UNIVERSITÀ DEGLI STUDI DI UDINE

Dottorato di Ricerca in Scienze e Biotecnologie Agrarie

Ciclo XXVIII

Coordinatore: prof. Mauro Spanghero

TESI DI DOTTORATO DI RICERCA

Characterisation of the pan-genome of *Vitis vinifera* using Next Generation Sequencing

DOTTORANDO
Dott. Gabriele Magris

SUPERVISORE
Prof. Michele Morgante

CO-SUPERVISORE
Dott. Fabio Marroni

ANNO ACCADEMICO 2014/2015

Contents

Acronyms I

Summary III

1 INTRODUCTION 1

1.1 NOVABREED project 1

1.2 Grapevine domestication history 1

1.3 Grapevine genome 3

1.4 Next Generation Sequencing 4

1.5 Structural variation analysis 8

1.6 Transposable elements and plant pan-genome 14

2 OBJECTIVES 19

3 MATERIALS AND METHODS 20

3.1 Plant material 20

3.2 Library preparation and sequencing 20

3.3 SNP analysis 22

3.4 SV dataset simulation 25

3.5 Benchmarked software packages for the detection of deletions 26

3.6 Identification of deletions 27

3.7 Transposable element annotation 29

3.8 Dating of Long Terminal Repeat insertion events 30

3.9 Detection of insertions 31

3.10 Identification of insertions....33
3.11 Gene annotation analysis....35
3.12 Validation of structural variants....36
3.13 Validation of SVs on *de novo* assembly....38
3.14 RNA sequencing experiments....38
4 RESULTS....40
4.1 Sequencing....40
4.2 Single Nucleotide Polymorphism (SNP) analysis....45
4.3 SV simulation....59
4.4 Simulation of deletions....61
4.5 Simulation of insertions....69
4.6 Structural variants detection in *V. vinifera*....70
4.7 Identification of deletions....72
4.8 Identification of insertions....84
4.9 SV validation through a PCR-based assay....92
4.10 SV validation based on the comparison to the *de novo* assembly....93
4.11 Transcriptome analysis....95
4.12 Grapevine pan-genome....98
4.13 Grapevine population structure based on SV....100
5 DISCUSSION....103
6 REFERENCES....118
7 APPENDIX....136
8 ACKNOWLEDGEMENTS....140

Acronyms

aCGH: array Comparative Genomic Hybridization
BAM: Binary Alignment/Map
ChIP-Seq: Chromatin Immunoprecipitation Sequencing
CNV: Copy Number Variant
CV: Cross-validation
DHH: DNA transposon Helitron Helitron
DIRS: *Dictyostelium* Intermediate Repeat Sequence
DOC: Depth Of Coverage
DTA: DNA transposon TIR hAT
DTC: DNA transposon TIR CACTA
DTH: DNA transposon TIR PIF-Harbinger
DTM: DNA transposon TIR Mutator
DTT: DNA transposon TIR Tc1-Mariner
DTX: DNA transposon TIR unknown
DXX: DNA transposon unknown unknown
FDR: False Discovery Rate
FPKM: Fragments Per Kilobase of transcript per Million mapped reads
Gb: Giga base pairs
GBS: Genotyping By Sequencing
GO: Gene Ontology
IBS: Identity By State
LINE: Long Interspersed Nuclear Element
LRR: Leucine-Rich Repeat
LTR: Long Terminal Repeat
MAF: Minor Allele Frequency
Mb: Mega base pairs
MYA: Million Years Ago
NAHR: Non-allelic Homologous Recombination
NBS: Nucleotide Binding Site
NGS: Next Generation Sequencing
PAV: Presence-Absence Variation

PCoA: Principal Coordinates Analysis
PE: Paired End
PEM: Paired-End Mapping
PLE: *Penelope*-Like Elements
PPV: Positive Predictive Value
RIL: Retrotransposon LINE L1
RLC: Retrotransposon LTR Copia
RLG: Retrotransposon LTR Gypsy
RLR: Retrotransposon LTR Retrovirus
RLX: Retrotransposon LTR unknown
RXX: Retrotransposon unknown unknown
RTA: Representative Transcript Assembly
SAM: Sequence Alignment/Map
SBS: Sequencing By Synthesis
SNP: Single Nucleotide Polymorphism
SPET: Single Primer Enrichment Technology
SRA: Short Read Archive
SV: Structural Variant
TE: Transposable Element
TIR: Terminal Inverted Repeat
TRF: Tandem Repeat Finder
UPGMA: Unweighted Pair Group Method with Arithmetic mean
VCF: Variant Call Format
XXX: unclassified transposable element

Summary

The present research work was carried out in the frame of the ERC-funded project NOVABREED, whose aim is to characterise the dispensable fraction of the pan-genome of *Vitis vinifera* and *Zea mays*, by developing and using bioinformatics tools for the analysis of DNA and RNA Next Generation Sequencing (NGS) data.

In the present work we focused on grapevine and characterised the genomes of 128 individuals. Single Nucleotide Polymorphism (SNP) markers were used to explore the grapevine population structure and to assess the genetic relationships between individuals. A total of approximately 9 million SNPs were obtained. Grapevine is characterised by an ancient and complex history of domestication: from the first centre of domestication (in the Caucasian region) grapevine spread to Central Europe, where secondary domestication centres were discovered (Grassi F et al., 2003; Arroyo-Garcia R et al., 2006; Myles S et al., 2011). ADMIXTURE (Alexander DH et al., 2009) was used to infer the population structure. At K=3 the population was divided in three main groups, in line with the observations of Negrul AM (1946), with a high proportion of admixed varieties. At higher K values we obtained a subdivision of the population into smaller clusters of varieties, linked by different degrees of relationship. The population structure was confirmed by the Principal Coordinates Analysis (PCoA) of the pairwise genetic distances. The estimated distances reflected the geographical distances in the prevalent area of current cultivation: the first major component explained a gradient of separation between the varieties from East to the West. The second major component explained a gradient of removal of domesticated varieties from wild forms.

In order to gain knowledge about the composition of the dispensable fraction, we investigated the Structural Variants (SVs) in 50 grapevine varieties. Structural variants ranging in size between 1 Kb and 25 Kb were identified based on the paired-end mapping information derived from the alignment of short reads to the reference genome sequence of *Vitis vinifera* (Jaillon O et al., 2007). For the detection of deletions we integrated the results obtained using two different tools: DELLY (Rausch T et al., 2012) and GASV (Sindi S et al., 2009). For the detection of insertions we used a pipeline developed by our research group. Overall, we identified a total of 18,551 deletions and 54,254 insertions amounting to 101.94 Mb and 329.9 Mb, respectively. The excess of structural variants with low allele frequencies supported the supposed recent origin of SVs. A high fraction of SVs (on average 61.26 Mb, in each variety) appeared in heterozygous condition. Thus, extensive regions of the grapevine genomes are hemizygous, an estimate confirmed also by the *de novo* assembly of different genomes.

A great fraction of small SVs is induced by the movement of transposable elements (TEs). An annotation pipeline was developed as part of the present project to characterise the SVs shaping the dispensable fraction. Approximately 65% of the deletions were annotated to a transposable element superfamily. As required by the pipeline used for the detection of insertions, 95% of the insertions were classified as TE. Transposable elements of class I (moving through a 'copy-and-paste' mechanism) contributed to a greater fraction of SVs (54% and 79% of the deletions and insertions, respectively), while DNA transposons of class II (moving via a 'cut-and-paste' mechanism) contributed only for a smaller fraction (11% and 16% of the deletions and insertions, respectively).

While promoting genetic variability, TEs may also disrupt genes, promoter or enhancer sequences or alter the status of epigenetic marks, such as cytosine

methylation. We observed that Gypsy TEs accumulated in pericentromeric heterochromatic regions: regions poor in coding sequences while rich in repetitive sequences. In the same regions an increase of the methylation levels was observed in the CG and CHG contexts. In addition, SVs affected overall 10,899 genes: a number of genes significantly lower than what expected by chance. Genes belonging to the *nucleotide binding* category (in part related to disease resistance) were mostly influenced by SVs. Lastly, transcriptomic analysis in three different tissues (leaves, berries and tendrils) of five varieties revealed that genes disrupted in the exonic regions showed a lower than average expression and a higher than average probability of being non-expressed, while genes affected by SVs in introns had a higher than average expression.

With the present work we created a detailed catalogue of structural variants in grapevine. We investigated the genome-wide distribution of SVs in a high number of varieties and estimated the pan-genome total size of *Vitis vinifera*. The fraction of genetic diversity captured by our set of 50 varieties makes us confident to have comprehensively described the pan-genome of this crop. Lastly, by measuring the pan-genome saturation we observed that a reduced number of varieties was sufficient to explain a great fraction of the dispensable portion: approximately 35 varieties explained more than 95% of the total SVs identified in grapevine.

