygenic risk scores in assessing risk of nine solid and hematologic cancers in European descendants. *International Journal of Cancer* (2020) 147(12):3416-23. doi: https://doi.org/10.1002/ijc.33176.

25. Dong J, Buas MF, Gharahkhani P, Kendall BJ, Onstad L, Zhao S, et al. Determining Risk of Barrett's Esophagus and Esophageal Adenocarcinoma Based on Epidemiologic Factors and Genetic Variants. *Gastroenterology* (2018) 154(5):1273-81.e3. Epub 2017/12/17. doi: 10.1053/j.gastro.2017.12.003. PubMed PMID: 29247777; PubMed Central PMCID: PMCPMC5880715.

26. Zhang YD, Hurson AN, Zhang H, Choudhury PP, Easton DF, Milne RL, et al. Assessment of polygenic architecture and risk prediction based on common variants across fourteen cancers. *Nature Communications* (2020) 11(1):3353. doi: 10.1038/s41467-020-16483-3.

27. Florez H, Singh S. Online dashboard and data analysis approach for assessing COVID-19 case and death data. *F1000Res* (2020) 9:570-. doi: 10.12688/f1000research.24164.1. PubMed PMID: 32884676.

28. Simba H, Ngcobo S. Are Pandemics Gender Neutral? Women's Health and COVID-19. *Frontiers in Global Women's Health* (2020) 1(8). doi: 10.3389/fgwh.2020.570666.

29. Simba H, Ngcobo S. OPINION | Covid-19 and the impact on women. *News24* (2020) 2 December 2020.

Chapter 6: Conclusions and implications for future research

Eliminating the burden of EC needs a multidisciplinary approach. A lot still needs to be done to understand the risk factors driving development and progression, as well as the pathobiology of EC, particularly in African populations. Our study identified gaps in the literature that can be addressed by future studies. These include:

1. More and larger studies analysing the association between genetic variants and ESCC in African populations.
2. Investment in WGS and GWAS studies on EC in Africa. We are aware of two GWAS studies which are currently underway under the Johannesburg Cancer Study (South Africa), and the ESCC African Prevention Research network (ESCCAPE) study (Malawi, Tanzania, Kenya).
3. More studies on ESCC transcriptomics in African populations that can elucidate ESCC pathobiology.
4. Further assessment of novel pathways identified in our gene expression analysis.
5. More studies on squamous cell dysplasia and its progression to ESCC. Additionally, the potential impact of endoscopic surveillance in ESCC "hotspot" areas.
6. Incorporate standardised methods for data and biological sample collection, analysis, and reporting. We found this to be a major issue in our critical appraisal of the studies in our systematic reviews.
7. Assess the new and emerging environmental and lifestyle risk factors identified in our systematic review including esophageal injury due to hot food and beverage consumption, and oral health. Further studies on the role of esophageal injury on ESCC as well as squamous dysplasia are also needed.
8. Interaction studies to investigate the multifactorial etiology of ESCC.
9. Polygenic risk scores for both precancerous lesions and tumours.

EC has a multifactorial and complex etiology, and this study provided compelling evidence on the role of genetic, environmental, and lifestyle, factors, as well as EC pathobiology. Investigating the complex interplay of risk factors will go a long way in elucidating the disease effect, particularly focusing on what is driving its development, why there is a distinct geographical delineation in incidence, and why it is presenting in younger people in African populations. Understanding EC requires an integrated approach to determine which risk factors interact with each other. We recommend that future studies incorporate study designs that will analyze the multifactorial etiology of EC, particularly in African populations. The lack of studies in African populations limits our understanding of the disease. EC constitutes a severe public health burden in high incidence areas and requires research prioritization as well as national and international intervention strategies.

List of Appendices

Appendix File: Published Systematic Review

frontiers in Genetics

SYSTEMATIC REVIEW
published: 02 August 2019
doi: 10.3389/fgene.2019.00642

Systematic Review of Genetic Factors in the Etiology of Esophageal Squamous Cell Carcinoma in African Populations

Hannah Simba [1]***, Helena Kuivaniemi*** [2]***, Vittoria Lutje*** [3]***, Gerard Tromp*** [2,4,5,6,7] ***and Vikash Sewram*** [1]*

[1] African Cancer Institute, Division of Health Systems and Public Health, Department of Global Health, Faculty of Medicine and Health Sciences, Stellenbosch University, Cape Town, South Africa, [2] Division of Molecular Biology and Human Genetics, Department of Biomedical Sciences, Faculty of Medicine and Health Sciences, Stellenbosch University, Cape Town, South Africa, [3] Cochrane Infectious Diseases Group, Liverpool, United Kingdom, [4] Bioinformatics Unit, South African Tuberculosis Bioinformatics Initiative, Stellenbosch University, Cape Town, South Africa, [5] DST–NRF Centre of Excellence for Biomedical Tuberculosis Research, Stellenbosch University, Cape Town, South Africa, [6] South African Medical Research Council Centre for Tuberculosis Research, Stellenbosch University, Cape Town, South Africa, [7] Centre for Bioinformatics and Computational Biology, Stellenbosch University, Stellenbosch, South Africa

Background: Esophageal squamous cell carcinoma (ESCC), one of the most aggressive cancers, is endemic in Sub-Saharan Africa, constituting a major health burden. It has the most divergence in cancer incidence globally, with high prevalence reported in East Asia, Southern Europe, and in East and Southern Africa. Its etiology is multifactorial, with lifestyle, environmental, and genetic risk factors. Very little is known about the role of genetic factors in ESCC development and progression among African populations. The study aimed to systematically assess the evidence on genetic variants associated with ESCC in African populations.

Methods: We carried out a comprehensive search of all African published studies up to April 2019, using PubMed, Embase, Scopus, and African Index Medicus databases. Quality assessment and data extraction were carried out by two investigators. The strength of the associations was measured by odds ratios and 95% confidence intervals.

Results: Twenty-three genetic studies on ESCC in African populations were included in the systematic review. They were carried out on Black and admixed South African populations, as well as on Malawian, Sudanese, and Kenyan populations. Most studies were candidate gene studies and included DNA sequence variants in 58 different genes. Only one study carried out whole-exome sequencing of 59 ESCC patients. Sample sizes varied from 18 to 880 cases and 88 to 939 controls. Altogether, over 100 variants in 37 genes were part of 17 case-control genetic association studies to identify susceptibility loci for ESCC. In these studies, 25 variants in 20 genes were reported to have a statistically significant association. In addition, eight studies investigated changes in cancer tissues and identified somatic alterations in 17 genes and evidence of loss of heterozygosity, copy number variation, and microsatellite instability. Two genes were assessed for both genetic association and somatic mutation.

OPEN ACCESS

Edited by:
Solomon Fiifi Ofori-Acquah,
University of Ghana, Ghana

Reviewed by:
Clara S. Tang,
The University of Hong Kong,
Hong Kong
Marco Matejcic,
University of Southern California,
United States

****Correspondence:***
Vikash Sewram
vsewram@sun.ac.za

Specialty section:
This article was submitted
to Evolutionary and
Population Genetics,
a section of the journal
Frontiers in Genetics

Received: *16 November 2018*
Accepted: *18 June 2019*
Published: *02 August 2019*

Citation:
Simba H, Kuivaniemi H, Lutje V,
Tromp G and Sewram V (2019)
Systematic Review of Genetic
Factors in the Etiology of Esophageal
Squamous Cell Carcinoma
in African Populations.
Front. Genet. 10:642.
doi: 10.3389/fgene.2019.00642

Conclusions: Comprehensive large-scale studies on the genetic basis of ESCC are still lacking in Africa. Sample sizes in existing studies are too small to draw definitive conclusions about ESCC etiology. Only a small number of African populations have been analyzed, and replication and validation studies are missing. The genetic etiology of ESCC in Africa is, therefore, still poorly defined.

Keywords: esophageal squamous cell carcinoma, genetic association, somatic variant, germline mutation, sequence variants, systematic review, African populations

INTRODUCTION

Esophageal cancer is an aggressive and fatal cancer of the 18digestive tract. It accounts for an estimated 455,800 new cases and 400,200 deaths per year globally, making it the eighth most common cancer in the world (Murphy et al., 2017). The malignant tumors are characterized by two major subtypes: esophageal squamous cell carcinoma (ESCC), which is the more common type and contributes 90%, and esophageal adenocarcinoma (EAC) (Kaz and Grady, 2014; Abnet et al., 2017). ESCC presents with poor prognosis and low survival rate (<5%) in low resource settings (Yazbeck et al., 2016; Murphy et al., 2017). The asymptomatic development of ESCC results in diagnosis at late stage for patients and is characterized by dysphagia. At this stage, treatment is limited to palliative care.

ESCC is endemic in specific geographic locations worldwide and has the most divergence in cancer incidence globally, with high prevalence reported in East Asia, Southern Europe, as well as in Eastern and Southern Africa (Abnet et al., 2017). This peculiar distribution draws questions on the specificity of certain risk factors to particular populations. The African ESCC corridor, which includes Ethiopia, Rwanda, Burundi, Malawi, Kenya, Uganda, Tanzania, and South Africa, is an ESCC hotspot region (Munishi et al., 2015; Schaafsma et al., 2015). It has also been reported that in Sub-Saharan Africa, ESCC develops in younger patients than in other regions (Kayamba et al., 2015).

The etiology of esophageal carcinoma is multifactorial. The risk factors reported worldwide comprise several lifestyle and environmental and genetic factors (Pink et al., 2011; Sewram et al., 2014; Chen et al., 2015; Sewram et al., 2016; Huang and Yu, 2018). Growing evidence supports the hypothesis that genomic alterations and epigenetic modifications contribute to tumor development (Baba et al., 2017). ESCC has both an inherited and cellular genetic basis (Abnet et al., 2017; Coleman et al., 2018). Familial syndromes associated with increased risk of malignancy include tylosis and Fanconi anemia (Abnet et al., 2017). The majority of genetic studies on ESCC have been case-control association studies analyzing single-nucleotide polymorphisms (SNPs) in various candidate genes. However, the reproducibility of these studies has been low. Some of the more common SNPs associated with ESCC have been identified in the aldehyde dehydrogenase 2 family gene (*ALDH2)* and an acetaldehyde dehydrogenase gene *(ADH1B)* (Abnet et al., 2017). Variants in these genes have been shown to increase susceptibility to ESCC development, and they are also associated with alcohol consumption (Abnet et al., 2017). Two meta-analyses published in 2018 reported associations between the genes *MTHFR* and *GSTT1* and esophageal cancer development (He et al., 2018; Kumar and Rai, 2018). However, the meta-analyses were done on predominantly Asian and Western populations. In recent years, the focus of ESCC research in the Western and Asian countries has shifted from candidate gene studies to genome-wide association studies (GWAS) and whole-exome sequencing (WES) to identify variants associated with ESCC. Combined analysis of different study designs has provided a better understanding of ESCC etiology in Asian populations (Abnet et al., 2017). Genes with variants implicated in the development of ESCC in these populations include phospholipase c epsilon 1 *(PLCE1)*, caspase 8 *(CAP8)*, tumor protein 53 *(TP53)*, and human leukocyte antigen *(HLA)* (Abnet et al., 2017).

The genetic etiology of ESCC in Africa is not well understood, since there have been very few studies on ESCC in African populations. This is in part due to the unavailability of adequate research infrastructure. A lack of comprehensive assessment and validation of existing evidence through systematic reviews has also contributed to this knowledge gap. A number of small studies on African populations have yielded varied associations between genetic variants and ESCC. There is, therefore, a need to systematically assess the current evidence in order to map out the contribution of genetic factors in the development of ESCC in African populations using critically appraised data.

The aim of the current systematic review was to assess all genetic (cross-sectional, case-control, and cohort) studies reporting on germline and somatic variants where risk factor estimates were calculated. This was achieved through the following: 1) critical appraisal of African literature on association of genetic factors to ESCC development; 2) comprehensive analysis of genetic (germline and somatic) variants in the reported studies; 3) data synthesis through pooled analysis, if feasible; and 4) comparison of genetic variants identified in African populations to those reported in other geographic regions.

MATERIALS AND METHODS

We followed the Preferred Reporting Items for Systematic Reviews and Meta-Analyses guidelines (PRISMA) (Little et al., 2009). However, because PRISMA is not a quality assessment tool, other instruments were used to assess quality control.

Data Sources and Search Strategy

We carried out a literature search on all published African ESCC studies up to April 2019. We developed a comprehensive set of search terms subjectively and iteratively. We searched the following electronic bibliographic databases without time or language limits: Medline (PubMed), Embase (OViD), Scopus, African Index Medicus, and Africa-wide information (EbsCOHost). We also checked the reference lists of potentially relevant articles for additional citations and used the "related citations" search key in PubMed to identify similar papers.

We checked Medline (PubMed) to identify controlled vocabulary (MeSH) terms related to esophageal cancer and also identified text keywords based on our knowledge of the field (**Table 1**). Medline search terms were modified for other electronic databases to conform to their search functions.

Screening for eligible studies was carried out by two authors (HS and HK). First, the two authors read the titles and abstracts independently and then met to finalize an initial list. Full articles of the studies selected based on the initial screening were read and assessed for inclusion to the systematic review. **Figure 1** shows the outline for selection of eligible studies.

Quality Control and Data Extraction

Quality of the methodology used in the published studies was assessed using a quality assessment tool adapted from the STrengthening the REporting of Genetic Association studies (STREGA) statement (Little et al., 2009). The quality assessment for genetic association studies to identify ESCC susceptibility loci included reporting on power calculations, detailed population characteristics for cases, description of ESCC diagnosis, screening of cases and controls, reporting a measure of association using odds ratios, adjustment of population stratification, assessment of genotyping error, reporting the Hardy–Weinberg equilibrium, correction for multiple testing, and reporting of National Center for Biotechnology Information (NCBI) rs numbers for variants (**Table S1**).

TABLE 1 | Medline (PubMed) search strategy to identify published African ESCC literature.

#1	Search cancer or carcinoma or neoplasm* Field: Title/Abstract
#2	Search Esophageal or oesophageal Field: Title/Abstract
#3	#1 and #2
#4	Search "Esophageal cancer" Field: Title/Abstract
#5	Search "oesophageal cancer" or "oesophageal neoplasm*" Field: Title/Abstract
#6	Search "Esophageal Neoplasms"[Mesh]
#7	Search "Esophageal Neoplasms" Field: Title/Abstract
#8	Search "Esophageal squamous cell carcinoma" or "oesophageal squamous cell carcinoma" or ESCC Field: Title/Abstract
#9	Search ((((#3) OR #4) OR #5) OR #6) OR #7 OR #8
#10	Search "Africa"[Mesh]
#11	Search algeria OR angola OR benin OR botswana OR burkina faso OR burundi OR cameroon OR cape verde OR central african republic OR chad OR comoros OR congo OR "Democratic Republic of Congo" OR DRC OR djibouti OR equatorial guinea OR egypt OR eritrea OR ethiopia OR gabon OR gambia OR ghana OR guinea OR bissau OR ivory coast OR (Côte d' Ivoire) OR jamahiriya OR kenya OR lesotho OR liberia OR Libya OR madagascar OR malawi OR mali OR mauritania OR mauritius OR mayotte OR morocco OR mozambique OR namibia OR niger OR nigeria OR principe OR reunion OR rwanda OR "Sao Tome" OR senegal OR seychelles OR "Sierra Leone" OR somalia OR "South Africa" OR st helena OR sudan OR swaziland OR tanzania OR togo OR tunisia OR uganda OR zaire OR zambia OR zimbabwe OR "Central Africa" OR "West Africa" OR "East Africa" OR "Southern Africa" OR "South Africa" Field: Title/Abstract
#12	Search (#10) or #11
#13	Search (#9) AND #12

For somatic mutation studies, quality assessment included the following: description of ESCC diagnosis, reporting of tissues used [cancerous (Ca) and normal neighboring tissue (NET)], detailed population characteristics, variant classification and type, confirmation of variants identified, reporting of amino acid change, and use of pathogenicity scoring (**Table S2**).

Data extraction was carried out by two authors (HS and HK) using data extraction forms. Two separate extraction forms were prepared for the germline (genetic susceptibility) and somatic mutation studies. The data extraction form for the genetic susceptibility studies included the following: description of the population (age, sex, sample size, smoking, and alcohol use for cases and controls separately), genotyping method, statistical analysis test, minor allele frequency (MAF), genotype frequency, haplotype frequency, and environmental association frequency. The somatic mutation study extraction form had the same variables excluding gene–environment interaction frequency and haplotype frequency.

The South African Admixed Population is reported as mixed ancestry in the tables according to how it was reported in the articles.

Data Analysis

A meta-analysis could not be performed as there were only two SNPs analyzed in more than one study and even those were analyzed in only two independent studies. For a meta-analysis to be carried out, SNPs have to be assessed in at least three separate case-control studies. *TP53* in the somatic variant studies was analyzed in four separate studies, but two of the studies had cases only with no controls, and the remaining two assessed different parts of the gene. The results of this systematic review will, therefore, be reported in a descriptive manner.

We were able to find rs numbers for most of the variants even if the authors of the original studies did not report them and have included them in the tables of this systematic review. We used the canonical SNP identifier (rs number) and dbSNP (version 152; April 2019) database at NCBI (https://www.ncbi.nlm.nih.gov/snp/) for this. We also determined the locus positions of the microsatellite markers reported in a study by Naidoo et al. (2005) using the primer-BLAST database at NCBI (https://www-ncbi-nlm-nih-gov.ez.sun.ac.za/tools/primer-blast).

To determine the linkage disequilibrium (LD) measures between the SNPs reported in the same genes, we obtained the imputed data set from the Thousand Genomes project (1000 Genomes Release Phase 3 2013-05-02) and used bcftools to extract all individuals from African populations, not including African Americans, and the 77 SNPs discussed here using all synonyms (alternative rs IDs) for SNPs (Auton et al., 2015). We obtained a dataset of 504 individuals and 67 SNPs. We computed all pair-wise r^2-values using PLINK (v1.09) (Danecek et al., 2011; Chang et al., 2015).

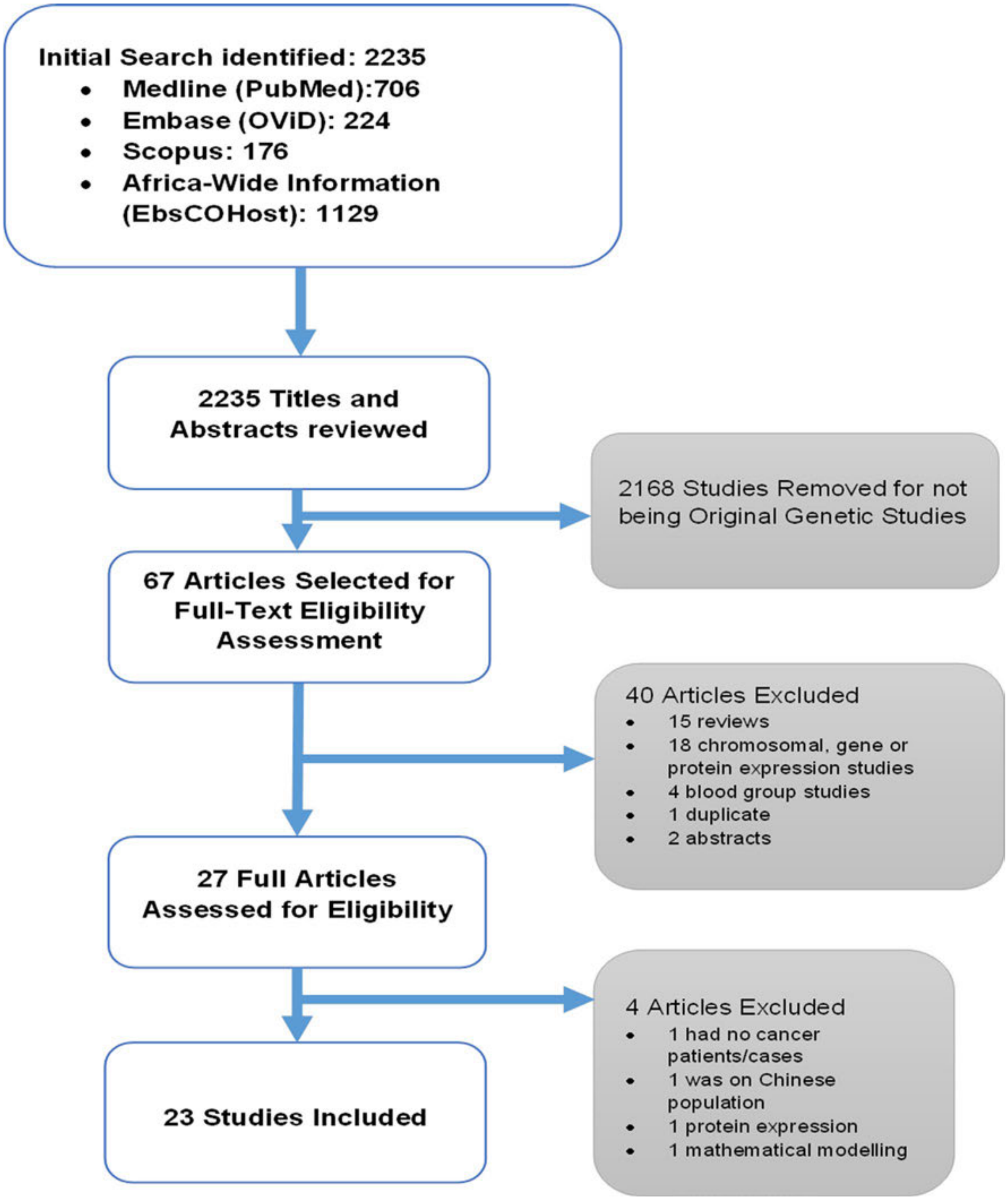

Initial Search identified: 2235
• Medline (PubMed):706
• Embase (OViD): 224
• Scopus: 176
• Africa-Wide Information (EbsCOHost): 1129
2235 Titles and Abstracts reviewed
2168 Studies Removed for not being Original Genetic Studies
67 Articles Selected for Full-Text Eligibility Assessment
40 Articles Excluded
• 15 reviews
• 18 chromosomal, gene or protein expression studies
• 4 blood group studies
• 1 duplicate
• 2 abstracts
27 Full Articles Assessed for Eligibility
4 Articles Excluded
• 1 had no cancer patients/cases
• 1 was on Chinese population
• 1 protein expression
• 1 mathematical modelling
23 Studies Included

FIGURE 1 | Outline of the systematic review.

RESULTS

Systematic Review Outline

The selection process for all the included studies is shown in **Figure 1**. The initial database search identified 2,235 articles. Titles and abstracts of these articles were reviewed, and 2,168 studies were removed for not being original genetic studies. The 67 articles that remained were selected for full-text eligibility assessment. This process resulted in the removal of 40 articles: 15 review articles, 18 chromosomal, gene or protein expression studies, 4 blood group studies, 1 duplicate, and 2 abstracts. A total of 27 full articles were then assessed for eligibility, and four articles were removed for not meeting the criteria, as follows: one study had no cancer patients/cases (Adams et al., 2003), one focused on the Chinese population (Li et al., 2016), while one focused on protein expression (Jaskiewicz and De Groot, 1994;

Huang and Yu, 2018), and the other was a mathematical model study (Uys and Van Helden, 2003). In the end, 23 studies were included and analyzed in the systematic review.

Study Characteristics

The characteristics of all the genetic susceptibility and somatic variant studies included are shown in **Tables 2** and **3**, respectively. The 23 studies included in the study were published between 1990 and 2019. There were 17 genetic susceptibility and eight somatic variant studies. Two studies reported on both genetic susceptibility and somatic variants.

Genetic Susceptibility Studies

The 17 genetic susceptibility studies (**Table 2**) were all case-control studies (Dietzsch et al., 2003; Vos et al., 2003; Dandara et al., 2005; Li et al., 2005; Zaahl et al., 2005; Chelule et al., 2006; Dandara et al., 2006; Li et al., 2008; Li et al., 2010; Bye et al., 2011; Matejcic et al., 2011; Bye et al., 2012; Eltahir et al., 2012; Strickland et al., 2012; Vogelsang et al., 2012; Matejcic et al., 2015; Chen et al., 2019) published between 2003 and 2019. Sixteen articles reported on the South African population and one article on the Sudanese population. The majority (13/17; 76%) of the studies reported on the main subject characteristics (ethnicity, sex, age, and type of clinical assessment). Sample sizes for ESCC patients ranged from 18 to 880 with six of the studies having over 200 patient samples. Sample sizes for controls ranged from 88 to 939 with nine of the studies having over 200 control samples. It is difficult to estimate the total number of patients analyzed in these 17 studies, since it appears that the same authors used the same sample set for different SNPs in different publications. Our assessment showed that Bye et al. (2011) and Bye et al. (2012) used the same participants. In addition, studies by Li et al. (2005) and Li et al. (2008) used the same participants as Dandara et al. (2005). The remaining 12 studies do not seem to have any obvious sample overlap.

Altogether, 16 out of 17 studies clinically assessed for ESCC through histology. None of the studies clinically assessed controls for ESCC with the exception of one study (Strickland et al., 2012), which assessed controls using a brush biopsy. Nine studies reported on smoking and alcohol consumption status for all participants (Dandara et al., 2005; Li et al., 2005; Dandara et al., 2006; Li et al., 2008 Li et al., 2010; Bye et al., 2012; Vogelsang et al., 2012; Matejcic et al., 2015; Chen et al., 2019), while three (Bye et al., 2011; Matejcic et al., 2011; Strickland et al., 2012) reported those risk factors for only the ESCC patients.

The Hardy−Weinberg equilibrium deviation was assessed in 11 (65%) studies; however, only six (35%) of the studies reported power calculations, and three (18%) studies reported the evaluation of a genotyping error. Detailed characteristics of the study population were reported in 12 of the studies for cases and 10 for controls. Correction for multiple testing was reported in only seven (41%) studies. NCBI rs numbers were reported in eight (47%) studies. Our quality assessment scoring had 11 items (**Table S1**), and each item had a weight of 1 point; therefore, total maximum quality score was 11. Overall, only seven of the 17 (41%) studies scored half or above half (5.5). The highest score was 9 (Vogelsang et al., 2012; Chen et al., 2019), and the lowest score was 1 (Vos et al., 2003; Zaahl et al., 2005).

Somatic Variant Studies

Somatic variant studies (**Table 3**) constituted of eight studies published between 1990 and 2016 (Victor et al., 1990; Gamieldien et al., 1998; Dietzsch and Parker, 2002; Dietzsch et al., 2003; Vos et al., 2003; Naidoo et al., 2005; Patel et al., 2011; Liu et al., 2016). A total of 455 patients were assessed, with the control group comprising 200 NET and 146 blood samples. Of the 455 patient samples, one was reported to be an adenocarcinoma from one study; therefore, the exact ESCC patient population was 454. The study populations were from South Africa, Kenya, and Malawi.

Clinical diagnosis of ESCC was determined by histology in five (75%) studies, and the remaining three did not report on how clinical assessment was done. Four (50%) studies reported using both cancer tissue and NET for assessment. Three of these studies had an equal number of cancer tissue and NET samples. Two (25%) studies did not have any control samples, and the remaining two (25%) studies collected blood samples only as controls. Only two studies reported on smoking and alcohol consumption status. On patient characteristics, age and sex were reported in six (75%) of the studies. Variant classification and type were reported in all of the studies, but confirmation of results was reported in only two studies. No studies used pathogenicity scoring. Amino acid change was also reported in only two of the studies. Our quality assessment score had seven items (**Table S2**), and each item had a weight of 1 point; therefore, total maximum score for the quality assessment was 7. Overall, six of the eight (75%) studies scored half or above half (3.5). The highest score was 6 (Gamieldien et al., 1998), and the lowest score was 0 (Victor et al., 1990).

Description of Genes Studied

A total of 58 genes were investigated in the 23 studies, which were selected for the systematic review, with 37 genes studied in the genetic susceptibility studies and 23 in the somatic variant studies. Two genes were investigated in both studies. In addition, the somatic studies investigated six genetic loci without specific gene names. A summary of SNPs analyzed in the genetic susceptibility studies is shown in **Table 4**. Over 100 SNPs were analyzed, and 25 SNPs were reported to be associated with ESCC (four SNPs using p values only, and 21 SNPs using p values and odds ratios). The 25 SNPs were in 20 genes: *ADH1B, ADH3, ALDH2, AR, CASP8, CHEK2, CP, CYP2E1, CYP3A5, GSTT2B, MGMT, MLH3, MSH3, NAT2, PTGS2 (also known as COX-2), PLCE1, PMS1, RUNX1, SLC11A1, and TP53*. The associations with all 25 SNPs were identified in South African populations, while none were found in the Sudanese population.

Table 5 shows a summary of the pathways for the 20 genes. All the genes encode for proteins. Three of the genes, *ADH1B, ADH3,* and *ALDH2*, are involved in alcohol metabolism (Li et al., 2008; Bye et al., 2011). Three mismatch repair genes, *MLH3, MSH3,* and *PMS1*, play a role in genomic integrity (Vogelsang et al., 2012). They are reported to also play a role in carcinogenesis. MGMT is involved in cell defense against mutagens, and mutations in the gene are reported to be associated with cancer formation (Bye et al., 2011). *NAT2* and *GSTT2B* play a role in the activation and deactivation of drugs and carcinogens, with reports of mutations

TABLE 2 | Characteristics of genetic susceptibility studies for ESCC in African populations.

Study (PMID)	Location	Year	Population	Age, y (SD)		Sample size		Sex, cases n (%)		Sex, ctrl n (%)		Clinical assessment		Analysis method	Smoking n (%)		Alcohol n (%)	
				Cases	Ctrl	Cases	Ctrl	Male	Female	Male	Female	Cases	Ctrl		Cases	Ctrl	Cases	Ctrl
Bye et al., 2011 (21926110)	South Africa	2011	Black	59.8 (11.3)	–	358	477	182 (50.8)	176 (49.2)	–	–	Histology	–	TaqMan Assay	228 (63.7)	–	228 (63.7)	–
			Mixed ancestry	60.5 (10.6)	–	201	427	131 (65.2)	70 (34.8)	–	–	Histology	–	TaqMan Assay	189 (94.1)	–	163 (81.1)	–
Bye et al., 2012 (22865593)	South Africa	2012	Black	59.8 (11.3)	48.8 (16.7)	407	849	199 (48.9)	208 (51.1)	335 (39.5)	511 (60.2)	Histology	–	TaqMan Assay and KASP	242 (59.5)	333 (39.2)	253 (62.2)	452 (53.2)
			Mixed ancestry	60.6 (10.6)	46.7 (16.8)	257	860	165 (64.2)	91 (35.4)	309 (35.9)	551 (64.1)	Histology	–	TaqMan Assay and KASP	240 (93.4)	597 (69.4)	212 (82.5)	419 (48.7)
Chelule et al., 2006 (17264406)	South Africa	2006	Black	18–74[1]	18–74	70	261	–	–	–	–	Histology	–	PCR-RFLP	–	–	–	–
Chen et al., 2019 (30753320)	South Africa	2019	Black[7]	60.2 (11.3)	48.9 (16.8)	591	852	284 (48.1)	307 (51.9)	342 (40.1)	507 (59.5)	Histology	–	TaqMan Assay	364 (61.6)	338 (39.7)	370 (62.6)	458 (53.7)
			Black[8]	58.2 (10.2)	50.0 (15.5)	880	939	545 (61.9)	332 (37.7)	240 (25.6)	698 (74.3)	Histology		iPLEX and TaqMan Assays	598 (68.0)	333 (35.5)	473 (53.8)	633 (67.4)
Dandara et al., 2005 (15978331)	South Africa	2005	Black	–	–	142	178	–	–	–	–	Histology	–	PCR-RFLP	179	162	171	160
			Mixed ancestry	–	–	99	94					Histology		PCR-RFLP				
Dandara et al., 2006 (16272171)	South Africa	2006	Black	61.23	61.85	145	194	85 (59)	60 (41)	111 (57)	83 (43)	Histology	–	PCR-RFLP	95 (65)	123 (63)	98 (68)	127 (65)
			Mixed ancestry	61.49	69.53	100	94	78 (78)	22 (22)	45 (48)	49 (52)	Histology	–	PCR-RFLP	93 (93)	74 (79)	73 (73)	45 (48)
Dietzsch et al., 2003 (12925954)	South Africa	2003	Black and mixed ancestry	59.6	58.7	58[2]	226	44	14	167	59	–	–	PCR and PAGE				
Eltahir et al., 2012 (23053979)	Sudan	2012				18	235					Histology		PCR-RFLP				
Li et al., 2005 (15899651)	South Africa	2005	Black and mixed ancestry	61.1 (10.5)	65.7 (10.2)	189	198	–	–	–	–	Histology	–	PCR-SSCP and DNA sequencing	144 (76)	122 (62)	133 (70)	114 (58)
Li et al., 2008 (18254707)	South Africa	2008	Black[3]	–	–	142	178	–	–	–	–	Histology	–	PCR- RLFP	179	162	71	160
			Mixed[3] ancestry			101	100					Histology		PCR-RFLP				
Li et al., 2010 (20540773)	South Africa	2010	Black[3]	61.23	61.85	145	194	85 (59)	60 (41)	111 (57)	83 (43)	Histology	–	PCR-RFLP	95 (65)	123 (63)	98 (68)	127 (65)
			Mixed[3] ancestry	61.49	69.53	100	94	78 (78)	22 (22)	45 (48)	49 (52)	Histology	–	PCR- RFLP	93 (93)	74 (79)	73 (73)	45 (48)
Matejcic et al., 2011 (22216261)	South Africa	2011	Black	–	–	330	479	–	–	–	–	Histology	–	TaqMan assay and gel electrophoresis	210	–	204	–
			Mixed ancestry	–	–	232	428	–	–	–	–	Histology	–	TaqMan assay and gel electrophoresis	216	–	189	–
Matejcic et al., 2015 (26447020)	South Africa	2015	Black	59.6 (10.7)	56.7 (15.0)	463	480	229 (49)	234 (51)	235 (49)	245 (51)	Histology	–	TaqMan assay	280 (60)	222 (46)	286 (62)	278 (58)
			Mixed ancestry	60.7 (10.3)	57.7 (14.3)	269	288	177 (66)	92 (34)	178 (62)	110 (38)	Histology	–	TaqMan Assay	250 (93)	226 (78)	215 (80)	172 (60)
Strickland et al., 2012 (21901748)	South Africa	2012	Black	59/66[4]	–	96	88	48	48	–	–	Histology	Brush biopsy	HEX SSCP and DNA sequencing	58	–	58	–
Vogelsang et al., 2012 (22623965)	South Africa	2012	Black	59.8 (11.3)	56.1 (16.2)	345[5]	344	166 (48.1)	179 (51.9)	120 (34.9)	224 (65.1)	Histology	–	Allele-specific quantitative PCR	209 (60.6)	117 (34.0)	160 (46.4)	92 (26.7)
			Mixed ancestry	60.7 (10.2)	56.8 (16.5)	205[6]	266	136 (66.3)	69 (33.7)	82 (30.8)	184 (69.2)	Histology	–	Allele-specific quantitative PCR	189 (92.2)	162 (60.9)	118 (57.6)	38 (14.3)

(Continued)

TABLE 2 | Continued

Study (PMID)	Location	Year	Population	Age, y (SD)		Sample size		Sex, cases n (%)		Sex, ctrl n (%)		Clinical assessment		Analysis method	Smoking n (%)		Alcohol n (%)	
				Cases	Ctrl	Cases	Ctrl	Male	Female	Male	Female	Cases	Ctrl		Cases	Ctrl	Cases	Ctrl
Vos et al., 2003 (12550754)	South Africa	2003	Black	57 (11)	57 (11)	74	118	–	–	–	–	Histology	–	SSCP and DNA sequencing	–	–	–	–
Zaahl et al., 2005 (15860357)	South Africa	2005	Mixed ancestry	–	–	105	110	82	23	43	67	Histology	–	SSCP and DNA sequencing	–	–	–	–

[1]Only range of age was reported for the combined group of cases and controls.

[2]57 had ESCC.

[3]Same population as in Dandara et al. (2005) study.

[4]59+/–13 for male (n = 48) and 66+/– (n = 48) for female patients.

[5]326 had ESCC.

[6]182 had ESCC.

[7]Western and Eastern Cape Province Black Population.

[8]Gauteng Province Black Population.

Ctrl, controls; ESCC, esophageal squamous cell carcinoma; HEX, heteroduplex; KASP, competitive allele specific PCR; PAGE, polyacrylamide gel electrophoresis; PCR, polymerase chain reaction; RFLP, restriction fragment length polymorphism; SD, Standard deviation; SSCP, single-strand conformation polymorphism.

TABLE 3 | Characteristics of studies on somatic changes in ESCC in African populations.

Study (PMID)	Country	Year	Population	Sample size			Age, y (SD)	Sex n (%)		Clinical assessment		Analysis method	Smoking n (%)	Alcohol n (%)
				Ca	NET	Blood	Cases	Male	Female	Ca	NET			
Dietzsch and Parker, 2002 (12435113)	South Africa	2002	Black	33	33	–	57.4	23 (70)	10 (30)	Histology	–	PCR and DNA sequencing analysis	–	–
Dietzsch et al., 2003 (12925954)	South Africa	2003	Black and mixed ancestry	58[1]	58	–	59.6	29 (67)	14 (33)	–	–	PCR and PAGE	–	–
Gamieldien et al., 1998 (9808520)	South Africa	1998	Black	76	9	50	57 (11)	49 (65)	27 (35)	Histology	Histology	PCR and HEX-SSCP	–	–
Liu et al., 2016 (29148985)	Malawi	2016	Malawian	59	–	59	56	27 (45.8)	31 (52.5)	Histology	-	WES	24 (40.7)	14 (23.7)
Naidoo et al., 2005 (15735161)	South Africa	2005	South African	100	100	–	56	53 (54)	45 (46)	Histology	Histology	PCR	–	–
Patel et al., 2011 (22040862)	Kenya	2011	Kenyan	28	–	–	56.03 (12.30)	13 (46)	15 (54)	–	–	PCR and DNA sequencing	6 (21)	10 (36)
Victor et al., 1990 (2199031)	South Africa	1990	Black and mixed ancestry	27	–	–	–	–	–	–	–	PCR and dot blot hybridization	–	–
Vos et al., 2003 (12550754)	South Africa	2003	South African	74	–	37	–	–	–	Histology	–	SSCP and DNA sequencing	–	–

Ca, cancer tissue; HEX-SSCP, heteroduplex single-strand conformation polymorphism; NET, neighboring tissue; PAGE, polyacrylamide gel electrophoresis; PCR, polymerase chain reaction; WES, whole exome sequencing.

[1]57 had ESCC and 1 had adenocarcinoma.

TABLE 4 | Summary of studies investigating genetic susceptibility of ESCC in African populations.

Gene	Variant (rs number)	Study (PMID)	Population	ESCC		Controls		Effect allele	Findings and Comments[2]
				n	MAF	n	MAF		
ADH1B	rs1229984 (Arg48His)	Bye et al., 2011 (21926110)	Black South African	358	0	477	0		Not informative
		Bye et al., 2011 (21926110)	Mixed ancestry South African	201	0.054	427	0.098	A	OR = 0.52 (0.32–0.86) p = 0.009
ADH2	ADH2*1/*2/*3	Li et al., 2008 (18254707)	Black South African	142	0.01	174	0.01		Not informative
		Li et al., 2008 (18254707)	Mixed ancestry South African	96	0.03	94	0.03		Not informative
ADH3	ADH3*1/*2	Li et al., 2008 (18254707)	Black South African	141	0.46	174	0.32		NS
		Li et al., 2008 (18254707)	Mixed ancestry South African	96	0.38	94	0.31	*2	OR = 1.80; p = 0.0004
ADH7	rs1573496 (Gly92Ala)	Bye et al., 2011 (21926110)	Black South African	358	0	477	0.001		Not informative
		Bye et al., 2011 (21926110)	Mixed ancestry South African	201	0.014	427	0.02		NS
ALDH2	rs671 (Glu504Lys)	Bye et al., 2011 (21926110)	Black South African	358	0	477	0		Not informative
		Bye et al., 2011 (21926110)	Mixed ancestry South African	201	0	427	0		Not informative
	rs441 (-261 C/T)	Bye et al., 2011 (21926110)	Black South African	358	0.154	477	0.145		NS
		Bye et al., 2011	Mixed ancestry South African	201	0.18	427	0.194		NS
	rs886205 (+82 A/G)	Bye et al., 2011 (21926110)	Black South African	358	0.247	477	0.252		NS
		Bye et al., 2011 (21926110)	Mixed ancestry South African	201	0.402	427	0.489	G	OR = 0.70 (0.55–0.89); p = 0.004
	ALDH2*1/*2	Li et al., 2008 (18254707)	Black South African	142	0.10	174	0.04	*2	OR = 2.35; p = 0.008
		Li et al., 2008 (18254707)	Mixed ancestry South African	101	0.03	1004	0.04		Not informative
	rs4767364 (A/G)	Chen et al., 2019 (30753320)	Black South African[5]	880	0.12	939	0.11		NS
ALS2CR12	rs13016963 (G/A)	Chen et al., 2019 (30753320)	Black South African[4]	591	0.35	852	0.35		NS
		Chen et al., 2019 (30753320)	Black South African[5]	880	0.39	939	0.38		NS
	rs10201587 (A/G)	Chen et al., 2019 (30753320)	Black South African[5]	880	0.38	939	0.39		NS
AR	CAG-repeat in exon 1	Dietzsch et al., 2003 (12925954)	Black South African males	29		109			NS
		Dietzsch et al., 2003 (12925954)	Mixed ancestry South African males	15		58			NS
	GGC-repeat in exon 1	Dietzsch et al., 2003 (12925954)	Black South African males	29		109		$(GGC)_{\leq 16}$	OR = 2.7 (1.14–6.36); p = 0.018
		Dietzsch et al., 2003 (12925954)	Mixed ancestry South African males	15		58			NS
ATP1B2/ TP53	rs1642764 (C/T)	Chen et al., 2019 (30753320)	Black South African[4]	591	0.21	852	0.20		NS
		Chen et al., 2019 (30753320)	Black South African[5]	880	0.18	939	0.18		NS
	rs1641511 (A/G)	Chen et al., 2019 (30753320)	Black South African[5]	880	0.39	939	0.42		NS
C20orf54	rs13042395	Bye et al., 2012 (22865593)	Black South African	407	0.002	849	0.005		Not informative
		Bye et al., 2012 (22865593)	Mixed ancestry South African	257	0.067	860	0.068		NS

(Continued)

TABLE 4 | Continued

Gene	Variant (rs number)	Study (PMID)	Population	ESCC		Controls		Effect allele	Findings and Comments[2]
				n	MAF	n	MAF		
CASP8	rs1045485 (Asp302His)	Bye et al., 2011 (21926110)	Black South African	358	0.154	477	0.152		NS
		Bye et al., 2011 (21926110)	Mixed ancestry South African	201	0.169	427	0.126	C	OR = 1.42 (1.01–1.98); p = 0.040
	rs3834129 (-652 6N ins/del)	Bye et al., 2011 (21926110)	Black South African	358	0.518	477	0.502		NS
		Bye et al., 2011 (21926110)	Mixed ancestry South African	201	0.385	427	0.386		NS
	rs10931936 (C/T)	Chen et al., 2019 (30753320)	Black South African[4]	591	0.19	852	0.20		NS
		Chen et al., 2019 (30753320)	Black South African[5]	880	0.22	939	0.20		NS
CHEK2	rs4822983 (C/T)	Chen et al., 2019 (30753320)	Black South African[4]	591	0.46	852	0.39	T	OR = 1.32 (1.12–1.56); p = 0.001
		Chen et al., 2019 (30753320)	Black South African[5]	880	0.43	939	0.42		NS
	rs1033667 (C/T)	Chen et al., 2019 (30753320)	Black South African[4]	591	0.44	852	0.38	T	OR = 1.30 (1.10–1.53) P = 0.002
		Chen et al., 2019 (30753320)	Black South African[5]	880	0.42	939	0.39		NS
CP	rs34053109 (C/G)	Strickland et al., 2012 (21901748)	Black South African	84	0	85	0.01		Not informative
	rs17838834 (T/C)	Strickland et al., 2012 (21901748)	Black South African	90	0.33	85	0.23		NS
	rs701749 (C/T)	Strickland et al., 2012 (21901748)	Black South African	79	0.01	78	0.02		Not informative
	rs17838833 (delT)	Strickland et al., 2012 (21901748)	Black South African	79	0.01	78	0		Not informative
	rs17838832 (T/C)	Strickland et al., 2012 (21901748)	Black South African	80	0.33	78	0.3		NS
	rs34334174 (C/T)	Strickland et al., 2012 (21901748)	Black South African	80	0.14	78	0.08		NS
	5'UTR-308G/A	Strickland et al., 2012 (21901748)	Black South African	52	0.05	64	0	A	p = 0.012; sample size very small
	rs17838831 (A/G)	Strickland et al., 2012 (21901748)	Black South African	53	0.21	64	0.22		NS
	rs138512757 (Thr83)	Strickland et al., 2012 (21901748)	Black South African	92	0.02	84	0.01		Not informative
	rs35438054 (Val223)	Strickland et al., 2012 (21901748)	Black South African	95	0.01	85	0.01		Not informative
	rs797045480 (Val246Ala)	Strickland et al., 2012 (21901748)	Black South African	95	0.01	85	0		Not informative
	rs34067682 (IVS4-14C/T)	Strickland et al., 2012 (21901748)	Black South African	84	0.12	83	0.12		NS
	rs34624984 (Arg367Cys)	Strickland et al., 2012 (21901748)	Black South African	94	0.02	86	0.01		Not informative
	rs34237139 (Tyr425)	Strickland et al., 2012 (21901748)	Black South African	91	0.01	87	0		Not informative
	rs35272481 (IVS7+9T/C)	Strickland et al., 2012 (21901748)	Black South African	91	0.01	87	0		Not informative
	rs701753 (D544E)	Strickland et al., 2012 (21901748)	Black South African	95	0.23	81	0.27		NS
	rs147192657 (Gly633 T/C)	Strickland et al., 2012 (21901748)	Black South African	88	0.07	84	0	C	p = 0.0004
	rs16861582 (IVS15-12T/C)	Strickland et al., 2012 (21901748)	Black South African	93	0,44	88	0.41		NS

(Continued)

TABLE 4 | Continued

Gene	Variant (rs number)	Study (PMID)	Population	ESCC		Controls		Effect allele	Findings and Comments[2]
				n	MAF	n	MAF		
CYP2E1	CYP2E1*1 (c1)/ CYP2E1*5 (c2)	Chelule et al., 2006 (17264406)	Black South African	30	0.04	331	0.06		Limited power
	-1053C/T	Li et al., 2005 (15899651)	Black and Mixed ancestry South African	189	0.01	198	0.02		NS
	-1293G/A	Li et al., 2005 (15899651)	Black and Mixed ancestry South African	189	0.01	198	0.03		NS
	7632T/A	Li et al., 2005 (15899651)	Black and Mixed ancestry South African	189	0.18	198	0.07	A	OR = 5.90 (3.25–10.7); p = 0.001 for genotype distribution
CYP3A5	CYP3A5*1	Dandara et al., 2005 (15978331)	Black South African	142	0.627	178	0.638		NS
		Dandara et al., 2005 (15978331)	Mixed ancestry South African	99	0.384	94	0.287		NS
	CYP3A5*3 (6986A/G)	Dandara et al., 2005 (15978331)	Black South African	142	0.155	178	0.138		NS
		Dandara et al., 2005 (15978331)	Mixed ancestry South African	99	0.475	94	0.590	G	OR = 0.60 (0.39–0.94); p = 0.025
	CYP3A5*6 (1490G/A)	Dandara et al., 2005 (15978331)	Black South African	142	0.190	178	0.213		NS
		Dandara et al., 2005 (15978331)	Mixed ancestry South African	99	0.136	94	0.122		NS
	CYP3A5*7 (27131-32insT; frameshift)	Dandara et al., 2005 (15978331)	Black South African	142	0.028	178	0.011		NS
		Dandara et al., 2005 (15978331)	Mixed ancestry South African	99	0.005	94	0		Not informative
	CYP3A5 all variants	Dandara et al., 2005 (15978331)	Black South African	142	0.373	178	0.441		NS
		Dandara et al., 2005 (15978331)	Mixed ancestry South African	99	0.616	94	0.713		OR = 0.65 (0.42–0.99); p = 0.045
FAS	rs1800682 (-670 G > A)	Bye et al., 2011 (21926110)	Black South African	358	0.219	477	0.225		NS
		Bye et al., 2011 (21926110)	Mixed ancestry South African	201	0.356	427	0.406		NS
	rs2234767 (-1377 G > A)	Bye et al., 2011 (21926110)	Black South African	358	0.096	477	0.072		NS
		Bye et al., 2011 (21926110)	Mixed ancestry South African	201	0.139	427	0.183		NS
FASL	rs763110 (-844 T > C)	Bye et al., 2011 (21926110)	Black South African	358	0.192	477	0.189		NS
		Bye et al., 2011 (21926110)	Mixed ancestry South African	201	0.416	427	0.386		NS
GSTP1	rs1695 (Ile105Val)	Matejcic et al., 2011	Black South African	325	0.518	474	0.534		NS
	rs1695 (Ile105Val)	Matejcic et al., 2011	Mixed ancestry South African	229	0.454	428	0.438		NS
	rs1695 (Ile105Val)	Li et al., 2010 (20540773)	Black South African		0.39		0.37		NS
	rs1695 (Ile105Val)	Li et al., 2010 (20540773)	Mixed ancestry South African		0.38		0.41		NS
	rs1138272 (Ala114Val)	Li et al., 2010 (20540773)	Black South African		0.22		0.07		NS
	rs1138272 (Ala114Val)	Li et al., 2010 (20540773)	Mixed ancestry South African		0.19		0.03		NS
GSTT1	Deletion allele	Matejcic et al., 2011 (22216261)	Black South African	311	0.574	462	0.554		NS
		Matejcic et al., 2011 (22216261)	Mixed ancestry South African	217	0.493	414	0.495		NS

(Continued)

TABLE 4 | Continued

Gene	Variant (rs number)	Study (PMID)	Population	ESCC		Controls		Effect allele	Findings and Comments[2]
				n	MAF	n	MAF		
GSTT2B	Deletion allele	Matejcic et al., 2011 (22216261)	Black South African	320	0.336	461	0.371		NS
		Matejcic et al., 2011 (22216261)	Mixed ancestry South African	226	0.418	425	0.501		OR = 0.71 (0.57–0.90); p = 0.004
MGMT	rs12917 (Leu84Phe)	Bye et al., 2011 (21926110)	Black South African	358	0.189	477	0.195		NS
		Bye et al., 2011 (21926110)	Mixed ancestry South African	201	0.222	427	0.168		OR = 1.41 (1.05–1.91); p = 0.023
MLH1	rs13320360 (c.546-191T/C)	Vogelsang et al., 2012 (22623965)	Black South African	343	0.15	340	0.17		NS
		Vogelsang et al., 2012 (22623965)	Mixed ancestry South African	203	0.07	264	0.06		NS
MLH3	rs28756991 (Arg797His)	Vogelsang et al., 2012 (22623965)	Black South African	345	0.11	342	0.12		NS
		Vogelsang et al., 2012 (22623965)	Mixed ancestry South African	205	0.09	264	0.4	G	OR = 2.07 (1.04–4.12); p = 0.038
MSH2	rs17217772 (Asn127Ser)	Vogelsang et al., 2012 (22623965)	Black South African	341	0.06	343	0.06		NS
		Vogelsang et al., 2012 (22623965)	Mixed ancestry South African	204	0.03	264	0.03		NS
	rs10188090 (c.2635-765G/A)	Vogelsang et al., 2012 (22623965)	Black South African	343	0.09	342	0.10		NS
		Vogelsang et al., 2012 (22623965)	Mixed ancestry South African	205	0.31	265	0.33		NS
	rs3771280 (c.1510+118T/C)	Vogelsang et al., 2012 (22623965)	Black South African	344	0.11	339	0.12		NS
		Vogelsang et al., 2012 (22623965)	Mixed ancestry South African	202	0.35	266	0.37		NS
MSH3	rs26279 (Ala1045Thr)	Vogelsang et al., 2012 (22623965)	Black South African	341	0.40	344	0.43		NS
		Vogelsang et al., 2012 (22623965)	Mixed ancestry South African	204	0.38	263	0.32	A	OR = 2.71 (1.34–5.50); $p = 5.71\times10^{-3}$
	rs1428030 (c.1341-12568A/G)	Vogelsang et al., 2012 (22623965)	Black South African	342	0.29	342	0.27		NS
		Vogelsang et al., 2012 (22623965)	Mixed ancestry South African	201	0.23	264	0.20		NS
	rs1805355 (Pro231Pro)	Vogelsang et al., 2012 (22623965)	Black South African	343	0.28	339	0.29		NS
		Vogelsang et al., 2012 (22623965)	Mixed ancestry South African	203	0.24	265	0.22		NS
NAT1	rs1057126 (1088T > A NAT1*10)	Matejcic et al., 2015 (26447020)	Black South African	463	54.8	480	57.7		NS
		Matejcic et al., 2015 (26447020)	Mixed ancestry South African	269	43.4	288	40.1		NS
	rs15561 (1095C > A NAT1*10, NAT1*3)	Matejcic et al., 2015 (26447020)	Black South African	463	55.7	480	57.7		NS
		Matejcic et al., 2015 (26447020)	Mixed ancestry South African	269	46.5	288	43		NS
NAT2	rs1799930 (590G/A NAT2*6)	Matejcic et al., 2015 (26447020)	Black South African	463	24.7	480	21.4		NS
		Matejcic et al., 2015 (26447020)	Mixed ancestry South African	269	22.4	288	22		NS
	rs1801280 (341T/C NAT2*5)	Matejcic et al., 2015 (26447020)	Black South African	463	27.1	480	29		NS
		Matejcic et al., 2015 (26447020)	Mixed ancestry South African	269	25.2	288	33.2	C	0R = 0.57 (0.38–0.87) p = 0.01
	rs1799931 (857G/A NAT2*7)	Matejcic et al., 2015 (26447020)	Black South African	463	0.01	480	0.05		Not informative
		Matejcic et al., 2015 (26447020)	Mixed ancestry South African	269	0.05	288	0.04		NS

(Continued)

TABLE 4 | Continued

Gene	Variant (rs number)	Study (PMID)	Population	ESCC		Controls		Effect allele	Findings and Comments[2]
				n	MAF	n	MAF		
	rs1801279 (191G/A NAT2*14)	Matejcic et al., 2015 (26447020)	Black South African	463	0.053	480	0.063		NS
		Matejcic et al., 2015 (26447020)	Mixed ancestry South African	269	0.038	288	0.023		NS
UNC5CL	rs10484761 (G/A)	Bye et al., 2012 (22865593)	Black South African	407	0.467	849	0.477		NS
		Bye et al., 2012 (22865593)	Mixed ancestry South African	257	0.354	860	0.314		NS
PTGS2	rs20417 (-765 G/C)	Bye et al., 2011 (21926110)	Black South African	358	0.471	477	0.513		NS
		Bye et al., 2011 (21926110)	Mixed ancestry South African	201	0.376	427	0.321		NS
	rs689466 (-1195 A/G)	Bye et al., 2011 (21926110)	Black South African	358	0.064	477	0.053		NS
		Bye et al., 2011 (21926110)	Mixed ancestry South African	201	0.103	427	0.155	G	OR = 0.63 (0.43–0.91); p = 0.014
PDE4D	rs10052657 (C/A)	Bye et al., 2012 (22865593)	Black South African	407	0.137	849	0.128		NS
		Bye et al., 2012 (22865593)	Mixed ancestry South African	257	0.175	860	0.155		NS
PLCE1	rs2274223 (His1927Arg)	Bye et al., 2012 (22865593)	Black South African	407	0.416	849	0.403		NS
		Bye et al., 2012 (22865593)	Mixed ancestry South African	257	0.437	860	0.40		NS
	rs17417407 (Arg548Leu)	Bye et al., 2012 (22865593)	Black South African	407	0.166	849	0.211	T	OR = 0.74 (0.60–0.93); p = 0.008
		Bye et al., 2012 (22865593)	Mixed ancestry South African	257	0.174	860	0.18		NS
	rs1438095332 (5′UTR 14 bp indel)	Bye et al., 2012 (22865593)	Black South African	321	0.234	456	0.242		NS
	rs199781223 (Gly1199Ser)	Bye et al., 2012 (22865593)	Black South African	321	0.053	449	0.045		NS
	rs3765525[3] (Ile1777Thr)	Bye et al., 2012 (22865593)	Black South African	316	0.472	452	0.463		NS
	rs58539480 (Pro1890Leu)	Bye et al., 2012 (22865593)	Black South African	307	0.073	429	0.064		NS
	rs17417407 (G/T)	Chen et al., 2019 (30753320)	Black South African[4]	591	0.17	852	0.21	T	OR = 0.76 (0.60–0.95); p = 0.014
		Chen et al., 2019 (30753320)	Black South African[5]	880	0.19	939	0.19		NS
	rs7084339 (G/A)	Chen et al., 2019 (30753320)	Black South African[5]	880	0.48	939	0.46		NS
	rs3765524 (T/C)	Chen et al., 2019 (30753320)	Black South African[4]	591	0.47	852	0.47		NS
		Chen et al., 2019 (30753320)	Black South African[5]	880	0.48	939	0.46		NS
	rs2274223 (A/G)	Chen et al., 2019 (30753320)	Black South African[4]	591	0.42	852	0.40		NS
		Chen et al., 2019 (30753320)	Black South African[5]	880	0.41	939	0.43		NS
	rs11187850 (A/G)	Chen et al., 2019 (30753320)	Black South African[5]	880	0.21	939	0.19		NS
PMS1	rs5742938 (c.-21+639G/A)	Vogelsang et al., 2012 (22623965)	Black South African	345	0.18	344	0.15		NS
		Vogelsang et al., 2012 (22623965)	Mixed ancestry South African	203	0.43	266	0.48	A	OR = 1.73 (1.07–2.79); p = 0.027
	rs13404927 (c.699+3331G/A)	Vogelsang et al., 2012 (22623965)	Black South African	342	0.18	339	0.19		NS
		Vogelsang et al., 2012 (22623965)	Mixed ancestry South African	204	0.14	264	0.12		NS

(Continued)

TABLE 4 | Continued

Gene	Variant (rs number)	Study (PMID)	Population	ESCC		Controls		Effect allele	Findings and Comments[2]
				n	MAF	n	MAF		
RUNX1	rs2014300 (A/G)	Bye et al., 2012 (22865593)	Black South African	407	0.378	849	0.403		NS
		Bye et al., 2012 (22865593)	Mixed ancestry South African	257	0.438	860	0.370	G	OR = 1.33 (1.09–1.63); p = 0.0055
	rs2014300 (A/G)	Chen et al., 2019 (30753320)	Black South African[4]	591	0.38	852	0.40		NS
		Chen et al., 2019 (30753320)	Black South African[5]	880	0.36	939	0.36		NS
	rs2834718 (T/A)	Chen et al., 2019 (30753320)	Black South African[5]	880	0.33	939	0.33		NS
SLC11A1	-237C/T	Zaahl et al., 2005 (15860357)	Mixed ancestry South African	105	0.029	110	0.1		$p < 0.004$
	-8G/A	Zaahl et al., 2005 (15860357)	Mixed ancestry South African	105	0.004	110	0.009		NS
	IVSI-28C/T	Zaahl et al., 2005 (15860357)	Mixed ancestry South African	105	0.028	110	0.0004		$p < 0.05$
	GT-repeat	Zaahl et al., 2005 (15860357)	Mixed ancestry South African		0.171		0.191		NS
SULT1A1	638G/A in Exon 7	Dandara et al., 2006 (16272171)	Black South African	145	0.42	194	0.37		NS[1]
		Dandara et al., 2006 (16272171)	Mixed ancestry South African	100	0.40	94	0.29		NS
TMEM173	rs13181561 (A/G)	Chen et al., 2019 (30753320)	Black South African[5]	880	0.48	939	0.49		NS
	rs13153461 (G/A)	Chen et al., 2019 (30753320)	Black South African[4]	591	0.04	852	0.05		NS
TP53	16-bp insertion in intron 3	Vos et al., 2003 (12550754)	Black South African	74	0.108	118	0.364		
	rs200073907 (Exon 4 codon 34)	Vos et al., 2003 (12550754)	Black South African	74	0.115	118	0.102		NS
	rs750578863 (Exon 4 codon 36)	Vos et al., 2003 (12550754)	Black South African	73	0.089	115	0.143		NS
	Arg72Pro	Vos et al., 2003 (12550754)	Black South African	73	0.356	115	0.409		$p < 0.05$
	Arg72Pro	Eltahir et al., 2012 (23053979)	Sudanese	25	0.49	235	0.51		NS
	rs1800371 (G/A)	Chen et al., 2019 (30753320)	Black South African[4]	591	0.02	852	0.03		NS
		Chen et al., 2019 (30753320)	Black South African[5]	880	0.03	939	0.02		NS
XBP1	rs2239815 (C/T)	Chen et al., 2019 (30753320)	Black South African[4]	591	0.21	852	0.16	T	OR = 1.41 (1.15–1.74) p = 0.001
		Chen et al., 2019 (30753320)	Black South African[5]	880		939			NS

[1] *Increased risk among smokers with SULT1A1*2/*2 genotype, but sample size was small.*

[2] *When OR > 1, effect allele = increased risk; when OR < 1, effect allele = protective effect.*

[3] *rs3765525 has been merged into rs959421.*

[4] *Western and Eastern Cape Province Black Population.*

[5] *Gauteng Province Black Population.*

being associated with carcinogenesis (Matejcic et al., 2015). Genes regulating cell apoptosis are *TP5, CHEK2, and CASP8* (Vos et al., 2003; Bye et al., 2011; Eltahir et al., 2012; Chen et al., 2019). *TP53* and *CHEK2* are also involved in gene expression and DNA repair. Regulation of gene expression is facilitated by *PLCE1* and *SLC11A1* (Zaahl et al., 2005; Bye et al., 2012). The *AR* gene regulates the sex hormones, androgens (Dietzsch et al., 2003), while *CYP2E1* and *CYP3A5* are involved in steroid, cholesterol, and lipid synthesis (Dandara et al., 2005; Li et al., 2005; Chelule et al., 2006). *CYP2E1* also metabolizes drugs and has been implicated in carcinogenesis. *CP* facilitates transportation of iron from organs into the blood cells; *RUNX1* plays a role in hematopoiesis and *PTGS2* in inflammation and mitogenesis (Bye et al., 2011; Bye et al., 2012; Strickland et al., 2012).

Nine of the 25 associated SNPs were from small studies with fewer than 150 cases and controls. These SNPs are in the following

TABLE 5 | Biological pathways for genetic susceptibility studies showing putative association with ESCC in African populations.

Gene	Full name	Pathway
ADH1B	Alcohol dehydrogenase 1B (class I), beta polypeptide	Ethanol metabolism
ADH3	Alcohol dehydrogenase ADH3	Metabolizes ethanol into acetaldehyde
ALDH2	Aldehyde dehydrogenase 2 family member	Alcohol metabolism. Implicated in increased susceptibility for cancer
AR	Androgen receptor	Regulates binding of androgens on androgen receptor
CASP8	Caspase 8	Cell apoptosis
CHEK2	Checkpoint kinase 2	Tumor suppressor gene. Mutations associated with predisposition to carcinogenesis
CP	Ceruloplasmin	Peroxidation of iron through its transportation from organs and tissue into blood
CYP2E1	Cytochrome P450 family 2 subfamily E member 1	Drug metabolism and catalysis and synthesis of cholesterol, steroids, and other lipids. Implicated in cancer development
CYP3A5	cytochrome P450 family 3 subfamily A member 5	Involved in drug metabolism and in the synthesis of cholesterol, steroids, and other lipids
GSTT2B	Glutathione S-transferase theta 2B (gene/ pseudogene)	Conjugation of glutathione to electrophilic and hydrophobic compounds. Plays a role in carcinogenesis
MGMT	O-6-methylguanine-DNA methyltransferase	DNA repair and defense from alkylating agents which cause mutagenesis and toxicity. Implicated in several cancers.
MLH3	MutL homolog 3	Maintenance of genomic integrity following cell division and DNA replication. Germline mutations implicated in cancer and somatic mutations implicated in microsatellite instability
MSH3	MutS homolog 3	Forms heterodimers with MSH2. Involved in mismatch repair and implicated in cancer development.
NAT2	N-acetyltransferase 2	Activation and deactivation of arylamine and hydrazine drugs and carcinogens. Implicated in high cancer incidence and drug toxicity.
PTGS2	Prostaglandin-endoperoxide synthase 2	A dioxygenase and a peroxidase involved in both inflammation and mitogenesis
PLCE1	Phospholipase C epsilon 1	Regulation of cell growth, differentiation, and gene expression.
PMS1	PMS1 homolog 1, mismatch repair system component	Mismatch repair gene. Mutations implicated in cancer development.
RUNX1	Runt related transcription factor 1	Development of hematopoiesis
SLC11A1	Solute carrier family 11 (proton-coupled divalent metal ion transporter), member 1	Regulation of gene expression.
TMEM173	Transmembrane protein 173	Regulation of the innate immune response to viral and bacterial infections. Role in tumorigenesis still inadequate
TP53	Tumor protein 53	Regulation of gene expression, cell cycle, apoptosis, and DNA repair.
XBP1	X-box binding protein 1	Regulation of genes involved in endoplasmic reticulum protein synthesis, folding, glycosylation, redox metabolism, autophagy, lipid biogenesis, and vesicular trafficking. Associated with development of cancer.

six genes: *ADH3, AR, CP, CYP3A5, SLC11A1,* and *TP53*. Because of the small sample size, the reliability and replicability of these results are uncertain. Sixteen of the SNPs came from studies with at least 150 cases and controls, and one study with 142 cases. These sample sizes could potentially give reliable and replicable results. The 16 SNPs were from the following genes: *ADH1B, ALDH2, CASP8, CHEK2, CYP2E1, GSTT2B, MGMT, MLH3, MSH3, NAT2, PLCE1, PMS1, PTGS2, and RUNX1*.

Two of the 16 SNPs are in the *ALDH2* gene and were analyzed in two different studies. However, it is not clear whether these two SNPs are the same because, while one study reported the NCBI rs number (rs886205) (Bye et al., 2011), the other study did not (Li et al., 2008).The two SNPs reported very different MAF, and opposite odds ratios of 2.35 and 0.70 demonstrating increased risk and a protective effect, respectively.

Six of the 16 SNPs were reported to reduce the risk of ESCC, and they are the following: *ADH1B* (Arg48His; rs1229984), *ALDH2* (+82 A > G; rs886205), *GSTT2B* (deletion allele), *NAT2* (341T > C; rs1801280), *PTGS2* (-1195 A > G; rs689466), and *PLCE1* (Arg548Leu; rs17417407). The remaining 10 SNPs were reported to increase the risk of ESCC: *ALDH2* (ALDH2*1/*2), *CASP8* (Asp302His; rs1045485), *CHEK2* (rs4822983 C > T, and rs1033667, C > T), *CYP2E1* (7632T > A), *MGMT* (Leu84Phe; rs12917), *MLH3* (Arg797His; rs28756991), *MSH3* (Ala1045Thr; rs26279), *PMS1* (c.-21+639G > A; rs5742938), and *RUNX1* (rs2014300). Eleven of the 16 SNPs showed association in the South African Admixed population, while only four showed association in the Black South African population and one in a combined South African population. All the studies used PCR-based methods for genotyping. Using the 1000 Genomes Database, r^2 analysis was carried out on SNPs reported in the same gene, to assess the LD between the SNPs. Thirteen pairs of SNPs in *MHS2, CP, MSH3, PLCE1,CHEK2,* and *NAT1* genes had $r^2 > 0.45$, shown in **Figure 2** and **Table S3**.

Altogether 44 somatic changes were reported in the following 22 genes: *AR*, *CCND1*, *CDKN2A*, *COL1A2*, *EFGR*, *EP300*, *FAT1*, *FAT2*, *FAT3*, *FAT4*, *FBXW7*, *JAG1*, *KMT2C (MLL3)*, *KMT2D (MLL2)*, *MUC2*, *NFE2L2*, *NOTCH1*, *NOTCH3*, *PIK3CA*, *SERPINB4*, *TP53*, and *TP63*, and six genetic loci without specific gene names (**Table 6**). The specific locus positions with the corresponding microsatellite markers are as follows: 2p (D2S123), 3p13 (D3S659), 3p24.2-25 (D3S1255), 4q12 (Bat 25), 2p21-p16.3 (Bat 26), and 1p12-13.3 (Bat 40). These variants were reported in the South African (20 variants), Kenyan (three variants), and Malawian (21 variants) populations. While the majority of the studies used PCR-based methods, a more recent study used WES as the analysis method (Liu et al., 2016). A total of 18 of the 22 genes with somatic variants in cancer tissue were

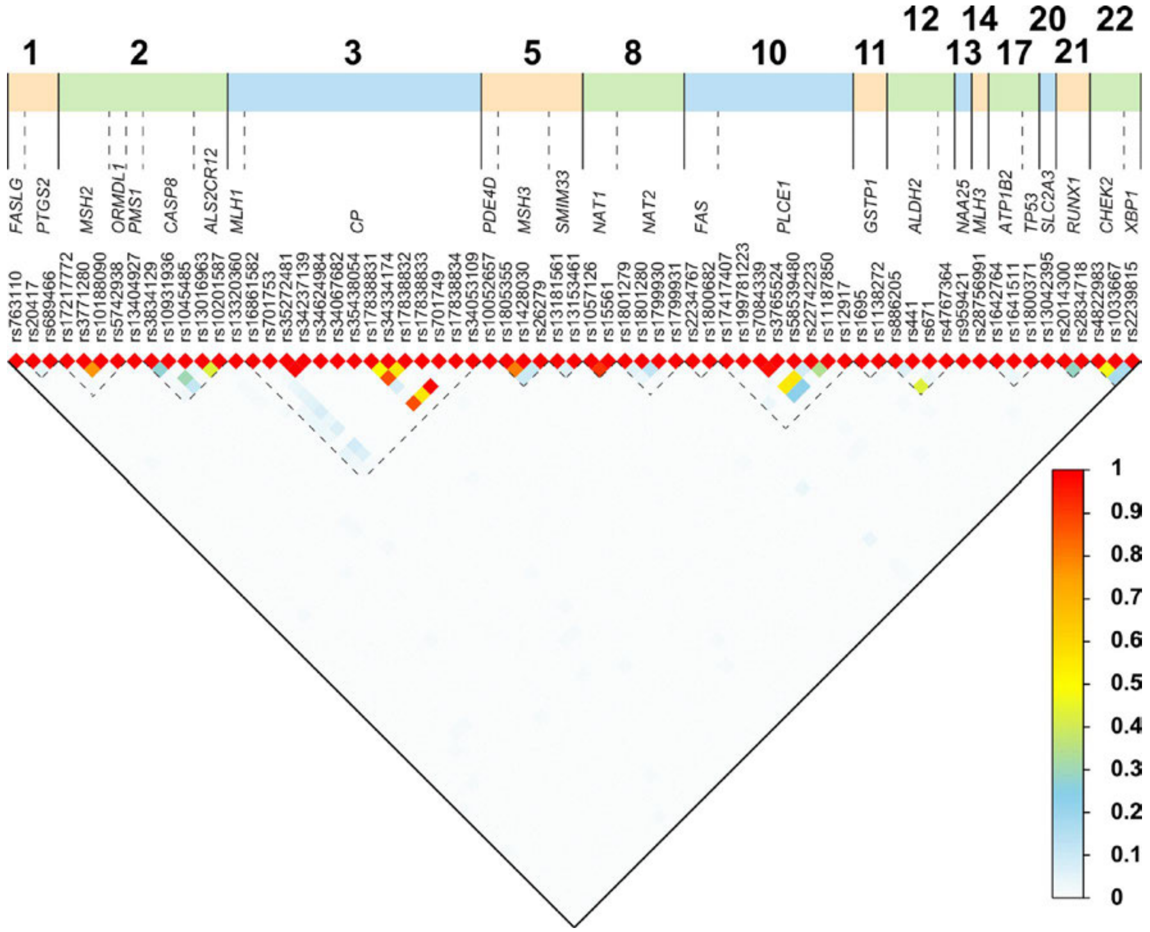


FIGURE 2 | Linkage disequilibrium (LD) plot for paired SNPs. We obtained the rs numbers of the variants from dbSNP (version 152; April 2019; (https://www.ncbi.nlm.nih.gov/snp/)) and used the canonical SNP identifier. To determine the LD between the SNPs, we obtained the imputed data set from the Thousand Genomes project (1000 Genomes Release Phase 3 2013-05-02) and used bcftools to extract all individuals from African populations not including African Americans, and the 77 SNPs discussed here using all synonyms (alternative rs IDs) for SNPs (Auton et al., 2015). We obtained a dataset of 504 individuals and 67 SNPs. We computed all pair-wise r2 using PLINK (v1.09) (Danecek et al., 2011; Chang et al., 2015).

discovered using WES. Statistical significance was not reported for any of the 44 variants. The most common type of somatic variants was missense mutations, reported in 14 of the 22 genes (64%) (Patel et al., 2011; Liu et al., 2016). Other somatic changes included copy number gains (14%), copy number losses (5%), deletions (14%), insertions (14%), and frameshift mutations (14%). In three studies (Dietzsch and Parker, 2002; Dietzsch et al., 2003; Naidoo et al., 2005), microsatellite instability and loss of heterozygosity (LOH) were reported (14%).

Table 7 shows a summary of the pathways in the 22 genes reporting somatic changes. Five genes, *AR, EP300, KMT2D, KMT2C,* and *TP53*, play a role in the regulation of transcription (Gamieldien et al., 1998; Dietzsch et al., 2003; Vos et al., 2003; Patel et al., 2011; Liu et al., 2016). The encoded protein for the *AR* gene functions as a steroid hormone activated transcription factor, while KMT2D has a role in methylation. Both *TP53* and *EP300* have been implicated in a number of cancers (Gamieldien et al., 1998; Vos et al., 2003; Patel et al., 2011; Liu et al., 2016). *TP53* additionally functions in DNA repair, gene expression, and apoptosis. The mismatch repair genes also facilitate DNA repair (Naidoo et al., 2005). *CCND1, CDKN2A, FAT1/2/3/4,* and *Ras* genes are all reported to be involved in cell cycle pathways including regulation of mitotic events, cell proliferation, and cell growth and death (Victor et al., 1990; Gamieldien et al., 1998; Liu et al., 2016). *NOTCH1* and *NOTCH3* both facilitate cell and tissue development (Liu et al., 2016). *JAG1* plays a role in hematopoiesis while *NFE2L2* is involved in response to inflammation including production of free radicals (Liu et al., 2016). *PIK3CA* is an oncogene implicated in tumor development while *SERPINB4* modulates response against tumor cells (Liu et al., 2016). *EGFR* and *COL1A2* genes encode for epidermal growth factor and type 1 collagen, respectively (Dietzsch and Parker, 2002; Liu et al., 2016). *FBXW7* is a tumor suppressor involved in ubiquitin degradation (Liu et al., 2016). *MUC2* facilitates the formation of a mucous barrier that protects the gut lumen (Liu et al., 2016). *TP63* gene is involved in tissue and organ development including skin and heart, and in adult stem cell regulation (Liu et al., 2016).

TABLE 6 | Summary of studies investigating somatic changes linked to ESCC in African patients.

Gene	Study (PMID)	Population	Findings
AR	Dietzsch et al., 2003 (12925954)	Black and mixed ancestry South African	LOH at CAG locus
CCND1	Liu et al., 2016 (29148985)	Malawian	Enriched copy number gains
CDKN2A	Gamieldien et al., 1998 (9808520)	Black South African	Insertions Deletions Frameshift mutations
	Liu et al., 2016 (29148985)	Malawian	Copy number losses
COL1A2	Dietzsch and Parker, 2002 (12435113)	Black South African	LOH (promoter and 1st intron) No evidence of MSI or allelic amplification
EFGR	Liu et al., 2016 (29148985)	Malawian	Copy number gains
EP300	Liu et al., 2016 (29148985)	Malawian	Missense mutations
FAT1	Liu et al., 2016 (29148985)	Malawian	Nonsense mutations
FAT2	Liu et al., 2016 (29148985)	Malawian	Missense mutations
FAT3	Liu et al., 2016 (29148985)	Malawian	Missense mutations
FAT4	Liu et al., 2016 (29148985)	Malawian	Missense mutations
FBXW7	Liu et al., 2016 (29148985)	Malawian	Frameshift mutations
JAG1	Liu et al., 2016 (29148985)	Malawian	Missense mutations
KMT2C (MLL3)	Liu et al., 2016 (29148985)	Malawian	Missense mutations
KMT2D (MLL2)	Liu et al., 2016 (29148985)	Malawian	Nonsense mutations
Mismatch repair genes	Naidoo et al., 2005 (15735161)	South African	LOH and MSI at: • D2S123 (2p) • D3S659 (3p13) • D3S1255 (3p3p24.2-25) • Bat 25 (4q12) • Bat 26 (2p2p21-p16.3) • Bat 40 (1p12-13.3)
MUC2	Liu et al., 2016 (29148985)	Malawian	Missense mutations
NFE2L2	Liu et al., 2016 (29148985)	Malawian	Missense mutations
NOTCH1	Liu et al., 2016 (29148985)	Malawian	Missense mutations
NOTCH3	Liu et al., 2016 (29148985)	Malawian	Missense mutations
PIK3CA	Liu et al., 2016 (29148985)	Malawian	Missense mutations
Ras genes	Victor et al., 1990 (2199031)	South African	No mutations found in codon 12, 13 or 61
SERPINB4	Liu et al., 2016 (29148985)	Malawian	Missense mutations
TP53	Liu et al., 2016 (29148985)	Malawian	Missense and nonsense mutations
	Gamieldien et al., 1998 (9808520)	Black South African	Exon 5–8 frameshift mutations: point mutations, deletions and insertions
	Patel et al., 2011 (22040862)	Kenyan	Exon 5–8 mutations: missense, nonsense and deletions
	Vos et al., 2003 (12550754)	South African	16-bp insertion in intron 3
	Vos et al., 2003 (12550754)	South African	Exon 4 polymorphism in codons 34, 36 and 72 LOH (16-bp repeat locus)
TP63	Liu et al., 2016 (29148985)	Malawian	Copy number gains

LOH, loss of heterozygosity; MSI, microsatellite instability.

Interaction Studies

Combinations of specific genotypes with environmental factors were also reported to be associated with ESCC in a number of studies (**Table 2**). The main two environmental factors studied were smoking and alcohol consumption. The interaction between smoking and alcohol status and specific genotypes was measured and reported as frequency (percentage) and assessed using p values and odds ratios in nine genetic susceptibility studies (Dandara et al., 2005; Li et al., 2005; Li et al., 2010; Dandara et al., 2006; Li et al., 2008; Li et al., 2010; Bye et al., 2011; Matejcic et al., 2011; Vogelsang et al., 2012; Matejcic et al., 2015). Four studies showed statistically significant associations between both alcohol and smoking status and variants in the *CYP3A5, CYP2E1, GST*, and *NAT2* genes (Dandara et al., 2005; Li et al., 2005; Matejcic et al., 2015). *SULT1A1* variants were associated with smoking status only (Dandara et al., 2006). Other interaction studies included wood/charcoal use and mutations in the *GST* genes (Li et al., 2010), as well as red and white meat intake and SNPs in *NAT1/2* genes (Matejcic et al., 2015).

DISCUSSION

General Systematic Review Findings

In this study, we systematically evaluated the genetic variants reported to be associated with ESCC in African populations providing the first systematic review on genetic factors of ESCC in this region. Of all studies that have been published on genetic association to ESCC in the African populations, only 23 fit our selection criteria. It was clear from the beginning that there is a dearth of information on this topic. Our analysis showed that 25 germline SNPs were reported to be associated with ESCC in the South African population. However, none of these SNPs were

TABLE 7 | Biological pathways for somatic changes studies showing putative association with ESCC in African populations.

Gene	Full name	Pathway
AR	Androgen receptor gene	Regulation of gene expression and the protein functions as a steroid-hormone activated transcription factor.
CCND1	Cyclin D1	Regulators of CDK kinases and mitotic events. Mutations and overexpression of the gene has been associated with cancer development.
CDKN2A	Cyclin dependent kinase inhibitor 2A	A tumor suppressor gene which regulates the cell cycle. Commonly inactivated in a variety of tumors.
CHEK2		
COL1A2	Collagen type I, alpha 2 chain	Encodes for type I collagen, which is an abundant connective tissue protein and part of extracellular matrix.
EFGR	Epidermal growth factor receptor	Encodes for the growth factor epidermal growth factor receptor.
EP300	E1A binding protein p300	Encodes the adenovirus E1A-associated cellular p300 transcriptional co-activator protein which functions in transcription regulation. Mutations have been implicated in tumorigenesis.
FAT1/2/3/4	FAT atypical cadherin 1/2/3/4	Human homologues of the *Drosophila* FAT genes. Putative tumor suppressor involved in cell proliferation during *Drosophila* development.
FBXW7	F-box and WD repeat domain containing 7	Encodes an F-Box protein which binds directly to cyclin E and potentially targets cyclin E for ubiquitin-mediated degradation.
JAG1	Jagged 1	Encodes for the human homolog of the *Drosophila* jagged 1 protein which is involved in hematopoiesis.
KMT2C (MLL3)	Lysine methyltransferase 2C	The gene is member of the myeloid/lymphoid or mixed-lineage leukemia (MLL) family. It encodes a nuclear protein involved in transcriptional regulation.
KMT2D (MLL2)	Lysine methyltransferase 2D	Methylation of histones and transcriptional regulation.
Mismatch repair genes	Mismatch repair genes	DNA repair. Mutations have been implicated in cancer.
MUC2	Mucin 2, oligomeric mucus/gel-forming	Formation of insoluble mucous barrier that protects the gut lumen.
NFE2L2	Nuclear factor, erythroid 2 like 2	Encodes for proteins involved in response to inflammation including free radical production.
NOTCH1	NOTCH1	Development of cell and tissue. Mutations have been reported to be linked with tumorigenesis.
NOTCH3	NOTCH3	The third discovered human homologue of the *Drosophila* melanogaster type I membrane protein notch. Involved in intercellular signaling pathways in neural development.
PIK3CA	Phosphatidylinositol-4,5-bisphosphate 3-kinase catalytic subunit alpha	Oncogenic and implicated in cancer development.
Ras genes	Rat sarcoma	Regulation of cell signaling pathways, and cell growth and death.
SERPINB4	Serpin family B member 4	Inactivation of granzyme M, an enzyme that kills tumor cells. Highly expressed in tumor cells.
TP53	Tumor protein p53	Regulates transcription, expression of target genes, thereby inducing cell cycle arrest, apoptosis, senescence, DNA repair, or changes in metabolism. Implicated in a number of cancers.
TP63	Tumor protein p63	Involved in the following processes in skin development and maintenance, adult stem/progenitor cell regulation, heart development, and premature aging.

repeated in three or more independent studies; hence, a meta-analysis was not possible. Additionally, only three (*ALDH2, PLCE* and *CYP2E1*) of the 20 genes were analyzed in two independent studies, but testing for different SNPs. We determined that it was unlikely that the two *ALDH2* SNPs analyzed were the same SNPs. This is because the MAFs were significantly different and, while one SNP had a protective effect (reduced risk), the other increased risk. The lack of studies re-assessing the same genetic variants poses a major hurdle in validating existing evidence on the association between genetic variants and ESCC development. This makes resolving the genetic etiology of ESCC in African populations difficult.

Genetic Susceptibility to ESCC

Of the 25 SNPs from the genetic susceptibility studies that showed an association to ESCC, we concluded that results on 16 SNPs had the potential to be reliable and reproducible due to the larger sample sizes. Ten of the SNPs were reported to increase the risk of ESCC, while six were reported to reduce the risk. However, it was noted that the majority (11) of these SNPs showed association in the South African Admixed population and the studies did not report controlling for population stratification. This is a highly admixed population (Chimusa et al., 2013), in which the predominant ancestral lines are Khoesan (32–43%), Bantu-speaking Africans (20–36%), European (21–28%), and Asian (9–11%) (De Wit et al., 2010). This diverse population is a result of South Africa's colonial and trade history, and constitutes 9% of the total South African population (De Wit et al., 2010). Genetic variability can also be seen in the Black South African population (Chimusa et al., 2013). Without controlling for population stratification, the reproducibility of these results is questionable. It is, however, important to note that the majority of these studies were carried out several years ago, and information on population stratification and methods to detect it may not have been available as yet.

Re-examination of common SNPs from the Chinese population was done in three of the studies (Bye et al., 2011; Bye et al., 2012; Chen et al., 2019), but the findings were not conclusive. It is possible that there may be population-specific differences influencing the genetic etiology of ESCC in the African populations. This may also point to the role of environmental factors contributing to the genetic susceptibility to ESCC through gene-environment interactions.

Somatic Changes in ESCC

Forty-four somatic variants were reported, but only two were significantly associated with ESCC. The paucity of information was also evident in the somatic variant studies. There were significantly fewer studies (8) on somatic variants than on genetic susceptibility (17). The molecular profiling of tumors is of great importance as it is relevant in the development of targeted cellular therapeutics. One gene (*CDKN2A*) was analyzed in two studies, but these studies focused on a different variant. Another gene, TP53, was analyzed in four studies, but two studies analyzed different parts of the gene, and two had no control data. It was evident, however, that the WES study provided with a wider variety of genetic variants associated with ESCC (Liu et al., 2016). The WES study overall had the largest number of genetic variants of all the 23 studies and was able to identify variants in an unbiased manner.

Common Limitations Among the African Studies

There were no GWAS among the studies we analyzed, but reports from the Chinese and European studies demonstrated that GWAS are able to successfully identify common genetic variants associated with ESCC (Abnet et al., 2017). To date, GWAS has successfully identified more than 700 loci for cancer risk. However, these studies have been predominantly done in populations of European ancestry (80%), with African and Latin American populations contributing less than 1% (Van Loon et al., 2018). A shift to WES and GWAS on the African populations might, therefore, yield better results in identifying variants that play a role in ESCC development. The African Esophageal Cancer Consortium, which was initiated in 2016 by African investigators and International partners, released a call to action to, among other priority activities, increase molecular research on esophageal cancer in Africa, particularly GWAS and genomic profiling (Van Loon et al., 2018).

