

Analysis of schizophrenia susceptibility variants identified by GWAS: A bioinformatics and molecular genetics approach

by

Michelle Coffee



*Thesis presented in partial fulfilment of the requirements for the degree
Master of Science in Genetics at Stellenbosch University*

Supervisor: Prof L. Warnich
Co-Supervisor: Prof D.J.H. Niehaus

December 2014

The financial assistance of the National Research Foundation (DAAD-NRF) towards this research is hereby acknowledged. Opinions expressed and conclusions arrived at, are those of the author and are not necessarily to be attributed to the DAAD-NRF.

DECLARATION

By submitting this thesis/dissertation 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.

Date: December 2014

SUMMARY

Described as one of the costliest and most debilitating disorders, schizophrenia has proven to be among the greatest challenges for medical researchers. The disorder poses difficulties on all levels: from genotype to phenotype. Even though it is known that there is a substantial genetic contribution to schizophrenia susceptibility (~80%), it is unknown whether this is due to common variants, rare variants, epigenetic factors, polymorphisms in regulatory regions of the genome or a combination of all these factors. Over the past few decades, many approaches have been employed to elucidate the genetic architecture of schizophrenia, with the latest and most promising being genome wide association studies (GWAS). However, nearly a decade after the first GWAS, the limitations are increasingly being recognised and new avenues need to be explored. Studies have recently started to focus on the analysis of non-coding regions of the genome since these regions harbour the majority of variants identified in GWAS thus far.

This study aimed to use recently developed programs that utilize data from large scale studies such as previous GWAS, the Encyclopaedia of DNA Elements (ENCODE), 1000 Genomes, HapMap and Functional Annotation of the Mammalian Genome (FANTOM) to establish a simple, yet effective bioinformatics pipeline for the identification and assessment of variants in regulatory regions. Using the established workflow, 149 single nucleotide polymorphisms (SNPs) in regulatory regions were implicated in schizophrenia susceptibility, with the most significant SNP being rs200981. Pathway and network analysis using the Database for Annotation, Visualization and Integrated Discovery (DAVID) and GeneMANIA respectively indicated that the most frequently affected genes were involved in immune responses or neurodevelopmental processes, which support previous findings. Yet, novel findings of this study implicated processes crucial for DNA packaging (from DNA level to chromatin level).

The second part of the study used restriction fragment length polymorphism analysis of polymerase chain reaction-amplified fragments (PCR-RFLP) to genotype ten of the most significant SNPs (identified by bioinformatic analyses in the first part of the study) in a South African Xhosa cohort of 100 cases and 100 controls, while bi-directional Sanger sequencing was used to confirm the presence of these SNPs. Statistical analyses revealed two haplotypes of regulatory variants, rs200483-rs200485-rs2517611 ($p = 0.0385$; OR = 1.71; 95% CI = 1.01-2.91) and rs200981-rs2517611-rs3129701 ($p = 0.041$; OR = 0.51; 95% CI = 0.27-0.98) associated with schizophrenia susceptibility. Bioinformatic analysis indicated that these

haplotypes affect DNA packaging, which supported the findings of the first part of the study and could implicate epigenetic processes.

The findings of this study support the importance of regulatory variants in schizophrenia susceptibility. This study also showed the importance of combining GWAS data with additional analyses in order to better understand complex diseases. It is hoped that these findings could fuel future research, specifically in genetically unique populations.

OPSOMMING

Skisofrenie kan beskryf word as een van die duurste en mees ernstige siektes en bly steeds een van die grootste uitdagings vir mediese navorsers. Hierdie versteuring behels probleme op alle vlakke: van genotipe tot fenotipe. Alhoewel dit bekend is dat daar 'n aansienlike genetiese bydrae tot skisofrenie vatbaarheid is (~ 80%), is dit onbekend of dit is as gevolg van algemene variasies, skaars variasies, epigenetiese faktore, variasies in regulerende gebiede van die genoom of 'n kombinasie van al hierdie faktore. Oor die afgelope paar dekades is verskeie benaderings gebruik om die genetiese samestelling van skisofrenie te bestudeer, met die nuutste en mees belowende synde genoom-wye assosiasie studies (GWAS). Byna 'n dekade na die eerste GWAS, word die beperkinge egter toenemend erken en nuwe navorsingstrategieë moet gebruik word. Studies het onlangs begin om meer te fokus op die analise van nie-koderende areas van die genoom aangesien hierdie areas die meerderheid van die variasies behels wat tot dusver in GWAS geïdentifiseer is.

Hierdie studie het gepoog om onlangs ontwikkelde programme, wat gebruik maak van die data van grootskaalse studies soos vorige GWAS, die “Encyclopaedia of DNA Elements” (ENCODE), “1000 Genomes”, “HapMap” en “Functional Annotation of the Mammalian Genome” (FANTOM), te implementeer om sodoende 'n eenvoudige, maar doeltreffende bioinformatika pyplyn vir die identifisering en evaluering van variante in regulerende gebiede, te vestig. Deur die gebruik van die gevestigde bioinformatika pyplyn, is 149 enkel nukleotied polimorfismes (SNPs) in regulerende gebiede in skisofrenie vatbaarheid betrek, met rs200981 wat die mees betekenisvol was. Pad- en netwerk-analise met die onderskeidelike hulp van die “Database for Annotation, Visualization and Integrated Discovery” (DAVID) en “GeneMANIA”, het aangedui dat die gene wat die meeste geïmpak was, betrokke is by immuunreaksies en neuro-ontwikkeling. Hierdie bevindinge ondersteun vorige studies. Tog het nuwe bevindinge van hierdie studie prosesse geïmpak wat uiters noodsaaklik is vir DNS verpakking (van DNS- tot chromatien-vlak).

Die tweede deel van die studie het restriksie fragment lengte polimorfisme analise van polimerase ketting reaksie geamplifiseerde fragmente (PKR-RFLP) gebruik om tien van die belangrikste SNPs (wat geïdentifiseer is deur bioinformatiese ontledings in die eerste deel van die studie) in 'n Suid-Afrikaanse Xhosa studiegroep van 100 skisofrenie gevalle en 100 kontroles te genotipeer, terwyl tweerigting Sanger volgordebepaling gebruik is om die teenwoordigheid van hierdie SNPs te bevestig. Statistiese analise het aangedui dat twee

haplotipes van regulerende variante, rs200483-rs200485-rs2517611 ($p = 0.0385$; OR = 1.71; 95% CI = 1.01-2.91) en rs200981-rs2517611-rs3129701 ($p = 0.041$; OR = 0.51; 95% CI = 0.27-0.98), geassosieer is met skisofrenie vatbaarheid. Bioinformatiese analise het aangedui dat hierdie haplotipes DNS verpakkingsprosesse affekteer, wat die bevindinge van die eerste deel van hierdie studie ondersteun en moontlik epigenetiese prosesse impliseer.

Die bevindinge van hierdie studie ondersteun die belangrikheid van regulerende variante in skisofrenie vatbaarheid. Hierdie studie het ook bewys hoe belangrik dit is om GWAS data met addisionele analises te kombineer om sodoende komplekse siektes beter te verstaan. Daar word gehoop dat hierdie resultate tot meer studies in die toekoms sal lei, spesifiek in genetiese unieke bevolkings.

ACKNOWLEDGEMENTS

I would hereby like to express my gratitude to the following people and institutions:

The DAAD-NRF for financial assistance.

Stellenbosch University for providing me with additional bursaries as well as the resources and facilities to complete this project.

DAAD for providing me with the opportunity to receive additional assistance with this project at the DKFZ in Heidelberg, Germany.

My supervisor, Prof Louise Warnich for her guidance, patience and support – particularly during the last few months.

My co-supervisor, Prof Dana Niehaus, for providing the samples and clinical data of the Xhosa cohort.

Mrs Lundi Korkie (Laboratory Manager, Lab 231, Department of Genetics, Stellenbosch University) for all of her help and never-ending moral support (and for telling me I CAN do it!).

Prof Lize van der Merwe for assistance with the statistical analyses (specifically the single SNP and haplotype association analyses).

Dr Naveed Ishaque for his supervision during my time at the DKFZ.

Dr Nathaniel McGregor and Dr Britt Drögemöller (both Postdoctoral fellows in Lab 231, Department of Genetics, Stellenbosch University) for all their help with this project.

My friend of nearly 20 years, Ciska Cockrell, for her moral support and proofreading my work.

My mother, Ruth Coffee. I wrote so many different versions of something good to say and nothing seemed enough for everything you have done for me... (*Thank you*)[∞]

TABLE OF CONTENTS

DECLARATION	i
SUMMARY	ii
OPSOMMING	iv
ACKNOWLEDGEMENTS	vi
TABLE OF CONTENTS	vii
LIST OF FIGURES	x
LIST OF TABLES	xii
LIST OF SYMBOLS AND ABBREVIATIONS	xiv
INTRODUCTION	1
LITERATURE REVIEW	5
2.1 SCHIZOPHRENIA: THE DISORDER	5
2.1.1 General description and a brief history.....	5
2.1.2 Symptoms, onset & diagnosis	7
2.1.3 Treatment.....	9
2.1.4 Aetiology	11
2.2 INITIAL GENETIC STUDIES.....	14
2.3 GENOME WIDE ASSOCIATION STUDIES	15
2.4 THE NON-CODING GENOME	17
2.5 DATA MINING AND BIOINFORMATICS	21
2.6 SCHIZOPHRENIA IN THE SOUTH AFRICAN POPULATION	22
2.7 AIMS & OBJECTIVES	25
IDENTIFICATION OF SINGLE NUCLEOTIDE POLYMORPHISMS IN REGULATORY REGIONS IMPLICATED IN SCHIZOPHRENIA SUSCEPTIBILITY	27
3.1 ABSTRACT.....	27
3.2 INTRODUCTION.....	28
3.3 MATERIALS & METHODS.....	29
3.3.1 Strategy.....	29
3.3.2 Data Sets	30
3.3.3 Significant SNPs from GWAS	31

3.3.4 SNPs in linkage disequilibrium	31
3.3.5 SNP annotation	32
3.3.6 SNPs in regulatory regions	32
3.3.7 Affected regulatory elements and related genes	33
3.3.8 Pathway Analysis	35
3.3.9 Network Analysis	36
3.4 RESULTS.....	37
3.4.1 SNP Annotation.....	37
3.4.2 SNPs in regulatory regions	39
3.4.3 Affected regulatory elements.....	40
3.4.4 Pathway analysis.....	45
3.4.5 Network analysis	45
3.5 DISCUSSION	47
3.5.1 Significant variants	47
3.5.2 Affected genes	51
3.5.3 Pathway and network analysis.....	56
3.6 CONCLUSION	59
INVESTIGATION OF REGULATORY POLYMORPHISMS IN SCHIZOPHRENIA SUSCEPTIBILITY IN A SOUTH AFRICAN XHOSA COHORT	62
4.1 ABSTRACT	62
4.2 INTRODUCTION.....	62
4.3 MATERIALS AND METHODS	65
4.3.1 Patient Samples.....	65
4.3.2 Genotyping	65
4.3.3 Statistical & Bioinformatic Analyses	70
4.4 RESULTS.....	71
4.5 DISCUSSION	74
4.6 CONCLUSION	78
CONCLUSION AND FUTURE DIRECTIONS.....	80
5.1 LIMITATIONS OF THIS STUDY	81
5.2 FUTURE DIRECTIONS.....	81

REFERENCES	84
APPENDIX 1: Conference presentations and Academic Exchanges	113
APPENDIX 2: Scripts used for bioinformatics analysis	114
APPENDIX 3: Supplementary Material – Bioinformatics Results	115
APPENDIX 4: Supplementary Material – Association Study Results	124
APPENDIX 5: Protocols, Reagents And Solutions	128

LIST OF FIGURES

- Figure 2.1.** Allelic spectrum of schizophrenia. Common variants with small effect size are indicated toward the lower right corner of the graph (the red and yellow circles), while rare variants with a substantial effect are indicated toward the upper left corner of the graph (the light blue diamonds). As the risk allele prevalence increases, the genotypic relative risk of the allele decreases. (Figure adapted from Kim et al., 2011; Sullivan et al., 2012.) *Reprinted with permission from Oxford University press and Nature Publishing Group.* 12
- Figure 3.1.** Flowchart of the strategy used to identify and analyse variants in non-coding regions. For the identification of SNPs in regulatory regions, a RegulomeDB score above 3b indicates that a variant is likely to affect binding. 30
- Figure 3.2.** Genome Variation Server (GVS) SNP Annotation with SeattleSeq-137..... 37
- Figure 3.3.** Circle plot depicting the interactions between various SNPs associated with schizophrenia susceptibility for the YRI population, across all cell lines. The cluster of interactions between histone protein genes on chromosome 6 is circled in yellow. Stronger interactions are indicated with thicker red lines. 44
- Figure 3.4.** Network analysis with GeneMANIA. Nodes representing genes that fit the most significant network “ER to Golgi transport vesicle membrane” are highlighted in red. 46
- Figure 4.1.** The locations of the selected SNPs on chromosome 6. The figure is not drawn to scale..... 66
- Figure 4.2.** Minor allele frequencies of the genotyped SNPs in the South African Xhosa cases and controls. The significant haplotypes associated with schizophrenia susceptibility are rs200483-rs200485-rs2517611 and rs200981-rs2517611-rs3129701. All the SNPs occur on chromosome 6, with the exception of the last two SNPs (rs2535629 and rs4687552) which are located on chromosome 3.. 72
- Figure S1.** Full network analysis using GeneMANIA. Query genes (which all mapped to specific pathways using DAVID) are indicated with black nodes. Additional genes (identified by GeneMANIA), which are related to the query genes and fit to the implicated network, are indicated with grey nodes. 122

Figure S2. The gene that had the highest score (80.32) in terms of associations with other genes, was *HLA-DRB5* (circled in red). The associated genes are highlighted in grey and black..... 123

Figure S3. Comparison of minor allele frequencies between the genotyped SNPs in the South African Xhosa cohort and the YRI, CEU and CHB populations according to 1000 Genomes and HapMap. Where no bars are indicated the MAF was 0 or no data was available 124

Figure S4. RFLP gel photos and corresponding chromatograms for each SNP: (a) rs200981 (b) rs17693963 (c) rs13211507 (d) rs2535629 and (e) rs4687552..... 125

Figure S4 (cont.). RFLP gel photos and corresponding chromatograms for each SNP: (f) rs200485 (g) rs200483 (h) rs3129701 (i) rs2517611 and (j) rs2021722. 126

LIST OF TABLES

Table 2.1. Diagnostic criteria for schizophrenia (DSM-V, American Psychiatric Association, 2013) <i>Reprinted with permission from the Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition, (Copyright 2013). American Psychiatric Association.</i>	8
Table 3.1. Scoring system used by RegulomeDB (Boyle <i>et al.</i> , 2012).....	33
Table 3.2. Annotation of functional SNPs.	38
Table 3.3. Top 10 significant SNPs in regulatory regions.	39
Table 3.4. Top 10 predicted affected genes using RegulomeDB.....	40
Table 3.5. Top 10 predicted affected genes using rSNPBase.	40
Table 3.6. Affected regulatory elements and genes due to SNPs in regulatory regions.	41
Table 3.7. Pathway analysis of genes affected by significant variants in regulatory regions.	45
Table 4.1. SNPs selected for genotyping. The regulatory SNPs correspond to and occur in LD with the original SNPs from GWAS.	66
Table 4.2. The forward (F) and reverse (R) primers and reaction conditions that were used to amplify each SNP.	68
Table 4.3. The restriction enzymes and conditions used to detect each of the SNPs.	69
Table 4.4. Results of association analysis with schizophrenia susceptibility.	73
Table 4.5. Results of association analysis with negative symptoms.....	73
Table 4.6. Features affected by haplotypes associated with schizophrenia susceptibility.....	73
Table S1. Most significant SNPs from previous GWAS.	115
Table S2. SNPs in regulatory regions (with a RegulomeDB score between 1a and 3b).	120
Table S3. The top 10 features identified by network analysis with GeneMANIA that fit the query genes	121

Table S4. Results of single SNP analysis of genotyped SNPs with formal thought disorder.
.....127

Table S5. Reagents for RE digest of the PCR products (for RFLP analysis). (In all instances the buffer was 1X, total volume was 20 µl and dH₂O was added)127

Table S6. Concentration of PCR product required according to template size being sequenced.....128

LIST OF SYMBOLS AND ABBREVIATIONS

&	and
~	approximately
©	copyright
°C	degrees Celsius
\$	dollar
=	equal to
<	less than
≤	less than or equal to
μg	microgram
μl	microliter
μM	micromolar
%	percentage
®	registered trademark
X	times
™	trademark
3'	3-prime end
5'	5-prime end
A	adenine
<i>ACTRIA</i>	ARP1 actin-related protein 1 homolog A, contractin alpha gene
ADRs	adverse drug reactions

AIDS	Acquired Immunodeficiency Syndrome
ALA	alanine
ARG	arginine
ASN	asparagine
<i>ATP13A1</i>	ATPase type 13A1 gene
ATPase	adenylpyrophosphatase
<i>BAG4</i>	BCL2-associated athanogene 4 gene
BBID	Biological Biochemical Image Database
BED	Browser Extensible Data
bp	base pairs
BPRS	Brief Psychiatric Rating Scale
BSA	bovine serum albumin
<i>BTN3A2</i>	butyrophilin, subfamily 3, member A2 gene
C	cytosine
<i>C10orf26</i>	chromosome 10 open reading frame 26 gene
<i>CACNA1C</i>	calcium channel, voltage-dependent, L type, alpha 1C subunit gene
<i>CACNB2</i>	calcium channel, voltage-dependent, beta 2 subunit gene
CAGE	Cap Analysis of Gene Expression
<i>CALHM1</i>	calcium homeostasis modulator 1 gene
<i>CALHM2</i>	calcium homeostasis modulator 2 gene
<i>CALHM3</i>	calcium homeostasis modulator 3 gene

CAMs	cell adhesion molecules
CDCV	common disease – common variant
CDRV	common disease – rare variant
CDX2	caudal type homeobox 2
CEU	Northern Europeans from Utah
CHB	Han Chinese in Beijing, China
ChIP-chip	chromatin immunoprecipitation chip
ChIP-seq	chromatin immunoprecipitation sequencing
Chr	chromosome
<i>CHRNA7</i>	Cholinergic receptor, nicotinic, alpha 7 gene
CI	confidence interval
<i>CNNM2</i>	cyclin M2 gene
CNS	central nervous system
CNVs	copy number variants
<i>COMT</i>	catechol-O-methyltransferase gene
<i>CSMD1</i>	CUB and Sushi multiple domains 1 gene
CTCF	CCCTC-binding factor
CYS	cysteine
DAVID	Database for Annotation, Visualization and Integrated Discovery
<i>DISC1</i>	disrupted in schizophrenia 1 gene
DNA	deoxyribonucleic acid
DNase	deoxyribonuclease

DNase-Seq	DNase I hypersensitive sites sequencing
dNTPs	deoxynucleoside triphosphate
<i>DRD2</i>	dopamine receptor D ₂ gene
DSM-V	Diagnostic and Statistical Manual of Mental Disorders, 5 th Edition
e.g.	<i>Exempli gratia</i>
E2F-1	E2F transcription factor 1
<i>EDG4</i>	endothelial differentiation, lysophosphatidic acid G-protein-coupled receptor, 4
ENCODE	Encyclopaedia of DNA Elements
EP300	E1A binding protein p300
eQTL	expression quantitative trait loci
ER	endoplasmic reticulum
<i>et al.</i>	<i>Et alii</i>
etc.	<i>Et cetera</i>
F	forward primer
<i>F2</i>	coagulation factor II
FANTOM5	Functional Annotation of the Mammalian Genome, 5 th Edition
FASTA	FAST-All
FDR	false discovery rate
<i>FLJ12442</i>	5'-nucleotidase domain containing 2
<i>FTO</i>	fat mass and obesity associated gene

g	gram
G1/S	1 st gap phase / synthesis gap phase
GABA	gamma-aminobutyric acid
GAD	glutamic acid decarboxylase
GATA1	GATA binding protein 1
gDNA	genomic deoxyribonucleic acid
GEO	Gene Expression Omnibus
GLN	glutamine
GLU	glutamic acid
GLY	glycine
GO	Gene Ontology
<i>GRIN1</i>	glutamate receptor ionotropic NMDA 1 gene
<i>GRIN2B</i>	glutamate receptor ionotropic NMDA 2B gene
<i>GRIN2C</i>	glutamate receptor ionotropic NMDA 2C gene
GTF2F1	general transcription factor IIF, polypeptide 1
GVS	Genome Variation Server
GWAS	genome-wide associations study
<i>H. sapiens</i>	<i>Homo sapiens</i>
H3meR17	histone 3 (methyl)arginine 17
Hbp1	HMG-box transcription factor 1
HEY1	hes-related family bHLH transcription factor with YRPW motif 1
<i>HIST1H2AG</i>	histone cluster 1, H2ag gene

<i>HIST1H2AH</i>	histone cluster 1, H2ah gene
<i>HIST1H2AL</i>	histone cluster 1, H2al gene
<i>HIST1H2BJ</i>	histone cluster 1, H2bj gene
<i>HIST1H2BK</i>	histone cluster 1, H2bk gene
<i>HIST1H2BPS2</i>	histone cluster 1, H2b, Pseudogene 2
<i>HIST1H4I</i>	histone cluster 1, H4i gene
HIV	Human Immunodeficiency Virus
HLA	Human leukocyte antigen
<i>HLA-A</i>	human leukocyte antigen A gene
<i>HLA-B</i>	human leukocyte antigen B gene
<i>HLA-C</i>	human leukocyte antigen C gene
<i>HLA-DPB1</i>	HLA, class II, DP beta 1 gene
<i>HLA-DQA1</i>	HLA, class II, DQ alpha 1 gene
<i>HLA-DQB1</i>	HLA, class II, DQ beta 1 gene
<i>HLA-DRA</i>	HLA class II, DR alpha chain gene
<i>HLA-DRB1</i>	HLA, class II, DR beta 1 gene
<i>HLA-DRB5</i>	HLA, class II, DR beta 5 gene
<i>HLA-G</i>	human leukocyte antigen G gene
<i>HLA-H</i>	human leukocyte antigen H gene
HMG	high-mobility group
HNF4A	Hepatocyte nuclear factor 4 alpha
<i>HSP90AB1</i>	heat shock protein 90kDa alpha (cytosolic), class B member 1 gene
HWE	Hardy-Weinberg equilibrium

ICD-10	International Classification of Diseases 10 th Revision
ID	identification
ILE	isoleucine
Inc.	incorporation
<i>IRX3</i>	iroquois homeobox 3 gene
ISC	International Schizophrenia Consortium
<i>ITIH4</i>	inter-alpha-trypsin inhibitor heavy chain family, member 4 gene
JPT	Japanese in Tokyo, Japan
kb	kilobase pairs
<i>KCNN3</i>	potassium intermediate/small conductance calcium-activated channel, subfamily N, member 3 gene
KEGG	Kyoto Encyclopedia of Genes and Genomes
<i>KIAA0892</i>	MAU2 Sister Chromatid Cohesion Factor
L	litre
LD	linkage disequilibrium
LEU	leucine
<i>LSM1</i>	U6 snRNA-associated Sm-like protein LSml gene
Ltd	limited
LYS	lysine

MAF	minor allele frequency
Mb	megabase pairs
MET	methionine
Mg ²⁺	magnesium ion
MgCl ₂	magnesium chloride
MGS	Molecular Genetics of Schizophrenia
MHC	major histocompatibility complex
MHC I	major histocompatibility complex class I
MHC II	major histocompatibility complex class II
min	minutes
<i>MIR137</i>	microRNA 137 gene
mM	millimolar
<i>MnSOD</i>	superoxide dismutase 2, mitochondrial gene
mRNA	messenger RNA
ncRNA	non-coding RNA
ng	nanogram
<i>NFATC2</i>	nuclear factor of activated T-cells, cytoplasmic, calcineurin-dependent 2 gene
NFIC	nuclear factor I/C
NFKB1	nuclear factor of kappa light polypeptide gene enhancer in B-cells 1
NHGRI	National Human Genome Research Institute
NK	natural killer

nm	nanometer
NMDA	<i>N</i> -methyl- <i>D</i> -aspartate
NR1H2::RXRA	nuclear receptor subfamily 1, group H, member 2 - retinoid X receptor, alpha duplex
NR2F1	nuclear receptor subfamily 2, group F, member 1
NRF	National Research Foundation
<i>NRG1</i>	neuregulin 1 gene
<i>NRGN</i>	neurogranin gene
<i>NT5C2</i>	5'-nucleotidase, cytosolic II gene
Oct-1	octamer-binding transcription factor 1
OR	odds ratio
OSSE	Online Sample Size Estimator
<i>p</i>	probability
PANSS	Positive and Negative Syndrome Scale
PANTHER	Protein Analysis Through Evolutionary Relationships
PCR	polymerase chain reaction
PGC	Psychiatric Genome-Wide Association Study Consortium
PHE	phenylalanine
PHP	Hypertext Preprocessor
POLR2A	DNA-directed RNA polymerase II subunit RPB1
PolyPhen	Polymorphism Phenotyping

POU2F2	POU class 2 homeobox 2
<i>PPP1R1B</i>	protein phosphatase 1, regulatory subunit 1B gene
<i>PPP2R2B</i>	protein phosphatase 2, regulatory subunit B, beta gene
PRO	proline
Prof	Professor
Pty	private company
Reg	regulatory
RFLP	restriction fragment length polymorphism
RNA	ribonucleic acid
RNase-Seq	ribonuclease sequence-specificity
rpm	revolutions per minute
rs	Reference SNP
SANS	Scale for the Assessment of Negative Symptoms
SAPS	Scale for the Assessment of Positive Symptoms
SB	sodium borate
SDS	sodium dodecyl sulfate
SER	serine
SIFT	Sorting Intolerant from Tolerant
SIN3A	SIN3 transcription regulator family member A
<i>SLC3A2</i>	Solute Carrier Family 3 (Amino Acid Transporter Heavy Chain) Member 2 gene

SNAP	SNP Annotation and Proxy Search Tool
SNP	single nucleotide polymorphism
SPI1	SPI1 transcription factor
SSTAR	Semantic catalogue of Samples, Transcription initiation And Regulators
STAT1	Signal Transducer And Activator Of Transcription 1
STAT3	Signal Transducer And Activator Of Transcription 3
Taq	<i>Thermus aquaticus</i>
TAF1	TATA box binding protein (TBP)-associated factor
TAF7	TAF7 RNA polymerase II, TATA box binding protein (TBP)-associated factor
TBP	TATA-binding protein
<i>TCF4</i>	transcription factor 4 gene
TD	tardive dyskinesia
TF	transcription factor
TFAP2A	transcription factor AP-2 alpha (activating enhancer binding protein 2 alpha)
TFAP2C	transcription factor AP-2 gamma (activating enhancer binding protein 2 gamma)
TFBS	transcription factor binding site
THR	threonine
T_m	melting temperature
<i>TNFα</i>	tumour necrosis factor alpha gene
<i>TNNC1</i>	troponin C type 1 gene

TPM	tags per million
TRAP	Transcription Factor Affinity Prediction
<i>TRIM26</i>	Tripartite motif containing 26 gene
TYR	tyrosine
U	unit (enzyme quantity)
<i>UBC</i>	ubiquitin C gene
UK	United Kingdom
<i>USMG5</i>	up-regulated during skeletal muscle growth 5 gene
UTR	untranslated region
v	version
V	volts
<i>VARSL</i>	valyl-tRNA synthetase 2, mitochondrial gene
<i>WDR51A</i>	WD repeat domain 51A gene
w/v	weight per volume
www	world wide web
YRI	Yoruba in Ibadan, Nigera
YY1	Yin Yang 1 factor

CHAPTER 1

INTRODUCTION

The initial definition of the central dogma of molecular biology seemed simple: DNA encodes for RNA, which in turn encodes proteins. However, since this definition was first articulated by Francis Crick in 1958 (Crick, 1958 as cited in Crick 1970), it has become known that gene expression involves many more factors than previously thought. For example, the non-coding genome which was first considered “junk DNA” is now considered to have a significant role in gene regulation and expression (Ward & Kellis, 2012). This has significant implications for the way in which research studies should be designed, performed and interpreted, particularly regarding complex diseases.

One such a complex disorder is schizophrenia. Genetic predisposition to the disorder is supported by the prevalence of schizophrenia among related individuals, with heritability estimates as high as 80% (Girard *et al.*, 2011). The incidence of schizophrenia in the general population is approximately 1%, while this percentage increases significantly between relatives (Gottesman & Shields, 1967). The genetic contribution to this disorder became apparent during the 1960's (Heston, 1966), which fuelled the search for a candidate gene involved in schizophrenia. While many linkage and candidate gene studies successfully associated a number of variants in/near genes associated with the disorder, it eventually became clear that there is not merely one variant associated with the disorder. According to the SchizophreniaGene/SZGene Database (June 2014, www.szgene.org), more than 1 700 studies regarding schizophrenia have been completed, in which 1008 genes and 8 788 polymorphisms have been researched. There are three key points that should be remembered with these numbers: firstly, not all findings were significant and many of the findings to date have only been replicated once or not at all; the studies tend to rely on small sample sizes and finally, due to technological or financial restraints, only a few markers were assessed for each gene (Kim *et al.*, 2011). In 2005, the first successful GWAS was performed (Klein *et al.*, 2005) and opened the doors for a new hypothesis free approach in schizophrenia research. Though many significant variants have been identified in a number of large GWAS, the pitfalls and gaps of these studies have lead researchers to start looking toward additional avenues to explore. While GWAS explore the whole genome and is effective in identifying common variants with small effect sizes, this does not include rare variants with large effect sizes, nor epigenetic factors (Korte & Farlow, 2013). Furthermore, many of the identified

common variants occur in non-coding regions – a part of the genome that has until recently been very rarely explored. However, in order to effectively study the way in which genetic factors are involved in complex diseases, researchers are starting to realise the potential of including functional interpretation of regulatory variation in association studies (Gaffney, 2013).

The release of ENCODE data has made it significantly easier to study non-coding regions of the human genome (The ENCODE Project Consortium, 2012). In September 2003 this project was launched by the United States National Human Genome Research Institute (NHGRI), with the primary goal of unravelling the non-coding genome and characterizing functional elements. Nine years and approximately \$400 million later, the ENCODE findings spanned a staggering 57 pages in the September 2012 issue of *Nature* (Volume 489, Issue 7414) alone. The amount of data generated by more than 1800 genome-wide experiments was equivalent to the amount obtained if the human genome had to be sequenced 1700 times. Despite these surreal figures, this is still just the beginning. For example, only a mere 10% of the estimated 1400 human transcription factors were analysed with chromatin immunoprecipitation sequencing (ChIP-Seq) data (Dunham, 2013).

With the advent of next generation sequencing techniques and large projects such as that of the ENCODE consortium, vast amounts of data are continuously being generated. Yet, instead of an increased understanding of the genetic mechanisms underlying many diseases, this has led to a research bottleneck. According to Curtis (2013), no next generation sequencing results can be interpreted as biologically meaningful without being followed up. Some of the ways that these results can be of value are by mining variant databases, evaluating the biological consequences and replicating findings of previous studies. Consequently, there is still an immense amount of research that needs to be performed in order to address many biological questions. This suggests that research focus should be adjusted from data generation to data mining and thorough analysis. By combining the findings of projects such as ENCODE and FANTOM, 5th Edition (FANTOM5) with what is already known about the genetic architecture of the human genome, this could provide clues to unravelling the intricate networks involved in complex disorders, such as schizophrenia.

The urgency to resolve the genetic complexity of schizophrenia is due to the heavy burden of this disorder on the affected individuals, caregivers and communities. This is even worse in developing countries and it has been estimated that between 4 and 5 million African

individuals suffer from some form of psychiatric disorder (Purgato *et al.*, 2012). The most recent statistics available for mental health in the South African population (released in 2008), indicated that approximately 30% of South African individuals suffer from some form of psychiatric illness (Stein *et al.*, 2008) and that neuropsychiatric disorders are third only to HIV/AIDS and infectious diseases in terms of the national burden of diseases (Seedat *et al.*, 2009). However, while HIV/AIDS is considered an epidemic and receives a considerable amount of time and money which is dedicated to research and treatment, neuropsychiatric disorders lag behind severely (Lund *et al.*, 2010).

Taking these factors into account, it seems rather tragic that despite modern technology and the research efforts that are being put forth around the world, there is still no cure or effective treatment for schizophrenia. Therefore, by using novel strategies and the multitude of available data as described in this study, it is hoped that clues can be provided to aid in the elucidation of the complex genetic architecture of this disorder.

CHAPTER 2

LITERATURE REVIEW

2.1 SCHIZOPHRENIA: THE DISORDER

2.1.1 General description and a brief history

The Diagnostic and Statistical Manual of Mental Disorders V (DSM-V, American Psychiatric Association, 2013) defines schizophrenia as a debilitating mental disorder characterized by disturbances in behaviour and cognition. This disorder affects approximately 24 million people worldwide (World Health Organisation, 2013), with no signs of decreasing incidences.

Schizophrenia was first described in the late nineteenth century by German psychiatrist Emil Kraepelin (1856-1926) as “dementia praecox” which translates to “dementia of the young” (Walker *et al.*, 2004). This name was very much representative of what is still true today: onset of the illness is commonly observed between the ages of 20 and 25. Kraepelin also described subtypes of the disorder such as “paranoid”, “catatonic” and “hebephrenic” (disorganised) schizophrenia, but it was clear from the beginning that diagnosis of the disorder would be difficult. By the early twentieth century, the more modern term “schizophrenia” was proposed by Eugen Bleuler (1857-1939). This term is derived from the Greek “schizo” and “phren” which means “to split” and “mind” respectively (Walker *et al.*, 2004). Therefore, by definition, schizophrenia means that there is a split in the mind of an individual suffering from the disorder, which directly relates to the DSM-V definition of disruptions in cognition. However, this description of a split mind has often been misunderstood by those that are unfamiliar with the disorder as a split personality disorder, which is incorrect. Even though Bleuler classified symptoms such as hallucinations, delusions, depression and mania as indicators of schizophrenia, he also stated that these symptoms overlapped with other mental illnesses. This had a significant influence on the diagnosis of the disorder. Furthermore, he labelled the broad range of symptoms with varying severity as the “group of schizophrenias”, which is reflective of the highly heterogeneous nature of the disorder today (Bleuler, 1911 as cited in Walker *et al.*, 2004). Nonetheless, the description of schizophrenia that has probably had the most significant impact on diagnosis was given by Kurt Schneider in 1959. He defined key symptoms of

schizophrenia which mainly distinguished between different types of hallucinations and delusions (e.g. thought broadcasting, thought echoing, thought withdrawal and thought intrusion).

Despite the fact that these descriptions of schizophrenia date back to more than a century, the illness still remains poorly understood (Tandon *et al.*, 2008). Statistics indicate that it is the third most debilitating disorder, relative to other mental, neurological and substance abuse disorders (Collins *et al.*, 2011). Furthermore, schizophrenia has been estimated to decrease an affected individuals' life expectancy by 10 to 15 years due to a myriad of accompanying problems such as health issues and an increased likelihood to commit suicide (Sullivan *et al.*, 2012; Mowry & Gratten, 2013). Due to the burden of the disorder as well as the depression that often accompanies it, suicide attempts are estimated to be at 50%, while 7-10% of the individuals succeed (American Psychiatric Association, 2013).

The devastating effects of schizophrenia extend far beyond the affected individual; it is a psychological, socio-economic and family burden. Furthermore, diagnosis is problematic and no specific biomarkers have been identified to date. The disorder also appears to have an obscure pathogenesis and treatment tends to be ineffective (Cacabelos, 2013). The economic implications are mainly due to factors such as health complications, disability, destitution, and low employment rates (Mowry & Gratten, 2013). As mentioned previously, schizophrenics find it difficult to secure employment and will usually require some sort of assistance from family or friends.

Even though there are no exact measures of the burden of schizophrenia, a good indication of the financial burden is provided by a 2012 report by the Personal Social Services Research Unit at the London School of Economics which estimated the yearly costs in the United Kingdom to be £11.8 billion (Knapp *et al.*, 2004). This is due to a number of direct costs (e.g. inpatient treatment and medication) and indirect costs (e.g. due to job loss, homelessness, expenses of caregivers), making it one of the most expensive mental disorders worldwide (Rössler *et al.*, 2005). Furthermore, according to Lauber *et al.* (2005), if the time and effort of primary caregivers could be converted to a monetary unit this amount would equal the costs of inpatient treatment.

2.1.2 Symptoms, onset & diagnosis

Delusions, hallucinations, bizarre behaviour and disordered speech are often part of the diverse symptoms that an individual can exhibit (American Psychiatric Association, 2013); however diagnosis is complicated not only by the range of these symptoms, but also by the severity of the symptoms.

