Genome and Transcriptome Sequencing of *Vitis vinifera* cv Pinotage

Beatrix Coetzee

Dissertation presented for the degree of

Doctor of Philosophy

in Science

at

Stellenbosch University

Department of Genetics, Faculty of Science

Supervisor: Prof Johan T Burger

Co-supervisor: Dr Hans J Maree

March 2018

Declaration

By submitting this 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.

Beatrix Coetzee

March 2018

Summary

Examining the genetic basis of natural phenotypic variation, and the transfer of this knowledge to a breeding program for improved crop cultivars or livestock races, is a major goal for biological sciences. As grapevine (*Vitis vinifera*) is one of the most important crop plants in the world, research into its genetics is imperatave, both in terms of sustainable food production and the vast economic impact of the wine industry. Grapevine displays a great level of intra-species phenotypic diversity in viticultural and oenological traits, between cultivars. Understanding this genetic diversity is an important step towards developing improved grapevine cultivars, but also the conservation of the important traditional cultivars.

Vitis vinifera cv Pinotage is an artificial Pinot noir/ Cinsaut cross, created with the South African climate and growing conditions in mind. Today it is a commercial cultivar, used for the production of premium wines, deeply rooted in the South African wine culture and history. This study focused on the next-generation sequencing and bioinformatic analysis of the Pinotage genome and transcriptome.

A *de novo* assembly strategy was followed to produce the first Pinotage draft genome sequence. Sequencing read data were also aligned to the available reference Pinot noir genome, and from this alignment the Pinotage/ Pinot noir variant density, determined. This was followed by a more in-depth focus on a number of functional gene clusters with more than 50% of their genes influenced by these variants.

Furthermore, this is the first research to lend scientific support to the current wine trend of exclusive, superior wines produced from old vineyards. These old-vine wines are assumed to have a deeper character and more flavour. To explore the role of genetics and differential gene expression in this phenomenon, RNA-seq data were used to survey and compare the leaf and berry transcriptomes of young and old Pinotage vines, at harvest. Differential gene expression between young and old vines was studied, and the involvement of these genes in fruit ripening, discussed. A general trend towards delayed ripening in older vines was observed. This suggests that the berries remain attached to the vine for a longer period, thereby allowing more time for flavour compounds to accumulate.

In the final part of the study, the Pinotage genome and transcriptome data were combined to identify Pinotage genes present in neither the reference Pinot noir PN40024 nor ENTAV115.

These genes were classified as both structural and regulatory genes and it was shown that genes involved in the stress response network are a major gene class contributing to the genetic differences between Pinotage and Pinot noir. A plant species is constantly challenged by various biotic and abiotic stresses and it is an evolutionary investment to diversify genes responsible for stress response, to be able to efficiently overcome these stresses. The information generated in this study will aid in grapevine breeding programs for sustainable production of high quality wine in a changing environment.

Opsomming

Die ondersoek na die genetiese basis van natuurlike fenotipiese variasie, en die oordrag van hierdie kennis na 'n teelprogram vir verbeterde gewaskultivars of vee-rasse, is 'n belangrike doelwit vir biologiese wetenskappe. Aangesien wingerd (*Vitis vinifera*) een van die belangrikste gewasplante ter wêreld is, is navorsing in sy genetika noodsaaklik, beide in terme van volhoubare voedselproduksie en die wye ekonomiese impak van die wynbedryf. Wingerd vertoon 'n groot vlak van fenotipiese diversiteit in die spesie, in wingerd- en wynboukundige eienskappe, tussen kultivars. Om hierdie genetiese diversiteit te verstaan, is 'n belangrike stap in die ontwikkeling van verbeterde wingerdkultivars, maar ook die bewaring van die belangrike tradisionele kultivars.

Vitis vinifera kultivar Pinotage is 'n kunsmatige Pinot noir/ Cinsaut kruising, geskep met die Suid-Afrikaanse klimaat en groeitoestande in gedagte. Vandag is dit 'n kommersiële kultivar, wat gebruik word vir die produksie van gehalte wyne, diep gewortel in die Suid-Afrikaanse wynkultuur en -geskiedenis. Hierdie studie het gefokus op die volgende-generasie-volgordebepaling en bioinformatiese analise van die Pinotage-genoom en transkriptoom. 'n *De novo*-samestellingstrategie is gevolg om die eerste Pinotage konsep-genoomvolgorde te produseer. Opvolgingsleesdata is ook in lyn gebring met die beskikbare verwysings Pinot noir-genoom en vanaf hierdie belyning is die Pinotage/ Pinot noir-variantdigtheid bepaal, gevolg deur 'n meer in-diepte fokus op 'n aantal funksionele geen-groepe met meer as 50% van hul gene beïnvloed deur hierdie variante.

Verder is dit die eerste navorsing wat wetenskaplike ondersteuning verleen aan die huidige wyn-tendens van eksklusiewe, uitstekende wyne geproduseer van ou wingerde. Hierdie ou-wingerdwyne word veronderstel om 'n dieper karakter en meer geur te hê. Om die rol van genetika en differensiële geenuitdrukking in hierdie verskynsel te ondersoek, is RNS-opeenvolgings-data gebruik om die blaar- en korrel transkriptome van jong en ou Pinotage-wingerdstokke, tydens oestyd, te ondersoek en te vergelyk. Differensiële geenuitdrukking tussen jong en ou wingerdstokke is bestudeer, en die betrokkenheid van hierdie gene in rypwording word bespreek. 'n Algemene neiging tot vertraagde rypwording in ouer wingerde is waargeneem. Dit dui daarop dat die korrels vir 'n langer tydperk aan die wingerdstok bly, en dat meer geurverbindings in die korrels kan versamel.

In die laaste gedeelte van die studie is die Pinotage-genoom en transkriptoomdata gekombineer om Pinotage variëteit-spesifieke gene te identifiseer, wat nie in die verwysing genoom Pinot noir PN40024 of ENTAV115 voorkom nie. Hierdie gene is geklassifiseer as beide strukturele en regulatoriese gene en dit is gewys dat gene wat betrokke is by die stresresponsnetwerk, 'n belangrike geenklas is wat bydra tot die genetiese verskille tussen Pinotage en Pinot noir. 'n Plantspesie word voortdurend uitgedaag deur verskeie biotiese en abiotiese stres en dit is 'n evolusionêre belegging om gene wat verantwoordelik is vir stresrespons te diversifiseer, om hierdie stres doeltreffend te oorkom. Die inligting wat in hierdie studie gegenereer is, sal van nut wees in wingerdbouprogramme vir die volhoubare produksie van hoë kwaliteit wyn in 'n veranderende omgewing.

Acknowledgements

I would like to express my sincere gratitude to the following people and institutions, without whom this study would not have been possible:

- Prof Johan Burger for allowing me the opportunity to be part of his research group and for all the guidance and intellectual input, during all of my postgraduate studies, but especially this study.
- Dr Hano Maree for guidance and mentorship during all of my postgraduate studies, being my host at the ARC, and intellectual input into this study.
- Drs Marike Visser and Rachelle Bester for their intellectual input into this project, many hours of critical discussions, support and friendship.
- Dr Caroli Diener for critical reading of this thesis, and endless encouragement, support and friendship throughout my studies.
- Ms Kristin Oosthuizen for the many hours we worked together, friendship and support, and proofreading of this thesis.
- Prof Delledonne's group (University of Verona, Italy), especially Dr Andrea Minio for hosting me, and assisting with the genome and transcriptome assembly.
- All the colleagues of the Vitis lab with whom I worked with during my postgraduate studies.
- Grondves and Kanonkop wine estate for allowing us to collected sample material.
- Charles Visser (Vititec), Koos du Toit (Kanonkop) and Grondves staff for assisting with sample collection.
- All computations were performed using the University of Stellenbosch Central Analytical Facilities' HPC2: http://www.sun.ac.za/hpc
- National Research Foundation (NRF) and Department of Science and Technology (DST) Professional Development Program (PDP) for personal funding.
- Stellenbosch University for personal funding.
- Agricultural Research Council for hosting me as a PDP student.
- Pinotage Association for funding this project.
- My friends and family for your love, support and encouragement.

Thank you.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

3’UTR	Three prime untranslated region
5’-end	Five prime end
5’UTR	Five prime untranslated region
aa	Amino acid
ABA	Abscisic acid
ABF2	ABA binding factor 2
ACC	1-aminocyclopropane-1-carboxylate
ACO	1-aminocyclopropane-1-carboxylate oxidase
ACS	1-aminocyclopropane-1-carboxylate synthase
Avr	avirulence genes
BLAST	Basic local alignment search tool
BR	Brassinosteroid
Cas	CRISPR associated protein
CC	Coiled-coil
cDNA	complementary deoxyribonucleic acid
CNV	Copy number variation
CPC	Coding Potential Calculator
CREs	*Cis*-regulatory elements
CRIBI	Centro di Ricerca Interdipartimentale per le Biotecnologie Innovative
CRISPR	Clustered regularly-interspaced short palindromic repeats
CRKs	Cysteine-rich Receptor-like Kinases
cv	cultivar
DEGs	Differentially expressed genes
DIN	DNA integrity number
DNA	Deoxyribonucleic acid
DNase	Deoxyribonuclease
dsDNA	double-stranded deoxyribonucleic acid
DUF26	Domain of Unknown Function 26 genes
dUTP	2´-Deoxyuridine, 5´-Triphosphate
e-value	expect value
EC	Enzyme commission
ENTAV	Etablissement National Technique pour l’Amolioration de la Viticulture, France
ERFs	Ethylene responsive transcription factors
ETI	Effector-triggered immunity
FAIR	Findable, Accessible, Interoperable, Reusable
FAO	Food and Agriculture Organization of the United Nations
FPKM	Fragments Per Kilobase of transcript per Million mapped reads
GARM	Genome Assembler, Reconciliation and Merging
gDNA	Genomic deoxyribonucleic acid
GO	Gene Ontology

GrapeIS	Grape Information System
grip22	Grapevine ripening induced protein 22
HR	Hypersensitive response
Indels	Insertion and/or deletion
KEGG	Kyoto Encyclopedia of Genes and Genomes
LRR	Leucine rich repeat
MAPK	Mitogen-activated protein kinase
(M/P)AMPs	Microbial- or pathogen-associated molecular patterns
MAPK	MAP kinase
MAPKK	MAP kinase kinase
MAPKKK	MAP kinase kinase kinase
mRNA	messenger Ribonucleic acid
NBS	Nucleotide-binding site
NCBI	National Center for Biotechnology Information
NGS	Next-generation sequencing
OIV	International Organisation of Vine and Wine
PAV	Presence/absence variation
PCR	Polymerase chain reaction
phi X	*Enterobacteria phage phiX174*
PR	Pathogenesis related
PR-genes	Pathogenesis related genes
PRRs	Pattern recognition receptors
PTI	PAMP-triggered immunity
p-value	Calculated probability
Q	Phred quality score
QTLs	Quantitative trait loci
R	Resistance
RBOH	Respiratory burst oxidase
RIN	RNA integrity number
RLKs	Receptor-like protein kinases
R-mediated	Resistance mediated
RNA	Ribonucleic acid
RNase	Ribonuclease
RNA-seq	Ribonucleic acid sequencing
ROS	Reactive oxygen species
R-proteins	Resistance proteins
RT-qPCR	Reverse transcription-qualitative polymerase chain reaction
SAR	Systemic acquired resistance
SAURs	SMALL AUXIN UP RNAs
SAWIS	South African Wine Industry Information and Systems
SMRT	Single Molecule Real-Time
SNP	Single nucleotide polymorphism
SPRI	Solid Phase Reversible Immobilization

Subsp.	Subspecies
TALENs	Transcription activator-like effector nucleases
TFs	Transcription factors
TIR	Toll and Interleukin 1 receptor
TM	Transmembrane
USA	United States of America
v	Version
vcf	Variant calling format
WOSA	Wines of South Africa
ZFNs	Zinc finger nucleases
ZMWs	Zero-mode waveguides

List of Chemicals

Ca^{2+}	Free calcium ions
CTAB	Cetyltrimethylammonium bromide
ddH_2O	Double distilled water
EDTA	Ethylenediaminetetraacetic acid
NaCl	Sodium chloride
PVP-10	Polyvinylpyrrolidone, average molecular weight 10,000
Tris-HCl	Tris-hydrochloride

List of Units

°Brix	degrees Brix
bp	basepairs
°C	degrees Celsius
g	gram
g/L	gram per litre
Ha	Hectare
Kb	Kilobases
M	Molar
Mb	Megabases
mg	microgram
mM	milliMolar
min	minutes
µg	microgram
µl	microlitre
ng	nanogram
nt	nucleotides
v/v	volume per volume
w/v	weight per volume

Chapter 1: Introduction

1.1 Research Context and Rationale

There is an increasing need to characterize the genomes of agriculturally-important species; the major food, feed and biofuel production crops. The development of genomic tools and resources is essential to complement traditional breeding for faster genetic improvement to increase yield, quality and stress tolerance. Knowledge of genetic diversity and relationships within a crop is crucial for the effective utilization and exploitation of plant genetic resources.

In an environment where climate and natural pressures are rapidly increasing and the wine market becomes exceedingly competitive, genetic research on grapevine cannot lag behind. In 2007, the grapevine (Pinot noir PN40024) genome was released (Jaillon et al. 2007), the first for a fruit crop. The availability of the grapevine genome sequence, together with advances in next-generation sequencing technologies, opens up opportunities for analysis of the grapevine genome, and since 2007 great progress has been made in understanding the grapevine genome.

Grapevine (*Vitis* spp., family *Vitaceae*) is the most cultivated fruit crop in the world. It is a woody perennial, widely grown in temperate regions. Worldwide, more than 7 million hectares are under grape cultivation, producing almost 75 million tonnes of grapes annually (FAO:http://www.fao.org). Grapevine has various uses as fresh fruit, raisins, grape juice, jam and wine, of which wine production is undeniably the largest industry (OIV: http://www.oiv.int). Production values from 2014 show that wine is the 7th largest processed commodity produced in the world. In 2016, 259 million hectolitres of wine were produced, with the top-ranking producers being Italy, France and Spain (OIV: http://www.oiv.int).

South Africa's climate, especially the westernmost part of South Africa, is ideal for viticulture. A victual station for passing ships was established at the Cape during the 1650s and in 1655 the first vineyard was planted there. Since then certain areas and farms became known for their wines. Under the auspices of the Wine of Origin Scheme, there are currently six officially demarcated geographical grapevine production units in South Africa, corresponding to provincial borders, namely Western, Eastern and Northern Cape and Free State, Kwazulu-Natal and Limpopo. The Western Cape is further divided into six production regions (WOSA:

http://www.wosa.co.za/). South Africa is the 8th largest wine producer in the world, responsible for 4% of the world's total wine volume. In the 2016/17 growing season, South Africa was ranked 14th in the world in terms of area, with more than 100 000 hectares of wine grapes, of which white-wine cultivars constituted 55.2% and red-wine cultivars 44.8% (SAWIS: http://www.sawis.co.za/).

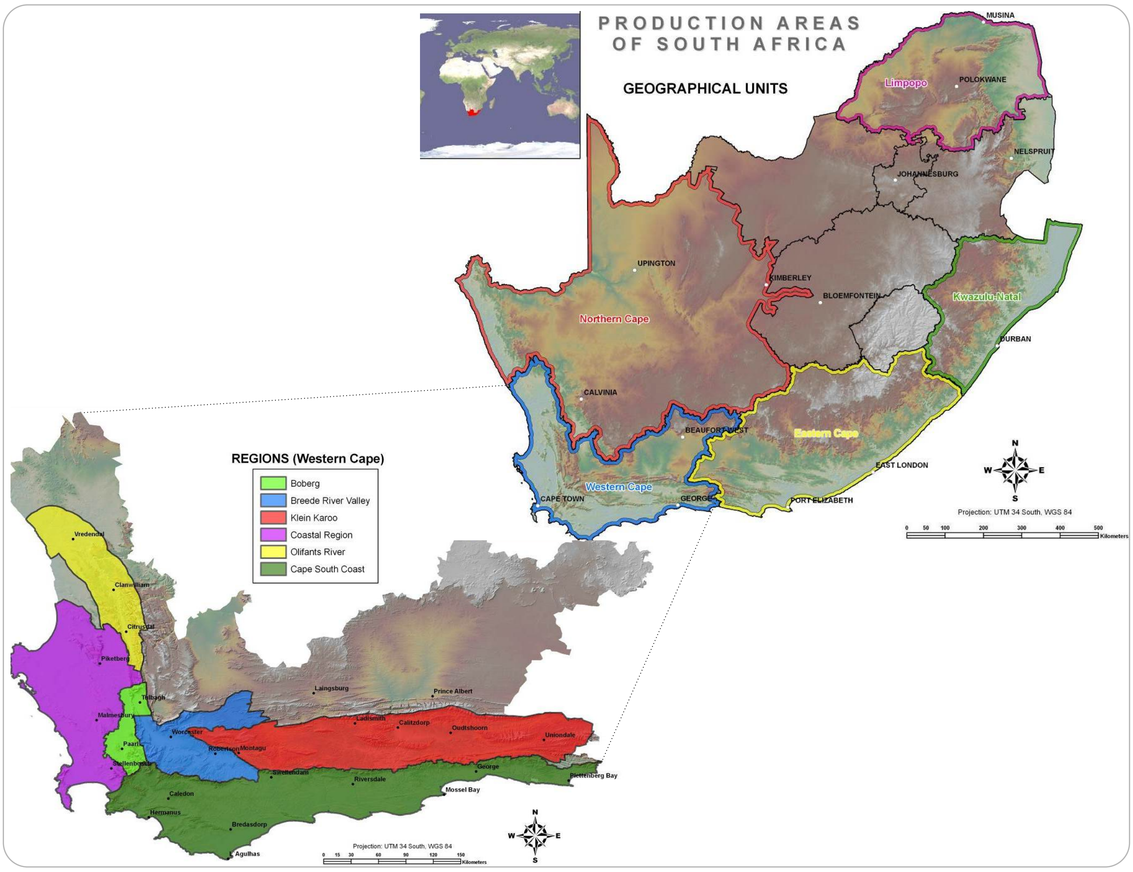


Figure 1.1: *Grapevine production areas in South Africa, showing the geographical units and regional divisions. The regions are further divided into 27 districts and 77 wards. (Adapted from: http://www.wosa.co.za/The-Industry/Winegrowing-Areas/Winelands-of-South-Africa/ and Wine of Origin Booklet 2016)*

Pinotage is a red-wine *Vitis vinifera* cultivar, bred in South Africa as a cross between Pinot noir and Cinsaut (Vivier and Pretorius 2000). Today, Pinotage is a successful commercial cultivar, used for the production of premium wines. It makes up 7.9% of the total vineyard area planted in South Africa (WOSA: http://www.wosa.co.za/). The characteristics of this South African flagship grapevine cultivar are different from the reference Pinot noir, e.g. Pinotage has a thicker berry skin and produce a darker red wine. This poses the question as to how the genome of Pinot noir and Pinotage differ, and are there genes that are not shared between these cultivars?

In South Africa, as in other grape-growing areas, there is a newfound interest in old vines and vineyards, and the exceptional wines made from them. These wines are generally accepted as having more depth and complexity than young-vineyard wines. The term "old vine" tends to be used on wine labels as an indication of a superior, high-quality wine (Heyns 2013; Easton 2016; Fridjhon 2016; Hawkins 2016; Hooke 2016; Beavers 2016; Van Wyk 2016; Szabo 2017). However, there is only anecdotal evidence that these wines are truly of higher standard. This study is the first scientific research into the so-called "old-vine" wine character, to determine whether there is any significant difference in gene expression between young and old vines, at the time of harvest. Gene expression of 40-year old and 7-year old vines growing in a commercial Pinotage vineyard, used for the production of such premium wines, were analysed as a starting point to elucidate the origins of this old-vine character.

1.2 Aims and Objectives

The key focus of this project was to study the genetics of Pinotage, in order to determine the genetic basis underlying the character of this cultivar. The scope of the project was further broadened to also include the study of gene expression levels of young and old Pinotage vines, in both berries and leaves, to gain insight into the old-vine character of old-vineyard wines. These aims were achieved using next-generation sequencing and the latest bioinformatic tools.

The following objectives were determined in order to achieve the aims:

1. *De novo* draft assembly of the Pinotage genome.
    a. Obtain high quality DNA and RNA from Pinotage vines and perform DNA and RNA sequencing.
    b. Assess suitable *de novo* assembly software and parameters and do a *de novo* assembly of the Pinotage genome.
    c. Analyse the genomic variance between Pinotage and the reference Pinot noir.
2. Compare the leaf and berry transcriptomes of young and old Pinotage vines, at the time of harvest.
    a. Obtain high quality RNA from the leaves and berries of young and old Pinotage vines and perform mRNA sequencing.
    b. Perform a reference-based transcriptome expression analysis of Pinotage vine leaves and berries.
    c. Identify and classify genes differentially expressed between young and old vines.

3. Discover Pinotage genes not present in the Pinot noir reference genome.
    a. Perform a *de novo* transcriptome assembly of the Pinotage transcriptome and compare the transcriptome and genome data to study Pinotage genes.

1.3 Chapter Layout

This dissertation is divided into six chapters. Each chapter is individually introduced. A complete reference list and supplementary tables and information to the research chapter are is provided at the end of the thesis.

Chapter 1: Introduction

A general introduction to the study and its significance, including aims and objectives, and an overview of the chapter layout are provided.

Chapter 2: Literature Review

An overview of the literature pertaining to this study is provided, including the origin and history of *Vitis vinifera* domestication, and overview of plant genome sequencing and the challenges thereof. The databases and resources available for grapevine genetic research and gene classification are listed and briefly discussed.

Chapter 3: *De novo* Assembly of the Pinotage Draft Genome

The next-generation sequencing and assembly strategy used to obtain a draft Pinotage genome sequence is discussed. A comparison of the Pinotage sequence data and the reference Pinot noir genome is also provided.

Chapter 4: The Pinotage Leaf and Berry Transcriptome in Young and Old Vines

The transcriptomes of Pinotage leaves and berries at harvest were surveyed, and a reference-based differential expression analysis between young and old vines, performed. The significance of the differential expressed genes in berry ripening is discussed. A number of putative novel gene loci were also identified.

Manuscript entitled “The Pinotage leaf and berry transcriptome in young and old vines” in preparation, to be submitted to a peer-reviewed journal.

Chapter 5: Pinotage *De novo* Transcriptome Assembly

Transcriptome data were *de novo* assembled and compared to the Pinotage genome data. Pinotage genes not found in Pinot noir were highlighted and discussed in the context of the stress response network they are involved in.

Manuscript entitled "Pinotage *De novo* Transcriptome Assembly" in preparation, to be submitted to a peer-reviewed journal. This manuscript will contain the results discussed in Chapter 5 together with a brief discussion of the Pinotage genome sequencing and assembly (Chapter 3).

Chapter 6: Conclusion

A concluding summary of the main results is provided, along with the strengths and limitations of this study and proposals for future research.

1.4 References

Beavers K (2016) What the heck is old vine wine? Here's everything you need to know. In: VinePair. https://vinepair.com/wine-geekly/what-the-heck-is-old-vine-wine-heres-everything-you-need-to-know/. Accessed 20 Jun 2017

Easton S (2016) Old vines – do they make the best wines? In: WineWisdom. http://www.winewisdom.com/articles/techie/old-vines-do-they-make-the-best-wines/. Accessed 20 Jun 2017

Fridjhon M (2016) How to save SA's old vines; SA Wine Ratings, News, Opinion & Analysis. In: Winemag.co.za. http://winemag.co.za/michael-fridjhon-how-to-save-sas-old-vines/. Accessed 20 Jun 2017

Hawkins B (2016) The Real Story Behind "Old Vine" Wines in South Africa. In: Explore Sideways. https://exploresideways.com/whats-the-deal-with-old-vines-and-why-do-they-matter-to-the-south-african-wine-industry/. Accessed 20 Jun 2017

Heyns E (2013) Old vines new opportunities? In: Wineland Mag. http://www.wineland.co.za/old-vines-new-opportunities/. Accessed 20 Jun 2017

Hooke H (2016) Defining old vines. In: Real Rev. https://www.therealreview.com/2016/12/01/defining-old-vines/. Accessed 20 Jun 2017

Jaillon O, Aury J-M, Noel B, Policriti A, Clepet C, Casagrande A, Choisne N, Aubourg S, Vitulo N, Jubin C et al. (2007) The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. Nature 449:463–467. doi: 10.1038/nature06148

Szabo J (2017) Growing old in South Africa: The old vines project. In: WineAlign. http://www.winealign.com/articles/2017/06/01/growing-old-in-south-africa-the-old-vines-project. Accessed 20 Jun 2017

Van Wyk E (2016) Aftertaste old vines : lifestyle wines. Prejudice 16:64–65.

Vivier M and Pretorius I (2000) Genetic improvement of grapevine: tailoring grape varieties for the third millennium - a review. South Afr J Enol Vitic 21:5–26.

Online resources

Food and Agriculture Organization of the United Nations (FAO): http://www.fao.org

International Organisation of Vine and Wine (OIV): http://www.oiv.int

South African Wine Industry Information and Systems (SAWIS): http://www.sawis.co.za/

Wines of South Africa (WOSA): http://www.wosa.co.za/

Chapter 2: Literature Review

This chapter provides a broad overview of the literature pertaining to this study. The history of grapevine domestication and the impact thereof on the grapevine genome is discussed. A brief overview is given of the contribution of genome sequencing to crop improvement, and the challenges of plant genome sequencing and assembly. Finally, the databases and resources available for grapevine genetic research, gene functional classification and metabolic network analyses, are listed and discussed. Literature pertaining to specific research chapters is discussed in the introduction sections of these chapters: the history of Pinotage in South Africa (Chapter 3), fruit ripening and an introduction to the so-called “old-vine” character (Chapter 4) and the phenotypic differences between Pinotage and Pinot noir and influence of secondary metabolites on the aroma profile of wine (Chapter 5).

2.1 History of Grapevine Domestication

Winemaking and grape cultivation share ancient historical connections with human cultural development and are inseparable parts of the culture and history of many countries. Grapevine is one of the first fruit crops domesticated by humans, and is probably strongly linked to the production of wine. Domestication most likely occurred during the Neolithic period (6000 BC) in the South Caucasus (Azerbaijan, Armenia, and Georgia) and the eastern Anatolian (Turkey and Iran) regions (Figure 2.1) (Alleweldt and Possingham 1988; Vivier and Pretorius 2000; McGovern et al. 2003; This et al. 2006; Reynolds 2010; Imazio et al. 2013; McGovern 2013).

Following initial domestication, cultivated grapevine was spread by humans to the Near and Middle East and Central Europe. There is evidence to suggest that separate secondary domestication events occurred in these regions. From there, the European cultivated *V. vinifera* subsp*. vinifera* was dispersed to North America, Africa, South America and Australia, mainly as a result of European colonization (Alleweldt and Possingham 1988; Vivier and Pretorius 2000; McGovern et al. 2003; This et al. 2006; Reynolds 2010; Imazio et al. 2013; McGovern 2013). Today grapevine is cultivated on every arable continent, mainly in areas with a temperate climate (Vivier and Pretorius 2000) between the 30^{o} and 50^{o} latitudes (Figure 2.1).

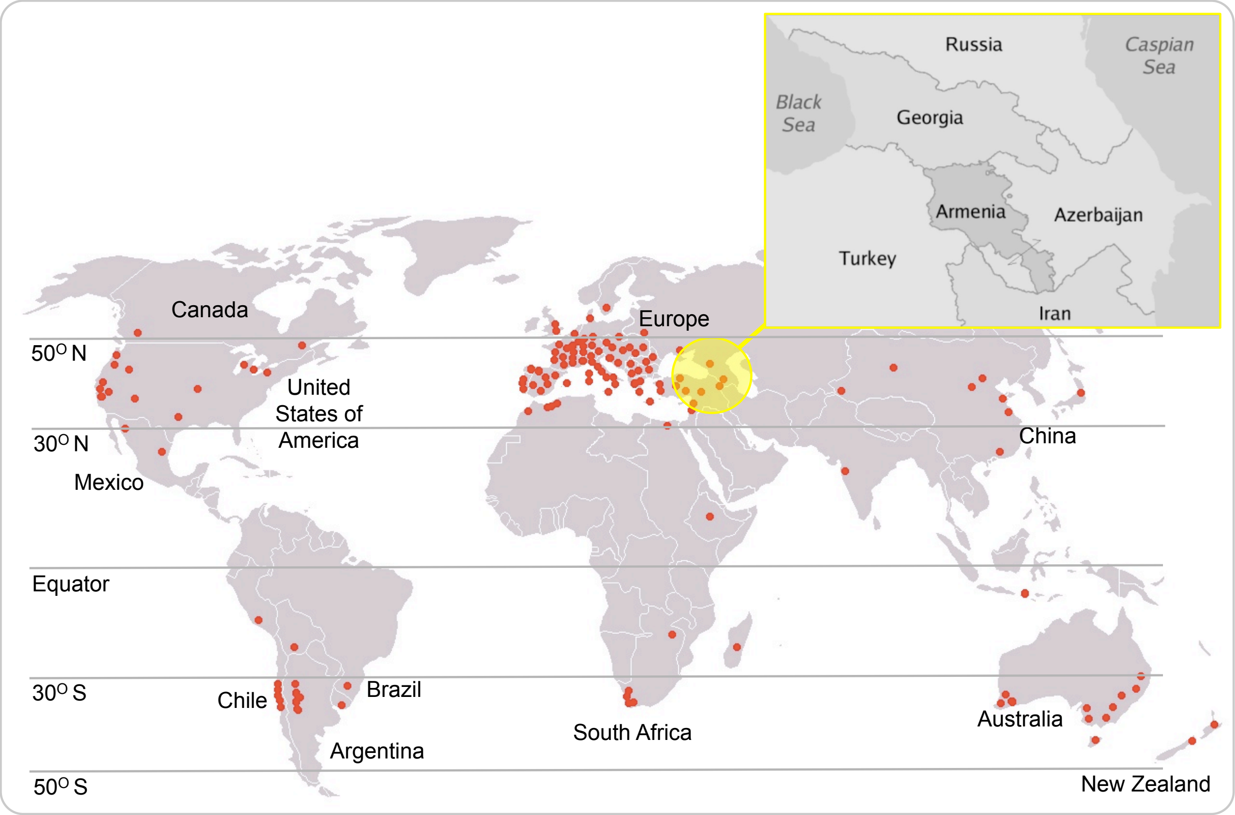


Figure 2.1: *Centre of grapevine domestication (yellow circle, insertion showing the modern-day country borders) and current wine-producing areas (red dots). Grapevine grows mainly in areas with a temperate climate, between the 30^o and 50^o latitudes. (Adapted from: http://drinkwire.liquor.com/post/okanagan-spirits-world-class-canadian-distillery-in-bc-wine-country#gs.7r_CE6k and https://www.slideshare.net/joelbutlermw/turkeys-indigenous-wine-varieties-11812*).

Grapevine belongs to the genus *Vitis* which includes ~60 species. The genus can be divided into three major groups, namely species native to North America, Europe and Asia (Figure 2.2). However, almost all the wine produced in the world derives from the European grapevine, *Vitis vinifera*. Within the species *vinifera* two subspecies exist, subspecies *vinifera* (also called *sativa) and sylvestris.* Subspecies *vinifera* was domesticated from the wild *sylvestris* ancestor. During the domestication process the physiology of grapevine evolved rapidly to produce berries with higher sugar content and larger, more consistent yields. Wild grapevine is dioecious, but most modern cultivars grow as hermaphroditic, self-fertile plants (This et al. 2006).

All *Vitis* species have 38 chromosomes (n=19), and are inter-fertile (Jaillon et al. 2007; Velasco et al. 2007). This characteristic allows the extensive use of hybridization in grapevine breeding, either natural or viticultural hybridization, most often to combine the fruit production characteristics of domesticated species and the hardiness of wild species. In the 1860s the

European grape cultivation was nearly wiped out in a matter of years by phylloxera, a soil-borne aphid imported from North America (Alleweldt and Possingham 1988; This et al. 2006). Native North American grapevine species are resistant to phylloxera, and today species such as *V. riparia and V. rupestris* are commonly used as rootstocks onto which *V. vinifera* cultivars are grafted to manage phylloxera.

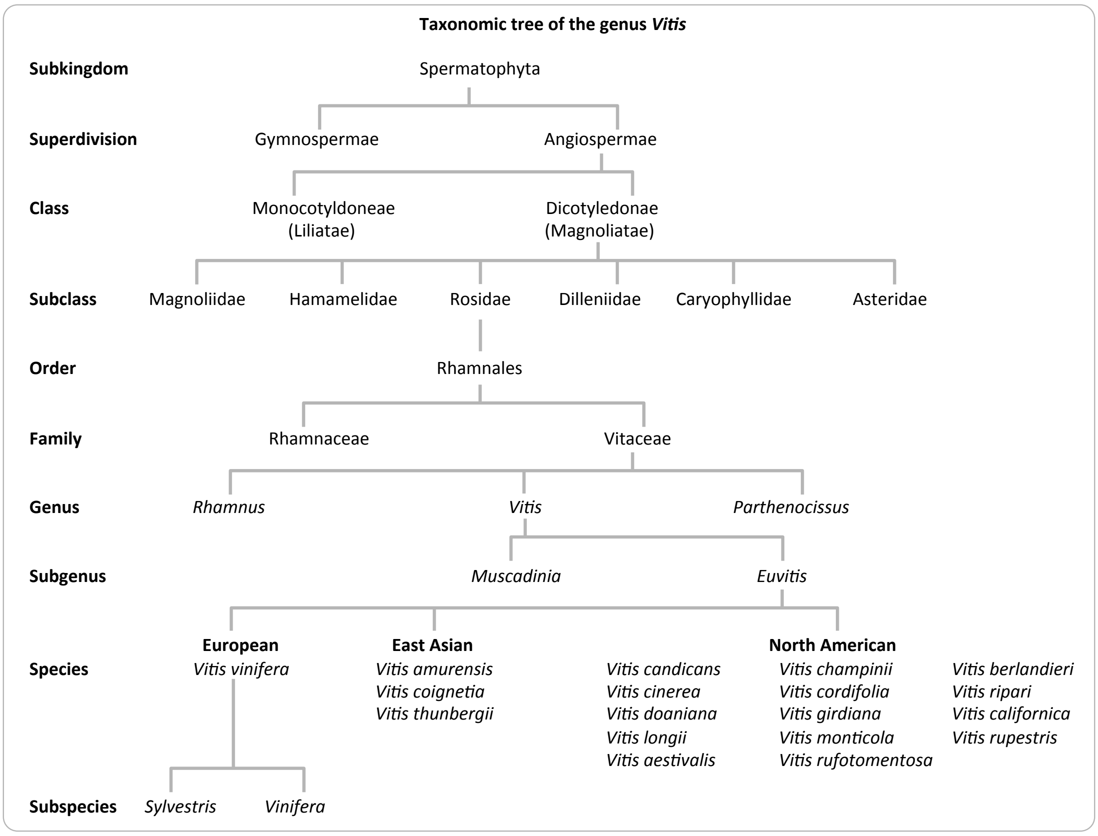


Figure 2.2: *Taxonomic tree of the modern grapevine, Vitis vinifera subsp. vinifera. Not all taxonomic entries are shown. Adapted from: Vivier and Pretorius (2000) and Vitis International Variety Catalogue database (http://www.vivc.de/index.php?r=aboutvivc%2Ftaxonomictree).*

The increase in popularity of wine drove the development of individual grapevine cultivars and the unique taste of their wines. The different climates and growing conditions together with human selective pressures have shaped the properties associated with the modern-day popular cultivars. Nowadays, most new cultivars arise from crosses between existing cultivars, to harness the positive characteristics of both parents. For example, the grapevine cultivar Pinotage was created in 1925 from a Pinot noir X Cinsaut cross in South Africa (discussed in Chapter 3, Section 3.1). More than 24,000 names and synonyms currently exist for grapevine cultivars, however only ±5,000 are true distinctive cultivars (Alleweldt and Possingham 1988). The *Vitis* International Variety Catalogue (http://www.vivc.de/) was established in 1984

(Lacombe et al. 2015) and hosts an encyclopaedic database with more than 23,000 cultivar names, breeding lines and *Vitis* species listed, including synonyms, country of origin, ampelographic information, susceptibility/resistance to diseases, etc. However, only a few popular cultivars are extensively grown for the global wine market. Nevertheless, there is an increased interest in the use of local cultivars to create boutique wines with a unique style, rather than just reproducing traditional old-world wine styles from the popular cultivars.

2.2 Impact of Domestication on the Grapevine Genome

A plant's genome is drastically reshaped during domestication, culminating in a genome with significantly reduced diversity in certain areas, but also enrichment for putative beneficial genomic changes, within genic and non-genic areas. Interestingly, many changes that played an important role in domestication are not within the gene coding region, but rather in the *cis*regulatory elements (CREs), controlling the expression of genes (Shi and Lai 2015; Swinnen et al. 2016). Data from genome-wide studies also suggest that it is not only absolute changes to the genome sequence, but also epigenetic modifications that can play a significant role in crop domestication (Shi and Lai 2015). Epigenetic modifications to a regulatory element can have a drastic impact by promoting or supressing gene expression.

Transposable elements contribute to somatic mutations, both beneficial and deleterious, as they randomly insert into the genome and certainly play a major role in grapevine genetic diversity and evolution (This et al. 2006; Benjak et al. 2008; Imazio et al. 2013). Probably the best-known example of variation due to the insertion of a transposable element is the colour mutation in grapevine. White cultivars originated from red cultivars by two independent mutations: the insertion of a gypsy-type transposon (Gret1) in the regulatory element of the *VvMybA1* gene, and a single nucleotide polymorphism (SNP) in *VvMybA2*. *VvMybA* genes encode for transcription factors in the MYB family. These transcription factors regulate the expression of the anthocyanin gene, the colour pigment, in grapevine. The mutations in the *VvMybA* genes are responsible for the loss of berry skin colour in homozygous vines (Kobayashi et al. 2004; Yakushiji et al. 2006; Fournier-Level et al. 2010; Shimazaki et al. 2011; Péros et al. 2015), while different allele combinations of these genes give rise to the colour variations in grapevine cultivars.

During the process of domestication, sexually propagated plants experience more severe genetic bottlenecks (greater reduction in genetic diversity) than vegetatively propagated crops.

As grapevine is a vegetatively propagated crop, it sustained a high level of genetic diversity in domesticated vines (Myles et al. 2011) and is highly heterozygous (Jaillon et al. 2007). Due to this extensive heterozygosity, seed-germinated offspring display diverse characteristics and cause erratic yields. Vegetative propagation, on the other hand, is easy, preserves the existing traits in a specific cultivar and allows for unique phenotypes arising from somatic mutations to be preserved. However, clonal propagation can also lead to the accumulation of recessive deleterious variants, with domesticated grape accessions containing up to 5.2% more deleterious mutations than wild individuals (Zhou et al. 2017).

2.3 Plant Genome Sequencing

2.3.1 Contribution of plant genome sequencing to crop improvement

The application of next-generation sequencing (NGS) technologies and subsequent bioinformatic analyses have already revolutionized breeding strategies to achieve faster, more efficient genetic improvement of crops (Bolger et al. 2014b; Barabaschi et al. 2016; Scossa et al. 2016; Batley and Edwards 2016; Scheben et al. 2016; Bevan et al. 2017). Breeding of crops can be time consuming, as the plants need to reach physiological maturity before assessment of the marketable product, for example the fruit or grain, is possible. In particular, fruit trees have a long generation time and juvenile phase and, together with large plant size, adds to the challenges of fruit tree breeding (Iwata et al. 2016). Marker-assisted selection greatly improved this process by allowing early genetic selection of breeding stock. The availability of genome sequences for several agronomically important crops, together with more re-sequencing data, provides genome information for *en mass* genome-wide marker development. Re-sequencing NGS data also allow for more rapid characterization of genetic diversity within a plant, and with a continued decrease in cost, will replace traditional genetic markers and genotyping techniques (Bolger et al. 2014b; Barabaschi et al. 2016; Scossa et al. 2016; Batley and Edwards 2016; Scheben et al. 2016; Bevan et al. 2017).

The first plant genome, that of *Arabidopsis thaliana*, was published in 2000 while the genome sequence of rice, the first for a crop plant, appeared in 2002 (Goff et al. 2002; Yu et al. 2002). Grapevine was the first fruit crop sequenced (Jaillon et al. 2007) (discussed in Section 2.4.1). Triggered by the fast pace of advancement in NGS technologies, many more plant genome sequences have since been produced. Currently more than 230 plant genome sequences are available (plaBi database: http://www.plabipd.de/); mostly food, fuel and fibre crops and model plant species. However, these genome sequences are of varying quality and stages of completeness, and very few draft genome sequences have been finished to a similar quality level as that of *Arabidopsis, maize or rice*. Currently, the whole-genome shotgun sequencing and assembly strategy, albeit much faster and cost-effective, does not offer the same quality and completeness of genome sequences previously obtained with Sanger sequencing and map-based approaches. Nevertheless, a high-quality draft is not a requirement for genome studies. Low coverage genome re-sequencing of more cultivars or genotypes is now possible at a reasonable cost, and a significant amount of insight can still be gained from these genomes. NGS data from re-sequencing projects is well suited for the discovery of genomic variants, genetic diversity and assessment, and marker development (Bolger et al. 2014b; Barabaschi et al. 2016; Scossa et al. 2016; Batley and Edwards 2016; Scheben et al. 2016; Bevan et al. 2017).

However, for genome data to contribute to crop breeding and horticulture, it is necessary to identify the genes and genetic variants underlying traits and/or phenotypic variation within the species that are of agronomic importance (Scossa et al. 2016; Batley and Edwards 2016; Scheben et al. 2016; Bevan et al. 2017; Yuan et al. 2017). In most genome projects, high-throughput functional assignment is performed by searches for orthologs in well-characterized genomes, often model plants such as *Arabidopsis*. Various software and applications exist to perform sequence similarity searches for functional assignment (some of these tools are discussed in Section 2.4.3, Table 2.4).

Although genetic diversity contained in the genomes, cultivated lineages and wild relatives, will continue to be the basis of any plant breeding program, new biotechnology techniques provide alternatives to standard breeding practices (Bolger et al. 2014b; Barabaschi et al. 2016; Scossa et al. 2016; Batley and Edwards 2016; Scheben et al. 2016; Bevan et al. 2017; Yuan et al. 2017). The possibility of creating genetically modified crops has already been proved feasible. Mostly *trans*- and *cis*genics have been used in genetic engineering, but more recently improved techniques for genome-editing were introduced (Chialva et al. 2016). Developing these technologies would not have been possible without the known genome sequences of the organisms to be engineered.

Transgenic technology is the isolation of a gene derived from one species and the random insertion thereof into the genome of another species. On the other hand, *cis*genic techniques rely on the transfer of genes or regulatory sequences between genotypes within the same or sexually compatible species (Cardi 2016). Genome-editing has definite advantages over traditional breeding, as well as *cis*- and transgenics, most importantly the speed and precision whereby these edits can be made to a genome. For such precision editing, a reliable genome sequence is essential. Genome-editing is based on techniques that create breaks in double-stranded DNA. Sequence-specific nucleases including zinc finger nucleases (ZFNs) and transcription activator-like effector nucleases (TALENs) have previously been used. More recently, the CRISPR/Cas9 system (CRISPR: clustered regularly-interspaced short palindromic repeats; Cas: CRISPR associated protein) became available (Chialva et al. 2016). A database with genomic sites suitable for CRISPR/Cas9 in grapevine has already been developed (Wang et al. 2016). Protoplasts with an edit to a gene conferring powdery mildew resistance (Malnoy et al. 2016) and targeted mutagenesis (Ren et al. 2016), both in the grapevine cultivar Chardonnay, were also generated and proved to be feasible.