One of the main deficiencies in the studies was that the majority of the genetic susceptibility studies did not report a power calculation, or a genotyping error, and this may have resulted in studies being underpowered and with increased type II error. Few studies reported correction for multiple testing; however, many of the studies were not analyzing multiple variants at the same time. The lack of correction for multiple testing, therefore, is not a reflection on the methodological quality. Very few studies reported NCBI rs numbers. In most studies, the diagnosis of ESCC in patients was adequately defined with no ambiguity on the number of patients with ESCC. There were, however, three studies that combined samples from patients with squamous cell and adenocarcinoma into one case group, which could introduce bias (Dietzsch et al., 2003; Eltahir et al., 2012; Vogelsang et al., 2012).

It is important to note that rs numbers were poorly documented in the majority of the studies assessed in this systematic review. Additionally, in many of these studies, the positions of the SNPs using genome coordinates were not reported, hence making it difficult to locate the SNPs. In the absence of an rs number, we recommend that authors report the position using genome coordinates and the version of the genome used as a reference.

The somatic variant studies also had adequately defined ESCC diagnosis for the majority of the studies. While the variant classification and type were reported by most studies, there was no confirmation of the results (except for two studies). Overall, for both the germline and somatic variant studies, the quality of reporting for the majority of the studies was not adequate. Other important limitations and biases are the lack of controlling for population stratification and small sample sizes in the study populations, which may have led to unreliable results.

Limitations of the Systematic Review

While we did a comprehensive search in four of the main literature databases, it is possible that we could have missed some non-English studies on African populations. Because of the lack of replication and validation studies, we could not carry out a meta-analysis in the current study. Furthermore, we did not re-analyze the data and relied on reported p values and odds ratios for descriptive analysis.

CONCLUSIONS

While this review has highlighted a number of genes that may be potentially associated with ESCC in the African populations, limitations such as lack of reproducibility, quality of reporting, and quality of assessment remain a major concern. The implications of having these inconsistencies and lack of reproducibility are that the genetic etiology of ESCC in Africa will continue to be unclear. The region lags behind in contributing to genetic knowledge and literature on ESCC. Importantly, any preventative, diagnostic, or therapeutic interventions cannot be effectively identified or applied in these populations.

The identification of genetic markers of esophageal cancer susceptibility has clear translational benefits to African populations in understanding the underlying disease risk and heritability. Benefits include the utilization of genetic information to improve risk prediction, which can be translated into prevention and screening programs relevant and specific to the African population. These studies also play a role in identifying and quantifying the interactions of modifiable environmental risk factors, which interact with these genetic variants, and hence provide a platform for better targeted interventions. The ability to sufficiently translate genetic research on the African population is dependent on more genetic studies done on the population.

Our recommendations are that more and larger genetic studies be done on the African populations, particularly focusing on WES and GWAS approaches. This will require multinational collaborations between the African countries.

ETHICS STATEMENT

The study was approved by the Stellenbosch University Health Research Ethics Committee as part of the Doctoral Studies of HS (HREC Reference #: S18/10/250).

AUTHOR CONTRIBUTIONS

VL, VS, and HS carried out literature searches. HS, VS, and HK appraised the articles, summarized the results, prepared the tables and figures, and drafted the manuscript. VS and VL reviewed the articles and edited the manuscript. VS and HK conceptualized the idea for the research, obtained funding, supervised the project, and wrote sections of the manuscript. VL provided specialist expertise and knowledge, and critically reviewed the manuscript. GT carried out the r^2 analyses, prepared the r^2 figure and table, and critically reviewed and revised the manuscript. All authors approved the final version of the manuscript.

FUNDING

This work was supported by the African Cancer Institute, Faculty of Medicine and Health Sciences, Stellenbosch University. HS acknowledges the Beit Trust Hardship Fund for providing a Doctoral Scholarship in part aid of tuition and registration fees and the Collaboration for Evidence-based Healthcare and Public Health in Africa (CEBHA+), as part of the Research Networks for Health Innovation in Sub-Saharan Africa Funding Initiative of the German Federal Ministry of Education and Research. GT was supported by the South African Tuberculosis Bioinformatics Initiative (SATBBI), a Strategic Health Innovation Partnership grant from the South African Medical Research Council and South African Department of Science and Technology.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2019.00642/full#supplementary-material

REFERENCES

Abnet, C. C., Arnold, M., and Wei, W. Q. (2017). Epidemiology of esophageal squamous cell carcinoma. *Gastroenterology* 154, 360–373. doi: 10.1053/j.gastro.2017.08.023

Adams, C. H., Werely, C. J., Victor, T. C., Hoal, E. G., Rossouw, G., and Van Helden, P. D. (2003). Allele frequencies for glutathione S-transferase and N-acetyltransferase 2 differ in African population groups and may be associated with oesophageal cancer or tuberculosis incidence. *Clin. Chem. Lab. Med.* 41, 600–605. doi: 10.1515/CCLM.2003.090

Auton, A., Brooks, L. D., Durbin, R. M., Garrison, E. P., Kang, H. M., Korbel, J. O., et al. (2015). A global reference for human genetic variation. *Nature* 526, 68–74. doi: 10.1038/nature15393

Baba, Y., Yamamura, K., Nakagawa, S., Mima, K., Ishimoto, T., Iwatsuki, M., et al. (2017). Abstract 4930: genetic and epigenetic characteristics of esophageal cancer tissues with microbiome fusobacterium nucleatum. *Cancer Res.* 77, 4930–4930. doi: 10.1158/1538-7445.AM2017-4930

Bye, H., Prescott, N. J., Lewis, C. M., Matejcic, M., Moodley, L., Robertson, B., et al. (2012). Distinct genetic association at the PLCE1 locus with oesophageal squamous cell carcinoma in the South African population. *Carcinogenesis* 33, 2155–2161. doi: 10.1093/carcin/bgs262

Bye, H., Prescott, N. J., Matejcic, M., Rose, E., Lewis, C. M., Parker, M. I., et al. (2011). Population-specific genetic associations with oesophageal squamous cell carcinoma in South Africa. *Carcinogenesis* 32, 1855–1861. doi: 10.1093/carcin/bgr211

Chang, C. C., Chow, C. C., Tellier, L. C., Vattikuti, S., Purcell, S. M., and Lee, J. J. (2015). Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* 4, 7. doi: 10.1186/s13742-015-0047-8

Chelule, P. K., Pegoraro, R. J., Gqaleni, N., and Dutton, M. F. (2006). The frequency of cytochrome P450 2E1 polymorphisms in Black South Africans. *Dis. Markers* 22, 351–354. doi: 10.1155/2006/980392

Chen, W. C., Bye, H., Matejcic, M., Amar, A., Govender, D., Khew, Y. W., et al. (2019). Association of genetic variants in CHEK2 with oesophageal squamous cell carcinoma in the South African Black population. *Carcinogenesis* 40, 513−520. doi: 10.1093/carcin/bgz026

Chen, X., Winckler, B., Lu, M., Cheng, H., Yuan, Z., Yang, Y., et al. (2015). Oral microbiota and risk for esophageal squamous cell carcinoma in a high-risk area of China. *PLoS One* 10, e0143603. doi: 10.1371/journal.pone.0143603

Chimusa, E. R., Daya, M., Moller, M., Ramesar, R., Henn, B. M., Van Helden, P. D., et al. (2013). Determining ancestry proportions in complex admixture scenarios in South Africa using a novel proxy ancestry selection method. *PLoS One* 8, e73971. doi: 10.1371/journal.pone.0073971

Coleman, H. G., Xie, S. H., and Lagergren, J. (2018). The Epidemiology of Esophageal Adenocarcinoma. *Gastroenterology* 154, 390–405. doi: 10.1053/j.gastro.2017.07.046

Dandara, C., Ballo, R., and Parker, M. I. (2005). CYP3A5 genotypes and risk of oesophageal cancer in two South African populations. *Cancer Lett.* 225, 275–282. doi: 10.1016/j.canlet.2004.11.004

Dandara, C., Li, D. P., Walther, G., and Parker, M. I. (2006). Gene-environment interaction: the role of SULT1A1 and CYP3A5 polymorphisms as risk modifiers for squamous cell carcinoma of the oesophagus. *Carcinogenesis* 27, 791–797. doi: 10.1093/carcin/bgi257

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., Depristo, M. A., et al. (2011). The variant call format and VCFtools. *Bioinformatics* 27, 2156–2158. doi: 10.1093/bioinformatics/btr330

De Wit, E., Delport, W., Rugamika, C. E., Meintjes, A., Moller, M., Van Helden, P. D., et al. (2010). Genome-wide analysis of the structure of the South African Coloured Population in the Western Cape. *Hum. Genet.* 128, 145–153. doi: 10.1007/s00439-010-0836-1

Dietzsch, E., Laubscher, R., and Parker, M. I. (2003). Esophageal cancer risk in relation to GGC and CAG trinucleotide repeat lengths in the androgen receptor gene. *Int. J. Cancer* 107, 38–45. doi: 10.1002/ijc.11314

Dietzsch, E., and Parker, M. I. (2002). Infrequent somatic deletion of the 5'region of the COL1A2 gene in oesophageal squamous cell cancer patients. *Clin. Chem. Lab. Med.* 40, 941–945. doi: 10.1515/CCLM.2002.165

Eltahir, H. A., Adam, A. A., Yahia, Z. A., Ali, N. F., Mursi, D. M., Higazi, A. M., et al. (2012). p53 Codon 72 arginine/proline polymorphism and cancer in Sudan. *Mol. Biol. Rep.* 39, 10833–10836. doi: 10.1007/s11033-012-1978-0

Gamieldien, W., Victor, T. C., Mugwanya, D., Stepien, A., Gelderblom, W. C., Marasas, W. F., et al. (1998). p53 and p16/CDKN2 gene mutations in esophageal tumors from a high-incidence area in South Africa. *Int. J. Cancer* 78, 544–549. doi: 10.1002/(SICI)1097-0215(19981123)78:5<544::AID-IJC3>3.0.CO;2-T

He, F., Liu, C., Zhang, R., Hao, Z., Li, Y., Zhang, N., et al. (2018). Association between the Glutathione-S-transferase T1 null genotype and esophageal cancer susceptibil ty: a meta-analysis involving 11,163 subjects. *Oncotarget* 9, 15111–15121. doi: 10.18632/oncotarget.24534

Huang, F. L., and Yu, S. J. (2018). Esophageal cancer: risk factors, genetic association, and treatment. *Asian J. Surg.* 41, 210–215. doi: 10.1016/j.asjsur.2016.10.005

Jaskiewicz, K., and De Groot, K. M. (1994). p53 gene mutants expression, cellular proliferation and differentiation in oesophageal carcinoma and non-cancerous epithelium. *Anticancer Res.* 14, 137–140.

Kayamba, V., Bateman, A. C., Asombang, A. W., Shibemba, A., Zyambo, K., Banda, T., et al. (2015). HIV infection and domestic smoke exposure, but not human papillomavirus, are risk factors for esophageal squamous cell carcinoma in Zambia: a case-control study. *Cancer Med.* 4, 588–595. doi: 10.1002/cam4.434

Kaz, A. M., and Grady, W. M. (2014). Epigenetic biomarkers in esophageal cancer. *Cancer Lett.* 342, 193–199. doi: 10.1016/j.canlet.2012.02.036

Kumar, P., and Rai, V. (2018). MTHFR C677T polymorphism and risk of esophageal cancer: an updated meta-analysis. *Egypt. J. Med. Hum. Genet.* 19, 273–284. doi: 10.1016/j.ejmhg.2018.04.003

Li, D.-P., Dandara, C., Walther, G., and Parker, M. I. (2008). Genetic polymorphisms of alcohol metabolising enzymes: their role in susceptibility to oesophageal cancer. *Clin. Chem. Lab. Med.* 46, 323–328. doi: 10.1515/CCLM.2008.073

Li, D., Dandara, C., and Parker, M. I. (2005). Association of cytochrome P450 2E1 genetic polymorphisms with squamous cell carcinoma of the oesophagus. *Clin. Chem. Lab. Med.* 43, 370–375. doi: 10.1515/CCLM.2005.067

Li, D., Dandara, C., and Parker, M. I. (2010). The 341C/T polymorphism in the GSTP1 gene is associated with increased risk of oesophageal cancer. *BMC Genet.* 11, 47. doi: 10.1186/1471-2156-11-47

Li, M., Yu, X., Zhang, Z. Y., Wu, C. L., and Xu, H. L. (2016). Interaction of XRCC1 Arg399Gln polymorphism and alcohol consumption influences susceptibility of esophageal cancer. *Gastroenterol. Res. Pract.* 2016, 9495417. doi: 10.1155/2016/9495417

Little, J., Higgins, J. P., Ioannidis, J. P., Moher, D., Gagnon, F., Von Elm, E., et al. (2009). STrengthening the REporting of Genetic Association Studies (STREGA)–an extension of the STROBE statement. *Genet. Epidemiol.* 33, 581–598. doi: 10.1002/gepi.20410

Liu, W., Snell, J. M., Jeck, W. R., Hoadley, K. A., Wilkerson, M. D., Parker, J. S., et al. (2016). Subtyping sub-Saharan esophageal squamous cell carcinoma by comprehensive molecular analysis. *JCI Insight* 1, e88755. doi: 10.1172/jci.insight.88755

Matejcic, M., Li, D., Prescott, N. J., Lewis, C. M., Mathew, C. G., and Parker, M. I. (2011). Association of a deletion of GSTT2B with an altered risk of oesophageal squamous cell carcinoma in a South African population: a case-control study. *PLoS One* 6, e29366. doi: 10.1371/journal.pone.0029366

Matejcic, M., Vogelsang, M., Wang, Y., and Parker, M. I. (2015). Erratum to: NAT1 and NAT2 genetic polymorphisms and environmental exposure as risk factors for oesophageal squamous cell carcinoma: a case-control study. *BMC Cancer* 15, 658. doi: 10.1186/s12885-015-1681-3

Munishi, M. O., Hanisch, R., Mapunda, O., Ndyetabura, T., Ndaro, A., Schüz, J., et al. (2015). Africa's oesophageal cancer corridor: do hot beverages contribute? *Cancer Causes Control* 26, 1477–1486. doi: 10.1007/s10552-015-0646-9

Murphy, G., Mccormack, V., Abedi-Ardekani, B., Arnold, M., Camargo, M. C., Dar, N. A., et al. (2017). International cancer seminars: a focus on esophageal squamous cell carcinoma. *Ann Oncol.* 28, 2086–2093. doi: 10.1093/annonc/mdx279

Naidoo, R., Ramburan, A., Reddi, A., and Chetty, R. (2005). Aberrations in the mismatch repair genes and the clinical impact on oesophageal squamous carcinomas from a high incidence area in South Africa. *J. Clin. Pathol.* 58, 281–284. doi: 10.1136/jcp.2003.014290

Patel, K., Mining, S., Wakhisi, J., Ghe t, T., Tommasino, M., Martel-Planche, G., et al. (2011). TP53 mutations, human papilloma virus DNA and inflammation markers in esophageal squamous cell carcinoma from the Rift Valley, a high-incidence area in Kenya. *BMC Res. Notes* 4, 469. doi: 10.1186/1756-0500-4-469

Pink, R. C., Bailey, T. A., Iputo, J. E., Sammon, A. M., Woodman, A. C., and Carter, D. R. (2011). Molecular basis for maize as a risk factor for esophageal cancer in a South African population via a prostaglandin E2 positive feedback mechanism. *Nutr. Cancer* 63, 714–721. doi: 10.1080/01635581.2011.570893

Schaafsma, T., Wakefield, J., Hanisch, R., Bray, F., Schüz, J., Joy, E. J. M., et al. (2015). Africa's oesophageal cancer corridor: geographic variations in incidence correlate with certain micronutrient deficiencies. *PloS One* 10, e0140107–e0140107. doi: 10.1371/journal.pone.0140107

Sewram, V., Sitas, F., O'connell, D., and Myers, J. (2014). Diet and esophageal cancer risk in the Eastern Cape Province of South Africa. *Nutr. Cancer* 66, 791–799. doi: 10.1080/01635581.2014.916321

Sewram, V., Sitas, F., O'connell, D., and Myers, J. (2016). Tobacco and alcohol as risk factors for oesophageal cancer in a high incidence area in South Africa. *Cancer Epidemiol.* 41, 113–121. doi: 10.1016/j.canep.2016.02.001

Strickland, N. J., Matsha, T., Erasmus, R. T., and Zaahl, M. G. (2012). Molecular analysis of ceruloplasmin in a South African cohort presenting with oesophageal cancer. *Int. J. Cancer* 131, 623–632. doi: 10.1002/ijc.26418

Uys, P., and Van Helden, P. D. (2003). On the nature of genetic changes required for the development of esophageal cancer. *Mol. Carcinog.* 36, 82–89. doi: 10.1002/mc.10100

Van Loon, K., Mwachiro, M. M., Abnet, C. C., Akoko, L., Assefa, M., Burgert, S.L., et al. (2018). The African esophageal cancer consortium: a call to action. *J. Glob. Oncol.* 4, 1–9. doi: 10.1200/JGO.17.00163

Victor, T., Du Toit, R., Jordaan, A. M., Bester, A. J., and Van Helden, P. D. (1990). No evidence for point mutations in codons 12, 13, and 61 of the ras gene in a high-incidence area for esophageal and gastric cancers. *Cancer Res.* 50, 4911–4914.

Vogelsang, M., Wang, Y., Veber, N., Mwapagha, L. M., and Parker, M. I. (2012). The cumulative effects of polymorphisms in the DNA mismatch repair genes and tobacco smoking in oesophageal cancer risk. *PLoS One* 7, e36962. doi: 10.1371/journal.pone.0036962

Vos, M., Adams, C. H., Victor, T. C., and Van Helden, P. D. (2003). Polymorphisms and mutations found in the regions flanking exons 5 to 8 of the TP53 gene in a population at high risk for esophageal cancer in South Africa. *Cancer Genet. Cytogenet.* 140, 23–30. doi: 10.1016/S0165-4608(02)00638-6

Yazbeck, R., Jaenisch, S. E., and Watson, D. I. (2016). From blood to breath: new horizons for esophageal cancer biomarkers. *World J. Gastroenterol.* 22, 10077–10083. doi: 10.3748/wjg.v22.i46.10077

Zaahl, M. G., Warnich, L., Victor, T. C., and Kotze, M. J. (2005). Association of functional polymorphisms of SLC11A1 with risk of esophageal cancer in the South African Colored population. *Cancer Genet. Cytogenet.* 159, 48–52. doi: 10.1016/j.cancergencyto.2004.09.017

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Supplementary Table S1. Quality Assessment of Genetic Susceptibility Studies

Study	Power calculations reported	Description of ESCC diagnosis	Screening of Controls for ESCC	Detailed population characteristics for cases	Detailed population characteristics for controls	Adjustments for population stratification	NCBI rs numbers	Assessment of HWE	Assessment of genotyping error	Reported data as risk ratios	Correction for multiple testing	Quality score (0 to 11)
Bye et al 2012	Yes	Yes	No	Yes	Yes	No	Yes	Yes	No	Yes	Yes	8
Bye et al 2011	Yes	Yes	No	Yes	No	No	Yes	Yes	No	Yes	Yes	7
Chen et al 2019	Yes	Yes	No	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes	9
Chelule at al 2006	No	Yes	No	No	No	No	No	Yes	No	No	No	2
Dandara at al 2005	No	Yes	No	No	No	No	No	No	No	Yes	No	2
Dandara at al 2006	No	Yes	No	Yes	Yes	No	No	Yes	No	Yes	No	5
Dietzsch et al 2003	No	Yes	No	Yes	Yes	No	No	No	No	Yes	No	4
Eltahir et al 2012	No	Yes	No	No	No	No	No	No	No	Yes	No	2
Li et al 2005	No	Yes	No	Yes	Yes	No	No	Yes	No	Yes	No	5
Li et al 2010	Yes	Yes	No	Yes	Yes	No	Yes	Yes	No	Yes	Yes	8
Li et al 2008	No	Yes	No	Yes	Yes	No	No	No	Yes	Yes	No	5
Matejcic et al 2011	Yes	Yes	No	Yes	Yes	No	Yes	Yes	No	Yes	Yes	8
Matejcic 2015	No	Yes	No	Yes	Yes	No	Yes	Yes	No	Yes	Yes	7
Strickland et al 2012	No	Yes	Yes	Yes	No	No	Yes	Yes	No	No	No	5
Vogelsang et al 2012	Yes	Yes	No	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes	9
Vos et al 2003	No	Yes	No	No	No	No	No	No	No	No	No	1
Zaahl et al 2005	No	Yes	No	No	No	No	No	No	No	No	No	1

ESCC, esophageal squamous cell carcinoma; HWE, Hardy Weinberg equilibrium; NCBI, National Center for Biotechnology Information (https://www.ncbi.nlm.nih.gov/): rs-numbers can be obtained from the NCBI dbSNP database, which is available at https://www.ncbi.nlm.nih.gov/snp/.

Supplementary Table S2. Quality Assessment of Somatic Variant Studies

Study	Description of ESCC diagnosis	Tissues used: Cancerous & Normal neighbouring tissue, or blood	Detailed population characteristics	Variant classification and type	Confirmation of results	Amino acid change reported	Use of pathogenicity scoring described	Quality score (0 to 7)
Dietzsch et al 2003	No	Yes	Yes	Yes	No	NA	NA	3
Dietzsch et al 2002	Yes	Yes	Yes	Yes	No	NA	NA	4
Gamieldien et al 1998	Yes	Yes	Yes	Yes	Yes	Yes	No	6
Liu et al 2016	Yes	Yes	Yes	Yes	No	No	No	4
Naidoo et al 2005	Yes	Yes	Yes	Yes	No	NA	NA	4
Patel et al 2011	Yes	No	Yes	Yes	No	Yes	No	4
Victor et al 1990	No	No	No	NA	NA	NA	NA	0
Vos et al 2003	Yes	Yes	No	Yes	Yes	Yes	No	5

ESCC, esophageal squamous cell carcinoma; NA, not applicable

Supplementary Table S3. Summary of SNPs with r^2>0.20.

Gene Symbol	Chromosome A	BP A[1]	SNP A	Chromosome B	BP B[1]	SNP B	R^2
MSH2	2	47690411	rs3771280	2	47709153	rs10188090	0.787
CASP8	2	202097531	rs3834129	2	202143928	rs10931936	0.256
CASP8/ ALS2CR12	2	202143928	rs10931936	2	202162811	rs13016963	0.315
ALS2CR12	2	202162811	rs13016963	2	202202791	rs10201587	0.409
CP	3	148919880	rs35272481	3	148919962	rs34237139	1
CP	3	148939861	rs17838831	3	148939929	rs34334174	0.515
CP	3	148939861	rs17838831	3	148939933	rs17838832	0.884
CP	3	148939861	rs17838831	3	148940142	rs17838834	0.884
CP	3	148939929	rs34334174	3	148939933	rs17838832	0.579
CP	3	148939929	rs34334174	3	148940142	rs17838834	0.580
CP	3	148939933	rs17838832	3	148940142	rs17838834	1
MSH3	5	79966029	rs1805355	5	80008704	rs1428030	0.809
NAT1	8	18080644	rs1057126	8	18080651	rs15561	0.908
PLCE1	10	96043732	rs7084339	10	96058298	rs3765524	0.970
PLCE1	10	96043732	rs7084339	10	96066341	rs2274223	0.572
PLCE1	10	96043732	rs7084339	10	96068480	rs11187850	0.224
PLCE1	10	96058298	rs3765524	10	96066341	rs2274223	0.585
PLCE1	10	96058298	rs3765524	10	96068480	rs11187850	0.233
PLCE1	10	96066341	rs2274223	10	96068480	rs11187850	0.349
ALDH2	12	112204427	rs886205	12	112521448	rs4767364	0.404
RUNX1	21	36357861	rs2014300	21	36360884	rs2834718	0.283
CHEK2	22	29115066	rs4822983	22	29130300	rs1033667	0.470

[1] Genome coordinate for the SNP

Appendix Table 2A1

Appendix Table 2A1. Quality Assessment of Genetic Susceptibility Studies

Study	Power calculations reported	Description of ESCC diagnosis	Screening of Controls for ESCC	Detailed population characteristics for cases	Detailed population characteristics for controls	Adjustments for population stratification	NCBI rs numbers	Assessment of HWE	Assessment of genotyping error	Reported data as risk ratios	Correction for multiple testing	Quality score(0 to 11)
Bye et al 2012	Yes	Yes	No	Yes	Yes	No	Yes	Yes	No	Yes	Yes	8
Bye et al 2011	Yes	Yes	No	Yes	No	No	Yes	Yes	No	Yes	Yes	7
Chen et al 2019	Yes	Yes	No	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes	9
Chelule at al 2006	No	Yes	No	No	No	No	No	Yes	No	No	No	2
Dandara at al 2005	No	Yes	No	No	No	No	No	No	No	Yes	No	2
Dandara at al 2006	No	Yes	No	Yes	Yes	No	No	Yes	No	Yes	No	5
Dietzsch et al 2003	No	Yes	No	Yes	Yes	No	No	No	No	Yes	No	4
Eltahir et al 2012	No	Yes	No	No	No	No	No	No	No	Yes	No	2
Li et al 2005	No	Yes	No	Yes	Yes	No	No	Yes	No	Yes	No	5
Li et al 2010	Yes	Yes	No	Yes	Yes	No	Yes	Yes	No	Yes	Yes	8
Li et al 2008	No	Yes	No	Yes	Yes	No	No	No	Yes	Yes	No	5
Matejcic et al 2011	Yes	Yes	No	Yes	Yes	No	Yes	Yes	No	Yes	Yes	8
Matejcic 2015	No	Yes	No	Yes	Yes	No	Yes	Yes	No	Yes	Yes	7
Strickland et al 2012	No	Yes	Yes	Yes	No	No	Yes	Yes	No	No	No	5
Vogelsang et al 2012	Yes	Yes	No	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes	9
Vos et al 2003	No	Yes	No	No	No	No	No	No	No	No	No	1
Zaahl et al 2005	No	Yes	No	No	No	No	No	No	No	No	No	1

ESCC, esophageal squamous cell carcinoma; HWE, Hardy Weinberg equilibrium; NCBI, National Center for Biotechnology Information (https://www.ncbi.nlm.nih.gov/): rs-numbers can be obtained from the NCBI dbSNP database, which is available at https://www.ncbi.nlm.nih.gov/snp/

Appendix Table 2A2

Table 2A2. Quality Assessment of Somatic Variant Studies

Study	Description of ESCC diagnosis	Tissues used: Cancerous & Normal neighbouring tissue, or blood	Detailed population characteristics	Variant classification and type	Confirmation of results	Amino acid change reported	Use of pathogenicity scoring described	Quality score(0 to 7)
Dietzsch et al 2003	No	Yes	Yes	Yes	No	NA	NA	3
Dietzsch et al 2002	Yes	Yes	Yes	Yes	No	NA	NA	4
Gamieldien et al 1998	Yes	Yes	Yes	Yes	Yes	Yes	No	6
Liu et al 2016	Yes	Yes	Yes	Yes	No	No	No	4
Naidoo et al 2005	Yes	Yes	Yes	Yes	No	NA	NA	4
Patel et al 2011	Yes	No	Yes	Yes	No	Yes	No	4
Victor et al 1990	No	No	No	NA	NA	NA	NA	0
Vos et al 2003	Yes	Yes	No	Yes	Yes	Yes	No	5

ESCC, esophageal squamous cell carcinoma; NA, not applicable

Appendix Table 2A3

Table 2A3: Summary of SNPs with r^2>0.20.

Gene Symbol	Chromosome A	BP A[1]	SNP A	Chromosome B	BP B[1]	SNP B	R^2
MSH2	2	47690411	rs3771280	2	47709153	rs10188090	0.787
CASP8	2	202097531	rs3834129	2	202143928	rs10931936	0.256
CASP8/ ALS2CR12	2	202143928	rs10931936	2	202162811	rs13016963	0.315
ALS2CR12	2	202162811	rs13016963	2	202202791	rs10201587	0.409
CP	3	148919880	rs35272481	3	148919962	rs34237139	1
CP	3	148939861	rs17838831	3	148939929	rs34334174	0.515
CP	3	148939861	rs17838831	3	148939933	rs17838832	0.884
CP	3	148939861	rs17838831	3	148940142	rs17838834	0.884
CP	3	148939929	rs34334174	3	148939933	rs17838832	0.579
CP	3	148939929	rs34334174	3	148940142	rs17838834	0.580
CP	3	148939933	rs17838832	3	148940142	rs17838834	1
MSH3	5	79966029	rs1805355	5	80008704	rs1428030	0.809
NAT1	8	18080644	rs1057126	8	18080651	rs15561	0.908
PLCE1	10	96043732	rs7084339	10	96058298	rs3765524	0.970
PLCE1	10	96043732	rs7084339	10	96066341	rs2274223	0.572
PLCE1	10	96043732	rs7084339	10	96068480	rs11187850	0.224
PLCE1	10	96058298	rs3765524	10	96066341	rs2274223	0.585
PLCE1	10	96058298	rs3765524	10	96068480	rs11187850	0.233
PLCE1	10	96066341	rs2274223	10	96068480	rs11187850	0.349
ALDH2	12	112204427	rs886205	12	112521448	rs4767364	0.404
RUNX1	21	36357861	rs2014300	21	36360884	rs2834718	0.283
CHEK2	22	29115066	rs4822983	22	29130300	rs1033667	0.470

[1] Genome coordinate for the SNP

Appendix Table 3A1

Table 3A1: Quality assessment of reporting and methodology in included studies

Study (PMID)	**Is the sample representative of patients in the population as a whole?**	**Are the patients at a similar point in the course of their condition/illness?**	**Has bias been minimised in relation to cases and controls?**	**Are confounding factors identified and strategies to deal with them stated?**	**Description of ESCC diagnosis**	**Were controls screened and tested for ESCC?**	**Were response rates reported?**	**Were outcomes measured in a reliable way?**	**Was appropriate statistical analysis used?**	**Score**
	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Total (Y)
Asombang 2016 (26973419)	N	U	Y	U	Y	Y	N	Y	Y	5
Astini et al 1990 (2083189)	N	U	Y	N	Y	N	N	Y	Y	4
Dandara et al 2005 (15978331)	Y	U	Y	N	Y	Y	N	Y	Y	6
Dandara et al 2006 (16272171)	Y	U	Y	Y	Y	N	N	Y	Y	6
Kayamba 2015 (25641622)	N	U	Y	Y	Y	N	N	Y	Y	5
Kgomo et al 2017 (29177066)	N	U	Y	Y	Y	Y	N	Y	Y	6
Leon et al 2017 (28594883)	N	U	Y	Y	Y	N	N	Y	Y	5
Li et al 2005 (15899651)	Y	U	Y	Y	Y	N	N	Y	Y	6
Machoki et al 2015	N	U	Y	N	Y	N	N	Y	Y	4

Matejcic et al 2015 (26447020)	Y	U	Y	Y	Y	N	N	Y	Y	6
Matsha et al 2006 (17176219)	Y	U	N	Y	Y	N	N	N	Y	4
Matsha b et al 2006 (16607430)	Y	N	NA	N	Y	NA	N	Y	N	3
Menya et al 2019 (30117158)	Y	Y	Y	Y	Y	N	N	Y	Y	7
Menya b et al 2019 (30582155)	Y	U	Y	Y	Y	N	N	Y	Y	6
Middleton et al 2019 (30496610)	Y	N	Y	Y	Y	N	N	Y	Y	6
Ocama et al 2008 (19357755)	N	U	U	Y	Y	N	N	Y	Y	4
Okello et al 2016 (27400987)	N	U	U	Y	Y	Y	N	Y	Y	5
Pacella-Norman 2002 (12087462)	Y	U	U	Y	Y	U	N	Y	Y	5
Parkin et al 1994 (7827583)	Y	U	U	Y	Y	U	N	Y	Y	5
Patel et al 2013 (24490085)	Y	U	Y	Y	N	N	N	Y	Y	5
Sammon 1992 (1735077)	N	U	Y	N	Y	N	N	N	Y	4
Sammon et al 1998 (9690530)	N	U	Y	U	Y	N	N	N	Y	3
Schaafsma et al 2015 (26448405)	Y	N	NA	Y	U	NA	NA	U	Y	3
Segal et al 1988 (3219281)	Y	U	Y	N	N	N	N	U	Y	3

Sewram et al 2014 (24877989)	Y	U	Y	Y	N	N	Y	Y	Y	6
Sewram et al 2016 (26900781)	Y	U	Y	Y	Y	N	Y	Y	Y	7
Shewaye et al 2015	Y	U	N	U	Y	N	N	Y	Y	4
Sitas et al 2007 (17331260)	U	N	U	Y	N	U	N	U	Y	2
van Rensburg et al 1985 (3970816)	Y	U	Y	Y	N	N	N	U	Y	4
Vizcaino et al 1995 (7669592)	Y	N	U	Y	Y	U	N	Y	Y	5
Vogelsang et al 2012 (22623965)	Y	U	Y	Y	Y	N	N	Y	Y	6

ESCC; esophageal squamous cell carcinoma

Appendix Figure 3A2

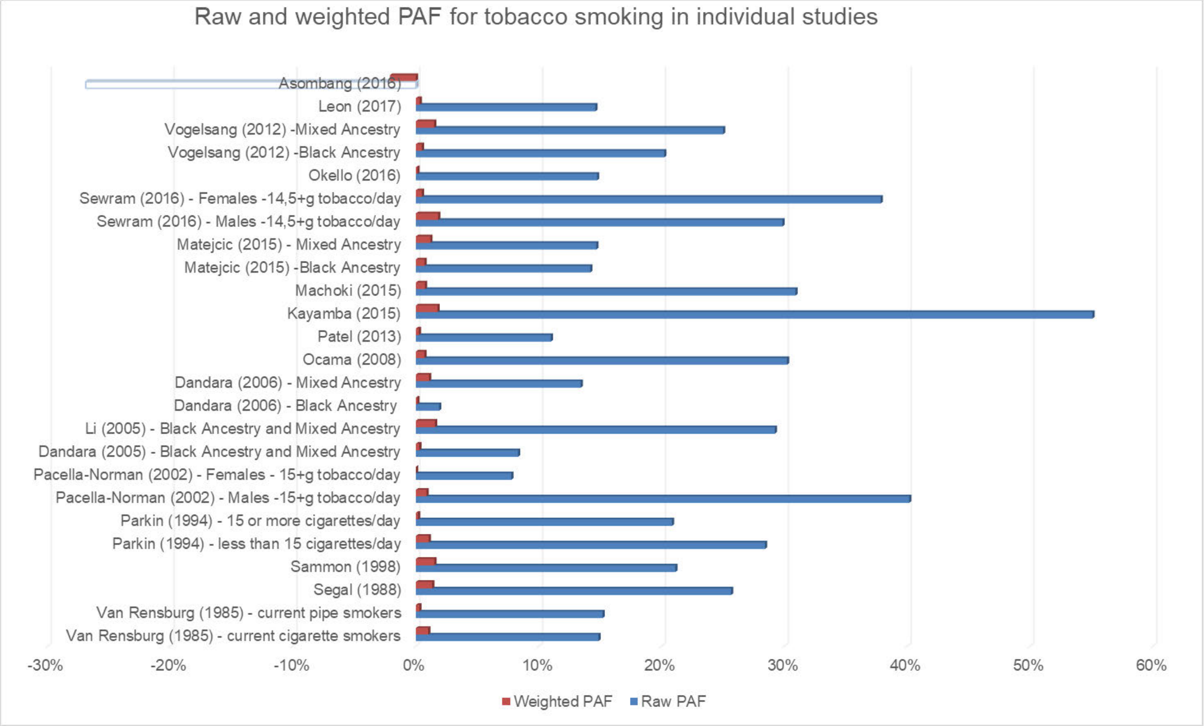


Figure 3A2: Bar graph showing raw and weighted PAF values for tobacco smoking in individual studies. PAF; population attributable fraction

Appendix Figure 3A3

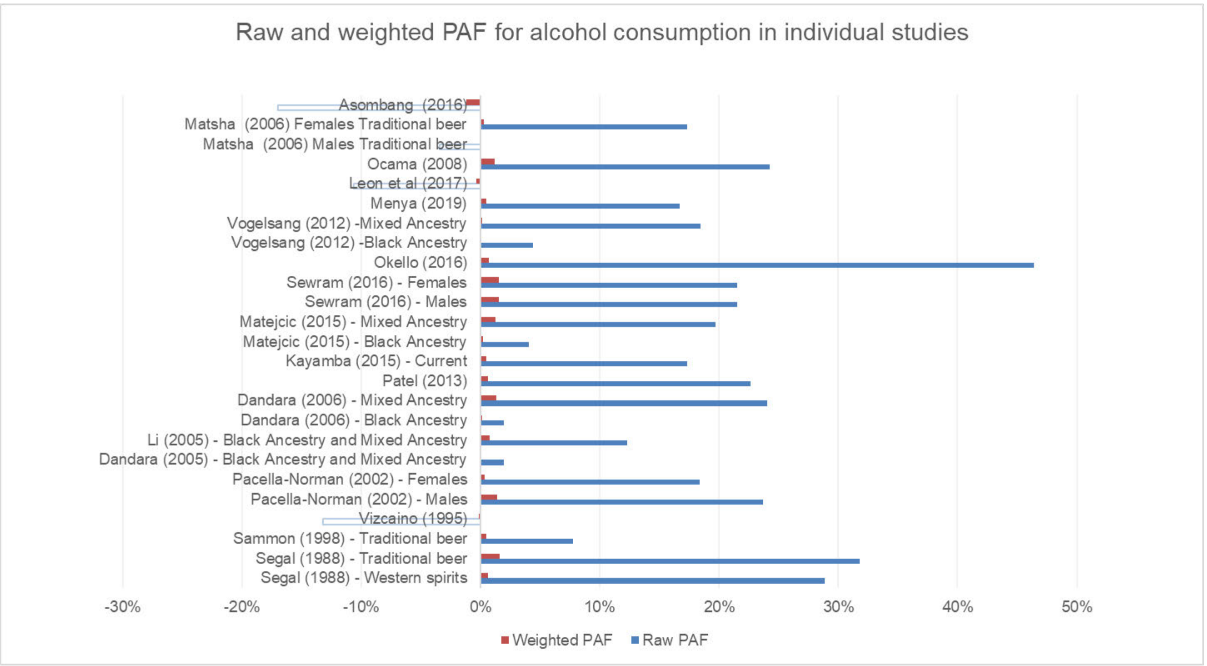


Figure 3A3: Bar graph showing raw and weighted PAF values for alcohol consumption in individual studies. PAF; population attributable fraction

Appendix Figure 3A4

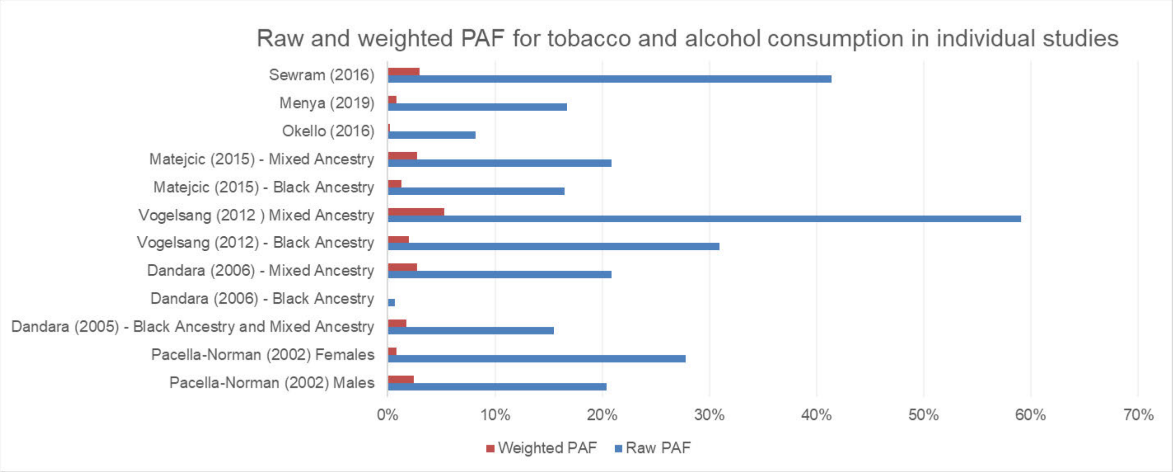


Figure 3A4: Bar graph showing raw and weighted PAF values for tobacco and alcohol consumption in individual studies. PAF; population attributable fraction.

Appendix Figure 3A5

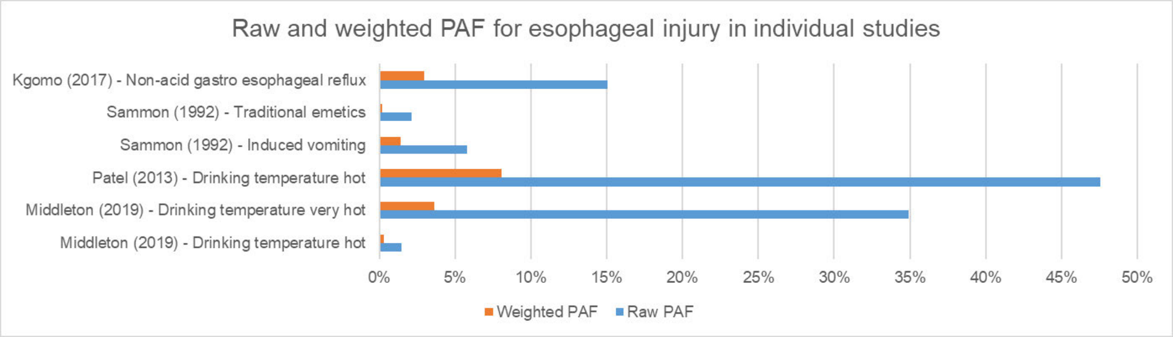


Figure 3A5: Bar graph showing raw and weighted PAF values for esophageal injury in individual studies. PAF; population attributable fraction.

Appendix Figure 3A6

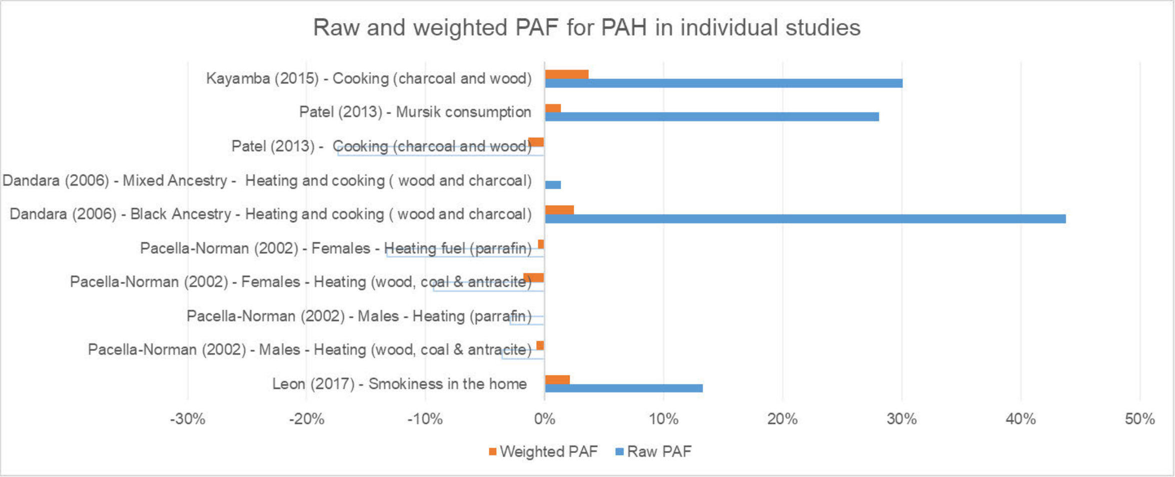


Figure 3A6: Bar graph showing raw and weighted PAF values for PAH in individual studies. PAF; population attributable fraction, PAH; polycyclic aromatic hydrocarbons

Appendix Figure 3A7

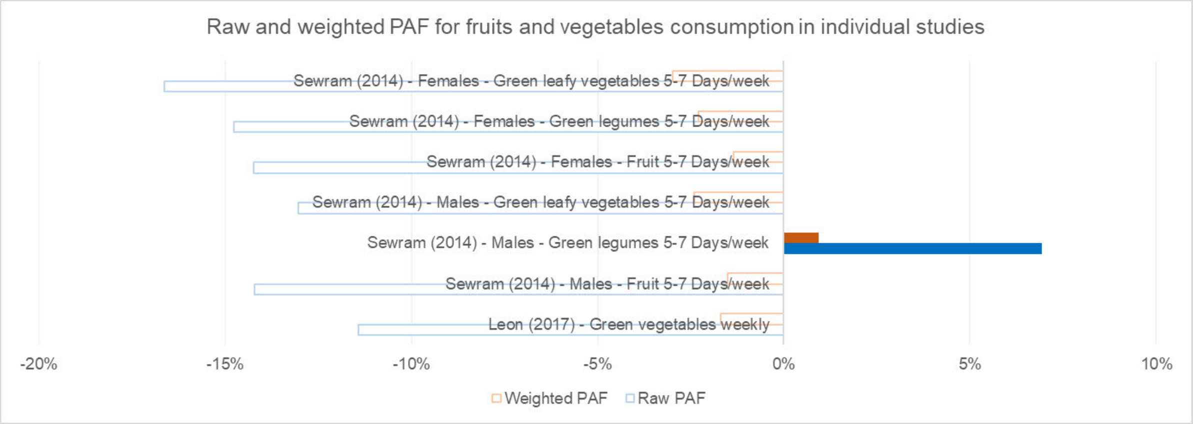


Figure 3A7: Bar graph showing raw and weighted PAF values for fruits and vegetables individual studies. PAF; population attributable fraction, PAH; polycyclic aromatic hydrocarbons

Appendix 4A1: Quality Control reports for included studies

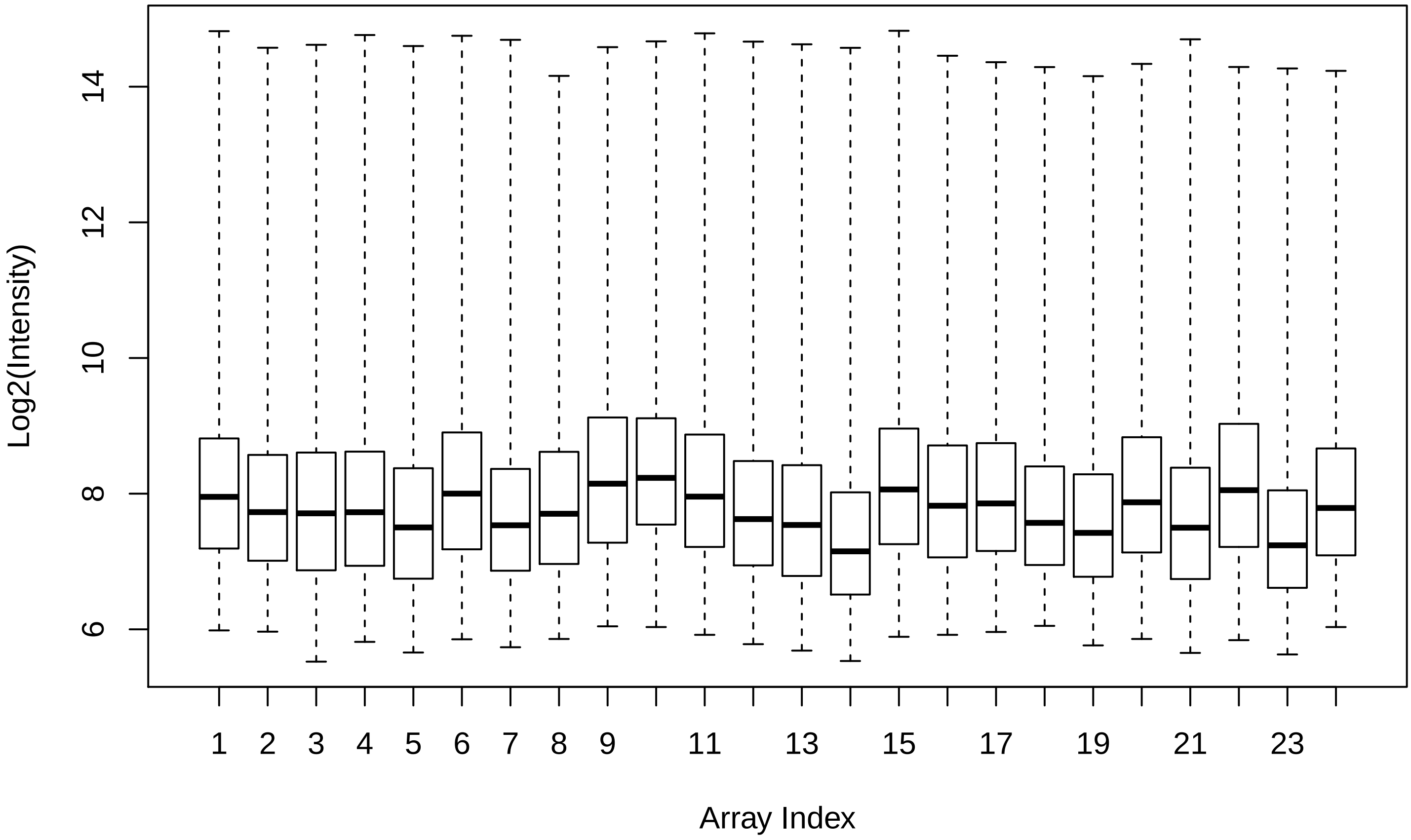
AffyBatch QC Report

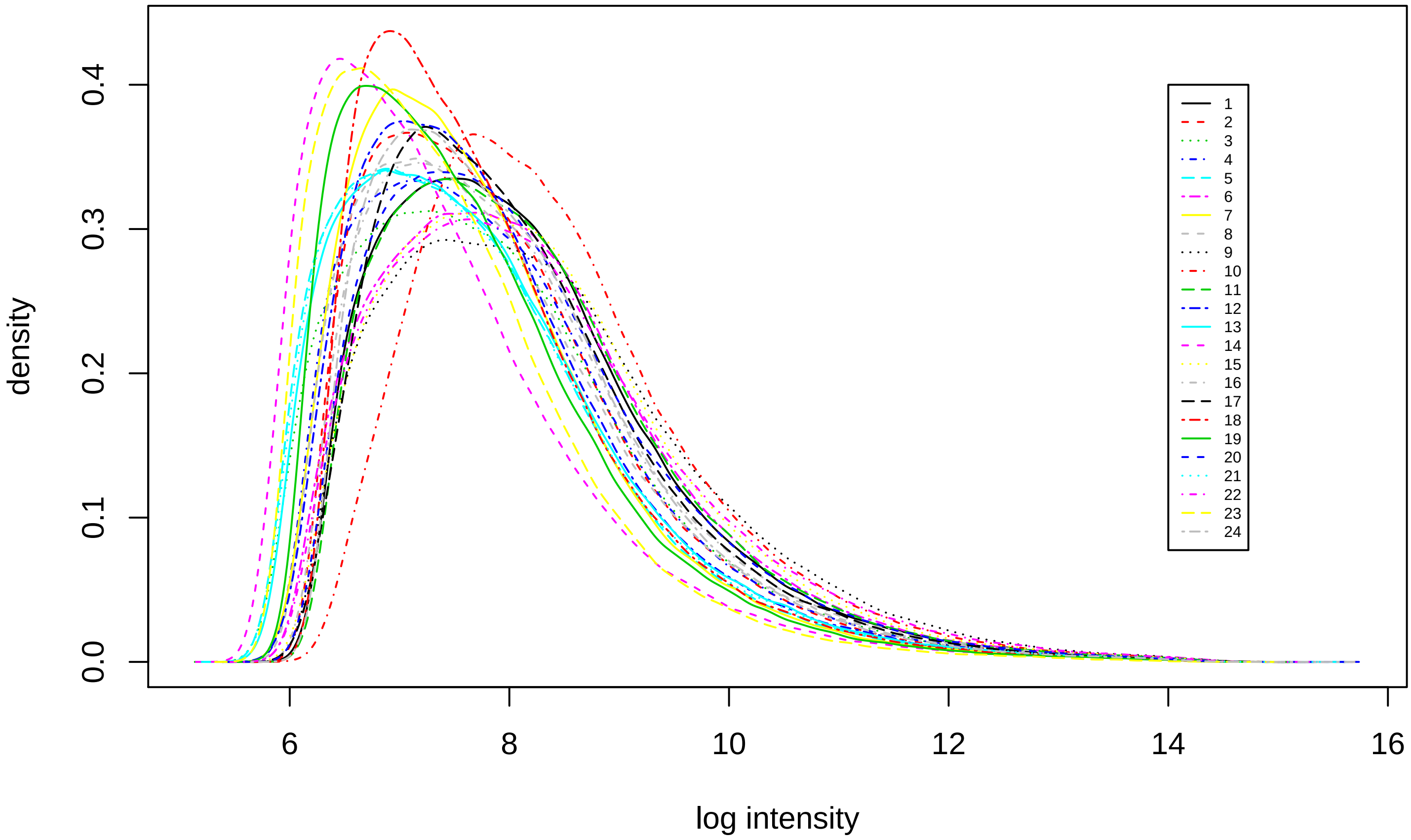
GSE1420

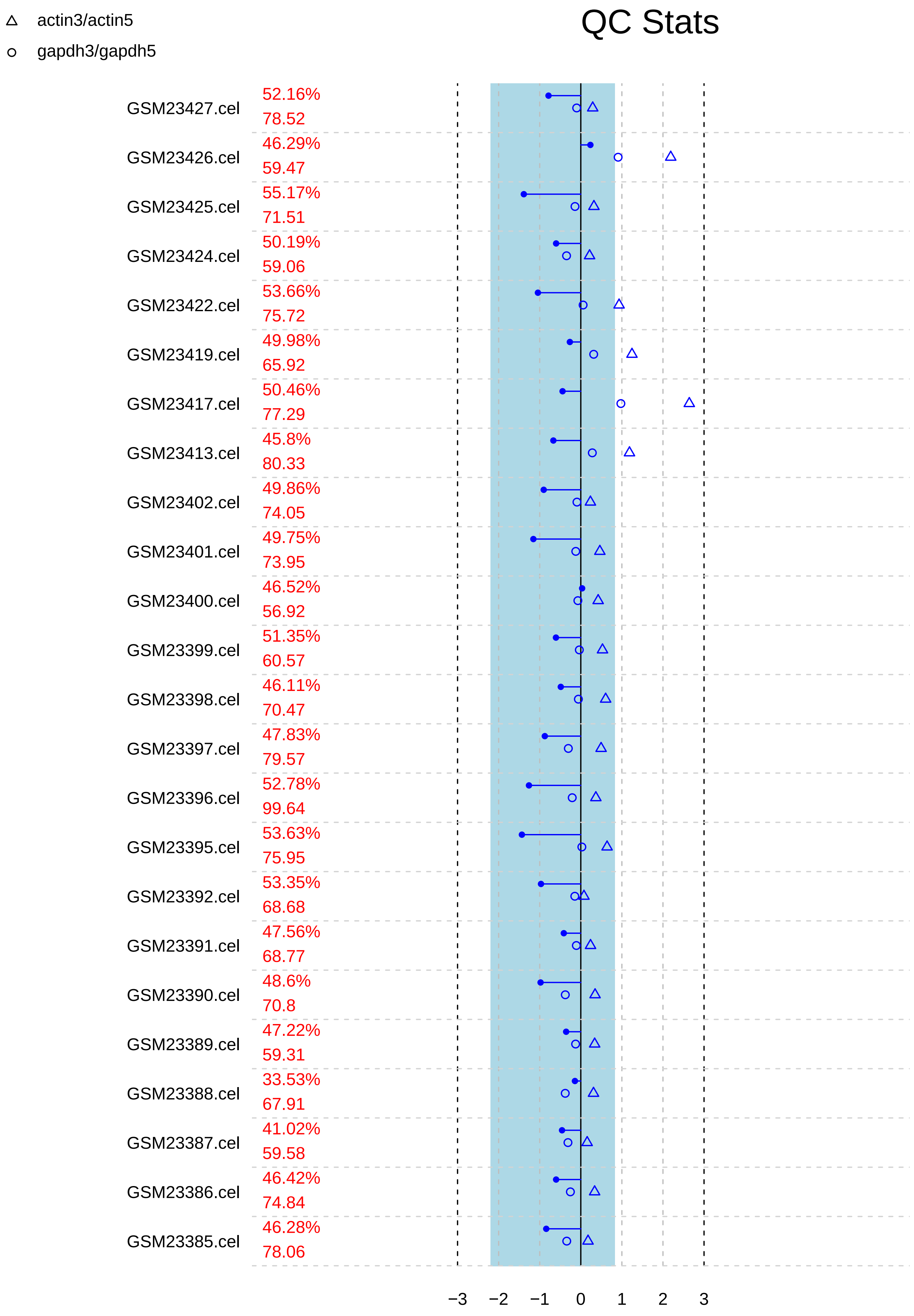
Array Index	Array Name
1	GSM23385.cel
2	GSM23386.cel
3	GSM23387.cel
4	GSM23388.cel
5	GSM23389.cel
6	GSM23390.cel
7	GSM23391.cel
8	GSM23392.cel
9	GSM23395.cel
10	GSM23396.cel
11	GSM23397.cel
12	GSM23398.cel
13	GSM23399.cel
14	GSM23400.cel
15	GSM23401.cel
16	GSM23402.cel
17	GSM23413.cel
18	GSM23417.cel
19	GSM23419.cel
20	GSM23422.cel
21	GSM23424.cel
22	GSM23425.cel
23	GSM23426.cel
24	GSM23427.cel

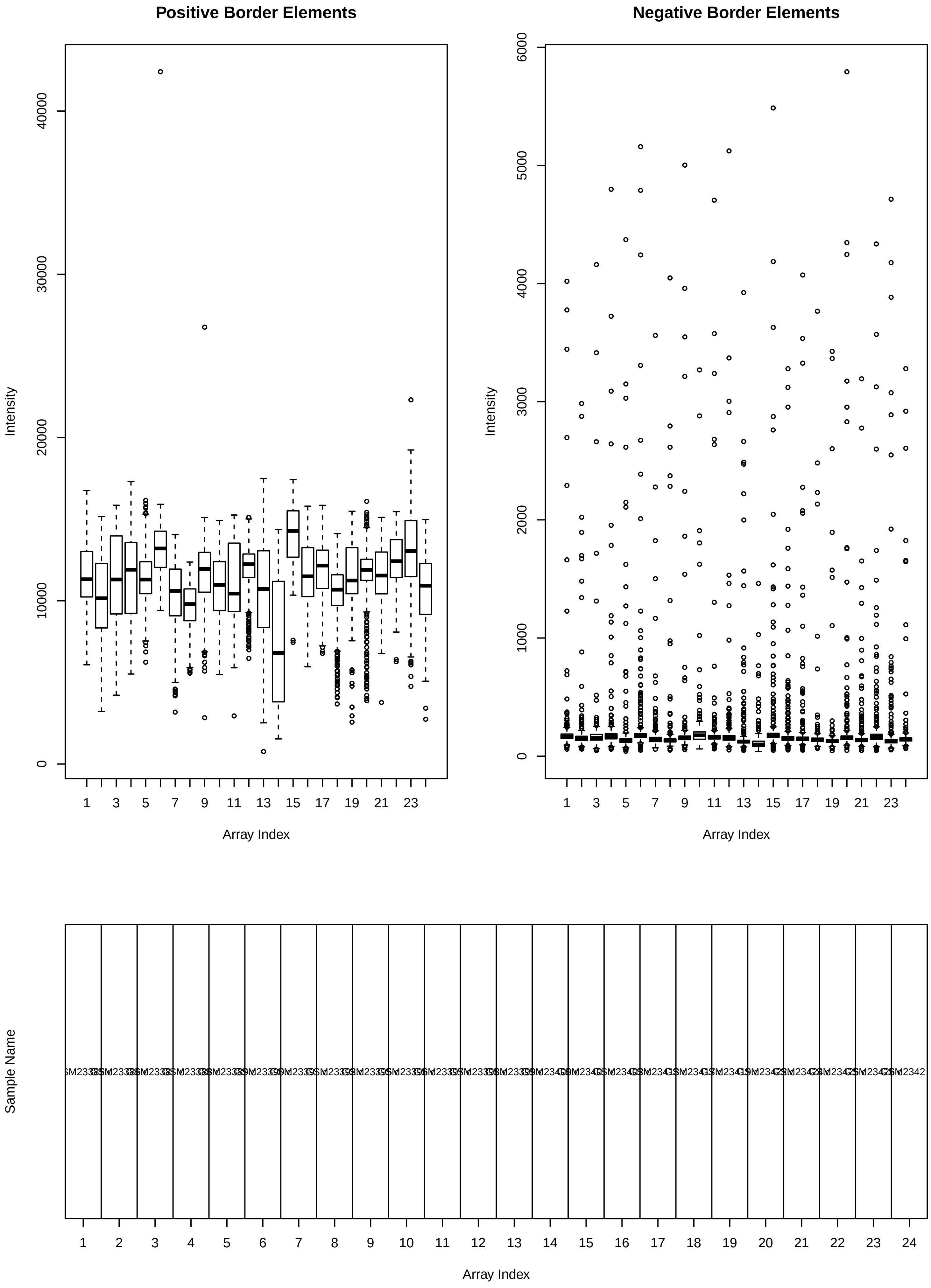
Thu Oct 1 16 01 31 2020

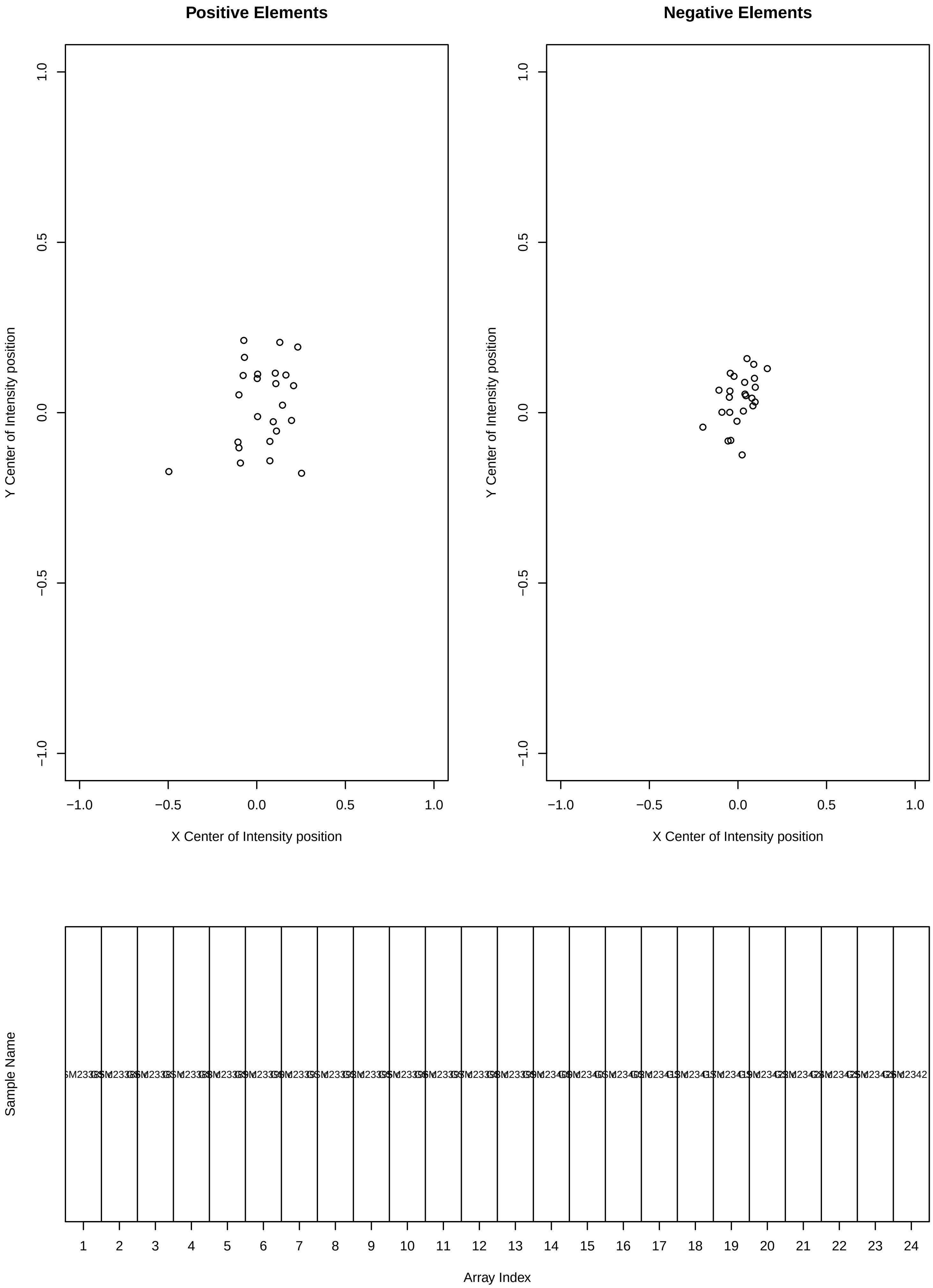
Produced by AffyQCReport R Package

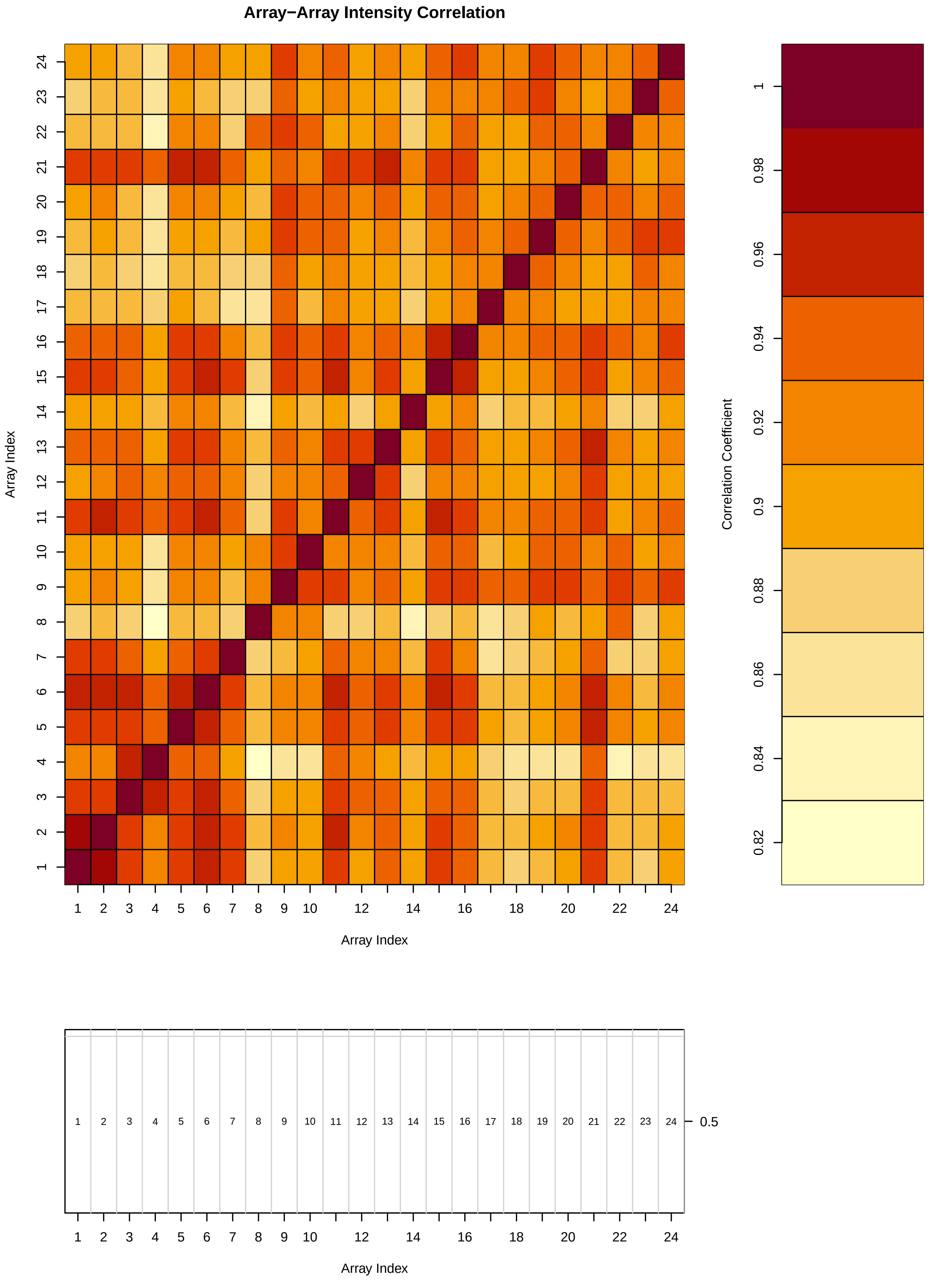

Log2(Intensity)
6
8
10
12
14
1
2
3
4
5
6
7
8
9
11
13
15
17
19
21
23
Array Index


density
0.0
0.1
0.2
0.3
0.4
6
8
10
12
14
16
log intensity
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24


△ actin3/actin5
○ gapdh3/gapdh5
QC Stats
GSM23427.cel
52.16%
78.52
GSM23426.cel
46.29%
59.47
GSM23425.cel
55.17%
71.51
GSM23424.cel
50.19%
59.06
GSM23422.cel
53.66%
75.72
GSM23419.cel
49.98%
65.92
GSM23417.cel
50.46%
77.29
GSM23413.cel
45.8%
80.33
GSM23402.cel
49.86%
74.05
GSM23401.cel
49.75%
73.95
GSM23400.cel
46.52%
56.92
GSM23399.cel
51.35%
60.57
GSM23398.cel
46.11%
70.47
GSM23397.cel
47.83%
79.57
GSM23396.cel
52.78%
99.64
GSM23395.cel
53.63%
75.95
GSM23392.cel
53.35%
68.68
GSM23391.cel
47.56%
68.77
GSM23390.cel
48.6%
70.8
GSM23389.cel
47.22%
59.31
GSM23388.cel
33.53%
67.91
GSM23387.cel
41.02%
59.58
GSM23386.cel
46.42%
74.84
GSM23385.cel
46.28%
78.06
−3
−2
−1
0
1
2
3


Positive Border Elements
Intensity
0
10000
20000
30000
40000
Array Index
1
3
5
7
9
11
13
15
17
19
21
23
Negative Border Elements
Intensity
0
1000
2000
3000
4000
5000
6000
Array Index
Sample Name
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
Array Index


Positive Elements
Negative Elements
Y Center of Intensity position
X Center of Intensity position
−1.0
−0.5
0.0
0.5
1.0
Sample Name
Array Index
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24


Array−Array Intensity Correlation
Array Index
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
1
2
3
4
5
6
7
8
9
10
12
14
16
18
20
22
24
Array Index
Correlation Coefficient
1
0.98
0.96
0.94
0.92
0.9
0.88
0.86
0.84
0.82
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
0.5
1
2
3
4
5
6
7
8
9
10
12
14
16
18
20
22
24
Array Index

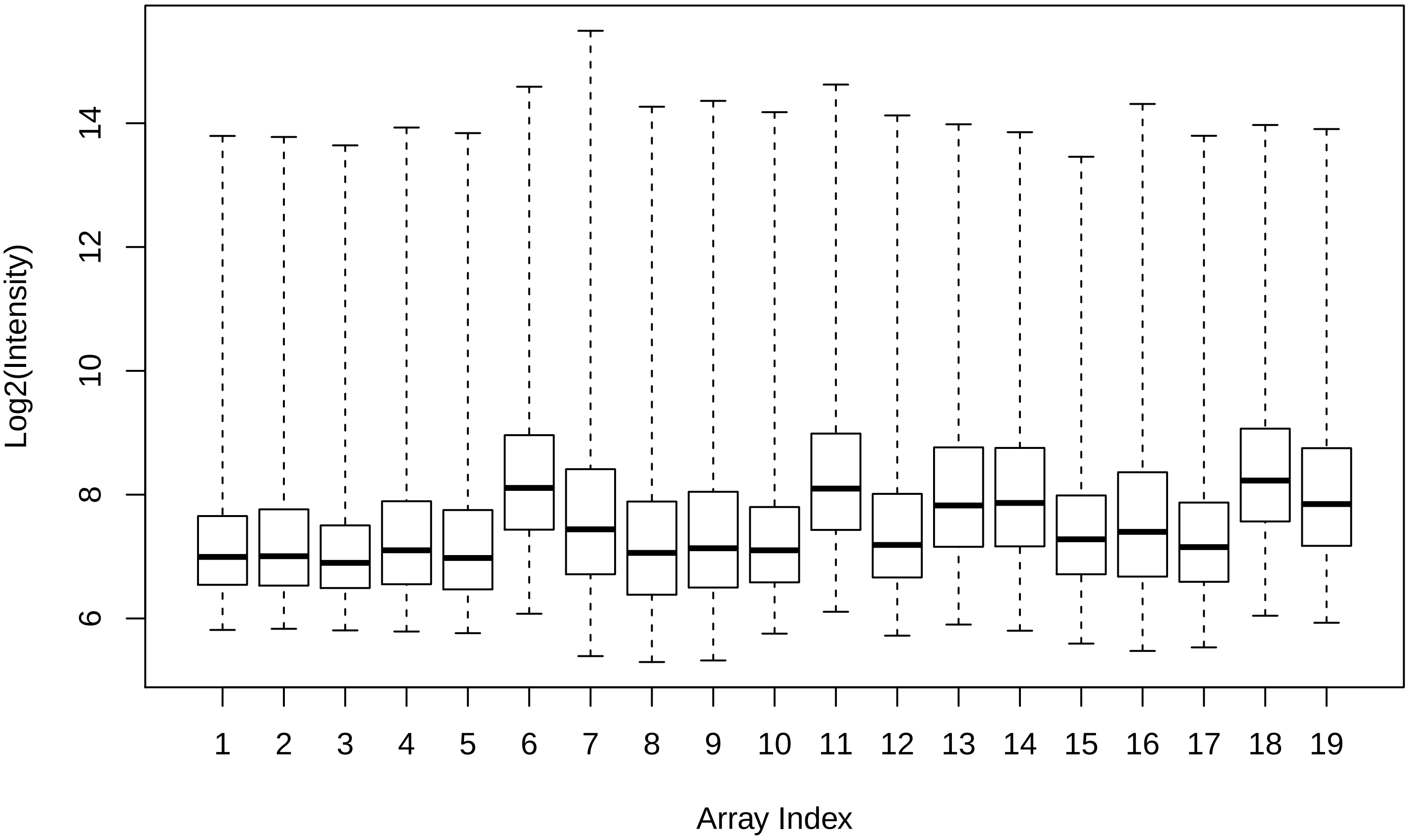
AffyBatch QC Report

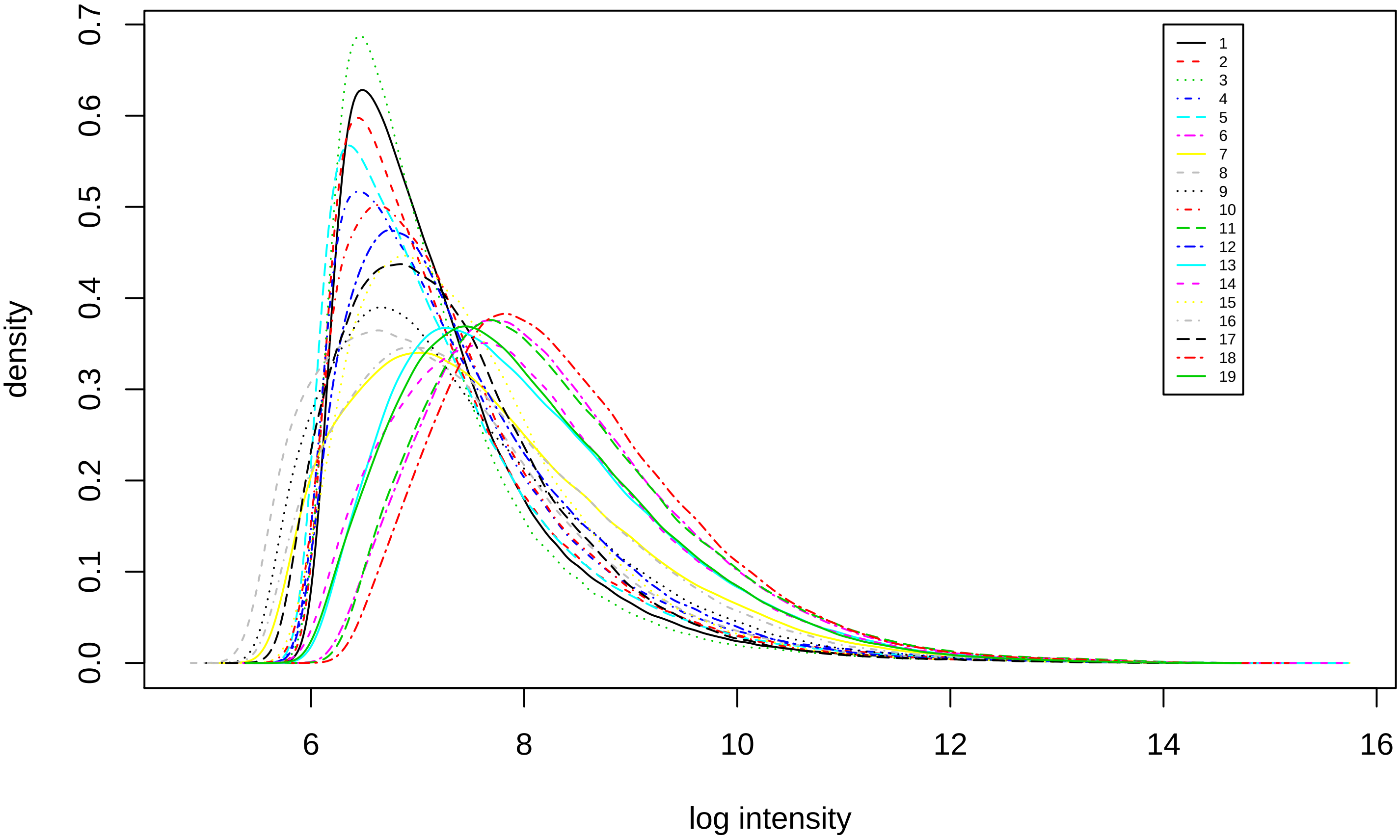
GSE13083

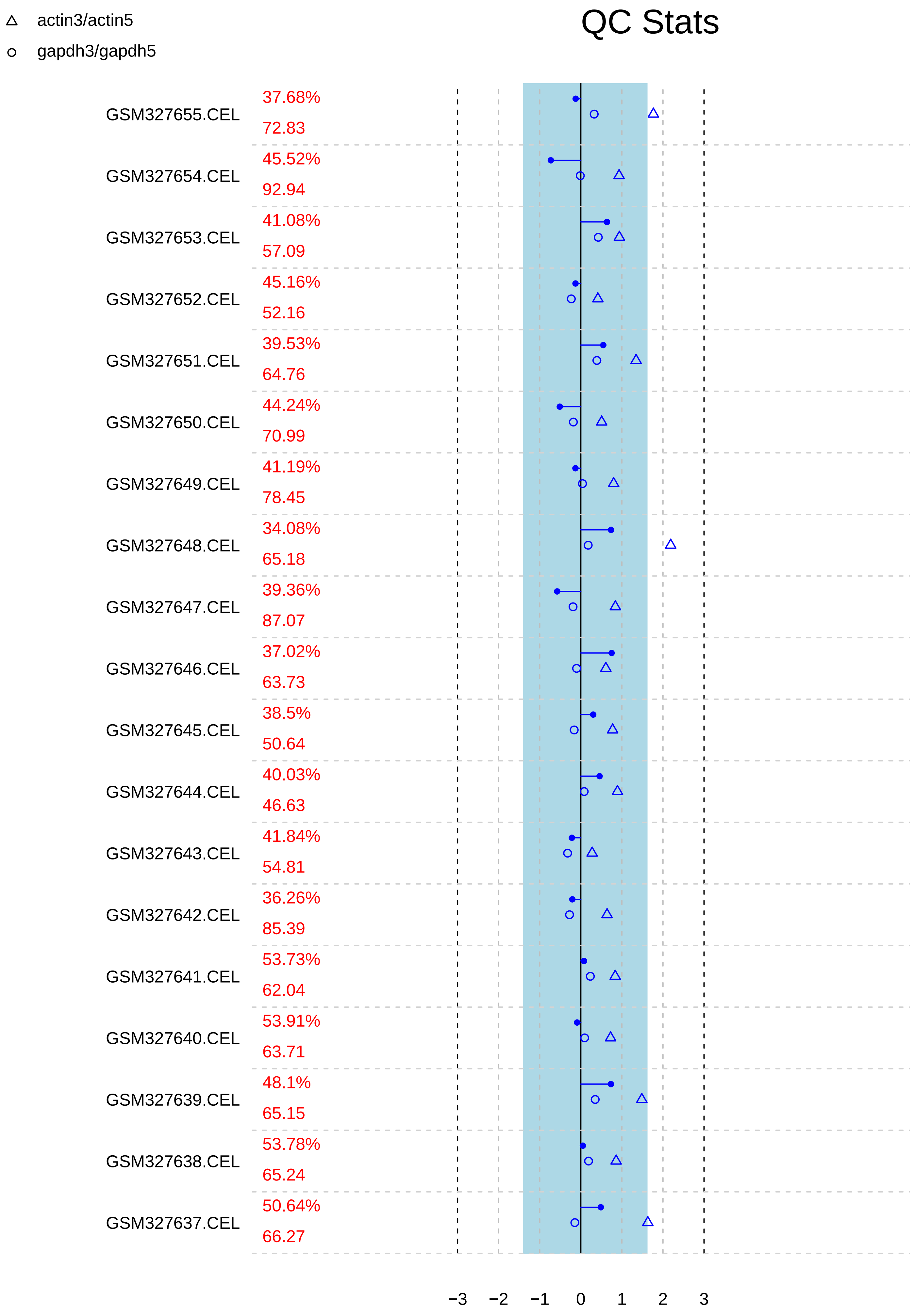
Array Index	Array Name
1	GSM327637.CEL
2	GSM327638.CEL
3	GSM327639.CEL
4	GSM327640.CEL
5	GSM327641.CEL
6	GSM327642.CEL
7	GSM327643.CEL
8	GSM327644.CEL
9	GSM327645.CEL
10	GSM327646.CEL
11	GSM327647.CEL
12	GSM327648.CEL
13	GSM327649.CEL
14	GSM327650.CEL
15	GSM327651.CEL
16	GSM327652.CEL
17	GSM327653.CEL
18	GSM327654.CEL
19	GSM327655.CEL