Distinctions between negative, positive and general symptoms have only been made since the late 1980's (Harvey & Walker, 1987). Positive symptoms are usually an exaggeration of normal functions such as hallucinations, delusions and grandiosity, while negative symptoms are synonymous with loss of normal functions which manifest as emotional and social withdrawal and a lack of spontaneity. Finally, general psychopathological symptoms include anxiety, tension, depression and uncooperativeness, to name a few (American Psychiatric Association, 2013). In addition to the broad range of symptoms associated with schizophrenia, the severity of these symptoms can vary and are measured by a multitude of scales such as the Positive and Negative Syndrome Scale (PANSS; Kay *et al.*, 1987), the Brief Psychiatric Rating Scale (BPRS; Overall & Gorham, 1962), Scale for the Assessment of Negative Symptoms (SANS) and Scale for the Assessment of Positive Symptoms (SAPS; Andreasen, 1990). These scales provide a comprehensive means of assessing schizophrenia symptoms, which is fundamental for the accurate diagnoses of disorders comprising a continuous range of clinical symptoms (Niehaus *et al.*, 2005). SANS and SAPS were used to measure symptom severity in the South African Xhosa cohort used in this study (Chapter 4). SANS can be divided into five subgroups of negative symptoms, namely affective blunting, anhedonia/asociality, lack of attention, avolition/apathy and alogia. Similarly, SAPS can be divided into four subgroups of positive symptoms, namely delusions, hallucinations, bizarre behaviour and formal thought disorder. Each of the subgroups is then rated on a scale between 0 (absent) and 5 (severe) (Andreasen, 1990).

Today, the DSM-V (American Psychiatric Association, 2013) is most commonly used as a guideline for licenced health officials to diagnose individuals who suffer from schizophrenia. However, there are alternative diagnostic systems such as the International Classification of Diseases (ICD-10). The criteria used for diagnosis according to the DSM-V are summarised in Table 2.1.

Table 2.1. Diagnostic criteria for schizophrenia (DSM-V, American Psychiatric Association, 2013) *Reprinted with permission from the Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition, (Copyright 2013). American Psychiatric Association.*

A.

Two (or more) of the following, each present for a significant portion of time during a one month period (or less if successfully treated). At least one of these must be (1), (2), or (3):

1. Delusions.
2. Hallucinations.
3. Disorganized speech (e.g., frequent derailment or incoherence).
4. Grossly disorganized or catatonic behaviour.
5. Negative symptoms (i.e., diminished emotional expression).

B.

For a significant portion of the time since the onset of the disturbance, level of functioning in one or more major areas, such as work, interpersonal relations, or self-care, is markedly below the level achieved prior to the onset (or when the onset is in childhood or adolescence, there is failure to achieve expected level of interpersonal, academic, or occupational functioning).

C.

Continuous signs of the disturbance persist for at least 6 months. This 6-month period must include at least 1 month of symptoms (or less if successfully treated) that meet Criterion A (i.e., active-phase symptoms) and may include periods of prodromal or residual symptoms. During these prodromal or residual periods, the signs of the disturbance may be manifested by only negative symptoms or by two or more symptoms listed in Criterion A present in an attenuated form (e.g., odd beliefs, unusual perceptual experiences).

D.

Schizoaffective disorder and depressive or bipolar disorder with psychotic features have been ruled out because either 1) no major depressive or manic episodes have occurred concurrently with the active-phase symptoms, or 2) if mood episodes have occurred during active-phase symptoms, they have been present for a minority of the total duration of the active and residual periods of the illness.

E.

The disturbance is not attributable to the physiological effects of a substance (e.g., a drug of abuse, a medication) or another medical condition.

F.

If there is a history of autism spectrum disorder or a communication disorder of childhood onset, the additional diagnosis of schizophrenia is made only if prominent delusions or hallucinations, in addition to the other required symptoms of schizophrenia, are also present for at least 1 month (or less if successfully treated).

Even though the prevalence of schizophrenia is more or less equal between male and female individuals, males tend to have an earlier onset of the disorder as well as a more debilitating course of the illness (American Psychiatric Association, 2013). Onset is usually observed in late teenage or early adult years; however there are cases of individuals showing symptoms much earlier or later on. According to the DSM-V, onset can be preceded by a “pre-morbid personality” during which an individual may exhibit unusual behaviour, whereas other individuals might appear completely normal before experiencing their first psychotic episode. It is, however, more likely that onset of initial clinical symptoms will be preceded by increasingly bizarre behaviour and that there would have been subtle signs of the disorder as early as childhood (American Psychiatric Association, 2013).

2.1.3 Treatment

Even though there is currently no cure for schizophrenia, treatment regimens to alleviate symptoms are available. Nonetheless, it has been estimated that more than 50% of affected individuals are not receiving appropriate treatment (World Health Organisation, 2013). More than 74% of patients do not fully adhere to treatment regimens, which may be attributed in part to adverse drug reactions (ADRs), and relapse rates are extremely high as a result of the ineffective antipsychotics currently available (Mowry & Gratten, 2013). In a study completed by Lieberman *et al.* (2005), 74% of individuals, in a cohort of 1 493 American schizophrenia patients (including Caucasian, Indian, Alaska Native, Asian and Native Hawaiian individuals as well as individuals of mixed race), did not complete antipsychotic treatment due to intolerable ADRs or inefficacy of the antipsychotics prescribed. The individuals received one of five different antipsychotic drugs for a period of 18 months. The medication with the lowest discontinuation rates was Olanzapine; however it was also associated with greater weight gain and altered lipid and glucose metabolism, which emphasizes the limited effectiveness of these medications. It is therefore vitally important to improve existing antipsychotic medication.

Schizophrenia is most commonly treated with two main classes of antipsychotic medication: first generation and second generation antipsychotics (Xu *et al.*, 2013). First generation antipsychotics emerged in the 1950s and are more effective in the treatment of positive symptoms (Tandon *et al.*, 2010). These agents act by blocking dopamine receptors, thereby reducing dopamine activity. However, ADRs induced by these medications can easily occur

and can be more debilitating than schizophrenia itself. The main side-effect comprises disturbances within motor control and can result in tardive dyskinesia (TD), pseudoparkinsonism, restlessness and dystonia. Even though symptoms such as parkinsonism could be reversible, TD and dystonia last much longer and cannot always be alleviated (Arranz & De Leon, 2007). TD, specifically, occurs in a staggering 20-30% of patients who are treated with first-generation antipsychotics (Srivastava *et al.*, 2006). Due to the harmful effects associated with these first-generation antipsychotics, psychiatrists have opted for second-generation antipsychotics which target a number of different neurotransmitter receptors and can be used for the treatment of positive as well as negative symptoms (Lieberman *et al.*, 2005). Even though second-generation antipsychotics are particularly effective in the treatment of drug-resistant schizophrenia with motor side-effects being a less frequent ADR, other major side-effects can still occur, such as metabolic syndrome, weight gain, cardiomyopathies and agranulocytosis (Cacabelos *et al.*, 2011; Meltzer, 2012). In addition to antipsychotic treatment, patients may undergo psychotherapy or cognitive-behavioural therapy, but more often than not, individuals will have to be admitted to a hospital for treatment. Treatment can result in relatively brief periods of remission, however this is a lifelong disorder characterised by fluctuations in a declining mental state (Tandon *et al.*, 2009).

Despite the fact that schizophrenia is a complex disorder, patients are initially treated according to a standardized regimen which does not always account for the broad heterogeneity of the condition. The type and dosage of medication is then gradually adjusted to suit an individual as far as possible, based on their response to a medication up to that point (Müller *et al.*, 2013). Taking into consideration the heterogeneous nature of schizophrenia as well as variable treatment response and the number of antipsychotics that are available today, this is no easy task (Tandon *et al.*, 2010). Therefore, by understanding the aetiology of schizophrenia and the genetics involved in antipsychotic response, the cost and time to optimize treatment and reduce ADRs can possibly be eliminated.

The field of pharmacogenetics / pharmacogenomics actively researches alternatives to the current failing system of treatment in schizophrenia and aims to move toward personalised medicine (Zandi & Judi, 2010). Treatment tailored for the individual patient will not only alleviate the symptoms of schizophrenia, but could potentially eliminate ADRs. However, progress in the pharmacogenetic research of schizophrenia is hindered by the current lack in knowledge of the underlying mechanisms that contribute to the clinical manifestation of the

disorder (Cacabelos *et al.*, 2011). This obstacle is increasingly being acknowledged, with many major pharmaceutical companies opting to move away from traditional drug-focused research to understanding disease aetiology (Mowry & Gratten, 2013).

2.1.4 Aetiology

The aetiology of schizophrenia remains poorly understood. It is generally accepted that schizophrenia is a neurodevelopmental disorder and furthermore that the clinical manifestations observed are as a result of both genetic and environmental factors (Williams *et al.*, 2009). Various chemical and anatomical changes have been observed in affected individuals, including impaired cortical structures, ventricular enlargement and altered chemical receptors (Stachowiak *et al.*, 2013). Due to disorder complexity and insufficient information regarding the mechanism underlying prognosis, a number of hypotheses have been proposed: neurodevelopmental defects, the polygenic inheritance model, perinatal epigenetic effects, immune dysfunction, seasonal infection and finally, dysfunctions of a number of chemical receptors including the dopamine, choline, serotonin, gamma-aminobutyric acid (GABA) and *N*-methyl-*D*-aspartate (NMDA) receptors (Cacabelos, 2013).

Pedigree analyses have indicated that schizophrenia is inherited in a non-Mendelian manner (Tandon *et al.*, 2008). The complexity of schizophrenia is highlighted by the myriad of loci that have been identified by linkage and candidate gene association studies, however, it is suspected that the majority of genes involved in the disorder are yet to be identified (Girard *et al.*, 2011). Based on the plethora of studies and the inability to replicate many of the findings, it is unlikely that one or a few genetic polymorphisms are responsible for schizophrenia susceptibility. It is expected that multiple variants in numerous genes are involved - potentially hundreds or thousands of variants (Harrison & Weinberger, 2005; Curtis, 2013). Furthermore, this hypothesis is supported by the major overlaps between the symptoms (and in some instances common genetic variants) of schizophrenia and other mental disorders such as autism and bipolar disorder (Gejman *et al.*, 2011; Cross-Disorder Group of the Psychiatric Genomics Consortium, 2013). This indicates that there is not one causal variant specific to one disorder, but rather a complex network of common variants that contribute to complex phenotypes associated with different disorders. Alternatively, there is also the possibility that individual variants such as copy number variants (CNVs) could have a substantial effect on disease aetiology. CNVs are duplications or deletions of genomic

segments ranging between 1 kb to several million bases in length (Cantor & Geschwind, 2008), and may be very rare ($p \ll 0.01$) (Haraldsson *et al.*, 2011; Curtis, 2013). Not only does the surfeit of expected associated polymorphisms complicate diagnosis of schizophrenia, but also the ability to develop effective treatments.

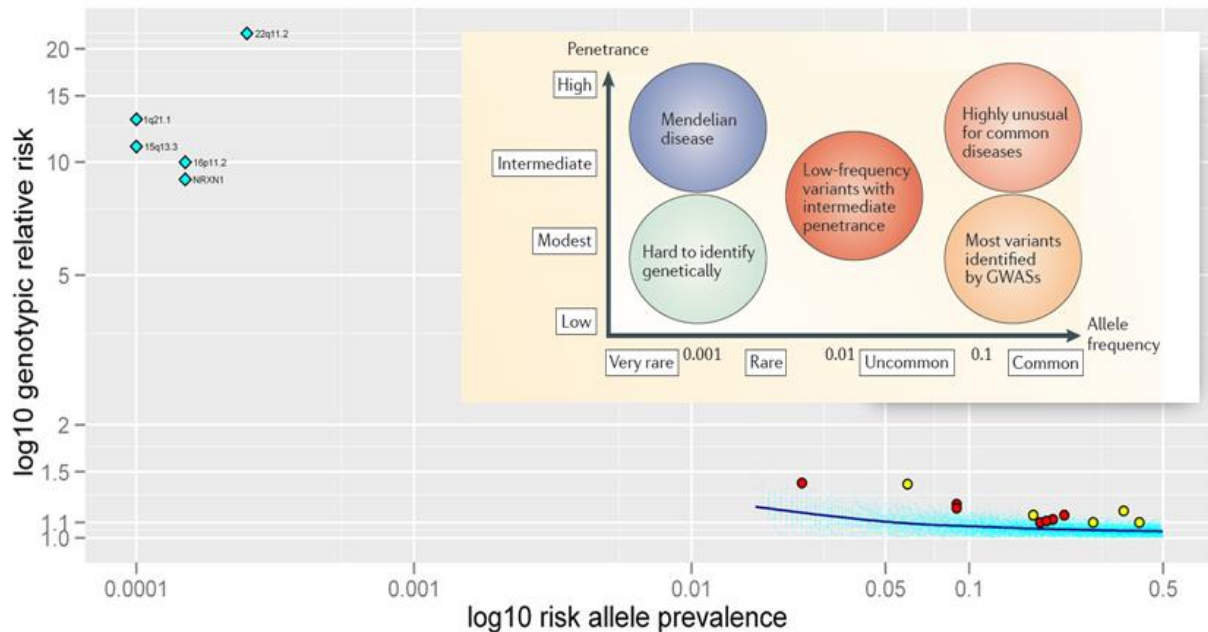


Figure 2.1. Allelic spectrum of schizophrenia. Common variants with small effect size are indicated toward the lower right corner of the graph (the red and yellow circles), while rare variants with a substantial effect are indicated toward the upper left corner of the graph (the light blue diamonds). As the risk allele prevalence increases, the genotypic relative risk of the allele decreases. (Figure adapted from Kim *et al.*, 2011; Sullivan *et al.*, 2012.) Reprinted with permission from Oxford University press and Nature Publishing Group.

These two abovementioned hypotheses, namely the common disease – common variant (CDCV) and common disease – rare variant (CDRV) hypothesis, are at opposite ends of the spectrum (as indicated in Fig. 2.1) and there is much debate regarding the validity of both hypotheses (Sullivan *et al.*, 2012). Arguments for the common variant hypothesis include the extensive number identified by GWAS; believed to be sufficient enough to capture most of the genetic variation present in complex disorders. Furthermore, variations in endophenotypes, which is commonly observed in schizophrenia, are likely due to common and not rare variants (Greenwood *et al.*, 2013). That said, the rare variant hypothesis is supported by the evolutionary theory that states that disease alleles should be rare, deleterious variants that tend to be sporadic and rare CNVs have been implicated in a number of complex psychiatric disorders (Gibson, 2012). It is likely that both hypotheses are partially correct

and that the intricate nature of schizophrenia can be attributed to both rare and common variants (Owen *et al.*, 2010; Mulle, 2012; Wright *et al.*, 2014).

Considering non-genetic contributors to disease, advanced paternal age, prenatal infection, perinatal complications, season of birth and a history of drug abuse have been implicated in current literature (Tandon *et al.*, 2008; Mulle, 2012). The mechanism in which older paternal age affects the risk of schizophrenia is not completely understood, but increased polymorphisms due to aberrant spermatogenesis has been suggested (Byrne *et al.*, 2003). Potentially harmful prenatal and perinatal events include smoking during pregnancy, malnutrition, toxoplasmosis and viral infections such as influenza and rubella (Tandon *et al.*, 2008; 2009). Although the precise mechanism of action for these factors is unclear, it is thought that viral infection during pregnancy leads to an irregular immune response which can cause disruptions in foetal brain development (Ashdown *et al.*, 2006), while smoking and malnutrition causes unnecessary stress which can lead to a hyperdopaminergic state (Lipska *et al.*, 1993). Regarding the season of birth, winter births have been associated with an increased risk for schizophrenia. However, as with the aforementioned environmental risks, this risk factor is not fully understood either (Davies *et al.*, 2003).

It is known that critical neurobiological and cognitive processes in the brain are influenced by epigenetic processes to some extent (Dempster *et al.*, 2013) and it is likely that epigenetic mechanisms mediate the interaction between genetic and environmental factors involved in schizophrenia (Maric & Svrakic, 2012). Numerous studies have found evidence to support this hypothesis. For instance, hyper-methylation has been shown to not only be associated with schizophrenia (Guidotti *et al.*, 2007), but also to significantly worsen symptoms (Petronijevic *et al.*, 2008). Furthermore, methylation patterns of the dopamine receptor D₂ gene (*DRD2*) were shown to differ considerably between monozygotic twins, where only one twin was affected by schizophrenia (Petronis *et al.*, 2003). Finally, it is also possible that histone modifications and changes in chromatin conformation could contribute to the downregulation of several metabolic genes (Akbarian *et al.*, 2005). Post-mortem analysis of the prefrontal cortex of 41 individuals with schizophrenia and 41 matched control individuals indicated that high levels of histone 3 methylation, H3-(methyl)arginine 17 (H3meR17), caused down-regulation of the aforementioned metabolic genes (Akbarian *et al.*, 2005). However, the epigenetic role of histones in schizophrenia has not been explored to the same extent as with other disorders and most of the studies have only focussed on methylation (Dempster *et al.*, 2013), leaving a large gap with regards to other epigenetic factors.

It is important to reiterate the fact that the abovementioned genetic, environmental and epigenetic factors are, in most cases, merely possible explanations contributing to an understanding of the molecular aetiology of schizophrenia and that an absolute rationale remains elusive.

2.2 INITIAL GENETIC STUDIES

The first landmark study to provide evidence of inheritance involved in schizophrenia was completed by Heston (1966). Children with the disorder, born to parents with and without schizophrenia, but adopted by individuals without the disorder, were compared. The results indicated that approximately 10% of children born to schizophrenic parents eventually developed symptoms of the illness, whilst control individuals remained asymptomatic. Another adoption study, wherein the biological relatives of adopted individuals were compared, found that the rate of schizophrenia was much higher among the biological relatives than the adopted relatives (Kety *et al.*, 1976). Together, these studies along with others established the hypothesis of a genetic contribution to schizophrenia. However, the extent of this contribution was still unknown. Twin studies have shown that monozygotic twins who share 100% of their genetic material, have a 48% risk of schizophrenia, while dizygotic twins who share only half of their genetic material have a much lower risk of 17% (Riley & Kendler, 2006). Finally, to put the statistics into perspective, a meta-analysis of 12 published twin studies of schizophrenia estimated the heritability of schizophrenia to be 81%, with an 11% environmental risk (Sullivan *et al.*, 2003).

The results of these landmark studies initiated the search for candidate genes associated with schizophrenia susceptibility. The strongest candidates that have emerged from these studies are primarily involved in chemical signalling or neuronal growth and development. Some of the most significant genes identified through linkage or candidate gene studies include disrupted in schizophrenia 1 (*DISC1*) on chromosome 1, catechol-O-methyltransferase (*COMT*) on chromosome 22, neuregulin 1 (*NRG1*) on chromosome 8 and protein phosphatase 1, regulatory subunit 1B (*PPP1R1B*) on chromosome 17 (Karayiorgou *et al.*, 1995; Blouin *et al.*, 1998; Millar *et al.*, 2000; Li *et al.*, 2006). A translocation breakpoint that disrupts *DISC1* was found to co-segregate with schizophrenia in a large Scottish family (Millar *et al.*, 2000), while a functional polymorphism in the *COMT* gene was shown to increase prefrontal dopamine catabolism, which impairs cognition (Egan *et al.*, 2001).

Furthermore, *NRG1* is expressed in the central nervous system and *NRG1* mutant mice showed a behavioural phenotype similar to schizophrenia mouse models (Stefansson *et al.*, 2002), while post-mortem studies have indicated reduced expression of *PPP1R1B* in the dorsolateral prefrontal cortex of individuals with schizophrenia (Li *et al.*, 2006). Importantly, even though variations in these candidate genes might have been associated with schizophrenia or a subset of phenotypes associated with schizophrenia, the exact manner in which these variants could increase susceptibility to the disorder remains unknown in most cases.

Despite the number of associated loci, only a few genes have consistently been associated with schizophrenia across multiple studies and populations (Guan *et al.*, 2012) and tend to include genes in the major histocompatibility complex (MHC) region on chromosome 6 (discussed in sections 2.3 and 2.4). This is mainly due to weak statistical associations, limited sample sizes, samples that are ethnically heterogeneous, inability to replicate findings in independent studies and, most importantly, the potentially polygenic inheritance of schizophrenia susceptibility (Palmatier *et al.*, 2004). More recently, GWAS have been favoured to find common variants and have proven to be more successful for schizophrenia genetics (Kim *et al.*, 2011; Ripke *et al.*, 2014).

2.3 GENOME WIDE ASSOCIATION STUDIES

As opposed to hypothesis-driven candidate gene studies which have yielded inconsistent results for schizophrenia, GWAS have identified a number of significant and, in some cases, replicated associations (Collins *et al.*, 2012; Wright, 2014). The approach used by GWAS has been particularly useful in identifying causal variants, due to the fact these studies use technologies that take a snapshot of the entire genome, making it possible to analyse large numbers of randomly sampled variants (hundreds of thousands to millions) concurrently in numerous individuals' genomes (Bertram, 2008; Réthelyi *et al.*, 2013).

Initially, GWAS struggled to identify variants associated with schizophrenia passing the threshold p -value ($p \leq 5 \times 10^{-8}$) for genome-wide significance, due to limited sample sizes. This led to a number of collaborations, resulting in consortia such as the Psychiatric GWAS Consortium (PGC) (Ripke *et al.*, 2011), International Schizophrenia Consortium (ISC) (Purcell *et al.*, 2009) and Molecular Genetics of Schizophrenia (MGS) (Shi *et al.*, 2009).

Although not all studies have found polymorphisms passing the threshold for genome-wide significance, there have been a number of important findings. The consortium with the most significant impact on schizophrenia research has been the PCG, with three major studies within the last three years. The first study, published in 2011, identified five novel loci of which microRNA 137 (*MIR137*) had the strongest association with schizophrenia. Interestingly, four of the other strongly associated loci, transcription factor 4 (*TCF4*), *CACNA1C*, CUB and sushi multiple domains 1 (*CSMD1*) and chromosome 10 open reading frame 26 (*C10orf26*), were predicted targets of miR-137 (Ripke *et al.*, 2011). The second GWAS performed by the PGC identified 22 susceptibility loci, of which 13 were novel (Ripke *et al.*, 2013). The most significant finding of this study was the implication of genes involved in calcium signalling, namely calcium channel, voltage-dependent, L type, alpha 1C subunit (*CACNA1C*), calcium channel, voltage-dependent, beta 2 subunit (*CACNB2*), ARP1 actin-related protein 1 homolog A, centractin alpha (*ACTRIA*), cyclin M2 (*CNNM2*), troponin C type 1 (*TNNC1*) and calcium homeostasis modulator 1, 2 and 3 (*CALHM1*, *CALHM2*, *CALHM3*). Additionally, the strongest SNP association occurred in the extended MHC region – a region that has consistently been associated with schizophrenia across multiple studies and populations. Furthermore, this study emphasized the contribution of common variants to the disorder, with a predicted 6 300 – 10 200 SNPs possibly contributing to schizophrenia aetiology and accounting for at least 32% of the variance in risk (Ripke *et al.*, 2013). The most recent study by the PGC and largest in the history of psychiatric research, identified a staggering 108 genetic loci associated with schizophrenia, of which 83 were novel (Ripke *et al.*, 2014). The most important finding of this study was the association of *DRD2*, a gene that has previously been associated with schizophrenia, particularly treatment response (Lencz *et al.*, 2006). Additional loci of importance included those involved in immune responses, which supported previous findings of a link between the immune system and schizophrenia (Ripke *et al.*, 2014). Other noteworthy GWAS findings to date include variants spanning the major histocompatibility complex (MHC) region on chromosome 6, as well as intronic SNPs of *TCF4* and SNPs upstream of the neurogranin gene (*NRGN*) (Shi *et al.*, 2009; Stefansson *et al.*, 2009). While the consistent association of the MHC region is in agreement with the hypothesis of an immune component in schizophrenia, the implication of genes such as *TCF4* and *NRGN* support the notion of aberrant pathways in brain development (Stefansson *et al.*, 2009).

The SNPs identified in the abovementioned studies passed the significance threshold with p -values ranging from 10^{-8} to 10^{-14} and as described, these SNPs are situated near genes involved in various important biological processes. Therefore, even though no conclusive findings have been established regarding the aetiology of the disorder, the regions that have been identified in these studies have offered insights into the pathophysiology of schizophrenia (Owen *et al.*, 2010). Additionally, a few structural abnormalities have also passed the genome-wide significance threshold with duplication and deletion sizes in the range of 0.02 – 9 Mb including up to 49 genes at a time (Mowry & Gratten, 2013). This includes genes known to have roles in neuronal development, neuroprotection or the regulation of proteins in neurons and once again highlights the involvement of multiple genes.

Many variants that have been identified, however, were not associated with any particular gene(s). This indicates the possible involvement of variants situated outside of the coding region of protein-coding genes and warrants further attention to the non-coding regions of the genome.

2.4 THE NON-CODING GENOME

When considering the significance of the non-coding genome, there are a few facts that should be kept in mind. Firstly, the roles of non-coding transcripts are increasingly being recognised in RNA processing and gene regulation; sequence comparisons among species have indicated several non-coding transcripts that are highly conserved pointing toward functionality of these transcripts and finally, many variants in non-coding regions have been implicated in a variety of disorders (Alexander *et al.*, 2010). Therefore, even though it has been long thought that 99% of the genome was simply “junk DNA”, it has become increasingly clear that some of these non-protein coding sequences have critical regulatory functions.

Identifying mutations that affect the expression of genes is a crucial task considering that these variants remain mostly unknown. However, variants that affect regulation of genes are extremely difficult to characterise since *cis*-acting regulatory variants can occur in many different locations ranging from introns and promoters to control elements that are far away (Bray, 2008). SNPs located in introns can alter transcription, splicing or mRNA stability. Transcription can also be affected by SNPs occurring in promoters, while variants in the

3'UTR can significantly impact translation. In each of these cases it is clear that variants in regulatory regions will ultimately affect the protein products (Harrison & Weinberger, 2005; Perkins *et al.*, 2005). One example of a regulatory polymorphism, specifically in the African population, is a SNP that occurs in the promoter region of the tumour necrosis factor α (*TNF- α*) gene. This results in a new transcription factor binding site (TFBS) for octamer-binding transcription factor 1 (Oct-1), which subsequently up-regulates the expression of *TNF- α* and causes an increased risk for cerebral malaria (Prokunina & Alarcón-Riquelme, 2004). This illustrates the importance of considering the roles of affected TFBS for complex disorders. Finally, in 2007 Dina *et al.* indicated that variations within the intronic regions of the fat mass and obesity associated (*FTO*) gene contributes to obesity. However, direct connections between these variants and expression of the *FTO* gene were difficult to establish. Recently, Smemo *et al.* (2014) identified iroquois homeobox 3 (*IRX3*) as a functional distal target of *FTO*. It was shown that the *FTO* intronic region contains a number of *cis*-regulatory elements which mediate functional interactions of *IRX3*. This finding highlighted the significance of non-coding variants as well as the importance of long-range interactions between genomic regions.

It has been suggested that the inconsistent findings of linkage studies might be attributed to non-coding RNA (ncRNA) regulation of schizophrenia susceptibility genes (Perkins *et al.*, 2005). Variants that occur in the regulatory regions could cause a disruption in gene regulation, resulting in altered transcription or translation of genes. Additionally, ncRNA regulation of gene expression is possibly a key process during brain development, which would also support the idea of schizophrenia as a neurodevelopmental disorder (Perkins *et al.*, 2005). In addition to this, the possible implications of regulatory variants in schizophrenia susceptibility is supported by the increasing number of studies that have identified variants associated with schizophrenia within regulatory regions (Palmatier *et al.*, 2004; Stephens *et al.*, 2009; Zhao *et al.*, 2007). A prime example of regulatory polymorphisms conferring risk of schizophrenia can be found in the case of glutamic acid decarboxylase 67 (*GAD67*) – an enzyme involved in GABA synthesis. It has been demonstrated that increased hypermethylation of the promoter in *GADI*, the gene that codes for *GAD67*, may cause a decreased expression of the enzyme, while a SNP in the 5' region of *GADI* may be involved in the onset of schizophrenia during childhood (Costa *et al.*, 2003; Addington *et al.*, 2005; Dong *et al.*, 2005). Lee *et al.* (2012) suggest that variants conferring increased risk for schizophrenia are located across the entire genome and that the architecture

of this disorder is polygenic. In fact, it is estimated that more than 80% of variants are located in non-coding regions, which emphasizes the importance of examining these regions (Manolio *et al.*, 2009). However, large research consortia have mainly focussed on the coding regions and candidate genes closest to these variants, leaving the remaining 99% of the genome unexplored.

An evaluation of the distribution of SNPs with an effect on gene expression in the human genome indicated that 30-60% of SNPs are located in promoter regions and are usually located 100bp from the transcription start site (Maston *et al.*, 2006). This indicates that regulatory elements such as promoters might be a “hot spot” for mutations that contribute to human diseases. Furthermore, Maurano *et al.* (2012) determined that most of the variation identified by GWAS occurs in regulatory regions that are marked by DNase hypersensitive sites. This is of particular importance when studying a disorder such as schizophrenia with potential disruptions during early development, since 88% of DNase hypersensitive sites are active during foetal development. Furthermore, considering the importance of *cis*-regulatory sequences to mediate transcription factor (*trans*-regulatory sequences) binding and transcription for precise and correct gene regulation during development and function, non-coding regions have become an attractive research area to understand disease susceptibility (Cowie *et al.*, 2013). One feature of *cis*-regulatory sequences that is of particular interest to the study of diseases is the fact that these sequences can control gene expression in a spatial and/or temporal manner. Therefore, in the event of a variant occurring in such a regulatory sequence, the effect will be restricted to a specific tissue or stage of development (Cowie *et al.*, 2013). Finally, regulatory variants could also affect epigenetic DNA and chromatin marks, which could provide more clues to understanding the interaction between genetic and environmental factors in schizophrenia risk (Sadec *et al.*, 2014).

The establishment of the ENCODE database (Rosenbloom *et al.*, 2012) has enabled further investigation of the human genome. In 2012, the ENCODE data were published in a series of 30 research articles in three journals, namely *Nature* (Volume 489, Issue 7414), *Genome Research* (Volume 22, Issue 9), and *Genome Biology* (Volume 13, Issue 9). According to the findings of the consortium, 80.4% of the human genome, including over 70 000 promoter regions and almost 400 000 enhancer regions is involved in some kind of biological process and by their definition, qualifies as a “functional element” (The ENCODE Project Consortium, 2012). This was a milestone event in the understanding of the human genome and has caused much anticipation about the potential for improving the understanding of

human diseases. Even though the large amount of ENCODE data has endless possibilities for the study of human disease, the highlight of this data is two-fold in that (1) it is possible to map many elements in one cell type or (2) one element in many cell types. The former is possible by mapping transcription factors and modified histones across chromatin in a specific cell type, while the latter is done by applying DNase-Seq or RNase-Seq to detect elements across many different cell types (Dunham, 2013).

One of the 30 ENCODE series research articles, published by Schaub *et al.* (2012), demonstrated the potential of the ENCODE data by using it to identify functional SNPs in regulatory regions and linking these variants to a number of phenotypes. This study also demonstrated that the functional SNP is often one that occurs in LD with the associated SNP reported by a GWAS, instead of the reported SNP itself and in many instances the functional SNPs were located as far as 170 kb from the lead SNP. One important example from this study was demonstrated with the functional SNP rs1333047 which occurs in a region previously associated with coronary artery disease. While there was no supporting evidence for the functional role of the initial reported lead SNP, rs1333049, rs1333047 (which is in strong LD with rs1333049) overlapped DNase hypersensitivity and ChIP-seq peaks for Signal Transducer And Activator Of Transcription 1 and 3 (*STAT1* and *STAT3*). Furthermore, strong LD between these SNPs was only observed in European populations in which the lead SNP was initially associated with coronary artery disease, but not in African populations, for which the associations with the lead SNPs could not be replicated (Schaub *et al.*, 2012). This has important implications for the interpretation of GWAS results in general and highlighted the fact that due to differences in LD patterns between populations, a SNP might be associated with a specific phenotype in one population but not another (Schaub *et al.*, 2012)

Another landmark project has been that of the FANTOM consortium - the largest consortium to focus on human gene expression. Early in 2014, the 5th edition of this consortium (FANTOM5) published a broad atlas of gene expression in human primary cells, tissues and cell lines. Researchers used Cap Analysis of Gene Expression (CAGE), a technology used to identify active genes, in order to monitor the activity of enhancers and promoters in various cells. A staggering 44 000 enhancers and 180 000 promoters were identified and provides the most complete overview to date of networks that regulate transcription in specific cell types (Forrest *et al.*, 2014).

Despite the progress that is being made regarding the non-coding regions of the genome and how they could affect gene expression, very little is still known about the various TFBS, the factors that bind to these sites and the genes that are regulated by these regions.

2.5 DATA MINING AND BIOINFORMATICS

Due to the increase in GWAS over the past few years, the development of many databases that contain all of the findings and information of these studies have been developed. An example of such a database is HuGE Navigator, a database of human genetic epidemiology that is continuously updated and contains records of GWAS among other types of studies (Yu *et al.*, 2008). The availability of such databases provides a more thorough and time-efficient way of collecting data from various different studies, which is particularly useful when looking at many large studies.

The findings of both the ENCODE (<http://genome.ucsc.edu/ENCODE/>) and FANTOM (<http://fantom.gsc.riken.jp/5/>) projects have been made publically available and provides researchers with a multitude of options to consider as well as an abundance of data with which to work. In addition to the data that is available, many platforms have recently been developed that integrate the findings of these studies with data from other independent studies to facilitate data analyses. One such an example is Haploreg, which was developed subsequent to the findings of the ENCODE consortium (Ward & Kellis, 2012). This tool not only provides an attractive way to access the ENCODE data, but combines the findings of ENCODE with population based data from projects such as HapMap and 1000 Genomes. By using this approach, a SNP can effectively be analysed for regulatory potential in specific populations. Another tool, GWAS3D, has combined the findings of GWAS with the ENCODE data to analyse an entire regulatory region and indicate whether the SNP of interest in that region was identified by previous GWAS or is in fact a SNP in LD with the SNP identified in the GWAS (Li *et al.*, 2013).

In addition to the integrative platforms for the analysis of the ENCODE and FANTOM5 data, the same approach has also been used to facilitate the analysis of data from other large projects such as HapMap (The International HapMap Consortium, 2003) and 1000 Genomes (The 1000 Genomes Project Consortium, 2010). For example, the SNP Annotation and Proxy Search (SNAP) Tool provides researchers with a convenient way of working with

large sets of SNPs (Johnson *et al.*, 2008). Instead of evaluating one SNP at a time, it is possible to evaluate an entire list of SNPs in order to identify all SNPs that are in LD with those SNPs in specific populations.

These are only a few of the examples of the multitude of tools that are available for data mining and analysis. If used correctly, these tools seem to provide researchers with endless possibilities and provide novel strategies to elucidate the underlying genetic mechanisms of many complex disorders.

2.6 SCHIZOPHRENIA IN THE SOUTH AFRICAN POPULATION

A brief overview of 12 schizophrenia GWAS revealed that nine studies focussed on individuals of Caucasian descent (Purcell *et al.*, 2009; Stefansson *et al.*, 2009; Ripke *et al.*, 2011; Steinberg *et al.*, 2011; Bergen *et al.*, 2012; Betcheva *et al.*, 2012; Rietschel *et al.*, 2012; Strange *et al.*, 2012; Lencz *et al.*, 2013), two included a Han-Chinese population (Shi *et al.*, 2011; Yue *et al.*, 2011) and only one study included African American individuals (combined with Caucasian individuals) (Shi *et al.*, 2009). Based on these numbers, it is clear that African patients have been under-represented in GWAS, despite the high levels of genetic variation in African groups (Schuster *et al.*, 2010) and the fact that schizophrenia accounts for approximately 10% of the health burden in sub-Saharan Africa (Lopez *et al.*, 2006). Furthermore, it is estimated that of the ~8% of GWAS to include African individuals, only a small fraction have focussed on specific populations in Africa (Drögemöller *et al.*, 2011). Although an estimated 90% of individuals with the disorder are from developing countries (Clark *et al.*, 2011; World Health Organisation, 2013), research in developing countries tends to focus more on current epidemics such as HIV, tuberculosis or malaria (Drögemöller *et al.*, 2011). Finally, results of studies such as the abovementioned GWAS cannot necessarily be extrapolated to African genomes, which further reduces the number of potentially significant findings for Sub-Saharan African individuals suffering from schizophrenia.

To date, the phenotype of schizophrenia among South African Xhosa patients has been well studied (Koen *et al.*, 2006; Niehaus *et al.*, 2005, 2008). Even though little is known about what might cause schizophrenia in Xhosa individuals, it is thought that unique cultural factors could contribute to onset of the disorder, specifically in young males (Le Roux *et al.*, 2007). It has been shown that initiation rites of males that form a pivotal part of Xhosa culture could

act as a trigger for the onset of schizophrenia or relapse of a psychotic episode (Le Roux *et al.*, 2007). Additionally, it is likely that cultural factors could also have an impact on how psychiatric disorders are perceived within African populations (Yen & Wilbraham, 2003). Moreover, stigma associated with mental disorders not only places an unnecessary burden on individuals suffering from schizophrenia, but also hinders treatment.

Only a few studies have researched the genetic basis of schizophrenia in South African individuals and such studies have mostly focussed on the Afrikaner population (Hall *et al.*, 2002; Abecasis *et al.*, 2004; Wiehahn *et al.*, 2004; Hsu *et al.*, 2007; Savitz *et al.*, 2007; Roos *et al.*, 2009; Xu *et al.*, 2009; Rodriguez-Murillo *et al.*, 2014). Nonetheless, it is known that Afrikaners are genetically more similar to Europeans than Xhosa or other Sub-Saharan individuals and although the phenotype might at first glance appear to be similar between patients, studies have shown that there are indeed clinical differences between Xhosa and Afrikaner patients (Niehaus *et al.*, 2005). Therefore it is crucial to include other South African populations, more representative of the majority of the nation, in psychiatric genetics research.