2.3.2 Challenges in plant genome sequencing and assembly

Plants present a number of challenges for genome sequencing and assembly compared to animal genomes. They display great diversity both in terms of size and structure of their genomes, and although the genome sizes of plants and animals are comparable, plants generally have more complex genome structures than most animal species (Gregory 2005; Gregory et al. 2007; Feuillet et al. 2011).

Analysis of the grapevine genome suggests that dicotyledons underwent an ancient hexaploidization event (Jaillon et al. 2007). Many plants species underwent more recent whole genome duplication events and as much as 80% of plants may have polyploid genomes (Meyers et al. 2006). Transposon activity causes genome rearrangements and duplications, and together with whole genome duplication events, are the main origin of gene family expansion and pseudogenes (Barabaschi et al. 2012). Paralogous genes in a family and pseudogenes may have nearly identical sequences, posing an assembly challenge, because the sequencing reads can map with equal likelihood to multiple reference genome positions, and it can be difficult or impossible to differentiate between alleles and paralogous family members (Morrell et al. 2011). Furthermore, due to the repetitive nature of transposable elements themselves, they exacerbate the problem. Plant genomes contain abundant transposable and repetitive elements; for example, it is estimated that these make up 41.4% of the grape genome sequence (Jaillon et al. 2007).

Plant genomes can be highly heterozygous (Feuillet et al. 2011) with high sequence variation between alleles that complicates haploid assembly. A method to overcome this challenge is to develop inbred lines or double haploids of the plant to be sequenced, eliminating heterozygosity. This, however, is both time consuming and costly, and in some cases inbreeding methods may fail to eliminate heterozygosity (Bolger et al. 2014b; Barabaschi et al. 2016; Scossa et al. 2016; Batley and Edwards 2016; Scheben et al. 2016; Bevan et al. 2017). As mentioned earlier, grapevine is an example of a plant with a high heterozygosity level. But researchers succeeded in generating an inbred line, called PN40024, and after successive generations of selfing, it was estimated to be 93% homozygous, greatly simplifying the genome assembly (sequencing of grapevine PN40024 is further discussed in Section 2.4.1) (Jaillon et al. 2007).

Most plants cells contain a great number of chloroplast and mitochondrion organelles. The nucleic acids of these plastid genomes are co-extracted with the nuclear DNA during DNA extraction, and as they are more abundant, their presence may skew the depth of coverage levels. Reads aligning to the plastids cannot simply be discarded from the dataset, due to the presence of plastid remnants in the nuclear genome (Bolger et al. 2014b; Barabaschi et al. 2016; Scossa et al. 2016; Batley and Edwards 2016; Scheben et al. 2016; Bevan et al. 2017). Another challenge of plant genome sequencing is that, due to the presence of secondary metabolites and polyphenolics in plant material, it can be very difficult to extract a sufficient amount of high-quality DNA necessary for the construction of NGS libraries (Salzman et al. 1999; Gambino et al. 2008; Aubakirova et al. 2014). The consequences of the aforementioned challenges on the genome sequencing in this project, are discussed in Chapter 3.

Despite these challenges, plants do have advantages over animals in the field of genomics. Unlike most animals, plants can be clonally propagated and many species' seeds can be preserved indefinitely, effectively immortalizing genotypes of interest. A genotype can therefore be sequenced once, but phenotyped repeatedly, in different environments. Furthermore, some plants can be maintained as inbred lines or double haploids, avoiding the complexities of assembling a highly heterozygous genome (Bolger et al. 2014b; Barabaschi et al. 2016; Scossa et al. 2016; Batley and Edwards 2016; Scheben et al. 2016; Bevan et al. 2017).

2.3.3 New technology offers solutions to plant genome sequencing challenges

As new technologies and improvements to NGS are developed and increasingly used for genome sequencing, it is expected that high quality reference genomes will quickly become available for many grapevine cultivars and species. One improvement to the bioinformatic aspect of genome assembly is the development of assemblers that are ploidy aware, i.e. they are able to assemble both alleles when heterozygous. To achieve the best assembly possible, sequencing read depth must be high enough and have a uniform coverage across the genome. This allows assemblers to recognize alternative alleles based on lower average read depth, and group them into contigs specific for each of the two alleles. The total size of the assembled contigs should therefore exceed the expected genome size.

Currently, the short read-length is still a major constraint in NGS. *De novo* assemblies will greatly improve with the availability of longer read-lengths. Substantially longer read lengths are generated using PACBIO's SMRT (Single Molecule Real-Time) DNA sequencing technology (PACBIO®: http://www.pacb.com/smrt-science/smrt-sequencing/) (Eid et al. 2009; Roberts et al. 2013). For this technology, the reported average read length is currently more than 10,000nt. SMRT sequencing is based on real-time imaging of the fluorescent signal produced by labelled nucleotides as they are incorporated into the growing DNA strand build on the template. The strength of this technology is the very strong light detection capability; the light signal from a single fluorophore can be detected, allowing a single template-strand. This eliminates the need for beforehand duplication of the template strands, with associated disadvantages. At the core of this SMRT technology is the sequencing cell, consisting of tens of thousands of small wells with a waveguide at the bottom, called zero-mode waveguides (ZMWs). The ZMWs are illuminated from below, but the light's wavelength is too large to allow it to pass efficiently through the waveguide. Attenuated light then penetrates only the lower 20-30nm of the ZMW, creating a very small detection volume. A single DNA template-polymerase complex is then immobilized at the bottom of each ZMW and nucleotides added. Each of the four nucleotides is labelled with a different coloured fluorophore. As the nucleotide is incorporated into the growing DNA chain by the polymerase, it is held close to the bottom of the ZMW, in the detection volume, and the fluorophore emits a light signal that is then detected. Due to the small detection volume, background noise is greatly reduced. The fluorophore is then cleaved, and the polymerase can incorporate the next nucleotide in the chain (PACBIO®: http://www.pacb.com/smrt-science/smrt-sequencing/). The SMRT sequencing technology, together with tools specifically developed for this type of sequencing data, have already been proven to improve plant genome assemblies (Chin et al. 2016).

Another recent development is that of long-range scaffolding technologies, such as optical maps. For optical maps, specific sequence motifs are fluorescently labelled and the DNA is then stretched to a linear configuration. The pattern of fluorescent labels can then be visualized by fluorescence microscopy, creating a high density "barcode" along the DNA, with known distances between specific sequence motifs. This information can then be integrated with the sequencing data to orientate and anchor assembled contigs.

The long read length of SMRT sequencing is especially useful in plant genome assembly (Bellec et al. 2016) and together with a scaffolding technology, it can greatly improve a *de novo* assembly to characterize complex structural variations and/or genomic rearrangements between the genotype of interest and the reference sequence, or to construct a *de novo* assembly where no reference is available. These technologies are being used to obtain a high-quality Chardonnay genome sequence (Minio et al. 2017).

2.4 Resources and Tools for Grapevine Genomic Research

The techniques used in biological research have changed significantly in the last decade, and the generation of various types of large datasets are now routine. Access to these datasets is promoted by FAIR (Findable, Accessible, Interoperable, Reusable) principles (Wilkinson et al. 2016), which put specific emphasis on standardization and organization of data to automate data mining. Researchers are increasingly applying high-throughput experimental techniques, generating large datasets, for example "omics" technologies such as genomics or transcriptomics that make use of NGS, making FAIR especially applicable in the field of life sciences and NGS datasets.

Although a substantial amount of grapevine genetic data exists, associated datasets are not necessarily in a standardized format and/or not readily accessible. These datasets are typically stored in public repositories, but to fully exploit the data it should be organized in a central information platform in a standardized format (according to FAIR principles). This will allow for the integration and comparison of different experiments and datasets, allowing researchers a holistic view of available data (Adam-Blondon et al. 2016). Such an information platform should host the complete experimental dataset and metadata, including genotypes, phenotypes, development stages, mutants, growth conditions, etc. In 2016 a strategy for the development of such a system, the Grape Information System (GrapeIS), was launched (Adam-Blondon et al. 2016).

2.4.1 Grapevine genome sequencing and available databases

Vitis vinifera cv Pinot noir was the first fruit crop genome sequenced. As previously discussed, commercial grapevine cultivars are highly heterozygous, complicating reliable genome assembly when applying a shotgun sequencing and *de novo* assembly strategy. The PN40024 line, derived from Pinot noir through repeated back crossing, and estimated to be 93% homozygous, was developed to reduce the complexity of genome assembly. Sequencing was performed using a whole-genome shotgun strategy. A library of bacterial artificial clones was sequenced with Sanger sequencing. The grapevine genome is 487Mb in size, diploid and consists of 19 chromosome pairs. The 2,093 assembled supercontigs were grouped into 33 "chromosomes" (NCBI Bioproject: PRJEA18785). Among these, "Random chromosomes" contain contigs that could be assigned to a chromosome but their exact position on the chromosome could not be determined. "Chromosome unknown" contains supercontigs that could not be assigned to a chromosome. Contigs in the "random" and "unknown" chromosomes were joined together by a stretch of 500 unknown nucleotides ("Ns"). The first assembly had 8X coverage and a predicted 30,434 genes (Jaillon et al. 2007). Since the first assembly release, more data were added and the current assembly has 12X coverage. At the time of writing, the latest annotation release, V2.1, contains 31,845 genes (Table 2.1).

Table 2.1: Number of predicted genes in the grapevine annotation versions. All annotations are available from CRIBI (Centro di Ricerca Interdipartimentale per le Biotecnologie Innovative, University of Padua, Italy; http://genomes.cribi.unipd.it/grape/).

	8X	**12X**		
	V1	V0	V1	V2
Number of genes	30,434	26,346	29,971	31,845
Number of transcripts	-	-	-	55,564
Gene identifier (Identifier prefix)	Vv	GSVIV JGVv PDVv	VIT	VIT_2
Reference	(Jaillon et al. 2007)	-	-	(Vitulo et al. 2014)

Since 2005, NGS platforms became available that offer high-throughput and cost-efficient sequencing. The heterozygous Pinot noir clone, ENTAV115 (ENTAV: Etablissement National Technique pour l'Amolioration de la Viticulture, France), was sequenced, employing a combination of a Sanger shotgun strategy and NGS (Velasco et al. 2007). Genome data for ENTAV115 were assembled into 66,164 contigs (NCBI Bioproject: PRJEA18357).

The genome of a table grape cultivar, Sultanina (Thomson Seedless) was published in 2014 (Di Genova et al. 2014) using only an NGS approach. A novel *de novo* assembly strategy for heterozygous genomes was implemented and a draft genome of 466Mb was produced (NCBI Bioproject: PRJNA207665). More than 82% of the genes annotated in the reference genome could be identified, together with a large number of structural variants, insertions and deletions (indels) and SNPs.

More recently the genome of Tannat, a red-wine cultivar from South East of France, has been sequenced (Da Silva et al. 2013) and the genome sequencing and assembly of Cabernet Sauvignon is underway (Minio et al. 2017). In the NCBI database the assembled genome data for four Georgian *Vitis vinifera* cultivars (NCBI Bioproject: PRJDB5761), as well as two non-*vinifera* species, Boerner (*V. riparia* X *V. cinerea* cross, NCBI Bioproject: PRJEB5934) and Norton (*V. aestivalis*, NCBI Bioproject: PRJNA302606) are also available.

Table 2.2 lists some of the publicly available databases that host grapevine genome data. The PN40024 reference sequence is available from Genoscope and CRIBI, while the other above-mentioned data are available from NCBI. Most of these databases have genome browsers and other grapevine NGS datasets available for data mining. Three platforms hosting genetic data for plants, including grapevine, are also included in Table 2.2.

2.4.2 Grapevine gene functional classification and pathway analysis

Having genome and/or transcriptome sequence information of a crop cultivar is only the first step in understanding the relationship between intra-species genetic and phenotypic variation. Assigning and classifying gene function is a key step in the interpretation of genome and/or transcriptome sequence data. Table 2.3 lists some of the resources available with grapevine gene names coupled to function. For example, GrapeCyc was extensively used in this study (Chapter 4, Section 4.3.2) to identify enzymes encoded by specific genes. These tools are in addition to the previously mentioned databases that also include functional annotations, for example CRIBI and Genoscope.

As genes and gene products do not function as isolated units, but rather as integrated metabolic networks, it is important to represent interactions between them to allow an overview of the interconnection of these genes. For example, enzymes encoded by different genes can all be involved in one biochemical pathway to produce a metabolite, or transcription factors that suppress or promote expression of other genes. When performing gene analysis, for example differential expression, it is important to analyse not only individual genes, but also genes as part of their respective pathways. For example, minor differential regulation in a number of genes may have a major impact on final metabolite concentration. The tools shown in Table 2.3 can also be used for grapevine metabolic network and/or functional enrichment analysis. However, our knowledge of gene interactions and metabolic pathways are far from complete. To further complicate such analyses, different tools use different functional annotations, gene identifiers, and representations of data.

Table 2.2: Grapevine and plant genomic databases, hosting the grapevine genome sequence, annotations and NGS datasets pertaining to grapevine and other plant species.

Grapevine databases	Website address	Description	Reference
BIOWINE	https://alpha.dmi.unict.it/biowine/	A database for functional analysis of Sicilian grapevine cultivars. A number of NGS datasets are hosted in the BIOWINE database.	(Pulvirenti et al. 2015)
CRIBI	http://genomes.cribi.unipd.it/grape/	Genomics and Bioinformatics group, University of Padua, Italy. Hosts the newest grapevine genome annotations.	
Genoscope	http://www.genoscope.cns.fr/externe/GenomeBrowser/Vitis/	Hosts the 12X grapevine genome assembly by a French-Italian public consortium (INRA, Genoscope, University of Milan, University of Udine and University of Padua).	
IGGP	http://www.vitaceae.org	The International Grape Genome Program (IGGP) was founded in 2001 to facilitate international collaboration and development of grapevine research resources.	
VTCdb	http://vtcdb.adelaide.edu.au/ Home.aspx	A grapevine database with gene co-expression data, include capabilities for functional enrichment and visualization of co-expression networks.	(Wong et al. 2013)
Plant databases	**Website address**	**Description**	**Reference**
Ensembl plants	http://plants.ensembl.org/	Hosts a number of plant genome assemblies, annotations, genome browsers and other tools.	
Gramene	http://www.gramene.org/	Database for comparisons and functional genomics of crop and model plant species.	(Gupta et al. 2016)
NCBI (not plant specific)	https://www.ncbi.nlm.nih.gov/genome/401	Hosts genome and assembly information of grapevine and other organisms. NCBI also supports its own grapevine genome annotation: https://www.ncbi.nlm.nih.gov/genome/annotation_euk/Vitis_vinifera/101/#AssembliesReport	

Table 2.3: Tools for grapevine metabolic network and enrichment analysis. The tools are used to map, integrate and visualize various types of biological data on molecular and genetic pathways.

Platform	Website address	Description	Reference
Pathview	https://pathview.uncc.edu/about	Map and visualize a variety of biological data on grapevine pathway diagrams, uses KEGG data.	
vespucci	http://vespucci.colombos.fmach.it/	Analyse and visualize gene expression values of the grapevine gene expression compendium.	
VitisNet	https://www.sdstate.edu/vitisnet-molecular-networks-grapevine	VitisNet hosts annotated metabolic networks of grapevine.	(Grimplet et al. 2009)
VitisPathways	http://momtong.rit.edu/cgi-bin/VitisPathways/vitispathways.cgi/	Enrichment analysis of grapevine metabolic pathways using the VitisNet and GrapeCyc pathway designations.	
GrapeCyc	http://www.plantcyc.org/databases/grapecyc/7.0	Version 7 comprises of 3,191 reactions and 5,791 enzymes, contained in 511 metabolic pathways. Annotation is based on Genoscope.	
PMN	http://www.plantcyc.org/	Release 12 (May 2017) hosts 76 plant species/taxon-specific metabolic pathway databases and one multi-species reference database called PlantCyc.	

Table 2.4: Platforms and tools for high-throughput gene/protein functional assignment.

Platform	Website address	Annotation tool	Website address
Gene Ontology	http://www.geneontology.org/	Blast2GO	https://www.blast2go.com/
KEGG	http://www.genome.jp/kegg/	BlastKOALA	http://www.kegg.jp/blastkoala/
MapMan	http://mapman.gabipd.org	Mercator	http://www.plabipd.de/portal/web/guest/mercator-sequence-annotation

2.4.3 Sequence similarity-based gene functional annotation

One caveat is that most of the above-mentioned tools require a list of known gene identifiers to import in these metabolic networks. However, some platforms allow for batch import of unknown nucleotide or amino acid sequences to annotate and map to networks. Three such platforms: Gene Ontology, Kyoto Encyclopedia of Genes and Genomes, and MapMan are discussed here (Table 2.4).

Gene Ontology (GO) analysis (Table 2.4) is widely used for gene functional annotation and classification (Ashburner et al. 2000; The Gene Ontology Consortium 2008), as it provides standardized terminology for describing gene function. The functions of gene products are defined in three categories, namely molecular function (activities of gene products), cellular component (location of gene product) and biological process (pathways or larger processes involving the activities of multiple gene products). Gene ontologies also include relationships between these gene functions. Such grouping of genes based on functional similarity can enhance biological interpretation. It can also be useful to map genes on metabolic networks to give an indication of biochemical processes the genes of interest are involved in. A tool specifically designed for assigning GOs to unknown sequences is Blast2GO (Conesa and Götz 2008). The CRIBI *Vitis vinifera* release V2.1 contains the latest gene ontology (GO) assignments for the genes in the annotation. Of the 31,845 genes in the V2.1 CRIBI annotation, 26,529 have GO terms assigned to them in the CRIBI GO annotation.

The Kyoto Encyclopedia of Genes and Genomes (KEGG, Table 2.4) release 82.1 (June 1, 2017) hosts 517 reference metabolic pathway maps, and genes and enzymes linked to these pathways, including 5,217 organisms (383 eukaryotes, 4260 bacteria, 252 archaea and 317 viruses). Entries on the pathways are identified with EC (Enzyme commission) and K (KEGG) numbers. *Vitis vinifera* (KEGG reference number: T01084) contains 135 metabolic pathways, 25,843 genes coding for proteins and 2,316 RNA genes. The grapevine annotation in KEGG is also based on the Genoscope identifiers, but KEGG uses Entrez gene identifiers (Entrez is the National Center for Biotechnology Information (NCBI) website search engine). The CRIBI V2.1 grapevine annotation has 11,479 genes that are assigned an EC number, i.e. the protein products of these genes function as enzymes. BlastKOALA (KEGG Orthology And Links Annotation, Table 2.4) can be used to assign unknown sequences a K-number and position them on a metabolic map.

Mercator (May et al. 2008) is a tool hosted on the PlaBi database used for the classification of unknown protein or gene sequences into MapMan (Thimm et al. 2004) functional categories (Table 2.4). Mercator uses BLAST alignment to various databases (TAIR, SwissProt/UniProt plant proteins, JGI Chlamy, TIGR5 rice proteins, Clusters of orthologous eukaryotic genes database (KOG), Conserved domain database and InterproScan) to assign nucleotide or protein sequences to 35 primary (a total of 1,307 bins) functional bins. Each of these bins can be further divided into secondary bins and sub-classifications. MapMan displays these bin assignments of genes onto various metabolic pathway diagrams.

All three of these annotation tools, Blast2GO, BlastKOALA and Mercator, rely on BLAST (Altschul et al. 1990) for similarity searches against known sequences in the respective databases. However, MapMan was specifically designed for the analysis of plant metabolic processes (Klie and Nikoloski 2012) and the functional bin classification system proved to be the most informative. Consequently, it was implemented in this study (Chapter 4 and 5).

2.4.4 Other plant databases

In addition to the above-mentioned annotation tools, a number of tools exist to perform functional enrichment and inter-species comparisons (Table 2.5). Mostly agriculturally important and model plant species are included in these databases. An example is functional cluster analysis (a functional cluster is defined as genes with similar function / involvement in the same biochemical pathway, located in close proximity on the chromosomes) possible with the PLAZA database, as implemented in this study (Chapter 3).

Table 2.5: A restricted list of databases for the genomes and gene ontologies of agricultural crops and model plants. These include genome browsers and tools for inter-species comparisons and metabolic network, enrichment and differential expression analysis.

Platform	Website address
AgriGO	http://bioinfo.cau.edu.cn/agriGO/
Crop ontology	http://www.cropontology.org/ontology/VITIS/Vitis
Ensembl plants	http://plants.ensembl.org/Vitis_vinifera/Info/Index
Gramene	http://www.gramene.org/
plaBi database	http://www.plabipd.de/portal/web/guest/home1
PlantGDB (VvGDB)	http://www.plantgdb.org/VvGDB/
PLAZA	http://bioinformatics.psb.ugent.be/plaza/
Phytozome	https://phytozome.jgi.doe.gov/pz/portal.html#
transPLANT	http://www.transplantdb.eu/

2.5 Concluding remarks

Grapevine is one of the fruit crops humans most successfully domesticated and spread worldwide. However, grapevine is very disease prone, most likely due to inbreeding depression and subsequent loss of resistance genes during the intensive domestication process. Grapevine is susceptible to bacterial, viral and fungal diseases and insect pests, and is consequently among the most heavily sprayed of all crops (Myles et al. 2011). And considering that grapevine has a relatively narrow climate range for optimum production and quality, its production faces a challenge from global climate change. There is also increased pressure to secure sustainable food sources for the ever-growing human population, and arable land must be used responsibly and optimally.

The genetic improvement of crops is exceedingly important to address these challenges, and genomics are now offering breeders new tools and techniques to allow great steps forward in plant breeding. However, currently in South Africa, grapevine cultivar breeding programmes are focused on table grapes and limited wine grape breeding is performed. Continued genetic research will not only help our understanding of the process of grapevine domestication, but will also facilitate genetic conservation and adaptation of grapevine in a changing environment. However, the question remains whether modern molecular technologies for improving grapevine will win the race against increased environmental and biotic pressures and loss of genetic diversity.

2.6 References

Adam-Blondon A-F, Alaux M, Pommier C, Cantu D, Cheng Z-M, Cramer GR, Davies C, Delrot S, Deluc L, Gaspero GD et al. (2016) Towards an open grapevine information system. Hortic Res 3:16056. doi: 10.1038/hortres.2016.56

Alleweldt G and Possingham JV (1988) Progress in grapevine breeding. Theor Appl Genet 75:669–673. doi: 10.1007/BF00265585

Altschul SF, Gish W, Miller W, Myers EW and Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215:403–410. doi: 10.1016/S0022-2836(05)80360-2

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT et al. (2000) Gene Ontology: tool for the unification of biology. Nat Genet 25:25–29. doi: 10.1038/75556

Aubakirova K, Omasheva M, Ryabushkina N, Tazhibaev T, Kampitova G and Galiakparov N (2014) Evaluation of five protocols for DNA extraction from leaves of Malus sieversii, Vitis vinifera, and Armeniaca vulgaris. Genet Mol Res GMR 13:1278–1287. doi: 10.4238/2014.February.27.13

Barabaschi D, Guerra D, Lacrima K, Laino P, Michelotti V, Urso S, Valè G and Cattivelli L (2012) Emerging knowledge from genome sequencing of crop species. Mol Biotechnol 50:250–266. doi: 10.1007/s12033-011-9443-1

Barabaschi D, Tondelli A, Desiderio F, Volante A, Vaccino P, Valè G and Cattivelli L (2016) Next generation breeding. Plant Sci 242:3–13. doi: 10.1016/j.plantsci.2015.07.010

Batley J and Edwards D (2016) The application of genomics and bioinformatics to accelerate crop improvement in a changing climate. Curr Opin Plant Biol 30:78–81. doi: 10.1016/j.pbi.2016.02.002

Bellec A, Courtial A, Cauet S, Rodde N, Vautrin S, Beydon G, Arnal N, Gautier N, Fourment J, Prat E et al. (2016) Long read sequencing technology to solve complex genomic regions assembly in plants. J Gener Seq Appl 2–6. doi: 10.4172/2469-9853.1000128

Benjak A, Forneck A and Casacuberta JM (2008) Genome-wide analysis of the “cut-and-paste” transposons of grapevine. PLOS ONE 3:e3107. doi: 10.1371/journal.pone.0003107

Bevan MW, Uauy C, Wulff BBH, Zhou J, Krasileva K and Clark MD (2017) Genomic innovation for crop improvement. Nature 543:346–354. doi: 10.1038/nature22011

Bolger ME, Weisshaar B, Scholz U, Stein N, Usadel B and Mayer KF (2014) Plant genome sequencing — applications for crop improvement. Curr Opin Biotechnol 26:31–37. doi: 10.1016/j.copbio.2013.08.019

Cardi T (2016) Cisgenesis and genome editing: combining concepts and efforts for a smarter use of genetic resources in crop breeding. Plant Breed 135:139–147. doi: 10.1111/pbr.12345

Chialva C, Eichler E, Muñoz C and Lijavetzky D (2016) Development and use of biotechnology tools for grape functional analysis. Grape Wine Biotechnol 75–101.

Chin C-S, Peluso P, Sedlazeck FJ, Nattestad M, Concepcion GT, Clum A, Dunn C, O’Malley R, Figueroa-Balderas R, Morales-Cruz A et al. (2016) Phased diploid genome assembly with Single Molecule Real-Time Sequencing. Nat Methods 13:1050–1054. doi: 10.1038/nmeth.4035

Conesa A and Götz S (2008) Blast2GO: A comprehensive suite for functional analysis in plant genomics. Int J Plant Genomics. doi: 10.1155/2008/619832

Da Silva C, Zamperin G, Ferrarini A, Minio A, Molin AD, Venturini L, Buson G, Tononi P, Avanzato C, Zago E et al. (2013) The high polyphenol content of grapevine cultivar Tannat berries is conferred primarily by genes that are not shared with the reference genome. Plant Cell 25:4777–4788. doi: 10.1105/tpc.113.118810

Di Genova A, Almeida AM, Muñoz-Espinoza C, Vizoso P, Travisany D, Moraga C, Pinto M, Hinrichsen P, Orellana A and Maass A (2014) Whole genome comparison between table and wine grapes reveals a comprehensive catalog of structural variants. BMC Plant Biol 14:7. doi: 10.1186/1471-2229-14-7

Eid J, Fehr A, Gray J, Luong K, Lyle J, Otto G, Peluso P, Rank D, Baybayan P, Bettman B et al. (2009) Real-Time DNA Sequencing from Single Polymerase Molecules. Science 323:133–138. doi: 10.1126/science.1162986

Feuillet C, Leach JE, Rogers J, Schnable PS and Eversole K (2011) Crop genome sequencing: lessons and rationales. Trends Plant Sci 16:77–88. doi: 10.1016/j.tplants.2010.10.005

Fournier-Level A, Lacombe T, Le Cunff L, Boursiquot J-M and This P (2010) Evolution of the VvMybA gene family, the major determinant of berry colour in cultivated grapevine (Vitis vinifera L.). Heredity 104:351–362. doi: 10.1038/hdy.2009.148

Gambino G, Perrone I and Gribaudo I (2008) A Rapid and effective method for RNA extraction from different tissues of grapevine and other woody plants. Phytochem Anal 19:520–525. doi: 10.1002/pca.1078

Goff SA, Ricke D, Lan T-H, Presting G, Wang R, Dunn M, Glazebrook J, Sessions A, Oeller P, Varma H et al. (2002) A draft sequence of the rice genome (Oryza sativa L. ssp. japonica). Science 296:92–100. doi: 10.1126/science.1068275

Gregory TR (2005) The C-value enigma in plants and animals: A review of parallels and an appeal for partnership. Ann Bot 95:133–146. doi: 10.1093/aob/mci009

Gregory TR, Nicol JA, Tamm H, Kullman B, Kullman K, Leitch IJ, Murray BG, Kapraun DF, Greilhuber J and Bennett MD (2007) Eukaryotic genome size databases. Nucleic Acids Res 35:D332–D338. doi: 10.1093/nar/gkl828

Grimplet J, Cramer GR, Dickerson JA, Mathiason K, Van Hemert J and Fennell AY (2009) VitisNet: “Omics” integration through grapevine molecular networks. PloS One 4:e8365. doi: 10.1371/journal.pone.0008365

Gupta P, Naithani S, Tello-Ruiz MK, Chougule K, D’Eustachio P, Fabregat A, Jiao Y, Keays M, Lee YK, Kumari S et al. (2016) Gramene database: Navigating plant comparative genomics resources. Curr Plant Biol 7:10–15. doi: 10.1016/j.cpb.2016.12.005

Imazio S, Maghradze D, Lorenzis GD, Bacilieri R, Laucou V, This P, Scienza A and Failla O (2013) From the cradle of grapevine domestication: molecular overview and description of Georgian grapevine (Vitis vinifera L.) germplasm. Tree Genet Genomes 9:641–658. doi: 10.1007/s11295-013-0597-9

Iwata H, Minamikawa MF, Kajiya-Kanegae H, Ishimori M and Hayashi T (2016) Genomics-assisted breeding in fruit trees. Breed Sci 66:100–115. doi: 10.1270/jsbbs.66.100

Jaillon O, Aury J-M, Noel B, Policriti A, Clepet C, Casagrande A, Choisne N, Aubourg S, Vitulo N, Jubin C et al. (2007) The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. Nature 449:463–467. doi: 10.1038/nature06148

Klie S and Nikoloski Z (2012) The choice between MapMan and Gene Ontology for automated gene function prediction in plant science. Front Genet. doi: 10.3389/fgene.2012.00115

Kobayashi S, Goto-Yamamoto N and Hirochika H (2004) Retrotransposon-induced mutations in grape skin color. Science 304:982–982. doi: 10.1126/science.1095011

Lacombe T, Audeguin L, Boselli M, Bucchetti B, Cabello F, Chatelet P, Crespan M, D'Onofrio C, Dias JE, Ercisli S et al. (2015) Grapevine European Catalogue: Towards a comprehensive list. VITIS - J Grapevine Res 50:65.

Malnoy M, Viola R, Jung M-H, Koo O-J, Kim S, Kim J-S, Velasco R and Nagamangala Kanchiswamy C (2016) DNA-Free genetically edited grapevine and apple protoplast using CRISPR/Cas9 ribonucleoproteins. Front Plant Sci. doi: 10.3389/fpls.2016.01904

May P, Wienkoop S, Kempa S, Usadel B, Christian N, Rupprecht J, Weiss J, Recuenco-Munoz L, Ebenhöh O, Weckwerth W et al. (2008) Metabolomics- and proteomics-assisted genome annotation and analysis of the draft metabolic network of Chlamydomonas reinhardtii. Genetics 179:157–166. doi: 10.1534/genetics.108.088336

McGovern PE (2013) Ancient wine: The search for the origins of viniculture. Princeton University Press

McGovern PE, Fleming SJ and Katz SH (2003) The Origins and Ancient History of Wine: Food and Nutrition in History and Antropology. Routledge

Meyers LA, Levin DA and Geber M (2006) On the abundance of polyploids in flowering plants. Evolution 60:1198–1206. doi: 10.1554/05-629.1

Minio A, Lin J, Gaut BS and Cantu D (2017) How Single Molecule Real-Time Sequencing and haplotype phasing have enabled reference-grade diploid genome assembly of wine grapes. Front Plant Sci. doi: 10.3389/fpls.2017.00826

Morrell PL, Buckler ES and Ross-Ibarra J (2011) Crop genomics: advances and applications. Nat Rev Genet. doi: 10.1038/nrg3097

Myles S, Boyko AR, Owens CL, Brown PJ, Grassi F, Aradhya MK, Prins B, Reynolds A, Chia J-M, Ware D et al. (2011) Genetic structure and domestication history of the grape. Proc Natl Acad Sci 108:3530–3535. doi: 10.1073/pnas.1009363108

Péros J-P, Launay A, Berger G, Lacombe T and This P (2015) MybA1 gene diversity across the Vitis genus. Genetica 143:373–384. doi: 10.1007/s10709-015-9836-3

Pulvirenti A, Giugno R, Distefano R, Pigola G, Mongiovi M, Giudice G, Vendramin V, Lombardo A, Cattonaro F and Ferro A (2015) A knowledge base for Vitis vinifera functional analysis. BMC Syst Biol 9:S5. doi: 10.1186/1752-0509-9-S3-S5

Ren C, Liu X, Zhang Z, Wang Y, Duan W, Li S and Liang Z (2016) CRISPR/Cas9-mediated efficient targeted mutagenesis in Chardonnay (Vitis vinifera L.). Sci Rep. doi: 10.1038/srep32289

Reynolds AG (2010) Managing Wine Quality: Viticulture and Wine Quality. Elsevier

Roberts RJ, Carneiro MO and Schatz MC (2013) The advantages of SMRT sequencing. Genome Biol 14:405. doi: 10.1186/gb-2013-14-7-405

Salzman RA, Fujita T, Zhu-Salzman K, Hasegawa PM and Bressan RA (1999) An Improved RNA Isolation Method for Plant Tissues Containing High Levels of Phenolic Compounds or Carbohydrates. Plant Mol Biol Report 17:11–17. doi: 10.1023/A:1007520314478

Scheben A, Yuan Y and Edwards D (2016) Advances in genomics for adapting crops to climate change. Curr Plant Biol 6:2–10. doi: 10.1016/j.cpb.2016.09.001

Scossa F, Brotman Y, de Abreu e Lima F, Willmitzer L, Nikoloski Z, Tohge T and Fernie AR (2016) Genomics-based strategies for the use of natural variation in the improvement of crop metabolism. Plant Sci 242:47–64. doi: 10.1016/j.plantsci.2015.05.021

Shi J and Lai J (2015) Patterns of genomic changes with crop domestication and breeding. Curr Opin Plant Biol 24:47–53. doi: 10.1016/j.pbi.2015.01.008

Shimazaki M, Fujita K, Kobayashi H and Suzuki S (2011) Pink-Colored Grape Berry Is the Result of Short Insertion in Intron of Color Regulatory Gene. PLOS ONE 6:e21308. doi: 10.1371/journal.pone.0021308

Swinnen G, Goossens A and Pauwels L (2016) Lessons from domestication: Targeting cis-regulatory elements for crop improvement. Trends Plant Sci 21:506–515. doi: 10.1016/j.tplants.2016.01.014

The Gene Ontology Consortium (2008) The Gene Ontology project in 2008. Nucleic Acids Res 36:D440–D444. doi: 10.1093/nar/gkm883

Thimm O, Bläsing O, Gibon Y, Nagel A, Meyer S, Krüger P, Selbig J, Müller LA, Rhee SY and Stitt M (2004) MapMan: a user-driven tool to display genomics data sets onto diagrams of metabolic pathways and other biological processes. Plant J 37:914–939. doi: 10.1111/j.1365-313X.2004.02016.x

This P, Lacombe T and Thomas MR (2006) Historical origins and genetic diversity of wine grapes. Trends Genet 22:511–519. doi: 10.1016/j.tig.2006.07.008

Velasco R, Zharkikh A, Troggio M, Cartwright DA, Cestaro A, Pruss D, Pindo M, FitzGerald LM, Vezzulli S, Reid J et al. (2007) A high quality draft consensus sequence of the genome of a heterozygous grapevine variety. PLOS ONE 2:e1326. doi: 10.1371/journal.pone.0001326

Vitulo N, Forcato C, Carpinelli EC, Telatin A, Campagna D, D’Angelo M, Zimbello R, Corso M, Vannozzi A, Bonghi C et al. (2014) A deep survey of alternative splicing in grape reveals changes in the splicing machinery related to tissue, stress condition and genotype. BMC Plant Biol 14:99. doi: 10.1186/1471-2229-14-99

Vivier M and Pretorius I (2000) Genetic improvement of grapevine: tailoring grape varieties for the third millennium - a review. South Afr J Enol Vitic 21:5–26.

Wang Y, Liu X, Ren C, Zhong G-Y, Yang L, Li S and Liang Z (2016) Identification of genomic sites for CRISPR/Cas9-based genome editing in the Vitis vinifera genome. BMC Plant Biol 16:96. doi: 10.1186/s12870-016-0787-3

Wilkinson MD, Dumontier M, Aalbersberg IjJ, Appleton G, Axton M, Baak A, Blomberg N, Boiten J-W, Santos LB da S, Bourne PE et al. (2016) The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 3:160018. doi: 10.1038/sdata.2016.18

Wong DC, Sweetman C, Drew DP and Ford CM (2013) VTCdb: a gene co-expression database for the crop species Vitis vinifera (grapevine). BMC Genomics 14:882. doi: 10.1186/1471-2164-14-882

Yakushiji H, Kobayashi S, Goto-Yamamoto N, Tae Jeong S, Sueta T, Mitani N and Azuma A (2006) A skin color mutation of grapevine, from black-skinned Pinot Noir to white-skinned Pinot Blanc, is caused by deletion of the functional VvmybA1 allele. Biosci Biotechnol Biochem 70:1506–1508. doi: 10.1271/bbb.50647

Yu J, Hu S, Wang J, Wong GK-S, Li S, Liu B, Deng Y, Dai L, Zhou Y, Zhang X et al. (2002) A draft sequence of the rice genome (Oryza sativa L. ssp. indica). Science 296:79–92. doi: 10.1126/science.1068037

Yuan Y, Bayer PE, Batley J and Edwards D (2017) Improvements in genomic technologies: Application to crop genomics. Trends Biotechnol 35:547–558. doi: 10.1016/j.tibtech.2017.02.009

Zhou Y, Massonnet M, Sanjak J, Cantu D and Gaut BS (2017) The evolutionary genomics of grape (Vitis vinifera ssp. vinifera) domestication. bioRxiv 146373. doi: 10.1101/146373

Online resources

CRIBI (Centro di Ricerca Interdipartimentale per le Biotecnologie Innovative), University of Padua, Italy: http://genomes.cribi.unipd.it/grape/

PACBIO®: http://www.pacb.com/smrt-science/smrt-sequencing/

plaBi database: http://www.plabipd.de/

Vitis International Variety Catalogue: http://www.vivc.de/

Chapter 3: *De novo* Assembly of the Pinotage Draft Genome

3.1 Introduction

Grapevine breeders are continuously aiming to improve existing cultivars and to develop new cultivars for the growing viticulture market. Most new cultivars today arise from crosses between existing cultivars, or hybridization with other *Vitis* species, to harness the positive characteristics of both parents. One such cultivar is the South African bred red wine cultivar, Pinotage. Pinotage is the result of a viticultural cross between Pinot noir and Cinsaut (Cinsaut was called Hermitage in South Africa, hence the name Pinotage) performed in 1925 by Professor A.I. Perold, at Stellenbosch University, South Africa.

Pinot noir is a popular red wine grape, widely cultivated in grape-growing regions worldwide. It is a noble cultivar, originating in France, and is known to produce high-quality wines (Bowers et al. 1999). Pinot noir is also planted in South Africa, but there was a need for a cultivar that can better withstand the hot and dry South African conditions. Therefore Cinsaut, known for its heat-tolerance and higher yield, was selected as the other crossing parent (Pinotage Association: online resources). Cinsaut is an ancient cultivar and its exact origins are unknown. This cultivar is called by many different names (the *Vitis* International Variety Catalogue lists 101 synonyms for Cinsaut) in different regions, e.g. Hermitage in South Africa, Cinsualt in France and Black Malvoisie in California. It is widely planted in the South of France and is almost exclusively used to blend with other cultivars.

The Pinot noir (presumably used as mother) and Cinsaut (presumably used as father) yielded 4 seeds. The four seeds were planted and the young vine later grafted onto rootstocks. One of the grafted plants performed remarkably well and was selected as the mother material of all Pinotage vines (Pinotage Association: online resources). The first Pinotage wine was only made in 1941, while commercial planting of Pinotage started in 1943. The name “Pinotage” first appeared on a wine label in 1961 (Pinotage Association: online resources). However, the first wines made from this cultivar did not fare so well, as it had an intense acetone flavour. Research on Pinotage has focused on development of vinification techniques suitable for the unique characteristics of Pinotage, and today winemakers tend to do pre-fermentation maceration at cooler temperatures (Marais 2003b; Marais 2003c; Marais 2003a; De Beer et al. 2017) to limit the formation of volatile esters that convey the acetone flavour.

The oldest existing Pinotage vineyard is a 66-year-old 0.47ha untrained vineyard in Stellenbosch, South Africa (De Waal Wines: online resources). Today, Pinotage is a popular cultivar in South Africa and is used for the production of premium wines. It is the third most planted red wine cultivar in South Africa, representing 7.9% of total vineyard area (WOSA: online resources), and is an integral part of the South African viticulture and winemaking history. Besides South Africa, Pinotage is also planted in New Zealand and Brazil (Anderson and Aryal 2015).

Although the grapevine genome was published in 2007 (Jaillon et al. 2007), there is an increasing awareness that one reference genome sequence is not sufficient to encompass all the variability within a species. Consequently, there is a need for the sequencing of additional genomes of other varieties/cultivars or genotypes. A complete assembled and annotated genome sequence, ideally with the position of variants and genetic markers indicated, is the ultimate genomic resource for genetic studies and applied genetics such as crop breeding. Since the publication of the grapevine genome, more grapevine sequencing projects have been launched (grapevine genome sequencing is discussed in Section 2.4.1). Building on this ever-growing list of grapevine genome sequencing projects, this is the first report of the genome sequencing of *Vitis vinifera* cv Pinotage. Pinotage was chosen for this study due to its importance in South African viticulture, but also to leverage the genetic data available for Pinot noir. Due to its close relation to Pinotage, the Pinot noir genome sequence would be highly suitable to contrast and compare to the Pinotage genome. The next-generation sequencing and assembly strategies used to obtain a draft Pinotage genome, are explained, and the sequence variant distribution between Pinotage and Pinot noir, analysed. This genomic and variant information can aid in the identification of agronomically important genes and accelerate genetic studies and new clone/cultivar selection programs.

3.2 Methods and Materials

3.2.1 Sample collection and DNA extraction

Canes and leaves were harvested from five vines (*V. vinifera* cv Pinotage, clone 6) Stellenbosch, South Africa. These vines were established from virus-free meristem cultures. The sample material was pooled, the bark removed from the canes, and phloem harvested. Thin cane shavings (a combination of phloem and xylem material) were also collected. The phloem, phloem/xylem, and leaf material were stored at -80°C.

DNA was extracted from phloem, phloem and xylem combined, and leaves. A modified cetyltrimethylammonium bromide (CTAB) method was used. Frozen sample material (1g) was powdered in liquid nitrogen and incubated for 15 min in CTAB buffer (2% [w/v] CTAB, 2.5% [w/v] PVP-10, 100mM Tris-HCl pH 8, 1.4M NaCl, 20mM EDTA pH 8 and 3% [v/v] β-mercaptoethanol) at 65°C. DNA was treated with 20mg RNase A (Thermo Scientific) and incubated for 15 min at 37°C, followed by three chloroform-isoamyl alcohol (24:1) extraction steps. The DNA was then precipitated with isopropanol and washed with 70% ethanol. DNA quality was assessed by gel-electrophoresis and quantified using the Trinean Xpose (Gentbrugge, Belgium).

3.2.2 DNA library preparation and sequencing

DNA sequencing was performed at the Agricultural Research Council, Biotechnology platform, Pretoria, South Africa. Before preparing the libraries, the DNA quality was assessed using the Agilent 2100 Bioanalyzer (Agilent Technologies, Santa Clara, USA).

The first set of DNA sequencing libraries were prepared using the NEBNext® Ultra™ II DNA Library Preparation Kit for Illumina® (NEBNext® Ultra™ kit: online resources). This included a fragmentation step using the NEBNext® dsDNA fragmentase enzyme that generates dsDNA breaks in a time-dependent manner.