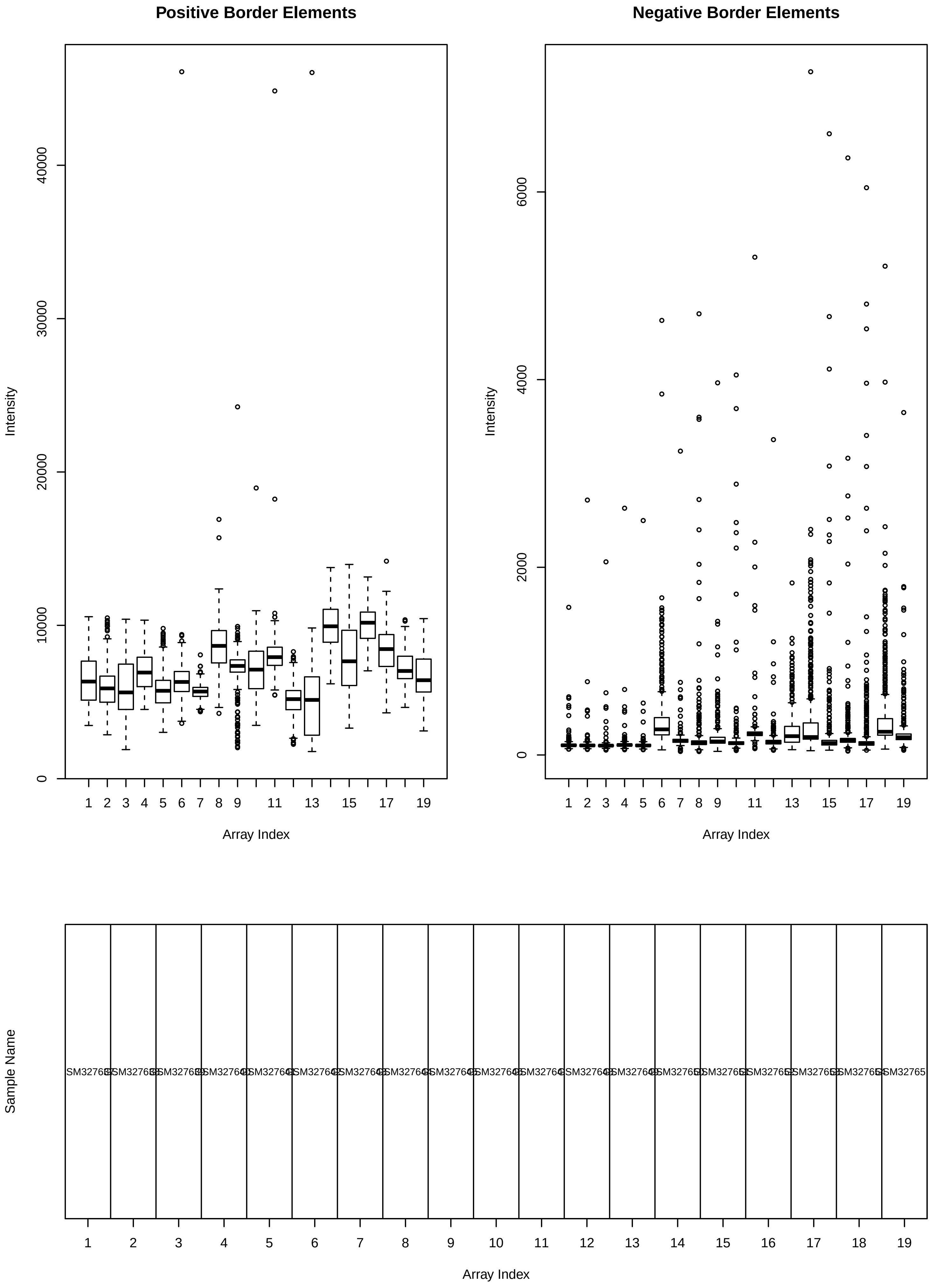
Thu Oct 1 15 51:03 2020

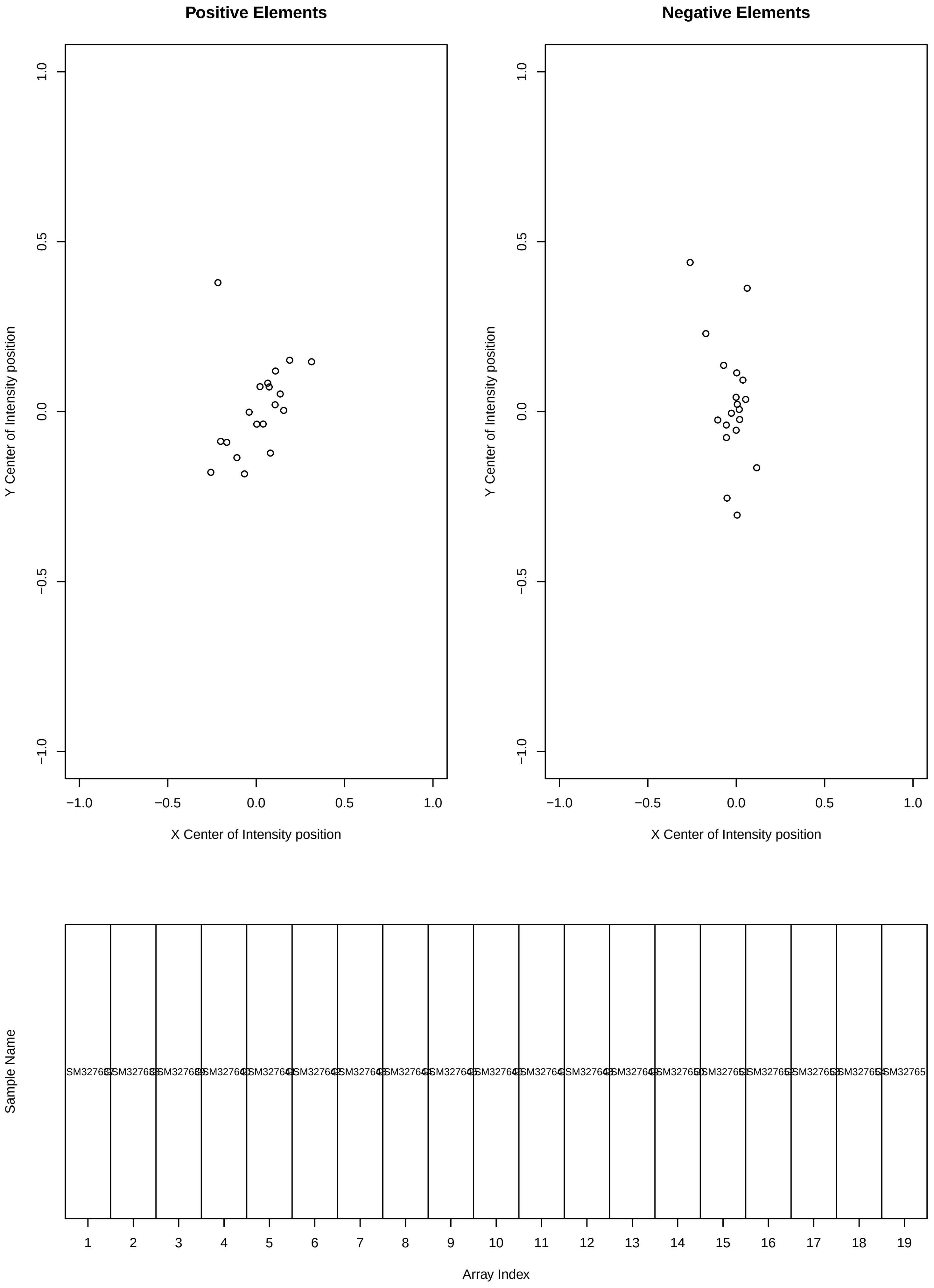
Produced by AffyQCReport R Package

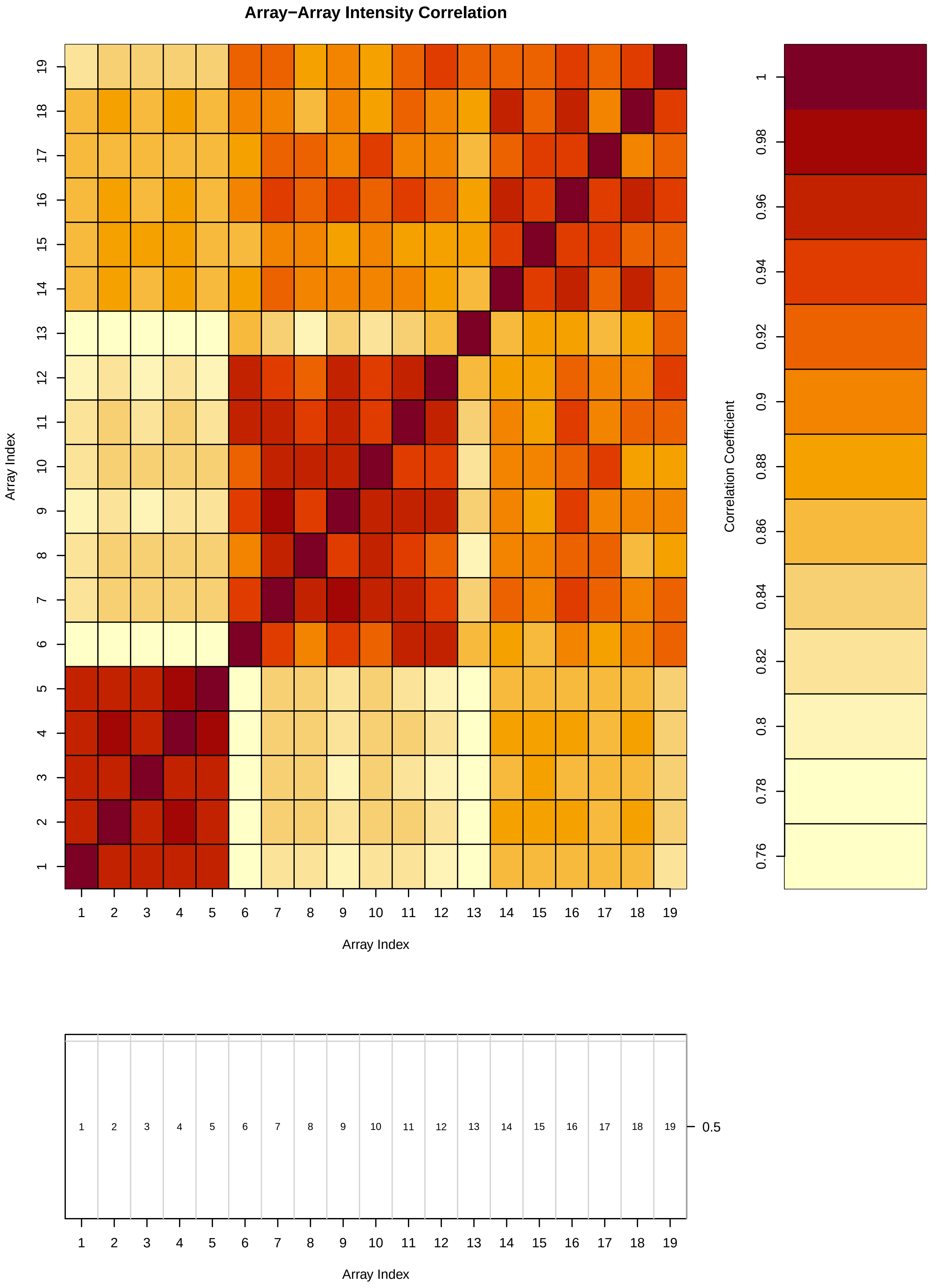

Log2(Intensity)
6
8
10
12
14
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
Array Index


density
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
6
8
10
12
14
16
log intensity
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19


△ actin3/actin5
○ gapdh3/gapdh5
QC Stats
GSM327655.CEL
37.68%
72.83
GSM327654.CEL
45.52%
92.94
GSM327653.CEL
41.08%
57.09
GSM327652.CEL
45.16%
52.16
GSM327651.CEL
39.53%
64.76
GSM327650.CEL
44.24%
70.99
GSM327649.CEL
41.19%
78.45
GSM327648.CEL
34.08%
65.18
GSM327647.CEL
39.36%
87.07
GSM327646.CEL
37.02%
63.73
GSM327645.CEL
38.5%
50.64
GSM327644.CEL
40.03%
46.63
GSM327643.CEL
41.84%
54.81
GSM327642.CEL
36.26%
85.39
GSM327641.CEL
53.73%
62.04
GSM327640.CEL
53.91%
63.71
GSM327639.CEL
48.1%
65.15
GSM327638.CEL
53.78%
65.24
GSM327637.CEL
50.64%
66.27
−3
−2
−1
0
1
2
3


Positive Border Elements
Intensity
0
10000
20000
30000
40000
1 2 3 4 5 6 7 8 9 11 13 15 17 19
Array Index
Negative Border Elements
Intensity
0
2000
4000
6000
1 2 3 4 5 6 7 8 9 11 13 15 17 19
Array Index
Sample Name
SM32763
GSM32763
GSM32763
GSM32764
GSM32764
GSM32764
GSM32764
GSM32764
GSM32764
GSM32764
GSM32764
GSM32764
GSM32764
GSM32765
GSM32765
GSM32765
GSM32765
GSM32765
GSM32765
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19
Array Index


Positive Elements
Negative Elements
Y Center of Intensity position
X Center of Intensity position
1.0
0.5
0.0
−0.5
−1.0
Sample Name
SM327637
GSM327638
GSM327639
GSM327640
GSM327641
GSM327642
GSM327643
GSM327644
GSM327645
GSM327646
GSM32764
GSM327648
GSM327649
GSM327650
GSM327651
GSM327652
GSM327653
GSM327654
GSM32765
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
Array Index


Array−Array Intensity Correlation
Array Index
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
Correlation Coefficient
0.76
0.78
0.8
0.82
0.84
0.86
0.88
0.9
0.92
0.94
0.96
0.98
1
0.5

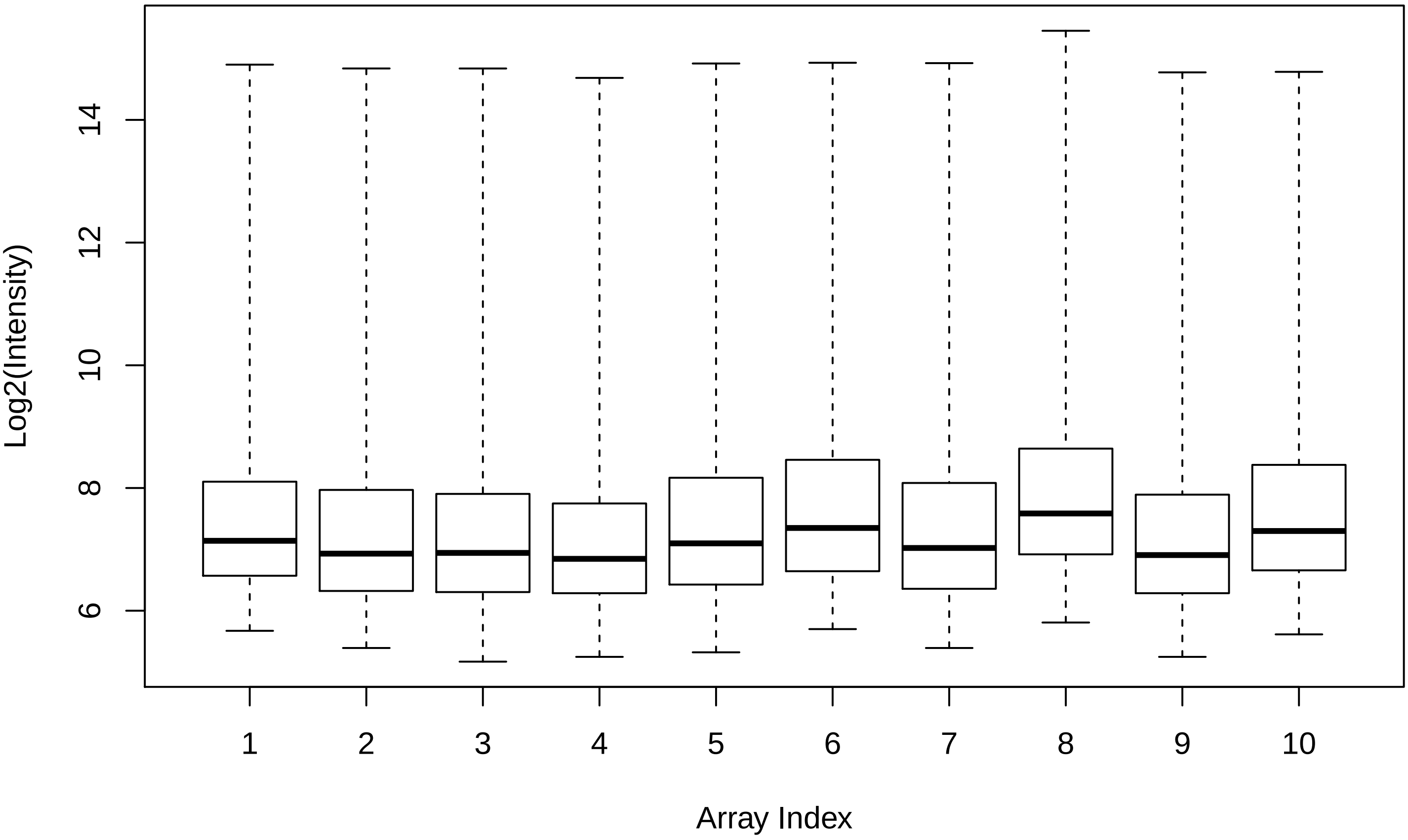
AffyBatch QC Report

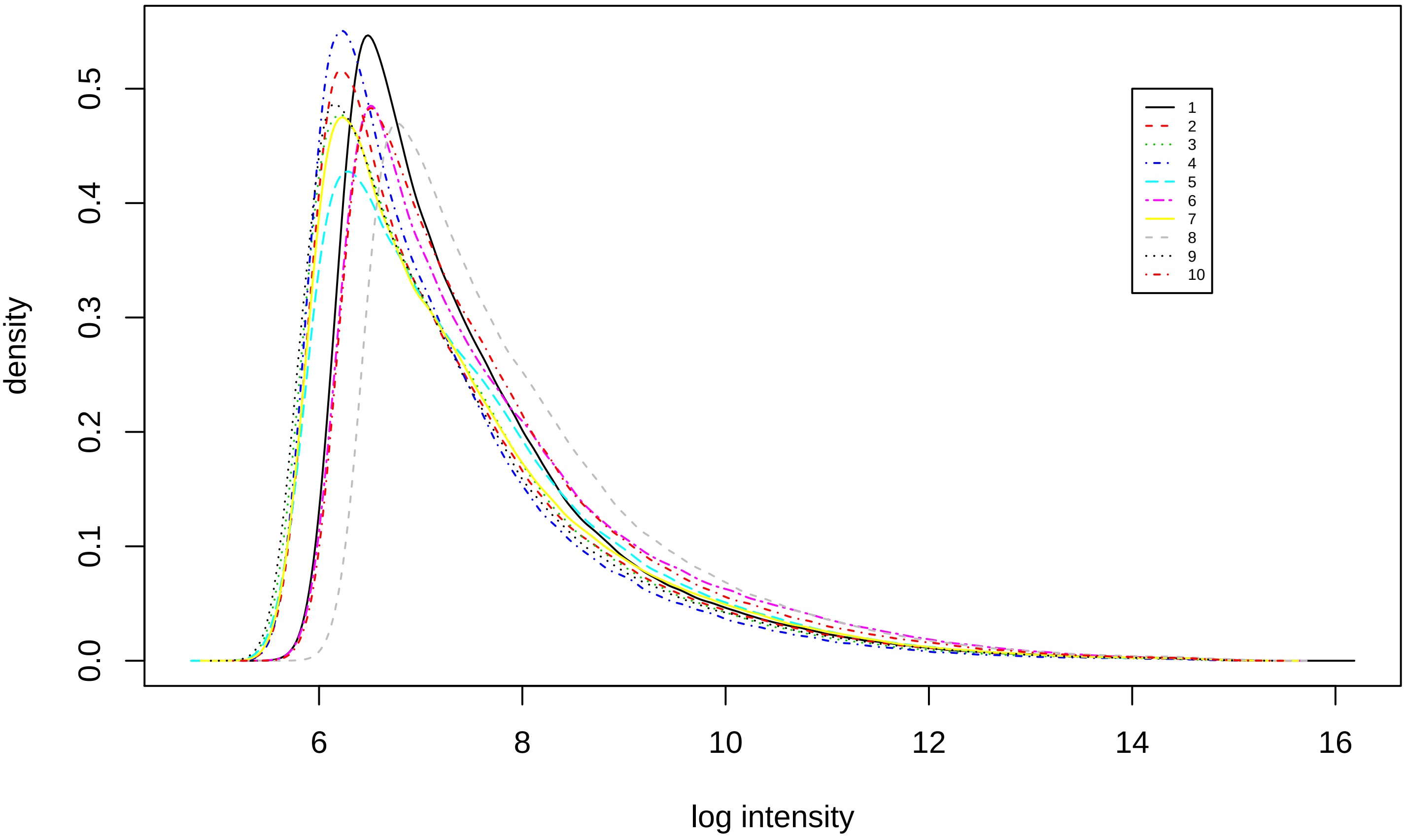
GSE17351

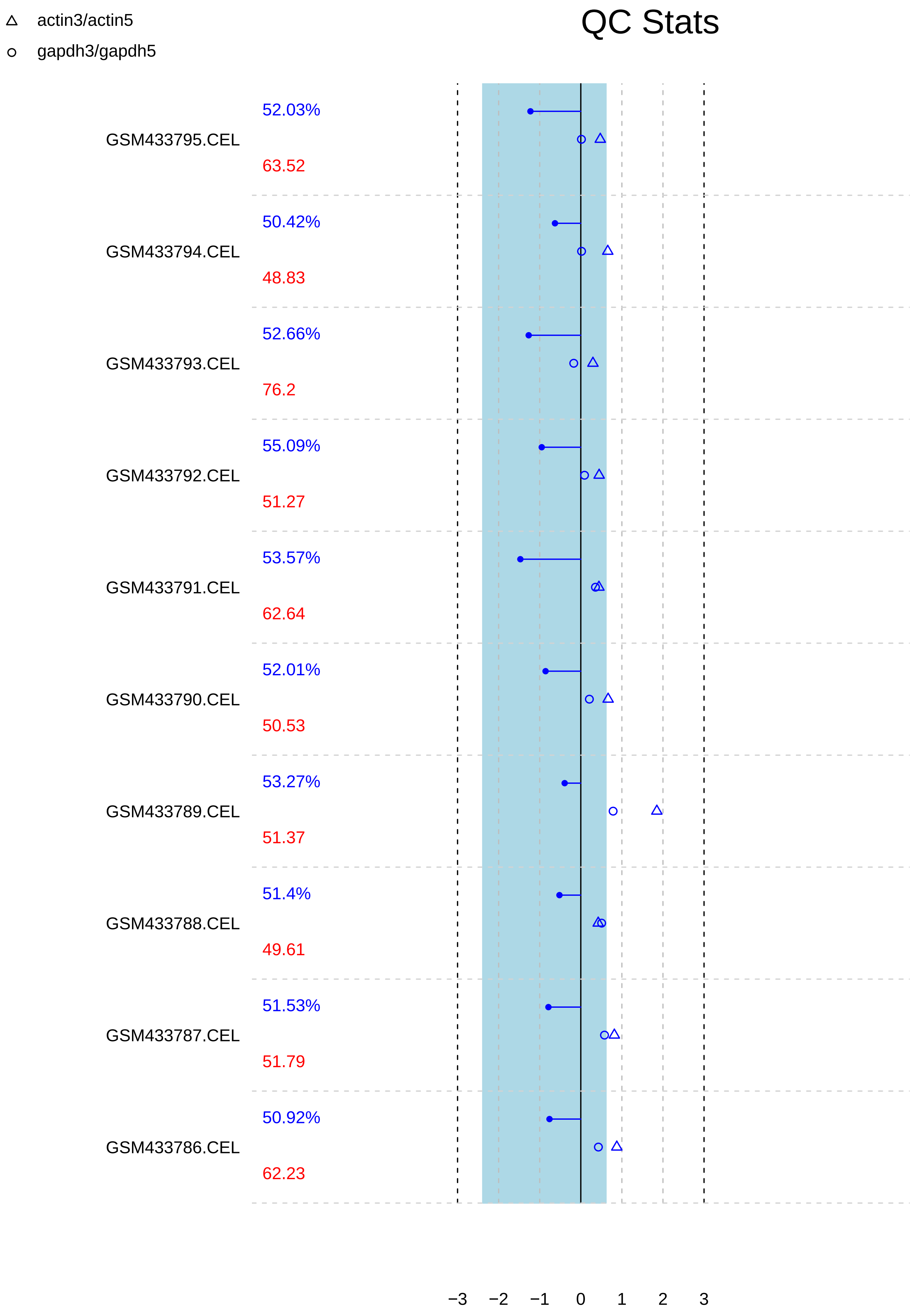
Array Index	Array Name
1	GSM433786.CEL
2	GSM433787.CEL
3	GSM433788.CEL
4	GSM433789.CEL
5	GSM433790.CEL
6	GSM433791.CEL
7	GSM433792.CEL
8	GSM433793.CEL
9	GSM433794.CEL
10	GSM433795.CEL

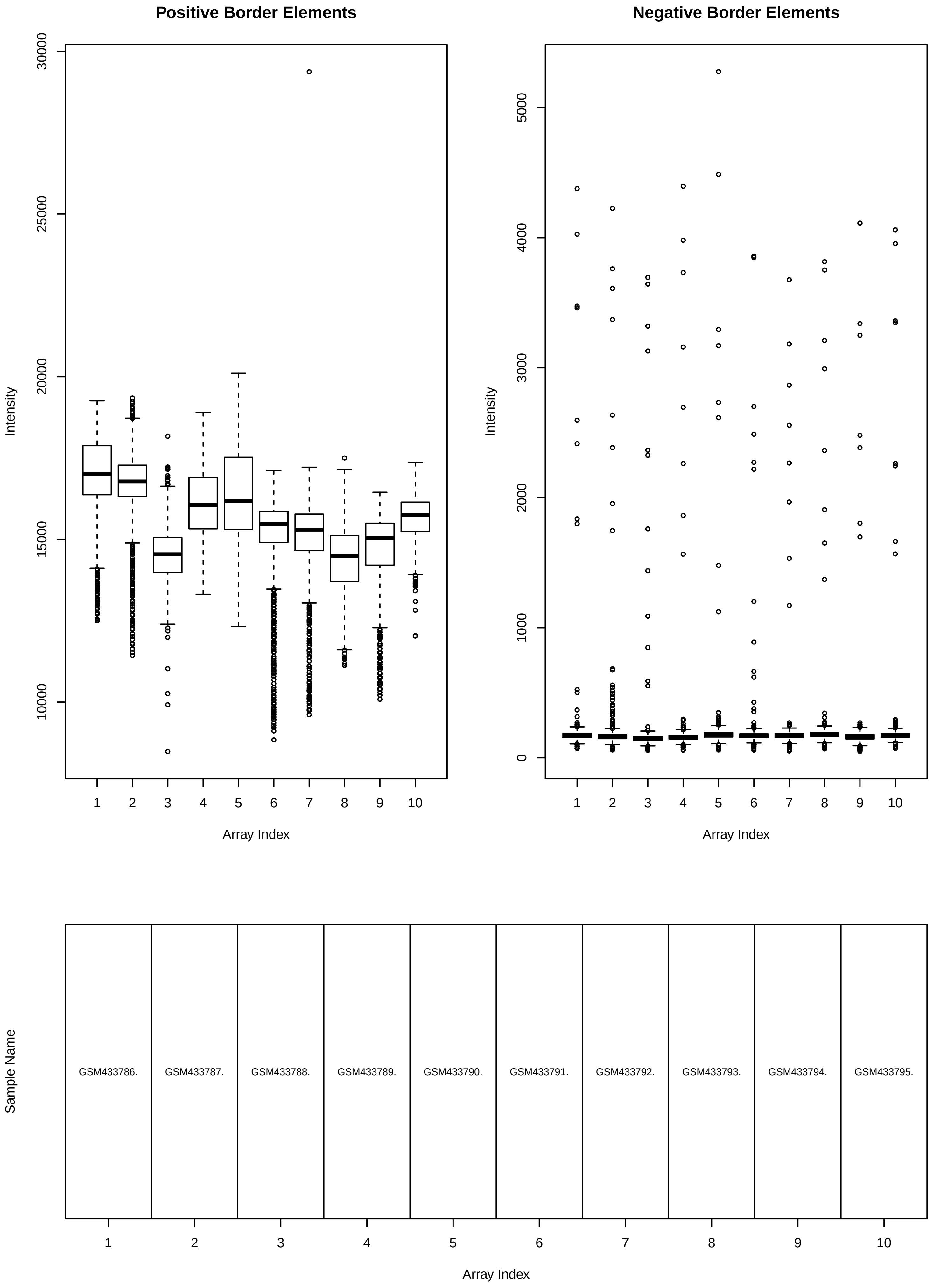
Thu Oct 1 16:12:23 2020

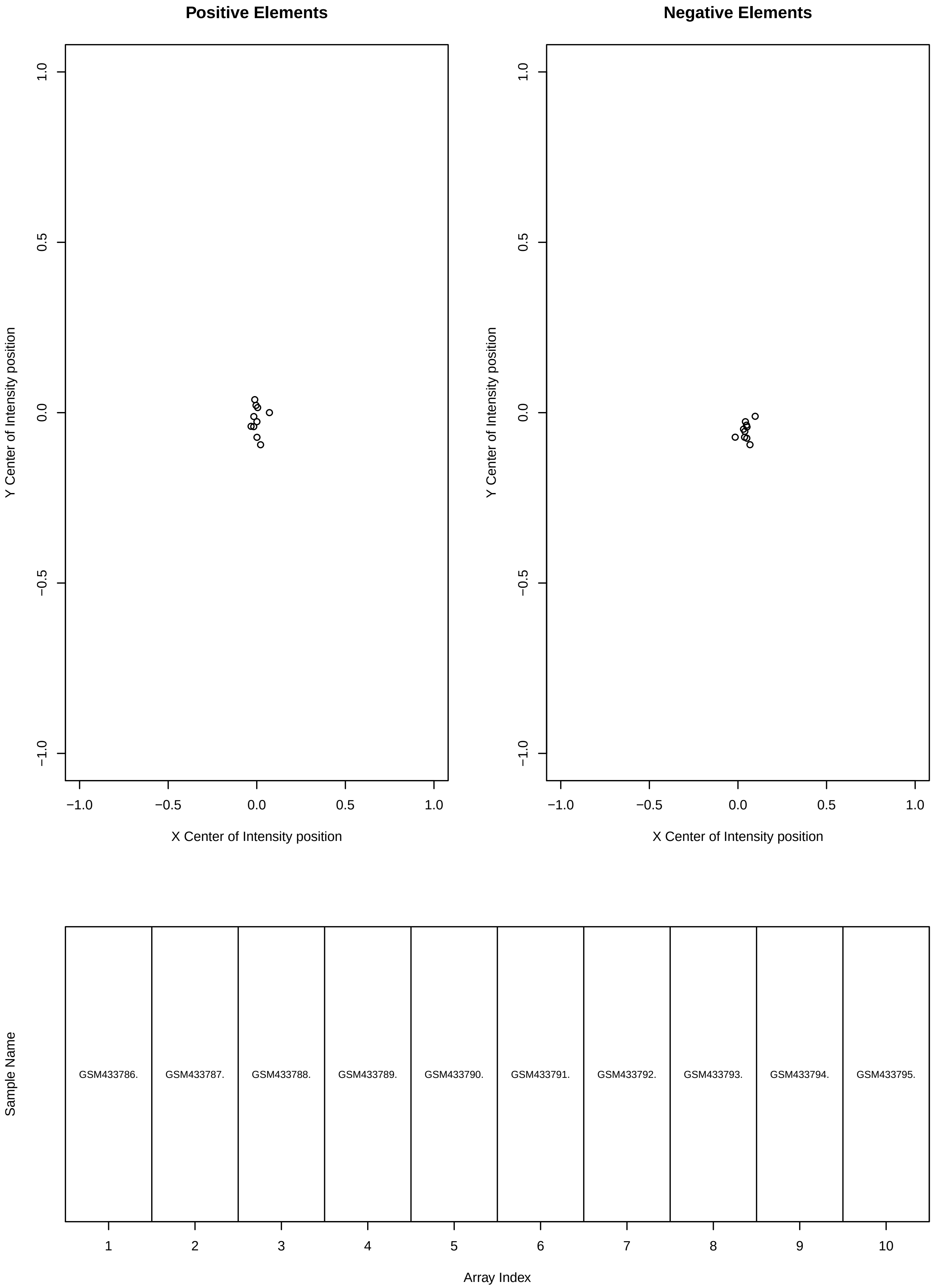
Produced by AffyQCReport R Package

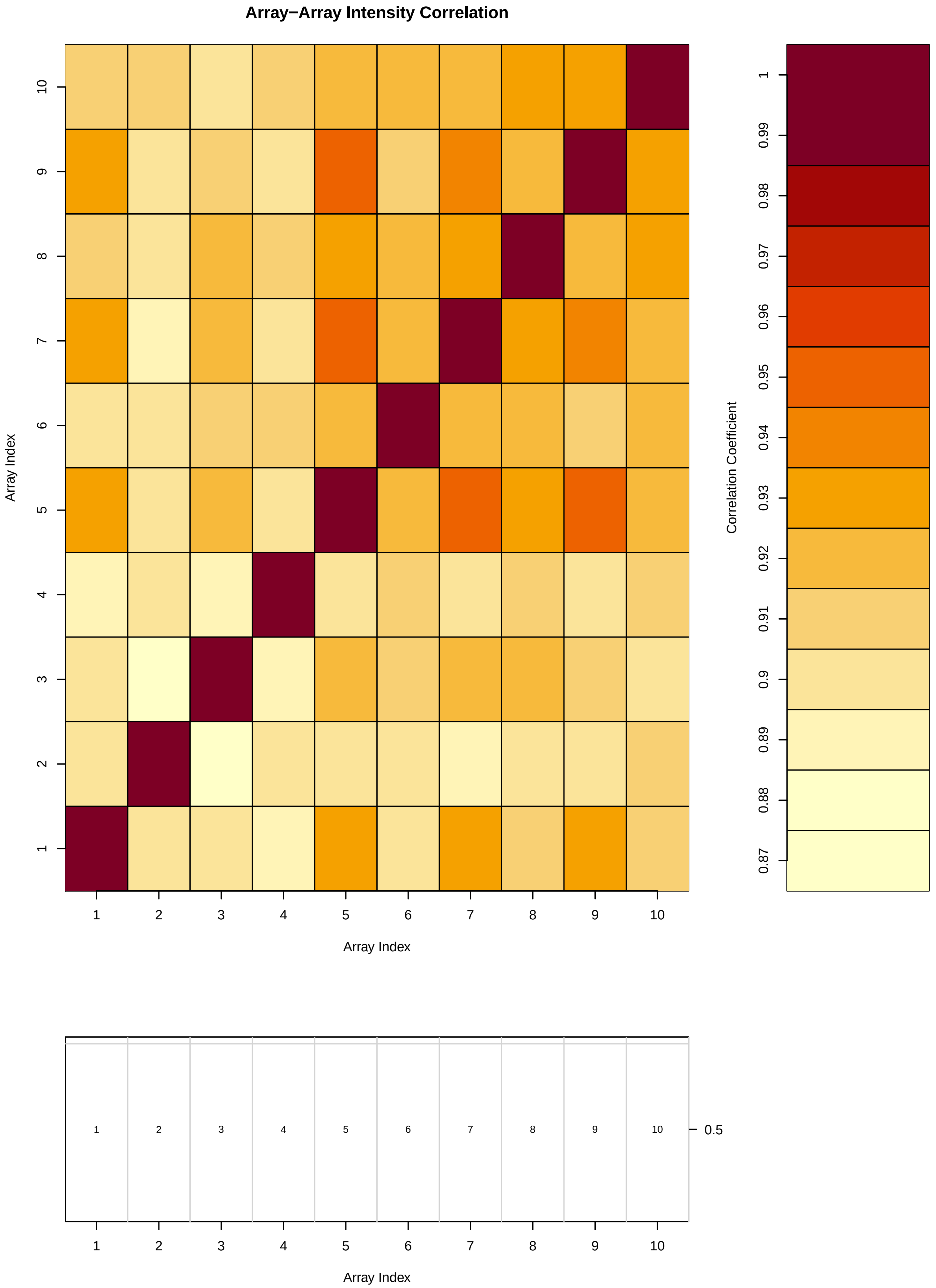

Log2(Intensity)
6
8
10
12
14
1
2
3
4
5
6
7
8
9
10
Array Index


density
0.0
0.1
0.2
0.3
0.4
0.5
6
8
10
12
14
16
log intensity
1
2
3
4
5
6
7
8
9
10


△ actin3/actin5
○ gapdh3/gapdh5
QC Stats
GSM433795.CEL
52.03%
63.52
GSM433794.CEL
50.42%
48.83
GSM433793.CEL
52.66%
76.2
GSM433792.CEL
55.09%
51.27
GSM433791.CEL
53.57%
62.64
GSM433790.CEL
52.01%
50.53
GSM433789.CEL
53.27%
51.37
GSM433788.CEL
51.4%
49.61
GSM433787.CEL
51.53%
51.79
GSM433786.CEL
50.92%
62.23
−3
−2
−1
0
1
2
3


Positive Border Elements
Intensity
30000
25000
20000
15000
10000
1
2
3
4
5
6
7
8
9
10
Array Index
Negative Border Elements
Intensity
5000
4000
3000
2000
1000
0
1
2
3
4
5
6
7
8
9
10
Array Index
Sample Name
GSM433786.
GSM433787.
GSM433788.
GSM433789.
GSM433790.
GSM433791.
GSM433792.
GSM433793.
GSM433794.
GSM433795.
1
2
3
4
5
6
7
8
9
10
Array Index


Positive Elements
Negative Elements
Y Center of Intensity position
X Center of Intensity position
−1.0
−0.5
0.0
0.5
1.0
Sample Name
GSM433786.
GSM433787.
GSM433788.
GSM433789.
GSM433790.
GSM433791.
GSM433792.
GSM433793.
GSM433794.
GSM433795.
1
2
3
4
5
6
7
8
9
10
Array Index


Array−Array Intensity Correlation
Array Index
1
2
3
4
5
6
7
8
9
10
Array Index
1
2
3
4
5
6
7
8
9
10
Correlation Coefficient
1
0.99
0.98
0.97
0.96
0.95
0.94
0.93
0.92
0.91
0.9
0.89
0.88
0.87
1
2
3
4
5
6
7
8
9
10
0.5
1
2
3
4
5
6
7
8
9
10
Array Index

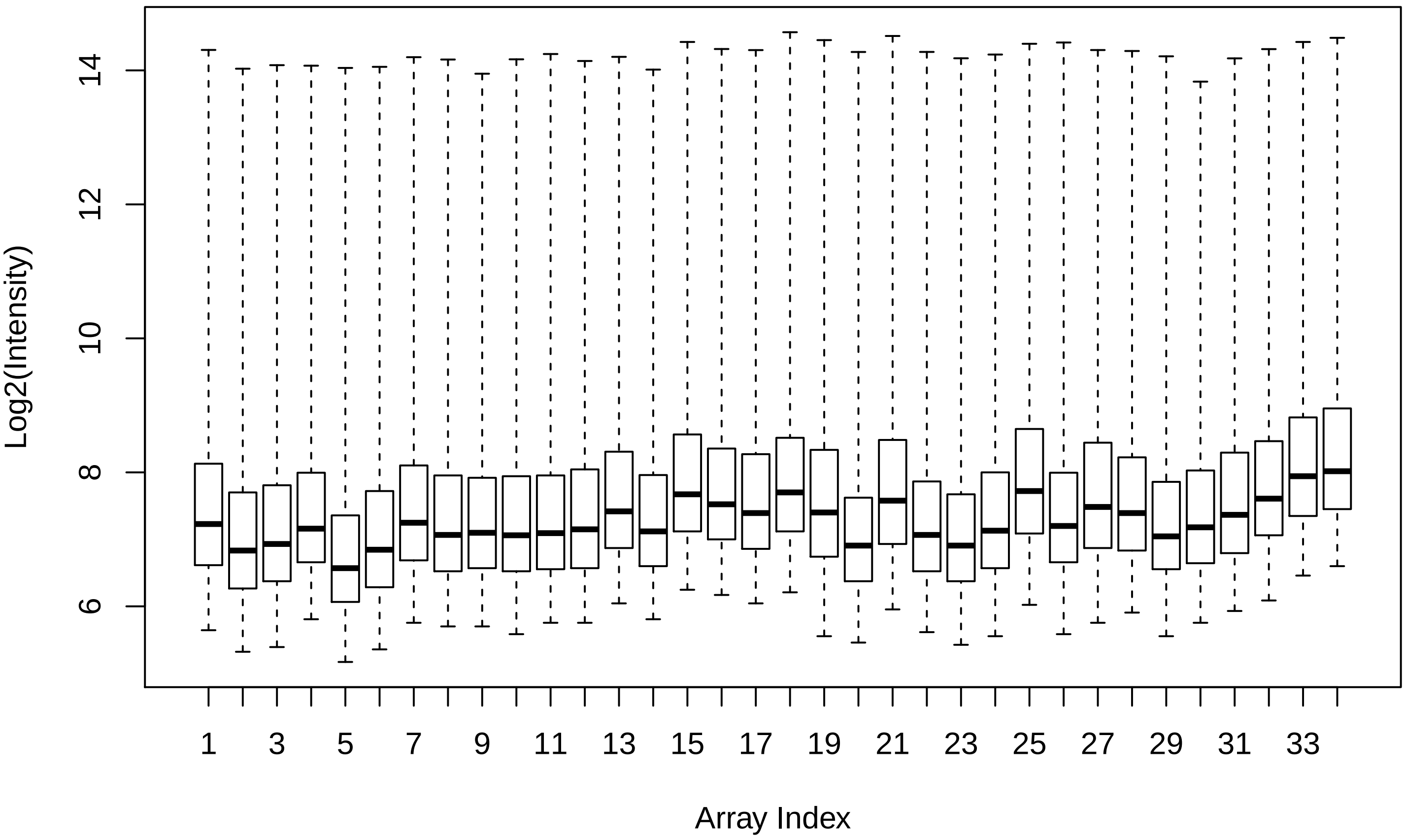
AffyBatch QC Report

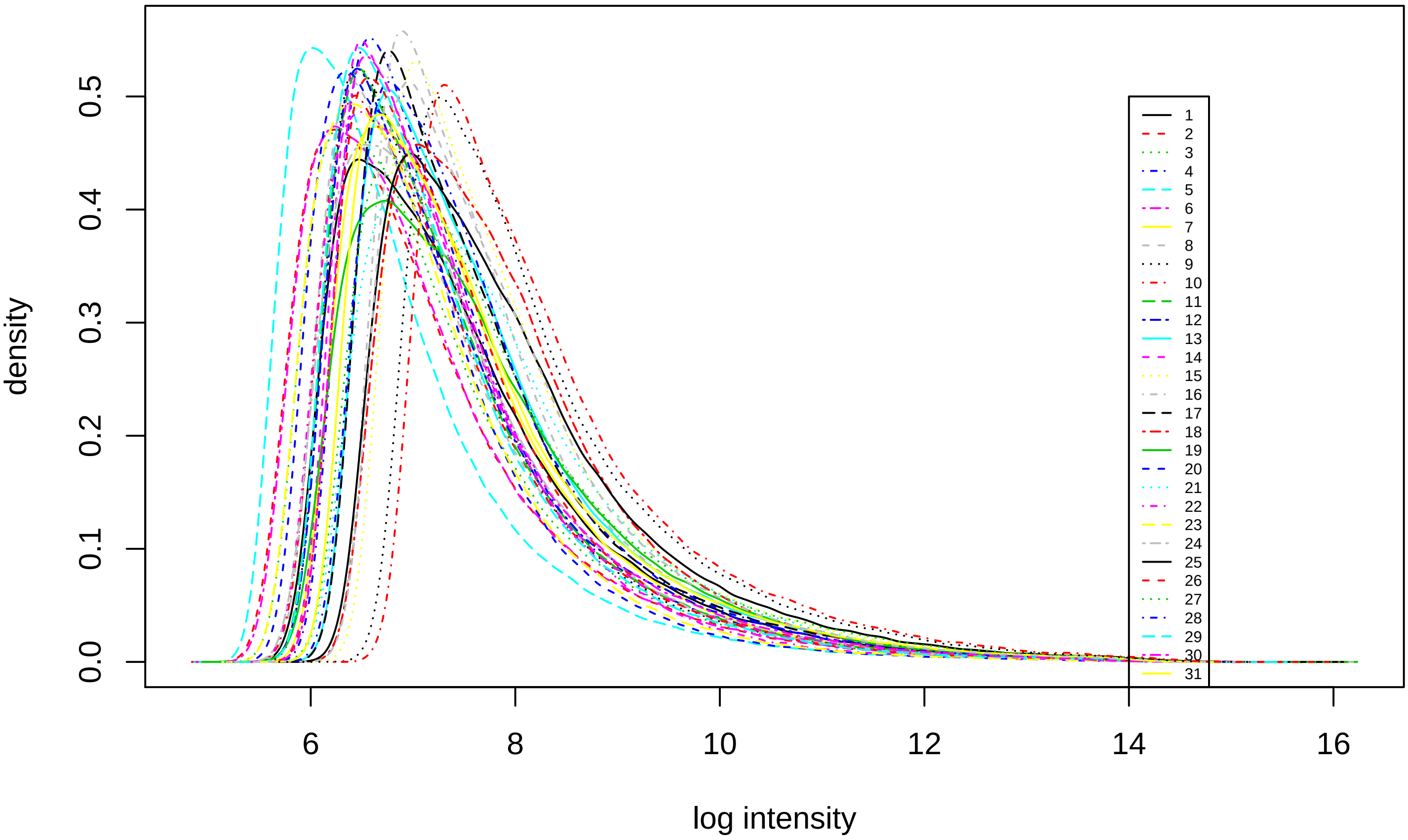
GSE20347

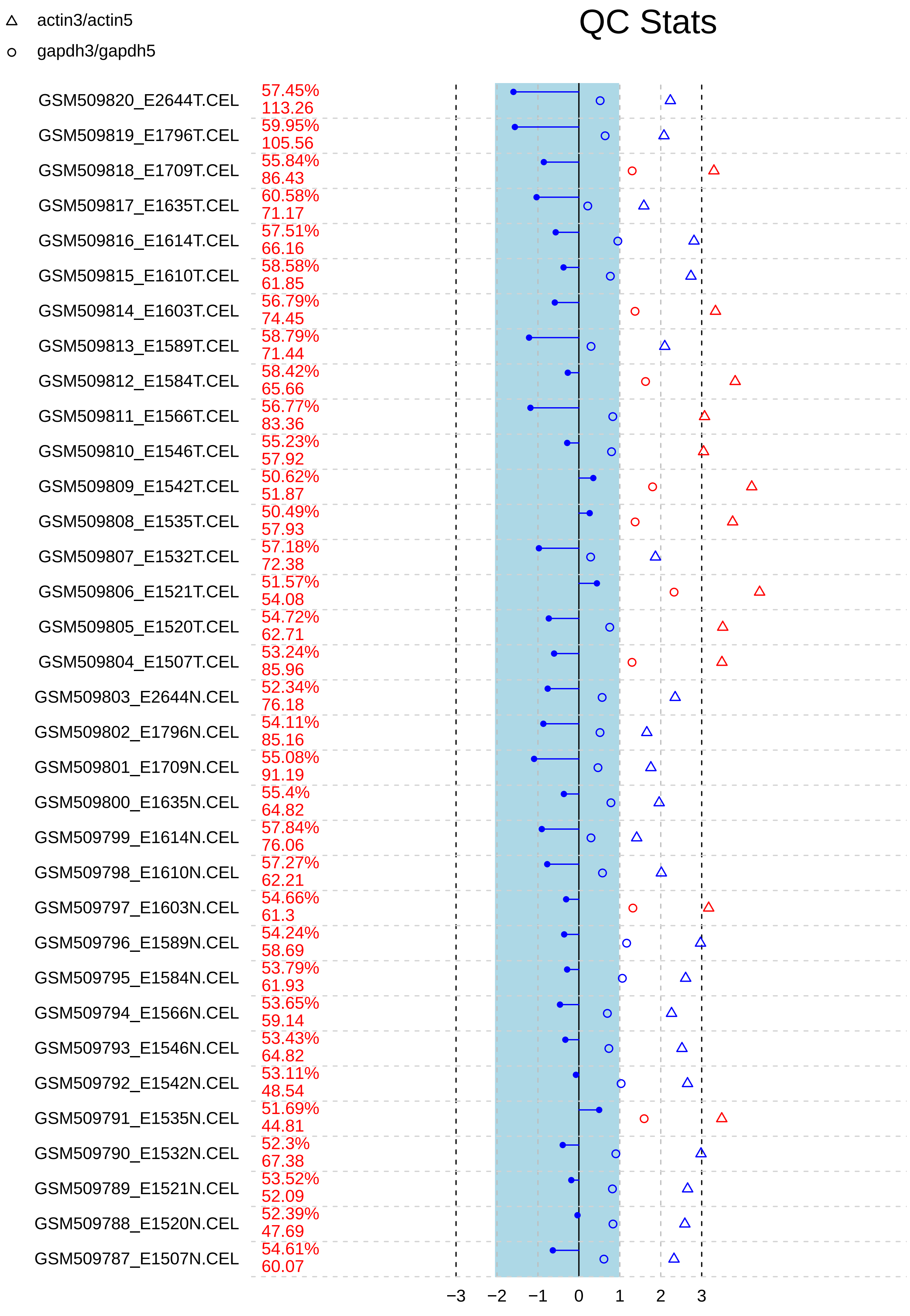
Array Index	Array Name
1	GSM509787_E1507N.CEL
2	GSM509788_E1520N.CEL
3	GSM509789_E1521N.CEL
4	GSM509790_E1532N.CEL
5	GSM509791_E1535N.CEL
6	GSM509792_E1542N.CEL
7	GSM509793_E1546N.CEL
8	GSM509794_E1566N.CEL
9	GSM509795_E1584N.CEL
10	GSM509796_E1589N.CEL
11	GSM509797_E1603N.CEL
12	GSM509798_E1610N.CEL
13	GSM509799_E1614N.CEL
14	GSM509800_E1635N.CEL
15	GSM509801_E1709N.CEL
16	GSM509802_E1796N.CEL
17	GSM509803_E2644N.CEL
18	GSM509804_E1507T.CEL
19	GSM509805_E1520T.CEL
20	GSM509806_E1521T.CEL
21	GSM509807_E1532T.CEL
22	GSM509808_E1535T.CEL
23	GSM509809_E1542T.CEL
24	GSM509810_E1546T.CEL
25	GSM509811_E1566T.CEL
26	GSM509812_E1584T.CEL
27	GSM509813_E1589T.CEL
28	GSM509814_E1603T.CEL
29	GSM509815_E1610T.CEL
30	GSM509816_E1614T.CEL
31	GSM509817_E1635T.CEL
32	GSM509818_E1709T.CEL
33	GSM509819_E1796T.CEL
34	GSM509820_E2644T.CEL

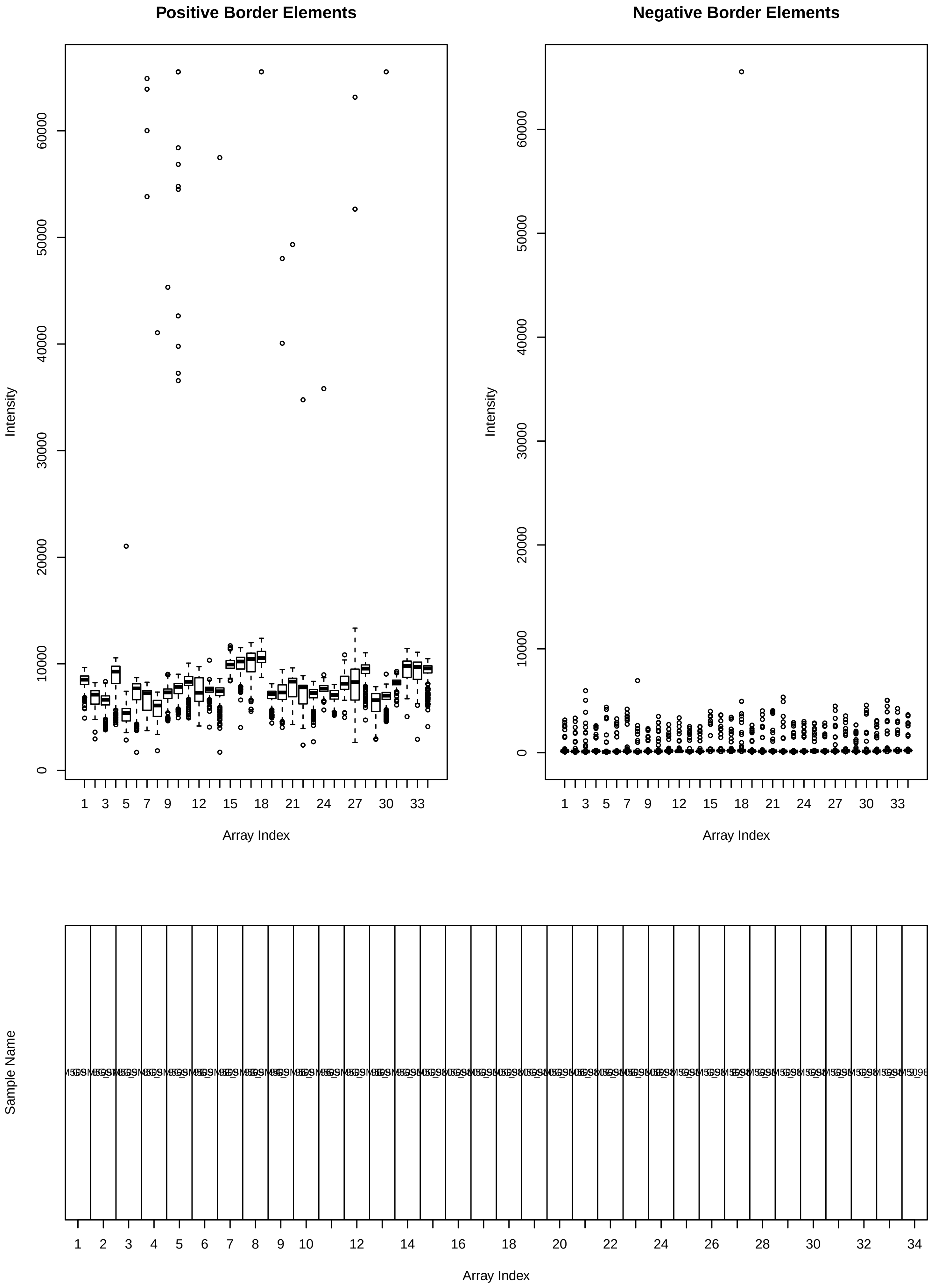
Thu Oct 1 16 24:55 2020

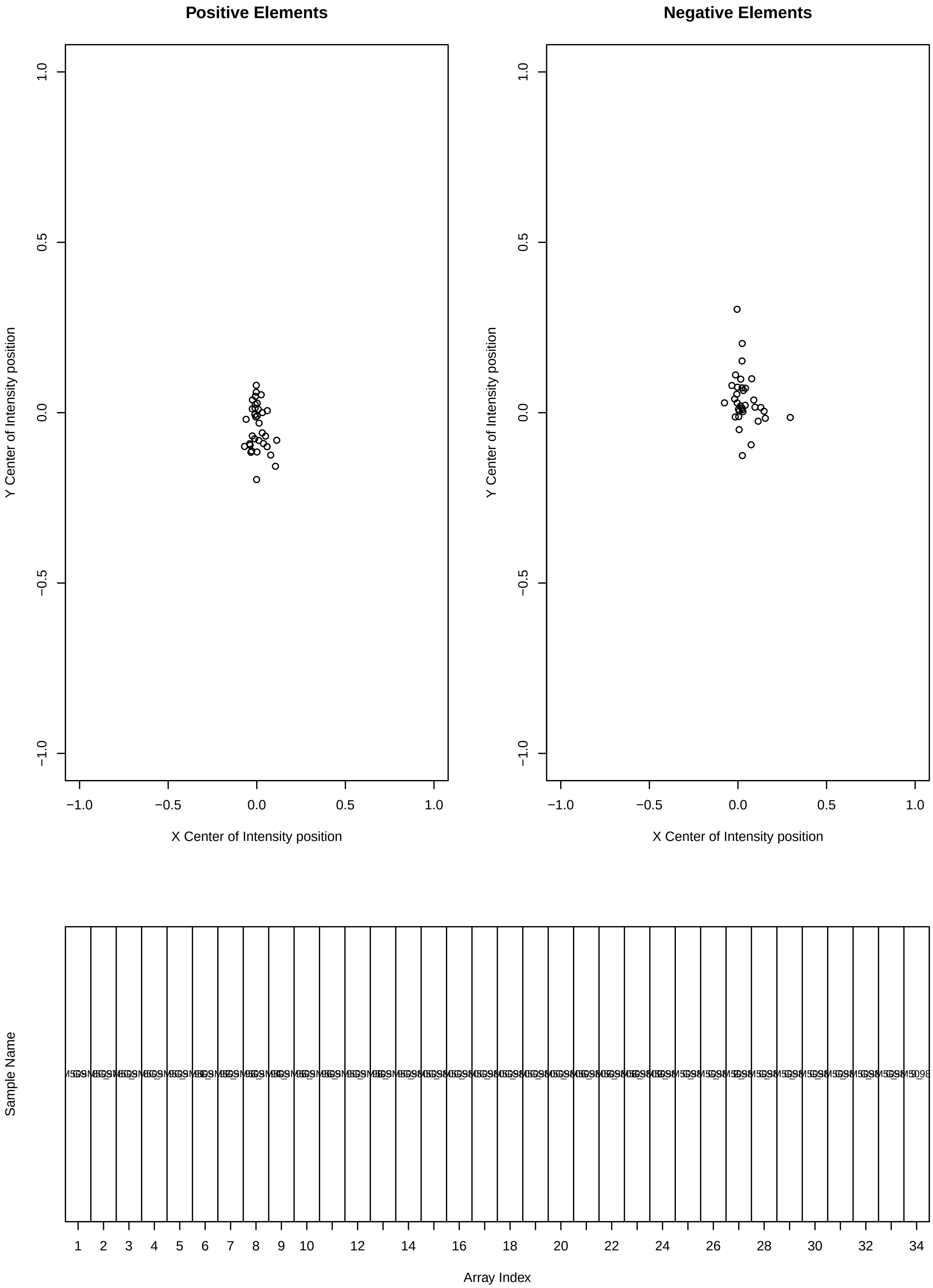
Produced by AffyQCReport R Package

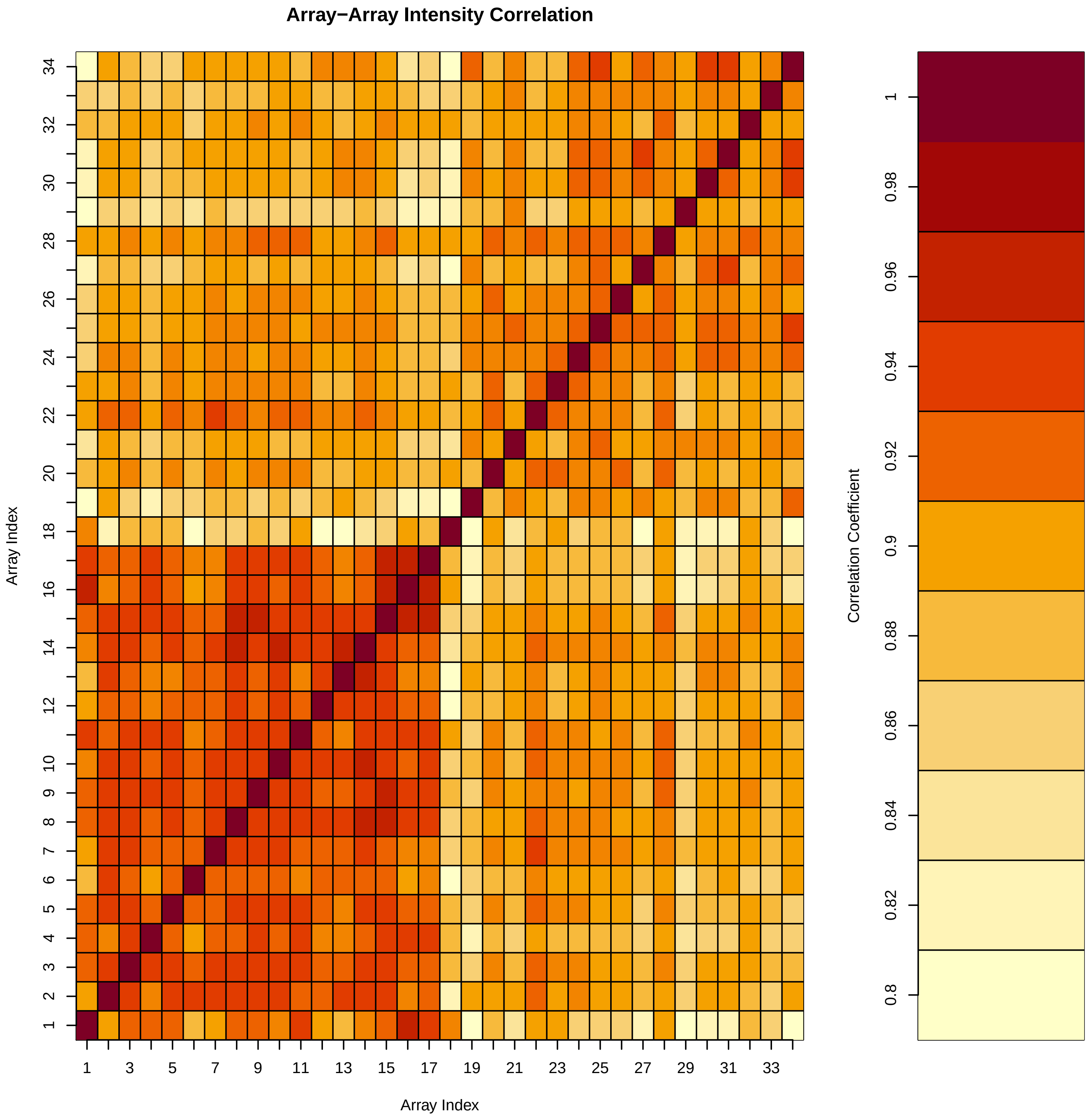

Log2(Intensity)
6
8
10
12
14
1
3
5
7
9
11
13
15
17
19
21
23
25
27
29
31
33
Array Index

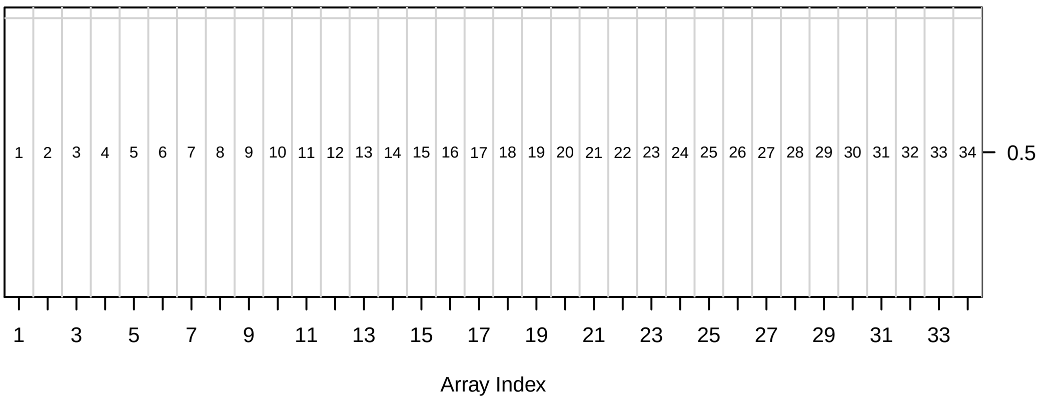

density
0.0
0.1
0.2
0.3
0.4
0.5
6
8
10
12
14
16
log intensity
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31


QC Stats
△ actin3/actin5
○ gapdh3/gapdh5
GSM509820_E2644T.CEL
57.45%
113.26
GSM509819_E1796T.CEL
59.95%
105.56
GSM509818_E1709T.CEL
55.84%
86.43
GSM509817_E1635T.CEL
60.58%
71.17
GSM509816_E1614T.CEL
57.51%
66.16
GSM509815_E1610T.CEL
58.58%
61.85
GSM509814_E1603T.CEL
56.79%
74.45
GSM509813_E1589T.CEL
58.79%
71.44
GSM509812_E1584T.CEL
58.42%
65.66
GSM509811_E1566T.CEL
56.77%
83.36
GSM509810_E1546T.CEL
55.23%
57.92
GSM509809_E1542T.CEL
50.62%
51.87
GSM509808_E1535T.CEL
50.49%
57.93
GSM509807_E1532T.CEL
57.18%
72.38
GSM509806_E1521T.CEL
51.57%
54.08
GSM509805_E1520T.CEL
54.72%
62.71
GSM509804_E1507T.CEL
53.24%
85.96
GSM509803_E2644N.CEL
52.34%
76.18
GSM509802_E1796N.CEL
54.11%
85.16
GSM509801_E1709N.CEL
55.08%
91.19
GSM509800_E1635N.CEL
55.4%
64.82
GSM509799_E1614N.CEL
57.84%
76.06
GSM509798_E1610N.CEL
57.27%
62.21
GSM509797_E1603N.CEL
54.66%
61.3
GSM509796_E1589N.CEL
54.24%
58.69
GSM509795_E1584N.CEL
53.79%
61.93
GSM509794_E1566N.CEL
53.65%
59.14
GSM509793_E1546N.CEL
53.43%
64.82
GSM509792_E1542N.CEL
53.11%
48.54
GSM509791_E1535N.CEL
51.69%
44.81
GSM509790_E1532N.CEL
52.3%
67.38
GSM509789_E1521N.CEL
53.52%
52.09
GSM509788_E1520N.CEL
52.39%
47.69
GSM509787_E1507N.CEL
54.61%
60.07
−3
−2
−1
0
1
2
3


Positive Border Elements
Negative Border Elements
Intensity
0
10000
20000
30000
40000
50000
60000
1
3
5
7
9
12
15
18
21
24
27
30
33
Array Index
Sample Name
1
2
3
4
5
6
7
8
9
10
12
14
16
18
20
22
24
26
28
30
32
34


Positive Elements
Negative Elements
Y Center of Intensity position
X Center of Intensity position
1.0
0.5
0.0
−0.5
−1.0
Sample Name
Array Index
1
2
3
4
5
6
7
8
9
10
12
14
16
18
20
22
24
26
28
30
32
34


Array−Array Intensity Correlation
Array Index
1
3
5
7
9
11
13
15
17
19
21
23
25
27
29
31
33
Array Index
1
2
3
4
5
6
7
8
9
10
12
14
16
18
20
22
24
26
28
30
32
34
Correlation Coefficient
1
0.98
0.96
0.94
0.92
0.9
0.88
0.86
0.84
0.82
0.8


1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34
0.5
1
3
5
7
9
11
13
15
17
19
21
23
25
27
29
31
33
Array Index

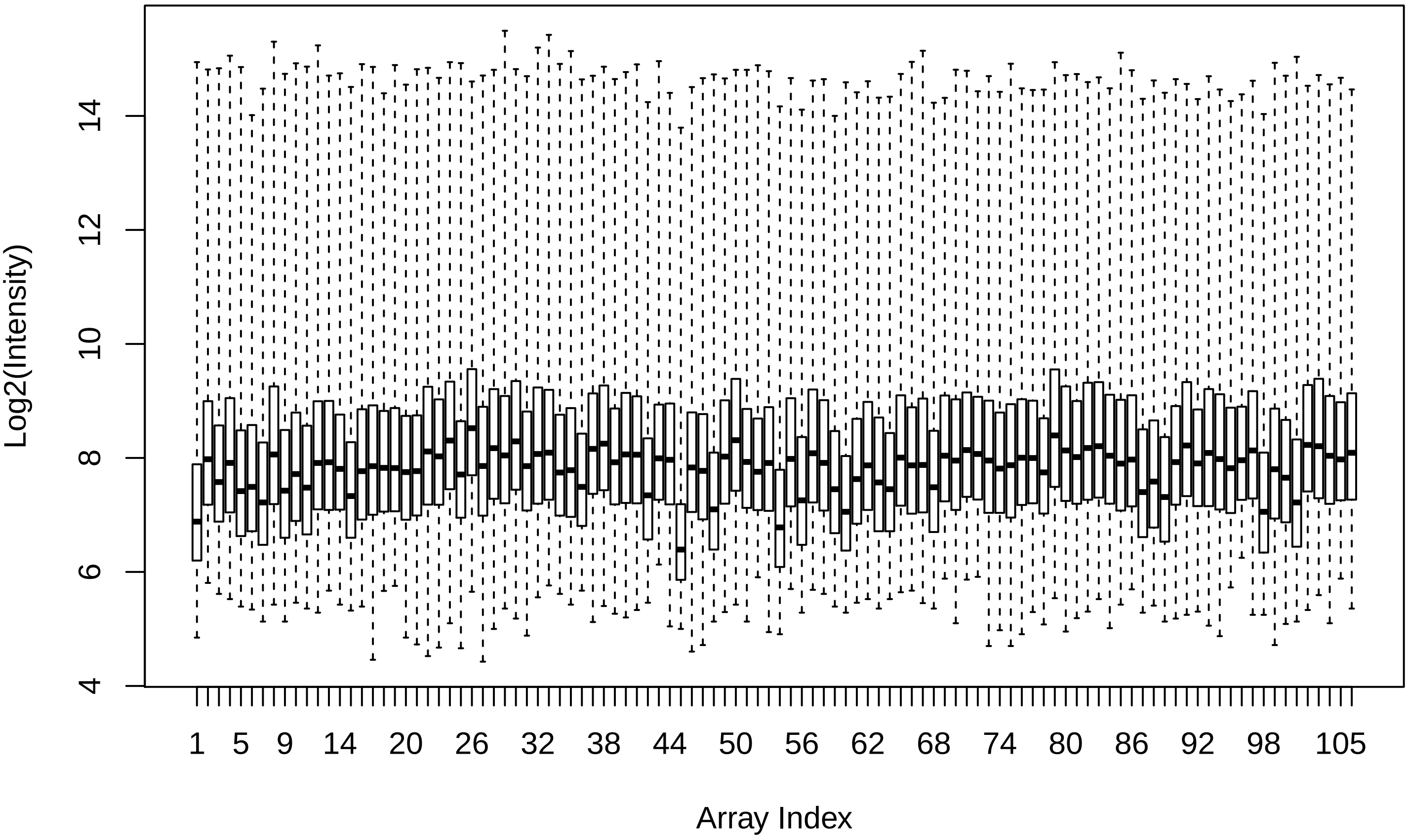
AffyBatch QC Report

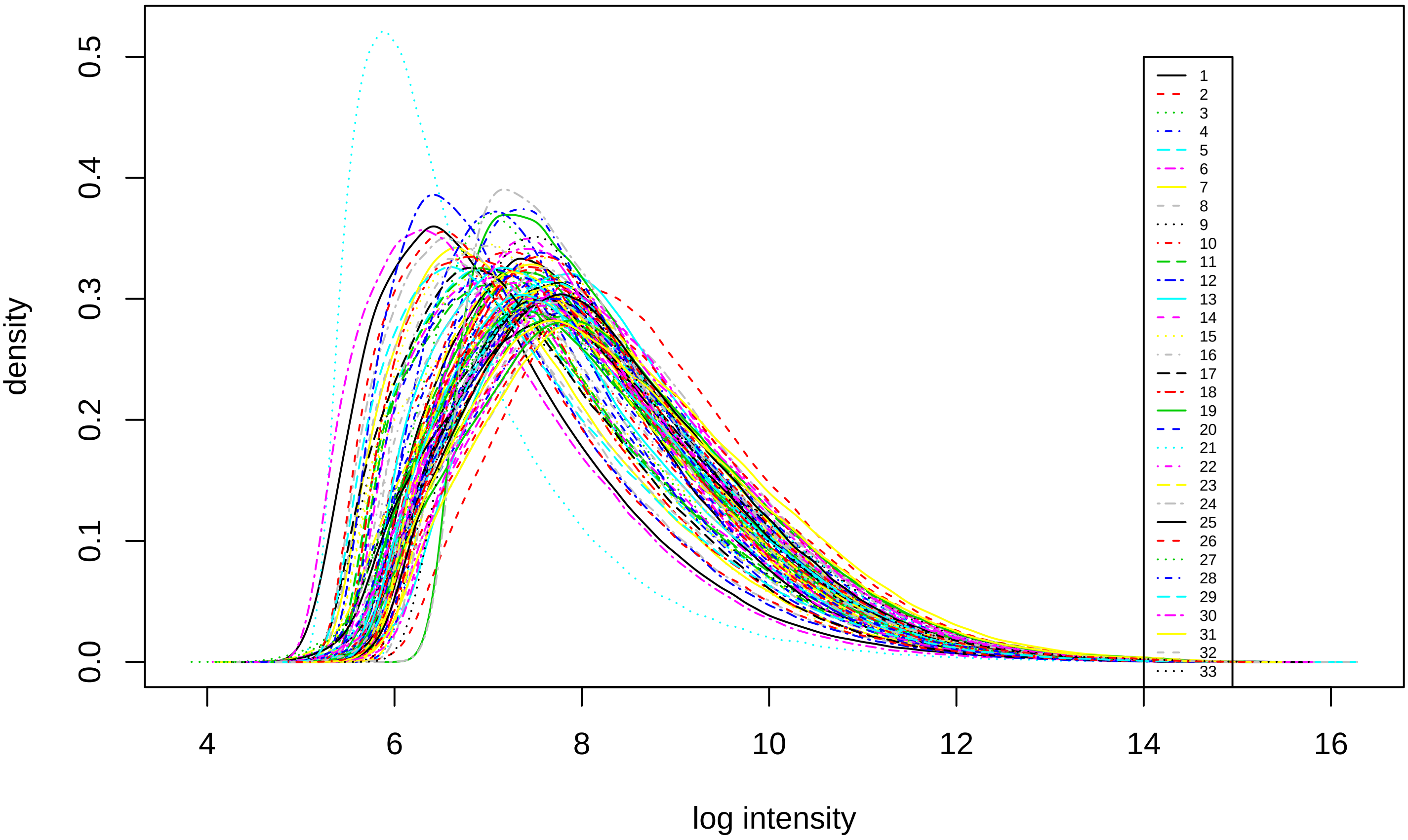
GSE23400A

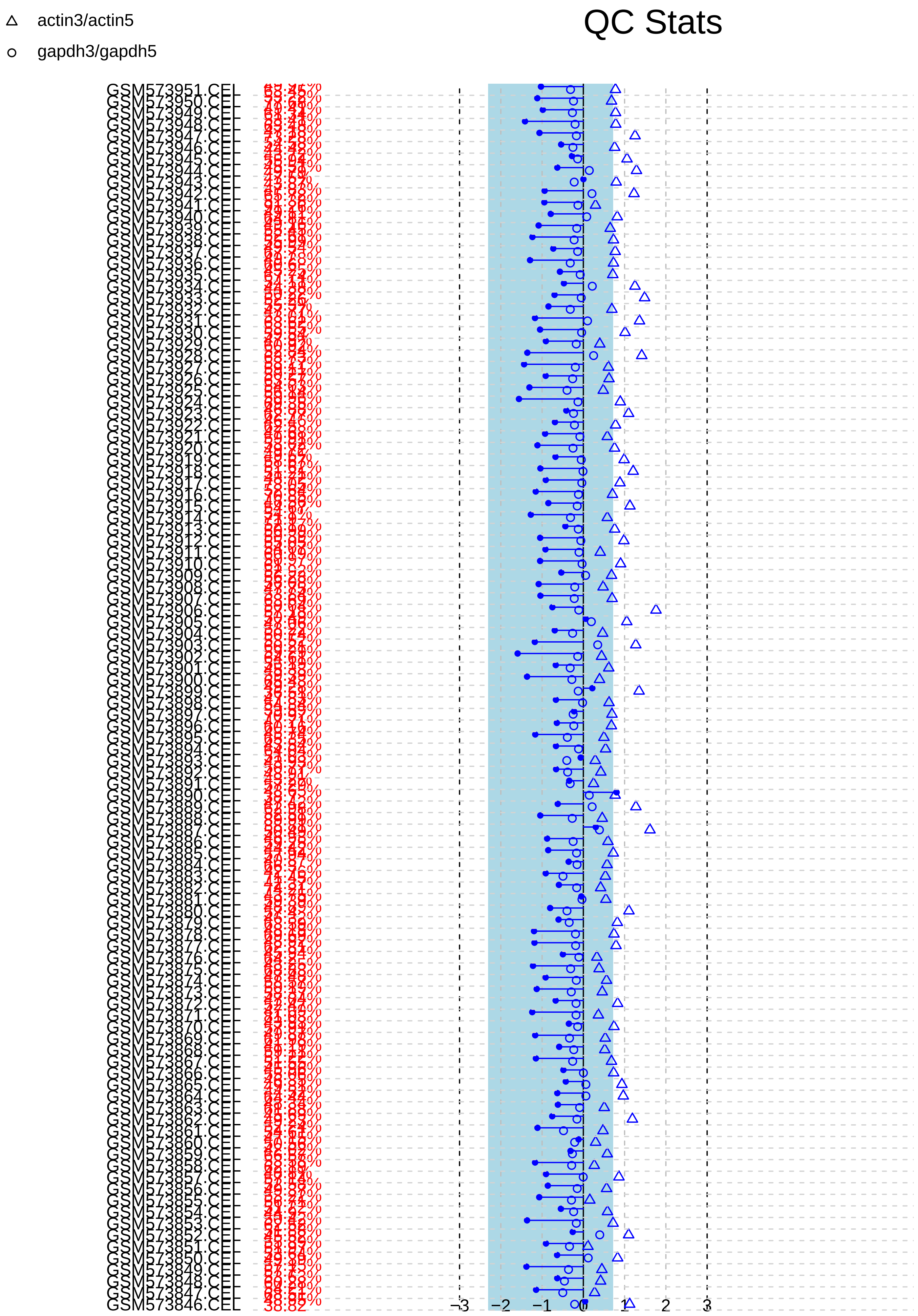
Array Index	Array Name

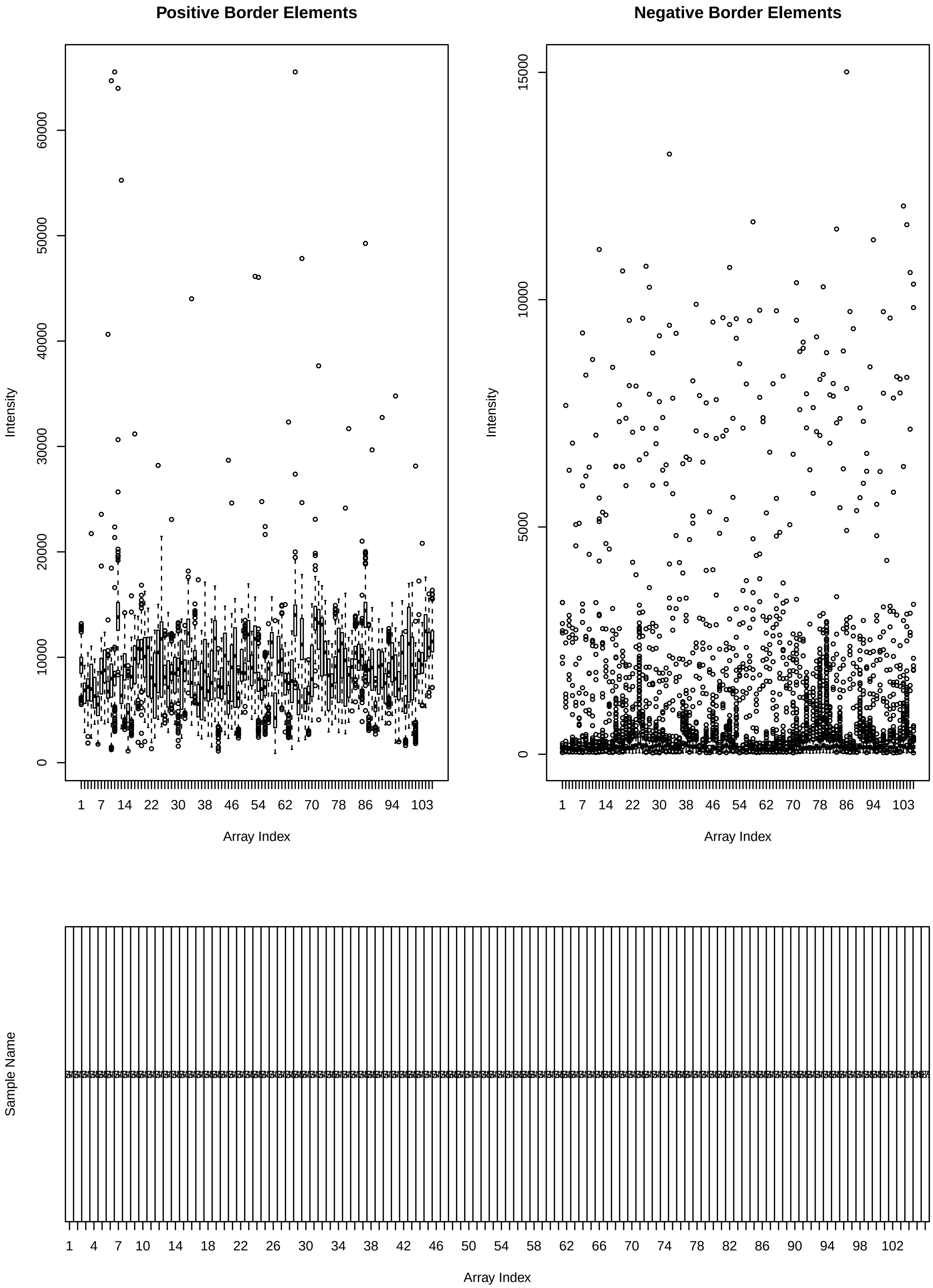
Sun Oct 11 12:49:48 2020

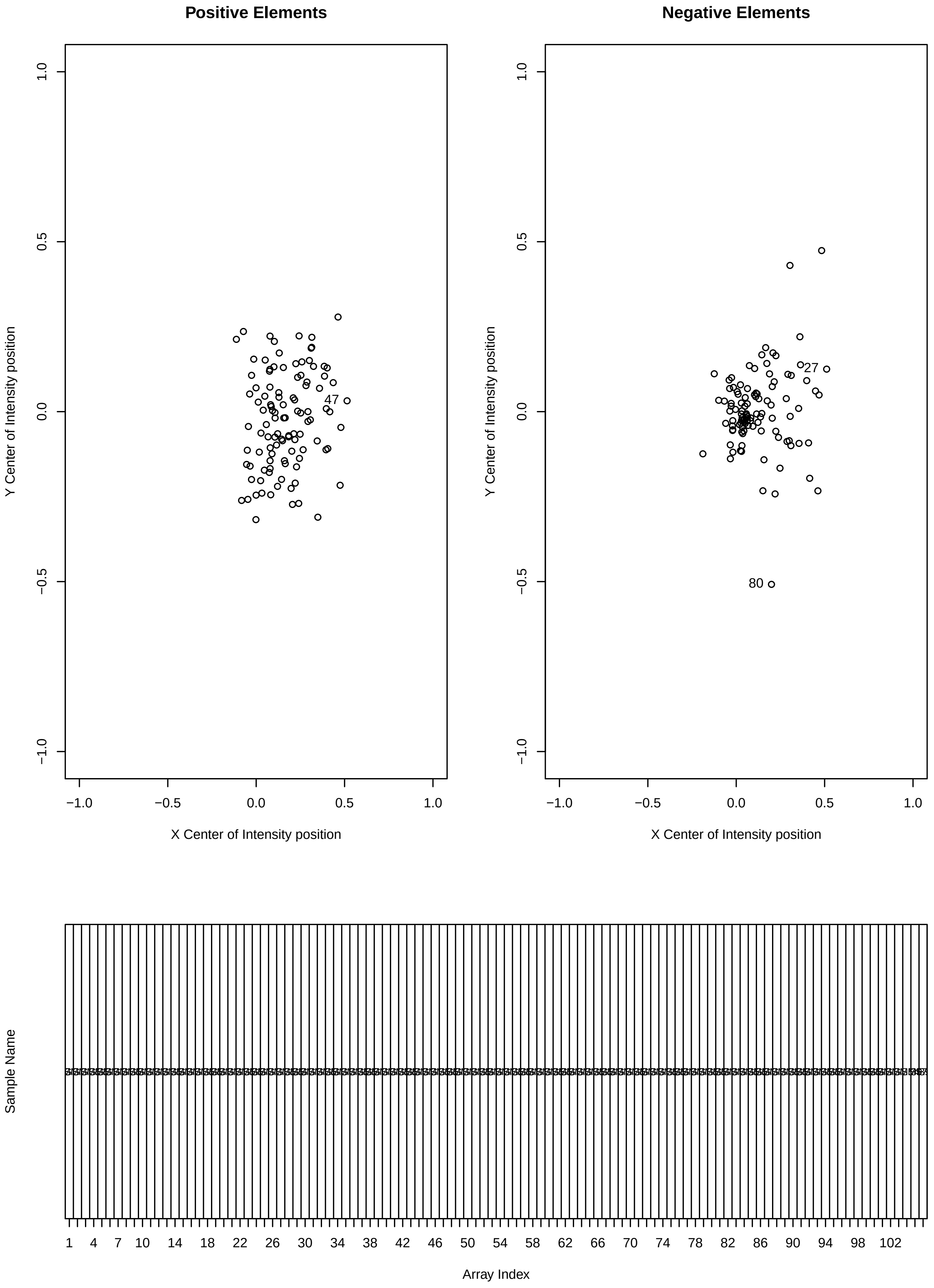
Produced by AffyQCReport R Package

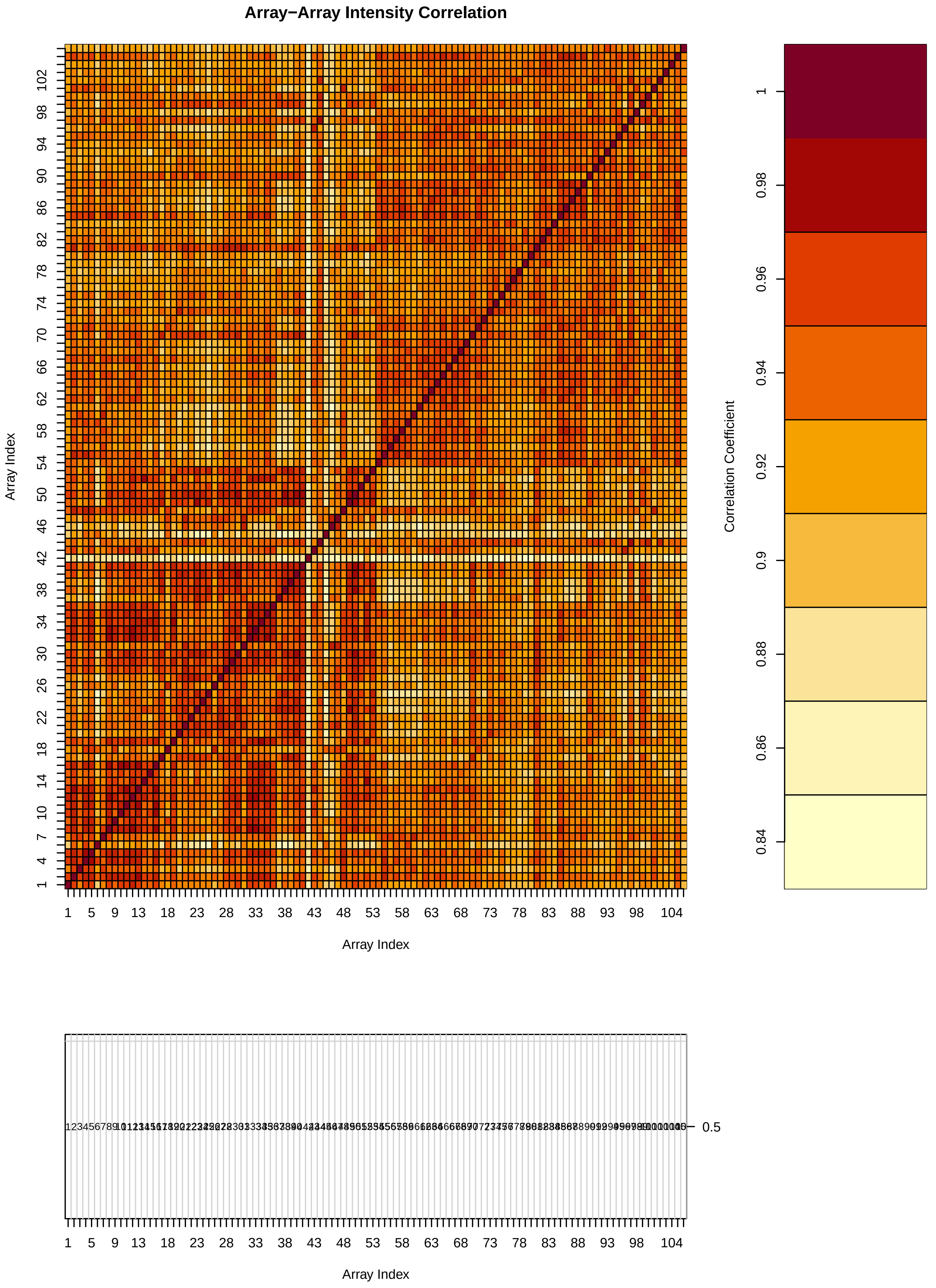

Log2(Intensity)
4
6
8
10
12
14
1
5
9
14
20
26
32
38
44
50
56
62
68
74
80
86
92
98
105
Array Index


density
0.0
0.1
0.2
0.3
0.4
0.5
4
6
8
10
12
14
16
log intensity
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33