Genetic studies conducted in the Xhosa population regarding schizophrenia risk, have investigated variants in the potassium intermediate/small conductance calcium-activated channel, subfamily N, member 3 gene (*KCNN3*) and protein phosphatase 2, regulatory subunit B, beta (*PPP2R2B*), both known to have a role in the central nervous system and superoxide dismutase 2, mitochondrial (*MnSOD*) (Laurent *et al.*, 2003; Hitzeroth *et al.*, 2007). However, association analyses failed to link any of these genes to schizophrenia risk in the Xhosa population. A series of studies conducted by Riley *et al.* (1996a; 1996b; 2000; 2002) investigated regions on chromosomes 6, 9, 12, 13, 15, 17 and 22 in the South African Bantu-speaking population, which includes Xhosa individuals. The most significant findings were the association of polymorphisms in the coding regions of the alpha-7 acetylcholine receptor gene (*CHRNA7*) as well as the subunits that code for NMDA receptors, glutamate receptor ionotropic NMDA 1, 2B and 2C (*GRIN1*, *GRIN2B*, *GRIN2C*). Nonetheless, Riley *et al.* (1996a; 1996b) was unable to identify a significant association between various chromosome 6p and chromosome 22 markers and schizophrenia in the Bantu-speaking individuals. These regions have consistently been associated in genetic studies of schizophrenia that included individuals of European descent and emphasizes the fact that findings from one population are not necessarily applicable to another population. However, this is not always the case. In 2012, Wright *et al.*, identified associations between *COMT*

genetic variants and schizophrenia susceptibility in a Xhosa cohort of 238 schizophrenia patients and 240 healthy individuals. This supported previous associations of *COMT* with schizophrenia in other populations (Shifman *et al.*, 2002; Chen *et al.*, 2004)

African populations are particularly important in the study of complex diseases due to greater levels of genetic diversity and genetic adaptations as a result of environmental effects amongst others. Furthermore, these groups have extensive population substructure and less linkage disequilibrium (LD) between loci in comparison with other populations (Campbell & Tishkoff, 2008). Of the approximately 52 million individuals that constitute the South African population (Statistics South Africa, 2013), approximately 9 million individuals belong to the Xhosa population, which makes it the second largest ethnic group in the country (Department of Government Communication and Information System, 2013). The importance of including Xhosa individuals in genetic studies is demonstrated by the fact that they are an ethnically homogenous group, which means that confounding factors and, in the case of schizophrenia, phenotypic heterogeneity, is limited (Niehaus *et al.*, 2005). Furthermore, due to the fact that African populations are the oldest, valuable information can be obtained through genetic studies that include groups such as the Xhosa population.

Given the current statistics for South African individuals with psychiatric disorders, the state of mental health care and the lack of understanding of these disorders, it is clear that this area requires a great deal of research. Early diagnosis is vital for improved treatment outcomes; therefore identifying the genetic causes of schizophrenia can aid in understanding the disorder and thereby facilitate accurate and early diagnoses in the future. Additionally, understanding the genetic architecture of schizophrenia may allow for treatment to be optimised which would improve treatment response and limit the current ADRs associated with antipsychotic medication.

2.7 AIMS & OBJECTIVES

In an effort to contribute to the elucidation of the underlying genetic components of schizophrenia, this study used a novel approach based on data-mining and the utilization of bioinformatics tools to study the variants involved in schizophrenia susceptibility. Non-coding regions were examined instead of only investigating genes most closely linked to significant variants. The strength of this study lied in the abundant data available for analyses as well as the unique, clinically well characterised, Xhosa cohort investigated. It was hoped that by broadening the genetic focus to include non-coding regions, a more comprehensive overview of the genetic architecture of schizophrenia would be determined.

The aim of this study was to identify regulatory regions which could potentially be involved in schizophrenia susceptibility and to determine which genes are regulated by these regions.

The main objectives of this study included the following:

- Mine existing databases containing GWAS data (e.g. HuGENavigator) as well as the literature, to identify variants that have significantly ($P \leq 5 \times 10^{-8}$) been associated with schizophrenia susceptibility.
- Identify variants that are in linkage disequilibrium (LD) with the abovementioned polymorphisms, using databases containing known variation such as HapMap and the 1000 Genomes Project.
- Determine whether any of the identified variants occur within regulatory regions with the use of ENCODE or other relevant data.
- Characterise any variation in regulatory regions with the use of predictive programs to determine the effect that the variants have on the function of the regulatory regions.
- Identify genes that are regulated by these regulatory regions using ENCODE data.
- Perform pathway and network analysis based on the predicted affected genes.
- Genotype the most significant variants in the South African Xhosa cohort.
- Perform statistical analyses to determine a possible association between significant variants and schizophrenia susceptibility within the Xhosa population.

CHAPTER 3

IDENTIFICATION OF SINGLE NUCLEOTIDE POLYMORPHISMS IN REGULATORY REGIONS IMPLICATED IN SCHIZOPHRENIA SUSCEPTIBILITY

3.1 ABSTRACT

Schizophrenia is a debilitating mental disorder, which remains poorly understood. Despite the increase in GWAS focussing on this disorder over the past few years, most significant associations have been for variants in non-coding regions. This study aimed to analyse variants in these regions that could be associated with schizophrenia susceptibility. Variants associated with the disorder by GWAS as well as all variants occurring in LD with these were identified. Using a range of freely available data and webtools (ENCODE, FANTOM5, RegulomeDB, rSNPBase, GWAS3D), variants were analysed for regulatory potential as well as potential effects on motifs, protein binding and gene expression. Finally, pathway and network analysis (using DAVID and GeneMANIA) were performed. This study provided evidence for regulatory variants in regions previously associated with schizophrenia risk, but also identified novel variants with strong regulatory evidence that could be involved in schizophrenia susceptibility. The most significant of these was rs200981, a variant with regulatory effects on genes located in the major histocompatibility (MHC) region as well as a number of histone protein genes in the extended MHC region. This provides support for the hypothesis of an immune component conferring an increased risk for schizophrenia. Additionally, this variant was shown to affect processes crucial to DNA, nucleosome and chromatin packaging. Furthermore, many of the variants identified to have regulatory potential have been implicated in neurodevelopment, which in addition provides supporting evidence for this hypothesis. It is suggested that further research should focus on two areas of interest: (i) interactions between the genetic architecture of immune responses and dysregulation in neurodevelopment and (ii) aberrant DNA packaging involved in schizophrenia susceptibility.

3.2 INTRODUCTION

Over the past few years there has been a significant increase in the number of GWAS to identify genetic variants that are associated with a wide variety of disease and traits. However, the majority of variants that have been associated with any specific disorder seem to occur in non-coding regions. While most GWAS initially focussed on the closest genes to variants in these regions, the release of the ENCODE data has made it possible to focus more specifically on the genes that are actually affected by variants in non-coding regions, whether it is due to proximal or distal interactions.

Since the release of the ENCODE data there has been a surge of disease studies focussing on the link between variants in non-coding regions and disease phenotype. It has been shown, for example, that variants which were associated with a disorder in a GWAS, but with no supporting functional evidence, could occur in strong LD with a SNP that overlaps with DNase hypersensitive sites or ChIP-seq peaks. A study published by Schaub *et al.* (2012), demonstrated the potential of the ENCODE data by using it to identify functional SNPs in regulatory regions and linking these variants to a number of phenotypes. This study also demonstrated that the functional SNP is often one that occurs in LD with the associated SNP reported by a GWAS, instead of the reported SNP itself. These findings, combined with the traditional identification of functional variants in protein-coding regions, could provide better insights into the molecular mechanisms by which certain genes are affected by polymorphisms.

Due to the increase in next generation sequencing studies, there has also been an increase in the amount of data generated. As the analysis of large datasets is rather time-consuming, this has inevitably caused somewhat of a bottleneck in research outputs. In an effort to make data mining and handling more efficient, many tools currently exist that can be applied to any study of human diseases and disorders. GWAS databases are usually the first tools to be accessed when identifying variants associated with a specific disorder. While many databases exist, HuGE Navigator and the NHGRI Catalog of Published GWAS are two of the most popular, convenient and user-friendly databases which are regularly updated. As of 1 July 2014, the NHGRI Catalog contains information from 1926 studies and 13403 SNPs. Additional databases useful for the analysis of human variation include HapMap and 1000 Genomes. However, tools that combine the information of both these databases, such as the

SNAP tool, provide an even more convenient alternative when working with large sets of data.

Traditional analysis of variants associated with human disease have relied on applications such as Sorting Intolerant from Tolerant (SIFT) and Polymorphism Phenotyping (PolyPhen) which annotate variants in protein-coding regions. However useful this may be, the results of these applications might be better understood if combined with tools that are capable of annotating variants in non-coding regions as well. Such bioinformatic tools include RegulomeDB, Haploreg, GWAS3D and rSNPBase. These applications act as integrative platforms which combine information from the ENCODE studies with those of HapMap and 1000 Genomes in order to provide a more comprehensive overview of the variants involved in human disorders. Importantly, due to the plethora of tools available, there is currently no set protocol for the analysis of large sets of SNPs in non-coding regions. It is likely that as current tools are improved, the most effective methods for analysis will become more apparent.

This study will use the abovementioned databases and bioinformatic applications to identify variants specifically in regulatory regions which are associated with schizophrenia susceptibility. This disorder is still poorly understood and is difficult to diagnose. While the aetiology of schizophrenia eludes researchers, it is expected that a multitude of variants will be involved in the genetic architecture. This disorder serves as a perfect case study to test the hypothesis of linking variants in coding regions to variants in non-coding regions, due to the plethora of associated variants in the latter. Furthermore, an intricate network of linked variants might explain the complexity of the disorder as well as the general inability to replicate findings between studies.

This study provides a relatively simple method for identifying and analysing schizophrenia associated variants that occur in regulatory regions. Furthermore, this method could in theory be applied to the study of any disorder or trait.

3.3 MATERIALS & METHODS

3.3.1 Strategy

An overview of the strategy used for this study is summarized in Fig. 3.1.

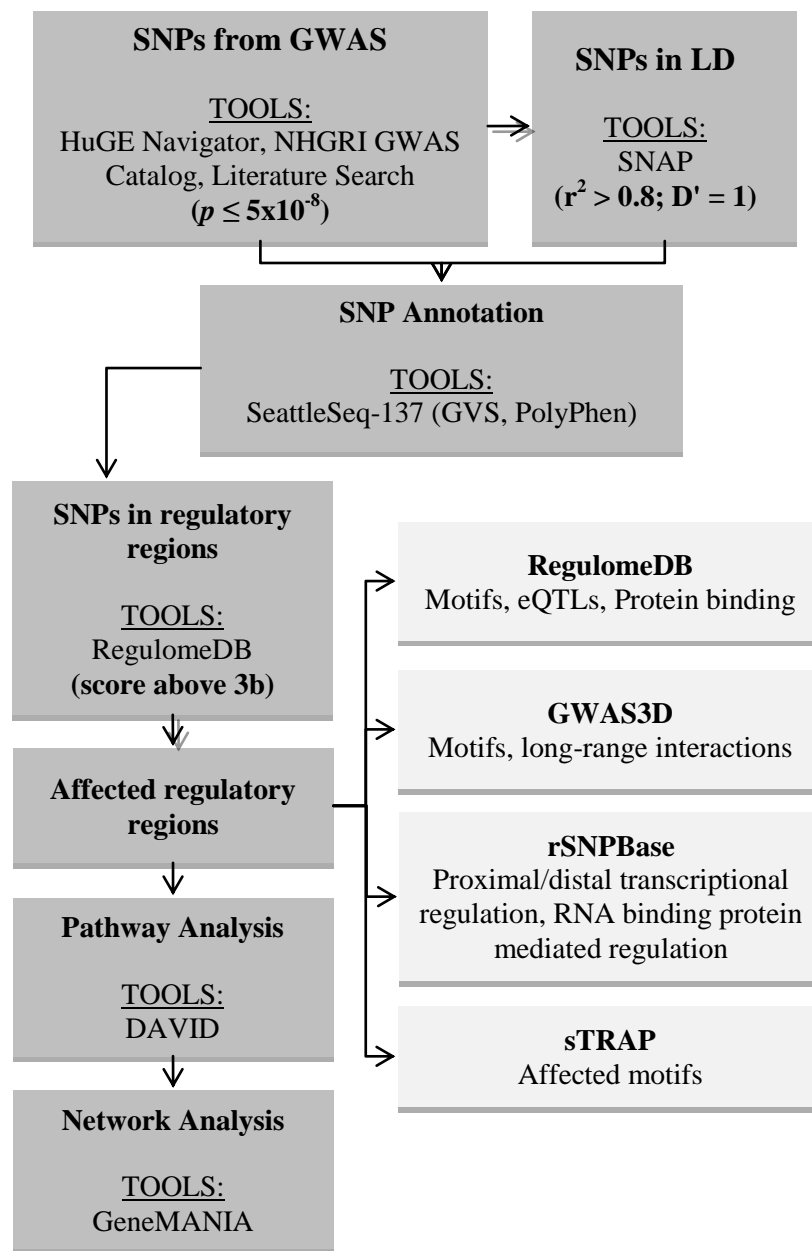


Figure 3.1. Flowchart of the strategy used to identify and analyse variants in non-coding regions. For the identification of SNPs in regulatory regions, a RegulomeDB score above 3b indicates that a variant is likely to affect binding.

3.3.2 Data Sets

Publically available data was used for this study. GWAS data from the PGC (Ripke *et al.*, 2011), ISC (Purcell *et al.*, 2009) and MGS (Shi *et al.*, 2009) as well as eight additional

studies were used (Accessed February 2013, with the exception of one GWAS by Lencz *et al.*, 2013 which was accessed August 2013). Furthermore, data from the 1000 Genomes and HapMap consortia were also accessed for each of the populations corresponding to the populations used in the respective GWAS.

3.3.3 Significant SNPs from GWAS

All significant SNPs from previous schizophrenia GWAS were identified using HuGE Navigator (Yu, *et al.* 2008; Accessed February 2013), the NHGRI Catalog of Published GWAS (Hindorff *et al.*, 2009; Accessed February 2013) as well as a literature search, using terms such as “schizophrenia susceptibility”, “schizophrenia GWAS” and “schizophrenia susceptibility GWAS”. Due to the multiple tests performed for a GWAS, there is a greater likelihood of false positive associations. Therefore, in order to determine appropriate genome-wide significance, a stringent threshold value is crucial and has been estimated to be approximately $P \leq 5 \times 10^{-8}$ (Risch & Merikangas, 1996). Only variants surpassing this threshold were selected.

3.3.4 SNPs in linkage disequilibrium

To identify all SNPs that occur in LD with the significant SNPs obtained from GWAS, the SNAP search tool was used (Johnson *et al.*, 2008; Accessed April 2013). This web-based tool utilizes data from both the 1000 Genomes Project (The 1000 Genomes Project Consortium, 2010) as well as HapMap (The International HapMap Consortium, 2003). It is extremely important to evaluate SNPs that are in LD, because often the reported significant SNP is actually in LD with a functional SNP. These databases contain extensive information regarding human variation, for example 1000 Genomes (The 1000 Genomes Project Consortium, 2010) contains information for approximately 38 million SNPs. Taking into consideration that the significant SNPs from previous GWAS were significant in different populations, analyses of LD in various populations had to be done. The list of significant SNPs was divided into three groups according to the populations in which the variants were identified, namely European, Asian and African American. These groups of SNPs were each uploaded to SNAP separately to search for SNPs that occur in LD with the significant SNPs identified by GWAS, in the Northern Europeans from Utah (CEU), Yoruba in Ibadan,

Nigeria (YRI) and finally, Han Chinese in Beijing, China and Japanese in Tokyo, Japan (CHB+JPT) populations. For each set of SNPs and in each population, 1000 Genomes, HapMap release 21, release 22 and 3 was searched. Threshold values included an r^2 -value of 0.8, to select for SNPs in strong LD, and distance limit of 500 kb. Even though SNPs that are separated by larger distances could be in LD, this distance limit is the largest possible limit when using SNAP. This is due to the fact that the HapMap consortium only did pairwise LD calculations for SNPs within 500 kb of one another and that LD signal tends to decrease significantly beyond 500 kb (Johnson *et al.*, 2008). Finally, the list of resulting SNPs was filtered for SNPs in LD with a D' value of 1.

3.3.5 SNP annotation

SeattleSeq Annotation 137 (National Heart, Lung, and Blood Institute, 2013; Accessed May 2013) was used for functional annotation of the significant SNPs from previous GWAS as well as all SNPs occurring in LD with these SNPs. As with all the other programs used for analysis, the annotation based on the human genome reference sequence GRCh37 was selected. The list of SNPs to be annotated was uploaded in a custom format, which included chromosome, location, reference allele and first allele. The resulting file contained information from various programs related to annotation. The results of the Genome Variation Server (GVS) database were used to create an overview of the SNPs involved. Furthermore, results of PolyPhen (Adzhubei *et al.*, 2010) were used to identify functional variants. Variants predicted to be “probably damaging” or “possibly damaging” were selected for further analyses.

3.3.6 SNPs in regulatory regions

Based on the results of SNP annotation that indicated most SNPs associated with schizophrenia susceptibility are not located in any known genes, it was necessary to determine which of the SNPs identified in GWAS and those in LD occur in regulatory regions. RegulomeDB (<http://regulome.stanford.edu/>; Accessed May 2013) is a tool which assists in the identification and interpretation of regulatory polymorphisms by using ENCODE datasets amongst others, and providing scores to identify potentially functional variants. Prediction methods that are used by RegulomeDB are based on the potential effects

of variants on binding motifs and DNase footprinting that indicates which proteins will bind to a specific site. The data that is used include chromatin state, expression quantitative trait loci (eQTL) and CHIP-seq information as well as manually curated regulatory regions. Finally, scores are based on available eQTL, TF binding, TF motif, DNase footprint and DNase peak information. Scores between 1a and 3b indicate that a variant is likely to affect binding and were therefore included for further analyses. Importantly, RegulomeDB was also used due to the scoring system (specific for regulatory polymorphisms) that is automatically implemented in this application (Table 3.1), which served as a means of prioritizing significant variants.

Table 3.1. Scoring system used by RegulomeDB (Boyle *et al.*, 2012).

SCORE	DATA AVAILABLE TO SUPPORT SCORE	
1a	eQTL + TF binding + matched TF motif + matched DNase Footprint + DNase peak	} Likely to affect binding and linked to expression of a gene target
1b	eQTL + TF binding + any motif + DNase Footprint + DNase peak	
1c	eQTL + TF binding + matched TF motif + DNase peak	
1d	eQTL + TF binding + any motif + DNase peak	
1e	eQTL + TF binding + matched TF motif	
1f	eQTL + TF binding / DNase peak	
2a	TF binding + matched TF motif + matched DNase Footprint + DNase peak	} Likely to affect binding
2b	TF binding + any motif + DNase Footprint + DNase peak	
2c	TF binding + matched TF motif + DNase peak	
3a	TF binding + any motif + DNase peak	} Less likely to affect binding
3b	TF binding + matched TF motif	
4	TF binding + DNase peak	} Minimal binding evidence
5	TF binding or DNase peak	
6	other	

3.3.7 Affected regulatory elements and related genes

Apart from determining the potential effects of variants with RegulomeDB, tools such as GWAS3D (Li *et al.*, 2013; <http://jjwanglab.org/gwas3d>; Accessed February 2014) and rSNPBase (Guo *et al.*, 2013; <http://rsnp.psych.ac.cn/>; Accessed May 2014) were also used to

identify proximal and distal interactions for all variants with a RegulomeDB score between 1a and 3b. GWAS3D is an integrative web server that uses information based on genome-wide associations, chromatin states and chromosome interactions to analyse human regulatory variants (Li *et al.*, 2013). rSNPBase is a database of curated regulatory SNPs (rSNPs) which have been annotated with reference to experimentally validated regulatory elements (Guo *et al.*, 2013). While GWAS3D is effective in identifying proximal and long-range interactions between a regulatory variant and a genome region, rSNPBase is somewhat more effective and precise in identifying which type of regulation is being affected (proximal and distal transcriptional regulation and post-transcriptional regulation) and can identify the potentially regulated genes. Analysis with GWAS3D was done by uploading the list of SNPs in single SNP ID (dbSNP) format. The option for “without GWAS association statistics” was selected, since this had been done previously. The list of SNPs were analysed for regulatory effects in the various relevant populations, using the Hapmap I+II+III or 1000 Genomes Pilot 1 SNP data set. All other default values were used. For analysis with rSNPBase, the “list search” option was used, after which the list of regulatory SNP IDs were uploaded.

Additionally, the Transcription Factor Affinity Prediction (TRAP) tool (Thomas-Chollier *et al.*, 2011), which is a unix-based application, was also used to determine which SNPs affect TFBS. More specifically, sTRAP was used which investigates how potential TFBS are affected by differences in sequences. The advantages of sTRAP include the fact that the program does not rely on a threshold value, but rather that all positions in a given sequence are used to determine the affinity for a transcription factor in the presence of a SNP and that the program relies on collections provided by the JASPAR (Bryne *et al.*, 2008) and TRANSFAC® (Matys *et al.*, 2006) databases. While a web-based version of this tool is available, the R-package of sTRAP is more effective when analysing large amounts of variants at once. The program requires two files: firstly a FASTA file with the SNP of interest as well as a select number of bp before and after the SNP, and secondly a PAIRS file that simply defines which sequences containing the “wild type” or “mutated” alleles go together. For the first file, all of the SNPs from previous GWAS as well as those in LD were analysed. The list of SNPs had to be in a Browser Extensible Data (BED) file. Thereafter, BED files representing the genomic region 15 bp before and after each SNP were also generated (in separate files). Subsequently, FASTA files were generated for each SNP as well as the defined regions before and after each SNP. Finally, these FASTA files had to be

combined; therefore the script “weave_fasta_for_sTRAP.pl” was written (Appendix 2). Using this file, along with the PAIRS file, sTRAP could analyse each SNP.

Importantly, in order to determine whether any TFBS were enriched for schizophrenia specifically, a set of controls were also analysed with sTRAP. Four additional, random disorders/traits were selected including Bipolar Disorder, Hypertension, Diabetes and Height. For each of these disorders, the most significant SNPs from GWAS as well as the SNPs in LD were identified and then analysed with sTRAP. Additionally, SNPs surrounding the previously analysed SNPs associated with schizophrenia susceptibility were also analysed with sTRAP. For this, the “slop” function in BedTools was used to identify all SNPs in the 50bp before and after each of the SNPs associated with schizophrenia susceptibility before the SNPs could be analysed with sTRAP. Once the final results were obtained, the affected motif identifiers had to be searched on the Jaspas Core Database (<http://jaspar.genereg.net/>; Accessed September & October 2013). Only motifs related to *H. sapiens* were selected.

Finally, the expression of the most frequently affected genes was assessed using FANTOM data. While there are many tools that incorporate FANTOM data, the Semantic catalogue of Samples, Transcription initiation And Regulators (SSTAR) was used for this study (<http://fantom.gsc.riken.jp/5/sstar/>; Accessed May 2014). Each gene was searched on the database, specifically in *H. sapiens* (Taxonomy ID: 9606). After the results were obtained for all cell lines, only results for cells related to the human brain were selected. Six cell lines were relevant, namely: (i) Smooth Muscle Cells - Brain Vascular, donor2.CNhs11900.11315-117D1, (ii) Smooth Muscle Cells - Brain Vascular, donor1.CNhs10863.11234-116D1, (iii) Smooth Muscle Cells - Brain Vascular, donor3.CNhs12004.11391-118C5, (iv) brain, adult, pool1.CNhs10617.10012-101C3, (v) brain, adult, donor1.CNhs11796.10084-102B3, (vi) brain, fetal, pool1.CNhs11797.10085-102B4. The average expression level, in tags per million (TPM), of each affected gene was calculated over the six cell lines.

3.3.8 Pathway Analysis

Pathway analysis was performed using the Database for Annotation, Visualization and Integrated Discovery version 6.7 (DAVID v6.7; <http://david.abcc.ncifcrf.gov/>; Accessed June 2014). The advantage of using DAVID is that the tool relies on novel functional annotation algorithms/tools and can easily group large gene sets into functional clusters. This tool relies

on key concepts such as the “DAVID Gene Concept” which uses a graph theory evidence-based method to identify and group genes or proteins from a variety of public genomic resources and the “Term/Gene Co-Occurrence Probability” concept which analyses the co-occurrence of functions with sets of genes. This facilitates biological interpretation of significant findings. Therefore, DAVID was used in order to determine which pathways are overrepresented by the predicted affected genes. Functional annotation clustering (using the DAVID Functional Annotation Clustering Tool) was performed by uploading a list of genes consisting of the genes affected by damaging SNPs (as predicted by Polyphen) as well as the genes affected by variants in non-coding regions with a RegulomeDB score of 3b or higher (as predicted by RegulomeDB and rSNPBase). All default checked items were unchecked and only the tools related to pathway analysis, namely Biological Biochemical Image Database (BBID), BioCarta, Kyoto Encyclopedia of Genes and Genomes (KEGG), Protein Analysis Through Evolutionary Relationships (PANTHER) and REACTOME were checked. Pathways specific to *H. sapiens* were selected. Ideally, only annotation clusters with an enrichment score of 1.3 or higher would be considered significant since a score of 1.3 is equivalent to non-log scale 0.05, however it is also important that clusters with lower enrichment scores should still be considered for further analyses since these clusters might contain individual genes of importance (Huang *et al.*, 2009a; 2009b). Finally, DAVID uses the Benjamini-Hochberg method to correct for multiple comparisons.

3.3.9 Network Analysis

In addition to performing the required pathway analysis of the genes predicted to be involved, network analysis was also performed using GeneMANIA (<http://www.genemania.org/>; Accessed June 2014) in order to determine whether any of the predicted pathways fit within a specific network. GeneMANIA is a publically available web-based tool which identifies additional genes related to a set of query genes. This is made possible using co-expression data from Gene Expression Omnibus (GEO), genetic and protein interactions data from BioGRID and I2D, and finally pathway interaction data from Pathway Commons. As of June 2014, GeneMANIA has a database of 2,104 association networks containing 535,774,338 interactions mapped to 161,629 genes from eight organisms (<http://www.genemania.org/>, June 2014).

Furthermore, GeneMANIA uses one of two weighting methods (Warde-Farley *et al.*, 2010). The first is used for longer gene sets, during which GeneMANIA learns from the uploaded gene set and integrates as few new genes into the predicted network as possible, resulting in a more accurate network prediction. However, in the event of shorter gene lists being used (as in this study), the second method is used which is similar to the first weighting method, but GeneMANIA will learn from Gene Ontology (GO) biological process annotations instead of the actual gene set.

For network analysis, only genes identified in selected pathways according to DAVID were uploaded to GeneMANIA and searched within *H. sapiens*. All default parameters were left checked. GeneMANIA also uses Benjamini-Hochberg multiple testing correction and only GO terms with a false discovery rate (FDR) of < 0.05 are reported.

3.4 RESULTS

3.4.1 SNP Annotation

A total of 64 SNPs passing the threshold for genome-wide significance ($p \leq 5 \times 10^{-8}$) from previous studies that investigated variants associated with schizophrenia susceptibility, were identified (Table S1). Furthermore, 1670 SNP in LD with the SNPs from GWAS were identified. Annotation of the combined total of 1734 SNPs with SeattleSeq-137 is summarised in Fig. 3.2.

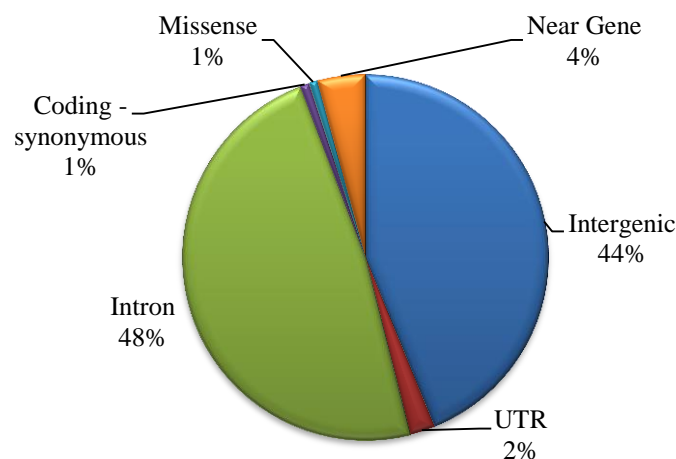


Figure 3.2. Genome Variation Server (GVS) SNP Annotation with SeattleSeq-137.

Table 3.2. Annotation of functional SNPs.

RS ID	CHR	POSITION	SNP	GVS FUNCTION	AMINO ACID CHANGE	PROTEIN POSITION	POLYPHEN	PHASTCONS SCORE	GENE
rs200484	6	27775674	A/G	missense	LEU,PRO	4/127	benign	0.497	<i>HIST1H2BL</i>
rs34788973	6	27879200	A/C	missense	ALA,SER	300/358	probably damaging	0.998	<i>ORB2B2</i>
rs61742093	6	27879982	A/G	missense	ILE,THR	39/358	possibly damaging	0.994	<i>ORB2B2</i>
rs1635	6	28227604	A/C	missense	THR,ASN	152/403	benign	0.003	<i>NKAPL</i>
rs16893892	6	28239873	A/G	missense	TYR,CYS	59/480	probably damaging	0.173	<i>ZNF187</i>
rs33932084	6	28268824	A/G	missense	ASN,SER	398/810	benign	0.992	<i>PGBD1</i>
rs2232423	6	28366151	A/G	missense	MET,THR	11/612	benign	0	<i>ZSCAN12</i>
rs9262143	6	30652781	C/T	missense	GLY,ARG	339/614	probably damaging	1	<i>PPP1R18</i>
rs3132580	6	30920124	A/G	missense	GLU,LYS	1295/1394	benign	0	<i>DPCR1</i>
rs1150752	6	32064726	C/T	missense	THR,ALA	302/4243	benign	0.376	<i>TNXB</i>
rs7775397	6	32261252	G/T	missense	LYS,GLN	400/564	probably damaging	0	<i>C6orf10</i>
rs1265754	6	32303692	A/T	missense	ILE,PHE	150/564	unknown	1	<i>C6orf10</i>
rs2306899	8	38095662	C/T	missense	THR,MET	186/712	possibly damaging	0.236	<i>DDHD2</i>

Only 13 SNPs were predicted to cause missense mutations. However, of these 13 only six were predicted to be probably or possibly damaging by PolyPhen (Table 3.2).

3.4.2 SNPs in regulatory regions

Annotation of the SNPs in non-coding regions with RegulomeDB predicted 149 SNPs to have some type of regulatory potential. While Table 3.3 only contains the ten most significant SNPs that are most likely to affect binding and are linked to the expression of a gene, the remaining SNPs with a score of 1a-3b are summarized in Table S2.

Table 3.3. Top 10 significant SNPs in regulatory regions.

LEAD VARIANT	LEAD VARIANT LOCATION	CHR	P-VALUE	REG VARIANT	REG VARIANT LOCATION	SNP	SCORE ^a	DISTANCE ^b (bp)
rs4687552	52838401	3	1.16E-08	rs2535629	52833218	G/A	1b	5183
rs17693963	27710164	6	3.08E-11	rs200483	27774823	G/A	1d	-64659
rs13211507	28257376		27774823		482553			
rs17693963	27710164	6	3.08E-11	rs200485	27775696	G/C	1b	-65532
rs13211507	28257376		27775696		481680			
rs17693963	27710164	6	3.08E-11	rs200981	27833173	A/G	1a	-123009
rs13211507	28257376		27833173		424203			
rs2523722	30165272	6	1.47E-16	rs3094071	30231767	G/A	1d	-66495
rs2021722	30174130		30231767		-57637			
rs886424	30782001	6	4.54E-08	rs886424	30782001	C/T	1b	0
				rs1059612	30708954	C/T	1d	73047
rs16887244	38031344	8	1.27E-10	rs16887343	38226276	A/G	1b	-194932
rs1488935	38133792		38226276		-92484			
rs2905424	19473444	19	3.44E-09	rs4808967	19640523	T/C	1b	-167079
				rs4808203	19568658	T/C	1b	-95214

^aScore was predicted by RegulomeDB.

^bDistance of regulatory variant from lead SNP.

3.4.3 Affected regulatory elements

The resulting affected regulatory elements by the top 10 variants in regulatory regions as predicted by RegulomeDB, rSNPBase and sTRAP are summarized in Table 3.6.

Furthermore, the top 10 genes predicted to be affected by all of the identified SNPs in regulatory regions (Table S2) by RegulomeDB and rSNPBase, as well as the corresponding expression levels in specific cell lines are summarized in Tables 3.4 and 3.5 respectively.

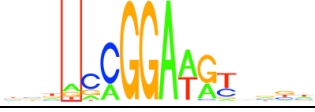

Table 3.4. Top 10 predicted affected genes using RegulomeDB.




GENE	CHROMOSOME	NUMBER OF SNPs	ESTIMATED EXPRESSION (TPM)
<i>HLA-A</i>	6	26	162.48
<i>USMG5</i>	10	18	223.82
<i>BTN3A2</i>	6	12	2.46
<i>HLA-H</i>	6	8	0.24
<i>HLA-DQA1</i>	6	8	6.73
<i>KIAA0892</i>	19	7	79.15
<i>HLA-DRB1</i>	6	7	1.38
<i>EDG4</i>	19	7	2.07
<i>ATP13A1</i>	19	7	28.59
<i>HLA-G</i>	6	6	0.33

Table 3.5. Top 10 predicted affected genes using rSNPBase.

GENE	CHROMOSOME	NUMBER OF SNPs	ESTIMATED EXPRESSION (TPM)
<i>CNNM2</i>	10	24	7.50
<i>HIST1H2AH</i>	6	23	208.33
<i>ZSCAN31</i>	6	21	3.87
<i>MIR3143</i>	6	21	NA
<i>HIST1H2BK</i>	6	20	39.41
<i>HIST1H2AG</i>	6	20	208.33
<i>HIST1H2BJ</i>	6	17	39.68
<i>GATAD2A</i>	19	17	13.18
<i>HIST1H4I</i>	6	15	594.76
<i>NT5C2</i>	10	13	22.50

Table 3.6. Affected regulatory elements and genes due to SNPs in regulatory regions.

REGULATORY VARIANT ^{a,d}	SNP	eQTL ^a	ASSOCIATED GENES ^b	BOUND PROTEIN ^a	MOTIFS ^c	POSITION WEIGHT MATRIX (PWM) ^a
rs2535629	G/A	<i>ITIH4, FLJ12442, WDR51A</i>	<i>ITIH3, PBRM1, GNL3, NEK4</i>	CTCF, RAD21	c-Ets-1	
rs200483	G/A	<i>HLA-H, HLA-A, Hs.158943, BTN3A2</i>	<i>HIST1H2AI, HIST1H3H, HIST1H4PS1, HIST1H2BJ, HIST1H2AG, HIST1H4I, HIST1H2BK, HIST1H2AH, MIR3143, ZNF391, ZSCAN31, HIST1H2AD, HIST1H3D, HIST1H2BF, HIST1H4E, PRSS16, ANKRD17, HIST1H4H, ING5, HIST1H4B, HIST1H3B, HIST1H1C, HIST1H2BE, HIST1H2BG, HIST1H2AE, HIST1H3G, HIST1H2APS4, HIST1H2BI</i>	POLR2A, CTCF, HEY1, MAX, CEBPB, ETS1, TAF7, MXI1	Smad3	

rs200485	G/C	<i>Hs.158943</i>	<i>HIST1H2BL, HIST1H2AI, HIST1H3H, HIST1H2BD, HIST1H2BJ, HIST1H2AG, HIST1H3B, HIST1H2AB, HIST1H4E, HIST1H4I, HIST1H2BK, HIST1H2AH, MIR3143, ZSCAN31, HIST1H2AD, HIST1H3D, HIST1H2BF, PRSS16, TOB2P1, SETD1B, NCOA3, ANKRD17, HIST1H2AC, HIST1H4H, HIST1H4B, HIST1H1C, HIST1H2BC, HIST1H2BE, HIST1H2BG, HIST1H2AE, HIST1H3G, HIST1H2APS4, HIST1H2BI, WDR74, CCT2</i>	POLR2A, BCL3, HEY1, RDBP, CDX2, GATA1, TAF1, SP1, GTF2F1, RFX5, CEPBP, NFYA, NFYB, POU2F2, TAF7, SIN3A, REST,	SOX9	
rs200981	A/G	<i>Hs.158943</i>	<i>HIST1H2AL, HIST1H2BPS2, HIST1H2AG, HIST1H2BK, HIST1H2AH, MIR3143, ZSCAN31, OR2B6, SCAND3, HIST1H2BG, HIST1H2AE, HIST1H4I, LPCAT1, HIST1H1E, HIST1H2BD, HIST1H2APS4, HIST1H2BI, HSP90AB1, RNU5E-1, HIST1H2AB, HIST1H4G, HIST1H3F, HIST1H2BH, TRIM27, RSPH3</i>	POLR2A, GATA1, HEY1, E2F1, E2F4, EP300, TBP, NFKB1, GTF2F1, YY1, CDX2, TAF1, SPI1, POU2F2, JUND, SIN3A, TAF7	E2F-1 <u>NR2F1, HNF4A,</u> <u>NR1H2::RXRA,</u> <u>NFIC</u>	
rs3094071	G/A	<i>HLA-A</i>		TFAP2A, TFAP2C, JUN, JUND, EP300	Hbp1	

Finally, circle plots are easily generated by GWAS3D, which depict the intricate interactions between various regions of the genome. Since analysis of the variants between the different populations (CEU, CHB+JPT, YRI) showed subtle differences between the numbers of affected regulatory elements, the results for the YRI population were used for further analysis (Fig. 3.3).

