UNIVERSITÀ DEGLI STUDI DI UDINE

Dottorato di Ricerca in Scienze e Biotecnologie Agrarie
Ciclo XXVII
Coordinatore: prof. Mauro Spanghero

TESI DI DOTTORATO DI RICERCA

The Rpv3 locus in grapevine: DNA variation and relevance for conventional breeding

DOTTORANDA
Serena Foria

SUPERVISORE
Gabriele Di Gaspero

ANNO ACCADEMICO 2014/2015

Summary

This thesis is composed of two chapters. In the first chapter, I report the sequencing and assembly of a haplotype of the *Rpv3* locus that confers race-specific resistance to *Plasmopara viticola*. Full-length sequences of the alleles were reconstructed for all genes in the locus and cloned from the resistant haplotype into an expression vector. *Agrobacterium*-mediated transformation of embryogenic calli from susceptible grapevines has been initiated for the functional characterization of each allele. During my PhD, I also worked on a grapevine breeding program for stacking major resistance genes against downy and powdery mildews and selecting high-quality resistant varieties via conventional breeding. Based on these experience, a review article entitled "Molecular grape breeding techniques" written by me and my supervisor will be published as a chapter in the book "Grape breeding programs for the wine industry" edited by Andrew G. Reynolds for the Woodhead Publishing series in Food Science, Technology and Nutrition. This article is also reported in the second chapter of this thesis.

Abstract of Chapter 1

Haplotypes at the *Rpv3* locus control the onset of hypersensitive responses to *Plasmopara viticola*. A few of them have been introgressed from North American grapevines into varieties of *Vitis vinifera* during historical breeding. We sequenced and assembled 105 kbp of one of these resistance haplotypes, shared by the descent group of ‘Seibel 4614’, which amounts to dozens of varieties obtained by phenotypic selection during the past century. We used complementary strategies to assemble and annotate the sequence of the resistance haplotype: restriction-based BAC fingerprinting, BAC sequencing, Illumina resequencing of the nuclear genome in a homozygous individual, RNA-Seq in the same individual during the course of pathogen infection, and PCR-based cloning of tandem duplicate genes. The comparison with 50 pure *vinifera* diploid genomes – resequenced with Illumina technology – provided an inventory of DNA variation, including gene content, nucleotide diversity, and structural variation. In this locus, gene density is low. Synteny is conserved with the exception of an extra copy of one NB-LRR gene in the wild haplotype. In quantitative

terms, most variation is accounted for by insertions of transposable elements in the intergenic space. Full-length alleles of all genes in the locus were isolated from the resistance haplotype and cloned into an expression vector for *Agrobacterium*-mediated grapevine transformation.

Abstract of Chapter 2

Genome sequencing was a quantum leap for conventional breeding in grapevine, which disclosed an unprecedented amount of DNA sequence information – intensifying the exploitation of genomics tools by breeders and opening the door to innovative strategies. A transition from empirical breeding to precision breeding is now ongoing. The greatest challenge ahead is to disentangle and sculpt those traits most relevant to the wine industry. The ultimate goal is the design and assembly of an ideal genome in a new variety composed of desirable chromosomal segments – each one contributing a favourable haplotype for a target trait or for setting an appropriate genetic background.

Abbreviations

BAC	bacterial artificial chromosome
DM	downy mildew
HR	hypersensitive response
LRR	leucine-rich repeat
NB-ARC	nucleotide-binding adaptor shared by APAF-1
NB-LRR	nucleotide-binding LRR
NGS	next generation sequencing
Rpv	resistance to *Plasmopara viticola*
SNP	single nucleotide polymorphism
TIR	Toll/interleukin-1 receptor
TNL	TIR-NB-LRR
TE	transposable element

Chapter 1

1 Introduction

1.1 Downy mildew resistance in grapevine

Downy mildew (DM) is one of the most widespread diseases in viticulture, affecting cultivated varieties of *Vitis vinifera*. DM resistance is a quantitative trait in the genus *Vitis*. Bellin et al. (2009) proved that a major component of DM resistance, the hypersensitive response (HR), behaves as a dominant Mendelian trait. In fact, HR-based resistance segregated with a 1:1 ratio in a population derived from the cross of the heterozygous resistant variety Bianca and the homozygous susceptible variety Chardonnay. The phenotypic variance due to the ability of mounting HR is explained by the major locus *Rpv3*, located on chromosome 18. HR starts during the initial stages of pathogen infection, restricting mycelium growth in the mesophyll, and reducing the release of sporangiophores. HR is visible to the naked eye as necrotic spots on the surface of the lower epidermis of young leaves. DM resistance is found in some accessions of North American grapevines. Native *Vitis* species have acquired a form of host resistance to *Plasmopara viticola,* a pathogen that is also native to the same regions. *P. viticola* establishes pathogenesis on leaves of these resistant accessions, by breaking into the mesophyll air spaces through stomata and by establishing biotrophy through the formation of haustoria. Pathogenesis is counteracted by an early response in resistant hosts that comes into force within 6-12 hours post infection and limits further growth of the pathogen in the next few days. Wild grapevine populations that grew in natural conditions have adapted to withstand pathogen populations that lived in the same environment. Downy mildew is caused in North America by a complex of cryptic species, with host specialization and geographic differentiation. According to Rouxel et al. (2012), what we know as *P. viticola* consists of highly heterogenous populations in which four different lineages are distinguishable by molecular markers, host range, and geographic distribution. Host plants are locked in evolutionary arms race with their pathogens, leading to the evolution of the host population under the pressure of the pathogen. Not surprising considering that, under such conditions, local

wild populations of grapevines retained selectively advantageous mutations responsible for host resistance against local variants of the pathogen, not against the entire population. Known DM resistances in North American *Vitis* operate in a race-specific manner. In this vertical resistance, host resistant grapevines behave so when challenged with some variants of the pathogen, but they are unable to mount any response when challenged with other variants. Contrary to natural conditions, all cultivated varieties of the European grapevine *V. vinifera* represent an easy prey for a recently introduced pathogen. This condition is exacerbated by the lack of evolution in a vegetatively propagated crop.

1.2 Sources of DM resistance and exploitation in conventional breeding

P. viticola has grapevines of North American species as a natural host. The vast majority of North American grapevines develop symptoms of DM under conditions conducive to the disease. The host range of *P. viticola* has expanded from wild to cultivated grapevines since the 17th century, when European settlers planted vineyards with European varieties in Central America, in Southwestern US missions and across the Atlantic coast. *P. viticola* and *V. vinifera* became sympatric in all viticulture regions, after the worldwide spread of the pathogen.

DM resistance is a trait present in several North American grapevines, in the species *Vitis labrusca, V. cinerea, V. riparia, V. rupestris, V. aestivalis, V. berlandieri, V. lincecumii* and *Muscadinia rotundifolia*. Many accessions within these species attracted the attention of viticulturists and breeders. In most cases, DM resistant vines had unpleasant fruit. Even their use as valuable donors of resistance for hydridisation breeding was abandoned soon, due to the slow progress in the improvement of their berry chemistry.

In the history of modern breeding, Barrett (1958) discerned a definite pattern of evolution in the use of DM resistant North American germplasm. Genealogy of French hybrids revealed a bottleneck in early stages of breeding, when a few valuable parents were used in a rush to save European vineyards. The combination between robust DM resistance and acceptable fruit composition naturally occurred in the North American germplasm with extreme rarity. Illuminating the restless search for valuable breeding

material is Munson's description of his life-long exploration of the Midwest (Munson 1909).

The use of resistant accessions of *V. labrusca* and *V. riparia* for breeding initiated in the mid-1800s, but it yielded little success in terms of fruit quality in backcross generations. A better compromise for level of resistance and fruit quality was obtained in descendents of resistant accessions of *V. cinerea*, *V. aestivalis* and *V. rupestris*. Due to some dilution of resistance in backcross generations, breeders also performed extensive intercrossing between descendents of different resistant lineages. This complicated the genealogy of modern resistant varieties, compared to a regular scheme of backcrossing. A few lineages of DM resistant descendents became the material of choice for breeding resistant varieties in Europe. One of these is the lineage of Villard blanc, which was particularly appreciated for its level and consistency of DM resistance, high yield, and wine attributes. Today we know that this resistance is mainly conferred by the *Rpv3* locus, likely introgressed from the species *V. rupestris*.

M. rotundifolia has a more recent history of utilization in grapevine breeding. The obtaining of fertile progeny from its hybridization with *V. vinifera*, despite the different karyotype, opened the door to the transfer of valuable traits in the background of cultivated varieties. An accession of *M. rotundifolia* has a dual resistance to DM and PM mildews, independently controlled by two genes, closely located on chr12 (Feechan et al. 2013). Similar to *Vitis* species, DM resistance in *M. rotundifolia* is based on a HR. The causal gene is *Rpv1,* and it encodes a TIR-NB-LRR protein. *Rpv1* is a host resistance gene, but it is claimed to confer broad spectrum resistance, based on the observed effectivity against all isolates tested so far.

Other sources of DM resistance were identified in Asian wild species. Accessions of the species *V. amurensis*, that is also cold hardy, were the material of choice for Chinese breeding. Resistant vines are able to mount a response against *P. viticola* similar to that mounted by resistant North American grapevines. These characteristics attracted breeders' attention, who started to cross *V. amurensis* with *V. vinifera* after World War II. Venuti et al. (2013) genetically mapped a major locus on chr14, responsible for DM resistance, called *Rpv12,* that explained 79 % of the phenotypic variation in descendents of the hybrid grapes 'Zarya severa' and 'Michurinets'. The *Rpv12* locus is

populated by a cluster of NB-LRR genes belonging to the coiled-coil class. Schwander et al. (2012) identified an independent source of DM resistance from the same species *V. amurensis.* The major locus responsible for DM resistance, called *Rpv10,* was genetically mapped on chr9 in descendents of the hybrid grape 'Severnyi'. The *Rpv10* locus is also populated by a cluster of NB-LRR genes belonging to the coiled-coil class. Asian grapevines have evolved into resistant vines in the absence of pathogen pressure, which was not present in Asia until 100 years ago. The historical presence in those regions of other species of *Plasmopara*, such as *P. cissii* and *P. amurensis*, suggested the hypothesis that broad spectrum resistance in *V. amurensis* against those potential pathogens may extend to the recently introduced *P. viticola*.

1.3 Relevance of the Villard Blanc lineage in modern breeding

Hybridization breeding for DM resistance started in North America approximately 150 years ago and proceeded with French breeding. The initial choice of resistant parents was restricted to those with palatable fruit flavors. Several rounds of phenotypic selection led the breeders to the unintentional selection of a few resistant haplotypes at *Rpv3* locus, creating a few lineages of resistant varieties. Genealogy and haplotype analysis with *Rpv3* markers revealed that seven haplotypes were conserved in five groups of DM resistant varieties and were traced back to the founders of these lineages (Di Gaspero et al. 2012). The most frequent haplotype is Rpv3$^{299\text{-}279}$, which is present in 106 accessions that are commonly descended from Villard blanc. In independent studies, Rpv3$^{299\text{-}279}$ has been genetically associated with DM resistance in progeny of Bianca and Regent. The ancestry of this haplotype was traced three generations back from Bianca to the Seibel selections created in the 1800s, far back to the oldest known living progenitor of this lineage, which is 'Seibel 4614' (Figure 1). 'Seibel 4614' gave rise to five recorded offspring, but 'Seibel 6468' is the only one that disseminated the resistant haplotype *Rpv3*$^{299\text{-}279}$. Villard blanc alone, an offspring of 'Seibel 6468', became the 3rd most widely planted white variety in France in the 1960's. All this evidence lends support to the hypothesis that the resistant haplotype *Rpv3*$^{299\text{-}279}$ has been circulating in Europe for more than 150 years. It conferred resistance to many varieties cultivated in several European countries for decades, some of which covering a significant acreage.

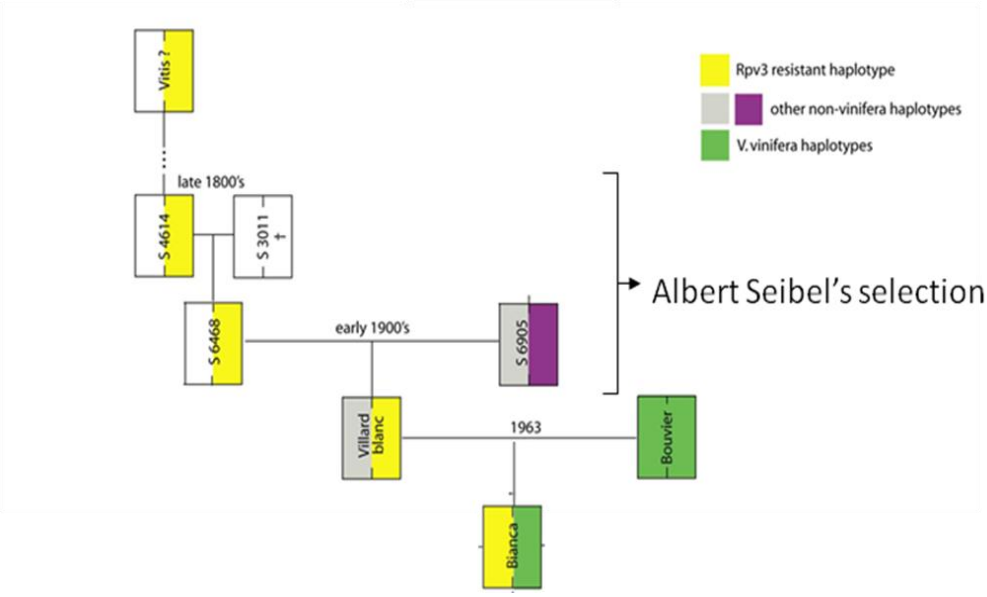


Figure 1. Pedigree of Bianca. The diagrams represent *Rpv3* haplotypes in diploid individuals.

1.4 Characterization of the Rpv3 locus

The HR component of DM resistance controlled by the *Rpv3* locus was genetically mapped in the variety Bianca to chr18 into an interval delimited by markers UDV305 and VMC7f2 (Bellin et al 2009). Before the start of my PhD, a total of 4,221 offspring of Bianca were used to restrict the genetic interval of the locus. Individuals that inherited recombinant chromosomes from Bianca in the *Rpv3* locus were phenotyped for DM resistance.

1.5 Rpv3 sequencing: state of art at the beginning of my PhD

At the beginning of my PhD, several resources were available: (i) 28 informative recombinants of Bianca in the *Rpv3* locus, (ii) a high quality reference genome of *V. vinifera* "PN40024", (iii) a partial physical map of the *Rpv3* resistant haplotype consisting of 7 large-insert BAC clones, (iv) next-generation sequencing (NGS) reads of each BAC clone, (i) CLC *de novo* assembly of each clone (Figure 2).

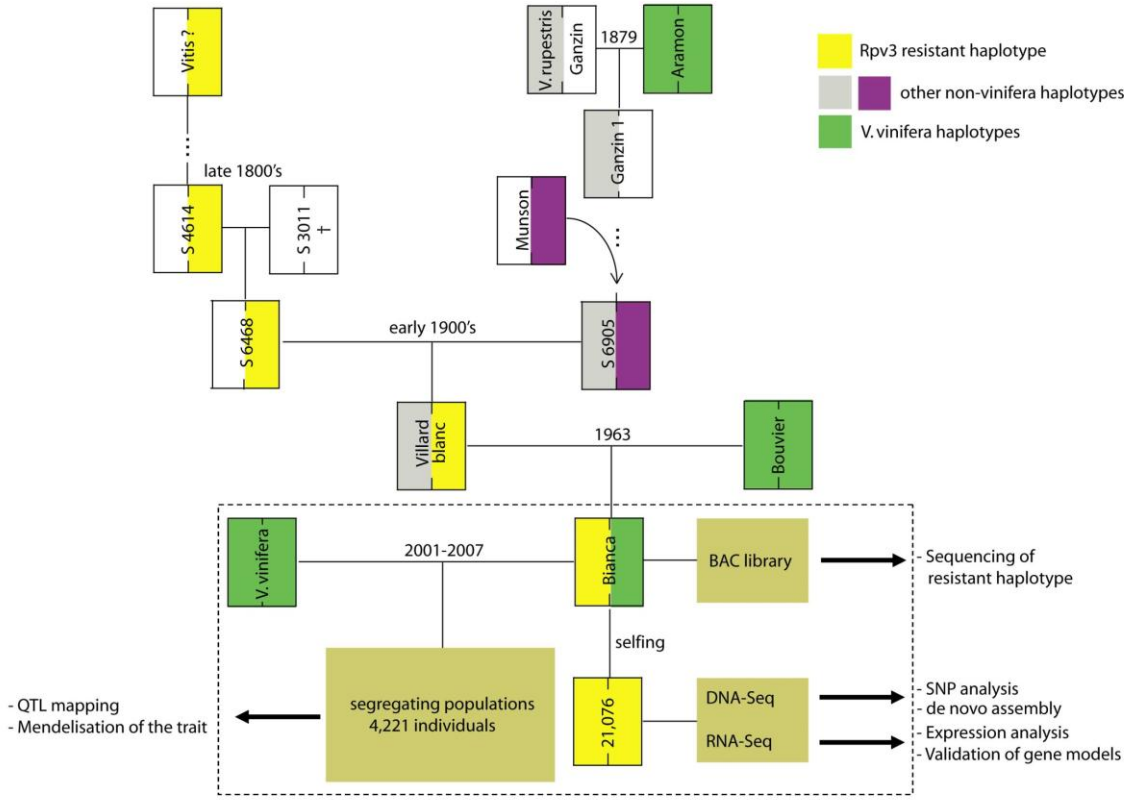


Figure 2. Materials and resources for reconstructing the sequence of the *Rpv3* resistance haplotype

The state of art at the beginning of my PhD was the following:

1. The BAC library of the cultivar Bianca was fully exploited to cover the resistance haplotype with large-inserts. Seven clones harbored inserts of the resistance haplotype and covered the *Rpv3* locus (indicated with yellow line in Figure 3). The genetic markers developed on these BAC and the availability of 28 recombinants allowed to restrict the locus within an interval spanned by two BAC contigs. A gap of unknown size was present between these two BAC contigs. No other BAC clone was recovered from the library using markers designed on the sequences bordering the gap.

2. Sequencing and *de novo* assembly of BAC clones produced two pseudomolecules corresponding to the BAC contigs spanning the locus, that were tentatively ordered using the reference genome PN40024 as a guide. In spite of the availability of a high-quality reference genome for *V. vinifera*, this tool proved of limited utility to assist in

the assembly of the resistance haplotype, due to the high interspecific DNA diversity. The BAC contig 46G16-55E09 contains a large transposable element (TE) that is not present in the allelic region of the haplotype of PN40024, and a gene that is duplicated in the *Rpv3* locus. No unique regions were found to be shared between the BAC contig and the PN40024 reference in order to unambiguously orient the BAC contig by synteny. BAC contig 46G16-55E09 was tentatively oriented by using the information of three recombinants and three SNP markers. One SNP marker was developed on one BAC-end sequence of BAC_46G16, the others within BAC_55E09. SNP 55E09_6a and SNP_55E09_2b were scored in amplicons that were intended to be amplified in recombinant individuals from the target region corresponding to BAC_55E09, but they may have actually originated from a highly similar duplicated region in the adjacent BAC contig. No other unique region was found in order to design more specific markers. Uncertainty in the orientation of this BAC contig caused uncertainty in the definition of the upper genetic border of the locus.

3. Candidate genes and TEs were identified in the sequence of BAC contigs. This provided clues on the nature of the region, characterized by the abundance of high-copy DNA and the presence of duplicated genes of the NB-LRR gene family. One of these gene copies was partially reconstructed, because it was only partially covered by the BAC-end of one BAC clone bordering the gap. The presence of a gap of unknown size within a NB-LRR gene cluster raised concern about the number of other gene copies possibly present in the resistance haplotype.

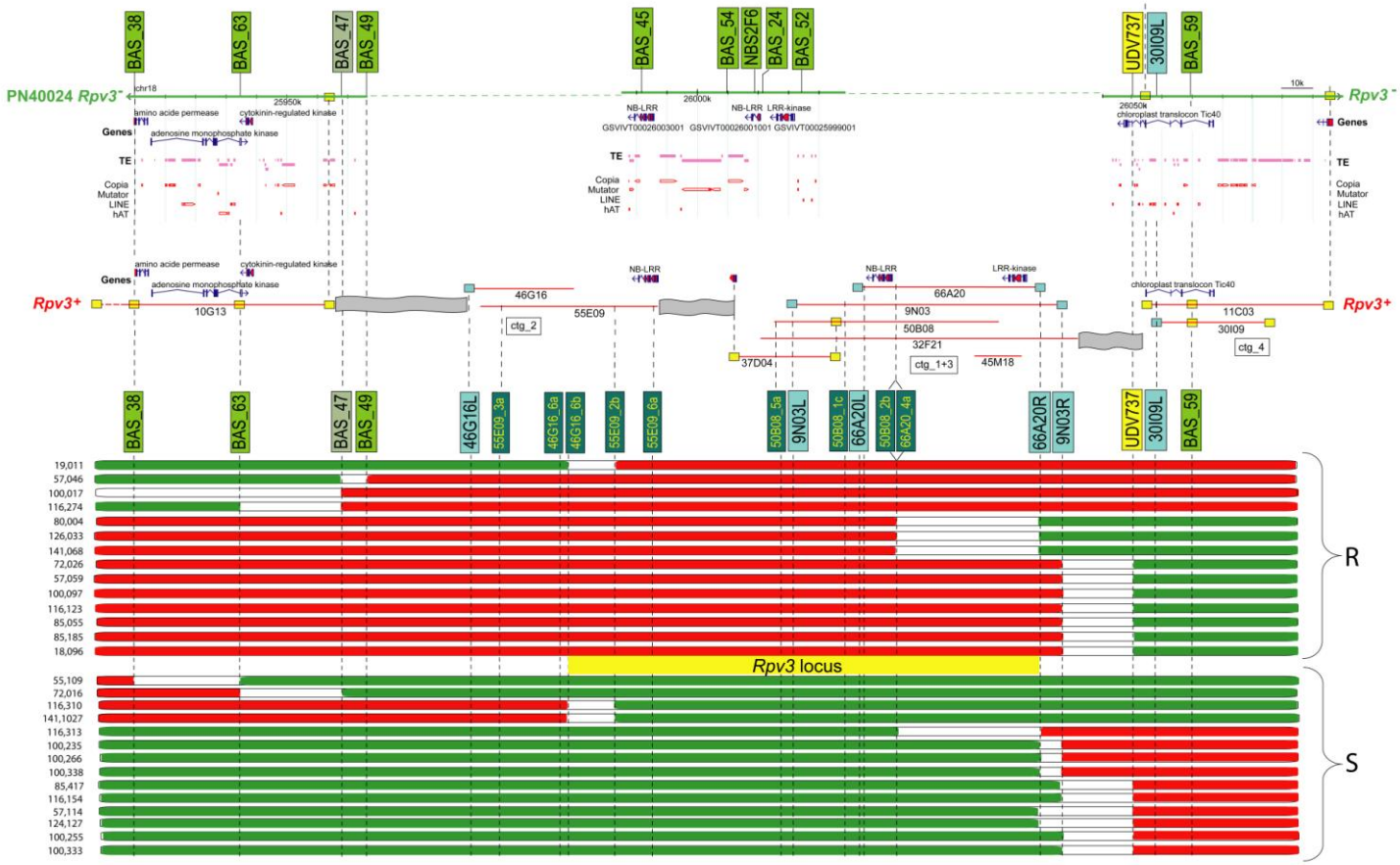


Figure 3. State of art of the reconstruction of the *Rpv3* locus at the beginning of my PhD. Physical map of the *Rpv3+* allele is shown in the middle panel. Physical contigs of BAC clones (red horizontal lines) were genetically ordered and oriented based on the relative position of the markers. Clones were assembled in sequence contigs. The physical and genetic location of SSR markers (in yellow background), SNP markers developed on the reference genome sequence (in green background) and SNP markers developed on the *Rpv3+* allele sequence (in cyan background for BAC end markers and olive green background for internal markers) are interconnected by vertical dotted lines. Predicted gene model are drawn in red/blue. In the lower part of the figure, diagram of the haplotypes of 28 informative recombinants that have a crossing-over in this region. The red section indicates the portion of the resistant homologue (*Rpv3+*), the green section indicates the portion of the susceptible homologue (*Rpv3-*) which recombined in the germ cells of the *Rpv3+/Rpv3-* parent. The haplotypes are grouped based on the phenotype of the corresponding seedling (resistant, R; susceptible, S). The genetic interval of the locus is indicated by the horizontal yellow bar. In the upper part of the panel, alignment with the PN40024 *Rpv3-* allele. Sequence gaps are indicated by grey waves.

1.6 Objectives of this thesis

The main purposes of my PhD thesis were:

(i) to complete the sequence of the downy mildew resistance locus *Rpv3*. To this end, using Illumina technology, we planned to sequence an individual homozygous for the resistance haplotype at the *Rpv3* locus and combine whole-genome assembly with the assembly of seven BAC clones that partially span the region;

(ii) to obtain the full-length coding sequence of candidate genes in the resistance haplotype. We planned to deep sequence messenger RNA extracted from infected leaves of an individual homozygous for the resistance haplotype at *Rpv3* locus in order to monitor all transcripts during the onset of HR;

(iii) to clone the candidate genes from the resistance haplotype into an expression vector. In order to proceed with a functional analysis of candidate genes, we planned to prepare constructs for *Agrobacterium*-mediated transformation of embryogenic calli of susceptible varieties;

(iv) to analyse DNA variation between the resistance haplotype – introgressed from an unknown species – and haplotypes in susceptible varieties of *V. vinifera*. To this end, we had Illumina reads available from resequencing of 50 varieties.

2 Materials and Methods

2.1 Plant material

The variety Bianca was forced to self by wrapping inflorescences in paper bags one week prior to anthesis. Seeds were extracted from rotting berries in mid October. Flesh was manually removed from seeds. Seeds were cleansed in 1.5 % hydrogen peroxide for 24 h and rinsed in tap water. Seeds endured cold stratification for 3 months. Seed germination was conducted in a glasshouse at 25°C. Seedlings were coded with the cross code “21” followed by the comma-separated ID number for each

individual. "21,076$^{Rpv3+/Rpv3+}$" is a homozygous individual for the resistant haplotype at the *Rpv3* locus used for further analysis.

2.2 DNA analysis

Young leaves of 21,076$^{Rpv3+/Rpv3+}$ were collected and DNA extraction was done with different protocols. Three libraries were prepared for Illumina sequencing.

2.2.1 Paired-end library

For the paired-end library, DNA was extracted with a CTAB protocol (Doyle and Doyle, 1990) obtaining a concentration of 148,2 ng/µl (The Qubit® 2.0 Fluorometer). DNA was size selected by gel electrophoresis by cutting different bands in the interval of 600-1300 bp. A paired-end library was constructed following the Truseq DNA PE protocol and run on an Illumina HiSeq2000.

2.2.2 Mate-pair library

For the mate-pair library, DNA was extracted with PowerPlant® Pro DNA Isolation Kit MO-Bio Kit with a final concentration of 563 ng/µl (The Qubit® 2.0 Fluorometer). DNA was size selected by gel electrophoresis in the interval of 350-3000 bp. A mate-pair library was constructed using Nextera Mate Pair GEL FREE protocol and run on an Illumina HiSeq2000.

2.2.3 MiSeq libraries

The same DNA used for the mate-pair library was also used for preparing a library of different insert size for MiSeq sequencing. DNA was size selected by gel electrophoresis by cutting two bands at approximately 500 bp and 600 bp. Two libraries were constructed following the Nextera DNA Sample Preparation protocol and run on an Illumina HiSeq2000.

2.2.4 Read filtering and trimming

Mitochondrial and chloroplastic reads were filtered out and PCR duplicate reads were removed with ERNE-FILTER (erne.sourceforge.net). Illumina adapters were removed using cutadapt (http://code.google.com/p/cutadapt/). Reads were trimmed for quality with ERNE-FILTER. Reads longer than 50 bp were retained for further analyses. Reads

were aligned to the grapevine genome reference with GATK. Aligned paired-reads were used to estimate the distribution of insert size in all libraries. Data were extracted from BAM file using Picard tools.

2.3 De novo assembly

A *de novo* assembly was generated with ALLPATHS (Butler et al. 2008) with default settings and the option HAPLOIDIFY set at TRUE. Short-insert paired-end reads, long-insert mate-pairs and MiSeq reads were used as an input to the assembler.

2.4 Gene prediction

Genes encoded by the assembled sequence of the *Rpv3* locus were identified by BlastN search against the 196,000 *V. vinifera* ESTs available at the National Center for Biotechnology Information (NCBI; http://www.ncbi.nlm.nih.gov). Similarities with known proteins were identified with BlastX (Altschul et al., 1990). *Ab initio* gene prediction and identification of exon–intron structure were carried out with FGENESH (http://www.softberry.com) .

2.5 Structural variation

The *de novo* assembly of 21,076$^{Rpv3+/Rpv3+}$ and the *V. vinifera* reference genome were compared using GEvo (https://genomevolution.org/CoGe/GEvo.pl) with the alignment algorithm (B)LastZ and the following parameters: word size 8, gap start penality 400, gap extend penality 30.

2.6 Read alignment and SNP-calling

Alignment of Illumina reads against a reference genome was carried out with BWA. For different kinds of analyses, we used either the *V. vinifera* reference genome (strain PN40024, GenBank Assembly ID GCA_000003745.2) or the *de novo* assembly of 21,076$^{Rpv3+/Rpv3+}$. Local realignment around indels was performed with the RealignerTargetCreator and IndelRealigner tools of the GATK package version 2.1-13 (McKenna et al. 2010). SNPs were called using default parameters of Unified Genotyper of GATK. The set of raw SNPs was filtered using the following parameters: minimum coverage >0.5 fold the average coverage, maximum coverage <3 fold the

average coverage, GATK Phred scaled quality score (QUAL) > 100, GATK Strand Bias (SB) < 0. SNPs called against the *V. vinifera* reference genome were also filtered out if they laid in regions of high-copy DNA, as identified by ReAS, in transposable elements and in the intervals spanning microsatellite DNA.

2.7 Gene expression

Cuttings were taken from dormant wood of 21,076$^{Rpv3+/Rpv3+}$. The basipetal end of the cuttings was treated with NAA and placed in hot beds for rooting. Rooted cuttings were transferred in potted soil and grown in a glasshouse. The 4th and the 5th leaves beneath the shoot apex were detached from shoots at the stage of 12 leaves. Discs of 1 cm diameter were cut with a cork borer from leaves of 21,076$^{Rpv3+/Rpv3+}$. In order to induce the expression of genes involved in the early response to downy mildew infection, the discs were placed on wet paper in Petri dishes and sprayed with a suspension of an avrRpv3+ isolate of *Plasmopara viticola* at 150,000 sporangia ml^{-1}. Discs were incubated in a growth chamber at 20°C and collected after 12, 24, 48, 96h post-inoculation, respectively, frozen in liquid nitrogen and stored at -80°C. All the discs collected at different time points were pooled for mRNA extraction.