QC Stats
actin3/actin5
gapdh3/gapdh5
GSM573951.CEL
GSM573950.CEL
GSM573949.CEL
GSM573948.CEL
GSM573947.CEL
GSM573946.CEL
GSM573945.CEL
GSM573944.CEL
GSM573943.CEL
GSM573942.CEL
GSM573941.CEL
GSM573940.CEL
GSM573939.CEL
GSM573938.CEL
GSM573937.CEL
GSM573936.CEL
GSM573935.CEL
GSM573934.CEL
GSM573933.CEL
GSM573932.CEL
GSM573931.CEL
GSM573930.CEL
GSM573929.CEL
GSM573928.CEL
GSM573927.CEL
GSM573926.CEL
GSM573925.CEL
GSM573924.CEL
GSM573923.CEL
GSM573922.CEL
GSM573921.CEL
GSM573920.CEL
GSM573919.CEL
GSM573918.CEL
GSM573917.CEL
GSM573916.CEL
GSM573915.CEL
GSM573914.CEL
GSM573913.CEL
GSM573912.CEL
GSM573911.CEL
GSM573910.CEL
GSM573909.CEL
GSM573908.CEL
GSM573907.CEL
GSM573906.CEL
GSM573905.CEL
GSM573904.CEL
GSM573903.CEL
GSM573902.CEL
GSM573901.CEL
GSM573900.CEL
GSM573899.CEL
GSM573898.CEL
GSM573897.CEL
GSM573896.CEL
GSM573895.CEL
GSM573894.CEL
GSM573893.CEL
GSM573892.CEL
GSM573891.CEL
GSM573890.CEL
GSM573889.CEL
GSM573888.CEL
GSM573887.CEL
GSM573886.CEL
GSM573885.CEL
GSM573884.CEL
GSM573883.CEL
GSM573882.CEL
GSM573881.CEL
GSM573880.CEL
GSM573879.CEL
GSM573878.CEL
GSM573877.CEL
GSM573876.CEL
GSM573875.CEL
GSM573874.CEL
GSM573873.CEL
GSM573872.CEL
GSM573871.CEL
GSM573870.CEL
GSM573869.CEL
GSM573868.CEL
GSM573867.CEL
GSM573866.CEL
GSM573865.CEL
GSM573864.CEL
GSM573863.CEL
GSM573862.CEL
GSM573861.CEL
GSM573860.CEL
GSM573859.CEL
GSM573858.CEL
GSM573857.CEL
GSM573856.CEL
GSM573855.CEL
GSM573854.CEL
GSM573853.CEL
GSM573852.CEL
GSM573851.CEL
GSM573850.CEL
GSM573849.CEL
GSM573848.CEL
GSM573847.CEL
GSM573846.CEL
38.82%
−3
−2
−1
0
1
2
3


Positive Border Elements
Intensity
0
10000
20000
30000
40000
50000
60000
1
7
14
22
30
38
46
54
62
70
78
86
94
103
Array Index
Negative Border Elements
Intensity
0
5000
10000
15000
1
7
14
22
30
38
46
54
62
70
78
86
94
103
Array Index
Sample Name
1
4
7
10
14
18
22
26
30
34
38
42
46
50
54
58
62
66
70
74
78
82
86
90
94
98
102
Array Index


Positive Elements
Negative Elements
Y Center of Intensity position
1.0
0.5
0.0
−0.5
−1.0
X Center of Intensity position
−1.0
−0.5
0.0
0.5
1.0
47
27
80
Sample Name
Array Index
1
4
7
10
14
18
22
26
30
34
38
42
46
50
54
58
62
66
70
74
78
82
86
90
94
98
102


Array−Array Intensity Correlation
Array Index
1
4
7
10
14
18
22
26
30
34
38
42
46
50
54
58
62
66
70
74
78
82
86
90
94
98
102
1
5
9
13
18
23
28
33
38
43
48
53
58
63
68
73
78
83
88
93
98
104
Array Index
Correlation Coefficient
1
0.98
0.96
0.94
0.92
0.9
0.88
0.86
0.84
0.5
1
5
9
13
18
23
28
33
38
43
48
53
58
63
68
73
78
83
88
93
98
104
Array Index

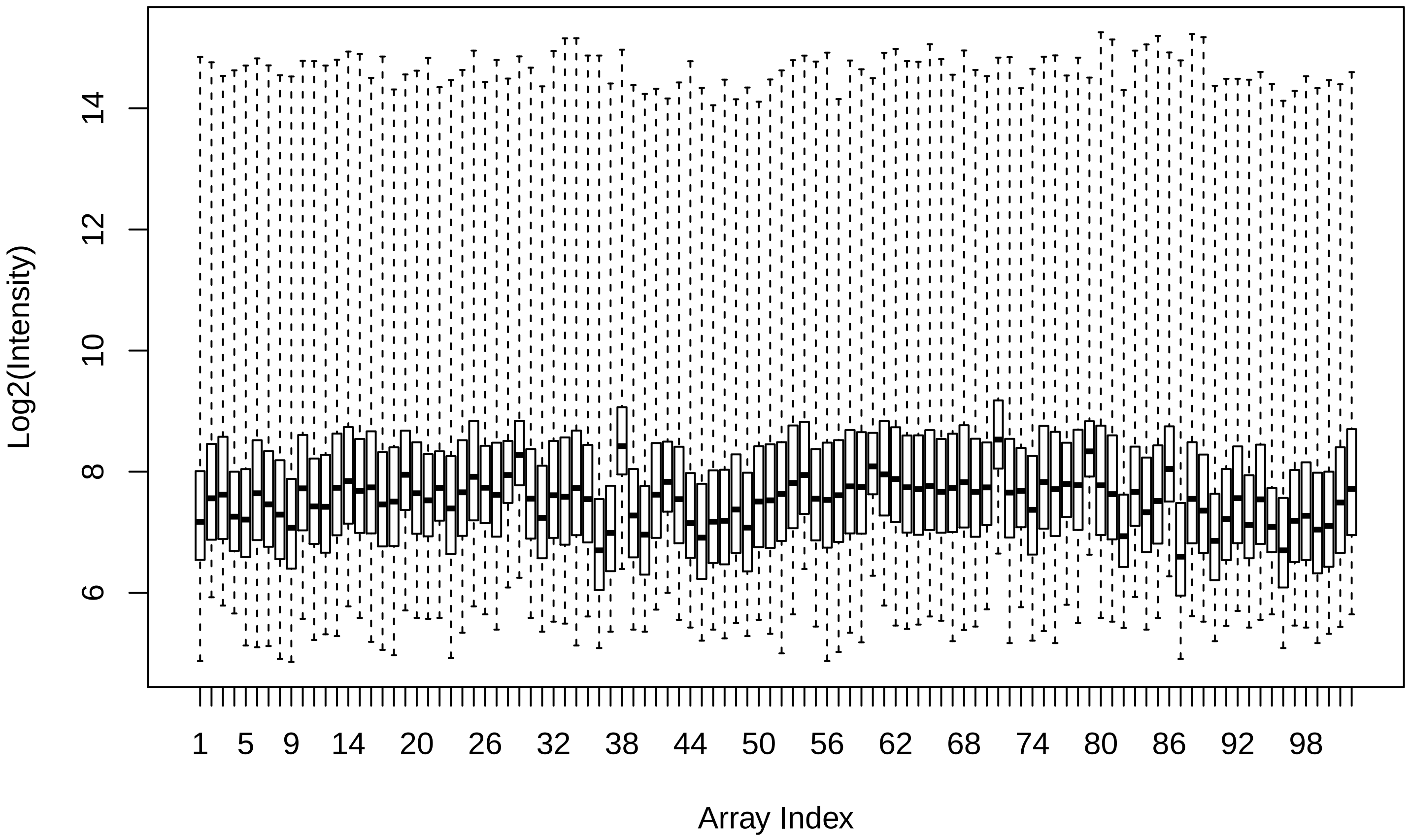
AffyBatch QC Report

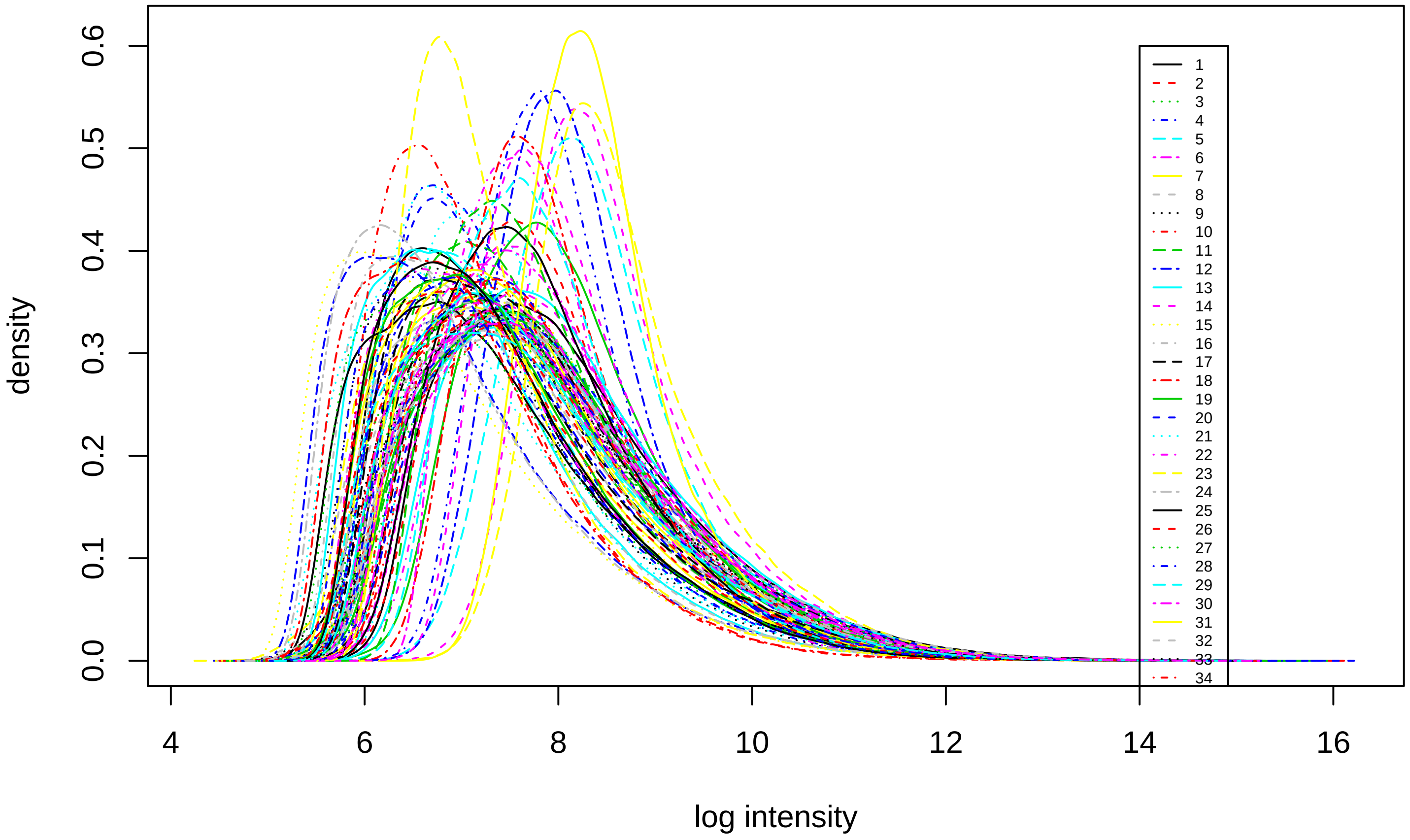
GSE23400B

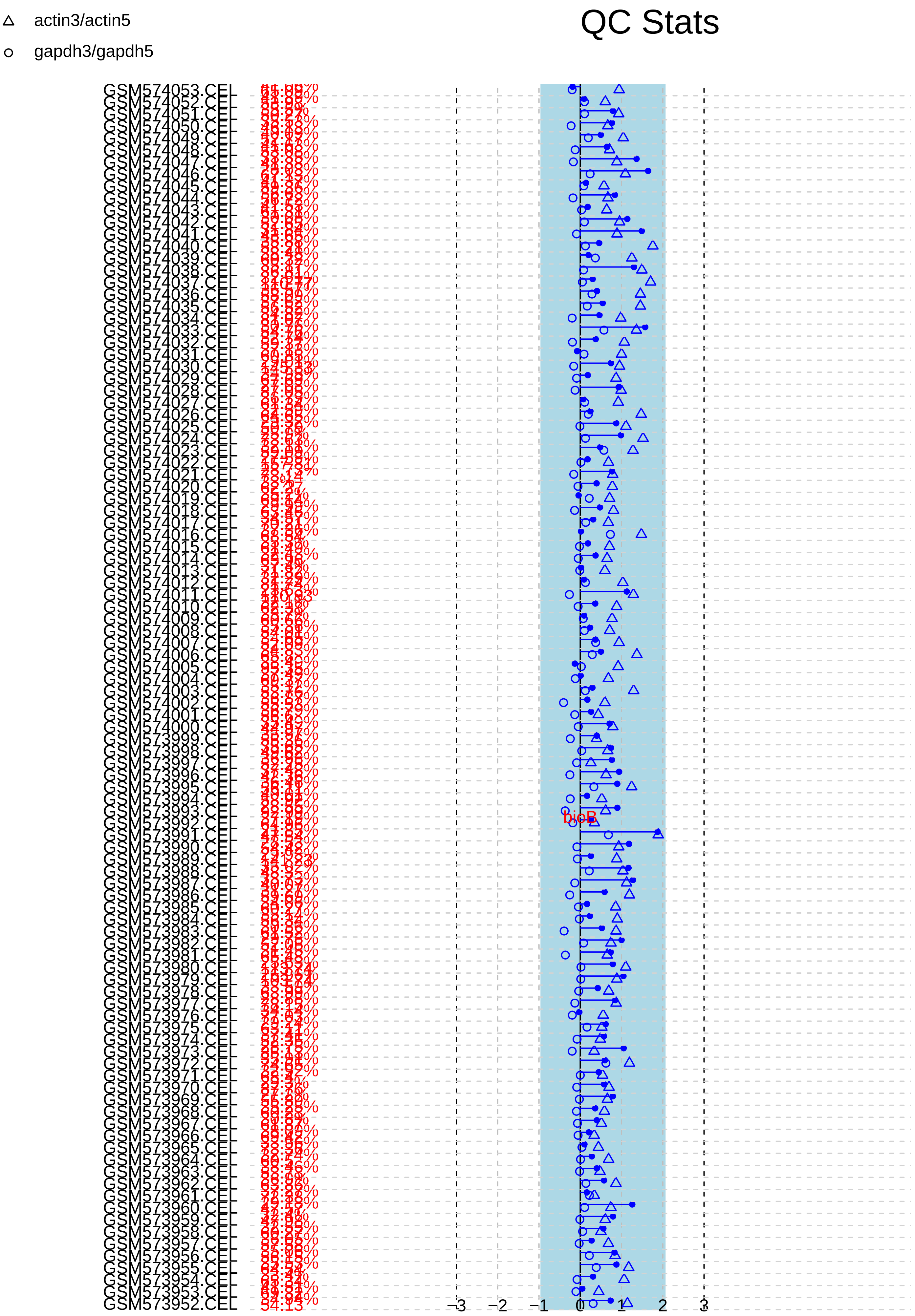
Array Index	Array Name

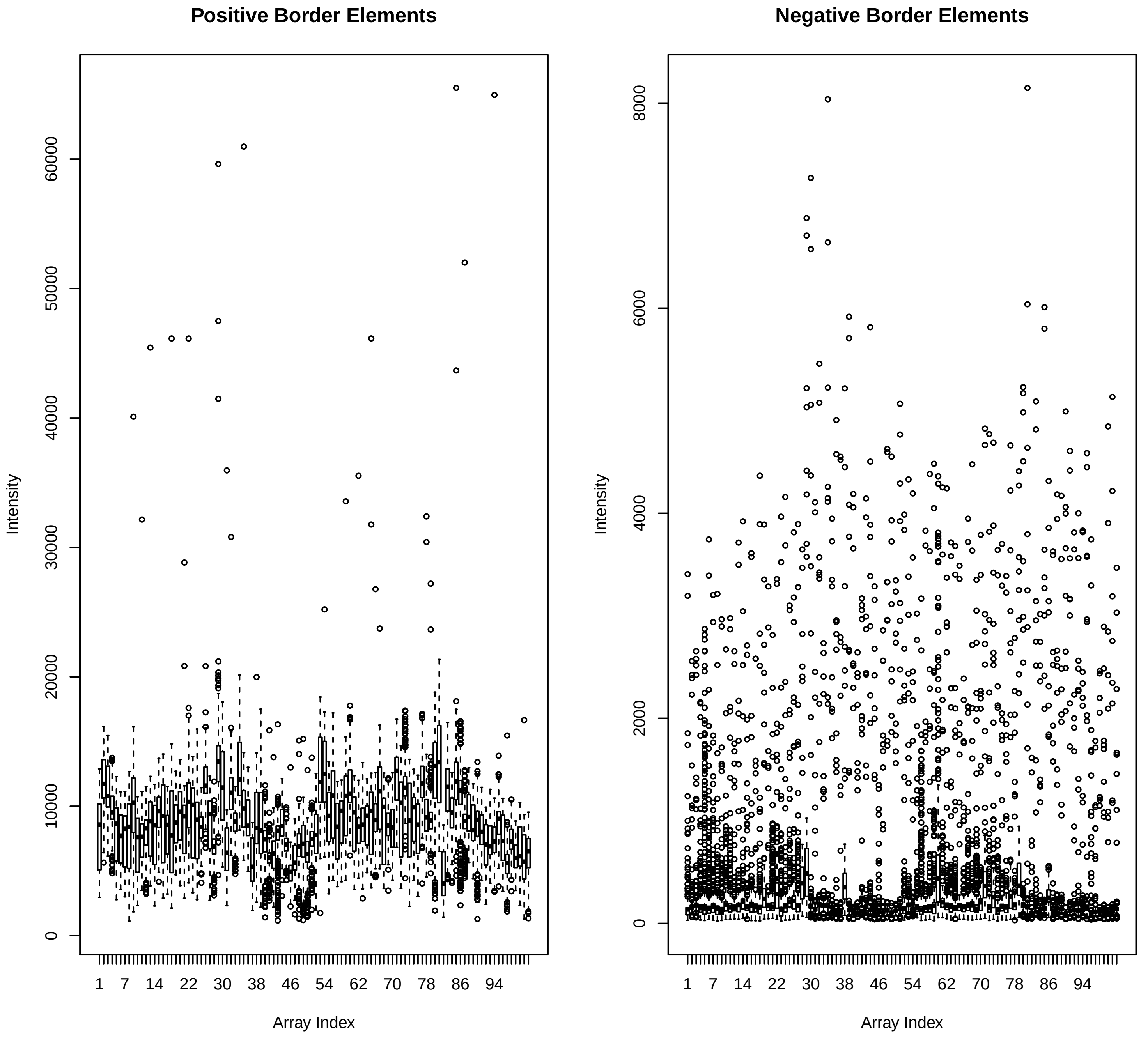
Sun Oct 11 13:11:16 2020

Produced by AffyQCReport R Package

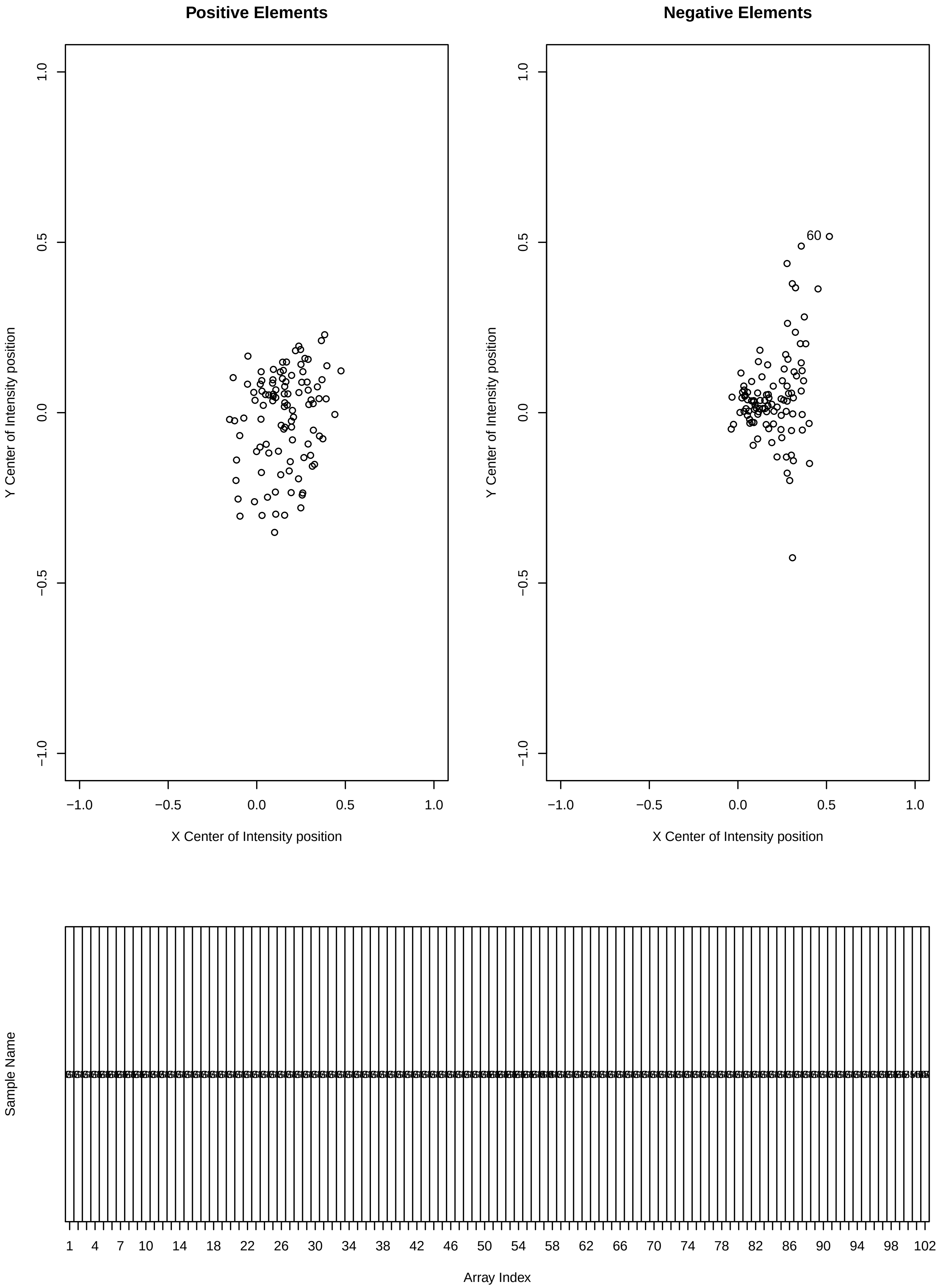

Log2(Intensity)
6
8
10
12
14
1
5
9
14
20
26
32
38
44
50
56
62
68
74
80
86
92
98
Array Index

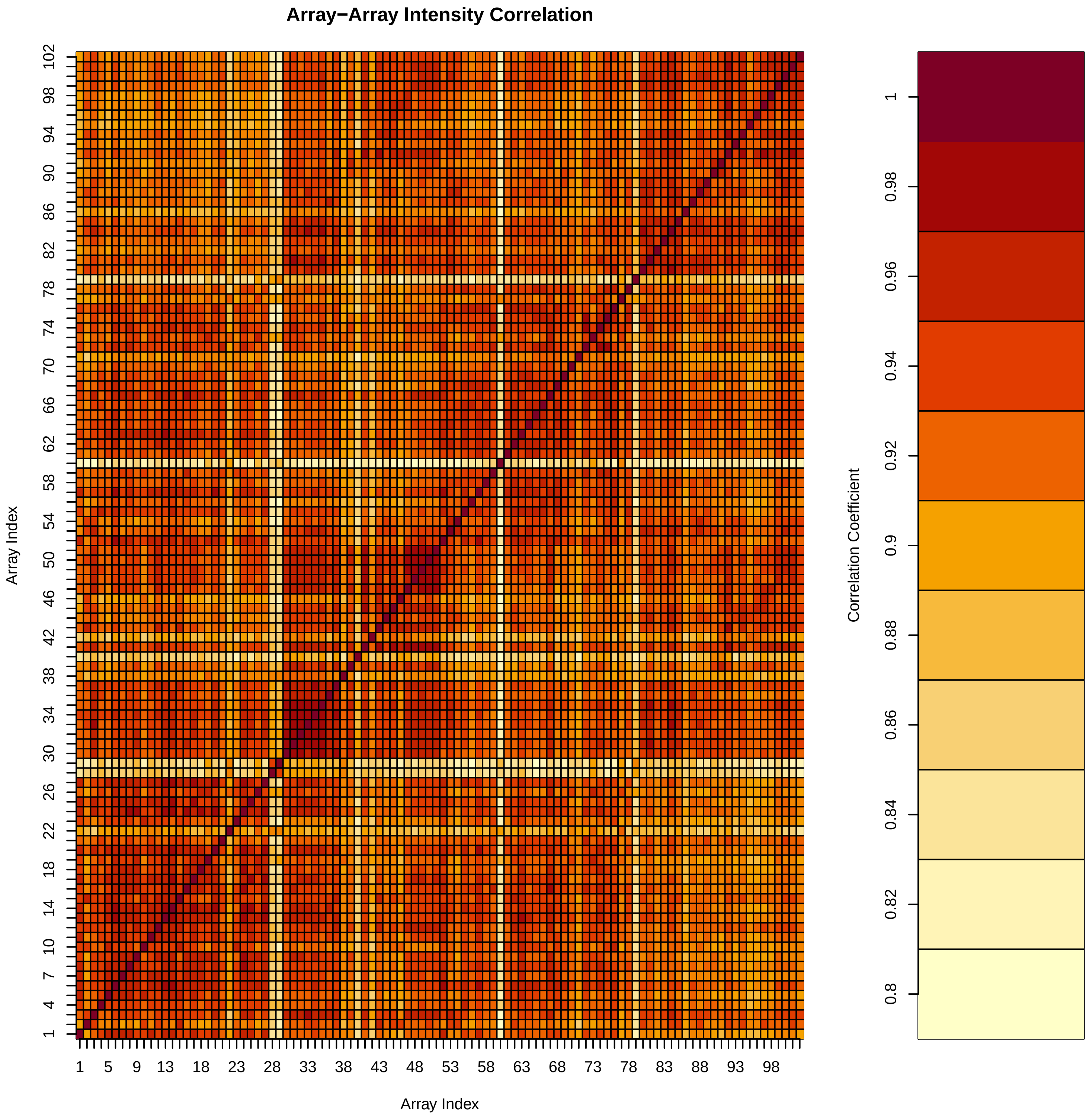

density
0.0
0.1
0.2
0.3
0.4
0.5
0.6
4
6
8
10
12
14
16
log intensity
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34

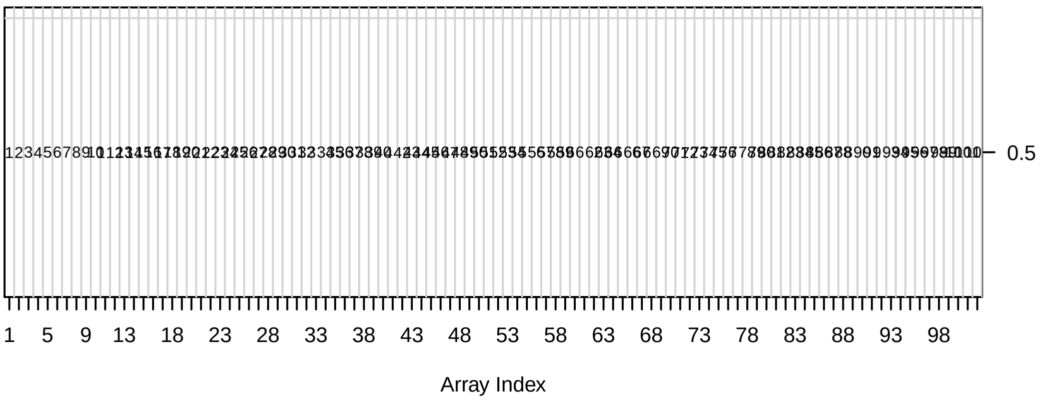

QC Stats
△ actin3/actin5
○ gapdh3/gapdh5
bioB
−3
−2
−1
0
1
2
3
GSM574053.CEL
GSM574052.CEL
GSM574051.CEL
GSM574050.CEL
GSM574049.CEL
GSM574048.CEL
GSM574047.CEL
GSM574046.CEL
GSM574045.CEL
GSM574044.CEL
GSM574043.CEL
GSM574042.CEL
GSM574041.CEL
GSM574040.CEL
GSM574039.CEL
GSM574038.CEL
GSM574037.CEL
GSM574036.CEL
GSM574035.CEL
GSM574034.CEL
GSM574033.CEL
GSM574032.CEL
GSM574031.CEL
GSM574030.CEL
GSM574029.CEL
GSM574028.CEL
GSM574027.CEL
GSM574026.CEL
GSM574025.CEL
GSM574024.CEL
GSM574023.CEL
GSM574022.CEL
GSM574021.CEL
GSM574020.CEL
GSM574019.CEL
GSM574018.CEL
GSM574017.CEL
GSM574016.CEL
GSM574015.CEL
GSM574014.CEL
GSM574013.CEL
GSM574012.CEL
GSM574011.CEL
GSM574010.CEL
GSM574009.CEL
GSM574008.CEL
GSM574007.CEL
GSM574006.CEL
GSM574005.CEL
GSM574004.CEL
GSM574003.CEL
GSM574002.CEL
GSM574001.CEL
GSM574000.CEL
GSM573999.CEL
GSM573998.CEL
GSM573997.CEL
GSM573996.CEL
GSM573995.CEL
GSM573994.CEL
GSM573993.CEL
GSM573992.CEL
GSM573991.CEL
GSM573990.CEL
GSM573989.CEL
GSM573988.CEL
GSM573987.CEL
GSM573986.CEL
GSM573985.CEL
GSM573984.CEL
GSM573983.CEL
GSM573982.CEL
GSM573981.CEL
GSM573980.CEL
GSM573979.CEL
GSM573978.CEL
GSM573977.CEL
GSM573976.CEL
GSM573975.CEL
GSM573974.CEL
GSM573973.CEL
GSM573972.CEL
GSM573971.CEL
GSM573970.CEL
GSM573969.CEL
GSM573968.CEL
GSM573967.CEL
GSM573966.CEL
GSM573965.CEL
GSM573964.CEL
GSM573963.CEL
GSM573962.CEL
GSM573961.CEL
GSM573960.CEL
GSM573959.CEL
GSM573958.CEL
GSM573957.CEL
GSM573956.CEL
GSM573955.CEL
GSM573954.CEL
GSM573953.CEL
GSM573952.CEL
54.13


Positive Border Elements
Intensity
0
10000
20000
30000
40000
50000
60000
1
7
14
22
30
38
46
54
62
70
78
86
94
Array Index
Negative Border Elements
Intensity
0
2000
4000
6000
8000
1
7
14
22
30
38
46
54
62
70
78
86
94
Array Index

Sample Name
1
4
7
10
14
18
22
26
30
34
38
42
46
50
54
58
62
66
70
74
78
82
86
90
94
98
102
Array Index


Positive Elements
Negative Elements
Y Center of Intensity position
X Center of Intensity position
1.0
0.5
0.0
−0.5
−1.0
60
Sample Name
Array Index
1
4
7
10
14
18
22
26
30
34
38
42
46
50
54
58
62
66
70
74
78
82
86
90
94
98
102


Array−Array Intensity Correlation
Array Index
1
4
7
10
14
18
22
26
30
34
38
42
46
50
54
58
62
66
70
74
78
82
86
90
94
98
102
1
5
9
13
18
23
28
33
38
43
48
53
58
63
68
73
78
83
88
93
98
Array Index
Correlation Coefficient
1
0.98
0.96
0.94
0.92
0.9
0.88
0.86
0.84
0.82
0.8


0.5
1
5
9
13
18
23
28
33
38
43
48
53
58
63
68
73
78
83
88
93
98
Array Index

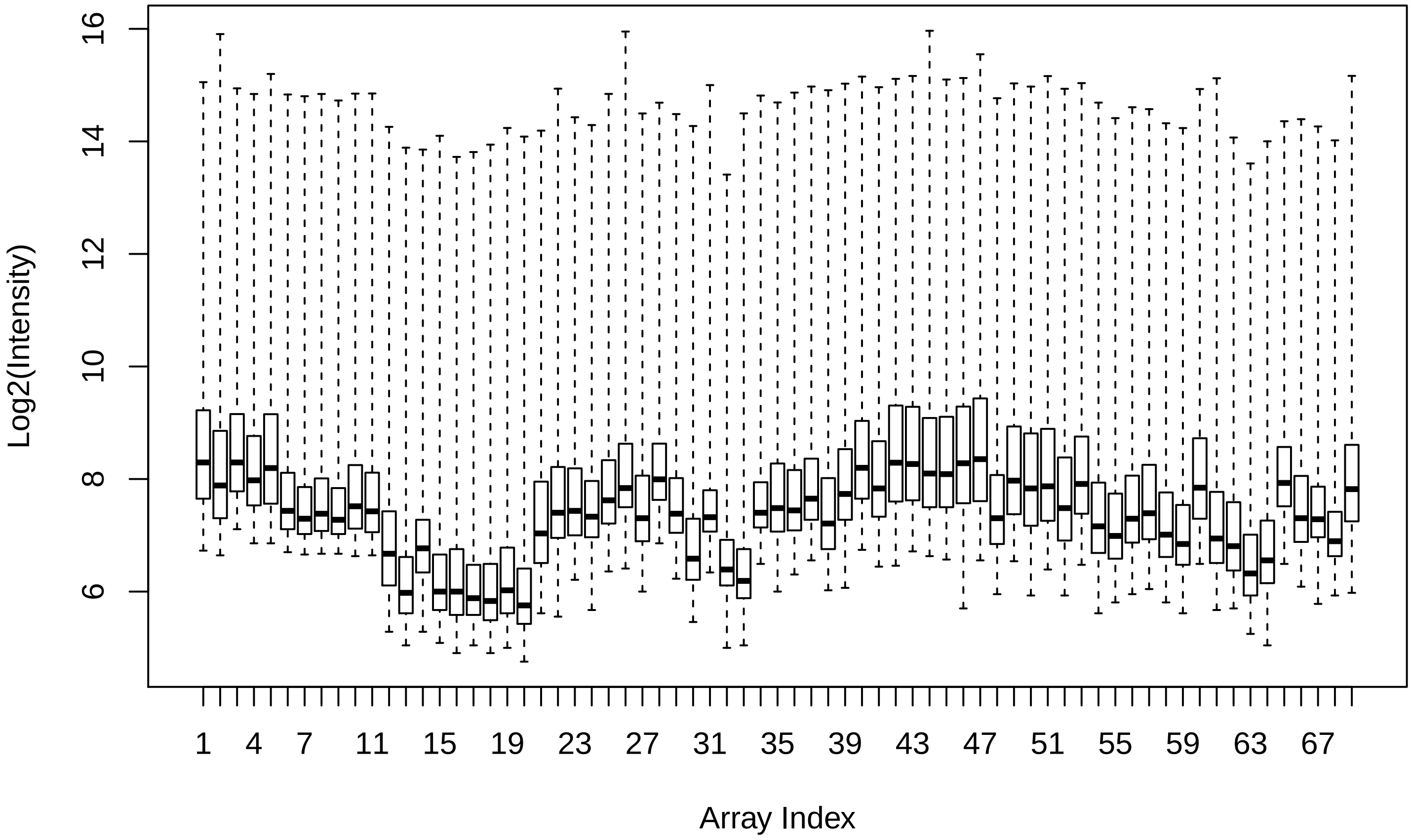
AffyBatch QC Report

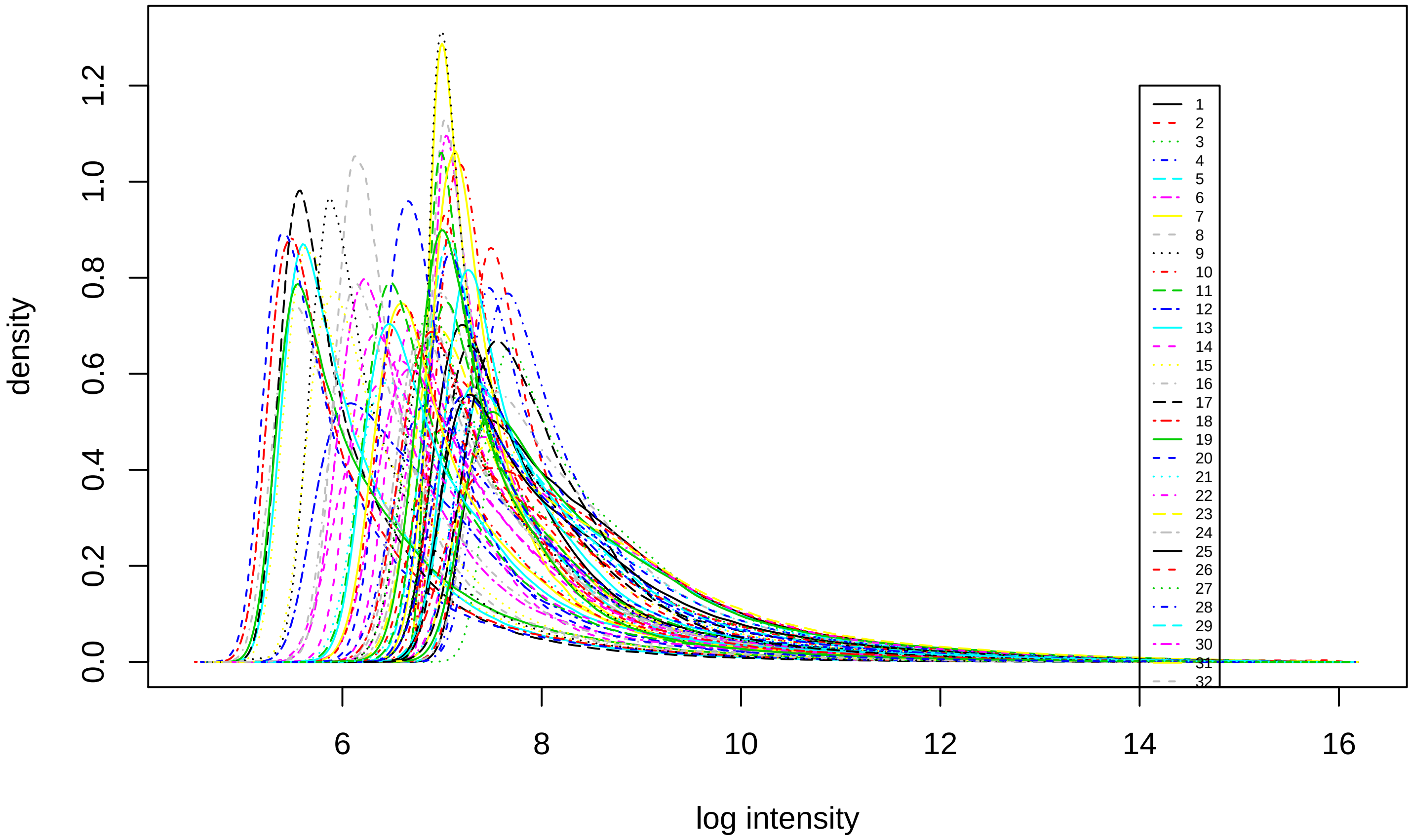
GSE26886

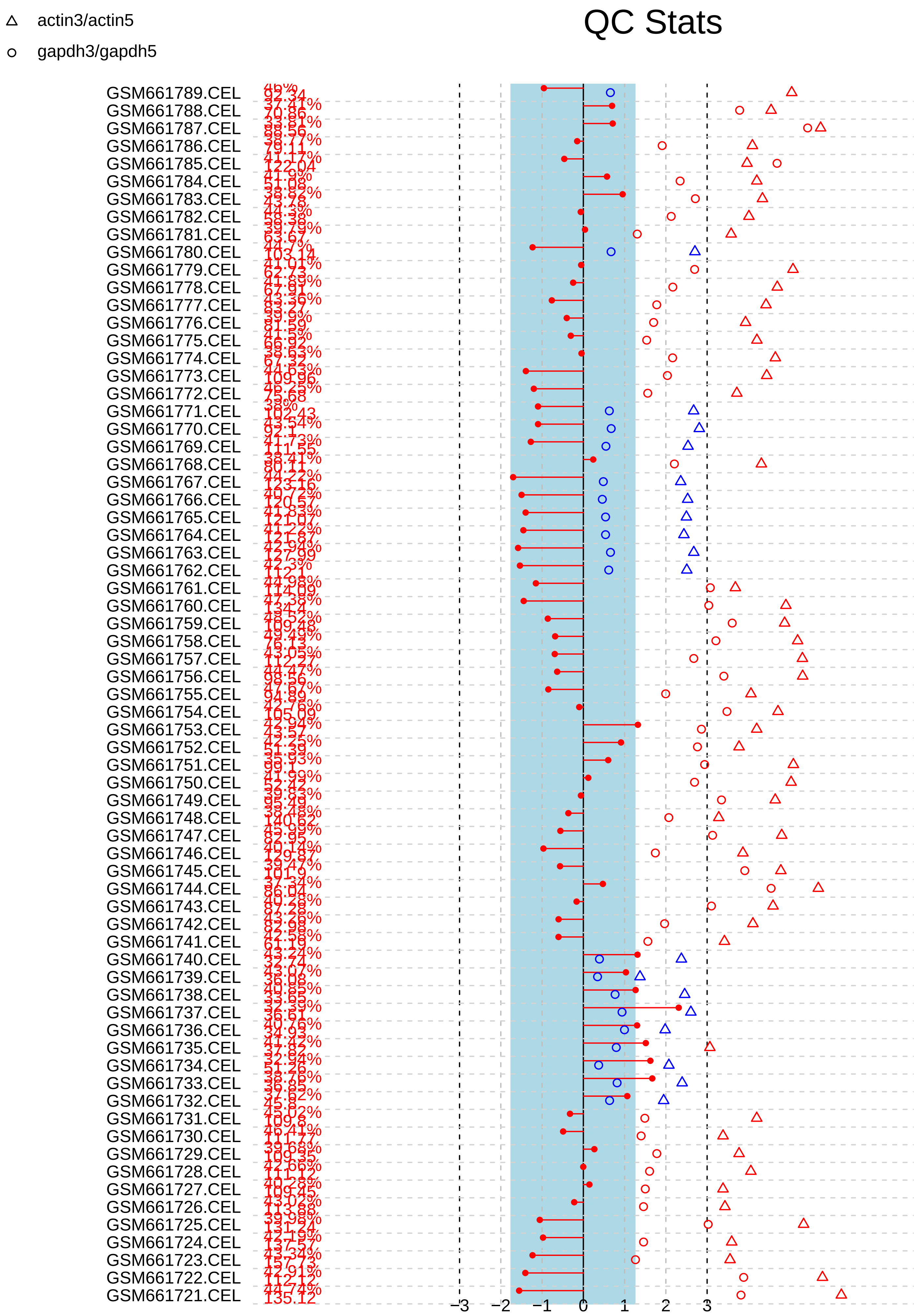
Array Index	Array Name
	GSM66 72 CE
2	GSM66 722 CE
3	GSM66 723 CE
4	GSM66 724 CE
5	GSM66 725 CE
6	GSM66 726 CE
7	GSM66 727 CE
8	GSM66 728 CE
9	GSM66 729 CE
0	GSM66 730 CE
	GSM66 73 CE
2	GSM66 732 CE
3	GSM66 733 CE
4	GSM66 734 CE
5	GSM66 735 CE
6	GSM66 736 CE
7	GSM66 737 CE
8	GSM66 738 CE
9	GSM66 739 CE
20	GSM66 740 CE
2	GSM66 74 CE
22	GSM66 742 CE
23	GSM66 743 CE
24	GSM66 744 CE
25	GSM66 745 CE
26	GSM66 746 CE
27	GSM66 747 CE
28	GSM66 748 CE
29	GSM66 749 CE
30	GSM66 750 CE
3	GSM66 75 CE
32	GSM66 752 CE
33	GSM66 753 CE
34	GSM66 754 CE
35	GSM66 755 CE
36	GSM66 756 CE
37	GSM66 757 CE
38	GSM66 758 CE
39	GSM66 759 CE
40	GSM66 760 CE
4	GSM66 76 CE
42	GSM66 762 CE
43	GSM66 763 CE
44	GSM66 764 CE
45	GSM66 765 CE
46	GSM66 766 CE
47	GSM66 767 CE
48	GSM66 768 CE
49	GSM66 769 CE
50	GSM66 770 CE
5	GSM66 77 CE
52	GSM66 772 CE
53	GSM66 773 CE
54	GSM66 774 CE
55	GSM66 775 CE
56	GSM66 776 CE
57	GSM66 777 CE
58	GSM66 778 CE
59	GSM66 779 CE
60	GSM66 780 CE
6	GSM66 78 CE
62	GSM66 782 CE
63	GSM66 783 CE
64	GSM66 784 CE
65	GSM66 785 CE
66	GSM66 786 CE
67	GSM66 787 CE
68	GSM66 788 CE
69	GSM66 789 CE

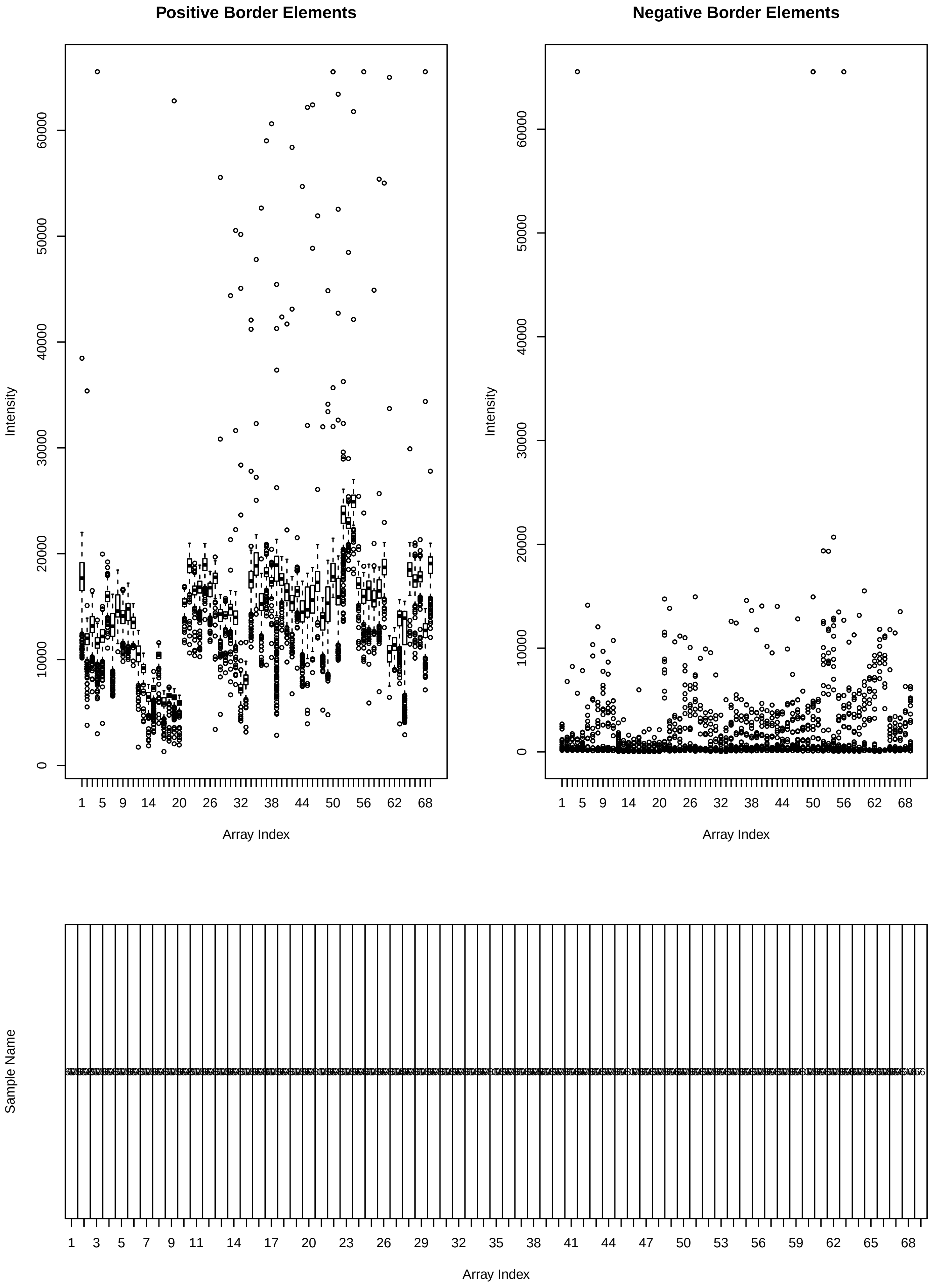
Thu Oct 1 20 09:55 2020

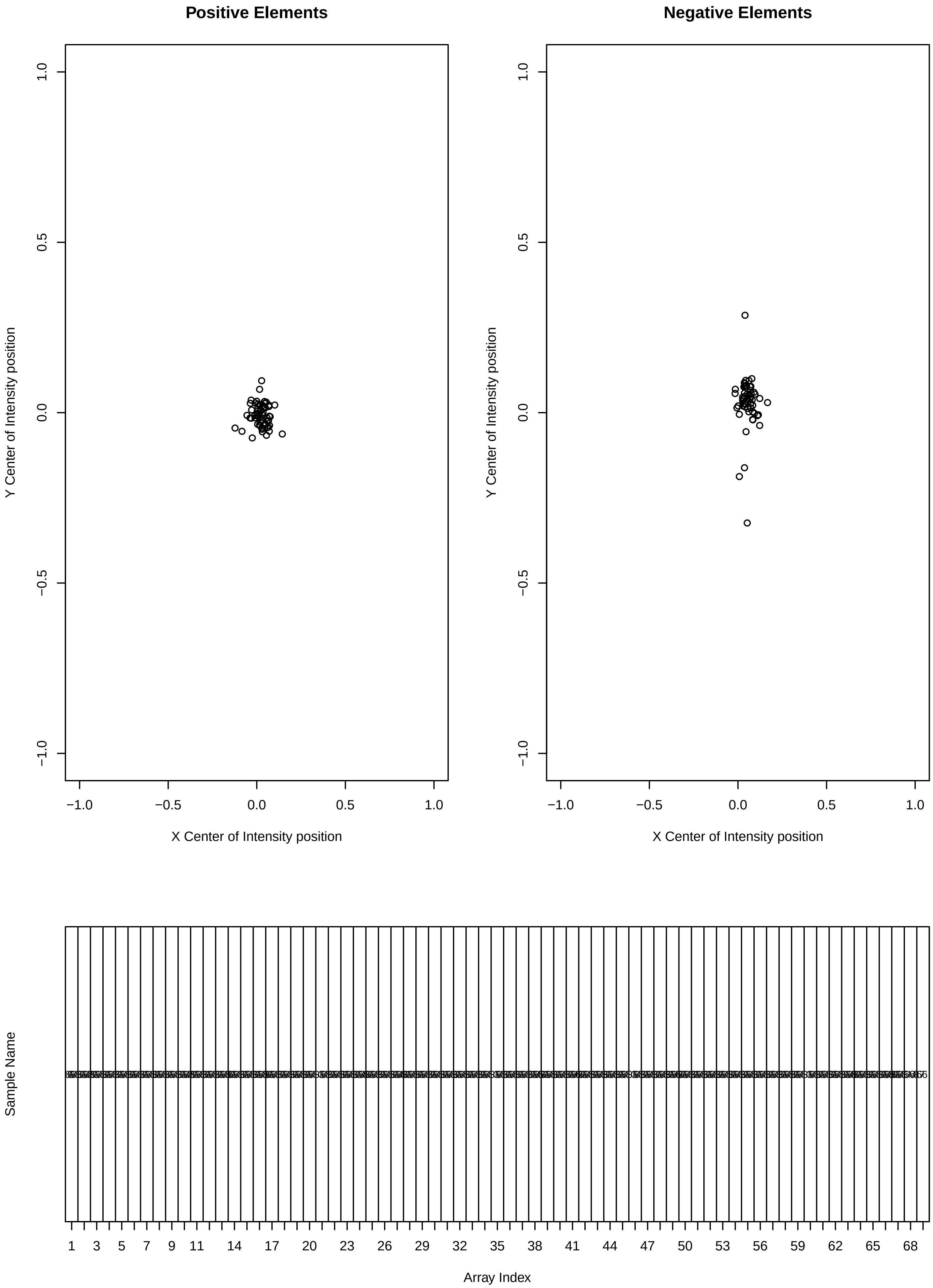
Produced by AffyQCReport R Package

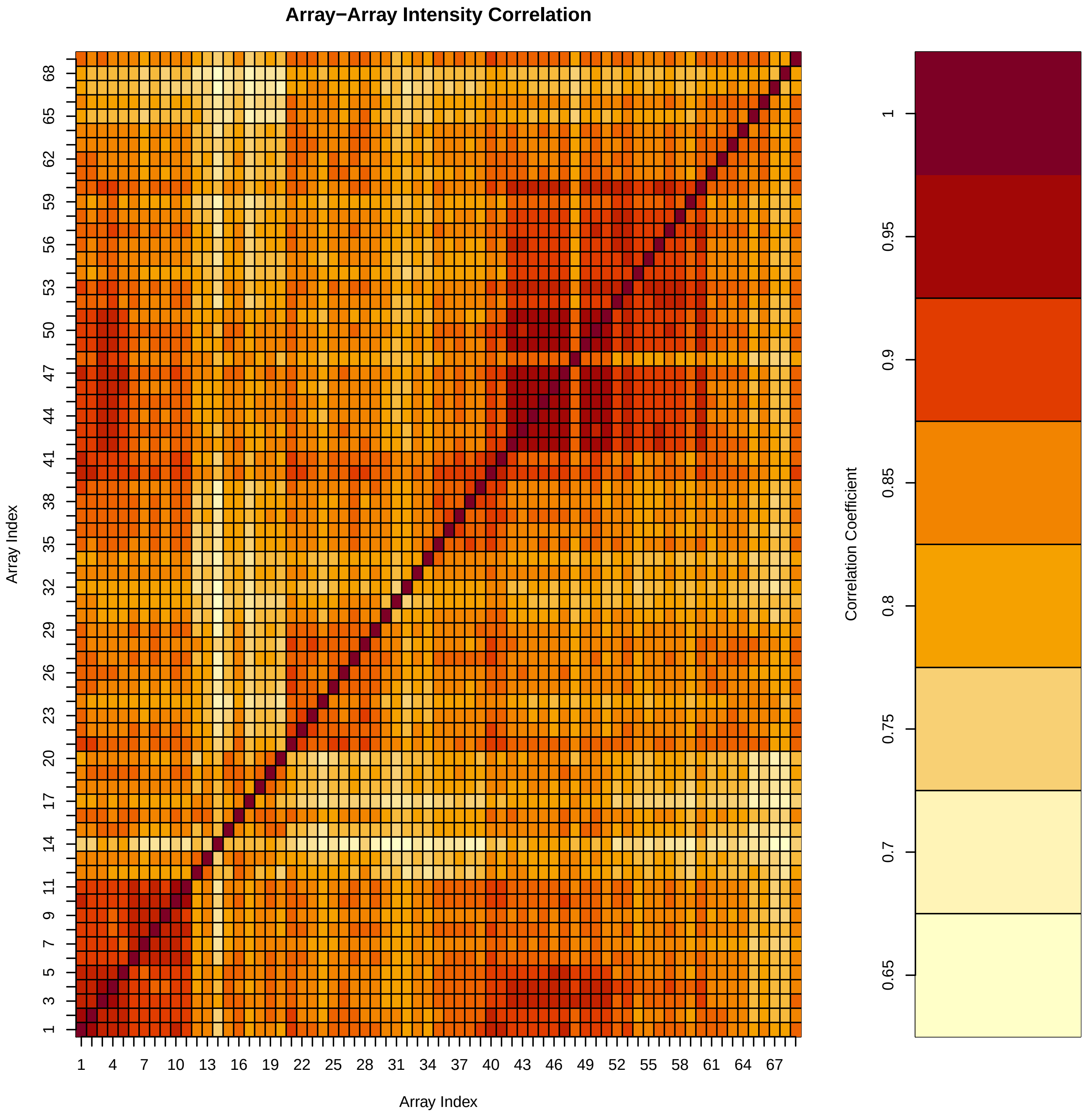

Log2(Intensity)
16
14
12
10
8
6
1
4
7
11
15
19
23
27
31
35
39
43
47
51
55
59
63
67
Array Index

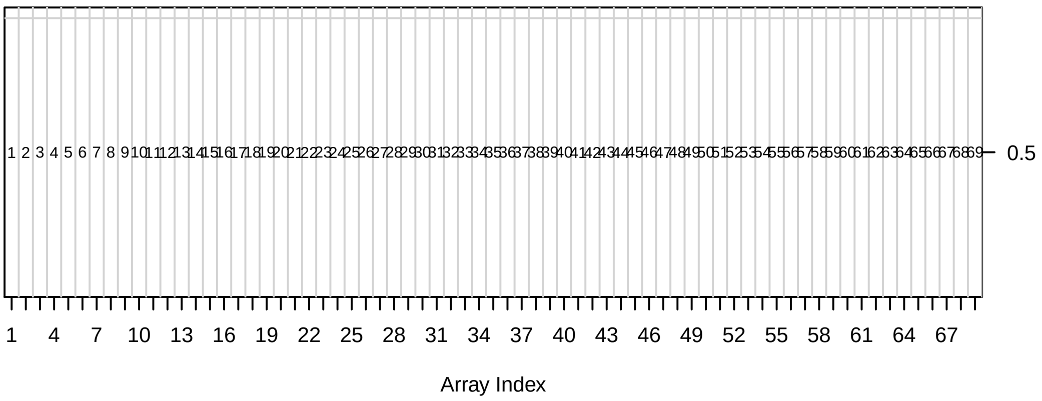

density
1.2
1.0
0.8
0.6
0.4
0.2
0.0
6
8
10
12
14
16
log intensity
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32


QC Stats
△ actin3/actin5
○ gapdh3/gapdh5
GSM661789.CEL
GSM661788.CEL
GSM661787.CEL
GSM661786.CEL
GSM661785.CEL
GSM661784.CEL
GSM661783.CEL
GSM661782.CEL
GSM661781.CEL
GSM661780.CEL
GSM661779.CEL
GSM661778.CEL
GSM661777.CEL
GSM661776.CEL
GSM661775.CEL
GSM661774.CEL
GSM661773.CEL
GSM661772.CEL
GSM661771.CEL
GSM661770.CEL
GSM661769.CEL
GSM661768.CEL
GSM661767.CEL
GSM661766.CEL
GSM661765.CEL
GSM661764.CEL
GSM661763.CEL
GSM661762.CEL
GSM661761.CEL
GSM661760.CEL
GSM661759.CEL
GSM661758.CEL
GSM661757.CEL
GSM661756.CEL
GSM661755.CEL
GSM661754.CEL
GSM661753.CEL
GSM661752.CEL
GSM661751.CEL
GSM661750.CEL
GSM661749.CEL
GSM661748.CEL
GSM661747.CEL
GSM661746.CEL
GSM661745.CEL
GSM661744.CEL
GSM661743.CEL
GSM661742.CEL
GSM661741.CEL
GSM661740.CEL
GSM661739.CEL
GSM661738.CEL
GSM661737.CEL
GSM661736.CEL
GSM661735.CEL
GSM661734.CEL
GSM661733.CEL
GSM661732.CEL
GSM661731.CEL
GSM661730.CEL
GSM661729.CEL
GSM661728.CEL
GSM661727.CEL
GSM661726.CEL
GSM661725.CEL
GSM661724.CEL
GSM661723.CEL
GSM661722.CEL
GSM661721.CEL
−3
−2
−1
0
1
2
3


Positive Border Elements
Negative Border Elements
Intensity
0
10000
20000
30000
40000
50000
60000
Array Index
1
5
9
14
20
26
32
38
44
50
56
62
68
Sample Name
Array Index
1
3
5
7
9
11
14
17
20
23
26
29
32
35
38
41
44
47
50
53
56
59
62
65
68


Positive Elements
Negative Elements
Y Center of Intensity position
X Center of Intensity position
1.0
0.5
0.0
−0.5
−1.0
Sample Name
Array Index
1
3
5
7
9
11
14
17
20
23
26
29
32
35
38
41
44
47
50
53
56
59
62
65
68


Array−Array Intensity Correlation
Array Index
1
3
5
7
9
11
14
17
20
23
26
29
32
35
38
41
44
47
50
53
56
59
62
65
68
1
4
7
10
13
16
19
22
25
28
31
34
37
40
43
46
49
52
55
58
61
64
67
Array Index
Correlation Coefficient
1
0.95
0.9
0.85
0.8
0.75
0.7
0.65


1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58 59 60 61 62 63 64 65 66 67 68 69
0.5
1
4
7
10
13
16
19
22
25
28
31
34
37
40
43
46
49
52
55
58
61
64
67
Array Index

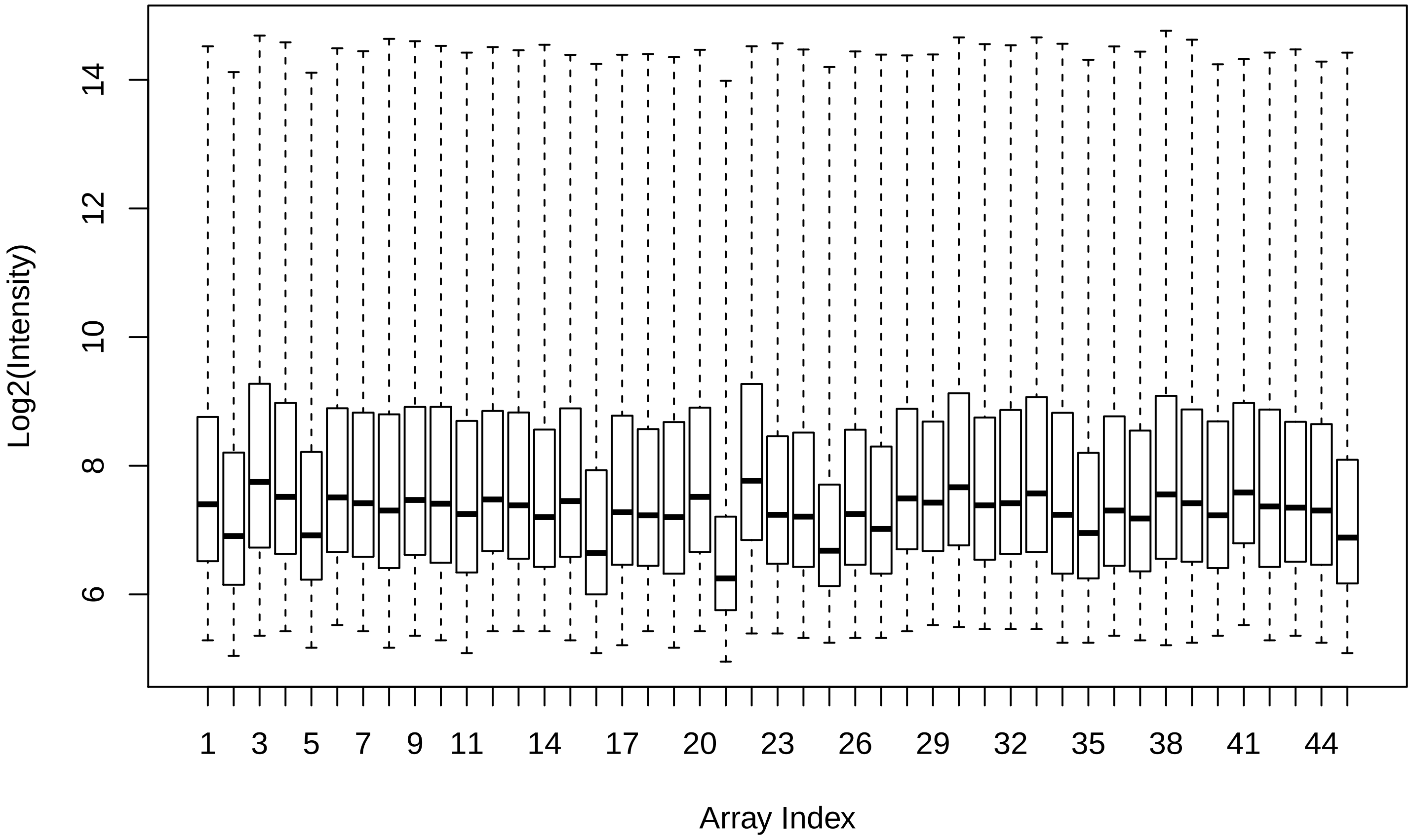
AffyBatch QC Report

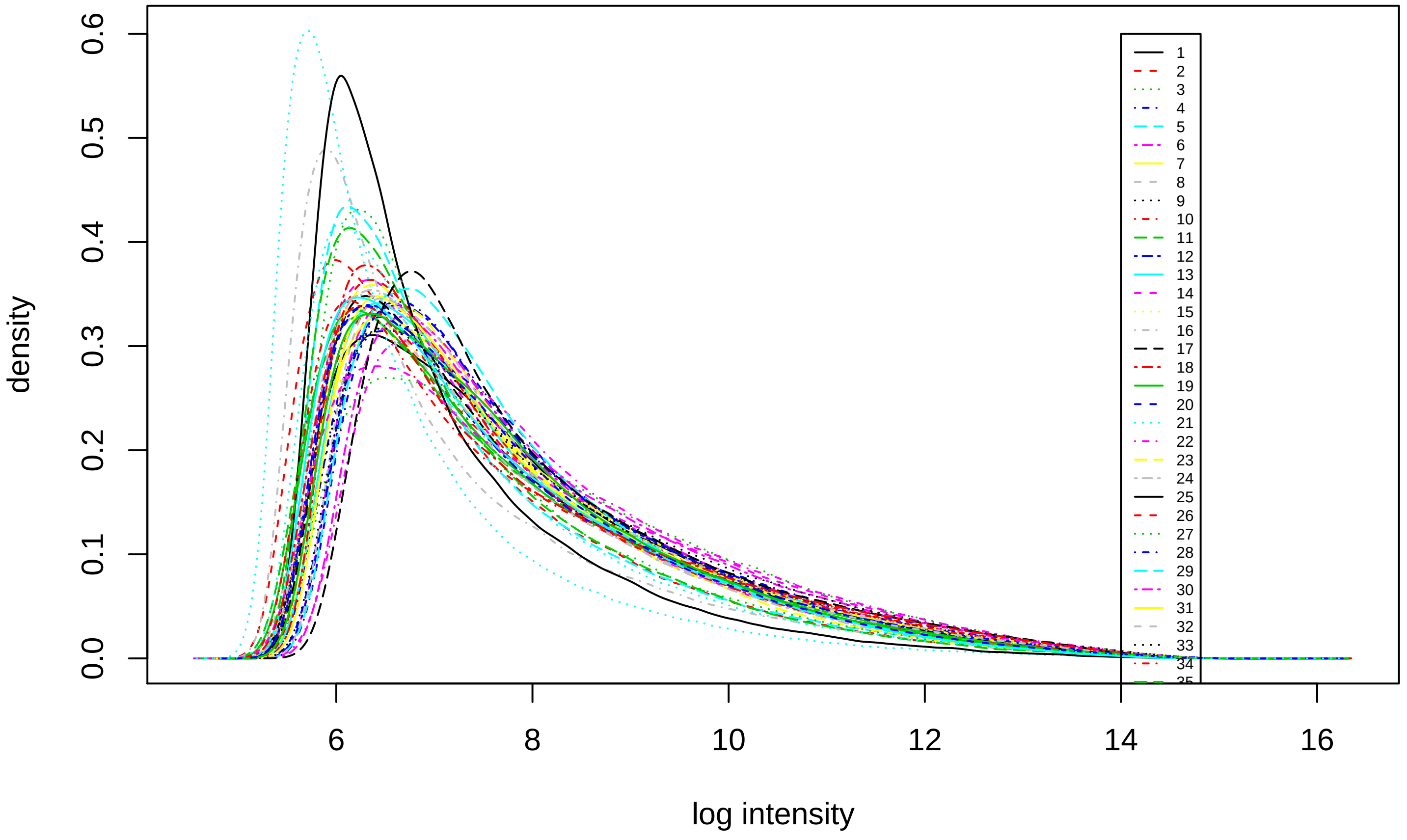
GSE29001

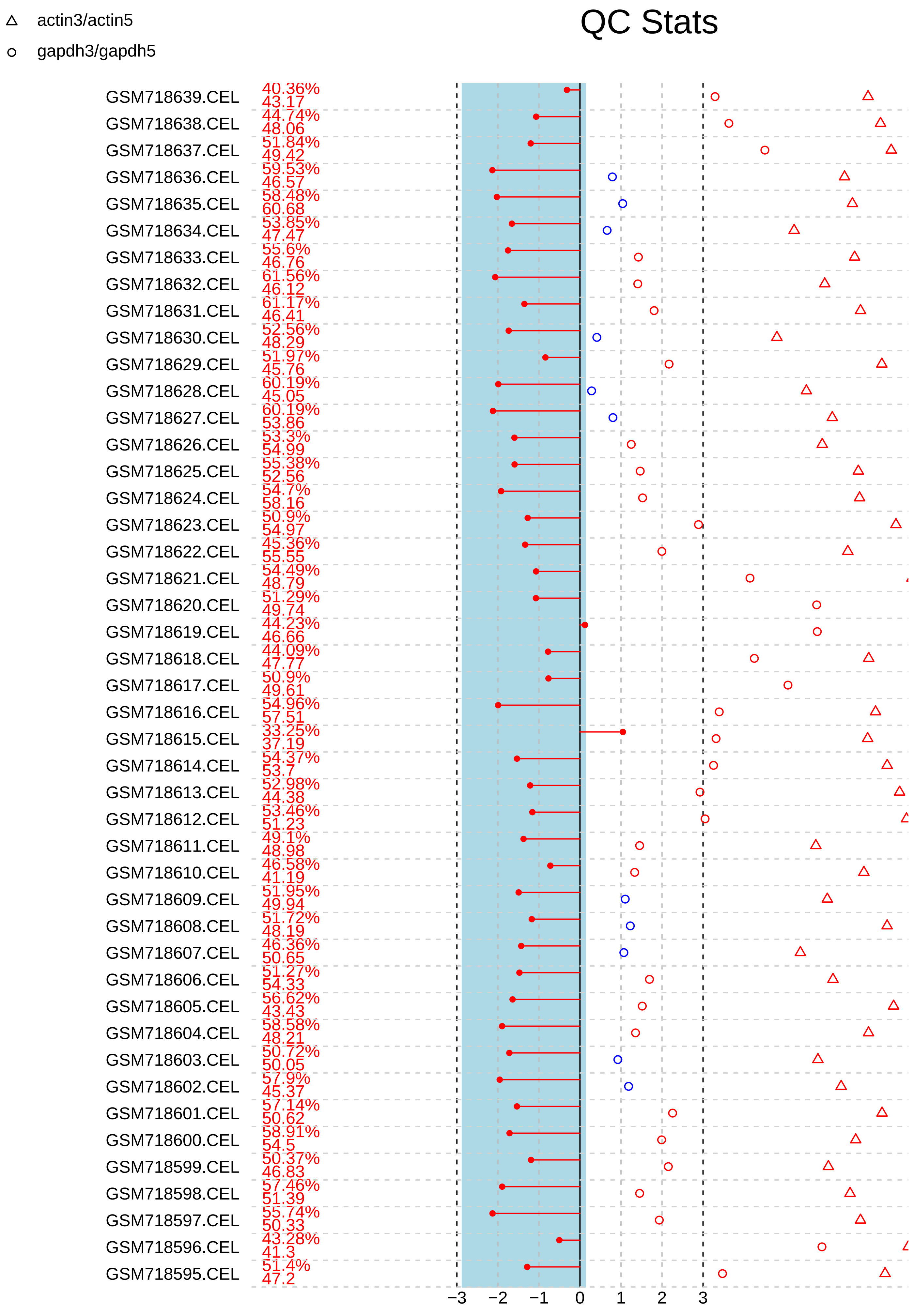
Array Index	Array Name
	GSM7 8595.CEL
2	GSM718596.CEL
3	GSM718597.CEL
	GSM718598.CEL
5	GSM718599.CEL
6	GSM718600.CEL
7	GSM718601.CEL
8	GSM718602.CEL
9	GSM718603.CEL
10	GSM71860 .CEL
11	GSM718605.CEL
12	GSM718606.CEL
13	GSM718607.CEL
1	GSM718608.CEL
15	GSM718609.CEL
16	GSM718610.CEL
17	GSM718611.CEL
18	GSM718612.CEL
19	GSM718613.CEL
20	GSM71861 .CEL
21	GSM718615.CEL
22	GSM718616.CEL
23	GSM718617.CEL
2	GSM718618.CEL
25	GSM718619.CEL
26	GSM718620.CEL
27	GSM718621.CEL
28	GSM718622.CEL
29	GSM718623.CEL
30	GSM71862 .CEL
31	GSM718625.CEL
32	GSM718626.CEL
33	GSM718627.CEL
3	GSM718628.CEL
35	GSM718629.CEL
36	GSM718630.CEL
37	GSM718631.CEL
38	GSM718632.CEL
39	GSM718633.CEL
0	GSM71863 .CEL
1	GSM718635.CEL
2	GSM718636.CEL
3	GSM718637.CEL
	GSM718638.CEL
5	GSM718639.CEL

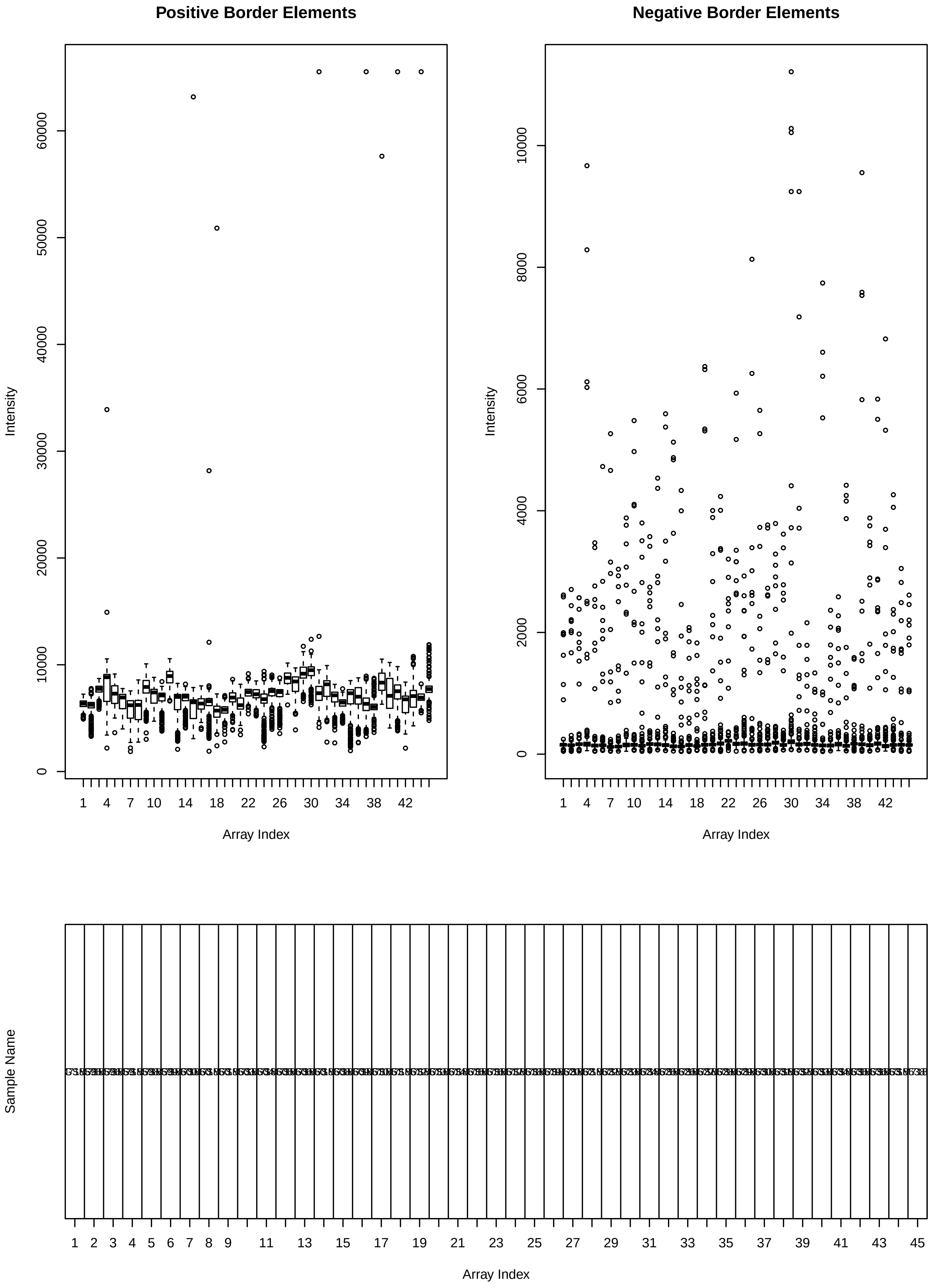
Thu Oct 1 16 54:31 2020

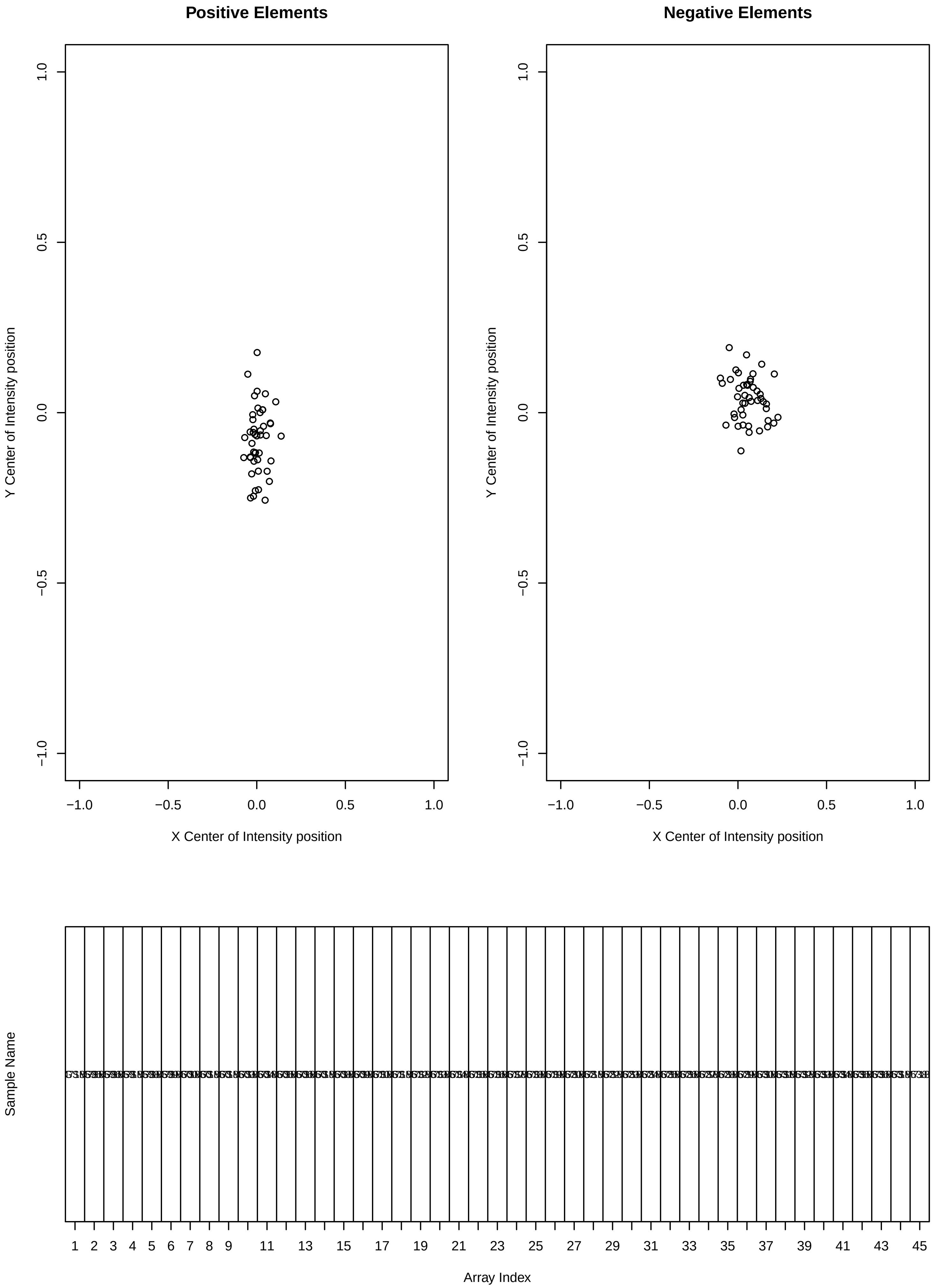
Produced by AffyQCReport R Package

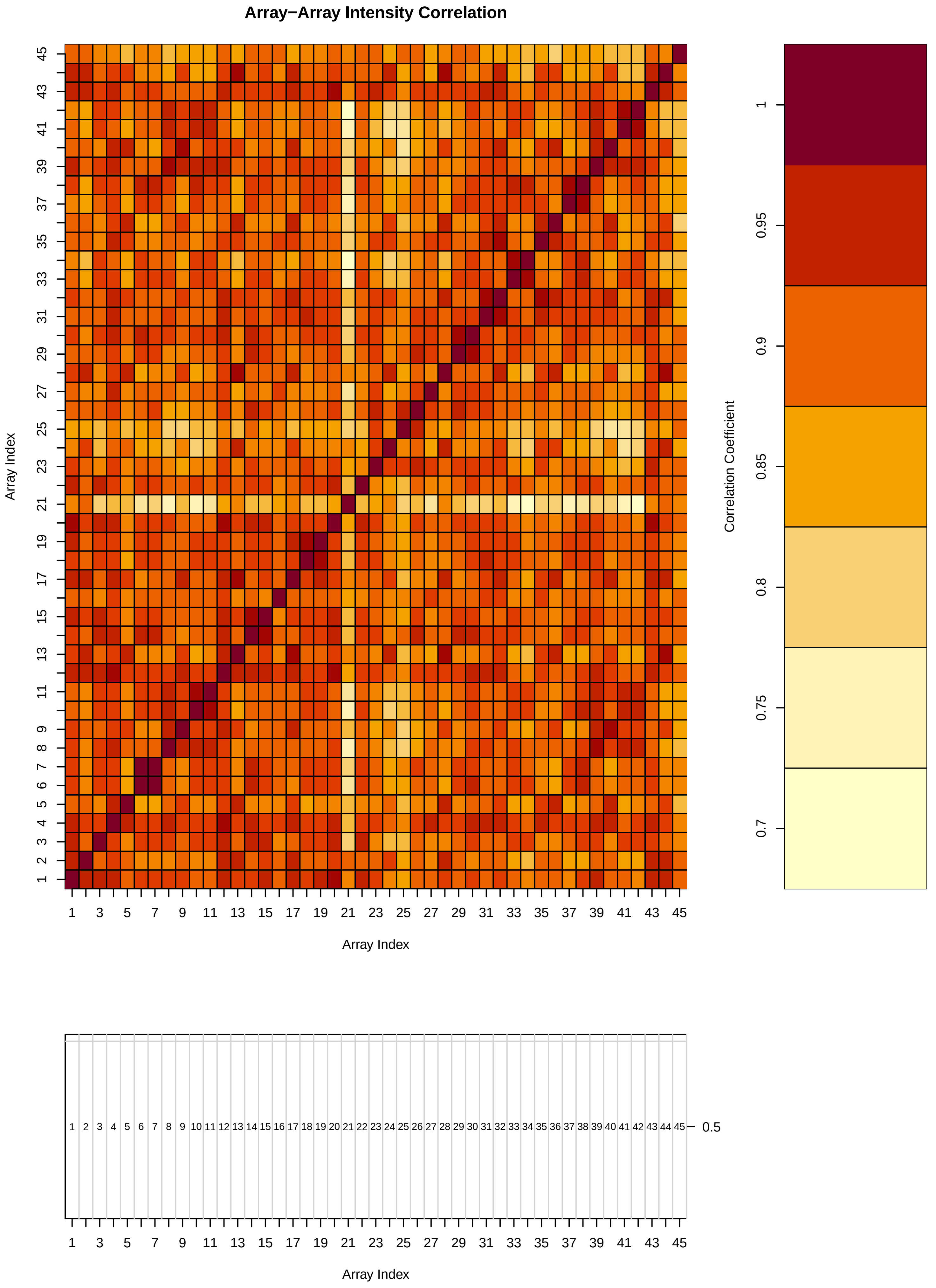

Log2(Intensity)
6
8
10
12
14
1
3
5
7
9
11
14
17
20
23
26
29
32
35
38
41
44
Array Index


density
0.0
0.1
0.2
0.3
0.4
0.5
0.6
6
8
10
12
14
16
log intensity
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35