A second set of sequencing libraries was created with the Illumina Nextera® DNA Library Preparation Kit (Nextera® DNA library kit: online resources). Included in the kit is the Nextera® transposome, used to tagment gDNA, a process that fragments and tags the DNA with adapter sequences, in a single step.

The same DNA sample was used to create a third set of sequencing libraries, with the Illumina TruSeq® paired-end PCR-free kit (Illumina TruSeq® library kit: online resources). Fractionation was performed using a Covaris® (Woburn, Massachusetts, USA).

For all three sequencing sets, size selection was performed using SPRI (Solid Phase Reversible Immobilization) beads (Beckman Coulter®, Brea, California, USA) and three libraries, with insert sizes 300, 500 and 700bp respectively, were prepared. All libraries were sequenced using the Illumina HiSeq2500 (version 4 chemistry) to generate 125bp paired-end reads.

3.2.3 DNA sequence quality assessment and trimming

Reads were assessed for quality with FastQC (Andrews et al. 2011). Trimmomatic (Bolger et al. 2014a) was used to remove adaptor sequences. The first 19nt of the reads showed a nucleotide composition imbalance, and were removed. Reads were trimmed from the 5'-end for a minimum average Phred score of Q20 over a window of 3nt. Only sequences with a minimum length of 50nt and unbroken read pairs were retained.

An in-house script was used to assess the insert size distribution of each library. The script automates alignment of reads to the *V. vinifera* chloroplast and mitochondrion (NCBI: NC007957 and NC012119) genomes using Bowtie2 (Langmead and Salzberg 2012), allowing only one mismatch per 20nt seed length (Parameters: -N 1), and not imposing paired-end distances. The template length can then be determined from the alignment for each read pair, and the average template length calculated for the library.

3.2.4 Assembly of DNA sequencing data

Error correction was performed on the reads using the SOAPdenovo error correction module (SOAPec_v2.01, Beijing Genomics Institute, http://soap.genomics.org.cn/index.html), with default parameters. The error-corrected reads were assembled using SOAPdenovo v2.04 (Luo et al. 2012) using default parameters and a k-mer minimum of 79 and maximum of 121. Multiple assemblies including different combinations of libraries were performed and the best assemblies were selected to continue analysis (Figure 3.1).

3.2.5 Sample collection and RNA extraction

The same vines used for the genome sequencing (Section 3.2.1), were sampled in the following growing season for RNA extraction. Canes were collected, bark removed and phloem scrapings harvested. The phloem tissue was stored at -80°C.

One gram of frozen phloem material was powdered in liquid nitrogen with a mortar and pestle and total RNA extracted using the protocol from Reid et al. (2006). Total RNA (15µg) was treated with RQ1 RNase-free DNase (Promega, Madison, USA) in 50µl reactions according to the manufacturer's instructions. After incubation at 37°C for 30 min, the reaction volume was adjusted to 500µl with 10mM Tris-HCl (pH 8.5). An acidic phenol extraction was performed, followed by a chloroform-isoamyl alcohol (24:1) extraction. The RNA was precipitated with ethanol and sodium acetate (2.5 volumes 100% ethanol and 0.1 volumes 3M sodium acetate, pH 5.2). After a wash step with 70% ethanol, pellets were dried and resuspended in 30µl ddH$_2$O. To ascertain RNA integrity, an aliquot of RNA extract was analysed by gel-electrophoresis and quantified using the Trinean Xpose (Gentbrugge, Belgium).

3.2.6 RNA library preparation and sequencing

The RNA was shipped on dry ice to the sequencing facility (Agricultural Research Council, Biotechnology platform, Pretoria, South Africa) and the quality assessed using the Agilent RNA Nano 6000 kit on the Agilent 2100 Bioanalyzer (Agilent Technologies, Santa Clara, USA). A ribosome depleted RNA library was prepared according to the manufacturer's instructions for the Illumina TruSeq® Stranded Total RNA library preparation kit with Ribo-Zero^{TM} Plant (Illumina TruSeq® Stranded Ribo-Zero^{TM} Plant kit: online resources) and sequenced on the Illumina HiSeq2500 generating 125bp paired-end reads.

3.2.7 RNA sequence quality assessment and trimming

RNA sequence data was assessed for quality using FastQC (Andrews et al. 2011). Trimmomatic (Bolger et al. 2014a) was used to remove adaptor sequences. The first 9nt from each read were also removed, as these nucleotides showed a nucleotide composition imbalance. Reads were trimmed from the 5'-end for an average minimum Phred score of Q20 over a sliding window of 3nt. Only unbroken read pairs and reads with a minimum length of 20nt, were retained.

3.2.8 Assembly of RNA sequence data

Quality trimmed RNA sequence data from the Ribo-Zero sequencing library were *de novo* assembled into putative transcripts using Trinity (Grabherr et al. 2011) with default parameters (Haas et al. 2013), specifying that it is a strand-specific RNA-seq library generated with the dUTP method (Parameters: --SS_lib_type RF).

3.2.9 Merging RNA and DNA assemblies

Scaffolds and contigs from the different DNA assemblies (Section 3.2.4) were merged in a step-by-step approach using GARM (Genome Assembler, Reconciliation and Merging) (Soto-Jimenez et al. 2014). The genome assembly with the highest N$_{50}$ was chosen as the scaffolding assembly. Contigs from two more genome assemblies were merged with the first, and finally the contigs obtained from the Trinity assembly of Ribo-Zero data (Section 3.2.8) were also merged (Figure 3.1).

Only contigs with a BLAST (Altschul et al. 1990) alignment e-value of 0.001 or less against land plants (taxid: 3193) in the NCBI nucleotide database (National Center for Biotechnology Information, https://blast.ncbi.nlm.nih.gov), were selected for further analysis. These contigs represent the draft genome sequence of Pinotage.

3.2.10 Variant analysis

The trimmed and error-corrected reads from all the DNA libraries (Section 3.2.3) were aligned to the Pinot noir PN40024 reference grapevine genome (http://genomes.cribi.unipd.it/grape/) using Bowtie2 (Langmead and Salzberg 2012) (Parameters: -N 1 --no-mixed --no-discordant --no-unal -X 1000), allowing only proper pairs. Alignments of the respective libraries were performed separately and each chromosome was treated individually. The resulting output files from the different libraries were combined per chromosome and filtered with Samtools (Li et al. 2009), allowing only reads with a mapping quality score of more than 30 (Parameters: samtools view -q30), to avoid subsequent calling of erroneous variants due to low-quality mapping or collapsed repetitive sequences.

The bam alignment files (output from Samtools) were converted to variant calling format (vcf) files and filtered using bcftools (part of the Samtools package). Only variants with a quality score higher than 30 and covered by an alignment depth of two or more, were allowed to pass (Parameters: bcftools filter -i 'QUAL>30 && DP>1'). Filtered vcf files for the 19 chromosomes were combined and processed by SnpEff (Cingolani et al. 2012) to call variants (single nucleotide polymorphisms (SNPs) and short insertions or deletions (indels)). The reference V2.1 grapevine annotation (http://genomes.cribi.unipd.it/grape/V2.1.gff3) was included in the variant calling to allow for classification of the variants in regions (intergenic, exon, intron, splice-site, 5’UTR and 3’UTR) and impact (low, moderate, high, modifier). From the SnpEff vcf output file, the number of variants per 10,000nt interval was calculated with vcftools (Parameters: vcftools --SNPdensity 10000) (Danecek et al. 2011). The variants with high impact effects were used for further analysis (Section 3.2.11).

To calculate the depth of coverage over the reference genome, the Bowtie2 alignment files were subjected to the same criteria as for the variant calling; allowing only reads with a quality score and a mapping quality score higher than 30 (Parameters: samtools depth -q30 -Q30). The mean depth of coverage was calculated over the same genomic intervals as for the variant calling, using bedtools (Parameters: bedtools coverage –mean). (Quinlan laboratory, University of Utah, http://bedtools.readthedocs.io/en/latest/index.html).

3.2.11 Functional cluster analysis

Chromosome positions for the predicted *Vitis vinifera* functional gene clusters were retrieved from the PLAZA (Proost et al. 2009; Van Bel et al. 2011) database (online resources). Functional gene clusters are predicted by the proximity of functionally related genes. Functionality of genes can be based upon different functional annotations, for example gene ontology, InterPro annotations or MapMan functional bin assignments. MapMan-based functional clustering was selected, since it was also used in other parts of this study (Chapter 4 and 5). The experiment with the least strict parameters (PLAZA functional clustering experiment 17), with a predicted 445 functional clusters, was selected.

PLAZA uses the 12X V0 *Vitis vinifera* annotation from Genoscope, whereas the latest V2.1 annotation from CRIBI was used in this study. Therefore, the gene names associated with the functional clusters were not used, but rather just the chromosome positions of the clusters. Names of genes with high impact variants were retrieved from the SnpEff output, their chromosomal positions obtained from the V2.1 grapevine annotation (http://genomes.cribi.unipd.it/grape/V2.1.gff3), and matched to the PLAZA functional clusters to assign them to a functional cluster where possible.

3.3 Results and Discussion

3.3.1 DNA extraction, library preparation and sequencing

Woody perennials, like grapevine, have high concentrations of polysaccharides, polyphenolics and other components that can bind and co-precipitate with nucleic acids during nucleic acid extraction, thereby influencing the quality and yield, or interfering with downstream applications (Salzman et al. 1999; Gambino et al. 2008; Aubakirova et al. 2014). In this study, a standard CTAB DNA extraction method was used and optimized for extraction of high quality DNA with minimal co-purified contaminants. The optimal protocol had only one nucleic acid precipitation step with isopropanol and included an RNase step. Different tissue types including leaves, petioles, phloem scrapings and a combination of phloem and xylem were tested.

Extraction from the phloem/xylem combination yielded a lower DNA concentration, but with less co-purified contaminants, and was used for further analyses.

Three library preparation protocols were used for genome sequencing, two with enzymatic fractionation (NEBNext® dsDNA fragmentase and Nextera® transposome) and one using mechanical fractionation (Covaris). Library size selection was performed with either SPRIselect® or AMPure® XP beads. Three libraries, with insert sizes 300, 500 and 700bp respectively, were prepared during all three sequencing rounds. During the first sequencing round, the 700bp insert library failed and was discarded. After sequencing, the insert size distribution of each library was assessed by aligning the sequencing data to the *V. vinifera* chloroplast and mitochondrion and calculating the distance between reads of a pair. It is important to have an as uniform as possible insert size distribution, since the assembler will anchor contigs together to create scaffolds, using the known distance between reads of a pair. Even though the library preparation, fractionation and size selection can be optimised for the desired insert size, it might be necessary to calculate the actual library insert size, after sequencing. The insert sizes calculated were used in the subsequent *de novo* assembly (Table 3.1).

More than 359 million read pairs were generated. Strict quality filtering criteria were used, retaining only the portion of a read with a minimum average Phred score of Q20 over a window of 3nt. Broken pairs and reads shorter than 50nt were discarded. After read error correction, only 48.34% read pairs were retained (Table 3.1).

Table 3.1: The number of sequencing read pairs obtained for the libraries in the respective sequencing rounds, calculated insert size distribution and number of read pairs retained after trimming, filtering and error correction.

			Number of read pairs		
Sequencing round	**Library insert size**	**Determined insert size**	**Raw read pairs**	**After trimming and filtering**	**After error correction**
Round 1	300	106	26,321,156	10,347,082	10,222,262
	500	194	38,194,181	28,129,755	27,557,700
Round 2	300	294	76,041,281	26,141,350	26,007,214
	500	422	77,728,504	16,080,441	16,016,188
	700	145	19,15,458	3,938,891	3,926,281
Round 3	300	272	39,717,744	34,074,665	33,648,821
	500	488	57,045,467	42,120,979	41,785,400
	700	657	25,623,135	14,923,814	14,775,066
Total			**359,826,926**	**175,756,977**	**173,938,932**

3.3.2 RNA extraction, library preparation and sequencing

RNA was extracted from phloem material, collected from the same vines that were used for genome sequencing. Extracted RNA had an RNA integrity number (RIN) of 7.8 and was used for a Ribo-Zero library preparation and sequencing. A total of 277,090,011 sequencing reads were received and after quality trimming and filtering, 244,888,249 (88.38%) were retained.

3.3.3 *De novo* genome assembly

Some genomes, especially those of plants, are intrinsically more difficult to assemble than others, due to their size, high levels of heterozygosity, polyploidy, and the presence of repeats and transposable elements (the challenges of plant genome sequencing are discussed in greater detail in Section 2.3.2). A reference-based approach was successfully used to assemble multiple *Arabidopsis thaliana* lines (Gan et al. 2011). However, in the case of grapevine the genome organization is more complex and cultivars are more divergent. Therefore, a *de novo* assembly approach was selected for the Pinotage genome assembly.

Several different assemblies of the Pinotage genome were attempted. The influence of different assembly parameters and addition of sequencing libraries from the three sequencing rounds, on the assembly outcome, were tested. The selected assemblies were merged to produce the final draft genome sequence of Pinotage. This strategy is depicted in Figure 3.1. It is important to note that the best assembly is not the assembly with the highest N_{50} or largest number of contigs, but rather those with the fewest assembly errors, e.g. erroneously concatenated contigs (Ekblom and Wolf 2014). Therefore, not only the N_{50} of the assemblies was evaluated, but also the number of transcripts that could be successfully mapped to the assembled contigs (discussed in Chapter 5, Section 5.2.3).

The first assembly (Assembly A, Figure 3.1) was performed by including data from all the sequencing libraries from the three sequencing rounds. However, due to the wide distribution of insert sizes in some of the libraries, including these can be detrimental to the assembly. Therefore, for the second assembly (Assembly B, Figure 3.1) only libraries with the tightest insert distribution (the 300 and 500bp libraries) were selected. Although more contigs were obtained in Assembly B, due to the limited sequencing depth provided by these two libraries, the contig lengths did not supersede those of Assembly A, as expected. For the third assembly (Assembly C, Figure 3.1) all the libraries, except those with the widest insert size distribution (the 700bp libraries) were included. Assembly C produced the largest number of contigs, but the most fragmented assembly. From all the different assembly attempts, those described here (Assemblies A, B and C) allow the largest number of transcripts to align.

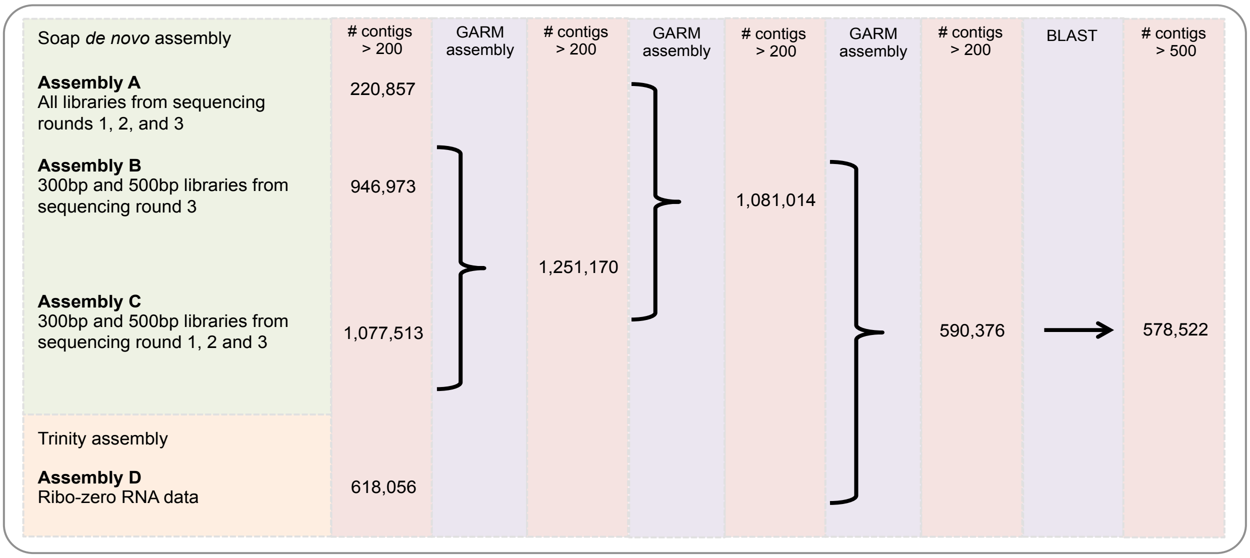


***Figure 3.1:** De novo assembly strategy used to obtain Pinotage draft genome. Different assemblies (performed with SOAPdenovo and Trinity for DNA and RNA, respectively) were merged with GARM. At each step, the number of contigs retained is indicated. The contigs were subjected to BLAST and non-plant hits and contigs shorter than 500nt were discarded.*

Different assemblers, parameters and sequencing libraries all have their strong suites when assembling. Therefore, one approach to genome assembly is to perform different assemblies, and then merge them. In this study, GARM (Genome Assembler, Reconciliation and Merging), a pipeline suitable for merging assemblies from different assembly attempts (different software and parameters), was used. The assemblies were merged using GARM, as indicated in Figure 3.1, producing more than 1 million contigs.

Furthermore, RNA-seq data was generated from the same vines used for DNA sequencing. A *de novo* RNA-seq assembly was performed with Trinity (Assembly D, Figure 3.1). The RNA-seq library was prepared from total RNA, reducing only the ribosomal RNA content in the sample by ribosome depletion, thereby including premature RNAs (before splicing the gene introns) and other functional RNAs. This is as opposed to RNA-seq library preparation performed by polyA selection, enriching for mature RNAs, therefore representing only expressed genes (e.g. the transcriptome data used in Chapter 4). Including the sequencing data from the Ribo-Zero library can therefore improve the genome assembly. Contigs from this assembly were merged with the contigs from the SOAP assemblies using GARM. The contigs from the DNA SOAP assemblies and those from the RNA Trinity assembly were collapsed to 590,376 contigs (Figure 3.1).

These contigs were subjected to BLAST to remove non-plant hits, and contigs smaller than 500nt were discarded. The remaining 578,522 contigs have an N_{50} of 2,366 and the longest is 59,856nt. These contigs represent the draft genome sequence of Pinotage. Despite a reasonable estimated sequencing depth of ~87X, the short read length, and lack of mate-pair or other scaffolding data, prevented a high resolution of the repetitive regions, resulting in a fragmented assembly.

3.3.4 Alignment of Pinotage sequence data to the Pinot noir reference genome

In order to evaluate genome coverage, the Pinotage sequence read data were aligned to the reference Pinot noir PN40024 genome sequence. The alignment results are summarized in Table 3.2 (average depth of coverage and % of chromosome covered), and Figure 3.2A is a visual representation of the alignment (the mean coverage depth over 10kb intervals). Of the total number of reads, 50.21% could be aligned to the 19 Pinot noir reference chromosomes. The remaining reads might be from areas dissimilar to Pinot noir, from the chloroplast or mitochondrial genomes or the 14 unanchored random chromosomes (excluded from this analysis due to the high number of unknown nucleotides [Ns] used to connect the contigs in these unanchored chromosomes), or contaminating sequences. The reads might also have been disqualified from the assembly due to low base quality (below quality score of 30) or mapping quality (below mapping quality score of 30). It is important to note that the calculation of the mapping quality score takes into account the uniqueness of the mapping; therefore, if a read originated from a repetitive region it would have been discarded, to ensure only high quality, unique mappings.

The 19 reference chromosomes have a total length of 426,176,009nt, of which 89.05% were covered with at least one read. At 96.1% and 76.54% coverage, chromosomes 6 and 9 were the most and least covered chromosomes, respectively. The average coverage depth is 42.62X, varying between 36.29X in chromosome 9 and 51.51X in chromosome 17. In another study, only 54.7% of the PN40024 reference sequence was covered by Corvina sequencing data, but a higher read depth threshold of three, was used (Venturini et al. 2013).

Table 3.2: Statistics of the alignment of Pinotage sequencing data to the reference Pinot noir PN40024 sequence. The average read depth and % of chromosome covered were calculated from read alignment data. Variants were called with bcftools and analysed with SnpEff.

		Pinotage sequencing data alignment statistics		
Chromosome number	**Total reference length (nt)**	**Average depth of coverage (X)**	**% of chromosome covered**	**Variant density (1 in X bp)**
1	23,037,639	41.27	92.10	91
2	18,779,844	46.49	92.56	112
3	19,341,862	42.18	88.14	112
4	23,867,706	41.12	90.90	93
5	25,021,643	38.17	87.38	98
6	21,508,407	50.66	96.10	124
7	21,026,613	44.87	90.79	102
8	22,385,789	49.44	95.85	106
9	23,006,712	36.29	76.54	98
10	18,140,952	42.69	86.74	104
11	19,818,926	42.09	83.69	95
12	22,702,307	41.43	83.18	87
13	24,396,255	40.54	86.69	103
14	30,274,277	39.80	87.72	109
15	20,304,914	42.99	91.83	115
16	22,053,297	40.63	87.44	120
17	17,126,926	51.51	95.64	122
18	29,360,087	43.82	93.34	123
19	24,021,853	38.80	86.75	119
Total/Average	**426,176,009**	**42.62**	**89.05**	**106**

The visual representation of coverage depth across the 19 chromosomes (Figure 3.2A) highlights islands of high and low Pinotage/Pinot noir sequence similarity. The areas of the reference chromosomes not covered with Pinotage reads might be due to low similarity, i.e. Pinotage areas inherited from the Cinsaut parent. Similarly, structural variation can cause areas of low or no alignment depth. Although a mapping quality threshold of 30 was implemented, ensuring unique mappings, highly repetitive areas and transposons can still be a reason for areas with greater read depth.

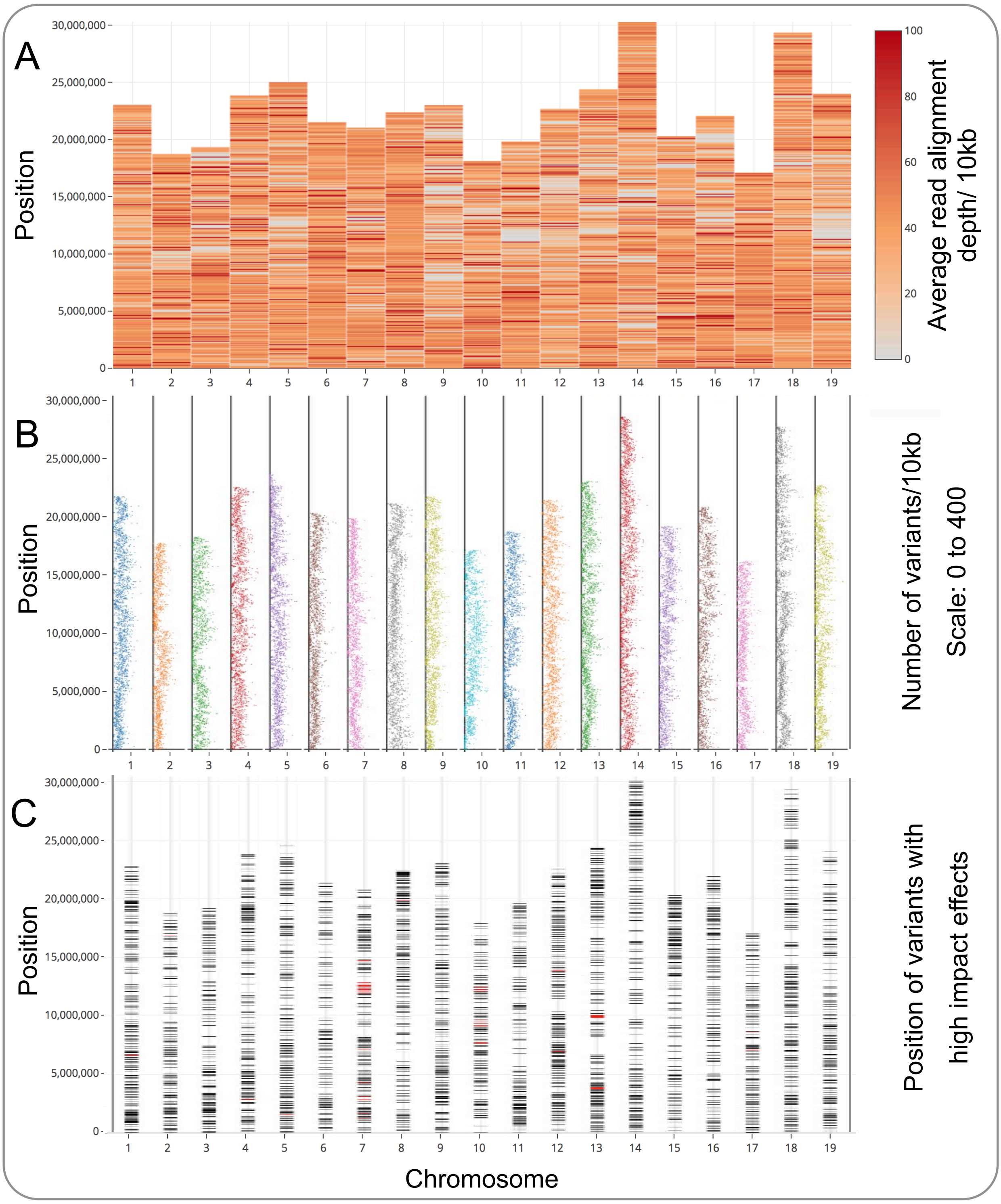


Figure 3.2: *Pinotage DNA sequence data aligned to the 19 chromosomes of the reference Pinot noir PN40024 and variants called with SnpEff. A) Heatmap of average read alignment depth over 10kb intervals. B) Scatter plots of total number of variants in same 10kb intervals. C) Chromosomal positions of 4,387 genes affected by high impact variants. Genes within functional clusters, where more than 50% of genes within the cluster contain one or more high impact variant, are indicated in red (also listed in Table 3.3).*

3.3.5 Pinotage/Pinot noir variant analysis

Large-scale genome-wide variant discovery is one of the opportunities provided by NGS and reference-based read mapping. The frequency of variants in a genome depends on the domestication and breeding history of the organism, as well as the reproduction system and mutation frequency. When genetic variants confer an advantage to the organism, they are subjected to positive selection pressure, and in the case of a crop or ornamental plant, this selection pressure may be human-driven. Heritable genomic diversity is conferred by two variant types: SNPs and structural variation such as copy number (CNVs), also called presence/absence variations (PAVs) (Cardone et al. 2016a; Scossa et al. 2016).

SNPs are the most abundant variant type (Taillon-Miller et al. 1998), and due to their relatively dense and uniform distribution along chromosomes and easy high-throughput genotyping, they are the most commonly used genetic marker in plants (Marrano et al. 2017). SNPs are used in many applications such as diversity analysis, genetic maps, cultivar identification and marker-assisted breeding.

Pinotage sequence data aligned to the 19 PN40024 reference chromosomes were used to identify variants (SNPs and short indels). Figure 3.2B is a visual representation of the number of variants distributed along the length of the chromosomes in 10kb intervals. Table 3.2 indicates the average variant density for the chromosomes. Some areas of low variant number (Figure 3.2B) correspond to areas of high read depth (Figure 3.2A), in other words, portions of Pinotage/Pinot noir similarity. And the converse is also true, i.e. high variant numbers/low read depth corresponding to low similarity areas. However, there was not always a correlation between variant number and read depth.

A total of 4,008,173 variants were found between the Pinotage data and the reference Pinot noir PN40024, of which more than 92% are SNPs, the remainder being insertions and deletions. The Pinotage/PN40024 variant density ranged from the lowest density in chromosome 6 (1 variant in 124bp) to the highest density in chromosome 12 (1 variant in 87bp), with an average variant density of 1 in 106bp across all 19 chromosomes (Table 3.2, Supplementary data 3.1).

In the case of heterozygous (polymorphic) variants (65.9% of variants predicted in this study), Pinotage likely inherited one allele from Pinot noir and the other from Cinsaut. However, Pinotage is homozygous for 34.1% of the identified variants, i.e. both alleles differ from Pinot noir. These variants might arise from novel mutations in Pinotage (not inherited from either parent), but it should be taken into consideration that the reference Pinot noir clone PN40024 is not the exact clone used for the Pinot noir X Cinsaut crossing used to generate Pinotage, and some of these variants might not be truly homozygous in Pinotage.

A number of studies have reported on grapevine sequence diversity; however, only a few used re-sequencing data mapped to the reference PN40024 genome for genome-wide SNP density analysis, and is therefore comparable to this study. The genome-wide average Pinotage/Pinot noir variant density found here (1 in 106bp), is lower than reported among 11 Eurasian and 5 Euramerican cultivars (1 in 23bp) (Dong et al. 2010), Tunisian grapevine cultivars (1 in 33bp) (Riahi et al. 2013), 11 ancient cultivars and wild vines (1 in 64bp) (Lijavetzky et al. 2007) and nine *V. vinifera* and *V. riparia* cultivars (1 in 78bp) (Salmaso et al. 2005). Furthermore, the total number of Pinotage/PN40024 variants identified in this study (4,008,173) is lower than the 4,740,493 identified between four table grape cultivars (Autumn royal, Italia, Redglobe and Thomson Seedless) and PN40024 (Cardone et al. 2016b). A lower Pinotage/Pinot noir variant density is to be expected due to their close genetic relationship.

Conversely, less variants are reported for Corvina (646,982) (Venturini et al. 2013), Sultanina (1,193,566) (Di Genova et al. 2014) and Tannat (2,087,275) (Da Silva et al. 2013). Caution should however be exercised when comparing these SNP density data, because analytical methods, types of data (whole genome DNA sequencing or RNA-seq), thresholds for inclusion/exclusion of a SNP and regions included (genes/ exons only/ complete genome etc.) vary greatly between studies. Furthermore, the definition of the term "SNP" is sometimes used ambiguously in literature, some only including true single nucleotide variants, others also including short insertions or deletions.

It is possible to identify structural variations, such as CNVs and PAVs, by read mapping. As these variations form large insertions or deletions between the genome of interest and the reference, locating reads that mapped at a distance not compatible with the library insert size is an indication of a structural variant. The DNA sequencing libraries used in this study did not have a narrow and uniform insert size distribution, complicating this method of structural variant detection. Therefore, to identify genes possibly absent from Pinot noir, a more robust approach was used, aligning *de novo* assembled transcripts from an independent experiment to confirm their presence in the Pinotage genome. This part of the study is discussed in Chapter 5.

3.3.6 High impact variants in gene functional clusters

Due to the conservation of protein functions, it is expected that most of the identified variant effects will not be in the coding region of genes. Only 2.8% of variants had an effect on exons, while 24.7% of effects were upstream of a gene, 23.1% downstream, 24% in intergenic areas and 21.4% within an intron (Supplementary data 3.1).

High impact variant effects are those that have a high probability of influencing regulatory elements of genes, causing frameshift mutations or mutations causing non-conservative amino acid substitutions or premature stop codons. A total of 7789 high impact effects were identified, located in 4,387 genes. The chromosome positions of these affected genes are shown in Figure 3.2C.

It is evident that areas dense with high impact variant effects exist along the chromosomes. To further explore these high impact variant dense areas, variant effects within functional clusters were considered. A functional gene cluster is predicted by the physical proximity of functionally related genes. When multiple genes within a functional cluster are impacted by sequence variants, it is highly likely that the functionality of that cluster is different from that in the reference genome, either reduced, promoted, altered, or any combination thereof. A total of 255 functional clusters contain genes affected by high impact variants (Supplementary data table 3.1). Of these, 22 clusters had more than 50% of genes contained in them affected by high impact variants (Table 3.3, Figure 3.2C).

The functional clusters are involved in diverse metabolic pathways, as indicated by their MapMan bin classifications. Interestingly, eight of these clusters are located on chromosome 7, and although chromosome 7 was not the chromosome with the highest variant density (chromosome 7 had a predicted density of 1 in 102nt, Table 3.2), it is possible that a large portion of chromosome 7 is highly divergent from Pinot noir and more similar to Cinsaut.

Cluster CH_vvi_197 contains genes grouped in the “signalling kinase: Domain of Unknown Function 26 (DUF26)” MapMan bin. Among novel gene loci not annotated in the current genome annotation (Chapter 4, Section 4.3.3), five loci, although on different chromosomes, were also assigned to this MapMan bin. This data suggests that DUF26 is one of the key signalling kinases classes that differ between Pinotage and Pinot noir.

Three clusters (CH_vvi_385, CH_vvi_197, CH_vvi_44) are involved in signalling receptor kinases, located on chromosomes 4, 7 and 10 respectively. Signalling receptor kinases play an important role in the plant stress response network, and genes involved in these pathways were identified as a major gene class dissimilar between Pinotage and Pinot noir (Chapter 5, Section 5.3.4).

Table 3.3: Functional clusters[#] with more than 50% of genes containing a high impact variant. Clusters are ordered in the table in terms of % genes containing high impact variants. (Complete table: Supplementary data table 3.1)

Functional cluster[#]	Chromosome number	High impact variants*	In cluster[$]	MapMan bin	MapMan bin description
CH_vvi_385	7	6	4	30.2.11	Signalling receptor kinases leucine rich repeat XI
CH_vvi_184	7	8	6	26.3	Misc gluco-, galacto- and mannosidases
CH_vvi_132	7	6	5	26.9	Misc glutathione S transferases
CH_vvi_349	7	36	34	35.2	Not assigned unknown
CH_vvi_182	7	4	4	26.1	Misc misc2
CH_vvi_252	4	4	4	17.1.1.1.10	Hormone metabolism abscisic acid synthesis-degradation synthesis
CH_vvi_301	13	11	15	26.28	Misc GDSL-motif lipase
CH_vvi_169	12	2	3	11.3.8	Lipid metabolism Phospholipid synthesis phosphatidylserine decarboxylase
CH_vvi_246	17	2	3	5.1	Fermentation aldehyde dehydrogenase
CH_vvi_419	7	2	3	17.6.1.1	Hormone metabolism gibberellin synthesis-degradation copalyl diphosphate synthase
CH_vvi_422	2	2	3	9.4	Mitochondrial electron transport / ATP synthesis alternative oxidase
CH_vvi_224	7	4	6	33.1	Development storage proteins
CH_vvi_273	1	4	6	26.22	Misc short chain dehydrogenase/reductase (SDR)
CH_vvi_279	17	4	6	27.3.66	RNA regulation of transcription Pseudo ARR transcription factor family
CH_vvi_197	10	8	12	30.2.17	Signalling receptor kinases DUF26
CH_vvi_96	8	5	8	26.4.1	Misc beta 1,3 glucan hydrolases glucan endo-1,3-beta-glucosidase
CH_vvi_112	5	3	5	26.1	Misc misc2
CH_vvi_170	12	3	5	34.99	Transport misc
CH_vvi_44	10	9	15	30.2.25	Signalling receptor kinases wall associated kinase
CH_vvi_149	10	4	7	34.16	Transport ABC transporters and multidrug resistance systems
CH_vvi_99	7	6	11	20.1.7	Stress biotic PR-proteins
CH_vvi_23	13	8	15	29.5.1	Protein degradation subtilases

[#] Functional clusters as predicted for *Vitis vinifera* in PLAZA functional clustering experiment 17.

[*] Number of genes located within the cluster boundaries, with a high impact variant (as determined by read alignment and variant calling with SnpEff).

[$] Number of genes currently identified as belonging in the functional cluster.

Multiple genes, having an additive effect, govern a large percentage of important crop traits. Likewise, single genes may also have a pleiotropic effect on more than one trait. Therefore, identifying the genes responsible for specific phenotypes is difficult, and despite considerable advances made in the field of crop genomics, a large amount of details about genetic control of phenotypes is still unsolved. Nevertheless, the results presented in this chapter are a stepping-stone towards the better understanding of grapevine genetics, specifically the genetic differences between Pinotage and Pinot noir.

3.4 Conclusion

As DNA sequencing technology becomes more accessible, it is no longer sufficient to have only one reference genome representing a species. Varietal differences have been exploited for centuries in agriculture and are the foundation of a breeding program to enhance crops. It is crucial to capture these varietal differences through genome sequencing of crop species variants.

This is the first report of genome sequencing and assembly for *Vitis vinifera* cv Pinotage. The Pinotage sequencing data were aligned to the Pinot noir reference genome to assess the degree of chromosome coverage and variation in sequencing depth along the length of the chromosomes. The alignment information was also used to detect variants between Pinotage and Pinot noir. Numerous variants (SNPs and indels) were identified, and the positions of high impact variants in functional clusters were analysed. Twenty-two functional clusters with more than 50% of genes containing a high impact variant were identified. Three of these clusters are classified as “signalling kinases”, indicating that this might be a gene category that differs most between Pinotage and Pinot noir. These results are confirmed in Chapter 5, where this gene category was also highlighted as a variable gene category between these cultivars.

The Pinotage genome sequence played an important confirmative role in a later part of this study. In Chapter 5, the alignment of Pinotage transcriptome data to the genome sequence in order to identify genes found in the Pinotage genome, is discussed. The genome nucleic acid extraction, sequencing and assembly (this chapter), and that of the transcriptome sequencing (Chapter 5) were performed completely separate. The genome and transcriptome data were aligned, and only transcripts corroborated by both datasets were selected for further analysis.

In future studies, this draft genome sequence of Pinotage will serve as a stepping-stone towards a more complete Pinotage genome sequence and a more in-depth analysis of Pinotage, and grapevine genetics. Genome and genetic variant data is an important resource to assist in the elucidation of the underlying basis for phenotypic differences between grapevine cultivars. However, having genome sequence information of a crop cultivar is only the first step in understanding the relationship between intra-species genetic and phenotypic variation. Gene sequences are not the only drivers of an organism's phenotypic outcome. Influences on gene expression, both environmental cues and intrinsic signals such as epigenetic modification, also play a critical role in determining phenotype. This prompted the next part of this study, namely an investigation of gene expression in Pinotage leaves and berries (Chapter 4).

Supplementary data

Supplementary data 3.1: SNPeff report (alignment of Pinotage DNA sequencing data against 19 chromosomes of Pinot noir PN40024).

Supplementary data table 3.1: Functional clusters (as predicted for *Vitis vinifera* in PLAZA functional clustering experiment 17) containing high impact variants. Clusteres are ordered in the table in terms of % genes containing high impact variants. 225 functional clusters are shown.

3.5 References

Altschul SF, Gish W, Miller W, Myers EW and Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215:403–410. doi: 10.1016/S0022-2836(05)80360-2

Anderson K and Aryal NR (2015) Which winegrape varieties are grown where? A global empirical picture. doi: http://dx.doi.org/10.20851/winegrapes

Andrews S, Lindenbaum P, Howard B and Ewels P (2011) FastQC. doi: Available at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc

Aubakirova K, Omasheva M, Ryabushkina N, Tazhibaev T, Kampitova G and Galiakparov N (2014) Evaluation of five protocols for DNA extraction from leaves of Malus sieversii, Vitis vinifera, and Armeniaca vulgaris. Genet Mol Res GMR 13:1278–1287. doi: 10.4238/2014.February.27.13

Bolger AM, Lohse M and Usadel B (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30:2114–2120. doi: 10.1093/bioinformatics/btu170

Bowers J, Boursiquot J-M, This P, Chu K, Johansson H and Meredith C (1999) Historical genetics: the parentage of Chardonnay, Gamay, and other wine grapes of Northeastern France. Science 285:1562–1565. doi: 10.1126/science.285.5433.1562

Cardone MF, Bergamini C, D'Addabbo P, Alkan C, Catacchio CR, Anaclerio F, Chiatante G, Marra A, Giannuzzi G, Perniola R et al. (2016a) Genomics technologies to study structural variations in the grapevine genome. BIO Web Conf 7:01016. doi: 10.1051/bioconf/20160701016

Cardone MF, D'Addabbo P, Alkan C, Bergamini C, Catacchio CR, Anaclerio F, Chiatante G, Marra A, Giannuzzi G, Perniola R et al. (2016b) Inter-varietal structural variation in grapevine genomes. Plant J 88:648–661. doi: 10.1111/tpj.13274

Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, Land SJ, Lu X and Ruden DM (2012) A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly (Austin) 6:80–92. doi: 10.4161/fly.19695

Da Silva C, Zamperin G, Ferrarini A, Minio A, Molin AD, Venturini L, Buson G, Tononi P, Avanzato C, Zago E et al. (2013) The high polyphenol content of grapevine cultivar Tannat berries is conferred primarily by genes that are not shared with the reference genome. Plant Cell 25:4777–4788. doi: 10.1105/tpc.113.118810

Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST et al. (2011) The variant call format and VCFtools. Bioinformatics 27:2156–2158. doi: 10.1093/bioinformatics/btr330

De Beer D, Joubert E, Marais J and Manley M (2017) Maceration before and during fermentation: effect on Pinotage wine phenolic composition, total antioxidant capacity and objective colour parameters. South Afr J Enol Vitic 27:137–150. doi: 10.21548/27-2-1614

Di Genova A, Almeida AM, Muñoz-Espinoza C, Vizoso P, Travisany D, Moraga C, Pinto M, Hinrichsen P, Orellana A and Maass A (2014) Whole genome comparison between table and wine grapes reveals a comprehensive catalog of structural variants. BMC Plant Biol 14:7. doi: 10.1186/1471-2229-14-7

Dong Q-H, Cao X, Yang G, Yu H-P, Nicholas KK, Wang C and Fang J-G (2010) Discovery and characterization of SNPs in Vitis vinifera and genetic assessment of some grapevine cultivars. Sci Hortic 125:233–238. doi: 10.1016/j.scienta.2010.03.023

Ekblom R and Wolf JBW (2014) A field guide to whole-genome sequencing, assembly and annotation. Evol Appl 7:1026–1042. doi: 10.1111/eva.12178

Gambino G, Perrone I and Gribaudo I (2008) A Rapid and effective method for RNA extraction from different tissues of grapevine and other woody plants. Phytochem Anal 19:520–525. doi: 10.1002/pca.1078

Gan X, Stegle O, Behr J, Steffen JG, Drewe P, Hildebrand KL, Lyngsoe R, Schultheiss SJ, Osborne EJ, Sreedharan VT et al. (2011) Multiple reference genomes and transcriptomes for Arabidopsis thaliana. Nature 477:419–423. doi: 10.1038/nature10414

Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q et al. (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol 29:644–652. doi: 10.1038/nbt.1883

Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, Couger MB, Eccles D, Li B, Lieber M et al. (2013) De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nat Protoc 8:1494–1512. doi: 10.1038/nprot.2013.084

Jaillon O, Aury J-M, Noel B, Policriti A, Clepet C, Casagrande A, Choisne N, Aubourg S, Vitulo N, Jubin C et al. (2007) The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. Nature 449:463–467. doi: 10.1038/nature06148

Langmead B and Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. Nat Methods 9:357–359. doi: 10.1038/nmeth.1923

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R and 1000 Genome Project Data Processing Subgroup (2009) The Sequence Alignment/Map format and SAMtools. Bioinforma Oxf Engl 25:2078–2079. doi: 10.1093/bioinformatics/btp352

Lijavetzky D, Cabezas JA, Ibáñez A, Rodríguez V and Martínez-Zapater JM (2007) High throughput SNP discovery and genotyping in grapevine (Vitis vinifera L.) by combining a re-sequencing approach and SNPlex technology. BMC Genomics 8:424. doi: 10.1186/1471-2164-8-424

Luo R, Liu B, Xie Y, Li Z, Huang W, Yuan J, He G, Chen Y, Pan Q, Liu Y et al. (2012) SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. GigaScience 1:1–6. doi: 10.1186/2047-217X-1-18

Marais DJ (2003a) Literature overview of Pinotage research. In: Wineland Mag. http://www.wineland.co.za/literature-overview-of-pinotage-research/. Accessed 9 Jun 2017

Marais J (2003b) Effect of different wine-making techniques on the composition and qualit of Pinotage wine: Juice/skin mixing practices. 24:

Marais J (2003c) Effect of different wine-making techniques on the composition and quality of Pinotage wine: Low-temperature skin contact prior to fermentation. South Afr J Enol Vitic 24:70–75.