1 INTRODUCTION

1.1 NOVABREED project

The present PhD work is part of the ERC-funded project, called NOVABREED, whose aim is to study the most variable and less characterised portions of the plant genomes. The goal of the project is to investigate genome-wide two main plant species, *Vitis vinifera* and *Zea mays*, paying particular attention to the intergenic and repetitive regions, often discarded because considered junk DNA regions. In the last decades, the increasing availability of sequenced genomes revealed the presence of high levels of genetic diversity among the individuals of a species. While human and primate genomes have been deeply investigated, in plants little advances have been done. With an extensive genome-wide analysis of grapevine and maize, the NOVABREED project aims to gain knowledge of the genetic diversity shaping the genomes of both species and to understand the molecular mechanisms at the heart of the diversity in living organisms.

1.2 Grapevine domestication history

In the present work we focused our attention on the *V. vinifera* genome and investigated the genetic diversity at the population level. Grapevine is one of the most economically important and widely cultivated crops, with an ancient history of domestication. Evidences of domestication date back to 8-10 thousand years (Neolithic Age) in the Eurasian region, from the dioecious *V. vinifera* L. subsp. *sylvestris* (Levadoux L, 1956). The Caucasian region (from South Caucasus to northern Mesopotamia) is considered the first domestication

centre of grapevine (Figure 1), where *V. vinifera* L. subsp. *sativa* and *V. vinifera* L. subsp. *sylvestris* coexist. Domesticated cultivars differ from the wild grapevines for several characters. Among them the hermaphroditism represented an essential character of domestication: while wild grapevines are dioecious, the domesticated grapevines are hermaphroditic, a shift essential for grapevine productivity. From the first centre of domestication, grapevine spread to neighbouring regions towards the East Mediterranean Basin, to South Italy, Spain, France until Central Europe, mainly through the rivers route (Danube, Rhone, Rhine) (McGovern PE, 2003; This P et al., 2006).

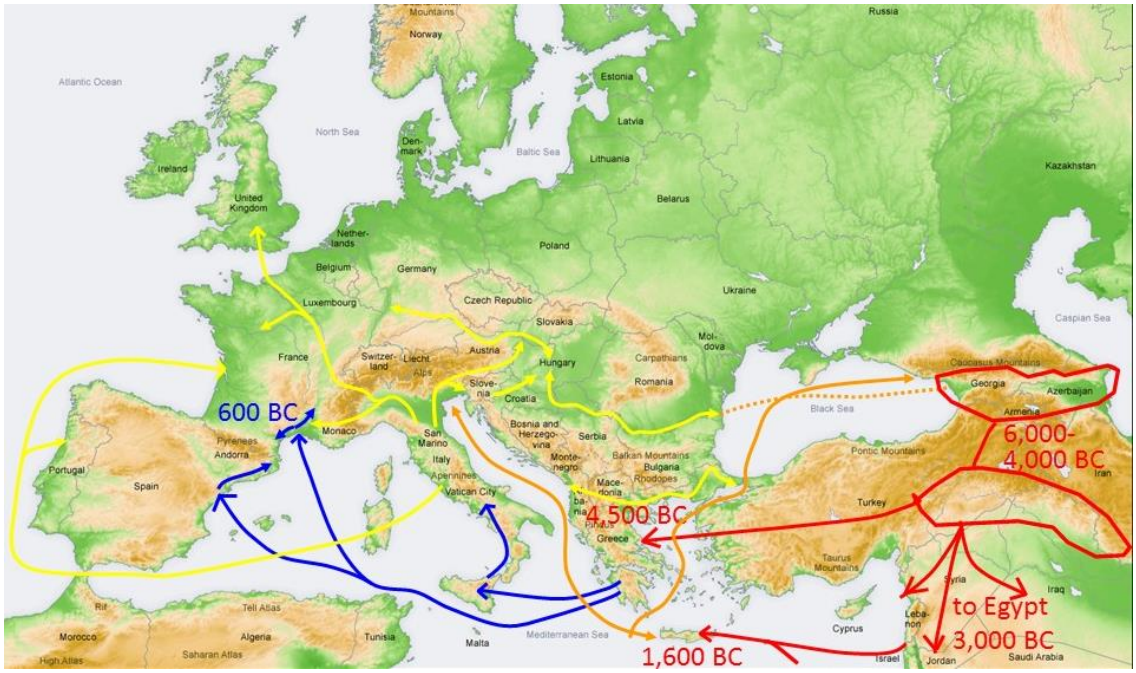


Figure 1: Domestication and diffusion routes of viticulture.

Several studies on grapevine genetic diversity revealed the existence of secondary domestication centres in the Mediterranean area or gene flow between introduced varieties and spontaneous grapes. The West and Central Europe grapevine varieties showed evidences of introgression from local wild (*sylvestris*) grapevines (Grassi F et al., 2003; Arroyo-Garcia R et al., 2006; Myles S

et al., 2011). Based on the migratory routes, in 1946 the Russian ampelographer Negrul AM proposed a classification of the grapevine varieties into three main eco-geographical groups, each with a distinct genetic origin from different wild grapevine populations. Each group, or *Proles*, was in turn divided into subgroups, or *sub-proles*. The groups and subgroups were described as follow: *Proles orientalis*, divided into *caspica* and *antasiatica*, spread in the area between the Caspian Sea and Central Asia; *Proles pontica*, diffused from Eastern Europe to Georgia, divided in turn into *balcanica* and *georgica*; lastly, *Proles occidentalis*, divided in *gallica* and *pyrenaica*, spread in Central and Western Europe (Negrul AM, 1946). Compared to *pontica* and *orientalis*, grapevine varieties belonging to *Proles occidentalis* are characterised by smaller berries and stronger resistance to cold temperatures.

1.3 Grapevine genome

Vitis vinifera is a dicotyledonous perennial species, mainly propagated vegetatively, characterised by highly heterozygous individuals. The grapevine genome was first assembled in 2007 by the French-Italian Public Consortium for Grapevine Genome Characterization (Jaillon O et al., 2007). High levels of heterozygosity in cultivated varieties hinder the assembly procedure. Therefore, a near homozygous individual (93% of homozygosity), derived from an accidental cross between Pinot Noir and Helfensteiner (itself derived from the cross Pinot Noir x Schiava Grossa) followed by successive selfings, was used for the genome assembly. A draft sequence of 487 Mb was assembled: a genome size comparable with the *Populus trichocarpa* genome (485 Mb) and the *Oryza sativa* genome (389 Mb). By combining different analysis approach, 30,434 genes were annotated, while on average 41.4% of the grapevine genome consisted of repetitive and transposable elements. By exploring conserved gene order in paralogous regions within chromosome triplets, Jaillon O and

colleagues (2007) concluded that three ancestral genomes contributed to the haploid grapevine content.

1.4 Next Generation Sequencing

In the last decades DNA sequencing has undergone rapid advances (Clyde 2007). Historically, DNA sequencing relied on the capillary-based biochemical Sanger method, which consisted in the electrophoretic separation of chain-termination products. The increasing request for low-cost sequencing has driven, in the last years, the development of new approaches, namely Next Generation Sequencing (NGS) technologies. Different NGS commercial platforms have been developed for DNA sequencing: Roche/454 (Margulies M et al., 2005), Solexa (Illumina) (Turcatti G et al., 2007), AB SOLiD and Polonator (Shendure J et al., 2005), and HeliScope (Harris TD et al., 2008). Through massive parallelization of the sequencing process, NGS platforms have increased drastically on one side the speed of sequencing, producing thus an enormous volume of data, and on the other side reduced the costs by several orders of magnitude: a single sequencing run generates hundreds of gigabases of nucleotide sequences (Mardis ER, 2008).

Illumina developed one of these NGS technologies. The first Solexa (Illumina) sequencer, the Genome Analyzer, was introduced in 2006 and since then consistent improvements have been made. The fast development of the sequencing technologies led to a drastic change of the sequencing performances. While the first sequencer produced 1 Giga base (Gb) of data in a single run, in approximately 2.5 days (Bentley DR et al., 2008), the HiSeq 2500 is nowadays able to produce up to 1,000 Gb of data in one single run, in approximately six days, with 2 billion of reads produced per flow-cell.

In common with other technologies, the Illumina protocol used for the library preparation may be summarized as follows (Figure 2):

- The genomic DNA template is fragmented randomly into small fragments, either through sonication or nebulization. Fragments are then end-repaired in order to generate blunt ends, and a single A is added to the 3' blunt end of the DNA fragments.
- Universal oligonucleotide adapters are ligated at both ends of the DNA fragments. The ligation is supported by the overhang of a single T at the 3' end of the adapter sequences.
- The library obtained is then immobilized on the surface of a flow-cell. The adapter sequences are complementary to the anchors spread over the flow-cell surface allowing thus the ligation. The DNA fragments are then amplified through a bridge amplification step in order to generate clusters of amplicons, constituting the templates for the sequencing reaction.