△ actin3/actin5
○ gapdh3/gapdh5
QC Stats
GSM718639.CEL
40.36%
43.17
GSM718638.CEL
44.74%
48.06
GSM718637.CEL
51.84%
49.42
GSM718636.CEL
59.53%
46.57
GSM718635.CEL
58.48%
60.68
GSM718634.CEL
53.85%
47.47
GSM718633.CEL
55.6%
46.76
GSM718632.CEL
61.56%
46.12
GSM718631.CEL
61.17%
46.41
GSM718630.CEL
52.56%
48.29
GSM718629.CEL
51.97%
45.76
GSM718628.CEL
60.19%
45.05
GSM718627.CEL
60.19%
53.86
GSM718626.CEL
53.3%
54.99
GSM718625.CEL
55.38%
52.56
GSM718624.CEL
54.7%
58.16
GSM718623.CEL
50.9%
54.97
GSM718622.CEL
45.36%
55.55
GSM718621.CEL
54.49%
48.79
GSM718620.CEL
51.29%
49.74
GSM718619.CEL
44.23%
46.66
GSM718618.CEL
44.09%
47.77
GSM718617.CEL
50.9%
49.61
GSM718616.CEL
54.96%
57.51
GSM718615.CEL
33.25%
37.19
GSM718614.CEL
54.37%
53.7
GSM718613.CEL
52.98%
44.38
GSM718612.CEL
53.46%
51.23
GSM718611.CEL
49.1%
48.98
GSM718610.CEL
46.58%
41.19
GSM718609.CEL
51.95%
49.94
GSM718608.CEL
51.72%
48.19
GSM718607.CEL
46.36%
50.65
GSM718606.CEL
51.27%
54.33
GSM718605.CEL
56.62%
43.43
GSM718604.CEL
58.58%
48.21
GSM718603.CEL
50.72%
50.05
GSM718602.CEL
57.9%
45.37
GSM718601.CEL
57.14%
50.62
GSM718600.CEL
58.91%
54.5
GSM718599.CEL
50.37%
46.83
GSM718598.CEL
57.46%
51.39
GSM718597.CEL
55.74%
50.33
GSM718596.CEL
43.28%
41.3
GSM718595.CEL
51.4%
47.2
−3
−2
−1
0
1
2
3


Positive Border Elements
Intensity
0
10000
20000
30000
40000
50000
60000
Array Index
1
4
7
10
14
18
22
26
30
34
38
42
Negative Border Elements
Intensity
0
2000
4000
6000
8000
10000
Array Index
1
4
7
10
14
18
22
26
30
34
38
42
Sample Name
Array Index
1
2
3
4
5
6
7
8
9
11
13
15
17
19
21
23
25
27
29
31
33
35
37
39
41
43
45


Positive Elements
Negative Elements
Y Center of Intensity position
X Center of Intensity position
1.0
0.5
0.0
−0.5
−1.0
Sample Name
Array Index
1 2 3 4 5 6 7 8 9 11 13 15 17 19 21 23 25 27 29 31 33 35 37 39 41 43 45


Array−Array Intensity Correlation
Array Index
1
3
5
7
9
11
13
15
17
19
21
23
25
27
29
31
33
35
37
39
41
43
45
Correlation Coefficient
1
0.95
0.9
0.85
0.8
0.75
0.7
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45
0.5

AffyBatch QC Report

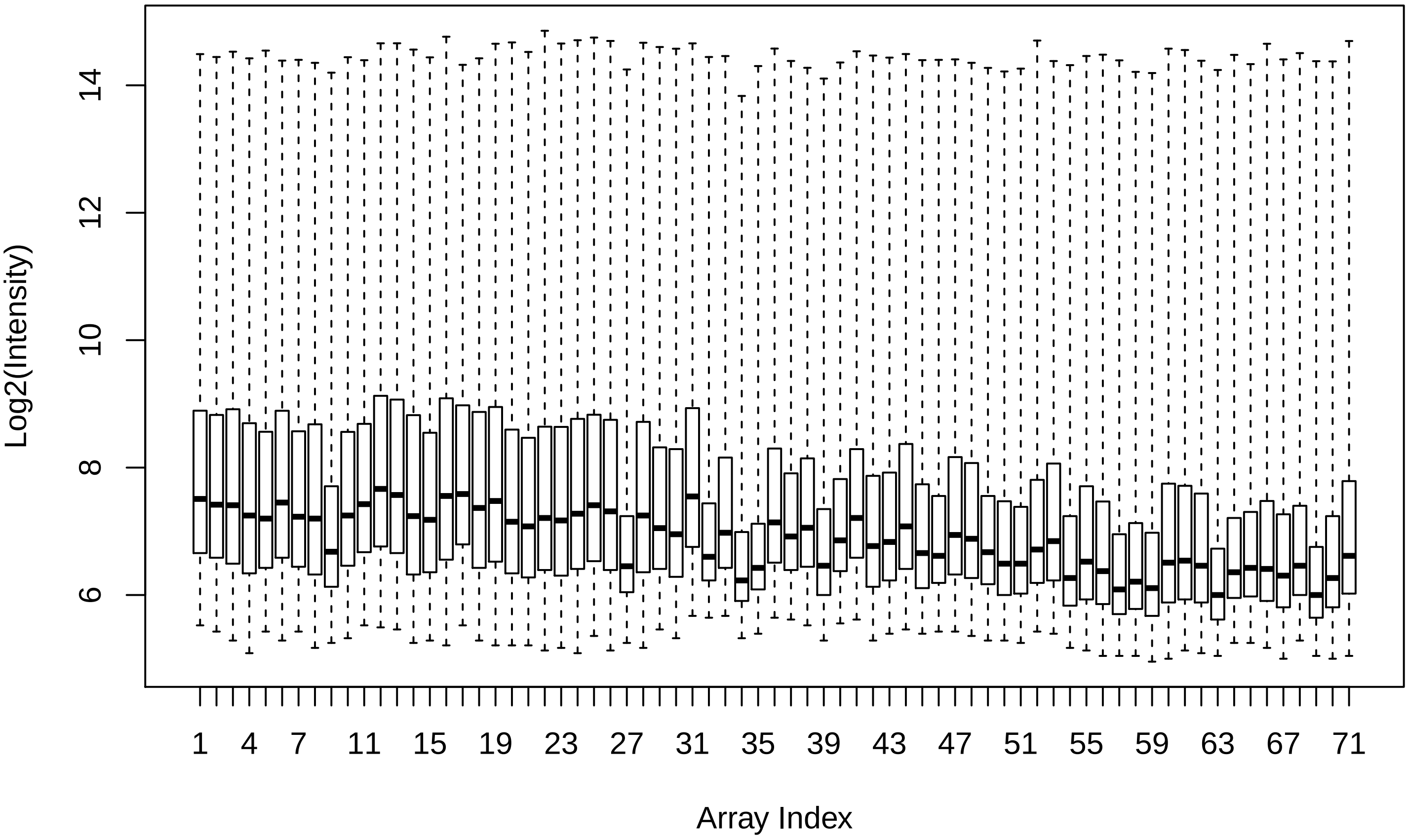
GSE

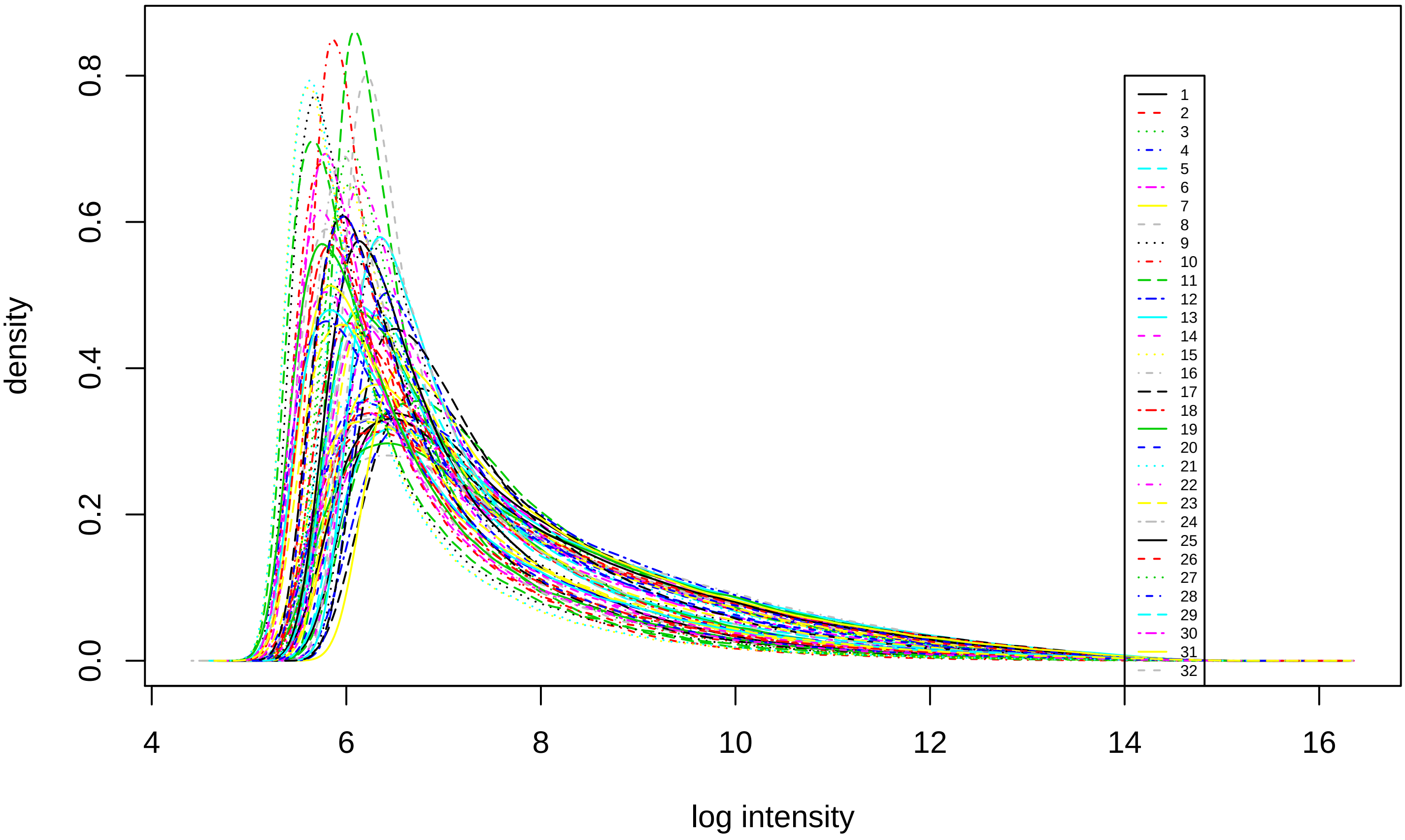
GSE33426

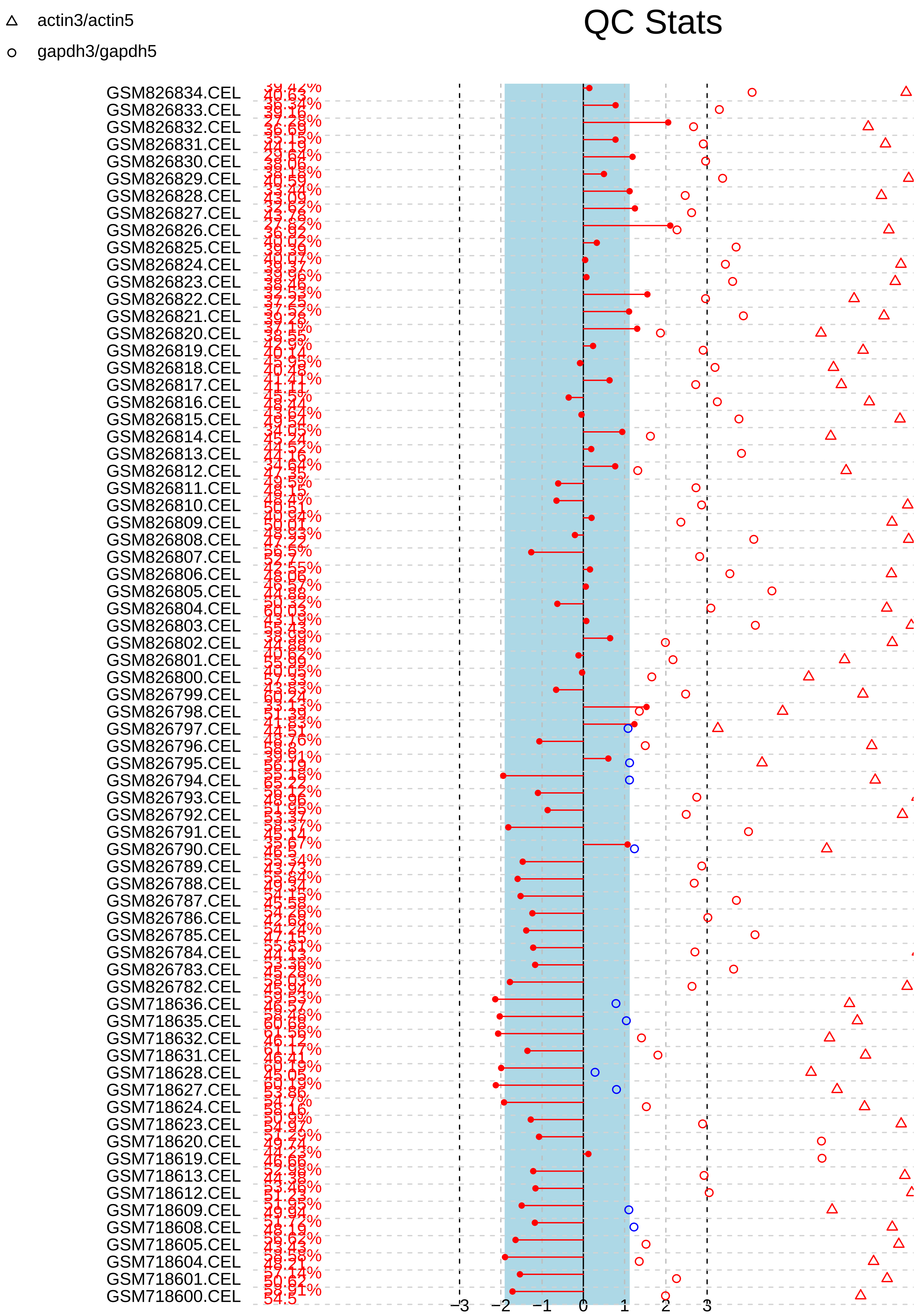
Array Index	Array Name
	GSM7 8600 CE
2	GSM7 860 CE
3	GSM7 8604 CE
4	GSM7 8605 CE
5	GSM7 8608 CE
6	GSM7 8609 CE
7	GSM7 86 2 CE
8	GSM7 86 3 CE
9	GSM7 86 9 CE
0	GSM7 8620 CE
	GSM7 8623 CE
2	GSM7 8624 CE
3	GSM7 8627 CE
4	GSM7 8628 CE
5	GSM7 863 CE
6	GSM7 8632 CE
7	GSM7 8635 CE
8	GSM7 8636 CE
9	GSM826782 CE
20	GSM826783 CE
2	GSM826784 CE
22	GSM826785 CE
23	GSM826786 CE
24	GSM826787 CE
25	GSM826788 CE
26	GSM826789 CE
27	GSM826790 CE
28	GSM82679 CE
29	GSM826792 CE
30	GSM826793 CE
3	GSM826794 CE
32	GSM826795 CE
33	GSM826796 CE
34	GSM826797 CE
35	GSM826798 CE
36	GSM826799 CE
37	GSM826800 CE
38	GSM82680 CE
39	GSM826802 CE
40	GSM826803 CE
4	GSM826804 CE
42	GSM826805 CE
43	GSM826806 CE
44	GSM826807 CE
45	GSM826808 CE
46	GSM826809 CE
47	GSM8268 0 CE
48	GSM8268 CE
49	GSM8268 2 CE
50	GSM8268 3 CE
5	GSM8268 4 CE
52	GSM8268 5 CE
53	GSM8268 6 CE
54	GSM8268 7 CE
55	GSM8268 8 CE
56	GSM8268 9 CE
57	GSM826820 CE
58	GSM82682 CE
59	GSM826822 CE
60	GSM826823 CE
6	GSM826824 CE
62	GSM826825 CE
63	GSM826826 CE
64	GSM826827 CE
65	GSM826828 CE
66	GSM826829 CE
67	GSM826830 CE
68	GSM82683 CE
69	GSM826832 CE
70	GSM826833 CE
7	GSM826834 CE

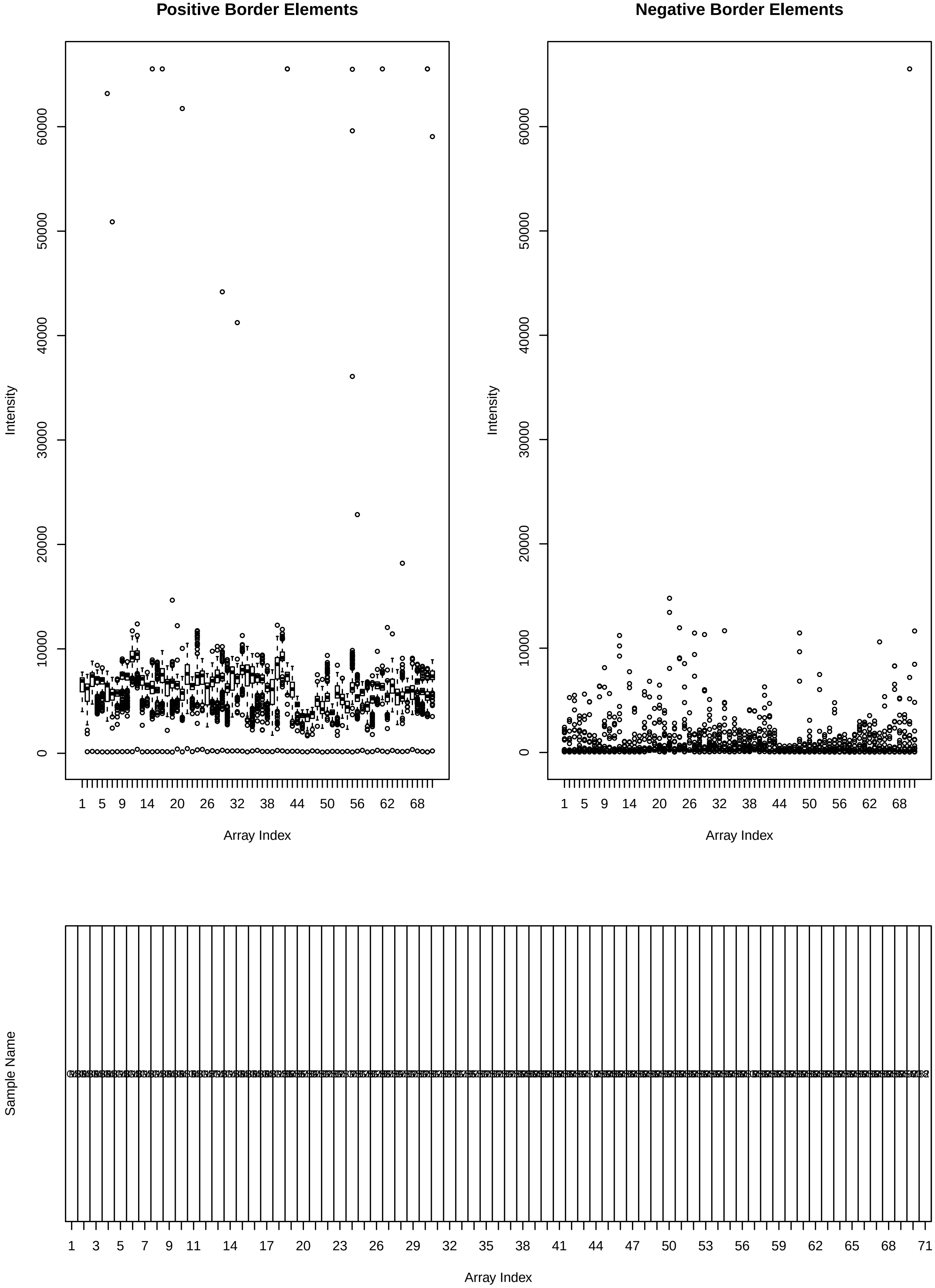
Thu Oct 1 16 59:53 2020

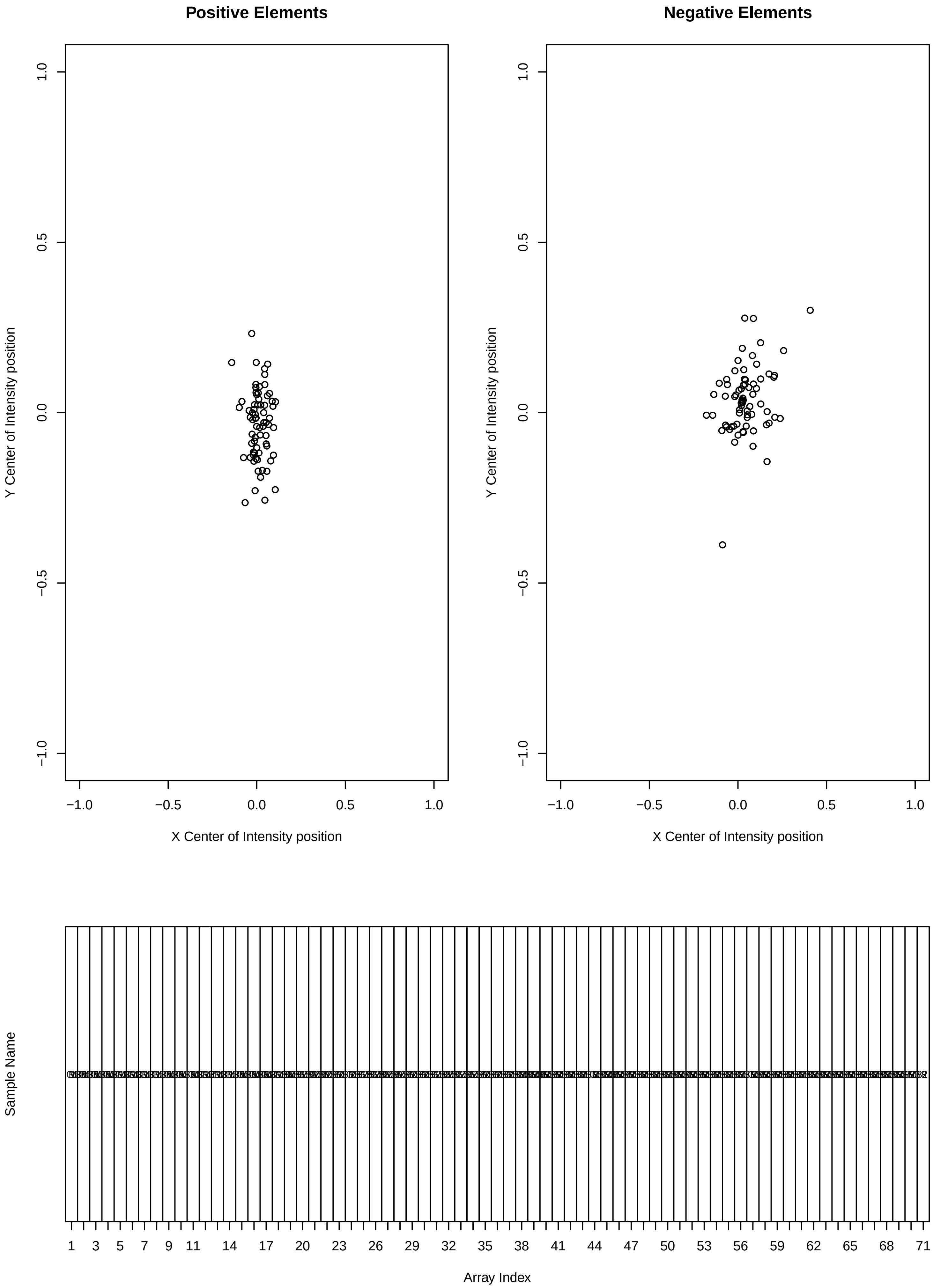
Produced by AffyQCReport R Package

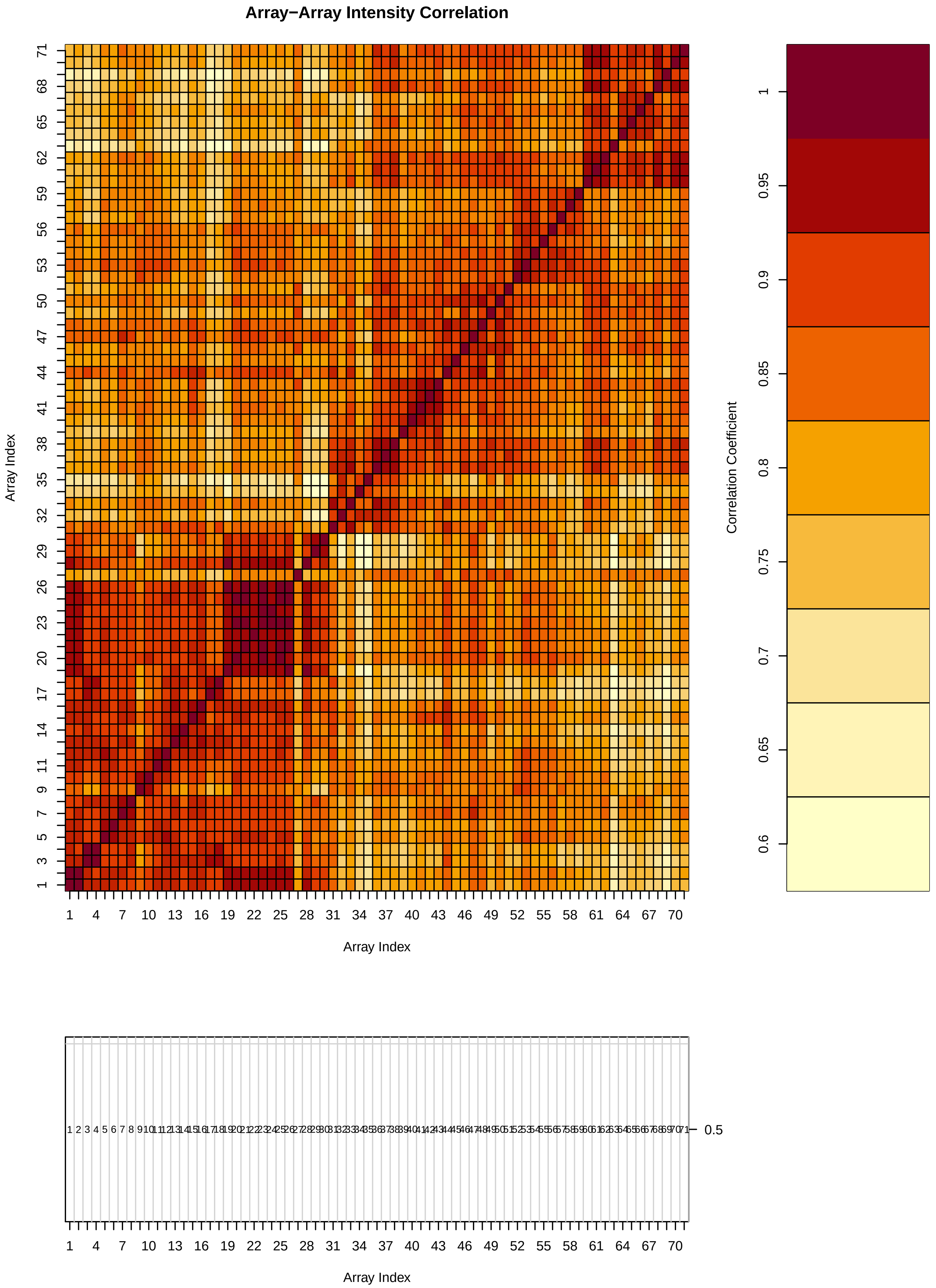

Log2(Intensity)
6
8
10
12
14
1
4
7
11
15
19
23
27
31
35
39
43
47
51
55
59
63
67
71
Array Index


density
0.0
0.2
0.4
0.6
0.8
4
6
8
10
12
14
16
log intensity
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32


QC Stats
actin3/actin5
gapdh3/gapdh5
GSM826834.CEL
GSM826833.CEL
GSM826832.CEL
GSM826831.CEL
GSM826830.CEL
GSM826829.CEL
GSM826828.CEL
GSM826827.CEL
GSM826826.CEL
GSM826825.CEL
GSM826824.CEL
GSM826823.CEL
GSM826822.CEL
GSM826821.CEL
GSM826820.CEL
GSM826819.CEL
GSM826818.CEL
GSM826817.CEL
GSM826816.CEL
GSM826815.CEL
GSM826814.CEL
GSM826813.CEL
GSM826812.CEL
GSM826811.CEL
GSM826810.CEL
GSM826809.CEL
GSM826808.CEL
GSM826807.CEL
GSM826806.CEL
GSM826805.CEL
GSM826804.CEL
GSM826803.CEL
GSM826802.CEL
GSM826801.CEL
GSM826800.CEL
GSM826799.CEL
GSM826798.CEL
GSM826797.CEL
GSM826796.CEL
GSM826795.CEL
GSM826794.CEL
GSM826793.CEL
GSM826792.CEL
GSM826791.CEL
GSM826790.CEL
GSM826789.CEL
GSM826788.CEL
GSM826787.CEL
GSM826786.CEL
GSM826785.CEL
GSM826784.CEL
GSM826783.CEL
GSM826782.CEL
GSM718636.CEL
GSM718635.CEL
GSM718632.CEL
GSM718631.CEL
GSM718628.CEL
GSM718627.CEL
GSM718624.CEL
GSM718623.CEL
GSM718620.CEL
GSM718619.CEL
GSM718613.CEL
GSM718612.CEL
GSM718609.CEL
GSM718608.CEL
GSM718605.CEL
GSM718604.CEL
GSM718601.CEL
GSM718600.CEL
54.5
−3
−2
−1
0
1
2
3


Positive Border Elements
Negative Border Elements
Intensity
0
10000
20000
30000
40000
50000
60000
Array Index
1
5
9
14
20
26
32
38
44
50
56
62
68
Sample Name
1
3
5
7
9
11
14
17
20
23
26
29
32
35
38
41
44
47
50
53
56
59
62
65
68
71


Positive Elements
Negative Elements
Y Center of Intensity position
X Center of Intensity position
1.0
0.5
0.0
−0.5
−1.0
Sample Name
Array Index
1
3
5
7
9
11
14
17
20
23
26
29
32
35
38
41
44
47
50
53
56
59
62
65
68
71


Array−Array Intensity Correlation
Array Index
1
3
5
7
9
11
14
17
20
23
26
29
32
35
38
41
44
47
50
53
56
59
62
65
68
71
1 4 7 10 13 16 19 22 25 28 31 34 37 40 43 46 49 52 55 58 61 64 67 70
Correlation Coefficient
0.6
0.65
0.7
0.75
0.8
0.85
0.9
0.95
1
0.5
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58 59 60 61 62 63 64 65 66 67 68 69 70 71

arrayQualityMetrics report for eset.oligoGeneCore

GSE34619

- **Section 1: Between array comparison**
  - Distances between arrays
  - Principal Component Analysis
- **Section 2: Array intensity distributions** GSE
  - Boxplots
  - Density plots
- **Section 3: Variance mean dependence**
  - Standard deviation versus rank of the mean
- **Section 4: Individual array quality**
  - MA plots

+ Array metadata and outlier detection overview

Section 1: Between array comparison

- Figure 1: Distances between arrays.

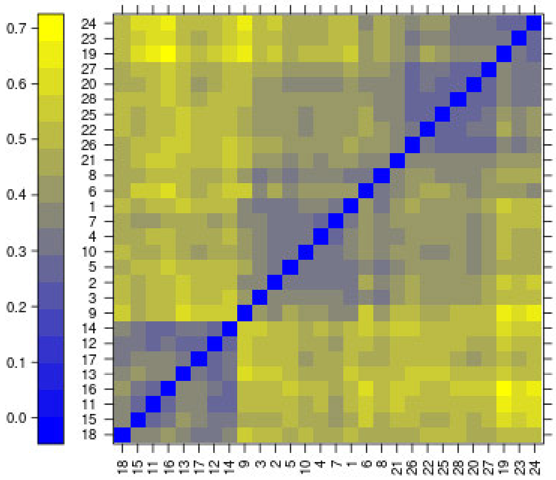


Figure 1 (PDF file) shows a false color heatmap of the distances between arrays. The color scale is chosen to cover the range of distances encountered in the dataset. Patterns in this plot can indicate clustering of the arrays either because of intended biological or unintended experimental factors (batch effects). The distance d_{ab} between two arrays a and b is computed as the mean absolute difference (L_1-distance) between the data of the arrays (using the data from all probes without filtering). In formula, d_{ab} = mean | M_{ai} - M_{bi} |, where M_{ai} is the value of the i-th probe on the a-th array. Outlier detection was performed by looking for arrays for which the sum of the distances to all other arrays, $S_a = \Sigma_b d_{ab}$ was exceptionally large. No such arrays were detected.

+ Figure 2: Outlier detection for Distances between arrays.

- Figure 3: Principal Component Analysis.

array
sampleNames
index

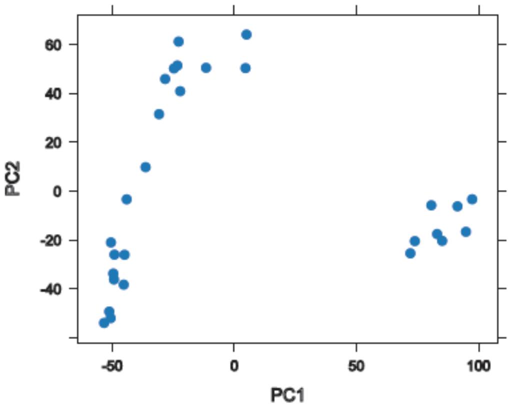


Figure 3 (PDF file) shows a scatterplot of the arrays along the first two principal components. You can use this plot to explore if the arrays cluster, and whether this is according to an intended experimental factor (you can indicate such a factor by color using the 'intgroup' argument), or according to unintended causes such as batch effects. Move the mouse over the points to see the sample names.
Principal component analysis is a dimension reduction and visualisation technique that is here used to project the multivariate data vector of each array into a two-dimensional plot, such that the spatial arrangement of the points in the plot reflects the overall data (dis)similarity between the arrays.

Section 2: Array intensity distributions

- Figure 4: Boxplots.

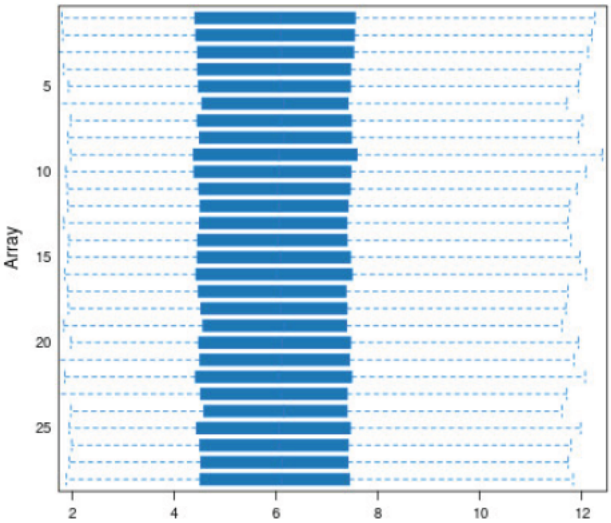


Figure 4 (PDF file) shows boxplots representing summaries of the signal intensity distributions of the arrays. Each box corresponds to one array. Typically, one expects the boxes to have similar positions and widths. If the distribution of an array is very different from the others, this may indicate an experimental problem. Outlier detection was performed by computing the Kolmogorov-Smirnov statistic K_a between each array's distribution and the distribution of the pooled data.

+ Figure 5: Outlier detection for Boxplots.

- Figure 6: Density plots.

array
sampleNames
index

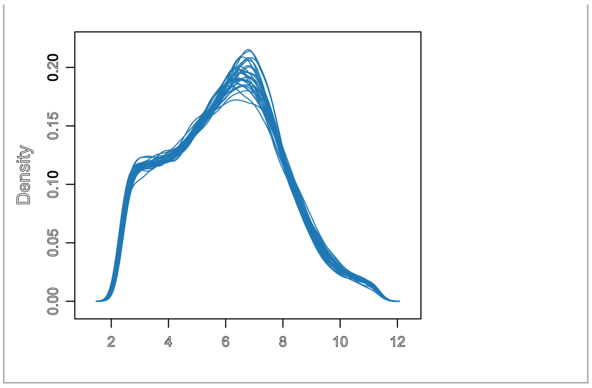


Figure 6 (PDF file) shows density estimates (smoothed histograms) of the data. Typically, the distributions of the arrays should have similar shapes and ranges. Arrays whose distributions are very different from the others should be considered for possible problems. Various features of the distributions can be indicative of quality related phenomena. For instance, high levels of background will shift an array's distribution to the right. Lack of signal diminishes its right right tail. A bulge at the upper end of the intensity range often indicates signal saturation.

Section 3: Variance mean dependence

- Figure 7: Standard deviation versus rank of the mean.

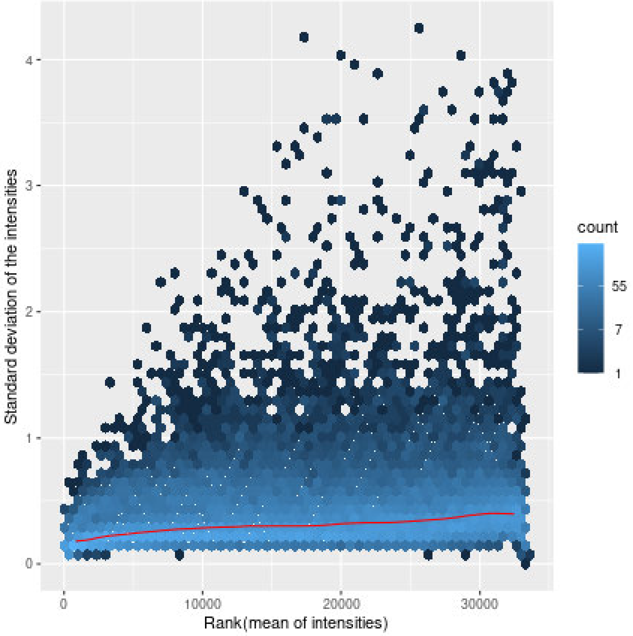


Figure 7 (PDF file) shows a density plot of the standard deviation of the intensities across arrays on the *y*-axis versus the rank of their mean on the *x*-axis. The red dots, connected by lines, show the running median of the standard deviation. After normalisation and transformation to a logarithm(-like) scale, one typically expects the red line to be approximately horizontal, that is, show no substantial trend. In some cases, a hump on the right hand of the x-axis can be observed and is symptomatic of a saturation of the intensities.

Section 4: Individual array quality

- Figure 8: MA plots.

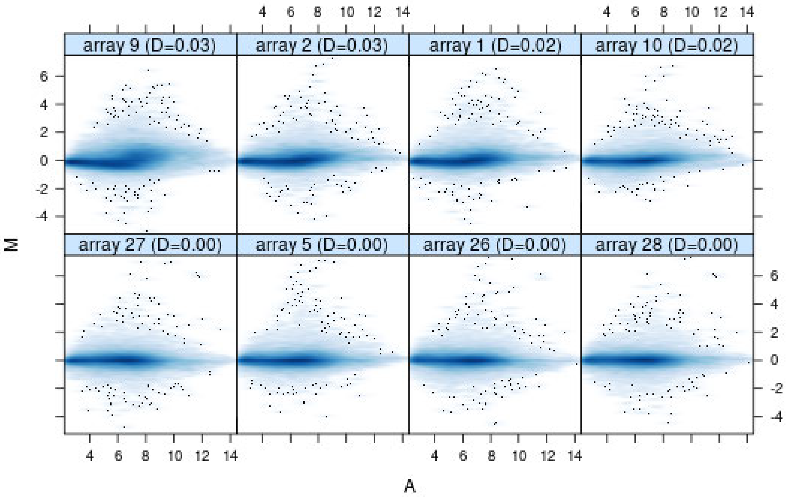


Figure 8 (PDF file) shows MA plots. M and A are defined as:
M = $\log_2(I_1) - \log_2(I_2)$
A = $1/2 (\log_2(I_1)+\log_2(I_2))$,
where I_1 is the intensity of the array studied,and I_2 is the intensity of a "pseudo"-array that consists of the median across arrays. Typically, we expect the mass of the distribution in an MA plot to be concentrated along the M = 0 axis, and there should be no trend in M as a function of A. If there is a trend in the lower range of A, this often indicates that the arrays have different background intensities; this may be addressed by background correction. A trend in the upper range of A can indicate saturation of the measurements; in mild cases, this may be addressed by non-linear normalisation (e.g. quantile normalisation).
Outlier detection was performed by computing Hoeffding's statistic D_a on the joint distribution of A and M for each array. Shown are first the 4 arrays with the highest values of D_a, then the 4 arrays with the lowest values. The value of D_a is shown in the panel headings. 0 arrays had D_a>0.15 and were marked as outliers. For more information on Hoeffing's *D*-statistic, please see the manual page of the function `hoeffd` in the `Hmisc` package.

+ Figure 9: Outlier detection for MA plots.

This report has been created with arrayQualityMetrics 3.42.0 under R version 3.6.0 (2019-04-26).

(Page generated on Tue Oct 6 01:08:39 2020 by hwriter)

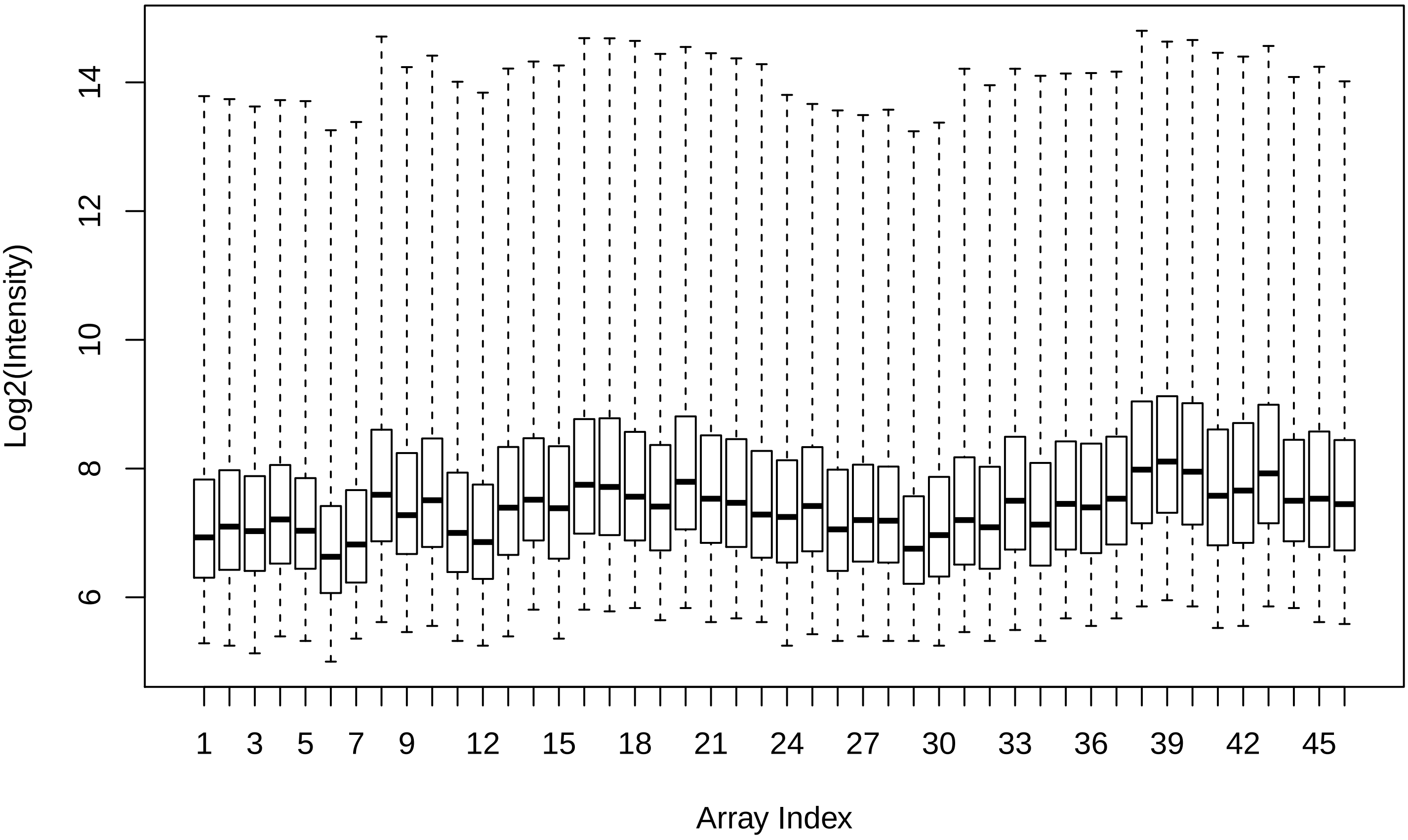
AffyBatch QC Report

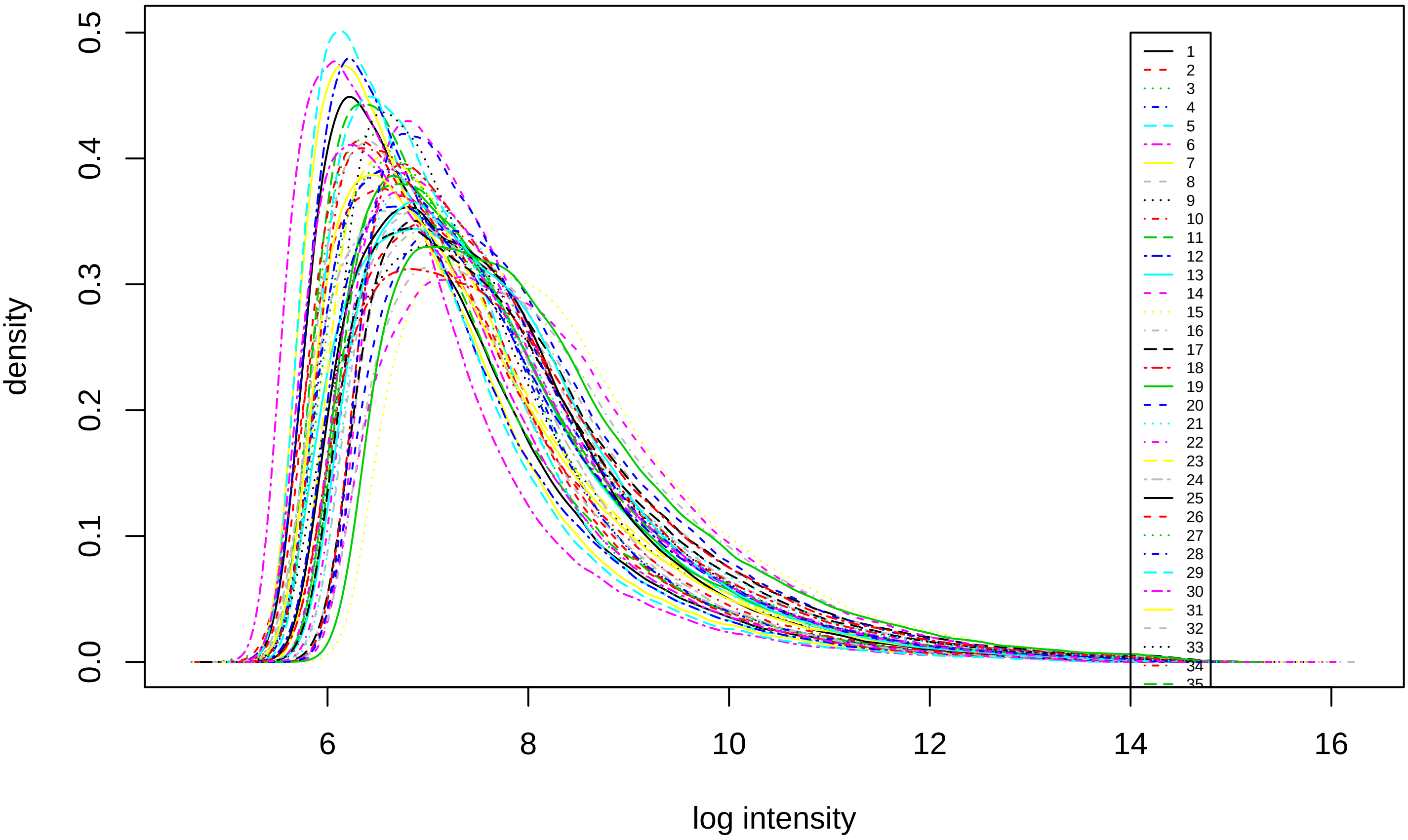
GSE36223

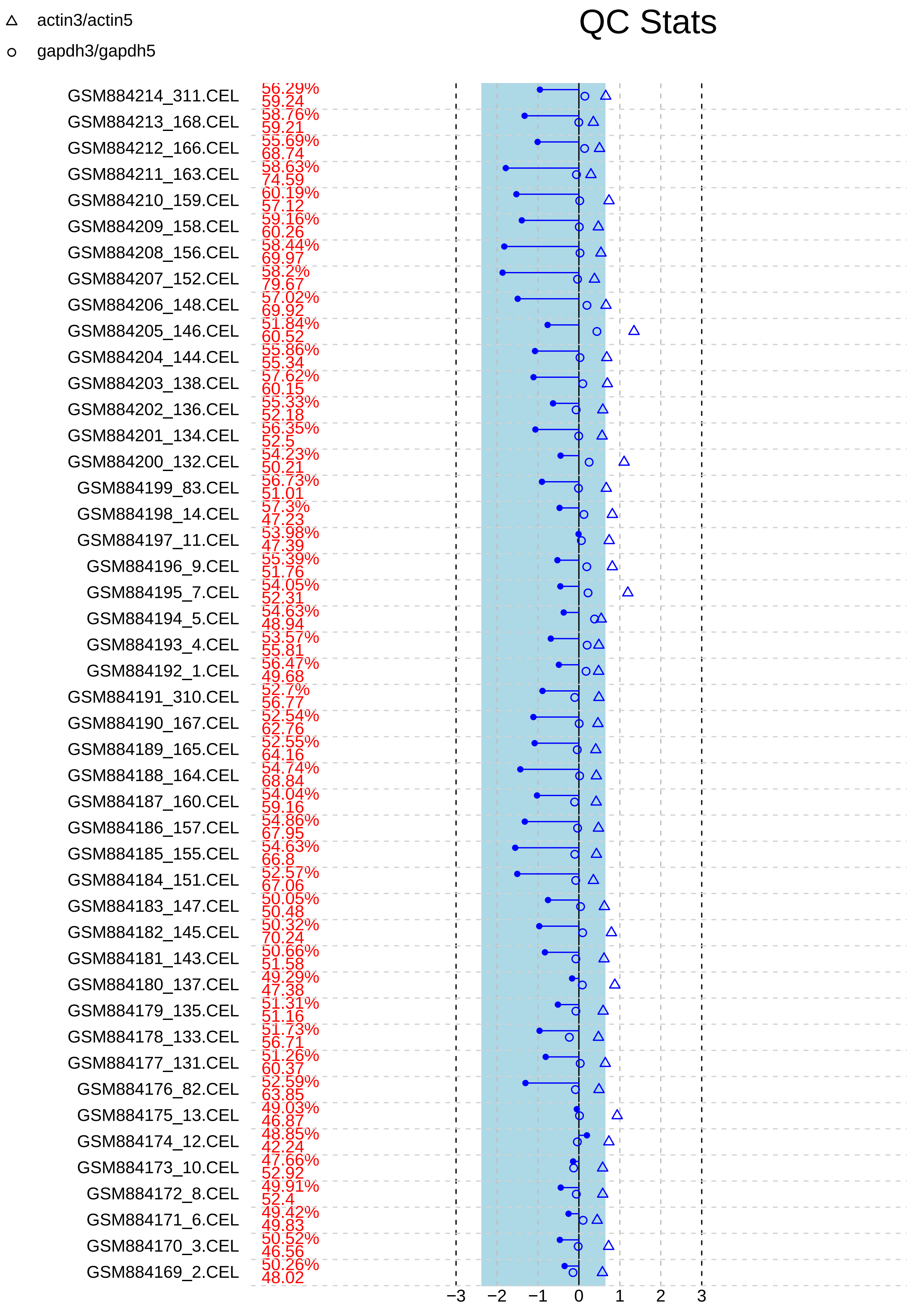
Array Index	Array Name
1	GSM88 169_2.CEL
2	GSM88 170_3.CEL
3	GSM88 171_6.CEL
	GSM88 172_8.CEL
5	GSM88 173_10.CEL
6	GSM88 17 _12.CEL
7	GSM88 175_13.CEL
8	GSM88 176_82.CEL
9	GSM88 177_131 CEL
10	GSM88 178_133 CEL
11	GSM88 179_135 CEL
12	GSM88 180_137 CEL
13	GSM88 181_1 3 CEL
1	GSM88 182_1 5 CEL
15	GSM88 183_1 7 CEL
16	GSM88 18 _151 CEL
17	GSM88 185_155 CEL
18	GSM88 186_157 CEL
19	GSM88 187_160 CEL
20	GSM88 188_16 CEL
21	GSM88 189_165 CEL
22	GSM88 190_167 CEL
23	GSM88 191_310 CEL
2	GSM88 192_1.CEL
25	GSM88 193_ .CEL
26	GSM88 19 _5.CEL
27	GSM88 195_7.CEL
28	GSM88 196_9.CEL
29	GSM88 197_11.CEL
30	GSM88 198_1 .CEL
31	GSM88 199_83.CEL
32	GSM88 200_132 CEL
33	GSM88 201_13 CEL
3	GSM88 202_136 CEL
35	GSM88 203_138 CEL
36	GSM88 20 _1 CEL
37	GSM88 205_1 6 CEL
38	GSM88 206_1 8 CEL
39	GSM88 207_152 CEL
0	GSM88 208_156 CEL
1	GSM88 209_158 CEL
2	GSM88 210_159 CEL
3	GSM88 211_163 CEL
	GSM88 212_166 CEL
5	GSM88 213_168 CEL
6	GSM88 21 _311 CEL

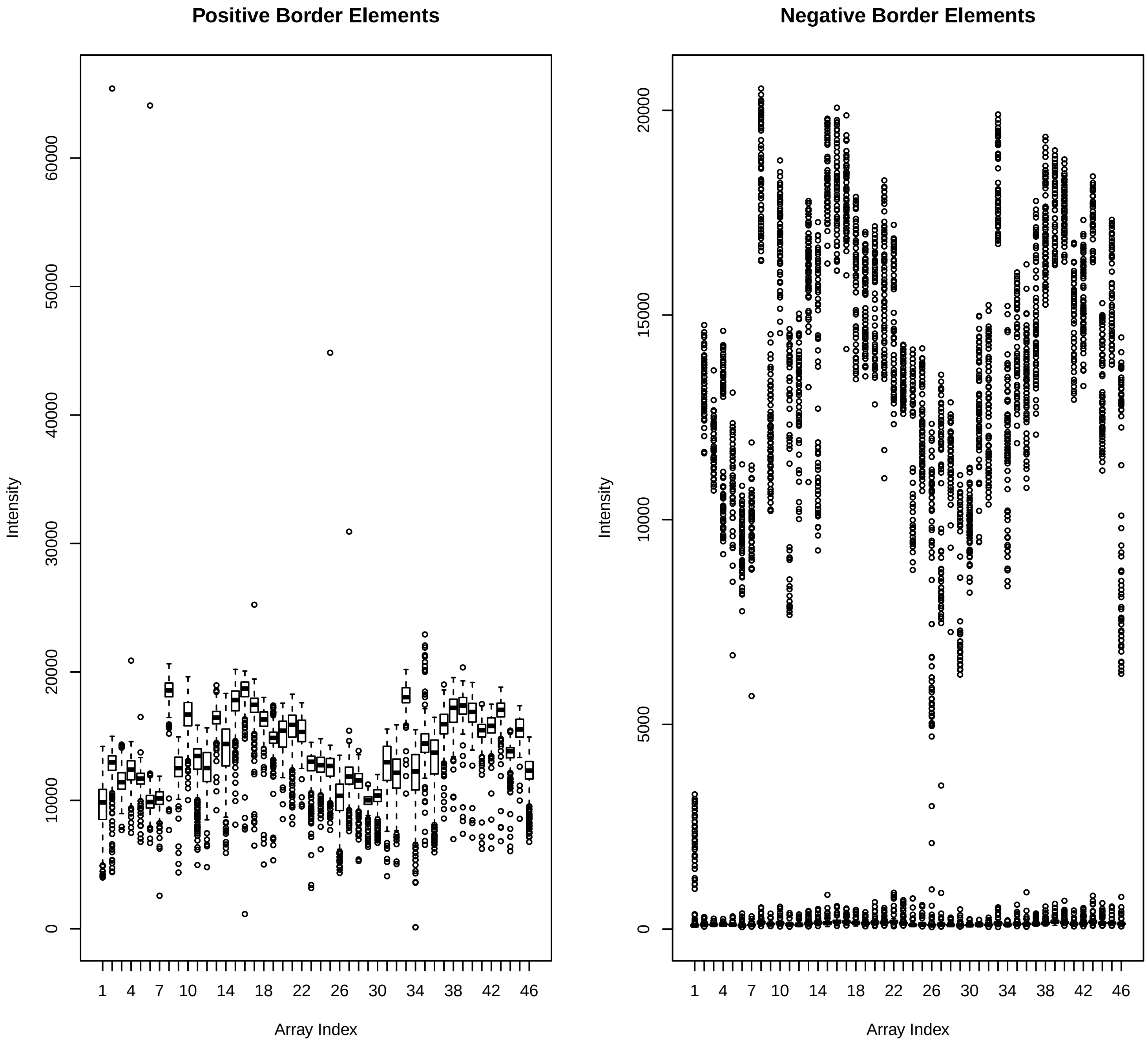
Thu Oct 1 17 07:34 2020

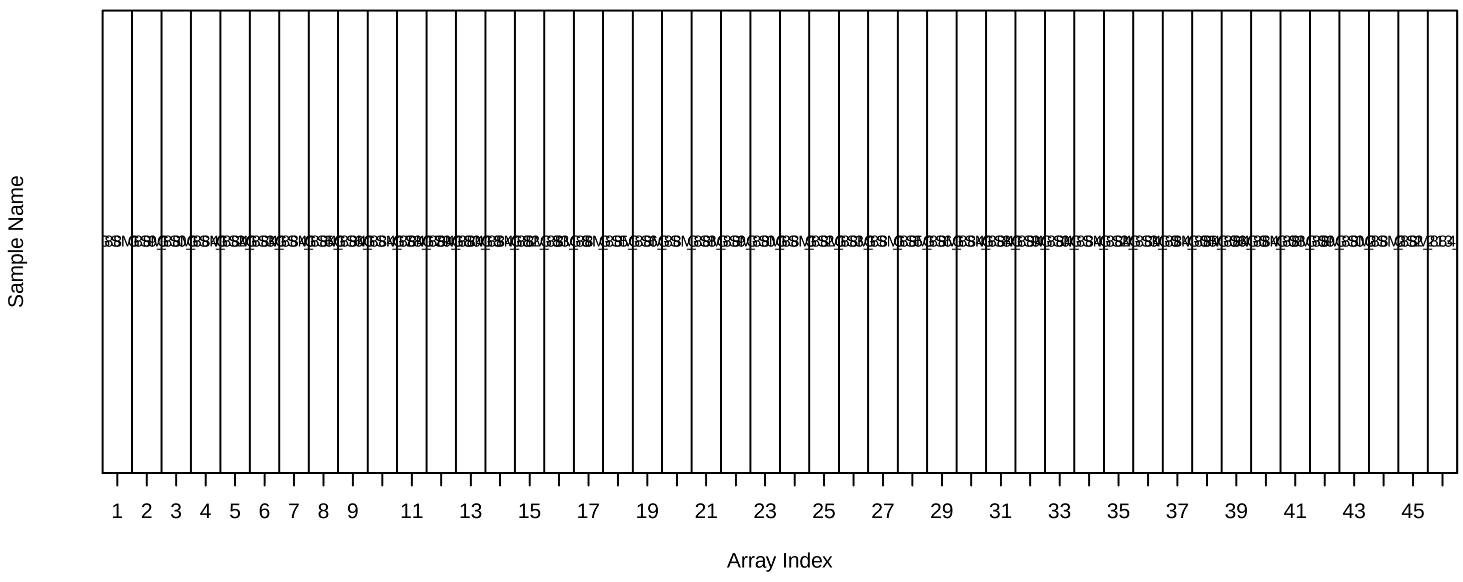
Produced by AffyQCReport R Package

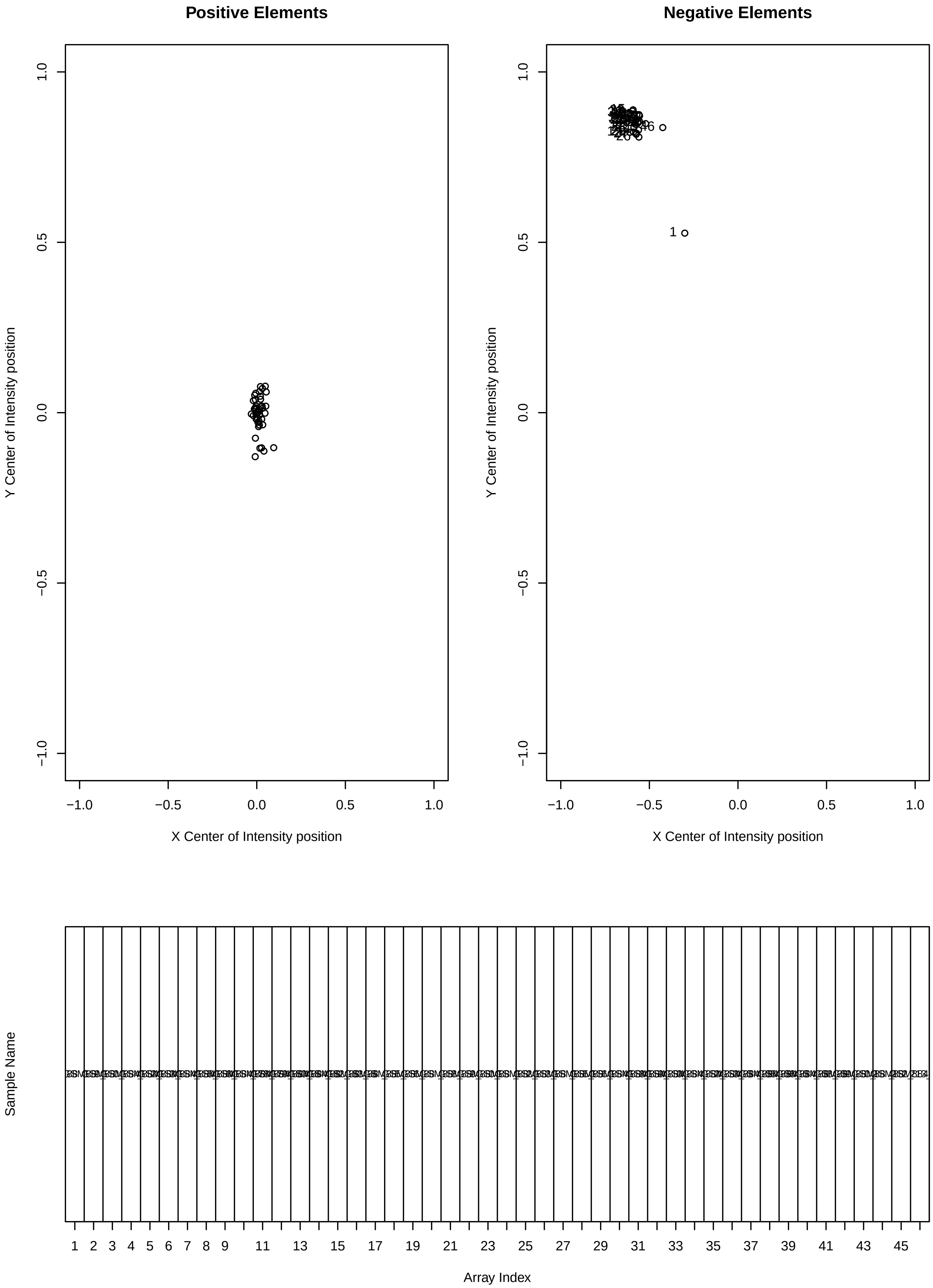

Log2(Intensity)
6
8
10
12
14
1
3
5
7
9
12
15
18
21
24
27
30
33
36
39
42
45
Array Index

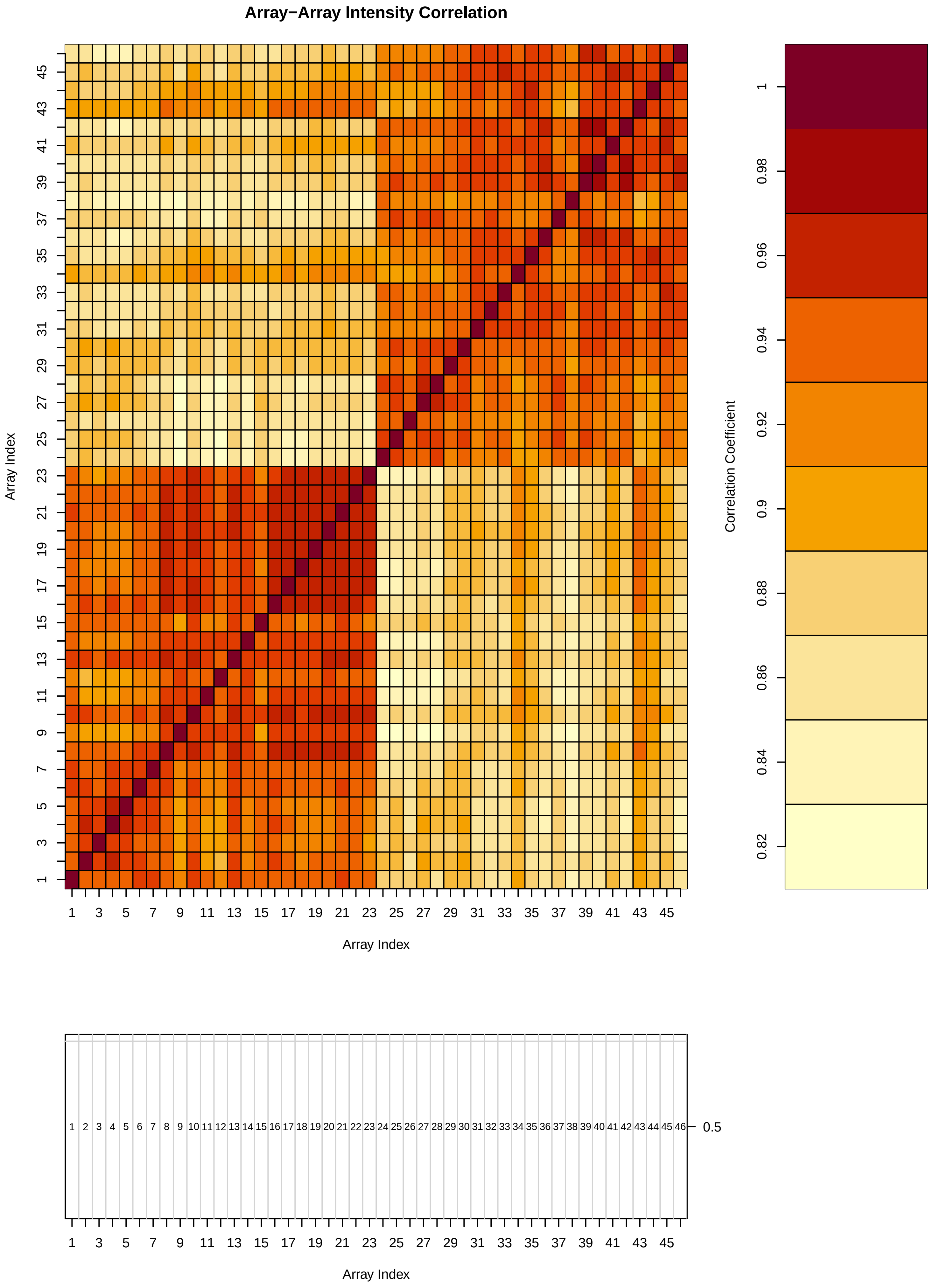

density
0.0
0.1
0.2
0.3
0.4
0.5
6
8
10
12
14
16
log intensity
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35


QC Stats
△ actin3/actin5
○ gapdh3/gapdh5
GSM884214_311.CEL 56.29% 59.24
GSM884213_168.CEL 58.76% 59.21
GSM884212_166.CEL 55.69% 68.74
GSM884211_163.CEL 58.63% 74.59
GSM884210_159.CEL 60.19% 57.12
GSM884209_158.CEL 59.16% 60.26
GSM884208_156.CEL 58.44% 69.97
GSM884207_152.CEL 58.2% 79.67
GSM884206_148.CEL 57.02% 69.92
GSM884205_146.CEL 51.84% 60.52
GSM884204_144.CEL 55.86% 55.34
GSM884203_138.CEL 57.62% 60.15
GSM884202_136.CEL 55.33% 52.18
GSM884201_134.CEL 56.35% 52.5
GSM884200_132.CEL 54.23% 50.21
GSM884199_83.CEL 56.73% 51.01
GSM884198_14.CEL 57.3% 47.23
GSM884197_11.CEL 53.98% 47.39
GSM884196_9.CEL 55.39% 51.76
GSM884195_7.CEL 54.05% 52.31
GSM884194_5.CEL 54.63% 48.94
GSM884193_4.CEL 53.57% 55.81
GSM884192_1.CEL 56.47% 49.68
GSM884191_310.CEL 52.7% 56.77
GSM884190_167.CEL 52.54% 62.76
GSM884189_165.CEL 52.55% 64.16
GSM884188_164.CEL 54.74% 68.84
GSM884187_160.CEL 54.04% 59.16
GSM884186_157.CEL 54.86% 67.95
GSM884185_155.CEL 54.63% 66.8
GSM884184_151.CEL 52.57% 67.06
GSM884183_147.CEL 50.05% 50.48
GSM884182_145.CEL 50.32% 70.24
GSM884181_143.CEL 50.66% 51.58
GSM884180_137.CEL 49.29% 47.38
GSM884179_135.CEL 51.31% 51.16
GSM884178_133.CEL 51.73% 56.71
GSM884177_131.CEL 51.26% 60.37
GSM884176_82.CEL 52.59% 63.85
GSM884175_13.CEL 49.03% 46.87
GSM884174_12.CEL 48.85% 42.24
GSM884173_10.CEL 47.66% 52.92
GSM884172_8.CEL 49.91% 52.4
GSM884171_6.CEL 49.42% 49.83
GSM884170_3.CEL 50.52% 46.56
GSM884169_2.CEL 50.26% 48.02
−3
−2
−1
0
1
2
3


Positive Border Elements
Intensity
0
10000
20000
30000
40000
50000
60000
1
4
7
10
14
18
22
26
30
34
38
42
46
Array Index
Negative Border Elements
Intensity
0
5000
10000
15000
20000
1
4
7
10
14
18
22
26
30
34
38
42
46
Array Index


Sample Name
1
2
3
4
5
6
7
8
9
11
13
15
17
19
21
23
25
27
29
31
33
35
37
39
41
43
45
Array Index


Positive Elements
Negative Elements
Y Center of Intensity position
X Center of Intensity position
−1.0
−0.5
0.0
0.5
1.0
1
6
26
Sample Name
Array Index
1 2 3 4 5 6 7 8 9 11 13 15 17 19 21 23 25 27 29 31 33 35 37 39 41 43 45


Array−Array Intensity Correlation
Array Index
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31 33 35 37 39 41 43 45
Array Index
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31 33 35 37 39 41 43 45
Correlation Coefficient
0.82 0.84 0.86 0.88 0.9 0.92 0.94 0.96 0.98 1
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46
0.5
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31 33 35 37 39 41 43 45
Array Index

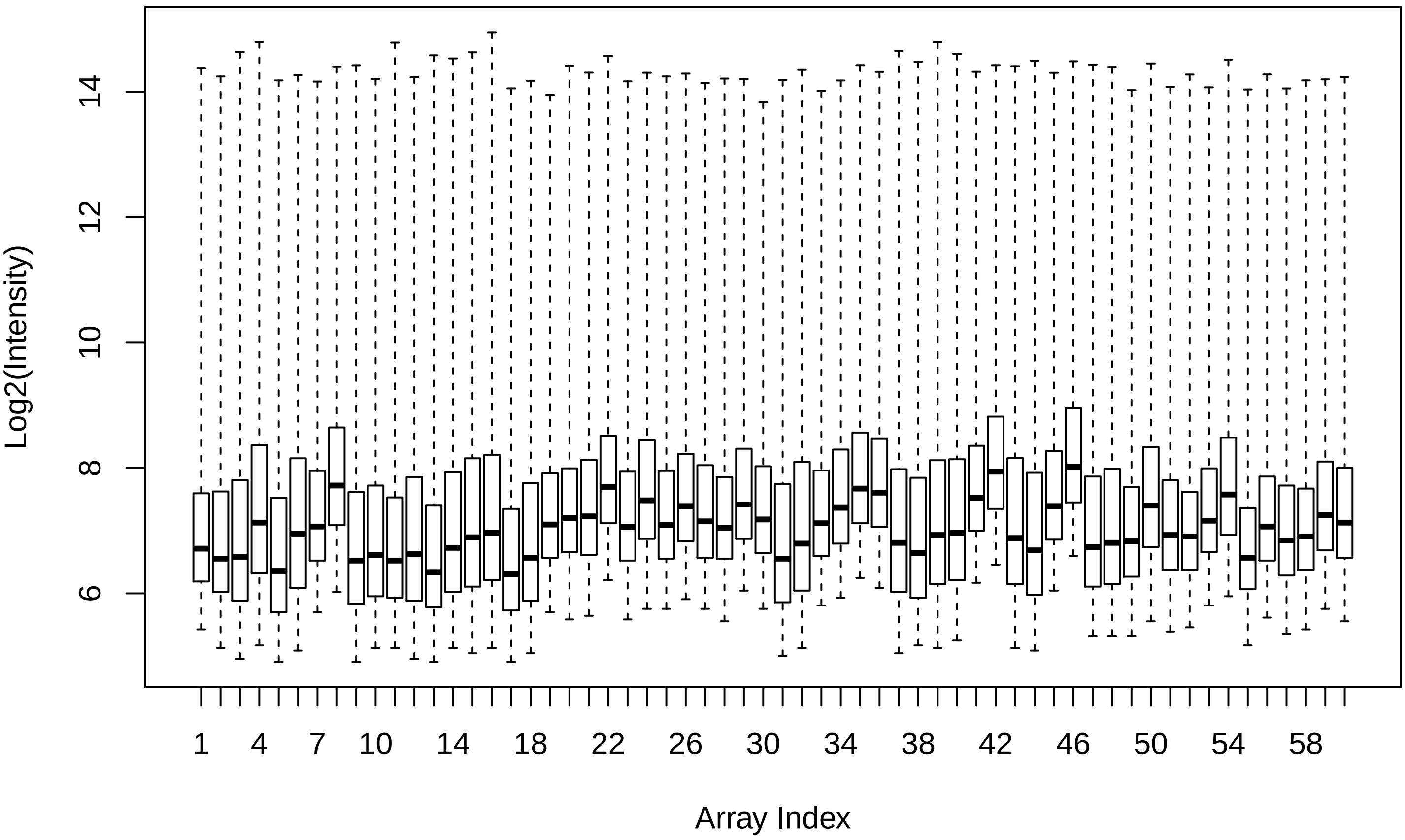
AffyBatch QC Report

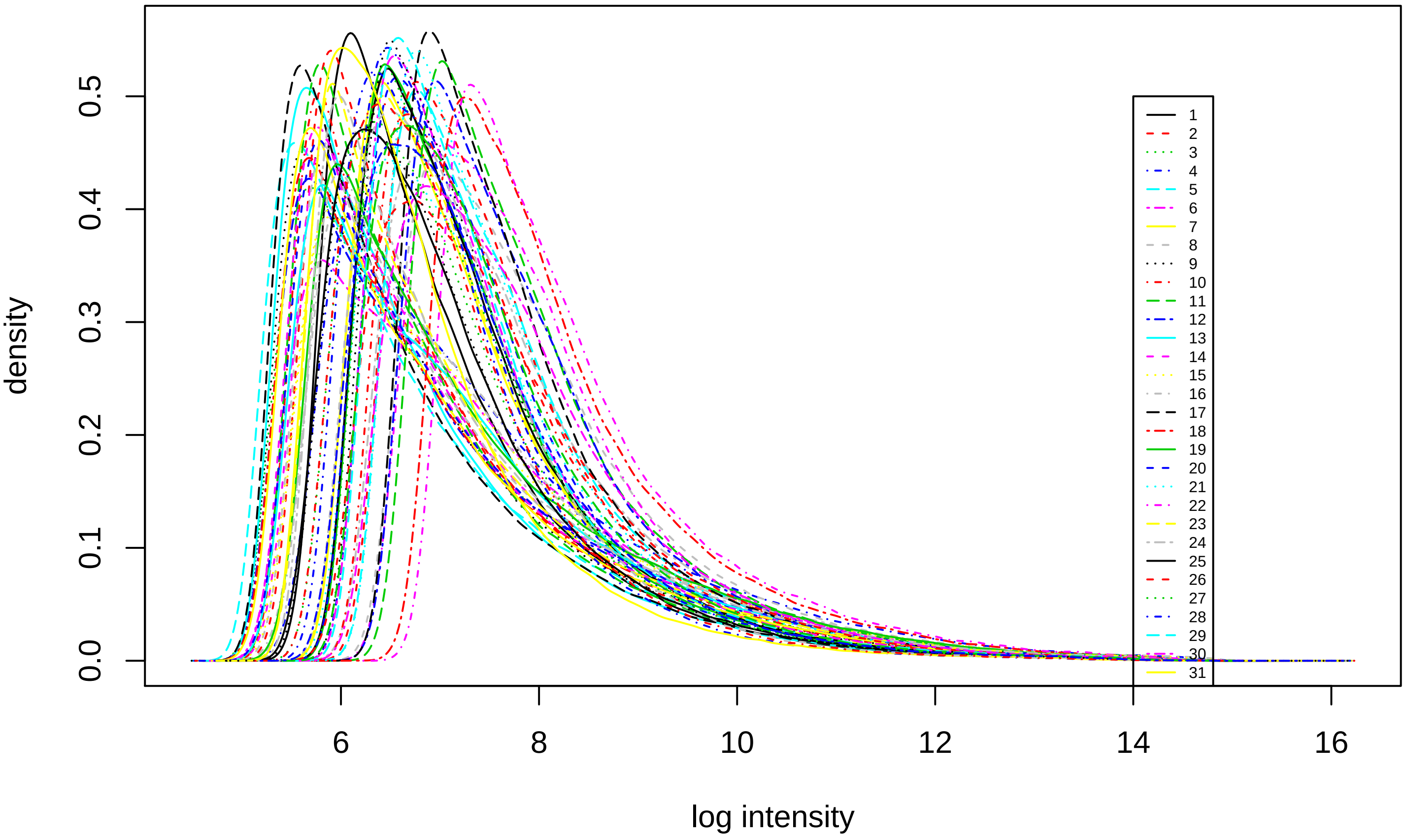
GSE38129

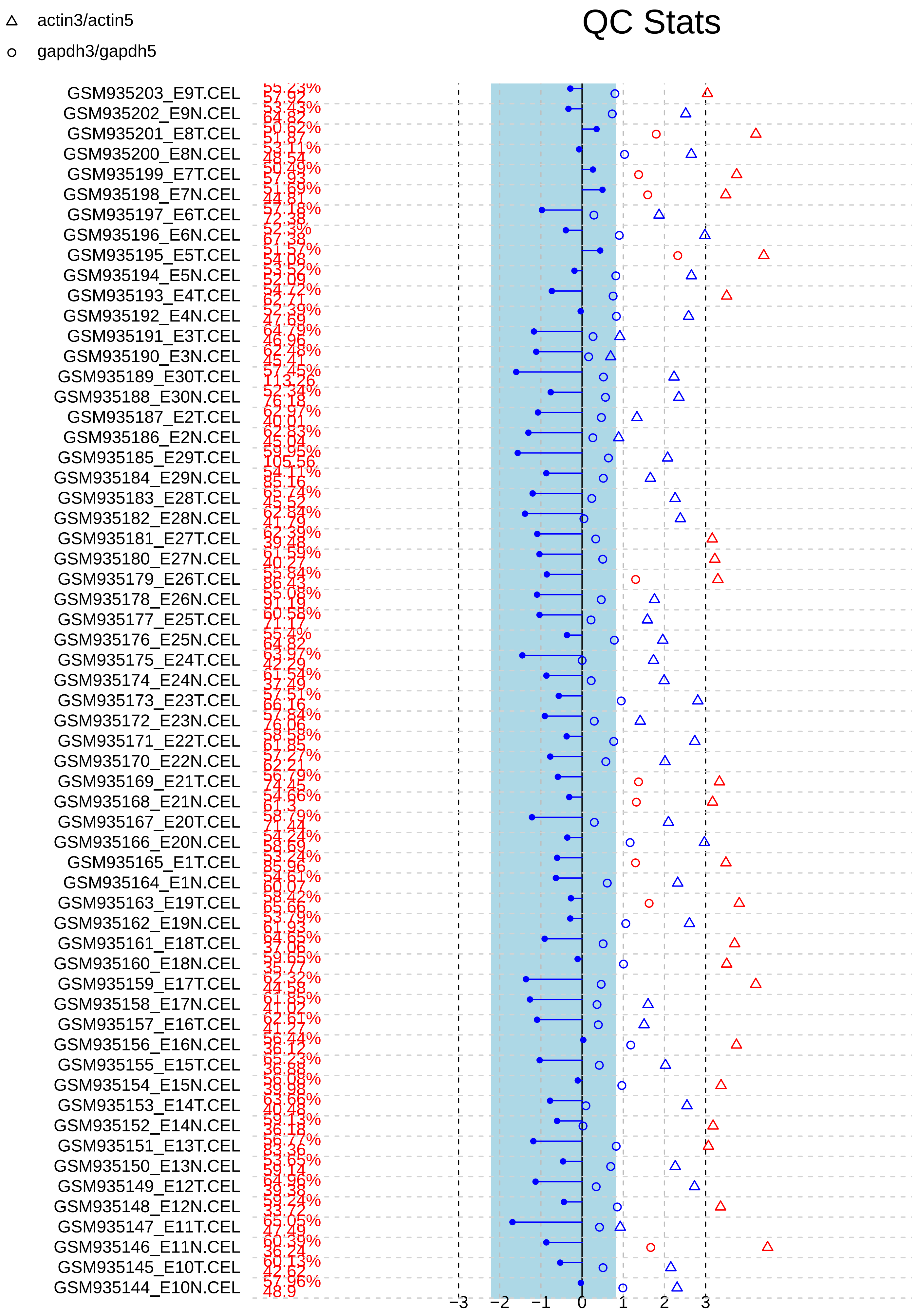
Array Index	Array Name
	GSM935 44_E 0N CE
3	GSM935 46_E N CE
4	GSM935 47_E CE
5	GSM935 48_E 2N CE
6	GSM935 49_E 2 CE
7	GSM935 50_E 3N CE
8	GSM935 5 _E 3 CE
9	GSM935 52_E 4N CE
0	GSM935 53_E 4 CE
	GSM935 54_E 5N CE
2	GSM935 55_E 5 CE
3	GSM935 56_E 6N CE
4	GSM935 57_E 6 CE
5	GSM935 58_E 7N CE
6	GSM935 59_E 7 CE
7	GSM935 60_E 8N CE
8	GSM935 6 _E 8 CE
9	GSM935 62_E 9N CE
20	GSM935 63_E 9 CE
2	GSM935 64_E N CE
22	GSM935 65_E CE
23	GSM935 66_E20N CE
24	GSM935 67_E20 CE
25	GSM935 68_E2 N CE
26	GSM935 69_E2 CE
27	GSM935 70_E22N CE
28	GSM935 7 _E22 CE
29	GSM935 72_E23N CE
30	GSM935 73_E23 CE
3	GSM935 74_E24N CE
32	GSM935 75_E24 CE
33	GSM935 76_E25N CE
34	GSM935 77_E25 CE
35	GSM935 78_E26N CE
36	GSM935 79_E26 CE
37	GSM935 80_E27N CE
38	GSM935 8 _E27 CE
39	GSM935 82_E28N CE
40	GSM935 83_E28 CE
4	GSM935 84_E29N CE
42	GSM935 85_E29 CE
43	GSM935 86_E2N CE
44	GSM935 87_E2 CE
45	GSM935 88_E30N CE
46	GSM935 89_E30 CE
47	GSM935 90_E3N CE
48	GSM935 9 _E3 CE
49	GSM935 92_E4N CE
50	GSM935 93_E4 CE
5	GSM935 94_E5N CE
52	GSM935 95_E5 CE
53	GSM935 96_E6N CE
54	GSM935 97_E6 CE
55	GSM935 98_E7N CE
56	GSM935 99_E7 CE
57	GSM935200_E8N CE
58	GSM93520 _E8 CE
59	GSM935202_E9N CE
60	GSM935203_E9 CE