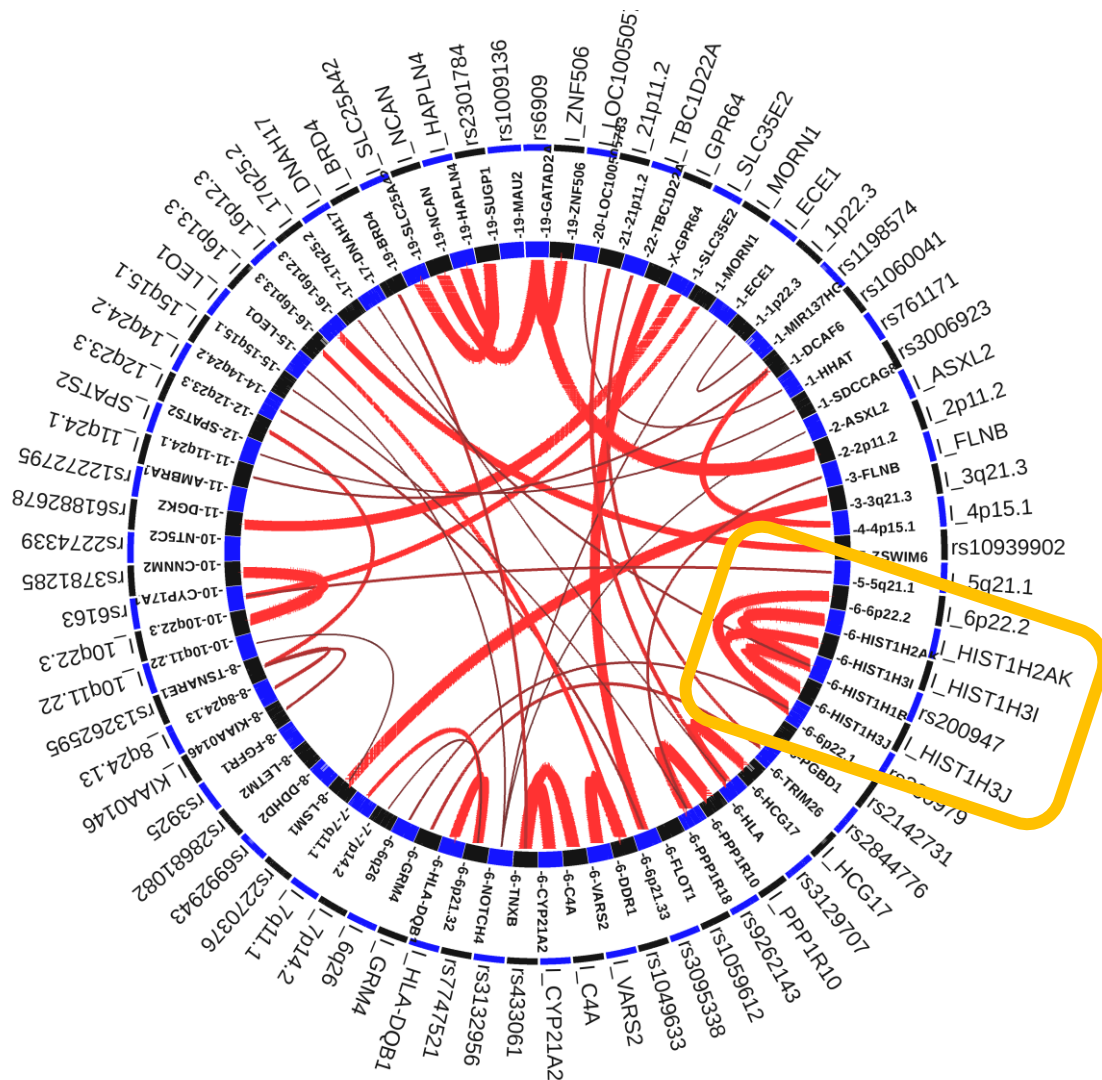


Figure 3.3. Circle plot depicting the interactions between various SNPs associated with schizophrenia susceptibility for the YRI population, across all cell lines. The cluster of interactions between histone protein genes on chromosome 6 is circled in yellow. Stronger interactions are indicated with thicker red lines.

Importantly, analysis with sTRAP not only indicated which TFBS were affected by the presence of a specific SNP, but also showed similar results between the affected TFBS of schizophrenia and the control disorders/traits.

3.4.4 Pathway analysis

Only one cluster was obtained for the set of query genes (Table 3.7); however this cluster had a relatively high enrichment score (4.69).

The terms in the cluster, as identified by the indicated pathway analysis tools in column 1, are sorted according to *p*-value which is calculated using a modified Fisher's exact test. The count indicates the number of query genes that fit the term.

Table 3.7. Pathway analysis of genes affected by significant variants in regulatory regions.

CLUSTER 1: ENRICHMENT SCORE = 4.69				
Pathway Analysis Tool	Term	Count	<i>p</i>-value	Benjamini
KEGG	Allograft rejection	10	2.79E-09	1.17E-07
KEGG	Graft-versus-host disease	10	6.04E-09	1.69E-07
KEGG	Type I diabetes mellitus	10	1.22E-08	2.57E-07
KEGG	Autoimmune thyroid disease	10	7.44E-08	1.25E-06
KEGG	Antigen processing and presentation	11	5.67E-07	7.94E-06
KEGG	Viral myocarditis	9	1.38E-05	1.66E-04
KEGG	Asthma	6	7.97E-05	8.36E-04
KEGG	Cell adhesion molecules (CAMs)	10	2.18E-04	2.04E-03
KEGG	Intestinal immune network for IgA production	6	9.97E-04	8.35E-03
REACTOME	Signaling in Immune system	13	1.91E-02	2.27E-01
KEGG	Natural killer cell mediated cytotoxicity	4	3.56E-01	9.41E-01
KEGG	Hematopoietic cell lineage	3	3.40E-01	9.53E-01

3.4.5 Network analysis

The results of GeneMANIA were ranked according to the FDR value (or *q*-value), which is the significance value after correction for multiple testing. According to this value, the most significant GO term was “endoplasmic reticulum (ER) to Golgi transport vesicle membrane” (Fig. 3.4), with a FDR of 8.38×10^{-42} . However, when focussing on the network terms with the most genes that fit into the network, “antigen processing and presentation of exogenous peptide antigen” would be considered the most significant term, with 24 genes matching this

3.5 DISCUSSION

The goal of this study was to analyse variants within non-coding regions of the genome, which could affect gene regulation and potentially contribute to schizophrenia susceptibility. Significant SNPs from previous GWAS, as well as the corresponding variants in LD, were analysed using a bioinformatics pipeline that was established during this study. Although bioinformatic analyses are largely predictive, many of the datasets used in this study comprised experimentally validated data. Furthermore, the advantage of using bioinformatic tools lies in the possibility to analyse large datasets to create a comprehensive overview of variants associated with a specific disorder.

Considering the fact that most variants are located in non-coding regions, the standard approach of analysing the effects of non-synonymous mutations on the gene products of protein coding genes seems inadequate when studying complex disorders. Therefore this study used a different approach, whereby non-coding variants were analysed. However, the most effective analysis will include non-coding as well as protein coding regions.

3.5.1 Significant variants

The first significant finding of this study is the fact that 12 of the 13 functional variants, as predicted by PolyPhen, occurred on chromosome 6 (Table 3.2). This was not surprising, since this is the most consistently linked region to schizophrenia susceptibility in the human genome (Purcell *et al.*, 2009; Shi *et al.*, 2009; Stefansson *et al.*, 2009). Furthermore, functional annotation of the variants predicted to be associated with the disorder highlighted the need to analyse non-coding regions of the genome. This was proved by the fact that only 13 of the initial 1734 variants (identified by previous GWAS and linkage disequilibrium analysis in this study), that were analysed were predicted to cause missense mutations, of which a mere six SNPs had “probably damaging” or “possibly damaging” effects (Table 3.2).

Of the top 10 SNPs in regulatory regions, rs200981 was the most significant, with a RegulomeDB score of 1a. Furthermore, this SNP appeared to affect one nearby gene, *Hs.158943* (an associated eQTL, but with unknown function) and affected a total of 25 distal genes (Table 3.6). This SNP is located in the MHC region on chromosome 6. It is important to reiterate that this specific region contains the most significant and best replicated GWAS findings associated with schizophrenia to date (Purcell *et al.*, 2009; Shi *et al.*, 2009;

Stefansson *et al.*, 2009). Furthermore, this SNP predominantly affected extended MHC histone protein genes, which have previously been linked to schizophrenia and are mainly involved in the regulation of DNA transcription and repair (Müller & Schwarz, 2010). This highlights the possibility of mutations in fundamental regulatory processes that could contribute to schizophrenia vulnerability.

SNP rs200981 is in LD with two SNPs, rs17693963 and rs13211507, which have previously been associated with schizophrenia susceptibility in two separate GWAS in the European population (Stefansson *et al.*, 2009; Bergen *et al.*, 2012). However, as with many GWAS findings, there was no functional evidence to support the association of these two SNPs, only the regions that it was located in. In contrast, rs200981 was found to overlap with an E2F transcription factor 1 (E2F-1) DNase footprint region as well as a number of ChIP-seq peaks (for POLR2A, GATA1, HEY1, E2F1, E2F4, EP300, TBP, NFKB1, GTF2F1, YY1, CDX2, TAF1, SPI1, POU2F2, JUND, SIN3A and TAF7), which indicates affected protein binding (Table 3.6). One of the affected proteins of particular interest was E1A binding protein p300 (EP300) – a protein which is critical for the regulation of cell growth and division. Furthermore, this protein also ensures normal development before and after birth (Eckner *et al.*, 1994). Importantly, this gene has previously been linked to schizophrenia in a European cohort and more recently in the largest psychiatric GWAS to date, where more than 150 000 individuals were included (Severinsen, 2006; Ripke *et al.*, 2014). Additional analysis of affected TFBS with sTRAP, indicated that TF affinity was affected at four other sites, namely NR2F1, HNF4A, NR1H2::RXRA and NFIC (Table 3.6). The largest change in TF affinity was for NR2F1, with a 23-fold increase for this TF at the related TFBS. Further analysis with rSNPBase, indicated a large number of additional genes that were potentially also affected by rs200981 (Table 3.6). Importantly, the histone cluster 1, H2b, Pseudogene 2 (*HIST1H2BPS2*) was affected through proximal transcriptional regulation, the histone cluster 1H2al gene (*HIST1H2AL*) by RNA binding protein mediated regulation and the rest of the genes by distal transcriptional regulation. Despite the surplus of information obtained for this regulatory SNP, the exact mechanism by which it could affect the predicted regulatory elements and genes remains unclear. Considering that E2F-1 belongs to the E2F family of transcription factors, which play an important role in the 1st gap phase / synthesis gap phase (G1/S) transition of cell cycle control and the action of tumor suppressor proteins (Bieda *et al.*, 2006), a variant affecting E2F-1 could have many extended effects on a number of proteins and genes. Furthermore, a study aimed at identifying all binding sites of E2F-1 used

chromatin immunoprecipitation chip (ChIP-chip) analysis to show that E2F-1 binds to more than 20% of all human promoters (Bieda *et al.*, 2006). It was suggested that E2F-1 is involved in the regulation of a large fractions of genes and therefore variants affecting this factor could have many implications.

Finally, network analysis of the genes predicted to be affected by rs200981 using GeneMANIA revealed association with functions such as “nucleosomes”, “DNA bending complex”, “DNA packaging complex”, “nucleosome assembly”, “chromatin assembly” and “protein-DNA complex assembly”, amongst others. These findings are in accordance with a recent study that suggested that aberrations in common underlying molecular processes were involved in schizophrenia susceptibility (Luo *et al.*, 2013). This study used the top schizophrenia susceptibility genes identified to date and performed protein-protein interaction network analysis on the proteins encoded by these genes. Additionally, the findings were subsequently confirmed using transcriptome data from a clinical sample. The highly interconnected protein network identified in this study highlighted that top schizophrenia susceptibility genes are involved in common fundamental molecular processes. This might also explain the heterogeneous nature of schizophrenia and the fact that many different pathways have been associated with the disorder (Luo *et al.*, 2013).

Two more variants, namely rs200485 and rs200483, which were also among the top 10 regulatory polymorphisms affected the binding of many of the same proteins, but the one that is probably most important is DNA-directed RNA polymerase II subunit RPB1 (POLR2A). The *POLR2A* gene codes for the largest subunit of RNA polymerase II, which is crucial for mRNA synthesis – a fundamental process in the human body. Both rs200485 and rs200483 appeared to affect the same histone protein genes associated with the extended MHC region as rs200981 – and were also in LD with the same two lead SNPs as mentioned above (rs17693963 and rs13211507). Even though rs200485 and rs200483 are located within 1000 bp of each other, these two SNPs are not only located very far from rs200981, but also the associated lead SNPs rs17693963 and rs13211507 (refer to Table 3.3 for distances in bp). This highlights the importance of analysing variants that occur in LD with GWAS findings, instead of merely looking at nearby variants or genes.

The significance of the second most important variant, rs886424, was two-fold. Not only did this SNP have a very high RegulomeDB score (1b), but it is a lead SNP from a previous GWAS. This SNP affects human leukocyte antigen (HLA) genes as well as other genes such

as valyl-tRNA synthetase 2, mitochondrial (*VARSL*) and butyrophilin, subfamily 3, member A2 (*BTN3A2*) which are also located in the MHC region. The variant also overlapped with a ChIP-seq peak for *POLR2A*, drawing attention to this factor once again. Additionally, the 10th most significant variant on the list, rs1059612, occurs in LD with rs886424 and was predicted to affect *HLA-A*, which highlights the importance of this variant as well as the involvement of HLA genes and possibly the immune system in schizophrenia susceptibility. Furthermore, rs1059612 also overlapped with ChIP-seq peaks for CCCTC-binding factor (*CTCF*) and *RAD21*, which recently received strong support with regards to association with schizophrenia susceptibility (Juraeva *et al.*, 2014). These two factors are known to interact, particularly in neurons (Guo *et al.*, 2012; Monahan *et al.*, 2012). The significance of CTCF is that it is a transcriptional regulator that modulates chromatin changes and has been shown to be a key regulator of neuronal differentiation (Phillips and Corces, 2009; Hirayama *et al.*, 2012). Oddly enough, rs1059612 and rs886424 were not predicted to affect any genes according to the findings of rSNPBase, despite the strong evidence given by RegulomeDB. Though, this does not mean that these two variants are not involved in schizophrenia susceptibility. Considering the available tools that can be used to assess the regulatory potential of a variant and the fact that these tools are still in their infancy (and constantly updated), it is possible that the link between rs1059612 and rs886424 and a known gene could have been missed.

The last significant SNP on chromosome 6 was rs3094071, which occurred in LD with two SNPs previously identified in two separate GWAS, namely rs2523722 and rs2021722. Once again, there was no functional evidence to support the association of either of these two SNPs with schizophrenia in the previous studies, however the findings of this study indicated an overlap between rs3094071 and an eQTL for *HLA-A*. Additional supporting evidence for rs3094071 include an overlap with a high-mobility group (HMG) box transcription factor 1 (Hbp1) DNase footprint region as well as overlaps with ChIP-Seq peaks for TFAP2A, TFAP2C, JUN, JUND and EP300.

Besides the variants on chromosome 6, there were also significant regulatory polymorphisms located on chromosomes 3 (rs2535629), 8 (rs16887343) and 19 (rs4808967 and rs4808203). Importantly, both variants on chromosome 19 occur in LD with the same SNP identified in a GWAS, namely rs2905424. These two regulatory variants seemed to affect the same eQTL (*ATP13A1*, *EDG4*, *KIAA0892*), despite showing overlaps with different DNase footprint

regions and ChIP-Seq peaks. This indicates that if these genes are indeed affected, it would most likely be due to a cumulative effect between these variants and not merely due to one SNP in one regulatory region. Finally, rs2535629 and rs16887343 were predicted to affect genes such as inter-alpha-trypsin inhibitor heavy chain family, member 4 (*ITIH4*), 5'-nucleotidase domain containing 2 (*FLJ12442*), WD repeat domain 51A (*WDR51A*), U6 snRNA-associated Sm-like protein LSm1 gene (*LSMI*) and BCL2-associated athanogene 4 gene (*BAG4*), which have previously been associated with schizophrenia (So *et al.*, 2010; Ripke *et al.*, 2011; Shi *et al.*, 2011).

3.5.2 Affected genes

3.5.2.1 The HLA and histone protein genes

Analysis of the genes potentially affected by SNPs in regulatory regions revealed that HLA genes were the most common targets according to RegulomeDB prediction, with *HLA-A* being affected by 26 different SNPs. The HLA genes as well as the other genes potentially affected by the regulatory SNPs are listed in Table 3.4.

Even though the first evidence linking this region to schizophrenia dates back as early as 1974 (Cazzullo *et al.*, 1974), the hypothesis of an immune component which contributes to schizophrenia susceptibility has gained more support through GWAS over the past few years (Debnath *et al.*, 2013). Chromosome 6p22.1 has consistently passed the genomewide significance threshold in several of these GWAS (Purcell *et al.*, 2009; Shi *et al.*, 2009; Stefansson *et al.*, 2009). However, the specific HLA variant(s) involved remain unclear. This is partly due to the vast number of genes located in this region as well as the blocks with variants in high LD (Sinkus *et al.*, 2013).

A recent study investigated the differences in expression levels of several HLA genes in the post-mortem hippocampus of individuals with schizophrenia and negative controls (Sinkus *et al.*, 2013). The selected genes all had potential brain specific functions and included MHC class I (MHC I) genes (*HLA-A*, *-B*, *-C* and *-G*) as well as MHC class II (MHC II) genes (*HLA-DRA*, *-DQA1* and *-DRB5*). MHCI molecules are expressed in nearly all cells and play a central role in the immune system. Furthermore, these molecules are crucial for the maintenance of dendrites in the central nervous system (CNS) during development (Boulanger, 2009), and could possibly explain the link between schizophrenia and a

dysregulation in development. Dendritic maintenance involves strengthening important synaptic connections while removing those that are defective (Shatz, 2002). In the event of over-expression of the MHCI genes, excessive maintenance may be caused which could possibly explain the decreased density of dendritic spines observed in subjects with schizophrenia (Rosoklija *et al.*, 2007; Shatz, 2002). Additionally, MHCII proteins are involved in synapse maintenance and are expressed on microglia, which form part of the innate immune system in the brain (Gehrmann *et al.*, 1995). An increased expression of MHCII genes is associated with activated microglia, which is in accordance with elevated levels of activated microglia in individuals with schizophrenia (Rosoklija *et al.*, 2007). In addition to a role in development, MHC genes are also involved in synaptic plasticity in the adult brain. Reduced synaptic function and dendritic spine density are both characteristics that have been associated with schizophrenia, thus supporting the association with MHC genes.

The fact that *HLA-A* was predicted by RegulomeDB to be affected the most of all the HLA genes is supported by the findings of Sinkus *et al.* (2013), where *HLA-A* was found to be the most highly expressed MHCI gene in the post-mortem hippocampal tissue samples of individuals with schizophrenia. These high levels of expression are supported by the FANTOM data, which indicated that *HLA-A* has high expression levels in the brain (162.48 TPM). Furthermore, variants affecting this gene have been implicated in several major independent GWAS, in different populations, which strengthens the support for the involvement of *HLA-A* in schizophrenia.

Another significant HLA gene among the top 10 affected genes was *HLA-DQA1*, which is part of MHC II and has a central role in the immune system. The main function of the HLA-DQA1 protein is to present peptides from extracellular proteins. A study by Shi *et al.* (2009) also implicated this gene in schizophrenia in a cohort consisting of European and African American individuals. *HLA-DRB1* has also been studied extensively in schizophrenia research. The findings have been both positive and negative, with some studies indicating an association between this gene and schizophrenia susceptibility (Li *et al.*, 2001; Wright *et al.*, 2001), while others have been unable to confirm this link (Halley *et al.*, 2013). However, this could possibly be due to the fact that different populations and samples sizes were used for each study.

As indicated by the findings of rSNPBase and GWAS3D, the histone protein genes of the extended MHC region appear to be strong candidates (based on frequency) for schizophrenia susceptibility. Furthermore, three of the genes predicted to be affected, namely histone cluster 1 H2ah, H2ag and H4i (*HIST1H2AH*, *HIST1H2AG* and *HIST1H4I*) show extremely high levels of expression in the brain with 208.33, 208.33 and 594.76 TPM respectively. Variants affecting histones highlight the possibility of aberrant gene regulation in fundamental processes, such as DNA packaging, that could confer an increased risk for this disorder. In the event that histone binding to DNA promoter regions would be affected, TF binding would also be affected and thereby hinder gene expression (Deutsch *et al.*, 2008). Specifically with regards to schizophrenia, studies have shown abnormal nucleo-histone staining patterns in peripheral neutrophils of schizophrenia patients (Issidorides *et al.*, 1975; Issidorides *et al.*, 1978; Sharma, 2005). Furthermore, a post-mortem study of the prefrontal cortex of schizophrenia cases and controls that looked at altered histone protein expression, concluded that even though there was no distinct association between these proteins and schizophrenia susceptibility, there were clear differences between histone protein expression in different subtypes of schizophrenia (Akbarian *et al.*, 2005). The interactions between histone protein genes on chromosome 6 are illustrated by the circle plot in Fig. 3.3. As the name suggests, GWAS3D focusses on variants of GWAS and therefore the genes in Fig. 3.3 are the closest genes to the variants identified in previous GWAS. However, these genes are in LD with the histone protein genes identified in this study. Once again this emphasizes the importance of identifying exactly which genes are affected by regulatory variants as opposed to merely identifying the closest genes. The histone protein genes predicted to be affected most frequently by rSNPBase (*HIST1H2AH*, *HIST1H2BK*, *HIST1H2AG*, *HIST1H2BJ*, *HIST1H4I*) have all only been identified in one study - a GWAS linking chromosome 6 to schizophrenia susceptibility (Shi *et al.*, 2009). Considering the importance of these proteins and the severe implications that variants affecting these genes would have, further analysis of such variants would be warranted.

Although the abovementioned study linked histone protein genes to schizophrenia susceptibility, the mechanism by which these genes could cause an increased risk for schizophrenia remains unknown. Nonetheless, the link between histone disruption and schizophrenia susceptibility might point towards more than the involvement of aberrant gene regulation conferring an increased risk for the disorder. Histones have a particularly important role in protection against infections - it can interfere with bacterial cell membranes

as well as gene expression in microbes (Kawasaki & Iwamuro, 2008). It has been shown that histones neutralize bacterial endotoxins in human placenta, which furthermore supports the notion of disrupted immune function in individuals with schizophrenia, particularly early in development (Kim *et al.*, 2002). Finally, it is also known that there is a positive correlation between the incidence of autoimmune disorders and schizophrenia risk (Müller & Schwarz, 2010).

3.5.2.2 Other genes on chromosome 6

The HLA genes were not the only genes located in the MHC region that was predicted to be affected. The results indicated that *BTN3A2*, a gene located in the extended MHC region and belonging to the butyrophilin (BTN) family and immunoglobulin superfamily, could possibly be affected by 14 different regulatory polymorphisms. The protein product of this gene, along with the proteins from the rest of the BTN family, plays an important role in T-cell responses during an immune response. A study by Bamne *et al.* (2012) looked at immune factors specifically in African American individuals. A total of five SNPs in the MHC region were significantly associated with schizophrenia, of which two occurred within *BTN3A2* and one occurred in the 3'UTR of *BTN3A2*. These results support the predictions of RegulomeDB and also indicate a possible gene involved in schizophrenia susceptibility in African individuals.

3.5.2.3 Genes on chromosome 10 and 19

One of the most significant genes affected by 18 variants in regulatory regions was *USMG5* on chromosome 10 - a gene with a critical role in the maintenance of ATP synthase in mitochondria (Meyer *et al.*, 2007). Furthermore, *USMG5* is shown to have high levels of expression in the brain (Martens *et al.*, 2006), which was confirmed by analysing FANTOM data for cell lines related to the brain in this study. This gene showed some of the highest expression levels in the brain with an average of 223.82 TPM. Importantly, *USMG5* was recently associated in a large GWAS including European individuals (Ripke *et al.*, 2013). However, it was not listed as one of the main findings, but rather as one of the genes in the genomic region defined by LD surrounding a different lead SNP. This is a prime example of important findings that are overlooked in GWAS.

Additionally, *CNNM2* was predicted to be affected by 24 different SNPs according to rSNPBase (Table 3.5). Interestingly, even though this gene shows moderate expression in the brain (7.50 TPM) according to FANTOM data, this gene has been linked to schizophrenia susceptibility in a large GWAS of European individuals as well as a study including a Han Chinese population (Ripke *et al.*, 2011; Guan *et al.*, 2012). Furthermore, both these studies suggested an association between the 5'-nucleotidase, cytosolic II gene (*NT5C2*) and schizophrenia susceptibility as well. This gene was also among the top 10 affected genes (according to rSNPBase) and was predicted to be affected by 13 different SNPs (Table 3.5).

Chromosome 19 harbours three of the genes predicted to be affected, namely MAU2 Sister Chromatid Cohesion Factor (*KIAA0892/MAU2*), endothelial differentiation, lysophosphatidic acid G-protein-coupled receptor, 4 (*EDG4*) and ATPase type 13A1 (*ATP13A1*). *KIAA0892* is crucial for association of the cohesin complex with chromatin during interphase. This gene has an important role during sister chromatid cohesion and normal progression through prometaphase (Braunholz *et al.*, 2012). *EDG4* is a lysophosphatidic acid receptor gene which codes for an important extracellular signalling molecule that is involved in common cellular processes such as calcium mobilization, cell survival and mitogenesis (Contos *et al.*, 2002). Finally, *ATP13A1* is involved in nucleotide binding, cation transport and adenylypyrophosphatase (ATPase) activity (Weingarten *et al.*, 2012). A study aimed at analysing developmental expression of ATPase mRNA in mice showed that genes belonging to this family might be of significance in the nervous system. *ATP13A1* was found to have high expression levels during organogenesis, while an isoform of this gene was shown to be highly significant during neuronal development (Weingarten *et al.*, 2012). It is important to note that none of these genes on chromosome 19 have been associated with schizophrenia. Furthermore, these genes are also involved in basic, common processes in the body, which points toward fundamental mutations that occur early during development, which could confer an increased risk for schizophrenia.

Interestingly, even though there was not a consensus between the predictions of RegulomeDB and rSNPBase, the most frequently affected genes were located on the same three chromosomes (6, 10 and 19). Chromosome 6 has long been a chromosome of interest for schizophrenia association studies; however the latter two chromosomes have only recently gained more interest.

3.5.3 Pathway and network analysis

At first glance, the terms in Cluster 1 of the DAVID pathway analysis seemed to be unrelated to schizophrenia despite high p -values for most of the terms in the cluster. However, following a literature search of the top terms, links with schizophrenia became more apparent. For example, the first two terms namely “Allograft rejection” and “Graft-versus-host disease” is mainly due to an immune response by the body and has previously been linked to schizophrenia in a study that analysed variants that support the autoimmune hypothesis of schizophrenia (Carter, 2011). Importantly, most of the pathways enriched for the selected genes were identified in this study as well. Furthermore, processes such as “Allograft rejection” and “Graft-versus-host disease” support previous findings of increased episodes of psychosis following kidney and liver transplantations (Abbott *et al.*, 2003; Goralczyk *et al.*, 2010).

Even though a number of studies have focussed on the overlap between type II diabetes and schizophrenia, very few studies have focussed on type I diabetes (KEGG p -value = 1.22E-08). In a nationwide study conducted in Finland, it was shown that individuals with schizophrenia had a decreased incidence of type I diabetes, although the genetic mechanism of this phenomenon was not explored (Juvonen *et al.*, 2007). In this study however, the overlap between these two disorders appeared to be due to the HLA genes (*HLA-DQB1*, *HLA-DRB1*, *HLA-A*, *HLA-DRB5*, *HLA-C*, *HLA-DPB1*, *HLA-DQA1*, *HLA-G*, *HLA-DRA*), which were all involved in the “Type I diabetes mellitus” pathway.

Importantly, the pathway for cell adhesion molecules (CAMs) was enriched for 10 genes, with a p -value of 2.18E-04. Neural CAMs are critical for processes such as neurulation, axonal outgrowth and neuronal connectivity during neurodevelopment. Postmortem examination of schizophrenia cases and controls indicated that the expression of the embryonic form of neural CAMs was decreased in individuals with the disorder and was suggested to alter synaptic plasticity in schizophrenic brains (Barbeau *et al.*, 1995). This once again supports the hypothesis that schizophrenia might be a neurodevelopmental disorder.

Furthermore, even though viral myocarditis was indicated to be enriched for a subset of the query genes ($n = 9$; p -value = 1.38E-05), this pathway has not previously been associated with schizophrenia. However, myocarditis due to long-term clozapine treatment, which is an atypical drug used to treat schizophrenia, is known to occur (Lang *et al.*, 2008).

The findings for the involvement of natural killer (NK) cells in schizophrenia have been inconsistent. While there is strong evidence linking these cells with the disorder, the exact mechanism is still being debated. McDaniel *et al.* (1992) showed that there was no difference in NK cell activity between schizophrenia cases and controls. While Yovel *et al.* (2000) and Vodjgani *et al.* (2005) managed to show that there was an increase in NK cell activity among individuals with schizophrenia, DeLisi *et al.* (1983) and Abdeljaber *et al.* (1994) successfully proved that NK cells were suppressed in individuals with the disorder. Additionally, low NK cell activity has also been observed in autism, a disorder that significantly overlaps with schizophrenia (Vojdani *et al.*, 2008). Either way, NK cells are critical to the innate immune system and specifically respond to microbial infections (Farak & Caligiuri, 2006). Therefore, these findings still support the hypothesis that schizophrenia is a pathogenic autoimmune disorder.

Finally, network analysis indicated that two networks might be of importance, namely “ER to Golgi transport vesicle membrane” and “antigen processing and presentation of exogenous peptide antigen”. The first network has appeared during pathway analysis of schizophrenia in two previous studies, however no strong associations were established (Lee *et al.*, 2013; Purcell *et al.*, 2014). In contrast to this, numerous pathway studies showed an enrichment for antigen processing pathways presentation among schizophrenia associated genes (Fellerhof and Wank, 2009; Xu *et al.*, 2012; Aberg *et al.*, 2013; Luo *et al.*, 2013). In fact, in a study that used RNA-Seq analysis to link dysregulation of the immune system to schizophrenia, “antigen processing and presentation” was the most enriched pathway in 218 differentially expressed genes (Xu *et al.*, 2012). Once again, this ties previous findings as well as the findings of this study to an immune response conferring an increased risk for schizophrenia.

Analysis of co-expressed genes within GeneMANIA indicated that many of the HLA genes are expressed together. However, the ubiquitin C gene (*UBC*) was co-expressed with a number of HLA genes as well as the Solute Carrier Family 3 (Amino Acid Transporter Heavy Chain) Member 2 gene (*SLC3A2*) and the heat shock protein 90kDa alpha (cytosolic), class B member 1 gene (*HSP90AB1*) (Fig. S1). Furthermore, analysis of genetic interaction data indicated that *UBC* interacts with coagulation factor II (*F2*) as well as nuclear factor of activated T-cells, cytoplasmic, calcineurin-dependent 2 (*NFATC2*) (Fig. S1). The importance of *UBC* is that it is a critical “switchboard” gene, which has shown to be involved in a number of diseases. In a study by Lee *et al.* (2011), *UBC* was abnormally expressed in samples from individuals with schizophrenia and was shown to interact with marker genes of

schizophrenia, bipolar disorder and major depression. Variants affecting this gene could therefore explain the overlaps that are often seen in the genetics of these disorders. Finally, FANTOM data showed the highest expression levels for *UBC* compared to the other genes identified in this study, with an average expression of 811.89 TPM in the brain. These expression levels are extremely high and would warrant further attention in future studies.

Additionally, an HLA gene was once again the most significant, with *HLA-DRB5* shown to interact with the most genes. This gene was also part of the “antigen processing and presentation of exogenous peptide antigen” network. While there is no evidence supporting the role of *HLA-DRB5* in schizophrenia, this gene has been identified as a candidate susceptibility gene for Alzheimer’s disease (Lambert *et al.*, 2013). It is known that there are many significant genetic overlaps between schizophrenia and Alzheimer’s disease (Farooqui *et al.*, 2013), which might indicate that there are many common genes that are yet to be identified. While there may be not evidence linking this gene to schizophrenia, the “antigen processing and presentation of exogenous peptide antigen” network support the notion of immunity playing a role in schizophrenia. More importantly, this network has also been associated with schizophrenia in previous studies (Xu *et al.*, 2012; Aberg *et al.*, 2013; Luo *et al.*, 2013), suggesting a significant dysregulation in this network that could confer an increased risk for schizophrenia.

The hypothesis that schizophrenia might be an autoimmune disease was first conceived in the early 1960’s (Burch, 1964). Circumstantial evidence that point towards autoimmunity include the presence of an autoimmune disorder in many individuals with schizophrenia, association with the MHC region and improvement of psychotic symptoms with immunosuppression (Jones *et al.*, 2005). This hypothesis has been strongly criticized due to the widely held belief that autoantibodies would not be able to cross the blood-brain barrier to the nervous system (Jones *et al.*, 2005). However, it is possible that infectious agents can cross this barrier and invade the central nervous system directly (Benros *et al.*, 2012). Additionally, it is also possible that inflammation and infection can affect the brain through many different pathways and interactions (Benros *et al.*, 2012). Further support for this hypothesis is that it might explain the high rate of perinatal obstetric complications observed among individuals with the disorder (Kirch, 1993). However, the exact etiological mechanism through which schizophrenia and autoimmune responses overlap is still unknown. It is suggested that the HLA and histone protein genes identified in this study

might have a role in an autoimmune response. Though, it is unlikely that these genes act in isolation and the interactions would therefore need considerable attention.

The limitations of this study lie within the tools used. Even though bioinformatics provide endless possibilities for the discovery and analysis of disease-associated variants, the tools used for analysis of non-coding regions could be considered to still be in an early developmental phase, compared to most other tools. This means that many significant findings could possibly have been overlooked. Additionally, there is not always consensus among the results of different tools, making it difficult to distinguish between results that are not significant and results that were potentially missed, but are still significant. Furthermore, despite the power of bioinformatic tools, it is still not possible to predict the exact mechanisms by which a variant or group of variants exactly affect the regulation and expression of a gene. Even though it is possible to link variants to various regions, it would be crucial to understand the detailed mechanism by which gene expression is affected in order to understand the association with specific disorders.

Finally, one of the biggest hurdles specifically in the context of this study is the fact that most of these findings would be applicable to individuals of European or Asian descent, since the original GWAS findings used for the initial analyses were from studies that included individuals from these populations. It would therefore be essential to study these variants in an African population as well.

3.6 CONCLUSION

The significance of this study is two-fold. Firstly, functional evidence was provided for genomic regions that have previously been linked to schizophrenia by GWAS. This was achieved by analysing variants with regulatory potential and the effects that these variants might have. Many of the findings of previous GWAS have been linked to schizophrenia, despite functional evidence to support the findings. However, this study was able to show that in many cases there was evidence supporting the association of variants in LD with GWAS SNPs to schizophrenia susceptibility. Secondly, novel findings of this study include the identification of three genes on chromosome 19 (*KIAA0892*, *EDG4* and *ATP13A1*) that were affected by a number of SNPs in regulatory regions, as well as the regulatory variant rs200981 on chromosome 6 that might be implicated in schizophrenia susceptibility. Finally,

although the identification of the histone protein genes that could be affected in schizophrenia were not novel, these genes have not been studied extensively and would require further attention.

Throughout this study, evidence was gained for a multitude of regulatory factors that could be involved in schizophrenia susceptibility. However, the pieces of this intricate puzzle still need to be put together. This will only be possible with the continuous development of bioinformatics tools in conjunction with experimental studies. The findings of this study point towards mutations that occur during early stages of development in pathways related to fundamental processes in the human body. This was illustrated by variants shown to affect essential genes and regulatory elements, such as the histone protein genes in the extended MHC region, which are crucial for transcriptional regulation, DNA repair and factors such as *POLR2A* which is crucial for mRNA synthesis and finally *EP300*, which is essential for normal development before and after birth. Further evidence pointing towards aberrant regulation of higher level processes, was that the sTRAP results corresponded largely between schizophrenia as well as the control disorders/traits. However, there seemed to be a substantial amount of evidence linking immune function and neuronal development to schizophrenia. It is possible that both of these events are crucial: while mutations in neuronal development could confer an increased risk for schizophrenia, an aberrant immune response could lead to onset of this disorder. Therefore, it is suggested that these two processes should be studied together, rather than in isolation. Additionally, the possibility that DNA packaging might be implicated in schizophrenia susceptibility based on novel findings, would warrant further attention.