2.7.1 RNA extraction and sequencing

Total RNA was extracted using the Spectrum™ Plant Total RNA Kit, according to the protocol B of the manufacturer's instructions (Sigma-Aldrich, St. Louis, MO). Quality of RNA was assessed with a 2100 Agilent Bioanalyzer and the quantity was measured with The Qubit® 2.0 Fluorometer. An amount of 2 µg of total RNA was used for library preparation following the Illumina library preparation protocol TrueSeq v2.0. RNA was fragmented into 500 bp fragments and mRNA was purified twice using polyT beads. A 100-bp paired-end read run was performed on a Illumina HiSeq2000.

2.7.2 RNA-Seq alignment and estimation of transcript abundances

Illumina adapters were removed using cutadapt (http://code.google.com/p/cutadapt/). Reads were trimmed for quality with ERNE-FILTER. Alignment was performed with TopHat 2.0.5 (Trapnell C. et al.,2012) with the following parameters: mate-inner-dist 500, min-anchor-length 20, max-multihits 50.

As previously described for DNA-Seq, alignment of Illumina reads against a reference genome was carried using either the *V. vinifera* reference genome (strain PN40024, GenBank Assembly ID GCA_000003745.2) or the *de novo* assembly of 21,076$^{Rpv3+/Rpv3+}$. Quantification of gene and transcript expression was estimated with Cufflinks. Using the *V. vinifera* genome as a reference, transcript abundance refers to the set of curated genes in the CRIBI annotation. Using the *de novo* assembly of 21,076$^{Rpv3+/Rpv3+}$, expression levels refer to the transcripts identified by Cufflinks.

2.8 BAC library

Paired-end reads of nine clones selected from a BAC library of the variety Bianca were available from previous work (Copetti D., PhD dissertation). They partially covered the *Rpv3* locus and belonged to the resistance haplotype. Each BAC was re-assembled with ABySS (Simpson et al., 2009). The assembly was performed with these parameters: k=75, b=1000, p=0.95, s=500.

2.9 Gene cloning

2.9.1 Primer design

Different primer combinations were designed for cloning and for sequencing the cloned fragments. For TA cloning, primers were designed on 5' and 3' UTR regions of each gene. For sequencing the entire genes, we used an approach of primer walking. For Gateway cloning, primers were synthesized with 3'-terminal *attB* sites. All primers were designed with Primer3 (v.0.4.0). Primer sequences are reported in Table 1.

Table 1. Primer sequences per gene cloning and full-length sequencing

Name	Primer forward	Primer reverse
TNL2	ATCGGCACTGCCAAAGTAAT	TGCATTCCTTCTGCTTCTCC
TNL3	TTTGGTCCCAGCACCTGTAT	TTGTTGTGAGACTTGGGTTCA
LRR-k	GTGGCCAATACGCATAAAGC	AGTTTCCATTGTGCCCATTC
TNL_internal1	CTTGATGCCAATTGCTGAGA	
TNL_internal2	TCGAAGCGAGACATTTTTCA	
TNL_internal3	CGCTTGGAAGACTCTTGAGC	
TNL_internal4	CTGCAAAGGCTTGTGTGGTA	
TNL_internal5	TTCCTCAATTTTCTTATGAATGG	
TNL_internal6	AGAGGCTCATTCCTGCCATA	
TNL_internal7	CAAAAGCCATTGCAGCAGTA	
TNL_internal8	TCCAACTGATTTCTCCGTTCTT	
LRR-k_internal1	CACATGAGACCCCGAAAAAT	
LRR-k_internal2	ATGAGTCCCTGCAAAACAGC	
LRR-k_internal3	CTGTGCCTCCACTTCCAATC	
LRR-k_internal4	TCAAAACTGTTCAGCGGAGA	
LRR-k_internal5	GCTGTTAGATTCGGTTTTGATG	
LRR-k_internal6	CGGGAACCAAACAGAGGATA	
LRR-k_internal7	AGCGGGAATTCAGAACTGCT	
TNL2_*attB*	*GGGGACCACTTTGTACAAGAAAGCTGGG*TATCGGCACTGCCAAAGTAAT	*GGGGACAAGTTTGTACAAAAAAGCAGGC*TTGCATTCCTTCTGCTTCTCC
TNL3_*attB*	*GGGGACCACTTTGTACAAGAAAGCTGGG*TTTTGGTCCCAGCACCTGTAT	*GGGGACAAGTTTGTACAAAAAAGCAGGC*TTTGTTGTGAGACTTGGGTTCA
LRR-k_*attB*	*GGGGACCACTTTGTACAAGAAAGCTGGG*TGTGGCCAATACGCATAAAGC	*GGGGACAAGTTTGTACAAAAAAGCAGGC*TAGTTTCCATTGTGCCCATTC
LA/LB	GCAGTTCCCTACTCTCGC	CATCAGAGATTTTGAGACAC

2.9.2 TA cloning

PCR amplifications were performed in a 10-µl volume using KAPA HiFi HotStart Ready Mix (Kapa Biosystems) and 0.2 µM of each primer, and run in Geneamp 9700 PCR system (Applied Biosystems, Foster City, CA) under the following conditions: 95°C for 1 minute, 10 cycles of 98°C for 20 seconds, 67°C for 15 seconds, 72°C for 4 minutes (-0,5°C each cycle), 25 cycles of 98°C for 20 seconds, 62°C for 15 seconds, 72°C for 4 minutes and 72°C for 10 minutes. The PCR products were separated on a 1 % agarose gel by electrophoresis. The bands were excised and collected in 1.5 ml tubes. Purification from agarose was performed with the reagents provided by the pCR-XL-TOPO cloning kit (Invitrogen). Samples were cloned in pCR-XL-TOPO vector and transformed into competent *E. coli*, following the manufacturer's instructions.

2.9.3 Screening and sequencing

2.9.3.1 Colony PCR and extraction of plasmid DNA

Colonies were picked up with a toothpick and dipped into a tube with PCR mix. Colony PCR was performed in a 10-µl volume using 0.2 µM of M13 primers and HotMaster Taq DNA polymerase (5PRIME) according to the manufacturer's instruction, under the following condition: 94°C for 10 minutes, 14 cycles of 94°C for 30 seconds, 62°C for 30 seconds, 65°C for 3 minutes, and 30 cycles of 94°C for 30 seconds, 55°C for 30 seconds, 65°C for 3 minutes and 65°C for 10 minutes. PCR products were analyzed on 1 % agarose gel by electrophoresis to check the length of the insert. For extraction of plasmid DNA, positive colonies were grown overnight in liquid 17 mM KH_2PO_4, 72 mM K_2HPO_4, 10.8 g/L tryptone, 21.6 g/L yeast extract, 3.6 ml/L glycerol medium supplemented with 50 µg/mL ampicillin. Two ml of bacterial culture were centrifuged and liquid medium was discarded. Lysis of bacterial cells was performed at 95°C for 3 min in 50 mM Tris (pH=7.5) and 62.5 mM EDTA supplemented with 0.20 mg of lysozyme. Crude lysate was incubated for 5 minutes in wet ice. Sedimentation of cell debris was obtained by centrifugation. Plasmid DNA in the supernatant was precipitated with isopropanol, rinsed in 70 % ethanol and resuspended in 10 mM Tris HCl 0.1 mM EDTA supplemented with 1 µg of RNase. Plasmid DNA was precipitated

again with ethanol and ammonium acetate, rinsed in 70% ethanol and resuspended in water.

2.9.3.2 Sequencing

Using M13 and internal primers with a BigDye Terminator Secquencing Kit, sequencing reaction was performed under the following conditions: 35 cycles of 96°C for 10 seconds, 50°C for 5 seconds, 60°C for 4 minutes. Sequencing reactions were run on an ABIPrism 3730xl DNA analyzer (Applied Biosystems). Reads were aligned against gene sequences using CLC Genomic Benchwork. Aligned pherograms were visually inspected to exclude the presence of variant positions. Based on 100 % identity of the sequenced insert with the target gene sequence, one colony was selected per target gene.

2.9.4 Gateway cloning

Plasmid DNA from positive colonies was used as a template for Gateway cloning. PCR was performed to produce *attB*-gene sequences. PCR was conducted in a 10-µl volume using KAPA HiFi HotStart Ready Mix (Kapa Biosystems) and 0.2 µM of *attB*-primers and pCR-XL-TOPO plasmid DNA extracted from the selected colonies. PCR conditions were the same as described in 1.9.2.

BP recombination was carried out using 400 ng of *attB*-gene PCR products, 150 ng of donor vector pDONR207 (Figure 4, panel A) and according to the manufacturer's instruction (Gateway technology, Invitrogen). DH5α™ *E. coli* strain was transformed with the entry vector and grown on LB medium supplemented with 20 µg/mL gentamicin. Colony PCR was performed using 0.2 µM of LA and LB primers flanking the cloning site as described in 1.9.3. PCR products were analyzed on 1 % agarose gel by electrophoresis to check the length of the insert. For extraction of plasmid DNA, positive colonies were grown overnight in liquid LB medium supplemented with 20 µg/mL gentamicin. Two ml of bacterial culture were centrifuged and liquid medium was discarded. Bacterial pellet was resuspended in 50 mM Tris (pH=8), 10mM EDTA, 10 µg/ml RNaseA buffer. Alkaline lysis was performed with a 200 mM NaOH, 1 % SDS solution. Circular DNA was renatured and SDS precipitated with 3M potassium acetate (pH=5.5). Plasmid DNA was precipitated with isopropanol, rinsed in 70 % ethanol and

resuspended in water. Sequencing reaction was performed with LA and LB primers flanking the cloning site and with internal primers as described in 1.9.3. Sanger reads were aligned against gene sequences using CLC Genomic Benchwork. Aligned pherograms were visually inspected to exclude nucleotide misincorportation. Based on 100 % identity of the sequenced insert with the target gene sequence, one colony was selected per target gene. In order to obtain high-quality plasmid DNA, the selected colony was grown overnight in liquid LB medium supplemented with 20 µg/mL gentamicin and plasmid DNA was extracted with the GenElute™ Plasmid Miniprep Kit (Sigma).

The insert was transferred by recombination from the entry clone into the destination vector pK2GW7 (Figure 4, panel B). LR reaction was performed according to the manufacturer's instruction (Gateway technology, Invitrogen). DH5α™ *E. coli* strain was transformed with the expression clone and grown on LB medium with 100 µg/mL of spectinomycin. Three colonies per target genes were grown overnight in liquid LB medium supplemented 100 µg/mL of spectinomycin. Plasmid DNA was extracted by alkaline lysis as described for the entry vector. Sequencing reaction was performed with LA-LB primers flanking the cloning site and carried out as described for the entry vector. Reads were aligned against gene sequences using CLC Genomic Benchwork in order to assess the full-length transfer of the insert.

One colony per target gene was delivered to a facility of grapevine genetic engineering for *Agrobacterium*-mediated transformation of embryogenic calli.

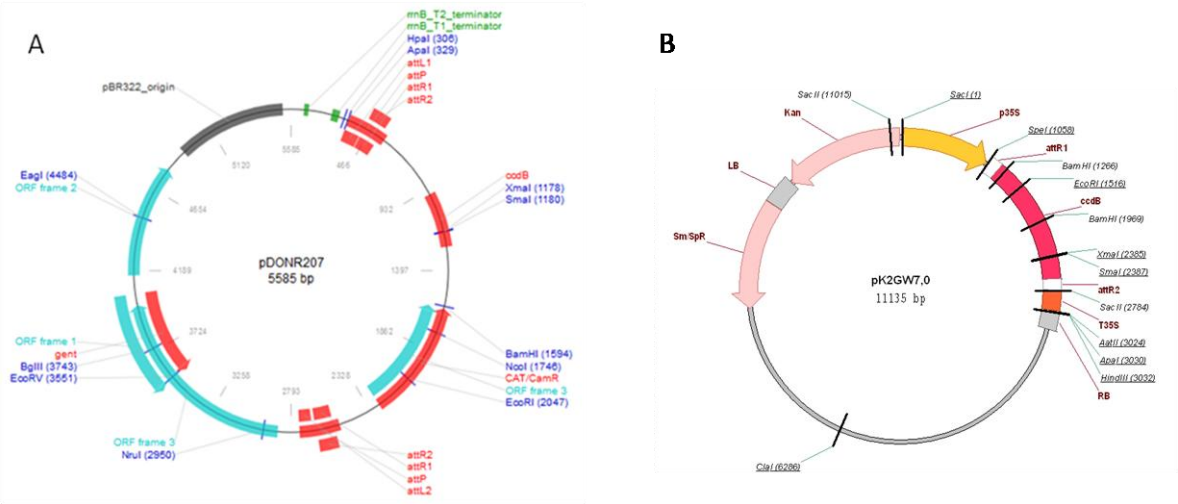


Figure 4. Vector map. In panel A, pDONR207. This vector consists of: multiple cloning site with ccdB gene, a lethal gene for selection of recombinant colonies and attP nucleotide to allow recombination, T1 and T2 terminator, origin of replication, resistance gene to gentamicin antibiotic. In panel B, destination vector pK2G7W. This vector has a 35S Cauliflower Mosaic Virus (CaMV) promoter/terminator and kanamycin resistance cassette and ccdB gene.

3 Results

3.1 Illumina Sequencing of 21,076 Rpv3+/Rpv3+

3.1.1 Selection of the homozygous individual 21,076 Rpv3+/Rpv3+

Genomic DNA was extracted from seedlings of an S1 population of Bianca at the stage of three true leaves. We identified a resistant seedling homozygous at the *Rpv3* locus using the molecular markers UDV305 and UDV737, bordering the *Rpv3* locus. The homozygous individual was coded "21,076$^{Rpv3+/Rpv3+}$".

3.1.2 DNA sequencing

3.1.2.1 Filtering and alignment of Illumina reads

Three different DNA libraries were prepared with the aim of obtaining fragments with different insert size and reads with different length: (i) short insert paired-end reads of 100 bp, (ii) long insert mate-pair reads of 100 bp, (iii) short insert paired-end reads of 300 bp. The mode of insert size distribution was estimated following read alignment against the reference genome. The modal class was 492 bp for library (i) with an additional peak at 203 bp; 2,249 bp for library (ii); and 300 bp with an additional minor peak at 380 bp for library (iii) (Figure 5).

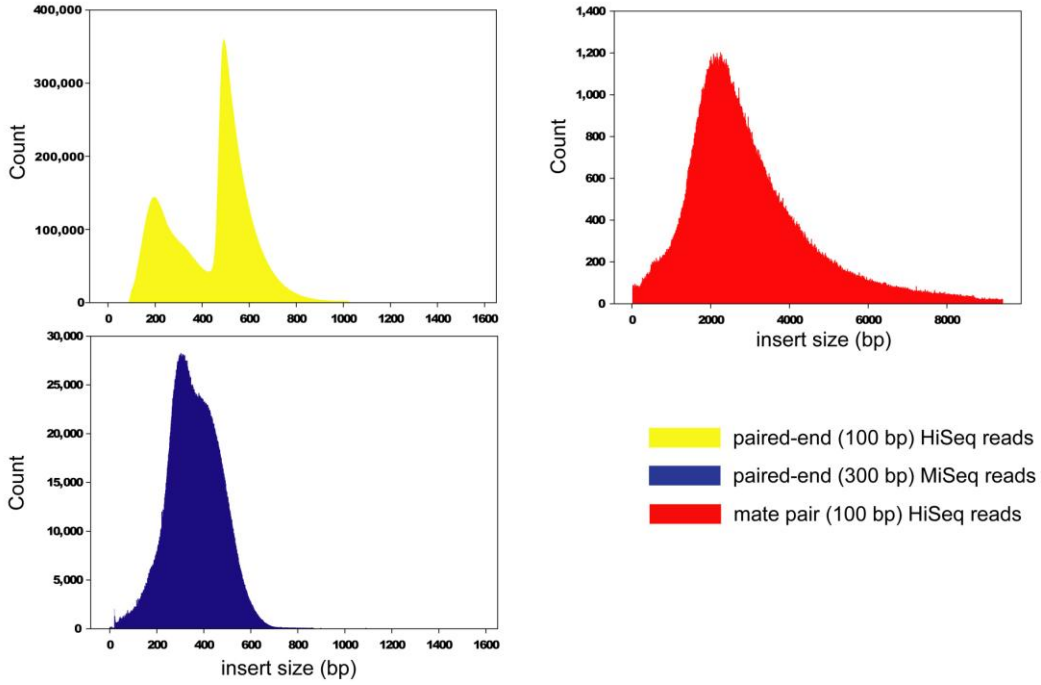


Figure 5. Plots of insert size. NGS libraries were prepared from genomic DNA of 21,076$^{Rpv3+/Rpv3+}$. Paired-end reads and mate-pairs were aligned against PN40024 genome reference. Data were extracted from BAM files using Picard tools.

Raw FASTQ data were processed for adapter removal, quality trimming e filtering for contaminants and duplicates. The metrics are reported in Table 2 .

Table 2. Metrics of Illumina reads before and after trimming and filtering.

Library	Raw reads	Bp before filtering and trimming	Trimmed and filtered reads	Bp after trimming and filtering	Insert size (mode,
Paired-end	353,335,622	35,333,562,200	298,891,888	28,573,893,485	492
Mate pair	151,742,094	15,174,209,400	35,747,954	2,765,895,809	2,249
“MiSeq”	19,286,682	5,786,004,600	16,311,690	4,442,271,215	300

3.1.3 RNA sequencing

3.1.3.1 Filtering and alignment of Illumina reads

Another library was prepared from leaf RNA extracted from 21,076$^{Rpv3+/Rpv3+}$ after induction of HR by *P. viticola*. Raw FASTQ data were processed for adapter removal, quality trimming and filtering for contaminants. The metrics are reported in Table 3.

Table 3. Metrics of Illumina reads before and after trimming and filtering.

	Raw reads	Bp before filtering and trimming	Trimmed and filtered reads	Bp after trimming and filtering
RNA library	36,234,334	35,894,776	35,701,128	27,453,975

3.1.3.2 Reads mapping

Processed reads were mapped to two different reference sequences: (i) the *de novo* assembly of 21,076$^{Rpv3+/Rpv3+}$, (ii) the *V. vinifera* reference genome PN40024. Total number of reads mapped are reported in Table 4.

Table 4. TopHat metrics after RNA read mapping to different references.

RNA reads / reference genome	Paired	Single	TOT mapped	Unmapped
21,076$^{Rpv3+/Rpv3+}$ / PN40024	20,116,212	7,337,763	27,453,975	10,875,348
21,076$^{Rpv3+/Rpv3+}$ / 21,076$^{Rpv3+/Rpv3+}$	17,070,574	6,581,017	23,651,591	12,685,633

3.2 De novo assembly of 21,076 $^{Rpv3+/Rpv3+}$ genome

The 21,076$^{Rpv3+/Rpv3+}$ *de novo* genome assembly is based on ALLPATHS assembly. ALLPATHS used approximately 200 million original fragment reads and 100 million jumping reads. At K=25 scale, 40.0 % of the genome was estimated to be repetitive. Details of the number of reads per library and the fraction of assembled reads are given in Table 5.

Table 5. Statistics of the libraries used by ALLPATHS

Fragment type	Library	N_reads	%_used	N_pairs
Fragment read	MiSeq 500-bp insert	6,773,758	80.9	3,047,157
Fragment read	MiSeq 600-bp insert	9,493,210	78.9	3,798,966
Fragment read	HiSeq paired-end	160,000,000	49.3	24,258,274
Fragment read	HiSeq mate-pair	24,707,568	60	5,129,675
Fragment read	total	200,974,536	53.1	36,234,072
Jump read	HiSeq mate-pair	100,744,694	45	17,664,348

Legend:

n_reads: number of reads in input

%_used: % of reads assembled

n_pairs: number of valid pairs assembled

ALLPATHS produced scaffolds for a total genome length of approximately 450 Mbp, pretty close to the estimated reference genome size. The total number of contigs is 70,355 and the total length of the contigs amounted to 322 Mbp. The total number of scaffolds is 9,821. Half of the genome is contained in scaffolds that are longer than 163 kb and in contigs longer than 8 kb. (Table 6)

Table 6. Metrics of *de novo* assembly.

Statistics	
Estimated genome size	448.7 Mbp
Number of contigs	70,355
Number of contigs per Mbp	156,8
Total contig length	322.7 Mbp
Number of scaffolds	9,821
Number of scaffolds per Mb	21,89
N50 contig size in kb	8,0
N50 scaffold size in kb	163,0

The heterozygosity of the genome had little impact on scaffold length. Paired-end reads were aligned against the 21,076$^{Rpv3+/Rpv3+}$ scaffolds. SNPs were called with GATK

and heterozygous SNPs were plotted against scaffolds. Figure 6 shows the 30 and the 100 largest scaffolds of the 21,076$^{Rpv3+/Rpv3+}$ assembly and the heterozygosity of 21,076$^{Rpv3+/Rpv3+}$ and PN40024 along those regions. Half of the 30 largest scaffolds were assembled from heterozygous regions of 21,076$^{Rpv3+/Rpv3+}$. As many as 42 of the 100 largest scaffolds were assembled from heterozygous regions of 21,076$^{Rpv3+/Rpv3+}$. The nearly homozygous genome of PN40024 is heterozygous across only two of these 100 scaffolds, corresponding to 2.4 % of the genome size sampled by the 100 largest scaffolds.

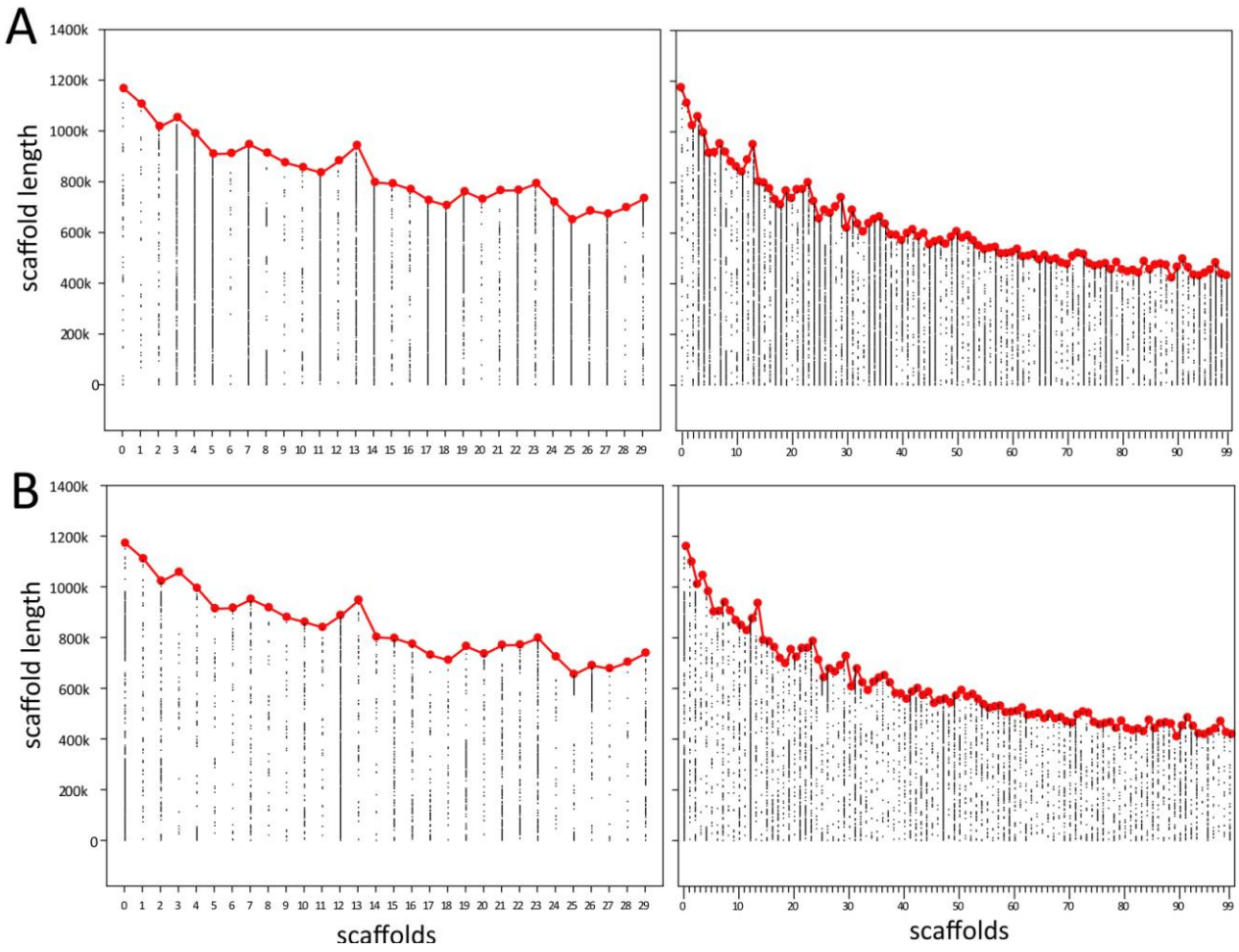


Figure 6. Scaffold size and heterozygosity of the 21,076$^{Rpv3+/Rpv3+}$ genome. Heterozygous SNPs are plotted as dots. Continuous lines of dots indicate heterozygous regions in 21,076$^{Rpv3+/Rpv3+}$ (panel A) and in PN40024 (Panel B) for the 30 (left) and the 100 (right) largest scaffolds. Isolated dots indicated false SNPs that filtering parameters failed to sort out. Red dots indicate scaffold size in kbp.

In order to assess the accuracy of the de novo assembly, we more closely inspected five cases among the 30 largest scaffolds (Figure 7): (i) scaffold_6 assembled from a

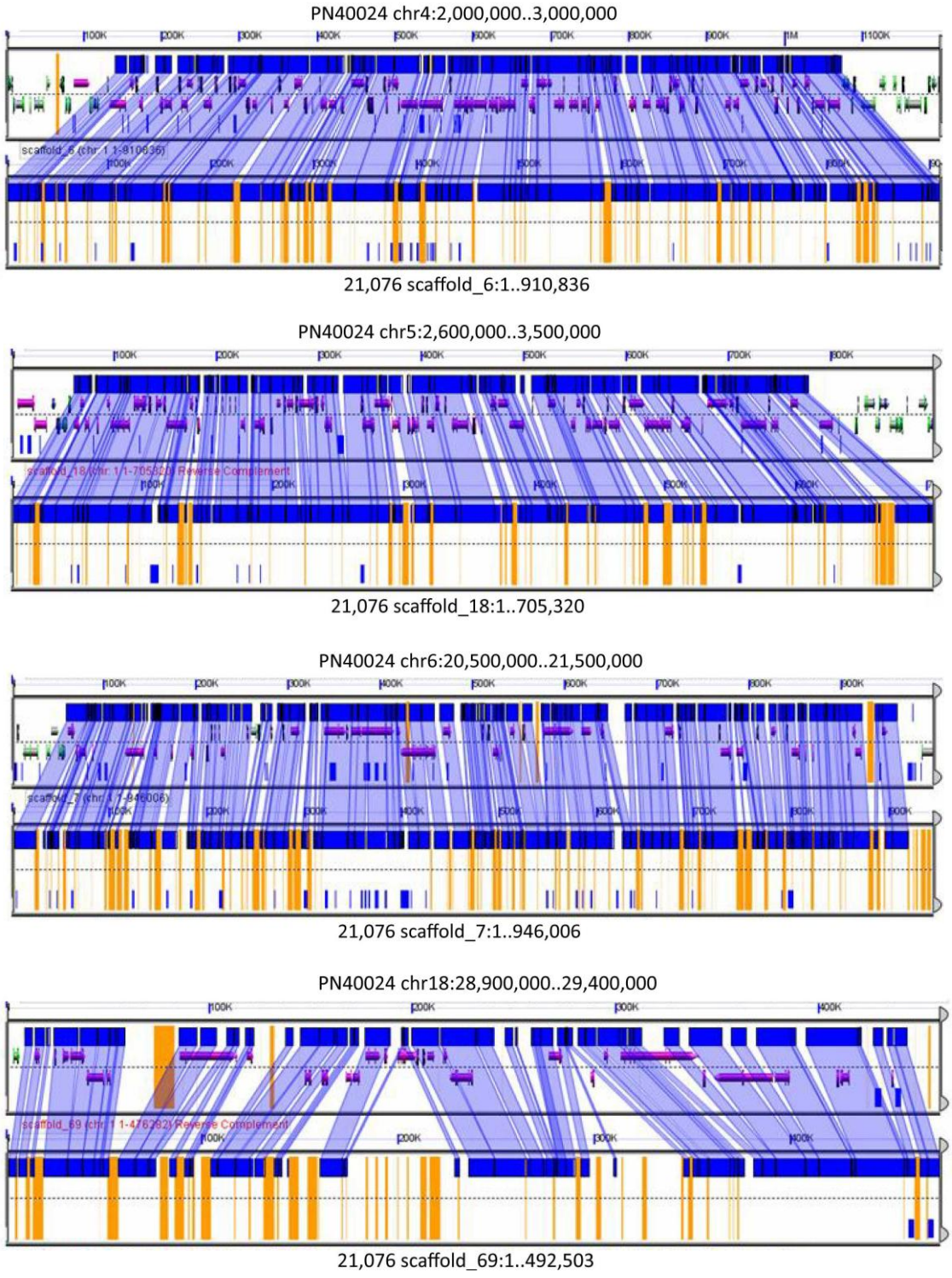


Figure 7. Alignment between 21,076$^{Rpv3+/Rpv3+}$ scaffolds and PN40024 chromosome pseudomolecules. Graphs were created using GEvo. Blue connectors show the shared regions between the two haplotypes. Orange bars indicate sequence gaps between contigs.

homozygous region of 21,076$^{Rpv3+/Rpv3+}$ and corresponding to a homozygous region of PN40024 on chr4; (ii) scaffold_18 assembled from a region of 21,076$^{Rpv3+/Rpv3+}$ that is heterozygous for two different *V. vinifera* haplotypes (see paragraph 3.10) and

corresponding to a homozygous region of PN40024 on chr5; scaffold_7 assembled from a region of 21,076$^{Rpv3+/Rpv3+}$ that is heterozygous for a *V. vinifera* haplotype and an introgressed haplotype and corresponding to a homozygous region of PN40024 on chr6; scaffold_69 assembled from a homozygous region of 21,076$^{Rpv3+/Rpv3+}$ for an introgressed haplotype; scaffold_12 assembled from a homozygous region of 21,076$^{Rpv3+/Rpv3+}$ and corresponding to a heterozygous region in PN40024.