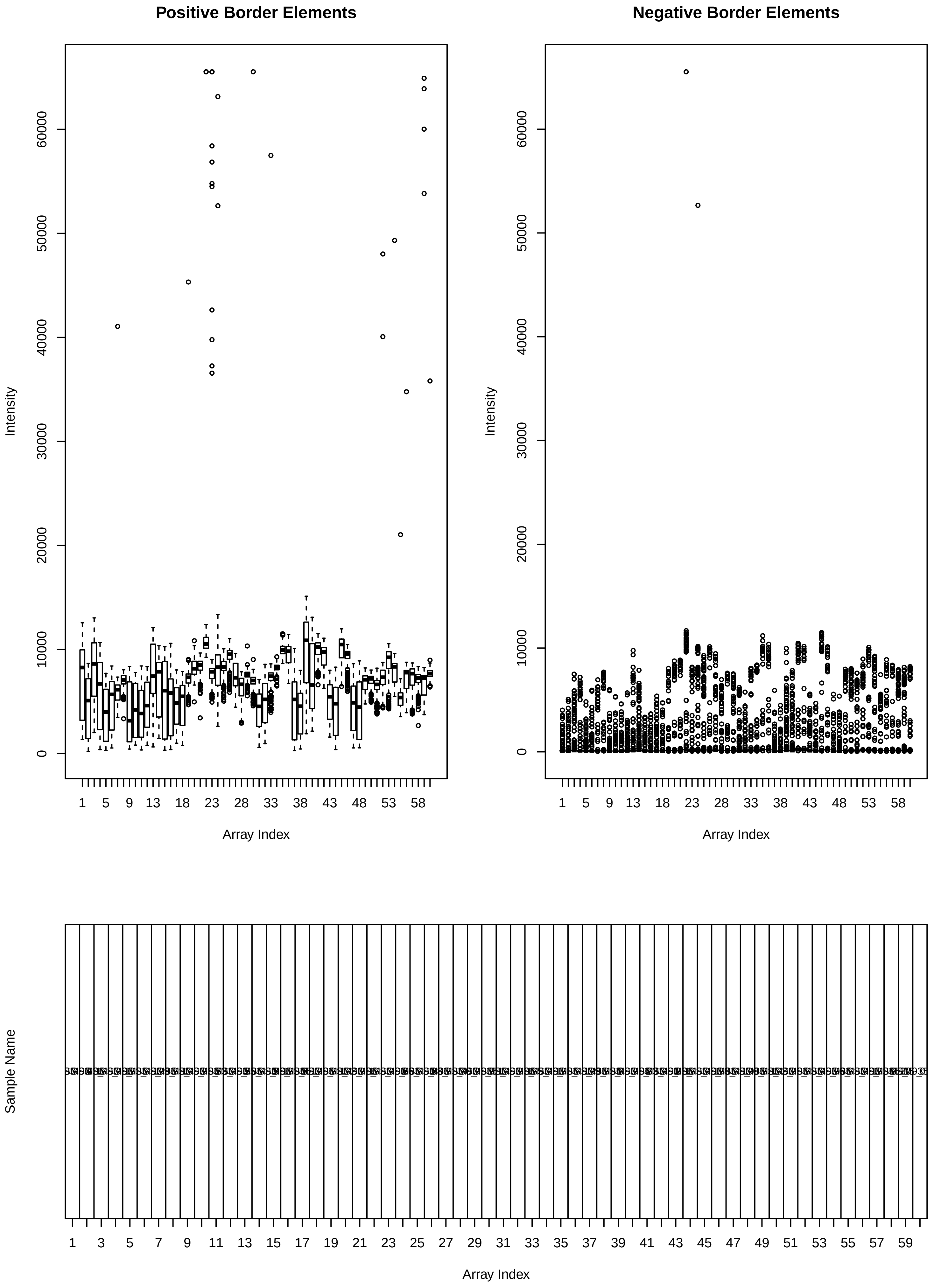
Thu Oct 1 17 26:56 2020

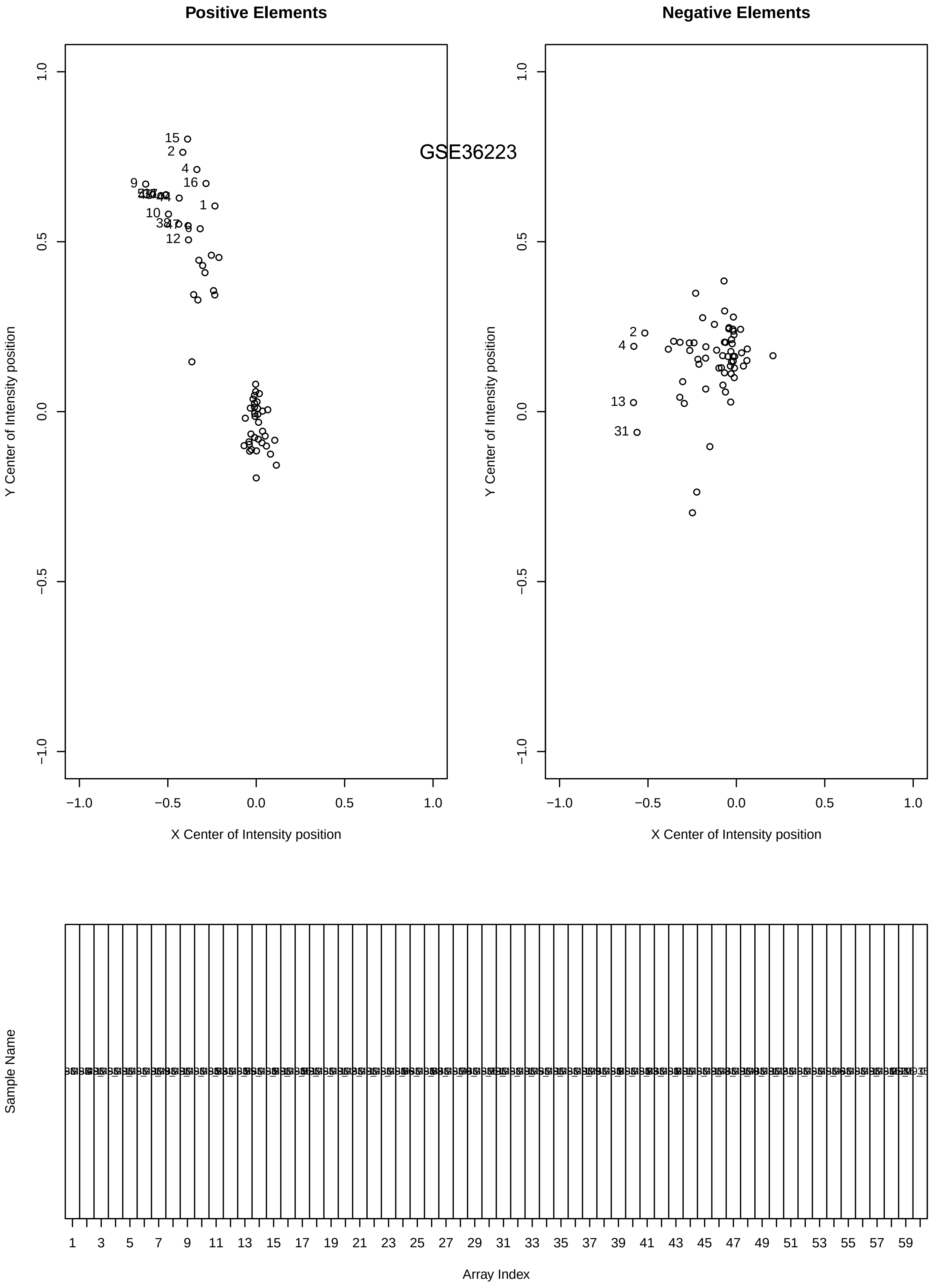
Produced by AffyQCReport R Package


Log2(Intensity)
6
8
10
12
14
1
4
7
10
14
18
22
26
30
34
38
42
46
50
54
58
Array Index

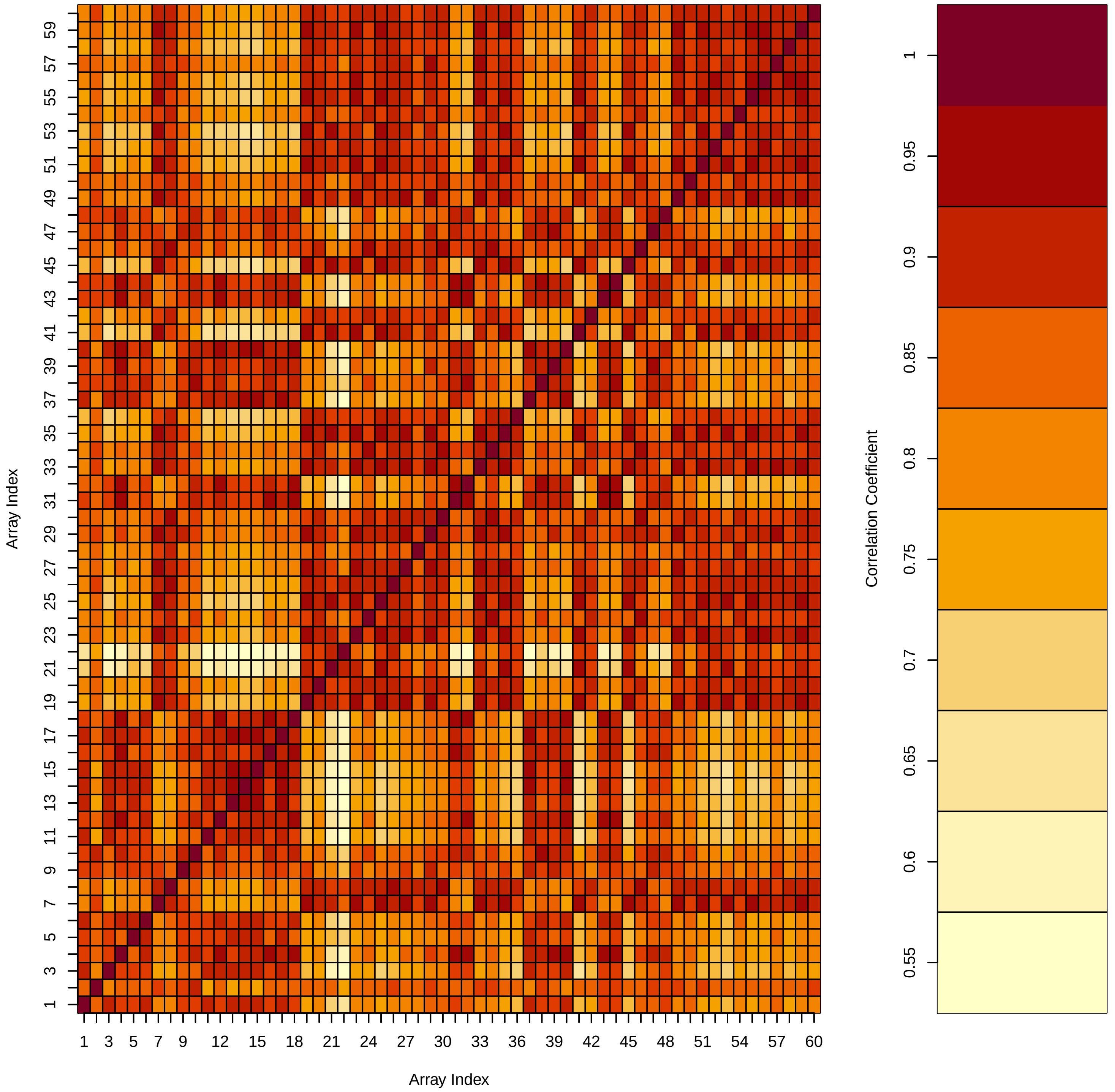

density
0.0
0.1
0.2
0.3
0.4
0.5
6
8
10
12
14
16
log intensity
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31

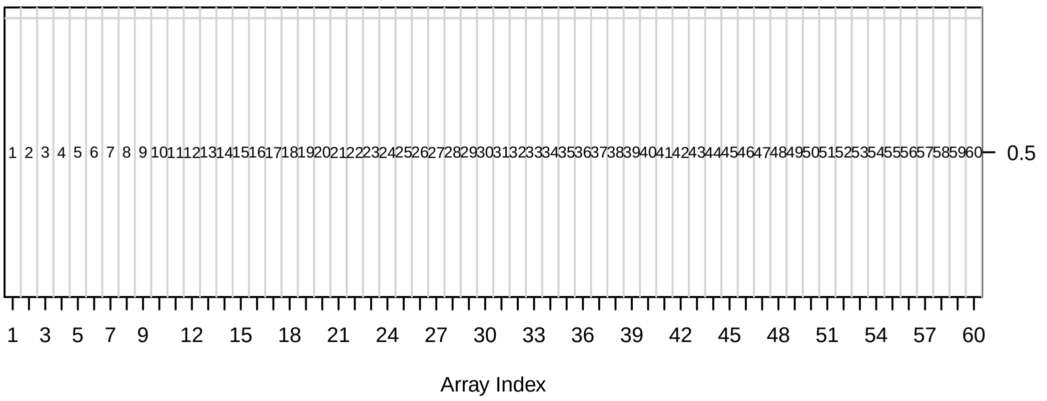

QC Stats
△ actin3/actin5
○ gapdh3/gapdh5
GSM935203_E9T.CEL
55.23%
57.92
GSM935202_E9N.CEL
53.43%
64.82
GSM935201_E8T.CEL
50.62%
51.87
GSM935200_E8N.CEL
53.11%
48.54
GSM935199_E7T.CEL
50.49%
57.93
GSM935198_E7N.CEL
51.69%
44.81
GSM935197_E6T.CEL
57.18%
72.38
GSM935196_E6N.CEL
52.3%
67.38
GSM935195_E5T.CEL
51.57%
54.08
GSM935194_E5N.CEL
53.52%
52.09
GSM935193_E4T.CEL
54.72%
62.71
GSM935192_E4N.CEL
52.39%
47.69
GSM935191_E3T.CEL
64.79%
46.96
GSM935190_E3N.CEL
62.48%
45.41
GSM935189_E30T.CEL
57.45%
113.26
GSM935188_E30N.CEL
52.34%
76.18
GSM935187_E2T.CEL
62.97%
40.01
GSM935186_E2N.CEL
62.83%
45.04
GSM935185_E29T.CEL
59.95%
105.56
GSM935184_E29N.CEL
54.11%
85.16
GSM935183_E28T.CEL
65.74%
45.52
GSM935182_E28N.CEL
62.84%
41.79
GSM935181_E27T.CEL
62.39%
39.48
GSM935180_E27N.CEL
61.59%
40.27
GSM935179_E26T.CEL
55.64%
86.43
GSM935178_E26N.CEL
55.08%
91.19
GSM935177_E25T.CEL
60.58%
71.17
GSM935176_E25N.CEL
55.4%
64.82
GSM935175_E24T.CEL
63.97%
42.29
GSM935174_E24N.CEL
61.54%
37.49
GSM935173_E23T.CEL
57.51%
66.16
GSM935172_E23N.CEL
57.84%
76.06
GSM935171_E22T.CEL
58.58%
61.85
GSM935170_E22N.CEL
57.27%
62.21
GSM935169_E21T.CEL
56.79%
74.45
GSM935168_E21N.CEL
54.66%
61.3
GSM935167_E20T.CEL
58.79%
71.44
GSM935166_E20N.CEL
54.24%
58.69
GSM935165_E1T.CEL
53.24%
85.96
GSM935164_E1N.CEL
54.61%
60.07
GSM935163_E19T.CEL
58.42%
65.66
GSM935162_E19N.CEL
53.79%
61.93
GSM935161_E18T.CEL
64.65%
37.06
GSM935160_E18N.CEL
59.65%
35.77
GSM935159_E17T.CEL
62.32%
44.58
GSM935158_E17N.CEL
61.85%
41.02
GSM935157_E16T.CEL
62.61%
41.27
GSM935156_E16N.CEL
56.44%
36.12
GSM935155_E15T.CEL
65.23%
36.88
GSM935154_E15N.CEL
56.08%
39.98
GSM935153_E14T.CEL
63.66%
40.48
GSM935152_E14N.CEL
59.13%
36.18
GSM935151_E13T.CEL
55.77%
83.36
GSM935150_E13N.CEL
53.65%
59.14
GSM935149_E12T.CEL
64.96%
39.38
GSM935148_E12N.CEL
59.24%
33.72
GSM935147_E11T.CEL
65.05%
47.49
GSM935146_E11N.CEL
60.39%
36.24
GSM935145_E10T.CEL
60.13%
42.62
GSM935144_E10N.CEL
57.96%
48.9
−3
−2
−1
0
1
2
3


Positive Border Elements
Negative Border Elements
Intensity
0
10000
20000
30000
40000
50000
60000
Array Index
1
5
9
13
18
23
28
33
38
43
48
53
58
Sample Name
Array Index
1
3
5
7
9
11
13
15
17
19
21
23
25
27
29
31
33
35
37
39
41
43
45
47
49
51
53
55
57
59


GSE36223
Positive Elements
Negative Elements
Y Center of Intensity position
X Center of Intensity position
1.0
0.5
0.0
−0.5
−1.0
15
2
4
9
16
1
10
38
47
6
12
2
4
13
31
Sample Name
Array Index
1
3
5
7
9
11
13
15
17
19
21
23
25
27
29
31
33
35
37
39
41
43
45
47
49
51
53
55
57
59

Array−Array Intensity Correlation





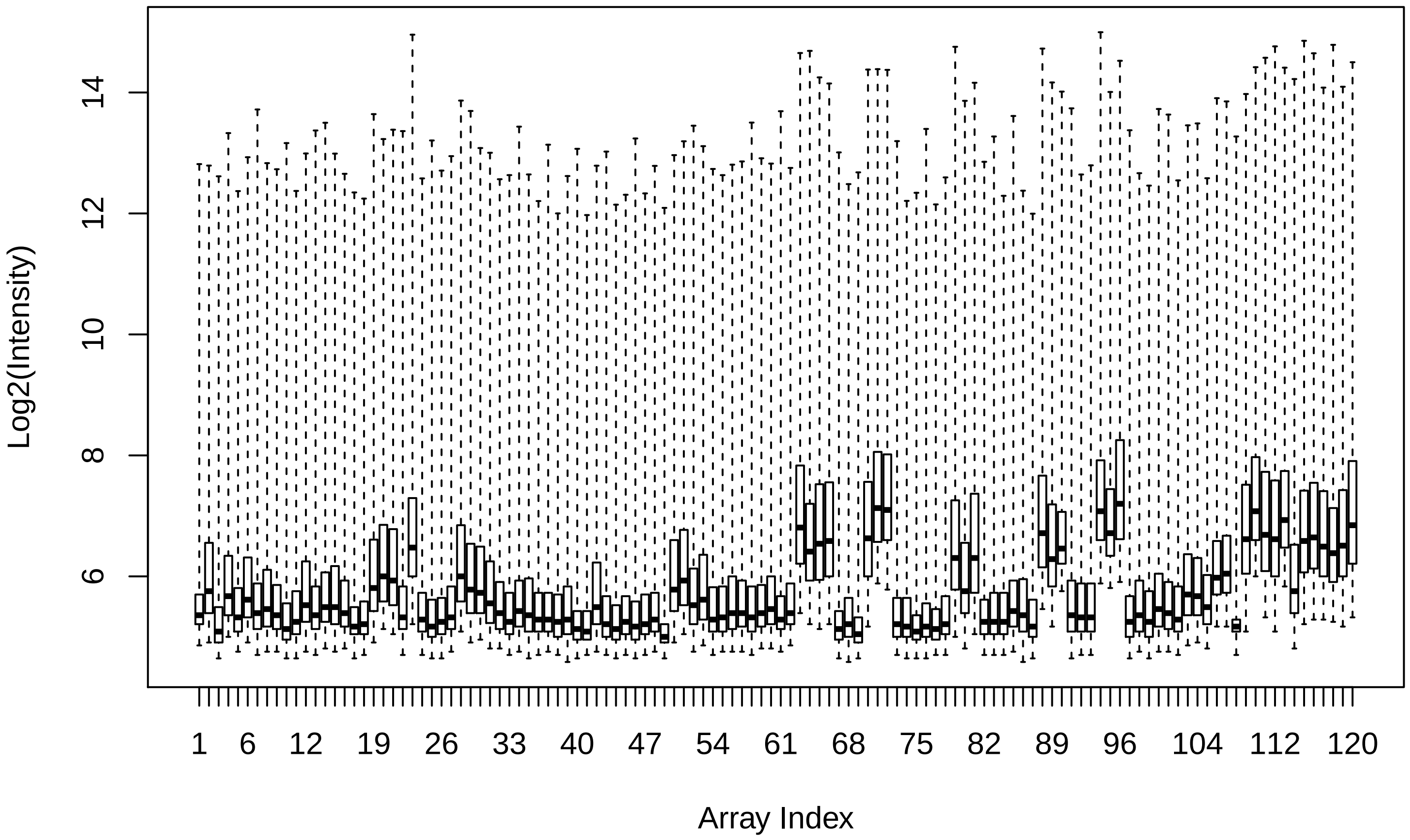
AffyBatch QC Report

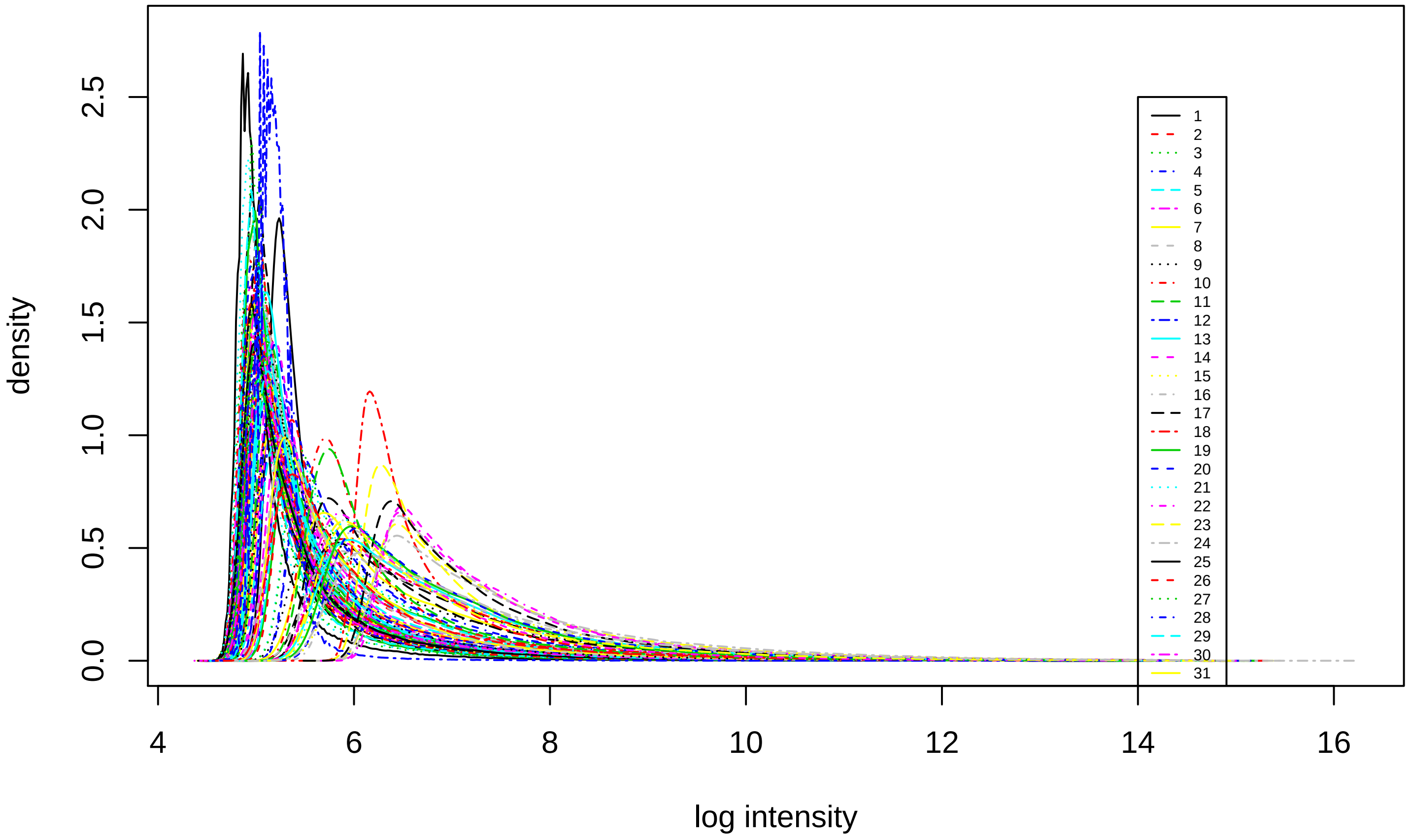
GSE39491

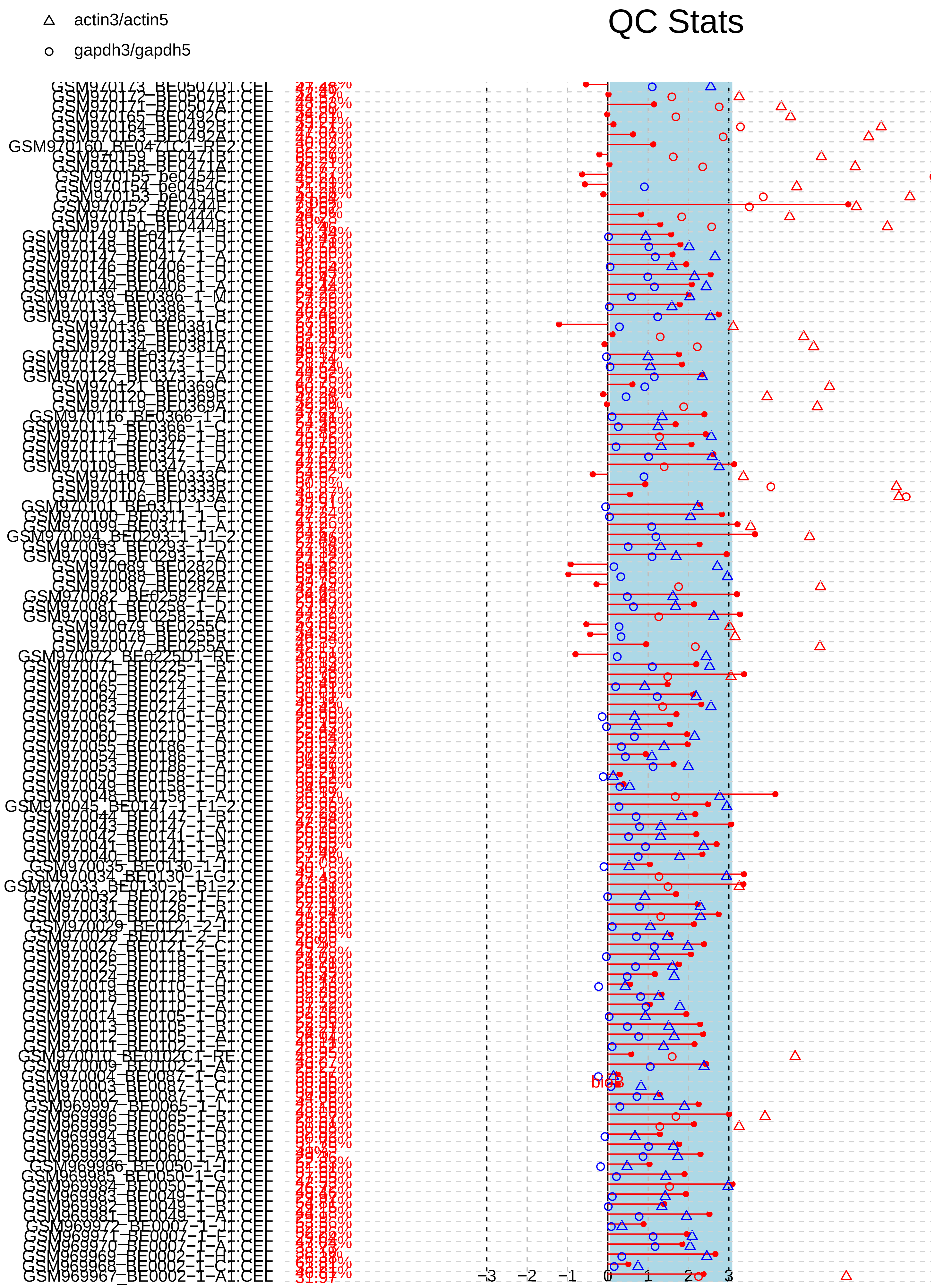
Array Index	Array Name

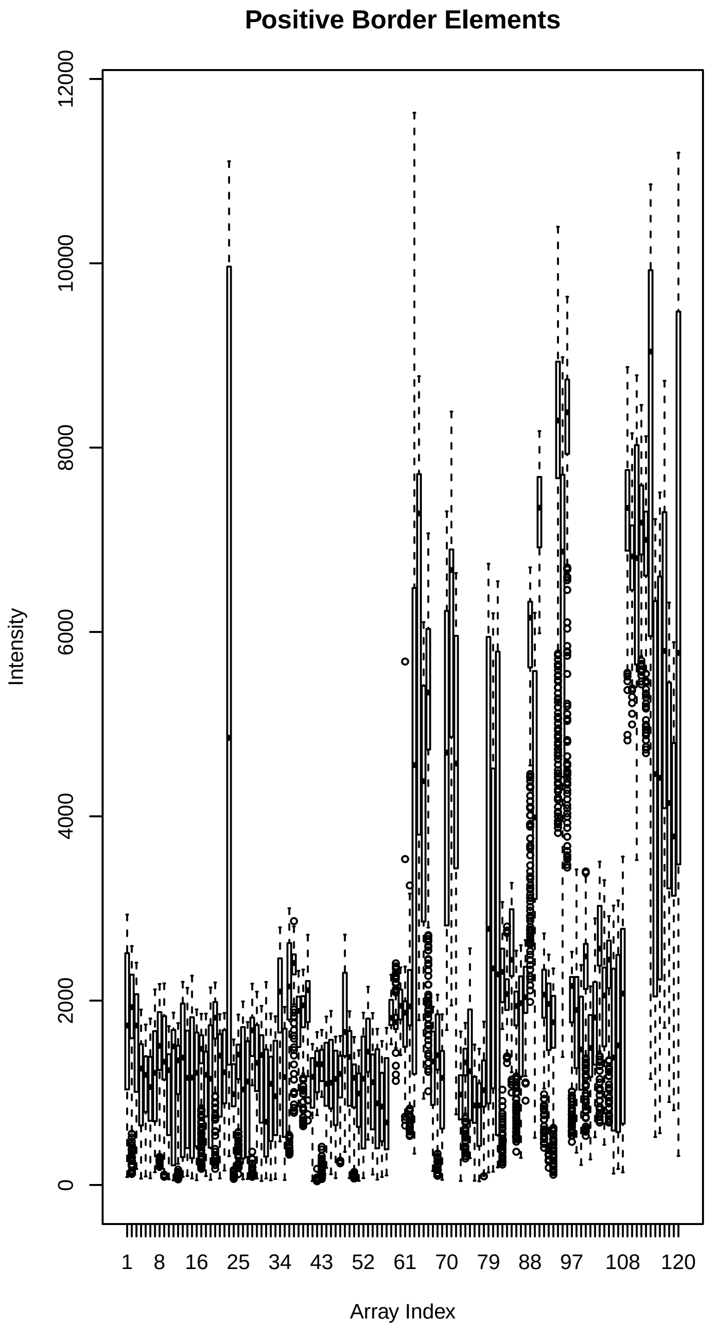
Thu Oct 1 17:44:07 2020

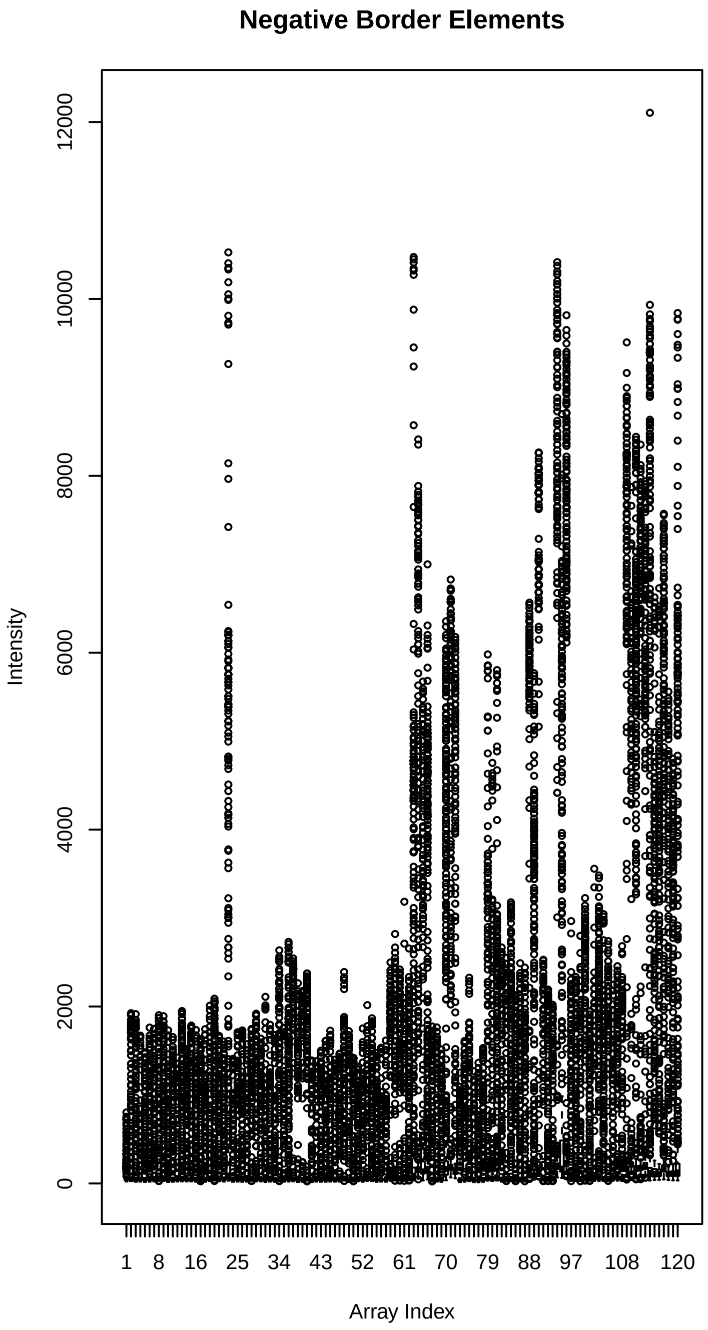
Produced by AffyQCReport R Package

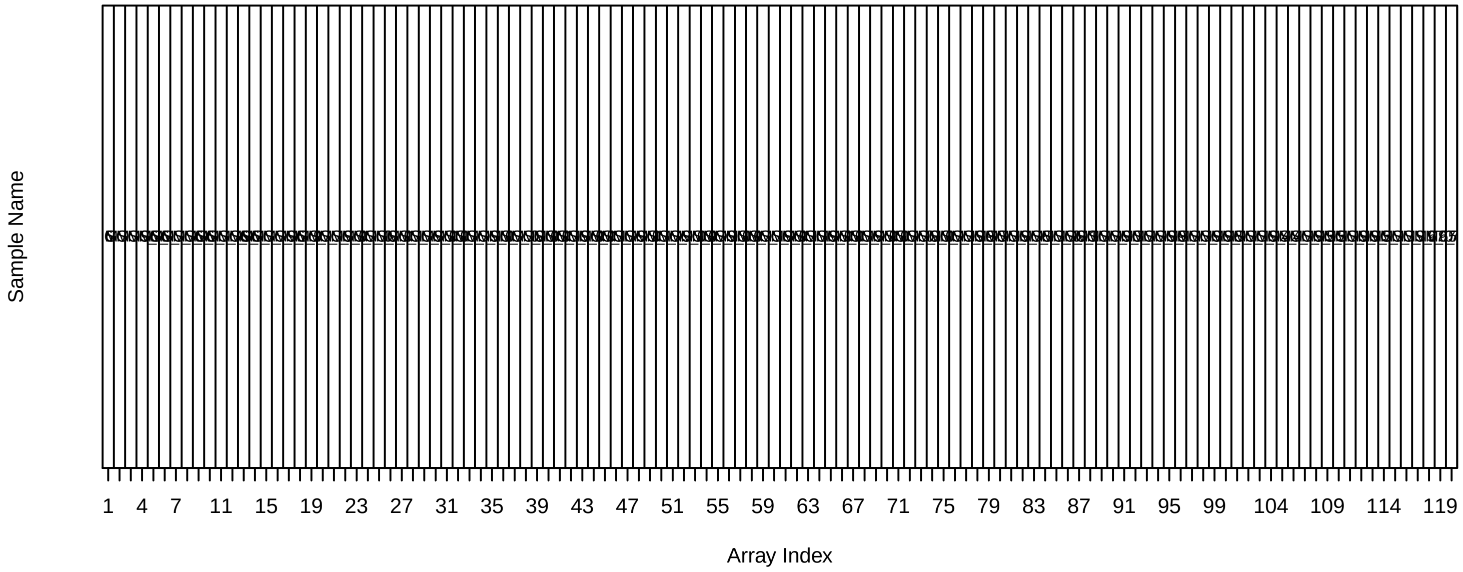

Log2(Intensity)
6
8
10
12
14
1
6
12
19
26
33
40
47
54
61
68
75
82
89
96
104
112
120
Array Index

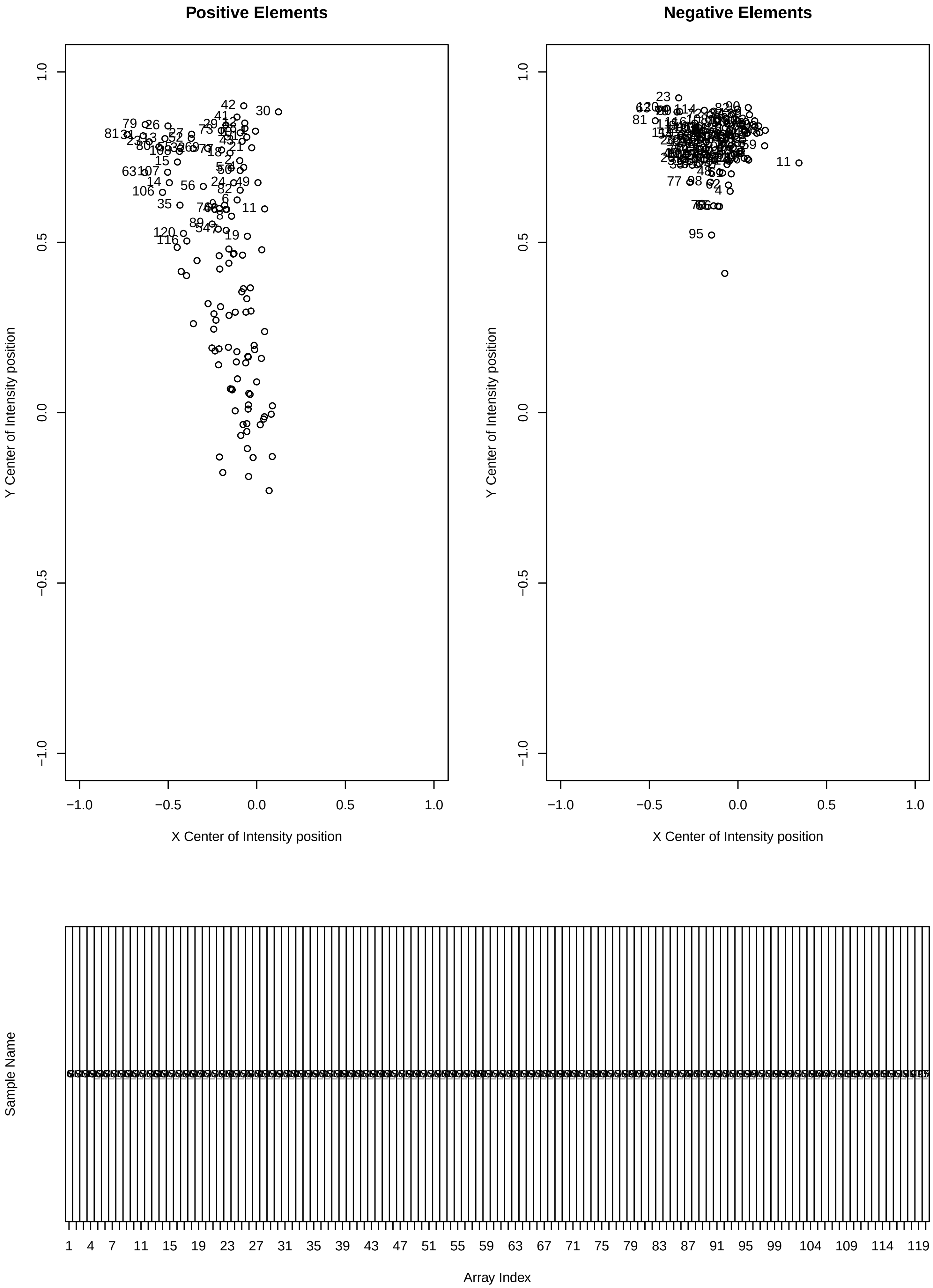

density
0.0
0.5
1.0
1.5
2.0
2.5
4
6
8
10
12
14
16
log intensity
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31

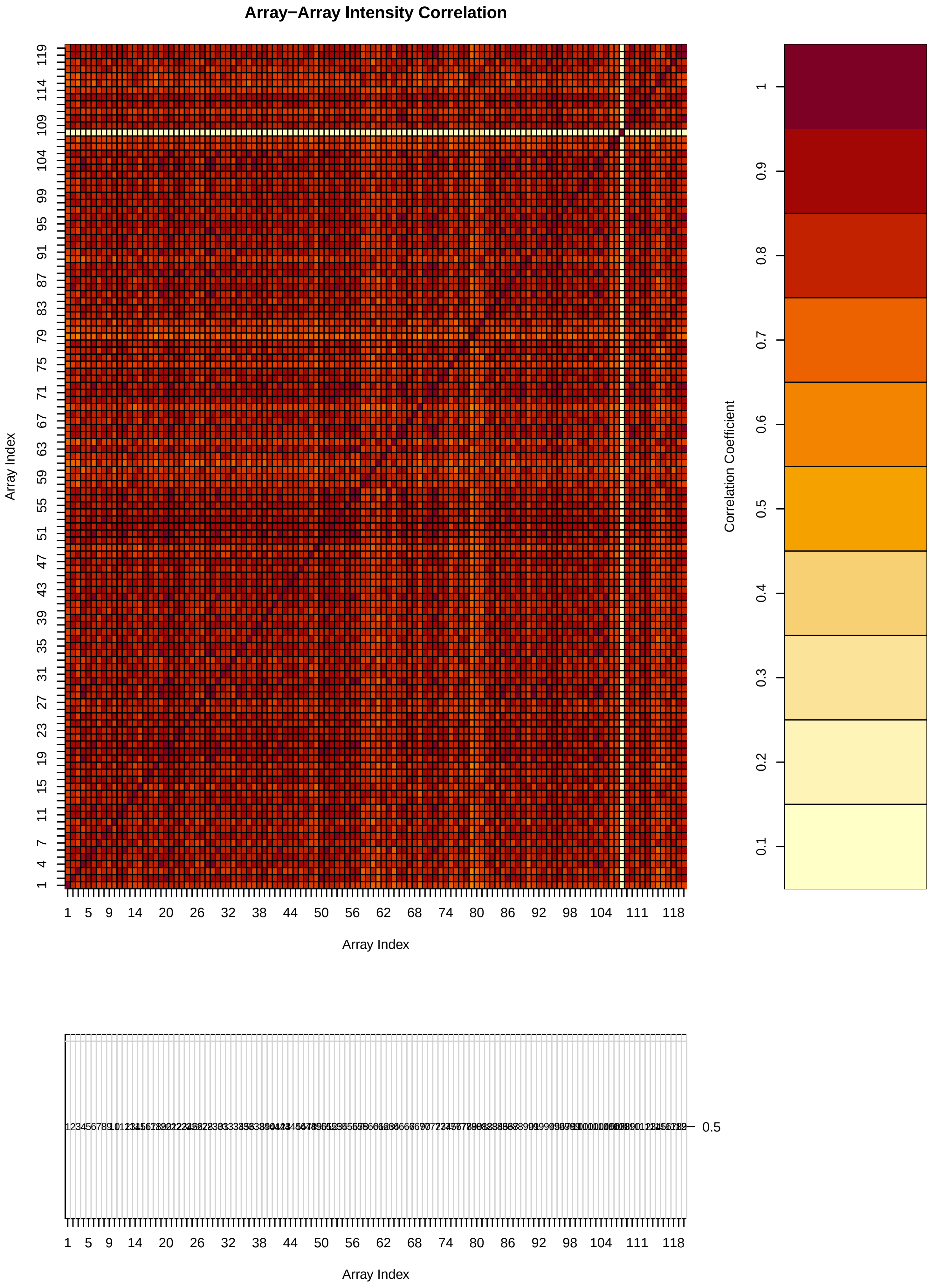

QC Stats
△ actin3/actin5
○ gapdh3/gapdh5
GSM970173_BE0507D1.CEL
GSM970172_BE0507B1.CEL
GSM970171_BE0507A1.CEL
GSM970165_BE0492C1.CEL
GSM970164_BE0492B1.CEL
GSM970163_BE0492A1.CEL
GSM970160_BE0471C1-RE2.CEL
GSM970159_BE0471B1.CEL
GSM970158_BE0471A1.CEL
GSM970155_be0454E1.CEL
GSM970154_be0454C1.CEL
GSM970153_be0454B1.CEL
GSM970152_BE0444D1.CEL
GSM970151_BE0444C1.CEL
GSM970150_BE0444B1.CEL
GSM970149_BE0417-1-H1.CEL
GSM970148_BE0417-1-D1.CEL
GSM970147_BE0417-1-A1.CEL
GSM970146_BE0406-1-H1.CEL
GSM970145_BE0406-1-D1.CEL
GSM970144_BE0406-1-A1.CEL
GSM970139_BE0386-1-M1.CEL
GSM970138_BE0386-1-C1.CEL
GSM970137_BE0386-1-B1.CEL
GSM970136_BE0381C1.CEL
GSM970135_BE0381B1.CEL
GSM970134_BE0381A1.CEL
GSM970129_BE0373-1-H1.CEL
GSM970128_BE0373-1-D1.CEL
GSM970127_BE0373-1-A1.CEL
GSM970121_BE0369C1.CEL
GSM970120_BE0369B1.CEL
GSM970119_BE0369A1.CEL
GSM970116_BE0366-1-I1.CEL
GSM970115_BE0366-1-C1.CEL
GSM970114_BE0366-1-B1.CEL
GSM970111_BE0347-1-H1.CEL
GSM970110_BE0347-1-D1.CEL
GSM970109_BE0347-1-A1.CEL
GSM970108_BE0333C1.CEL
GSM970107_BE0333B1.CEL
GSM970106_BE0333A1.CEL
GSM970101_BE0311-1-G1.CEL
GSM970100_BE0311-1-F1.CEL
GSM970099_BE0311-1-A1.CEL
GSM970094_BE0293-1-J1-2.CEL
GSM970093_BE0293-1-D1.CEL
GSM970092_BE0293-1-A1.CEL
GSM970089_BE0282D1.CEL
GSM970088_BE0282B1.CEL
GSM970087_BE0282A1.CEL
GSM970082_BE0258-1-F1.CEL
GSM970081_BE0258-1-D1.CEL
GSM970080_BE0258-1-A1.CEL
GSM970079_BE0255C1.CEL
GSM970078_BE0255B1.CEL
GSM970077_BE0255A1.CEL
GSM970072_BE0225D1-RE.CEL
GSM970071_BE0225-1-B1.CEL
GSM970070_BE0225-1-A1.CEL
GSM970065_BE0214-1-C1.CEL
GSM970064_BE0214-1-B1.CEL
GSM970063_BE0214-1-A1.CEL
GSM970062_BE0210-1-D1.CEL
GSM970061_BE0210-1-B1.CEL
GSM970060_BE0210-1-A1.CEL
GSM970055_BE0186-1-D1.CEL
GSM970054_BE0186-1-B1.CEL
GSM970053_BE0186-1-A1.CEL
GSM970050_BE0158-1-H1.CEL
GSM970049_BE0158-1-D1.CEL
GSM970048_BE0158-1-A1.CEL
GSM970045_BE0147-1-F1-2.CEL
GSM970044_BE0147-1-B1.CEL
GSM970043_BE0147-1-A1.CEL
GSM970042_BE0141-1-N1.CEL
GSM970041_BE0141-1-B1.CEL
GSM970040_BE0141-1-A1.CEL
GSM970035_BE0130-1-J1.CEL
GSM970034_BE0130-1-G1.CEL
GSM970033_BE0130-1-B1-2.CEL
GSM970032_BE0126-1-F1.CEL
GSM970031_BE0126-1-B1.CEL
GSM970030_BE0126-1-A1.CEL
GSM970029_BE0121-2-A1.CEL
GSM970028_BE0121-2-F1.CEL
GSM970027_BE0121-2-C1.CEL
GSM970026_BE0118-1-E1.CEL
GSM970025_BE0118-1-B1.CEL
GSM970024_BE0118-1-A1.CEL
GSM970019_BE0110-1-H1.CEL
GSM970018_BE0110-1-D1.CEL
GSM970017_BE0110-1-A1.CEL
GSM970014_BE0105-1-D1.CEL
GSM970013_BE0105-1-B1.CEL
GSM970012_BE0105-1-A1.CEL
GSM970011_BE0102-1-E1.CEL
GSM970010_BE0102C1-RE.CEL
GSM970009_BE0102-1-A1.CEL
GSM970004_BE0087-1-G1.CEL
GSM970003_BE0087-1-C1.CEL
GSM970002_BE0087-1-A1.CEL
GSM969997_BE0065-1-L1.CEL
GSM969996_BE0065-1-B1.CEL
GSM969995_BE0065-1-A1.CEL
GSM969994_BE0060-1-D1.CEL
GSM969993_BE0060-1-B1.CEL
GSM969992_BE0060-1-A1.CEL
GSM969986_BE0050-1-J1.CEL
GSM969985_BE0050-1-G1.CEL
GSM969984_BE0050-1-A1.CEL
GSM969983_BE0049-1-D1.CEL
GSM969982_BE0049-1-B1.CEL
GSM969981_BE0049-1-A1.CEL
GSM969972_BE0007-1-J1.CEL
GSM969971_BE0007-1-F1.CEL
GSM969970_BE0007-1-A1.CEL
GSM969969_BE0002-1-H1.CEL
GSM969968_BE0002-1-C1.CEL
GSM969967_BE0002-1-A1.CEL
31.97%
bioB
−3
−2
−1
0
1
2
3


Positive Border Elements
Intensity
0
2000
4000
6000
8000
10000
12000
1
8
16
25
34
43
52
61
70
79
88
97
108
120
Array Index


Negative Border Elements
Intensity
0
2000
4000
6000
8000
10000
12000
1
8
16
25
34
43
52
61
70
79
88
97
108
120
Array Index


Sample Name
1
4
7
11
15
19
23
27
31
35
39
43
47
51
55
59
63
67
71
75
79
83
87
91
95
99
104
109
114
119
Array Index


Positive Elements
Negative Elements
Y Center of Intensity position
X Center of Intensity position
1.0
0.5
0.0
−0.5
−1.0
42
30
41
79
26
81
31
27
73
29
52
80
108
15
18
21
63
107
57
56
14
106
24
49
82
35
6
11
74
89
54
120
116
19
23
81
114
11
48
77
98
62
4
75
95
Sample Name
Array Index
1
4
7
11
15
19
23
27
31
35
39
43
47
51
55
59
63
67
71
75
79
83
87
91
95
99
104
109
114
119


Array−Array Intensity Correlation
Array Index
1
4
7
11
15
19
23
27
31
35
39
43
47
51
55
59
63
67
71
75
79
83
87
91
95
99
104
109
114
119
1
5
9
14
20
26
32
38
44
50
56
62
68
74
80
86
92
98
104
111
118
Array Index
Correlation Coefficient
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0.5
1
5
9
14
20
26
32
38
44
50
56
62
68
74
80
86
92
98
104
111
118
Array Index

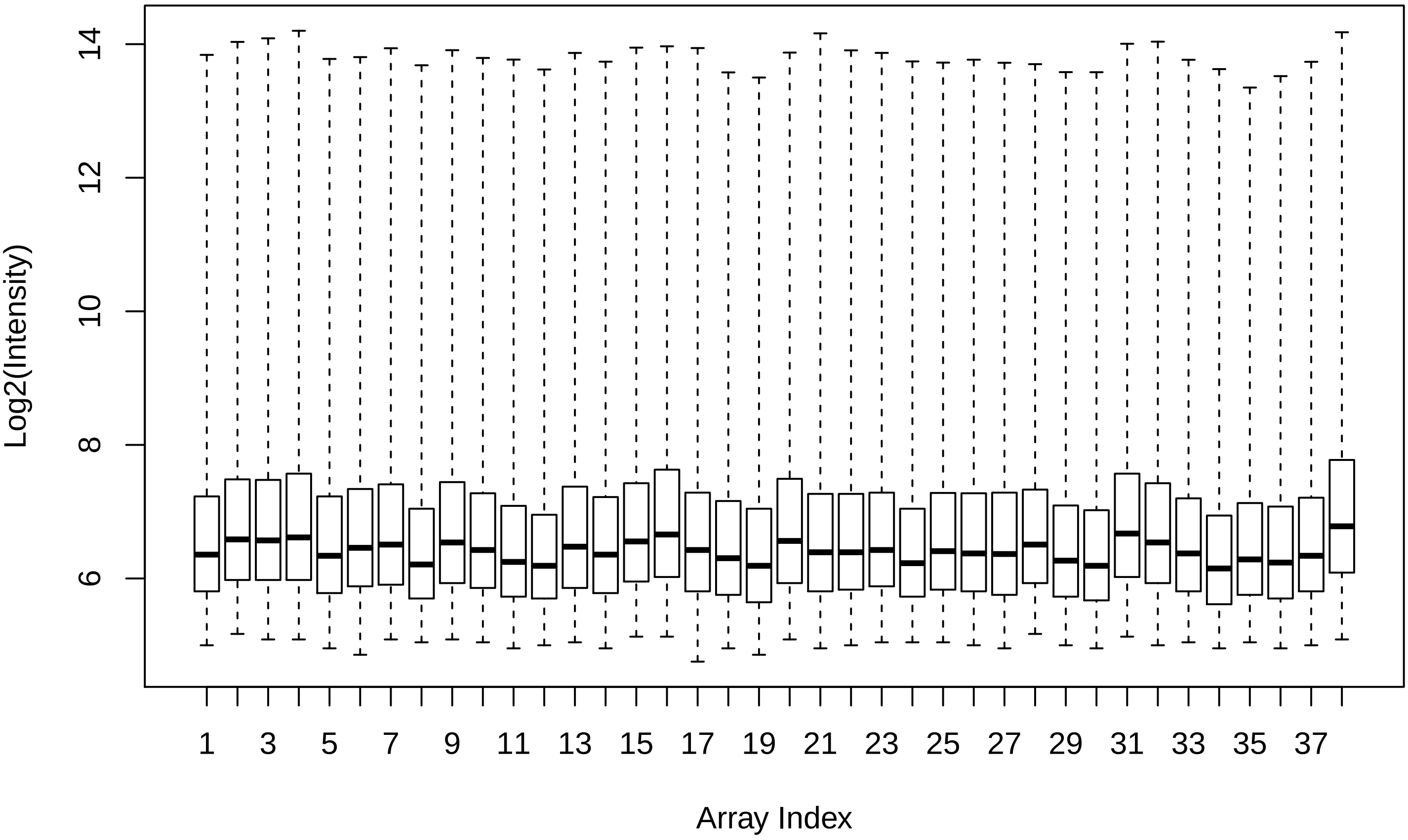
AffyBatch QC Report

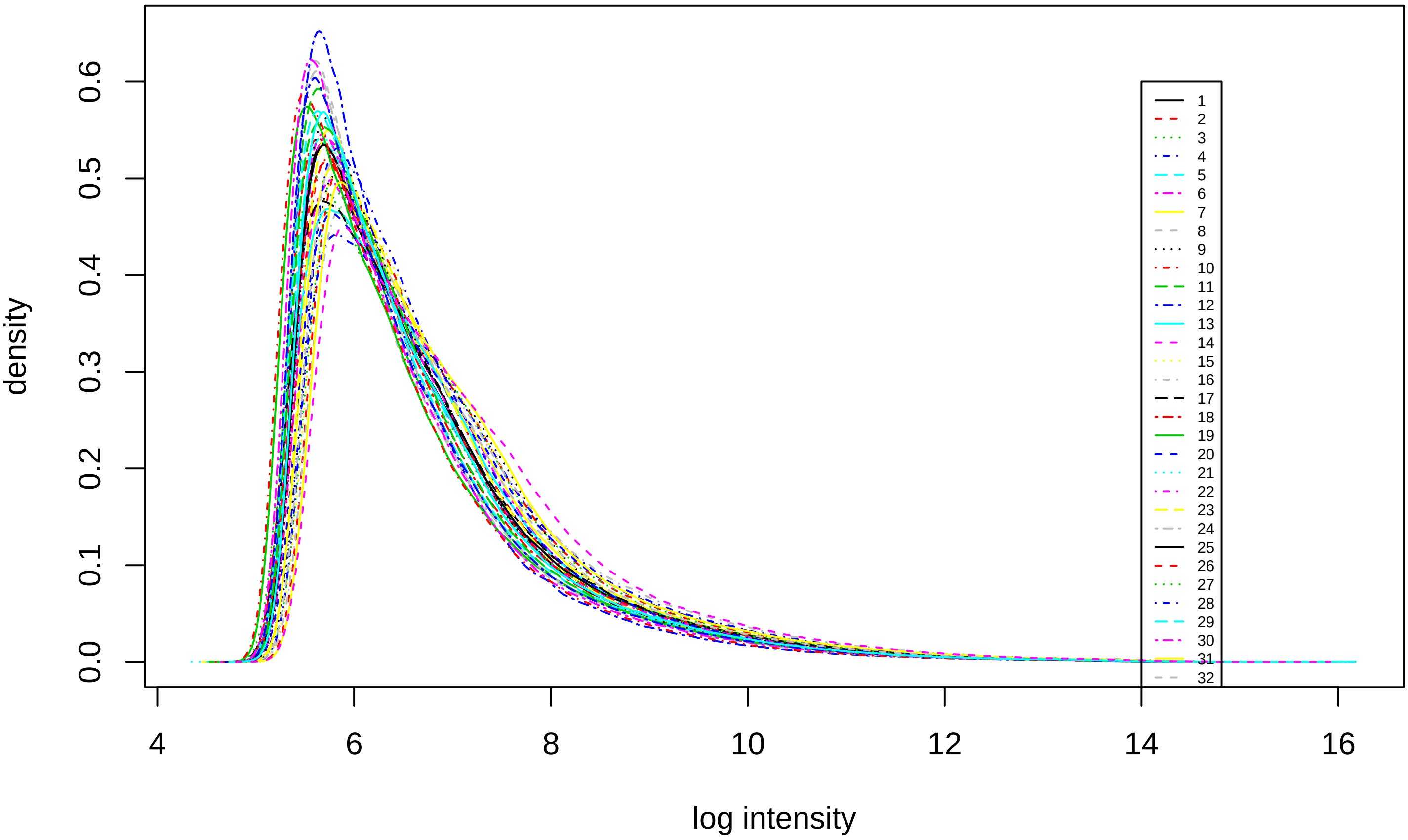
GSE45670

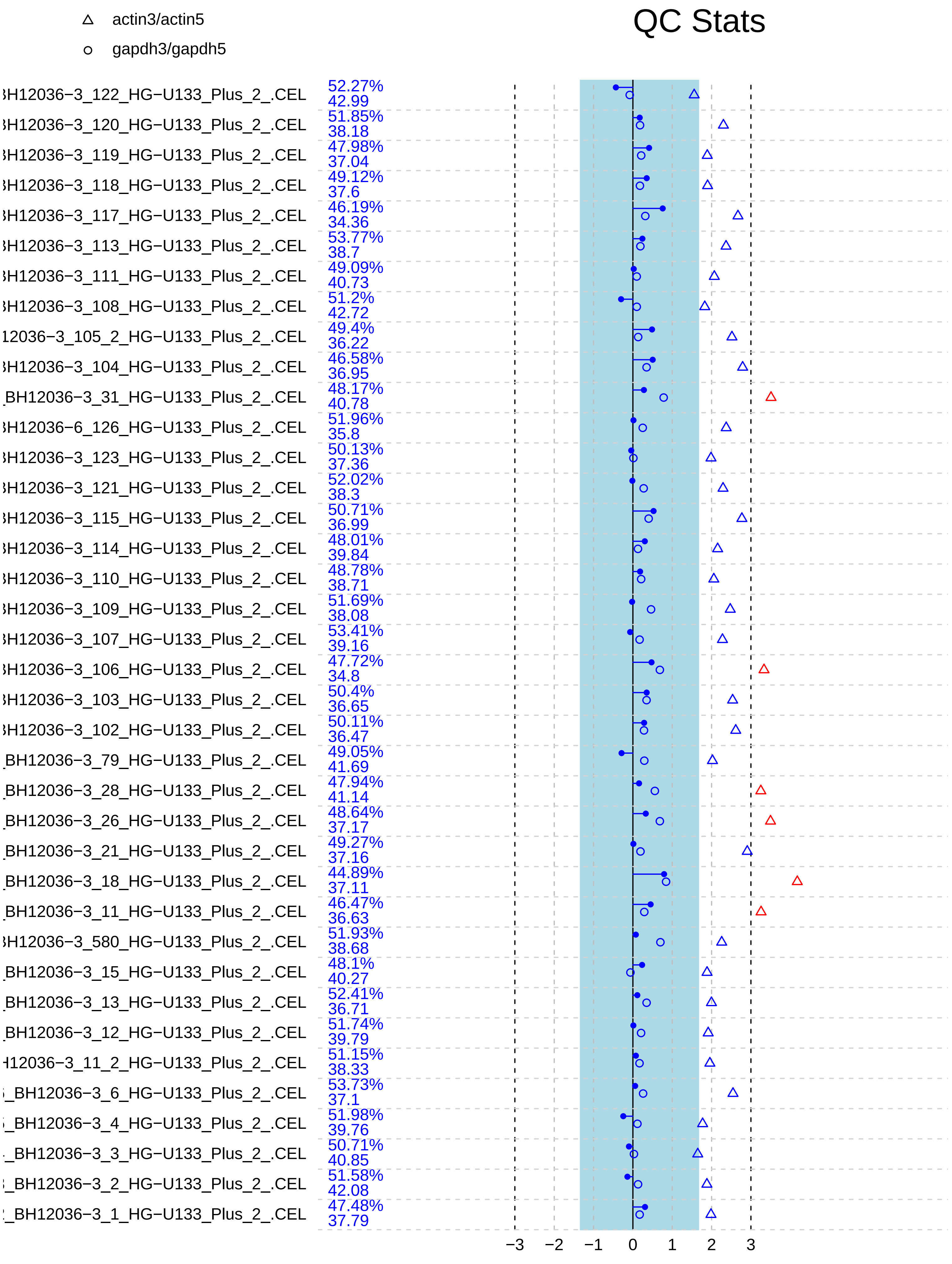
Array Index	Array Name
1	GSM1111662_BH12036−3_1_HG−U133_Plus_2_.CEL
2	GSM1111663_BH12036−3_2_HG−U133_Plus_2_.CEL
3	GSM111166 _BH12036−3_3_HG−U133_Plus_2_.CEL
	GSM1111665_BH12036−3_ _HG−U133_Plus_2_.CEL
5	GSM1111666_BH12036−3_6_HG−U133_Plus_2_.CEL
6	GSM1111667_BH12036−3_11_2_HG−U133_Plus_2_.CEL
7	GSM1111668_BH12036−3_12_HG−U133_Plus_2_.CEL
8	GSM1111669_BH12036−3_13_HG−U133_Plus_2_.CEL
9	GSM1111670_BH12036−3_15_HG−U133_Plus_2_.CEL
10	GSM1111671_BH12036−3_580_HG−U133_Plus_2_ CEL
11	GSM1111672_BH12036−3_11_HG−U133_Plus_2_.CEL
12	GSM1111673_BH12036−3_18_HG−U133_Plus_2_.CEL
13	GSM111167 _BH12036−3_21_HG−U133_Plus_2_.CEL
1	GSM1111675_BH12036−3_26_HG−U133_Plus_2_.CEL
15	GSM1111676_BH12036−3_28_HG−U133_Plus_2_.CEL
16	GSM1111677_BH12036−3_79_HG−U133_Plus_2_.CEL
17	GSM1111678_BH12036−3_102_HG−U133_Plus_2_ CEL
18	GSM1111679_BH12036−3_103_HG−U133_Plus_2_ CEL
19	GSM1111680_BH12036−3_106_HG−U133_Plus_2_ CEL
20	GSM1111681_BH12036−3_107_HG−U133_Plus_2_ CEL
21	GSM1111682_BH12036−3_109_HG−U133_Plus_2_ CEL
22	GSM1111683_BH12036−3_110_HG−U133_Plus_2_ CEL
23	GSM111168 _BH12036−3_11 _HG−U133_Plus_2_ CEL
2	GSM1111685_BH12036−3_115_HG−U133_Plus_2_ CEL
25	GSM1111686_BH12036−3_121_HG−U133_Plus_2_ CEL
26	GSM1111687_BH12036−3_123_HG−U133_Plus_2_ CEL
27	GSM1111688_BH12036−6_126_HG−U133_Plus_2_ CEL
28	GSM1111689_BH12036−3_31_HG−U133_Plus_2_.CEL
29	GSM1111690_BH12036−3_10 _HG−U133_Plus_2_ CEL
30	GSM1111691_BH12036−3_105_2_HG−U133_Plus_2_.CEL
31	GSM1111692_BH12036−3_108_HG−U133_Plus_2_ CEL
32	GSM1111693_BH12036−3_111_HG−U133_Plus_2_ CEL
33	GSM111169 _BH12036−3_113_HG−U133_Plus_2_ CEL
3	GSM1111695_BH12036−3_117_HG−U133_Plus_2_ CEL
35	GSM1111696_BH12036−3_118_HG−U133_Plus_2_ CEL
36	GSM1111697_BH12036−3_119_HG−U133_Plus_2_ CEL
37	GSM1111698_BH12036−3_120_HG−U133_Plus_2_ CEL
38	GSM1111699_BH12036−3_122_HG−U133_Plus_2_ CEL

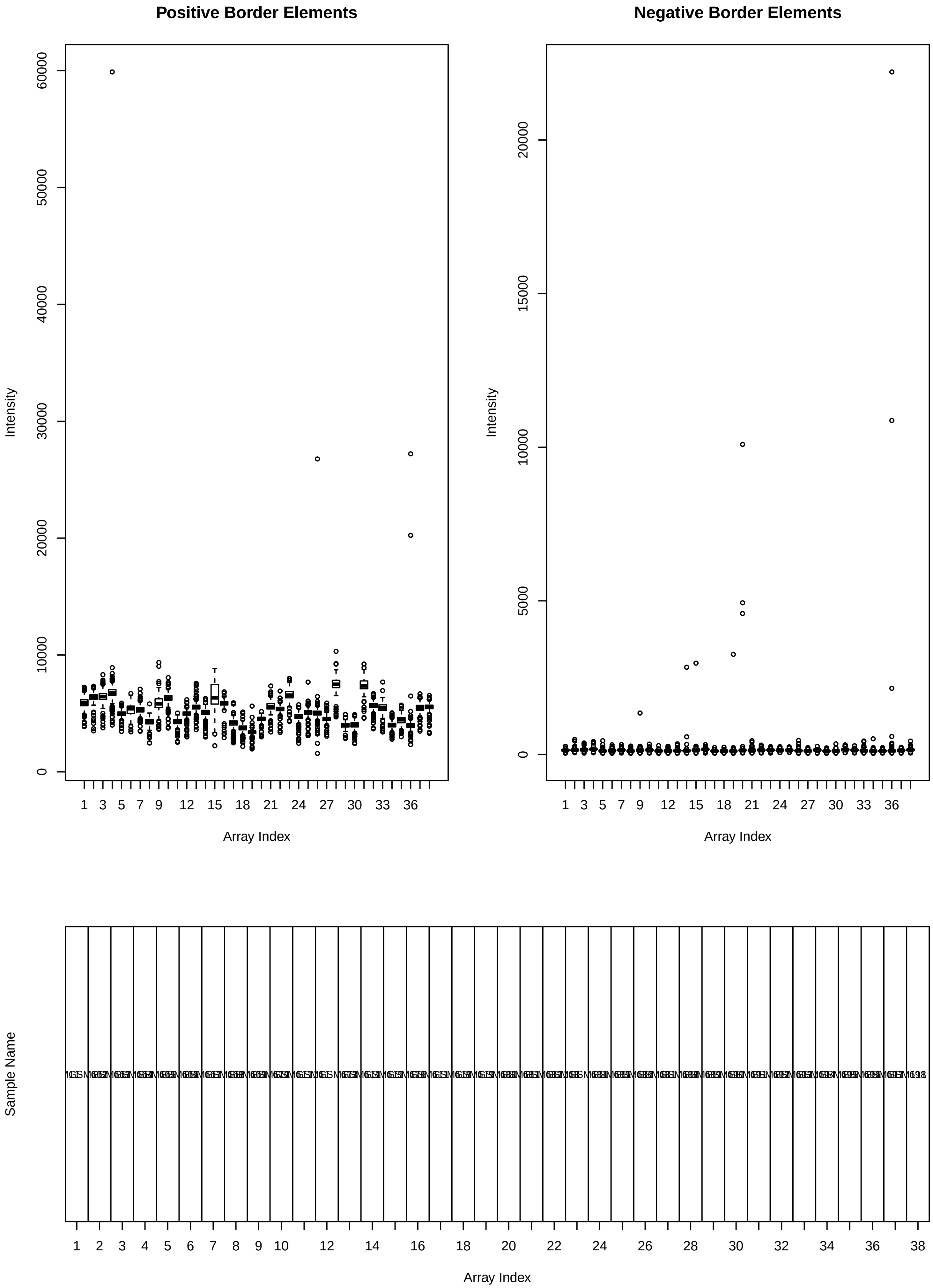
Thu Oct 1 19 55:32 2020

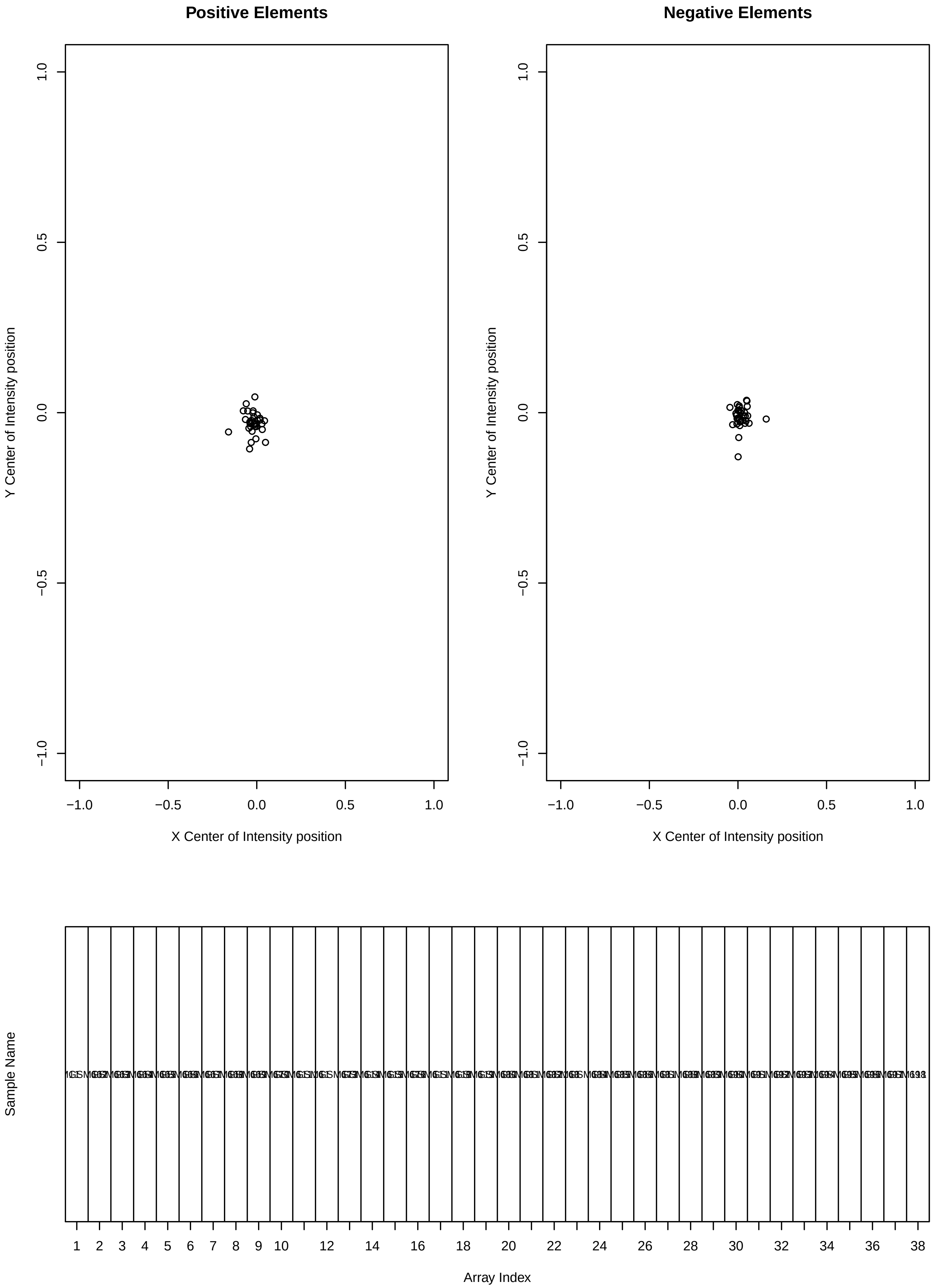
Produced by AffyQCReport R Package


Log2(Intensity)
6
8
10
12
14
1
3
5
7
9
11
13
15
17
19
21
23
25
27
29
31
33
35
37
Array Index

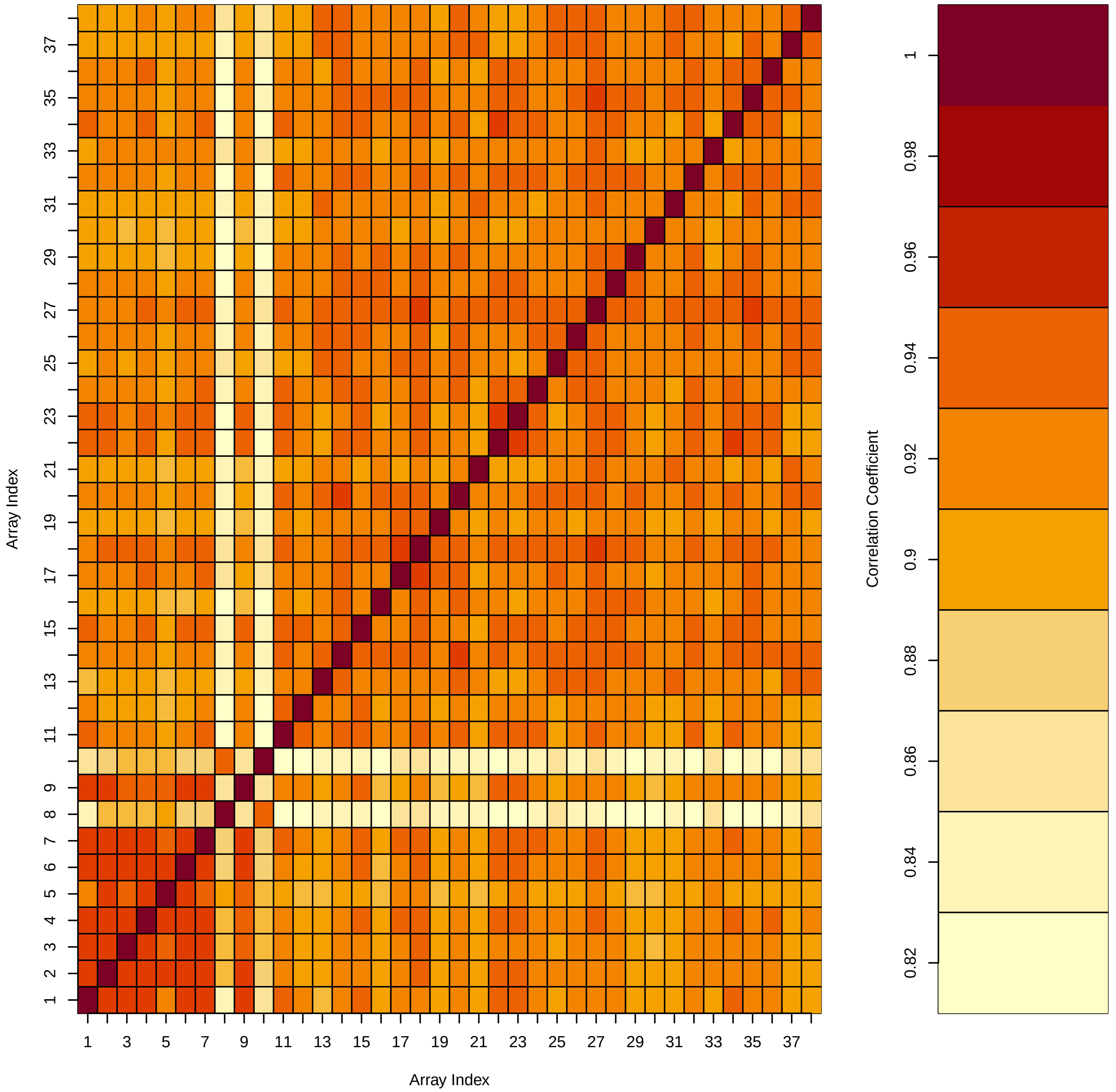

density
0.0
0.1
0.2
0.3
0.4
0.5
0.6
4
6
8
10
12
14
16
log intensity
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32

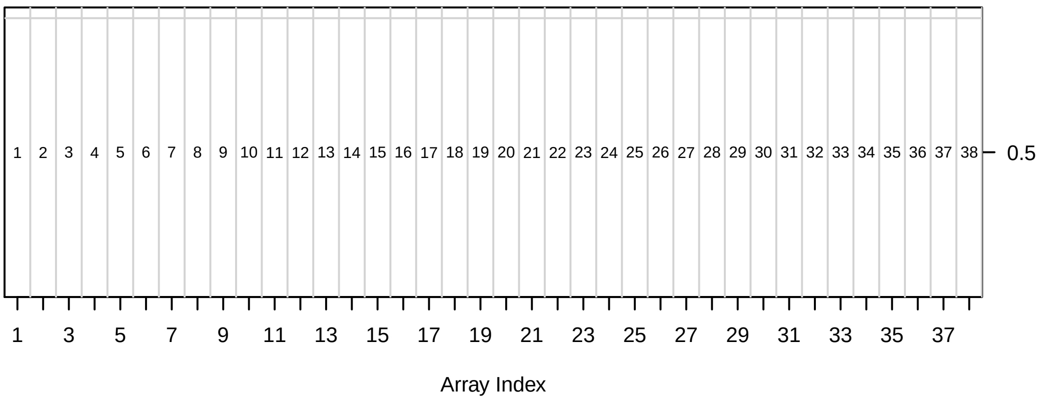

△ actin3/actin5
○ gapdh3/gapdh5
QC Stats
BH12036−3_122_HG−U133_Plus_2_.CEL 52.27% 42.99
BH12036−3_120_HG−U133_Plus_2_.CEL 51.85% 38.18
BH12036−3_119_HG−U133_Plus_2_.CEL 47.98% 37.04
BH12036−3_118_HG−U133_Plus_2_.CEL 49.12% 37.6
BH12036−3_117_HG−U133_Plus_2_.CEL 46.19% 34.36
BH12036−3_113_HG−U133_Plus_2_.CEL 53.77% 38.7
BH12036−3_111_HG−U133_Plus_2_.CEL 49.09% 40.73
BH12036−3_108_HG−U133_Plus_2_.CEL 51.2% 42.72
12036−3_105_2_HG−U133_Plus_2_.CEL 49.4% 36.22
BH12036−3_104_HG−U133_Plus_2_.CEL 46.58% 36.95
_BH12036−3_31_HG−U133_Plus_2_.CEL 48.17% 40.78
BH12036−6_126_HG−U133_Plus_2_.CEL 51.96% 35.8
BH12036−3_123_HG−U133_Plus_2_.CEL 50.13% 37.36
BH12036−3_121_HG−U133_Plus_2_.CEL 52.02% 38.3
BH12036−3_115_HG−U133_Plus_2_.CEL 50.71% 36.99
BH12036−3_114_HG−U133_Plus_2_.CEL 48.01% 39.84
BH12036−3_110_HG−U133_Plus_2_.CEL 48.78% 38.71
BH12036−3_109_HG−U133_Plus_2_.CEL 51.69% 38.08
BH12036−3_107_HG−U133_Plus_2_.CEL 53.41% 39.16
BH12036−3_106_HG−U133_Plus_2_.CEL 47.72% 34.8
BH12036−3_103_HG−U133_Plus_2_.CEL 50.4% 36.65
BH12036−3_102_HG−U133_Plus_2_.CEL 50.11% 36.47
_BH12036−3_79_HG−U133_Plus_2_.CEL 49.05% 41.69
_BH12036−3_28_HG−U133_Plus_2_.CEL 47.94% 41.14
_BH12036−3_26_HG−U133_Plus_2_.CEL 48.64% 37.17
_BH12036−3_21_HG−U133_Plus_2_.CEL 49.27% 37.16
_BH12036−3_18_HG−U133_Plus_2_.CEL 44.89% 37.11
_BH12036−3_11_HG−U133_Plus_2_.CEL 46.47% 36.63
BH12036−3_580_HG−U133_Plus_2_.CEL 51.93% 38.68
_BH12036−3_15_HG−U133_Plus_2_.CEL 48.1% 40.27
_BH12036−3_13_HG−U133_Plus_2_.CEL 52.41% 36.71
_BH12036−3_12_HG−U133_Plus_2_.CEL 51.74% 39.79
H12036−3_11_2_HG−U133_Plus_2_.CEL 51.15% 38.33
_BH12036−3_6_HG−U133_Plus_2_.CEL 53.73% 37.1
_BH12036−3_4_HG−U133_Plus_2_.CEL 51.98% 39.76
_BH12036−3_3_HG−U133_Plus_2_.CEL 50.71% 40.85
_BH12036−3_2_HG−U133_Plus_2_.CEL 51.58% 42.08
_BH12036−3_1_HG−U133_Plus_2_.CEL 47.48% 37.79
−3
−2
−1
0
1
2
3


Positive Border Elements
Negative Border Elements
Intensity
60000
50000
40000
30000
20000
10000
0
Array Index
1 3 5 7 9 12 15 18 21 24 27 30 33 36
Intensity
20000
15000
10000
5000
0
Array Index
1 3 5 7 9 12 15 18 21 24 27 30 33 36
Sample Name
Array Index
1 2 3 4 5 6 7 8 9 10 12 14 16 18 20 22 24 26 28 30 32 34 36 38


Positive Elements
Negative Elements
Y Center of Intensity position
X Center of Intensity position
1.0
0.5
0.0
−0.5
−1.0
Sample Name
Array Index
1
2
3
4
5
6
7
8
9
10
12
14
16
18
20
22
24
26
28
30
32
34
36
38

Array−Array Intensity Correlation





arrayQualityMetrics report for eset.oligoGeneCore

GSE75241

+ Array metadata and outlier detection overview

Section 1: Between array comparison

- Figure 1: Distances between arrays.

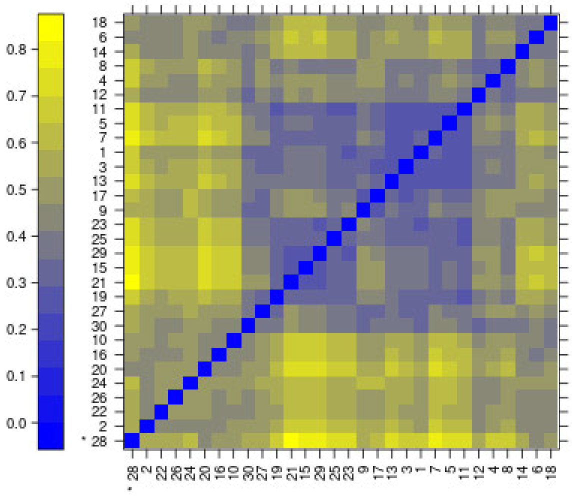


Figure 1 (PDF file) shows a false color heatmap of the distances between arrays. The color scale is chosen to cover the range of distances encountered in the dataset. Patterns in this plot can indicate clustering of the arrays either because of intended biological or unintended experimental factors (batch effects). The distance d_{ab} between two arrays a and b is computed as the mean absolute difference (L$_1$-distance) between the data of the arrays (using the data from all probes without filtering). In formula, d_{ab} = mean | M_{ai} - M_{bi} |, where M_{ai} is the value of the i-th probe on the a-th array. Outlier detection was performed by looking for arrays for which the sum of the distances to all other arrays, S_a = Σ_b d_{ab} was exceptionally large. One such array was detected, and it is marked by an asterisk, *.

+ Figure 2: Outlier detection for Distances between arrays.

- Figure 3: Principal Component Analysis.

array
sampleNames
index

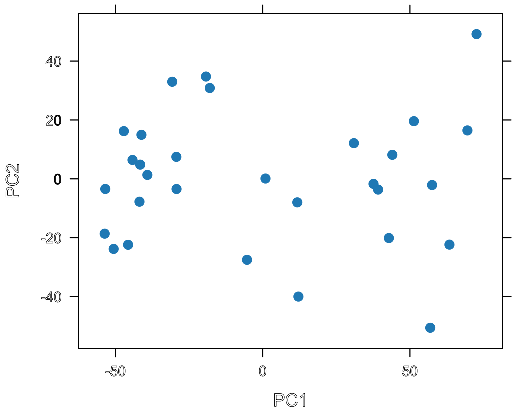


Figure 3 (PDF file) shows a scatterplot of the arrays along the first two principal components. You can use this plot to explore if the arrays cluster, and whether this is according to an intended experimental factor (you can indicate such a factor by color using the 'intgroup' argument), or according to unintended causes such as batch effects. Move the mouse over the points to see the sample names.
Principal component analysis is a dimension reduction and visualisation technique that is here used to project the multivariate data vector of each array into a two-dimensional plot, such that the spatial arrangement of the points in the plot reflects the overall data (dis)similarity between the arrays.