Marrano A, Birolo G, Prazzoli ML, Lorenzi S, Valle G and Grando MS (2017) SNP-discovery by RAD-sequencing in a germplasm collection of wild and cultivated grapevines (V. vinifera L.). PLoS ONE 12:1–19. doi: 10.1371/journal.pone.0170655

Proost S, Bel MV, Sterck L, Billiau K, Parys TV, Peer YV de and Vandepoele K (2009) PLAZA: A comparative genomics resource to study gene and genome evolution in plants. Plant Cell Online 21:3718–3731. doi: 10.1105/tpc.109.071506

Reid KE, Olsson N, Schlosser J, Peng F and Lund ST (2006) An optimized grapevine RNA isolation procedure and statistical determination of reference genes for real-time RT-PCR during berry development. BMC Plant Biol 6:27. doi: 10.1186/1471-2229-6-27

Riahi L, Zoghlami N, Fournier-Level A, Dereeper A, Cunff LL, Laucou V, Mliki A and This P (2013) Characterization of single nucleotide polymorphism in Tunisian grapevine genome and their potential for population genetics and evolutionary studies. Genet Resour Crop Evol 60:1139–1151. doi: 10.1007/s10722-012-9910-y

Salmaso M, Faes G, Segala C, Stefanini M, Salakhutdinov I, Zyprian E, Toepfer R, Grando MS and Velasco R (2005) Genome diversity and gene haplotypes in the grapevine (Vitis vinifera L.), as revealed by single nucleotide polymorphisms. Mol Breed 14:385–395. doi: 10.1007/s11032-005-0261-7

Salzman RA, Fujita T, Zhu-Salzman K, Hasegawa PM and Bressan RA (1999) An Improved RNA Isolation Method for Plant Tissues Containing High Levels of Phenolic Compounds or Carbohydrates. Plant Mol Biol Report 17:11–17. doi: 10.1023/A:1007520314478

Scossa F, Brotman Y, de Abreu e Lima F, Willmitzer L, Nikoloski Z, Tohge T and Fernie AR (2016) Genomics-based strategies for the use of natural variation in the improvement of crop metabolism. Plant Sci 242:47–64. doi: 10.1016/j.plantsci.2015.05.021

Soto-Jimenez LM, Estrada K and Sanchez-Flores A (2014) GARM: genome assembly, reconciliation and merging pipeline. Curr Top Med Chem 14:418–424.

Taillon-Miller P, Gu Z, Li Q, Hillier L and Kwok P-Y (1998) Overlapping Genomic Sequences: A Treasure Trove of Single-Nucleotide Polymorphisms. Genome Res 8:748–754. doi: 10.1101/gr.8.7.748

Van Bel M, Proost S, Wischnitzki E, Movahedi S, Scheerlinck C, Peer YV de and Vandepoele K (2011) Dissecting plant genomes with the PLAZA comparative genomics platform. Plant Physiol pp.111.189514. doi: 10.1104/pp.111.189514

Venturini L, Ferrarini A, Zenoni S, Tornielli GB, Fasoli M, Santo SD, Minio A, Buson G, Tononi P, Zago ED et al. (2013) De novo transcriptome characterization of Vitis vinifera cv. Corvina unveils varietal diversity. BMC Genomics 14:41. doi: 10.1186/1471-2164-14-41

Online resources

Bedtools: http://bedtools.readthedocs.io/en/latest/index.html

CRIBI: http://genomes.cribi.unipd.it/grape/

De Waal Wines:
http://www.dewaal.co.za/Page.aspx?PAGEID=3357&Type=About&CLIENTID=3512&Title=DE%20WAAL%20DIFFERENCE

Illumina TruSeq Library kit:
http://research.lunenfeld.ca/ngs/truseq_dna_pcrfree_sampleprep_guide_15036187_a.pdf

Illumina TruSeq® Stranded Total RNA Library Preparation Kit with Ribo-Zero™Plant:
https://www.illumina.com/content/dam/illumina-marketing/documents/products/datasheets/datasheet_truseq_stranded_rna_plant.pdf

NEBNext® Ultra™ kit: https://www.neb.com/protocols/2015/09/16/protocol-for-use-with-nebnext-ultra-ii-dna-library-prep-kit-for-illumina-e7645

Nextera DNA library kit: https://support.illumina.com/content/dam/illumina-support/documents/documentation/chemistry_documentation/samplepreps_nextera/nexteradna/nextera-dna-library-prep-reference-guide-15027987-01.pdf

PlaBi database: http://www.plabipd.de/portal/web/guest/home1

PLAZA database: http://bioinformatics.psb.ugent.be//plaza/versions/plaza_v3_dicots/functionalcluster/chrom_clusters/vvi/1

Pinotage Association: http://www.pinotage.co.za/index.php/about-pinotage/the-pinotage-story/origin-of-the-grape

PN40024 reference grapevine genome: http://genomes.cribi.unipd.it/grape/

Wines of South Africa (WOSA): http://www.sawis.co.za/info/download/Book_2016_statistics_year_english_final.pdf

Chapter 4: The Pinotage Leaf and Berry Transcriptome in Young and Old Vines

4.1 Introduction

The development and maturation of grapevine berries have received considerable scientific attention because of the importance of this process in plant physiology and the significance of the fruit as an agricultural commodity. Grapevine berries are non-climacteric fruit and ripening is controlled by fluctuations in hormone levels (Coombe and McCarthy 2000; Robinson and Davies 2000; Conde et al. 2007; Kuhn et al. 2014; Fortes et al. 2015). Berry maturation immediately precedes harvesting and is the crucial phase that determines the type and concentration of flavour components in the grapes, and ultimately the quality of wine that can be made from these grapes. An important characteristic of grapevine is its ability to accumulate and store large quantities of sugar, and also flavour and aroma compounds, in its berries.

The availability of the grapevine genome sequence (Jaillon et al. 2007; Velasco et al. 2007), and microarray and next-generation sequencing (NGS) technologies have advanced large-scale mRNA expression profiling in grapevine (Jain 2012; Gapper et al. 2014). Grapevine berry development and transcriptomes have been studied using microarrays (Waters et al. 2005; Deluc et al. 2007; Grimplet et al. 2007; Pilati et al. 2007; Zamboni et al. 2010; Fortes et al. 2011; Lijavetzky et al. 2012; Dal Santo et al. 2013; Agudelo-Romero et al. 2013; Cramer et al. 2014) and NGS (Zenoni et al. 2010; Sweetman et al. 2012; Fasoli et al. 2012; Degu et al. 2014; Ghan et al. 2017). Microarrays and NGS have also been used to characterise the transcriptomes of new cultivars, mutants and clones (Venturini et al. 2013; Guo et al. 2016; Royo et al. 2016; Pervaiz et al. 2016; Muñoz-Espinoza et al. 2016; Grimplet et al. 2017).

In this study NGS was used to characterise the transcriptome of a local *Vitis vinifera* cultivar, Pinotage. The RNA-seq data was also used in the first research on the contribution of genetics to the so-called “old-vine” wine character. In the international wine market there is a newfound interest in old vines and the artisanal wines crafted from them. Wines produced from older vines are generally accepted as having more depth and complexity than those produced from younger vineyards, and this term is used on wine labels to indicate a wine of high quality, with an intense and full flavour. Nonetheless, no formal classification exists at what age a vine becomes an “old vine”, and it largely depends on the history of vineyards and winemaking in the area. For example, in old world wine production areas many vineyards in excess of 100 years old may exist. In contrast, in a new world growing area, a vine of 50 years might be considered old (Heyns 2013; Easton 2016; Fridjhon 2016; Hawkins 2016; Hooke 2016; Beavers 2016; Van Wyk 2016; Szabo 2017). In South Africa, the economic life of a vineyard is an average of 20 to 25 years, and vines are generally considered to be old when they reach 35 years (Easton 2016). In terms of hectare coverage of South African vineyards older than 35 years, Chenin blanc is the most represented cultivar, followed by Sultana and Pinotage (http://iamold.withtank.com/home/).

The unique character of wine produced from old vines may be the result of various complex factors. Old vines are usually grown as bushvines under dryland conditions and these vines naturally produce lower yields of grapes than younger, more vigorous vines. This allows for the concentration of more flavour and aroma compounds in the berries. Older vines are more adapted to the specific climate and soil type of their growing environment, and will therefore show more distinct regional characteristics and a greater expression of the specific *terroir*. They also have established interactions with microorganisms and insects in their environment. Moreover, older vines have deeper, better-established root systems that can serve as a buffer in dry conditions. Molecular factors such as mutations and epigenetic modifications acquired over the lifespan of the vine might also contribute to this old-vine character (Heyns 2013; Easton 2016; Fridjhon 2016; Hawkins 2016; Hooke 2016; Beavers 2016; Van Wyk 2016; Szabo 2017). However, describing wines made from old vines as having more depth and character is subjective, and to our knowledge no scientific research has been done to prove which compounds differ between young- and old-vine wines. Changes in gene expression and hormone levels play a role in initiating and controlling ripening and influence flavour and aroma compounds accumulating in the berry. Still, it is unknown to what extent gene expression levels differ between young and old vines.

To study this distinctive old-vine character, gene expression profiling of both berries and leaves from young and old Pinotage vines at harvest, were performed. Vine material was sampled from a commercial Pinotage vineyard where young and old vines are inter-planted. This data was used to form an overall picture of gene expression in leaves and berries of Pinotage vines. Novel gene loci, not present in the current reference genome annotation, were identified and differential gene expression between young and old vines studied. The possible roles these differentially expressed genes play in hormone metabolism and biochemical changes linked to fruit ripening, were explored.

4.2 Methods and Materials

4.2.1 Sample collection

Sample material was collected from nine young and nine old Pinotage (*Vitis vinifera* cv Pinotage) vines (Figure 4.1) at harvest time, January 2016. The vineyard (Pinotage clone 6, grafted onto Richter 99 rootstock) is situated in Stellenbosch, South Africa, and was established in 1976 as bushvines, without irrigation. During the lifespan of the vineyard, several old vines displaying poor vigour and growth were replaced with new vines, of the same scion/rootstock combination. The new scion material was derived from a virus-free meristem culture of plant tissue from one of the old vines in the vineyard. The new vines are therefore genetic clones of the older vines. At the time of sampling, the old vines were 40 years and the young vines seven years old. The vineyard had a high prevalence of grapevine leafroll disease.

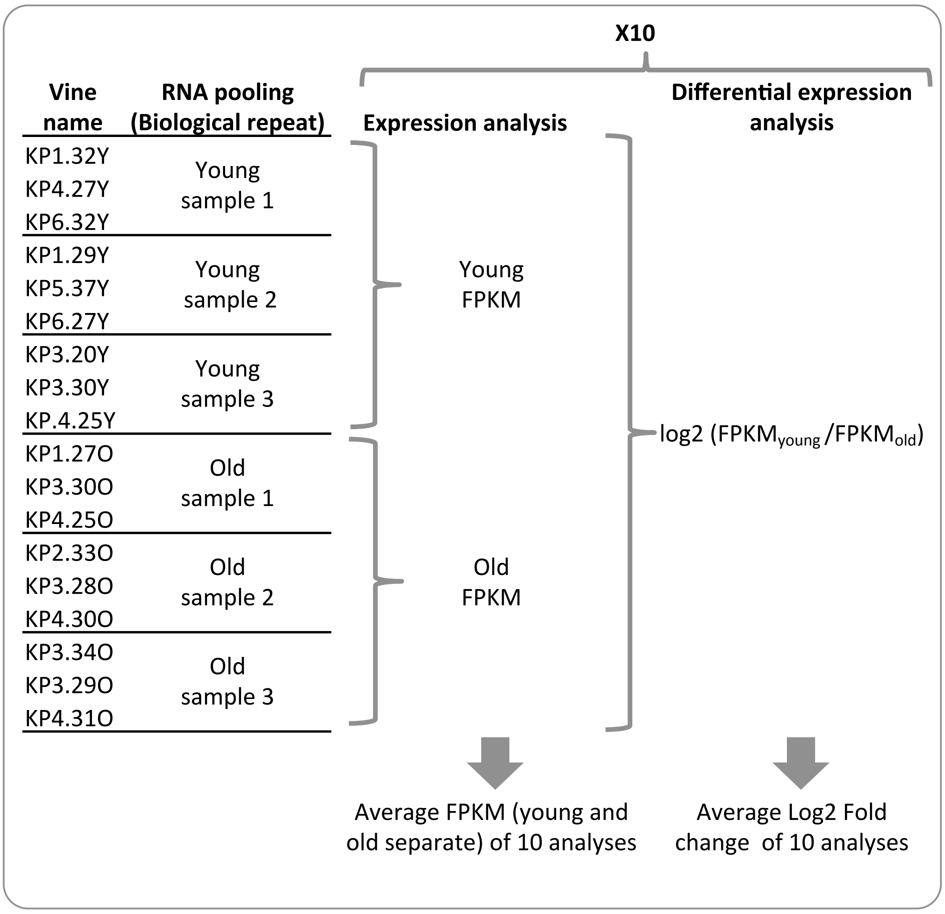
Three berry clusters from each vine were randomly selected and ten berries were collected from each cluster. The seeds were removed the berries pressed through cheesecloth to release the juice. The collected juice was evaluated for total sugar (°Brix) and titratable acids at the department of Viticulture and Oenology, Stellenbosch University. The remaining berry flesh and skins were stored at -80°C.

Three young fully expanded (mature) leaves were collect from three randomly selected canes from each vine, and stored at -80°C. All of the samples were collected on the same day and time. Selection of samples was randomized to minimise variation in developmental stages in berry and leave samples from young and old vines.

4.2.2 RNA extraction

One gram of frozen material (leaf and berry separately) was powdered in liquid nitrogen with a mortar and pestle and total RNA extracted using the protocol from Reid et al. (2006). Total RNA (15µg) was treated with RQ1 RNase-free DNase (Promega, Madison, USA) in 50µl reactions according to the manufacturer's instructions. After incubation at 37°C for 30 min, the reaction volume was adjusted to 500µl with 10mM Tris-HCl (pH 8.5). An acidic phenol extraction was performed, followed by a chloroform-isoamyl alcohol (24:1) extraction. The RNA was precipitated with ethanol and sodium acetate (2.5 volumes 100% ethanol and 0.1 volumes 3M sodium acetate, pH 5.2). After a wash step with 70% ethanol, pellets were dried and resuspended in 30µl ddH_2O.

To ascertain the RNA integrity, an aliquot of RNA extract from each vine was analysed by gel-electrophoresis and quantified using the Trinean Xpose (Gentbrugge, Belgium). RNA from individual vines was pooled in groups of three to form six berry and six leaf samples (three young and three old each, Figure 4.1). Each RNA sample contained an equal amount of DNase-treated RNA (3µg) from each vine in a total volume of 50µl H_2O.



Vine name	RNA pooling (Biological repeat)
KP1.32Y KP4.27Y KP6.32Y	Young sample 1
KP1.29Y KP5.37Y KP6.27Y	Young sample 2
KP3.20Y KP3.30Y KP.4.25Y	Young sample 3
KP1.27O KP3.30O KP4.25O	Old sample 1
KP2.33O KP3.28O KP4.30O	Old sample 2
KP3.34O KP3.29O KP4.31O	Old sample 3

X10
Expression analysis
Differential expression analysis
Young FPKM
Old FPKM
log2 ($FPKM_{young}$ /$FPKM_{old}$)
Average FPKM (young and old separate) of 10 analyses
Average Log2 Fold change of 10 analyses

Figure 4.1: *Vine names and the RNA pooling, as well as the bioinformatic analysis strategy used in this study. The same strategies were followed for berries and leaves, resulting in four groups (young- and old-vine berries, and young- and old-vine leaves), each with three biological replicates. The expression and differential expression analysis were performed ten times. Final gene expression values used were the average FPKM over the ten analyses per gene per sample group. Differential expression is calculated as* log_2 $FPKM_{young}/FPKM_{old}$ *and final differential expression values used were the average* log_2 $FPKM_{young}/FPKM_{old}$ *over the ten analyses per gene per sample group. FPKM: Fragments Per Kilobase of transcript per Million mapped reads.*

4.2.3 RNA Library preparation and sequencing

The RNA was shipped on dry ice to the sequencing facility (Agricultural Research Council, Biotechnology platform, Pretoria, South Africa) and the quality assessed using the Agilent RNA Nano 6000 kit on the Agilent 2100 Bioanalyzer (Agilent Technologies, Santa Clara, USA). RNA libraries were prepared using the Illumina TruSeq stranded mRNA protocol (Illumina TruSeq stranded mRNA protocol: online resources) according to the manufacturer's instructions and sequenced on the Illumina HiSeq2000 (San Diego, USA), generating 125nt paired-end reads.

4.2.4 RNA sequence quality trimming and filtering

Sequencing data were assessed for quality with FastQC (Andrews et al. 2011). Trimmomatic (Bolger et al. 2014a) was used to remove adaptor sequences. The first 9nt of the reads showed a nucleotide composition imbalance, and were removed. Sequence reads were scanned from the 5'-end for a minimum average Phred score of Q20 over a sliding window of 3nt. Only unbroken pairs and reads with a minimum length of 20nt were retained.

4.2.5 Gene expression in berries and leaves

Twenty million quality trimmed reads were randomly selected from each sample with fastq-tools (subsampling module; Daniel Jones; http://homes.cs.washington.edu/~dcjones/fastq-tools/). A reference-based expression analysis was performed with the RNA-seq Tuxedo pipeline (Trapnell et al. 2012) for the four sample groups (young- and old-vine berries, and young- and old-vine leaves). Read pairs were aligned with Tophat (Parameters: -r 150 -N 5 -I 10000 --library-type fr-firststrand --segment-mismatches 3 --read-gap-length 3 --read-edit-dist 5 -m 1 --mate-std-dev 50 --no-mixed --no-discordant) to the reference grapevine genome (http://genomes.cribi.unipd.it/grape/). The alignment was integrated with the reference annotation version 2.1 (http://genomes.cribi.unipd.it/grape/V2.1.gff3). Gene expression levels are reported as FPKM (Fragments Per Kilobase of transcript per Million mapped reads) for each of the four sample groups. This expression analysis was repeated ten times, each time with a randomly selected subsample of 20 million read pairs from each sample. Subsampling was performed with read replacement. The final gene expression values used, were the average FPKM over the ten analyses (Figure 4.1).

4.2.6 Differentially expressed genes in young and old vines

Differential gene expression between young and old leaf and berry samples were performed using Cuffdiff (Tuxedo pipeline). Cuffdiff reports a $\log_2$ fold change value for each gene in the reference annotation (can be 0 if no reads aligned) and also novel loci not present in the reference annotation. The $\log_2$ fold change value is calculated as $\log_2$ $\text{FPKM}_{\text{young}}/\text{FPKM}_{\text{old}}$.

A gene was accepted as differentially expressed if it had a significant $\log_2$ fold change value between young and old samples (adjusted p-value of less than 0.05) in seven or more of the repeat analyses. The differential expression analysis was repeated ten times with the output of the ten expression analyses (Figure 4.1). The average $\log_2$ fold change value per gene (average for the ten Cuffdiff repeat analyses) was used in further analyses.

4.2.7 Identification of novel gene loci

A gene locus was identified as novel if the area covered by RNA-seq reads did not correspond with a known gene locus in the V2.1 annotation, and had an average FPKM of 10 or higher (in the Tuxedo pipeline output) for at least one of the four groups (young- and old-vine berries, and young- and old-vine leaves). The novel loci coordinates were extracted from the Tuxedo output and used to obtain the fasta sequences for these loci from the grapevine genome sequence using bedtools (Quinlan laboratory, University of Utah, http://bedtools.readthedocs.io/en/latest/index.html). A flanking region of 1200nt upstream of each locus was included in the fasta sequence. Fasta sequences containing a continuous stretch of 20 or more unknown nucleotides (“Ns”) were removed and the remaining sequences uploaded to Coding Potential Calculator (CPC, http://cpc.cbi.pku.edu.cn/) (Kong et al. 2007). Sequences with a coding potential score of more than 1 (strong coding potential), were regarded as true novel loci. The original fasta sequences (without the flanking regions) from these loci were uploaded to Mercator for assignment into MapMan functional bins.

4.2.8 Gene functional assignment

The transcripts from the differentially expressed genes (DEGs) identified by Cuffdiff were retrieved from the V2.1 grapevine annotation (Vitulo et al. 2014) (http://genomes.cribi.unipd.it/grape/) and assigned to MapMan functional bins (Thimm et al. 2004; Usadel et al. 2009) (http://mapman.gabipd.org/) with Mercator (May et al. 2008), allowing multiple hits, and a BLAST cut-off of 1. All transcripts from a single gene that hit the same bin were collapsed and counted only once. The same procedure was followed to assign the putative novel genes to MapMan functional bins.

The V2.1 Blast2GO annotation file was downloaded from the CRIBI website (http://genomes.cribi.unipd.it/grape/) and enzyme commission (EC) numbers for genes were retrieved from this file. Enzyme names and pathways were retrieved from GrapeCyc (https://www.plantcyc.org/databases/grapecyc/7.0).

4.3 Results and Discussion

4.3.1 Sample collection, RNA extraction and sequencing

Berry and leaf material were collected from nine young and nine old vines during harvest time (January 2016) from a commercial vineyard (*Vitis vinifera* cv Pinotage) in Stellenbosch, South Africa. The feasibility of this study was based on the availability of an old Pinotage vineyard where vines from the same clone are inter-planted. Separate RNA extractions from leaves and berries of all 18 vines were performed. To limit the influence of environmental, inter-plant and technical variations, the RNA of three grapevines were pooled to form one biological replicate (Figure 4.1). RNA integrity numbers of between 7.10 and 8 were obtained for the 12 samples.

During sample preparation the juice from the berries were collected. Juice was mixed in equal ratios using the same pooling strategy as for the RNA. The titratable acid (g/L) and sugar (°Brix) content were measured. The juice from young-vine samples had a significantly higher sugar (average 22.33 °Brix) and lower acid (average 4.24g/L) concentration than that of old-vine samples (average sugar 20.4 °Brix and average acid 5.48g/L; Figure 4.2).

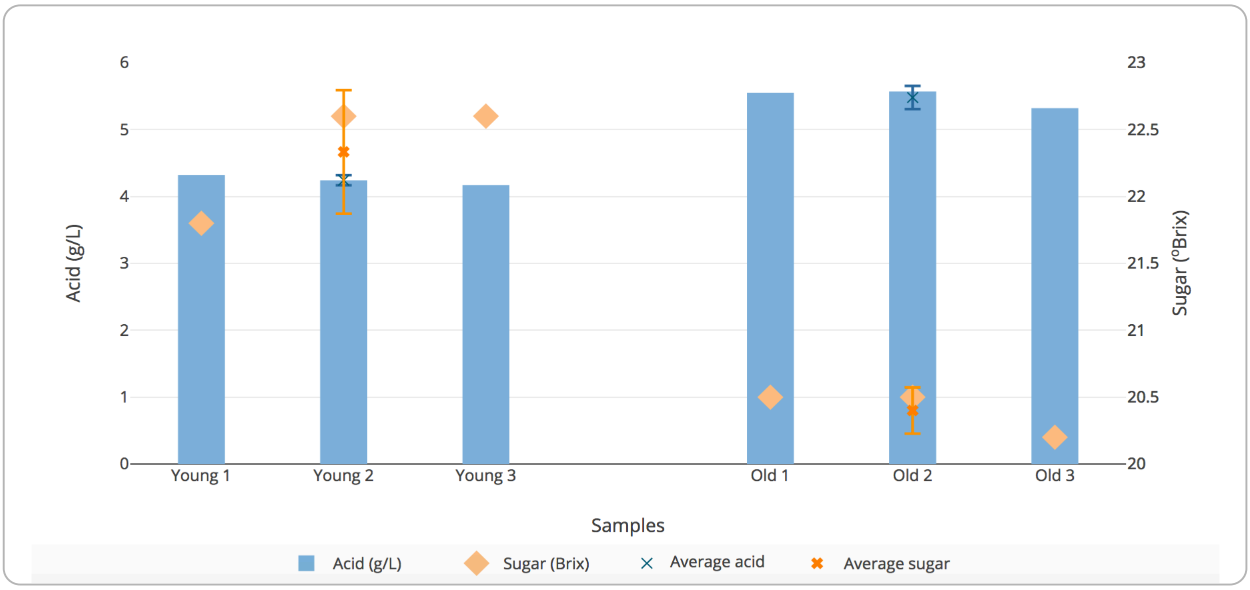


Figure 4.2: *Grape juice titratable acid (g/L) and sugar (°Brix) content of three young- and three old-vine samples (each consisting of the combined juice of three vines), at the time of harvest.*

4.3.2 Gene expression in berries and leaves

This is the first study of the berry and leaf transcriptome of *Vitis vinifera* cv Pinotage. Due to the large variation in read pair numbers between samples, we decided on a subsampling strategy, selecting 20 million read pairs per sample library. Bioinformatic analysis (expression and differential expression analysis) was performed ten times. For each gene that appear in the V2.1 *Vitis vinifera* genome annotation, the corresponding FPKM values from the Tuxedo pipeline output were averaged over the ten technical repeats. Average FPKM values were analysed to give an indication of how representative the transcriptome data is of the complete Pinotage transcriptome (i.e. how many genes are covered). Figure 4.3 gives an overview of the number of genes covered at a range of FPKM values in berries and leaves.

A total of 94 and 56 genes in berries and leaves, respectively, were highly expressed with an FPKM of 1000 or higher. Enzyme commission (EC) numbers are available for 32 (berries, Supplementary data table 4.1) and 36 (leaves, Supplementary data table 4.2) of these genes. The gene products without EC numbers might function as structural components, non-coding regulatory RNAs or may not yet have been annotated as an enzyme.

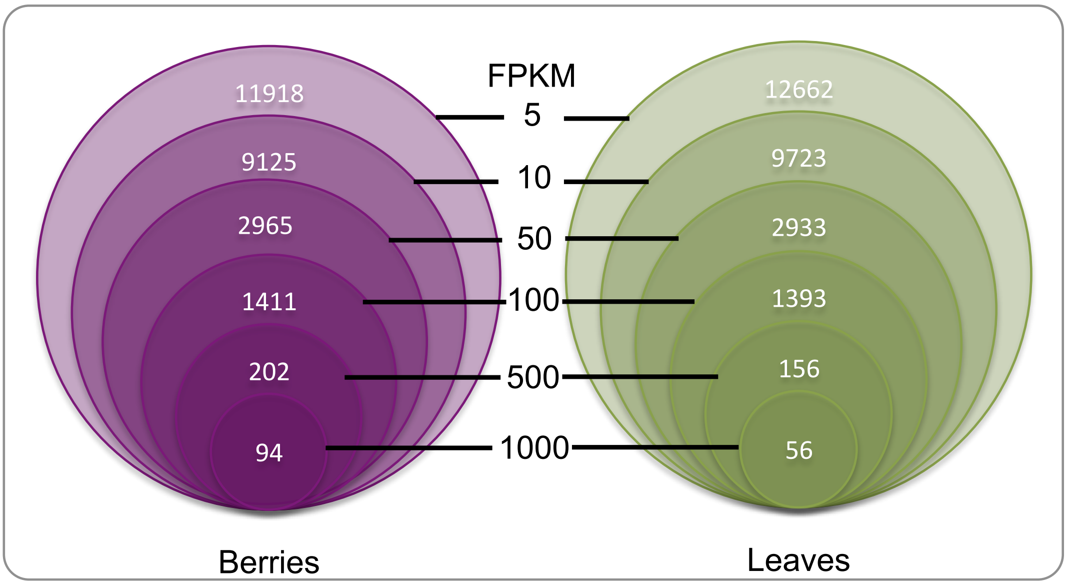


Figure 4.3: *Number of genes (from the V2.1 CRIBI Vitis vinifera annotation) covered at a range of 5 to 1000 FPKM in berries and leaves. At each FPKM value at least one of the groups (young or old) is covered at that FPKM or higher.*

The only enzyme encoded for by highly expressed genes in both berries and leaves is 2-alkenal reductase (EC:1.3.1.74), an enzyme from the oxidoreductase family that catalyzes electron transfer between molecules using NAD$^+$ as a co-factor (GrapeCyc). A number of genes highly expressed in berries code for enzymes associated with cell wall modification and breakdown, such as pectinesterase and cellulose. Glycosidases, namely glucan endo-1,3-β-D-glucosidase and fructose-bisphosphate aldolase, are enzymes involved in the breakdown of complex sugars (GrapeCyc) (Supplementary data table 4.1). Furthermore, genes coding for enzymes involved in secondary metabolism, jasmonate O-methyltransferase, flavonol synthase and caffeate O-methyltransferase, are amongst the most highly expressed genes in leaves. Four highly expressed genes in leaves code for serine endopeptidases (EC:3.4.21.0, Supplementary data table 4.2). Serine endopeptidases are involved in various plant physiological processes, for example the hypersensitive response to pathogens, senescence and signal transduction (Antão and Malcata 2005).

A FPKM bottom threshold value of ten was deemed biologically significant and genes with an FPKM of ten or higher were selected for further analysis. Of the 31,845 gene loci in the V2.1 CRIBI annotation, more than 50% were covered at an average FPKM of 10 or more in at least one of the groups (Figure 4.4). A total of 2,821 genes were expressed in both leaves and berries. In leaves, 598 more genes were expressed than in berries. This is to be expected, since young actively growing leaves were sampled, while the berries were in the final ripening phase and metabolic processes are expected to decline. Additionally, reads aligned to 647 loci not covered in the V2.1 grapevine annotation (Figure 4.4).

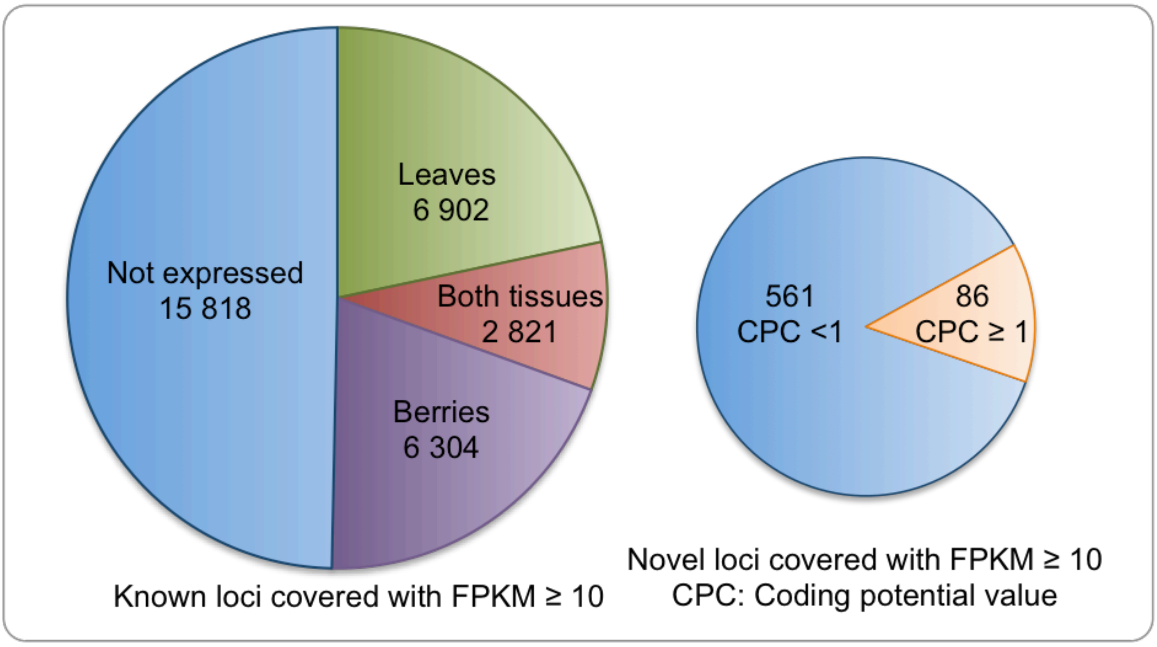


Figure 4.4: *A number of 16,027 genes (berries and leaves combined), out of the total of 31,845 genes in the V2.1 annotation, were covered at an FPKM ≥10 in at least one of the four groups (young- and old-vine berries, and young- and old-vine leaves).* §*Known loci in V2.1 annotation;* $^{\#}$*Loci not in V2.1 annotation.*

4.3.3 Identification of novel gene loci

Identification of novel gene loci on an already annotated reference genome sequence is one of the major advantages of RNA-seq analysis. This is the first RNA-seq study of the berry and leaf transcriptomes of Pinotage vines. A total of 647 putative novel gene loci were identified (Figure 4.4), covered with an average FPKM of 10 or higher in at least one of the sample groups. The sequences of these loci were retrieved from the grapevine reference genome sequence, and after removing sequences with more than 20 continuous unknown nucleotides, 559 sequences remained. Of these sequences, 86 had a strong coding potential (score of higher than 1 assigned to them with CPC) and were deemed true novel loci. This estimate of novel gene loci was conservative, given the strict criteria used for the analysis (high FPKM and coding potential threshold). In a similar study (Venturini et al. 2013), 180 novel gene loci were found in the grape cultivar Corvina. Considering that the reference grapevine sequence is Pinot noir, a parent of Pinotage, it is to be expected that Pinot noir and Pinotage will be less genetically divergent and fewer novel genes will be found in Pinotage. The dispersion of the identified novel loci throughout the genome is shown in Figure 4.5.

This section focuses on novel loci not annotated in the reference genome. Some degree of similarity is necessary for read mapping to a reference sequence. Therefore, for a more robust look at genes only found in the Pinotage genome, this RNA-seq dataset was used in a *de novo* transcriptome assembly and compared to the Pinotage draft genome sequence. This part of the study is discussed in Chapter 5.

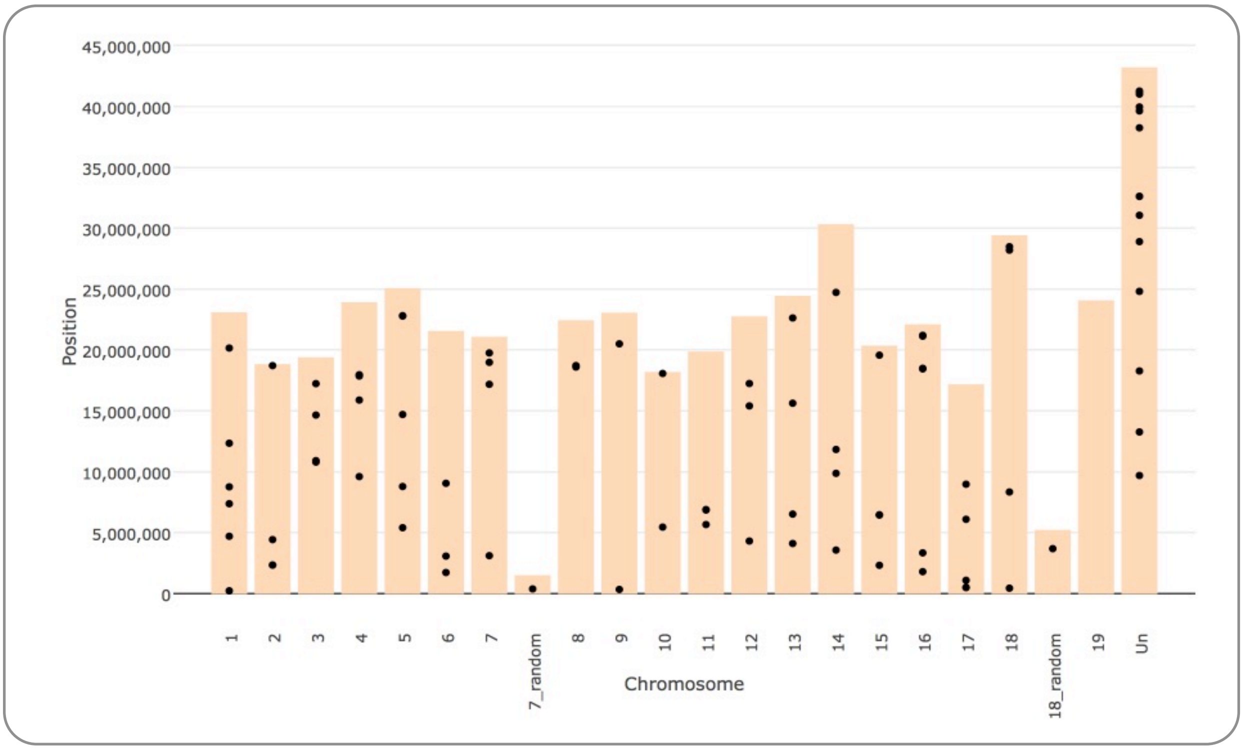


Figure 4.5: *Positions of the 86 novel identified loci on the grapevine chromosomes. Orange bars represent total chromosome length and black dots represent the position of the novel loci on the chromosomes. "Random" chromosomes denotes contigs assigned to a chromosome, but location or orientation could not be determined. Chromosome_un denotes contigs not assigned to a specific chromosome. Random chromosomes without novel loci are not shown.*

To give an indication of the roles of these putative novel genes in plant metabolism, they were classified into the 35 MapMan functional bins using Mercator (Figure 4.6). Interestingly, one of the novel genes assigned to the "cell wall" bin hit a grapevine ripening-related protein (grape ripening-induced protein, grip22). Davies and Robinson (2000) found seven cell wall proteins and ten proteins involved in stress response that accumulate during ripening, and called them grape ripening-induced proteins (grips). Grip22 is involved in stress response and there are four genes labelled as grip22 in the current grapevine annotation (VIT_206s0004g02540, VIT_206s0004g02550, VIT_206s0004g02560 and VIT_206s0004g02570), all located on chromosome 6. The novel locus is predicted to be between VIT_206s0004g02550 and VIT_206s0004g02560 at position 3,066,928 to 3,067,588. VIT_206s0004g02560 is highly expressed in berries (greater than 1000 FPKM) from both young- and old-vine samples, while the putatively novel grip is only expressed in leaves. An *in silico* translation of this area from the Pinot noir PN40024 genome sequence was performed, and the Pinotage *de novo* assembled transcript and amino acid translation (discussed in Chapter 5) retrieved. The Pinot noir translation contains two premature stop codons, which is most likely the reason it is not present in the reference annotation. The Pinotage transcript translation has 100% identity to a grip22-like protein from *Vitis quinquangularis* (Genbank accession: AMB38758).

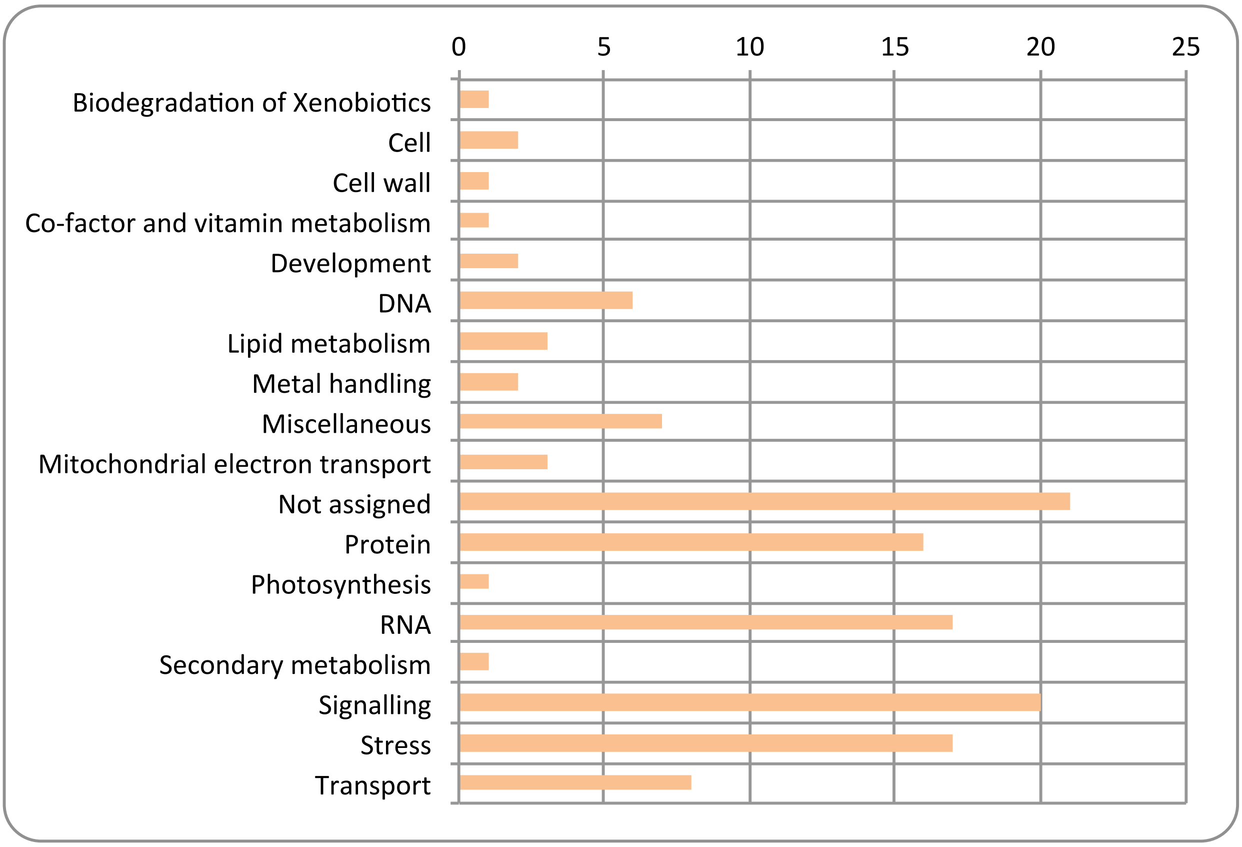


***Figure 4.6:** Mercator classification of 86 identified novel genes into the 35 primary MapMan functional bins. Bins with no hits were omitted. Bars represent number of genes in each bin.*

The MapMan secondary bins with the highest number of putative novel genes are: 11 pathogenesis-related proteins in the "stress" bin, five genes in the "protein degradation ubiquitin E3" bin, 14 transcription factors assigned to "RNA regulation of transcription" and five signalling receptor kinases called "Domain of Unknown Function 26" (DUF26). DUF26, also called Cysteine-rich Receptor-like Kinases (CRKs), is a large subfamily in the receptor-like protein kinase (RLKs) family. In plants, RLKs serve as signalling molecules that regulate plant development and stress response programs. DUF26 specifically plays an important role in pathogen defence response (Wrzaczek et al. 2010). “Stress”, “protein degradation” and “RNA regulation of transcription” are also the MapMan bins with the highest number of Pinotage *de novo* assembled transcripts (Chapter 5).

Of the ten putative novel genes that are only expressed in young-vine berries, one F1-ATPase functions in mitochondrial electron transport, one is involved in the secondary metabolism of simple phenols, and two are β-1,3 glucan hydrolases (the remaining six were classified as “unknown”). Seven putative novel genes were differentially expressed between young and old vines. Of these genes, one each is involved in “biodegradation of xenobiotics”, “UDP glucosyl and glucuronyl transferases” and “development” and was down-regulated in young vines, while a gene classified as “protein degradation AAA type” was up-regulated. Three DEGs were classified as “unknown”.