All scaffolds assembled from 21,076$^{Rpv3+/Rpv3+}$ are collinear with the chromosome pseudomolecules of PN40024, irrespectively of the homozygous/heterozygous status of the region in 21,076$^{Rpv3+/Rpv3+}$. In particular, the two haplotypes that are expected to diverge substantially in the region covered by scaffold_7 were assembled in a single sequence that is collinear with the reference genome. Blast search of scaffold_7 against all other scaffolds did not identify any assembled allelic scaffold. The most interesting case is scaffold_12 in 21,076$^{Rpv3+/Rpv3+}$ corresponding to a heterozygous region in the genome of PN40024. The allelic region in PN40024 is entirely assembled into small unanchored scaffolds that were temporarily placed in chr_Unkwown (Figure 7).

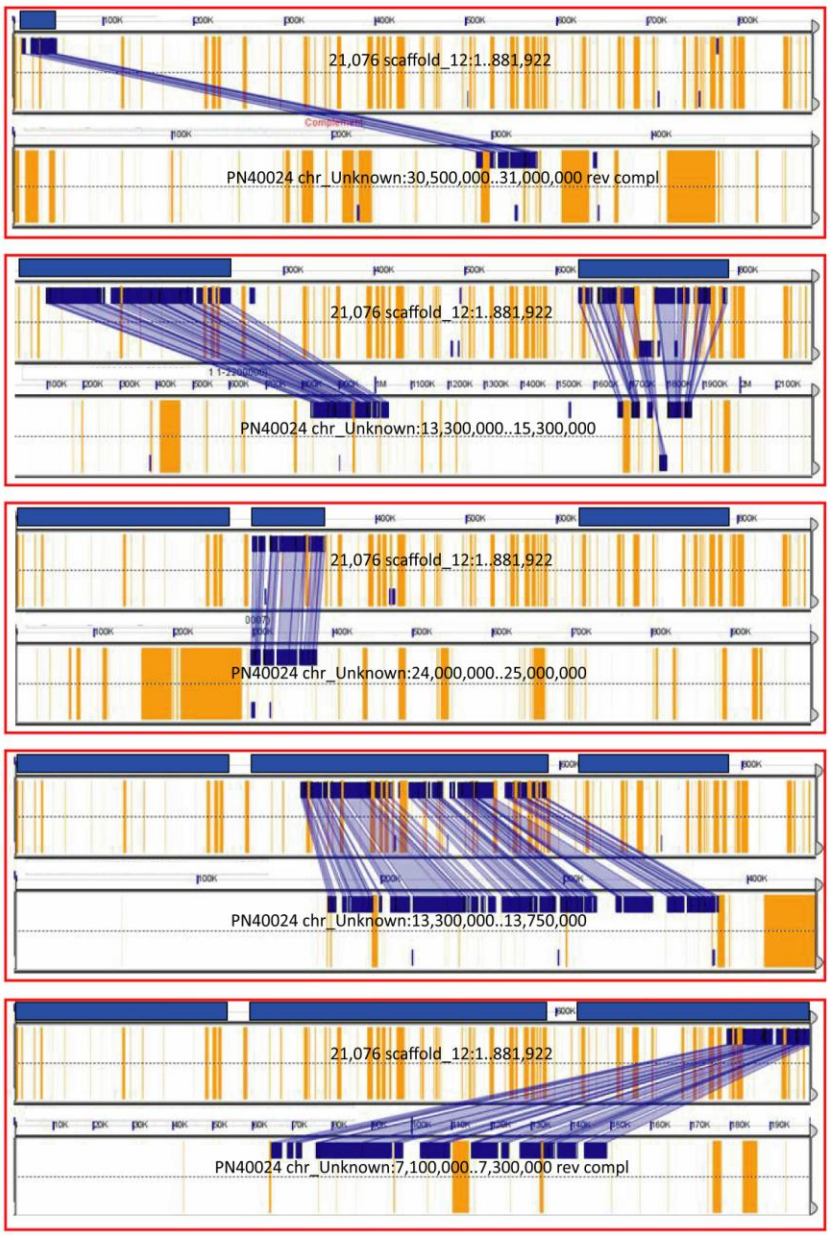


Figure 7. Alignments between 21,076$^{Rpv3+/Rpv3+}$ scaffold_12 and five intervals of the PN40024 chromosome pseudomolecule chr_Unknown. Graphs were created using GEvo. Blue connectors show the shared regions between the two haplotypes. Orange bars indicate sequence gaps between contigs.

3.2.1 Assembly of the *Rpv3* locus

The ALLPATHS scaffolds were identified, ordered and oriented to build the pseudomolecules based on the relative position of genetic markers, BAC-end sequences and unique regions within BACs previously anchored to the locus (Table 7). The upper border of the locus was defined by marker UDV736 and supported by 17 recombinants, available in pseudo-testcross populations. The lower border of the

locus was defined by marker UDV737 and supported by 18 recombinants. In a previous work (Copetti D., PhD dissertation), BAC-end sequences and amplicons from unique regions within BACs were also used for developing SNP markers, which were also used as a bait to catch scaffolds.

Primer sequences of the markers UDV736 and UDV737 were used for BlastN search of all scaffolds resulting from the assembly. Marker UDV736 is located on scaffold sc_19 with a length of 760 kb. Marker UDV737 is located on sc_777 that has a size of 167 kb, separated by a gap of unknown size. BAC-end sequences 10G13_L, 10G13_R and 55E09_R were located on scaffold sc_19 as well, downstream the marker UDV736. BAC-end sequences 37D04_L, 66A20_R,32F21_L and 11C03_R were all located on scaffold sc_777 upstream the marker UDV737. BAC-end sequences 55E09_L and 46G16_L allowed to identify the scaffold sc_2088, with a length of 54 kb. Scaffold sc_1742 was found with the BAC-end sequence 11C03_L, and this scaffold is 75 kb long.

Table 7. Type of marker and position within the scaffolds.

Marker	Marker type	Scaffold	Scaffold size (kb)	Marker position (bp)
UDV736	SSR	sc_19	760	594,283
10G13_R	SNP in BAC ends	sc_19	760	644,643
10G13_L	SNP in BAC ends	sc_19	760	723,370
55E09_R	SNP in BAC ends	sc_19	760	737,442
55E09_L	SNP in BAC ends	sc_2088	54	25,180
46G16_L	SNP in BAC ends	sc_2088	54	31,411
37D04_L	SNP in BAC ends	sc_777	167	4,243
66A20_R	SNP in BAC ends	sc_777	167	85,493
32F21_L	SNP in BAC ends	sc_777	167	97,716
UDV737	SSR	sc_777	167	110,533
11C03_R	SNP in BAC ends	sc_777	167	114,867
11C03_L	SNP in BAC ends	sc_1742	75	3,644

3.2.2 BAC assemblies

Paired-end Illumina reads from nine clones from a BAC library of the variety Bianca were available at the beginning of my PhD. We reassembled them into contigs using ABySS. In Table 8, reported are the number of contigs per BAC, the total size of the contigs per BAC, and the estimated size of each BAC based on the alignment of BAC Illumina reads on the assembly of 21,076$^{Rpv3+/Rpv3+}$. To determine the relative order and the orientation of BAC contigs, we used partial overlapping between BAC contigs previously assembled by CLC, BAC contigs assembled by ABySS, and scaffolds of 21,076$^{Rpv3+/Rpv3+}$. One supercontig per each BAC was obtained. Then, overlapping regions were searched between BAC supercontigs, resulting in two pseudomolecules. The first one is composed by the overlapping BACs 55E09 and 46G16, the second one by the overlapping BACs 37D04, 32F21, 9N03, 50B08 and 66A20. 10G13 and 11C03 do not overlap other BACs and are located at either extremity of the *Rpv3* locus (Figure 8).

Table 8. Metrics assembly of BAC clones.

BAC	N. of contigs	Assembled BAC size	Size of the region covered by BAC reads on the 21,076$^{Rpv3+/Rpv3+}$assembly
10G13	21	81 kb	78,7 kb
55E09	41	41,4 kb	63,7 kb
46G16	51	31,5 kb	37,0 kb
37D04	22	25,1 kb	18,9 kb
32F21	41	79,8 kb	96,4 kb
9N03	57	78,2 kb	84,2 kb
50B08	27	57,7 kb	34,5 kb
66A20	53	52,8 kb	74,1 kb
11C03	10	78,8 kb	71,9 kb

3.2.3 Reconstruction of the *Rvp3* locus

Overlapping in the locus between 21,076$^{Rpv3+/Rpv3+}$ scaffolds and BAC supercontigs was used in order to bridge sequence gaps in either assembly (Figure 8). Three gaps were present between adjacent scaffolds. Three gaps were present between groups of overlapping BACs. BAC clone 10G13 is entirely included in sc_19. BAC_55E09 partially

overlaps BAC_46G16. The terminus of BAC_55E09 opposite to the end overlapping BAC_46G16 has a 100 % sequence identity with the terminus of sc_19 across the terminal 17.4 kb. The gap between BACs 10G13 and 55E09 was 18.8 kb, based on their projection on sc_19. The opposite terminus of BAC_55E09 as well as the part of BAC 46G16 overhanging BAC_55E09 have a 100 % sequence identity with one terminus of sc_2088, extending up to 32 kbs inwards into sc_2088. Thus, BAC_55E09 and BAC_46G16 bridge the gap between scaffolds sc_19 and sc_2088, and orient sc_2088. Likewise, sc_19 and sc_2088 allowed us to correctly orient BAC_55E09 and BAC_46G16 with respect to a provisional orientation that was based on two SNP markers 55E09_6a and 55E09_2b (Copetti D., PhD dissertation). SNP 55E09_6a and SNP_55E09_2b were scored in amplicons that were intended to be amplified in recombinant individuals from the target region corresponding to BAC_55E09, but they may have actually originated from a highly similar duplicated region corresponding to BAC 50B08. The terminus of BAC_37D04 overhanging the BACs 50B08, 66A20, 9N03, and 32F21 has a 100 % sequence identity across the terminal 16 kbp of sc_2088, opposite to the terminus covered by 46G16. The gap between BACs 46G16 and 37D04 was 6 kb, based on their projection on sc_2088. On the opposite side, BAC_37D04 extends inwards into sc_777 for 4.5 kb with a 100 % sequence identity. BACs 50B08, 9N03, and 32F21 also bridge sc_2088 and sc_777, whereas BAC_66A20 is entirely included in sc_777. Finally, BAC_11C03 spans the terminal 68 kb of sc_777, opposite to the terminus covered by BACs 37D04, 50B08, 66A20, 9N03, and 32F21, and extends for 4.1 kb into sc_1742. The gap between the overhang of BACs 32F21 and BAC_11C03 was 17.4 kb, based on their projection on sc_777.

To sum up, all gaps between 21,076$^{Rpv3+/Rpv3+}$ scaffolds were bridged by BAC supercontigs, and vice versa. The presence of repetitive DNA was the cause of interruption of scaffolding in the 21,076$^{Rpv3+/Rpv3+}$ assembly. Two gaps between 21,076$^{Rpv3+/Rpv3+}$ scaffolds were caused by the presence of transposable elements which were otherwise assembled in each BAC where they were present as single-copy DNA. The third gap between sc_777 and sc_1742 was caused by the presence of a TNL gene which has high similarity with several duplicated copy contained in sc_1742 and downstream. Physical gaps between adjacent BAC contigs, previously identified by my predecessor (Copetti D., PhD dissertation), amounted to 42.2 kb.

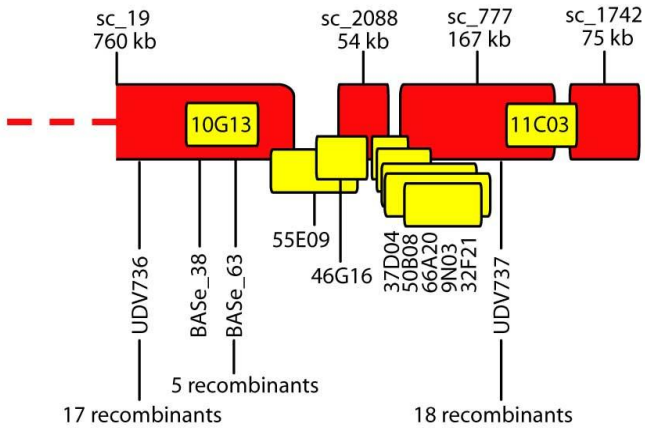


Figure 8. Reconstruction of the *Rpv3* locus from BAC and whole-genome assemblies. ALLPATHS scaffolds of 21,076$^{Rpv3+/Rpv3+}$ are indicated by red blocks. In yellow, indicated are supercontigs of BAC clones. In this diagram, the *Rpv3* locus is defined upstream by SSR-marker UDV736 and downstream by SSR-marker UDV737.

3.3 Genetic interval on the resistance haplotype

In order to define the boundaries of the locus, genetic markers previously designed on unique regions of BAC clones were aligned to the assembled sequence. The information about the recombination at these markers was integrated with the phenotypic data of recombinants from the cross “Chardonnay x Bianca” (Figures 9 and 10). Markers SNP_46G16L and SNP_9N03R narrowed the locus. The SNP_9N03L cosegregated with the trait. The reorientation of the BAC contig 55E09-46G16 allowed us to locate the BAC end 46G16_L towards the telomeric end of the chromosome and place 46G16 in the middle of sc_2088. A SNP within 46G16_L narrowed the locus, excluding the functional role of important candidate genes present in the BAC contig 55E09-46G16. In the discarded region upstream 46G16L, a TNL disease resistance gene homolog and a TIR gene are predicted to encode full-length proteins, but according to the phenotypic data, the alleles carried on the resistance haplotype are not associated with downy mildew resistance. Three recombinants excluded that the *Rpv3* gene is located upstream 46G16_L. Seven recombinants excluded that the *Rpv3* gene is located downstream 9N03_R. The current interval on the resistance haplotpe is 105 kb. This restricts the number of candidate genes to TNL2a, TNL2b and LRR-kinase.

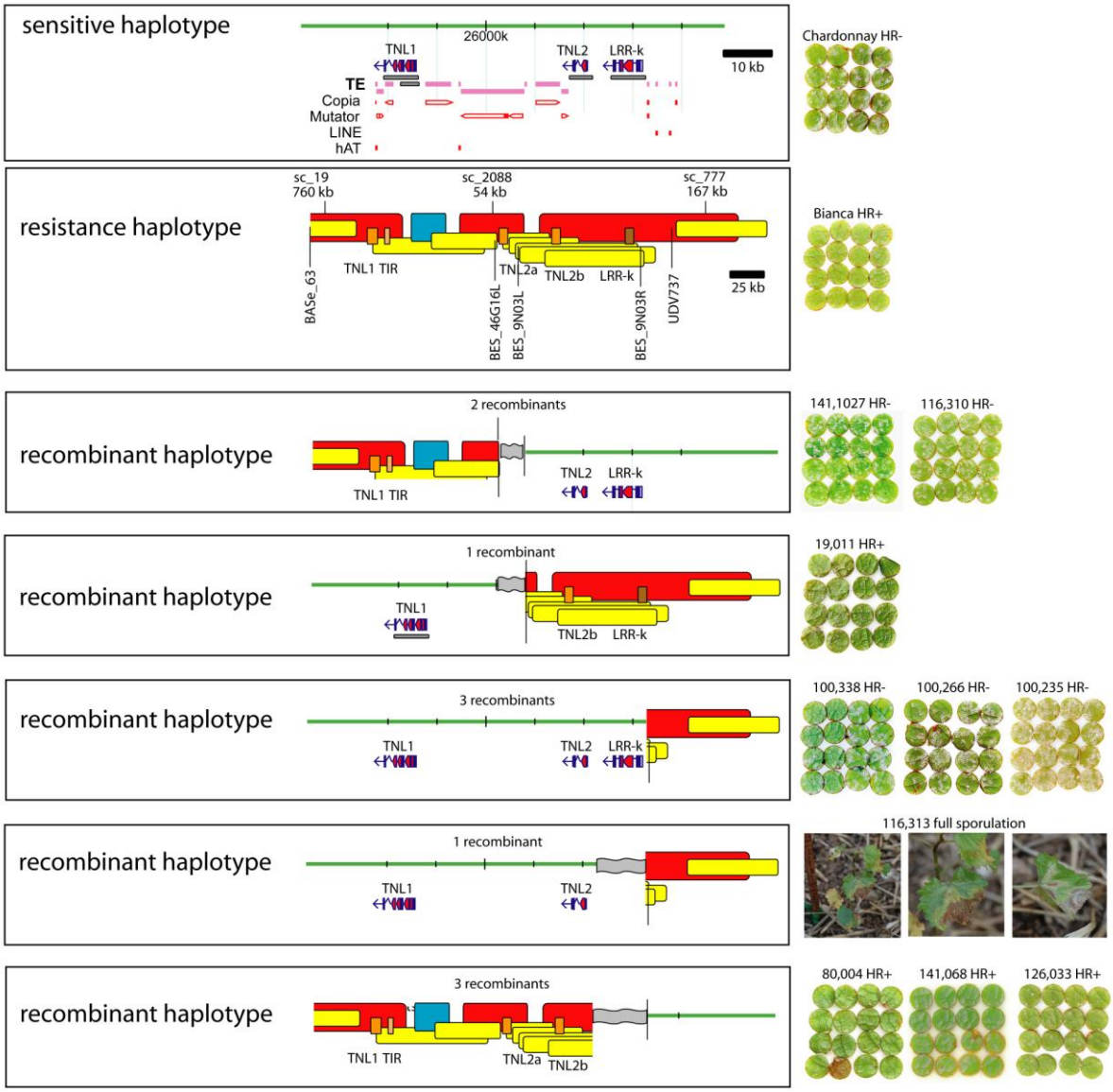


Figure 9. The *Rpv3* locus. Diagram of a sensitive (PN40024) and the resistance haplotype are shown on top. Recombinant haplotypes in 10 offspring of Bianca are shown with the position of crossovers, as defined by the genetic markers SNP_46G16L, SNP_9N03 and SNP_9N03R. Orange and brown rectangles show predicted genes on the pseudomolecule of the resistance haplotype, which is composed by whole-genome assembled scaffolds (in red) and BAC supercontigs (in yellow). All offspring are heterozygous for the recombinant haplotype shown in this figure and a vinifera sensitive haplotype donated by Chardonnay. The phenotype of each offspring is shown on the right-hand side. A transposable element of 20 kb (paragraph 3.4.1.5) is indicated with the blue rectangle.

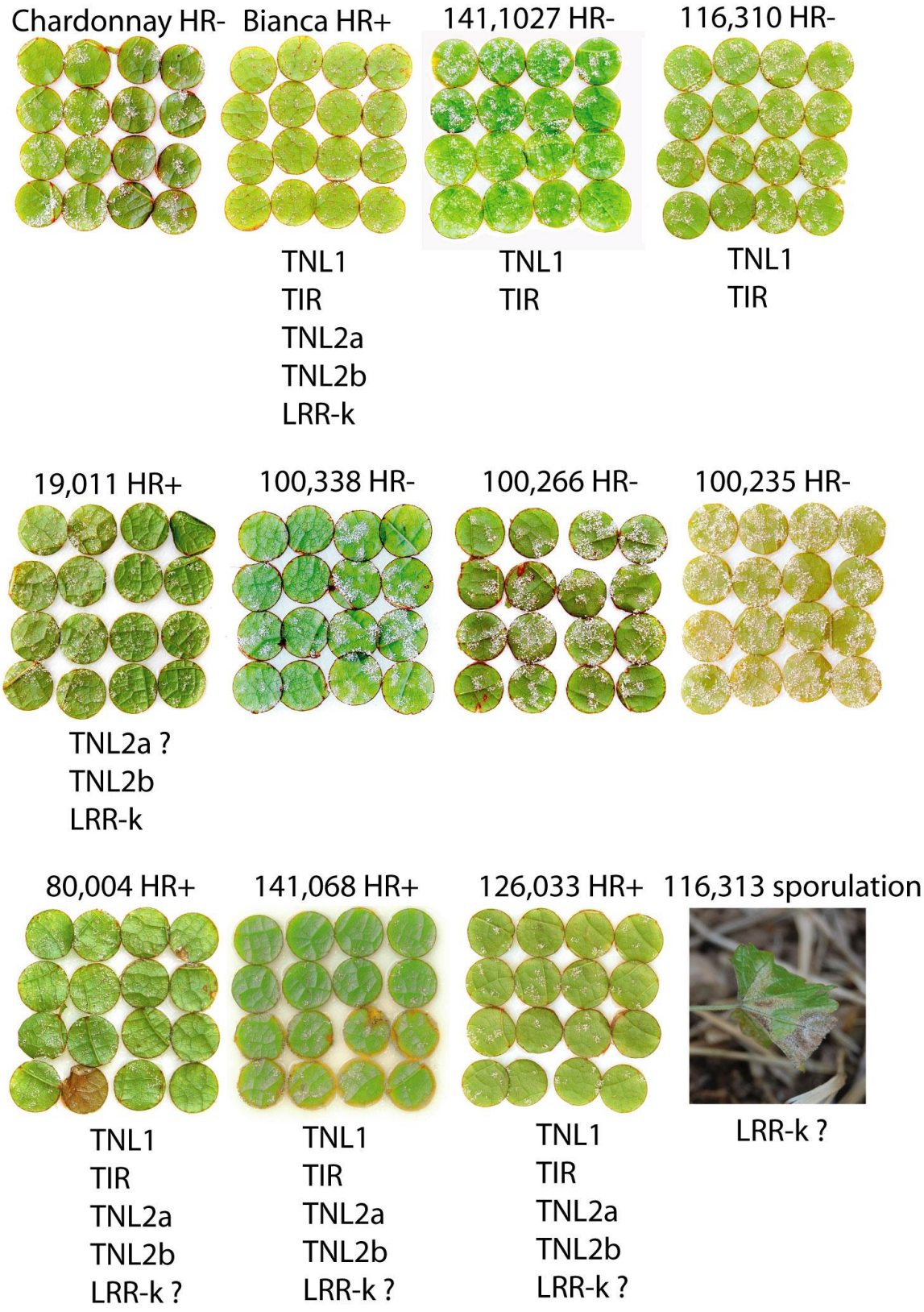


Figure 10. Phenotypes of Bianca offspring carrying recombinant haplotypes in the Rpv3 locus. Leaf discs were inoculated with P. viticola and pictures were taken at 96 hpi. Each individual is identified by a code number (above the discs) and the presence/absence of HR is

reported beside the ID code. Below each set of leaf discs, reported are the genes of the seedling present in the segment of the resistance haplotype inherited from Bianca

3.4 Structural annotation

Structural annotation of the current genetic interval (Figure 9) revealed the presence of three genes in the *Rpv3* locus. Two of them are disease resistance genes that encode for TIR-NB-LRR protein receptors (TNL) and the other one encodes for a LRR-kinase, a transmembrane protein kinase. To characterize the intron-exon structure of these genes, two approaches were used: *ab initio* gene prediction and RNA-Seq to validate the gene models.

3.4.1 Gene prediction

3.4.1.1 Ab initio

Gene prediction was carried out with FGENESH and confirmed by BlastX against full-length proteins. In the upper part of Figure 12 shown are the gene models. The two TNL genes have coding regions with a similar structure composed by five exons and four introns. The LRR-kinase has a coding sequence with only two exons. All three genes are encoded by the negative strand.

3.4.1.2 RNA-Seq

RNA was extracted from infected leaves of 21,076$^{Rpv3+/Rpv3+}$ (Figure 11) and sequenced. Paired-end 100-bp reads were aligned on assembly of 21,076$^{Rpv3+/Rpv3}$. In the region delimited by markers 46G16_L and 9N03_R, gene expression analysis revealed that the three predicted genes are expressed during the early response to *Plasmopara viticola* infection. No other region within this genetic interval showed evidence of transcription.

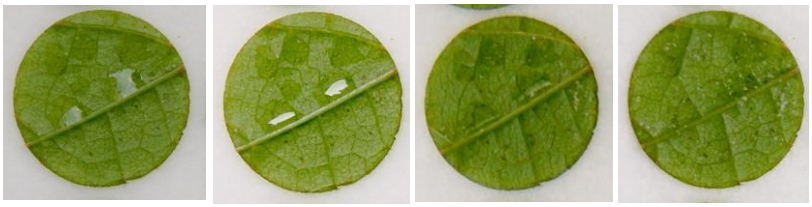


Figure 11. Infection of leaf discs with *Plasmopara viticola*.

A split read alignment was performed by the software TopHat. A first round of read alignment to the reference defined potential exons that are joined together in a consensus made of regions covered by reads. Reads were then split and re-align taking into account all possible splicing sites. There is evidence of intron splicing when two segments of the same read are aligned on adjacent genomic regions. Shown in the lower part of figure 12, is the alignment of RNA reads against the predicted genes with the graphical viewer Tablet. In panel A, reads are aligned against the TNL2a sequence, showing a gene with five exons and confirming the structure predicted by FGENESH. Four events of splicing take place during mRNA maturation. There is also evidence of alternative splicing by skipping exon 2, visible by split reads – represented by red junctions – between the first and the third exon. A few reads also support limited retention of the first intron. TNL2b is shown in Panel B and has the same number of exons as in TNL2a, with four events of splicing. Contrary to TNL2a, there is no evidence of exon skipping or intron retention for TNL2b. Coverage of the 5'-UTR and the first exon of TNL2b is higher than the remainder of the coding sequence. No SNP was detected in the RNA reads aligned to these regions.

The expression of LRR-kinase is shown in panel C. RNA-Seq confirmed the presence of two exons in accord with the *ab initio* gene prediction. Read alignment indicates a long 5'-UTR region.

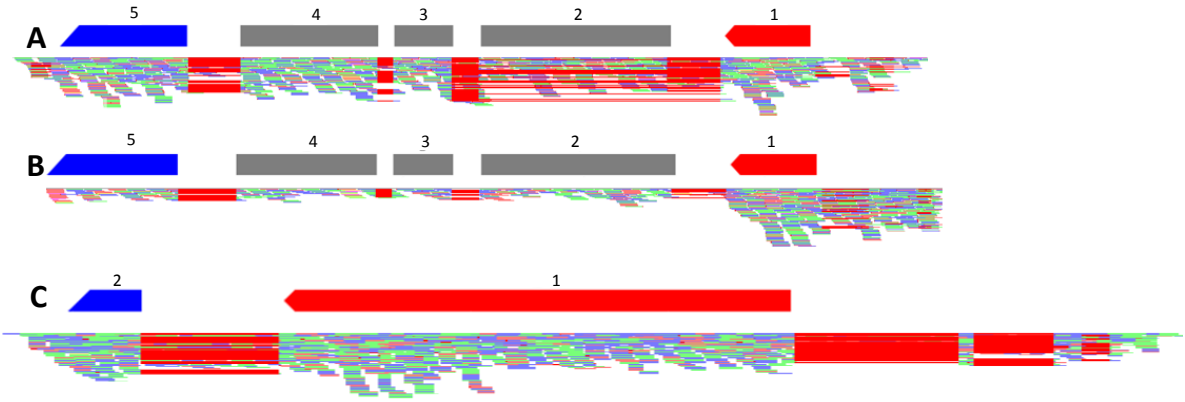


Figure 12. Gene models: *ab initio* prediction versus RNA-seq. In the upper part of each panel shown is *ab initio* gene prediction by FGenesh of TNL2a (panel A), TNL2b (panel B), LRR-kinase (panel C). The blocks indicate exons. In blue the terminal exon, in red the first exon, in gray internal exons. All three genes are encoded by the negative strand. The lower part of each panel shows RNA reads aligned against gene sequences. Split reads spanning introns are connected across introns by red lines. In panel A, red lines connecting split reads from the first exon to the third exon indicate alternative splicing by exon skipping.

3.4.1.3 Alignment of known proteins

Sequences of genes present in the *de-novo* assembly of the locus were aligned against the NCBI protein database with BlastX algorithm. Analysis of amino acid identity revealed that TNL2a may encode a predicted protein showing similarity with functionally characterized NB-LRR across 1130 amino acids, while TNL2b may encode a predicted protein of 1131 amino acids. NCBI annotation and CRIBI annotation of the PN40024 genome sequence were interrogated to find predicted gene models that translate into predicted proteins similar to TNL2a and TNL2b. The most similar predicted protein in PN40024 is encoded by a gene referred to as LOC100255177 in NCBI and VIT_18s0041g01340 in CRIBI annotation. The protein predicted by NCBI is 1212 amino acids long, while CRIBI prediction suggests the translated amino acid sequence is as short as 1019 amino acids. Similarity with NCBI protein database was searched in order to also find related proteins in other plants. Tobacco TMV resistance gene *N* shows high similarity with TNL2a and TNL2b in *Rpv3* locus. Gene *N* encodes a protein of 1144 amino acids.

The LRR-kinase in *Rpv3* locus is 973 amino acids long and showed the highest similarity with the receptor-like protein kinase HSL1-like of PN40024 corresponding to the gene ID LOC100260256 in the NCBI annotation and to the CRIBI annotation VIT_18s0041g01350, located at chr18: 26,025,959..26,031,870.

3.4.1.4 Nucleotide variation between duplicated TNL genes

A pairwise comparison of nucleotide variation between TNL genes in the *Rpv3* locus, including the copy located immediately upstream of the genetic border, was conducted domain-by-domain. The comparison was conducted between paralogs in PN40024, between paralogs in the resistance haplotype, and between allelic counterparts in PN40024 and the resistance haplotype (Table 9).

TNL1, the copy located upstream of the locus border, and TNL2, the copy in the *Rpv3* locus, differ slightly in PN40024 for the number of synonymous substitutions per synonymous site (Ks) in the region encoding the TIR (Ks=0.0096). The ratio of (K_a) to the number of synonymous substitutions per synonymous site (K_s) is 1.43. The paralogs TNL1 and TNL2 have a Ka/Ks ratio of 0.33-0.55 in the resistance haplotype, depending of which copy of TNL2 in considered. Allelic comparison of TNL1 between PN40024 and the resistance haplotype gave a Ka/Ks ratio of 0.56. Ka/Ks ratio is slightly higher in the allelic comparison of TNL2 (0.99).

In the region encoding the NB-ARC domain, Ka/Ks ratios stay within a narrow range of variation comprised between 0.46 and 0.62 for all comparisons between alleles and between paralogs in either haplotype. In the region encoding the LRRs, Ka/Ks ratios are much higher, ranging 1.66-2.01 in comparisons between alleles and between paralogs in either haplotype.

Table 9. Ks, Ka, and Ka/Ks ratio between TNL genes in PN40024 and 21,076$^{Rpv3+/Rpv3+}$. TNL1 and TNL2 in PN40024 (in green background) correspond, respectively, to NCBI gene models LOC100251613 and LOC100255177, CRIBI gene models VIT_18s0041g01330 and VIT_18s0041g01340 located at chr18:25,981,032..25,986,117(complement) and chr18:26,016,614..26,022,006.