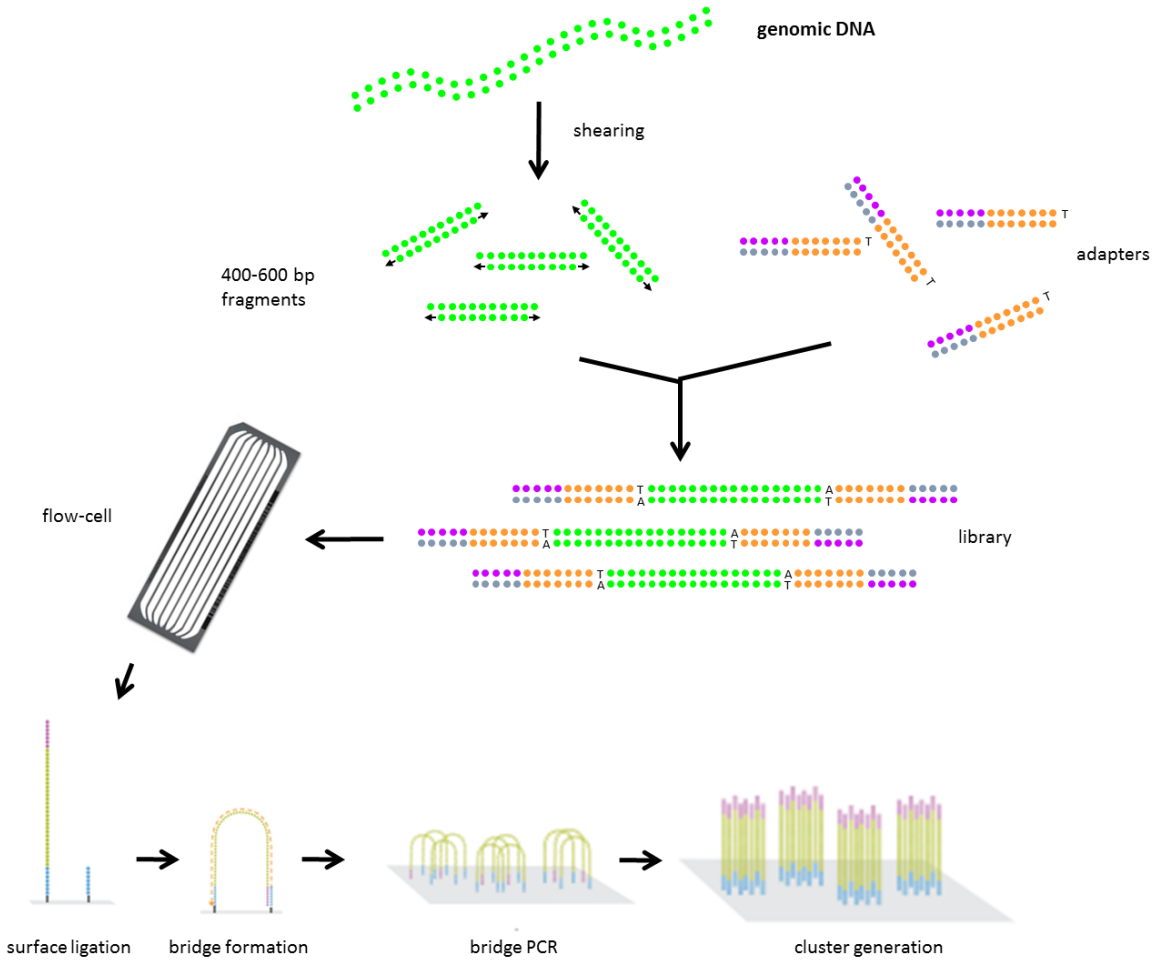


Figure 2: NGS library preparation workflow.

Sequencing by synthesis (SBS) occurs on single stranded sequences. The sequencing of the forward strand is mediated by the hybridization of a primer complementary to the adapter sequences which activate the sequencing mechanism. DNA polymerase and a mixture of four fluorescently labelled reversible dye terminators are then added to the solution: only the terminators complementary to the template sequence are bound. Further synthesis is prevented by the reversible terminator. Reagents present in excess are then washed away, a laser excites the fluorescent tag and images are recorded. Once the identity of the bases of the cluster is determined, the fluorescent tags are cleaved and removed and the reversible dye terminators are unblocked: a new

sequencing cycle can begin: the process is repeated until the end of the sequencing run (Figure 3).

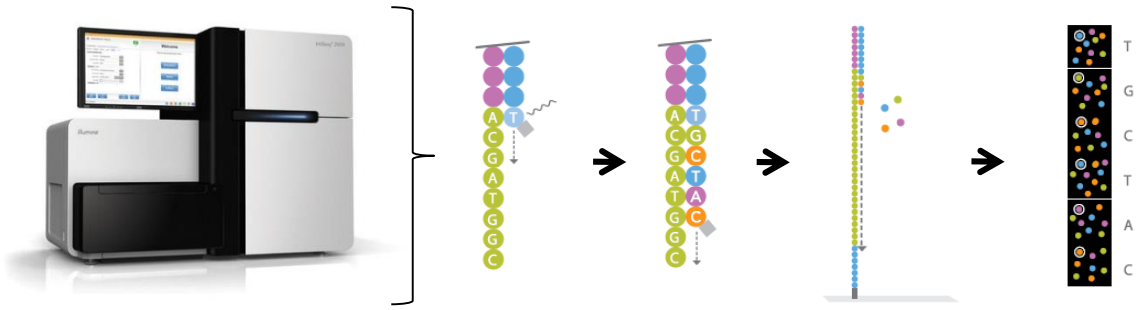


Figure 3: Sequencing by synthesis.

Sequenced reads may be employed in several ways: reads may be aligned or mapped to an available reference genome for different purposes, or used as bricks for the *de novo* assembly of genomes (Horner DS et al., 2010; Magi A et al., 2010). Besides single read sequencing, NGS technologies offer the possibility to sequence both ends of the template sequences. The so called paired-end method enables a genome-wide screening of a wide range of structural variants and chromosome rearrangements (Korbel JO et al., 2007). Furthermore, paired-end sequencing improves the alignment and the assembly of genomes, with a better resolution of the repeated regions.

NGS technologies have been employed in several fields of research. NGS allows both the complete genome resequencing, as well the reduced representation sequencing. Resequencing of whole genomes (whereby short reads are compared to a reference genome) enables the identification of Single Nucleotide Polymorphisms (SNPs) and the detection of structural variants. Through bisulfite-treated DNA sequencing, NGS allows the exploration of the methylation levels inside the genomes. Furthermore, NGS enables the analysis of gene expression and microRNA profiling, through RNA and small-RNA

sequencing, or the genome-wide exploration of the DNA-protein interactions, via Chromatin ImmunoPrecipitation-Sequencing (ChIP-Seq). Based on Chromosome Conformation Capture, the Hi-C method enables the genome-wide discovery of chromatin interactions. While, Genotyping By Sequencing (GBS) may be employed for the analysis of large number of SNPs, through a highly multiplexed approach based on restriction enzymes (in order to reduce genome complexity). Lastly, environmental samples may be investigated through metagenomic sequencing.

Crop genome sequencing benefited from the development and improvement of next generation sequencing technologies. NGS paved the way to the identification of new molecular markers and genes influencing agronomically important traits (Varshney RK et al., 2009; Edwards D & Batley J, 2010). Whole-genome sequencing enabled a better understanding of the genome complexity, with single nucleotide resolution (Rastogi K et al., 2013).

1.5 Structural variation analysis

In the past it was thought that the intra-species DNA sequence variation was mostly due to Single Nucleotide Polymorphisms (SNPs) (Sachidanandam R et al., 2001), but several works demonstrated that the human genomes differed more as a result of Structural Variation (SV) than SNPs (Iafrate AJ et al., 2004; Sebat J et al., 2004; Scherer SW et al., 2007; Hurles ME et al., 2008; Conrad DF et al., 2010).

SV is a broad term traditionally used to describe chromosomal alterations which involve DNA sequences longer than 1 Kb (Feuk L, 2006; Freeman JL et al., 2006).

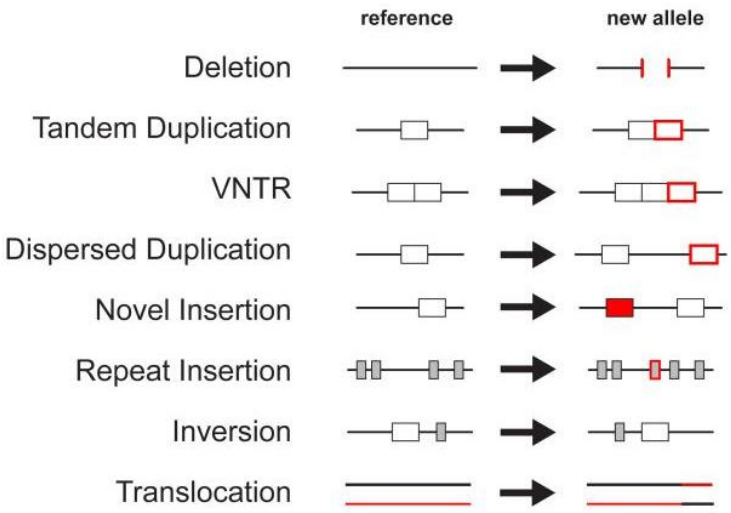


Figure 4: Structural variants observed in the sample (new allele) compared to the reference genome (Hurles ME et al., 2008).

As reported in Figure 4, SV includes both balanced alterations (for example translocations or inversions) and unbalanced variants (for example deletions, insertions or duplications), which alter the DNA copy number and are generally referred as Copy Number Variants (CNVs). Extreme CNVs are usually referred as Presence-Absence Variation (PAV), sequences present in the genome of an individual, but missing in another genome.