CHAPTER 4

INVESTIGATION OF REGULATORY POLYMORPHISMS IN SCHIZOPHRENIA SUSCEPTIBILITY IN A SOUTH AFRICAN XHOSA COHORT

4.1 ABSTRACT

Despite the value of this cohort, the South African Xhosa population has been significantly underrepresented in psychiatric genetic studies. This study aimed to serve as a proof of concept of the validity of the regulatory variants identified by means of bioinformatic analyses in schizophrenia susceptibility in a Xhosa cohort of 100 cases and 100 controls. Six SNPs with high regulatory potential as well as four corresponding SNPs from previous GWAS (which occurred in LD with the regulatory SNPs) were selected for genotyping. PCR-RFLP was used for genotyping, while bi-directional Sanger sequencing was used to confirm the functionality of the assay. Statistical analyses were performed with SNPStats and R statistical packages. Even though no single SNP was significantly associated with schizophrenia susceptibility, two haplotypes of regulatory variants (rs200483-rs200485-rs2517611; $p = 0.0385$ and rs200981-rs2517611-rs3129701; $p = 0.041$) were significantly associated with risk for the disorder in this cohort. These haplotypes were predicted to affect DNA packaging. A further two haplotypes were associated with positive symptoms, while two regulatory SNPs (rs2535629; $p = 0.0325$ and rs2517611; $p = 0.0456$) were significantly associated with negative symptoms. It is hoped that these results as well as the findings of future studies could act in conjunction with GWAS to uncover the mechanisms by which common variants in regulatory regions contribute to schizophrenia susceptibility.

4.2 INTRODUCTION

Despite the decades of research and different approaches that have been used, the aetiology of schizophrenia still eludes researchers. In a series of articles published in *Nature* in April 2014 (Volume 508, Issue 7494), it is clear that despite the multitude of studies and significant findings thus far, schizophrenia remains a complex and devastating disorder. Since 2011, more than 100 genome-wide significant hits associated with almost 700 genes have been identified (Wright, 2014). However, with each association, more unresolved questions

emerge, making research efforts seem like an infinite loop. The most recent and largest schizophrenia GWAS to date was recently published by the PGC and provided irrefutable evidence for the genetic contribution to schizophrenia. A staggering 108 loci were associated with the disorder, of which 83 were novel (Ripke *et al.*, 2014). Two of the most significant findings of this study were firstly, the association of *DRD2* which codes for the dopamine receptor DRD2 and has therapeutic relevance and secondly, the association of the MHC region (as well as genes outside of this region, but with an immune function) that support the hypothesis of aberrant immune responses in schizophrenia (Ripke *et al.*, 2014). The findings of this study are an indication of the sheer size of psychiatric genetics and the progress being made.

Initially, the promise of GWAS seemed endless. The first GWAS, performed in 2005, screened for 116 204 SNP in 96 cases and 50 controls in a study of age-related macular degeneration (Klein *et al.*, 2005). Since then, GWAS have increased substantially in power and size. Between 2005 and 2012, more than 1000 GWAS, replication studies and meta-analyses were published, with the promise of personalised therapy in the near future. However, in the last few years, it has become clear that GWAS is not necessarily the best approach for genetic studies of complex disorders. In a review by Klein *et al.* (2012), three important factors of GWAS were assessed, namely sample size, sample characterization and effect size. While it is generally thought that bigger sample sizes lead to more accurate findings, large GWAS tend to be successful at identifying many variants with small effect sizes, with the majority of odds ratio values around 1. Finally, regarding sample characterization, it is possible that individuals with symptoms of one disorder could wrongfully be included in a study that analyses a completely different disorder. Another flaw of GWAS is that due to the extensive corrections and stringent thresholds for significance, many variants with small effects can easily be missed (Sadec *et al.*, 2014). When studying complex psychiatric illnesses, the question arises whether studying something as simple as SNPs could provide clues to the intricate nature of the disorder. In a study that tested the polygenic model of schizophrenia, the conclusion was that approximately one third of the heritability observed in the disorder can be attributed to the cumulative effect of hundreds or thousands of common SNPs (Purcell *et al.*, 2009). While the significance of SNPs is undoubted, it is more accurate to assume that many of these nucleotides along with copy number variations in a specific pathway(s) would interact to cause a disorder such as schizophrenia. Furthermore, these causal networks would have to be studied in multiple

populations in order to clarify what leads to susceptibility of a disorder worldwide. However, the point is not to discredit the validity of GWAS, but rather to focus on the additional studies that need to be done, such as identifying rare variants with large effect sizes, variants in regulatory regions and performing pathway and network analyses. Finally, GWAS tend to include individuals of European or Asian descent and no GWAS regarding schizophrenia has included African individuals to date (with the exception of African American individuals). This leaves a considerable gap in understanding the disorder as well as effective treatment for all ethnicities.

One such group that has been overlooked in genetic studies is the South African Xhosa population. Xhosa individuals form the second largest indigenous African group in South Africa (Statistics South Africa, 2013). This group is considered to be culturally and genetically homogenous, due to geographical influences and therefore provide valuable data for genetic research (Le Roux *et al.*, 2007). Additionally, it would also be extremely important to study genetic variation among the other populations within Africa. As indicated by May *et al.* (2013), there are distinct genetic differences between the southern African Bantu speakers and the Yoruban and Luhya populations from western and central Africa respectively, indicating that the inclusion of only one or a few African populations would not be sufficient to capture the entirety of genetic variation among all the various African genomes.

Therefore, the aim was to perform a pilot study using a small cohort of South African Xhosa cases and controls to address two issues of current research: firstly to genotype selected variants in regulatory regions of the genome instead of variants in protein-coding genes and secondly, to use a cohort of South African Xhosa individuals to fill the void regarding the inclusion of these individuals in genetics studies. Finally, genotyping the variants identified by bioinformatics analyses in Chapter 3 was particularly important since it is usually recommended that bioinformatics results be validated with experimental work (McInerney, 2002; Fernandes *et al.*, 2012).

4.3 MATERIALS AND METHODS

4.3.1 Patient Samples

A subset of 100 Xhosa schizophrenia patients, which form part of a larger cohort of South African Xhosa patients recruited from various in- and outpatient hospital services and clinics in the Western Cape, was used. The two selection criteria that were used to choose individuals included (1) diagnosis with schizophrenia according to the DSM-IV and (2) all four grandparents of each individual had to be of Xhosa ancestry. Patients were also assessed with comprehensive tools, including the Diagnostic Interview for Genetic Studies version 2 and the SANS and SAPS. The age of the schizophrenia cases ranged between 13 and 51 years of age, with the average age being 24. Regarding gender, 77% of the cases were male and 23% female. Due to mutual cultural and environmental backgrounds, this group is clinically homogeneous, and therefore provides valuable data for the improved understanding of genetic factors in African schizophrenia patients. A subset of 100 unrelated healthy Xhosa controls, which are also part of a larger group, was available and were matched for age and gender. Genomic DNA (gDNA), extracted from whole blood using the Miller *et al.* (1988) method, was available for each patient and control individual. This project forms part of a larger long-term study, *Genetics of Schizophrenia*, for which ethical approval was obtained from the Stellenbosch University Human Research Committee (Reference number: 1907/005- and 2001/050). Informed written consent was obtained from all individuals or, where necessary, the primary caregivers.

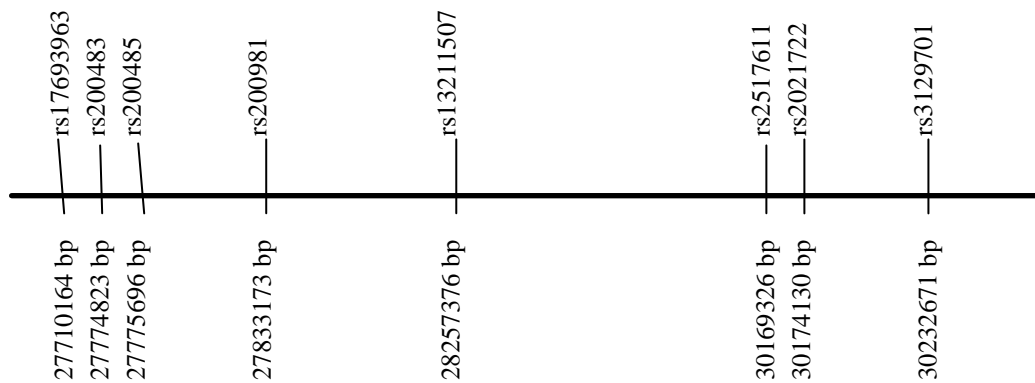
4.3.2 Genotyping

4.3.2.1 SNP Selection

Based on the bioinformatics analyses in Chapter 3, ten SNPs were selected for genotyping in the South African Xhosa cohort to determine whether any of the regulatory SNPs or corresponding SNPs from previous GWAS are associated with schizophrenia susceptibility in this cohort. Six SNPs, based on high regulatory potential (indicated by RegulomeDB scores) and the four corresponding SNPs from the original GWAS that occurred in LD with these six SNPs were selected for genotyping (Table 4.1). The majority of the SNPs are on chromosome 6, with the locations indicated in Fig. 4.1.

Table 4.1. SNPs selected for genotyping. The regulatory SNPs correspond to and occur in LD with the original SNPs from GWAS.

GWAS SNP	Chromosome	Regulatory SNP	RegulomeDB Score
rs17693963	6	rs200981	1a
rs13211507		rs200485	1b
		rs200483	1d
rs2021722	6	rs3129701	1f
		rs2517611	1f
rs4687552	3	rs2535629	1b

**Figure 4.1.** The locations of the selected SNPs on chromosome 6. The figure is not drawn to scale.

4.3.2.2 Primer Design

Primers had to be designed for all but one SNP, rs2021722, which had been designed prior to this study (Coffee, 2012). The primers used for amplification of the regions containing the SNPs in regulatory regions as well as the corresponding SNPs from previous GWAS (in the event that a regulatory variant is a SNP in LD with a SNP from a GWAS) are summarized in Table 4.2. All the primers were designed using PrimerQuest and analyzed using OligoAnalyzer (Integrated DNA Technologies Inc., 2012). Subsequently, the primers were synthesized by Integrated DNA Technologies (Integrated DNA Technologies, Inc., Iowa, USA).

4.3.2.3 PCR Optimization

All reactions, except for rs2021722, had to be optimized. For optimal amplification by PCR, the reaction cycles and conditions were optimized using the Veriti® 96-Well Thermal Cycler

(Applied Biosystems[®], California, USA) to determine optimal temperature by performing a gradient PCR and the GeneAmp[®] PCR Systems 2700 (Applied Biosystems[™], California, USA) for further optimization and final PCR reactions. Reagents were supplied by Bioline (Bioline[™], London, UK) and Kapa Biosystems (Cape Town, South Africa). All PCR reactions were performed using 30 ng DNA, 1X reaction buffer, 1.5 mM MgCl₂, 0.4 μM dNTPs and 0.5U BIOTAQ Polymerase, with a total reaction volume of 25 μl. For all reactions, 0.4 μM forward primer and 0.4 μM reverse primer were used, with the exception of the amplification of the amplicons for rs17693963 and rs2021722 which used 0.2 μM of each primer. In order to amplify the region for SNP rs200981, 4 μl Betaine (Sigma-Aldrich (Pty) Ltd, Aston Manor, South Africa) was required in each reaction. Due to non-specific amplification, KAPATaq Ready Mix (Kapa Biosystems, Cape Town, South Africa) was used for SNP rs200483. For this analysis, the amount of reagents were used according to the Standard PCR Protocol for 20 μl reactions (10 μl ReadyMix with Mg²⁺, 0.4 μM forward primer, 0.4 μM reverse primer, 30 ng DNA template and 7.4 μl water).

The reaction cycles for all amplifications consisted of an initial denaturation at 94°C for 3 minutes; followed by varying 40 cycles (except for rs200483 which used 35 cycles) of denaturation at 94°C for 15 seconds, annealing at the respective melting temperatures (T_m , indicated in Table 4.2) for 15 seconds and extension at 72°C for 30 seconds; concluding with a final extension step at 72°C for 5 minutes. Thereafter, each amplicon (mixed with 5 μl Cresol loading dye) was loaded onto a 1.5% (w/v) agarose gel stained with 0.5 μg/ml ethidium bromide to facilitate visualisation. Electrophoresis was performed in 1X sodium borate (SB) buffer at 200 V for 25 minutes. Additionally, a 100 bp molecular weight marker (Hyperladder IV, Bioline[™], London, UK) was loaded to determine the sizes of the amplified fragments. Finally, the fragments were visualized under an ultra-violet light with the Bio-Rad Molecular Imager[®] Gel Doc[™] XR+ System with Image Lab[™] v5.1 Software (Bio-Rad Laboratories, Inc., California).

Table 4.2. The forward (F) and reverse (R) primers and reaction conditions that were used to amplify each SNP.

SNP rs ID	Amplicon Size (bp)	T _m (°C)	Primer Sequence (5'-3')
rs17693963	630	65	F: CTGGCTATCAAGAGCGAAACT R: TCACGTTGGTCTGAGGAAAC
rs200483	498	58	F: GCGATCTCAGCTCACTGTAAA R: GAGCATCCACGGGTTTCTAA
rs200485	358	60	F: CGCTTCTTGCCATCCTTCTT R: AGCTGTGCGATTGGCTTAC
rs200981	327	65	F: GGGATAATGCGGGTCTTCTT R: CGCGCCCAGTATTGACTATAA
rs13211507	303	65	F: CCACCACCAGACACATCTTAC R: CTGGTCTCAAACCTCCTTGTCTC
rs2517611	379	60	F: CGGGTTGAGGTGATTGATTCT R: CCAACTGGAAGCACTAATGGA
rs2021722	678	62	F: TACTCCATTTAAGGGCTCCTGGGT R: GGAGATCCTAAGTCAGTTGGTCTAGGT
rs3129701	676	60	F: GGGTGTGAGCAGTGGAAATA R: GGAATCCTGGTCTGTGTTCTT
rs2535629	484	58	F: GGTGCTGATTACCTGCTCTAA R: TGTACAGACTCCCTCCTCAA
rs4687552	791	62	F: TGAGGACAACGAGGATGAGA R: TGGCTGGCTAAAGGCTTAC

4.3.2.4 PCR-RFLP Genotyping Assays

All SNPs were genotyped by means of PCR-RFLP, using the restriction enzymes identified by RestrictionMapper v.3 (RestrictionMapper, 2009) and “*In silico* Simulation of Molecular Biology Experiments” (Bikandi, 2004) (Table 4.3). In all instances, restriction enzyme digests were performed using the amounts of PCR product, restriction enzyme and where necessary BSA, as indicated in Table S5. The respective incubation times and temperatures for each SNP are indicated in Table 4.3. Once the restriction enzyme digests were completed, the resulting fragments were separated by electrophoresis on the appropriate gel (Tables 4.3, S5).

The resulting digest product did not need to be mixed with loading dye in the case of FastDigest enzymes, since the FastDigest Green Buffer was used, which allows direct loading

of the product onto gels. However, in all other instances, 10 µl cresol loading dye was added to 20 µl of digest product and loaded onto the gel. The gels were stained with 0.5 µg/ml ethidium bromide and electrophoresis was performed in a 1X SB buffer at 160 V for the appropriate time (Table 4.3). A 100 bp molecular weight marker (Hyperladder IV, Bioline™, London, UK) was also loaded onto each gel to determine the sizes of the amplified fragments. In order to visualize the fragments, the Bio-Rad Molecular Imager® Gel Doc™ XR+ System with Image Lab™ v5.1 Software (Bio-Rad Laboratories, Inc., California) was used.

Table 4.3. The restriction enzymes and conditions used to detect each of the SNPs.

SNP	Chr	Restriction Enzyme	Temperature / Incubation time	SNP Alleles ^a	Fragment sizes (bp)	Gel % (w/v) / Electrophoresis run time
rs17693963	6	<i>Hpy188I</i>	37°C 16 h	A C	374, 245, 11 266, 245, 108, 11	3% agarose 1 h 30 min
rs200483	6	<i>NlaIII</i>	37°C 30 min	G A	218, 159, 121 280, 218	3% agarose 40 min
rs200485	6	<i>BseNI</i>	65°C 30 min	G C	215, 86 109, 106, 86	3% agarose 40 min
rs200981	6	<i>StyI</i>	37°C 30 min	T C	327 200, 127	3% agarose 40 min
rs13211507	6	<i>BccI</i>	37°C 16 h	T C	303 205, 98	3% agarose 40 min
rs2517611	6	<i>Mva1269I</i>	37°C 30 min	T C	219, 160 379	3% agarose 40 min
rs2021722	6	<i>AluI</i>	37°C 30 min	C T	386, 170, 122 277, 170, 122, 109	3% agarose 40 min
rs3129701	6	<i>AjiI</i>	37°C 12 h	C T	453, 196, 27 649, 27	3% agarose 40 min
rs2535629	3	<i>PstI</i>	37°C 30 min	T C	352, 132 299, 132, 53	3% agarose 40 min
rs4687552	3	<i>HinfI</i>	37°C 30 min	T C	421, 370 791	3% agarose 40 min

^aThe reference allele is listed first.

4.3.2.5 Bi-directional Sanger sequencing

To ensure the functionality of the genotyping assay, bi-directional Sanger sequencing was performed to confirm sample genotypes. PCR products of two samples (one homozygote and one heterozygote) for each SNP were purified using the SureClean protocol (BiolineTM, London, UK; Appendix 5). Once the samples were cleaned, the concentrations of the purified product were measured using the Nanodrop spectrophotometer (NanoDrop[®] ND-100, Nanodrop Technologies Inc., Wilmington, Delaware, USA) at a wavelength of 260 nm. Thereafter, the sequencing reaction was performed as described in the Big Dye[®] Terminator v3.1 Cycle Sequencing Kit Manual (Applied BiosystemsTM, California, USA; Appendix 5). The samples were then analysed using capillary electrophoresis at the Central Analytical Facility, Stellenbosch University on a 3130XI Genetic Analyzer (Applied BiosystemsTM, California, USA). Once bi-directional sequencing was completed, the presence of the SNP was confirmed using Finch TV[®] v1.4.0 (Geospiza, 2006).

4.3.3 Statistical & Bioinformatic Analyses

The Power Calculator tool of An Online Sample Size Estimator (OSSE[®], 2006; <http://osse.bii.a-star.edu.sg/>; Accessed September 2014) was used to determine the power of this study, given the sample size and allele frequencies. In order to calculate the range of power, calculations were done with the two SNPs with the lowest and highest minor allele frequencies (MAFs), namely rs13211507 and rs2021722 respectively.

Due to known genetic differences between various populations, the MAFs of the ten selected SNPs were compared to the allele frequencies observed in the CEU, YRI and CHB populations, using 1000 Genomes and HapMap data.

For each genotyped SNP, descriptive statistics and Hardy-Weinberg equilibrium (HWE) analysis were done with SNPStats (Solé *et al.*, 2006; http://bioinfo.iconcologia.net/SNPstats_web; Accessed September 2014). This online tool allows for a convenient and accurate overview of all clinical and genotype data. It implements R packages for statistical analyses (Solé *et al.*, 2006). The genotype and allele frequencies were determined in the Xhosa schizophrenia cases and controls. Genotypes were tested for deviations from HWE and the *p*-value cut off was 0.01. Remaining statistical analyses were performed in R (R Core Team, 2014), using the R packages *genetics* (Warnes

et al., 2013) and *haplo.stats* (Sinnwell & Schaid, 2013), to determine whether there was an association between any of the genotyped SNPs or SNP haplotypes and schizophrenia susceptibility. Analyses were also done to determine possible associations between the genotyped SNPs / SNP haplotypes and negative or positive symptoms, using SANS and SAPS as well as the subgroups of these symptoms. Logistic regression was used to model schizophrenia susceptibility, which is a dichotomous outcome, while general linear models were used for numerical data. Confidence intervals, effects sizes and *p*-values were derived from the aforementioned models and associations with *p*-values < 0.05 were accepted as significant. Allelic combinations and frequencies were modelled with functions from *haplo.stats*. Modelling was used in order to avoid false positives and adjust for known or suspected confounding factors such as age and gender.

Finally, all variants significantly associated with schizophrenia susceptibility were analysed with GeneMANIA to identify affected biological features, by performing network analysis for the predicted affected genes (<http://www.genemania.org/>; Accessed September 2014). The analysis was done similar to the description in section 3.3.9. All features with a FDR < 10^{-5} were accepted as significant. This was only done for variants significantly associated with schizophrenia susceptibility, since this was the main focus of this project.

4.4 RESULTS

Statistical power of the study using OSSE, revealed power to detect significance was between 16.8% and 20.7%, depending on the frequency of the SNP that was being investigated. These results were based on SNPs rs13211507 and rs2021722 which had the lowest and highest minor allele frequencies respectively and therefore gave an indication of the lowest and highest power to detect significance.

The ten selected SNPs were successfully genotyped in the Xhosa cohort and bi-directional Sanger sequencing confirmed that the genotyping assay worked (Fig. S4).

Comparisons of the MAFs of each SNP between the Xhosa schizophrenia cases and controls are illustrated in Fig. 4.2. Additionally, comparisons of MAFs of these SNPs in different populations, using the 1000 Genomes and HapMap data are depicted in Fig. S3. Comparisons with these population data indicated distinct differences between the MAFs of the selected SNPs in the Xhosa cohort compared to MAFs in the CEU, CHB and YRI

populations. The SNPs identified in previous GWAS, namely rs17693963, rs13211507 and rs4687552, were lower in the YRI and Xhosa populations than in the CEU population.

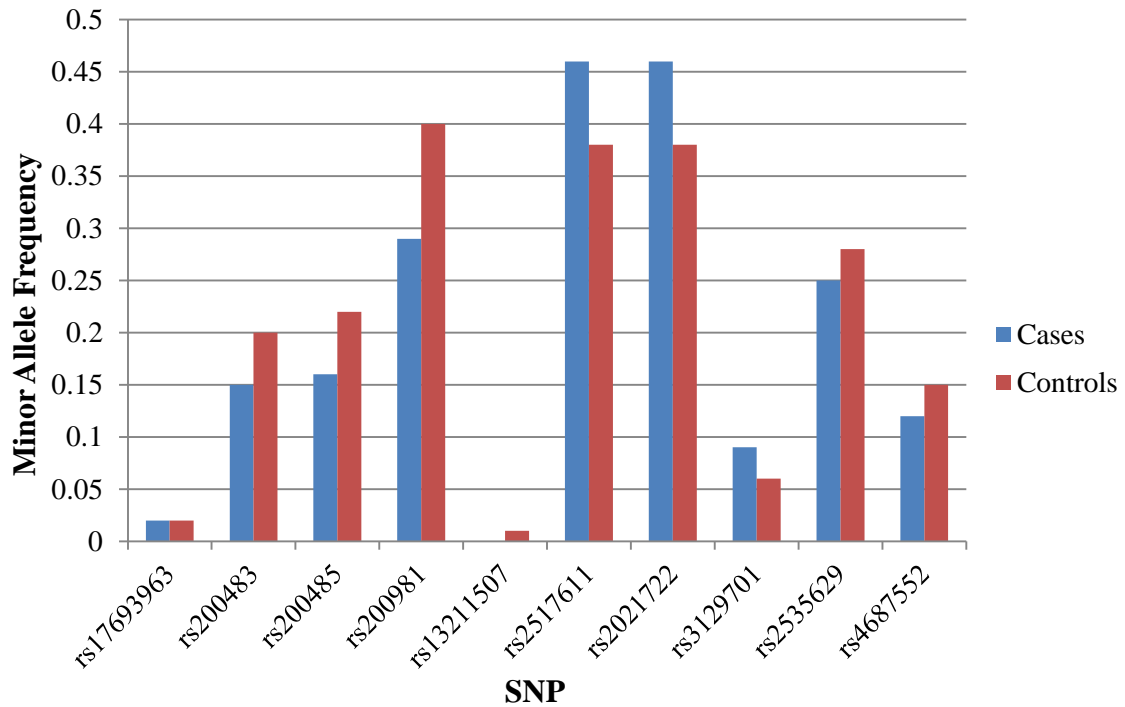


Figure 4.2. Minor allele frequencies of the genotyped SNPs in the South African Xhosa cases and controls. The significant haplotypes associated with schizophrenia susceptibility are rs200483-rs200485-rs2517611 and rs200981-rs2517611-rs3129701. All the SNPs occur on chromosome 6, with the exception of the last two SNPs (rs2535629 and rs4687552) which are located on chromosome 3.

Analysis of HWE revealed that all SNPs were in HWE. When adjusted for age and gender, single SNP analysis indicated that no SNPs were significantly associated with schizophrenia risk. However, two haplotypes were significantly associated with susceptibility for the disorder (Table 4.4). Significant results of association analysis with negative symptoms are summarised in Table 4.5 and indicates that two single SNPs were associated with these symptoms. In contrast to this, no single SNPs or SNP haplotypes were significantly associated with positive symptoms. However, a significant single SNP association with the formal thought disorder subgroup of positive symptoms was identified (Table S4). No significant associations were identified with the subgroups of negative symptoms.

Table 4.4. Results of association analysis with schizophrenia susceptibility.

Outcome	SNPs / specific allele			<i>p</i> -value	Frequency	OR	95% CI	
Schizophrenia Susceptibility	rs200483 G	rs200485 G	rs2517611 C	0.0385	0.36	1.71	1.01	2.91
Schizophrenia Susceptibility	rs200981 C	rs2517611 T	rs3129701 C	0.041	0.22	0.51	0.27	0.98

Table 4.5. Results of association analysis with negative symptoms.

Outcome	SNP	Inheritance Model	Comparison	<i>p</i> -value	Effect	95% CI	
SANS	rs2535629	Dominant	TT versus CC+CT	0.0325	-0.456	-0.805	-0.106
SANS	rs2517611	Heterozygous	CT versus CC+TT	0.0456	-0.453	-0.808	-0.098

Finally, analysis of the genes affected by the two significant SNP haplotypes (rs200483-rs200485-rs2517611 and rs200981-rs2517611-rs3129701) associated with schizophrenia susceptibility using GeneMania, implicated significant processes involved in DNA packaging, from DNA level to chromatin level, in schizophrenia susceptibility (Table 4.6). The most significantly affected feature was “nucleosome” for both of the haplotypes, with FDRs of 1.69×10^{-15} and 1.25×10^{-13} .

Table 4.6. Features affected by haplotypes associated with schizophrenia susceptibility.

Feature	FDR	
	rs200483-rs200485-rs2517611	rs200981-rs2517611-rs3129701
nucleosome	1.69E-15	1.25E-13
DNA bending complex	1.69E-15	1.25E-13
DNA packaging complex	3.75E-14	1.67E-12
nucleosome assembly	4.45E-11	7.70E-10
chromatin assembly	3.75E-10	4.90E-09
nucleosome organization	1.32E-09	1.43E-08
protein-DNA complex assembly	1.32E-09	1.43E-08
chromatin assembly or disassembly	4.95E-09	4.53E-08
DNA packaging	7.57E-09	6.51E-08
protein-DNA complex subunit organization	1.25E-08	1.00E-07
DNA conformation change	3.06E-07	1.69E-06
chromatin	7.55E-07	2.43E-06
protein-DNA complex	1.05E-06	4.99E-06

4.5 DISCUSSION

The goal of this study was to genotype ten of the most significant variants (including the corresponding SNPs from previous GWAS) identified in the previous chapter, in a South African Xhosa cohort of 100 schizophrenia cases and 100 controls as a pilot study to validate the bioinformatics approach used. As mentioned previously, there is currently a large void with regards to genetic studies that include African individuals. The largest psychiatric GWAS to date, which was performed by the PGC, has provided valuable clues to the genetic architecture of schizophrenia, but predominantly included individuals of northern European descent (Ripke *et al.*, 2014). Yet, in order to improve and optimize treatments for individuals with schizophrenia worldwide, all populations would have to be represented in genetic studies – a major flaw that was recognised by the PGC itself (Reardon, 2014).

As previously discussed, there are significant genetic differences between different populations. Therefore, the MAFs of the genotyped SNPs were compared to the MAFs observed in the YRI, CEU and CHB populations according to 1000 Genomes and HapMap data. Fig. S3 clearly indicates the differences observed between the African population and European and Asian individuals. In most instances the allele frequencies observed in the African and/or Xhosa population differed significantly from the frequencies observed in the European and Asian populations. Regarding SNPs rs17693963, rs13211507, rs2535629, rs4687552, rs3129701 the MAFs were much lower in the CEU and/or the CHB populations. It is well known that African genomes exhibit greater genetic diversity, less linkage disequilibrium between loci and extensive population substructure (Campbell & Tishkoff, 2008). Moreover, Africa has a complex population history due to the drastic variation in environmental factors and the adaptation of inhabitants to these factors. This provides valuable information for genetic studies of disease susceptibility. Furthermore, the Out of Africa model suggests that modern humans originated in Africa before migrating to other parts of the world (Tishkoff & Verrelli, 2003). Nonetheless, to date only a small fraction of the ~2000 different ethno-linguistic groups in Africa has been included in GWAS. In a study by Yu *et al.* (2002), the patterns of SNPs particularly in non-coding regions were analysed. Furthermore, the differences between patterns within Africans and between Africans and Eurasians were compared. The average nucleotide diversity among African individuals was almost twice as high as what was observed among Europeans and Asians. This clearly

indicates that Africans differ much more among each other than other populations and highlights the need for genetic studies in these groups.

Once all of the variants were genotyped, analysis for HWE revealed that all SNPs were in HWE. Single SNP analysis indicated that no SNPs were significantly associated with risk for the disorder, yet haplotype analysis identified two combinations of SNPs that were indeed associated with schizophrenia susceptibility (Table 4.4). All of the SNPs involved in the two different haplotype combinations were regulatory SNPs (as determined by bioinformatic analyses in Chapter 3) and one of the combinations (rs200981-rs2517611-rs3129701; $p = 0.041$; OR = 0.51; 95% CI = 0.27-0.98) contained the most significant SNP identified in the previous chapter, which was rs200981. Importantly, none of these SNPs have been associated with schizophrenia susceptibility except for rs200483 and rs200981 (Nielsen *et al.*, 2014). However, both these SNPs were only associated very recently (August 2014) in one other study and rs200981 was also identified as the most significantly associated variant (Nielsen *et al.*, 2014). Network analysis indicated that both haplotypes associated with schizophrenia susceptibility affected genes involved in functions such as “nucleosome”, “DNA bending complex”, “DNA packaging complex”, “nucleosome assembly”, “chromatin assembly” and “nucleosome organization” amongst others (Table 4.6). These findings are in accordance with the findings of Luo *et al.* (2013), where genes involved in nucleosome assembly were significantly associated with schizophrenia susceptibility. By modifying chromatin packaging, accessibility to gene promoters can be affected, which in turn can have extensive implications for gene regulation (Sharma *et al.*, 2005). Furthermore, chromatin remodelling is known to be particularly important during neural development and plasticity, which highlights the severity of abnormalities during this process (Hsieh & Gage, 2005; Lessard *et al.*, 2007). In fact, alterations in chromatin packaging could affect gene expression at every stage of neurogenesis (Hsieh & Gage, 2005). This emphasizes the findings of the previous chapter, which pointed towards aberrant DNA packaging as well as affected histone protein genes, which might confer an increased risk for schizophrenia. Furthermore, these findings also highlight the importance of focussing on gene regulation when studying complex disorders.

Additional statistical analyses were aimed at identifying associations between any of the genotyped SNPs and negative or positive symptoms of schizophrenia based on SANS or SAPS. This was particularly important considering the complications in schizophrenia diagnosis. One of the many limitations of schizophrenia studies is that the disorder is usually

studied as a whole instead of focussing on the various endophenotypes (Greenwood *et al.*, 2013). In fact, a recent study showed that schizophrenia is more likely a heterogeneous group of eight different inheritable disorders with distinct symptoms (Arnedo *et al.*, 2014). While this study was not equipped to extensively study these endophenotypes, information regarding positive and negative symptoms was available for the Xhosa cases. The positive symptoms which are measured by SAPS include delusions, hallucinations, formal thought disorder and bizarre behaviour, while negative symptoms measured by SANS include affective blunting, avolition/apathy, anhedonia/asociality, alogia and lack of attention (Andreasen, 1990). Two SNPs, rs2535629 ($p = 0.0325$) and rs2517611 ($p = 0.0456$) were associated with negative symptoms based on a dominant and heterozygous inheritance model respectively (Table 4.5). Once again, both these SNPs were regulatory variants. The importance of these findings is that negative symptoms are particularly difficult to treat. Therefore, by understanding which genetic variants (particularly regulatory variants) could be associated with these symptoms, this could aid in developing future research strategies to optimize treatments for these symptoms. In contrast to this, there were no significant single SNP or SNP haplotype associations with overall positive symptoms of the disorder.

In addition to associations with positive and negative symptoms, the subgroups were also used for analysis. Interestingly, the top SNP (rs200981) was associated with a subgroup of positive symptoms, formal thought disorder ($p = 0.0447$; effect size = -0.59; 95% CI = -1.13-(-0.05)), according to a heterozygous inheritance model (Table S4). An association with formal thought disorder is particularly significant since this is one of central symptoms of schizophrenia (Wang *et al.*, 2012). Despite the prevalence of this symptom and the fact that twin and adoption studies have suggested that genetic factors might be involved, not many studies have focussed on the genetics of formal thought disorder (Levy *et al.*, 2010). To date, only one GWAS meta-analysis of this symptom has been performed with data from European-American individuals. The study identified 61 SNPs associated with formal thought disorder with $p < 10^{-4}$ (Wang *et al.*, 2012) and supports the notion of genetic factors contributing to this symptom.

None of the original SNPs identified by previous GWAS were significantly associated with schizophrenia susceptibility based on single SNP or haplotype analyses. This once again emphasizes the importance of including different populations in genetic studies. The original SNPs from GWAS, namely rs17693963, rs13211507, rs4687552, rs2021722 were identified in European individuals. In all instances, the MAFs of these SNPs were lower in the YRI and

Xhosa population compared to the CEU population, except for rs2021722 which had a much higher MAF in the YRI population (Fig. S3). This clearly shows that variants identified in these populations are not necessarily of importance for individuals of other ethnicities. This also applies to the regulatory variants for which the MAFs differed significantly between the different populations and it is suggested that these SNPs be genotyped in additional populations as well.

The lack of association between individual genotyped SNPs and schizophrenia susceptibility could be due to a number of reasons. Firstly, it is possible that associations could not be detected due to the small size of the Xhosa cohort. Secondly, as previously discussed, many of the MAFs (particularly of the SNPs identified in previous GWAS) were extremely low (Fig. 4.2). Finally, considering the fact that the original SNPs identified in GWAS and the regulatory SNPs in LD with these SNPs were identified in European individuals, these SNPs would not necessarily be significant in an African population.

Limitations of this study include the small sample size, validity of the diagnosed cases and controls as well as the fundamental understanding (and therefore the diagnosis) of schizophrenia. Firstly, due to the small cohort size, this had implications for the power to detect significant associations. It is therefore possible that significant associations with the genotyped SNPs could have been missed due to the lack of power of this study. Nonetheless, it is possible to detect significant associations in small cohorts, as demonstrated by Shimada *et al.* (2000) in a study of acute myocardial infarction (AMI) in a cohort of 164 Japanese individuals (81 cases and 83 controls). Despite the small cohort, an association was still identified between a SNP, C(-260) > T, in the promoter region of CD14 and AMI ($p = 0.004$; OR = 3.8). This shows that smaller cohorts can successfully be used in association studies. Secondly, due to the current uncertainty surrounding the definition of the disorder as well as major overlaps in symptoms of other psychiatric illnesses, this has significant implications for diagnosis. It is therefore possible, that a diagnosed individual might in fact suffer from a different, yet similar disorder. Furthermore, individuals with psychosis are often not able to provide healthcare officials with a complete/accurate medical history (Duwe & Turetsky, 2002). There are numerous cases of patients being wrongfully diagnosed due to overlaps between symptoms of schizophrenia and other mental disorders. For instance, psychotic bipolar disorder is often misdiagnosed as schizophrenia (Gonzalez-Pinto *et al.*, 1998; Duwe & Turetsky, 2002). This is even more common in individuals of African descent. Importantly, a study by Strakowski *et al.* (1996) determined that psychotic symptoms

reflective of schizophrenia were more frequently misdiagnosed in African American individuals than compared to Caucasians. Finally, it is also a possibility that control subjects could develop the disorder at a later stage. Still, considering that the prevalence of schizophrenia is approximately 1% in the general population, the chances of having such a control subject in the cohort would be rare. Finally, due to the heterogeneous nature of schizophrenia, it would be more effective to study the genetic factors associated with the subtypes of schizophrenia, rather than looking at the disorder as a whole. However, this seems to be a difficult situation: by understanding the genetics of the disorder, accurate diagnosis could be facilitated, but in order to study the genetics of schizophrenia subtypes, individuals would have to be carefully and accurately diagnosed in the first place.

4.6 CONCLUSION

This study successfully served as a pilot study to demonstrate the significance of regulatory variants in schizophrenia susceptibility and symptoms. Furthermore, the findings validated the bioinformatic results of the previous chapter. Statistical analyses significantly associated two haplotypes with schizophrenia susceptibility in a South African Xhosa cohort. Additionally, a number of SNPs and SNP haplotypes were significantly associated with the positive and negative symptoms of schizophrenia. None of the genotyped variants that were identified in previous GWAS were significantly associated with schizophrenia susceptibility in this cohort, which proved (1) significant variants associated with a disorder in one population cannot necessarily be extrapolated to other populations and (2) in many instances the reported associated variant is in fact in LD with the actual variant of importance. This study also highlighted the importance of verifying bioinformatic findings with experimental work, since not all variants identified in the previous chapter were significant in the Xhosa cohort. However, it is possible that the identified regulatory variants could be significant in a larger Xhosa cohort or in a different population group. Therefore, future work would require genotyping regulatory polymorphisms in a larger Xhosa cohort and ideally also other population groups. Due to the positive results of this study, custom TaqMan® OpenArray® Genotyping Plates have been designed to genotype additional regulatory SNP (as identified in Chapter 3) in a larger cohort of Xhosa cases and controls.