TIR domain

Seq 1	Seq 2	SynDif	SynPos	Ks	NSynDif	NSynPos	Ka	Ka/Ks
TNL1	TNL2	1	104.92	0.0096	5	369.08	0.0137	1.43
TNL1	TNL1	4	104.33	0.0394	8	369.67	0.022	0.56
TNL2	TNL2a	2	104.58	0.0194	7	369.42	0.0192	0.99
TNL2	TNL2b	2	104.67	0.0194	7	369.33	0.0192	0.99
TNL1	TNL2a	5	104	0.0497	6	370	0.0164	0.33
TNL1	TNL2b	5	104.08	0.0496	10	369.92	0.0275	0.55
TNL2a	TNL2b	0	104.58	0	4	369.42	0.0109	-

NB-ARC domain

Seq 1	Seq 2	SynDif	SynPos	Ks	NSynDif	NSynPos	Ka	Ka/Ks
TNL1	TNL2	29.17	190	0.1718	56.83	710	0.0846	0.49
TNL1	TNL1	25.67	189.67	0.1492	51.33	710.33	0.076	0.51
TNL2	TNL2a	14.5	189.17	0.0809	30.5	710.83	0.0442	0.55
TNL2	TNL2b	4	190.17	0.0213	9	709.83	0.0128	0.60
TNL1	TNL2a	16	188.83	0.0899	38	711.17	0.0554	0.62
TNL1	TNL2b	11.5	189.83	0.0632	22.5	710.17	0.0324	0.51
TNL2a	TNL2b	16.5	189	0.0928	29.5	711	0.0427	0.46

LRR domain

Seq 1	Seq 2	SynDif	SynPos	Ks	NSynDif	NSynPos	Ka	Ka/Ks
TNL1	TNL2	16.75	379.67	0.0455	99.25	1303.33	0.0803	1.76
TNL1	TNL1	12.5	376.17	0.034	81.5	1306.83	0.0651	1.91
TNL2	TNL2a	16	378.92	0.0435	93	1304.08	0.0749	1.72
TNL2	TNL2b	7.33	379.25	0.0196	55.67	1303.75	0.044	2.24
TNL1	TNL2a	18	375.42	0.0495	102	1307.58	0.0824	1.66
TNL1	TNL2b	14.5	375.75	0.0396	98.5	1307.25	0.0794	2.01
TNL2a	TNL2b	15.67	375.33	0.0429	98.33	1307.67	0.0792	1.85

sensitive haplotype PN40024

resistant haplotype 21078

3.4.1.5 Transposable elements

The assembled sequence of the resistance haplotype was also analyzed for the content in transposable elements. A large transposable element of 20,895 bp was found in the region spanned by the BAC contig 55E09-46G16. The sequence of this TE

was assembled from the BAC clone 55E09. The same TE caused an interruption in the process of scaffolding in the whole-genome assembly of 21,076$^{Rpv3+/Rpv3+}$, leaving a gap between scaffolds sc_19 and sc_2088. BlastN search against all 21,076$^{Rpv3+/Rpv3+}$ scaffolds showed that the *de novo* assembly failed to reconstruct this element in scaffolds larger than 900 bp. Alignment of this TE against chr_18 of PN40024 showed that the element is not present in the allelic region of PN40024, but it is repeated for its entirety in two other locations on chromosomes 4 and 14. The coordinates of the TE on chromosome 4 are 15,904,878 – 15,927,139, while the coordinates on chromosome 14 are 11,491,395 – 11,514,659. Similarity between the TE in the resistance haplotype *of Rpv3* and in chromosome 14 of PN40024 was shown in Figure 13 with a dot plot graph.

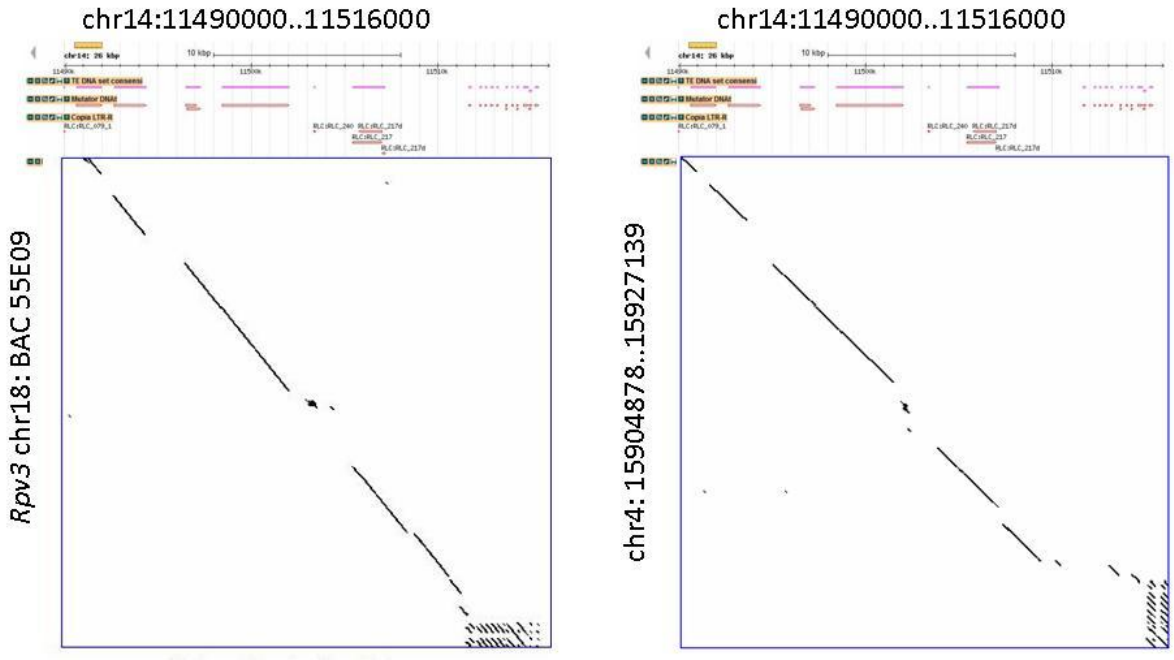


Figure 13. Structural organization of the transposable element in the *Rpv3* locus and colinearity with TEs in PN40024 on chr14 and chr4. On top of the dot plot, it is shown the Gbrowse visualization of the interval in PN40024 chr14. For *Rpv3*, illustrated is the extracted portion of BAC_55E09 containing the TE.

A closer inspection of the sequence of the transposable element with the Gbrowse of PN40024 revealed that this large block is composed by smaller repeated sequences (in the upper part of the Figure 13). Two different classes of TE are included in the block: Mutator and Copia. High similarity was observed between the Mutator elements in the *Rpv3* locus and the corresponding sequence on chr 14 with 98 % of nucleotide

identity, while 97 % of identity was found between the Copia element in the two regions. The k-mer profile in PN40024 (Figure 14) indicates that each element is present in the genome in multiple copies. However, the k-mer profile in the intervening regions and the colinearity across the entire region in the dot plot suggests that the entire block is ectopically duplicated in two locations of the PN40024 genome as well as in the resistance haplotype of the *Rpv3* locus.

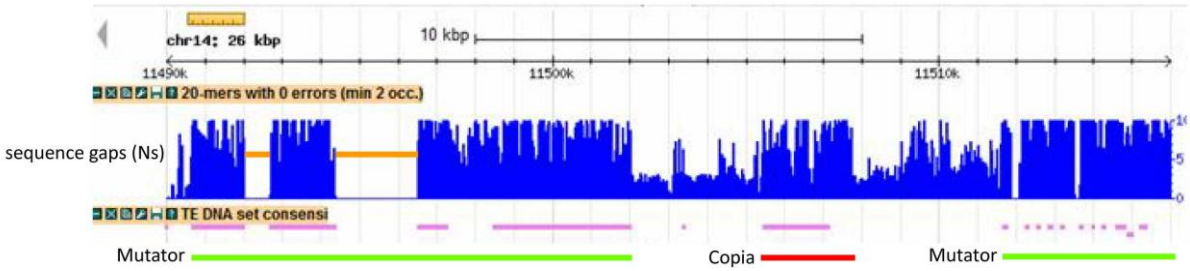


Figure 14. K-mer profile across the TE_chr14:11,491,395..11,514,659 in PN40024. Orange bars indicate sequence gaps.

Other TEs were annotated in the *Rpv3* locus in the region populated by the candidate genes. These TEs are discussed in detail in the paragraph of the analysis of structural variation.

3.5 Functional annotation

3.5.1 Predicted proteins

NB-LRR proteins and LRR-kinases mediate the recognition of pathogen effectors activating host defence. Genes present in the *Rpv3* locus belong to the TNL class of NB-LRR because they code for a TIR (Toll/interleukin 1-like receptor) domain at N-terminus. Another candidate encodes a receptor carrying a transmembrane domain, linking putative extracellular LRR and cytosolic kinase domains. In *Arabidopsis thaliana*, TNL genes were classified into groups and subgroups depending on number of introns, exons and encoded protein motifs (Meyers et al. 2003). Proteins predicted in the *Rpv3* locus were compared with this set of classified genes. The highest similarity was found with At5g45250 gene, the prototype of the TNL-B subclass. At5g45250 gene is better known as *RPS4*, a functionally characterized gene conferring resistance to bacterial pathogens that express the type III effector AvrRps4. This gene

model shows a structure with five exons, the first one encodes a TIR domain, the second a NB-ARC domain, the third and the forth exon encodes NL-LRR and LRR domain respectively, and the last one encodes the C-terminal domain (Figure 15). RPS4 is targeted to the nucleus. The TIR domain of RPS4 is involved in TIR-TIR domain heterodimerization with RRS1 conferring resistance to multiple pathogens (Sohn et al 2014).

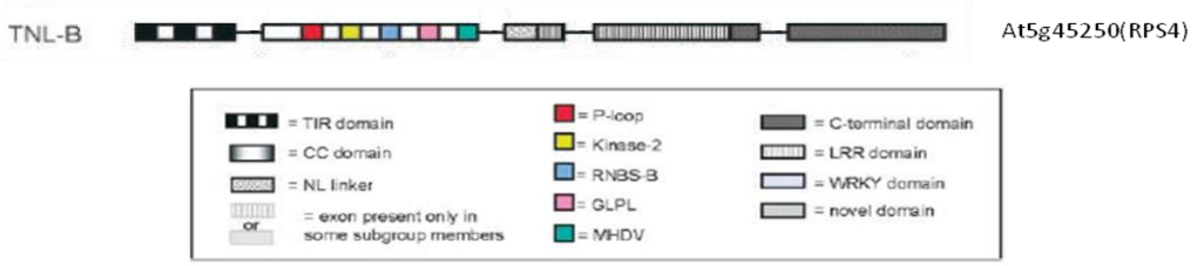


Figure 15. Prototype of the TNL-B subclass of NB-LRR genes in *Arabidopsis thaliana*.

The rice *Xa21* gene encodes a transmembrane receptor protein of 1025 amino acid with the same intron-exon structure and the same domains as in the LRR-kinase of the *Rpv3* locus. *Xa21* is composed of two exons, the first encodes a kinase domain at the N-terminus of the protein protruding into the cytoplasm, the second one encodes extracellular LRRs that are involved in the recognition of the pathogen outside the cell membrane. The two domains are linked by a short transmembrane domain.

3.5.2 Conserved domains

NB-LRR genes code for a number of conserved domains. Each domain contains conserved motifs. Predicted amino acid sequences of TNL genes in the *Rpv3* locus of 21,076$^{Rpv3+/Rpv3+}$, were aligned with ClustalW algorithm along with allelic TNLs in PN40024 and the tobacco TMV resistance protein N (Thompson et al. 1994), with the aim of identifying conserved motifs and amino acid changes. Domains and motifs were identified in the amino acid sequences using the Pfam database and the web-tool LRRfinder (www.lrrfinder.com). Domains and motifs are shown in the aligned amino acid sequences (Figures 16, 17 and 18). The TIR domain predicted for TNL alleles and paralogs in the *Rpv3* locus is invariably 144 amino acids long in 21,076$^{Rpv3+/Rpv3+}$, whereas the tobacco N protein has an additional Ser in position 156, extending the

total length of the TIR domain to 145 amino acids. A trinucleotide repeat $(TCT)_n$ is present in TNL2 between the ATG start codon and the first triplet coding for the TIR domain. This trinucleotide repeat is in step with the reading frame and it codes for a homopolymeric serine repeat. Variation in the number of repeats of the core motif occurs between the TNL2a and TNL2b paralogs in 21,076$^{Rpv3+/Rpv3+}$, resulting in a variable length of the homoserine tract. The homoserine is seven amino acids long in TNL2a. The same length of the homoserine tract is encoded by the allele of TNL2 present in PN40024. The paralog TNL2b 21,076$^{Rpv3+/Rpv3+}$ codes in for a homopolymeric amino acid repeat with two additional serines. Four motifs are conserved in the TIR domain and are named from TIR-1 to TIR-4. The corresponding amino acid sequences are underlined in Figure 16. TIR-1 and TIR-2 are highly conserved between allelic and paralogous TNL2 in PN40024 and 21,076$^{Rpv3+/Rpv3+}$. Amino acid substitutions in these motifs are observed only in the interspecific comparison of grapevine TNL2 with the tobacco N protein. Compared to tobacco N, TNL2 has three conservative substitutions at the C-terminus of the TIR-1 and four substitutions in the TIR-2 motif. The TIR-3 motif is more variable. Both TNL2 paralogs in 21,076$^{Rpv3+/Rpv3+}$ differ from the allele in PN40024 for the last three amino acids at the C-terminus of the TIR-3 motif. A Lys-Lys-Gln polymer in the protein encoded by PN40024 is substituted with a Arg-Asn-His polymer in the proteins encoded by 21,076$^{Rpv3+/Rpv3+}$. Arg-Asn-Gln is the polymer in the TIR-3 motif of the N protein. In TIR-4, TNL2 paralogs in 21,076$^{Rpv3+/Rpv3+}$ have one non-conservative substitution each, compared to the allele in PN40024. The first substitution is in position 153 of TNL2b with Cys replacing Ser, and the second one is in position 158 of TNL2a with Asp replacing Glu.

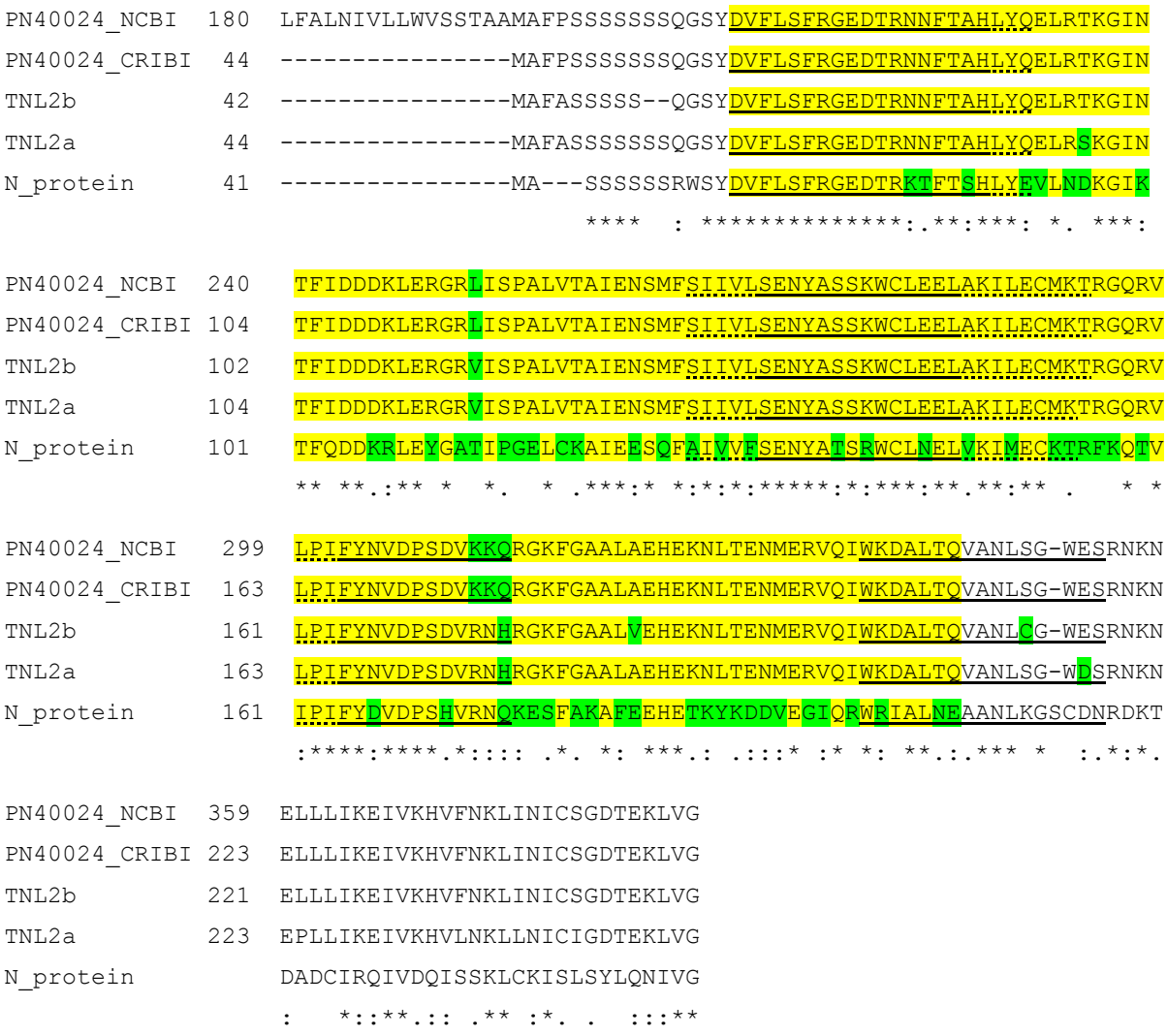


Figure 16. TIR domain. Comparison of the TIR domain between TNL genes present in the *Rpv3* locus of PN40024 and 21,076$^{Rpv3+/Rpv3+}$ and in protein N in tobacco. In yellow, indicated is the TIR domain; in green, the amino acid substitutions. Conserved motifs TIR-1 to TIR-4 are underlined in the TIR domain. Numbers beside gene names count the last amino acid in the raw.

Nucleotide binding site (NBS) or NB-ARC domain plays a central role in the activation of NB-LRR proteins. NB-ARC can bind different nucleotides, either ADP or ATP, and change the conformation of the protein depending on the presence of the pathogen. Several motifs are conserved in the NB-ARC domain, and they underlined in Figure 17. P-loop is 23 amino acids long and highly conserved, except for a non-conservative substitution in TNL2a of 21,076$^{Rpv3+/Rpv3+}$ with Arg replacing Gly, compared to the paralogous TNL2b in 21,076$^{Rpv3+/Rpv3+}$ and the allelic TNL2 in PN40024. Kinase-2 is 15 amino acid long with one Ala-to-Val conservative substitution in TNL2b of 21,076$^{Rpv3+/Rpv3+}$, compared to the paralogous TNL2a in 21,076$^{Rpv3+/Rpv3+}$ and the allelic

TNL2 in PN40024. The RNBS-B motif has a length of 25 amino acids. This motif is highly conserved between TNL2b in 21,076$^{Rpv3+/Rpv3+}$ and TNL2 in PN40024, with only one Ala-to-Thr conserved substitution. The RNBS-B motif is more diverse in TNL2b in 21,076$^{Rpv3+/Rpv3+}$, with one conservative and two non-conservative substitutions (Leu-to-Phe, Val-to-Ile, Trp-to-Gln). The GLPL motif is 16 amino acid long with a single Lys-to-Gln conservative substitution in TNL2a of 21,076$^{Rpv3+/Rpv3+}$. The C-terminal motifs in the NB-ARC are RNBS-D and MHDV. Both motifs have all amino acids conserved across alleles and paralogs of TNL2. Downstream the NB-ARC and just before the LRR repeats, there is another conserved domain of approximately 65 amino acids, highlighted in pink in Figure 17. This domain is initiated by the amino acids Gln-Phe-Val. The length of the linker is variable. Different gene annotations conflicted in the prediction of the amino acid sequence in the PN40024 TNL2 allele. According to NCBI prediction, the linker is composed of 56 amino acids, while CRIBI annotation predicted that is as short as 24 amino acids. The length of the linker in both TNL2 paralogs of 21,076$^{Rpv3+/Rpv3+}$ is 112 amino acids, more similar to the length in the tobacco protein N which is 91 amino acids.

```
PN40024_NCBI   359 ELLLIKEIVKHVFNKLINICSGDTEKLVGIDARIQEIKMRLRLESDDVGMIGIWGMGGIG
PN40024_CRIBI  223 ELLLIKEIVKHVFNKLINICSGDTEKLVGIDARIQEIKMRLRLESDDVGMIGIWGMGGIG
TNL2b          221 ELLLIKEIVKHVFNKLINICSGDTEKLVGIDARIQEIKMRLRLESDDVGMIGIWGMGGIG
TNL2a          223 EPLLIKEIVKHVLNKLLNICIGDTEKLVGIDARIREIKMRLCLESDDVRMIGIWGMGGIG
N_protein      221 DADCIRQIVDQISSKLCKISLSYLQNIVGIDTHLEKIESLLEIGINGVRIMGIWGMGGVG
                   :   *::**.:: .** :*. .  :::****:::.:*:  * :  :.* ::*******:*

PN40024_NCBI   410 KTTLARALYN------EISRQFEAHSFLEDVGKVLVNKGLIKLQQIFLYDLLEEKD---L
PN40024_CRIBI  274 KTTLARALYN------EISRQFEAHSFLEDVGKVLVNKGLIKLQQIFLYDLLEEKD---L
TNL2b          272 KTTLARALYN------EISRQFEAHSFLEDVGKVLVNKGLIKLQQIFLSDLLEEKD---L
TNL2a          274 KTTLARALYN------KISRQFEAHAFLEDVGKVLANEGLIKLQQIFLSSLLEEKY---L
N_protein      279 KTTIARAIFDTLLGRMDSSYQFDGACFLKDIKENKR--GMHSLQNALLSELLREKANYNN
                   ***:***:::      . * **:. .**:*: :     *: .**: :* .**.**

PN40024_NCBI   469 NTKGFTFIKARLHSKKALVVLDNVNDP-KILECLVGNWDWFGRGSRIIITARDKHLLIAH
PN40024_CRIBI  333 NTKGFTFIKARLHSKKALVVLDNVNDP-KILECLVGNWDWFGRGSRIIITARDKHLLIAH
TNL2b          331 NTKGFTSIKARLHSKKALVVLDNVNDP-KILECLVGNWDWFGRGSRIIITTRDKHLLISH
TNL2a          333 NMKGLTSIKARLHSKKVLVVLDNVNDP-TIFECLIGNQDWFGRGSRIIITARDKCLLISH
N_protein      339 EEDGKHQMASRLRSKKVLIVLDDIDNKDHYLEYLAGDLDWFGNGSRIIITTRDKHLIEKN
                   : .*   : :**:***.*:***::::    :* * *: ****.*******:*** *:  :
```

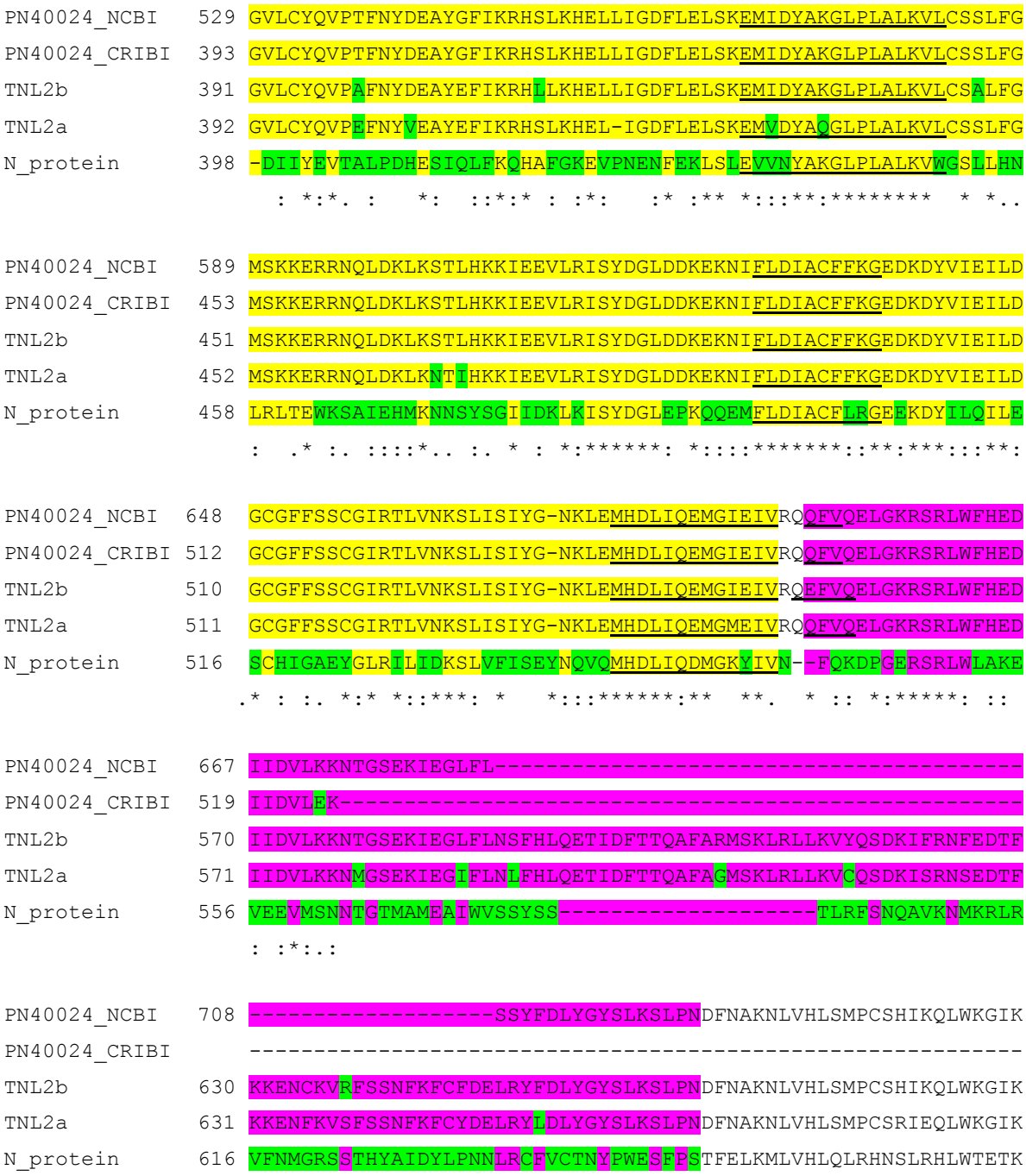


Figure 17. NB-ARC and NL-linker domains. NB-ARC is highlighted in yellow. Conserved motifs are underlined with this order starting from the top of the figure: P-loop, kinase-2, RNBS-B, GLPL, RNBS-D. NL-linker is highlighted in pink. Amino acid substitutions are highlighted in green. Numbers beside gene names indicate amino acid counts in each row.

The LRR domain is involved in the recognition of pathogen elicitors and consists of stretches of 20-30 amino acids that are rich in leucine. LRRs predicted by Pfam and LRRfinder are underlined in Figure 18. Irrespectively of the sensitivity of the algorithms in the identification of LRRs, all alleles and paralogs of TNL2 encode the same number

of LRRs and the predicted proteins have the same length across this domain. The LRR domain is longer and highly diverse in the tobacco N protein.

```
PN40024_NCBI  708   -------------------SSYFDLYGYSLKSLPNDFNAKNLVHLSMPCSHIKQLWKGIK
PN40024_CRIBI       ------------------------------------------------------------
TNL2b         630   KKENCKVRFSSNFKFCFDELRYFDLYGYSLKSLPNDFNAKNLVHLSMPCSHIKQLWKGIK
TNL2a         631   KKENFKVSFSSNFKFCYDELRYLDLYGYSLKSLPNDFNAKNLVHLSMPCSRIEQLWKGIK
N_protein     616   VFNMGRSSTHYAIDYLPNNLRCFVCTNYPWESFPSTFELKMLVHLQLRHNSLRHLWTETK

PN40024_NCBI  768   VLEKLKCMDLSHSKYLIETPNLSRVTNLERLVLEDCVSLCKVHPSLRDLKNLNFLSFKNC
PN40024_CRIBI 575   ----LKCMDLSHSKYLIETPNLSRVTNLERLVLEDCVSLCKVHPSLRDLKNLNFLSFKNC
TNL2b         690   VLEKLKCMDLSHSKYLIETPNLSRVTNLERLVLEDCVSLCKVHPSLKDLKNLKFLSLKNC
TNL2a         691   VLEKLKCMDLSHSKYLKETPNLSRVTNLERLVLEDCVSLCKVDPSLRDLKNLKFLSLKNC
N_protein     676   HLPSLRRIDLSWSKRLTRTPDFTGMPNLEYVNLYQCSNLEEVHHSLGCCSKVIGLYLNDC
                        *: :*** ** * .**::: :.*** : * :* .* :*. **   .::  * :::*

PN40024_NCBI  828   KMLKSLPSGPYDLKSLATLILSGCSKFEQFPENFGYLEMLKKLYADGTALRELPSSLSSL
PN40024_CRIBI 635   KMLKSLPSGPYDLKSLATLILSGCSKFEQFPENFGYLEMLKKLYADGTALRELPSSLSSL
TNL2b         750   KMLKSLPSGPYDLKSLATLILSGCSKFEQFPENFGNLEMLKKLYADGTALRELPSSLYSL
TNL2a         751   KMLKSLPSGPYDLKSLETLILSGCSKFEQFPENFGNLEMLKELYADGTALRELPSSLSSS
N_protein     734   KSLKRFP--CVNVESLEYLGLRSCDSLEKLPEIYGRMKPEIQIHMQGSGIRELPSSIFQY
                    * ** :*    ::::**  * * .*..:*::** :* ::   ::: :*:.:******: .

PN40024_NCBI  878   R----NLEILSFVGCKGPPSASWLFPRRSSNSTG------FILHNLSGLCSLRKLDLSDC
PN40024_CRIBI 685   R----NLEILSFVGCKGPPSASWLFPRRSSNSTG------FILHNLSGLCSLRKLDLSDC
TNL2b         800   R----NLEILSFVGCKGPPSASWLFPRRSSNSTG------FILHNLSGLCSLRKLDLSDC
TNL2a         801   R----NLVILSLEGCKGPPSASWWFPRRSSNSTG------FRLHNLSGLCSLRTLNLSYC
N_protein     794   KTHVTKLLLWNMKNLVALPSSICRLKSLVSLSVSGCSKLESLPEEIGDLDNLRVFDASDT
                    :    :* : .: .  . **:   :    * *..         .::..* .** :: *

PN40024_NCBI   933   NLSDETNLSCLVYLSSLKDLYLC----ENNFVTLPNLSRLSRLERFRLANCTRLQ-ELPD
PN40024_CRIBI  740   NLSDETNLSCLVYLSSLKDLYLC----ENNFVTLPNLSRLSRLERFRLANCTRLQ-ELPD
TNL2b          855   NLSDETNLSSLVFLSSLEGLDLS----GNNFVMLPNFSRLSCLKCFRLENCTRLQ-ELPD
TNL2a          856   NLSDETNLSSLVLLSSLKELDLC----GNNFVTLPNLSRLSRLKHFWLKHCTRLQ-ELPD
N_protein      852   LILRPP--SSIIRLNKLIILMFRGFKDGVHFEFPPVAEGLHSLEYLNLSYCNLIDGGLPE
                     :   .  *.:: *..*  * :       :*   *  . *  *: : *  *. ::  **:

PN40024_NCBI   983   LPSSIVQVDARNCTS--LKNVSLRNVQSFLLKNRVIWDLNFVLALEILTPGS--------
PN40024_CRIBI  790   LPSSIVQVDARNCTS--LKNVSLRNVQSFLLKNRVIWDLNFVLALEILTPGS--------
TNL2b          905   LPSSIVQVDARYCTS--LKNVSLRNVQSFPLKNHAFRSLNIIAALHMLTPGS--------
TNL2a          906   LPSSIGQVDARNCTS--LKNVSLRNVQSFPLKSRSFRVPNLVLALEMLTPGS--------
N_protein      912   EIGSLSSLKKLDLSRNNFEHLPSSIAQLGALQSLDLKDCQRLTQLPELPPELNELHVDCH
                       .*: .:.    :   ::::.   .*   *:.  :   : :  *  *.*
```

Figure 18. LRR domain. In yellow, highlighted is the LRR domain predicted by Pfam in TNL2 alleles and paralogs. Single LRR motifs identified by LRRfinder are underlined. Amino acid substitutions are highlighted in green.