Evidences of SV in humans increased gradually in the last decades, and it has been demonstrated that SVs lie at the heart of several diseases with a genetic aetiology (Hurles ME et al., 2008). Structural variants may influence gene expression in different ways, playing thus an important role in the human phenotypic variation. In human genome, SVs have been deeply investigated (Raphael BJ, 2012), while in plants few efforts have been made trying to understand the role of SVs (Saxena RK et al., 2014). Structural variants have been investigated in different species: in *Arabidopsis* (DeBolt S, 2010), in barley (Muñoz-Amatriaín M et al., 2013), in maize (Springer NM et al., 2009; Beló A et al., 2010), in melon (Sanseverino W et al., 2015), in soybean (McHale LK et al.,

2012), in rice (Xu X et al., 2012), in sorghum (Zheng L-Y et al., 2011) and in grapevine (Giannuzzi G et al., 2011; Di Genova A et al., 2014).

Historically, the detection of structural variants was based on whole-genome array Comparative Genomic Hybridization (aCGH) (Medvedev P et al., 2009). CGH arrays relied on the comparative hybridization of two differently labelled samples (a test and a reference) to a set of hybridization targets, in order to test the relative frequencies of DNA probe fragments between two samples (Figure 5). CGH array platforms were originally developed for the evaluation of differences between normal and solid tumour tissues (Kallioniemi A et al., 1992). However, CGH arrays presented various disadvantages. Microarrays were used for the detection of copy number variations only for sequences present in the reference genome (sequences required for the probe design), but could not identify balanced structural variants. Furthermore, no information about the genomic position of duplicated sequences was obtained, and the breakpoint resolution was usually low. Lastly, hybridization was limited in highly repeated sequences where CNVs occur very frequently (Alkan C et al., 2011; Raphael BJ, 2012; Saxena RK et al., 2014).

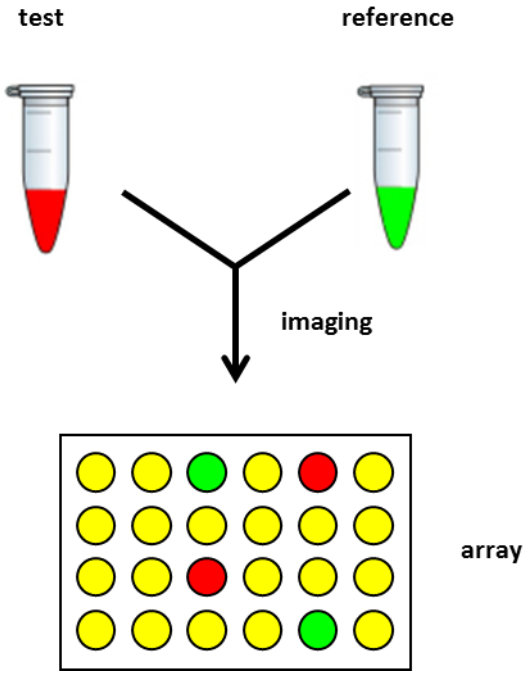


Figure 5: Comparative Genomic Hybridization array (aCGH). Two differently labelled samples are hybridized to a set of hybridization probes. Differences in the intensities of the two fluorophores reveal differences in copy number between the two samples.

Structural variation studies changed drastically with the advent of NGS technologies which replaced the microarray techniques (Alkan C et al., 2011). This introduced bioinformatics and computational challenges. The short length of the sequenced reads increased the difficulties of read mapping, especially in highly repeated regions. In addition, NGS platforms produced a higher throughput, at relatively low costs, but at the same time with lower accuracy (Shendure J & Ji H, 2008): this scenario required the development of bioinformatics tools to facilitate the mapping of a high number of short reads to a reference genome.

Four different methods for the SV discovery have been developed (Medvedev P et al., 2009; Mills RE et al., 2011). The first three strategies require an alignment of the reads to a reference genome followed by the analysis of discordant patterns which explain different classes of structural variation. The last

approach (*de novo* assembly) enables the identification of variants not present in the reference genome, since it doesn't rely on a reference sequence. The four methods are described below.

Paired-end mapping (PEM)

The read-pair technology takes advantage of the mapping information, mainly span and orientation, of paired-end reads aligned to a reference genome. Structural variants can be detected through "discordant" pairs, Paired Ends (PEs) mapping to the reference with anomalous orientation or with span inconsistent with the reference itself (Figure 6). Different SV categories may be detected. Deletions are identified by pair-reads mapping to the reference genome at a distance greater than the mean library insert size. Conversely, insertions are described by pair of reads mapping with a distance smaller than the expected one (only if the read pairs span the complete insertion). If the insertions have a size greater than the library insert size, the PEM method is unable to produce a signature. Inversions may also be detected: reads spanning an inversion breakpoint will map to the reference genome with an opposite orientation. The PEM method was first developed for BAC end sequences (Volik S et al., 2003) and only later employed for NGS analysis (Korbel JO et al., 2007). Nowadays, PEM is the most widely applied approach for the SV detection and several tools for the detection have been developed (Alkan C et al., 2011).

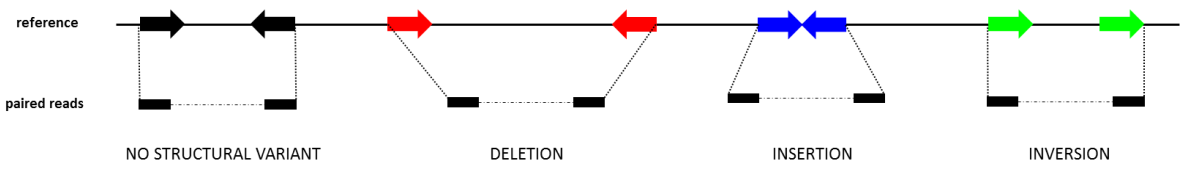


Figure 6: Paired-end method (PEM). Detection of different SV categories, based on the alignment of paired-end reads to the reference genome.

Depth of coverage (DOC)

Assuming the sequencing process to be uniform, the number of reads mapping to a region is expected to be proportional to the number of times the region appears in the donor (Medvedev P et al., 2009). The Depth Of Coverage (DOC) method can be applied for the detection of structural variants which alter the copy number of a sequence (Figure 7). By measuring the increase or decrease in sequence coverage, DOC enables the detection of duplications and deletions, respectively. Read-depth methods exploiting NGS data were developed to define SV rearrangements in cancer tissues (Campbell PJ et al., 2008). In contrast to PEM signatures, DOC signatures enable the detection of large SV events, with less breakpoint resolution power, while the method has less power in the detection of smaller events.

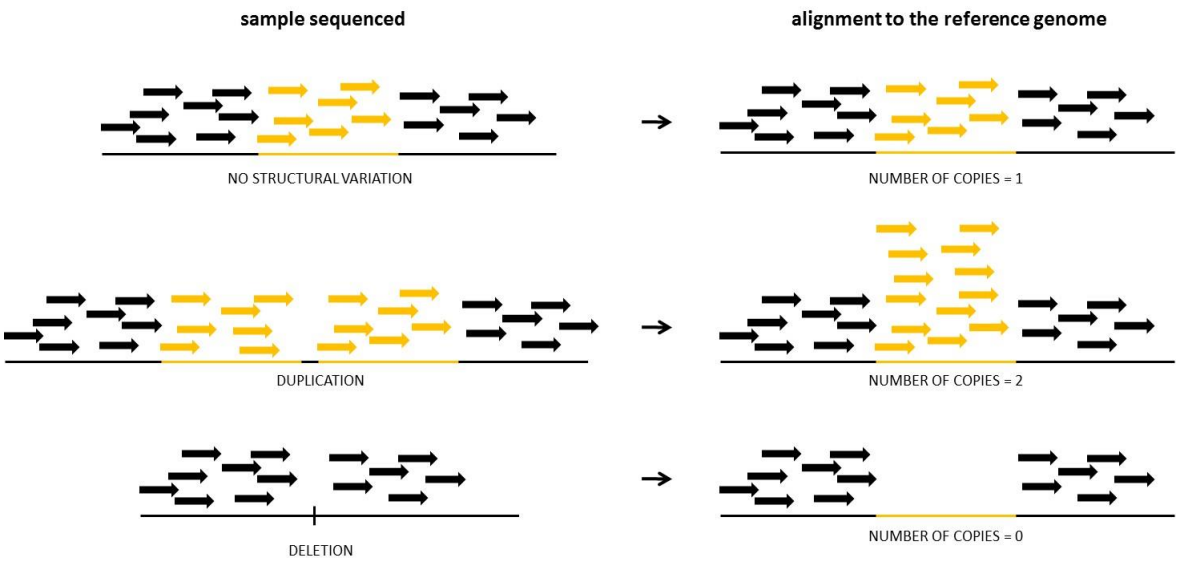


Figure 7: Depth of coverage (DOC) method. Reads obtained from the sequenced sample (left part of the picture) are aligned to the reference genome (right part of the picture). Realignment on reference genome enables the detection of duplications or deletions, based respectively on the increase or decrease of the local read alignment coverage.