4.3.4 Differentially expressed genes between young and old vines

Of the 16,027 total genes expressed in berries and leaves (Figure 4.4), 952 genes were differentially expressed between young and old vines (Figure 4.7). At an FPKM of 10 or higher, 3.3% of the genes that were expressed in the berries, and 7.6% in leaves, displayed differential expression between young and old vines. Five DEGs were present in both berries and leaves.

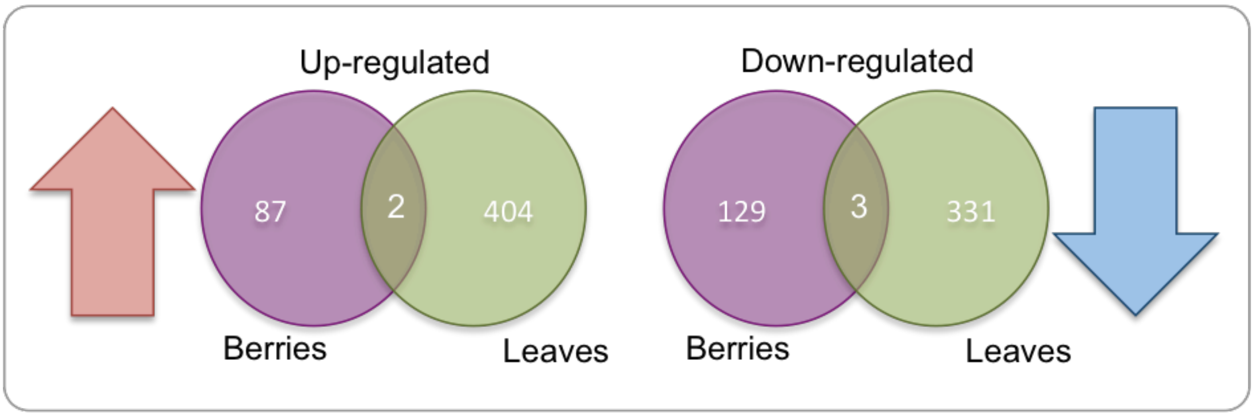


Figure 4.7: *Number of differentially expressed genes (DEGs) in berries and leaves. DEGs have an adjusted p-value of ≤ 0.5 in seven or more of the bioinformatic repeat analyses, and have an FPKM ≥ 10 in young and/or old groups.*

MapMan allows for specific metabolic pathways, where genes of interest are involved in, to be highlighted. The 952 DEGs were classified into MapMan functional bins to give an indication of their roles in plant metabolism (Figure 4.8). A total of 1,168 MapMan hits were obtained, of which 230 were classified in the “not assigned” bin. The remaining DEGs are mainly involved in the bins “cell”, “miscellaneous”, “protein”, “RNA”, “secondary metabolism”, “signalling”, “stress” and “transport”. Due to the importance of the grapevine berry ripening process in winemaking, genes differentially expressed between young and old vines and involved in fruit ripening were further investigated, with a particular focus on genes involved in hormone signalling and biochemical changes associated with berry ripening.

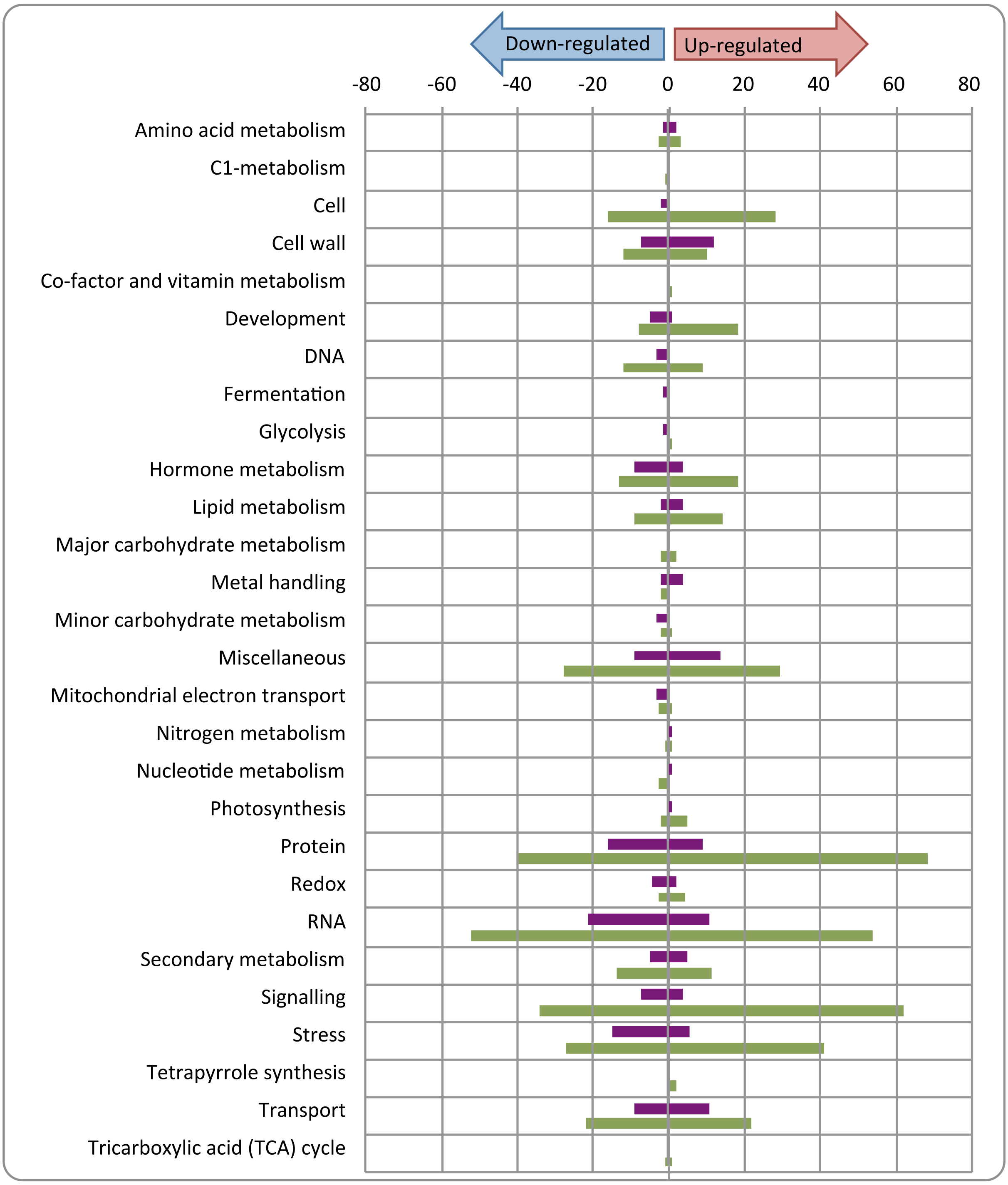


Figure 4.8: *Mercator classification of 925 DEGs into the 35 primary MapMan functional bins. Bins with no hits and the "not assigned" bin, representing 230 hits, were omitted. Bars represent number of genes in each bin. Down-regulated genes are shown to the left as negatives and up-regulated genes are shown to the right (regulation is in the context of young/old). Leaf DEGs are indicated in green and berry DEGs in purple. DEGs: Differentially expressed genes.*

4.3.5 Differentially expressed genes involved in hormone metabolism and signalling, and their influence on fruit ripening

Grapevine berry development has three phases (Figure 4.9). In the first phase of early fruit development, berry size increases, cell division and enlargement take place and organic acids and tannins accumulate. The second phase, called the herbaceous plateau, is a lag phase. Véraison marks the onset of the final phase, berry ripening, starting with colour and sugar accumulation and berry softening. During this ripening phase water and sugars accumulate in the vacuoles of the mesocarp cells, and anthocyanins accumulate in the berry skin. Organic acid metabolism (malic acid respiration) takes place, reducing the overall acid concentration in the berry. Secondary metabolites are produced, mainly in the berry skin, to protect against biotic and abiotic stresses (Coombe and McCarthy 2000; Robinson and Davies 2000; Conde et al. 2007; Kuhn et al. 2014; Fortes et al. 2015).

Grapevine berries undergo complex biochemical, physiological and molecular changes during fruit maturation and several multigenic families control the biosynthesis of molecules involved in grape berry ripening. A number of hormones participate in the control of berry ripening, specifically abscisic acid, ethylene and brassinosteroids as promoters of ripening (Coombe and McCarthy 2000; Robinson and Davies 2000; Conde et al. 2007; Fortes et al. 2015). In transcriptome analysis of the leaf and mature berry, 46 DEGs involved in hormone metabolism and signal transduction (Supplementary data table 4.3, Figure 4.9) were identified. Most DEGs were associated with auxin and ethylene metabolism, followed by those related to brassinosteroids.

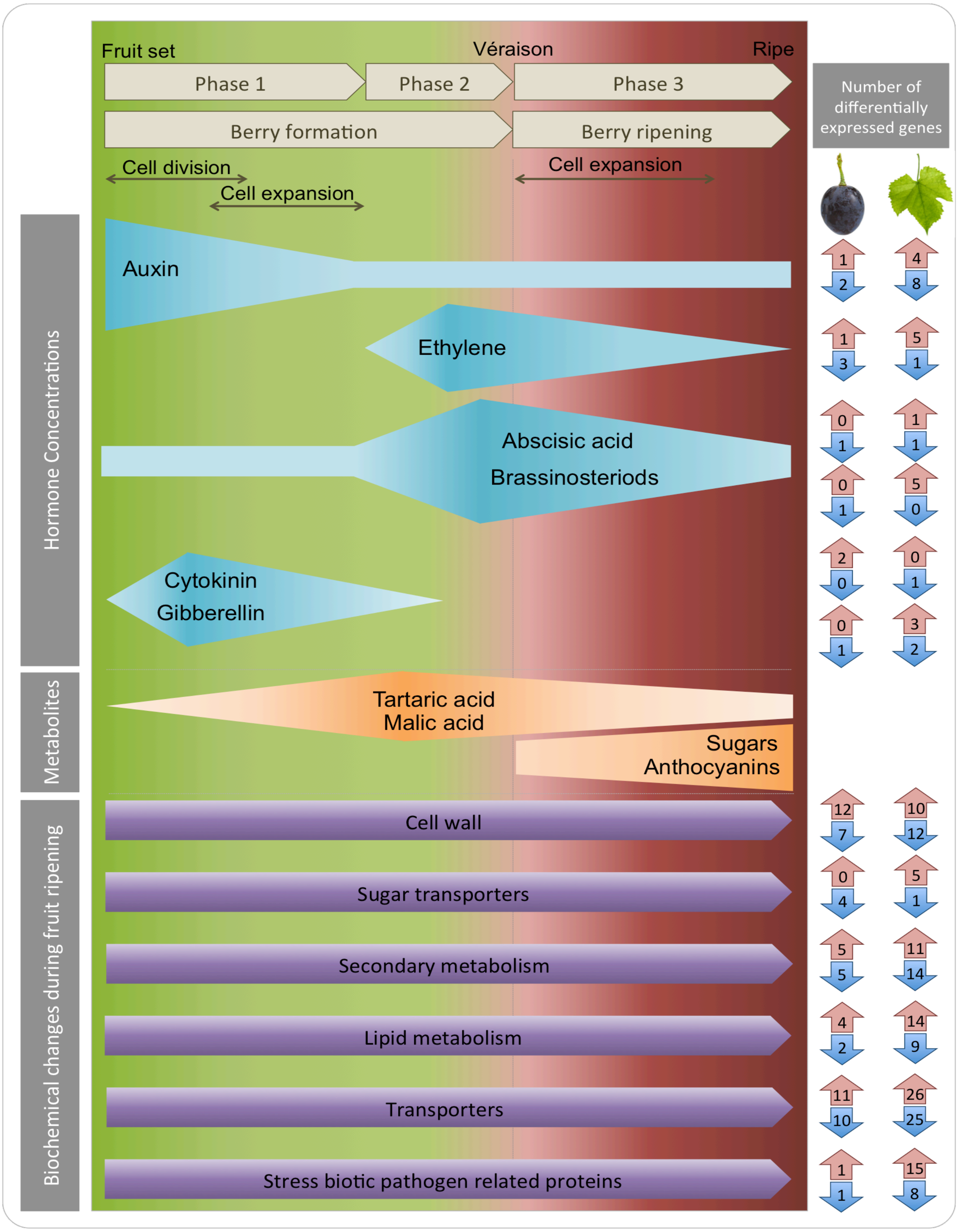


Figure 4.9: *Differential expression of genes involved in grapevine berry ripening: hormone metabolism (13 DEGs in berries and 31 in leaves) and biochemical changes occurring in the berry (62 in berries and 150 in leaves). Up-regulated genes are shown in red arrows and down-regulated genes in blue arrows. Hormonal concentration fluctuations, metabolite accumulation and development phases are shown based on Coombe and McCarthy 2000; Robinson and Davies 2000; Conde et al. 2007; Fortes et al. 2011; Kuhn et al. 2014.*

Fifteen DEGs related to auxin metabolism were identified, 12 in leaves and three in berries (Supplementary data table 4.3, Figure 4.9). Auxin concentration is high in the early stages of fruit development, immediately after fertilization of the ovary, and then gradually decline towards fruit maturity (Deluc et al. 2007; Pilati et al. 2007). The specific concentration balance between auxin and ethylene signals is critical for the tight regulation of berry ripening (Böttcher et al. 2013). Small Auxin Up RNAs (*SAURs*) are the largest family of early auxin response genes (Ren and Gray 2015). In berries, two genes from the *SAUR* gene family were down-regulated and one involved in auxin synthesis, up-regulated. In leaves, one *SAUR* gene was up-regulated. The other three up- and eight down-regulated auxin-associated genes were classified as non-specific polypeptides.

Grapevine is a non-climacteric fruit, and as such does not produce large amounts of ethylene. Nevertheless, ethylene plays an important role to regulate berry ripening (Cramer et al. 2014), by directly inducing ripening genes, but also the expression of ethylene responsive transcription factors (ERFs). There are 130 genes in the AP2/ERF transcription factor family (Licausi et al. 2010; Cramer et al. 2014). It is one of the biggest transcription factor families in plants, containing transcription factors with the APETALA2 (AP2) and the ETHYLENE RESPONSE FACTOR (ERF) domain (Licausi et al. 2010). In this study, 64 AP2/ERF transcription factors expressed in berries were found. Three of these transcription factors were significantly down-regulated in berries and three up-regulated in leaves of young vines compared to those of old vines (Supplementary data table 4.3, Figure 4.9).

Ethylene is a gaseous molecule and can diffuse freely between cells. However, long-distance ethylene responses can also be achieved by transport of its precursor. The precursor of ethylene, 1-aminocyclopropane-1-carboxylate (ACC), is synthesised from S-adenosyl-L-methionine by the enzyme 1-aminocyclopropane-1-carboxylate synthase (ACS) (Van de Poel and Van Der Straeten 2014). The enzyme 1-aminocyclopropane-1-carboxylate oxidase (ACO) then converts the ACC to ethylene. Just before véraison, ethylene concentration is at its highest in berries (Chervin et al. 2004; Pilati et al. 2007). One gene coding for ACS is up-regulated and one coding for ACO is down-regulated in leaves of young vines (Supplementary data table 4.3, Figure 4.9). Therefore, it is possible that a higher concentration of ACC can accumulate in the leaves of younger vines, which can then be transported to berries, where it is converted to ethylene to induce ripening, leading to a higher sugar and lower acid concentration at harvest (Figure 4.2).

Two genes related to abscisic acid (ABA) were identified, one down-regulated in berries and one up-regulated in leaves. One gene involved in ABA synthesis and degradation was down-regulated in leaves (Supplementary data table 4.3, Figure 4.9). Abscisic acid concentration is at its highest just before véraison (Symons et al. 2006) and serves as the signal triggering véraison (Pilati et al. 2007; Kuhn et al. 2014). It induces the ABA binding factor 2 (ABF2) transcription factor, which in turn induces the expression of genes involved in the phenylpropanoid pathway, mainly anthocyanin production. Abscisic acid also stimulates acid invertase activity and sugar transporters, thereby lowering the acid concentration and increasing the sugar concentration in the berry (Cramer et al. 2014). Differential expression of these genes is therefore directly involved in the final berry sugar/acid ratio observed at harvest (Figure 4.2).

Genes involved in brassinosteroid (BR) synthesis and signalling were also found to be differentially expressed in this study. Brassinosteroids are plant hormones essential for normal growth and development (Luan et al. 2016). There is peak in BR concentration before véraison, thereafter the concentration diminish as the fruit ripens (Symons et al. 2006). Two genes related to BR and three involved in BR signal transduction were up-regulated in the leaves of young vines, while one gene involved in BR signal transduction was down-regulated in berries, in comparison to old vines (Supplementary data table 4.3, Figure 4.9).

Cytokinins are involved in fruit set as well as growth and cell division. Higher concentrations of cytokinins are found at fruit set and then decrease from véraison to maturity (Fortes et al. 2015). Two DEGs involved in cytokinin synthesis/degradation was up-regulated in berries and one involved in cytokinin signal transduction down-regulated in leaves of young vines (Supplementary data table 4.3, Figure 4.9).

Gibberellins regulate a number of processes in plant development. They might not be directly involved in fruit ripening, but rather induce cell division and expansion (Fortes et al. 2015). One gene related to gibberellins was down-regulated in berries, and three up- and two down-regulated in leaves of young vines (Supplementary data table 4.3, Figure 4.9).

4.3.6 Differentially expressed genes responsible for biochemical changes during fruit ripening

The results from this study revealed 203 DEGs involved in biochemical changes that occur during fruit maturation (Figure 4.9), under the direction of hormone signalling. Fruit undergo complex biochemical changes during ripening and, in grapevine the transcriptome of the entire vine is modulated as the vine enters maturity (Fasoli et al. 2012). This includes changes in cell wall composition (fruit softening), sugar metabolism, secondary metabolism, expression of stress-related genes, lipid metabolism and transport (Deluc et al. 2007; Grimplet et al. 2007; Pilati et al. 2007; Fortes et al. 2011; Sweetman et al. 2012; Lijavetzky et al. 2012; Fasoli et al. 2012; Dal Santo et al. 2013; Agudelo-Romero et al. 2013; Shangguan et al. 2017). A number of DEGs, belonging to these gene categories that are modulated during ripening, were found in this study (Figure 4.9): cell wall synthesis and degradation (41 DEGs, Supplementary data table 4.4), sugar signalling and transport (ten DEGs, Supplementary data table 4.5), secondary metabolism (35 DEGs, Supplementary data table 4.6), lipid metabolism (29 DEGs, Supplementary data table 4.7), transporters (72 DEGs, Supplementary data table 4.8) and pathogenesis-related genes (25 DEGs, Supplementary data table 4.9).

Softening of the grapevine berry is a major indicator of the ripening process. Fruit softening occurs as a result of changes in the cell wall components. Plant cell walls are mainly composed of cellulose microfibrils embedded in a matrix of polysaccharides and cell wall-related proteins (Robinson and Davies 2000). Changes in ripening berries' cell walls have been shown to involve re-modelling of pectin, xyloglucan and cellulose networks (Nunan et al. 2001), mainly due to the action of expansins, pectin methylesterase, pectate lyase and xyloglucan endotransglycosylase/hydrolase enzymes (Shangguan et al. 2017). Genes coding for these enzymes are amongst those most highly expressed in the berry transcriptome of Pinotage (Section 4.3.2 and Supplementary data table 4.1).

From véraison to maturity there is an increase in arabinogalactan proteins (AGPs) and extensins (Moore et al. 2014). At the time of sampling, it appears that the transcriptome of berries from younger vines were more shifted towards fruit softening than berries from older vines. Seven genes involved in cell wall modification were up-regulated in berries from young vines (Supplementary data table 4.4). Four genes coding for cell wall proteins (AGPs) (Supplementary data table 4.4) and four genes involved in lignin biosynthesis (Supplementary data table 4.6) were down-regulated in berries from young vines.

Grapevine has the capacity to accumulate and store a high concentration of sugar in its berries. The berries import sucrose, where invertase enzymes then catalyse the conversion to hexoses, mainly glucose and fructose (Robinson and Davies 2000). Enzymes involved in sugar metabolism are highly expressed in Pinotage berries at harvest (Section 4.3.2 and

Supplementary data table 4.1). Of the genes classified in the bins “signalling: sugar and nutrient physiology” and “transporters sugar”, four were down-regulated in berries, five up-regulated in leaves and only one down-regulated in leaves (Supplementary data table 4.5).

After véraison protein content of grapes increase, mainly pathogen-related proteins (Monteiro et al. 2007; Ghan et al. 2015) of the class chitinase and thaumatin-like. Both these proteins have been shown to have anti-fungal properties. The function of PR-proteins in berry ripening is not clear, but may be expressed as increased early prevention against diseases as the berry matures (Robinson and Davies 2000). Among the DEGs related to pathogenesis, 14 were up- and seven down-regulated in leaves, while in berries only two genes, one up- and one down-regulated, were classified as PR-proteins (Supplementary data table 4.9).

The majority of DEGs involved in secondary metabolism (isoprenoids, flavonoids, simple phenol metabolism, Supplementary data table 4.6), lipid synthesis and degradation (Supplementary data table 4.7), transportation (Supplementary data table 4.8) and pathogenesis-related proteins (Supplementary data table 4.9) were found in leaves. This demonstrates the importance of the leaves in the berry ripening process, and that the key differences between observed phenotypes (in this case the berries of young and old vines producing different wines) might not be directly evident in the transcriptome of the organ investigated, but also in metabolites and signals from other parts of the plant.

The findings of this study demonstrate that genes involved in ripening are indeed differentially expressed between young and old vines. At the time of harvesting, the sugar content in young-vine berries were higher and the acid concentration lower (Figure 4.2), i.e. they were riper than old-vine berries. In general, genes involved in inducing ripeness were up-regulated in young vines. It is known that older vineyards consisting of bushvines have an uneven ripening (Heyns 2013) pattern and it is our hypothesis that ripeness is delayed in older vines, allowing for a longer time to accumulate secondary metabolites that contribute to the flavour and aroma of the wine. Therefore, wines made from older vineyards may have a more pronounced and deeper character.

Differentially expressed genes related to hormone metabolism (46 DEGs, Supplementary data table 4.3) and other gene categories commonly associated with fruit ripening (212 DEGs, Supplementary data table 4.4 to 4.9) were highlighted in this study. However, due to integrate and complex cross-talk between different genes and gene products, it is difficult to determine specific genes or gene categories that are differentially expressed and responsible for the phenotypic differences between young and old vines. Differential expression of a small number of genes can have a pleiotropic impact on the expression of many genes and their associated networks. Furthermore, not all transcripts are translated into functional gene products. Therefore, future studies that include proteomic and metabolomic (Gapper et al. 2014; Feussner and Polle 2015) analyses will be of great value to elucidate the difference between young and old vines.

4.4 Conclusion

This study provides the first RNA-seq analysis of the berry and leaf transcriptomes of *Vitis vinifera* cv Pinotage. These results provide insights into the gene expression patterns in leaves and mature berries at harvest time, and highlight the contribution of the leaf transcriptome to the ripening process.

This analysis also revealed 86 gene loci, currently not annotated on the PN40024 reference genome, where sequence reads from our transcriptomes align. One novel locus is located on chromosome 6, in-between a cluster of four genes coding for grape ripening-induced proteins, namely grip22. This is a further indication that each cultivar may contain a number of genes not present in the reference genome, and that one grapevine reference genome is not sufficient to represent all cultivars. To further investigate the genes that are present in the Pinotage genome, but absent in Pinot noir, a *de novo* transcriptome assembly was performed and compared with the grapevine reference genome and the assembled Pinotage genome. This part of the study is discussed in Chapter 5.

By studying the transcriptomes of young and old vines, differential gene expression that might contribute to the old-vine character in wines produced from old vineyards, was highlighted. Rigorous RNA pooling and bioinformatic analysis strategies were used, resulting in 925 high-confidence DEGs that demonstrated that the gene expression profile in leaves and berries of young and old vines do differ. Genes involved in critical fruit ripening steps, those associated with hormone signalling, cell wall structural changes, sugar metabolism and secondary metabolism, are differentially expressed between young and old vines. The results of this study suggest that older vines experience a delay in the ripening process and this, together with a decline in yield, might allow the accumulation of more flavour and aroma components in the berry.

The information presented here will contribute to the improvement of the existing grapevine genome annotation and extends the knowledge of grapevine to help establish it as a model plant for non-climacteric fruit ripening. Furthermore, it provides a new, genetic platform to study grapevine ageing and its influence on berry composition.

Supplementary data

Supplementary data table 4.1: Genes expressed in berries with an FPKM of 1000 and higher, with associated EC numbers and enzyme descriptions.

Supplementary data table 4.2: Genes expressed in leaves with an FPKM of 1000 and higher, with associated EC numbers and enzyme descriptions.

Supplementary data table 4.3: Differentially expressed genes classified in the MapMan bin "Hormone metabolism".

Supplementary data table 4.4: Differentially expressed genes classified in the MapMan bin "Cell wall".

Supplementary data table 4.5: Differentially expressed genes in sugar signalling and transport (MapMan bins "Signalling: sugar & nutrient physiology" and "Transport sugars").

Supplementary data table 4.6: Differentially expressed genes classified in the MapMan bin "Secondary metabolism".

Supplementary data table 4.7: Differentially expressed genes classified in the MapMan bin "Lipid metabolism".

Supplementary data table 4.8: Differentially expressed genes classified in the MapMan bin "Transporters".

Supplementary data table 4.9: Differentially expressed genes classified in the MapMan bin "Stress biotic PR-proteins"

4.5 References

Agudelo-Romero P, Erban A, Sousa L, Pais MS, Kopka J and Fortes AM (2013) Search for transcriptional and metabolic markers of grape pre-ripening and ripening and insights into specific aroma development in three Portuguese cultivars. PLOS ONE 8:e60422. doi: 10.1371/journal.pone.0060422

Andrews S, Lindenbaum P, Howard B and Ewels P (2011) FastQC. doi: Available at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc

Antão CM and Malcata FX (2005) Plant serine proteases: biochemical, physiological and molecular features. Plant Physiol Biochem 43:637–650. doi: 10.1016/j.plaphy.2005.05.001

Beavers K (2016) What the heck is old vine wine? Here's everything you need to know. In: VinePair. https://vinepair.com/wine-geekly/what-the-heck-is-old-vine-wine-heres-everything-you-need-to-know/. Accessed 20 Jun 2017

Bolger AM, Lohse M and Usadel B (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30:2114–2120. doi: 10.1093/bioinformatics/btu170

Böttcher C, Burbidge CA, Boss PK and Davies C (2013) Interactions between ethylene and auxin are crucial to the control of grape (Vitis vinifera L.) berry ripening. BMC Plant Biol 13:222. doi: 10.1186/1471-2229-13-222

Chervin C, El-Kereamy A, Roustan J-P, Latché A, Lamon J and Bouzayen M (2004) Ethylene seems required for the berry development and ripening in grape, a non-climacteric fruit. Plant Sci 167:1301–1305. doi: 10.1016/j.plantsci.2004.06.026

Conde C, Silva P, Fontes N, Dias ACP, Tavares RM, Sousa MJ, Agasse A, Delrot S and Gerós H (2007) Biochemical changes throughout grape berry development and fruit and wine quality. Food Glob Sci Books 1–22. doi: 10.1.1.549.9659

Coombe BG and McCarthy MG (2000) Dynamics of grape berry growth and physiology of ripening. Aust J Grape Wine Res 6:131–135. doi: 10.1111/j.1755-0238.2000.tb00171.x

Cramer GR, Ghan R, Schlauch KA, Tillett RL, Heymann H, Ferrarini A, Delledonne M, Zenoni S, Fasoli M and Pezzotti M (2014) Transcriptomic analysis of the late stages of grapevine (Vitis vinifera cv. Cabernet Sauvignon) berry ripening reveals significant induction of ethylene signaling and flavor pathways in the skin. BMC Plant Biol 14:370. doi: 10.1186/s12870-014-0370-8

Dal Santo S, Tornielli GB, Zenoni S, Fasoli M, Farina L, Anesi A, Guzzo F, Delledonne M and Pezzotti M (2013) The plasticity of the grapevine berry transcriptome. Genome Biol 14:r54. doi: 10.1186/gb-2013-14-6-r54

Davies C and Robinson SP (2000) Differential screening indicates a dramatic change in mRNA profiles during grape berry ripening. Cloning and characterization of cDNAsencoding putative cell wall and stress response proteins. Plant Physiol 122:803–812. doi: 10.1104/pp.122.3.803

Degu A, Hochberg U, Sikron N, Venturini L, Buson G, Ghan R, Plaschkes I, Batushansky A, Chalifa-Caspi V, Mattivi F et al. (2014) Metabolite and transcript profiling of berry skin during fruit development elucidates differential regulation between Cabernet Sauvignon and Shiraz cultivars at branching points in the polyphenol pathway. BMC Plant Biol 14:188. doi: 10.1186/s12870-014-0188-4

Deluc LG, Grimplet J, Wheatley MD, Tillett RL, Quilici DR, Osborne C, Schooley DA, Schlauch KA, Cushman JC and Cramer GR (2007) Transcriptomic and metabolite analyses of Cabernet Sauvignon grape berry development. BMC Genomics 8:429. doi: 10.1186/1471-2164-8-429

Easton S (2016) Old vines – do they make the best wines? In: WineWisdom. http://www.winewisdom.com/articles/techie/old-vines-do-they-make-the-best-wines/. Accessed 20 Jun 2017

Fasoli M, Dal Santo S, Zenoni S, Tornielli GB, Farina L, Zamboni A, Porceddu A, Venturini L, Bicego M, Murino V et al. (2012) The grapevine expression atlas reveals a deep transcriptome shift driving the entire plant into a maturation program. Plant Cell 24:3489–3505. doi: 10.1105/tpc.112.100230

Fortes AM, Agudelo-Romero P, Silva MS, Ali K, Sousa L, Maltese F, Choi YH, Grimplet J, Martinez- Zapater JM, Verpoorte R et al. (2011) Transcript and metabolite analysis in Trincadeira cultivar reveals novel information regarding the dynamics of grape ripening. BMC Plant Biol 11:149. doi: 10.1186/1471-2229-11-149

Fortes AM, Teixeira RT and Agudelo-Romero P (2015) Complex Interplay of Hormonal Signals during Grape Berry Ripening. Molecules 20:9326–9343. doi: 10.3390/molecules20059326

Fridjhon M (2016) How to save SA's old vines; SA Wine Ratings, News, Opinion & Analysis. In: Winemag.co.za. http://winemag.co.za/michael-fridjhon-how-to-save-sas-old-vines/. Accessed 20 Jun 2017

Gapper NE, Giovannoni JJ and Watkins CB (2014) Understanding development and ripening of fruit crops in an 'omics' era. Hortic Res 1:14034. doi: 10.1038/hortres.2014.34

Ghan R, Petereit J, Tillett RL, Schlauch KA, Toubiana D, Fait A and Cramer GR (2017) The common transcriptional subnetworks of the grape berry skin in the late stages of ripening. BMC Plant Biol 17:94. doi: 10.1186/s12870-017-1043-1

Ghan R, Van Sluyter SC, Hochberg U, Degu A, Hopper DW, Tillet RL, Schlauch KA, Haynes PA, Fait A and Cramer GR (2015) Five "omic" technologies are concordant in differentiating the biochemical characteristics of the berries of five grapevine (Vitis vinifera L.) cultivars. BMC Genomics 16:946. doi: 10.1186/s12864-015-2115-y

Grimplet J, Deluc LG, Tillett RL, Wheatley MD, Schlauch KA, Cramer GR and Cushman JC (2007) Tissue-specific mRNA expression profiling in grape berry tissues. BMC Genomics 8:187. doi: 10.1186/1471-2164-8-187

Grimplet J, Tello J, Laguna N and Ibáñez J (2017) Differences in flower transcriptome between grapevine clones are related to their cluster compactness, fruitfulness, and berry size. Front Plant Sci. doi: 10.3389/fpls.2017.00632

Guo D-L, Xi F-F, Yu Y-H, Zhang X-Y, Zhang G-H and Zhong G-Y (2016) Comparative RNA-Seq profiling of berry development between table grape 'Kyoho' and its early-ripening mutant 'Fengzao'. BMC Genomics. doi: 10.1186/s12864-016-3051-1

Hawkins B (2016) The Real Story Behind “Old Vine” Wines in South Africa. In: Explore Sideways. https://exploresideways.com/whats-the-deal-with-old-vines-and-why-do-they-matter-to-the-south-african-wine-industry/. Accessed 20 Jun 2017

Heyns E (2013) Old vines new opportunities? In: Wineland Mag. http://www.wineland.co.za/old-vines-new-opportunities/. Accessed 20 Jun 2017

Hooke H (2016) Defining old vines. In: Real Rev. https://www.therealreview.com/2016/12/01/defining-old-vines/. Accessed 20 Jun 2017

Jaillon O, Aury J-M, Noel B, Policriti A, Clepet C, Casagrande A, Choisne N, Aubourg S, Vitulo N, Jubin C et al. (2007) The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. Nature 449:463–467. doi: 10.1038/nature06148

Jain M (2012) Next-generation sequencing technologies for gene expression profiling in plants. Brief Funct Genomics 11:63–70. doi: 10.1093/bfgp/elr038

Kong L, Zhang Y, Ye Z-Q, Liu X-Q, Zhao S-Q, Wei L and Gao G (2007) CPC: assess the protein-coding potential of transcripts using sequence features and support vector machine. Nucleic Acids Res 35:W345–W349. doi: 10.1093/nar/gkm391

Kuhn N, Guan L, Dai ZW, Wu B-H, Lauvergeat V, Gomès E, Li S-H, Godoy F, Arce-Johnson P and Delrot S (2014) Berry ripening: recently heard through the grapevine. J Exp Bot 65:4543–4559. doi: 10.1093/jxb/ert395

Licausi F, Giorgi FM, Zenoni S, Osti F, Pezzotti M and Perata P (2010) Genomic and transcriptomic analysis of the AP2/ERF superfamily in Vitis vinifera. BMC Genomics 11:719. doi: 10.1186/1471-2164-11-719

Lijavetzky D, Carbonell-Bejerano P, Grimplet J, Bravo G, Flores P, Fenoll J, Hellín P, Oliveros JC and Martínez-Zapater JM (2012) Berry Flesh and Skin Ripening Features in Vitis vinifera as Assessed by Transcriptional Profiling. PLOS ONE 7:e39547. doi: 10.1371/journal.pone.0039547

Luan L-Y, Zhang Z-W, Xi Z-M, Huo S-S and Ma L-N (2016) Brassinosteroids regulate anthocyanin biosynthesis in the ripening of grape berries. South Afr J Enol Vitic 34:196–203.

May P, Wienkoop S, Kempa S, Usadel B, Christian N, Rupprecht J, Weiss J, Recuenco-Munoz L, Ebenhöh O, Weckwerth W et al. (2008) Metabolomics- and proteomics-assisted genome annotation and analysis of the draft metabolic network of Chlamydomonas reinhardtii. Genetics 179:157–166. doi: 10.1534/genetics.108.088336

Monteiro S, Piçarra-Pereira MA, Loureiro VB, Teixeira AR and Ferreira RB (2007) The diversity of pathogenesis-related proteins decreases during grape maturation. Phytochemistry 68:416–425. doi: 10.1016/j.phytochem.2006.11.014

Moore JP, Fangel JU, Willats WGT and Vivier MA (2014) Pectic-β(1,4)-galactan, extensin and arabinogalactan–protein epitopes differentiate ripening stages in wine and table grape cell walls. Ann Bot 114:1279–1294. doi: 10.1093/aob/mcu053

Muñoz-Espinoza C, Di Genova A, Correa J, Silva R, Maass A, González-Agüero M, Orellana A and Hinrichsen P (2016) Transcriptome profiling of grapevine seedless segregants during berry development reveals candidate genes associated with berry weight. BMC Plant Biol 16:104. doi: 10.1186/s12870-016-0789-1

Nunan KJ, Davies C, Robinson SP and Fincher GB (2001) Expression patterns of cell wall-modifying enzymes during grape berry development. Planta 214:257–264. doi: 10.1007/s004250100609

Pervaiz T, Haifeng J, Haider MS, Cheng Z, Cui M, Wang M, Cui L, Wang X and Fang J (2016) Transcriptomic analysis of grapevine (cv. Summer Black) leaf, using the Illumina platform. PLOS ONE 11:e0147369. doi: 10.1371/journal.pone.0147369

Pilati S, Perazzolli M, Malossini A, Cestaro A, Demattè L, Fontana P, Dal Ri A, Viola R, Velasco R and Moser C (2007) Genome-wide transcriptional analysis of grapevine berry ripening reveals a set of genes similarly modulated during three seasons and the occurrence of an oxidative burst at vèraison. BMC Genomics 8:428. doi: 10.1186/1471-2164-8-428

Reid KE, Olsson N, Schlosser J, Peng F and Lund ST (2006) An optimized grapevine RNA isolation procedure and statistical determination of reference genes for real-time RT-PCR during berry development. BMC Plant Biol 6:27. doi: 10.1186/1471-2229-6-27

Ren H and Gray WM (2015) SAUR Proteins as Effectors of Hormonal and Environmental Signals in Plant Growth. Mol Plant 8:1153–1164. doi: 10.1016/j.molp.2015.05.003

Robinson SP and Davies C (2000) Molecular biology of grape berry ripening. Aust J Grape Wine Res 6:175–188. doi: 10.1111/j.1755-0238.2000.tb00177.x

Royo C, Carbonell-Bejerano P, Torres-Pérez R, Nebish A, Martínez Ó, Rey M, Aroutiounian R, Ibáñez J and Martínez-Zapater JM (2016) Developmental, transcriptome, and genetic alterations associated with parthenocarpy in the grapevine seedless somatic variant Corinto bianco. J Exp Bot 67:259–273. doi: 10.1093/jxb/erv452

Shangguan L, Mu Q, Fang X, Zhang K, Jia H, Li X, Bao Y and Fang J (2017) RNA-sequencing reveals biological networks during table grapevine ('Fujiminori') fruit development. PLOS ONE 12:e0170571. doi: 10.1371/journal.pone.0170571

Sweetman C, Wong DC, Ford CM and Drew DP (2012) Transcriptome analysis at four developmental stages of grape berry (Vitis vinifera cv. Shiraz) provides insights into regulated and coordinated gene expression. BMC Genomics 13:691. doi: 10.1186/1471-2164-13-691

Symons GM, Davies C, Shavrukov Y, Dry IB, Reid JB and Thomas MR (2006) Grapes on steroids. Brassinosteroids are involved ingrape berry ripening. Plant Physiol 140:150–158. doi: 10.1104/pp.105.070706

Szabo J (2017) Growing old in South Africa: The old vines project. In: WineAlign. http://www.winealign.com/articles/2017/06/01/growing-old-in-south-africa-the-old-vines-project. Accessed 20 Jun 2017

Thimm O, Bläsing O, Gibon Y, Nagel A, Meyer S, Krüger P, Selbig J, Müller LA, Rhee SY and Stitt M (2004) MapMan: a user-driven tool to display genomics data sets onto diagrams of metabolic pathways and other biological processes. Plant J 37:914–939. doi: 10.1111/j.1365-313X.2004.02016.x

Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, Pimentel H, Salzberg SL, Rinn JL and Pachter L (2012) Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat Protoc 7:562–578. doi: 10.1038/nprot.2012.016

Usadel B, Poree F, Nagel A, Lohse M, Czedik-Eysenberg A and Stitt M (2009) A guide to using MapMan to visualize and compare Omics data in plants: a case study in the crop species, Maize. Plant Cell Environ 32:1211–1229. doi: 10.1111/j.1365-3040.2009.01978.x

Van de Poel B and Van Der Straeten D (2014) 1-aminocyclopropane-1-carboxylic acid (ACC) in plants: more than just the precursor of ethylene! Front Plant Sci. doi: 10.3389/fpls.2014.00640

Van Wyk E (2016) Aftertaste old vines : lifestyle wines. Prejudice 16:64–65.

Velasco R, Zharkikh A, Troggio M, Cartwright DA, Cestaro A, Pruss D, Pindo M, FitzGerald LM, Vezzulli S, Reid J et al. (2007) A high quality draft consensus sequence of the genome of a heterozygous grapevine variety. PLOS ONE 2:e1326. doi: 10.1371/journal.pone.0001326

Venturini L, Ferrarini A, Zenoni S, Tornielli GB, Fasoli M, Santo SD, Minio A, Buson G, Tononi P, Zago ED et al. (2013) De novo transcriptome characterization of Vitis vinifera cv. Corvina unveils varietal diversity. BMC Genomics 14:41. doi: 10.1186/1471-2164-14-41

Vitulo N, Forcato C, Carpinelli EC, Telatin A, Campagna D, D'Angelo M, Zimbello R, Corso M, Vannozzi A, Bonghi C et al. (2014) A deep survey of alternative splicing in grape reveals changes in the splicing machinery related to tissue, stress condition and genotype. BMC Plant Biol 14:99. doi: 10.1186/1471-2229-14-99

Waters E j., Alexander G, Muhlack R, Pocock K f., Colby C, O'neill B k., Høj P b. and Jones P (2005) Preventing protein haze in bottled white wine. Aust J Grape Wine Res 11:215–225. doi: 10.1111/j.1755-0238.2005.tb00289.x

Wrzaczek M, Brosché M, Salojärvi J, Kangasjärvi S, Idänheimo N, Mersmann S, Robatzek S, Karpiński S, Karpińska B and Kangasjärvi J (2010) Transcriptional regulation of the CRK/DUF26 group of Receptor-like protein kinases by ozone and plant hormones in Arabidopsis. BMC Plant Biol 10:95. doi: 10.1186/1471-2229-10-95

Zamboni A, Carli MD, Guzzo F, Stocchero M, Zenoni S, Ferrarini A, Tononi P, Toffali K, Desiderio A, Lilley KS et al. (2010) Identification of putative stage-specific grapevine berry biomarkers and omics data integration into networks. Plant Physiol 154:1439–1459. doi: 10.1104/pp.110.160275

Zenoni S, Ferrarini A, Giacomelli E, Xumerle L, Fasoli M, Malerba G, Bellin D, Pezzotti M and Delledonne M (2010) Characterization of transcriptional complexity during berry development in Vitis vinifera using RNA-seq. Plant Physiol 152:1787–1795. doi: 10.1104/pp.109.149716

Online resources

Bedtools: http://bedtools.readthedocs.io/en/latest/index.html

CPC: http://cpc.cbi.pku.edu.cn/

CRIBI: http://genomes.cribi.unipd.it/grape/

FastQC: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/

fastq-tools: http://homes.cs.washington.edu/~dcjones/fastq-tools/

GrapeCyc: https://www.plantcyc.org/databases/grapecyc/7.0

“I am old” homepage: http://iamold.withtank.com/home/

Illumina mRNA stranded library preparation protocol: https://support.illumina.com/content/dam/illumina-support/documents/documentation/chemistry_documentation/samplepreps_truseq/truseqstrandedmrna/truseq-stranded-mrna-sample-prep-guide-15031047-e.pdf

MapMan: http://mapman.gabipd.org/

Chapter 5: Pinotage *De novo* Transcriptome Assembly

5.1 Introduction

The Pinotage grapevine cultivar is the F1 progeny of a controlled cross between Pinot noir and Cinsaut (discussed in Chapter 3; Vivier and Pretorius 2000). The Pinot noir genome has been sequenced, but little is known about the genetic background and genome sequence of Cinsaut. Pinotage has remarkably different characteristics from both its parents (Orffer and Visser 2009). Pinotage is a productive cultivar, ripening earlier than Pinot noir. The berries generally have a higher sugar concentration and thicker skin, producing a darker wine that is rich in anthocyanins and tannins, with typical berry, cherry and plum flavours (Pinotage Association: http://pinotage.co.za/recognition/).