CHAPTER 5

CONCLUSION AND FUTURE DIRECTIONS

The aim of this study was to establish and apply a bioinformatics pipeline for the identification of variants in regulatory regions and to validate these findings in a pilot association study. The goal was successfully achieved by using two approaches. The first part of the study included bioinformatic analyses and started by identifying significant SNPs from previous GWAS as well as SNPs that occur in LD with these variants. The combined total of variants were annotated and analysed in terms of regulatory potential, using ENCODE data. Variants were characterised using a series of predictive programs such as RegulomeDB, GWAS3D and TRAP, while affected genes were predicted using RegulomeDB and rSNPBase. The expression of potentially affected genes in cell lines related to the brain was evaluated using FANTOM expression data. Finally, pathway and network analyses were performed to provide a comprehensive overview of the processes affected by the implicated regulatory variants, using DAVID and GeneMANIA. These analyses resulted in a bioinformatics pipeline that could theoretically be applied to the study of any disorder. Many of the identified SNPs, predicted affected genes and implicated pathways and networks supported previous findings, particularly with regards to the involvement of immune functions and neurodevelopmental processes. Additionally, many SNPs were novel findings and pointed towards the possible involvement of disrupted regulation during fundamental processes such as DNA packaging. One such novel finding was rs200981 and was the most significant variant in terms of regulatory potential. Therefore, regarding the bioinformatic analyses, the aims of this study were met.

The second part of the study was important in order to validate the findings of the abovementioned bioinformatic analyses in a pilot association study. For this association study a cohort of South African Xhosa cases and controls was chosen due to the fact that there is a large void with regards to the inclusion of African populations in schizophrenia genetics studies. Furthermore, this population group is considered genetically homogenous and provides valuable genetic information. The selected regulatory variants and corresponding SNPs from previous GWAS (with which these variants occurred in LD), were successfully genotyped in the entire cohort. Statistical analyses of associations between genotyped SNPs / SNP haplotypes and schizophrenia susceptibility and symptoms identified a number of significant associations. The most significant association was between two

haplotypes of regulatory variants (rs200483-rs200485-rs2517611 and rs200981-rs2517611-rs3129701) and schizophrenia susceptibility. Importantly, none of the SNPs that were previously associated with schizophrenia in GWAS of European individuals were significantly associated with schizophrenia in the Xhosa cohort. This supported the notion that it is necessary to replicate significant findings in different populations.

5.1 LIMITATIONS OF THIS STUDY

While significant associations between the genotyped SNPs / SNP haplotypes and schizophrenia susceptibility as well as positive and negative symptoms were identified, a small cohort with lower power to detect significance was used. This means that significant associations could have been missed. Nonetheless, this pilot study validated the bioinformatic results of the first part of the study. An additional limitation included the tools used for bioinformatic analyses. Many of these tools are still in the development phase and therefore significant associations could have been missed. Additionally, due to inconsistencies between these bioinformatic tools (as demonstrated by the different genes predicted to be affected by RegulomeDB and rSNPBase) numerous tools needed to be used.

5.2 FUTURE DIRECTIONS

Future work would include optimization of bioinformatics pipelines for the analysis of non-coding regions. In fact, new programs are being developed and made publically available at a rapid pace, with many of the programs used in this study having being developed within the last three years. It is expected that with the progress of projects such as ENCODE and FANTOM, the programs that utilize these databases will also improve, thereby increasing the reliability of the results. Furthermore, the bioinformatic results of this study will be validated in a larger South African Xhosa cohort. Custom TaqMan® OpenArray® Genotyping Plates were designed to genotype more of the regulatory variants identified in this study in a larger cohort of Xhosa individuals. The genotyping is currently being done and it is hoped that the findings would provide valuable information regarding schizophrenia susceptibility in this population. However, validation of these results will also be required in other populations, due to the differences in population genetics as demonstrated in this study. Regarding the

inclusion of African populations in large scale studies, it is crucial that the various African populations be included in GWAS studies, particularly when focussing on complex phenotypes. These populations provide valuable, genetically unique information that could aid in uncovering the intricate genetic architectures of many complex diseases. Furthermore, an advantage of including individuals from these populations is that the LD patterns observed in their genomes differ significantly. The wealth of genetic variation due to the multitude of ethnic groups should be explored for the enhancement of future studies. The 1000 Genomes Project is in fact already involved in resequencing 100 individuals from each of the following five countries: Malawi, Kenya, Gambia, Nigeria and Sierra Leone (The 1000 Genomes Project Consortium, 2014). Though this will inevitably aid in understanding the genetic diversity of African individuals, it would still be crucial for individuals of populations from different regions of Africa to be included in large scale disease studies in the future. There can be no hope of alleviating the burden of crippling disorders on individuals from Africa without acknowledging the contribution that these individuals could make to future research efforts with regards to genetic information.

In conclusion, this study met the aims and objectives set out at the beginning. The major goal of personalized medicine is not only to determine optimal medication based on an individual's unique genetic architecture, but more importantly to be able to predict an individual's risk for an illness and achieve an accurate diagnosis (Ozomaro *et al.*, 2013). Difficulties in studying schizophrenia and risk for this disorder have mainly been due to the heterogeneity of the disorder, interactions between genetic and environmental factors, side effects of prescribed medication and disease stage (Tomasik *et al.*, 2014). However, novel approaches as used in this study could provide a new perspective and potentially be more fruitful. It is hoped that the findings of this study will stimulate future research studies, thereby contributing to the understanding of the genetic factors that are involved in schizophrenia susceptibility in order to alleviate the burden of this disorder.

REFERENCES

REFERENCES

- Abbott, K.C., Agodoa, L.Y., and O'Malley, P.G. (2003). Hospitalized psychoses after renal transplantation in the United States: incidence, risk factors, and prognosis. *Journal of the American Society of Nephrology*. *14*(6), 1628-1635.
- Abdeljaber, M.H., Nair, M.P., Schork, M.A., and Schwartz, S.A. (1994). Depressed natural killer cell activity in schizophrenic patients. *Immunological Investigations*. *23*, 259-268.
- Abecasis, G.R., Burt, R.A., Hall, D., Bochum, S., Doheny, K.F., Lundy, S.L., Torrington, M., *et al.* (2004). Genomewide scan in families with schizophrenia from the founder population of Afrikaners reveals evidence for linkage and uniparental disomy on chromosome 1. *American Journal of Human Genetics*. *74*(3), 403-417.
- Aberg, K.A., Liu, Y., Bukszár, J., McClay, J.L., Khachane, A.N., Andreassen, O.A., Blackwood, D., *et al.* (2013). A comprehensive family-based replication study of schizophrenia genes. *JAMA Psychiatry*. *70*(6), 573-581.
- Addington, A.M., Gornick, M., Duckworth, J., Sporn, A., Gogtay, N., Bobb, A., Greenstein, D., *et al.* (2005). *GAD1* (2q31.1), which encodes glutamic acid decarboxylase (GAD67), is associated with childhood-onset schizophrenia and cortical gray matter volume loss. *Molecular Psychiatry*. *10*(6), 581-588.
- Adzhubei, I.A., Schmidt, S., Peshkin, L., Ramensky, V.E., Gerasimova, A., Bork, P., Kondrashov, A.S., *et al.* (2010). A method and server for predicting damaging missense mutations. *Nature Methods*. *7*(4), 248-249.
- Akbarian, S., Ruehl, M.G., Bliven, E., Luiz, L.A., Peranelli, A.C., Baker, S.P., Roberts, R.C., *et al.* (2005). Chromatin alterations associated with down-regulated metabolic gene expression in the prefrontal cortex of subjects with schizophrenia. *Archives of General Psychiatry*. *62*, 829-840.
- Alexander, R.P., Fang, G., Rozowsky, J., Snyder, M., and Gerstein, M.B. (2010). Annotating non-coding regions of the genome. *Nature Reviews. Genetics*. *11*(8), 559-571.

REFERENCES

American Psychiatric Association. (2013). *Diagnostic and Statistical Manual of Mental Health Disorders 5th Ed. (DSM-V)*. American Psychiatric Publishing, Virginia.

Andreasen, N.C. (1990). Methods for assessing positive and negative symptoms. *Modern Problems of Pharmacopsychiatry*. 24, 707-747.

Arnedo, J., Svrakic, D.M., del Val, C., Romero-Zaliz, R., Hernández-Cuervo, H., Molecular Genetics of Schizophrenia Consortium, Fanous, A.H., *et al.* (2014). Uncovering the hidden risk architecture of the schizophrenias: confirmation in three independent genome-wide association studies. *American Journal of Psychiatry*. Advanced Access. doi:10.1176/appi.ajp.2014.14040435.

Arranz, M.J., and de Leon, J. (2007). Pharmacogenetics and pharmacogenomics of schizophrenia: a review of last decade of research. *Molecular Psychiatry*. 12, 707-747.

Ashdown, H., Dumont, Y., Ng, M., Poole, S., Boksa, P., and Luheshi, G.N. (2006). The role of cytokines in mediating effects of prenatal infection in the fetus: implications for schizophrenia. *Molecular Psychiatry*. 11, 47-55.

Bamne, M., Wood, J., Chowdari, K., Watson, A.M., Celik, C., Mansour, H., Klei, L., *et al.* (2012). Evaluation of HLA polymorphisms in relation to schizophrenia risk and infectious exposure. *Schizophrenia Bulletin*. 38(6), 1149-1154.

Barbeau, D., Liang, J.J., Robitaille, Y., Quirion, R., and Srivastava, L.K. (1995). Decreased expression of the embryonic form of the neural cell adhesion molecule in schizophrenic brains. *Proceedings of the National Academy of Sciences of the United States of America*. 92(7), 2785-2789.

Benros, M.E., Mortensen, P.B., and Eaton, W.W. (2012). Autoimmune diseases and infections as risk factors for schizophrenia. *Annals of the New York Academy of Sciences*. 1262, 56-66.

Bergen, S.E., O'Dushlaine, C.T., Ripke, S., Lee, P.H., Ruderfer, D.M., Akterin, S., Moran, J.L., *et al.* (2012). Genome-wide association study in a Swedish population yields support for greater CNV and MHC involvement in schizophrenia compared with bipolar disorder. *Molecular Psychiatry*. 17, 880-886.

REFERENCES

- Bertram, L. (2008). Genetic research in schizophrenia: new tools and future perspectives. *Schizophrenia Bulletin*. *34*(5), 806-812.
- Betcheva, E.T., Yosifova, A.G., Mushiroda, T., Kubo, M., Takahashi, A., Karachanak, S.K., Zaharieva, I.T., *et al.* (2013). Whole-genome-wide association study in the Bulgarian population reveals *HHAT* as schizophrenia susceptibility gene. *Psychiatric Genetics*. *23*, 11-19.
- Bieda, M., Xu, X., Singer, M.A., Green, R., and Farnham, P.J. (2006). Unbiased location analysis of E2FI-binding sites suggests a widespread role for E2F1 in the human genome. *Genome Research*. *16*, 595-605.
- Bikandi, J., San Millán, R., Rementeria, A., and Garaizar, J. (2004). *In silico* analysis of complete bacterial genomes: PCR, AFLP-PCR, and endonuclease restriction. *Bioinformatics*. *20*, 798-9.
- Bioinformatics Institute Singapore. An Online Sample Size Estimator. (2014). [Online] Available: <http://osse.bii.a-star.edu.sg/> Accessed: September 2014.
- Blouin, J.L., Dombroski, B.A., Nath, S.K., Lasseter, V.K., Wolyniec, P.S., Nestadt, G., Thornquist, M., *et al.* (1998). Schizophrenia susceptibility loci on chromosomes 13q32 and 8p21. *Nature Genetics*. *20*, 70-73.
- Boulanger, L.M. (2009). Immune proteins in brain development and synaptic plasticity. *Neuron*. *64*, 93-109.
- Boyle, A.P., Hong, E.L., Hariharan, M., Cheng, Y., Schaub, M.A., Kasowski, M., Karczewski, K.J., *et al.* (2012). Annotation of functional variation in personal genomes using RegulomeDB. *Genome Research*. *22*(9), 1790-1797.
- Braunholz, D., Hullings, M., Gil-Rodríguez, M.C., Fincher, C.T., Mallozzi, M.B., Loy, E., Albrecht, M., *et al.* (2012). Isolated *NIPBL* missense mutations that cause Cornelia de Lange syndrome alter *MAU2* interaction. *European Journal of Human Genetics*. *20*(3), 271-276.
- Bray, N.J. (2008). Gene expression in the etiology of schizophrenia. *Schizophrenia Bulletin*. *34*(3), 412-418.

REFERENCES

Broad Communications. \$650 million commitment to Stanley Center at Broad Institute aims to galvanize mental illness research. [Online] Available at: <http://www.broadinstitute.org/news/5896> Accessed: July 2014.

Bryne, J.C., Valen, E., Tang, M.E., Marstrand, T., Winther, O., da Piedade, I., Krogh, A., *et al.* (2008). JASPAR, the open access database of transcription factor-binding profiles: New content and tools in the 2008 update. *Nucleic Acids Research*. 36, D102-D106.

Burch, P.R.D. (1964). Schizophrenia: some new aetiological considerations. *British Journal of Psychiatry*. 110, 818-896.

Byrne, M., Agerbo, E., Ewald, H., Eaton, W.W., and Mortensen, P.B. (2003). Parental age and risk of schizophrenia: a case-control study. *Archives of General Psychiatry*. 60, 673-678.

Cacabelos, R., Cacabelos, P., and Aliev, G. (2013). Genomics of schizophrenia and pharmacogenomics of antipsychotic drugs. *Open Journal of Psychiatry*. 3, 46-139.

Cacabelos, R., Hashimoto, R., and Takeda, M. (2011). Pharmacogenomics of antipsychotics efficacy for schizophrenia. *Psychiatry and Clinical Neuroscience*. 65, 3-19.

Campbell, M.C., and Tishkoff, S.A. (2008). African genetic diversity: implications for human demographic history, modern human origins, and complex disease mapping. *Annual Review of Genomics and Human Genetics*. 9, 403-433.

Cantor, R.M., and Geschwind, D.H. (2008). Schizophrenia: Genome, interrupted. *Neuron*. 58, 165-167.

Carter, C.J. (2011). Schizophrenia: a pathogenetic autoimmune disease caused by viruses and pathogens and dependent on genes. *Journal of Pathogens*. 2011(128318), 1-37.

Cazzullo, C.L., Smeraldi, E., and Penati, G. (1974). The leucocyte antigen system HLA as possible genetic marker of schizophrenia. *British Journal of Psychiatry*. 125, 612-615.

Chen, X., Wang, X., O'Neill, A.F., Walsh, D., and Kendler, K.S. (2004). Variants in the catechol-o-methyltransferase (*COMT*) gene are associated with schizophrenia in Irish high-density families. *Molecular Psychiatry*. 9, 962-967.

REFERENCES

- Cirulli, E.T., and Goldstein, D.B. (2010). Uncovering the roles of rare variants in common disease through whole-genome sequencing. *Nature Reviews Genetics*. *11(6)*, 415-425.
- Clark, S.L., Adkins, D.E., and van den Oord, E.J.C.G. (2011). Analysis of efficacy and side effects in CATIE demonstrates drug response subgroups and potential for personalized medicine. *Schizophrenia Research*. *132*, 114-120.
- Coffee, M. (2012). Bioinformatic and molecular genetic analysis of novel candidate schizophrenia susceptibility genes (*MIR137*, *TRIM26*, *CNNM2*, *NT5C2*, *STT3A*) in a South African first episode schizophrenia cohort. Honours Thesis. Stellenbosch University.
- Collins, A.L., Kim, Y., Sklar, P., O'Donovan, M.C., and Sullivan, P.F. (2012). Hypothesis-driven candidate genes for schizophrenia compared to genome-wide association results. *Psychological Medicine*. *42(3)*, 607-616.
- Collins, P.Y., Patel, V., Joestl, S.S., March, D., Insel, T.R., and Daar, A.S. (2011). Grand challenges in global mental health. *Nature*. *475*, 27-30.
- Contos, J.J.A., Ishii, I., Fukushima, N., Kingsbury, M.A., Ye, X., Kawamura, S., Brown, J.H., *et al.* (2002). Characterization of *Ipa2* (Edg4) and *Ipa1/Ipa2* (Edg2/Edg4) lysophosphatidic acid receptor knockout mice: signaling deficits without obvious phenotypic abnormality attributable to *Ipa2*. *Molecular and Cellular Biology*. *22(19)*, 6921-6929.
- Cooper, G.M., and Shendure, J. (2011). Needles in stacks of needles: finding disease-causal variants in a wealth of genomic data. *Nature*. *12*, 628–640.
- Costa, E., Grayson, D.R., and Guidotti, A. (2003). Epigenetic downregulation of GABAergic function in schizophrenia: potential for pharmacological intervention? *Molecular Interventions*. *3(4)*, 220-229.
- Cowie, P., Ross, R., and MacKenzie, A. (2013). Understanding the dynamics of gene regulatory systems; characterisation and clinical relevance of cis-regulatory polymorphisms. *Biology*. *2(1)*. 64-84.
- Crick, F. (1970). Central Dogma of Molecular Biology. *Nature*. *227*, 561-563.

REFERENCES

- Cross-Disorder Group of the Psychiatric Genomics Consortium. (2013). Identification of risk loci with shared effects on five major psychiatric disorders: a genome-wide analysis. *The Lancet*. *381*(9875), 1371-1379.
- Curtis, D. (2013). Consideration of plausible genetic architectures for schizophrenia and implications for analytical approaches in the era of next generation sequencing. *Psychiatric Genetics*. *23*, 1-10.
- Davies, G., Welham, J., Chant, D., Torrey, E.F., and McGrath, J. (2003). A systematic review and meta-analysis of northern hemisphere season of birth studies in schizophrenia. *Schizophrenia Bulletin*. *29*, 587-593.
- Debnath, M., Cannon, D.M., and Venkatasubramanian, G. (2013). Variation in the major histocompatibility complex (MHC) gene family in schizophrenia: associations and functional implications. *Progress in Neuro-Psychopharmacology & Biological Psychiatry*. *42*, 49-62.
- DeLisi, L.E., Ortaldo, J.R., Maluish, A.E., and Wyatt, R.J. (1983). Deficient natural killer cell (NK) activity and macrophage functioning in schizophrenic patients. *Journal of Neural Transmission*. *58*, 99-106.
- Dempster, E., Viana, J., Pidsley, R., and Mill, J. (2013). Epigenetic studies of schizophrenia: progress, predicaments, and promises for the future. *Schizophrenia Bulletin*. *39*(1), 11-16.
- Deutsch, S.I., Rosse, R.R., Mastropaolo, J., Long, K.D., and Gaskins, B.L. (2008). Epigenetic therapeutic strategies for the treatment of neuropsychiatric disorders: ready for prime time? *Clinical neuropharmacology*. *31*(2), 104-119.
- Dina, C., Meyre, D., Gallina, S., Durand, E., Körner, A., Jacobson, P., Carlsson, L.M.S. *et al.* (2007). Variation in *FTO* contributes to childhood obesity and severe adult obesity. *Nature Genetics*. *39*, 724-726.
- Dong, E., Agis-Balboa, R.C., Simonini, M.V., Grayson, D.R., Costa, E., and Guidotti, A. (2005). Reelin and glutamic acid decarboxylase67 promoter remodeling in an epigenetic methionine-induced mouse model of schizophrenia. *Proceedings of the National Academy of Sciences*. *102*, 12578-12583.

REFERENCES

Drögemöller, B.I., Wright, G.E.B., Niehaus, D.J.H., Emsley, R.A., and Warnich, L. (2011). Whole-genome resequencing in pharmacogenomics: moving away from past disparities to globally representative applications. *Pharmacogenomics*. *12*(12), 1717-1728.

Dunham, I. (2013). ENCODE-ing the future. [Online] Available: <http://genengnews.com/gen-articles/encode-ing-the-future/4803/> Accessed: May 2014.

Duwe, B.V, and Turetsky, B.I. (2002). Misdiagnosis of schizophrenia in a patient with psychotic symptoms. *Neuropsychiatry, Neuropsychology, and Behavioral Neurology*. *15*(4), 252-260.

Eaton, W.W., Byrne, M., Ewald, H., Mors, O., Chen, C.Y., Agerbo, E., and Mortensen, P.B. (2006). Association of schizophrenia and autoimmune diseases: linkage of Danish national registers. *American Journal of Psychiatry*. *163*(3), 521-528.

Eckner, R., Ewen, M.E., Newsome, D., Gerdes, M., DeCaprio, J.A., Lawrence, J.B., and Livingston, D.M. (1994). Molecular cloning and functional analysis of the adenovirus E1A-associated 300-kD protein (p300) reveals a protein with properties of a transcriptional adaptor. *Genes & Development*. *8*, 869-884.

Egan, M.F., Goldberg, T.E., Kolachana, B.S., Callicott, J.H., Mazzanti, C.M., Straub, R.E., Goldman, D., *et al.* (2001). Effect of *COMT* Val108/158 Met genotype on frontal lobe function and risk for schizophrenia. *Proceedings of the National Academy of Sciences of the United States of America*. *98*, 6917-6922.

Elert, E. (2014). Searching for schizophrenia's roots. *Nature*. *508*, S2-S3.

Farag, S.S., and Caligiuri, M.A. (2006). Human natural killer cell development and biology. *Blood Reviews*. *20*(3), 123-137.

Farooqui, T., Farooqui, A.A., Anderson, G., and Maes, M. (2013). Metabolic syndrome, Alzheimer disease, schizophrenia, and depression: role for leptin, melatonin, kynurenine pathways, and neuropeptides. *Metabolic Syndrome and Neurological Disorders*. DOI: 10.1002/9781118395318.ch13

REFERENCES

- Fellerhof, B., and Wank, R. (2009). Transporter associated with antigen processing and the chaperone tapasin: are non-classical *HLA* genes keys to the pathogenesis of schizophrenia? *Medical Hypotheses*. 72, 535-538.
- Fernandes, P., Jain, P., and Moita, C. (2012). Training experimental biologists in bioinformatics. *Advances in Bioinformatics*. 2012, 1-5.
- Forrest, A., Kawaji, H., Rehli, M., Baillie, J.K., de Hoon, M.J., Lassmann, T., Itoh, M., *et al.* (2014). A promoter-level mammalian expression atlas. *Nature*. 507(7493), 462-470.
- Gaffney, D.J. (2013). Global properties of functional complexity of human gene regulatory variation. *PLOS Genetics*. 9(5), 1-8.
- Gehrmann, J., Matsumoto, Y., and Kreutzberg, G.W. (1995). Microglia: intrinsic immuneffector cell of the brain. *Brain Research Reviews*. 20, 269-287.
- Gejman, P.V., Sanders, A.R., and Duan, J. (2011). The role of genetics in the etiology of schizophrenia. *Psychiatric Clinics of North America*. 33(1), 35-66.
- Gelernter, J. (2014). Genetics of complex traits in psychiatry. *Biological Psychiatry*. Advanced Access. DOI: <http://dx.doi.org/10.1016/j.biopsych.2014.08.005>.
- Gibson, G. (2012). Rare and common variants: twenty arguments. *Nature Reviews Genetics*. 13, 135-145.
- Girard, S.L., Gauthier, J., Noreau, A., Xiong, L., Zhou, S., Jouan, L., Dionne-Laporte, A., *et al.* (2011). Increased exonic *de novo* mutation rate in individuals with schizophrenia. *Nature Genetics*. 43(9), 860-864.
- Gonzalez-Pinto, A., Gutierrez, M., Mosquera, F., Ballesteros, J., Lopez, P., and Ezcurra, J. (1998). First-episode in bipolar disorder: misdiagnosis and psychotic symptoms. *Journal of Affective Disorders*. 50, 41-44.
- Goralczyk, A.D., Meier, V., Ramadori, G., Obed, A., and Lorf, T. (2010). Acute paranoid psychosis as sole clinical presentation of hepatic artery thrombosis after living donor liver transplantation. *BMC Surgery*. 10(7), 1-5.

REFERENCES

- Gottesman, I.I., and Shields, J. (1967). A polygenic theory of schizophrenia. *Proceedings of the National Academy of Science*. 58, 199-205.
- Greenwood, T.A., Swerdlow, N.R., Gur, R.E., Cadenhead, K.S., Calkins, M.E., Dobie, D.J., Freedman, R., *et al.* (2013). Genome-wide linkage analyses of 12 endophenotypes for schizophrenia from the Consortium on the Genetics of Schizophrenia. *The American Journal of Psychiatry*. 170(5), 521-32.
- Groenewald, P., Bradshaw, D., Daniels, J., Matzopoulos, R., Bourne, D., Blease, D., Zinyaktira, N., *et al.* (2008). Cause of death and premature mortality in Cape Town, 2001-2006. South African Medical Research Council Report.
- Guan, F., Wei, S., Feng, J., Zhang, C., Xing, B., Zhang, H., Gao, C., *et al.* (2012). Association study of a new schizophrenia susceptibility locus of 10q24.32-33 in a Han Chinese population. *Schizophrenia Research*. 138, 63–68.
- Guidotti, A., Ruzicka, W., Grayson, D.R., Veldic, M., Pinna, G., Davis, J.M., and Costa, E. (2007). S-adenosyl methionine and DNA methyltransferase-1 mRNA overexpression in psychosis. *Neuroreport*. 18, 57-60.
- Guo, L., Du, Y., Chang, S., Zhang, K., and Wang, J. (2013). rSNPBase: a database for curated regulatory SNPs. *Nucleic Acids Research*. 42(D1), D1033-D1039.
- Guo, Y., Monahan, K., Wu, H., Gertz, J., Varley, K.E., Li, W., Myers, R.M., *et al.* (2012). CTCF/cohesion mediated DNA looping is required for protocadherin a promoter choice. *Proceedings of the National Academy of Science*. 109, 21081–21086.
- Hall, D., Wijsman, E.M., Roos, J.L., Gogos, J.A., and Karayiorgou, M. (2002). Extended intermarker linkage disequilibrium in the Afrikaners. *Genome Research*. 12, 956-961.
- Halley, L., Doherty, M.K., Megson, I.L., McNamara, N., Gadja, A., and Wei, J. (2013). Search for schizophrenia susceptibility variants at the *HLA-DRB1* locus among a British population. *Immunogenetics*. 65(1), 1-7.
- Haraldsson, H.M., Ettinger, U., and Sigurdsson, E. (2011). Developments in schizophrenia genetics: From linkage to microchips, deletions and duplications. *Nordic Journal of Psychiatry*. 65(2), 82-88.

REFERENCES

- Harrison, P.J., and Weinberger, D.R. (2005). Schizophrenia genes, gene expression, and neuropathology: on the matter of their convergence. *Molecular Psychiatry*. *10*, 40-68.
- Harvey, P.D., and Walker, E.F. (1987). Positive and negative symptoms of psychosis: Description, research and future directions. Erlbaum, New Jersey. 341.
- Heston, L.L. (1966). Psychiatric disorders in foster home reared children of schizophrenic mothers. *The British Journal of Psychiatry*. *12(489)*, 819-825.
- Hindorff, L.A., Sethupathy, P., Junkins, H.A., Ramos, E.M., Mehta, J.P., Collins, F.S., and Manolio, T.A. (2009). Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proceedings of the National Academy of Sciences of the United States of America*. *106(23)*, 9362-9367.
- Hirayama, T., Tarusawa, E., Yoshimura, Y., Galjart, N., and Yagi, T. (2012). CTCF is required for neural development and stochastic expression of clustered Pcdh genes in neurons. *Cell Reports*. *2*, 345-357.
- Hitzeroth, A., Niehaus, D.J.H., Koen, L., Botes, W.C., Deleuze, J.F., and Warnich, L. (2007). Association between the *MnSOD* Ala-9Val polymorphism and development of schizophrenia and abnormal involuntary movements in the Xhosa population. *Progress in Neuro-Psychopharmacology and Biological Psychiatry*. *31(3)*, 664-672.
- Hsieh, J., and Gage, F.H. (2005). Chromatin remodeling in neural development and plasticity. *Current Opinion in Cell Biology*. *17*, 664-671.
- Hsu, R., Woodroffe, A., Lai, W.S., Cook, M.N., Mukai, J., Dunning, J.P., Swanson, D.J., *et al.* (2007). Nogo Receptor 1 (*RTN4R*) as a candidate gene for schizophrenia: analysis using human and mouse genetic approaches. *PLoS One*. *2(11)*, e1234.
- Huang, D.W., Sherman, B.T., and Lempicki, R.A. (2009a). Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene sets. *Nucleic Acids Research*. *37(1)*, 1-13.

REFERENCES

- Huang, D.W., Sherman, B.T., and Lempicki, R.A. (2009b). Systematic and integrative analysis of large gene lists using DAVID Bioinformatics Resources. *Nature Protocols*. *4*(1), 44-57.
- Issidorides, M.R., Stefanis, C.N., Varsou, E., and Katsorichis, T. (1975). Altered chromatin ultrastructure in neutrophils of schizophrenics. *Nature*. *258*, 612-614.
- Issidorides, M.R., Zioudrou, C., Lykouras, E., Stefanis, C.N. (1978). Drug-induced changes in chromatin ultrastructure and nuclear basic proteins of the neutrophils of chronic schizophrenics. *Progress in Neuropsychopharmacology*. *2*, 79-85.
- Johnson, A.D., Handsaker, R.E., Pulit, S.L., Nizzari, M.M., O'Donnell, C.J.O., and de Bakker, P.I.W. (2008). SNAP: a web-based tool for identification and annotation of proxy SNPs using HapMap. *Bioinformatics*. *24*(24), 2938-2939.
- Jones, A., Mowry, B.J., Pender, M.P., and Greer, J.M. (2005). Immune dysregulation and self-reactivity in schizophrenia: do some cases of schizophrenia have an autoimmune basis? *Immunology and Cell Biology*. *83*(1), 9-17.
- Juraeva, D., Haenisch, B., Zapatka, M., Frank, J., GROUP Investigators; PSYCH-GEMS SCZ Working Group, Witt, S.H., *et al.* (2014). Integrated pathway-based approach identifies association between genomic regions at *CTCF* and *CACNB2* and schizophrenia. *PLOS Genetics*. *10*(6), e1004345.
- Juvonen, H., Reunanen, A., Haukka, J., Muhonen, M., Suvisaari, J., Arajärvi, R., Partonen, T., *et al.* (2007). Incidence of schizophrenia in a nationwide cohort of patients with type 1 diabetes mellitus. *Archives of General Psychiatry*. *64*(8), 894-899.
- Karayiorgou, M., Morris, M.A., Morrow, B., Shprintzen, R.J., Goldberg, R., Borrow, J., Gos, A., *et al.* (1995). Schizophrenia susceptibility associated with interstitial deletions of chromosome 22q11. *Proceedings of the National Academy of Science*. *92*, 7612-7616.
- Kawasaki, H., and Iwamuro, S. (2008). Potential roles of histones in host defense as antimicrobial agents. *Infectious Disorders – Drug Targets*. *8*, 195-205.

REFERENCES

- Kay, S.R., Fiszbein, A., and Opler, L.A. (1987). The positive and negative syndrome scale (PANSS) for schizophrenia. *Schizophrenia Bulletin*. *13*(2), 261-276.
- Kety, S.S., Rosenthal, D., Wender, P.H., Schulsinger, F., and Jacobsen, B. (1976). Mental illness in the biological and adoptive families of adopted individuals who have become schizophrenic. *Behavior Genetics*. *6*(3), 219-225.
- Kim, H.S., Cho, J.H., Park, H.W., Yoon, H., Kim, M.S., Kim, S.C. (2002). Endotoxin-neutralizing antimicrobial proteins of the human placenta. *Journal of Immunology*. *168*, 2356-2364.
- Kim, Y., Zerwas, S., Trace, S.E., and Sullivan, P.F. (2011). Schizophrenia genetics: where next? *Schizophrenia Bulletin*. *37*(3), 456-463.
- Kirch, D.G. (1993). Infection and autoimmunity as etiologic factors in schizophrenia: a review and reappraisal. *Schizophrenia Bulletin*. *19*, 355-70.
- Klein, R.J., Zeiss, C., Chew, E.Y., Tsai, J.Y., Sackler, R.S., Haynes, C., Henning, A.K., *et al.* (2005). Complement factor h polymorphism in age-related macular degeneration. *Science*. *308*(5720), 385–389.
- Knapp, M., Mangalore, R., and Simon, J. (2004). The global costs of schizophrenia. *Schizophrenia Bulletin*. *30*(2), 279-293.
- Koen, L., Niehaus, D.J., De Jong, G., Muller, J.E., and Jordaan, E. (2006). Morphological features in a Xhosa schizophrenia population. *BMC Psychiatry*. *6*(47), 1-5.
- Korte, A., and Farlow, A. (2013). The advantages and limitations of trait analysis with GWAS: a review. *Plant Methods*. *9*(29), 1-20.
- Kraft, P., and Hunter, D.J. (2009). Genetic risk prediction - are we there yet? *The New England Journal of Medicine*. *360*(17), 1701–1703.
- Lambert, J., Ibrahim-Verbaas, C.A., Harold, D., Naj, A.C., Sims, R., Bellenguez, C., Jun, G., *et al.* (2013). Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. *Nature Genetics*. *45*, 1452-1458.

REFERENCES

- Lang, U.E., Willbring, M., von Golitschek, R., Schmeisser, A., Matschke, and K., Malte Tugtekin, S. (2008). Clozapine-induced myocarditis after long-term treatment: case presentation and clinical perspectives. *Journal of Psychopharmacology*. 22(5), 576-580.
- Lauber, C., Keller, C., Eichenberger, A., and Rössler, W. (2005). Family burden during exacerbation of schizophrenia: quantification and determinants of additional costs. *International Journal of Social Psychiatry*. 51(3), 259-264.
- Laurent, C., Niehaus, D., Bauché, S., Levinson, D.F., Soubigou, S., Pimstone, S., Hayden, M., *et al.*, (2003). CAG repeat polymorphism in *KCNN3* (HSKCa3) and *PPP2R2B* show no association or linkage to schizophrenia. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics*. 116B, 45-50.
- Le Roux, R., Niehaus, D.J.H., Koen, L., Seller, C., Lochner, C., and Emsley, R.A. (2007). Initiation rites as a perceived stressor for Isixhosa males with schizophrenia. *Transcultural Psychiatry*. 44, 292-299.
- Lee, S., Tsao, T.T., Yang, K., Lin, H., Kuo, Y., Hsu, C., Lee, W., *et al.* (2011). Construction and analysis of the protein-protein interaction networks for schizophrenia, bipolar disorder, and major depression. *BMC Bioinformatics*. 12, 1-15.
- Lee, S.H., DeCandia, T.R., Ripke, S., Yang, J., PGC-SCZ, ISC, MGS, *et al.* (2012). Estimating the proportion of variation in susceptibility to schizophrenia captured by common SNPs. *Nature Genetics*. 44(3), 247-252.
- Lee, Y.H., Kim, J., and Song, G.G. (2013). Pathway analysis of a genome-wide association study. *Gene*. 525(1), 107-115.
- Lencz, T., Guha, S., Liu, C., Rosenfeld, J., Mukherjee, S., DeRosse, P., John, M., *et al.* (2013). Genome-wide association study implicates *NDST3* in schizophrenia and bipolar disorder. *Nature Communications*. 1 – 10.
- Lencz, T., Robinson, D.G., Xu, K., Ekholm, J., Sevy, S., Gunduz-Bruce, H., Woerner, M.G., *et al.* (2006). *DRD2* promoter region variation as a predictor of sustained response to antipsychotic medication in first episode schizophrenia patients. *American Journal of Psychiatry*. 163, 529-531.

REFERENCES

- Lessard, J., Wu, J.I., Ranish, J.A., Wan, M., Winslow, M.M., Staahl, B.T., Wu, H., *et al.* (2007). An essential switch in subunit composition of a chromatin remodeling complex during neural development. *Neuron*. 55(2), 201-215.
- Levy, D.L., Coleman, M.J., Sung, H., Ji, F., Matthyse, S., Mendell, N.R., and Titone D. (2010). The genetic basis of thought disorder and language and communication disturbances in schizophrenia. *Journal of Neurolinguistics*. 23(3), 1-23.
- Li, C., Liao, H., Hung, T., and Chen, C. (2006). Mutation analysis of *DARPP-32* as a candidate gene for schizophrenia. *Schizophrenia Research*. 87(1-3), 1-5.
- Li, M., Zhang, H., Luo, X., Gao, L., Qi, X., Gourraud, P., and Su, B. (2013). Meta-analysis indicates that the European GWAS-identified risk SNP rs1344706 within *ZNF804A* is not associated with schizophrenia in Han Chinese population. *PLOS One*. 8(6), e65780.
- Li, M.J., Wang, L.Y., Xia, Z., Sham, P.C., and Wang, J. (2013). GWAS3D: detecting human regulatory variants by integrative analysis of genome-wide associations, chromosome interactions and histone modifications. *Nucleic Acids Research*. 41(W), W150-W158.
- Li, T., Underhill, J., Liu, X.H., Sham, P.C., Donaldson, P., Murray, R.M., Wright, P., *et al.* (2001). Transmission disequilibrium analysis of HLA class II *DRB1*, *DQA1*, *DQB1* and *DPB1* polymorphisms in schizophrenia using family trios from a Han Chinese population. *Schizophrenia Research*. 49, 73–78.
- Lieberman, J.A., Stroup, T.S., McEvoy, J.P., Schwartz, M.S., Rosenheck, R.A., and Perkins, D.O. (2005). Effectiveness of antipsychotic drugs in patients with chronic schizophrenia. *The New England Journal of Medicine*. 353(12), 1209-1223.
- Lipska, B.K., Jaskiw, G.E., and Weinberger, D.R. (1993). Postpubertal emergence of hyperresponsiveness to stress and to amphetamine after neonatal excitotoxic hippocampal damage: a potential animal model of schizophrenia. *Neuropsychopharmacology*. 9, 67-75.
- Lopez, A.D., Mathers, C.D., Ezzati, M., Jamison, D.T. and Murray, C.J.L. (2006). *Global burden of disease and risk factors*. Oxford University Press, New York.
- Lund, C., Kleintjes, S., Kakuma, R., Flisher, A.J. and MHaPP Research Programme Consortium. (2010). *Public sector mental health systems in South Africa: inter-provincial*

REFERENCES

comparisons and policy implications. *Social Psychiatry Psychiatric Epidemiology*. 45, 393-404.