Split-read approach

The split-read method enables the single-base-pair breakpoint resolution of different SV categories. Based on the signature of split reads, this approach defines the exact breakpoint of a structural variant. The application of this method is still limited, since it needs to cope with the alignment and correct location of short read fragments. However, different computational tools already exploit this information for a better resolution of the SV identification.

De novo assembly

With the *de novo* assembly of a genome, theoretically all SV categories may be identified. Combining *de novo* and local assembly, contig sequences are generated and then compared with the reference genome. *De novo* assembly still suffers from the resolution of highly repeated or duplicated sequences, reducing thus the ability in the detection of SVs that involve repeated sequences such as transposable elements.

1.6 Transposable elements and plant pan-genome

Chromosomal rearrangements may occur with different sizes. Two groups of structural variants may be described: small SVs ranging in size between 1 and 25/30 Kb, and larger SVs which extend to several Mb. Smaller SVs in higher plants are mainly influenced by the movement of Transposable Elements (TEs). Transposable elements were first described in plants in the '40s as controlling elements, based on their ability to influence the nearby genes' expression (McClintock B, 1956). Once, TEs were considered junk DNA sequences (Orgel LE & Crick FHC, 1980), but nowadays it is ascertained that TEs may play important roles and influence several cellular functions. Among plants, TEs in extant angiosperms are young and have been very active in the recent past, through bursts of activity (El Baidouri M & Panaud O, 2013; Oliver KR et al., 2013).

Compared to mammals, TEs in angiosperms are more active (Kejnovsky E et al., 2009), creating thus structural variation in the plant species, and widely influence genome size and structure, contributing up to 84% of the latter genomes (Kumar A & Bennetzen JL, 1999).

Based on the mechanism of transposition, two different classes of TEs are known (Wicker T et al., 2007; Levin HL & Moran J V., 2011): class I transposable elements, or retrotransposons, which move along the genome via a copy-and-paste mechanism (Figure 8A, Figure 8B); class II transposable elements, also known as DNA transposons, which move inside an individual through a cut-and-paste mechanism (Figure 8C).

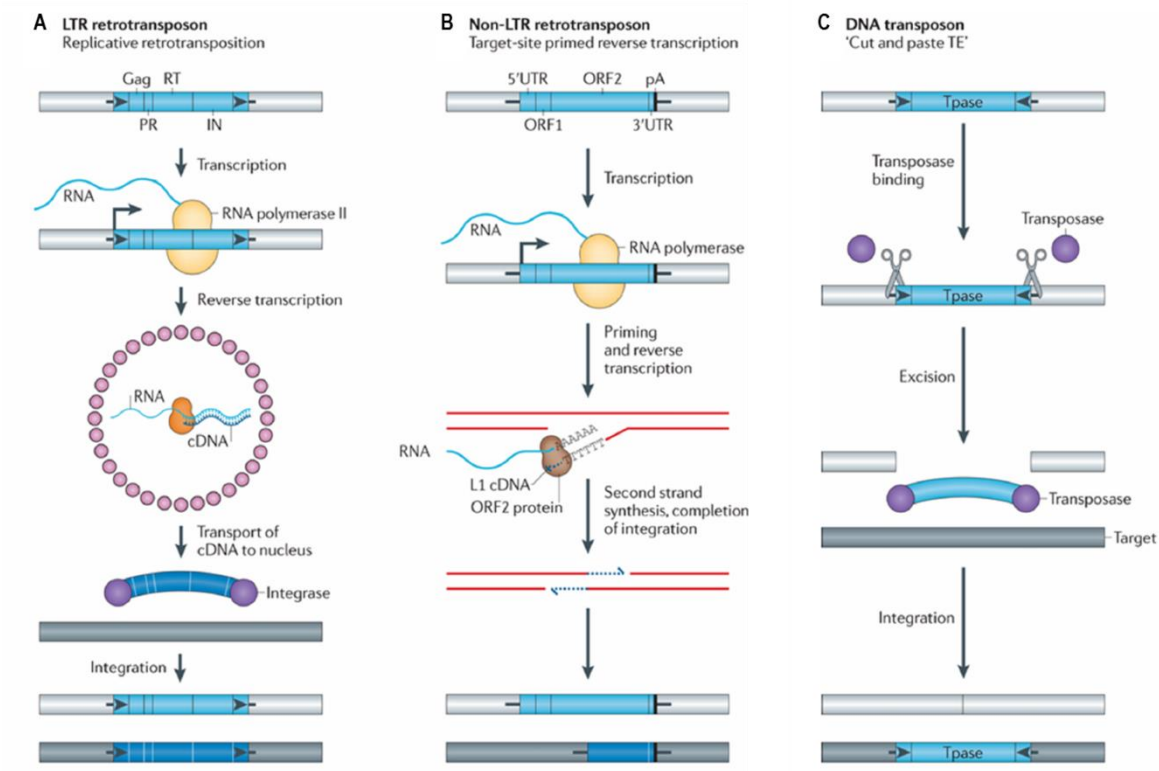


Figure 8: Mechanisms of transposon mobilization. A. & **B.** Copy-and-paste mechanism of the class I retrotransposons. **A.** Replicative strategy of the Long Terminal Repeat (LTR) retrotransposons. **B.** Transposition method of non-LTR retrotransposons. **C.** Cut-and-paste mechanism of the class II DNA transposons (modified from Levin HL & Moran J V., 2011).

Class I elements move in the genome through an RNA intermediate, reverse transcribed by a reverse transcriptase encoded by the TE itself. Among the retrotransposons, different orders are known: Long Terminal Repeat (LTR), Long and Short Interspersed Nuclear Element (LINE and SINE, respectively), *Dictyostelium* Intermediate Repeat Sequence (DIRS) and *Penelope-Like* Elements (PLE). Class II TEs code for a transposase, a protein required for their excision and consecutive insertion in another genomic position. Class II elements are divided in two subclasses, differentiated by the number of DNA sequences involved in the transposition process. Subclass I is represented by Terminal Inverted Repeat (TIR) elements and *Crypton* TEs (unique to fungi), which move via the classic cut and paste mechanism. On the other hand, subclass II transposable elements involve only one DNA strand (Wicker T et al., 2007). Two main orders have been described: *Helitron*, which replicate through a rolling-circle mechanism, and *Maverick* TEs. Orders are in turn divided in superfamilies. Although TE superfamilies within the same class share the same replication strategy, they differ by copious features. Superfamilies are then divided in families, based on the DNA sequence conservation (Wicker T et al., 2007).

Transposable elements influence wide portions of the eukaryotic genomes, and, as stated by Wicker T and colleagues (2007), LTR retrotransposons are the most abundant order in plants. TEs affect the structure and evolution of the plant genomes in several ways: transposable elements may modulate gene expression, by silencing or altering the expression of nearby genes; TEs may contribute to chromosomal rearrangements via non-homologous recombination; furthermore, TEs influence the local methylation patterns (Feschotte C et al., 2002; Eichler EE & Sankoff D, 2003; Bennetzen JL, 2005; Slotkin RK & Martienssen R, 2007; Lisch D, 2012). The grapevine berry colour is one of the striking examples of gene expression modulation mediated by TE. The berry colour ranges in a wide spectrum of colours, varying between white

and black. The coloured phenotype is controlled by two transcription factors, *MybA1* and *MybA2*, which control the anthocyanin biosynthesis pathway. White varieties, which lack the anthocyanin pigments, are characterised by an inactivation of the *MybA1* transcription factor, caused by the insertion of *Gret1* (transposable element of class I) in the promoter region of the gene. White varieties are homozygous for the *Gret1* insertion, while coloured cultivars carry at least one functional allele at the *MybA1* locus. As reported in 2004 by Kobayashi, 10 out of 10 white varieties carried the homozygous white haplotype, while all (9) red varieties were heterozygous at the colour locus, proving the coexistence of the inactivated *MybA1* gene (white haplotype) with the functional or coloured haplotype. These observations were confirmed also in a later work of Fournier-Level A and colleagues: out of 137 grapevine varieties, only nine varieties were homozygous at the colour locus (carrying both coloured haplotypes) and produced darker berries (Fournier-Level A et al., 2010).

Evidences of structural variation revealed that the genomes of individuals belonging to the same species were characterised by several alterations, encompassing both small variants due to the transposable element movement and larger ones, which altered the chromosomal structure. Based on these observations the concept of pan-genome, originally introduced for bacteria (Tettelin H et al., 2005), was extended to plants by Morgante M and colleagues in 2007. The pan-genome is composed of a core fraction, shared between all the individuals of the species, and a dispensable portion present only in some individuals, but not in all. While the core genome includes single copy sequences, thus the majority of genes, and few transposable elements shared by all individuals; the dispensable fraction is mostly composed of transposable elements and repeated sequences found in a specific location only in some individuals (Morgante M et al., 2007). Although the dispensable genome is not essential for survival, it might play an important role in shaping the genomes

structure and mediating the response to environmental stimuli (Marroni F et al., 2014).