Regarding the LRR-kinase, the amino acid sequence is highly conserved between the proteins encoded by 21,076$^{Rpv3+/Rpv3+}$ and by the allele in PN40024 (Figure 19). They differ for only a deletion of Asp in position 643 of the protein encoded by 21,076$^{Rpv3+/Rpv3+}$. A nuclear localization sequence (NLS) was identified in *Xa21*, a LRR-kinase functionally characterized in rice. NLS allows to the intercellular kinase domain of Xa21 to be cleaved, released in the cytoplasm, and targeted to the nucleus, where it interacts with a transcription factor (Park and Ronald 2012). None of the alleles of the LRR-kinase in the *Rpv3* locus has conserved amino acids corresponding to the NLS motif of Xa21.

```
PN40024 LRR-k 691   LLVVSYRNFKHNESYAENELEGGKEKDLKWKLESFHPVNFTAEDVCNLEEDNLIGSGGTG
21,076  LRR-k 690   LLVVSYRNFKH-ESYAENELEGGKEKDLKWKLESFHPVNFTAEDVCNLEEDNLIGSGGTG
Xa21          720   AAALAILSSLYLLITWHKRTKKGAPSRTSMKGHPLVSYSQLVKATDGFAPTNLLGSGSFG
                      .::  .  :     .:. : *  .  . * ..: . .  .: . .:   **:***. *
```

Figure 19. Nuclear localization sequence (NLS) in rice Xa21 and amino acid changes between grapevines. Xa21 and two alleles of the LRR-kinase in the *Rpv3* locus were compared with ClustalW. A small portion of the amino acid sequence is reported in this figure, including a variable site between *Rpv3* alleles (in green background), the NLS in *Xa21* composed of the sequence HKR-KK (in cyan background), and the beginning of the kinase domain (in yellow). The kinase domain was identified with Pfam.

3.5.3 Truncated proteins encoded by transcript variants

The TNL2a gene produces multiple transcripts via alternative splicing. As shown by split read alignment, alternative splicing involves the first intron and the second exon. Combining the evidence of RNA-seq analysis and protein prediction, three different transcript variants of TNL2a gene were identified (Figure 20): (i) a transcript with regular intron splicing encoding a full-length protein, (ii) a transcript with intron retention, (iii) a transcript with exon skipping. In panel B, the transcript with intron retention translates into a predicted protein carrying an additional amino acid encoded by the retained intron, downstream of the TIR domain encoded by the first exon. The second triplet of retained intron (frame +2) codes for a stop codon. In the alternative transcript shown in panel C, the exon encoding the NB-ARC domain is skipped along with intron one and two. This second exon is composed of 1075

nucleotides, and its splicing causes a +1 frameshift in the translation of exon 3. This causes a premature stop codon after 18 amino acids encoded by the third exon.

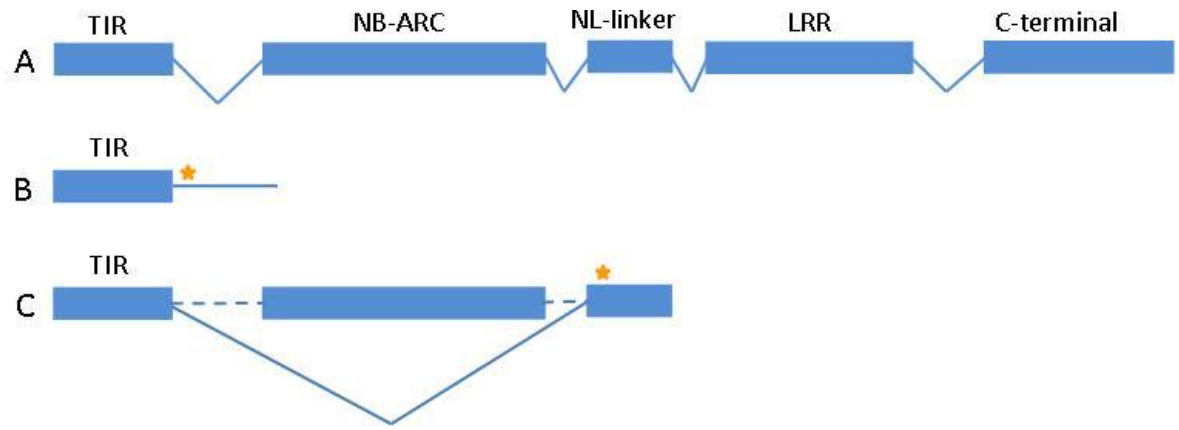


Figure 20. Transcript variants of the TNL2a gene. Spliced regions are shown as diagonal lines. Retained introns are shown as black lines. Stars indicate premature stop codons. In panel A, indicated is the full-length transcript. In panel B, indicated is the variant transcript with retention of the first intron. In panel C, indicated is the variant transcript with splicing of the first and second intron and skipping of the second exon.

3.6 Comparison between the Rpv3 haplotype and vinifera haplotypes

Assembled sequence of the resistance haplotype across the *Rpv3* locus is composed of a continuous nucleotide sequence of 718,212 bp. In order to compare the resistance haplotype with a susceptible haplotype in *V. vinifera*, the assembled sequence of the *Rpv3* locus was aligned against the genome reference PN40024. Comparison showed that the resistance haplotype spans a region that is much larger than 700kb in PN40024, with approximately 1.1 Mbp on chromosome 18 of PN40024, starting from 25,256,951 and ending in position 26,356,274 (Figure 21). GEvo comparative sequence alignment tool (genomevolution.org/CoGe/GEvo.pl) was used to identify shared regions between the two haplotypes as well as unique regions involved in presence/absence variation.

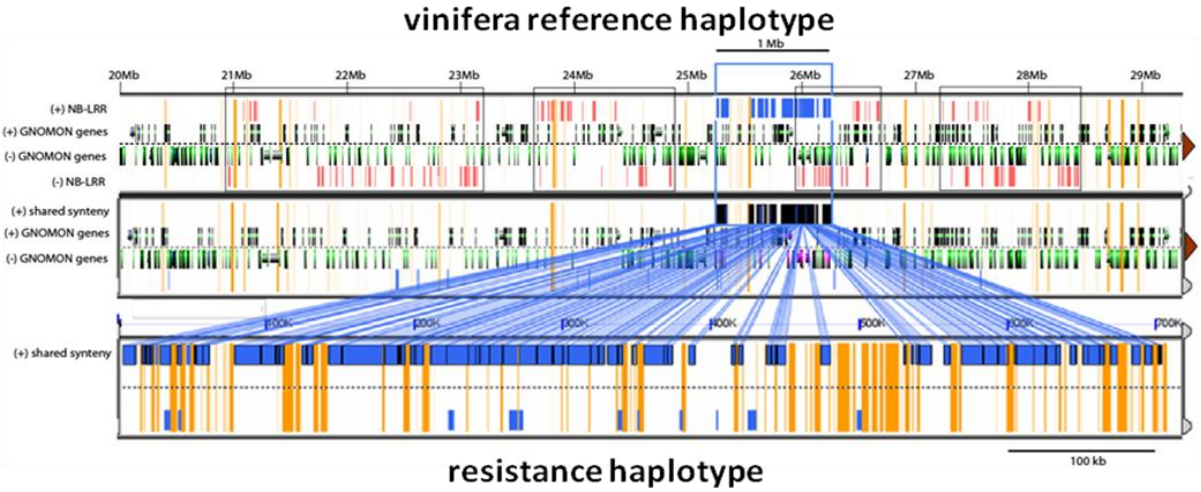


Figure 21. Assembled sequence of the resistance haplotype compared with the vinifera haplotype of PN40024. Graphic from GEvo shows in the lower part the *Rpv3* sequence and its counterpart on chromosome 18 of PN40024 in the upper part. The blue box shows the portion of the susceptible haplotype on PN40024 that correspond to the *Rpv3* locus. Blue connectors show the shared regions between the two haplotypes. Orange bars indicate sequence gaps.

3.7 Conserved DNA

Regions of conserved DNA represent a limited part in each haplotype, and it amounts to only 246 kb (Figure 22). Two main blocks are conserved between the two haplotypes with little structural variation. The first block starts from position 25,623,371 of the PN40024 genome and ends in position 25,974,875. The second starts from position 26,027,331 and ends in position 26,107,454. These two regions are preceded, interrupted and followed by clusters of duplicated genes and pseudogenes (boxed by dotted rectangles in Figure 22). In the conserved intervals, the two haplotypes have an average nucleotide identity of 96.4 % and 97.3 %, including SNPs and small indels. Regions of conserved DNA are populated by single-copy genes and most of them are conserved between the two haplotypes (in purple in the Figure 22). Based on gene prediction and tBlastx comparison with protein databases, the following genes are present in the region of conserved DNA upstream the TNL candidate genes: Acyl-CoA-N-acetyltransferase, stromal 70-kDa heat shock protein, protein phosphatase 1, amino acid permease, adenylate kinase, serine/threonine protein kinase. In the region downstream the *Rpv3* locus, two genes were predicted, a TIC 40 chloroplastic like and the first of the member of another TIR-NB-LRR gene

cluster. All the genes predicted in the conserved regions show colinearity with PN40024.

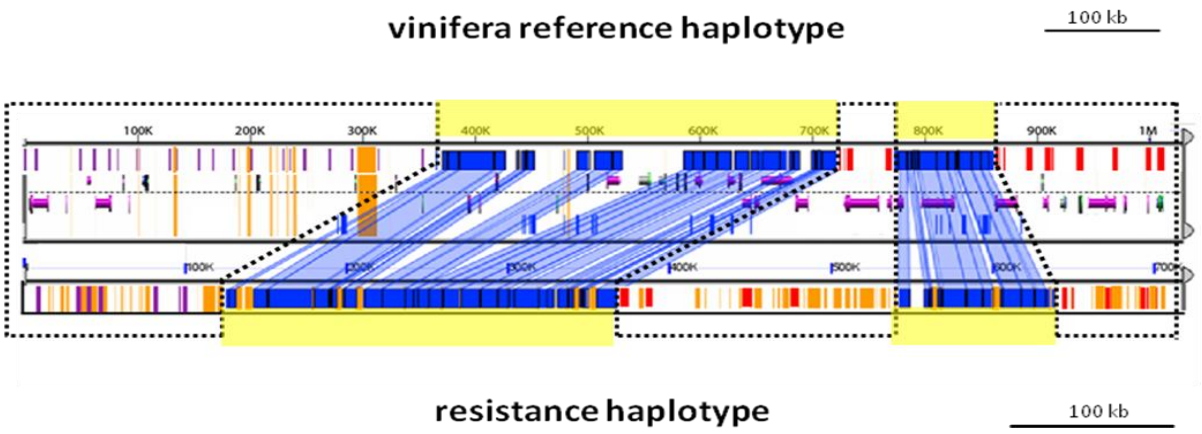


Figure 22. Conserved DNA between a susceptible haplotype and the resistance haplotype. The blue connectors show the regions of the PN40024 haplotype that are shared with the resistance haplotype. The purple elements represent gene models shared between the two haplotypes, while in green represented are non-shared genes. In red, represented are NB-LRR genes. Orange bars indicate sequence gaps.

3.8 Non-conserved DNA

Non-conserved DNA consists mostly of clusters of genes that are present in both haplotypes, but with a different number of gene-copies (Figure 23).

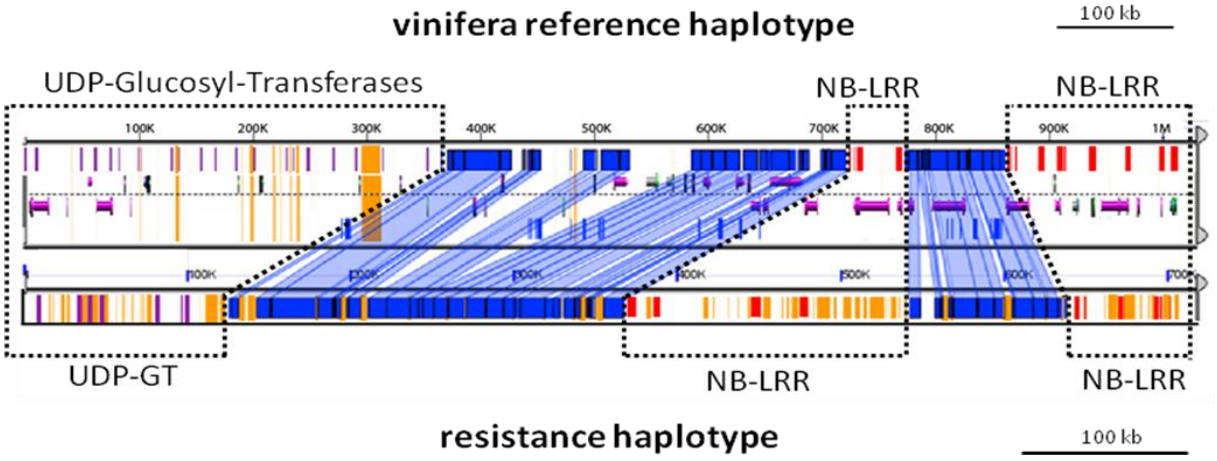


Figure 23. Non-conserved DNA between a susceptible haplotype and the resistance haplotype. The dotted boxes indicate the regions with structural variation between haplotypes. Purple and red ticks indicate duplicated genes. The blue connectors show the regions of the PN40024 haplotype that are shared with the resistance haplotype. Orange bars indicate sequence gaps.

The first region of non-conserved DNA extends from 25,256,948 to 25,623,371 in the susceptible haplotype and contains a cluster of UDP-Glucosyl-Transferases, spanning approximately 360 kb. Short portions of intergenic space are conserved between the two haplotypes and, in the susceptible haplotype, they flank a region of 200 kb that is deleted from the resistance haplotype. This region is indicated by the purple box in Figure 24. That's why the whole region containing the UDP-GT gene cluster is shorter in the resistance haplotype than in the vinifera haplotype and has a lower number of gene copies. The second cluster of genes that corresponds to genetic interval of the *Rpv3* locus in the resistance haplotype covers a region that is larger in the resistance haplotype than in the vinifera reference, with 170 kb versus only 67 kb. In this portion, a cluster of NB-LRR genes is present and the number of gene copies is lower in the vinifera haplotype than in the resistance haplotype. The third cluster of duplicate genes covers a region downstream the genetic interval of the *Rpv3* locus that is larger in the vinifera haplotype than in the resistance haplotype, and contains arrayed NB-LRR genes.

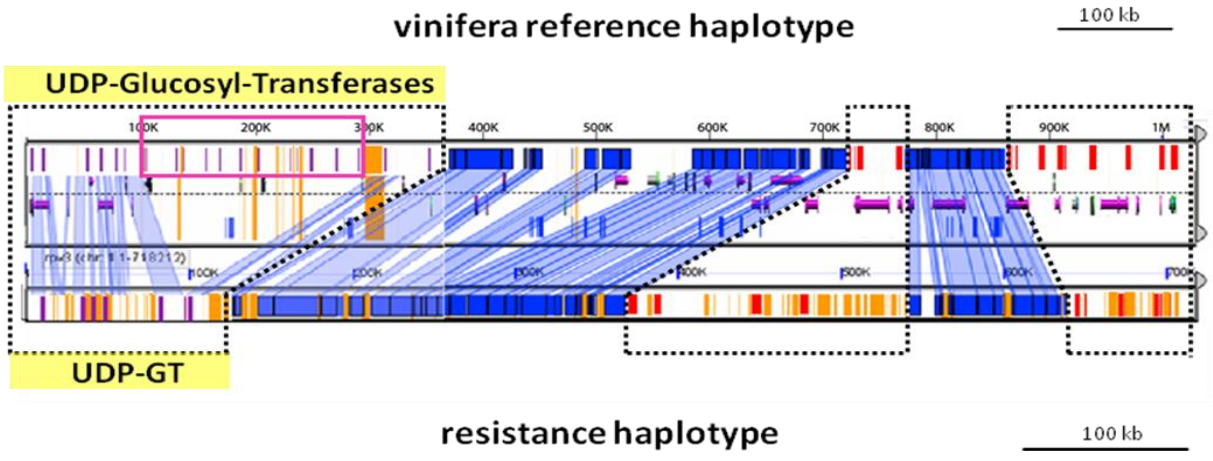


Figure 24. UDP-Glucosyl-Transferases gene cluster. Blue connectors indicate conserved regions. Pink box shows a region present in the vinifera haplotype and absent from the resistance haplotype, flanked by shared intergenic DNA, indicated by blue connectors. Orange bars indicate sequence gaps.

Extra-DNA content in the resistance haplotype corresponds to the region where the markers that co-segregate with the trait are located. Gene content was compared between the resistance haplotype and a susceptible haplotype. Differences were found in the number of gene copies (Figure 25). In the vinifera haplotype, two copies

of TNL genes are predicted, while in the resistance haplotype there are three full-length copies and an additional truncated copy, composed only by the TIR domain and lacking both the NB-ARC and LRR domains. One of these full-length copies (TNL1) falls short of the genetic interval of *Rpv3* and is allelic to the LOC100251613 TNL1 in PN40024. In the resistance haplotype, the TIR of TNL1 is duplicated over a short distance. Two of the full-length copies (TNL2a and TNL2b) fall within the genetic interval of *Rpv3* and are allelic to LOC100255177 TNL2 in PN40024. The block of DNA, including TNL2 and the Copia TE immediately upstream, is duplicated in the resistance haplotype compared to the PN40024 haplotype. This Copia is the only TE that is shared between the two haplotypes. The variation in DNA content is also accounted for by the presence or absence of several transposable elements in either haplotype. Four transposable elements were identified on the resistance haplotype that are absent from the *V. vinifera* haplotype. Four TEs are uniquely present in the vinifera haplotype, including a Copia inserted in the coding sequence of TNL1. In addition to the TE that caused the interruption in scaffolding between scaffolds sc_19 and sc_2088 and the Copia involved in the segmental duplication of TNL2, a Gypsy is inserted in the resistance haplotype upstream TNL2a, a Copia is present between TNL2a and TNL2b, and another Copia is inserted upstream the LRR-kinase.

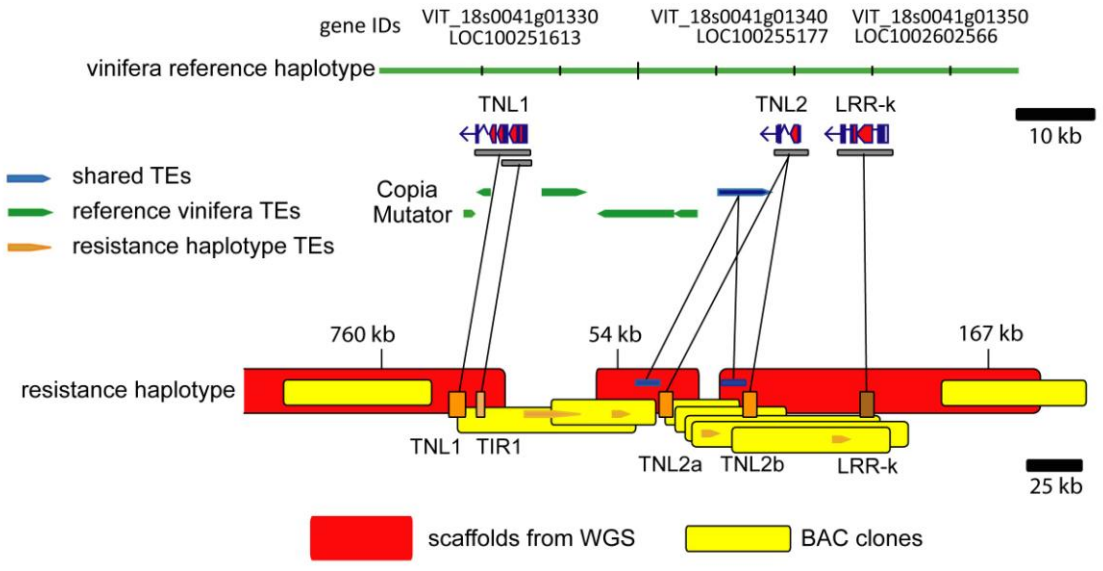


Figure 25. Extra-DNA content in the resistance haplotype. The green line in the upper part of the figure represents the *V. vinifera* haplotype in PN40024. In the lower part the resistance haplotype. Starting from the left-hand side in the diagram of the resistance haplotype, the orange rectangles represent TNL genes. The brown rectangle represents the LRR-kinase. The black lines connect shared genes and TEs shared showing the highest sequence identity

between haplotypes. Blue bars indicate TEs shared between the two haplotypes. In green, the TEs that are present only in the vinifera haplotype, in brown the TEs present only in the resistance haplotype.

3.9 Haplotype diversity in single-copy regions

Haplotype diversity was analyzed in single-copy regions by SNP analysis. Illumina reads of 50 varieties of *V. vinifera* and reads of 21,076$^{Rpv3+/Rpv3+}$ were aligned against the PN40024 reference genome. SNPs were called in shared regions identified by GEvo (Figure 26). SNP density between viniferas ranging from 0 to 7 SNPs/kb, depending on the genomic interval considered. SNP density in the resistance haplotype is well-above the range of variation observed in vinifera, consistently along the locus, and consistently with what is expected for a non-vinifera haplotype SNP density reaching a maximum of 10 SNPs/kb.

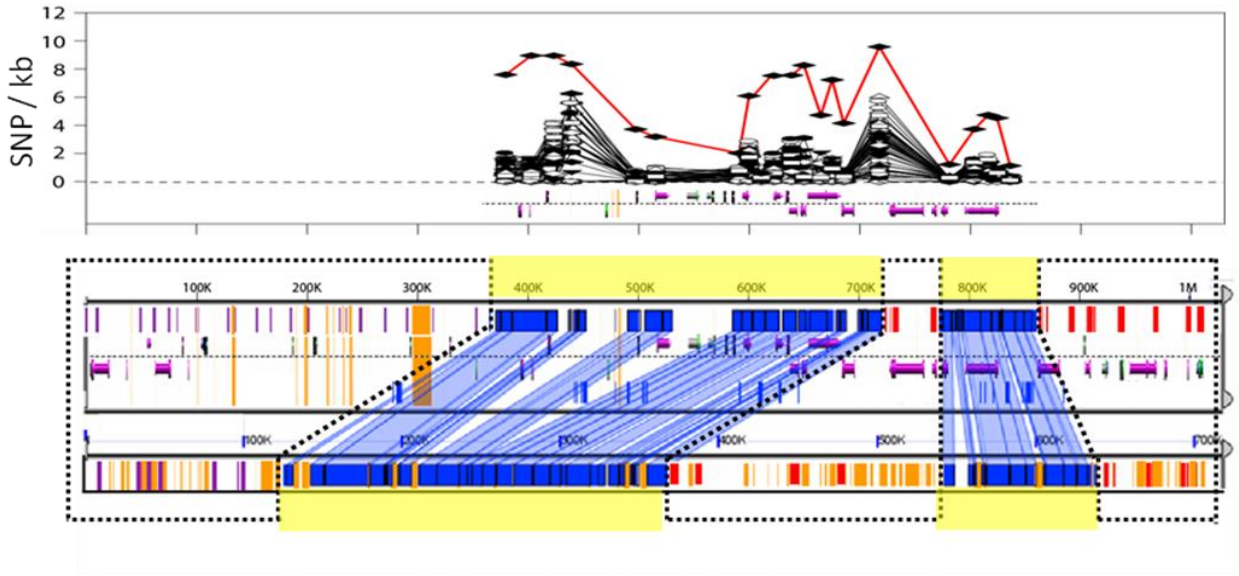


Figure 26. SNPs frequency in regions of conserved DNA in varieties of *Vitis vinifera* and in the resistance haplotype. In the upper panel, with the red line plotted is the SNPs density between the resistance haplotype and the reference genome PN40024. With black lines plotted are intraspecific SNP frequencies between varieties of vinifera and PN40024. SNP density is plotted as an average value in the middle position of each conserved fragment, as identified in the GEvo comparison between the reference PN40024 (upper part of the lower panel) and the resistance haplotype (lower part of the lower panel). The blue connectors show the regions of the PN40024 that are shared with the resistance haplotype.

In order to estimate the divergence between the resistance haplotype and susceptible haplotypes, we considered the total number of SNPs across the locus in each variety,

including 21,076$^{Rpv3+/Rpv3+}$, compared to the reference genome PN40024 in the conserved region of 246 kb. We also considered the number of private SNPs in each variety, that are SNPs uniquely present in a single variety compared to PN40024. The total number of SNPs in 21,076$^{Rpv3+/Rpv3+}$ is four-fold higher, or more, than in viniferas with a total number of 1,238 SNPs across the whole conserved region (Figure 27). Also the number of private SNPs is much higher in 21,076$^{Rpv3+/Rpv3+}$ than in viniferas in the conserved region, with a number of 718 private SNPs. Thus, 21,076$^{Rpv3+/Rpv3+}$ is highly diverse from varieties of *V. vinifera* in the region of the *Rpv3* locus.

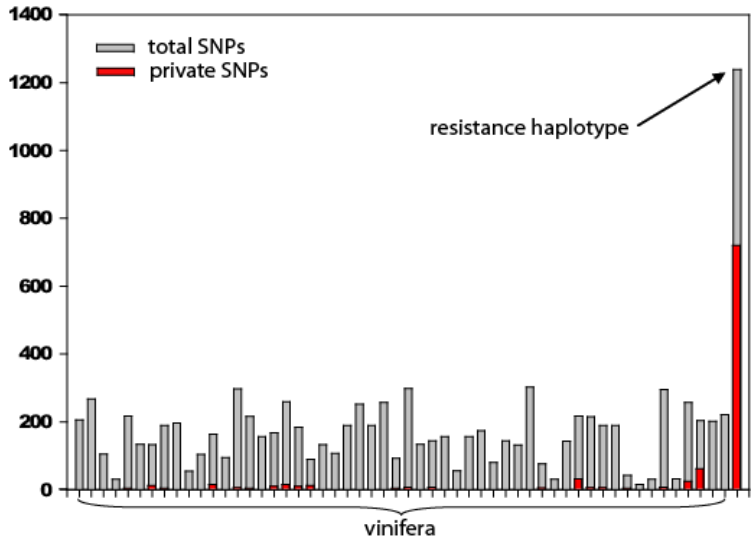


Figure 27. SNPs in varieties of *Vitis vinifera* and in 21,076$^{Rpv3+/Rpv3+}$ across 246 kb of conserved DNA in the *Rpv3* locus with respect to reference genome of PN40024. Total number of SNPs is reported by grey histograms, the fraction of private SNPs – that is polymorphic sites found only in a given variety – is indicated by red histograms.

3.10 Introgressed regions in 21,076 $^{Rpv3+/Rpv3+}$

The individual 21,076$^{Rpv3+/Rpv3+}$ derives from selfing of the variety Bianca, a hybrid grape with a complex pedigree including several North American *Vitis* species. The parent combination that generated Bianca includes one pure vinifera parent (Bouvier). Thus, wild haplotypes are expected to be present invariably in a heterozygous state in Bianca, combined with a haplotype of the European *V. vinifera*. The composition of the wild and vinifera genome was investigated in 21,076$^{Rpv3+/Rpv3+}$. We expect four cases in 21,076$^{Rpv3+/Rpv3+}$: (i) chromosomal segments homozygous for a vinifera haplotype, (ii) chromosomal segments heterozygous for two different vinifera haplotypes, (iii) chromosomal segments carrying an introgressed wild haplotype

combined with a vinifera haplotype on either homolog; (iv) chromosomal segments with a wild haplotype in a homozygous state. In order to identify the genome composition of this individual, we aligned Illumina reads of 21,076$^{Rpv3+/Rpv3+}$ against the *V. vinifera* genome reference PN40024 and called SNPs. Raw SNPs were filtered out for variable positions in repeated regions, transposable elements, small indels and SSR intervals. SNP density along the chromosomes is represented in the graph of Figure 28. Homozygous SNPs are plotted separately from heterozygous SNPs. Homozygous SNPs are indicated in red. Due to the process of selfing that generated 21,076$^{Rpv3+/Rpv3+}$ vast regions are homozygous for a haplotype different from the reference and appear in the graph as covered by homozygous SNPs only. Heterozygous SNPs are indicated in blue. Heterozygous regions of 21,076$^{Rpv3+/Rpv3+}$ sharing one haplotype with PN40024 appear in the graphs as covered by heterozygous SNPs only. Heterozygous regions of 21,076$^{Rpv3+/Rpv3+}$ for two haplotypes, both different from the PN40024 haplotype, appear in the graph as covered by a mixture of heterozygous and homozygous SNPs. Regions without SNPs, like the lower arm of chr2, represent homozygous segments identical to PN40024.