Luo, X., Huang, L., Jia, P., Li, M., Su, B., Zhao, Z., and Gan, L. (2013). Protein-protein interaction and pathway analyses of top schizophrenia genes reveal schizophrenia susceptibility genes converge on common molecular networks and enrichment of nucleosome (chromatin) assembly genes in schizophrenia susceptibility loci. *Schizophrenia Bulletin*. 40(1), 39-50.

Manolio, T.A., Collins, F.S., Cox, N.J., Goldstein, D.B., Hindorff, L.A., Hunter, D.J., McCarthy, M.I., *et al.* (2009). Finding the missing heritability of complex diseases. *Nature*. 461, 747-753.

Maric, N.P., and Svrakic, D.M. (2012). Why schizophrenia genetics needs epigenetics: a review. *Psychiatria Danubina*. 24(1), 2-18.

Martens, L., Van Damme, P., Van Damme, J., Staes, A., Timmerman, E., Ghesquière, B., Thomas, G.R., *et al.* (2006). The human platelet proteome mapped by peptide-centric proteomics: a functional protein profile. *Proteomics*. 5(12), 3193-3204.

Maston, G.A., Evans, S.K., and Green, M.R. (2006). Transcriptional regulatory elements in the human genome. *Annual Review of Genomics and Human Genetics*. 7, 29-59.

Matys, V., Kel-Margoulis, O.V., Fricke, E., Liebich, I., Land, S., Barre-Dirrie, A., Reuter, I., *et al.* (2006). TRANSFAC® and its module TRANSCompel®: transcriptional gene regulation in eukaryotes. *Nucleic Acids Research*. 34, D108-110.

Maurano, M.T., Humbert, R., Rynes, E., Thurman, R.E., Haugen, E., Wang, H., Reynolds, A.P., *et al.* (2012). Systematic localization of common disease-associated variation in regulatory DNA. *Science*. 337(6099), 190-1195.

May, A., Hazelhurst, S., Li, Y., Norris, S.A., Govind, N., Tikly, M., Hon, C., *et al.* (2013). Genetic diversity in black South Africans from Soweto. *BMC Genomics*. 14(644), 1-12.

McDaniel, J.S., Jewart, R.D., Eccard, M.B., Pollard, W.E., Caudle, J., Stipetic, M., Risby, E.D., *et al.* (1992). Natural killer cell activity in schizophrenia and schizoaffective disorder: a pilot study. *Schizophrenia Research*. 8(2), 125-128.

REFERENCES

McInerney, J.O. (2002). Bioinformatics in a post-genomics world – the need for an inclusive approach. *The Pharmacogenomics Journal*. 2, 207-208.

Mehta, U., Durrheim, D.N., Blockman, M., Kredo, T., Gounden, R., and Barnes, K.I. (2007). Adverse drug reactions in adult medical inpatients in a South African hospital serving a community with a high HIV/AIDS prevalence: prospective observational study. *British Journal of Clinical Pharmacology*. 65(3), 398-404.

Meltzer, H.Y. (2012). Clozapine: balancing safety with superior antipsychotic efficacy. *Clinical Schizophrenia and Related Psychoses*. 134-144.

Meyer, B., Wittig, I., Trifilieff, E., Karas, M., and Schagger, H. (2007). Identification of two proteins associated with mammalian ATP synthase. *Molecular and Cellular Proteomics*. 6, 1690-1699.

Millar, J.K, Wilson-Annan, J.C., Anderson, S., Christie, S., Taylor, M.S., Semple, C.A.M., Devon, R.S., *et al.* (2000). Disruption of two novel genes by a translocation co-segregating with schizophrenia. *Human Molecular Genetics*. 9(9), 1415–1423.

Miller, S.A., Dykes, D.D., and Polesky, H.F. (1988). A simple salting out procedure for extracting DNA from human nucleated cells. *Nucleic Acids Research*. 16(3), 1215.

Monahan, K., Rudnick, N.D., Kehayova, P.D., Pauli, F., Newberry, K.M., Myers, R.M., Maniatis, T. (2012). Role of CCCTC binding factor (CTCF) and cohesin in the generation of single-cell diversity of protocadherin-a gene expression. *Proceedings of the National Academy of Science*. 109, 9125–9130.

Mowry, B.J., and Gratten, J. (2013). The emerging spectrum of allelic variation in schizophrenia: current evidence and strategies for the identification and functional characterization of common and rare variants. *Molecular Psychiatry*. 18, 38-52.

Mulle, J.G. (2012). Schizophrenia genetics: progress, at last. *Current Opinion in Genetics & Development*. 22, 238-244.

Müller, D.J., Chowdhury, N.I., and Zai, C.C. (2013). The pharmacogenetics of antipsychotic-induced adverse events. *Current Opinion in Psychiatry*. 26(2), 144-150.

REFERENCES

- Müller, N., and Schwarz, M.J. (2010). Immune system and schizophrenia. *Current Immunology Reviews*. 6(3), 213-220.
- National Heart, Lung, and Blood Institute. (2013). SeattleSeq Annotation 137. [Online] Available: <http://snp.gs.washington.edu/SeattleSeqAnnotation137/>
- National Institute of Allergy and Infectious Diseases (NIAID), NIH. (2014). DAVID Bioinformatics Resources 6.7. [Online] Available: <http://david.abcc.ncifcrf.gov/> Accessed: June 2014.
- Nebert, D.W., Zhang, G., and Vesell, E.S. (2008). From human genetics and genomics to pharmacogenetics and pharmacogenomics: past lessons, future directions. *Drug Metabolism Reviews*. 40(2),187–224.
- Niehaus, D.J., du Plessis, S.A., Koen, L., Lategan, B.H., Steyn, J. Oosthuizen, P.P., Warnich, L., *et al.* (2008). Predictors of abnormal involuntary movement in an African schizophrenia population. *Journal of Neuropsychiatry and Clinical Neurosciences*. 20, 317-326.
- Niehaus, D.J.H., Koen, L., Laurent, C., Muller, J., Deleuze, J., Mallet, J., Seller, C., *et al.* (2005). Positive and negative symptoms in affected sib pairs with schizophrenia: Implications for genetic studies in an African Xhosa sample. *Schizophrenia Research*. 79, 239-249.
- Nielsen, C.S., Mørup-Lendal, M., Hansen, M.G., and Niemeier, S. (2014). *In Silico Prædiktion af Skizofreni-relaterede Varianter*. Thesis. Roskilde University.
- Overall, J.E. and Gorham, D.R. (1962). Brief psychiatric rating scale. *Psychological Reports*. 10, 799-812.
- Owen, M.J., Craddock, N., and O'Donovan, M.C. (2010). Suggestion of roles for both common and rare risk variants in genome-wide studies of schizophrenia. *Archives of General Psychiatry*. 67(7), 667-673.
- Ozomaro, U., Wahlestedt, C., and Nemeroff, C.B. (2013). Personalised medicine in psychiatry: problems and promises. *BMC Medicine*. 11, 132-167.

REFERENCES

- Palmatier, M.A., Pakstis, A.J., Speed, W., Paschou, P., Goldman, D., Odunsi, A., Okonofua, F., *et al.* (2004). *COMT* haplotypes suggest P2 promoter region relevance for schizophrenia. *Molecular Psychiatry*. *9(9)*, 859-70.
- Patel, V., Flisher, A.J., Nikapota, A., and Malhotra, S. (2008). Promoting child and adolescent mental health in low and middle income countries. *Journal of Child Psychology and Psychiatry*. *49(3)*, 313–334.
- Perkins, D.O., Jeffries, C., and Sullivan, P. (2005). Expanding the ‘central dogma’: the regulatory role of nonprotein coding genes and implications for the genetic liability to schizophrenia. *Molecular Psychiatry*. *10(1)*, 69-78.
- Petronijević, N.D., Radonjić, N.V., Ivković, M.D., Marinković, D., Piperski, V.D., Duricić, B.M., and Paunović, V.R. (2008). Plasma homocysteine levels in young male patients in the exacerbation and remission phase of schizophrenia. *Progress in Neuro-Psychopharmacology & Biological Psychiatry*. *32*, 1921-1926.
- Petronis, A., Gottesman, I.I., Kan, P., Kennedy, J.L., Basile, V.S., Paterson, A.D., and Pependikyte, V. (2003). Monozygotic twins exhibit numerous epigenetic differences: clues to twin discordance? *Schizophrenia Bulletin*. *29*, 169–178.
- Phillips, J.E., and Corces, V.G. (2009). CTCF: master weaver of the genome. *Cell*. *137*, 1194-1211.
- Prokunina, L., and Alarcón-Riquelme, M.E. (2004). Regulatory SNPs in complex diseases: their identification and functional validation. *Expert Reviews In Molecular Medicine*. *6(10)*, 1-15.
- Purcell, S.M., Moran, J.L., Fromer, M., Ruderfer, D., Solovieff, N., Roussos, P., O’Dushlaine, C., *et al.* (2014). A polygenic burden of rare disruptive mutations in schizophrenia. *Nature*. *506(7487)*, 185-190.
- Purcell, S.M., Wray, N.R., Stone, J.L., Visscher, P.M., O’Donovan, M.C., Sullivan, P.F., Sklar, P., *et al.* (2009). Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature*. *460*, 748-752.

REFERENCES

Purgato, M., Adams, C., and Barbu C. (2012). Schizophrenia trials conducted in African countries: a drop of evidence in the ocean of morbidity? *International Journal of Mental Health Systems*. 6(9), 1-10.

R Core Team. (2014). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Available: <http://www.R-project.org/>

Reardon, S. (2014). Gene-hunt gain for mental health. *Nature*. 511, 393.

Réthelyi, J., Benkovits, J., and Bitter, I. (2013). Genes and environments in schizophrenia: The different pieces of a manifold puzzle. *Neuroscience and Biobehavioral Reviews*. 37(10.1), 2424-2437.

Restriction Mapper. (2014). [Online] Available: <http://www.restrictionmapper.org/> Accessed July 2014.

Rietschel, M., Mattheisen, M., Degenhardt, F., GROUPE Investigators, Mühleisen, T.W., Kirsch, P., Esslinger, C., *et al.* (2012). Association between genetic variation in a region on chromosome 11 and schizophrenia in large samples from Europe. *Molecular Psychiatry*. 17(9), 906-917.

Riley, B.P., Rajogopalan, S., Mogudi-Carter, M. Jenkins, T., and Williamson, R. (1996a). No evidence for linkage of chromosome 6p markers to schizophrenia in southern African Bantu-speaking families. *Psychiatric Genetics*. 6, 41-49.

Riley, B., Mogudi-Carter, M., Jenkins, T., and Williamson, R. (1996b). No evidence for linkage of chromosome 22 markers to schizophrenia in southern African Bantu-speaking families. *American Journal of Medical Genetics*. 67, 515-522.

Riley, B.P., Makoff, A., Mogudi-Carter, M., Jenkins, T., Williamson, R., Collier, D., Murray, R. (2000). Haplotype transmission disequilibrium and evidence for linkage of the *CHRNA7* gene region to schizophrenia in Southern African Bantu families. *American Journal of Medical Genetics*. 96, 196-201.

Riley, B., Williamson, M., Collier, D., Wilkie, H., and Makoff, A. (2002). A 2-Mb map of a large segmental duplication overlapping the $\alpha 7$ -nicotinic acetylcholine receptor gene (*CHRNA7*) at human 15q13-q14. *Genomics*. 79(2), 197-209.

REFERENCES

- Riley, B., and Kendler, K.S., (2006). Molecular genetic studies of schizophrenia. *European Journal of Human Genetics*. *14*, 669–680.
- Ripke, S., Sanders, A.R., Kendler, K.S., Levinson, D.F., Sklar, P., Holmans, P.A., Lin, D., *et al.* (2011). Genome-wide association study identifies five new schizophrenia loci. *Nature Genetics*. *43*, 969-978.
- Ripke, S., O’Dushlaine, C., Chambert, K., Moran, J.L., Kähler, A.K., Akterin S., Bergen, S.E., *et al.* (2013). Genome-wide association analysis identifies 13 new risk loci for schizophrenia. *Nature Genetics*. *45(10)*, 1150-1159.
- Ripke, S., Neale, B.M., Corvin, A., Walters, J.T.R., Farh, K., Holmans, P.A., Lee, P., *et al.* (2014). Biological insights from 108 schizophrenia-associated genetic loci. *Nature*. *511*, 421-438.
- Risch, N., and Merikangas, K. (1996). The future of genetic studies of complex human diseases. *Science*. *273*, 1516–1517.
- Rodriguez-Murillo, L., Xu, B., Roos, J.L., Abecasis, G.R., Gogos, J.A. and Karayiorgou, M. 2014. Fine mapping on chromosome 13q32 – 34 and brain expression analysis implicates *MYO16* in schizophrenia. *Neuropsychopharmacology*. *39*, 934–943.
- Roos, J.L., Pretorius, H.W., and Karayiorgou, M. (2009). Clinical characteristics of an Afrikaner founder population recruited for a schizophrenia genetic study. *Annals of the New York Academy of Sciences*. *1151*, 85-101.
- Rosenbloom, K.R., Sloan, C.A., Malladi, V.S., Dreszer, T.R., Learned, K., Kirkup, V.M., Wong, M.C., *et al.* (2012) ENCODE Data in the UCSC Genome Browser: year 5 update. *Nucleic Acids Research*. *41(Database issue)*, D56-63.
- Rosoklija, G., Derkits, E., Serafimova, T., Dika, A., Mancevski, B., Stankov, A., Davceva, N., *et al.* (2007). Post mortem studies of dendritic abnormalities in schizophrenia and mood disorders. *Schizophrenia Bulletin*. *33*, 271-272.
- Rössler, W., Salize H.J., van Os, J. and Riecher-Rössler, A. (2005). Size of burden of schizophrenia and psychotic disorders. *European Neuropsychopharmacology*. *15*, 399-409.

REFERENCES

- Sadee, W., Hartmann, K., Seweryn, M., Pietrzak, M., Handelman, S.K., and Rempala, G.A. (2014). Missing heritability of common diseases and treatments outside the protein-coding exome. *Human Genetics*. *133(10)*, 1199-215.
- Savitz, J., Cupido, C.L., and Ramesar, R.K. (2007). Preliminary evidence for linkage to chromosome 1q31-32, 10q23.3, and 16p13.3 in a South African cohort with bipolar disorder. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics*. *144B*, 383-387.
- Schaub, M.A., Boyle, A.P, Kundaje, A., Batzoglou, S., and Snyder, M. (2012). Linking disease associations with regulatory information in the human genome. *Genome Research*. *22*, 1748-1759.
- Schizophrenia Research Forum. (2014). [Online] Available at: www.szgene.org Accessed: June 2014.
- Schuster, S.C., Miller, W., Ratan, A., Tomsho, L.P., Giardine, B., Kasson, L.R., Harris, R.S., *et al.* (2010). Complete Khoisan and Bantu genomes from southern Africa. *Nature*. *463*, 943-947.
- Seedat, S., Williams, D.R., Herman, A.A., Moomal, H., Williams, S.L., Jackson, P.B., Myer, L., *et al.* (2009). Mental health service use among South Africans for mood, anxiety and substance use disorders. *South African Medical Journal*. *99*, 346-352.
- Severinsen, J.E. (2006). Identification of susceptibility genes for bipolar affective disorder and schizophrenia on chromosome 22. *Danish Medical Bulletin*. *53(4)*, 456.
- Sharma, R.P. (2005). Schizophrenia, epigenetics and ligand-activated nuclear receptors: a framework for chromatin therapeutics. *Schizophrenia Research*. *72*, 79-90.
- Sharma, R.P., Grayson, D.R., Guidotti, A., and Costa, E. (2005). Chromatin, DNA methylation and neuron gene regulation — the purpose of the package. *Journal of Psychiatry and Neuroscience*. *30(4)*, 257-263.
- Shatz, C.J. (2002). Neural activity, immune genes and synaptic remodelling in brain development. *FASEB Journal*. *16*, A378-A379.

REFERENCES

Shi, J., Levinson, D.F., Duan, J., Sanders, A.R., Zheng, Y., Itsik, P., Dudbridge, F., *et al.* (2009). Common variants on chromosome 6p22.1 are associated with schizophrenia. *Nature*. *460*, 753-757.

Shi, Y., Li, Z., Xu, Q., Wang, T., Li, T., Shen, J., Zhang, F., *et al.*, (2011). Common variants on 8p12 and 1q24.2 confer risk of schizophrenia. *Nature Genetics*. *43*(12), 1224-1227.

Shifman, S., Bronstein, M., Sternfeld, M., Pisanté-Shalom, A., Lev-Lehman, E., Weizman, A. Reznik, I., *et al.* (2002). A highly significant association between a *COMT* haplotype and schizophrenia. *American Journal of Human Genetics*. *71*, 1296-1302.

Shimada, K., Watanabe, Y., Mokuno, H., Iwama, Y., Daida, H., and Yamaguchi, H. (2000). Common polymorphism in the promoter of the CD14 monocyte receptor gene is associated with acute myocardial infarction in Japanese men. *86*, 682–684.

Sinkus, M.L., Adams, C.E., Logel, J., Freedman, R., and Leonard, S. (2013). Expression of immune genes on chromosome 6p21.3-22.1 in schizophrenia. *Brain, Behavior, and Immunity*. *32*, 51-62.

Sinnwell, J.P., and Schaid, D.J. (2013). haplo.stats: Statistical Analysis of Haplotypes with Traits and Covariates when Linkage Phase is Ambiguous. R package version 1.6.8. Available: <http://CRAN.R-project.org/package=haplo.stats>

Smemo, S., Tena, J.J., Kim, K., Gamazon, E.R., Sakabe, N.J., Gómez-Marin, C., Aneas, I. *et al.* (2014). Obesity-associated variants within *FTO* form long-range functional connections with *IRX3*. *Nature*. *507*(7492), 371-375.

So, H.C., Fong, P.Y., Chen, R.Y., Hui, T.C., Ng, M.Y., Cherny, S.S., Mak, W.W., *et al.* (2010). Identification of Neuroglycan C and interacting partners of potential susceptibility genes for schizophrenia in a southern Chinese population. *American Journal of Medical Genetics*. *153B*(1), 103-113.

Sole, X., Guino, E., Valls, J., Iñesta, R., and Moreno, V. (2006). SNPStats: a web tool for the analysis of association studies. *Bioinformatics*. *22*(15), 1928-1929.

Department of Government Communication and Information System. Republic of South Africa. South Africa's People. [Online] Available:

REFERENCES

http://www.gcis.gov.za/sites/default/files/docs/resourcecentre/pocketguide/004_saspeople.pdf

Accessed: October 2013

South African Federation of Mental Health. (2014). [Online] Available: <http://www.safmh.org.za/infographics.htm> Accessed: February 2014.

Srivastava, V., Varma, P.G., Prasad, S., Semwal, P., Nimgaonkar, V.L., Lerer, B., Deshpande, S.N., *et al.* (2006). Genetic susceptibility to tardive dyskinesia among schizophrenia subjects: IV. Role of dopaminergic pathway gene polymorphisms. *Pharmacogenetics and Genomics*. *16*(2), 111-117.

Stachowiak, M.K., Kucinski, A., Curl, R., Syposs, C., Yang, Y., Narla, S., Terranova, C., *et al.* (2013). Schizophrenia: A neurodevelopmental disorder – Integrative genomic hypothesis and therapeutic implications from a transgenic mouse model. *Schizophrenia Research*. *143*(2-3), 367-376.

Statistics South Africa. (2013). [Online] Available: <http://beta2.statssa.gov.za/> Accessed: October 2013.

Stefansson, H., Ophoff, R.A., Steinberg, S., Andreassen, O.A., Cichon, S., Rujescu, D., Werge, T., *et al.* (2009). Common variants conferring risk of schizophrenia. *Nature*. *460*(7256), 744-747.

Stefansson, H., Sigurdsson, E., Steinthorsdottir, V., Bjornsdottir, S., Sigmundsson, T., Ghosh, S., Brynjolfsson, J., *et al.* (2002). Neuregulin 1 and susceptibility to schizophrenia. *American Journal of Human Genetics*. *71*, 877–892.

Stein, D.J., Seedat, S., Herman, A., Moomal, H., Heeringa, S.G., Kessler, R.C., and Williams, D.R. (2008). Lifetime prevalence of psychiatric disorders in South Africa. *The British Journal of Psychiatry*. *192*, 112-117.

Steinberg, S., de Jong, S., Irish Schizophrenia Genomics Consortium, Andreassen, O.A., Werge, T., Borglum, A.D., Mors, O., *et al.* (2011). Common variants at *VRK2* and *TCF4* conferring risk of schizophrenia. *Human Molecular Genetics*. *20*(20), 4076-4081.

REFERENCES

- Stephens, S.H., Logel, J., Barton, A., Franks, A., Schultz, J., Short M., Dickenson, J., *et al.* (2009). Association of the 5'-upstream regulatory region of the $\alpha 7$ nicotinic acetylcholine receptor subunit gene (*CHRNA7*) with schizophrenia. *Schizophrenia Research*. *109*, 102-112.
- Strakowski, S., Flaum, M., Amador, X., Bracha, H., Pandurangi, A., Robinson, D., and Tohen, M. (1996). Racial differences in the diagnosis of psychosis. *Schizophrenia Research*. *21*(2), 117-124.
- Strange, A., Riley, B.P., Spencer, C.C.A., Morris, D.W., Pirinen, M., O'Dushlaine, C.T., Su, Z., *et al.* (2012). Genome-wide association study implicates *HLA-C*01:02* as a risk factor at the major histocompatibility complex locus in schizophrenia. *Biological Psychiatry*. *72*(8), 620-628.
- Sullivan, P.F., Daly, M.J., and O'Donovan, M. (2012). Genetic architectures of psychiatric disorders: the emerging picture and its implications. *Nature Reviews Genetics*. *13*, 537-551.
- Sullivan, P.F., Kendler, K.S., and Neale, M.C. (2003). Schizophrenia as a complex trait. Evidence from a meta-analysis of twin studies. *Archives of General Psychiatry*. *60*, 1187-1192.
- Tandon, R., Keshavan, M.S., and Nasrallah, H.A. (2008). Schizophrenia, "Just the facts": What we know in 2008. Part 1: Overview. *Schizophrenia Research*. *100*, 4-19.
- Tandon, R., Nasrallah, H.A., and Keshavan, M.S. (2009). Schizophrenia, "Just the facts" 4. Clinical features and conceptualization. *Schizophrenia Research*. *110*(1-3), 1-23.
- Tandon, R., Nasrallah, H.A., and Keshavan, M.S. (2010). Schizophrenia, "Just the facts" 5. Treatment and prevention past, present, and future. *Schizophrenia Research*. *122*, 1-23.
- The 1000 Genomes Project Consortium. (2010). 1000 Genomes: A Deep Catalog of Human Genetic Variation. [Online] Available: <http://www.1000genomes.org/page.php?page=home>
- The 1000 Genomes Consortium. About 1000 Genomes. [Online]. Available: <http://www.1000genomes.org/about#ProjectSamples> Accessed: August 2014
- The ENCODE Project Consortium. (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature*. *489*, 57-74.

REFERENCES

The International HapMap Consortium. (2003). The international HapMap project. *Nature*. 426, 789-796.

Thomas-Chollier, M., Hufton, A., Heinig, M., O’Keeffe, S., Masri, N.E., Roider, H.G., Manke, T., *et al.* (2011). Transcription factor binding predictions using TRAP for the analysis of ChIP-seq data and regulatory SNPs. *Nature Protocols*. 6(12), 1860-1869.

Tishkoff, S.A., and Verrelli, B.C. (2003). Patterns of human genetic diversity: implications for human evolutionary history and disease. *Annual Review of Genomics and Human Genetics*. 4, 293-340.

Tomasik, J., Rahmoune, H., Guest, P.C. and Bahn, S. (2014). Neuroimmune biomarkers in schizophrenia. *Schizophrenia Research*. Advanced Access. DOI: <http://dx.doi.org/10.1016/j.schres.2014.07.025>.

University of Toronto. (2014). GeneMANIA. [Online] Available: <http://www.genemania.org/> Accessed: June & September 2014.

Vodjgani, M., Matloubi, H., Nasehi, A.A., Niknam, M.H., Kazemnejad, A., Salehi, E., Aboufazeli, T., *et al.* (2005). Increased natural killer cells activity in schizophrenic patients. *Iranian Journal of Immunology*. 2(2), 111-116.

Vojdani, A., Mumper, E., Granpeesheh, D., Mielke, L., Traver, D., Bock, K., Hirani, K., *et al.* (2008). Low natural killer cell cytotoxic activity in autism: the role of glutathione, IL-2 and IL-15. *Journal of Neuroimmunology*. 205(1-2), 148-154.

Walker, E., Kestler, L., Bollini, A., and Hochman, K.M. (2004). Schizophrenia: etiology and course. *Annual Review of Psychology*. 55, 401-430.

Wang, K., Zhang, Q., Liu, X., Wu, L., and Zeng, M. (2012). *PKNOX2* is associated with formal thought disorder in schizophrenia: a meta-analysis of two genome-wide association studies. *Journal of Molecular Neuroscience*. 48, 265-272.

Ward, L.D., and Kellis, M. (2012). HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. *Nucleic Acids Research*. 40(D1), D930–D934.

REFERENCES

- Warde-Farley, D., Donaldson, S.L., Comes, O., Zuberi, K., Badrawi, R., Chao, P., Franz, M., *et al.* (2010). The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function. *Nucleic Acids Research*. 38, W214-W220.
- Warnes, G., Gorjanc, G., Leisch, F., and Man, M. (2013). genetics: Population Genetics. R package version 1.3.8.1. Available: <http://CRAN.R-project.org/package=genetics>
- Weingarten, L.S., Dave, H., Li, H., and Crawford, D.A. (2012). Developmental expression of P5 ATPase mRNA in the mouse. *Cellular and Molecular Biology Letters*. 17(1), 153-170.
- Wiehahn, G.J., Bosch, G.P., Preez, R.R., Pretorius, H.W., Karayiorgou, M. and Roos, J.L. (2004). Assessment of the frequency of the 22q11 deletion in Afrikaner schizophrenic patients. *American Journal of Medical Genetics Part B (Neuropsychiatric Genetics)*. 129B, 20–22.
- Williams, H.J., Owen, M.J., and O'Donovan M.C. (2009). Schizophrenia genetics: new insights from new approaches. *British Medical Bulletin*. 91, 61-74.
- World Health Organisation. (2013). Schizophrenia. [Online] Available: http://www.who.int/mental_health/management/schizophrenia/en/ Accessed: 1 March 2013 & August 2014.
- Wright, G.E.B., Niehaus, D.J.H., Koen, L., Drögemöller, B.I., and Warnich, L. (2011). Psychiatric genetics in South Africa: cutting a rough diamond. *African Journal of Psychiatry*. 14, 355-366.
- Wright, G.E.B., Niehaus, D.J.H., van der Merwe, L., Koen, L., Korkie, L.J., Kinnear, C.J., Drögemöller, B.I., *et al.* (2012). Association of *MB-COMT* polymorphisms with schizophrenia-susceptibility and symptom severity in an African cohort. *Progress in Neuro-Psychopharmacology & Biological Psychiatry*. 39(1), 163-9.
- Wright, J. (2014). Unravelling complexity. *Nature*. 508, S6-7.
- Wright, P., Nimgaonkar, V.L., Donaldson, P.T., and Murray, R.M. (2001). Schizophrenia and HLA: a review. *Schizophrenia Research*. 47(1), 1-12.

REFERENCES

- Xu, B., Ionita-Laza, I., Roos, J.L., Boone, B., Woodrick, S., Sun, Y., Levy, S., *et al.* (2012). *De novo* gene mutations highlight patterns of genetic and neural complexity in schizophrenia. *Nature Genetics*. *44*(12), 1365-1369.
- Xu, B., Woodroffe, A., Rodriguez-Murillo, L., Roos, J.L., van Rensburg, E.J., Abecasis, G.R., Gogos, J.A., *et al.* (2009). Elucidating the genetic architecture of familial schizophrenia using rare copy number variant and linkage scans. *Proceedings of the National Academy of Science*. *106*, 16746-16751.
- Xu, Q., Wu, X., Xiong, Y., Xing, Q., He, L., and Qin, S. (2013). Pharmacogenomics can improve antipsychotic treatment in schizophrenia. *Frontiers of Medicine*. *7*(2), 180-190.
- Yen, J., and Wilbraham, L. (2003). Discourses of culture and illness in South African mental health care and indigenous healing, part II: African mentality. *Transcultural Psychiatry*. *40*(4), 562-584.
- Yngvadottir, B., Macarthur, D.G., Jin, H. and Tyler-smith, C. (2009). The promise and reality of personal genomics. *Genome Biology*. *10*(237), 1–4.
- Yovel, G., Sirota, P., Mazeh, D., Shakhbar, G., Rosenne, E. and Ben-Eliyahu, S. (2000). Higher natural killer cell activity in schizophrenic patients: the impact of serum factors, medication, and smoking. *Brain, Behavior, and Immunity*. *14*, 153–169.
- Yu, N., Chen, F., Ota, S., Jorde, L.B., Pamilo, P., Patthy, L., Ramsay, M., *et al.* (2002). Larger genetic differences within Africans than between Africans and Eurasians. *Genetics*. *161*, 269-274.
- Yu, W., Gwinn, M., Clyne, M., Yesupriya, A., and Khoury, M.J. (2008). A navigator for human genome epidemiology. *Nature Genetics*. *40*(2), 124-125.
- Yue, W., Wang, H., Sun, L., Tang, F., Liu, Z., Zhang, H., Li, W., *et al.* (2011). Genome-wide association study identifies a susceptibility locus for schizophrenia in Han Chinese at 11p11.2. *Nature Genetics*. *43*(12), 1228-1231
- Zandi, P.P., and Judy, J.T. (2010). The promise and reality of pharmacogenetics in psychiatry. *Psychiatric Clinics of North America*. *33*(1), 181-224.

REFERENCES

Zhao, X., Tang, R., Gao, B., Shi, Y., Zhou, J., Guo, S., Zhang, J., *et al.* (2007). Functional variants in the promoter region of chitinase 3-like 1 (*CHI3L1*) and susceptibility to schizophrenia. *American Journal of Human Genetics*. 80, 12-18.

APPENDICES

APPENDIX 1: Conference presentations and Academic Exchanges

Conference Poster Presentations

- M. Coffee, N. Ishaque, B.I. Drögemöller, D.J.H. Niehaus, L. Warnich. 2013. “Linking gene regulation and schizophrenia susceptibility variants: insights into novel treatment strategies”. 17th World Congress of Basic & Clinical Pharmacology (WCP), Cape Town, South Africa.

Academic Exchanges

- August – November 2013: DAAD Award for a Short-term Research Scholarship (4 months) tenable at a German University (within the DAAD-NRF In-Country Scholarship Programme). Visit completed at the Division of Theoretical Bioinformatics, German Cancer Research Center (DKFZ), Heidelberg, Germany. Supervised by Dr. Naveed Ishaque.
- February – May 2014: A*STAR SIPGA Award for a 3 month exchange in Singapore. Completed at the Biomolecular Function Discovery Division, Bioinformatics Institute, Singapore. Supervised by Dr Birgit Eisenhaber.