Aroma and bouquet are two of the most significant characteristics of a wine and are mainly due to the volatile compounds found in wine. The sensory profile of a wine is the result of a series of complex biochemical processes (Conde et al. 2007). Accumulation of volatile compounds begins in the grapes, forming the primary aroma of a wine. Secondary aromas, also called the bouquet, are derived from the volatile components formed during the winemaking and ageing process. Certain volatiles, referred to as impact odorants, are characteristic of particular wine varieties. Pinotage wines have a very distinctive aroma profile. Young wines of this cultivar are renowned for their fruity bouquet. The typical Pinotage bouquet does not appear in either the must or the grape, indicative that this typical character is formed during fermentation (Van Wyk et al. 1979; Weldegergis et al. 2011a; Weldegergis et al. 2011b). Isoamyl acetate, when present in relatively large concentrations, is responsible for this typical bouquet (Van Wyk et al. 1979) and Pinotage generally has a higher isoamyl acetate and isoamyl alcohol concentration than other South African red wines (Louw et al. 2009).

Aroma and bouquet are however not the only important characteristics of wine. Sweetness, alcohol concentration, acidity, tannin content, to name a few, all contribute to the flavour, body and mouth feel, and essentially the quality of the wine. There is more than a thousand compounds in wine that could influence these characteristics (Conde et al. 2007). Diverse factors affect the level of each of these chemical compounds in wine and complex biochemical networks produce these compounds. This, together with how the vine interacts with the environment, its suitability for certain climates and/or soil types, and how it responds to biotic and abiotic stressors, makes a cultivar unique. It is often difficult to establish the association between the underlying genetics of a vine and the interplay with environmental factors that shape the chemical constituents, and ultimately the resulting sensory properties of a wine. However, in an exceedingly competitive international and local wine market, it is imperative to invest in molecular grapevine research to understand the genetics of a vine and cultivar, to preserve its unique character.

In another part of the study (Chapter 4), 86 novel loci were identified that are not annotated in the current grapevine genome annotation. This prompted a further investigation into genes present in the Pinotage genome and absent in the reference genome. A *de novo* transcriptome assembled from stranded mRNA sequencing data, complemented with the available Pinotage genome data, was used to identify these genes. The relationship between the observed these genes and Pinotage/Pinot noir phenotypic differences were evaluated. It was shown that most of the Pinotage-specific genes are involved in the stress response network.

5.2 Methods and Materials

5.2.1 Sample collection, RNA extraction and sequencing

The same RNA-seq data as in Chapter 4 (Sections 4.2.1 to 4.4.4) were used. In short, RNA were extracted from 18 berry and leaf samples (nine young and nine old vines) and pooled in groups of three to form 12 samples (six berry and six leaf samples, Figure 4.1). RNA was prepared using the Illumina TruSeq stranded mRNA protocol according to the manufacturer’s instructions and sequenced on the Illumina HiSeq2000, generating 125nt paired-end reads.

5.2.2 De novo transcriptome assembly

Quality trimmed mRNA reads from all samples (six berry and six leaf samples) were combined and *de novo* assembled into putative transcripts using Trinity (Grabherr et al. 2011), with default parameters (Haas et al. 2013), specifying that a strand-specific RNA-seq library was generated with the dUTP method (Parameters --SS_lib_type RF).

The assembled transcripts were assessed with the EvidentialGene pipeline (Gilbert 2013), using default parameters. In brief, the EvidentialGene pipeline extracts the best, biologically meaningful transcript sequences by assessing coding potential and BLAST (Altschul et al. 1990) alignments to identify coding domains. EvidentialGene then clusters the transcripts and select the best representative transcript for each transcript cluster (unigene). From the EvidentialGene transcripts, only those longer than 150nt (50aa) with a positive orientation were selected. Next, these transcripts were subjected to BLAST (Altschul et al. 1990) against the NCBI nucleotide database (National Center for Biotechnology Information, https://blast.ncbi.nlm.nih.gov), restricting the search to land plants (taxid: 3193). Transcripts with an alignment e-value of less than 0.001 were selected for further analysis.

5.2.3 Comparison of genome and transcriptome data

The selected transcripts were aligned to the assembled Pinotage genome sequence (discussed in Chapter 3) using GMAP (Wu and Watanabe 2005), with a minimum trimmed coverage of 90%, minimum identity of 98% and allowing chimeras. The same parameters were used to align the transcripts to the Pinot noir PN40024 (Jaillon et al. 2007)(NCBI Bioproject: PRJEA18785) and ENTAV115 (Velasco et al. 2007) (NCBI Bioproject: PRJEA18357) genomes.

5.2.4 Transcript functional assignment

Transcripts aligning only to Pinotage were assigned to MapMan functional bins (Thimm et al. 2004; Usadel et al. 2009) (http://mapman.gabipd.org/) with Mercator (May et al. 2008), allowing only one hit per transcript and a BLAST cut-off of 1; using the same approach differentially expressed genes were assigned to functional bins (Chapter 4, Section 4.2.8).

5.3 Results and Discussion

5.3.1 Identifying Pinotage genes absent from the Pinot noir genome

A total of 479,005,854 read pairs were retained after quality trimming and filtering and were *de novo* assembled into 314,672 transcripts using Trinity. The EvidentialGene pipeline selected 45,999 transcripts, each being the most complete and longest representative transcript of a gene. Transcripts in the wrong orientation, too short, or having non-plant BLAST hits were removed (Table 4.1 and Table 4.2). The remaining 24,527 transcripts each represents 1.244 isoforms on average (as classified by EvidentialGene) and has an average length of 1,529nt.

Table 5.1: Number of transcripts remaining after each filtering step.

Trinity assembly	314,627
EvidentialGene	45,999
Transcripts in forward orientation	35,988
Transcripts coding for > 50 amino acids	35,527
Transcripts with plant BLAST hits	24,527

Table 5.2: Sources of non-plant BLAST hits.

	Number of BLAST hits
Bacteria	1,550
Fungi	552
Insects	451
Mammals	201
Viruses	49
Nematodes	42
Mites	37
Arthropods	34
Birds	27
Platyhelminthes	23
Molluscs	18
Amphibia	8
Other	134
No hit	7,874
Total	**11,000**

The remaining high quality transcripts were aligned to the Pinotage, as well as the PN40024 and ENTAV115 genomes (Figure 5.1), as an indication of the similarity between these genomes. A total of 19,491 transcripts aligned to the reference PN40024 genome, and an additional 2,209 to the ENTAV115 genome. A similar study using *de novo* transcriptome assembly of Corvina RNA-seq data recovered 15,161 known genes (Venturini et al. 2013).

A total of 16,661 transcripts were shared among all three genomes, while Pinotage share 1,716 with ENTAV115 and 1,156 with PN40024. PN40024 is a highly inbred homozygous line, and might have lost many genes during the self-breeding process, while ENTAV115 is a commercially grown heterozygous cultivar. It is therefore expected that Pinotage will share more gene similarity with ENTAV115 than with PN40024. A number of transcripts (2,466) did not align to Pinotage. Of the three genomes, PN40024 is the most complete and is assembled into 2,093 contigs (33 super-scaffolds), while ENTAV115 is assembled into 66,164 contigs. The

Pinotage genome (discussed in Chapter 3) is not as completely assembled as PN40024, and it is expected that a number of transcripts will not align to Pinotage.

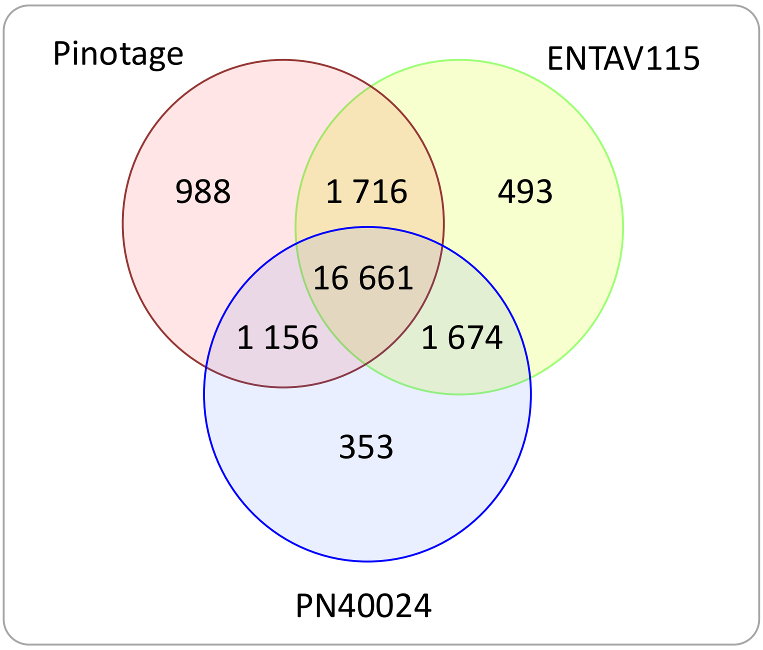


Figure 5.1: *Number of de novo assembled transcripts aligning to PN40024 (Bioproject: PRJEA18785), ENTAV115 (Bioproject: PRJEA18357) and the Pinotage draft genome (Chapter 3 of this study) sequences, respectively. Of the 24,527 predicted transcripts, a total of 22 886 aligned to the three genomes, of which 988 aligned exclusively to Pinotage.*

Plant genomes are complex, plastic and variable, and a single reference genome can not represent all the genes contained in a species. Therefore, the concept of a pan-genome is very applicable to plant genomes. The pan-genome includes all genes and other genetic elements contained in a group of individuals, for example in plants, all the cultivars or varieties of a species. It consists of a core genome containing genes shared by all the genotypes, and a dispensable genome composed of genetic elements only present in some genotypes, i.e. the unique genes or genes with significant sequence variation (Morgante 2006; Morgante et al. 2007; Marroni et al. 2014; Vernikos et al. 2015; Casacuberta et al. 2016; Cardone et al. 2016b). Intra-species variation is conferred by the dispensable part of the genome, giving rise to phenotypically distinct varieties or cultivars.

An estimated 8% of the grapevine genome is affected by variations, including copy number variations (CNVs) and present/absent variations (PAVs) (Cardone et al. 2016b). This study identified 988 transcripts (Supplementary data 5.1) that align only to Pinotage. The DNA for the genome and RNA for the transcriptome assembly in this study were sampled from different locations, and independently extracted, sequenced and assembled. This, and the rigorous filtering criteria used for transcript selection, led to the conclusion that the 988 genes that align with 98% identity to the Pinotage genome are true genes present in Pinotage while absent from the Pinot noir reference genomes, and not sequencing or assembly artefacts. These 988 genes (hereafter called Pinotage genes) are most likely not specific to the Pinotage genome, but are the genetic contribution of the Cinsaut crossing parent.

5.3.2 Classifying Pinotage genes

To further investigate the 988 identified genes, they were assigned to MapMan functional classification bins (Table 5.3). "Stress" is the largest MapMan bin, with 131 assigned transcripts. Other genes, although not classified in the "stress" bin, might also indirectly be involved in the greater stress response network. Since adaptation to environmental pressures, both biotic and abiotic, is critical for the evolvement of new varieties and one of the key drivers for new cultivar development, it is not surprising that a large number of genes that is responsible for intra-species variety, is part of the stress response network.

Abiotic and biotic stress responses consist of integrated signalling networks that modulate gene expression to produce more secondary metabolites and components for cell wall strengthening. Plants react to environmental stress conditions through a few conserved signalling pathways. Biotic stress response in plants can broadly be divided into two pathways, basal or non-host response, and R-mediated or host response; however, there is a substantial amount of cross-talk between these pathways (Ben Rejeb et al. 2014; Gill et al. 2015; Péros et al. 2015).

For an in-depth look at the function of these Pinotage genes, they are discussed in the following five sections: basal/non-host responses (Section 5.3.3), R-mediated/host-responses (Section 5.3.4), transcription factors (Section 5.3.5), the genes modulated by transcription factors (Section 5.3.6), and abiotic stress (Section 5.3.7). Figure 5.2 presents an overview of the biotic and abiotic stress response network in *Vitis vinifera* and indicate the metabolic steps that where the Pinotage genes may be involved.

Table 5.3: Mercator classification of the 988 Pinotage genes absent from the Pinot noir genome. Genes marked with * appear in Figure 5.2, genes marked with § appear in Figure 5.3.

Primary classification		Secondary classification	Number of Pinotage genes
Cell	35	Cycle	5
		Division	4
		Organisation	21
		Vesicle transport	5
Cell wall*	17	Cell wall proteins	4
		Cellulose synthesis	3
		Degradation	4
		Modification	2
		Pectin esterases	2
		Precursor synthesis	2
Development	15		
DNA	19	Repair	1
		Synthesis/chromatin structure	12
		Unspecified	6
Hormone metabolism	20	Abscisic acid*	6
		Auxin	3
		Brassinosteroid	3
		Ethylene*	5
		Gibberelin	3
Lipid metabolism	12	Exotics (steroids, squalene etc.)	3
		Fatty acid synthesis and elongation	4
		Phospholipid synthesis	1
		Lipid degradation	3
		Lipid transfer proteins etc.	1
Protein	111	Amino acid activation	3
		Assembly and co-factor ligation	4
		Degradation*	50
		Glycosylation	4
		Postranslational modification	21
		Synthesis	20
		Targeting	9
Redox (respiratory burst)*	12		
RNA	93	RNA binding	10
		Processing	8
		Regulation of transcription (17*)	70
		Transcription	5
Secondary metabolism*	13	Flavonoids (Anthocyanin production)§	2
		Isoprenoids (Terpenoid production)§	5
		Phenylpropanoids (Ligin production)§	4
		Sulfur-containing	1
		Wax	1
Stress	131	Abiotic - heat*	5
		Abiotic - unspecific *	8
		Biotic - unspecific*	24
		Biotic - Pathogenesis-related genes*	94
Signalling	69	G-proteins	5
		MAP kinases*	2
		Calcium*	10
		Receptor kinases (Leucine-rich repeat receptor kinase 18*)	45
		Other	7

Table 5.3 continued: Mercator classification of the 988 Pinotage genes absent from the Pinot noir genome. Genes marked with * appear in Figure 5.2, genes marked with § appear in Figure 5.3.

Primary classification		**Secondary classification**	**Number of Pinotage genes**
Transport	39	p- and v-ATPases	7
		Metal	5
		ABC transporters	7
		Other	2
Miscellaneous	98	UDP glucosyl and glucuronyl transferases	5
		Nitrilases	3
		Glutathione S transferases	3
		Cytochrome P450	5
		GDSL-motif lipase	4
		Other	78
Not assigned to ontology	304	Pentatricopeptide (PPR) repeat-containing protein	22
		DC1 domain containing protein	4
		Proline rich protein	3
		Other	9
		Unknown	266

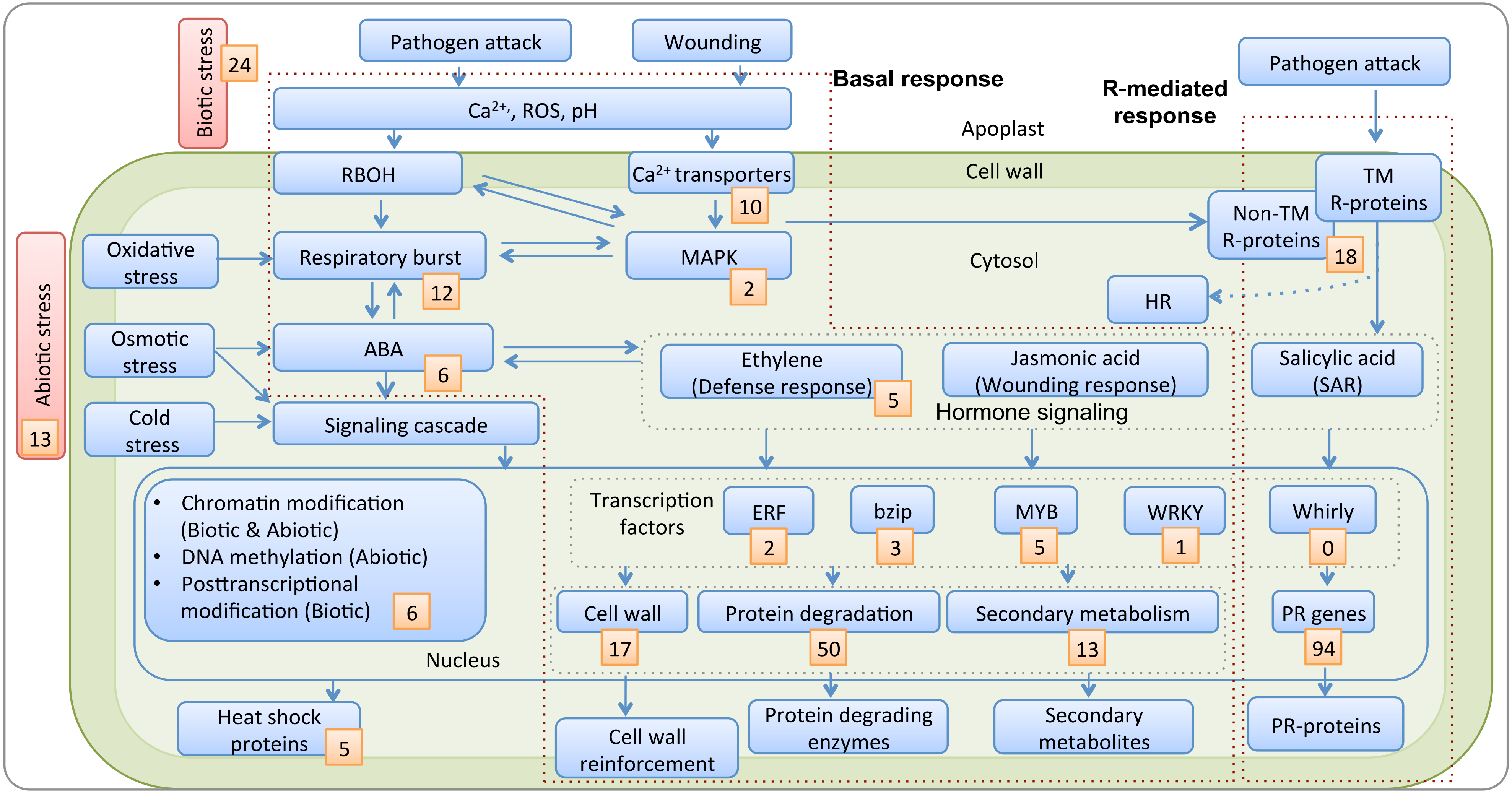


Figure 5.2: *The biotic and abiotic stress response network in Vitis vinifera. The two main pathways of biotic stress response, basal and R-mediated are indicated. The number of Pinotage genes involved in each step is indicated in orange blocks (corresponds to genes marked with * in Table 5.3). ABA: abscisic acid, HR: hypersensitive response, MAPK: mitogen-activated protein kinase, PR: pathogenesis-related, R: resistance, RBOH: Respiratory burst oxidase homologue proteins, SAR: systemic acquired resistance, TM: transmembrane.*

5.3.3 Basal or non-host resistance

Basal or non-host defence responses (Figure 5.2) are triggered by recognition of non-specific molecular signatures common to certain groups of pathogens, such as flagellin, a protein present in flagella of gram-negative bacteria. These elicitors are called microbial- or pathogen-associated molecular patterns (MAMPs or PAMPs) and are recognised by transmembrane pattern recognition receptors (PRRs). PRRs convey the message across the membrane that a pathogen is being detected and sets in motion the defence response cascade, called PAMP-triggered immunity (PTI). A pathogen might be successful in overcoming the basal defence by deploying effectors to interfere with PTI. These effectors can be recognised resistance proteins (R-proteins) inside the cell membrane, resulting in effector-triggered immunity (ETI) (discussed in Section 5.3.4) (Jones and Dangl 2006; Zipfel 2008).

Ten Pinotage genes were classified in the “calcium signalling” bin (Table 5.3, Figure 5.2). Pattern recognition receptors (PRRs) can use free calcium ions (Ca^{2+}) as secondary messengers in signal transduction pathways. Calcium ions will bind to phosphate and form an insoluble precipitate, which would interfere with phosphate-based metabolism; therefore cells actively export Ca^{2+} from the cytoplasm to organelles and the apoplast. The resulting Ca^{2+} osmotic difference between the cytoplasm and apoplast allows for the rapid generation of signals by changing the cytoplasmic Ca^{2+} levels through membrane-localized Ca^{2+}-channels. Downstream Ca^{2+} sensitive receptors can then further convey the message (Sewelam et al. 2016; Stael et al. 2012).

When stress is sensed, a respiratory burst occurs and a rapid accumulation of reactive oxygen species (ROS) is observed (Figure 5.2). High levels of ROS lead to the hypersensitive response (HR) and cell death, to combat the spread of infection. At lower levels, ROS mainly serve as signalling molecules (Lamb and Dixon 1997). Reactive oxygen species interacts with the mitogen-activated protein kinase (MAPK) cascade and abscisic acid (ABA) production. The MAPKs form a conserved signal transduction pathway, comprising of three kinases: MAP kinase kinase kinase (MAPKKK), a MAP kinase kinase (MAPKK), and a MAP kinase (MAPK) (Rodriguez et al. 2010; Wang et al. 2014b; Ben Rejeb et al. 2014). To date, 14 MAPK, five MAPKK, and 62 MAPKKK encoding genes were identified in the grapevine genome (Wang et al. 2014b; Wang et al. 2014a; Çakir and Kılıçkaya 2015).

Reactive oxygen species and the MAP kinase pathway stimulate the production of ABA (Figure 5.2) (Rodriguez et al. 2010; Sewelam et al. 2016). Abscisic acid plays a pivotal role in both biotic and abiotic stress signalling, and mediating the cross-talk between these responses. The role of ABA in plant defence is complex, and may vary according to the nature of the plant-pathogen interaction, as ABA can negatively or positively regulate the defence response (Bari and Jones 2009; Sewelam et al. 2016; Pandey et al. 2017). The ROS/MAPK/ABA interaction forms the core stress signalling system in plants and are tightly regulated. Pinotage genes involved in these steps were observed in our data and included respiratory burst (12), MAP kinase cascade (two) and ABA metabolism (six) (Table 5.3, Figure 5.2).

The presence of ABA activates the production of hormones in the cell (Figure 5.2). The three phytohormones predominantly involved in mediating the defence response in plants are: ethylene, jasmonic acid and salicylic acid. Ethylene activates the pathogen defence pathway (Bari and Jones 2009), and five genes involved in ethylene metabolism (Table 5.3, Figure 5.2) were identified. Jasmonic acid is mostly involved in wounding response and response to herbivores, while salicylic acid is a mediator in the systemic acquired resistance (SAR) pathway (discussed in Section 5.3.4).

5.3.4 R-mediated or host resistance

In R-mediated resistance (Figure 5.2), protein products of disease resistance genes (R-genes) act as immune receptors that sense the presence of pathogens by recognizing pathogen effectors (avirulence genes [Avr]). The R-proteins then activate the effector-triggered immunity (ETI) cascade. Most R-proteins contain an N-terminal nucleotide-binding site (NBS), responsible for nucleotide binding and signal transduction, and a leucine-rich repeat (LRR) domain responsible for pathogen-recognition specificity. They may also contain a transmembrane (TM) domain and a coiled-coil (CC) domain (Gaspero and Cipriani 2003; McHale et al. 2006; DeYoung and Innes 2006; Liu et al. 2007).

R-genes are classified into four main classes: Receptor-like kinase (RLK, serine/threonine kinases), NBS-LRR, LRR-TM and TM-CC (Gaspero and Cipriani 2003; McHale et al. 2006; DeYoung and Innes 2006; Liu et al. 2007). The RLKs and NBS-LRRs lack a transmembrane domain and reside inside the cell membrane (Martin et al. 2003). The NBS-LRR-encoding genes are the largest class of plant R-genes (Shao et al. 2016) and consist of two subgroups, the NBS-LRR-CC and NBS-LRR-TIR (Toll and Interleukin 1 receptor). These genes have a high level of intra-species variation (McHale et al. 2006). R-genes are required to recognise the diverse range of pathogens that might attack the plant. It is therefore not unexpected to find 18 receptor-kinases, containing an LRR domain (Table 5.3, Figure 5.2), amongst the Pinotage genes.

When the R-proteins sense the presence of a pathogen, they can induce a hypersensitive response (HR) that can bring about cell death to impair spreading of the pathogen, or they can launch the systemic acquired resistance (SAR, Figure 5.2). Systemic acquired resistance is a plant-wide response that occurs following localized exposure to a pathogen, and is established by the production of salicylic acid (Jones and Dangl 2006; Sels et al. 2008; Bari and Jones 2009; Wanderley-Nogueira et al. 2012). Salicylic acid activates the Whirly transcription factors, which in turn activate the pathogenesis-related genes (PR-genes) to produce PR-proteins (Desveaux et al. 2004).

During normal growth conditions, *PR*-genes are expressed at a basal level to generate a long-lasting, broad-spectrum disease resistance, but expression is rapidly increased by the presence of biotic or abiotic stress (Wanderley-Nogueira et al. 2012). PR-proteins are typically acidic, resistant to enzymatic degradation and have a low molecular mass (Ali et al. 2010). The majority of proteins present in the berry skin cell apoplast are stress-related proteins (Delaunois et al. 2013). In grapevine, the PR-protein classes present in berry skins differ between cultivars, even in the absence of pathogen pressures (Ghan et al. 2015). In this study, a large number (94) of the Pinotage genes were classified as PR-genes (Table 5.3, Figure 5.2).

5.3.5 The role of transcription factors in stress response

The stress-induced phytohormones, from both the basal and R-mediated responses, activate the expression of transcription factors (TFs, Figure 5.2) (Feller et al. 2011). Transcription factors are classified into the “regulation of transcription” MapMan bin. Table 5.4 provides the sub-classification of this bin and the number of Pinotage genes found in each transcription factor category.

Table 5.4: Mercator sub-classification of the 70 Pinotage genes in the "RNA Regulation of Transcription" MapMan bin (See Table 5.3). Transcription factors are classified into this MapMan bin. Genes marked with * appear in Figure 5.2.

Transcription factor family	Number of Pinotage genes
Ethylene-responsive element binding factors (ERF) *	2
Basic Helix-Loop-Helix (bHLH) / MYC	3
Basic domain-leucine zipper (bZip)*	3
C2H2 zinc finger	4
C3H zinc finger	1
CCAAT box binding factor	2
Homeobox	3
MYB domain*	5
MYB-related	5
WRKY domain*	1
Chromatin Remodeling Factors*	3
Histone acetyltransferases*	3
Unclassified and putative	17
Other	18

Five plant TF families have been shown to participate in the regulation of pathogen defence response: basic domain-leucine zipper (bZip), ethylene-responsive element binding factors (ERF), MYB, WRKY and Whirly (Rushton and Somssich 1998; Singh et al. 2002; Desveaux et al. 2005; Eulgem 2005). Eleven Pinotage genes classified as TFs involved in stress response were identified in this study (Table 5.4, Figure 5.2). Ethylene interacts primarily with ERF, while salicylic acid interacts with Whirly TFs to activate SAR (Figure 5.2). Together with transcription factors, a number of other factors exist that will bring about changes to DNA, which will influence the expression of genes. These changes includes posttranscriptional and posttranslational modification, DNA methylation and chromatin modifications (Dowen et al. 2012; Guerra et al. 2015; Probst and Mittelsten Scheid 2015; Asensi-Fabado et al. 2017). Three chromatin-remodelling factors and three histone acetyltransferases (Table 5.4, Figure 5.2), involved in chromatin modifications, were identified in this study.

5.3.6 Genes modulated by transcription factors

When expressed, the TFs employ their sequence-specific DNA binding ability to modulate gene expression, either promoting or repressing, of genes involved in defence against the stressor

(Feller et al. 2011), such as secondary metabolism, cell wall reinforcement, protein degradation, and PR-proteins (Figure 5.2).

Secondary metabolites play an important role in defending the plant against herbivores, pathogens and other abiotic stresses. They are also important in reproduction to make the flowers and fruit attractive to pollinators and seed dispersers (Pichersky and Gang 2000). Thirteen Pinotage-specific genes involved in secondary metabolism (Table 5.3, Figure 5.2) were identified in this study. Figure 5.3 presents an overview of primary metabolism in plants and how the primary metabolites feed into the secondary metabolism network, highlighting the three main classes of secondary metabolites; the alkaloids (nitrogen containing), the phenolics, and the terpenoids (Bourgaud et al. 2001).

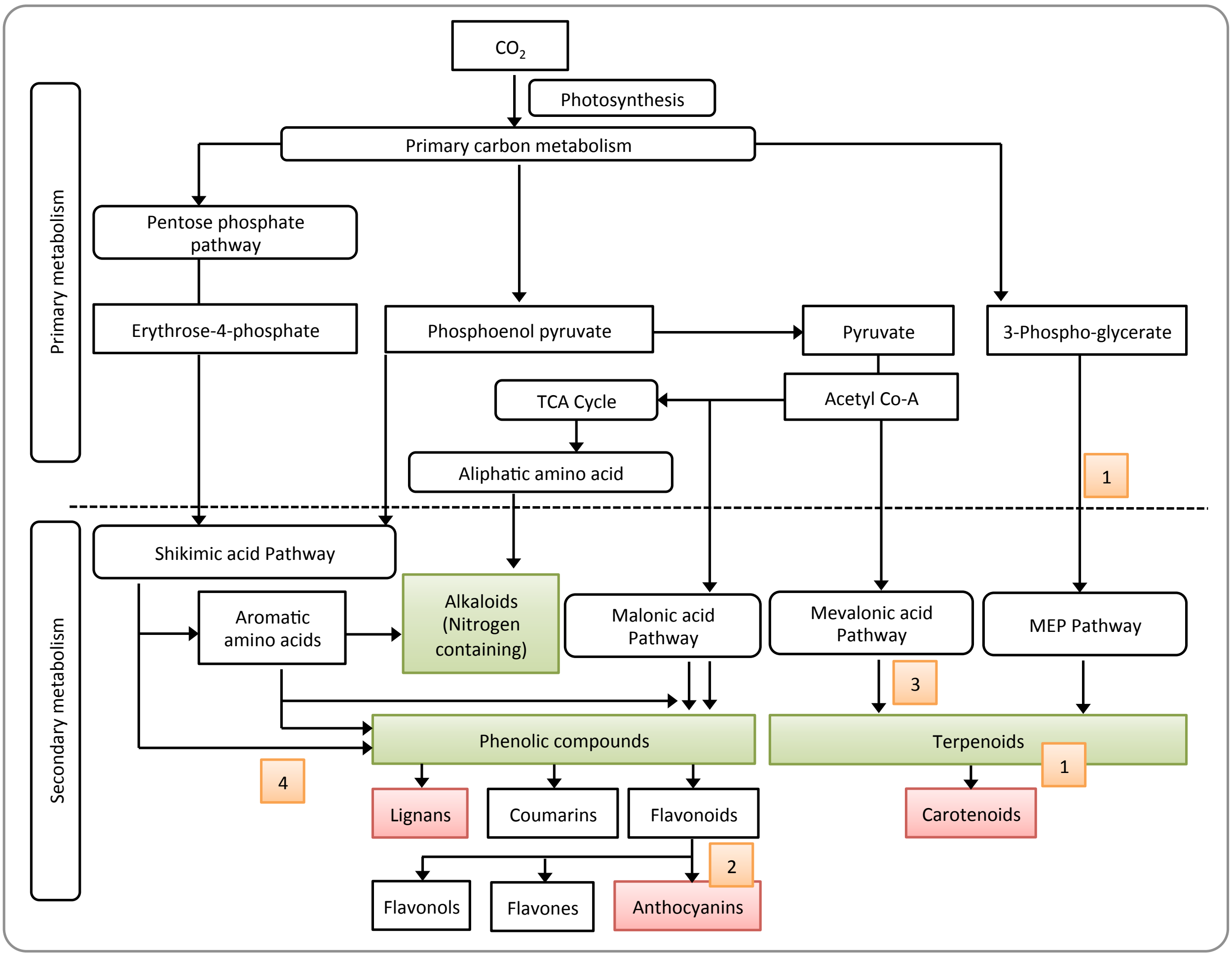


Figure 5.3: *Primary and secondary metabolic network in Vitis vinifera. The number of Pinotage genes assigned to each metabolic step, is indicated in orange blocks (corresponds to genes marked with § in Table 5.3). The three main classes of secondary metabolites, alkaloids, phenolics and terpenoids, are indicated in green; subclasses with Pinotage genes involved in their formation, in red.*

The phenylpropanoid pathway (pathway leading to the production of polyphenolics) includes the formation of lignans, coumarins and flavonoids. Four Pinotage genes involved in the formation of lignans, an important structural component of cell walls, and 2 involved in the conversion of flavonoids to anthocyanins, were identified (Figure 5.3). Anthocyanins are the purple, blue and red pigments and their accumulation in the berries is responsible for the colour in red and black cultivars (Treutter 2006).

Furthermore, five Pinotage genes involved in the metabolic steps leading to the formation of carotenoids, were identified (Figure 5.3). Carotenoids, a subclass of terpenoids, are essential pigments in photosynthetic organisms and their major function is the protection of the photosynthetic membranes. The accumulation of anthocyanin and carotenoid pigments in grape berry skin is an important parameter of berry quality, as this pigment is transferred to the wine during maceration and will have an impact on the wine colour (Young et al. 2012).

Another group of genes regulated by transcription factors, are those involved in cell wall biosynthesis. In reaction to a biotic attack, the plant will strengthen its first line of defence, the wax layer, cuticle and cell walls (Zipfel 2008). Seventeen genes were classified as being associated with the cell wall, including cell wall modification, synthesis of precursors, proteins and lignin biosynthesis (Table 5.3, Figure 5.2).

Fifty putative Pinotage genes involved in protein degradation were identified (Table 5.3, Figure 5.2), 16 containing a ubiquitin E3-ring domain and 15 the E3-SCF-box domain. The E3 ubiquitinases are pathogen-responsive genes and play a central role in modulating signalling pathways. The E3 ubiquitin ligase enzyme tag proteins for degradation by ligating them to a ubiquitin. Plant signalling pathways, including stress responses, are controlled by feedback loops. Ubiquitination and protein degradation provide a negative feedback loop by regulating the levels of R-proteins and other proteins involved in signalling or transcription. Protein degradation might also be involved in the removal of negative regulators/repressors of plant defence responses (Martin et al. 2003; Dreher and Callis 2007).

5.3.7 Abiotic stress

Heat-shock proteins play the role of chaperones to proteins, responsible for folding, correct assembly, translocation and degradation. These processes occur during normal cellular metabolism, but are especially critical when the plant experiences stress. Abiotic stresses cause proteins to fold incorrectly and to become dysfunctional. The role of heat-shock proteins to re-establish the correct protein conformation and facilitate the degradation of miss-folded proteins during stress, is of crucial importance for plant survival (Wang et al. 2004). Eight genes were classified as “abiotic stress” and five as “heat-shock proteins” (Table 5.3, Figure 5.2).

5.3.8 “Stress response” as a major gene category conferring inter-species variety

As discussed in chapter 4, the mature grape berry, ready for harvesting, is the most important part of the plant from an agronomic perspective. In grapevine, as in other plants, leaves are the major sites of carbon assimilation and sugar production. The sugars and sugar-derivatives are transported to the storage organs, in the case of grapevine, the berries. Therefore, Pinotage berries and leaves at harvest were included in this study.

Five technologies, namely microarray and RNA-seq (transcript abundance), nano-liquid chromatography-mass spectroscopy (proteins) and gas chromatography-mass spectroscopy (metabolites), used to study the mature grapevine berry skin, showed that these technologies are concordant in differentiating the biochemical characteristics of the berries from five grapevine cultivars (Ghan et al. 2015). Therefore, RNA-seq is a feasible method to analyse varietal diversity, and augmented with genomic data as in this study, makes for a powerful technique to study the genetic differences between cultivars. A *de novo* transcriptome assembly was performed and compared to the Pinotage and Pinot noir genome data, to gain insight as to how these cultivars differ. A total of 131 genes were classified in the “stress” MapMan bin, while another 132 genes in other bins are directly or indirectly involved in the stress response network.

Berry skin analysis showed that grapevine cultivars differ in terms of PR-proteins present (Ghan et al. 2015), and different cultivars can have distinct defence strategies against pathogens (Amrine et al. 2015). Sultanina genes for example were, other than hypothetical and transposon-related genes, mostly classified as related to disease resistance/defence response (Di Genova et al. 2014).

Regulation, expression and the number of genes involved in secondary metabolism, specifically phenylpropanoid and amino acid metabolism, shows inter-cultivar variation in grapevine (Ghan et al. 2015). Genome and RNA sequencing showed that the expansion of gene families involved in polyphenol biosynthesis confers the high polyphenolic content of Tannat wine (Da Silva et al.

2013). The differential regulation of this pathway is an important difference between Cabernet Sauvignon and Shiraz (Degu et al. 2014).

The same was observed in other plant species, two genotypes of *Eucalyptus* differ mainly in their stress signal transduction pathways (Villar et al. 2011) and a major group of genes that have structural variations between cultivated soybean (*Glycine max*) and its wild relatives (*Glycine soja*), are related to abiotic and biotic stress (Li et al. 2014). Plants produce a wide array of secondary metabolites (Pichersky and Gang 2000), which from an evolutionary perspective is important in order to cope with the ever-evolving pathogens, herbivores, pollinators and seed dispersers.

Overall, it is evident that genes coding for products involved in stress responses, both biotic and abiotic, are one of the key factors conveying varietal diversity between cultivars. The stress response network is a large and integrate network with which many proteins, transcription factors and other gene products are associated. Genes involved in stress responses might be among many others that confer varietal diversity. It is also important to note that the genes discussed in this and other studies are, besides stress response, also involved in a diverse range of other physiological processes during the life cycle of the plant, such as growth and development, flowering, fruit ripening and senescence.

5.4 Conclusion

This study demonstrated the feasibility of mRNA sequencing for the analysis of varietal diversity between a local grapevine cultivar Pinotage, and Pinot noir. The *de novo* transcriptome assembly approach allowed for 24,527 transcripts of the Pinotage genome to be constructed. A potential 988 Pinotage-specific genes, present in neither the Pinot noir ENTAV115 nor PN40024 genome sequences, were identified. Although the most genetic differences between Pinotage and Pinot noir would have been derived from the Cinsaut ancestry, it is possibly that the Pinotage genome may have indeed evolved unique genes, true Pinotage varietal genes. Sequencing of the Cinsaut genome will confirm the sources of Pinotage/Pinot noir genetic variation, and enable the identification of Pinotage varietal genes.

The results from this study agree with other studies in that the genes in the stress response network are an important gene-class conferring intra-species variation. Since plants are in a continuous struggle for survival against biotic and abiotic stressors, it is not surprising that genes in the stress response network are among those that evolve most rapidly and differ most between plants within a species. These genes include those encoding enzymes responsible for secondary metabolism. The accumulation of secondary metabolites, polyphenols and other volatiles are largely responsible for berry quality, especially in wine grapes. These novel stress-induced transcripts identified in Pinotage can serve as a valuable resource to explore candidate genes for enhanced stress tolerance in grapevine. These newly identified transcripts will also help pave the way for a more accurate and complete grapevine genome annotation.

Different wine cultivars have different characteristics they impart on the wines made from them. Wine consumption, and the appreciation of different cultivars has become imbedded as part of human culture. As new molecular research techniques become available, it is essential to study grapevine genetics. The assembled Pinotage transcriptome presented in this study provides further insight into the molecular mechanisms underlying cultivar variation and will assist in future breeding to preserve the unique Pinotage character.

Supplementary data

Supplementary data 5.1: Sequence data for the 988 Pinotage genes.

5.5 References

Ali K, Maltese F, Choi YH and Verpoorte R (2010) Metabolic constituents of grapevine and grape-derived products. Phytochem Rev 9:357–378. doi: 10.1007/s11101-009-9158-0

Altschul SF, Gish W, Miller W, Myers EW and Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215:403–410. doi: 10.1016/S0022-2836(05)80360-2

Amrine KCH, Blanco-Ulate B, Riaz S, Pap D, Jones L, Figueroa-Balderas R, Walker MA and Cantu D (2015) Comparative transcriptomics of Central Asian Vitis vinifera accessions reveals distinct defense strategies against powdery mildew. Hortic Res 2:15037. doi: 10.1038/hortres.2015.37

Asensi-Fabado M-A, Amtmann A and Perrella G (2017) Plant responses to abiotic stress: The chromatin context of transcriptional regulation. Biochim Biophys Acta BBA - Gene Regul Mech 1860:106–122. doi: 10.1016/j.bbagrm.2016.07.015

Bari R and Jones JDG (2009) Role of plant hormones in plant defence responses. Plant Mol Biol 69:473–488. doi: 10.1007/s11103-008-9435-0

Ben Rejeb I, Pastor V and Mauch-Mani B (2014) Plant responses to simultaneous biotic and abiotic stress: molecular mechanisms. Plants 3:458–475. doi: 10.3390/plants3040458

Bourgaud F, Gravot A, Milesi S and Gontier E (2001) Production of plant secondary metabolites: a historical perspective. Plant Sci 161:839–851. doi: 10.1016/S0168-9452(01)00490-3

Çakir B and Kılıçkaya O (2015) Mitogen-activated protein kinase cascades in Vitis vinifera. Front Plant Sci. doi: 10.3389/fpls.2015.00556

Cardone MF, D'Addabbo P, Alkan C, Bergamini C, Catacchio CR, Anaclerio F, Chiatante G, Marra A, Giannuzzi G, Perniola R et al. (2016) Inter-varietal structural variation in grapevine genomes. Plant J 88:648–661. doi: 10.1111/tpj.13274

Casacuberta JM, Jackson S, Panaud O, Purugganan M and Wendel J (2016) Evolution of Plant Phenotypes, from Genomes to Traits. G3 Genes Genomes Genet 6:775–778. doi: 10.1534/g3.115.025502

Conde C, Silva P, Fontes N, Dias ACP, Tavares RM, Sousa MJ, Agasse A, Delrot S and Gerós H (2007) Biochemical changes throughout grape berry development and fruit and wine quality. Food Glob Sci Books 1–22. doi: 10.1.1.549.9659

Degu A, Hochberg U, Sikron N, Venturini L, Buson G, Ghan R, Plaschkes I, Batushansky A, Chalifa-Caspi V, Mattivi F et al. (2014) Metabolite and transcript profiling of berry skin during fruit development elucidates differential regulation between Cabernet Sauvignon and Shiraz cultivars at branching points in the polyphenol pathway. BMC Plant Biol 14:188. doi: 10.1186/s12870-014-0188-4

Delaunois B, Colby T, Belloy N, Conreux A, Harzen A, Baillieul F, Clément C, Schmidt J, Jeandet P and Cordelier S (2013) Large-scale proteomic analysis of the grapevine leaf apoplastic fluid reveals mainly stress-related proteins and cell wall modifying enzymes. BMC Plant Biol 13:24. doi: 10.1186/1471-2229-13-24

Desveaux D, Maréchal A and Brisson N (2005) Whirly transcription factors: defense gene regulation and beyond. Trends Plant Sci 10:95–102. doi: 10.1016/j.tplants.2004.12.008

Desveaux D, Subramaniam R, Després C, Mess J-N, Lévesque C, Fobert PR, Dangl JL and Brisson N (2004) A "Whirly" transcription factor is required for salicylic acid-dependent disease resistance in Arabidopsis. Dev Cell 6:229–240.