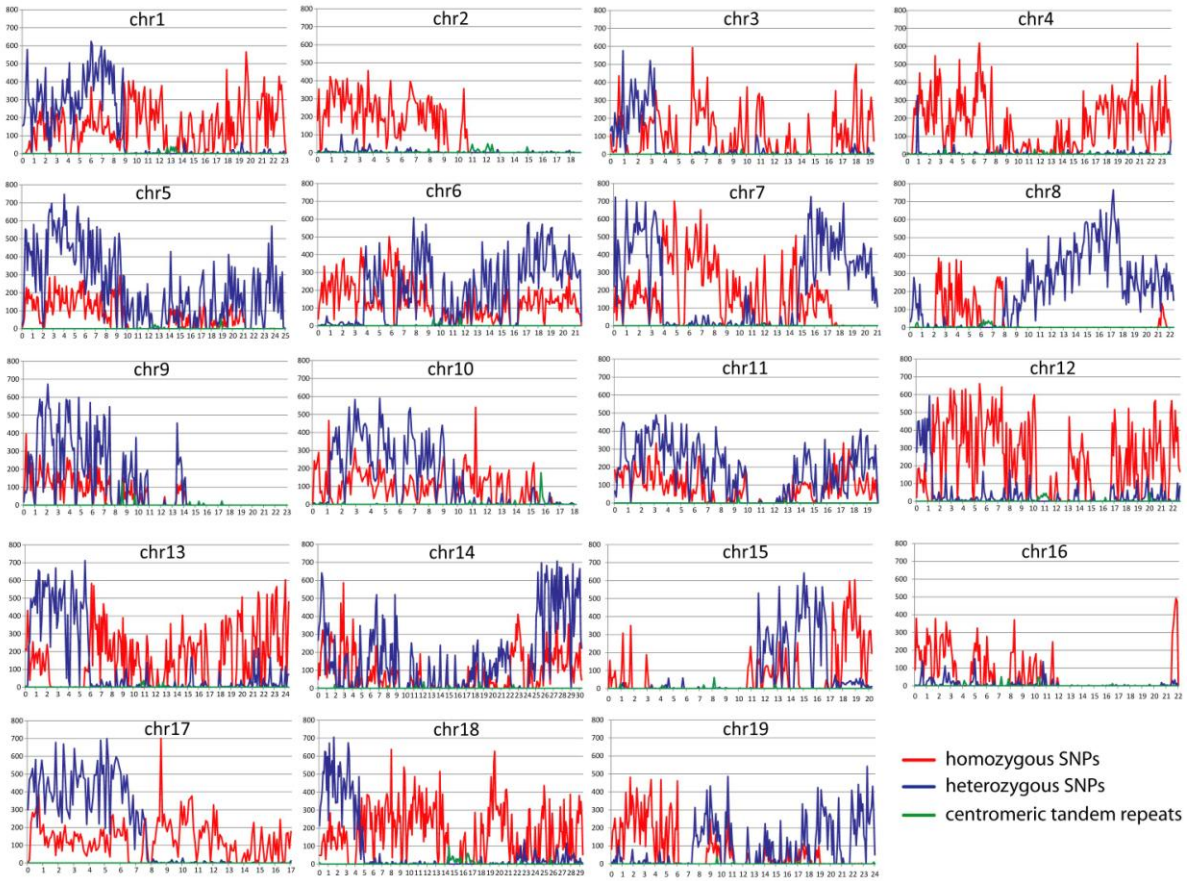


Figure 28. SNP density of a 21,076$^{Rpv3+/Rpv3+}$ compared to the reference genome PN40024. Reads were aligned against PN40024 using BWA. Variants were called using GATK with default parameters. SNPs in repetitive regions identified by ReAS and RepeatMasker, in annotated transposable elements, and in CNV regions were filtered out. Plots report the number of SNPs in adjacent windows of 100 kbp on the y-axis. The x-axis indicates chromosome length in Mbp. Homozygous SNPs are plotted in red, heterozygous SNPs are plotted in blue.

Based on SNP plots, we inferred the profile of homo-/heterozygosity per each chromosome (Figure 29). Homozygous chromosomal segments are indicated in red, and heterozygous regions are in blue. 21,076$^{Rpv3+/Rpv3+}$ is homozygous across the entirety of the length of chr2, chr4 and chr16. Conversely, 21,076$^{Rpv3+/Rpv3+}$ is heterozygous across the entirety of the length of chr5 and chr11. Homozygous segments amount to 54.7 % of the total length of the chromosomes. Heterozygous segments amount to 45.3 % of the total length of the chromosomes.

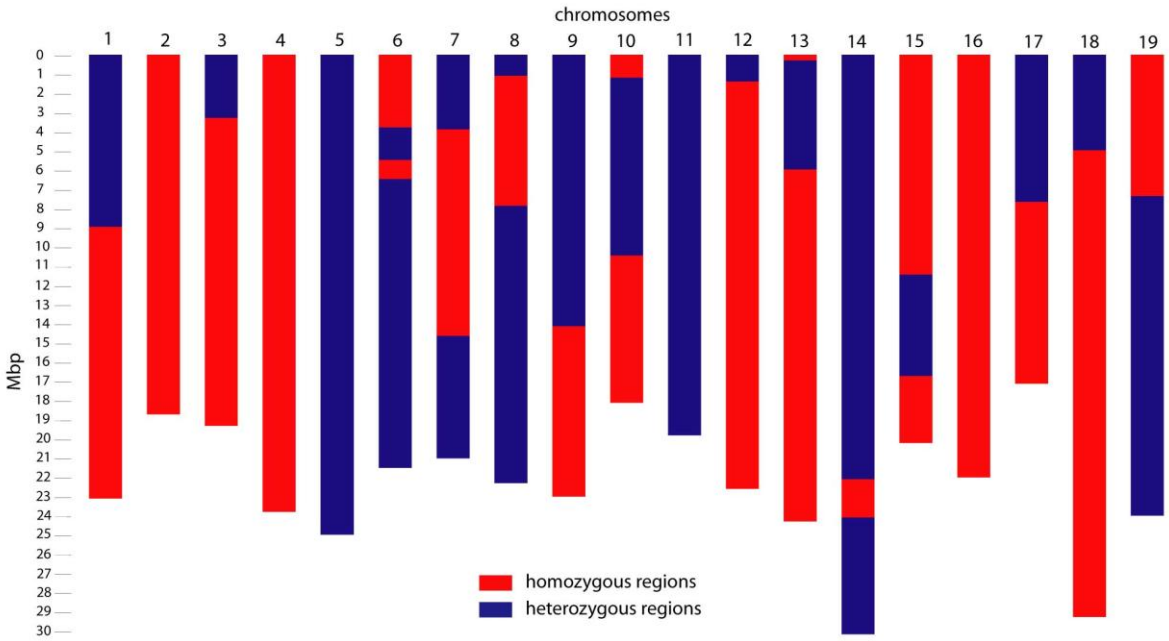


Figure 29. Homozygous (red) and heterozygous (blue) regions in the 21,076$^{Rpv3+/Rpv3+}$ genome. The y-axis indicates the length of the chromosomes in Mbp.

We sorted out the SNPs of 21,076$^{Rpv3+/Rpv3+}$ shared with any variety of *V. vinifera*, obtaining a genome-wide inventory of private SNPs of 21,076$^{Rpv3+/Rpv3+}$. If residual SNPs are private SNPs carried by unrelated haplotypes, we expect a non-random distribution. In fact, private SNPs are expected to cover those regions that have been introgressed in 21,076$^{Rpv3+/Rpv3+}$ from a non-vinifera grapevine. The chromosomal plot showing the genomic distribution of private SNPs is shown in Figure 30.

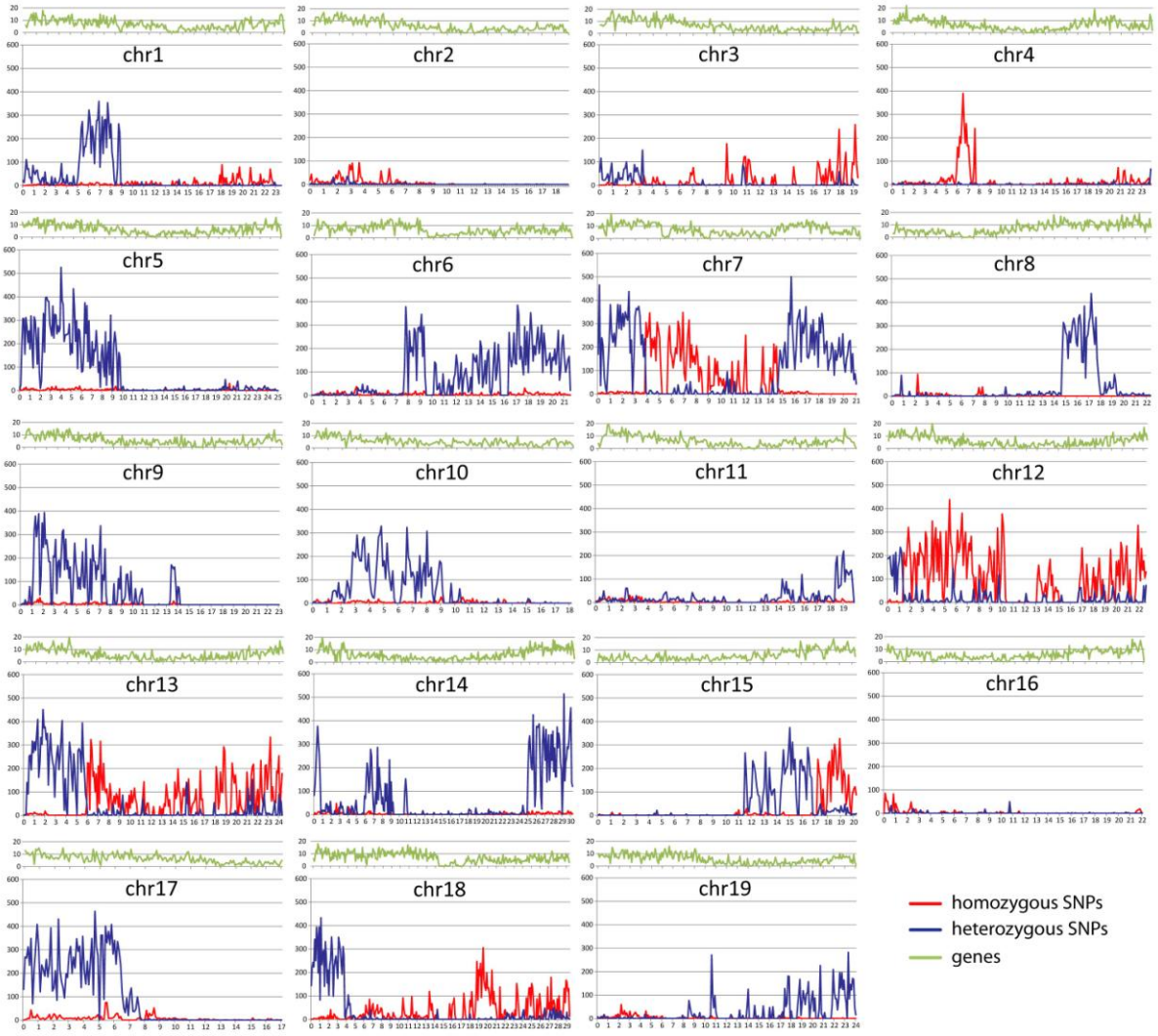


Figure 30. SNP density of 21,076$^{Rpv3+/Rpv3+}$ compared to the reference genome PN40024, retaining only private SNPs not shared with vinifera varieties. Plots report the number of SNPs in adjacent windows of 100 kbp on the y-axis. The x-axis indicates chromosome length in Mbp. Homozygous private SNPs are plotted in red, heterozygous private SNPs are plotted in blue. The stacked plot in green reports gene density along the chromosomes in adjacent windows of 100 kbp.

By combining the information on homo-/heterozygosity and the distribution of non-vinifera SNPs, we classified each chromosomal segment of 21,076$^{Rpv3+/Rpv3+}$ in four classes: vinifera/vinifera homozygous, vinifera/vinifera heterozygous, non-vinifera homozygous, vinifera/non-vinifera heterozygous (Table 12). Vinifera/vinifera homozygous segments account for 38.8 % of total chromosome length, vinifera/vinifera heterozygous segments for 19.1 %, non-vinifera homozygous segments for 15.9 %, and vinifera/non-vinifera heterozygous for 26.2 %. For instance, chr2 and chr16 are entirely homozygous for a vinifera haplotype, chr3 is partially

heterozygous and partially homozygous for vinifera haplotype. These three chromosomes are the only chromosomes with pure vinifera alleles. All others carry introgressed segments in either a heterozygous or homozygous state. Chromosomes 7, 12, 13 and 18 are particularly rich in introgressed DNA. Considering the whole genome of 21,076$^{Rpv3+/Rpv3+}$, 58 % of the haploid genome has exclusively vinifera alleles on both homologs, 70 % of the total length of the diploid genome (counting separately each homolog) is vinifera and 30 % is introgressed.

Table 12. Percentage of introgressed and vinifera genome in 21,076$^{Rpv3+/Rpv3+}$. For every chromosome indicated is the percentage of chromosomal segments homozygous for a vinifera haplotype, heterozygous for two vinifera haplotypes, homozygous for a introgressed haplotype, heterozygous for a vinifera and non-vinifera haplotype. In the last column, reported is the total length of chromosome segments based on the above mentioned classification.

	chr1	chr2	chr3	chr4	chr5	chr6	chr7	chr8	chr9	chr10
vinifera/vinifera homozygous	61	100	83	93	0	22	0	30	39	49
vinifera/vinifera heterozygous	22	0	17	0	61	13	0	55	5	16
non-vinifera homozygous	0	0	0	7	0	0	51	0	0	0
vinifera/non-vinifera heterozygous	17	0	0	0	39	64	49	15	57	35

	chr11	chr12	chr13	chr14	chr15	chr16	chr17	chr18	chr19
vinifera/vinifera homozygous	0	0	0	7	57	100	55	43	31
vinifera/vinifera heterozygous	92	0	0	57	0	0	0	0	13
non-vinifera homozygous	0	94	77	0	17	0	0	40	0
vinifera/non-vinifera heterozygous	8	6	23	37	26	0	45	17	56

	tot (nt)
vinifera/vinifera homozygous	166,000,000
vinifera/vinifera heterozygous	81,700,000
non-vinifera homozygous	67,900,000
vinifera/non-vinifera heterozygous	112,000,000
tot (haploid genome length)	427,600,000

The genomic profile illustrated in Figure 31 summarizes the regions introgressed from the wild genome into the vinifera background as a result of historical backcrosses that led to Bianca and, finally, of the selfing that led to 21,076$^{Rpv3+/Rpv3+}$. Indicated in yellow are the regions that were inherited from vinifera, while violet indicates the regions for which at least one haplotype was inherited from a wild ancestor. As much as 42 % of the haploid genome of 21,076$^{Rpv3+/Rpv3+}$ is non-vinifera. In Bianca, all these regions have a vinifera allele derived from Bouvier. Thus, 21 % or more of the total length of the diploid genome in Bianca (counting each homolog separately) is introgressed. Chr18, where the *Rpv3* locus is located, shows a heterozygous introgression across the first 4 Mbp, the remainder of the chromosome is homozygous for a vinifera haplotype across 13 Mbp in the middle of the chromosome, and homozygous for a introgressed haplotype in terminal segment of 9 Mbp, in which the resistance locus is included.

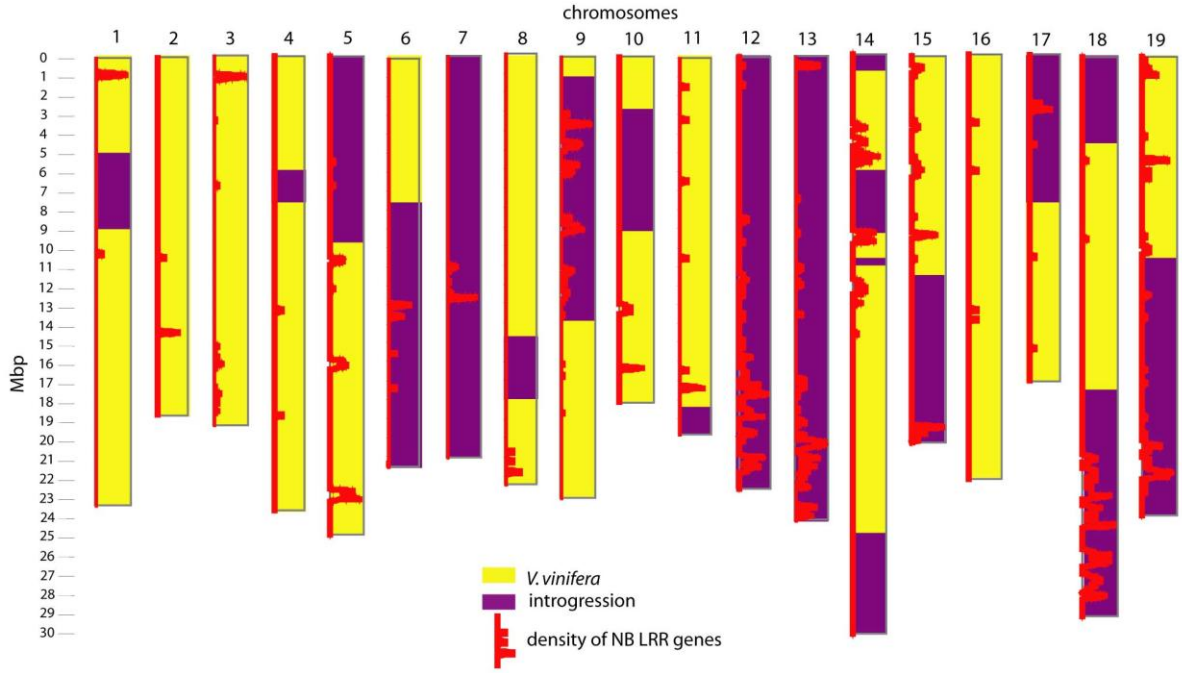


Figure 31. Extent of introgressed regions in 21,076$^{Rpv3+/Rpv3+}$. The regions of the genome that are introgressed from the wild ancestor are indicated in violet. Those in yellow are regions that derive from vinifera genotype. The y-axis indicates the length of chromosomes in Mbp. Density of NB-LRR genes in the grapevine reference genome is indicated by the red track (see the second part of the thesis for more details).

The location of introgressed regions was compared with the chromosomal distribution of NB-LRR genes. We considered all genes, pseudogenes and homologous gene fragments, which amount to a total of 821 features in the reference genome of PN40024 and show a non-random distribution. Of these, 708 are located on scaffolds

assigned to chromosomes. The largest NB-LRR gene clusters are located across vast segments of chromosomes 9, 12, 13, and 18. Notably, 21,076$^{Rpv3+/Rpv3+}$ retained an entire homolog of both chr12 and chr13 from its non-vinifera ancestors. The introgressed regions on both chr9 and chr18 spanned exactly the NB-LRR rich regions on those chromosomes. As a result of this, as many as 480 out of 708 NB-LRR genes are located across regions for which 21,076$^{Rpv3+/Rpv3+}$ has at least one non-vinifera haplotype. In spite of the non-pure vinfera genome accounting for 42 % of the haploid genome length, that 42 % contains 68 % of all NB-LRR genes, pseudogenes and homologous gene fragments. Assuming synteny across *Vitis* species, 21,076$^{Rpv3+/Rpv3+}$ and its parent Bianca are highly enriched for wild NB-LRR alleles across the genome, in addition to the *Rpv3* resistant allele.

3.11 Gene expression of candidate genes

3.11.1 Global gene expression

We aligned RNA-Seq reads obtained from inoculated leaves of 21,076$^{Rpv3+/Rpv3+}$ against the grapevine genome reference. Out of the 30,066 gene models predicted by the CRIBI V1 annotation, 21,752 (78.8 %) showed evidence of transcription. The relative abundance of transcripts was measures using Cufflinks and expressed in Fragments Per Kilobase of exon per Million fragments mapped (FPKM). The distribution of the level of expression of all transcribed genes is shown in Figure 32, panel A. The values of transcript abundance were $\log_2$-transformed ($\log_2$ FPKM). We identified 467 CRIBI V1 gene models that showed sequence similarity with NB-LRR genes. Of these, 367 (61.8 %) showed evidence of transcription in inoculated leaves of 21,076$^{Rpv3+/Rpv3+}$. The distribution of the level of expression of all transcribed NB-LRR genes is shown in Figure 32, panel B. The modal class of transcript abundance of the subset of NB-LRR genes has a lower value of $\log_2$ FPKM compared to the entire set of genes. RNA-Seq reads obtained from inoculated leaves of 21,076$^{Rpv3+/Rpv3+}$ were also aligned against the scaffolds of the *de novo* assembly of 21,076$^{Rpv3+/Rpv3+}$. A total of 21724 transcripts were identified by Cufflinks. The distribution of the values of transcript abundance is shown in Figure 32, panel C.

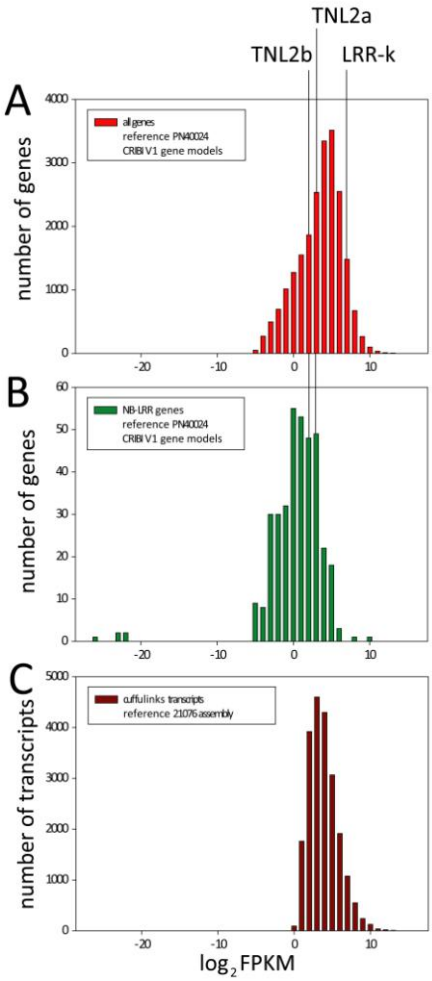


Figure 32. Distribution of transcript abundance. Genes were grouped in frequency classes of transcript abundance. FPKM data were log2-transformed. The class of grouped data is 1 FPKM. Panel A included all genes of the grapevine reference genome, panel B included the subset of NB-LRR genes, panel C included the transcripts predicted in the de-novo assembly of 21,076$^{Rpv3+/Rpv3+}$.

3.11.2 Transcript abundance of candidate genes in the *Rpv3* locus

The LRR-Kinase is expressed at a high level (166.1 FPKM). This LRR-Kinase falls in the 1st percentile of the most expressed genes. NB-LRR genes are expressed at much lower levels. TNL2a was expressed at 7.9 FPKM, falling in the 55th percentile of the most expressed genes. TNL2b was expressed at 3.6 FPKM, falling in the 68th percentile of the most expressed genes. The NB-LRR genes in the *Rpv3* locus are expressed at relatively high levels compared to other NB-LRR copies. TNL2a falls in 10th percentile of the most expressed NB-LRR genes. TNL2b falls in 22th percentile of the most expressed NB-LRR genes.

3.12 Cloning and characterization of candidate genes

In order to characterize the candidate genes for downy mildew resistance in the *Rpv3* locus, full-length sequences were amplified and cloned into an expression vector. Deep sequencing of transcript by RNA-Seq facilitated the correct identification of 5'- and 3'-UTR regions in order to target the most suitable regions for primer design and amplify each target gene selectively. Primers were designed in 5'- and 3'-UTRs. In Table 13, indicated are the primer combinations specific for the amplification of each candidate gene, the annealing sites of the primers, relative to the start and the stop codon, and the expected length of the amplicon based on the predicted gene model. Candidate genes have long coding sequences and are not expressed at high levels. TNL genes are also present in several copies with an high sequence identity. These adversities made it difficult to clone this kind of genes. To circumvent this problem, genes were amplified from genomic DNA or from a BAC clone using short and specific oligonucleotides primers. PCR products were preliminarily cloned into a TA vector. Following the identification of colonies carrying the target insert, the isolated gene was re-amplified from the recombinant plasmid DNA using longer oligonucleotide primers, suitable for cloning into an expression vector.

Table 13. Annealing sites of primer combinations for the specific amplification of each candidate gene. Annealing sites are indicated in terms of distance from the start and the stop codon.

Primer combination / Target gene	Start codon	Stop codon	Predicted size of the amplicon
TNL2a	-490 bp	+196 bp	5022 bp
TNL2b	-421 bp	+160 bp	4962 bp
LRR-kinase	-1531 bp	+103 bp	6680 bp

3.12.1 Gene amplification and TA cloning

PCR amplification from different templates was attempted for each candidate gene. All candidate genes are expressed at relatively low levels. Full-length amplification

from cDNA was weak and the amount of PCR products after agarose purification was limiting for cloning. The full-length sequence of two candidate genes (TNL2b and LLR-k) was present in two BAC clones. The third candidate gene (TNL2a) was truncated by the BAC end in the only BAC clone that partially covered its coding region. Thus, genomic DNA of 21,076$^{Rpv3+/Rpv3+}$ was used as a template for the amplification of TNL2a, whereas BAC clones were used as a template for the amplification of TNL2b and LRR-kinase. PCR was conducted using BAC DNA extracted from clone 50B08 and clone 66A20, respectively (Table 14).

Table 14. DNA template used for PCR amplification of candidate genes and length of the amplicons as assessed by gel electrophoresis before TA cloning.

Gene	Template	Approximated size of the amplicons
TNL2a	Genomic DNA 21,076$^{Rpv3+/Rpv3+}$	5000 bp
TNL2b	Bianca BAC_50B08	5000 bp
LRR-kinase	Bianca BAC_66A20	7000 bp

Several *E. coli* colonies were obtained after transformation with pCR-XL-TOPO vector. Colony PCR with M13 primers flanking the cloning site proved that a fragment of the expected size had been inserted in the vector of 16 out of 20 colonies for LRR-kinase, 80 out of 96 colonies for TNL2a and 9 out 15 colonies for TNL2b.

3.12.2 Gene sequencing

Inserts of the expected size were sequenced by Sanger sequencing. M13 primers flanking the cloning sites as well as eight internal primers were used for entirely covering the sequence of TNL genes. M13 primers flanking the cloning sites as well as seven internal primers were used for covering entirely the sequence of the LRR-kinase. Sequence analysis revealed three evidences. First, we assessed that the consensus sequence of the inserts was 100 % identical with the gene sequence reconstructed by *de novo* assembly of NGS reads. Second, TNL2a was amplified from genomic DNA but the primers were copy-specific because all cloned inserts obtained from these PCR

products aligned to the gene sequence of TNL2a. Third, we selected one colony per candidate gene with a insert 100 % identical to the consensus sequence, without nucleotide misincorporation.

3.12.3 Cloning in an expression vector

To proceed with the functional analysis of candidate genes, the full-length sequence was reamplified from the insert cloned into pCR-XL-TOPO vector and moved into the Gateway recombination system. A recombinase recognizes specific sequences, called *att*-sites, to catalyze recombination between PCR products and the donor vector. To this end, gene-specific 5'-UTR and 3'-UTR primers were extended at 5'-terminal end with *att*B sequences for directional cloning. An aliquot of the PCR product amplified with the *att*B-primers from recombinant pCR-XL-TOPO plasmid DNA is shown in Figure 33, prior to recombination in the donor vector pDONR207. PCR products were obtained in the following concentrations, TNL2a = 451,7 ng/µl, TNL2b = 452,2 ng/ µl, LRR-kinase= 446,3 ng/µl, which were suitable for performing the BP reaction with the donor vector pDONR207.

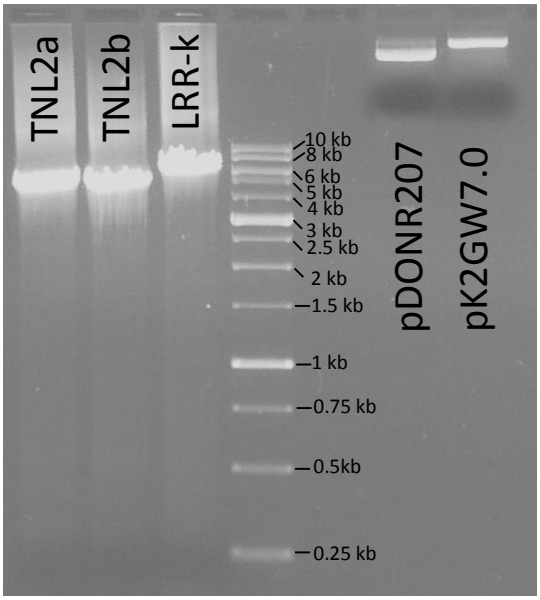


Figure 33. Gel electrophoresis of amplified candidate gene cloning into the donor vector pDONR207. Plasmid DNA of donor vector pDONR207 and destination vector pK2GW7.0 is loaded in adjacent lanes.

Several *E. coli* colonies were obtained after transformation with the entry vector pDONR207. We tested 16 colonies per gene. Colony PCR with LA-LB primers flanking

the cloning site proved that a fragment of the expected size had been inserted in the vector of all 16 colonies for LRR-kinase, 14 out of 16 colonies for TNL2a and 12 out 16 colonies for TNL2b. Inserts of the expected size were sequenced by Sanger sequencing, using LA-LB primers flanking the cloning site and internal primers. Sequence analysis provided support for selecting two colonies per candidate gene with an insert 100 % identical to the target gene sequence, without nucleotide misincorporation. Plasmid DNA of the entry vector of these colonies was used for LR reaction with the destination vector pK2G7W.

Several *E. coli* colonies were obtained after transformation with the expression clone pK2G7W. We tested 3 colonies per gene. Plasmid DNA was Sanger sequenced using LA-LB primers flanking the cloning site. Sanger reads of the insert ends were aligned against the target gene sequence using CLC Genomic Benchwork providing evidence for the full-length transfer of the insert into the expression clone. The expression clone pK2G7W has a 35S Cauliflower Mosaic Virus (CaMV) promoter/terminator and kanamycin resistance cassette and *ccdB* gene. One colony per target gene (TNL2a, TNL2b, and LRR-k) of the expression clone was delivered to a facility of grapevine genetic engineering for *Agrobacterium*-mediated transformation of embryogenic calli.

4 Discussion

P. viticola is a biotrophic oomycete that causes downy mildew in *V. vinifera* and in most accessions of other *Vitis* species. The *Rpv3* locus is a major determinant of downy mildew resistance in descendents of North American grapevines produced by intentional hybridization during historical breeding. A resistance haplotype was introgressed in highly *V. vinifera* backgrounds, after several generations of backcrosses and intercrosses. This haplotype is shared by numerous resistant varieties, recently released by European breeding programs. We have reconstructed a continuous sequence of this resistance haplotype across 1 Mbp. We have reduced the genetic interval of the region containing the causal *Rpv3* gene down to 105,6 Kbp and obtained the full-length coding sequence of the candidate genes in this interval. Next generation sequencing and whole-genome *de novo* assembly were used to achieve these goals. A *de novo* assembly of an entire genome from short-reads in a highly heterozygous species with approximately 40% of repetitive DNA remains a tough task.

We mastered the trick of obtaining a high-quality assembly across the *Rpv3* locus by sequencing a custom-made diploid genome – homozygous for the resistant haplotype – and by sequencing a set of libraries with variable insert size and read length.