2 OBJECTIVES

The main objective of the present project was the characterisation of the dispensable portion of *Vitis vinifera* pan-genome, mostly attributable to structural variants (SVs), and to investigate how genetic relationships, measured using SVs, compare to those estimated using SNPs. In an era where the genomes of several individuals may be re-sequenced rapidly and at relatively reduced costs, we first used SNP markers to gain knowledge on the genetic relationships and population structure in more than one hundred individuals. In the last decades next-generation sequencing enabled an accurate genome-wide analysis of the structural variation affecting genomes of any species (Korbel JO et al., 2007; Campbell PJ et al., 2008). Comparative sequencing revealed that the genomes of individuals belonging to a species are shaped by high levels of structural variation. The latter encompasses both smaller alterations mediated by the movement of transposable elements, as larger ones which modify the chromosomal structure. Thus, the genomic complement of a species is better explained by its pan-genome, composed by a core fraction, shared by all individuals and a dispensable portion, unique to some. We aimed to explore genome wide the small structural variants (ranging in size from 1 to 25 Kb) shaping the pan-genome of *Vitis vinifera*. The aim was to characterise the dispensable fraction of the pan-genome, through a combination of different analysis methods. Besides the discovery of insertions and deletions, and their accurate description, by mean of the transcriptome sequencing we investigated how SVs affected gene expression. With the results of the present work we gained knowledge about the composition of the dispensable portion and gave a first estimate of the grapevine pan-genome total size.

3 MATERIALS AND METHODS

3.1 Plant material

We analysed 128 *Vitis vinifera* varieties originating from different locations across the European and Asian regions (see Table 2 for details about the country of origin). Leaf tissues were collected from plants held in different collections. Plant material was collected for 54 varieties at the Experimental farm "A. Servadei", Udine (UD, Italy); 51 cultivars were sampled at the *Consiglio per la ricerca in agricoltura e l'analisi dell'economia agraria*, *Centro di ricerca per la viticoltura* (CREA, CRA-VIT), Conegliano (TV, Italy); 12 varieties were sampled at *Vivai Cooperativi Rauscedo*, Rauscedo (PN, Italy); 5 cultivars were sampled at the *Kmetijsko gozdarski zavod Nova Gorica*, Loze, Vipava (Slovenia). The raw read sequences of 6 additional grapevine varieties were instead acquired from the Sequence Read Archive (SRA) (Wheeler DL et al., 2005). The following varieties and runs were selected: Autumn royal, run SRR354199 (experiment SRX101831); Italia, runs SRR354198 and SRR769824 (experiments SRX101830, SRX247604); Red globe, runs SRR769829 and SRR354201 (experiments SRX247609, SRX101832); Tannat, run SRR863595 (experiment SRX283507); Sultanina, runs SRR931841, SRR931842, SRR931843, SRR931844, SRR931845, SRR931846 and SRR924196 (experiment SRX316886); Thompson seedless, runs SRR769825 and SRR354200 (experiments SRX247605, SRX101833).

3.2 Library preparation and sequencing

DNA paired-end libraries were generated from genomics DNA, according to the standard Illumina paired-end sample preparation guide (Illumina Inc., San Diego,

CA, USA), with slight modifications. DNA was extracted following a modified Zhang protocol (Zhang H-B et al., 1995) and then sheared by sonication. The resulting mixture, composed of fragments with sticky ends (both 3' and 5') and blunt ends was treated with T4 DNA polymerase and *Klenow enzyme*, in order to perform end repair, and an 'A' was added to the 3' ends of the obtained blunt fragments in order to facilitate the ligation of Illumina adaptors. Unligated adaptors were then removed and the obtained libraries validated. The NanoDrop ND-1000 UV-Vis Spectrophotemeter (Thermo Scientific, Wilmington, DE, USA) and Agilent 2100 Bioanalyzer (Agilent Technologies, Santa Clara, CA) were respectively used for the quantification and for the quality assessment. The libraries were then immobilized to the surface of the Illumina flow-cell via the Illumina Cluster Generation Station (cBot) and sequenced by synthesis in one lane of the Illumina HiSeq2500 at the Institute of Applied Genomics (Udine, Italy). Based on the library type, the sequencing was accomplished with 101, 126 or 133 cycles per read. The CASAVA 1.8.2 version of the Illumina pipeline was used to process raw data.

Raw sequences were quality trimmed and contaminant filtered using erne-filter version 1.2 (Del Fabbro C et al., 2013) and adapters were removed with cutadapt version 1.1 (Martin M, 2011). Short reads sequences were then mapped against the *Vitis vinifera* reference genome sequence (Jaillon O et al., 2007) using the software package BWA version 0.7.5a (Li H & Durbin R, 2009) with the default settings (seed length 32, mismatch penalty 3, gap open penalty 11, gap extension penalty 4). The output of the aligner in Sequence Alignment/Map (SAM) format was sorted and transformed to Binary Alignment/Map (BAM) file through the software package SAMtools version 0.1.18 (Li H et al., 2009). PCR duplicates were then removed with samtools rmdup command and uniquely aligned reads were selected for further analyses (reads mapping in multiple positions on the reference genome were discarded).

Insert size statistics were computed with the *CollectInsertSizeMetrics* command of the 1.88 version of the Picard suite (http://broadinstitute.github.io/picard).

The mean coverage of each individual was calculated dividing the total number of unique aligned bases by the number of covered positions. The physical coverage was computed as above, but considering the insert-size information, including thus the bases not sequenced, but comprised between the two sequenced reads.

To assess the quality of the sequencing data, KmerCounter (developed by our research group) was used (on trimmed reads) to measure the k-mers distribution with a k value of 16. Instead, the coverage profile was computed on the uniquely aligned reads, measuring the coverage of each single base.

3.3 SNP analysis

The software package GATK version 3.3-0 (McKenna A et al., 2010) was used for SNP calling. Alignment files generated by BWA (Li H & Durbin R, 2009) were parsed with the tools *CleanSam* and *FixMateInformation* of the Picard suite, to provide the correct input format to GATK. The GATK *RealignerTargetCreator* command was used to define intervals in proximity of indels, targeted for local realignment. *IndelRealigner* (with default settings) performed then the local realignment over the intervals defined by the previous tool. Lastly, the variant discovery tool *UnifiedGenotyper* was applied (with *heterozygosity* parameter set to 0.01) in order to call SNPs in each variety (Van der Auwera GA et al., 2002; DePristo MA et al., 2011).

The raw SNPs identified in the grapevine population were filtered by quality and coverage as follows:

- Only SNPs with a Phred-scaled quality score greater than 50 were selected.
- In each variety, only positions with a coverage ranging between 0.5 and 2.5 times the modal coverage value were considered for SNP calling.
- SNP positions were discarded if the number of non-informative varieties (with missing data) was greater than 50% the total number of varieties compared in the analysis.
- SNPs in regions characterised by repeats or microsatellites were removed from the analysis. Repetitive regions were identified and masked based on ReAS annotations (Li R et al., 2005), on Sputnik annotations (Abajian C, 1994) and on a hand curated database of transposable elements (Dario Copetti, PhD thesis).

We examined the population structure of *V. vinifera* with ADMIXTURE (Alexander DH et al., 2009) with different values of K (number of ancestral populations) varying between 2 and 15. Twenty independent runs of ADMIXTURE were carried out, with randomly generated seed for each run and for each value of K. The cross-validation (CV) method implemented in ADMIXTURE was used to identify the K value with the best predictive accuracy. In addition, the ΔK method (Evanno G et al., 2005) was employed to identify the true value of K. The mean Log likelihood, L(K), was calculated over 20 runs for each value of K. The first order rate of change of the Log likelihood was measured as the mean difference between two successive values of L(K): $L'(K) = L(K) – L(K-1)$. The second order rate of change of L(K) was estimate as the absolute value of the difference between two successive values of L'(K): $|L''(K)| = |L'(K+1) - L'(K)|$. Lastly, the ΔK value was calculated as the mean of the

L''(K) values divided by the standard deviation of L(K). ΔK was calculated as following: ΔK = |L''(K)| / s[L(K)]. The modal value of the ΔK distribution should be located at the real number of clusters K.

The genotypic distance between any two varieties was computed comparing the SNP genotype at each position. Genetic distance between two varieties at any given SNP was set to 0, 0.5, or 1 if they shared two, one or no alleles, respectively. We summarized the distances for each chromosome with the following equation:

$$\frac{(0.5 * \text{positions 1 allele shared}) + (1 * \text{positions 0 allele shared})}{\text{total SNP positions}}$$

Genome wide the genotypic distance measure was normalized dividing the single chromosome distance measure by the total number of variant positions in the chromosome. The principal coordinates of the genome wide genotypic distance were plotted with the R function cmdscale of the *stats* package (R Core Team, 2013).

In addition, pairwise haplotype distance was estimated as follows in windows of 100,000 non-repetitive bp: for each window the number of haplotypes shared between two varieties was inferred according to a slightly modified version of the method implemented by Wu GA and colleagues (Wu GA et al., 2014). Briefly, for each window, each position was classified as IBS0, IBS1, IBS2 or 0, if respectively, no alleles were shared, one allele was shared or both alleles were shared in heterozygous condition (IBS2) or in homozygous state (0). Genome wide the haplotype distance was normalized as accomplished for the genotypic distance measure.