Section 2: Array intensity distributions

- Figure 4: Boxplots.

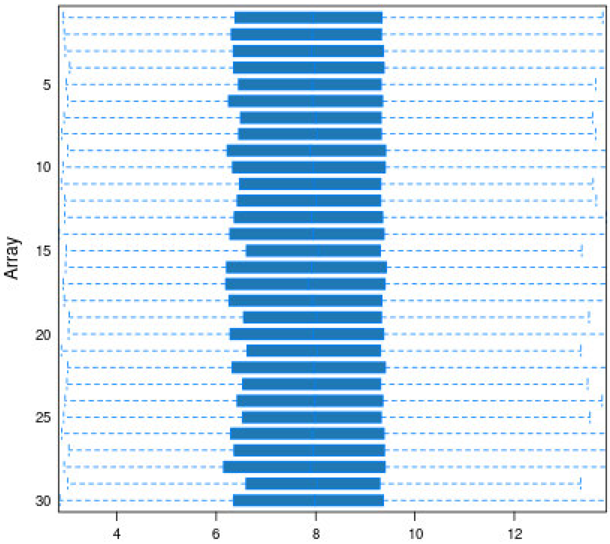


Figure 4 (PDF file) shows boxplots representing summaries of the signal intensity distributions of the arrays. Each box corresponds to one array. Typically, one expects the boxes to have similar positions and widths. If the distribution of an array is very different from the others, this may indicate an experimental problem. Outlier detection was performed by computing the Kolmogorov-Smirnov statistic K_a between each array's distribution and the distribution of the pooled data.

+ Figure 5: Outlier detection for Boxplots.

- Figure 6: Density plots.

array
sampleNames
index

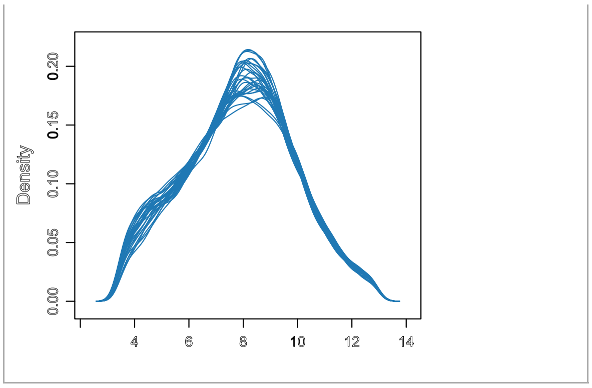


Figure 6 (PDF file) shows density estimates (smoothed histograms) of the data. Typically, the distributions of the arrays should have similar shapes and ranges. Arrays whose distributions are very different from the others should be considered for possible problems. Various features of the distributions can be indicative of quality related phenomena. For instance, high levels of background will shift an array's distribution to the right. Lack of signal diminishes its right right tail. A bulge at the upper end of the intensity range often indicates signal saturation.

Section 3: Variance mean dependence

- Figure 7: Standard deviation versus rank of the mean.

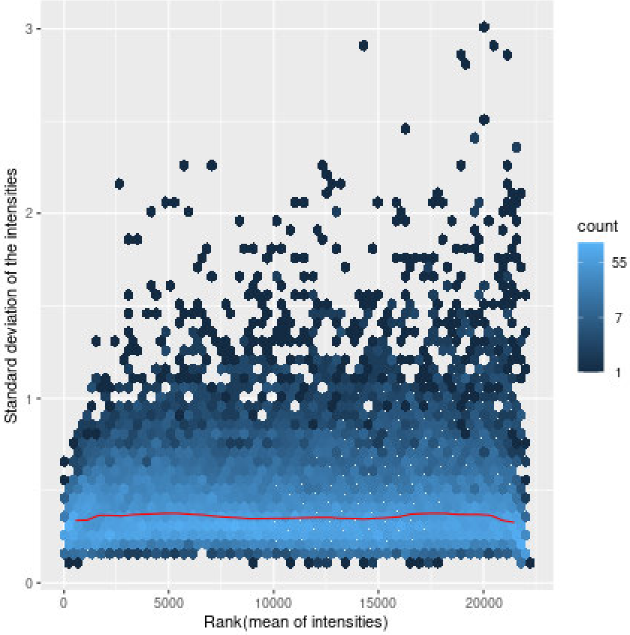


Figure 7 (PDF file) shows a density plot of the standard deviation of the intensities across arrays on the *y*-axis versus the rank of their mean on the *x*-axis. The red dots, connected by lines, show the running median of the standard deviation. After normalisation and transformation to a logarithm(-like) scale, one typically expects the red line to be approximately horizontal, that is, show no substantial trend. In some cases, a hump on the right hand of the x-axis can be observed and is symptomatic of a saturation of the intensities.

Section 4: Individual array quality

- Figure 8: MA plots.

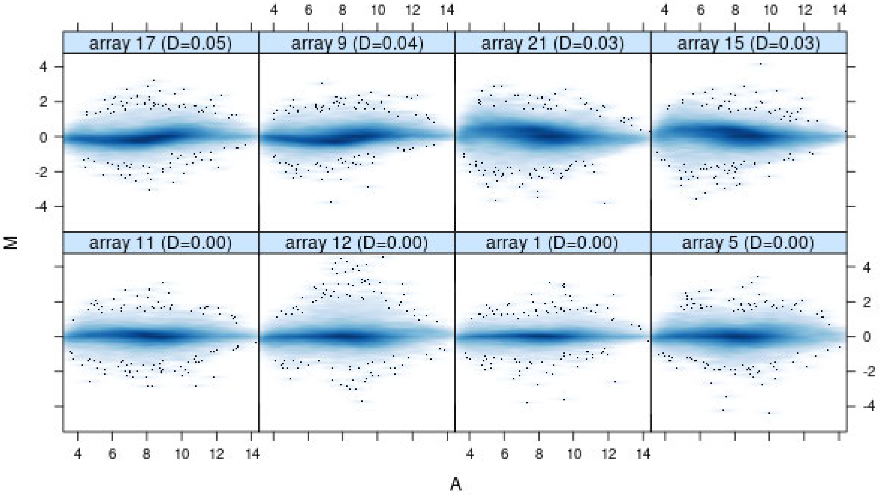


Figure 8 (PDF file) shows MA plots. M and A are defined as:
$M = \log_2(I_1) - \log_2(I_2)$
$A = 1/2 (\log_2(I_1)+\log_2(I_2))$,
where I_1 is the intensity of the array studied,and I_2 is the intensity of a "pseudo"-array that consists of the median across arrays. Typically, we expect the mass of the distribution in an MA plot to be concentrated along the M = 0 axis, and there should be no trend in M as a function of A. If there is a trend in the lower range of A, this often indicates that the arrays have different background intensities; this may be addressed by background correction. A trend in the upper range of A can indicate saturation of the measurements; in mild cases, this may be addressed by non-linear normalisation (e.g. quantile normalisation).
Outlier detection was performed by computing Hoeffding's statistic D_a on the joint distribution of A and M for each array. Shown are first the 4 arrays with the highest values of D_a, then the 4 arrays with the lowest values. The value of D_a is shown in the panel headings. 0 arrays had D_a>0.15 and were marked as outliers. For more information on Hoeffing's *D*-statistic, please see the manual page of the function `hoeffd` in the `Hmisc` package.

+ Figure 9: Outlier detection for MA plots.

This report has been created with arrayQualityMetrics 3.42.0 under R version 3.6.0 (2019-04-26).

(Page generated on Sun Oct 11 13:46:23 2020 by hwriter)

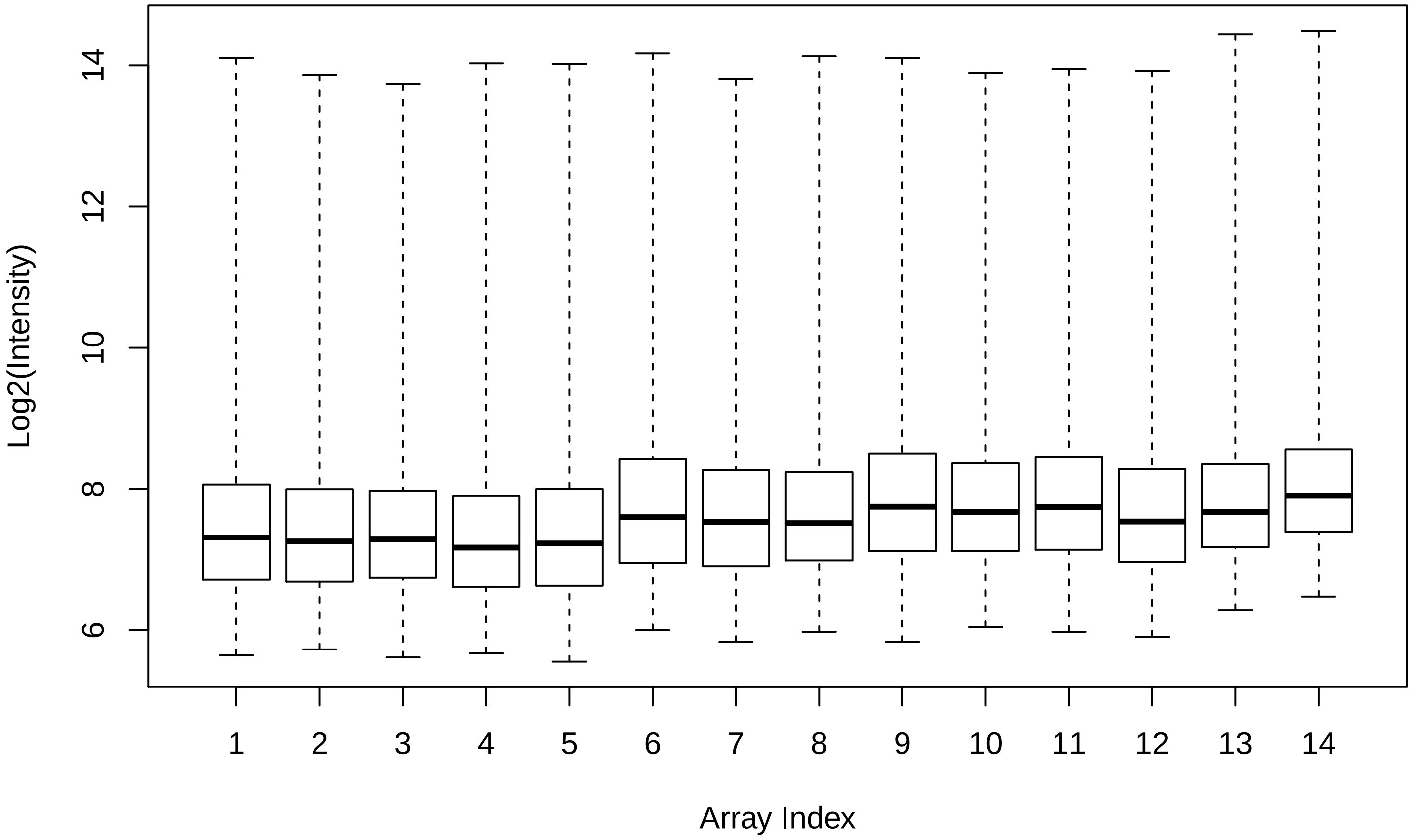
AffyBatch QC Report

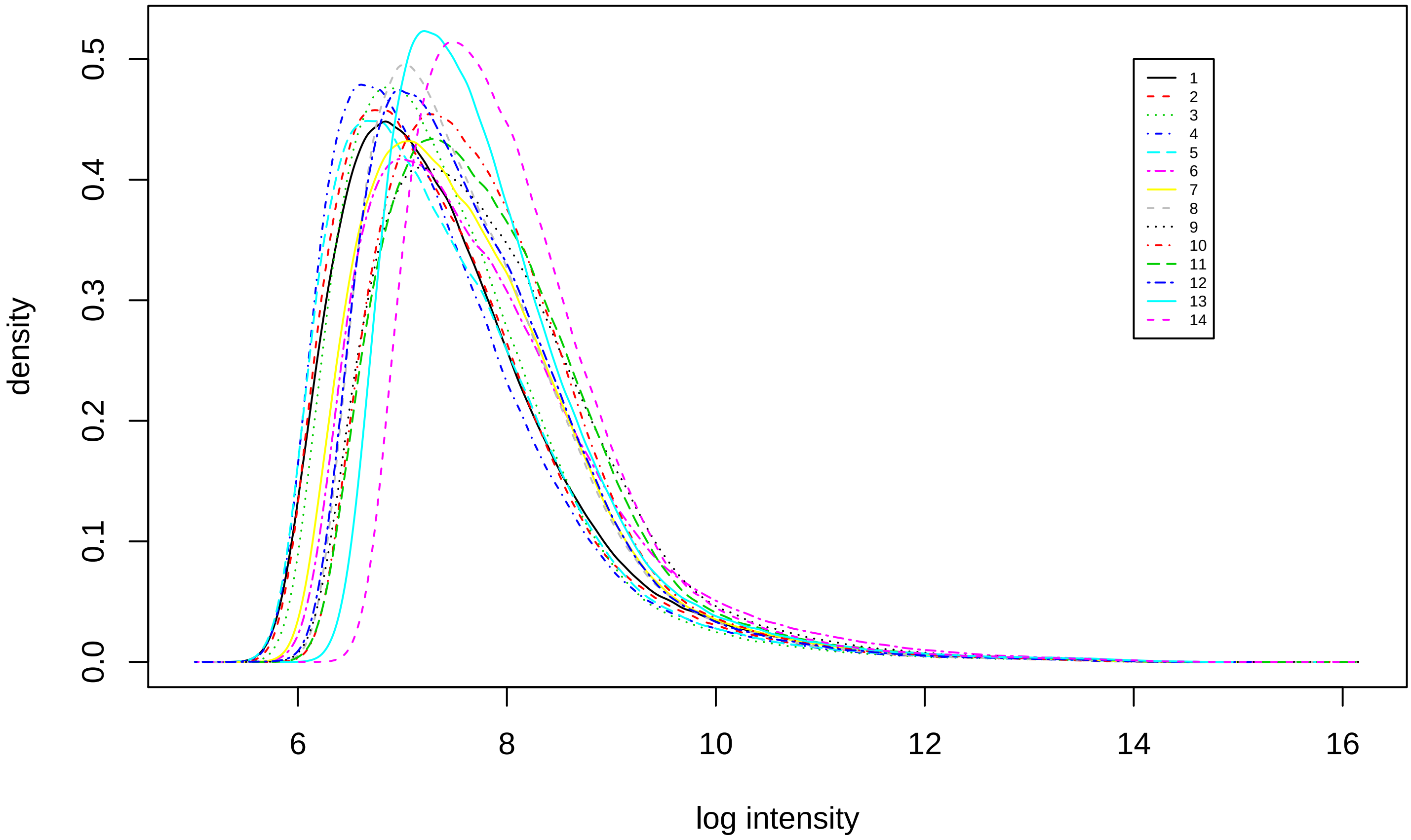
GSE77861

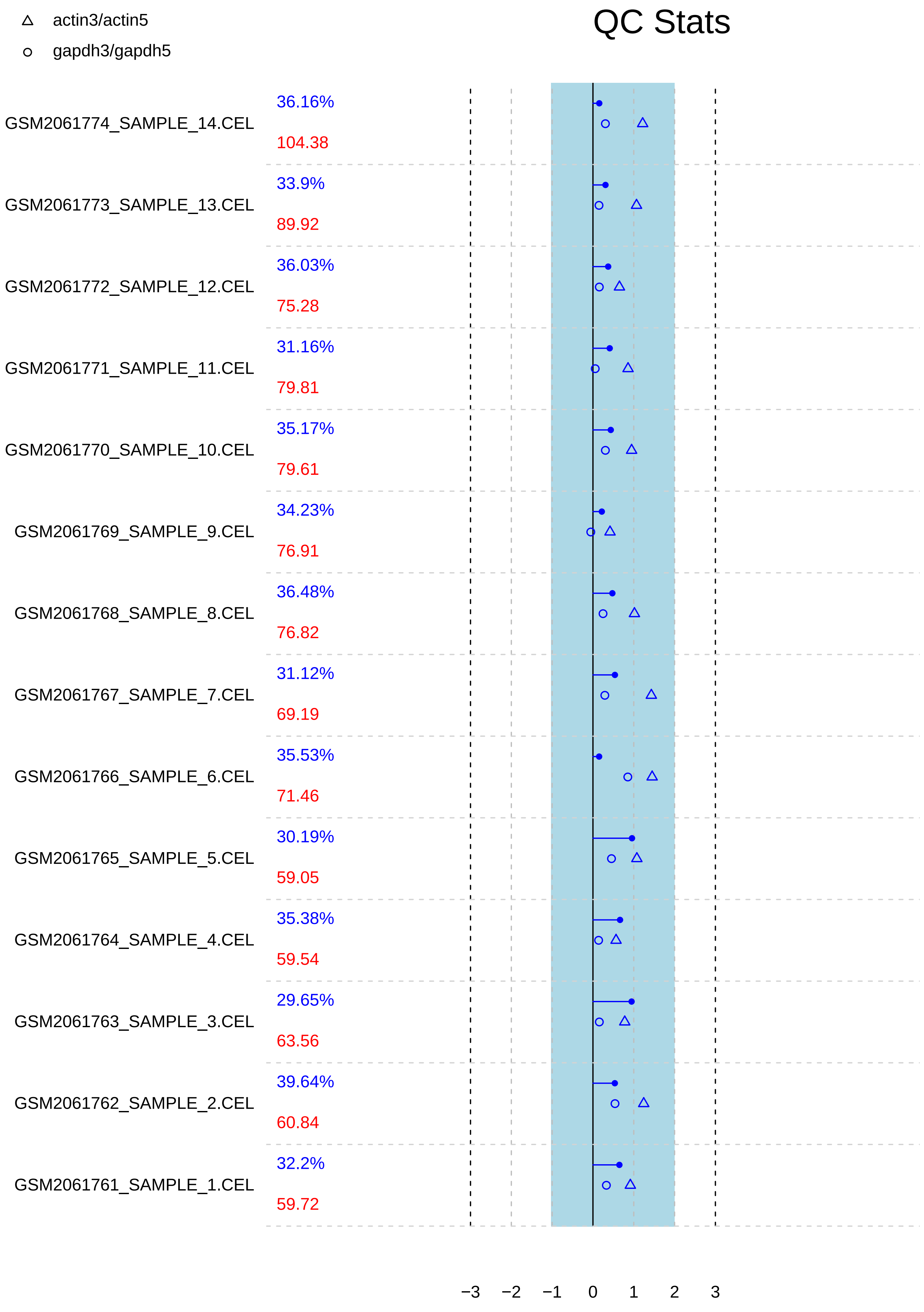
Array Index	Array Name
1	GSM2061761_SAMPLE_1.CEL
2	GSM2061762_SAMPLE_2.CEL
3	GSM2061763_SAMPLE_3.CEL
4	GSM2061764_SAMPLE_4.CEL
5	GSM2061765_SAMPLE_5.CEL
6	GSM2061766_SAMPLE_6.CEL
7	GSM2061767_SAMPLE_7.CEL
8	GSM2061768_SAMPLE_8.CEL
9	GSM2061769_SAMPLE_9.CEL
10	GSM2061770_SAMPLE_10.CEL
11	GSM2061771_SAMPLE_11.CEL
12	GSM2061772_SAMPLE_12.CEL
13	GSM2061773_SAMPLE_13.CEL
14	GSM2061774_SAMPLE_14.CEL

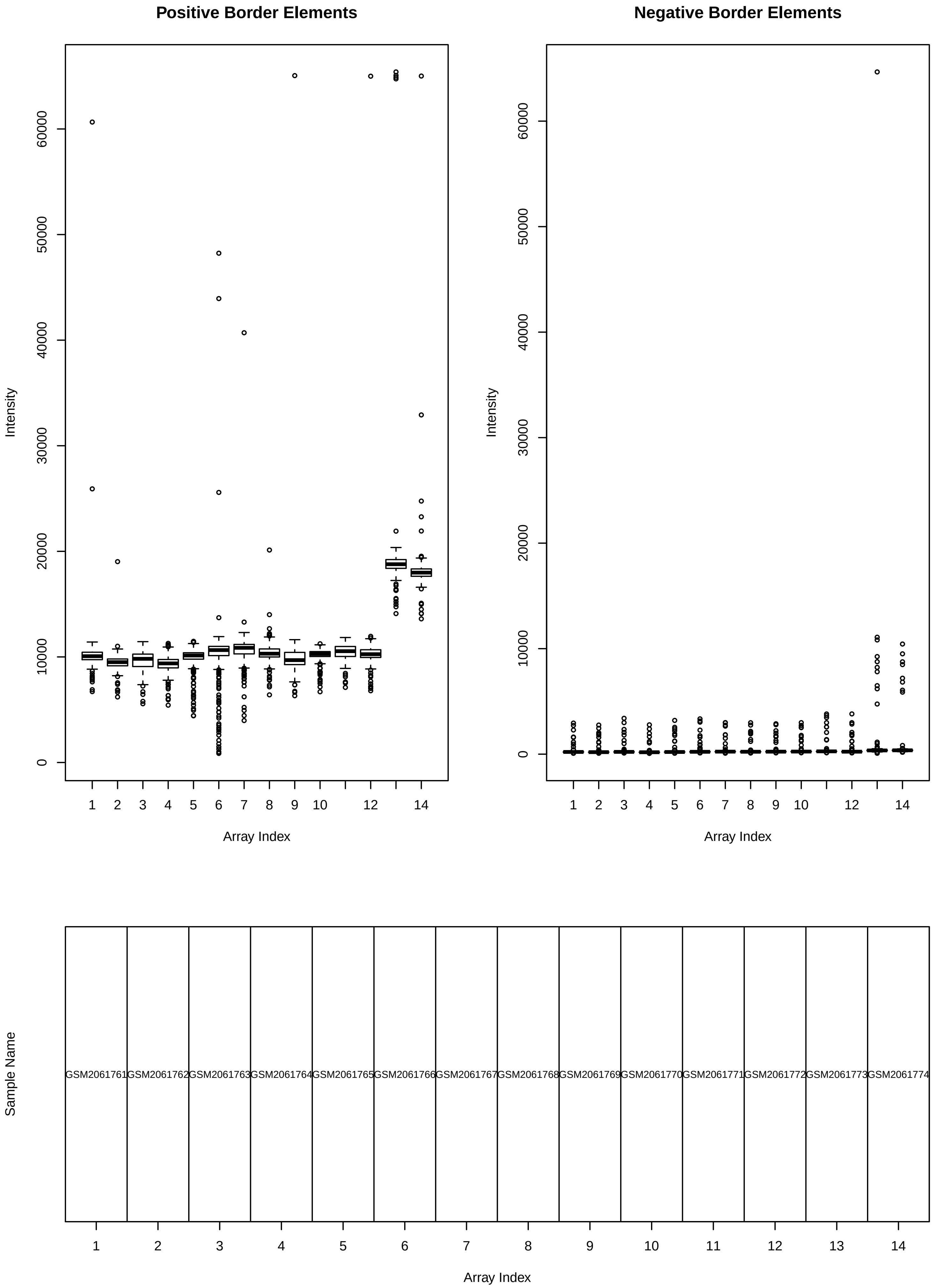
Thu Oct 1 17 59:08 2020

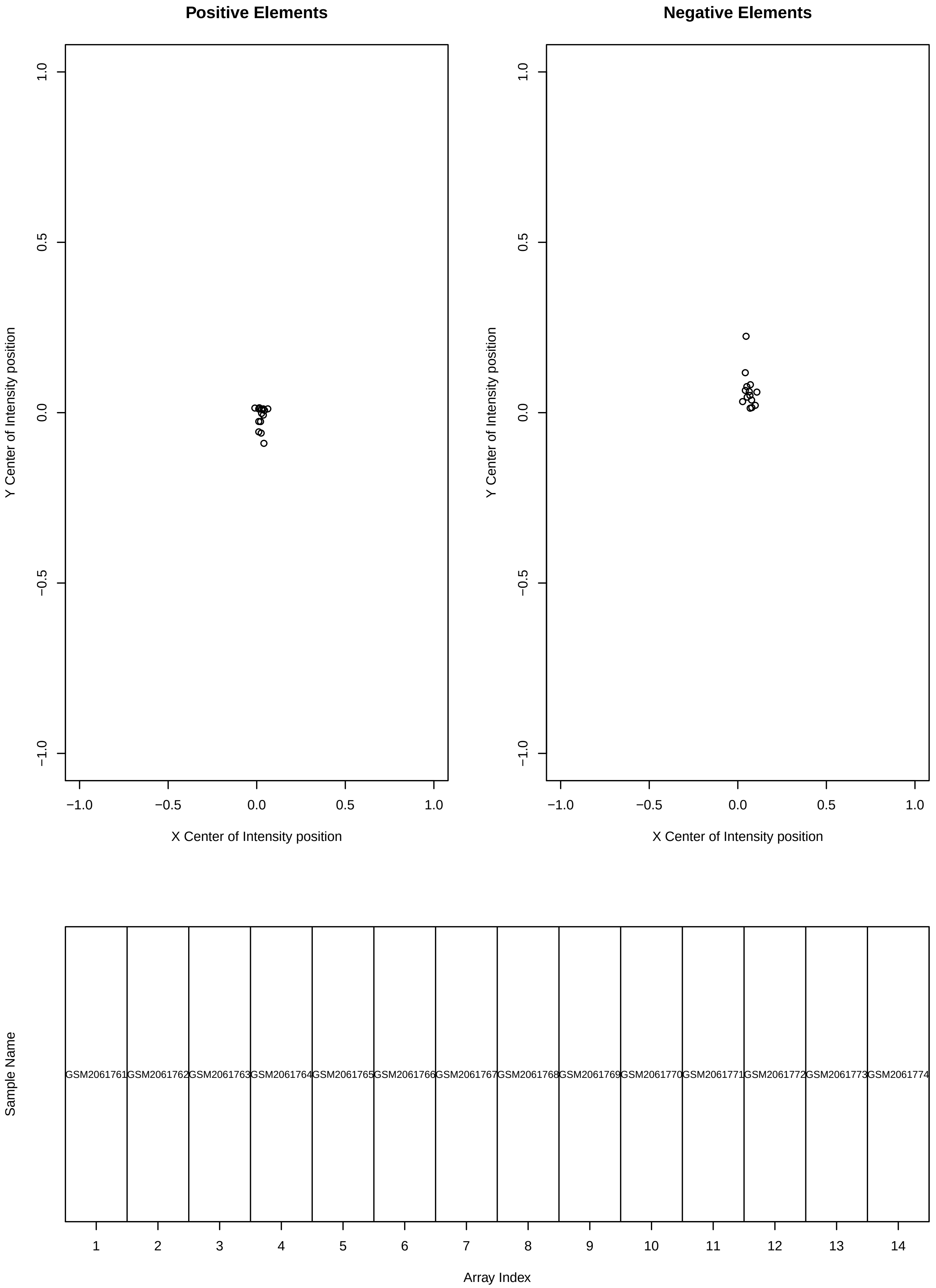
Produced by AffyQCReport R Package

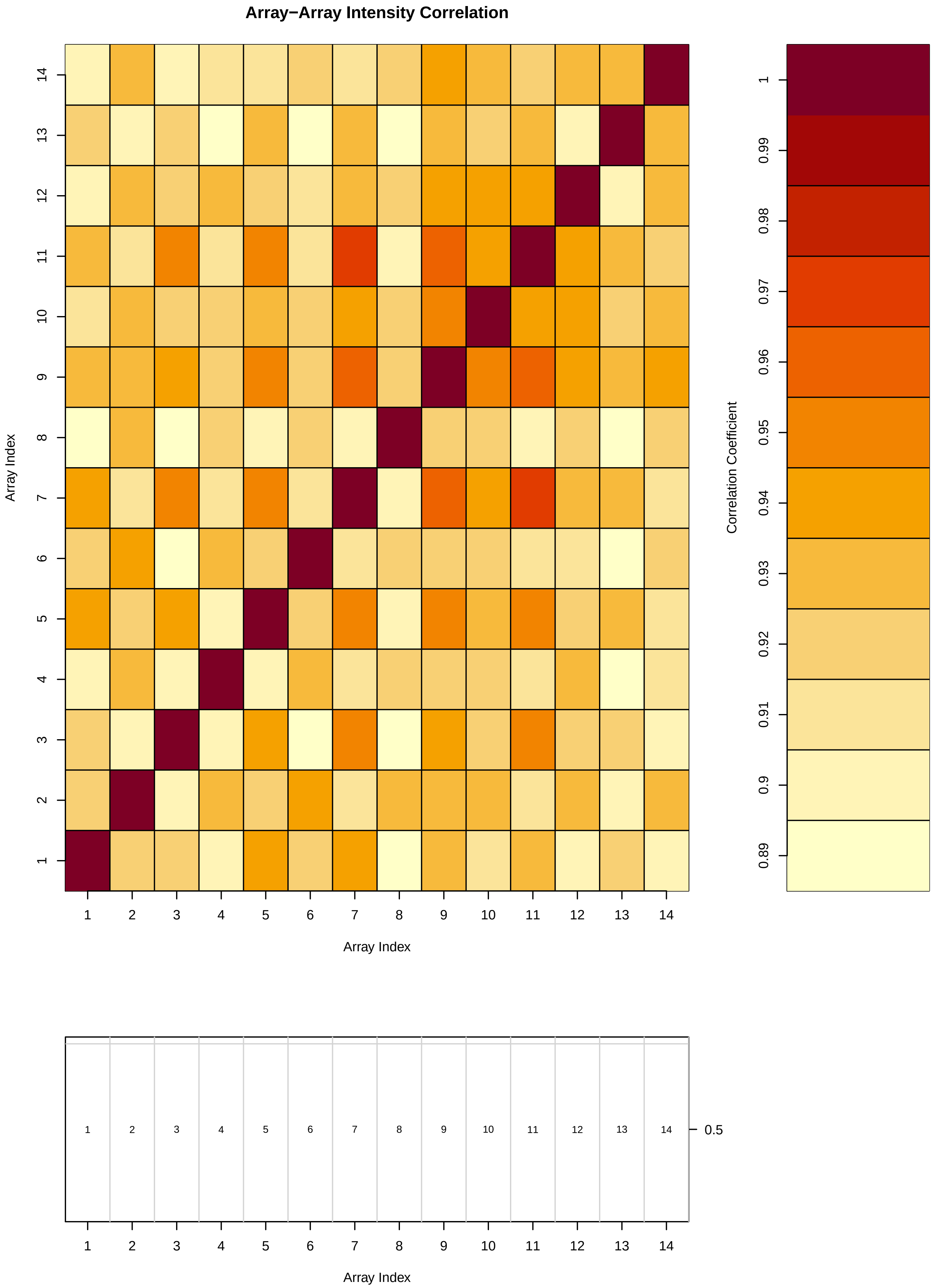

Log2(Intensity)
6
8
10
12
14
1
2
3
4
5
6
7
8
9
10
11
12
13
14
Array Index


density
0.0
0.1
0.2
0.3
0.4
0.5
6
8
10
12
14
16
log intensity
1
2
3
4
5
6
7
8
9
10
11
12
13
14


△ actin3/actin5
○ gapdh3/gapdh5
QC Stats
GSM2061774_SAMPLE_14.CEL
36.16%
104.38
GSM2061773_SAMPLE_13.CEL
33.9%
89.92
GSM2061772_SAMPLE_12.CEL
36.03%
75.28
GSM2061771_SAMPLE_11.CEL
31.16%
79.81
GSM2061770_SAMPLE_10.CEL
35.17%
79.61
GSM2061769_SAMPLE_9.CEL
34.23%
76.91
GSM2061768_SAMPLE_8.CEL
36.48%
76.82
GSM2061767_SAMPLE_7.CEL
31.12%
69.19
GSM2061766_SAMPLE_6.CEL
35.53%
71.46
GSM2061765_SAMPLE_5.CEL
30.19%
59.05
GSM2061764_SAMPLE_4.CEL
35.38%
59.54
GSM2061763_SAMPLE_3.CEL
29.65%
63.56
GSM2061762_SAMPLE_2.CEL
39.64%
60.84
GSM2061761_SAMPLE_1.CEL
32.2%
59.72
−3
−2
−1
0
1
2
3


Positive Border Elements
Negative Border Elements
Intensity
0
10000
20000
30000
40000
50000
60000
Array Index
1
2
3
4
5
6
7
8
9
10
12
14
Sample Name
GSM2061761
GSM2061762
GSM2061763
GSM2061764
GSM2061765
GSM2061766
GSM2061767
GSM2061768
GSM2061769
GSM2061770
GSM2061771
GSM2061772
GSM2061773
GSM2061774
1
2
3
4
5
6
7
8
9
10
11
12
13
14
Array Index


Positive Elements
Negative Elements
Y Center of Intensity position
X Center of Intensity position
1.0
0.5
0.0
−0.5
−1.0
Sample Name
GSM2061761
GSM2061762
GSM2061763
GSM2061764
GSM2061765
GSM2061766
GSM2061767
GSM2061768
GSM2061769
GSM2061770
GSM2061771
GSM2061772
GSM2061773
GSM2061774
1 2 3 4 5 6 7 8 9 10 11 12 13 14
Array Index


Array−Array Intensity Correlation
Array Index
1
2
3
4
5
6
7
8
9
10
11
12
13
14
Correlation Coefficient
0.89
0.9
0.91
0.92
0.93
0.94
0.95
0.96
0.97
0.98
0.99
1
0.5

arrayQualityMetrics report for eset.oligoGeneCore

GSE92396

- **Section 1: Between array comparison**
    - Distances between arrays
    - Principal Component Analysis
- **Section 2: Array intensity distributions**
    - Boxplots
    - Density plots
- **Section 3: Variance mean dependence**
    - Standard deviation versus rank of the mean
- **Section 4: Individual array quality**
    - MA plots

+ Array metadata and outlier detection overview

Section 1: Between array comparison

- Figure 1: Distances between arrays.

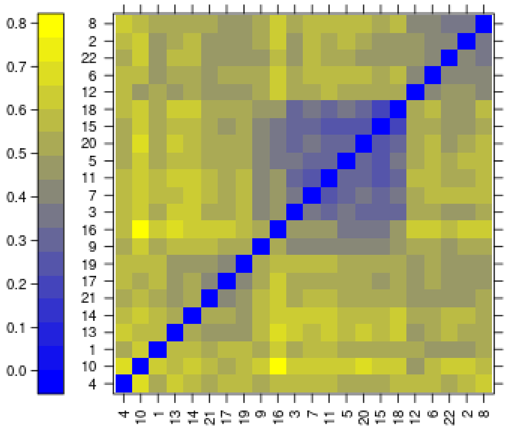


Figure 1 (PDF file) shows a false color heatmap of the distances between arrays. The color scale is chosen to cover the range of distances encountered in the dataset. Patterns in this plot can indicate clustering of the arrays either because of intended biological or unintended experimental factors (batch effects). The distance d_{ab} between two arrays *a* and *b* is computed as the mean absolute difference (L_1-distance) between the data of the arrays (using the data from all probes without filtering). In formula, d_{ab} = mean | M_{ai} - M_{bi} |, where M_{ai} is the value of the *i*-th probe on the *a*-th array. Outlier detection was performed by looking for arrays for which the sum of the distances to all other arrays, $S_a = \Sigma_b d_{ab}$ was exceptionally large. No such arrays were detected.

+ Figure 2: Outlier detection for Distances between arrays.

- Figure 3: Principal Component Analysis.

array
sampleNames
index

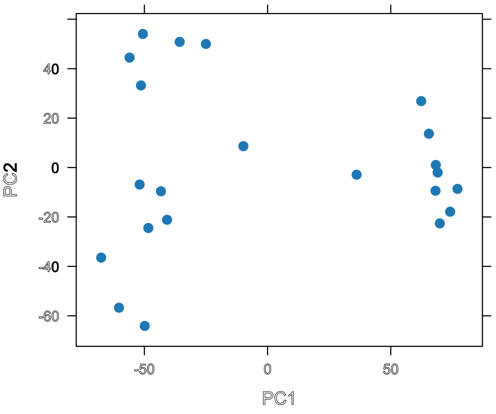


Figure 3 (PDF file) shows a scatterplot of the arrays along the first two principal components. You can use this plot to explore if the arrays cluster, and whether this is according to an intended experimental factor (you can indicate such a factor by color using the 'intgroup' argument), or according to unintended causes such as batch effects. Move the mouse over the points to see the sample names.
Principal component analysis is a dimension reduction and visualisation technique that is here used to project the multivariate data vector of each array into a two-dimensional plot, such that the spatial arrangement of the points in the plot reflects the overall data (dis)similarity between the arrays.

Section 2: Array intensity distributions

- Figure 4: Boxplots.

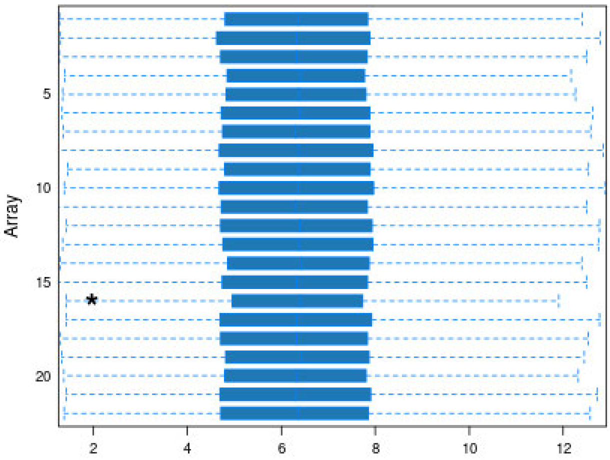


Figure 4 (PDF file) shows boxplots representing summaries of the signal intensity distributions of the arrays. Each box corresponds to one array. Typically, one expects the boxes to have similar positions and widths. If the distribution of an array is very different from the others, this may indicate an experimental problem. Outlier detection was performed by computing the Kolmogorov-Smirnov statistic K_a between each array's distribution and the distribution of the pooled data.

+ Figure 5: Outlier detection for Boxplots.

- Figure 6: Density plots.

array
sampleNames
index

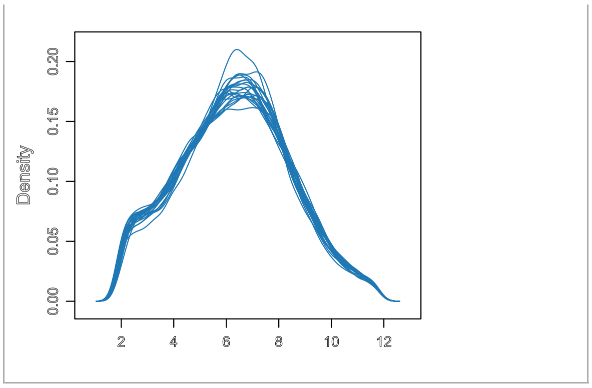


Figure 6 (PDF file) shows density estimates (smoothed histograms) of the data. Typically, the distributions of the arrays should have similar shapes and ranges. Arrays whose distributions are very different from the others should be considered for possible problems. Various features of the distributions can be indicative of quality related phenomena. For instance, high levels of background will shift an array's distribution to the right. Lack of signal diminishes its right right tail. A bulge at the upper end of the intensity range often indicates signal saturation.

Section 3: Variance mean dependence

- Figure 7: Standard deviation versus rank of the mean.

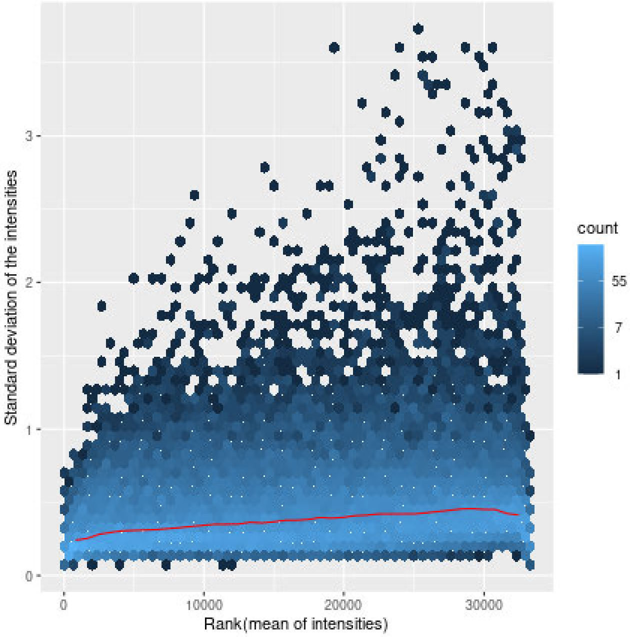


Figure 7 (PDF file) shows a density plot of the standard deviation of the intensities across arrays on the *y*-axis versus the rank of their mean on the *x*-axis. The red dots, connected by lines, show the running median of the standard deviation. After normalisation and transformation to a logarithm(-like) scale, one typically expects the red line to be approximately horizontal, that is, show no substantial trend. In some cases, a hump on the right hand of the x-axis can be observed and is symptomatic of a saturation of the intensities.

Section 4: Individual array quality

- Figure 8: MA plots.

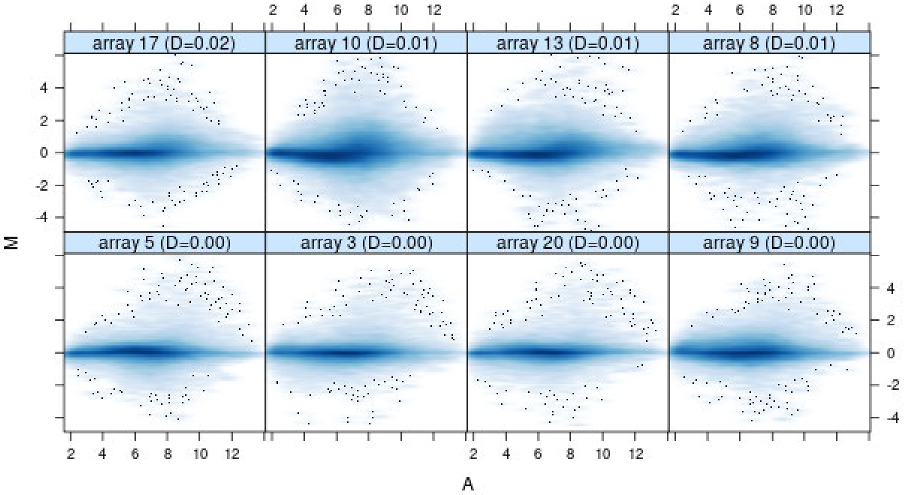


Figure 8 (PDF file) shows MA plots. M and A are defined as:
$M = \log_2(I_1) - \log_2(I_2)$
$A = 1/2 (\log_2(I_1)+\log_2(I_2))$,
where I_1 is the intensity of the array studied,and I_2 is the intensity of a "pseudo"-array that consists of the median across arrays. Typically, we expect the mass of the distribution in an MA plot to be concentrated along the M = 0 axis, and there should be no trend in M as a function of A. If there is a trend in the lower range of A, this often indicates that the arrays have different background intensities; this may be addressed by background correction. A trend in the upper range of A can indicate saturation of the measurements; in mild cases, this may be addressed by non-linear normalisation (e.g. quantile normalisation).
Outlier detection was performed by computing Hoeffding's statistic D_a on the joint distribution of A and M for each array. Shown are first the 4 arrays with the highest values of D_a, then the 4 arrays with the lowest values. The value of D_a is shown in the panel headings. 0 arrays had D_a>0.15 and were marked as outliers. For more information on Hoeffing's *D*-statistic, please see the manual page of the function `hoeffd` in the `Hmisc` package.

+ Figure 9: Outlier detection for MA plots.

This report has been created with arrayQualityMetrics 3.42.0 under R version 3.6.0 (2019-04-26).

(Page generated on Tue Oct 6 00:54:54 2020 by hwriter)

arrayQualityMetrics report for eset.oligoGeneCore

GSE100843

- **Section 1: Between array comparison**
  - Distances between arrays
  - Principal Component Analysis
- **Section 2: Array intensity distributions**
  - Boxplots
  - Density plots
- **Section 3: Variance mean dependence**
  - Standard deviation versus rank of the mean
- **Section 4: Individual array quality**
  - MA plots

+ Array metadata and outlier detection overview

Section 1: Between array comparison

- Figure 1: Distances between arrays.

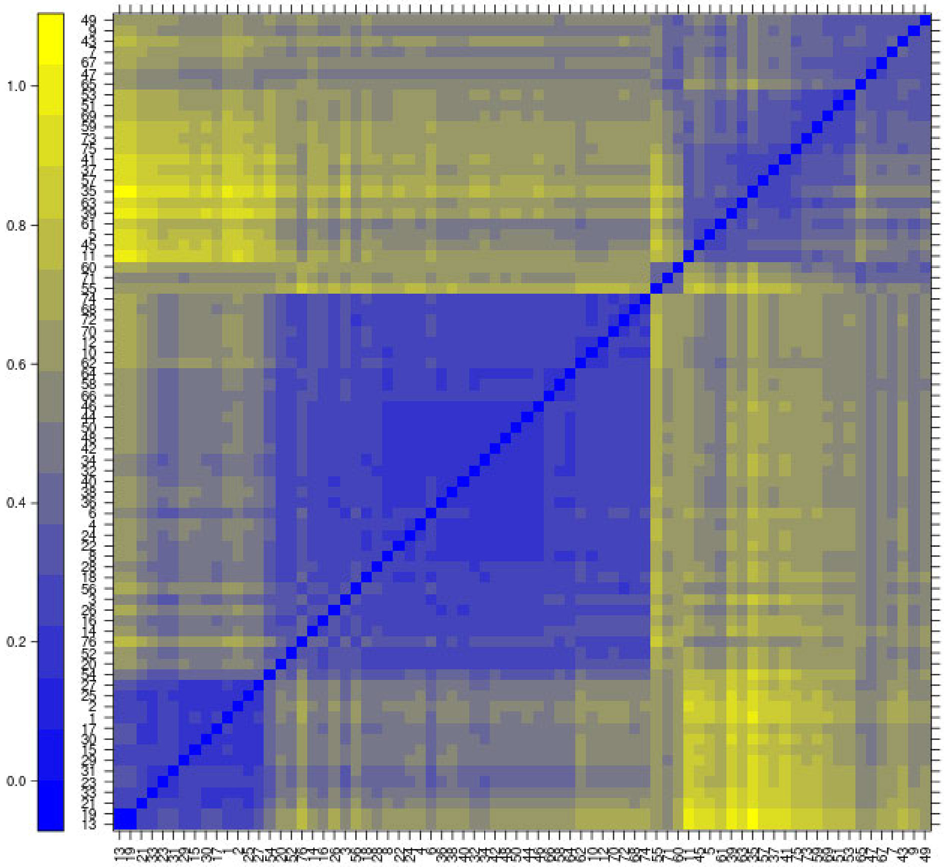


Figure 1 (PDF file) shows a false color heatmap of the distances between arrays. The color scale is chosen to cover the range of distances encountered in the dataset. Patterns in this plot can indicate clustering of the arrays either because of intended biological or unintended experimental factors (batch effects). The distance d_{ab} between two arrays a and b is computed as the mean absolute difference (L_1-distance) between the data of the arrays (using the data from all probes without filtering). In formula, d_{ab} = mean | M_{ai} - M_{bi} |, where M_{ai} is the value of the i-th probe on the a-th array. Outlier detection was performed by looking for arrays for which the sum of the distances to all other arrays, $S_a = \Sigma_b d_{ab}$ was exceptionally large. No such arrays were detected.

+ Figure 2: Outlier detection for Distances between arrays.

- Figure 3: Principal Component Analysis.

array
sampleNames

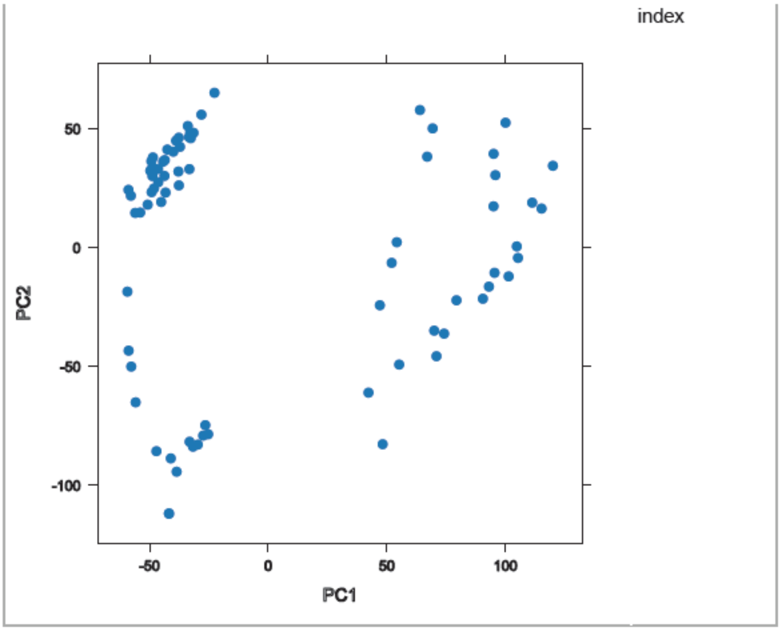


Figure 3 (PDF file) shows a scatterplot of the arrays along the first two principal components. You can use this plot to explore if the arrays cluster, and whether this is according to an intended experimental factor (you can indicate such a factor by color using the 'intgroup' argument), or according to unintended causes such as batch effects. Move the mouse over the points to see the sample names.
Principal component analysis is a dimension reduction and visualisation technique that is here used to project the multivariate data vector of each array into a two-dimensional plot, such that the spatial arrangement of the points in the plot reflects the overall data (dis)similarity between the arrays.

Section 2: Array intensity distributions

- Figure 4: Boxplots.

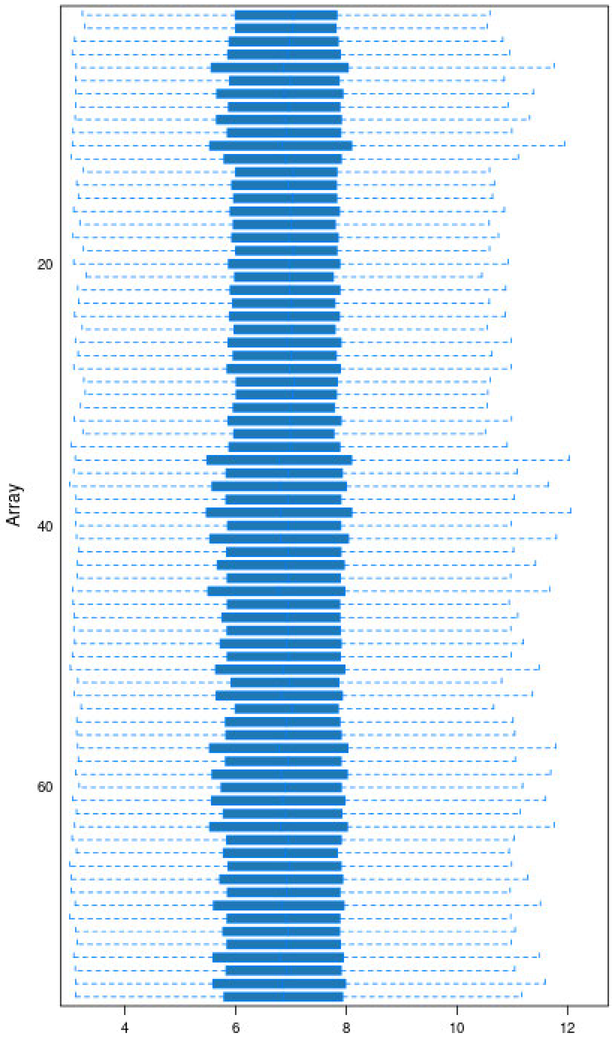


Figure 4 [(PDF file)] shows boxplots representing summaries of the signal intensity distributions of the arrays. Each box corresponds to one array. Typically, one expects the boxes to have similar positions and widths. If the distribution of an array is very different from the others, this may indicate an experimental problem. Outlier detection was performed by computing the Kolmogorov-Smirnov statistic K_a between each array's distribution and the distribution of the pooled data.

+ Figure 5: Outlier detection for Boxplots.

- Figure 6: Density plots.

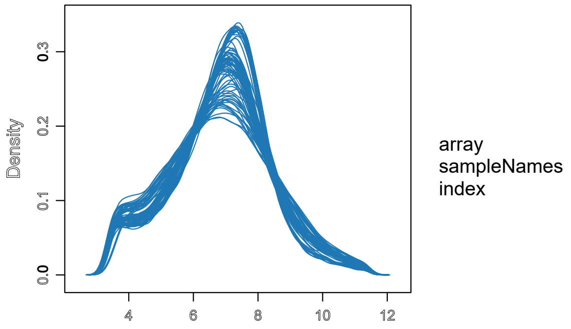


Figure 6 [(PDF file)] shows density estimates (smoothed histograms) of the data. Typically, the distributions of the arrays should have similar shapes and ranges. Arrays whose distributions are very different from the others should be considered for possible problems. Various features of the distributions can be indicative of quality related phenomena. For instance, high levels of background will shift an array's distribution to the right. Lack of signal diminishes its right right tail. A bulge at the upper end of the intensity range often indicates signal saturation.

Section 3: Variance mean dependence

- Figure 7: Standard deviation versus rank of the mean.

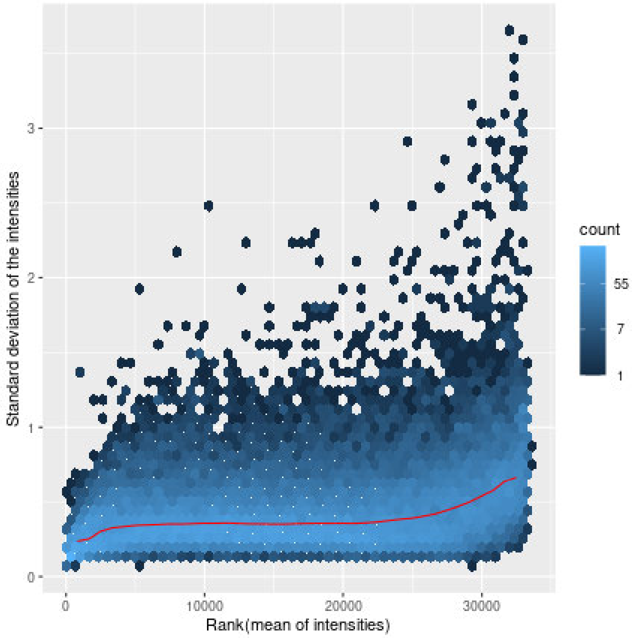


Figure 7 [(PDF file)] shows a density plot of the standard deviation of the intensities across arrays on the *y*-axis versus the rank of their mean on the *x*-axis. The red dots, connected by lines, show the running median of the standard deviation. After normalisation and transformation to a logarithm(-like) scale, one typically expects the red line to be approximately horizontal, that is, show no substantial trend. In some cases, a hump on the right hand of the x-axis can be observed and is symptomatic of a saturation of the intensities.

Section 4: Individual array quality

- Figure 8: MA plots.

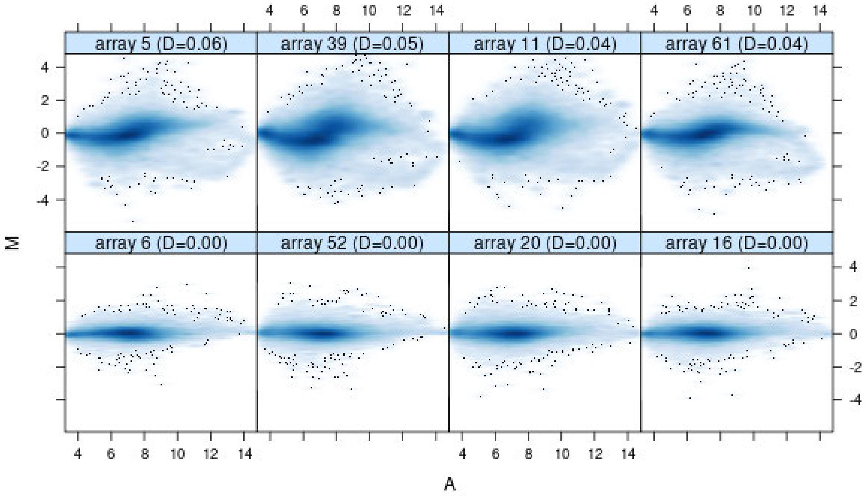


Figure 8 [(PDF file)] shows MA plots. M and A are defined as:
M = $\log_2(I_1) - \log_2(I_2)$
A = $1/2 (\log_2(I_1)+\log_2(I_2))$,
where I_1 is the intensity of the array studied,and I_2 is the intensity of a "pseudo"-array that consists of the median across arrays. Typically, we expect the mass of the distribution in an MA plot to be concentrated along the M = 0 axis, and there should be no trend in M as a function of A. If there is a trend in the lower range of A, this often indicates that the arrays have different background intensities; this may be addressed by background correction. A trend in the upper range of A can indicate saturation of the measurements; in mild cases, this may be addressed by non-linear normalisation (e.g. quantile normalisation).
Outlier detection was performed by computing Hoeffding's statistic D_a on the joint distribution of A and M for each array. Shown are first the 4 arrays with the highest values of D_a, then the 4 arrays with the lowest values. The value of D_a is shown in the panel headings. 0 arrays had D_a>0.15 and were marked as outliers. For more information on Hoeffing's *D*-statistic, please see the manual page of the function `hoeffd` in the `Hmisc` package.

+ Figure 9: Outlier detection for MA plots.

This report has been created with arrayQualityMetrics 3.42.0 under R version 3.6.0 (2019-04-26).

(Page generated on Tue Oct 6 00:36:48 2020 by hwriter)

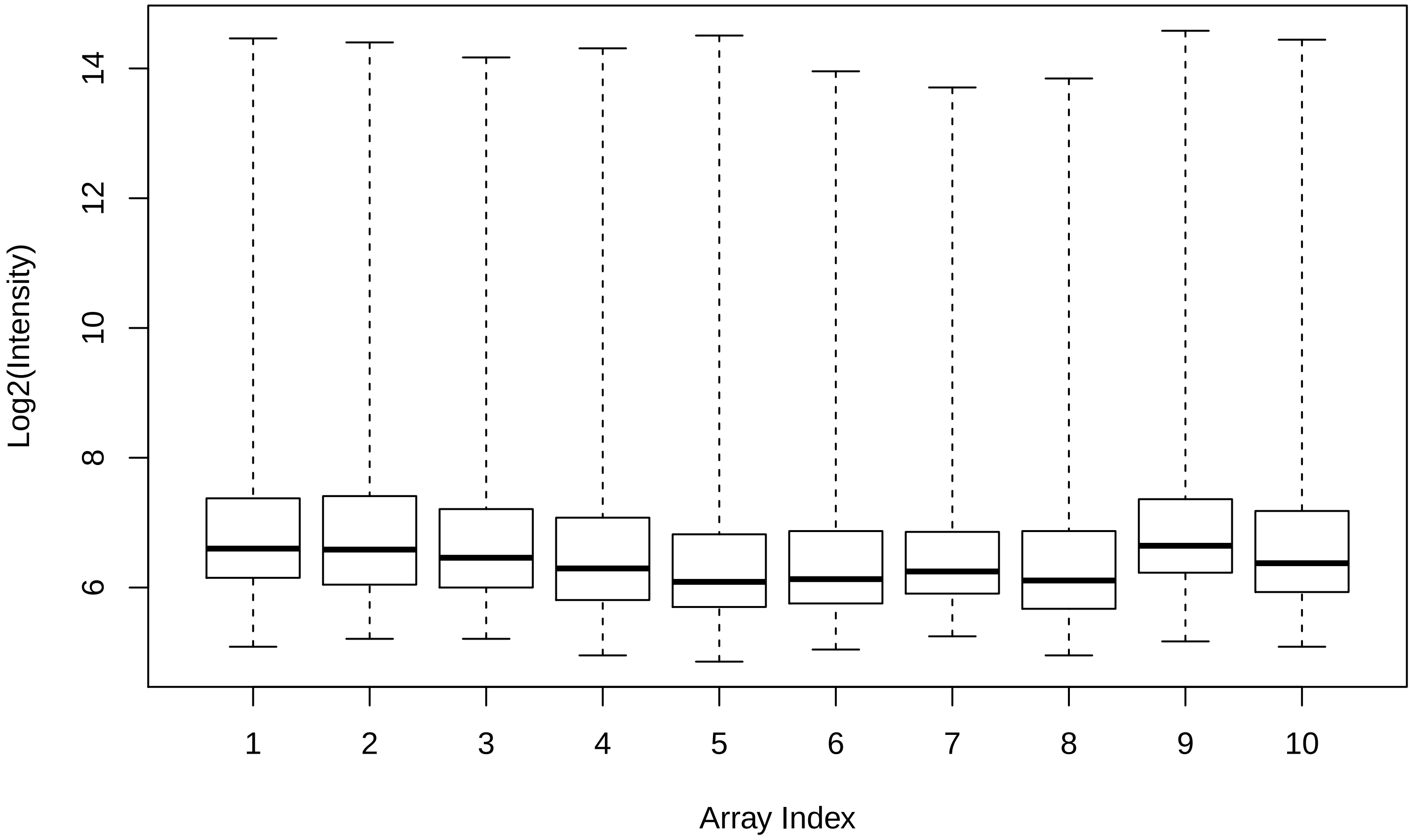
AffyBatch QC Report

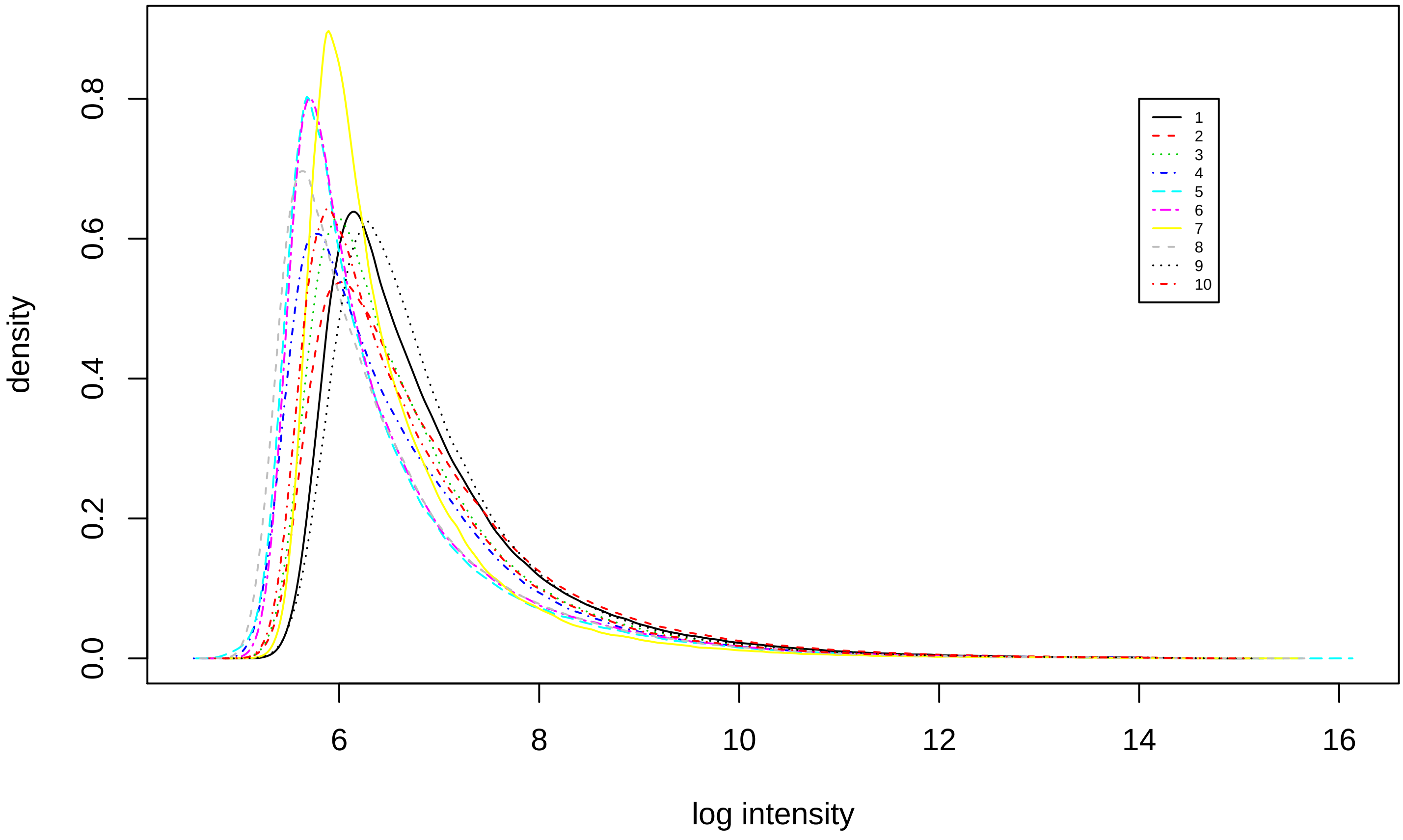
GSE100942

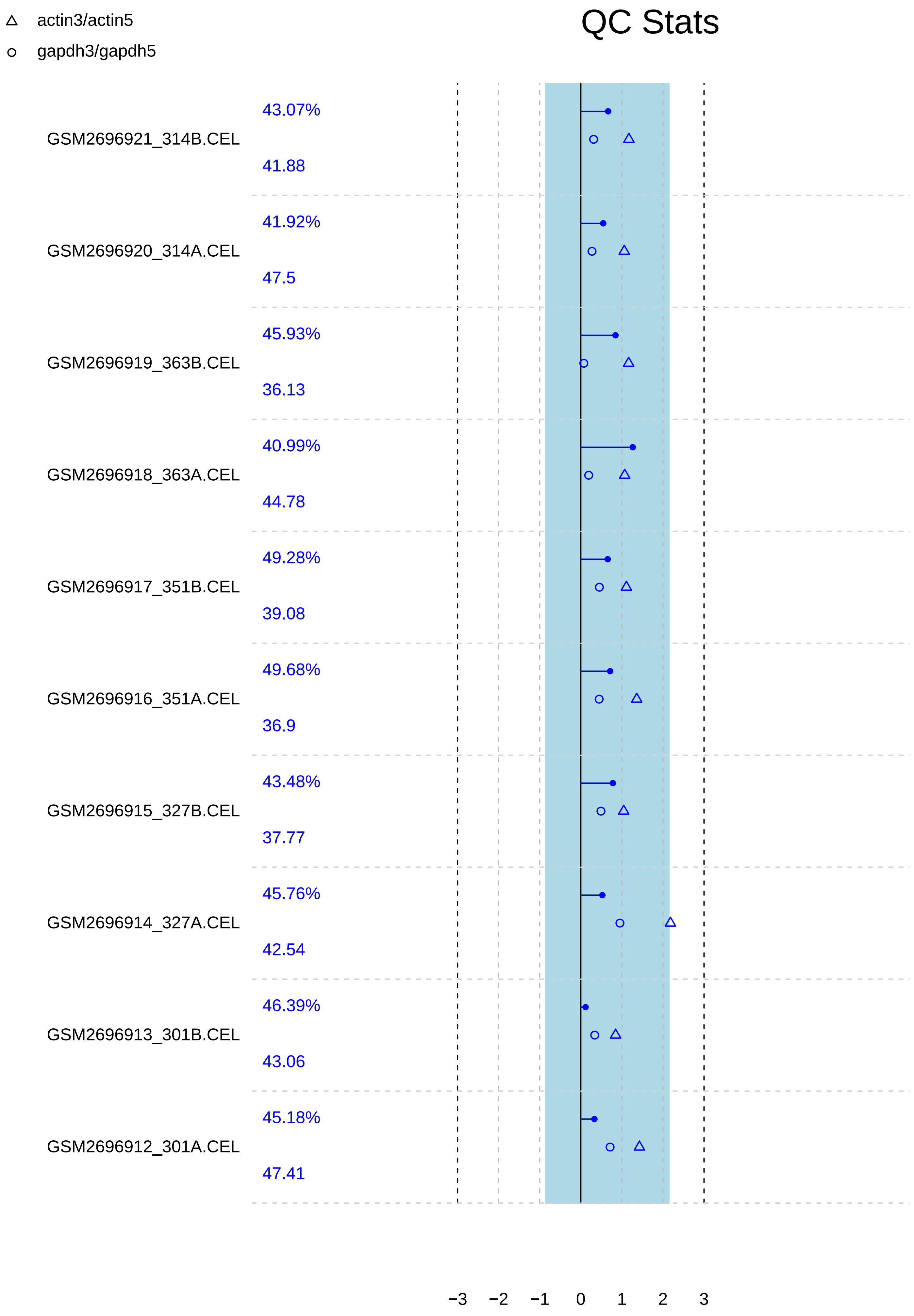
Array Index	Array Name
1	GSM2696912_301A.CEL
2	GSM2696913_301B.CEL
3	GSM2696914_327A.CEL
4	GSM2696915_327B.CEL
5	GSM2696916_351A.CEL
6	GSM2696917_351B.CEL
7	GSM2696918_363A.CEL
8	GSM2696919_363B.CEL
9	GSM2696920_314A.CEL
10	GSM2696921_314B.CEL

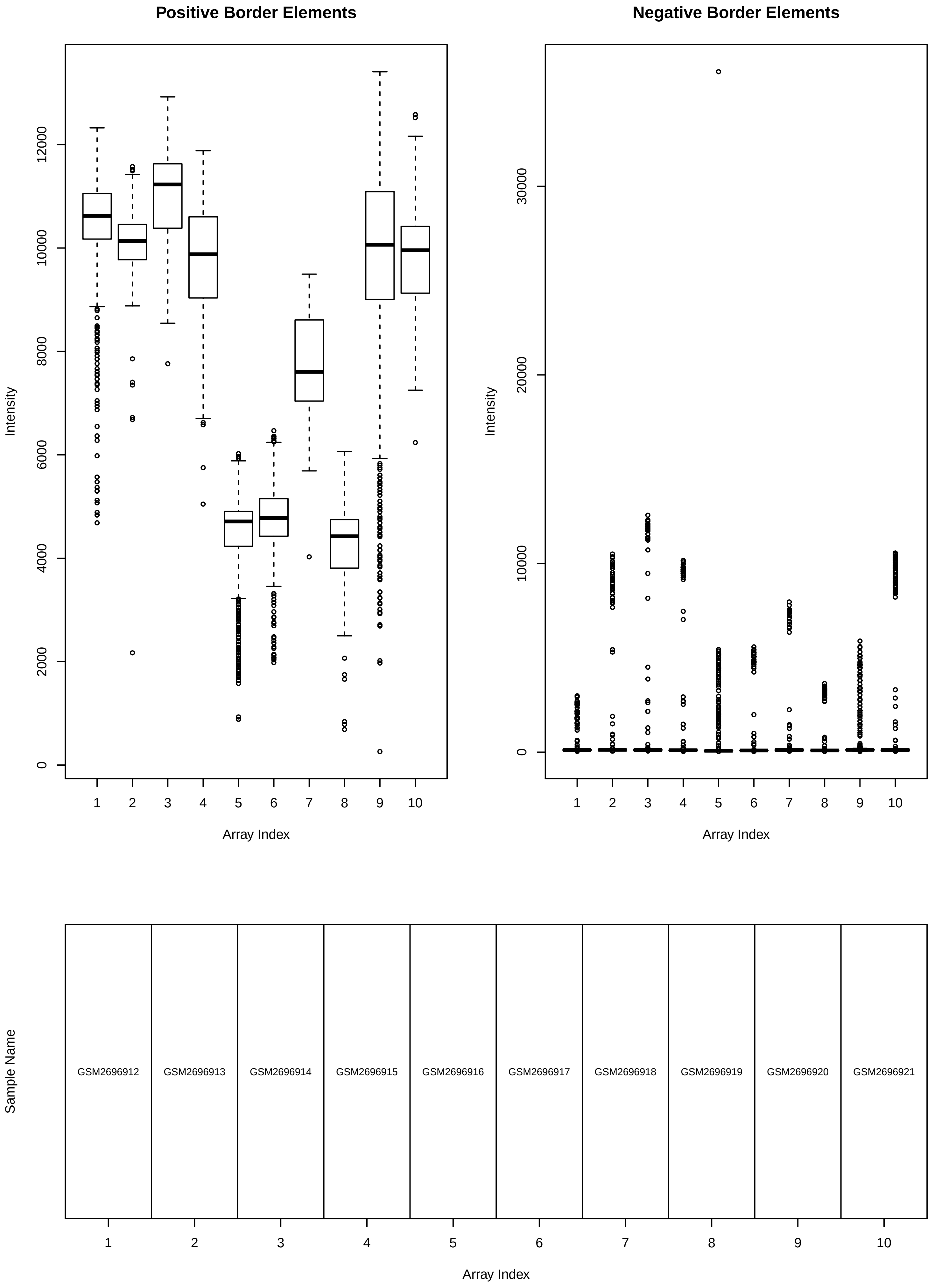
Thu Oct 1 15 38:35 2020

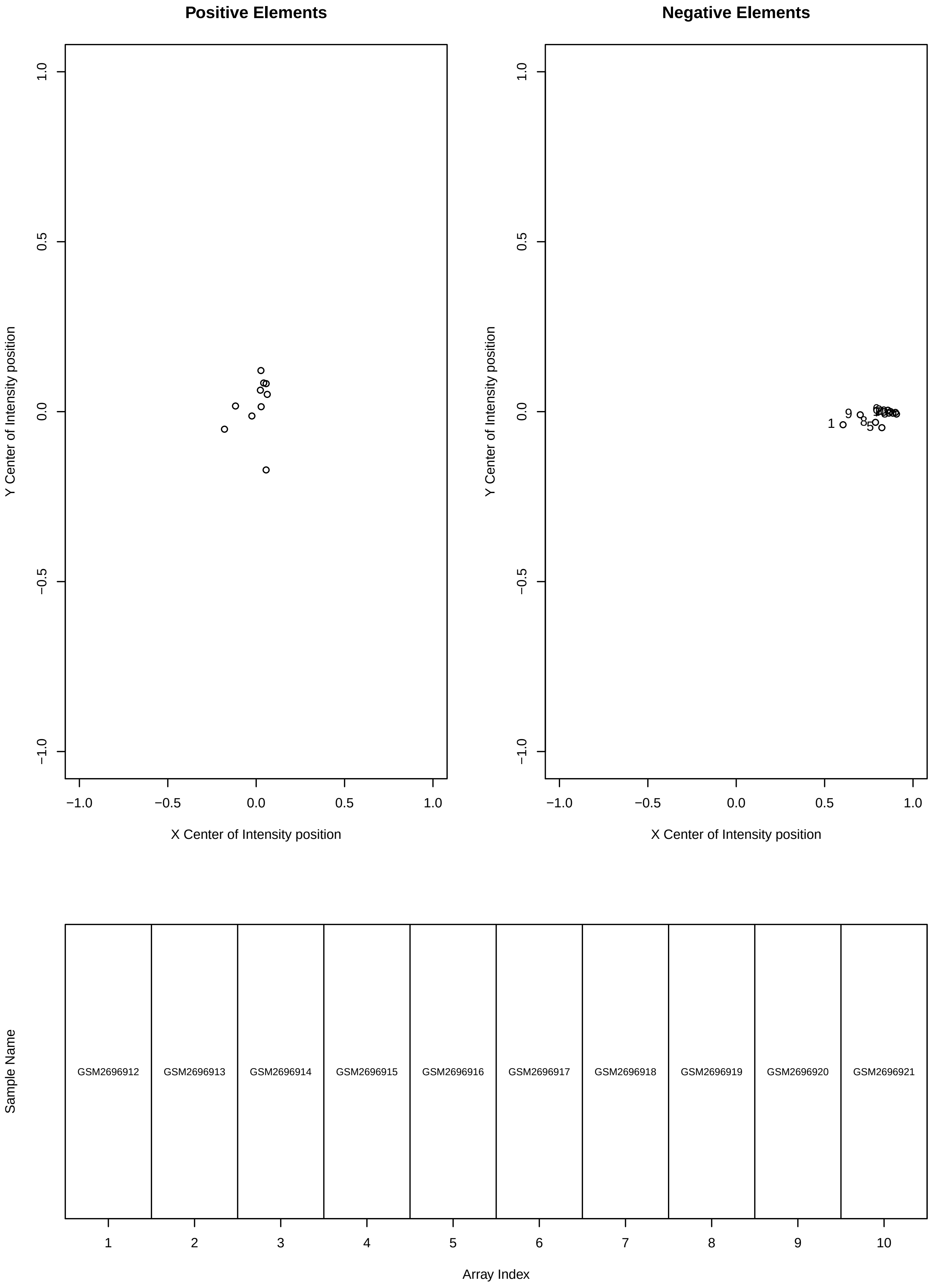
Produced by AffyQCReport R Package

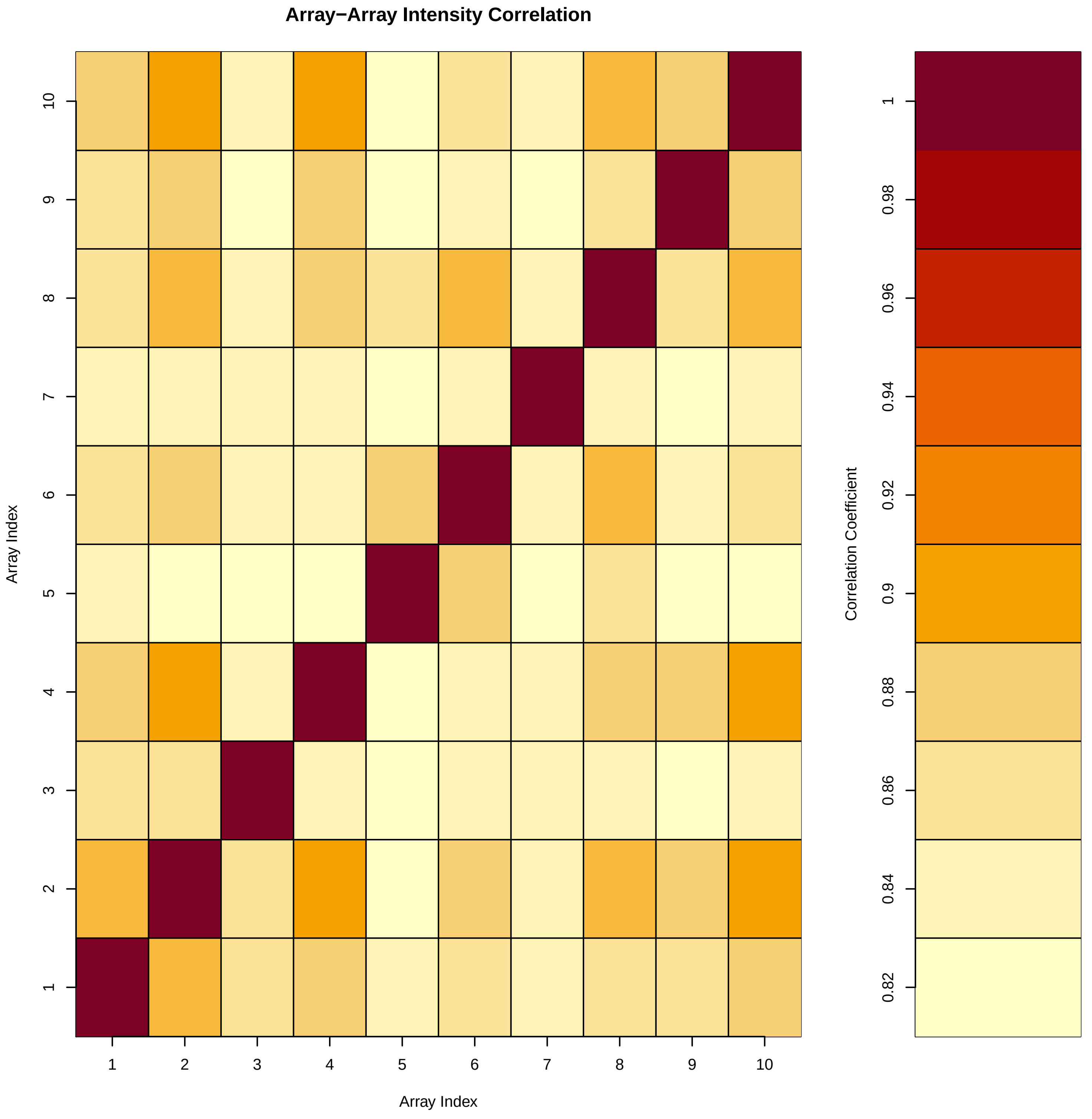

Log2(Intensity)
6
8
10
12
14
1
2
3
4
5
6
7
8
9
10
Array Index

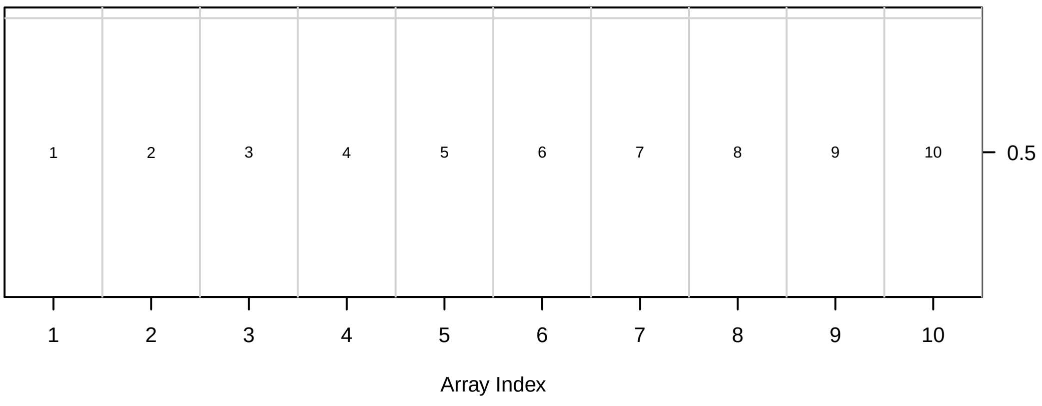

density
0.0
0.2
0.4
0.6
0.8
6
8
10
12
14
16
log intensity
1
2
3
4
5
6
7
8
9
10


△ actin3/actin5
○ gapdh3/gapdh5
QC Stats
GSM2696921_314B.CEL
43.07%
41.88
GSM2696920_314A.CEL
41.92%
47.5
GSM2696919_363B.CEL
45.93%
36.13
GSM2696918_363A.CEL
40.99%
44.78
GSM2696917_351B.CEL
49.28%
39.08
GSM2696916_351A.CEL
49.68%
36.9
GSM2696915_327B.CEL
43.48%
37.77
GSM2696914_327A.CEL
45.76%
42.54
GSM2696913_301B.CEL
46.39%
43.06
GSM2696912_301A.CEL
45.18%
47.41
−3
−2
−1
0
1
2
3


Positive Border Elements
Negative Border Elements
Intensity
0
2000
4000
6000
8000
10000
12000
Intensity
0
10000
20000
30000
1 2 3 4 5 6 7 8 9 10
Array Index
1 2 3 4 5 6 7 8 9 10
Array Index
Sample Name
GSM2696912
GSM2696913
GSM2696914
GSM2696915
GSM2696916
GSM2696917
GSM2696918
GSM2696919
GSM2696920
GSM2696921
1 2 3 4 5 6 7 8 9 10
Array Index


Positive Elements
Negative Elements
Y Center of Intensity position
X Center of Intensity position
1.0
0.5
0.0
−0.5
−1.0
1
9
8
5
Sample Name
GSM2696912
GSM2696913
GSM2696914
GSM2696915
GSM2696916
GSM2696917
GSM2696918
GSM2696919
GSM2696920
GSM2696921
1
2
3
4
5
6
7
8
9
10
Array Index


Array−Array Intensity Correlation
Array Index
Array Index
1
2
3
4
5
6
7
8
9
10
Correlation Coefficient
1
0.98
0.96
0.94
0.92
0.9
0.88
0.86
0.84
0.82


1
2
3
4
5
6
7
8
9
10
0.5
Array Index

Appendix Table 4A2

Table 4A2: Pairwise contrasts for EAC, BE and ESCC

		EAC model	BE models								ESCC models				
Dataset	Country	EAC	BE1	BE2	BE3	BE4	BE5	BE6	BE7	BE8*	ESCC1	ESCC3	ESCC4**	ESCC6	ESCC8
GSE45670	China										✓				✓
GSE77861	USA										✓				✓
GSE100942	Hong Kong										✓				✓
GSE17351	USA and Japan										✓				✓
GSE13083	USA			✓	✓	✓		✓	✓						
GSE1420	USA	✓													
GSE20347	China											✓			
GSE29001	China											✓			
GSE33426	China											✓		✓	
GSE36223	Poland		✓		✓		✓	✓		✓					
GSE38129	China											✓		✓	
GSE39491	USA		✓		✓		✓								
GSE23400	China												✓		

GSE23400B	China										✓		
GSE26886	Germany	✓	✓		✓		✓	✓	✓	✓			✓
GSE92396	USA	✓											
GSE100843	USA			✓	✓			✓	✓				
GSE34619	UK			✓	✓	✓	✓		✓				
GSE75241	Brazil											✓	✓

*Dataset GSE100843 was subsampled

**Both GSE23400 and GSE2344B were subsampled

Appendix Table 4A3

Table 4A3: All enriched pathways present and respective p-values in the analysis for BE, EAC and ESCC