Only three genes are present in the reduced genetic interval of the locus. Two of them are nucleotide-binding leucine rich repeat (NB-LRR) genes. This family of genes is involved in innate immunity and encodes a class of cytoplasmic receptors, sometimes carrying nuclear localization signals, that directly or indirectly detect pathogen effectors and trigger cell death-mediated defence responses (Jacob at al. 2013). In flax, the L locus encodes TIR-NB-LRR proteins that recognize effectors from the flax rust fungus, leading to cell death (Bernoux et al. 2011). In the grapevine *M. rotundifolia*, the only functionally characterized resistance genes *Rpv1* and *Run1* belong to this class.

The third gene present in the resistance haplotype of the *Rpv3* locus encodes a receptor-like kinase (RLK). RLK genes encode transmembrane receptors. Two functional groups of RLKs are recognized in plants: RLKs controlling growth and developmental processes, and RLKs regulating interactions with microbes, both pathogens and symbionts. The LRR-kinase in the *Rpv3* locus shows higher similarity with RLKs involved in plant-pathogen interactions, in particular with the resistance gene *Xa21* of *Oryza sativa* which confers resistance to the bacterium *Xanthomonas oryzae* pv. *oryzae*. The functional classes of candidate genes in the *Rpv3* resistant haplotype are also consistent with the expectation based on the phenotype of the resistant individuals. In fact, the differential phenotype between resistant and susceptible offspring is the ability versus the inability to trigger a localized HR in the primary site of infection, soon after the onset of cellular interactions between the pathogen and the host (Bellin et al. 2009). This response is usually triggered by the presence of an appropriate protein receptor, enabling the cell to signal the presence of pathogen effectors and to switch on the defense machinery.

4.1 Assembly of the Rpv3 locus

Assembly of the resistance haplotype was obtained by combining sequencing of 21,076$^{Rpv3+/Rpv3+}$ and sequencing of large insert clones of a BAC library. *De novo* assembly reconstructs a genome through two processes, assembly of contigs and

assembly of scaffolds, linking contigs together (Hunt et al. 2014). The distribution in size of contigs and scaffolds is dependent on several factors like genome coverage, read length, insert length, level of heterozygosity, and abundance of the repetitive fraction in the genome. The assembly of a diploid genome proves very difficult in conditions of high heterozygosity. Assemblers that deal with allelic polymorphisms in heterozygous species may produce other flaws. In some cases, genomic complexity is artifactually reduced because the *de novo* assembler collapses repeated sequences into a single genomic region (Alkan et al. 2010). Regions containing tandem duplicate genes belonging to multigene families can also be particularly challenging for the assembler, leading to loss of gene copies or to artifactual proliferation in the assembly. In our case, the target region and several other introgressed regions were heterozygous for two distantly related haplotypes, donated by different *Vitis* species, exacerbating these adversities. To overcome this problem, we did not sequence Bianca, but we produced by selfing a homozygous individual at the *Rpv3* locus.

Scaffold statistics are usually used as the sole metrics to measure assembly performance. Inflated assembly metrics may hide errors in scaffolding, derived from bridging together unrelated contigs. We inspected, in detail, a set of scaffolds assembled from genomic regions with different characteristics. The distribution of scaffold size and the inspection of scaffold synteny with the reference grapevine genome suggested that our approach of *de novo* assembly was appropriate and efficient. We obtained a total genome size very close to the expected size. Correct scaffolding was obtained for both homozygous and heterozygous regions, as well as for homozygous regions with extensive tandem duplication. The same approach based on the HAPLOIDIFY algorithm implemented in the ALLPATHS assembler was used by Di Genova et al. (2014) for the whole-genome assembly of the variety 'Sultanina', obtaining comparable metrics in terms of total scaffold length (448.7 Kb in 21,076$^{Rpv3+/Rpv3+}$ versus 466.7 in Sultanina). We obtained in 21,076$^{Rpv3+/Rpv3+}$ better metrics for scaffold N50 (163 kb versus 78 kb in Sultanina), probably due to a higher fraction of homozygous regions in 21,076$^{Rpv3+/Rpv3+}$ compared to Sultanina. In fact, homozygous segments amount to 54.7 % of the total length of the chromosomes in 21,076$^{Rpv3+/Rpv3+}$, a condition that is unlikely to occur in cultivated varieties. The genome of another grapevine variety 'Tannat' was assembled from short-reads using a

completely different approach (Da Silva et al. 2013). The metrics of the 'Tannat' assembly are not dissimilar from those obtained for Sultanina and 21,076$^{Rpv3+/Rpv3+}$ (total scaffold length 482 Mbp, scaffold N50 97.2 kb), although some uncommon biological features inferred from the assembly may raise some concern on its accuracy (i.e. unusually high homozygosity, high number of duplicated genes, high number of novel and varietal genes, etc.). However, whole-genome assembly alone would not have been sufficient for obtaining a single scaffold across such a complex region as the *Rpv3* locus. Neither BAC sequences alone would have been sufficient, because of uneven coverage of BAC clones over the region with abundant high-copy DNA and the difficulty in the design of unique markers for further BAC screening.

The combination of the two approaches complemented the weaknesses of one another. All gaps between 21,076$^{Rpv3+/Rpv3+}$ scaffolds in the *Rpv3* locus were bridged by BAC supercontigs, and vice versa. The presence of repetitive DNA was the cause of interruption of scaffolding in the 21,076$^{Rpv3+/Rpv3+}$ assembly. Gaps between two scaffolds in 21,076$^{Rpv3+/Rpv3+}$ were caused by the presence of transposable elements which were otherwise assembled in each BAC spanning the region because they were present in BAC clones as single-copy DNA. The third gap was caused by the presence of a TNL gene which has high similarity with several duplicated copies, contained in another TNL gene cluster located downstream of the *Rpv3* locus.

4.2 Candidate genes for downy mildew resistance

The resistance haplotype at the *Rpv3* locus corresponds to a region on chromosome 18 where present is a cluster of TIR-NB-LRR genes and a LRR-kinase. Two gene copies belonging to the TIR-NB-LRR class were found in the locus, TNL2a and TNL2b. They have highly similar nucleotide sequences and a conserved intron-exon structure. The counterpart in the reference vinifera genome PN40024 is a single-copy allele. The two copies in the resistance haplotype likely arose by a small segmental duplication. In fact, the 3'-end of the TNL2 allele in PN40024 is followed by a Copia TE. The entire block of DNA including the Copia element and TNL2 is duplicated over a short distance in the resistance haplotype.

TIR-NB-LRR genes usually encode receptors that activate a programmed cell death in infected cells, in order to restrict the development of the pathogen. Both TNL genes at

the *Rpv3* locus show the same gene structure, but with a different level of transcription during pathogen infection. TNL2a has a higher level of expression than TNL2b. RNA-Seq showed evidence of alternative transcripts in the expression of TNL2a, including a full-length transcript with regular intron splicing, a transcript with retention of the first intron that translates into a truncated protein downstream the TIR domain, and a transcript with exon skipping and a premature stop codon some 20 amino acids downstream the TIR domain. Alternative splicing is frequently reported for functional resistance genes. The role of alternative splicing is obscure in the NB-LRR genes, but in some cases it proved indispensable for expressing pathogen resistance. The *Arabidopsis* resistance gene *RPS4* requires two types of transcripts to mediate resistance against *Pseudomonas syringae*, the full-length transcript and a transcript with intron retention (Zhang et al. 2007). It is not known if this crucial role is played by the alternative transcript itself or by the truncated protein that the transcript is predicted to encode. It was a curious coincidence that *RPS4* is the prototype of the TNL-B subclass of NB-LRR genes, to which *Rpv3* TNL2a and TNL2b also belong.

Intron retention can have positive effects on the level of gene expression. The CC-NB-LRR gene *RCY1* in *Arabidopsis thaliana* is associated with resistance to Cucumber mosaic virus, and represents an example of intron retention. Plants transformed with genomic sequence of *RCY1* gene showed a higher concentration of the protein than plants transformed with the corresponding cDNA sequence. Accumulation at high levels is necessary for the expression of complete resistance (Sato et al. 2014). However, the alternative transcripts of our TNL2a encode truncated proteins. Feechan et al. (2013) argued that the putative TIR proteins generated from truncated transcripts of *Rpv1* and *Run1* lack NLS signals and, contrary to the full-length proteins, may be involved in cytoplasmic cell death signaling. Collectively, this set of evidence on the occurrence of alternative transcripts in the expression of this kind of functional resistance genes lends support to our choice of cloning genomic sequences rather than intron-less cDNA into an expression vector for testing the function of *Rpv3* candidate genes.

Ka/Ks ratios between TNL2 paralogs in the resistance haplotype of the *Rpv3* locus as well as between each paralog and the allelic counterpart in PN40024 revealed that

different domains within the genes were subjected to different selective pressure. The highly conserved NB-ARC domain is under purifying selection, with Ka/Ks ratios consistently lower than one in all pairwise comparisons. The LRR region is under diversifying selection, based on Ka/Ks ratios higher than one. For the TIR domain, Ka/Ks ratios are close to one. However, the first exon of TNL2a and TNL2b has a trinucleotide repeat of variable length, encoding either seven or five consecutive serine amino acids between the first Met and the N-terminal border of the TIR domain, in TNL2a and in TNL2b respectively. In humans, size variation in homopolymeric amino acid repeats are often associated with genetic disorders (Simon et al. 2009). The repeat tract length can easily mutate by DNA polymerase slippage. Amino acid repeats form an intrinsically disordered structure that is a protein or a region that does not form a stable tertiary structure. They are involved in protein-protein interactions that can lead to binding of kinases, transcription factors, DNA, and RNA (Gojobori et al. 2010).

The LRR-kinase is present as a single-copy gene in the resistance haplotype. It encodes a predicted protein with a high amino acid identity with its allelic counterpart in PN40024. It is highly expressed in 21,076$^{Rpv3+/Rpv3+}$ at the onset of HR. However, based on this level of amino acid conservation across all major functional domains in sensitive and resistant individuals, this gene is a weaker functional candidate, though still a positional candidate.

4.3 Cloning of candidate genes

The NB-LRR family is one of the largest gene families in plants. In *V. vinifera*, there are 459 NB-LRR genes, with 97 genes belonging to the TNL group (Yang et al. 2008). This number nearly doubles if pseudogenes and homologous gene fragments are included. The massive expansion of these kinds of genes makes it difficult to isolate single gene copies from highly similar duplicates out of genomic DNA. We had the opportunity to use BAC clones as a source DNA for cloning TNL2b and LRR-kinase. Each gene was present in a single BAC clone, isolated from other duplicates. This was not possible for TNL2a, because no BAC clone covered the entirety of its coding sequence.

The size of the coding sequence of TNL genes is remarkably larger than the average gene size, which is 3,300 nucleotides in grapevine. Cloning cDNA sequences is a short-

cut in other cases. The possible role of alternative transcripts for expressing resistance suggested that we should not use this short-cut, but rather clone genomic DNA. Since evidence from RNA-Seq shows that genes in the Rpv3 locus are expressed at moderated levels in native conditions, we cloned them under the 35S promoter.

Cloning the candidate gene under its native promoter would further complicate the isolation of the full sequence, extending the size of the fragment to be cloned.

Functional analysis of candidate genes implies cloning of the genes of interest into an expression vector. It was done in two steps. First, we cloned the candidate genes into a cloning vector. Colonies positive for the full-length coding sequence of the target gene were used as a template for cloning the insert into an expression vector. The Gateway system was chosen for cloning candidate genes in an expression vector, in our case pK2G7W. We chose this vector to have a constitutive expression of these candidate genes under a 35S Cauliflower Mosaic Virus (CaMV) promoter/terminator.

One colony per target gene (TNL2a, TNL2b, and LRR-k) of the expression clone was delivered to a facility of grapevine genetic engineering for *Agrobacterium*-mediated transformation of embryogenic calli of susceptible varieties. The length of time required for somatic embriogenesis of transformed calli will postpone the phenotypic evaluation of DM resistance in transformed adult plants beyond the conclusion of my PhD program.

4.4 Diversity of the resistance haplotype

In terms of nucleotide diversity, the resistance haplotype stands off the range of intraspecific variation observed in a comprehensive set of 50 varieties of *V. vinifera*. This level of SNP density, relative to intraspecific SNP density, is in accordance with the resistance haplotype being a DNA segment introgressed from a non-vinifera species of North American origin. The divergence time between North American species and the Asian lineage, including *V. vinifera,* is estimated to be 3.5-9.5 millions years (Zecca et al. 2012). On the contrary, diversification within the cultivated compartment of *V. vinifera* mostly occurred past domestication, within the last 7,000 years. In spite of being a clear outgroup, branching from the set of *V. vinifera* haplotypes, the resistance haplotype shares an important fraction of SNPs with *V. vinifera* varieties. These SNPs may be ancestral polymorphic sites, retained in diverged taxa due to incomplete

lineage sorting. This observation is not unexpected, as incomplete lineage sorting was already postulated by Venuti et al. (2013) for explaining similar results for other genomic regions in the comparison between *V. vinifera* and Asian lineages.

4.5 Extent of genome introgression in 21,076$^{Rpv3+/Rpv3+}$

The resequencing of 21,076$^{Rpv3+/Rpv3+}$, along with a comprehensive set of 50 varieties of *V. vinifera,* allowed us to map with unprecedented resolution all the introgressed regions in a DM resistance line. In addition to a large chromosome segment around the *Rpv3* locus, 42% of the haploid genome of 21,076$^{Rpv3+/Rpv3+}$ is non-vinifera. In the cultivated variety Bianca, all these regions have a vinifera allele derived from the vinifera parent Bouvier. Thus, 21 % or more of the total length of the diploid genome in Bianca (counting separately each homolog) is introgressed. This portion included three entire homologs in chromosomes 7, 12, and 13. This extent of introgression is quite surprising, given the number of backcross generations that should have diluted the initial contribution of the non-vinifera genome from the wild donor parent. This high residual content in introgressed segments might be simply due to chance. However, introgressed regions largely overlap with chromosomal segments populated by NB-LRR genes. The minor effect on DM resistance of wild alleles of NB-LRR paralogs scattered through the genome may have undergone undetected in QTL analyses, but it might have been captured by the severe phenotypic selection operated by breeders. The combination of this effect with linkage drag may explain the abundance of non-vinifera DNA in Bianca. The intercrosses between different resistant lines in the pedigree of Bianca – which were aimed at strengthening resistance – may have partially ruined the expected gain in the fraction of vinifera genome that breeders aimed to achieve through backcrosses to pure vinifera varieties.

5 References

Alkan C, Sajjadian S, Eichker EE (2010) Limitation of next-generation genome sequence assembly. Nat Methods 8: 61-65

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215(8): 403-410

Barrett HC (1958) The Best Parents in Breeding French Hybrid Grapes. Fruit Varieties and Horticultural Digest. 12: 39-42

Bellin D, Peressotti E, Merdinoglu D, Wiedemann-Merdinoglu S,Adam-Blondon AF, Cipriani G, Morgante M, Testolin R, Di Gaspero G (2009) Resistance to Plasmopara viticola in grapevine 'Bianca' is controlled by a major dominant gene causing localised necrosis at the infection site. Theor Appl Genet: 120:163–176

Bernoux M,Ve T, Williams S, Warren C, Hatters D, Valkov E, Zhang X, Ellis JG, Kobe B, Dodds PN (2011) Structural and functional analysis of a plant resistance protein TIR domain reveals interfaces for self-association, signaling, and autoregulation. Cell Host Microbe 9:200-211

Butler J, MacCallum I, Kleber M, Shlyakhter IA, Belmonte MK, Lander ES, Nusbaum C, Jaffe DB (2008) ALLPATHS:de novo assembly of whole-genome shotgun microreads. Genome Res 18: 810–820

Da Silva C , Zamperin G, Ferrarini A, Minio A, Dal Molin A, Venturini L, Buson G, Tononi P, Avanzato C, Zago E, Boido E, Dellacassa E, Gaggero C, Pezzotti M, Carrau F, Delledonne M (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

Di Gaspero G, Copetti D, Coleman C, Castellarin SD, Eibach R et al (2012) Selective sweep at the *Rpv3* locus during grapevine breeding for downy mildew resistance. Theor Appl Genet 124: 277–286

Di Genova A, Almeida A, Munoz-Espinoza C, Vizoso P, Travisany D, Moraga C, Pinto M, Hinrichsen P, Orellana A, Maass A (2014) Whole genome comparison between table and wine grapes reveals a comprehensive catalog of structural variants. BMC Plant Biol 14(1):7

Doyle JJ, Doyle JL (1990) Isolation of plant DNA from fresh tissue. Focus 12:13-15

Feechan A, Anderson C, Torregrosa L, Jermakow A, Mestre P, Wiedemann-Merdinoglu S, Merdinoglu D, Walker A, Cadle-Davidson L, Reisch B, Aubourg S, Bentahar N, Shrestha B, Bouquet A, Adam-Blondon AF, Thomas MR, Dry IB (2013)

Genetic dissection of a TIR-NB-LRR locus from the wild North American grapevine species Muscadinia rotundifolia identifies paralogous genes conferring resistance to major fungal and oomycete pathogens in cultivated grapevine. Plant J 76:661-674

Gojobori J, Ueda S (2010) Elevated evolutionary rate in genes with homopolymeric amino acid repeats constituting nondisordered structure. Mol Biol Evol 28:543-50

Hunt M, Newbold C, Berriman M, Otto TD (2014) A comprehensive evaluation of assembly scaffolding tools. Genome Biol 15:R42

Jacob F, Vernaldi S, Maekawa T (2013) Evolution and conservation of plant NLR functions. Frontiers of Immunology 4:297

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA (2010) The Genome Analysis Toolkit: a MapReduce framework for analyzing next generation DNA sequencing data. Genome Res 20: 1297–1303

Meyers BC, Kozik A, Griego A, Kuang H, Michelmore RW (2003) Genome-wide analysis of NBS-LRR-encoding genes in Arabidopsis. Plant Cell 15:809-834

Munson TV (1909) Foundations of American grape culture. T.V. Munson & Son, Denison

Park CJ, Ronald PC (2012) Cleavage and nuclear localization of the rice XA21 immune receptor. Nat Commun doi: 10.1038/ncomms1932.

Rouxel M, Mestre P, Comont G, Lehman BL, Schilder A, Delmotte F (2012) Phylogenetic and experimental evidence for host-specilized cryptic species in a biotrophic oomycete. New Phytol 197:251-263

Sato Y, Ando S, Takahashi H (2014) Role of intron-mediated enhancement on accumulation of an Arabidopsis NB-LRR Class R-protein that confers resistance to Cucumber mosaic virus. Plos One 9(6):e99041

Schwander F, Eibach R, Fechter I, Hausmann L, Zyprian E, Töpfer R (2012) *Rpv10*: a new locus from the Asian Vitis gene pool for pyramiding downy mildew resistance loci in grapevine. Theor Appl Genet 124: 163–176

Simon M, Hancock JM (2009) Tandem and cryptic amino acid repeats accumulate in disordered regions of proteins. Genome Biol 10(6):R59

Simpson JT, Wong K, Jackman SD, Schein JE, Jones SJ, Birol I (2009) ABySS: A parallel assembler for short read sequence data. Genome Res 19(6):1117-1123

Sohn KH, Segonzac C, Rallapalli G, Sarris PF, Woo JY, et al. (2014) The Nuclear Immune Receptor RPS4 Is Required for RRS1^{SLH1}-Dependent Constitutive Defense Activation in Arabidopsis thaliana. PLoS Genet 10(10): e1004655

Thompson JD, Higgins DG, Gibson TJ (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 22: 4673–80

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

Venuti S, Copetti D, Foria S, Falginella L, Hoffmann S, Bellin D, Cindrić P, Kozma P, Scalabrin S, Morgante M, Testolin R, Di Gaspero G (2013) Historical introgression of the downy mildew resistance gene Rpv12 from the Asian species *Vitis amurensis* into grapevine varieties. PLoS One 8:e61228

Yang S, Zhang X, Yue JX, Tian D, Chen JQ (2008) Recent duplications dominate NBS-encoding gene expansion in two woody species. Mol Genet Genomics 280:187-198

Zecca G, Abbott JR, Sun WB, Spada A, Sala F, Grassi F (2012) The timing and the mode of evolution of wild grapes (Vitis). Mol Phylogenet Evol 62:736-747

Zhang XC, Gassmann W (2007) Alternative splicing and mRNA levels of the disease resistance gene *RPS4* are induced during defense responses. Plant Physiol 145:1577–1587

Chapter 2

Book: Grapevine breeding programs for the wine industry, Andrew G. Reynolds (ed)

Chapter: Molecular grape breeding techniques

Gabriele Di Gaspero[1,2] *, Serena Foria*[1]

[1] Dipartimento di Scienze Agrarie e Ambientali, University of Udine, 33100 Udine, Italy

[2] Istituto di Genomica Applicata, Parco Scientifico e Tecnologico Luigi Danieli, 33100 Udine, Italy

Keywords

precision breeding, genotyping-by-sequencing, genomic selection

1 Introduction

1.1 The use of DNA sequence information for assisting conventional breeding

The selection of novel varieties in highly heterozygous species requires observations and measurements of innumerable phenotypic characters that appear with distinct variants and combinations in the progeny created from controlled crosses. DNA sequence information is now routinely used for assisting conventional breeding in the selection of parents and the most valuable progeny. Breeding programs have stepped up during the past decade as genetic information became increasingly available. Considerable progress has been achieved towards the marker-assisted selection (MAS) of characters controlled by major genes, such as disease and pest resistances, and

towards the efficient removal of linkage drag around introgressed chromosome segments carrying valuable wild alleles. A few loci have been discovered in *Vitis vinifera* that are important for wine quality attributes (i.e. those controlling the synthesis of anthocyanins, proanthocyanins, methoxypyrazines, and terpenes). However, a swift improvement of our capacity to predict wine quality from fruit composition is demanded before DNA markers can fruitfully help in the selection for oenological value. A breakthrough in the past couple of years was the advent of genotyping-by-sequencing (GBS). GBS removed previous limitations in the generation and scoring of unlimited numbers of markers for interrogating all possible loci explaining the phenotypic variation for a trait of interest. Advances in our understanding of genome architecture and population structure in natural and breeding germplasm will indicate future directions in the use of genome-wide selection for characters relevant to the wine industry. In this field, genomic selection (GS) – an approach borrowed from animal breeding – is still an option to test as a practical method for wine-related breeding values. At the present time, MAS of desirable traits – especially disease-resistance loci – and genome-wide analyses are trustworthy, auxiliary means available to the breeders for foreground and background selection. As long as fully integrated into breeding strategies, this knowledge will serve conventional breeding in the design and creation of novel winegrape cultivars that can successfully compete with traditional and genetically engineered varieties.

2 Technical outline

2.1 Global structure of genetic diversity

Little is known about the structure of genetic diversity in the grape germplasm, beyond the parentages of many varieties. The state-of-art technology before the advent of (NGS) provided partial insight into the historical effects of human activity on the selection of winegrape varieties. The decay of linkage disequilibrium in varieties of *V. vinif*era is as rapid as in undomesticated populations of the same species. The reduction of haplotype diversity is generally irrelevant, with only a couple of exceptions associated with berry colour and domestication syndrome (Fournier-Level

et al. 2010, Myles et al. 2011). This little knowledge of the genetic make-up of elite winegrapes – historically selected by ancient viticulturists and adopted by modern viticulture – has largely left to each breeder's sensibility the burden to find the best empirical way for breeding and selecting new varieties. As a result, only a few winegrape varieties intentionally bred in the past two centuries excelled enough to acquire popularity, even among those varieties that have exclusively high-quality *V. vinifera* in their ancestry (i.e. Müller-Thurgau). Breeding activity in this crop was highly focused on improving biotic resistances, usually introgressed from other grape species. The recurrent use of a few excellent donors of disease resistance led to little genetic diversity being captured in grape breeding germplasm (Di Gaspero et al. 2012).

2.2 Architecture of the grapevine genome

The exploitation of beneficial alleles in conventional breeding is dependent upon the chromosomal location of the interesting loci and their genome landscape. Molecular breeders should always consider that genetic variation is structured in haplotypes and haplotypes are portions of the chromosome structure. The grapevine genome is highly heterogeneous along each chromosome and among chromosomes (Figure 1).

2.3 Gene-rich and gene-poor chromosomal regions

Gene density varies along each chromosome showing an inverse relationship with density of TEs. Gene density is the lowest in pericentromeric regions of the grapevine genome (Figure 1). A gene-poor landscape extends symmetrically over millions of nucleotides in pericentromeric regions of some chromosomes (i.e. chr 4, chr16). In other chromosomes, genes are dislodged from small chromosome segments, corresponding exactly to the centrometric tandem repeats (i.e. chr3, chr8). Genetic diversity in pericentromeric regions is expected to suffer more constraints than in the rest of genome because of lower recombination rate. Linkage drag is also more persistent in those regions. In corn, 95 % of total recombination rate is restricted to slightly more than half of the genome, decreasing dramatically in pericentromeric regions, possibly as a consequence of high structural variation in TE-rich regions.

2.4 Chromosomal regions with expanded gene families

Several gene families have proliferated by local duplications in the grape genome. NB-LRR genes encode receptors for pathogen- and pest-surveillance systems and are present in clusters located in subtelomeric regions. Entire chromosome arms (i.e. in chromosomes 12, 13, 18) are densely populated by potential resistance genes against biotic threats. Several haplotypes carry functional alleles for disease and pest resistance across these regions in geographically and genetically unrelated accessions of grapevines. Some resistance haplotypes have been introgressed into *V. vinifera* for many decades and retained during backcross breeding by phenotypic selection, before the advent of molecular breeding. Fewer resistance haplotypes have naturally evolved within the population of *V. vinifera* (Coleman et al. 2009, Rouxel et al. 2013). Other gene families are present with a higher number of gene copies compared to other plant genomes. Some of them are important for berry composition and wine sensory attributes, i.e. flavonoid 3',5'-hydroxylases (Falginella et al. 2010) and stilbene synthases (Vannozzi et al. 2012). Most of these gene copies are in physical proximity and have evolved some sort of functional specialisation. MAS for haplotypes that span these large gene clusters will probably have a major impact on berry-related traits. By contrast, single-copy genes encoding for key enzymes in other metabolic pathways are dispersed across the chromosomes and may require a more meticulous MAS of each favourable allele (Figure 1).

2.5 Other components of chromosome structure

Telomeric repeats guide a multiprotein complex that distinguishes natural DNA ends from DNA double-strand breaks (DSB), thereby protecting chromosome ends from DNA repair mechanisms and preventing chromosome fusion. In the PN40024 reference genome, 31 of the expected 38 telomeric ends are present in the outermost ends of the scaffolds anchored at the termini of linkage groups. Another seven regions with telomeric repeats were assembled into scaffolds not yet assigned to chromosomes. Telomeric-like repeats are present not only in the chromosome ends but also at interstitial sites of some chromosomes. We identified 16 interstitial telomeric sequences (ITS) on eight chromosomes. Notably, chr7, chr9, and chr13 have three ITSs, chr2 and chr15 have two ITSs. The presence of ITSs is posited to be the

result of ancestral chromosome fusion, intrachromosomal rearrangements, and insertion of telomeric DNA within unstable sites during DSB repair. Once generated, ITSs are unstable regions that may undergo rearrangements including amplification, deletion and transposition/translocation, and they are possibly signatures of past DSBs in fragile sites. In humans, the ITSs most prone to cause chromosomal aberrations are those located in centromeric regions. Physical vicinity of ITSs and centromeric repeats occurs in a number of notable cases in the grapevine genome, two spots on each chromosome for chr2 and chr7, and a single spot on chromosomes 6, 9, and 15.

2.6 Chromosomal regions with selective sweeps

Significant selective sweeps were detected in winegrape varieties on chromosomes 2 and 17 (Myles et al. 2011). Genetic diversity is severely reduced in the terminal part of chr2, as a consequence of positive selection for the most common white-skinned haplotype at the MybA gene cluster (Figure 1). White-skinned winegrape varieties have low genetic diversity because they are invariantly homozygous across this chromosome segment, but this reduction is also noteworthy in red-skinned varieties, because the majority of them are heterozygous for a red haplotype and the white haplotype. Another strong signature of selective sweep is present on chr17. The only phenotypic trait known to be partially dependent upon genes located within this region is berry proanthocyanidin content (Figure 1).

2.7 Genomic tools and breeding strategies in the genome sequencing era

The grapevine nuclear genome has been entirely assembled in 2007 following a whole-genome shotgun Sanger sequencing of the highly-inbred line PN40024 (Jaillon et al. 2007). The reference sequence offered the framework against which to compare genome-wide polymorphisms present in natural and breeding germplasm by the exploitation of NGS. This has led to the development of high-density SNP chips (Myles et al. 2011, Le Paslier et al. 2013) and facilitated the application of genotyping-by-sequencing (GBS). The availability of these genomics tools has made technically feasible a number of breeding strategies for precise introgression of wild alleles, removal of linkage drag, combination of multiple favourable haplotypes, and selection

of the desired background in novel varieties – briefly, precision breeding. These strategies comprise marker-assisted backcrossing (MABC), marker-assisted background selection (MABS), advanced backcross QTL strategy (AB-QTL), and marker-assisted pyramidisation (MAP).

2.8 Marker-assisted backcrossing (MABC) and marker-assisted background selection (MABS)

The pace and precision of backcross breeding can be significantly improved when (i) tightly linked markers are available for relevant traits, (ii) favourable and unfavourable haplotypes are known for the relevant loci, (iii) many DNA polymorphisms are available to monitor the transmission of non-sister chromatids from one generation to the next and to map the location of each recombination event. Recent advances in high-throughput genotyping have removed the bottlenecks that previously limited the level of resolution and the haplotype information content required for mapping effectively the genetic backgrounds. The high number of SNPs included in the most advanced chips – and the extent of genetic diversity that they are able to capture – ensure that a large proportion of SNPs is informative irrespectively of the type of breeding material under screening. The grapevine community converged on the use of commonly developed and publicly available tools, thereby dumping the cost of SNP chips and facilitating cross-comparison of results as well as sharing of knowledge. SNP chip hybridizations and GBS experiments are now commonly outsourced from genotyping/sequencing facilities of private companies, saving money otherwise required to maintain the breeder's laboratory with modernized and cost-efficient technologies. Outsourcing saves labour and money, but it is not a replacement for the capability of the breeder's laboratory to design the experiment and elaborate crude data. With these capabilities, SNP haplotyping will allow breeders to move from MABC – which was so far aimed at reassembling a species-specific recipient genome under the guide of microsatellite markers – to MABS, which is aimed more ambitiously to select for favourable combinations of chromosomal segments donated by different *V. vinifera* varieties in addition to the desirable introgressed gene.