To measure the haplotype diversity, phase was inferred and missing genotypes were imputed based on localized haplotype clustering with Beagle version 3.2.2,

with 100 iterations (Browning BL & Browning SR, 2009). Haplotype diversity was then estimated in blocks of five consecutive markers.

For each window of 100 non-repetitive Kb, we further measured the pairwise linkage disequilibrium (LD) between SNPs positioned at less than five Kb. Only SNPs with a Minor Allele Frequency (MAF) above 0.2 were retained for the analysis. LD was computed with the R function LD of the *genetics* package (Warnes G et al., 2011).

Lastly, the mean SNP frequency of a 100 Kb window of non-repetitive bp was calculated in each variety by dividing the number of SNPs by the number of base-pairs with a coverage ranging between 0.5 and 2.5 times the modal value of the coverage and was then averaged over the grapevine population.

3.4 SV dataset simulation

To evaluate the performances of tools for the detection of SV, we simulated 1000 insertions and deletions in two different reference genomes: *Populus trichocarpa* (Tuskan GA et al., 2006) and *Vitis vinifera* (Jaillon O et al., 2007). Variants were simulated by randomly deleting 1000 repeated sequences, ranging in size between 1 and 25 Kb, from their original position in the reference genome and inserting them in a new randomly chosen position of the genome, creating thus a modified reference genome (*simulated SV genome*). Through the alignment of reads (obtained from the individual used to produce the original reference genome) to the *simulated SV* genome, insertions were expected where the 1000 repeated sequences had been removed, while deletions were expected where the sequences had been inserted (Figure 9).

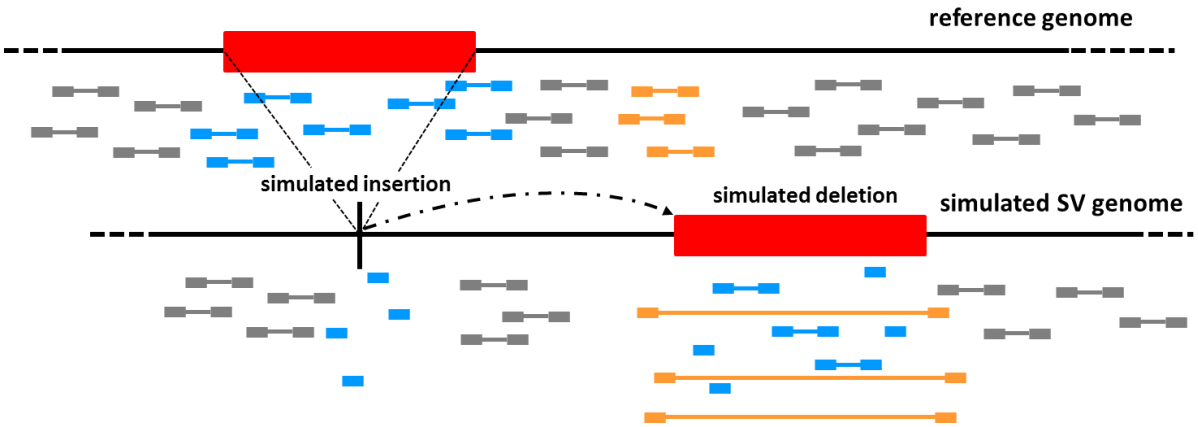


Figure 9: SV simulation outline. A simulated insertion (in the sample compared with the *simulated SV genome*) was created where the repeated sequence had been removed, while a simulated deletion was observed where the repeated sequence was introduced. Reads obtained from the individual that was used to produce the original reference genome sequence are displayed.

3.5 Benchmarked software packages for the detection of deletions

For the detection of deletions, four freely available tools were tested: CLEVER (Marschall T et al., 2012), DELLY (Rausch T et al., 2012), GASV (Sindi S et al., 2009) and Pindel (Ye K et al., 2009).

The 2.0rc3 version of CLEVER was used with the option *use_xa*, which enabled the interpretation of the XA tags in the alignment file, in order to get a better resolution of the deletions prediction. The software produced an output file in Variant Call Format (VCF) and only variants with size ranging between 1 Kb and 25 Kb were retained for analysis.

DELLY (version 0.3.3) was used with the default parameters (*map-qual* = 0, *mad-cutoff* = 5). The VCF output file was filtered selecting the deletions included in the 1 to 25 Kb size range and supported by at least two paired ends.

The 2.0 version of GASV was used. First, the *BAMToGASV* software was employed with default parameters to generate the GASV input file from the alignment file. Then, GASV algorithm was run with the *minClusterSize* option set to 2, which required at least two paired reads for the deletions prediction. Based on the information of discordantly mapped reads, GASV predicts the SV through a geometrical approach. Once identified the putative breakpoint, it draws a polygon and thus the breakpoint is reported as an interval. Therefore, the central points of the left and right polygons breakpoints were used for defining the deletion coordinates. Once again, the output file was filtered by keeping only the deletions ranging from 1 to 25 Kb in size. Furthermore, positions were discarded where GASV wasn't able to explain the data with a single structural variant (identified in the *Localization* field of the output file with the flag value -1, where the square root of the breakpoint region is reported).

Pindel version 0.2.5a3 was used with the *minimum_support_for_event* option set to 2, in order to require at least two supporting reads for the SV detection. The deletions with a size included between 1 Kb and 25 Kb and with a mean mapping quality above 20 were selected for further analysis.

3.6 Identification of deletions

The preferred method for detecting deletions in the whole *Vitis vinifera* dataset was the integration of results obtained by DELLY and GASV. DELLY was used with default parameters, and the deletions were selected in the size range between 1 Kb and 25 Kb, discarding those with a paired-end support lower than five and a median mapping quality below 20. GASV was run with default parameters, and once again the deletions ranging between 1 Kb and 25 Kb in size and with a support of at least five paired-end reads were selected. As previously explained, the central points of the SV intervals were used in order to

approximate the left and right breakpoint coordinates of each deletion identified by GASV. The deletions obtained by the two methods were then merged in each sample. If the coordinates of the deletions identified by the two methods overlapped on both extremities within 250 bp, the deletions were combined as a single event. The coordinates of DELLY were used for the overlapped deletions. Then, the total deletions identified in the 50 varieties were merged using a greater interval of 500 bp around the left and right breakpoints.

Lastly, the deletions were filtered for each sample based on the coverage information of the SV's flanking regions. The mean coverage of the unique aligned reads was computed separately in an interval of 500 bp spanning the left and right SV coordinates. The coverage information of the left and right region were then merged: positions of a variety with a total mean coverage greater than 2.5 or lower than 0.5 times the modal coverage value were considered non-informative.

In order to assign a genotype to the SV identified in each sample, the total dataset of deletions obtained was analysed using an internally developed Python script. For each deletion, the software retrieved from the alignment file of each variety the total number of reads supporting the deletion - **positive reads** - (reads mapping to the reference genome with a greater insert size than expected, spanning both deletion coordinates) and the total number of not supporting reads - **negative reads** - (reads supporting the reference genotype, mapping with an insert size similar to the expected one and spanning only one of the two SV breakpoints). The analysis was computed on regions of 500 bp flanking the left and right SV coordinates. The frequency of the variant was obtained dividing the number of positive reads (those supporting the SV) by the total number of reads (sum of positive and negative reads). Based on the allele frequency we refined the total dataset of deletions observed with DELLY and

GASV and assigned to each variety a genotype for the SV observed: the genotype was assigned only to samples with at least five reads (both positive and negative) in the regions flanking the SV. In any given individuals, the SVs with a positive to negative ratio below 0.25 were considered homozygous for reference; positions with a ratio between 0.25 and 0.75 were considered heterozygous; lastly, SVs with a ratio above 0.75 were identified as homozygous for alternate allele.

3.7 Transposable element annotation

The sequences of the identified deletions were extracted from the reference genome and used as query for the annotation of transposable elements. An annotation pipeline was developed by integrating several freely available tools. The database used for the annotation of deletions was composed of: 202 sequences obtained from RepBase (Jurka J et al., 2005), representing a non-redundant set of *Vitits vinifera* transposable elements, and 467 *V. vinifera* TE sequences, obtained from an internal database.

The annotation pipeline was composed of five progressive steps.

1. Tandem Repeats were masked in the deletions sequences with Tandem Repeat Finder (TRF) tool (Benson G, 1999) with the option *pattern size* < 170 bp. All deletions containing more than 80% of unknown (N) bases were removed from further analysis.
2. Stretches of 100 bp, corresponding to the extremities of the database sequences, were then aligned separately with RepeatMasker (Smit AFA, Hubley R & Green P, *RepeatMasker Open-3.0*) against 400 bp of the deletions extremities. A deletion was classified and annotated to a transposable superfamily if both ends mapped to both TE extremities belonging to the same superfamily.

3. LINE and solo-LTR were classified by aligning the deletions to the database of LINE and LTR elements using RepeatMasker. All deletions aligned for more than 80% of their length were classified at this step.
4. LTR_finder (Xu Z & Wang H, 2007) was used to discover novel retro-transposons missing in the TE database. Deletions were scanned, searching for LTR features. Deletions were classified as new LTR retrotransposons if the coordinates of the putative LTRs were located no more than 300 bp from both extremities of the deletions.
5. Lastly, all the deletions were classified using Teannot software, a package of the REPET suite (Flutre T et al., 2011). All the deletions were mapped to the database of known transposable elements taking advantage of different alignment programs such as BLASTER, RepeatMasker and CENSOR. An empirical statistical filter was used in order to discard false positive calls. Transposable element fragments belonging to the same TE were then concatenated with MATCHER and a *long joining* step was applied in order to merge relative distant fragments.