DeYoung BJ and Innes RW (2006) Plant NBS-LRR proteins in pathogen sensing and host defense. Nat Immunol 7:1243–1249. doi: 10.1038/ni1410

Di Genova A, Almeida AM, Muñoz-Espinoza C, Vizoso P, Travisany D, Moraga C, Pinto M, Hinrichsen P, Orellana A and Maass A (2014) Whole genome comparison between table and wine grapes reveals a comprehensive catalog of structural variants. BMC Plant Biol 14:7. doi: 10.1186/1471-2229-14-7

Dowen RH, Pelizzola M, Schmitz RJ, Lister R, Dowen JM, Nery JR, Dixon JE and Ecker JR (2012) Widespread dynamic DNA methylation in response to biotic stress. Proc Natl Acad Sci 109:E2183–E2191. doi: 10.1073/pnas.1209329109

Dreher K and Callis J (2007) Ubiquitin, hormones and biotic stress in Plants. Ann Bot 99:787–822. doi: 10.1093/aob/mcl255

Eulgem T (2005) Regulation of the Arabidopsis defense transcriptome. Trends Plant Sci 10:71–78. doi: 10.1016/j.tplants.2004.12.006

Feller A, Machemer K, Braun EL and Grotewold E (2011) Evolutionary and comparative analysis of MYB and bHLH plant transcription factors. Plant J 66:94–116. doi: 10.1111/j.1365-313X.2010.04459.x

Gaspero GD and Cipriani G (2003) Nucleotide binding site/leucine-rich repeats, Pto-like and receptor-like kinases related to disease resistance in grapevine. Mol Genet Genomics 269:612–623. doi: 10.1007/s00438-003-0884-5

Ghan R, Van Sluyter SC, Hochberg U, Degu A, Hopper DW, Tillet RL, Schlauch KA, Haynes PA, Fait A and Cramer GR (2015) Five "omic" technologies are concordant in differentiating the biochemical characteristics of the berries of five grapevine (Vitis vinifera L.) cultivars. BMC Genomics 16:946. doi: 10.1186/s12864-015-2115-y

Gilbert D (2013) Gene-omes built from mRNA seq not genome DNA.

Gill US, Lee S and Mysore KS (2015) Host versus nonhost resistance: distinct wars with similar arsenals. Phytopathology 105:580–587. doi: 10.1094/PHYTO-11-14-0298-RVW

Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q et al. (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol 29:644–652. doi: 10.1038/nbt.1883

Guerra D, Crosatti C, Khoshro HH, Mastrangelo AM, Mica E and Mazzucotelli E (2015) Post-transcriptional and post-translational regulations of drought and heat response in plants: a spider's web of mechanisms. Front Plant Sci. doi: 10.3389/fpls.2015.00057

Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, Couger MB, Eccles D, Li B, Lieber M et al. (2013) De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nat Protoc 8:1494–1512. doi: 10.1038/nprot.2013.084

Jaillon O, Aury J-M, Noel B, Policriti A, Clepet C, Casagrande A, Choisne N, Aubourg S, Vitulo N, Jubin C et al. (2007) The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. Nature 449:463–467. doi: 10.1038/nature06148

Jones JDG and Dangl JL (2006) The plant immune system. Nature 444:323–329. doi: 10.1038/nature05286

Lamb C and Dixon RA (1997) The oxidative burst in plant disease resistance. Annu Rev Plant Physiol Plant Mol Biol 48:251–275. doi: 10.1146/annurev.arplant.48.1.251

Li Y, Zhou G, Ma J, Jiang W, Jin L, Zhang Z, Guo Y, Zhang J, Sui Y, Zheng L et al. (2014) De novo assembly of soybean wild relatives for pan-genome analysis of diversity and agronomic traits. Nat Biotechnol 32:1045–1052. doi: 10.1038/nbt.2979

Liu J, Liu X, Dai L and Wang G (2007) Recent progress in elucidating the structure, function and evolution of disease resistance genes in plants. J Genet Genomics 34:765–776. doi: 10.1016/S1673-8527(07)60087-3

Louw L, Roux K, Tredoux A, Tomic O, Naes T, Nieuwoudt HH and van Rensburg P (2009) Characterization of selected South African young cultivar wines using FTMIR spectroscopy, gas chromatography, and multivariate data analysis. J Agric Food Chem 57:2623–2632. doi: 10.1021/jf8037456

Marroni F, Pinosio S and Morgante M (2014) Structural variation and genome complexity: is dispensable really dispensable? Curr Opin Plant Biol 18:31–36. doi: 10.1016/j.pbi.2014.01.003

Martin GB, Bogdanove AJ and Sessa G (2003) Understanding the functions of plant disease resistance proteins. Annu Rev Plant Biol 54:23–61. doi: 10.1146/annurev.arplant.54.031902.135035

May P, Wienkoop S, Kempa S, Usadel B, Christian N, Rupprecht J, Weiss J, Recuenco-Munoz L, Ebenhöh O, Weckwerth W et al. (2008) Metabolomics- and proteomics-assisted genome annotation and analysis of the draft metabolic network of Chlamydomonas reinhardtii. Genetics 179:157–166. doi: 10.1534/genetics.108.088336

McHale L, Tan X, Koehl P and Michelmore RW (2006) Plant NBS-LRR proteins: adaptable guards. Genome Biol 7:212. doi: 10.1186/gb-2006-7-4-212

Morgante M (2006) Plant genome organisation and diversity: the year of the junk! Curr Opin Biotechnol 17:168–173. doi: 10.1016/j.copbio.2006.03.001

Morgante M, Depaoli E and Radovic S (2007) Transposable elements and the plant pan-genomes. Curr Opin Plant Biol 10:149–155. doi: 10.1016/j.pbi.2007.02.001

Orffer PC and Visser C (2009) The origin of grapevine cultivars. In: Wineland Mag. http://www.wineland.co.za/the-origin-of-grapevine-cultivars/. Accessed 13 Jun 2017

Pandey S, Fartyal D, Agarwal A, Shukla T, James D, Kaul T, Negi YK, Arora S and Reddy MK (2017) Abiotic stress tolerance in plants: myriad roles of ascorbate peroxidase. Front Plant Sci. doi: 10.3389/fpls.2017.00581

Péros J-P, Launay A, Berger G, Lacombe T and This P (2015) MybA1 gene diversity across the Vitis genus. Genetica 143:373–384. doi: 10.1007/s10709-015-9836-3

Pichersky E and Gang DR (2000) Genetics and biochemistry of secondary metabolites in plants: an evolutionary perspective. Trends Plant Sci 5:439–445. doi: 10.1016/S1360-1385(00)01741-6

Probst AV and Mittelsten Scheid O (2015) Stress-induced structural changes in plant chromatin. Curr Opin Plant Biol 27:8–16. doi: 10.1016/j.pbi.2015.05.011

Rodriguez MCS, Petersen M and Mundy J (2010) Mitogen-Activated Protein Kinase signaling in plants. Annu Rev Plant Biol 61:621–649. doi: 10.1146/annurev-arplant-042809-112252

Rushton PJ and Somssich IE (1998) Transcriptional control of plant genes responsive to pathogens. Curr Opin Plant Biol 1:311–315.

Sels J, Mathys J, De Coninck BMA, Cammue BPA and De Bolle MFC (2008) Plant pathogenesis-related (PR) proteins: A focus on PR peptides. Plant Physiol Biochem 46:941–950. doi: 10.1016/j.plaphy.2008.06.011

Sewelam N, Kazan K and Schenk PM (2016) Global plant stress signaling: reactive oxygen species at the cross-road. Front Plant Sci. doi: 10.3389/fpls.2016.00187

Shao Z-Q, Xue J-Y, Wu P, Zhang Y-M, Wu Y, Hang Y-Y, Wang B and Chen J-Q (2016) Large-scale analyses of angiosperm nucleotide-binding site-leucine-rich repeat (NBS-LRR) genes reveal three anciently diverged classes with distinct evolutionary patterns. Plant Physiol pp.01487.2015. doi: 10.1104/pp.15.01487

Singh KB, Foley RC and Oñate-Sánchez L (2002) Transcription factors in plant defense and stress responses. Curr Opin Plant Biol 5:430–436. doi: 10.1016/S1369-5266(02)00289-3

Stael S, Wurzinger B, Mair A, Mehlmer N, Vothknecht UC and Teige M (2012) Plant organellar calcium signalling: an emerging field. J Exp Bot 63:1525–1542. doi: 10.1093/jxb/err394

Thimm O, Bläsing O, Gibon Y, Nagel A, Meyer S, Krüger P, Selbig J, Müller LA, Rhee SY and Stitt M (2004) MapMan: a user-driven tool to display genomics data sets onto diagrams of metabolic pathways and other biological processes. Plant J 37:914–939. doi: 10.1111/j.1365-313X.2004.02016.x

Treutter D (2006) Significance of flavonoids in plant resistance: a review. Environ Chem Lett 4:147. doi: 10.1007/s10311-006-0068-8

Usadel B, Poree F, Nagel A, Lohse M, Czedik-Eysenberg A and Stitt M (2009) A guide to using MapMan to visualize and compare Omics data in plants: a case study in the crop species, Maize. Plant Cell Environ 32:1211–1229. doi: 10.1111/j.1365-3040.2009.01978.x

Van Wyk CJ, Augustyn OPH, Wet P de and Joubert WA (1979) Isoamyl acetate — a key fermentation volatile of wines of Vitis Vinifera cv Pinotage. Am J Enol Vitic 30:167–173.

Velasco R, Zharkikh A, Troggio M, Cartwright DA, Cestaro A, Pruss D, Pindo M, FitzGerald LM, Vezzulli S, Reid J et al. (2007) A high quality draft consensus sequence of the genome of a heterozygous grapevine variety. PLOS ONE 2:e1326. doi: 10.1371/journal.pone.0001326

Venturini L, Ferrarini A, Zenoni S, Tornielli GB, Fasoli M, Santo SD, Minio A, Buson G, Tononi P, Zago ED et al. (2013) De novo transcriptome characterization of Vitis vinifera cv. Corvina unveils varietal diversity. BMC Genomics 14:41. doi: 10.1186/1471-2164-14-41

Vernikos G, Medini D, Riley DR and Tettelin H (2015) Ten years of pan-genome analyses. Curr Opin Microbiol 23:148–154. doi: 10.1016/j.mib.2014.11.016

Villar E, Klopp C, Noirot C, Novaes E, Kirst M, Plomion C and Gion J-M (2011) RNA-Seq reveals genotype-specific molecular responses to water deficit in eucalyptus. BMC Genomics 12:538. doi: 10.1186/1471-2164-12-538

Vivier M and Pretorius I (2000) Genetic improvement of grapevine: tailoring grape varieties for the third millennium - a review. South Afr J Enol Vitic 21:5–26.

Wanderley-Nogueira AC, Belarmino LC, Soares-Cavalcanti N da M, Bezerra-Neto JP, Kido EA, Pandolfi V, Abdelnoor RV, Binneck E, Carazzole MF and Benko-Iseppon AM (2012) An overall evaluation of the Resistance (R) and Pathogenesis-Related (PR) superfamilies in soybean, as compared with Medicago and Arabidopsis. Genet Mol Biol 35:260–271. doi: 10.1590/S1415-47572012000200007

Wang G, Lovato A, Liang Y h., Wang M, Chen F, Tornielli G b., Polverari A, Pezzotti M and Cheng Z m. (2014a) Validation by isolation and expression analyses of the mitogen-activated protein kinase gene family in the grapevine (Vitis vinifera L.). Aust J Grape Wine Res 20:255–262. doi: 10.1111/ajgw.12081

Wang G, Lovato A, Polverari A, Wang M, Liang Y-H, Ma Y-C and Cheng Z-M (2014b) Genome-wide identification and analysis of mitogen activated protein kinase kinase kinase gene family in grapevine (Vitis vinifera). BMC Plant Biol 14:219. doi: 10.1186/s12870-014-0219-1

Wang W, Vinocur B, Shoseyov O and Altman A (2004) Role of plant heat-shock proteins and molecular chaperones in the abiotic stress response. Trends Plant Sci 9:244–252. doi: 10.1016/j.tplants.2004.03.006

Weldegergis BT, De Villiers A and Crouch AM (2011a) Chemometric investigation of the volatile content of young South African wines. Food Chem 128:1100–1109. doi: 10.1016/j.foodchem.2010.09.100

Weldegergis BT, De Villiers A, McNeish C, Seethapathy S, Mostafa A, Górecki T and Crouch AM (2011b) Characterisation of volatile components of Pinotage wines using comprehensive two-dimensional gas chromatography coupled to time-of-flight mass spectrometry (GC × GC–TOFMS). Food Chem 129:188–199. doi: 10.1016/j.foodchem.2010.11.157

Wu TD and Watanabe CK (2005) GMAP: a genomic mapping and alignment program for mRNA and EST sequences. Bioinforma Oxf Engl 21:1859–1875. doi: 10.1093/bioinformatics/bti310

Young PR, Lashbrooke JG, Alexandersson E, Jacobson D, Moser C, Velasco R and Vivier MA (2012) The genes and enzymes of the carotenoid metabolic pathway in Vitis vinifera L. BMC Genomics 13:243. doi: 10.1186/1471-2164-13-243

Zipfel C (2008) Pattern-recognition receptors in plant innate immunity. Curr Opin Immunol 20:10–16. doi: 10.1016/j.coi.2007.11.003

Online resources

CRIBI: http://genomes.cribi.unipd.it/grape/

FastQC: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/

Illumina mRNA stranded library preparation protocol: https://support.illumina.com/content/dam/illumina-support/documents/documentation/chemistry_documentation/samplepreps_truseq/truseqstrandedmrna/truseq-stranded-mrna-sample-prep-guide-15031047-e.pdf

MapMan: http://mapman.gabipd.org/

NCBI: https://blast.ncbi.nlm.nih.gov

Pinotage Association: http://pinotage.co.za/recognition/

Chapter 6: Conclusion

Grapevine (*Vitis vinifera*) is one of the most widely grown fruit crops in the world. Today, the choice of grapevine cultivar is an important viticultural decision, due to strong consumer preferences for particular wine or table grapes. Wine production is the largest grapevine-related industry. New technologies such as NGS allow for whole-genome sequencing and are revolutionizing viticulture and winemaking, with a more precise understanding of the underlying genetic makeup of all the organisms involved in winemaking.

Genome re-sequencing is necessary for the description of a species' diversity and pan-genome. Although a grapevine reference genome is available, a single reference genome is not sufficient to represent the genetic diversity within a species, especially one that has undergone significant domestication over several millennia. Since intra-specific phenotypic diversity cannot be explained by genomic variation alone, analysis of differences in gene expression and regulation should also be conducted. In this study, NGS together with bioinformatic analysis were employed to unravel the genome and transcriptome of Pinotage, a *Vitis vinifera* cultivar with special importance in the South African viticultural and wine industry.

This project serves as a pilot study to explore the possibilities for genome sequencing of grapevine in a South African context. The sequencing and *de novo* assembly approaches used to construct the first draft genome sequence of Pinotage is described. A total of 578,522 contigs with an N_{50} of 2,366 were obtained. As a continuation of this project, the draft genome assembly can be further improved by including mate-pair library and/or additional scaffolding data, such as optical mapping data. A follow-on study can also use this genomic NGS data to assemble the Pinotage chloroplast and mitochondrion genomes to study plastid diversity. An in-depth analysis of sequence variation in the promotor/*cis* regulatory elements, together with discovery of splice variants and non-coding and regulatory RNA species in the RNA Ribo-Zero data, would give further insight into cultivar-specific gene expression profiles.

In addition to the genome assembly, the distribution of Pinotage/Pinot noir variants (SNPs and indels) was explored and an average variant density of 1 variant in 106 bp reported. Gene functional clusters influenced by high impact variants were highlighted. In particular, "signalling receptor kinases" are reported as a gene functional cluster influenced by these variants. Signalling receptor kinases play an important role in the stress response network, suggesting that this network is a notable difference between the Pinotage and Pinot noir genomes. This hypothesis was confirmed by the comparison of the Pinotage genome and transcriptome data.

This study also provides the first research into the unique character of old Pinotage vines, by comparing the leaf and berry transcriptomes of young and old vines at harvest. Although the term “old-vine” is increasingly used to denote a wine of high quality, the definition of wine quality and taste and/or flavour is subjective and prone to suggestion and expectation. To date, no scientific evidence exists to explain why better wines can be produced from older vines, to what extent genetic and/or environmental influences contribute to the quality of old-vine wines, or exactly what wine components confer this “old-vine” character.

The vine material was sampled from a commercial Pinotage vineyard where young and old vines are inter-planted. Field sampling allows for all the environmental cues that might influence the “old-vine” character to be integrated in the gene expression profile, while simultaneously excluding the effects a greenhouse study design might have had on the vines’ growth, development and ageing. Additionally, the experimental layout was designed to include an RNA pooling strategy and allowed for three biological repeats per analysis group to limit between-sample variation. Furthermore, ten repetitions of the differential expression analysis were performed, using a high FPKM threshold. This allowed for the identification of 925 high quality genes differentially expressed between young and old vines.

The results indicate that many of the identified differentially expressed genes are involved in metabolic pathways active during fruit ripening. Considering the hormonal control of fruit ripening, and differential expression of these genes, a general trend was observed towards delayed berry ripening in older vines. Berries of these vines also had a lower sugar concentration at harvest compared to young-vine berries. Combined these results would suggest that berries of old vines take longer to ripen, allowing for the accumulation of volatile aromas that influence berry flavour.

A number of follow-up experiments can be performed to complement the *in silico* analyses of the transcriptome data. Firstly, the differentially-expressed gene predictions between young and old vines could be validated with RT-qPCR, and RNA-seq could be performed on berries and leaves from the same vines in successive harvest seasons or at different time-points throughout growing seasons. To complement the transcriptome analyses, metabolite and hormone levels in the vine leaves and berries can be monitored throughout a growing season and the aroma and flavour composition of the wine produced from these vines, analysed. As epigenetic modifications can have a major influence on gene expression, these differences between young and old vines could be analysed by methylome sequencing.

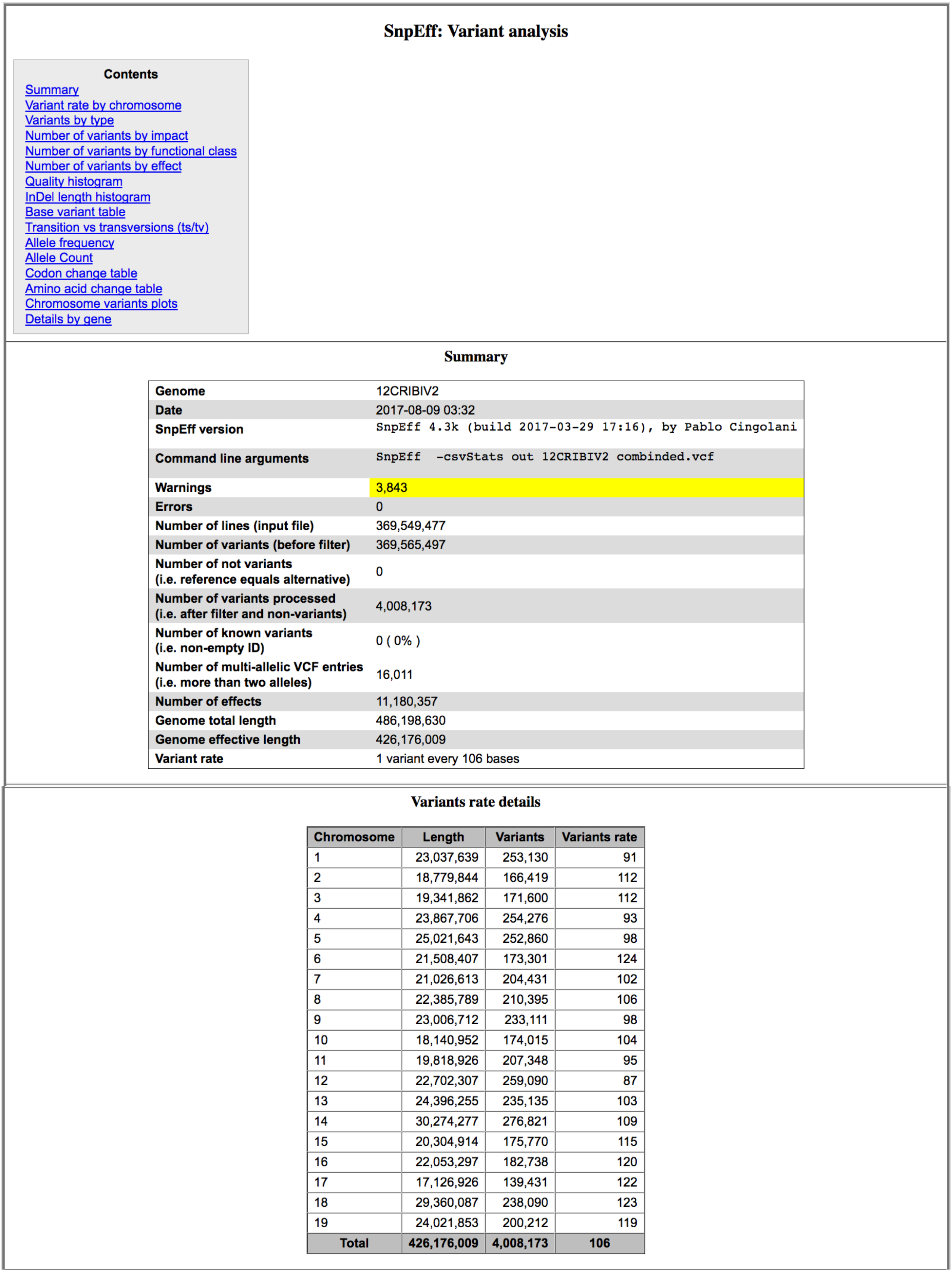
The Pinotage genome data generated were also compared with the transcriptome data. DNA and RNA for these analyses were independently extracted, sequenced and assembled. Transcripts were also not only compared to the highly inbred Pinot noir PN40024 genome, but also the commercial Pinot noir ENTAV115, to ensure the identified transcripts are truly different from Pinot noir. The resulting assemblies were compared to identify 988 genes that are present in the Pinotage genome, but absent from the Pinot noir genome.

A large number of these Pinotage genes were found to be involved in the stress response network. Various pathogenic and environmental stressors constantly challenge the plants' survival and they have developed an array of biochemical and physiological mechanisms to combat these stresses. From an evolutionary point of view, it makes sense to have different stress response networks to reduce the possibility for one stressor to drive the species to extinction. It is therefore reasonable to conclude that differences in the stress response networks can be a major contributor to varietal differences between cultivars. As an additional experiment, the presence of these genes in Pinotage can be validated, and their presence/ absence in Pinot noir and Cinsaut confirmed, to assess the origin of these genomic variations.

The data and knowledge generated in this study will ultimately contribute to the establishment of grapevine as a model system for ripening of non-climacteric fruit and fruit functional genomics, as well as promote the advancement of precision breeding of grapevine for improved traits, such as yield and quality for sustainable production of high-quality wine in a changing environment.

Supplementary Data

Supplementary data 3.1: SNPeff report



SnpEff: Variant analysis

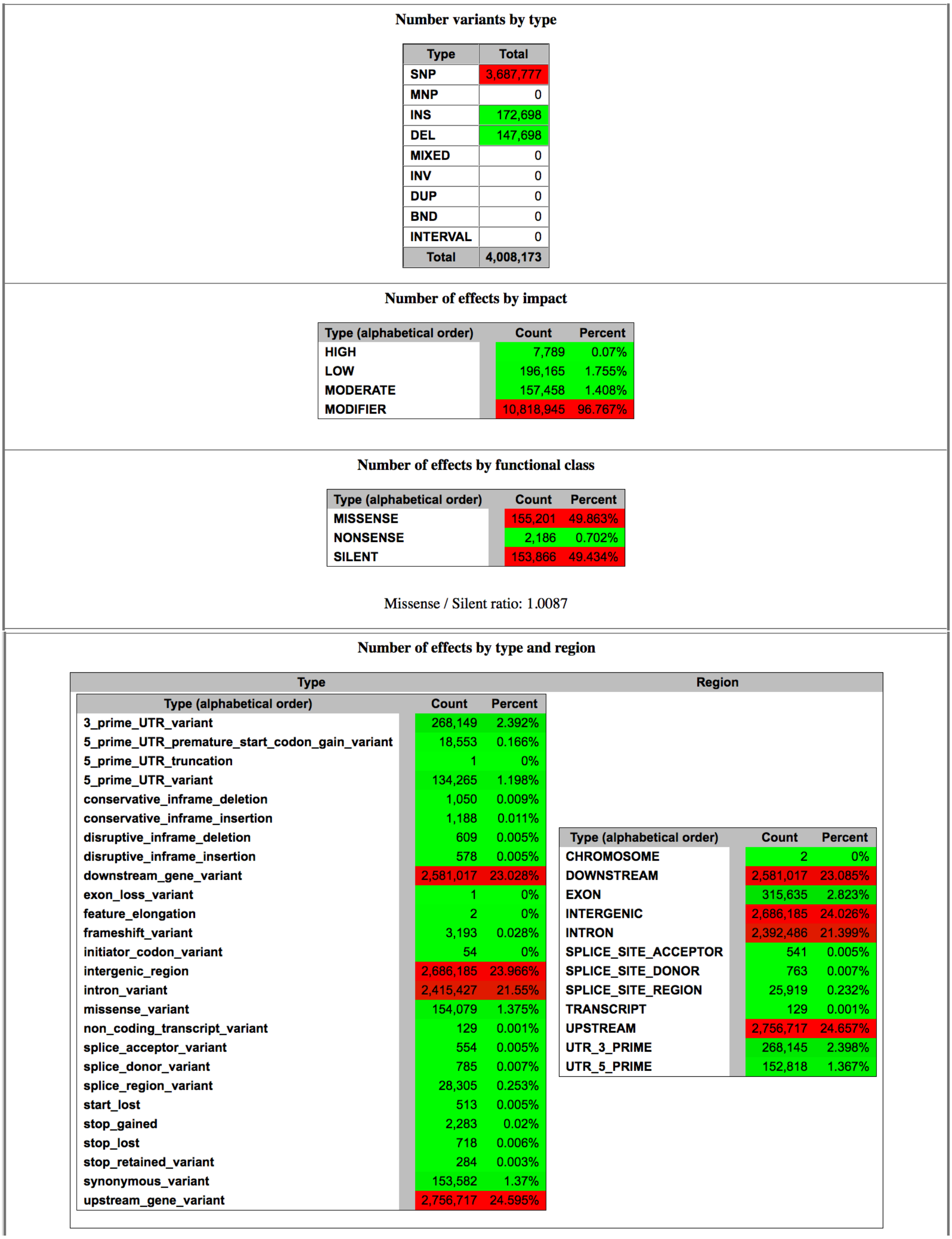
Contents

Summary

Genome	12CRIBIV2
Date	2017-08-09 03:32
SnpEff version	`SnpEff 4.3k (build 2017-03-29 17:16), by Pablo Cingolani`
Command line arguments	`SnpEff  -csvStats out 12CRIBIV2 combinded.vcf`
Warnings	3,843
Errors	0
Number of lines (input file)	369,549,477
Number of variants (before filter)	369,565,497
Number of not variants **(i.e. reference equals alternative)**	0
Number of variants processed **(i.e. after filter and non-variants)**	4,008,173
Number of known variants **(i.e. non-empty ID)**	0 (0%)
Number of multi-allelic VCF entries **(i.e. more than two alleles)**	16,011
Number of effects	11,180,357
Genome total length	486,198,630
Genome effective length	426,176,009
Variant rate	1 variant every 106 bases

Variants rate details

Chromosome	Length	Variants	Variants rate
1	23,037,639	253,130	91
2	18,779,844	166,419	112
3	19,341,862	171,600	112
4	23,867,706	254,276	93
5	25,021,643	252,860	98
6	21,508,407	173,301	124
7	21,026,613	204,431	102
8	22,385,789	210,395	106
9	23,006,712	233,111	98
10	18,140,952	174,015	104
11	19,818,926	207,348	95
12	22,702,307	259,090	87
13	24,396,255	235,135	103
14	30,274,277	276,821	109
15	20,304,914	175,770	115
16	22,053,297	182,738	120
17	17,126,926	139,431	122
18	29,360,087	238,090	123
19	24,021,853	200,212	119
Total	**426,176,009**	**4,008,173**	**106**



Number variants by type

Type	Total
SNP	3,687,777
MNP	0
INS	172,698
DEL	147,698
MIXED	0
INV	0
DUP	0
BND	0
INTERVAL	0
Total	4,008,173

Number of effects by impact

Type (alphabetical order)	Count	Percent
HIGH	7,789	0.07%
LOW	196,165	1.755%
MODERATE	157,458	1.408%
MODIFIER	10,818,945	96.767%

Number of effects by functional class

Type (alphabetical order)	Count	Percent
MISSENSE	155,201	49.863%
NONSENSE	2,186	0.702%
SILENT	153,866	49.434%

Missense / Silent ratio: 1.0087

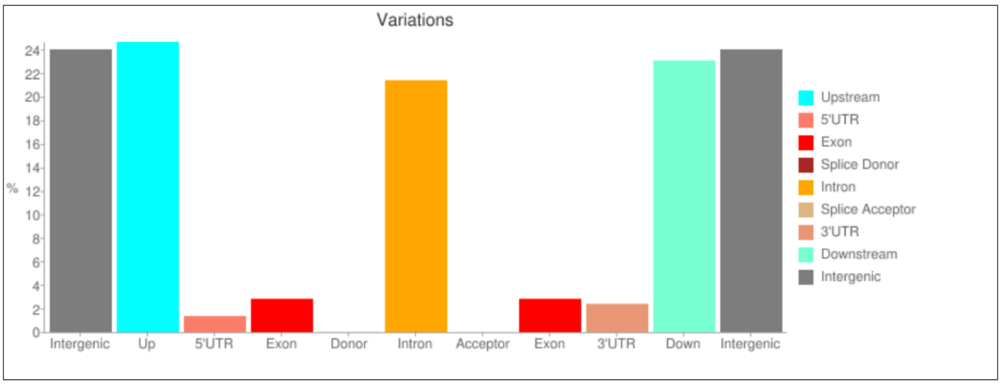
Number of effects by type and region

Type

Type (alphabetical order)	Count	Percent
3_prime_UTR_variant	268,149	2.392%
5_prime_UTR_premature_start_codon_gain_variant	18,553	0.166%
5_prime_UTR_truncation	1	0%
5_prime_UTR_variant	134,265	1.198%
conservative_inframe_deletion	1,050	0.009%
conservative_inframe_insertion	1,188	0.011%
disruptive_inframe_deletion	609	0.005%
disruptive_inframe_insertion	578	0.005%
downstream_gene_variant	2,581,017	23.028%
exon_loss_variant	1	0%
feature_elongation	2	0%
frameshift_variant	3,193	0.028%
initiator_codon_variant	54	0%
intergenic_region	2,686,185	23.966%
intron_variant	2,415,427	21.55%
missense_variant	154,079	1.375%
non_coding_transcript_variant	129	0.001%
splice_acceptor_variant	554	0.005%
splice_donor_variant	785	0.007%
splice_region_variant	28,305	0.253%
start_lost	513	0.005%
stop_gained	2,283	0.02%
stop_lost	718	0.006%
stop_retained_variant	284	0.003%
synonymous_variant	153,582	1.37%
upstream_gene_variant	2,756,717	24.595%

Region

Type (alphabetical order)	Count	Percent
CHROMOSOME	2	0%
DOWNSTREAM	2,581,017	23.085%
EXON	315,635	2.823%
INTERGENIC	2,686,185	24.026%
INTRON	2,392,486	21.399%
SPLICE_SITE_ACCEPTOR	541	0.005%
SPLICE_SITE_DONOR	763	0.007%
SPLICE_SITE_REGION	25,919	0.232%
TRANSCRIPT	129	0.001%
UPSTREAM	2,756,717	24.657%
UTR_3_PRIME	268,145	2.398%
UTR_5_PRIME	152,818	1.367%


Variations
%
0
2
4
6
8
10
12
14
16
18
20
22
24
Intergenic
Up
5'UTR
Exon
Donor
Intron
Acceptor
Exon
3'UTR
Down
Intergenic
Upstream
5'UTR
Exon
Splice Donor
Intron
Splice Acceptor
3'UTR
Downstream
Intergenic

Supplementary data table 3.1: Functional clusters (as predicted for *Vitis vinifera* in PLAZA functional clustering experiment 17) containing high impact variants. Clusters are ordered in the table in terms of % genes containing high impact variants. 225 functional clusters are shown.

Functional cluster	Chromo-some	High impact variants	Genes in cluster	MapMan bin	MapMan bin description
CH_vvi_385	7	6	4	30.2.11	Signalling.receptor kinases.leucine rich repeat XI
CH_vvi_184	7	8	6	26.3	Misc.gluco-, galacto- and mannosidases
CH_vvi_132	7	6	5	26.9	Misc.glutathione S transferases
CH_vvi_349	7	36	34	35.2	Not assigned.unknown
CH_vvi_252	4	4	4	17.1.1.1.10	Hormone metabolism.abscisic acid.synthesis-degradation.synthesis.9-cis-epoxycarotenoid dioxygenase
CH_vvi_182	7	4	4	26.1	Misc.misc2
CH_vvi_301	13	11	15	26.28	Misc.GDSL-motif lipase
CH_vvi_197	10	8	12	30.2.17	Signalling.receptor kinases.DUF26
CH_vvi_273	1	4	6	26.22	Misc.short chain dehydrogenase/reductase (SDR)
CH_vvi_279	17	4	6	27.3.66	RNA.regulation of transcription.Pseudo ARR transcription factor family
CH_vvi_224	7	4	6	33.1	Development.storage proteins
CH_vvi_169	12	2	3	11.3.8	Lipid metabolism.Phospholipid synthesis.phosphatidylserine decarboxylase
CH_vvi_246	17	2	3	5.1	Fermentation.aldehyde dehydrogenase
CH_vvi_422	2	2	3	9.4	Mitochondrial electron transport / ATP synthesis.alternative oxidase
CH_vvi_419	7	2	3	17.6.1.1	Hormone metabolism.gibberellin.synthesis-degradation.copalyl diphosphate synthase
CH_vvi_96	8	5	8	26.4.1	Misc.beta 1,3 glucan hydrolases.glucan endo-1,3-beta-glucosidase
CH_vvi_44	10	9	15	30.2.25	Signalling.receptor kinases.wall associated kinase
CH_vvi_170	12	3	5	34.99	Transport.misc
CH_vvi_112	5	3	5	26.1	Misc.misc2
CH_vvi_149	10	4	7	34.16	Transport.ABC transporters and multidrug resistance systems
CH_vvi_99	7	6	11	20.1.7	Stress.biotic.PR-proteins
CH_vvi_23	13	8	15	29.5.1	Protein.degradation.subtilases
CH_vvi_39	12	5	10	10.2.1	Cell wall.cellulose synthesis.cellulose synthase
CH_vvi_337	13	5	10	31.1	Cell.organisation
CH_vvi_358	14	4	8	30.2.11	Signalling.receptor kinases.leucine rich repeat XI
CH_vvi_212	6	4	8	30.2.11	Signalling.receptor kinases.leucine rich repeat XI
CH_vvi_306	10	2	4	20.1.7	Stress.biotic.PR-proteins
CH_vvi_151	15	2	4	26.6	Misc.O-methyl transferases
CH_vvi_408	15	2	4	26.1	Misc.cytochrome P450
CH_vvi_409	19	2	4	26.1	Misc.cytochrome P450
CH_vvi_287	4	2	4	16.1.5	Secondary metabolism.isoprenoids.terpenoids
CH_vvi_181	7	2	4	17.5.2	Hormone metabolism.ethylene.signal transduction
CH_vvi_377	7	2	4	11.9.2.1	Lipid metabolism.lipid degradation.lipases.triacylglycerol lipase
CH_vvi_310	3	1	2	13.2.4.3	Amino acid metabolism.degradation.branched chain group.valine

Functional cluster	Chromo-some	High impact variants	Genes in cluster	MapMan bin	MapMan bin description
CH_vvi_371	5	1	2	19.1	Tetrapyrrole synthesis.glu-trna synthetase
CH_vvi_367	6	1	2	13.1.7.7	Amino acid metabolism.synthesis.histidine.histidinol-phosphate aminotransferase
CH_vvi_87	10	10	21	26.8	Misc.nitrilases, *nitrile lyases, berberine bridge enzymes, reticuline oxidases, troponine reductases
CH_vvi_119	11	6	13	34.14	Transport.unspecified cations
CH_vvi_83	13	5	11	13.2.6.3	Amino acid metabolism.degradation.aromatic aa.tryptophan
CH_vvi_30	13	33	74	20.1	Stress.biotic
CH_vvi_238	8	4	9	30.2.11	Signalling.receptor kinases.leucine rich repeat XI
CH_vvi_29	19	35	79	26.9	Misc.glutathione S-transferases
CH_vvi_45	7	6	14	26.1	Misc.cytochrome P450
CH_vvi_416	4	3	7	33.99	Development.unspecified
CH_vvi_60	8	10	24	34.99	Transport.misc
CH_vvi_17	2	7	17	16.5.1.3.3	Secondary metabolism.sulfur-containing.glucosinolates.degradation.nitrilase
CH_vvi_7	4	10	25	30.1	Signalling.in sugar and nutrient physiology
CH_vvi_92	3	6	15	26.1	Misc.cytochrome P450
CH_vvi_34	13	4	10	27.3.37	RNA.regulation of transcription.AS2,Lateral Organ Boundaries Gene Family
CH_vvi_348	14	4	10	30.3	Signalling.calcium
CH_vvi_266	13	2	5	10.6.3	Cell wall.degradation.pectate lyases and polygalacturonases
CH_vvi_415	3	2	5	26.16	Misc.myrosinases-lectin-jacalin
CH_vvi_107	5	2	5	10.6.2	Cell wall.degradation.mannan-xylose-arabinose-fucose
CH_vvi_141	5	2	5	10.2	Cell wall.cellulose synthesis
CH_vvi_211	7	2	5	28.1.3	DNA.synthesis/chromatin structure.histone
CH_vvi_343	7	2	5	20.1	Stress.biotic
CH_vvi_265	9	2	5	10.6.3	Cell wall.degradation.pectate lyases and polygalacturonases
CH_vvi_68	1	3	8	3.5	Minor CHO metabolism.others
CH_vvi_203	2	3	8	2.2.2.1	Major CHO metabolism.degradation.starch.starch cleavage
CH_vvi_232	11	11	30	29.2.3	Protein.synthesis.initiation
CH_vvi_155	1	8	22	34.3	Transport.amino acids
CH_vvi_121	12	4	11	30.2.17	Signalling.receptor kinases.DUF 26
CH_vvi_215	11	5	14	30.2.11	Signalling.receptor kinases.leucine rich repeat XI
CH_vvi_8	9	7	21	34.16	Transport.ABC transporters and multidrug resistance systems
CH_vvi_48	3	3	9	16.8.5.1	Secondary metabolism.flavonoids.isoflavones.isoflavone reductase
CH_vvi_236	13	2	6	20.1	Stress.biotic
CH_vvi_102	3	2	6	2.2.2.1	Major CHO metabolism.degradation.starch.starch cleavage
CH_vvi_171	4	2	6	34.3	Transport.amino acids
CH_vvi_198	4	2	6	34.99	Transport.misc

Functional cluster	Chromo-some	High impact variants	Genes in cluster	MapMan bin	MapMan bin description
CH_vvi_106	9	2	6	13.1.6.4.1	Amino acid metabolism.synthesis.aromatic aa.tyrosine.arogenate dehydrogenase & prephenate dehydrogenase
CH_vvi_293	9	2	6	26.1	Misc.cytochrome P450
CH_vvi_331	1	1	3	21.2	Redox.ascorbate and glutathione
CH_vvi_318	11	1	3	16.4.1	Secondary metabolism.N misc.alkaloid-like
CH_vvi_267	12	1	3	23.3.3	Nucleotide metabolism.salvage.NUDIX hydrolases
CH_vvi_334	13	1	3	13.1.6	Amino acid metabolism.synthesis.aromatic aa
CH_vvi_342	13	1	3	28.1.3	DNA.synthesis/chromatin structure.histone
CH_vvi_158	14	1	3	13.1.6.1.10	Amino acid metabolism.synthesis.aromatic aa.chorismate.dehydroquinate/shikimate dehydrogenase
CH_vvi_437	14	1	3	20.2.99	Stress.abiotic.unspecified
CH_vvi_297	16	1	3	26.11	Misc.alcohol dehydrogenases
CH_vvi_307	16	1	3	11.9.2.1	Lipid metabolism.lipid degradation.lipases.triacylglycerol lipase
CH_vvi_391	16	1	3	29.5.9	Protein.degradation.AAA type
CH_vvi_353	18	1	3	16.8.3.1	Secondary metabolism.flavonoids.dihydroflavonols.dihydroflavonol 4-reductase
CH_vvi_157	19	1	3	8.2.99	TCA / org transformation.other organic acid transformatons.misc
CH_vvi_442	3	1	3	29.5.1	Protein.degradation.subtilases
CH_vvi_283	5	1	3	26.24	Misc.GCN5-related N-acetyltransferase
CH_vvi_240	6	1	3	16.5.1.3.3	Secondary metabolism.sulfur-containing.glucosinolates.degradation.nitrilase
CH_vvi_269	8	1	3	33.2	Development.late embryogenesis abundant
CH_vvi_388	8	1	3	26.22	Misc.short chain dehydrogenase/reductase (SDR)
CH_vvi_243	13	6	20	26.2	Misc.UDP glucosyl and glucoronyl transferases
CH_vvi_38	18	3	10	16.5.99.1	Secondary metabolism.sulfur-containing.misc.alliinase
CH_vvi_142	1	8	27	29.5.9	Protein.degradation.AAA type
CH_vvi_168	16	12	41	30.2	Signalling.receptor kinases
CH_vvi_10	10	7	24	17.5.1	Hormone metabolism.ethylene.synthesis-degradation
CH_vvi_412	5	7	24	11.9	Lipid metabolism.lipid degradation
CH_vvi_185	12	6	21	20.1	Stress.biotic
CH_vvi_150	14	6	21	20.1.7	Stress.biotic.PR-proteins
CH_vvi_18	14	4	14	34.2	Transport.sugars
CH_vvi_134	11	2	7	31.5.1	Cell.cell death.plants
CH_vvi_239	15	2	7	33.99	Development.unspecified
CH_vvi_62	16	2	7	26.18	Misc.invertase/pectin methylesterase inhibitor family protein
CH_vvi_375	5	2	7	30.11	Signalling.light
CH_vvi_340	7	2	7	29.5.11.4.2	Protein.degradation.ubiquitin.E3.RING
CH_vvi_53	9	2	7	17.6.1	Hormone metabolism.gibberelin.synthesis-degradation
CH_vvi_249	10	3	11	16.8.4	Secondary metabolism.flavonoids.flavonols
CH_vvi_199	12	3	11	11.3.5	Lipid metabolism.Phospholipid synthesis.diacylglycerol kinase