2.9 Advanced backcross QTL strategy (AB-QTL)

The frequent presence of favourable QTL alleles for biotic stresses in unadapted species – interfertile with cultivated varieties – and the wish to introgress these traits into cultivated germplasm have led in other crops to propose the concept of AB-QTL. AB-QTL combines QTL analysis and variety development, by designing a mapping/breeding scheme for the simultaneous identification and introgression of wild haplotypes. AB-QTL relies on segregating populations in which most of the wild-parent genome that donates the trait of interest has been purged in early segregating generations by phenotypic selection. This strategy has been more commonly adopted in grapevine – eventually tracing back the QTL haplotype in the pedigree – rather than using early segregating generations (F1, F2, BC1) for QTL mapping. Favourable QTL alleles identified in early generations often vanish in later backcross generations, once other donor genes that have epistatic interactions with the beneficial QTL alleles are removed from highly *V. vinifera* genetic backgrounds. QTL stability should also be carefully considered for fruit-related and phenology-related traits mapped in *V. vinifera*, before the linked markers are proposed to breeders. Non-additive genetic effects may partially explain the plethora of different QTL regions – and their variable relevance with the genetic background of the mapping populations – that appeared on journal articles in the last years. Breeders' faith in the use of MAS for fruit-related and phenology-related traits has often been shaken by the lack of validation of QTL-marker associations in a comprehensive sample of breeding germplasm.

2.10 Marker-assisted pyramidisation (MAP)

Simultaneous MAS for independent genes controlling the same trait, also referred to as (MAP), is conducted for enduring the desired phenotype or, in the case of pathogen resistance, for securing the trait from the possible effects of adaptive evolution in the population of the pathogen. The concept of MAP also extends to the assisted selection for multiple target traits. High-throughput genotyping has the highest utility when it assists breeders in assembling all desirable haplotypes into the same genome. A gene pyramiding scheme is usually implemented by intermating best AB-lines, each one carrying complementary genes/haplotypes of interest, then the progeny is screened for individuals that have inherited all beneficial alleles at target loci.

3 Relevance and role in current and future scientific and commercial work

3.1 Improvement for disease and pest resistance

Several examples of the successful use of molecular breeding are now available in winegrapes. However, MAS is routinely used only for the improvement of traits related to pathogen and pest resistance (reviewed in Töpfer et al. 2011). The reasons for this confined success are numerous. Pathogen and pest resistances are quantitative traits, but single loci account for the vast majority of the phenotypic variation observed in bi-parental populations. Significant effort has been put into mapping major loci at high resolution, thereby providing the community with tightly linked markers on both sides of the causal genes. The haplotypes of interest are widely present in the breeding material used by the community, beyond the original populations in which the loci were mapped. For grape disease-resistance, it has become a good practise to validate the markers across the germplasm before releasing them, which makes them trustworthy upon publication (Di Gaspero et al. 2012, Venuti et al. 2013). This also gives breeders a sense of relevance of the tagged haplotpyes in the breeding germplasm. A proof-of-concept for the superiority of MAS over phenotypic selection in the improvement of downy and powdery mildew resistance has been provided by the work done by the breeding team at Julius Kuhn Institute, Germany (Eibach et al. 2007). MAS is the only mean for pyramiding genes for a certain disease resistance. For both downy and powdery mildew, pyramidisation of resistance haplotypes from different grape species into resistant winegrape varieties has become a common practise (Schwander et al. 2012, Li et al. 2013, Venuti et al. 2013), because host-specialisation in natural populations of both pathogens is more extensive than commonly assumed (Brewer & Milgroom 2010, Rouxel et al. 2013). Equally remarkable is the breeding work done by Andrew Walker's team at the University of California in Davis for fighting Pierce's disease. They identified two different resistant alleles of *PdR1* – a major resistance gene against *Xylella fastidiosa* in *Vitis* species endemic to the Southwestern US – that are now introgressed into a wide

range of winegrape backgrounds over multiple generations, thanks to the assistance of tightly linked markers (Riaz et al. 2008).

3.2 Elimination of linkage drag

Traits associated with resistance to many biotic threats are necessarily introduced into *Vitis vinifera* from wild species. The elimination of linkage drag around the introgressed haplotypes has become a priority for reducing the contribution of the undomesticated genome. In backcross and intercross breeding for downy and powdery mildew, the use of the latest generation of breeding lines that carry the resistance genes *Rpv3*, *Rpv10*, *Rpv12*, *Run1*/*Rpv1*, *Ren1* and *Ren3* in a highly-*vinifera* genetic background generate progeny with oenological potential comparable to traditional varieties, as long as the population size is large enough to let parental alleles affecting wine quality shuffle into many combinations. In our empirical experience, seedlings with wine quality attributes as high as in their parents occur in the order of magnitude of one seedling out of a few thousands, regardless of the fact that the cross-combination involves only pure *V. vinifera* varieties or introgression lines with highly-*vinifera* background. Progress has also been made in conventional breeding for other resistance traits, originally present only in wild grapes. The winegrape selections resistant to Pierce's disease are among the brightest examples. Several other valuable wild haplotypes for disease and pest resistances are now being discovered, which will require intensive backcrossing before being introduced into breeding material.

3.3 Improvement of other traits and quest for genetic variation

Molecular breeding in grapevine has taken important steps to translate the accuracy of DNA-guided selection into practice. However, the number of reports on the successful incorporation of MAS into breeding programs lags behind the number of scientific publications reporting the identification of QTLs for traits potentially interesting to breeders. Most QTLs have been mapped in small-size bi-parental populations – appropriately generated for scoring the segregation of phenotypes – and the genetic variation revealed by the linked markers may not be detectable in other breeding germplasm. Fine mapping and haplotype analysis are increasingly used

as a validation step to ensure that the published markers maintain their predictive power in breeding germplasm (Di Gaspero et al. 2012, Venuti et al. 2013). Alternatively, a mixed approach of linkage/association mapping is used to assess the relevance in the germplasm of the haplotypes of candidate genes underlying mapped QTLs (Carrier et al. 2013). Haplotype analysis also assists breeders in their quest for novel genetic variation in breeding germplasm and in the wilderness at important loci. This has become a necessity for some practical applications, i.e. finding new sources of major genes for disease-resistance, because the genetic basis of the current breeding germplasm is narrow.

4 Future trends

4.1 Precision breeding

Precision breeding in grapevine should aim at assembling an ideal genome that is a mosaic of desirable chromosomal segments – donated by multiple ancestors – each one carrying a favourable haplotype for a target trait or providing a suitable genetic background. This accumulation of favourable alleles for loci with large effects on interesting traits should provide measurable genetic gain. The genetic gain in a specialty crop with a highly heterozygous genome diverge from the original concept developed for staple crops, in which quantitative traits associated with maximization of production and productivity are common targets in all breeding programs and they are expressed by measurable parameters. Estimation of genetic gain – which would be important to monitor efficiency of the process and to adjust actions and strategies accordingly – is difficult to conduct in winegrape breeding. Grape breeding has been an empirical activity in which the evaluation of many viticulturally and oenologically important traits in the candidate parents and in the progeny is left to the intuitive perception of the breeder. Large effort is still needed to overcome this limitation.

4.2 Haplotype mapping

Ancestral genomic segments are passed down from parents to kin and they are shared by descent across generations as discrete units (haplotypes), until being shuffled by genetic recombination. Kinship in high-quality varieties of *V. vinifera* implies that large

blocks of DNA along the chromosomes are conserved among varieties. Most winegrapes are derived from a few founders and are removed from them by a few generations, thereby grouping into a dozen of family groups (Bacilieri et al. 2013). Thus, most variation is structured into haplotypes. What breeders would need is a dynamic map showing the chromosomal segments in founder varieties and how these segments became fragmented in their descendents. Once the biological relevance of each haplotype or its frequency in highly regarded varieties have been assessed, the ideal genetic background can be planned as a mosaic of the most wanted haplotypes. Haplotype mapping in *V. vinifera* has the potential to provide the fundamental information to revolutionize breeding strategies. The intermating of major lineages of winegrapes, carrying interesting haplotypes that underlay quantitative variation in *V. vinifera*, followed by the production of inbred lines, should fix important traits in a few individuals. The removal of constraints for the generation of dihaploid plants and nearly homozygous lines may open the door to breakthroughs in conventional breeding in the years to come. The future availability of parental lines with fixed traits for wine sensory attributes could represent a paradigm shift in the next decades from outcrossing to hybrid breeding in grapevine.

4.3 Short-cycling vines

The long generation time is still a significant limitation for grape breeding. Juvenility and annual reproductive cycle are major constraints against a rapid and exhaustive evaluation of seedlings for berry-related characteristics and wine sensory attributes. Breeding can be significantly accelerated by the use of short-cycling dwarf mutants with precocious and continuous flowering, also known as microvines (Chaïb et al. 2010). Double homozygous plants were developed for precocious flowering and female flowers, which bloom within two months after seed germination, do not require emasculation prior to cross-pollination, and generate large progeny populations, themselves precociously flowering. Development of near-homozygous isogenic lines, rapid introgression of wild haplotypes, gene pyramidisation, and construction of a planned genetic background are now more rapidly feasible with the appropriate use of microvines.

4.4 Fruit and wine composition: so many metabolites, so little known

We are still a long way off from deciphering the complexity of wine aroma. Chemistry has to hurdle many obstacles in the discovery of all relationships between berry composition and wine sensory attributes. And genetics lags behind. Carotenoids, *S*-cysteine conjugates, glycoconjugates, unsaturated lipids, phenolic acids are all metabolites present in the grape berry that are capable to generate odorants. A few odorant volatiles in the wines were traced back to their precursors in the berry, and for even fewer the genetic control of their synthesis has been elucidated. Methoxypyrazines impart green pea and bell pepper characters. The 2-methoxy-3-isobutylpyrazine (IBMP) is the major methoxypyrazine in berries of Bordeaux cultivars and is released from its non-volatile precursor by the gene product of an S-adenosyl-methionine-dependent *O*-methyltransferase (Dunlevy et al. 2013, Guillaumie et al. 2013). Floral odours of many white winegrapes are imparted by monoterpenes and monoterpene alcohols that are synthesised under the control of several genes in the terpenoid pathway (Battilana et al. 2011, Martin et al. 2011). The volatiles 2-aminoacetophenone and methyl-anthranilate are responsible for the distinctive foxy aroma in some wild grapes and in their interspecific crosses. An alcohol acyltransferase catalyzes the formation of methyl-anthranilate from the substrates anthraniloyl-coenzyme A and methanol (Wang and De Luca 2005). All of these publications have demonstrated the contribution of terpenoid genes, *O*-methyltransferase and alcohol acyltransferase to the synthesis of key odorants that are important for varietal components of wine aroma, but this knowledge has yet to provide practical breeding with markers that effectively predict wine sensory attributes in new seedlings. What lies beneath the synthesis of other odorants conferring varietal characters, such as the pepper aroma of rotundone and the precursors of passion fruit/grapefruit thiols 3-mercaptohexan-1-ol and 3-mercaptohexyl acetate, remains completely unknown at the genetic level. Insights into the genetic control of metabolites that are important for the structure and colour of red wines (i.e. proanthocyanidins and anthocyanins) were provided by Patrice This' team at INRA Montpellier. QTL regions for proanthocyanidin content and degree of polymerisation of condensed tannins in berry skin and seeds were identified on

several chromosomes and associated with a bunch of structural genes and transcription factors (Figure 1). QTL regions for anthocyanin content, level of hydroxylation, and level of methoxylation were identified on chromosomes 1, 2, 6, and associated with key structural genes and transcription factors (Huang et al. 2012, Carrier et al. 2013, Huang et al. 2013). These papers tackle the complexity of quantitative genetics acting behind metabolite traits.

4.5 Genomics selection: think wide!

Genomic selection is an approach borrowed from livestock breeding – mainly dairy cattle – that simultaneously estimates the effect of each marker across the entire genome to predict the breeding value of individuals, theoretically capturing more genetic variation for small-effects underneath complex traits (Jonas & de Koning 2013). Contrary to MAS, the contribution of all genome-wide DNA polymorphisms to the breeding value is accounted for in the diagnostic model during the calibration of the system. Then all markers – not exclusively those linked to significant QTLs – are used to measure the genomic estimated breeding value of each individual. In plant science, GS has been initiated in cereals and forest trees. It is also becoming attractive to grape breeders, because it promises to help even in the selection of those traits for which the genetics basis remains obscure. However, the key factor in the utility and success of GS is the accurate measurement and prediction of breeding values, which is particularly critical for oenological potential and for species in which the genetic structure of the breeding material is not fixed. The application of GS to winegrape breeding should also take into consideration the fact that successful prediction of breeding values in livestock currently applies to traits affected mainly by additive genetic effects, while we still ignore the relevance of dominance, epistasis, and genetic x environment interactions on wine sensory attributes.

4.6 Genetics and breeding : who's ancillary to whom ?

Mapping and breeding strategies mutually benefit from a coordinated exploitation of the same genetic resourses. The role of geneticists as mere providers of molecular tools to breeders in the classical genetics-to-breeding approach is obsolete. Geneticists should consider the possibility to build on the breeders' knowledge and

get the best breeding germplasm – appropriately selected or generated on purpose – for QTL mapping and gene discovery, with the side-effect that the novel markers will more likely find a practical application in relevant breeding germplasm. In this breeding-to-genetics approach, QTLs and genes are mapped in custom-made populations generated with the most advanced lines that passed multiple cycles of phenotypic selection, in which the inheritance of traits of interest is largely purged from non-additive effects. This approach has been widely and successfully adopted for genetic mapping of disease resistances. Biparental populations encompass restricted allelic variation, allow to map one or a few segregating traits, and may not capture the entire genetic architecture of complex traits because of non-additive effects transmitted by the genetic background of those particular parents. Many traits with a more complex genetic architecture vary quantitatively among high-quality varieties. This subtle variation makes the difference for the success of a variety, and breeders wish to control the underlying haplotype variation during the cycles of background selection. The rapid decay of LD has vanished efforts of genome-wide association mapping for these traits, and the situation is not going to change until millions of SNP are mappable by GBS. For these kind of traits, the breeding-to-genetics path should lead to the development of a few large segregating populations, generated after trait selection and cycles of intermating between multiple parental lines. The reward for the investment required by the preparation of this segregating material should be provided by the advantage of mapping simultaneously and at higher resolution favourable alleles from different sources, the same used in the ongoing breeding program. In our view, genetics and breeding are different sides of the same coin, intimately interconnected as never before.

5 Conclusions

The first varieties newly bred and selected in the last decade with the fundamental aid of molecular and genomics information are now ready to enter the market. The applied value of molecular tools in the domain of winegrape breeding is likely to further increase in the near future, accelerating progress in the (i) characterization of genetic variation in natural and breeding germplasm, (ii) precise introgression of genes and QTLs, (iii) differentiation and pyramidisation of valuable genetic variation in

breeding material and selected varieties, (iv) identification of the best breeding stock and fewer meritorious lines that will be taken to the ultimate step of plot selection and large-scale vinification. With these prospects, we are pretty convinced that winegrape varieties bred and selected through genomics-assisted breeding can compete with traditional varieties and genetically engineered lines with the same chance to succeed in consumer choice, making a contribution to the sustainable development of viticulture.

5.1 Sources of further information and advice

An excellent and up-to-date review of grapevine molecular breeding is provided by the book chapter Töpfer et al. (2011). A general overview on the possibilities of precision breeding in crop improvement is given by Peleman and van der Voort (2003) in their seminal work about Breeding by Design™. More recent perspectives of the application of genomics-assisted breeding for grapevine improvement are explored by the review articles Di Gaspero and Cattonaro (2010) and Myles (2013).

6 References

Chaïb J, Torregrosa L, Mackenzie D, Corena P, Bouquet A, Thomas MR (2010) The grape microvine - a model system for rapid forward and reverse genetics of grapevines. Plant J 62:1083–1092

Di Gaspero G, Cattonaro F (2010) Application of genomics to grapevine improvement. Aust J Grape Wine Res 16:122–130

Jonas E, de Koning DJ (2013) Does genomic selection have a future in plant breeding? Trends Biotechnol doi: 10.1016/j.tibtech.2013.06.003 [Epub ahead of print]

Myles S (2013) Improving fruit and wine: what does genomics have to offer? Trends Genet 29:190–196

Peleman JD, van der Voort JR (2003) Breeding by design. Trends Plant Sci 8:330–334

Töpfer R, Hausmann L, Eibach R (2011) Molecular breeding. In Zapater MM, Adam-Blondon AF, Kole C (eds) Genetics, Genomics and Breeding of Grapes, Sciences Publishers, Enfield NH, USA. 160–185

Le Paslier M-C, Choisne N, Bacilieri R, Bounon R, Boursiquot J-M, Bras M, Brunel D, Di Gaspero G, Hausmann L, Lacombe T, Laucou V, A Launay A, Martinez-Zapater JM, Morgante M, Raj PS, Ponnaiah M, Quesneville H, Scalabrin S, Torres-Perez R, Adam-Blondon A-F (2013) The GrapeReSeq 18k Vitis genotyping chip. IX International Symposium on Grapevine Physiology and Biotechnology. April 21-26, 2013, La Serena, Chile

Features of the grapevine genomes and genetic diversity relevant to breeding

Bacilieri R, Lacombe T, Le Cunff L, Di Vecchi-Staraz M, Laucou V, Genna B, Péros JP, This P, Boursiquot JM (2013) Genetic structure in cultivated grapevines is linked to geography and human selection. BMC Plant Biol 13:25

Brewer MT, Milgroom MG (2010) Phylogeography and population structure of the grape powdery mildew fungus, *Erysiphe necator*, from diverse *Vitis* species. BMC Evol Biol 10:268

Fechter I, Hausmann L, Daum M, Sörensen TR, Viehöver P, Weisshaar B, Töpfer R (2012) Candidate genes within a 143 kb region of the flower sex locus in Vitis. Mol Genet Genomics 287:247–259

Jaillon O, Aury JM, Noel B, Policriti A, Clepet C, Casagrande A, Choisne N, Aubourg S, Vitulo N, Jubin C, Vezzi A, Legeai F, Hugueney P, Dasilva C, Horner D, Mica E, Jublot D, Poulain J, Bruyère C, Billault A, Segurens B, Gouyvenoux M, Ugarte E, Cattonaro F, Anthouard V, Vico V, Del Fabbro C, Alaux M, Di Gaspero G, Dumas V, Felice N, Paillard S, Juman I, Moroldo M, Scalabrin S, Canaguier A, Le Clainche I, Malacrida G, Durand E, Pesole G, Laucou V, Chatelet P, Merdinoglu D, Delledonne M, Pezzotti M, Lecharny A, Scarpelli C, Artiguenave F, Pè ME, Valle G, Morgante M, Caboche M, Adam-Blondon AF, Weissenbach J, Quétier F, Wincker P (2007) The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. Nature 449:463–467

Myles S, Boyko AR, Owens CL, Brown PJ, Grassi F, Aradhya MK, Prins B, Reynolds A, Chia JM, Ware D, Bustamante CD, Buckler ES (2011) Genetic structure and domestication history of the grape. Proc Natl Acad Sci USA 108:3530–3535

Rouxel M, Mestre P, Comont G, Lehman BL, Schilder A, Delmotte F (2013) Phylogenetic and experimental evidence for host-specialized cryptic species in a biotrophic oomycete. New Phytol 197:251-263

Metabolites and genetics of wine sensory attributes

Battilana, J, Emanuelli F, Gambino G, Gribaudo I, Gasperi F, Boss PK, Grando MS (2011) Functional effect of grapevine 1-deoxy-D-xylulose 5-phosphate synthase substitution K284N on Muscat flavour formation. J Exp Bot 62 5497–5508

Carrier G, Huang YF, Le Cunff L, Fournier-Level A, Vialet S, Souquet JM, Cheynier V, Terrier N, This P (2013) Selection of candidate genes for grape proanthocyanidin pathway by an integrative approach. Plant Physiol Biochem. doi: 10.1016/j.plaphy.2013.04.014

Fournier-Level A, Lacombe T, Le Cunff L, Boursiquot JM, 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

Guillaumie S, Ilg A, Réty S, Brette M, Trossat-Magnin C, Decroocq S, Léon C, Keime C, Ye T, Baltenweck-Guyot R, Claudel P, Bordenave L, Vanbrabant S, Duchêne E, Delrot S, Darriet P, Hugueney P, Gomès E (2013) Genetic analysis of the biosynthesis of 2-methoxy-3-isobutylpyrazine, a major grape-derived aroma compound impacting wine quality. Plant Physiol 162:604–615

Dunlevy JD, Dennis EG, Soole KL, Perkins MV, Davies C, Boss PK (2013) A methyltransferase essential for the methoxypyrazine-derived flavour of wine. Plant J 2013 Apr 29. doi: 10.1111/tpj.12224

Falginella L, Castellarin SD, Testolin R, Gambetta GA, Morgante M, Di Gaspero G (2010) Expansion and subfunctionalisation of flavonoid 3',5'-hydroxylases in the grapevine lineage. BMC Genomics 11:562

Huang YF, Doligez A, Fournier-Level A, Le Cunff L, Bertrand Y, Canaguier A, Morel C, Miralles V, Veran F, Souquet JM, Cheynier V, Terrier N, This P (2012) Dissecting

genetic architecture of grape proanthocyanidin composition through quantitative trait locus mapping. BMC Plant Biol 12:30

Huang YF, Bertrand Y, Guiraud JL, Vialet S, Launay A, Cheynier V, Terrier N, This P (2013) Expression QTL mapping in grapevine - revisiting the genetic determinism of grape skin colour. Plant Sci doi: 10.1016/j.plantsci.2013.02.011

Martin DM, Chiang A, Lund ST, Bohlmann J (2011) Biosynthesis of wine aroma: Transcript profiles of hydroxymethylbutenyl diphosphate reductase, geranyl diphosphate synthase, and linalool/nerolidol synthase parallel monoterpenol glycoside accumulation in Gewürztraminer grapes. Planta 236:919–929

Vannozzi A, Dry IB, Fasoli M, Zenoni S, Lucchin M (2012) Genome-wide analysis of the grapevine stilbene synthase multigenic family: genomic organization and expression profiles upon biotic and abiotic stresses. BMC Plant Biol 12(1):130

Wang J, DeLuca V (2005) The biosynthesis and regulation of biosynthesis of Concord grape fruit esters, including 'foxy' methylanthranilate. Plant J 44:606–619

Marker assisted selection

Di Gaspero G, Copetti D, Coleman C, Castellarin SD, Eibach R, Kozma P, Lacombe T, Gambetta G, Zvyagin A, Cindrić P, Kovács L, Morgante M, Testolin R (2012) Selective sweep at the *Rpv3* locus during grapevine breeding for downy mildew resistance. Theoretical & Applied Genetics 124:277–286

Eibach R, Zyprian E, Welter L, Töpfer R (2007) The use of molecular markers for pyramiding resistance genes in grapevine breeding. Vitis 46 120–124

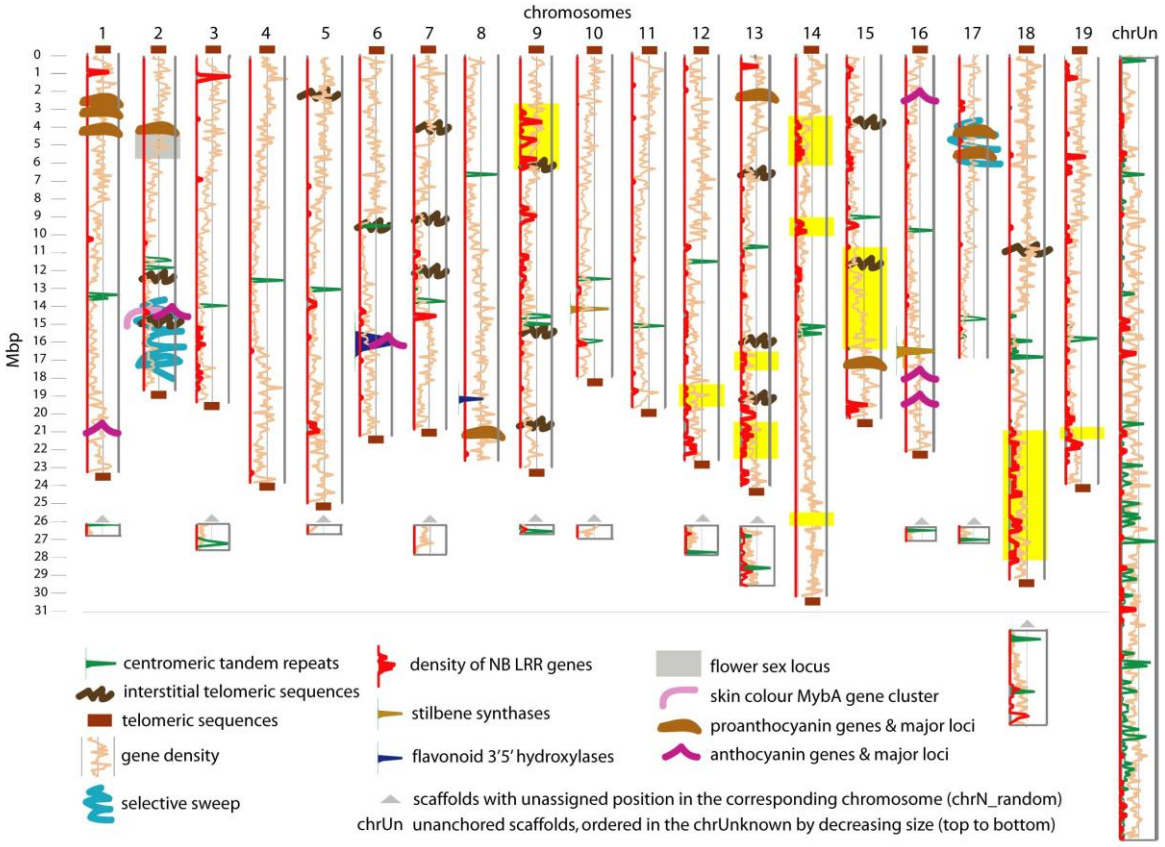
Li C, Erwin A, Pap D, Coleman C, Higgins AD, Kiss E, Kozma P, Hoffmann S, Ramming DW, Kovács LG (2013) Selection for *Run1*-*Ren1* dihybrid grapevines using microsatellite markers. Am J Enol Vitic 64:152–155

Riaz S, Tenscher AC, Rubin J, Graziani R, Pao SS, Walker MA (2008) Fine-scale genetic mapping of two Pierce's disease resistance loci and a major segregation distortion region on chromosome 14 of grape. Theor Appl Genet 117:671–681

Schwander F, Eibach R, Fechter I, Hausmann L, Zyprian E, Töpfer R (2012) *Rpv10*: a new locus from the Asian *Vitis* gene pool for pyramiding downy mildew resistance loci in grapevine. Theor Appl Genet 124:163–176

Venuti S, Copetti D, Foria S, Falginella L, Hoffmann S, Bellin D, Cindric P, Kozma P, Scalabrin S, Morgante M, Testolin R, Di Gaspero G (2013) Historical introgression of the downy mildew resistance gene *Rpv12* from the Asian species *Vitis amurensis* into grapevine varieties. PLos One 8(4): e61228

Figure 1 – Grapevine genome structure and relevant regions for molecular breeding. Prediction of structural characteristics is based on the 12X genome assembly of the PN40024 reference sequence. Chromosome length is indicated in million base pairs (Mbp). Telomeric sequences were searched by Blast using the eptamer $[TTTAGGG]_n$. Terminal telomeric sequences are indicated by *brown boxes* at the end of the chromosomes, interstitial telomeric sequences are indicated by *brown wavy lines*. The locations of centromeric repeats were predicted by Blast using the 107-nt monomer AGTACCGAAAAAGGGTCGAATCAGTGTGAGTACCGAAAAATGGTAGAATCCGGGCGAGTACCGGGA AAAGGTAGAATCCGTGCGAGTATCGAAAAACTGTCCGGGCG and indicated by the *green plot*. Regions with selective sweeps in cultivated varieties of *Vitis vinifera* are indicated by *cyan symbols* according to Myles et al. (2011). Density of genes (*peach plot*, scale 0 to 20 genes per 100 kbp, according to 29,970 genes of the V1 gene prediction) and NB-LRR genes, pseudogenes and homologous gene fragments (*red plot,* scale 0 to 20 genes per 100 kbp) is shown in adjacent windows of 100 kbp. *Yellow boxes* indicate major loci with disease- and pest-resistance haplotypes identified across clusters of NB-LRR genes (DM-*Rpv10* on chr9; DM/PM-*Run1/Rpv1* on chr12; PM-*Ren1* and phylloxera-*Rdv1* on chr13, DM-*Rpv8*, DM-*Rpv12*, PM-*Ren5*, and *Xylella fastidiosa-PdR1a* on chr14; PM-*Ren3* and *Agrobacterium-Rcg1* on chr15; DM-*Rpv2*, DM-*Rpv3*, PM-*Run2*, PM-*Ren4* on chr18; *Xiphinema index-XiR1* on chr19). The reported proanthocyanin genes are *LAR1*, *MybC2-L1* and *Trans-like* on chr1; *LDOX* on chr2; *Myb5a* on chr8; *CHI* on chr13; *MybPA1* on chr15; *LAR2* and *COBRA-like* on chr17 (Huang et al. 2012, Carrier et al. 2013). The reported anthocyanin genes are the *OMT* gene cluster associated with the level of methylation on chr1, the *MybA* gene cluster on chr2, the *F3'5'H* gene cluster on chr6, the *UFGT* gene associated with a *cis*-eQTL, the *anthoMATE* gene cluster associated with transport of acylated anthocyanidins, and the *ABCC1* ATP-binding cassette protein associated with transport of glucosylated anthocyanidins on chr16. The flower sex locus is indicated according to Fechter et al. (2012). Location of flavonoid 3',5'-hydroxylase and stilbene synthase gene clusters is indicated according to Falginella et al. (2010) and Vannozzi et al. (2012), respectively.


chromosomes
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
chrUn
Mbp
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
centromeric tandem repeats
interstitial telomeric sequences
telomeric sequences
gene density
selective sweep
density of NB LRR genes
stilbene synthases
flavonoid 3'5' hydroxylases
flower sex locus
skin colour MybA gene cluster
proanthocyanin genes & major loci
anthocyanin genes & major loci
scaffolds with unassigned position in the corresponding chromosome (chrN_random)
chrUn unanchored scaffolds, ordered in the chrUnknown by decreasing size (top to bottom)

Acknowledgments

I thank Nicoletta Felice and Eleonora di Centa for preparation of Illumina libraries; Irena Jurman and Federica Cattonaro for Illumina sequencing; Davide Scaglione and Michele Vidotto for assistance in *de novo* assembly; Dario Copetti for BAC clone assembly; Fabio Marroni, Mara Miculan and Simone Scalabrin for assistance in SNP analysis and analysis of gene expression; Giorgio Gambino for assistance in gene cloning; Michele Morgante and Raffaele Testolin for advice and guidance.

I also thank Laurent Torregrosa and Clèa Houel for the delightful time I spent at Supagro Montpellier.

I would especially like to thank my supervisor, Gabriele Di Gaspero, for his personal support, patience, insightful advice and guidance during this work.