	Datasets	EAC	BE1	BE2	BE3	BE4	BE5	BE6	BE7	BE9sub50	ESCC1	ESCC3	ESCC4sub80	ESCC6	ESCC8	
	Number of platforms	**2**	**3**	**3**	**3**	**4**	**3**	**4**	**3**	**4**	**5**	**4**	**2**	**3**	**6**	
Pathway Name	**Sample size**	**143**	**205**	**123**	**185**	**162**	**194**	**132**	**134**	**163**	**98**	**192**	**160**	**143**	**128**	**Contrast count**
Formation of the cornified envelope		2,23E-07	1,40E-05	4,86E-06	4,44E-04	1,35E-06	8,97E-03	3,86E-07	1,81E-07	1,79E-07		1,15E-03	3,26E-03	2,80E-03	3,70E-03	13
Smooth Muscle Contraction		9,00E-04	5,17E-05		7,03E-04	6,55E-03	1,00E-04	1,37E-04	5,94E-03	6,26E-05		3,95E-03	2,00E-05			10
Cytosolic sulfonation of small molecules			2,54E-04	7,72E-04	3,34E-04	1,77E-03	4,46E-04	2,32E-04	7,06E-04	6,11E-04						8
Assembly of collagen fibrils and other multimeric structures			4,08E-03		3,59E-03	3,28E-03		4,49E-03		1,19E-03	2,83E-03		3,28E-03	7,72E-04	1,39E-03	9
Transport of inorganic cations/anions and amino acids/oligopeptides		5,27E-04	6,17E-03	6,85E-03		7,49E-05		9,01E-03	3,27E-04	2,20E-05						7
BMAL1:CLOCK,NPAS2 activates circadian gene expression			3,27E-03			6,50E-03	6,35E-03	3,72E-03	7,43E-03	3,70E-03			2,56E-03			7
Collagen chain trimerization			9,05E-03		4,02E-03		1,49E-03	8,23E-03			1,92E-04	5,38E-05	1,04E-04	2,45E-05	7,87E-04	9
Defective GALNT3 causes familial hyperphosphatemic tumoral calcinosis (HFTC)		8,87E-03	4,40E-03	4,10E-03		3,84E-03	7,19E-03		7,65E-03	6,97E-03						7
Type I hemidesmosome assembly				1,97E-03				2,75E-03		9,90E-04	1,08E-03		7,35E-03	2,56E-03	5,66E-04	7
Neutrophil degranulation				6,17E-04		4,18E-03			3,11E-03	1,23E-03			5,16E-03			5

O-linked glycosylation of mucins				9,81E-03	3,82E-03		4,08E-04	7,05E-03	9,63E-03		8,77E-03				6
Nicotinate metabolism	5,06E-03		2,98E-03		6,26E-03			5,76E-03	6,44E-03						5
Glycogen storage diseases	7,11E-03		9,58E-03							1,31E-03	2,76E-03		1,27E-03	1,20E-03	6
Metabolism of Angiotensinogen to Angiotensins		9,40E-03	5,47E-03	5,98E-03	6,40E-03	4,32E-03			8,71E-03						6
Tight junction interactions		8,06E-03			7,68E-03		7,56E-03	5,10E-03	6,97E-03						5
Hormone-sensitive lipase (HSL)-mediated triacylglycerol hydrolysis			7,16E-03	7,47E-03	3,38E-03	7,19E-03		1,10E-03							5
TP53 regulates transcription of several additional cell death genes whose specific roles in p53-dependent apoptosis remain uncertain					9,38E-03		8,32E-03	6,23E-03	3,85E-03						4
Cargo concentration in the ER	7,28E-03	3,90E-03		3,23E-03		6,49E-04	2,46E-03								5
PPARA activates gene expression		3,03E-03		3,92E-03		4,90E-03	1,66E-03		7,11E-03						5
Antagonism of Activin by Follistatin										3,29E-03	6,07E-03	3,90E-05	6,33E-03	1,80E-03	5
Extracellular matrix organization	4,51E-03									1,73E-04		1,88E-03	3,58E-03		4
Surfactant metabolism		2,69E-03		4,49E-03		6,48E-03	3,15E-03								4
Glucose metabolism			5,77E-03		4,17E-03		5,17E-03		8,77E-03						4
Cyclin B2 mediated events										7,37E-03	2,78E-03		2,67E-03	5,55E-03	4
Glycoprotein hormones										3,29E-03	6,07E-03		6,33E-03	1,80E-03	4
Integrin cell surface interactions										4,43E-03	7,40E-04	5,65E-04	1,03E-03		4
Arachidonate production from DAG											8,28E-03	5,95E-03	5,70E-03	6,58E-03	4

Detoxification of Reactive Oxygen Species		9,85E-03		2,25E-03		2,41E-03									3
Fructose catabolism		8,75E-03		7,48E-03		8,65E-03									3
Metabolism of water-soluble vitamins and cofactors			8,13E-03								7,61E-03			1,08E-03	3
Glycosaminoglycan metabolism					2,77E-04						3,28E-03			2,78E-03	3
Ketone body metabolism					8,33E-03			8,94E-03							2
RAB geranylgeranylation					5,06E-03			4,57E-03							2
ECM proteoglycans										4,43E-03		5,65E-04		3,58E-03	3
Metabolism of amino acids and derivatives										3,17E-03	6,46E-03		7,98E-03		3
Apoptotic cleavage of cell adhesion proteins	3,56E-03						7,57E-03								2
Interleukin-4 and 13 signaling	9,26E-04											9,88E-03			2
Glycosphingolipid metabolism		3,24E-03					8,35E-03								2
Heparan sulfate/heparin (HS-GAG) metabolism						8,17E-03								2,88E-03	2
Cell-Cell communication								6,23E-03							1
Rho GTPase cycle								1,13E-03							1
Glycolysis								9,95E-03							1
Defective CHST6 causes MCDC1										3,60E-03		2,96E-03			2
Insulin-l ke Growth Factor-2 mRNA Binding Proteins (IGF2BPs/IMPs/VICKZs) bind RNA										4,60E-03				5,39E-03	2
Metabolism											2,98E-03			4,88E-03	2
Phase 1 - Functionalization of compounds											4,49E-03			3,46E-03	2

Signaling by MET											3,00E-03	2,76E-03			2
G2/M Checkpoints											1,15E-04		1,50E-03		2
TFAP2 (AP-2) family regulates transcription of other transcription factors											9,39E-03		8,23E-03		2
Peptide hormone metabolism	8,10E-03														1
RHO GTPases activate IQGAPs	8,02E-03														1
Diseases associated with glycosaminoglycan metabolism		8,29E-03													1
Translocation of GLUT4 to the plasma membrane			5,69E-03												1
PI3K/AKT activation			8,88E-03												1
Regulation of gene expression in beta cells			7,62E-03												1
FCERI mediated Ca+2 mobilization			3,68E-03												1
Amino acid transport across the plasma membrane			5,10E-03												1
Striated Muscle Contraction			8,57E-03												1
GPCR downstream signaling				2,97E-03											1
GPCR ligand binding				6,78E-03											1
Aflatoxin activation and detoxification				3,72E-03											1
Common Pathway of Fibrin Clot Formation						4,83E-03									1
Scavenging by Class A Receptors						1,49E-03									1
G0 and Early G1										6,66E-03					1
Signal Transduction										4,67E-03					1

Terminal pathway of complement	9,20E-03			1
Diseases of glycosylation	4,07E-03			1
RHO GTPases activate CIT	6,74E-03			1
RHO GTPases Activate ROCKs	6,81E-03			1
RHO GTPases activate PAKs	6,81E-03			1
Activation of the pre-replicative complex	9,32E-04			1
Removal of licensing factors from origins	9,32E-04			1
Ethanol oxidation	2,63E-03			1
Interferon gamma signaling	1,91E-03			1
E2F-enabled inhibition of pre-replication complex formation		7,41E-03		1
Metabolism of ingested SeMet, Sec, MeSec into H2Se		2,11E-03		1
Homologous DNA Pairing and Strand Exchange		9,63E-03		1
Amino acid synthesis and interconversion (transamination)		2,17E-03		1
Interferon alpha/beta signaling		9,81E-03		1
Phosphorylation of Emi1			8,14E-03	1
Transmembrane transport of small molecules			4,76E-03	1
Depolymerisation of the Nuclear Lamina			8,14E-03	1
M Phase			8,53E-03	1

Chondroitin sulfate/dermatan sulfate metabolism	2,74E-03		1
Biological oxidations	7,98E-03		1
Resolution of Sister Chromatid Cohesion	3,79E-03		1
Fatty acid, triacylglycerol, and ketone body metabolism	8,67E-03		1
Metabolism of lipids and lipoproteins	6,36E-03		1
Purine metabolism	1,92E-03		1
The role of GTSE1 in G2/M progression after G2 checkpoint	9,86E-03		1
Retinoid metabolism and transport	5,11E-03		1
Activation of ATR in response to replication stress		3,43E-03	1
Membrane Trafficking		3,92E-03	1
Keratan sulfate biosynthesis		4,10E-03	1
Vesicle-mediated transport		3,87E-03	1
Chk1/Chk2(Cds1) mediated inactivation of Cyclin B:Cdk1 complex		8,92E-03	1
Growth hormone receptor signaling		8,55E-03	1

Appendix Table 4A4

Table 4A4: Barrett's Esophagus mapping of enriched pathways to differentially expressed genes

Neutrophil degranulation	Glycosaminoglycan metabolism	Transport of inorganic cations/anions and amino acids/oligopeptides	Smooth Muscle Contraction	Formation of the cornified envelope	BMAL1:CLOCK,NPAS2 activates circadian gene expression	RAB geranylgeranylation	Glucose metabolism	Assembly of collagen fibrils and other multimeric structures	O-linked glycosylation of mucins	Hormone-sensitive lipase (HSL)-mediated triacylglycerol hydrolysis	Cytosolic sulfonation of small molecules	Nicotinate metabolism	TP53 regulates transcription of several additional cell death genes whose specific roles in p53-dependent apoptosis remain uncertain	Defective GALNT3 causes familial hyperphosphatemic tumoral calcinosis (HFTC)	Tight junction interactions	Metabolism of Angiotensinogen to Angiotensins	Ketone body metabolism			
R.HSA.6798695	R.HSA.1630316	R.HSA.425393	R.HSA.445355	R.HSA.6809371	R.HSA.1368108	R.HSA.8873719	R.HSA.70326	R.HSA.2022090	R.HSA.913709	R.HSA.163560	R.HSA.156584	R.HSA.196807	R.HSA.6803205	R.HSA.5083625	R.HSA.420029	R.HSA.2022377	R.HSA.74182	GENENAME	SYMBOL	logFC
.	.	X	X	.	.	.	X	.	.	.	.	.	.	.	.	.	.	calmodulin 1	CALM1	0.72
X	X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	heparanase	HPSE	0.90
X	X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	glucuronidase beta	GUSB	-0.59
X	X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	CD44 molecule (Indian blood group)	CD44	0.27

X	X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	beta-1,4-galactosyltransferase 1	B4GALT1	0.37
.	X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	solute carrier family 35 member D2	SLC35D2	-0.85
.	X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	heparan sulfate-glucosamine 3-sulfotransferase 3B1	HS3ST3B1	0.37
.	X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	carbohydrate sulfotransferase 6	CHST6	-0.99
.	X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	lumican	LUM	-1.55
.	X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	iduronate 2-sulfatase	IDS	0.53
.	X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	hexosaminidase subunit alpha	HEXA	-0.78
.	X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	decorin	DCN	-0.63
.	X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	exostosin glycosyltransferase 1	EXT1	-0.86
.	.	X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	solute carrier family 38 member 1	SLC38A1	-0.11
.	.	X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	solute carrier family 26 member 6	SLC26A6	-0.84
.	.	X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	solute carrier family 6 member 15	SLC6A15	0.71
.	.	X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	solute carrier family 6 member 20	SLC6A20	-1.84
.	.	X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	solute carrier family 38 member 2	SLC38A2	0.79
.	.	X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	solute carrier family 17 member 5	SLC17A5	-0.64
.	.	X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	solute carrier family 12 member 6	SLC12A6	0.92

.	.	X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	solute carrier family 20 member 1	SLC20A1	-0.88
.	.	X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	solute carrier family 15 member 2	SLC15A2	0.47
.	.	X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	solute carrier family 12 member 2	SLC12A2	-1.97
.	.	X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	solute carrier family 7 member 2	SLC7A2	0.74
.	.	X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	solute carrier family 1 member 4	SLC1A4	0.35
.	.	X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	solute carrier family 3 member 1	SLC3A1	-2.21
.	X	X	.	.	.	.	.	.	.	.	X	.	.	.	.	.	.	solute carrier family 26 member 2	SLC26A2	1.04
.	X	X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	solute carrier family 9 member A1	SLC9A1	-1.30
.	.	.	.	X	.	.	.	.	.	.	.	.	X	.	.	.	.	p53 apoptosis effector related to PMP22	PERP	1.10
.	.	.	.	.	.	.	.	.	.	.	.	.	X	.	.	.	.	BCL6 transcription repressor	BCL6	0.80
.	.	.	.	.	.	.	.	.	.	.	.	.	X	.	.	.	.	BCL2 like 14	BCL2L14	-1.37
X	.	.	.	.	.	X	.	.	.	.	.	.	.	.	.	.	.	RAB27A, member RAS oncogene family	RAB27A	-0.79
X	.	.	.	.	.	X	.	.	.	.	.	.	.	.	.	.	.	RAB7A, member RAS oncogene family	RAB7A	0.60
.	.	.	.	.	.	X	.	.	.	.	.	.	.	.	.	.	.	RAB20, member RAS oncogene family	RAB20	-1.15
.	.	.	.	.	.	X	.	.	.	.	.	.	.	.	.	.	.	RAB8B, member RAS	RAB8B	-0.32

																		oncogene family		
.	.	.	.	.	.	X	.	.	.	.	.	.	.	.	.	.	.	RAB21, member RAS oncogene family	RAB21	0.64
.	.	.	.	.	.	X	.	.	.	.	.	.	.	.	.	.	.	RAB29, member RAS oncogene family	RAB29	-0.21
.	.	.	.	.	.	X	.	.	.	.	.	.	.	.	.	.	.	RAB11A, member RAS oncogene family	RAB11A	0.83
.	.	.	.	.	.	X	.	.	.	.	.	.	.	.	.	.	.	RAB27B, member RAS oncogene family	RAB27B	-0.54
.	.	.	.	.	.	X	.	.	.	.	.	.	.	.	.	.	.	RAB3B, member RAS oncogene family	RAB3B	-1.22
.	.	.	.	.	.	X	.	.	.	.	.	.	.	.	.	.	.	RAB5A, member RAS oncogene family	RAB5A	0.47
X	.	.	.	X	.	.	.	.	.	.	.	.	.	.	.	.	.	calpain 1	CAPN1	0.83
.	.	.	.	X	.	.	.	.	.	.	.	.	.	.	.	.	.	kallikrein related peptidase 12	KLK12	1.77
.	.	.	.	X	.	.	.	.	.	.	.	.	.	.	.	.	.	kazrin, periplakin interacting protein	KAZN	1.34
.	.	.	.	X	.	.	.	.	.	.	.	.	.	.	.	.	.	kallikrein related peptidase 8	KLK8	1.39
.	.	.	.	X	.	.	.	.	.	.	.	.	.	.	.	.	.	plakophilin 4	PKP4	-0.32
.	.	.	.	X	.	.	.	.	.	.	.	.	.	.	.	.	.	small proline rich protein 3	SPRR3	3.19
.	.	.	.	X	.	.	.	.	.	.	.	.	.	.	.	.	.	desmoglein 3	DSG3	2.85
.	.	.	.	X	.	.	.	.	.	.	.	.	.	.	.	.	.	desmocollin 3	DSC3	3.03
.	.	.	.	X	.	.	.	.	.	.	.	.	.	.	.	.	.	calpain small subunit 1	CAPNS1	0.65
.	.	.	.	X	.	.	.	.	.	.	.	.	.	.	.	.	.	desmocollin 2	DSC2	1.18

.	.	.	.	.	.	.	.	X	.	.	.	.	.	.	.	.	.	lysyl oxidase like 2	LOXL2	-0.62
.	.	.	.	.	.	.	.	X	.	.	.	.	.	.	.	.	.	laminin subunit alpha 3	LAMA3	-1.22
.	.	.	.	.	.	.	.	X	.	.	.	.	.	.	.	.	.	integrin subunit beta 4	ITGB4	-1.11
.	.	.	.	.	.	.	.	X	.	.	.	.	.	.	.	.	.	integrin subunit alpha 6	ITGA6	-1.42
.	.	.	.	.	.	.	.	X	.	.	.	.	.	.	.	.	.	collagen type IV alpha 5 chain	COL4A5	-0.15
.	.	.	.	.	.	.	.	X	.	.	.	.	.	.	.	.	.	collagen type III alpha 1 chain	COL3A1	-1.09
.	.	.	.	.	.	.	.	X	.	.	.	.	.	.	.	.	.			0.70
.	.	.	.	.	.	.	.	X	.	.	.	.	.	.	.	.	.	dystonin	DST	0.70
.	.	.	.	.	.	.	.	X	.	.	.	.	.	.	.	.	.	collagen type I alpha 2 chain	COL1A2	-1.04
X	.	.	.	.	.	.	.	X	.	.	.	.	.	.	.	.	.	cathepsin B	CTSB	0.51
X	.	.	.	.	.	.	.	X	.	.	.	.	.	.	.	.	.	cathepsin S	CTSS	-1.34
X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	X	.	alanyl aminopeptidase, membrane	ANPEP	-3.46
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	X	.	aminopeptidase O (putative)	AOPEP	0.48
.	.	.	.	.	.	.	X	.	.	X	.	.	.	.	.	.	.	protein kinase cAMP-activated catalytic subunit beta	PRKACB	-0.69
.	.	.	.	.	.	.	.	.	.	X	.	.	.	.	.	.	.	abhydrolase domain containing 5, lysophosphatidic acid acyltransferase	ABHD5	1.26
.	.	.	.	.	.	.	.	.	.	X	.	.	.	.	.	.	.	monoglyceride lipase	MGLL	0.42
.	.	.	.	.	.	.	.	.	.	X	.	.	.	.	.	.	.	fatty acid binding protein 1	FABP1	-2.76

.	.	.	.	.	.	.	.	.	.	X	.	.	.	.	.	.	.	protein phosphatase 1 catalytic subunit beta	PPP1CB	1.06
X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	magnesium transporter 1	MAGT1	-0.25
X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	tetraspanin 14	TSPAN14	0.78
X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	potassium channel modulatory factor 1	KCMF1	0.61
X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	transmembrane protein 30A	TMEM30A	0.29
X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	calcineurin like phosphoesterase domain containing 1	CPPED1	0.83
X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	ras homolog family member F, filopodia associated	RHOF	-0.71
X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	toll interacting protein	TOLLIP	0.55
X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	calcium binding protein 39	CAB39	0.43
X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	yippee l ke 5	YPEL5	0.28
X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	ATPase H+ transporting V1 subunit D	ATP6V1D	0.76
X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	heme binding protein 2	HEBP2	0.69
X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	ATPase phospholipid transporting 11B (putative)	ATP11B	0.62
X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	GLI pathogenesis related 1	GLIPR1	0.20
X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	cytoskeleton associated protein 4	CKAP4	0.23

X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	IQ motif containing GTPase activating protein 2	IQGAP2	-2.54
X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	proteasome 26S subunit, non-ATPase 6	PSMD6	0.40
X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	transmembrane protein 63A	TMEM63A	-0.95
X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	synaptosome associated protein 29	SNAP29	0.41
X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	VAMP associated protein A	VAPA	0.22
X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	synaptosome associated protein 23	SNAP23	-0.36
X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	late endosomal/lysosomal adaptor, MAPK and MTOR activator 3	LAMTOR3	0.29
X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	r bonuclease T2	RNASET2	-0.96
X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	X-ray repair cross complementing 5	XRCC5	0.13
X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	serine/threonine kinase 10	STK10	-0.35
X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	signal recognition particle 14	SRP14	0.28
X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	Rho associated coiled-coil containing protein kinase 1	ROCK1	-0.40
X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	RAP2B, member of RAS	RAP2B	0.39

																		oncogene family		
X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	protein tyrosine phosphatase receptor type B	PTPRB	-0.62
X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	proteasome 20S subunit beta 7	PSMB7	0.24
X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	proteasome 20S subunit alpha 5	PSMA5	-0.35
X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	DnaJ heat shock protein family (Hsp40) member C3	DNAJC3	-0.47
X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	mitogen-activated protein kinase 1	MAPK1	0.70
X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	prolylcarboxypeptidase	PRCP	0.92
X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	phospholipase D1	PLD1	0.96
X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	polymeric immunoglobulin receptor	PIGR	-3.00
X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	serpin family B member 6	SERPINB6	-0.64
X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	serpin family A member 1	SERPINA1	-2.74
X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	inositol 1,4,5-triphosphate receptor associated 2	IRAG2	0.27
X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	lysosomal associated membrane protein 2	LAMP2	0.70
X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	isocitrate dehydrogenase (NADP(+)) 1	IDH1	-0.38
X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	heat shock protein family A (Hsp70) member 8	HSPA8	0.15

X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	gelsolin	GSN	0.61
X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	GM2 ganglioside activator	GM2A	1.55
X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	alpha-L-fucosidase 1	FUCA 1	-1.25
X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	serpin family B member 1	SERP INB1	0.95
X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	cytochrome b-245 alpha chain	CYBA	-0.98
X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	cystatin B	CSTB	1.24
X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	COPI coat complex subunit beta 1	COPB 1	-0.42
X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	cathepsin C	CTSC	-0.56
X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	CD59 molecule (CD59 blood group)	CD59	0.61
X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	CD47 molecule	CD47	0.31
X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	CD36 molecule	CD36	0.01
X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	beta-2-microglobulin	B2M	-0.86
X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	ras homolog family member A	RHOA	0.16
X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	acid phosphatase 3	ACP3	1.55
X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	arginase 1	ARG1	0.68
X	.	.	.	.	.	.	X	.	.	.	.	.	.	.	.	.	.	aldolase, fructose-bisphosphate A	ALDO A	0.46
X	.	.	.	.	.	.	X	.	.	.	.	.	.	.	.	.	.	glycogen phosphorylase B	PYGB	-0.76
.	.	.	.	.	.	.	.	.	X	.	.	.	.	X	.	.	.	mucin 4, cell surface associated	MUC4	0.49
.	.	.	.	.	.	.	.	.	X	.	.	.	.	X	.	.	.	mucin 13, cell surface associated	MUC1 3	-3.87

.	X	.	.	.	.	.	.	.	X	.	.	.	.	.	.	.	.	ST3 beta-galactoside alpha-2,3-sialyltransferase 1	ST3GAL1	0.68
.	.	.	.	.	.	.	.	.	X	.	.	.	.	.	.	.	.	polypeptide N-acetylgalactosaminyltransferase 10	GALNT10	-1.43
.	.	.	.	.	.	.	.	.	X	.	.	.	.	.	.	.	.	polypeptide N-acetylgalactosaminyltransferase 7	GALNT7	-1.39
.	.	.	.	.	.	.	.	.	X	.	.	.	.	.	.	.	.	ST6 N-acetylgalactosaminide alpha-2,6-sialyltransferase 4	ST6GALNAC4	-0.31
.	.	.	.	.	.	.	.	.	X	.	.	.	.	.	.	.	.	polypeptide N-acetylgalactosaminyltransferase 6	GALNT6	-2.36
.	.	.	.	.	.	.	.	.	X	.	.	.	.	.	.	.	.	polypeptide N-acetylgalactosaminyltransferase 1	GALNT1	0.66
.	.	.	.	.	.	.	.	.	X	.	.	.	.	.	.	.	.	glucosaminyl (N-acetyl) transferase 1	GCNT1	-1.67
.	.	.	.	.	.	.	.	.	.	.	.	X	.	.	.	.	.	poly(ADP-r bose) polymerase family member 8	PARP8	-0.55
.	.	.	.	.	.	.	.	.	.	.	.	X	.	.	.	.	.	nicotinamide N-methyltransferase	NNMT	-0.78
.	.	.	.	.	.	.	.	.	.	.	.	X	.	.	.	.	.	5'-nucleotidase ecto	NT5E	-2.22

.	.	.	.	.	X	.	.	.	.	.	.	X	.	.	.	.	.	nicotinamide phosphoribosyltransferase	NAMPT	1.40
.	.	.	.	.	X	.	.	.	.	.	.	.	.	.	.	.	.	trimethylguanosine synthase 1	TGS1	-0.39
.	.	.	.	.	X	.	.	.	.	.	.	.	.	.	.	.	.	basic helix-loop-helix family member e41	BHLHE41	-1.72
.	.	.	.	.	X	.	.	.	.	.	.	.	.	.	.	.	.	aryl hydrocarbon receptor nuclear translocator like 2	ARNTL2	1.10
.	.	.	.	.	X	.	.	.	.	.	.	.	.	.	.	.	.	nuclear receptor corepressor 1	NCOR1	-0.25
.	.	.	.	.	X	.	.	.	.	.	.	.	.	.	.	.	.	clock circadian regulator	CLOCK	0.29
.	.	.	.	.	X	.	.	.	.	.	.	.	.	.	.	.	.	RAR related orphan receptor A	RORA	1.84
.	.	.	.	.	X	.	.	.	.	.	.	.	.	.	.	.	.	RNA binding motif protein 4	RBM4	0.27
.	.	.	.	.	X	.	.	.	.	.	.	.	.	.	.	.	.	peroxisome proliferator activated receptor alpha	PPARA	-0.31
.	.	.	.	.	X	.	.	.	.	.	.	.	.	.	.	.	.	nuclear receptor subfamily 3 group C member 1	NR3C1	0.63
.	.	.	.	.	X	.	.	.	.	.	.	.	.	.	.	.	.	period circadian regulator 1	PER1	0.60
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	X	3-hydroxy-3-methylglutaryl-CoA synthase 2	HMGCS2	-2.67
.	.	.	.	.	.	.	X	.	.	.	.	.	.	.	.	.	.	solute carrier family 25 member 13	SLC25A13	-0.36
.	.	.	.	.	.	.	X	.	.	.	.	.	.	.	.	.	.	protein phosphatase	PPP2CA	0.39

																		2 catalytic subunit alpha		
.	.	.	.	.	.	.	X	.	.	.	.	.	.	.	.	.	.	protein phosphatase 1 regulatory subunit 3C	PPP1R3C	2.44
.	.	.	.	.	.	.	X	.	.	.	.	.	.	.	.	.	.	phosphoglycerate kinase 1	PGK1	0.58
.	.	.	.	.	.	.	X	.	.	.	.	.	.	.	.	.	.	aldolase, fructose-bisphosphate B	ALDOB	-2.07
.	.	.	.	.	.	.	X	.	.	.	.	.	.	.	.	.	.	enolase 1	ENO1	0.36
.	.	.	X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	sorbin and SH3 domain containing 1	SORBS1	-0.08
.	.	.	X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	myosin light chain 9	MYL9	-0.14
.	.	.	X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	tropomyosin 2	TPM2	-0.36
.	.	.	X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	tropomyosin 1	TPM1	-1.91
.	.	.	X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	talin 1	TLN1	-0.23
.	.	.	X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	myosin light chain kinase	MYLK	-0.52
.	.	.	X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	myosin heavy chain 11	MYH11	-0.50
.	.	.	X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	caldesmon 1	CALD1	-0.34
.	.	.	X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	annexin A6	ANXA6	-0.19
.	.	.	X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	annexin A1	ANXA1	1.66
.	.	.	X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	actin alpha 2, smooth muscle	ACTA2	-1.14
.	.	.	X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	actin gamma 2, smooth muscle	ACTG2	-0.73
.	.	.	.	.	.	.	.	.	.	.	X	.	.	.	.	.	.	3'(2'), 5'-bisphosphate nucleotidase 1	BPNT1	-0.53
.	.	.	.	.	.	.	.	.	.	.	X	.	.	.	.	.	.	sulfotransferase family 1C member 2	SULT1C2	-3.99

.	.	.	.	.	.	.	.	.	.	.	X	.	.	.	.	.	.	sulfotransferase family 1E member 1	SULT 1E1	-1.27
.	.	.	.	.	.	.	.	.	.	.	X	.	.	.	.	.	.	sulfotransferase family 1A member 1	SULT 1A1	-1.25
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	X	.	.	claudin 1	CLDN 1	0.42
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	X	.	.	claudin 18	CLDN 18	-3.53

Appendix Table 4A5:

Table 4A5: Esophageal Adenocarcinoma map of enriched pathways to differentially expressed genes

Extracellular matrix organization	Transport of inorganic cations/anions and amino acids/oligopeptides	Formation of the cornified envelope	Smooth Muscle Contraction	Interleukin-4 and 13 signaling	Cargo concentration in the ER	Peptide hormone metabolism	RHO GTPases activate IQGAPs	Nicotinate metabolism	Defective GALNT3 causes familial hyperphosphatemic tumoral calcinosis (HFTC)	Glycogen storage diseases	Apoptotic cleavage of cell adhesion proteins			
R.HSA.1474244	**R.HSA.425393**	**R.HSA.6809371**	**R.HSA.445355**	**R.HSA.6785807**	**R.HSA.5694530**	**R.HSA.2980736**	**R.HSA.5626467**	**R.HSA.196807**	**R.HSA.5083625**	**R.HSA.3229121**	**R.HSA.351906**	**GENENAME**	**SYMBOL**	**logFC**
.	X	.	X	.	.	.	X	.	.	.	.	calmodulin 1	CALM1	0.89
.	.	.	X	X	.	.	.	.	.	.	.	annexin A1	ANXA1	2.12
.	.	.	X	.	.	.	.	.	.	.	.	sorbin and SH3 domain containing 1	SORBS1	-0.15
.	.	.	X	.	.	.	.	.	.	.	.	myosin light chain 9	MYL9	-0.31
.	.	.	X	.	.	.	.	.	.	.	.	tropomyosin 2	TPM2	-0.97
.	.	.	X	.	.	.	.	.	.	.	.	tropomyosin 1	TPM1	-1.59
.	.	.	X	.	.	.	.	.	.	.	.	myosin light chain kinase	MYLK	-0.49
.	.	.	X	.	.	.	.	.	.	.	.			0.14
.	.	.	X	.	.	.	.	.	.	.	.	myosin heavy chain 11	MYH11	0.14
.	.	.	X	.	.	.	.	.	.	.	.	caldesmon 1	CALD1	-0.55
.	.	.	X	.	.	.	.	.	.	.	.	actin alpha 2, smooth muscle	ACTA2	-0.77
.	.	.	X	.	.	.	.	.	.	.	.			0.19

.	.	.	X	.	.	.	.	.	.	.	.	actin gamma 2, smooth muscle	ACTG2	0.19
.	.	X	.	.	.	.	.	.	.	.	X	desmoglein 3	DSG3	3.44
.	.	X	.	.	.	.	.	.	.	.	.	p53 apoptosis effector related to PMP22	PERP	1.22
.	.	X	.	.	.	.	.	.	.	.	.	kallikrein related peptidase 12	KLK12	2.57
.	.	X	.	.	.	.	.	.	.	.	.	kazrin, periplakin interacting protein	KAZN	1.32
.	.	X	.	.	.	.	.	.	.	.	.	kallikrein related peptidase 8	KLK8	1.52
.	.	X	.	.	.	.	.	.	.	.	.	small proline rich protein 3	SPRR3	4.24
.	.	X	.	.	.	.	.	.	.	.	.	desmocollin 2	DSC2	1.73
.	.	X	.	.	.	.	.	.	.	.	.	desmocollin 3	DSC3	3.29
.	.	.	.	.	X	.	.	.	.	.	.	transmembrane p24 trafficking protein 2	TMED2	0.32
.	.	.	.	.	X	.	.	.	.	.	.	golgi SNAP receptor complex member 2	GOSR2	-0.49
.	.	.	.	.	X	.	.	.	.	.	.	serpin family A member 1	SERPINA1	-1.85
.	.	.	.	.	X	.	.	.	.	.	.	coagulation factor V	F5	-2.19
	.	.	.	.	X	.	.	.	.	.	.	CD59 molecule (CD59	CD59	0.80

												blood group)		
.	.	.	.	.	X	.	.	.	.	.	.	cathepsin C	CTSC	-0.51
.	.	.	.	.	.	.	X	.	.	.	.	CAP-Gly domain containing linker protein 1	CLIP1	1.45
.	.	.	.	.	.	.	X	.	.	.	.	IQ motif containing GTPase activating protein 2	IQGAP2	-1.45
.	.	.	.	.	.	.	.	.	X	.	.	mucin 4, cell surface associated	MUC4	1.17
.	.	.	.	.	.	.	.	.	X	.	.	mucin 13, cell surface associated	MUC13	-3.49
.	X	.	.	.	.	.	.	.	.	.	.	solute carrier family 26 member 6	SLC26A6	-0.77
.	X	.	.	.	.	.	.	.	.	.	.	solute carrier family 6 member 15	SLC6A15	0.88
.	X	.	.	.	.	.	.	.	.	.	.	solute carrier family 6 member 20	SLC6A20	-1.77
.	X	.	.	.	.	.	.	.	.	.	.	solute carrier family 38 member 2	SLC38A2	0.61
.	X	.	.	.	.	.	.	.	.	.	.	solute carrier family 12 member 6	SLC12A6	1.11
.	X	.	.	.	.	.	.	.	.	.	.	solute carrier family 20 member 1	SLC20A1	-1.04

.	X	.	.	.	.	.	.	.	.	.	.	solute carrier family 15 member 2	SLC15A2	0.82
.	X	.	.	.	.	.	.	.	.	.	.	solute carrier family 12 member 2	SLC12A2	-1.75
.	X	.	.	.	.	.	.	.	.	.	.	solute carrier family 9 member A1	SLC9A1	-0.38
.	X	.	.	.	.	.	.	.	.	.	.	solute carrier family 26 member 2	SLC26A2	1.16
.	X	.	.	.	.	.	.	.	.	.	.	solute carrier family 3 member 1	SLC3A1	-1.56
.	.	.	.	.	.	.	.	.	.	X	.	protein phosphatase 1 regulatory subunit 3C	PPP1R3C	2.79
.	.	.	.	.	.	.	.	X	.	.	.	nicotinamide phosphor b osyltransferase	NAMPT	0.94
.	.	.	.	.	.	.	.	X	.	.	.	nicotinamide N-methyltransferase	NNMT	-1.37
.	.	.	.	.	.	.	.	X	.	.	.	5'-nucleotidase ecto	NT5E	-1.78
.	.	.	.	.	.	X	.	.	.	.	.	aminopeptidase O (putative)	AOPEP	0.50
.	.	.	.	.	.	X	.	.	.	.	.	endoplasmic reticulum oxidoreductase 1 alpha	ERO1A	1.46

.	.	.	.	.	.	X	.	.	.	.	.	transcription factor 7 like 2	TCF7L2	-0.65
.	.	.	.	.	.	X	.	.	.	.	.	inhibin subunit beta A	INHBA	-1.39
.	.	.	.	.	.	X	.	.	.	.	.	alanyl aminopeptidase, membrane	ANPEP	-1.86
.	.	.	.	.	.	X	.	.	.	.	.	GATA binding protein 4	GATA4	-1.19
X	.	.	.	.	.	.	.	.	.	.	.	junctional adhesion molecule 3	JAM3	-0.50
X	.	.	.	.	.	.	.	.	.	.	.	junctional adhesion molecule 2	JAM2	0.43
X	.	.	.	.	.	.	.	.	.	.	.	prolyl 3-hydroxylase 2	P3H2	0.54
X	.	.	.	.	.	.	.	.	.	.	.	asporin	ASPN	-1.50
X	.	.	.	.	.	.	.	.	.	.	.	ADAM metallopeptidase with thrombospondin type 1 motif 5	ADAMTS5	-0.58
X	.	.	.	.	.	.	.	.	.	.	.	ADAM metallopeptidase with thrombospondin type 1 motif 1	ADAMTS1	-0.58
X	.	.	.	.	.	.	.	.	.	.	.	calcium/calmodulin dependent serine protein kinase	CASK	-0.58
X	.	.	.	.	.	.	.	.	.	.	.	thrombospondin 1	THBS1	-1.19
X	.	.	.	.	.	.	.	.	.	.	.	transforming growth factor beta 2	TGFB2	-0.71

X	.	.	.	.	.	.	.	.	.	.	.	presenilin 1	PSEN1	0.48
X	.	.	.	.	.	.	.	.	.	.	.	kallikrein related peptidase 7	KLK7	1.43
X	.	.	.	.	.	.	.	.	.	.	.	prolyl 4-hydroxylase subunit alpha 1	P4HA1	-0.51
X	.	.	.	.	.	.	.	.	.	.	.	matrix metallopeptidase 15	MMP15	-0.98
X	.	.	.	.	.	.	.	.	.	.	.	matrix metallopeptidase 11	MMP11	-1.14
X	.	.	.	.	.	.	.	.	.	.	.	lumican	LUM	-1.81
X	.	.	.	.	.	.	.	.	.	.	.	latent transforming growth factor beta binding protein 2	LTBP2	-0.62
X	.	.	.	.	.	.	.	.	.	.	.	lysyl oxidase like 2	LOXL2	-1.33
X	.	.	.	.	.	.	.	.	.	.	.	laminin subunit alpha 3	LAMA3	-0.52
X	.	.	.	.	.	.	.	.	.	.	.	integrin subunit beta 8	ITGB8	0.47
X	.	.	.	.	.	.	.	.	.	.	.	integrin subunit beta 4	ITGB4	-0.59
X	.	.	.	.	.	.	.	.	.	.	.	integrin subunit alpha 6	ITGA6	-0.69
X	.	.	.	.	.	.	.	.	.	.	.	tenascin C	TNC	0.24
X	.	.	.	.	.	.	.	.	.	.	.	EGF containing fibulin extracellular matrix protein 1	EFEMP1	0.42
X	.	.	.	.	.	.	.	.	.	.	.	fibrillin 1	FBN1	-1.21

X	.	.	.	.	.	.	.	.	.	.	.	decorin	DCN	-0.33
X	.	.	.	.	.	.	.	.	.	.	.	cathepsin S	CTSS	-0.56
X	.	.	.	.	.	.	.	.	.	.	.	collagen type XI alpha 1 chain	COL11A1	-0.57
X	.	.	.	.	.	.	.	.	.	.	.	collagen type III alpha 1 chain	COL3A1	-1.15
X	.	.	.	.	.	.	.	.	.	.	.	CD47 molecule	CD47	0.34
X	.	.	.	.	.	.	.	.	.	.	.	dystonin	DST	1.08
X	.	.	.	.	.	.	.	.	.	.	.	CD44 molecule (Indian blood group)	CD44	0.57
X	.	X	.	.	.	.	.	.	.	.	.	calpain 1	CAPN1	0.82
X	.	X	.	.	.	.	.	.	.	.	.	calpain small subunit 1	CAPNS1	0.73
X	.	.	.	X	.	.	.	.	.	.	.	collagen type I alpha 2 chain	COL1A2	-1.85
.	.	.	.	X	.	.	.	.	.	.	.	zinc finger E-box binding homeobox 1	ZEB1	-0.73
.	.	.	.	X	.	.	.	.	.	.	.	signal transducer and activator of transcription 3	STAT3	0.45
.	.	.	.	X	.	.	.	.	.	.	.	signal transducer and activator of transcription 1	STAT1	-0.96

.	.	.	.	X	.	.	.	.	.	.	.	SRY-box transcription factor 2	SOX2	1.94
.	.	.	.	X	.	.	.	.	.	.	.	RAR related orphan receptor A	RORA	1.51
.	.	.	.	X	.	.	.	.	.	.	.	MCL1 apoptosis regulator, BCL2 family member	MCL1	0.54
.	.	.	.	X	.	.	.	.	.	.	.	BCL6 transcription repressor	BCL6	0.35
.	.	.	.	X	.	.	.	.	.	.	.	heat shock protein family A (Hsp70) member 8	HSPA8	0.46

Appendix Table 4A6

Table 4A6: Esophageal Squamous Cell Carcinoma map of enriched pathways to differentially expressed genes

Metabolism	Metabolism of amino acids and derivatives	Glycosaminoglycan metabolism	G2/M Checkpoints	Signaling by MET	Interferon alpha/beta signaling	Formation of the cornified envelope	Smooth Muscle Contraction	Metabolism of water-soluble vitamins and cofactors	Integrin cell surface interactions	O-linked glycosylation of mucins	Phase 1 - Functionalization of compounds	Homologous DNA Pairing and Strand Exchange	Collagen chain trimerization	Amino acid synthesis and interconversion (transamination)	Metabolism of ingested SeMet, Sec, MeSec into H2Se	E2F-enabled inhibition of pre-replication complex formation	Cyclin B2 mediated events	Glycoprotein hormones	Antagonism of Activin by Follistatin	TFAP2 (AP-2) family regulates transcription of other transcription factors	Glycogen storage diseases	Arachidonate production from DAG			
R.HSA.1430728	R.HSA.71291	R.HSA.1630316	R.HSA.69481	R.HSA.6806834	R.HSA.909733	R.HSA.6809371	R.HSA.445355	R.HSA.196849	R.HSA.216083	R.HSA.913709	R.HSA.211945	R.HSA.5693579	R.HSA.8948216	R.HSA.70614	R.HSA.2408508	R.HSA.113507	R.HSA.157881	R.HSA.209822	R.HSA.2473224	R.HSA.8866906	R.HSA.3229121	R.HSA.426048	GENENAME	SYMBOL	logFC
.	.	.	X	.	.	.	.	.	.	.	.	X	.	.	.	.	.	.	.	.	.	.	replication protein A1	RPA1	-0.49
.	.	.	X	.	.	.	.	.	.	.	.	X	.	.	.	.	.	.	.	.	.	.	replication factor C subunit 3	RFC3	-0.86
.	.	.	X	.	.	.	.	.	.	.	.	X	.	.	.	.	.	.	.	.	.	.	RAD1 checkpoint DNA exonuclease	RAD1	-0.71
.	.	.	X	.	.	.	.	.	.	.	.	X	.	.	.	.	.	.	.	.	.	.	nibrin	NBN	-0.67
.	.	.	X	.	.	.	.	.	.	.	.	X	.	.	.	.	.	.	.	.	.	.	MRE11 homolog, double strand break repair nuclease	MRE11	-0.79

.	.	.	X	.	.	.	.	.	.	.	.	X	.	.	.	.	.	.	.	.	.	.	ATR serine/threonine kinase	ATR	-0.98
.	.	.	X	.	.	.	.	.	.	.	.	X	.	.	.	.	.	.	.	.	.	.	checkpoint kinase 1	CHEK1	-1.34
.	.	.	X	.	.	.	.	.	.	.	.	.	.	.	.	X	X	.	.	.	.	.	cyclin dependent kinase 1	CDK1	-1.40
.	.	.	X	.	.	.	.	.	.	.	.	.	.	.	.	X	.	.	.	.	.	.	cyclin B1	CCNB1	-1.13
.	.	.	X	.	.	.	.	.	.	.	.	.	.	.	.	X	.	.	.	.	.	.			
.	.	.	X	.	.	.	.	.	.	.	.	.	.	.	.	.	X	.	.	.	.	.	cyclin B2	CCNB2	-1.30
.	.	.	X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	minichromosome maintenance 10 replication initiation factor	MCM10	-1.52
.	.	.	X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	G2 and S-phase expressed 1	GTSE1	-1.46
.	.	.	X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	H2B clustered histone 4	H2BC4	0.43
.	.	.	X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	tyrosine 3-monooxygenase/tryptophan 5-monooxygenase activation protein eta	YWHAH	-0.41
.	.	.	X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	nuclear receptor binding SET domain protein 2	NSD2	-0.79
.	.	.	X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.			
.	.	.	X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	minichromosome maintenance complex component 4	MCM4	-1.09
.	.	.	X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	minichromosome maintenance	MCM6	-1.10

																							complex component 6		
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	X	X	.	.	.	inhibin subunit beta A	INHBA	-2.61
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	X	.	.	.	follistatin	FST	-1.48
.	.	.	.	.	.	.	.	.	.	.	.	.	X	.	.	.	.	.	.	.	.	.	collagen type XI alpha 1 chain	COL11A1	-2.08
.	.	.	.	.	.	.	.	.	.	.	.	.	X	.	.	.	.	.	.	.	.	.	collagen type IV alpha 5 chain	COL4A5	-1.05
.	.	.	.	.	.	.	.	.	.	.	.	.	X	.	.	.	.	.	.	.	.	.	collagen type I alpha 2 chain	COL1A2	-2.64
.	.	.	.	.	.	.	.	.	.	.	.	.	X	.	.	.	.	.	.	.	.	.	collagen type III alpha 1 chain	COL3A1	-1.79
.	.	.	.	.	.	.	.	.	X	.	.	.	X	.	.	.	.	.	.	.	.	.			
.	.	.	.	.	.	.	.	.	X	.	.	.	.	.	.	.	.	.	.	.	.	.	junctional adhesion molecule 3	JAM3	0.57
.	.	.	.	.	.	.	.	.	X	.	.	.	.	.	.	.	.	.	.	.	.	.	junctional adhesion molecule 2	JAM2	1.21
.	.	.	.	.	.	.	.	.	X	.	.	.	.	.	.	.	.	.	.	.	.	.	integrin subunit alpha 8	ITGA8	1.07
.	.	.	.	.	.	.	.	.	X	.	.	.	.	.	.	.	.	.	.	.	.	.	thrombospondin 1	THBS1	-0.72
.	.	.	.	.	.	.	.	.	X	.	.	.	.	.	.	.	.	.	.	.	.	.	integrin subunit alpha 6	ITGA6	-1.67
.	.	.	.	.	.	.	.	.	X	.	.	.	.	.	.	.	.	.	.	.	.	.	fibrillin 1	FBN1	-0.43
.	.	.	.	.	.	.	.	.	X	.	.	.	.	.	.	.	.	.	.	.	.	.	tenascin C	TNC	-1.88
X	X	.	.	.	.	.	.	X	.	.	.	.	.	.	X	.	.	.	.	.	.	.	nicotinamide N-methyltransferase	NNMT	-0.71
X	X	.	.	.	.	.	.	.	.	.	.	.	.	.	X	.	.	.	.	.	.	.	cystathionine beta-synthase	CBS	-0.94
X	X	.	.	.	.	.	.	.	.	.	.	.	.	.	X	.	.	.	.	.	.	.	histamine N-methyltransferase	HNMT	1.17

X	.	X	.	.	.	.	.	.	.	X	.	.	.	.	.	.	.	.	.	.	.	.			
X	.	X	.	.	.	.	.	.	X	.	.	.	.	.	.	.	.	.	.	.	.	.	CD44 molecule (Indian blood group)	CD44	-0.78
X	.	X	.	.	.	.	.	.	X	.	.	.	.	.	.	.	.	.	.	.	.	.	lumican	LUM	-1.93
X	.	X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	xylosyltransferase 1	XYLT1	1.30
X	.	X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	carbohydrate sulfotransferase 12	CHST12	-0.57
X	.	X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	chondroitin sulfate N-acetylgalactosaminyltransferase 2	CSGALNACT2	-0.72
X	.	X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	carbohydrate sulfotransferase 11	CHST11	-0.75
X	.	X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.			
X	.	X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	heparanase	HPSE	0.93
X	.	X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	ATP binding cassette subfamily C member 5	ABCC5	-1.08
X	.	X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.			
X	.	X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	carbohydrate sulfotransferase 2	CHST2	-1.14
X	.	X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	carbohydrate sulfotransferase 1	CHST1	-0.95
X	.	X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	proline and arginine rich end leucine rich repeat protein	PRELP	1.13
X	.	X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	osteoglycin	OGN	2.21
X	.	X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.			
X	.	X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.			
X	.	X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	hexosaminidase subunit alpha	HEXA	-0.50

X	.	X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	hyaluronan synthase 2	HAS2	-0.61
X	.	X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.			
X	.	X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	glypican 3	GPC3	-0.89
X	.	X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	exostosin glycosyltransferase 1	EXT1	-0.96
X	.	X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	decorin	DCN	0.49
X	.	X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.			0.49
X	.	X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	solute carrier family 26 member 2	SLC26A2	1.35
X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	X	.	protein phosphatase 1 regulatory subunit 3C	PPP1R3C	3.47
X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	X	monoglyceride lipase	MGLL	2.20
X	.	.	.	.	.	.	.	X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	LMBR1 domain containing 1	LMBRD1	1.04
X	.	.	.	.	.	.	.	X	.	.	.	.	.	.	.	.	.	.	.	.	.	.			
X	.	.	.	.	.	.	.	X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	nicotinamide phosphoribosyltransferase	NAMPT	0.73
X	.	.	.	.	.	.	.	X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	solute carrier family 19 member 1	SLC19A1	-0.81
X	.	.	.	.	.	.	.	X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	propionyl-CoA carboxylase subunit alpha	PCCA	0.95
X	.	.	.	.	.	.	.	X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	5'-nucleotidase ecto	NT5E	-0.51
X	.	.	.	.	.	.	.	X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	biotinidase	BTD	0.70
X	.	.	.	.	.	.	.	X	.	.	.	.	.	.	.	.	.	.	.	.	.	.			
X	X	.	.	.	.	.	.	X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	5-methyltetrahydrofolate-homocysteine methyltransferase	MTR	-0.66

X	X	.	.	.	.	.	.	X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	serine hydroxymethyltransferase 1	SHMT1	1.13
X	X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	2-aminoethanethiol dioxygenase	ADO	-0.50
X	X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	agmatinase	AGMAT	-0.89
X	X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.			
X	X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	dehydrogenase E1 and transketolase domain containing 1	DHTKD1	-0.53
X	X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.			
X	X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	betaine--homocysteine S-methyltransferase 2	BHMT2	1.01
X	X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.			
X	X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	tryptophan 2,3-dioxygenase	TDO2	-1.69
X	X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	solute carrier family 6 member 8	SLC6A8	-0.81
X	X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.			
X	X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.			
X	X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	ribosomal protein L15	RPL15	0.68
X	X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.			
X	X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.			
X	X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.			
X	X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	glutamate-cysteine ligase modifier subunit	GCLM	-0.91
X	X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	glycine amidinotransferase	GATM	1.68

X	X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	iodothyronine deiodinase 2	DIO2	1.85
X	X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	branched chain amino acid transaminase 1	BCAT1	-1.84
X	X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	acyl-CoA dehydrogenase short/branched chain	ACADSB	0.82
X	X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	arginase 1	ARG1	0.96
X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.			
X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	nucleoporin 43	NUP43	-0.63
X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	lysophosphatidylcholine acyltransferase 4	LPCAT4	0.30
X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	membrane bound O-acyltransferase domain containing 2	MBOAT2	0.51
X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	oxysterol binding protein like 10	OSBPL10	0.88
X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	oxysterol binding protein like 8	OSBPL8	0.34
X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	trimethylguanosine synthase 1	TGS1	-0.77
X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.			
X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	ADP dependent glucokinase	ADPGK	-0.61
X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	SEH1 like nucleoporin	SEH1L	-0.48
X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	phospholipase A2 group XIIA	PLA2G12A	0.74
X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	RUN and FYVE	RUFY1	0.42

																							domain containing 1		
X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	stearoyl-CoA desaturase 5	SCD5	-0.44
X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	phospholipase B domain containing 1	PLBD1	1.02
X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	TBL1X receptor 1	TBL1XR1	-0.56
X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	fatty acid 2-hydroxylase	FA2H	0.53
X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	ELOVL fatty acid elongase 6	ELOVL6	1.21
X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	nucleoporin 37	NUP37	-0.51
X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.			
X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	LYR motif containing 4	LYRM4	-0.59
X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.			
X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	choline phosphotransferase 1	CHPT1	0.90
X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	StAR related lipid transfer domain containing 7	STARD7	-0.67
X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	G protein subunit gamma 12	GNG12	0.68
X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	nucleoporin 133	NUP133	-0.45
X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	fatty acyl-CoA reductase 2	FAR2	0.00
X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	NDC1 transmembrane nucleoporin	NDC1	-1.08
X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.			
X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	abhydrolase domain containing 10,	ABHD10	-0.47

																							depalmitoylase		
X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1-acylglycerol-3-phosphate O-acyltransferase 5	AGPAT5	-0.41
X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	pyruvate dehyrogenase phosphatase catalytic subunit 1	PDP1	-0.66
X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	carnitine O-octanoyltransferase	CROT	0.47
X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	UDP glucuronosyltransferase family 1 member A1	UGT1A1	1.35
X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.			
X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.			
X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.			
X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	cytidine/uridine monophosphate kinase 1	CMPK1	0.66
X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	acyl-CoA synthetase long chain family member 5	ACSL5	0.56
X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	mitochondrial pyruvate carrier 1	MPC1	1.12
X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	3-hydroxyacyl-CoA dehydratase 3	HACD3	-0.79
X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	inositol-3-phosphate synthase 1	ISYNA1	-0.67

X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	solute carrier family 25 member 37	SLC25A37	-0.41
X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	glycolipid transfer protein	GLTP	0.92
X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.			
X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	secretion associated Ras related GTPase 1B	SAR1B	0.53
X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	retinol dehydrogenase 11	RDH11	-0.39
X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	abhydrolase domain containing 5, lysophosphatidic acid acyltransferase	ABHD5	1.12
X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.			
X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	ORMDL sphingolipid biosynthesis regulator 2	ORMDL2	0.38
X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.			
X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	LDL receptor related protein 10	LRP10	0.71
X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	WW domain containing transcription regulator 1	WWTR1	0.53
X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.			
X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.			
X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	nucleoporin 62	NUP62	-0.76
X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	alpha-methylacyl-CoA racemase	AMACR	0.76
X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.			
X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	dimethylarginine	DDAH1	1.67

																							dimethylaminohydrolase 1		
X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.			
X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.			
X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	cyclin dependent kinase 19	CDK19	0.65
X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	5'-nucleotidase, cytosolic II	NT5C2	0.89
X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	pleckstrin homology domain containing A6	PLEKHA6	1.22
X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.			
X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	phospholipase A and acyltransferase 3	PLAAT3	-0.02
X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	GLI pathogenesis related 1	GLIPR1	-0.89
X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.			
X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.			
X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.			
X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.			
X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	G protein subunit beta 5	GNB5	-0.54
X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	acetyl-CoA acyltransferase 2	ACAA2	1.05
X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	Yes1 associated transcriptional regulator	YAP1	0.35
X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	solute carrier family 25 member 13	SLC25A13	-0.66
X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	leucine rich pentatricopeptide repeat containing	LRPPRC	-0.60

X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	ceramide transporter 1	CERT1	0.63
X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.			
X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	lysophosphatidylglycerol acyltransferase 1	LPGAT1	-0.57
X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	nucleoporin 93	NUP93	-0.48
X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	G protein subunit alpha 14	GNA14	1.05
X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	ATP binding cassette subfamily G member 1	ABCG1	-0.83
X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.			
X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	clock circadian regulator	CLOCK	0.90
X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	serine palmitoyltransferase long chain base subunit 2	SPTLC2	0.70
X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	phosphatidylglycerophosphate synthase 1	PGS1	-0.56
X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.			
X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	fatty acid desaturase 2	FADS2	-0.72
X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	WASP like actin nucleation promoting factor	WASL	0.59
X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	myotubularin related protein 3	MTMR3	0.64
X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	guanine monophosphate synthase	GMPS	-0.80
X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	inositol polyphosphate-4-	INPP4B	0.59

																							phosphatase type II B		
X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	succinate-CoA ligase GDP/ADP-forming subunit alpha	SUCLG1	0.42
X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.			
X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.			
X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.			
X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	alkylglycerone phosphate synthase	AGPS	-0.45
X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	phosphoinositide-3-kinase regulatory subunit 3	PIK3R3	-0.51
X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	LDL receptor related protein 8	LRP8	-1.22
X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.			
X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	uridine monophosphate synthetase	UMPS	-0.79
X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	uridine-cytidine kinase 2	UCK2	-0.89
X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	UDP glycosyltransferase 8	UGT8	-0.85
X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.			
X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	UDP-glucose ceramide glucosyltransferase	UGCG	0.36
X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	translocated promoter region, nuclear basket protein	TPR	-0.37
X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	TEA domain transcription factor 1	TEAD1	0.82

X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.			
X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.			
X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	sterol O-acyltransferase 1	SOAT1	-1.23
X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.			
X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	solute carrier family 5 member 1	SLC5A1	0.76
X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	regulator of solute carriers 1	RSC1A1	0.34
X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	ribonucleotide reductase catalytic subunit M1	RRM1	-0.37
X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	RAR related orphan receptor A	RORA	1.47
X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	retinol binding protein 1	RBP1	-1.88
X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.			
X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	RAB5A, member RAS oncogene family	RAB5A	0.92
X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.			
X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.			
X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	phosphoribosyl pyrophosphate synthetase 2	PRPS2	-0.48
X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	protein kinase cAMP-dependent type II regulatory subunit alpha	PRKAR2A	0.58
X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	protein kinase cAMP-activated	PRKACB	0.54

																							catalytic subunit beta		
X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	protein kinase AMP-activated non-catalytic subunit beta 2	PRKAB2	-0.44
X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	protein kinase AMP-activated catalytic subunit alpha 2	PRKAA2	1.71
X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	protein phosphatase 1 catalytic subunit beta	PPP1CB	0.92
X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	phosphoribosyl pyrophosphate amidotransferase	PPAT	-1.04
X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	mediator complex subunit 1	MED1	-0.54
X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	paraoxonase 2	PON2	-0.75
X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	phospholipase D1	PLD1	0.47
X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.			
X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	phosphorylase kinase regulatory subunit beta	PHKB	0.63
X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.			
X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	pyruvate dehydrogenase kinase 1	PDK1	-0.56
X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.			
X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.			
X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.			
X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.			

X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.			
X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.			
X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.			
X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.			
X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	lipase A, lysosomal acid type	LIPA	-0.28
X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	potassium voltage-gated channel subfamily B member 1	KCNB1	0.88
X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	inositol 1,4,5-trisphosphate receptor type 3	ITPR3	-1.30
X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	inositol 1,4,5-trisphosphate receptor type 1	ITPR1	0.81
X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.			
X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.			
X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	3-hydroxy-3-methylglutaryl-CoA synthase 2	HMGCS2	0.78
X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	hexokinase 2	HK2	-0.50
X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.			
X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.			
X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	glutathione S-transferase theta 1	GSTT1	0.67
X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	glutathione S-transferase mu 3	GSTM3	-0.07
X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.			-0.07
X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.			
X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	glutathione peroxidase 2	GPX2	-0.95
X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.			

X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	G protein subunit gamma 11	GNG11	0.27
X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	G protein subunit alpha q	GNAQ	0.99
X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	G protein subunit alpha i1	GNAI1	-0.64
X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.			
X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	phosphoribosylglycinamide formyltransferase, phosphoribosylglycinamide synthetase, phosphoribosylaminoimidazole synthetase	GART	-0.57
X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.			
X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	fatty acid binding protein 4	FABP4	-0.79
X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.			
X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	deoxyguanosine kinase	DGUOK	-0.69
X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	aldo-keto reductase family 1 member C2	AKR1C2	-1.15
X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.			0.52
X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	cholinergic receptor muscarinic 3	CHRM3	0.52
X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	CDP-diacylglycerol synthase 1	CDS1	0.82
X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	ectonucleoside triphosphate diphosphohydrolase 5 (inactive)	ENTPD5	0.76

X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.			
X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	scavenger receptor class B member 1	SCARB1	-0.83
X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	CD36 molecule	CD36	0.63
X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.			
X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	biliverdin reductase A	BLVRA	0.45
X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	ATP synthase peripheral stalk-membrane subunit b	ATP5PB	0.72
X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.			
X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	adenylate kinase 4	AK4	0.97
X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	adenylosuccinate synthase 2	ADSS2	-0.71
X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	acyl-CoA oxidase 1	ACOX1	1.45
X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.			
X	X	.	.	.	.	.	.	.	.	.	X	.	.	.	.	.	.	.	.	.	.	.			
X	.	.	.	.	.	.	.	.	.	.	X	.	.	.	.	.	.	.	.	.	.	.	mitochondrial amidoxime reducing component 2	MTARC2	0.89
X	.	.	.	.	.	.	.	.	.	.	X	.	.	.	.	.	.	.	.	.	.	.	cytochrome P450 family 39 subfamily A member 1	CYP39A1	0.27
X	.	.	.	.	.	.	.	.	.	.	X	.	.	.	.	.	.	.	.	.	.	.	prostaglandin-endoperoxide synthase 1	PTGS1	1.65
X	.	.	.	.	.	.	.	.	.	.	X	.	.	.	.	.	.	.	.	.	.	.			
X	.	.	.	.	.	.	.	.	.	.	X	.	.	.	.	.	.	.	.	.	.	.	cytochrome P450 family 3 subfamily A member 5	CYP3A5	2.40
X	.	.	.	.	.	.	.	.	.	.	X	.	.	.	.	.	.	.	.	.	.	.	alcohol dehydrogena	ADH1B	3.05

																							se 1B (class I), beta polypeptide		
X	.	.	.	.	.	.	.	.	.	.	X	.	.	.	.	.	.	.	.	.	.	.	cytochrome P450 family 3 subfamily A member 4	CYP3A4	0.70
X	X	.	.	.	.	.	.	.	.	.	.	.	.	X	.	.	.	.	.	.	.	.	glutamate-ammonia ligase	GLUL	0.46
X	X	.	.	.	.	.	.	.	.	.	.	.	.	X	.	.	.	.	.	.	.	.	glutaminase	GLS	-0.83
X	X	.	.	.	.	.	.	.	.	.	.	.	.	X	.	.	.	.	.	.	.	.	asparagine synthetase (glutamine-hydrolyzing)	ASNS	-0.87
X	X	.	.	.	.	.	.	.	.	.	.	.	.	X	.	.	.	.	.	.	.	.	aspartoacylase	ASPA	1.65
X	X	.	X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	proteasome 20S subunit beta 4	PSMB4	-0.65
X	X	.	X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.			
X	X	.	X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	proteasome 20S subunit beta 2	PSMB2	-0.93
.	.	.	.	.	.	.	.	.	.	X	.	.	.	.	.	.	.	.	.	.	.	.	polypeptide N-acetylgalactosaminyltransferase 12	GALNT12	2.19
.	.	.	.	.	.	.	.	.	.	X	.	.	.	.	.	.	.	.	.	.	.	.	core 1 synthase, glycoprotein-N-acetylgalactosamine 3-beta-galactosyltransferase 1	C1GALT1	0.69
.	.	.	.	.	.	.	.	.	.	X	.	.	.	.	.	.	.	.	.	.	.	.	polypeptide N-acetylgalactosaminyltransferase 10	GALNT10	-0.65
.	.	.	.	.	.	.	.	.	.	X	.	.	.	.	.	.	.	.	.	.	.	.	polypeptide N-acetylgalacto	GALNT7	0.23

																							saminyltransferase 7		
.	.	.	.	.	.	.	.	.	.	X	.	.	.	.	.	.	.	.	.	.	.	.	polypeptide N-acetylgalactosaminyltransferase 6	GALNT6	-1.48
.	.	.	.	.	.	.	.	.	.	X	.	.	.	.	.	.	.	.	.	.	.	.			0.32
.	.	.	.	.	.	.	.	.	.	X	.	.	.	.	.	.	.	.	.	.	.	.	mucin 4, cell surface associated	MUC4	0.32
.	.	.	.	.	.	.	.	.	.	X	.	.	.	.	.	.	.	.	.	.	.	.	polypeptide N-acetylgalactosaminyltransferase 1	GALNT1	0.72
.	.	.	.	.	.	.	.	.	.	X	.	.	.	.	.	.	.	.	.	.	.	.	polypeptide N-acetylgalactosaminyltransferase 2	GALNT2	-0.75
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	X	.	.	Cbp/p300 interacting transactivator with Glu/Asp rich carboxy-terminal domain 2	CITED2	2.18
.	.	.	.	.	X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	radical S-adenosyl methionine domain containing 2	RSAD2	-1.23
.	.	.	.	.	X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	XIAP associated factor 1	XAF1	-0.57
.	.	.	.	.	X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	SAM and HD domain containing deoxynucleoside triphosphate triphosphohydrolase 1	SAMHD1	0.88
.	.	.	.	.	X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	signal transducer and activator of	STAT1	-1.60

																							transcription 1		
.	.	.	.	.	X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	2'-5'-oligoadenylate synthetase 3	OAS3	-0.84
.	.	.	.	.	X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	2'-5'-oligoadenylate synthetase 2	OAS2	-0.81
.	.	.	.	.	X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	interferon regulatory factor 1	IRF1	-0.31
.	.	.	.	.	X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	interferon alpha and beta receptor subunit 2	IFNAR2	-0.46
.	.	.	.	.	X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	interferon induced protein with tetratricopeptide repeats 3	IFIT3	-1.06
.	.	.	.	.	X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	early growth response 1	EGR1	0.22
.	.	.	.	.	X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.			
X	.	.	.	X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.			
X	.	.	.	X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	RAP1A, member of RAS oncogene family	RAP1A	1.07
.	.	.	.	X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	tensin 3	TNS3	-0.73
.	.	.	.	X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	leucine rich repeats and immunoglobulin l ke domains 1	LRIG1	-0.11
.	.	.	.	X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	RAN binding protein 9	RANBP9	1.68
.	.	.	.	X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	serine peptidase inhibitor, Kunitz type 1	SPINT1	0.53
.	.	.	.	X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	protein tyrosine phosphatase	PTPN2	-0.69

																							non-receptor type 2		
.	.	.	.	X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	protein tyrosine kinase 2	PTK2	-0.71
.	.	.	.	X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	laminin subunit alpha 5	LAMA5	-0.68
.	.	.	.	X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	laminin subunit alpha 3	LAMA3	-1.50
.	.	.	.	X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	ADP ribosylation factor 6	ARF6	0.74
.	.	.	.	X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	GRB2 associated binding protein 1	GAB1	1.29
X	.	.	.	.	.	.	X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	calmodulin 1	CALM1	0.84
.	.	.	.	.	.	.	X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	sorbin and SH3 domain containing 1	SORBS1	1.74
.	.	.	.	.	.	.	X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.			0.72
.	.	.	.	.	.	.	X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	myosin light chain 9	MYL9	0.72
.	.	.	.	.	.	.	X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	tropomyosin 2	TPM2	0.36
.	.	.	.	.	.	.	X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.			0.36
.	.	.	.	.	.	.	X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	tropomyosin 1	TPM1	0.02
.	.	.	.	.	.	.	X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.			0.02
.	.	.	.	.	.	.	X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	myosin light chain kinase	MYLK	0.04
.	.	.	.	.	.	.	X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.			0.04
.	.	.	.	.	.	.	X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.			1.57
.	.	.	.	.	.	.	X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	myosin heavy chain 11	MYH11	1.57
.	.	.	.	.	.	.	X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	caldesmon 1	CALD1	-0.07
.	.	.	.	.	.	.	X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	annexin A1	ANXA1	1.86

.	.	.	.	.	.	.	X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	actin alpha 2, smooth muscle	ACTA2	0.02
.	.	.	.	.	.	.	X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.			0.02
.	.	.	.	.	.	.	X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.			1.23
.	.	.	.	.	.	.	X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	actin gamma 2, smooth muscle	ACTG2	1.23
X	.	.	.	.	.	X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	proprotein convertase subtilisin/kexin type 6	PCSK6	0.45
.	.	.	.	.	.	X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.			
.	.	.	.	.	.	X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	kall krein related peptidase 12	KLK12	2.11
.	.	.	.	.	.	X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	kazrin, periplakin interacting protein	KAZN	0.88
.	.	.	.	.	.	X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	kall krein related peptidase 8	KLK8	0.45
.	.	.	.	.	.	X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	small proline rich protein 3	SPRR3	2.81
.	.	.	.	.	.	X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	desmoglein 3	DSG3	0.60
.	.	.	.	.	.	X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	desmocollin 3	DSC3	-0.51
.	.	.	.	.	.	X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	desmocollin 2	DSC2	1.33
.	.	.	.	.	.	X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	calpain 1	CAPN1	0.31
.	.	.	.	.	.	X	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	calpain small subunit 1	CAPNS1	0.50

Appendix Table 4A7

Table 4A7: Subsampling substudy from nine repeats and top 50 pathways

Pathway ID	Pathway name	Trial1	Trial2	Trial3	Trial4	Trial5	Trial6	Trial7	Trial8	Trial9	Ztrial1	Ztrial2	Ztrial3	Ztrial4	Ztrial5	Ztrial6	Ztrial7	Ztrial8	Ztrial9	Z mean	P mean	Z variance	Consistency (Zmean/Zvariance)
R-HSA-6809371	Formation of the cornified envelope	1,65E-06	7,96E-07	8,57E-07	1,04E-06	8,34E-07	2,29E-07	2,36E-07	3,31E-07	1,84E-07	-4,65	-4,80	-4,78	-4,75	-4,79	-5,04	-5,04	-4,97	-5,08	-4,88	5,34E-07	2,45E-02	-199
R-HSA-445355	Smooth Muscle Contraction	7,91E-05	5,23E-05	1,92E-04	1,22E-04	1,18E-04	9,60E-05	1,41E-04	1,18E-04	1,12E-04	-3,78	-3,88	-3,55	-3,67	-3,68	-3,73	-3,63	-3,68	-3,69	-3,70	1,09E-04	8,55E-03	-433
R-HSA-446107	Type I hemidesmosome assembly	6,18E-04	6,69E-04	3,92E-04	5,71E-04	4,77E-04	6,73E-04	5,44E-04	7,56E-04	7,23E-04	-3,23	-3,21	-3,36	-3,25	-3,30	-3,21	-3,27	-3,17	-3,19	-3,24	5,92E-04	3,60E-03	-900
R-HSA-156584	Cytosolic sulfonation of small molecules	1,53E-03	3,96E-04	5,45E-04	1,45E-03	1,87E-03	1,12E-03	6,13E-04	1,06E-03	4,80E-04	-2,96	-3,36	-3,27	-2,98	-2,90	-3,06	-3,23	-3,07	-3,30	-3,12	8,90E-04	2,79E-02	-112
R-HSA-2022090	Assembly of collagen fibrils and other multimeric structures	2,34E-03	1,95E-03	1,21E-03	1,40E-03	1,57E-03	1,29E-03	1,06E-03	1,55E-03	1,46E-03	-2,83	-2,89	-3,03	-2,99	-2,95	-3,01	-3,07	-2,96	-2,98	-2,97	1,50E-03	5,52E-03	-537
R-HSA-6803205	TP53 regulates transcription of several additional cell death genes whose specific roles in p53-dependent apoptosis remain uncertain	2,84E-03	3,70E-03	2,53E-03	2,71E-03	3,17E-03	2,82E-03	2,78E-03	2,74E-03	2,77E-03	-2,77	-2,68	-2,80	-2,78	-2,73	-2,77	-2,77	-2,78	-2,77	-2,76	2,88E-03	1,33E-03	-2078
R-HSA-6798695	Neutrophil degranulation	7,83E-03	3,05E-03	2,94E-03	8,78E-03	7,45E-03	2,56E-03	1,91E-03	1,56E-03	2,70E-03	-2,42	-2,74	-2,75	-2,37	-2,43	-2,80	-2,89	-2,96	-2,78	-2,68	3,64E-03	4,73E-02	-57
R-HSA-	O-linked glycosylation of mucins	5,61E-03	6,99E-03	6,98E-03	6,14E-03	1,50E-04	7,76E-03	7,20E-03	5,75E-03	7,35E-03	-2,54	-2,46	-2,46	-2,50	-3,62	-2,42	-2,45	-2,53	-2,44	-2,60	4,65E-03	1,47E-01	-18

913709																							
R-HSA-1368108	BMAL1:CLOCK, NPAS2 activates circadian gene expression	2,32E-02	4,06E-03	2,43E-03	4,46E-03	2,52E-02	5,01E-03	5,05E-03	9,04E-03	4,79E-03	-1,99	-2,65	-2,82	-2,62	-1,96	-2,58	-2,57	-2,36	-2,59	-2,46	6,97E-03	8,90E-02	-28
R-HSA-6811440	Retrograde transport at the Trans-Golgi-Network	6,35E-03	7,50E-03	1,27E-02	6,70E-03	6,19E-03	9,09E-03	6,03E-03	1,00E-02	8,58E-03	-2,49	-2,43	-2,24	-2,47	-2,50	-2,36	-2,51	-2,33	-2,38	-2,41	7,91E-03	8,72E-03	-277
R-HSA-5423646	Aflatoxin activation and detoxification	6,98E-03	9,78E-03	7,00E-03	6,92E-03	6,83E-03	1,11E-02	6,99E-03	1,17E-02	8,63E-03	-2,46	-2,33	-2,46	-2,46	-2,47	-2,29	-2,46	-2,27	-2,38	-2,40	8,28E-03	6,71E-03	-357
R-HSA-420029	Tight junction interactions	1,16E-02	6,96E-03	9,33E-03	9,11E-03	1,16E-02	6,87E-03	1,17E-02	6,98E-03	9,29E-03	-2,27	-2,46	-2,35	-2,36	-2,27	-2,46	-2,27	-2,46	-2,35	-2,36	9,09E-03	6,90E-03	-342
R-HSA-6785807	Interleukin-4 and 13 signaling	1,06E-02	1,14E-02	9,61E-03	8,67E-03	1,18E-02	1,48E-02	6,71E-03	1,36E-02	7,12E-03	-2,30	-2,28	-2,34	-2,38	-2,26	-2,18	-2,47	-2,21	-2,45	-2,32	1,02E-02	1,04E-02	-224
R-HSA-5683826	Surfactant metabolism	1,97E-02	9,30E-03	9,35E-03	9,14E-03	9,90E-03	7,46E-03	1,05E-02	9,32E-03	1,97E-02	-2,06	-2,35	-2,35	-2,36	-2,33	-2,43	-2,31	-2,35	-2,06	-2,29	1,10E-02	1,82E-02	-126
R-HSA-74182	Ketone body metabolism	1,28E-02	1,27E-02	1,51E-02	1,48E-02	1,62E-02	1,26E-02	1,52E-02	1,28E-02	1,28E-02	-2,23	-2,23	-2,17	-2,18	-2,14	-2,24	-2,17	-2,23	-2,23	-2,20	1,38E-02	1,58E-03	-1395
R-HSA-8854521	Interaction between PHLDA1 and AURKA	1,69E-02	1,80E-02	1,87E-02	1,60E-02	1,69E-02	1,78E-02	1,93E-02	1,69E-02	1,69E-02	-2,12	-2,10	-2,08	-2,14	-2,12	-2,10	-2,07	-2,12	-2,12	-2,11	1,75E-02	5,87E-04	-3591
R-HSA-5627083	RHO GTPases regulate CFTR trafficking	1,86E-02	2,03E-02	2,10E-02	1,94E-02	1,75E-02	1,95E-02	1,82E-02	1,92E-02	2,10E-02	-2,08	-2,05	-2,03	-2,07	-2,11	-2,06	-2,09	-2,07	-2,03	-2,07	1,94E-02	6,71E-04	-3081
R-HSA-983695	Antigen activates B Cell Receptor (BCR) leading to generation of second messengers	1,98E-02	2,05E-02	1,91E-02	1,90E-02	1,89E-02	1,90E-02	2,09E-02	2,18E-02	1,79E-02	-2,06	-2,04	-2,07	-2,07	-2,08	-2,07	-2,04	-2,02	-2,10	-2,06	1,96E-02	6,32E-04	-3264
R-HSA-	Factors involved in megakaryocyte	2,42E-02	2,41E-02	1,11E-02	1,13E-02	2,32E-02	2,36E-02	2,64E-02	3,22E-02	2,42E-02	-1,97	-1,98	-2,29	-2,28	-1,99	-1,99	-1,94	-1,85	-1,97	-2,03	2,13E-02	2,28E-02	-89

983231	development and platelet production																						
R-HSA-163560	Hormone-sensitive lipase (HSL)-mediated triacylglycerol hydrolysis	2,96E-02	2,27E-02	2,47E-02	2,86E-02	2,81E-02	2,10E-02	2,29E-02	1,99E-02	1,87E-02	-1,89	-2,00	-1,96	-1,90	-1,91	-2,03	-2,00	-2,06	-2,08	-1,98	2,38E-02	4,93E-03	-402
R-HSA-425393	Transport of inorganic cations/anions and amino acids/oligopeptides	6,56E-04	1,03E-03	9,86E-01	7,56E-04	5,09E-04	9,33E-04	7,44E-04	9,86E-01	1,42E-03	-3,21	-3,08	2,20	-3,17	-3,29	-3,11	-3,18	2,20	-2,98	-1,96	2,51E-02	5,56E+00	0
R-HSA-2022377	Metabolism of Angiotensinogen to Angiotensins	4,66E-03	5,80E-03	1,05E-02	8,55E-03	4,66E-03	9,86E-01	4,68E-03	6,98E-03	1,04E-02	-2,60	-2,52	-2,31	-2,38	-2,60	2,20	-2,60	-2,46	-2,31	-1,95	2,54E-02	2,44E+00	-1
R-HSA-1989781	PPARA activates gene expression	5,09E-03	5,96E-03	9,86E-01	6,50E-03	6,10E-03	8,52E-03	7,65E-03	7,56E-03	8,97E-03	-2,57	-2,51	2,20	-2,48	-2,51	-2,39	-2,43	-2,43	-2,37	-1,94	2,60E-02	2,42E+00	-1
R-HSA-109582	Hemostasis	1,74E-02	3,28E-02	2,30E-02	1,79E-02	1,70E-02	2,50E-02	3,17E-02	6,18E-02	2,59E-02	-2,11	-1,84	-2,00	-2,10	-2,12	-1,96	-1,86	-1,54	-1,95	-1,94	2,61E-02	3,35E-02	-58
R-HSA-446728	Cell junction organization	3,17E-02	2,08E-02	2,78E-02	2,72E-02	3,45E-02	2,05E-02	3,47E-02	2,08E-02	2,77E-02	-1,86	-2,04	-1,91	-1,92	-1,82	-2,04	-1,82	-2,04	-1,92	-1,93	2,68E-02	8,39E-03	-230
R-HSA-5083625	Defective GALNT3 causes familial hyperphosphatemic tumoral calcinosis (HFTC)	6,80E-03	8,12E-03	8,16E-03	6,35E-03	9,86E-01	8,02E-03	7,02E-03	5,82E-03	8,13E-03	-2,47	-2,40	-2,40	-2,49	2,20	-2,41	-2,46	-2,52	-2,40	-1,93	2,69E-02	2,40E+00	-1
R-HSA-5676594	TNF receptor superfamily (TNFSF) members mediating non-canonical NF-kB pathway	3,04E-02	3,03E-02	3,40E-02	3,15E-02	3,42E-02	2,93E-02	3,53E-02	3,11E-02	3,06E-02	-1,87	-1,88	-1,82	-1,86	-1,82	-1,89	-1,81	-1,87	-1,87	-1,85	3,18E-02	8,34E-04	-2224
R-HSA-3229121	Glycogen storage diseases	9,86E-01	1,34E-02	1,34E-02	1,37E-02	1,16E-02	1,09E-02	1,35E-02	1,28E-02	1,11E-02	2,20	-2,22	-2,21	-2,21	-2,27	-2,29	-2,21	-2,23	-2,29	-1,75	4,03E-02	2,19E+00	-1

R-HSA-1475029	Reversible hydration of carbon dioxide	3,67E-02	3,78E-02	7,34E-02	3,54E-02	3,73E-02	3,84E-02	3,98E-02	3,73E-02	4,02E-02	-1,79	-1,78	-1,45	-1,81	-1,78	-1,77	-1,75	-1,78	-1,75	-1,74	4,09E-02	1,21E-02	-144
R-HSA-196807	Nicotinate metabolism	7,82E-03	1,74E-02	1,81E-02	1,71E-02	1,63E-02	9,86E-01	1,70E-02	1,63E-02	6,88E-03	-2,42	-2,11	-2,09	-2,12	-2,14	2,20	-2,12	-2,14	-2,46	-1,71	4,36E-02	2,17E+00	-1
R-HSA-1474228	Degradation of the extracellular matrix	6,59E-02	6,25E-02	3,75E-02	5,49E-02	5,48E-02	3,91E-02	2,53E-02	3,29E-02	3,68E-02	-1,51	-1,53	-1,78	-1,60	-1,60	-1,76	-1,95	-1,84	-1,79	-1,71	4,39E-02	2,34E-02	-73
R-HSA-70171	Glycolysis	4,13E-02	4,12E-02	4,37E-02	4,05E-02	4,13E-02	5,18E-02	3,93E-02	5,48E-02	4,58E-02	-1,74	-1,74	-1,71	-1,74	-1,74	-1,63	-1,76	-1,60	-1,69	-1,70	4,42E-02	3,09E-03	-552
R-HSA-70263	Gluconeogenesis	4,13E-02	4,12E-02	4,37E-02	4,05E-02	4,13E-02	5,18E-02	3,93E-02	5,48E-02	4,58E-02	-1,74	-1,74	-1,71	-1,74	-1,74	-1,63	-1,76	-1,60	-1,69	-1,70	4,42E-02	3,09E-03	-552
R-HSA-70921	Histidine catabolism	9,16E-02	3,17E-02	3,36E-02	3,61E-02	8,69E-02	3,64E-02	4,12E-02	4,22E-02	4,07E-02	-1,33	-1,86	-1,83	-1,80	-1,36	-1,79	-1,74	-1,73	-1,74	-1,69	4,59E-02	3,92E-02	-43
R-HSA-71387	Metabolism of carbohydrates	2,18E-02	5,88E-02	3,61E-02	6,64E-02	2,93E-02	7,24E-02	5,31E-02	7,72E-02	3,39E-02	-2,02	-1,57	-1,80	-1,50	-1,89	-1,46	-1,62	-1,42	-1,83	-1,68	4,67E-02	4,46E-02	-38
R-HSA-70614	Amino acid synthesis and interconversion (transamination)	4,58E-02	4,29E-02	3,28E-02	3,31E-02	3,55E-02	5,07E-02	8,10E-02	1,29E-01	5,04E-02	-1,69	-1,72	-1,84	-1,84	-1,81	-1,64	-1,40	-1,13	-1,64	-1,63	5,13E-02	5,42E-02	-30
R-HSA-350864	Regulation of thyroid hormone activity	1,98E-02	1,86E-02	1,64E-02	9,86E-01	1,98E-02	1,89E-02	1,76E-02	1,86E-02	1,98E-02	-2,06	-2,08	-2,14	2,20	-2,06	-2,08	-2,11	-2,08	-2,06	-1,61	5,41E-02	2,04E+00	-1
R-HSA-8866906	TFAP2 (AP-2) family regulates transcription of other transcription factors	5,88E-02	5,58E-02	6,02E-02	5,71E-02	6,17E-02	5,34E-02	6,09E-02	5,59E-02	5,47E-02	-1,56	-1,59	-1,55	-1,58	-1,54	-1,61	-1,55	-1,59	-1,60	-1,58	5,76E-02	6,49E-04	-2428
R-HSA-71384	Ethanol oxidation	6,00E-02	5,81E-02	6,78E-02	5,94E-02	6,35E-02	5,39E-02	5,92E-02	5,71E-02	6,69E-02	-1,55	-1,57	-1,49	-1,56	-1,53	-1,61	-1,56	-1,58	-1,50	-1,55	6,05E-02	1,43E-03	-1086
R-HSA-6794361	Interactions of neurexins and neuroligins at synapses	6,20E-02	5,33E-02	7,91E-02	4,64E-02	9,19E-02	5,29E-02	6,25E-02	6,24E-02	5,59E-02	-1,54	-1,61	-1,41	-1,68	-1,33	-1,62	-1,53	-1,53	-1,59	-1,54	6,19E-02	1,19E-02	-129
R-HSA-	Wax biosynthesis	5,94E-02	6,33E-02	6,90E-02	6,17E-02	6,23E-02	6,25E-02	6,68E-02	6,17E-02	6,11E-02	-1,56	-1,53	-1,48	-1,54	-1,54	-1,53	-1,50	-1,54	-1,55	-1,53	6,31E-02	5,56E-04	-2750

8848584																							
R-HSA-1430728	Metabolism	5,54E-02	7,84E-02	6,84E-02	7,37E-02	2,70E-02	1,00E-01	7,51E-02	8,70E-02	8,49E-02	-1,59	-1,42	-1,49	-1,45	-1,93	-1,28	-1,44	-1,36	-1,37	-1,48	6,94E-02	3,57E-02	-41
R-HSA-1445148	Translocation of GLUT4 to the plasma membrane	8,30E-02	7,88E-02	7,65E-02	7,04E-02	7,38E-02	6,43E-02	6,36E-02	6,15E-02	6,10E-02	-1,39	-1,41	-1,43	-1,47	-1,45	-1,52	-1,53	-1,54	-1,55	-1,48	7,00E-02	3,61E-03	-409
R-HSA-70326	Glucose metabolism	5,97E-02	6,51E-02	6,55E-02	6,76E-02	6,14E-02	7,35E-02	7,71E-02	9,89E-02	7,93E-02	-1,56	-1,51	-1,51	-1,49	-1,54	-1,45	-1,42	-1,29	-1,41	-1,47	7,14E-02	7,01E-03	-209
R-HSA-75876	Synthesis of very long-chain fatty acyl-CoAs	7,16E-02	7,81E-02	7,69E-02	7,03E-02	6,50E-02	7,88E-02	7,04E-02	7,83E-02	7,49E-02	-1,46	-1,42	-1,43	-1,47	-1,51	-1,41	-1,47	-1,42	-1,44	-1,45	7,37E-02	1,19E-03	-1215
R-HSA-1474244	Extracellular matrix organization	1,04E-01	3,55E-02	1,14E-01	5,97E-02	1,20E-01	9,04E-02	8,79E-02	4,69E-02	5,36E-02	-1,26	-1,80	-1,21	-1,56	-1,17	-1,34	-1,35	-1,68	-1,61	-1,44	7,46E-02	5,10E-02	-28
R-HSA-171052	LDL-mediated lipid transport	1,36E-01	6,04E-02	7,67E-02	5,91E-02	1,56E-01	4,25E-02	7,60E-02	6,34E-02	6,25E-02	-1,10	-1,55	-1,43	-1,56	-1,01	-1,72	-1,43	-1,53	-1,53	-1,43	7,64E-02	5,29E-02	-27
R-HSA-5626467	RHO GTPases activate IQGAPs	8,04E-02	8,02E-02	8,30E-02	8,17E-02	8,27E-02	7,80E-02	8,14E-02	7,75E-02	7,57E-02	-1,40	-1,40	-1,39	-1,39	-1,39	-1,42	-1,40	-1,42	-1,43	-1,40	8,00E-02	2,88E-04	-4882
R-HSA-556833	Metabolism of lipids and lipoproteins	7,46E-02	7,03E-02	4,80E-02	9,49E-02	4,62E-02	1,56E-01	1,05E-01	1,40E-01	7,99E-02	-1,44	-1,47	-1,66	-1,31	-1,68	-1,01	-1,26	-1,08	-1,41	-1,37	8,54E-02	5,40E-02	-25