Functional cluster	Chromo-some	High impact variants	Genes in cluster	MapMan bin	MapMan bin description
CH_vvi_228	5	3	11	30.3	Signalling.calcium
CH_vvi_97	13	7	26	20.2.1	Stress.abiotic.heat
CH_vvi_93	15	4	15	26.1	Misc.cytochrome P450
CH_vvi_81	12	20	76	20.1	Stress.biotic
CH_vvi_428	5	3	12	26.9	Misc.glutathione S transferases
CH_vvi_220	13	2	8	26.1	Misc.cytochrome P450
CH_vvi_213	18	2	8	30.2.11	Signalling.receptor kinases.leucine rich repeat XI
CH_vvi_221	18	2	8	26.1	Misc.cytochrome P450
CH_vvi_100	2	2	8	27.3.25	RNA.regulation of transcription.MYB domain transcription factor family
CH_vvi_187	6	2	8	26.2	Misc.UDP glucosyl and glucoronyl transferases
CH_vvi_136	8	2	8	26.1	Misc.cytochrome P450
CH_vvi_54	8	2	8	10.5.1.1	Cell wall.cell wall proteins.agps.AGP
CH_vvi_75	8	2	8	26.22	Misc.short chain dehydrogenase/reductase (SDR)
CH_vvi_56	9	2	8	26.28	Misc.GDSL-motif lipase
CH_vvi_247	1	1	4	26.28	Misc.GDSL-motif lipase
CH_vvi_407	1	1	4	26.1	Misc.cytochrome P450
CH_vvi_172	10	1	4	16.4.1	Secondary metabolism.N misc.alkaloid-like
CH_vvi_193	10	1	4	3.8.2	Minor CHO metabolism.galactose.alpha-galactosidases
CH_vvi_114	11	1	4	16.2.1.6	Secondary metabolism.phenylpropanoids.lignin biosynthesis.ccoaomt
CH_vvi_357	13	1	4	35.1.19	Not assigned.no ontology.C2 domain-containing protein
CH_vvi_248	14	1	4	26.28	Misc.GDSL-motif lipase
CH_vvi_333	14	1	4	26.2	Misc.UDP glucosyl and glucoronyl transferases
CH_vvi_191	16	1	4	10.8.1	Cell wall.pectin*esterases.PME
CH_vvi_277	18	1	4	31.4	Cell.vesicle transport
CH_vvi_386	18	1	4	30.2.11	Signalling.receptor kinases.leucine rich repeat XI
CH_vvi_230	2	1	4	26.3.5	Misc.gluco-, galacto- and mannosidases.glycosyl hydrolase family 5
CH_vvi_234	2	1	4	29.5.1	Protein.degradation.subtilases
CH_vvi_258	2	1	4	26.7	Misc.oxidases - copper, flavone etc
CH_vvi_394	5	1	4	11.3.7	Lipid metabolism.Phospholipid synthesis.cyclopropane-fatty-acyl-phospholipid synthase
CH_vvi_233	6	1	4	29.5.1	Protein.degradation.subtilases
CH_vvi_346	6	1	4	26.11	Misc.alcohol dehydrogenases
CH_vvi_401	6	1	4	30.2.17	Signalling.receptor kinases.DUF 26
CH_vvi_304	16	8	33	30.2.11	Signalling.receptor kinases.leucine rich repeat XI
CH_vvi_3	9	11	46	16.1.5	Secondary metabolism.isoprenoids.terpenoids
CH_vvi_12	2	5	21	10.2.1	Cell wall.cellulose synthesis.cellulose synthase
CH_vvi_5	9	8	34	30.2.99	Signalling.receptor kinases.misc
CH_vvi_146	1	2	9	35.1.26	Not assigned.no ontology.DC1 domain containing protein
CH_vvi_152	16	2	9	26.1	Misc.cytochrome P450
CH_vvi_427	3	2	9	26.8	Misc.nitrilases, nitrile lyases, berberine bridge enzymes, reticuline oxidases, troponine reductases
CH_vvi_103	7	2	9	34.15	Transport.potassium

Functional cluster	Chromo-some	High impact variants	Genes in cluster	MapMan bin	MapMan bin description
CH_vvi_26	19	17	78	26.1	Misc.cytochrome P450
CH_vvi_438	12	3	14	27.3.46	RNA.regulation of transcription.DNA methyltransferases
CH_vvi_14	5	3	14	27.1.19	RNA.processing.ribonucleases
CH_vvi_204	5	3	14	17.2.3	Hormone metabolism.auxin.induced-regulated-responsive-activated
CH_vvi_163	12	4	19	16.1	Secondary metabolism.isoprenoids
CH_vvi_382	1	12	58	35.1.5	Not assigned.no ontology.pentatricopeptide (PPR) repeat-containing protein
CH_vvi_77	5	6	29	31.1	Cell.organisation
CH_vvi_89	15	2	10	26.22	Misc.short chain dehydrogenase/reductase (SDR)
CH_vvi_57	18	2	10	30.2.17	Signalling.receptor kinases.DUF 26
CH_vvi_176	1	1	5	34.7	Transport.phosphate
CH_vvi_378	12	1	5	26.6	Misc.O-methyl transferases
CH_vvi_288	13	1	5	26.8	Misc.nitrilases, *nitrile lyases, berberine bridge enzymes, reticuline oxidases, troponine reductases
CH_vvi_384	13	1	5	34.16	Transport.ABC transporters and multidrug resistance systems
CH_vvi_69	19	1	5	26.3.5	Misc.gluco-, galacto- and mannosidases.glycosyl hydrolase family 5
CH_vvi_435	4	1	5	16.4.1	Secondary metabolism.N misc.alkaloid-like
CH_vvi_139	18	11	59	20.1.7	Stress.biotic.PR-proteins
CH_vvi_65	18	10	55	16.1	Secondary metabolism.simple phenols
CH_vvi_444	19	4	22	35.2	Not assigned.unknown
CH_vvi_445	4	2	11	27.3.99	RNA.regulation of transcription.unclassified
CH_vvi_201	7	2	11	16.1	Secondary metabolism.simple phenols
CH_vvi_135	5	11	61	20.1	Stress.biotic
CH_vvi_28	15	8	47	16.7	Secondary metabolism.wax
CH_vvi_24	16	4	24	17.5.2	Hormone metabolism.ethylene.signal transduction
CH_vvi_13	5	3	18	17.5.1	Hormone metabolism.ethylene.synthesis-degradation
CH_vvi_59	17	2	12	20.1.7.6.1	Stress.biotic.PR-proteins.proteinase inhibitors.trypsin inhibitor
CH_vvi_66	2	2	12	17.5.1	Hormone metabolism.ethylene.synthesis-degradation
CH_vvi_73	1	1	6	17.8.1	Hormone metabolism.salicylic acid.synthesis-degradation
CH_vvi_124	12	1	6	29.5.3	Protein.degradation.cysteine protease
CH_vvi_76	18	1	6	10.6.2	Cell wall.degradation.mannan-xylose-arabinose-fucose
CH_vvi_235	3	1	6	20.1	Stress.biotic
CH_vvi_166	8	1	6	30.2.17	Signalling.receptor kinases.DUF 26
CH_vvi_4	19	7	44	30.2.17	Signalling.receptor kinases.DUF 26
CH_vvi_286	5	3	19	35.2	Not assigned.unknown
CH_vvi_336	15	5	32	35.2	Not assigned.unknown
CH_vvi_128	12	4	26	30.2.11	Signalling.receptor kinases.leucine rich repeat XI
CH_vvi_298	14	4	26	21.6	Redox.dismutases and catalases
CH_vvi_82	12	2	13	10.6.3	Cell wall.degradation.pectate lyases and polygalacturonases
CH_vvi_98	14	2	13	29.5	Protein.degradation
CH_vvi_326	15	2	13	20.1	Stress.biotic

Functional cluster	Chromo-some	High impact variants	Genes in cluster	MapMan bin	MapMan bin description
CH_vvi_27	18	2	13	34.13	Transport.peptides and oligopeptides
CH_vvi_229	7	2	13	17.4.1	Hormone metabolism.cytokinin.synthesis-degradation
CH_vvi_43	19	5	34	30.2	Signalling.receptor kinases
CH_vvi_281	14	4	28	20.1.7	Stress.biotic.PR-proteins
CH_vvi_237	11	2	14	16.2.1.1	Secondary metabolism.phenylpropanoids.lignin biosynthesis.PAL
CH_vvi_179	14	2	14	29.2.1.2.2.34	Protein.synthesis.ribosomal protein.eukaryotic.60S subunit.L34
CH_vvi_356	15	2	14	20.1	Stress.biotic
CH_vvi_105	12	1	7	26.12	Misc.peroxidases
CH_vvi_316	14	1	7	31.2	Cell.division
CH_vvi_253	16	1	7	13.2.3.1.1	Amino acid metabolism.degradation.aspartate family.asparagine.L-asparaginase
CH_vvi_294	16	1	7	26.22	Misc.short chain dehydrogenase/reductase (SDR)
CH_vvi_278	18	1	7	26.2	Misc.UDP glucosyl and glucoronyl transferases
CH_vvi_78	18	1	7	26.28	Misc.GDSL-motif lipase
CH_vvi_209	3	1	7	17.1.3	Hormone metabolism.abscisic acid.induced-regulated-responsive-activated
CH_vvi_80	3	1	7	26.8	Misc.nitrilases, *nitrile lyases, berberine bridge enzymes, reticuline oxidases, troponine reductases
CH_vvi_223	1	8	58	18	Co-factor and vitamine metabolism
CH_vvi_177	16	2	15	29.5.1	Protein.degradation.subtilases
CH_vvi_19	4	2	15	17.8.1	Hormone metabolism.salicylic acid.synthesis-degradation
CH_vvi_205	6	2	15	29.5	Protein.degradation
CH_vvi_174	12	1	8	13.2.6.2	Amino acid metabolism.degradation.aromatic aa.tyrosine
CH_vvi_42	12	1	8	16.8.1.21	Secondary metabolism.flavonoids.anthocyanins.anthocyanin 5-aromatic acyltransferase
CH_vvi_178	14	1	8	30.3	Signalling.calcium
CH_vvi_36	18	1	8	17.7.1.5	Hormone metabolism.jasmonate.synthesis-degradation.12-Oxo-PDA-reductase
CH_vvi_111	19	1	8	34.16	Transport.ABC transporters and multidrug resistance systems
CH_vvi_305	2	1	8	21.2.2	Redox.ascorbate and glutathione.glutathione
CH_vvi_317	2	1	8	29.5.9	Protein.degradation.AAA type
CH_vvi_63	3	1	8	17.2.3	Hormone metabolism.auxin.induced-regulated-responsive-activated
CH_vvi_84	19	3	25	29.1.20	Protein.aa activation.phenylalanine-trna ligase
CH_vvi_16	18	2	17	33.1	Development.storage proteins
CH_vvi_130	5	2	17	20.1.7	Stress.biotic.PR-proteins
CH_vvi_22	10	2	18	16.1.5	Secondary metabolism.isoprenoids.terpenoids
CH_vvi_31	10	2	18	30.2	Signalling.receptor kinases
CH_vvi_21	11	2	18	10.7	Cell wall.modification
CH_vvi_127	15	1	9	29.5.1	Protein.degradation.subtilases
CH_vvi_118	5	1	9	20.1.7	Stress.biotic.PR-proteins
CH_vvi_123	9	1	9	31.1	Cell.organisation
CH_vvi_188	9	1	9	34.16	Transport.ABC transporters and multidrug resistance systems

Functional cluster	Chromo-some	High impact variants	Genes in cluster	MapMan bin	MapMan bin description
CH_vvi_20	14	4	38	20.2.99	Stress.abiotic.unspecified
CH_vvi_147	10	2	19	17.1	Hormone metabolism.abscisic acid
CH_vvi_138	9	2	20	30.2.11	Signalling.receptor kinases.leucine rich repeat XI
CH_vvi_327	1	1	10	33.99	Development.unspecified
CH_vvi_61	1	1	10	26.12	Misc.peroxidases
CH_vvi_64	1	1	10	20.2.99	Stress.abiotic.unspecified
CH_vvi_85	12	1	10	30.2.17	Signalling.receptor kinases.DUF 26
CH_vvi_41	14	1	10	3.1.1.2	Minor CHO metabolism.raffinose family.galactinol synthases.putative
CH_vvi_195	18	1	10	30.1	Signalling.in sugar and nutrient physiology
CH_vvi_129	2	2	21	34	Transport
CH_vvi_79	9	2	21	16.2	Secondary metabolism.phenylpropanoids
CH_vvi_74	18	7	75	20.1.7	Stress.biotic.PR-proteins
CH_vvi_148	3	7	77	27.3.41	RNA.regulation of transcription.B3 transcription factor family
CH_vvi_2	16	2	22	16.8.2	Secondary metabolism.flavonoids.chalcones
CH_vvi_345	14	1	11	29.3.99	Protein.targeting.unknown
CH_vvi_86	14	1	11	17.2.3	Hormone metabolism.auxin.induced-regulated-responsive-activated
CH_vvi_35	4	1	11	16.4.1	Secondary metabolism.N misc.alkaloid-like
CH_vvi_250	9	1	11	29.7.3	Protein.glycosylation.mannosyl-oligosaccharide alpha-1,2-mannosidase
CH_vvi_50	12	2	24	31.1	Cell.organisation
CH_vvi_32	14	1	12	13.1.5.3.1	Amino acid metabolism.synthesis.serine-glycine-cysteine group.cysteine.OASTL
CH_vvi_126	4	1	12	30.2.11	Signalling.receptor kinases.leucine rich repeat XI
CH_vvi_108	5	2	26	26.2	Misc.UDP glucosyl and glucoronyl transferases
CH_vvi_25	16	1	13	16.2.1.1	Secondary metabolism.phenylpropanoids.lignin biosynthesis.PAL
CH_vvi_194	19	4	56	20.1	Stress.biotic
CH_vvi_67	5	1	14	26.7	Misc.oxidases - copper, flavone etc
CH_vvi_210	6	1	15	28.1.3	DNA.synthesis/chromatin structure.histone
CH_vvi_47	9	1	15	29.2.1.2.2.7	Protein.synthesis.ribosomal protein.eukaryotic.60S subunit.L7
CH_vvi_95	12	1	16	33.1	Development.storage proteins
CH_vvi_11	6	1	20	26.8	Misc.nitrilases, *nitrile lyases, berberine bridge enzymes, reticuline oxidases, troponine reductases
CH_vvi_6	13	1	21	26.3	Misc.gluco-, galacto- and mannosidases
CH_vvi_222	3	1	25	35.1	Not assigned.no ontology
CH_vvi_71	5	1	29	26	Misc
CH_vvi_145	13	1	31	26.11.1	Misc.alcohol dehydrogenases.cinnamyl alcohol dehydrogenase
CH_vvi_338	13	1	32	16.8.3.1	Secondary metabolism.flavonoids.dihydroflavonols

Misc: Miscellaneous

Supplementary data table 4.1: Genes expressed in berries with an FPKM of 1000 and higher, with associated EC numbers and enzyme descriptions.

Gene	EC number	Enzyme
VIT_218s0001g01140	EC:1.11.1.7	Peroxidase
VIT_202s0025g03600	EC:1.11.1.9	Glutathione peroxidase
VIT_217s0000g10430	EC:1.2.1.12	Glyceraldehyde-3-phosphate dehydrogenase
VIT_218s0001g09990	EC:1.3.1.74	2-alkenal reductase [NAD(P)+]
VIT_213s0019g01430	EC:1.3.1.74	2-alkenal reductase [NAD(P)+]
VIT_201s0026g01460	EC:1.8.4.0	Electron transport/ With a disulfide as acceptor
VIT_218s0001g13250	EC:1.8.4.0	Electron transport/ With a disulfide as acceptor
VIT_218s0001g13240	EC:1.8.4.0	Electron transport/ With a disulfide as acceptor
VIT_202s0033g01120	EC:2.1.1.0	Methyltransferases
VIT_206s0061g00550	EC:2.4.1.207	Xyloglucan:xyloglucosyl transferase
VIT_211s0016g00200	EC:2.7.11.17	Calcium/calmodulin-dependent protein kinase.
VIT_216s0022g00960	EC:3.1.1.11	Pectinesterase
VIT_206s0061g00550	EC:3.2.1.0	Glycosidases
VIT_218s0001g12830	EC:3.2.1.0	Glycosidases
VIT_205s0094g00340	EC:3.2.1.14	Chitinase
VIT_205s0094g00350	EC:3.2.1.14	Chitinase
VIT_208s0007g05990	EC:3.2.1.39	Glucan endo-1,3-β-D-glucosidase
VIT_208s0007g06020	EC:3.2.1.39	Glucan endo-1,3-β-D-glucosidase
VIT_208s0007g06030	EC:3.2.1.39	Glucan endo-1,3-β-D-glucosidase
VIT_208s0007g06040	EC:3.2.1.39	Glucan endo-1,3-β-D-glucosidase
VIT_208s0007g06000	EC:3.2.1.39	Glucan endo-1,3-β-D-glucosidase
VIT_208s0007g06010	EC:3.2.1.39	Glucan endo-1,3-β-D-glucosidase
VIT_208s0007g06060	EC:3.2.1.39	Glucan endo-1,3-β-D-glucosidase
VIT_206s0061g00550	EC:3.2.1.4	Cellulase
VIT_201s0010g00990	EC:4.1.1.19	Arginine decarboxylase
VIT_200s1995g00010	EC:4.1.1.49	Phosphoenolpyruvate carboxykinase (ATP)
VIT_201s0010g00990	EC:4.1.1.50	Adenosylmethionine decarboxylase
VIT_208s0007g03830	EC:4.1.2.13	Fructose-bisphosphate aldolase
VIT_216s0022g01770	EC:4.2.1.11	Phosphopyruvate hydratase
VIT_203s0038g01930	EC:5.2.1.8	Peptidyl-prolyl cis-trans isomerase
VIT_218s0001g14400	EC:5.2.1.8	Peptidyl-prolyl cis-trans isomerase
VIT_208s0040g00040	EC:6.3.2.19	Ubiquitin—protein ligase

Supplementary data table 4.2: Genes expressed in leaves with an FPKM of 1000 and higher, with associated EC numbers and enzyme descriptions.

Gene	EC number	Enzyme
VIT_203s0017g01470	EC:1.1.1.158	UDP-N-acetylmuramate dehydrogenase
VIT_217s0000g05790	EC:1.13.11.19	Cysteamine dioxygenase
VIT_202s0012g00360	EC:1.14.11.23	Flavonol synthase
VIT_214s0128g00430	EC:1.3.1.74	2-alkenal reductase [NAD(P)+]
VIT_210s0003g02170	EC:1.3.1.74	2-alkenal reductase [NAD(P)+]
VIT_201s0137g00010	EC:1.6.5.3	NADH:ubiquinone reductase (H+-translocating)
VIT_212s0057g01250	EC:2.1.1.127	[Ribulose-bisphosphate carboxylase]-lysine N-methyltransferase
VIT_204s0023g02230	EC:2.1.1.141	Jasmonate O-methyltransferase
VIT_218s0001g12880	EC:2.1.1.141	Jasmonate O-methyltransferase
VIT_219s0135g00030	EC:2.1.1.68	Caffeate O-methyltransferase
VIT_200s0203g00180	EC:2.3.1.0	Transketolases
VIT_201s0011g01830	EC:2.3.1.43	Phosphatidylcholine—sterol O-acyltransferase
VIT_218s0001g15400	EC:2.4.1.12	Cellulose synthase (UDP-forming)
VIT_214s0128g00330	EC:2.7.1.36	Mevalonate kinase
VIT_214s0128g00330	EC:2.7.1.6	Galactokinase
VIT_214s0128g00430	EC:2.7.10.0	Protein-tyrosine kinases
VIT_208s0007g08300	EC:2.7.10.0	Protein-tyrosine kinases
VIT_218s0001g09780	EC:2.7.11.0	Protein-tyrosine kinases
VIT_208s0007g08300	EC:2.7.11.17	Calcium/calmodulin-dependent protein kinase
VIT_217s0000g04400	EC:2.7.11.25	Mitogen-activated protein kinase kinase kinase
VIT_214s0128g00430	EC:2.7.11.25	Mitogen-activated protein kinase kinase kinase
VIT_218s0001g09780	EC:2.7.7.49	RNA-directed DNA polymerase
VIT_205s0077g00656	EC:3.1.1.11	Pectinesterase
VIT_212s0059g01470	EC:3.1.1.3	Triacylglycerol lipase
VIT_202s0012g00460	EC:3.4.21.0	Serine endopeptidases
VIT_210s0003g02170	EC:3.4.21.0	Serine endopeptidases
VIT_206s0004g07990	EC:3.4.24.0	Serine endopeptidases
VIT_217s0000g06390	EC:3.4.24.0	Serine endopeptidases
VIT_217s0000g06390	EC:3.6.4.3	Microtubule-severing ATPase
VIT_212s0057g01240	EC:4.1.1.48	Indole-3-glycerol-phosphate synthase
VIT_202s0012g00360	EC:4.2.1.78	(S)-norcoclaurine synthase
VIT_214s0081g00700	EC:5.2.1.8	Peptidyl-prolyl cis-trans isomerase
VIT_219s0027g01660	EC:5.2.1.8	Peptidyl-prolyl cis-trans isomerase
VIT_204s0069g01146	EC:6.2.1.26	O-succinylbenzoate—coA ligase

Supplementary data table 4.3: Differentially expressed genes classified in the MapMan bin "Hormone metabolism".

MapMan bin		GeneID	FPKM old vines	FPKM young vines	Log_2FC
		Berries			
Abscisic acid	Related	VIT_207s0005g00140	10.62	4.84	-1.13
Auxin	Related	VIT_207s0031g02740	33.42	12.85	-1.38
	Related	VIT_211s0016g00500	20.07	8.66	-1.21
	Synthesis/Degradation	VIT_208s0007g02760	23.06	38.70	0.75
Brassinosteroids	Signal transduction	VIT_203s0038g03860	303.39	121.89	-1.32
Cytokinin	Synthesis/Degradation	VIT_218s0001g05990	19.22	35.42	0.88
	Synthesis/Degradation	VIT_218s0001g06060	116.56	202.68	0.80
Ethylene	Related	VIT_201s0011g02790	26.01	13.33	-0.97
	Signal transduction	VIT_212s0028g03270	263.41	99.24	-1.41
	Signal transduction	VIT_216s0013g00890	34.96	16.43	-1.09
	Signal transduction	VIT_218s0089g01030	59.91	13.30	-2.17
Gibberellin	Related	VIT_208s0040g01820	747.11	266.14	-1.49
Jasmonate	Signal transduction	VIT_206s0004g01510	136.50	253.14	0.89
		Leaves			
Abscisic acid	Related	VIT_203s0038g01650	49.45	69.96	0.50
	Synthesis/Degradation	VIT_212s0055g00060	153.26	62.51	-1.29
Auxin	Related	VIT_200s0251g00020	6.79	18.53	1.45
	Related	VIT_204s0023g00560	9.48	19.26	1.02
	Related	VIT_210s0597g00010	13.56	26.05	0.94
	Related	VIT_203s0038g01080	14.64	8.08	-0.86
	Related	VIT_203s0038g01150	661.84	432.37	-0.61
	Related	VIT_203s0038g01260	196.93	126.41	-0.64
	Related	VIT_204s0044g01200	79.73	54.34	-0.55
	Related	VIT_218s0001g05210	62.87	44.65	-0.49
	Related	VIT_218s0001g14330	279.07	194.85	-0.52
	Related	VIT_219s0014g03130	32.41	20.95	-0.63
	Synthesis/Degradation	VIT_205s0062g00740	14.40	23.37	0.70
	Signal transduction	VIT_211s0052g00440	373.54	246.94	-0.60
Brassinosteroid	Related	VIT_201s0011g02360	23.72	36.33	0.62
	Related	VIT_211s0016g03790	12.32	33.41	1.44
	Signal transduction	VIT_201s0010g03124	41.06	118.33	1.53
	Signal transduction	VIT_212s0121g00430	50.66	122.35	1.27
	Signal transduction	VIT_219s0027g01544	24.83	50.82	1.03
Cytokinin	Signal transduction	VIT_201s0026g01310	40.14	17.88	-1.17
Ethylene	Related	VIT_204s0023g02410	8.66	17.54	1.02
	Synthesis/Degradation[§]	VIT_201s0011g05650	288.73	205.51	-0.49
	Synthesis/Degradation*	VIT_202s0025g00360	17.33	25.68	0.57
	Signal transduction	VIT_204s0023g02410	8.66	17.54	1.02
	Signal transduction	VIT_208s0058g00050	7.81	13.85	0.83
	Signal transduction	VIT_219s0014g02240	20.18	35.47	0.81

	Related	VIT_205s0077g01120	196.49	324.91	0.73
	Related	VIT_208s0040g01820	32.02	19.67	-0.70
Gibberellin	Synthesis/Degradation	VIT_200s2197g00010	21.46	34.61	0.69
	Synthesis/Degradation	VIT_207s0151g01070	16.29	7.95	-1.04
	Signal transduction	VIT_205s0077g01120	196.49	324.91	0.73
Jasmonate	Synthesis/Degradation	VIT_213s0064g01500	29.15	18.73	-0.64

[§] 1-aminocyclopropane-1-carboxylate oxidase

*1-aminocyclopropane-1-carboxylate synthase

Supplementary data table 4.4: Differentially expressed genes classified in the MapMan bin "Cell wall".

MapMan bin	Gene ID	FPKM old vines	FPKM young vines	Log_2FC	
Berries					
Cellulose Synthesis	VIT_206s0061g01230	7.07	12.50	0.82	↑
Cellulose Synthesis	VIT_200s0469g00040	9.90	17.03	0.78	
Cellulose Synthesis	VIT_219s0015g00730	7.48	13.29	0.83	
Degradation	VIT_211s0118g00420	7.42	15.08	1.02	
Degradation Pectate Lyase	VIT_205s0051g00590	32.60	99.82	1.61	
Modification	VIT_201s0026g02620	82.95	135.38	0.71	
Modification	VIT_208s0007g00440	4.31	10.16	1.24	
Modification	VIT_211s0016g04720	4.50	10.49	1.22	
Modification	VIT_213s0067g02930	417.68	767.41	0.88	
Modification	VIT_214s0108g01020	7.45	21.16	1.51	
Modification	VIT_215s0021g02700	820.92	1906.93	1.22	
Modification	VIT_217s0053g00990	29.44	57.13	0.96	
Cell Wall Precursor Synthesis	VIT_217s0000g06960	33.77	18.47	-0.87	↓
Hemicellulose Synthesis	VIT_204s0023g01120	16.19	6.75	-1.26	
Cell Wall Proteins AGPs	VIT_203s0091g00420	39.86	23.32	-0.77	
Cell Wall Proteins AGPs	VIT_207s0129g00560	13.93	3.87	-1.85	
Cell Wall Proteins AGPs	VIT_208s0007g08020	156.22	57.45	-1.44	
Cell Wall Proteins AGPs	VIT_208s0040g01820	747.11	266.14	-1.49	
Pectin Esterases	VIT_215s0048g00500	92.60	43.76	-1.08	
Leaves					
Cell Wall Precursor Synthesis	VIT_202s0025g04610	13.27	32.73	1.30	↑
Cell Wall Precursor Synthesis	VIT_202s0025g04610	13.27	32.73	1.30	
Cellulose Synthesis	VIT_200s0469g00020	27.79	47.24	0.77	
Hemicellulose Synthesis	VIT_212s0057g01420	19.47	41.25	1.08	
Cell Wall Proteins AGPs	VIT_200s0302g00060	441.76	622.56	0.49	
Cell Wall Proteins AGPs	VIT_208s0007g07980	14.72	25.98	0.82	
Cell Wall Proteins LRR	VIT_201s0127g00550	11.13	17.08	0.62	
Degradation cellulases	VIT_214s0036g01040	22.81	32.27	0.50	
Degradation pectate Lyases	VIT_209s0002g08690	448.56	715.98	0.67	
Modification	VIT_213s0019g01650	7.27	11.77	0.70	
Cellulose Synthesis	VIT_219s0085g00670	74.69	51.21	-0.55	↓
Hemicellulose Synthesis	VIT_204s0023g01120	43.51	26.93	-0.69	
Cell Wall Proteins AGP	VIT_201s0010g03150	18.13	12.17	-0.58	
Cell Wall Proteins AGP	VIT_208s0040g01820	32.02	19.67	-0.70	
Cell Wall Proteins LRR	VIT_201s0026g01310	40.14	17.88	-1.17	
Degradation	VIT_203s0091g00890	76.07	48.61	-0.65	
Degradation	VIT_218s0001g02220	19.39	11.22	-0.79	
Degradation Pectate Lyase	VIT_200s0220g00140	487.99	305.08	-0.68	
Degradation Pectate Lyase	VIT_206s0004g02550	12.56	8.79	-0.51	
Modification	VIT_206s0004g02550	12.56	8.79	-0.51	
Pectin Esterases	VIT_210s0116g01600	35.36	24.54	-0.53	
Pectin Esterases	VIT_211s0016g00300	10.65	4.66	-1.20	

AGP: arabinogalactan proteins, LRR: Leaucine rich repeat

Supplementary data table 4.5: Differentially expressed genes in sugar signalling and transport (MapMan bins "Signalling: sugar & nutrient physiology" and "Transport sugars").

MapMan bin	Gene ID	FPKM old vines	FPKM young vines	Log$_2$FC	
		Berries			
Signalling: sugar & nutrient physiology	VIT_203s0017g01210	208.46	79.19	-1.40	
Signalling: sugar &nutrient physiology	VIT_207s0005g00870	65.73	16.20	-2.02	↓
Transport sugars	VIT_203s0063g02250	17.43	7.84	-1.15	
Transport sugars	VIT_205s0020g02170	21.38	10.07	-1.09	
		Leaves			
Signalling: sugar & nutrient physiology	VIT_205s0077g01120	196.49	324.91	0.73	
Signalling: sugar & nutrient physiology	VIT_210s0003g00790	6.89	10.70	0.64	
Signalling: sugar & nutrient physiology	VIT_217s0000g10480	30.99	52.15	0.75	↑
Signalling: sugar & nutrient physiology	VIT_218s0001g06170	95.98	141.24	0.56	
Transport sugars	VIT_209s0002g08690	448.56	715.98	0.67	
Signalling: sugar & nutrient physiology	VIT_219s0014g01730	868.41	537.27	-0.69	↓

Supplementary data table 4.6: Differential expressed genes classified in the MapMan bin “Secondary metabolism”.

MapMan bin	Gene ID	FPKM old vines	FPKM young vines	Log$_2$FC	
Berries					
Nitrogen containing	VIT_204s0210g00060	32.29	52.66	0.71	
Nitrogen containing	VIT_206s0004g05380	4.63	12.37	1.42	
Isoprenoids	VIT_205s0020g02130	9.41	16.59	0.82	↑
Wax	VIT_214s0006g02990	30.79	51.66	0.75	
Flavonoids	VIT_208s0105g00380	36.83	63.39	0.78	
Lignin	VIT_208s0040g01710	20.75	6.37	-1.70	
Lignin	VIT_208s0040g00780	14.53	5.46	-1.41	
Lignin	VIT_209s0070g00240	11.00	3.98	-1.47	↓
Lignin	VIT_204s0023g02900	14.97	3.74	-2.00	
Nitrogen containing	VIT_210s0003g05450	17.92	10.58	-0.76	
Leaves					
Lignin	VIT_208s0007g04060	31.40	61.43	0.97	
Isoprenoids	VIT_206s0009g03090	28.35	72.84	1.36	
Isoprenoids	VIT_218s0001g04120	40.83	58.25	0.51	
Flavonoids	VIT_201s0011g06520	8.02	13.74	0.78	
Flavonoids	VIT_205s0077g02190	19.46	33.39	0.78	
Flavonoids	VIT_218s0001g09560	7.86	16.87	1.10	↑
Flavonoids	VIT_206s0009g02810	14.30	21.53	0.59	
Flavonoids	VIT_217s0000g07200	37.87	72.77	0.94	
Flavonoids	VIT_203s0038g04710	14.76	36.35	1.30	
Unspecified	VIT_216s0050g01430	62.04	97.34	0.65	
Sulfur containing	VIT_202s0033g00850	8.83	14.90	0.75	
Isoprenoids	VIT_200s0169g00100	12.86	8.04	-0.68	
Isoprenoids	VIT_218s0001g02720	697.34	362.34	-0.94	
Isoprenoids	VIT_203s0132g00010	27.88	12.61	-1.15	
Isoprenoids	VIT_200s0169g00100	12.86	8.04	-0.68	
Simple phenols	VIT_218s0001g00850	26.74	17.42	-0.62	
Simple phenols	VIT_218s0001g02350	32.55	20.75	-0.65	
Simple phenols	VIT_218s0164g00170	772.90	495.97	-0.64	↓
Sulfur containing	VIT_207s0031g01730	206.72	112.56	-0.88	
Wax	VIT_218s0001g02720	697.34	362.34	-0.94	
Flavonoids	VIT_212s0134g00620	25.44	15.43	-0.72	
Flavonoids	VIT_210s0042g00870	33.49	17.95	-0.90	
Flavonoids	VIT_216s0100g00900	14.94	8.66	-0.79	
Flavonoids	VIT_212s0055g00060	153.26	62.51	-1.29	
Flavonoids	VIT_202s0012g00420	111.25	73.07	-0.61	

Supplementary data table 4.7: Differentially expressed genes classified in the MapMan bin "Lipid metabolism".

MapMan bin	Gene ID	FPKM old vines	FPKM young vines	Log$_2$FC	
		Berries			
Synthesis	VIT_205s0049g00800	19.55	77.60	2.00	↑
Synthesis	VIT_200s0271g00110	37.81	60.94	0.69	
Degradation	VIT_203s0063g00830	26.42	45.95	0.80	
Degradation	VIT_203s0063g00830	26.42	45.95	0.80	
Degradation	VIT_214s0066g00700	44.18	21.21	-1.06	↓
Degradation	VIT_209s0002g05730	24.53	12.93	-0.92	
		Leaves			
Synthesis	VIT_200s0357g00020	61.52	95.20	0.63	↑
Synthesis	VIT_204s0210g00110	244.04	346.70	0.51	
Synthesis	VIT_215s0021g00580	35.02	66.81	0.93	
Synthesis	VIT_214s0006g00580	11.45	16.68	0.54	
Degradation	VIT_216s0098g00510	10.70	32.88	1.62	
Degradation	VIT_209s0002g00590	31.05	60.79	0.97	
Degradation	VIT_202s0025g04620	42.24	67.48	0.68	
Degradation	VIT_203s0063g00710	53.28	100.14	0.91	
Degradation	VIT_216s0098g00180	8.59	31.67	1.88	
Degradation	VIT_216s0098g00200	34.95	104.94	1.59	
Degradation	VIT_216s0098g00400	16.43	50.16	1.61	
Degradation	VIT_216s0098g00420	17.67	58.20	1.72	
Degradation	VIT_217s0000g07450	60.41	117.22	0.96	
Degardation	VIT_203s0063g00710	53.28	100.14	0.91	
Synthesis	VIT_218s0001g02720	697.34	362.34	-0.94	↓
Synthesis	VIT_218s0001g02720	697.34	362.34	-0.94	
Synthesis	VIT_206s0004g02550	12.56	8.79	-0.51	
Synthesis	VIT_204s0008g01450	369.99	259.23	-0.51	
Synthesis	VIT_200s0207g00050	215.00	141.14	-0.61	
Degradation	VIT_204s0008g05340	941.48	606.86	-0.63	
Degradation	VIT_216s0013g01160	38.48	13.18	-1.55	
Degradation	VIT_216s0013g01350	111.32	73.64	-0.60	
Degradation	VIT_204s0008g05340	941.48	606.86	-0.63	

Supplementary data table 4.8: Differentially expressed genes classified in the MapMan bin “Transporters”.

MapMan bin	Gene ID	FPKM old vines	FPKM young vines	Log$_2$FC	
		Berries			
Nucleotides	VIT_208s0007g04550	7.12	13.46	0.92	↑
Nucleotides	VIT_210s0003g02840	14.09	30.83	1.13	
Metal	VIT_202s0025g00820	23.44	45.75	0.96	
Metal	VIT_208s0058g00740	8.25	15.74	0.93	
Major intrinsic proteins	VIT_206s0004g02850	5.99	10.43	0.80	
Calcium regulated channels	VIT_204s0069g00790	20.39	43.07	1.08	
Amino acids	VIT_203s0038g02860	19.71	32.30	0.71	
Nitrate	VIT_201s0026g01570	16.12	37.04	1.20	
Phosphate	VIT_200s0187g00160	244.65	490.60	1.00	
Envelope membrane	VIT_202s0025g04920	20.72	33.60	0.70	
Miscellaneous	VIT_208s0007g08200	10.79	21.18	0.97	
Vesicle transport	VIT_208s0032g01150	66.21	38.37	-0.79	↓
Metal	VIT_216s0013g00440	38.99	18.74	-1.06	
Metal	VIT_216s0013g00480	91.29	52.55	-0.80	
ABC transporters	VIT_216s0013g00450	117.27	66.98	-0.81	
Sugars	VIT_203s0063g02250	17.43	7.84	-1.15	
Sugars	VIT_205s0020g02170	21.38	10.07	-1.09	
Calcium	VIT_216s0013g00410	32.49	18.08	-0.85	
Amino acids	VIT_208s0007g08010	100.49	51.98	-0.95	
Mitochondrial membrane	VIT_200s2349g00010	183.24	89.03	-1.04	
Mitochondrial membrane	VIT_218s0001g07320	56.63	29.36	-0.95	
		Leaves			
Vesicle transport	VIT_202s0012g01330	13.64	21.41	0.65	↑
Vesicle transport	VIT_203s0063g02450	5.17	13.38	1.37	
Vesicle transport	VIT_208s0007g02340	116.99	168.48	0.53	
Vesicle transport	VIT_215s0045g00060	8.01	12.81	0.68	
P- and v-ATPases	VIT_212s0028g03277	17.36	26.50	0.61	
P- and v-ATPases	VIT_207s0141g00500	38.55	57.74	0.58	
Peptides and oligopeptides	VIT_200s0438g00030	15.77	28.25	0.84	
Peptides and oligopeptides	VIT_218s0001g13350	9.46	21.71	1.20	
ABC transporters	VIT_201s0011g04670	6.70	10.68	0.67	
ABC transporters	VIT_216s0098g00570	24.27	88.63	1.87	
Unspecified anions	VIT_200s0316g00020	9.79	20.67	1.08	
Unspecified anions	VIT_200s0336g00020	13.89	23.08	0.73	
Major intrinsic proteins	VIT_212s0028g03235	30.66	50.26	0.71	
Sugars	VIT_209s0002g08690	448.56	715.98	0.67	
Amino acids	VIT_201s0010g01540	35.97	52.46	0.54	
Phosphate	VIT_201s0182g00130	37.91	54.04	0.51	
Phosphate	VIT_205s0049g00920	6.41	13.97	1.12	
Phosphate	VIT_218s0122g00780	9.05	13.30	0.56	
Envelope membrane	VIT_210s0003g00300	18.95	39.38	1.06	
Envelope membrane	VIT_217s0000g08560	48.06	84.23	0.81	
Mitochondrial membrane	VIT_214s0128g00390	18.58	26.26	0.50	
Mitochondrial membrane	VIT_216s0039g00600	45.82	72.23	0.66	
Mitochondrial membrane	VIT_216s0039g00630	7.68	14.19	0.89	
Miscellaneous	VIT_201s0011g04430	97.12	138.49	0.51	
Miscellaneous	VIT_216s0039g00720	5.31	15.17	1.52	
Miscellaneous	VIT_218s0122g01330	17.75	25.58	0.53	
Vesicle transport	VIT_201s0127g00810	410.42	211.00	-0.96	↓
Vesicle transport	VIT_214s0006g00970	16.18	9.71	-0.74	
Vesicle transport	VIT_214s0083g00780	16.77	10.98	-0.61	
P- and v-ATPases	VIT_200s0288g00050	107.55	66.15	-0.70	

Metal	VIT_208s0058g00740	24.25	14.97	-0.70
Metal	VIT_212s0035g02230	163.32	109.88	-0.57
Unspecified cations	VIT_219s0085g00910	127.87	84.38	-0.60
Potassium	VIT_211s0016g04750	245.95	142.97	-0.78
Potassium	VIT_212s0134g00250	14.32	5.84	-1.30
ABC transporters	VIT_206s0061g00260	19.09	11.60	-0.72
ABC transporters	VIT_218s0166g00080	12.05	7.13	-0.76
Calcium	VIT_212s0057g00615	121.99	76.06	-0.68
Amino acids	VIT_203s0038g02680	40.30	23.33	-0.79
Amino acids	VIT_208s0007g08010	83.60	48.94	-0.77
Amino acids	VIT_213s0073g00050	86.93	50.54	-0.78
Amino acids	VIT_219s0027g01860	96.52	58.82	-0.71
Phosphate	VIT_200s0186g00110	104.78	73.83	-0.51
Phosphate	VIT_201s0011g02520	216.86	144.36	-0.59
Mitochondrial membrane	VIT_216s0039g00420	30.11	20.68	-0.54
Miscellaneous	VIT_200s0301g00030	42.17	27.51	-0.62
Miscellaneous	VIT_201s0011g03220	13.07	7.71	-0.76
Miscellaneous	VIT_208s0056g01070	82.25	46.54	-0.82
Miscellaneous	VIT_213s0019g05200	1365.89	935.73	-0.55
Miscellaneous	VIT_214s0006g02260	185.27	113.38	-0.71
Miscellaneous	VIT_219s0085g00910	127.87	84.38	-0.60

Supplementary data table 4.9: Differentially expressed genes classified in the MapMan bin "Stress biotic PR-proteins".

Mapman bin	Gene ID	FPKM old vines	FPKM young vines	Log$_2$FC	
Berries					
Stress biotic PR-proteins	VIT_205s0077g01670	10.90	48.75	2.16	↑
Stress biotic PR-proteins	VIT_200s0270g00120	29.36	9.82	-1.58	↓
Leaves					
Stress biotic PR-proteins	VIT_201s0010g03165	43.63	133.25	1.61	↑
Stress biotic PR-proteins	VIT_203s0038g01750	124.45	241.92	0.96	
Stress biotic PR-proteins	VIT_213s0139g00190	7.77	14.83	0.93	
Stress biotic PR-proteins	VIT_214s0036g00100	65.59	122.60	0.90	
Stress biotic PR-proteins	VIT_217s0000g06850	7.35	12.92	0.81	
Stress biotic PR-proteins	VIT_218s0001g06210	90.46	158.07	0.81	
Stress biotic PR-proteins	VIT_218s0089g00050	50.39	85.30	0.76	
Stress biotic PR-proteins	VIT_218s0089g00500	13.33	20.23	0.60	
Stress biotic PR-proteins	VIT_218s0122g01330	17.75	25.58	0.53	
Stress biotic PR-proteins	VIT_219s0027g00660	34.73	73.68	1.09	
Stress biotic PR-proteins	VIT_219s0027g01540	12.48	28.00	1.17	
Stress biotic PR-proteins	VIT_219s0027g01542	3.93	10.69	1.44	
Stress biotic PR-proteins	VIT_219s0027g01546	150.01	344.40	1.20	
Stress biotic PR-proteins	VIT_219s0027g01550	31.25	50.15	0.68	
Stress biotic PR-proteins	VIT_219s0027g01570	12.39	23.04	0.89	
Stress biotic PR-proteins	VIT_200s0160g00310	56.51	38.47	-0.55	↓
Stress biotic PR-proteins	VIT_200s1256g00010	615.06	433.03	-0.51	
Stress biotic PR-proteins	VIT_204s0044g00040	42.60	21.88	-0.96	
Stress biotic PR-proteins	VIT_206s0004g00370	101.57	61.01	-0.74	
Stress biotic PR-proteins	VIT_212s0055g00550	22.94	14.10	-0.70	
Stress biotic PR-proteins	VIT_215s0024g00440	436.37	277.33	-0.65	
Stress biotic PR-proteins	VIT_216s0050g02010	67.39	47.12	-0.52	
Stress biotic PR-proteins	VIT_219s0014g00600	101.39	72.35	-0.49	

PR Pathogenesis related

Supplementary data 5.1: Sequence data for the 988 Pinotage genes.

Available on request from beatrix@sun.ac.za