APPENDIX 2: Scripts used for bioinformatics analysis*weave_fasta_for_sTRAP.pl***Written by Dr Naveed Ishaque**

```

# use strict;
my $usage = "This program weaves fasta files for sTRAP\n\n\t$0
[dbSNP bed file, 4th col = ref, 5th col = alt] [PRE FASTA] [POST
FASTA]\n\n";
my $nex = shift or die "Please provide SNP CSV file\n\n$usage";
my $pre = shift or die "Please provide PRE FASTA
file\n\n$usage";
my $post = shift or die "Please provide POST FASTA
file\n\n$usage";
open (NEX_F, "$nex") or die "Cannot open NEXUS FILE
'$nex'\n\n$usage";
open (PRE_F, "$pre") or die "Cannot open PRE FASTA FILE
'$nex'\n\n$usage";
open (POST_F, "$post") or die "Cannot open POST FASTA FILE
'$nex'\n\n$usage";

my %iupac=("B" => ["C","G","T"], "D" => ["A","G","T"], "K" =>
["G","T"], "M" => ["A","C"], "R" => ["A","G"], "S" => ["C","G"], "V"
=> ["A","C","G"], "W" => ["A","T"], "Y" => ["C","T"]);

while (<NEX_F>){if (/^(.*?)\t(.*?)\t(.*?)\t(.*?)\t(.*?)\t(.*?)$/){my
($snp, $ref, $alt) = ($4,$5,$6);
my $pre_f_line= <PRE_F>;
$pre_f_line= <PRE_F>;
my $post_f_line= <POST_F>;
$post_f_line= <POST_F>;
chomp ($pre_f_line);
chomp ($post_f_line);
if ($alt =~ m/A/ || $alt =~ m/C/ ||$alt =~ m/G/ ||$alt =~ m/T/ )
{
print ">$snp"."_WT\n$pre_f_line$ref$post_f_line\n";
print ">$snp"."_MUT\n$pre_f_line$alt$post_f_line\n";
}
else {
#my @bases=$iupac{$alt};
foreach my $base (@{$iupac{$alt}}){
# warn "$snp $alt $base\n";
print
">$snp"."_$alt$base"."_WT\n$pre_f_line$ref$post_f_line\n";
print
">$snp"."_$alt$base"."_MUT\n$pre_f_line$base$post_f_line\n";
} } }
else {
die "invalid line in SNP NEXUS file: $_\n";
}
}
close (NEX_F); close (PRE_F); close (POST_F);

```

APPENDICES

APPENDIX 3: Supplementary Material – Bioinformatics Results

Table S1. Most significant SNPs from previous GWAS

SNP RS NUMBER	CHR	CLOSEST GENE	INITIAL PHASE	REPLICATION PHASE	P-VALUE (COMBINED)	ODDS RATIO	REFERENCE
rs2523722	6	<i>MHC, TRIM26</i>	1606 European cases, 1794 European controls	13195 European cases, 31021 European controls	1.47E-16	1.25	Strange <i>et al.</i> , 2012
rs114002140	6	<i>HLA-DRB9</i>	5001 Swedish cases, 6243 Swedish controls	7,452 cases, 20,404 controls and 581 trios	9.14E-14	1.167	Ripke <i>et al.</i> , 2013
rs7085104	10	<i>C10orf32-AS3MT</i>	5002 Swedish cases, 6243 Swedish controls	7,452 cases, 20,404 controls and 581 trios	3.68E-13	1.11	Ripke <i>et al.</i> , 2013
rs6461049	7	<i>MAD1L1</i>	5003 Swedish cases, 6243 Swedish controls	7,452 cases, 20,404 controls and 581 trios	5.93E-13	1.107	Ripke <i>et al.</i> , 2013
rs6932590	6	<i>PRSS16</i>	2,663 European cases, 13,498 European controls	10,282 European cases, 21,093 European controls	1.40E-12	1.16	Stefansson <i>et al.</i> , 2009
rs1198588	1	<i>MIR137</i>	5004 Swedish cases, 6243 Swedish controls	7,452 cases, 20,404 controls and 581 trios	1.72E-12	0.889	Ripke <i>et al.</i> , 2013
rs2021722	6	<i>TRIM26</i>	9,394 European cases, 12,462 European controls	8,442 European cases, 21,397 European controls	2.18E-12	1.15	Ripke <i>et al.</i> , 2011
rs1006737	12	<i>CACNA1C</i>	5005 Swedish cases, 6243 Swedish controls	7,452 cases, 20,404 controls and 581 trios	5.22E-12	1.103	Ripke <i>et al.</i> , 2013
rs1635	6	<i>NKAPL</i>	746 Han Chinese cases, 1,599 Han Chinese controls	4,027 Han Chinese cases, 5,603 Han Chinese controls	6.91E-12	0.78	Yue <i>et al.</i> , 2011
rs11038167	11	<i>TSPAN18</i>	746 Han Chinese cases, 1,599 Han Chinese controls	4,027 Han Chinese cases, 5,603 Han Chinese controls	1.09E-11	1.29	Yue <i>et al.</i> , 2011
rs1625579	1	<i>MIR137</i>	9,394 European cases, 12,462 European controls	8,442 European cases, 21,397 European controls	1.59E-11	1.12	Ripke <i>et al.</i> , 2011
rs835784	11	<i>TSPAN18</i>	746 Han Chinese cases, 1,599 Han Chinese	4,027 Han Chinese cases, 5,603 Han Chinese	2.73E-11	1.27	Yue <i>et al.</i> , 2011

APPENDICES

			controls	controls			
rs17693963	6	<i>MHC</i>	2111 European cases, 2535 European controls	11271 cases, 14601 controls	3.08E-11	1.24	Bergen <i>et al.</i> , 2012
rs13211507	6	<i>PGBD1</i>	2,663 European cases, 13,498 European controls	10,282 European cases, 21,093 European controls	3.80E-11	1.24	Stefansson <i>et al.</i> , 2009
rs1233710	6	<i>ZKSCAN4</i>	746 Han Chinese cases, 1,599 Han Chinese controls	4,027 Han Chinese cases, 5,603 Han Chinese controls	4.76E-11	0.79	Yue <i>et al.</i> , 2011
rs16887244	8	<i>LSM1</i>	3,750 Han Chinese cases, 6,468 Han Chinese controls	4,383 Han Chinese cases, 4,539 Han Chinese controls	1.27E-10	0.83	Shi <i>et al.</i> , 2011
rs17691888	10	<i>CACNB2</i>	5006 Swedish cases, 6243 Swedish controls	7,452 cases, 20,404 controls and 581 trios	1.27E-10	0.862	Ripke <i>et al.</i> , 2013
rs13219354	6	<i>PRSS16</i>	2,663 European cases, 13,498 European controls	10,282 European cases, 21,093 European controls	1.30E-10	1.2	Stefansson <i>et al.</i> , 2009
rs4129585	8	<i>TSNARE1</i>	5007 Swedish cases, 6243 Swedish controls	7,452 cases, 20,404 controls and 581 trios	2.19E-10	1.091	Ripke <i>et al.</i> , 2013
rs3131296	6	<i>HLA-E, NOTCH4</i>	2,663 European cases, 13,498 European controls	10,282 European cases, 21,093 European controls	2.30E-10	1.19	Stefansson <i>et al.</i> , 2009
rs12966547	18	<i>CCDC68</i>	9394 European cases, 12462 European controls	8442 European cases, 21397 European controls	2.60E-10	1.09	Ripke <i>et al.</i> , 2011
rs10789369	1	<i>x10NST00000415686.1</i>	5008 Swedish cases, 6243 Swedish controls	7,452 cases, 20,404 controls and 581 trios	3.64E-10	1.095	Ripke <i>et al.</i> , 2013
rs2142731	6	<i>PGBD1</i>	746 Han Chinese cases, 1,599 Han Chinese controls	4,027 Han Chinese cases, 5,603 Han Chinese controls	5.14E-10	0.79	Yue <i>et al.</i> , 2011
rs204999	6	<i>MHC</i>	1606 European cases, 1794 European controls	13195 European cases, 31021 European controls	5.37E-10	1.33	Strange <i>et al.</i> , 2012
rs11038172	11	<i>TSPAN18</i>	746 Han Chinese cases, 1,599 Han Chinese controls	4,027 Han Chinese cases, 5,603 Han Chinese controls	7.21E-10	1.25	Yue <i>et al.</i> , 2011

APPENDICES

rs6913660	6	<i>HIST1H2BJ</i>	2,663 European cases, 13,498 European controls	10,282 European cases, 21,093 European controls	1.10E-09	1.15	Stefansson <i>et al.</i> , 2009
rs12666575	7	<i>MAD1L1</i>	2111 European cases, 2535 European controls	11271 cases, 14601 controls	1.75E-09	0.89	Bergen <i>et al.</i> , 2012
rs7914558	10	<i>CNNM2</i>	9394 European cases, 12462 European controls	8442 European cases, 21397 European controls	1.82E-09	1.1	Ripke <i>et al.</i> , 2011
rs7940866	11	<i>SNX19</i>	5009 Swedish cases, 6243 Swedish controls	7,452 cases, 20,404 controls and 581 trios	1.83E-09	0.921	Ripke <i>et al.</i> , 2013
rs2312147	2	<i>VRK2</i>	7946 European cases, 19 036 European controls	9246 European cases, 22 356 European controls	1.90E-09	1.09	Steinberg <i>et al.</i> , 2011
rs12807809	11	<i>NRGN</i>	2,663 European cases, 13,498 European controls	10,282 European cases, 21,093 European controls	2.40E-09	1.15	Stefansson <i>et al.</i> , 2009
rs17504622	5	<i>ENST00000503048.1</i>	5010 Swedish cases, 6243 Swedish controls	7,452 cases, 20,404 controls and 581 trios	2.65E-09	1.238	Ripke <i>et al.</i> , 2013
rs2905424	19	<i>MAU2</i>	5011 Swedish cases, 6243 Swedish controls	7,452 cases, 20,404 controls and 581 trios	3.44E-09	1.092	Ripke <i>et al.</i> , 2013
rs11819869	11	<i>AMBRA1</i>	1169 European cases, European controls 3714	2569 European cases, 4088 European controls	3.89E-09	1.25	Rietschel <i>et al.</i> , 2012
rs9960767	18	<i>TCF4</i>	2,663 European cases, 13,498 European controls	10,282 European cases, 21,093 European controls	4.10E-09	1.23	Stefansson <i>et al.</i> , 2009
rs1488935	8	<i>WHSC1L1</i>	3,750 Han Chinese cases, 6,468 Han Chinese controls	4,383 Han Chinese cases, 4,539 Han Chinese controls	5.06E-09	0.87	Shi <i>et al.</i> , 2011
rs7527939	1	<i>HHAT</i>	188 European cases, 376 European controls	99 European cases, 328 European controls	6.49E-09	2.63	Betcheva <i>et al.</i> , 2013
rs2373000	2	<i>QPCT</i>	5012 Swedish cases, 6243 Swedish controls	7,452 cases, 20,404 controls and 581 trios	6.78E-09	1.087	Ripke <i>et al.</i> , 2013
rs4309482	18	<i>TCF4</i>	7946 European cases, 19 036 European controls	9246 European cases, 22 356 European controls	7.80E-09	1.09	Steinberg <i>et al.</i> , 2011

APPENDICES

rs6878284	5	<i>SLCO6A1</i>	5013 Swedish cases, 6243 Swedish controls	7,452 cases, 20,404 controls and 581 trios	9.03E-09	0.92	Ripke <i>et al.</i> , 2013
rs10489202	1	<i>BRP44</i>	3750 Han Chinese cases, 6468 Han Chinese controls	4383 Han Chinese cases, 4539 Han Chinese controls	9.50E-09	1.19	Shi <i>et al.</i> , 2011
rs13194053	6	<i>HIST1H2BJ</i>	2,681 European cases, 2,653 European controls, 1,286 African American cases, 973 African American controls	5,327 European cases, 16,424 European controls	9.54E-09	0.88	Shi <i>et al.</i> , 2009
rs11191580	10	<i>NT5C2</i>	9,394 European cases, 12,462 European controls	8,442 European cases, 21,397 European controls	1.11E-08	1.15	Ripke <i>et al.</i> , 2011
rs4687552	3	<i>ITIH3</i>	5014 Swedish cases, 6243 Swedish controls	7,452 cases, 20,404 controls and 581 trios	1.16E-08	1.086	Ripke <i>et al.</i> , 2013
rs12991836	2	<i>ZEB2</i>	5015 Swedish cases, 6243 Swedish controls	7,452 cases, 20,404 controls and 581 trios	1.19E-08	0.922	Ripke <i>et al.</i> , 2013
rs2949006	2	<i>FONG</i>	5016 Swedish cases, 6243 Swedish controls	7,452 cases, 20,404 controls and 581 trios	1.21E-08	1.102	Ripke <i>et al.</i> , 2013
rs4801131	18	<i>ENST00000565991.1</i>	5017 Swedish cases, 6243 Swedish controls	7,452 cases, 20,404 controls and 581 trios	1.22E-08	0.925	Ripke <i>et al.</i> , 2013
rs7004633	8	<i>MMP16</i>	9,394 European cases, 12,462 European controls	8,442 European cases, 21,397 European controls	1.45E-08	1.16	Ripke <i>et al.</i> , 2011
rs778371	2	<i>C2orf82</i>	5018 Swedish cases, 6243 Swedish controls	7,452 cases, 20,404 controls and 581 trios	1.51E-08	0.92	Ripke <i>et al.</i> , 2013
rs14403	1	<i>AKT3</i>	5019 Swedish cases, 6243 Swedish controls	7,452 cases, 20,404 controls and 581 trios	1.80E-08	0.91	Ripke <i>et al.</i> , 2013
rs11532322	12	<i>C12orf65</i>	5020 Swedish cases, 6243 Swedish controls	7,452 cases, 20,404 controls and 581 trios	2.28E-08	1.094	Ripke <i>et al.</i> , 2013
rs17512836	18	<i>TCF4</i>	9,394 European cases, 12,462 European controls	8,442 European cases, 21,397 European controls	2.35E-08	1.4	Ripke <i>et al.</i> , 2011
rs1538774	1	<i>SDCCAG8</i>	5021 Swedish cases,	7,452 cases, 20,404	2.53E-08	0.917	Ripke <i>et al.</i> , 2013

APPENDICES

rs11098403	4	<i>NDST3</i>	6243 Swedish controls 904 Ashkenazi Jewish cases, 1640 Ashkenazi Jewish controls	controls and 581 trios 10200 mixed population cases, 12991 mixed population controls	2.67E-08	1.15	Lencz <i>et al.</i> , 2013
rs548181	11	<i>STT3A</i>	9,394 European cases, 12,462 European controls	8,442 European cases, 21,397 European controls	2.91E-08	1.2	Ripke <i>et al.</i> , 2011
rs11995572	8	<i>Intergenic</i>	5022 Swedish cases, 6243 Swedish controls	7,452 cases, 20,404 controls and 581 trios	3.33E-08	1.12	Ripke <i>et al.</i> , 2013
rs171748	5	<i>ENST00000506902.1</i>	5023 Swedish cases, 6243 Swedish controls	7,452 cases, 20,404 controls and 581 trios	3.78E-08	1.078	Ripke <i>et al.</i> , 2013
rs7709645	5	<i>ZSWIM6</i>	2111 European cases, 2535 European controls	11271 cases, 14601 controls	3.80E-08	1.11	Bergen <i>et al.</i> , 2012
rs3800316	6	<i>unknown</i>	2,681 European cases, 2,653 European controls, 1,286 African American cases, 973 African American controls	5,327 European cases, 16,424 European controls	3.81E-08	0.86	Shi <i>et al.</i> , 2009
rs2910032	5	-	5024 Swedish cases, 6243 Swedish controls	7,452 cases, 20,404 controls and 581 trios	4.12E-08	0.925	Ripke <i>et al.</i> , 2013
rs10503253	8	<i>CSMD1</i>	9,394 European cases, 12,462 European controls	8,442 European cases, 21,397 European controls	4.14E-08	1.11	Ripke <i>et al.</i> , 2011
rs3800307	6	<i>unknown</i>	2,681 European cases, 2,653 European controls, 1,286 African American cases, 973 African American controls	5,327 European cases, 16,424 European controls	4.35E-08	0.89	Shi <i>et al.</i> , 2009
rs886424	6	<i>MHC</i>	2111 European cases, 2535 European controls	11271 cases, 14601 controls	4.54E-08	0.71	Bergen <i>et al.</i> , 2012
rs17662626	2	<i>PCGEM1</i>	9,394 European cases, 12,462 European controls	8,442 European cases, 21,397 European controls	4.65E-08	1.2	Ripke <i>et al.</i> , 2011

APPENDICES

Table S2. SNPs in regulatory regions (with a RegulomeDB score between 1a and 3b)

SNP	Score	SNP	Score	SNP	Score	SNP	Score
rs200981	1a	rs2916074	1f	rs2297787	2a	rs1233583	2b
rs886424	1b	rs2523722	1f	rs200949	2a	rs4129585	2c
rs4808967	1b	rs2523719	1f	rs938575	2b	rs926300	3a
rs4808203	1b	rs2517612	1f	rs9324383	2b	rs7709645	3a
rs2535629	1b	rs2517611	1f	rs876701	2b	rs718885	3a
rs200485	1b	rs2297786	1f	rs7921574	2b	rs71559051	3a
rs16887343	1b	rs2071278	1f	rs761171	2b	rs7067970	3a
rs3094071	1d	rs200995	1f	rs7527939	2b	rs67457459	3a
rs200483	1d	rs200954	1f	rs67340775	2b	rs59498392	3a
rs1059612	1d	rs200948	1f	rs61882678	2b	rs56401801	3a
rs9633712	1e	rs200484	1f	rs60935759	2b	rs13197175	3a
rs943037	1f	rs1926032	1f	rs55703821	2b	rs55657382	3a
rs9262143	1f	rs17115213	1f	rs4537271	2b	rs4656564	3a
rs9262141	1f	rs16887244	1f	rs433061	2b	rs35909544	3a
rs911186	1f	rs1264361	1f	rs425335	2b	rs35744819	3a
rs886422	1f	rs1264350	1f	rs401754	2b	rs35525740	3a
rs7745603	1f	rs1264322	1f	rs3800913	2b	rs34218844	3a
rs736408	1f	rs1264304	1f	rs3779921	2b	rs3094072	3a
rs7125907	1f	rs12413046	1f	rs3765879	2b	rs3094070	3a
rs6996860	1f	rs12412038	1f	rs3765872	2b	rs28681082	3a
rs6913660	1f	rs11777067	1f	rs3740400	2b	rs28360499	3a
rs6904596	1f	rs4428528	1f	rs35202262	2b	rs2747054	3a
rs4808964	1f	rs1150752	1f	rs34765154	2b	rs2717007	3a
rs4808200	1f	rs7766843	1f	rs34064842	2b	rs2678903	3a
rs4409766	1f	rs4434496	1f	rs3129984	2b	rs2142731	3a
rs3794993	1f	rs11191582	1f	rs3129697	2b	rs200489	3a
rs3781285	1f	rs11191580	1f	rs3095338	2b	rs1371833	3a
rs3134942	1f	rs11191558	1f	rs3006923	2b	rs13219181	3a
rs3132658	1f	rs11191557	1f	rs2905435	2b	rs13199906	3a
rs3131296	1f	rs11191548	1f	rs2270376	2b	rs13197633	3a
rs3131064	1f	rs11191515	1f	rs1846416	2b	rs13197176	3a
rs3131060	1f	rs11191514	1f	rs13262595	2b	rs12272795	3a
rs3130673	1f	rs11191479	1f	rs13212318	2b	rs11783967	3a
rs3130403	1f	rs10883832	1f	rs13199772	2b	rs1233593	3a
rs3129985	1f	rs10883808	1f	rs13195636	2b	rs10786729	3a
rs3129702	1f	rs10748839	1f	rs12221064	2b	rs1060041	3a
rs3129701	1f	rs10748836	1f	rs12219901	2b	rs10100894	3a
rs3095336	1f	rs1049633	1f	rs4280993	2b		
rs3094629	1f	rs4713074	2a	rs1742743	2b		
rs2965189	1f	rs2916068	2a	rs114633780	2b		

APPENDICES

Table S3. The top 10 features identified by network analysis with GeneMANIA that fit the query genes.

Feature	FDR	Genes in network	Genes in genome
ER to Golgi transport vesicle membrane	8.38E-42	19	31
ER to Golgi transport vesicle	5.13E-41	19	35
luminal side of membrane	5.13E-41	18	27
integral component of luminal side of endoplasmic reticulum membrane	5.13E-41	18	27
luminal side of endoplasmic reticulum membrane	5.13E-41	18	27
transport vesicle membrane	1.83E-37	19	49
antigen processing and presentation of exogenous peptide antigen	2.89E-37	24	166
antigen processing and presentation of exogenous antigen	5.41E-37	24	171
antigen processing and presentation of peptide antigen	1.34E-36	24	178
antigen processing and presentation	1.64E-34	24	216

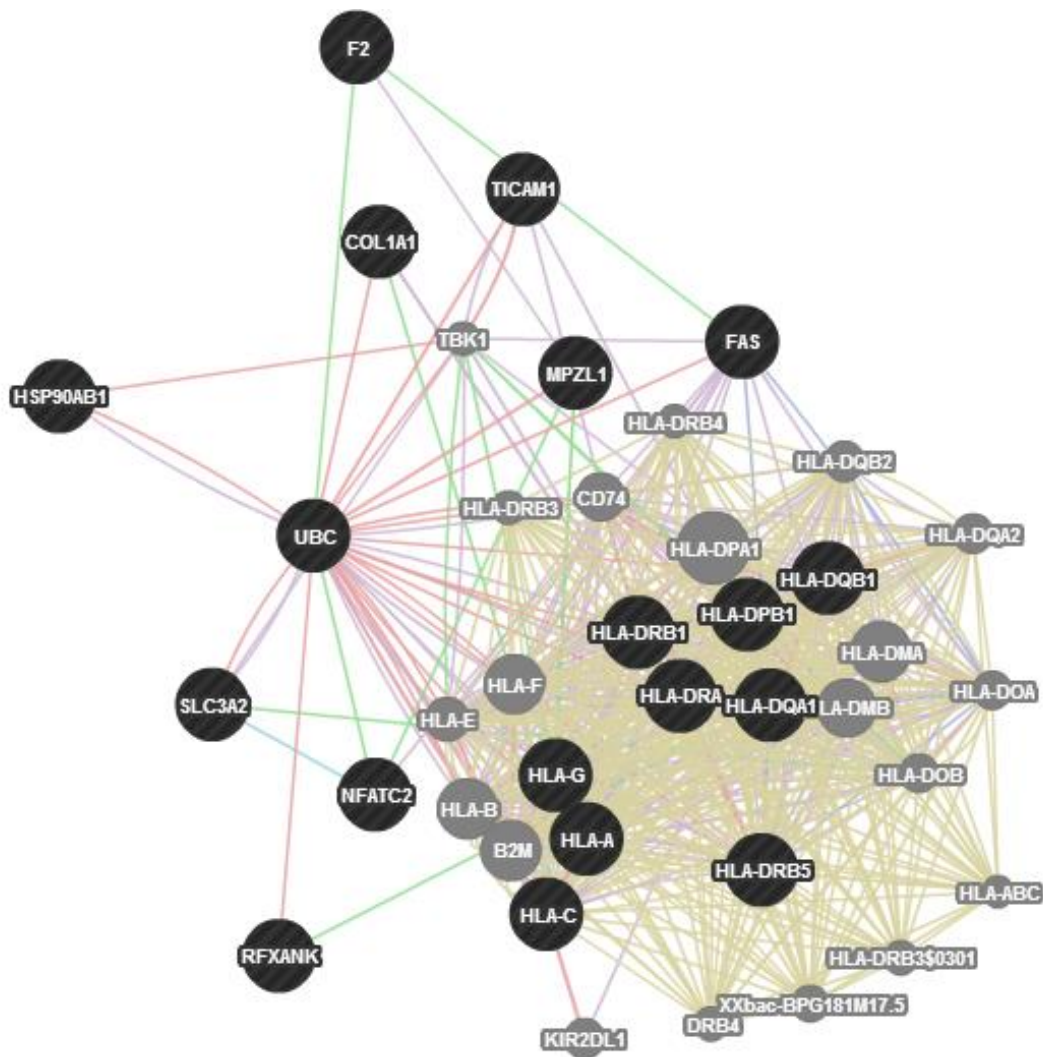


Figure S1. Full network analysis using GeneMANIA. Query genes (which all mapped to specific pathways using DAVID) are indicated with black nodes. Additional genes (identified by GeneMANIA), which are related to the query genes and fit to the implicated network, are indicated with grey nodes.

APPENDICES

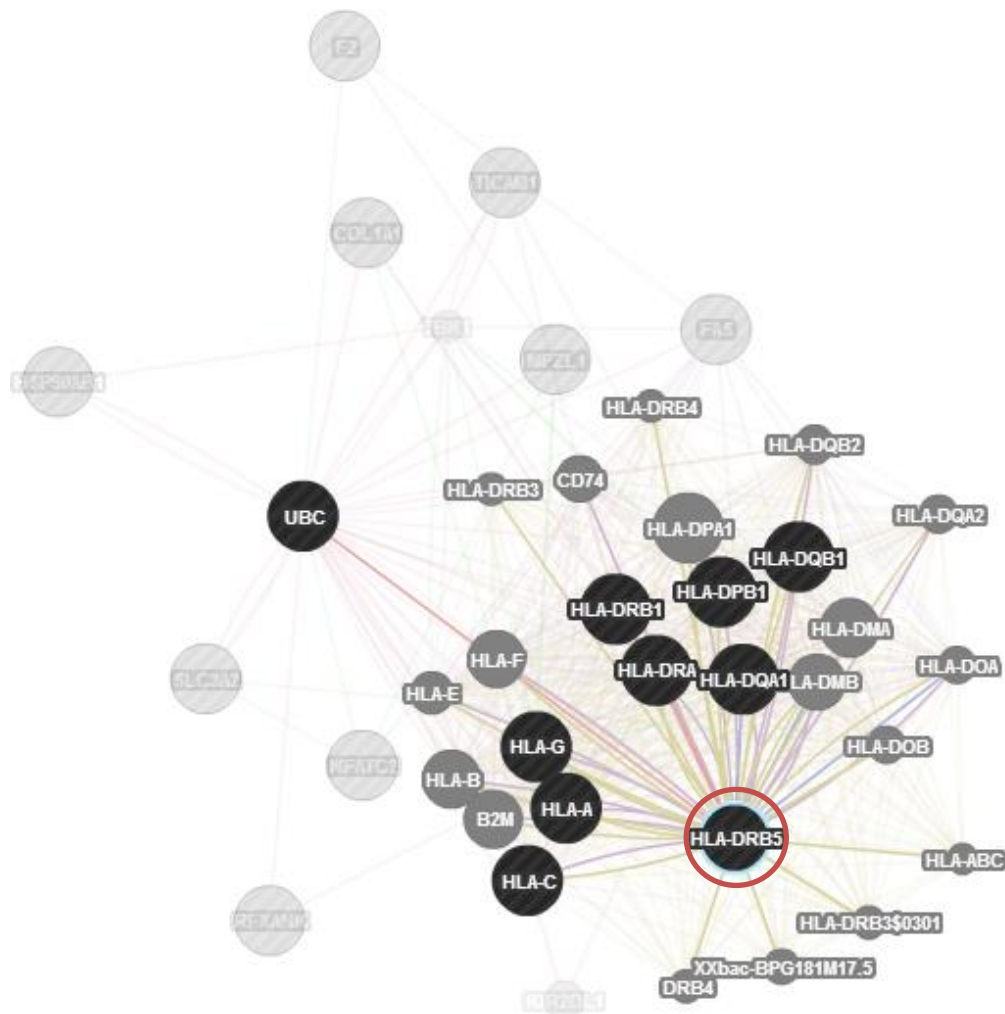


Figure S2. The gene that had the highest score (80.32) in terms of associations with other genes, was *HLA-DRB5* (circled in red). The associated genes are highlighted in grey and black.

APPENDIX 4: Supplementary Material – Association Study Results

Comparison of minor allele frequencies between South African Xhosa cohort and YRI, CEU and CHB populations:

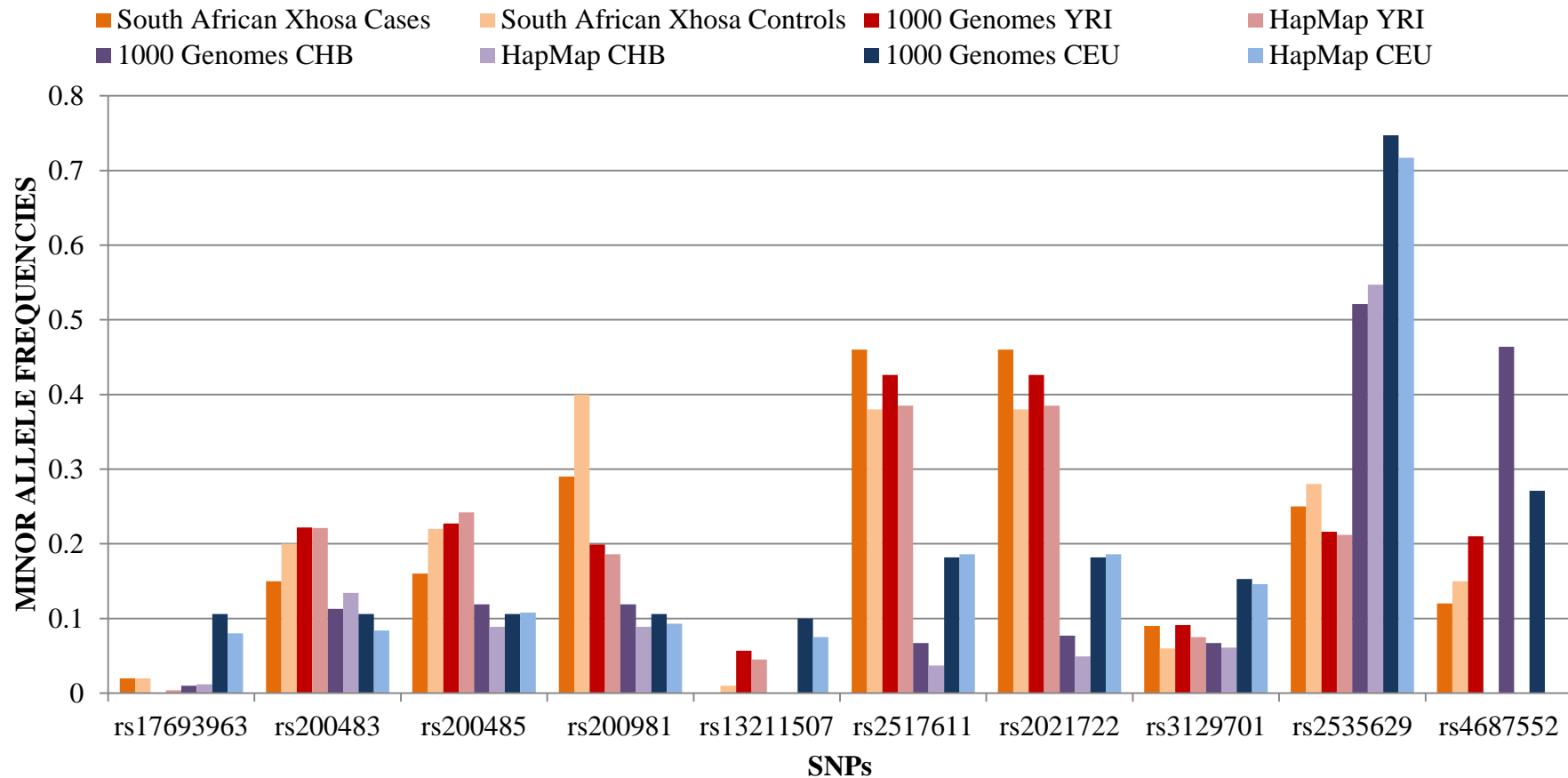


Figure S3. Comparison of minor allele frequencies between the genotyped SNPs in the South African Xhosa cohort and the YRI, CEU and CHB populations according to 1000 Genomes and HapMap. Where no bars are indicated the MAF was 0 or no data was available.

APPENDICES

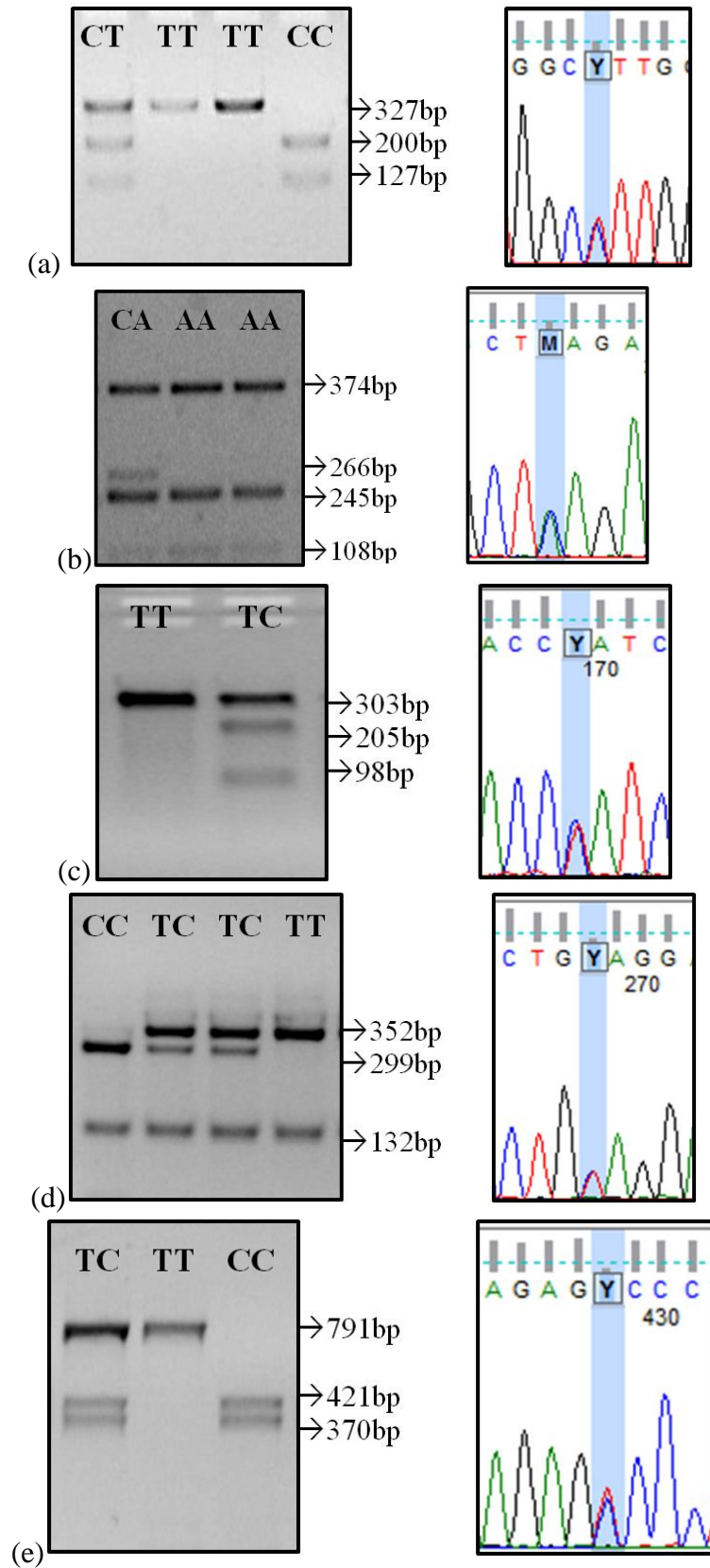


Figure S4. RFLP gel photos and corresponding chromatograms for each SNP: (a) rs200981 (b) rs17693963 (c) rs13211507 (d) rs2535629 and (e) rs4687552.

APPENDICES

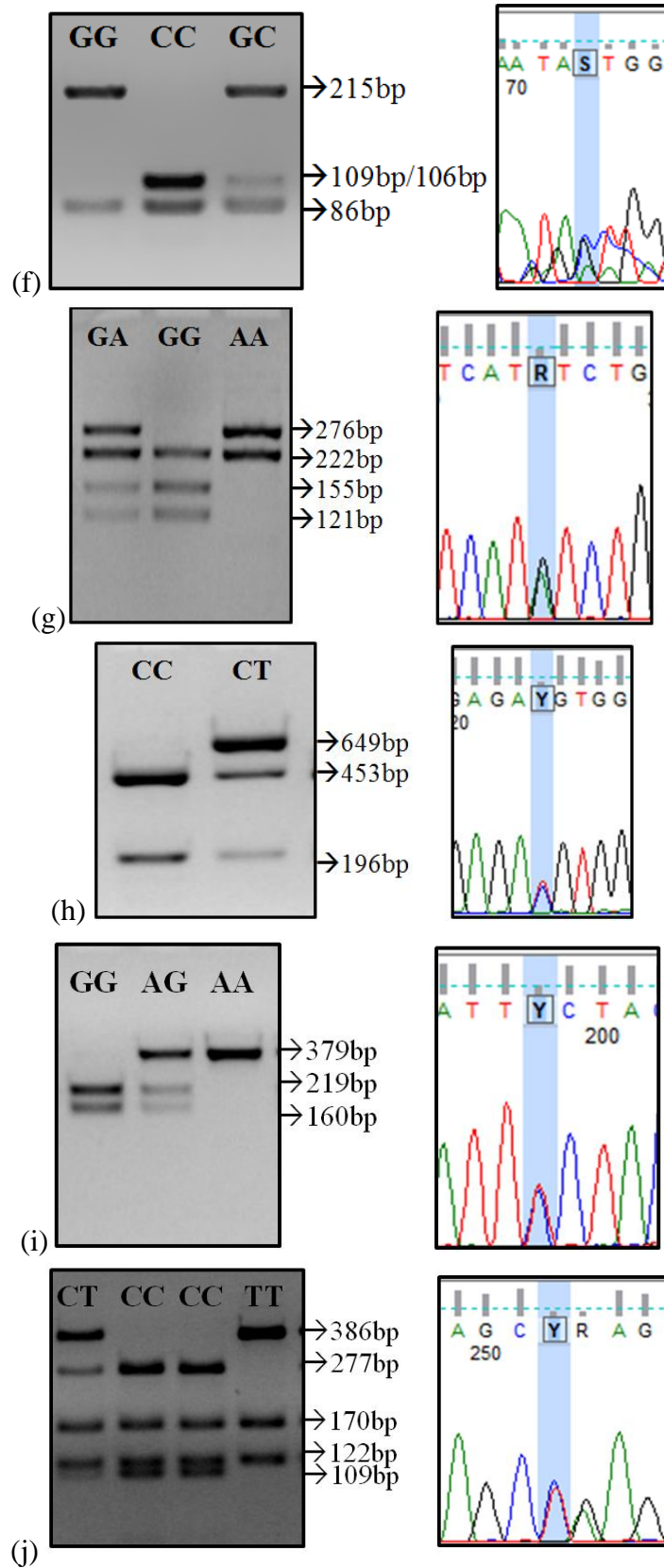


Figure S4 (cont.). RFLP gel photos and corresponding chromatograms for each SNP: (f) rs200485 (g) rs200483 (h) rs3129701 (i) rs2517611 and (j) rs2021722.

APPENDICES

Table S4. Results of single SNP analysis of genotyped SNPs with formal thought disorder.

Outcome	SNP	Inheritance Model	Comparison	<i>p</i> -value	Effect	95% CI	
Formal Thought Disorder	rs200981	Heterozygous	CT versus CC+TT	0.0447	-0.59	-1.13	-0.05

Table S5. Reagents for RE digest of the PCR products (for RFLP analysis). (In all instances the buffer was 1X, total volume was 20 µl and dH₂O was added)

SNP	Reagents		
	PCR product (µl)	Enzyme Amount (µl)	BSA
rs200981	5	0.9	-
rs17693963	5	0.5	-
rs13211507	5	0.3	0.2
rs2535629	5	0.9	-
rs4687552	5	0.8	-
rs200485	5	0.8	-
rs200483	5	0.9	-
rs3129701	5	0.6	-
rs2517611	5	0.9	-
rs2021722	5	0.9	-

APPENDIX 5: Protocols, Reagents And Solutions**SureClean Quick-Clean Protocol (Bioline)**

- Add 1X volume of Quick-Clean to the PCR product .
- Vortex.
- Incubate at room temperature for 10 minutes.
- Centrifuge at 13000 rpm for 10 minutes.
- Discard supernatant.
- Add 100 µl 70% ethanol.
- Vortex for 30 seconds.
- Centrifuge at 13000 rpm for 10 minutes.
- Discard supernatant.
- Air-dry.
- Resuspend pellet in 10 µl water.

Big Dye v3.1 Sequencing Chemistry (Applied Biosystems™)

- Dilute PCR product according to the size of the template to be sequenced (Table S6).
- In a 0.2 ml PCR tube, add 3 µl diluted PCR product, 3 µl diluted primer (3.3 pmol), 1.3 µl Big Dye reaction mix and 2.7 µl Half Dye mix.
- Place the PCR tubes in a thermocycler with the following conditions:
 - 94°C – 5 min
 - (94°C – 10 sec, 55°C – 10 sec, 60°C – 4 min) → 25X
 - 4°C – ∞
- Add 1 µl 2.2% SDS to each tube and place the tubes in the thermocycler with the following conditions:
 - 98°C = 5 min
 - 25°C = 10 min

Table S6. Concentration of PCR product required according to template size being sequenced.

Template	Quantity
PCR product:	
100 – 200 bp	1 – 3 ng
200 – 500 bp	3 – 10 ng
500 – 1 000 bp	5 – 20 ng
1 000 – 2 000 bp	10 – 40 ng
>2 000 bp	20 – 50 ng

APPENDICES

SB BUFFER**20X Stock (2 L)**

- 90 g Boric Acid (61.83 g/mol)
- 16 g NaOH (40.00 g/mol)
- BTV with dH₂O

1X SB

- 100 ml 20X SB
- 1900 ml dH₂O

CRESOL LOADING DYE**Stock**

- 10 mg Cresol powder
- 1 ml dH₂O

Loading Dye

- 3.4 g Sucrose
- 200 µl Cresol Stock
- 9.5 ml dH₂O

GELS**1.5% w/v Gel**

- 1.5 g Agarose powder
- 100 ml 1X SB
- 5 µl Ethidium Bromide

2% w/v Gel

- 2 g Agarose powder
- 100 ml 1X SB
- 5 µl Ethidium Bromide

3% w/v Gel

- 3 g Agarose powder
- 100 ml 1X SB
- 5 µl Ethidium Bromide