3.8 Dating of Long Terminal Repeat insertion events

Insertion dates of complete LTR retrotransposons were estimated according to the method developed by SanMiguel P and colleagues, considering the amount of divergence between the 5' and 3' LTRs (SanMiguel P et al., 1998). LTR_FINDER software was used to define, on the reference genome, the exact coordinates of the 5' and 3' LTRs within deletions, with the following parameters: max distance between LTRs *D* = 25000; min distance between LTRs *d* = 100; max LTR length *L* = 6000; min LTR length *l* = 50; LTR must have edge signal *E*; auto mask highly repeated regions *C*; length of exact match pairs *p* = 15; predict PBS based on a tRNA database *s*; predict protein domains *a*; signal status control *F* = 11110000000 (equivalent to requiring that both LTRs

have the TG signal at the 5' end and the CA signal at the 3' end, and not requiring any additional sequence signature). In order to increase the number of candidate LTR retrotransposons, the deletion breakpoint coordinates were extended by 400 bp at both extremities. Only LTRs defined at less than 500 bp from the breakpoint coordinates were retained for further analysis. LTR retrotransposons not involved in structural variation were discovered genome-wide with LTR_FINDER as described above, discarding all positions involved in SV in our grapevine population. LTR sequences of each retrotransposon were recovered from the reference sequence and aligned with the *stretcher* command of the EMBOSS suite (Rice P et al., 2000; Olson SA, 2002). The evolutionary distance (K) between two LTRs was calculated for each pairwise comparison with the *distmat* tool of EMBOSS, with the *nucmethod* option set to 2, in order to compute the distance measure with the Kimura's Two-Parameter method (Kimura M, 1980). Finally, the time of insertion (T) was estimated for each retrotransposon with the substitution rate (k) of $1.3E^{-08}$ (Ma J & Bennetzen JL, 2004), via the following equation: T=K/2*k, where K is the evolutionary distance. The substitution rate (k) is two time higher than the synonymous substitution rate observed by Gaut BS and colleagues for the *adh1* and *adh2* loci of grasses (Gaut BS et al., 1996).

3.9 Detection of insertions

For the discovery of the insertions in *Vitis vinifera*, a pipeline previously developed by our group was used (Sara Pinosio, unpublished results). Insertions of transposable elements were observed by means of the peculiar mapping pattern of the paired-end reads spanning an insertions site: while one read of the pair originates from the flanking region, its mate originates from either the 3' or 5' end of the inserted element. The former reads are expected to map to the reference genome with opposite orientation and creating thus a "wall" of

reads pointing toward the insertion site, while the SV internal reads should map to multiple positions of the genome (upper part of Figure 10). The insertion discovery tool used in the present work relied on a database of transposable elements, that included the deletion sequences identified previously and the above described database of transposable elements used for the annotation step.

The pipeline for the identification of TE insertions was structured in four steps (Figure 10):

1. The reads pointing toward the insertions site (creating thus a wall of reads), whose mate mapped to another genomic position, were *de novo* assembled using CAP3 (Huang X & Madan A, 1999), creating thus two consensus sequences of the insertion flanking regions. Cap3 was run with an overlap length cutoff *o* = 16, a clipping range *y* = 6, a base quality cutoff *c* = 6, an overlap similarity score cutoff *s* = 251, a max overhang percent length *h* = 100, a match score factor *m* = 40, a segment pair score cutoff *i* = 21 and a chain score cutoff *j* = 31.
2. The consensus sequences were then aligned to the reference genome of *Vitis vinifera* using blastn (Altschul SF et al., 1990) with default parameters. If the reconstructed consensus sequences mapped with opposite orientation and at distance lower than the mean sequenced library insert size, a putative insertion site was identified.
3. In order to characterise the transposable element in the putative insertion site, the mates of the anchored reads were used to reconstruct *de novo* the extremities of the TE with CAP3.
4. In order to validate the TE insertion, the consensus sequences were aligned to the database of transposable elements using

blastn: an insertion was called if both *de novo* reconstructed extremities mapped to both extremities of the same TE.

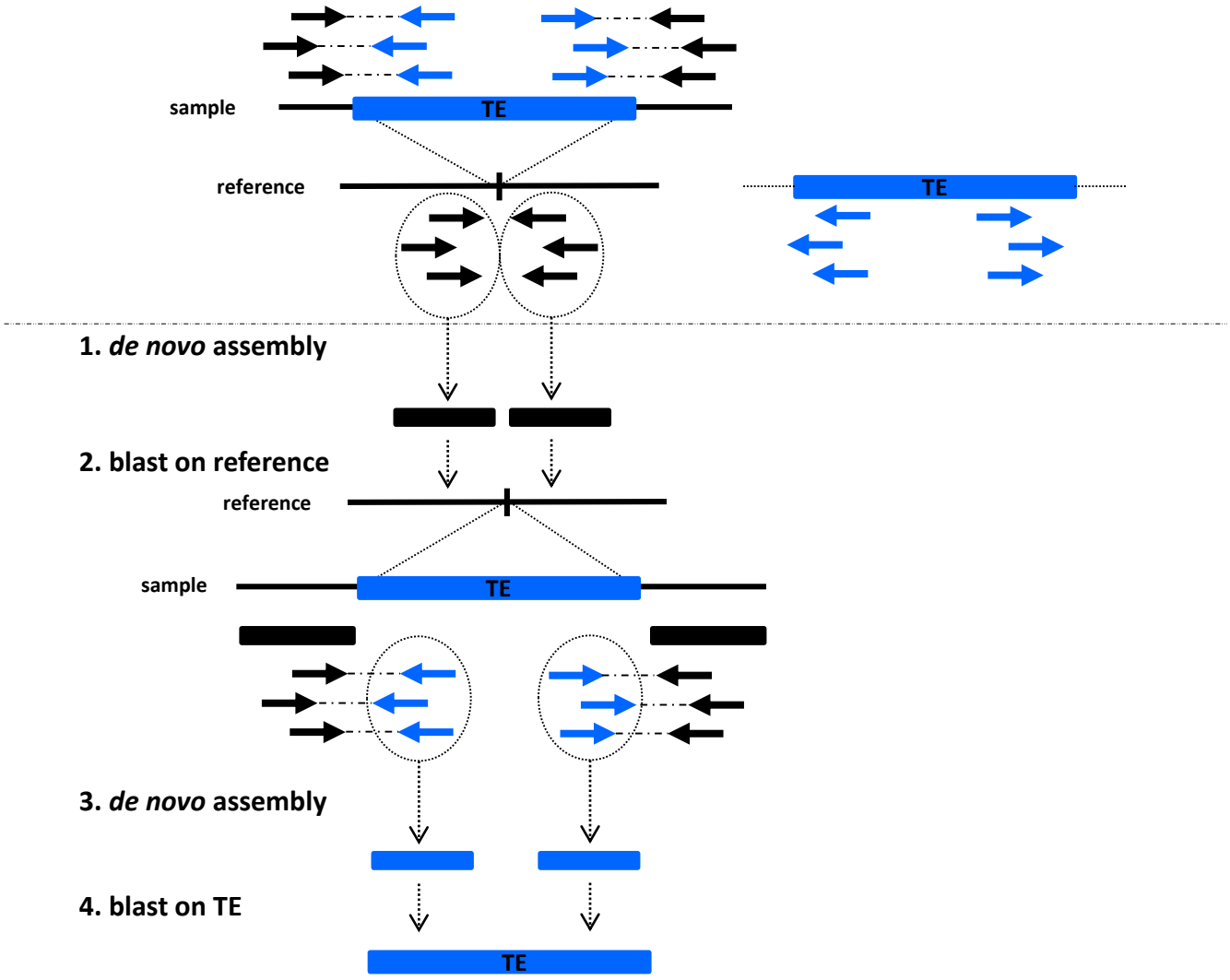


Figure 10: Insertion discovery pipeline. In the upper panel the mapping pattern of reads flanking an insertion in the sample with respect to the reference genome is reported. In the lower panel the pipeline workflow is outlined.

3.10 Identification of insertions

The pipeline previously described was employed for the identification of the insertions in the grapevine population. The insertions were searched separately in each variety and then merged across samples within an interval of 250 bp

around the breakpoint site: insertions identified in at least two samples with coordinates overlapping within the interval were combined as single event. Insertions were filtered in each sample based on coverage information. The mean coverage of the unique aligned reads was computed separately in an interval of 500 bp spanning left and right the SV breakpoint. The mean coverage of the two regions flanking the insertion breakpoint was then calculated. Only positions of a variety with a total mean coverage ranging between 0.5 and 2.5 times the modal coverage value of the variety were considered informative.

As described previously for the deletions, we employed a Python script to estimate the genotypic status of the insertions. The analysis was computed on regions of 500 bp flanking the left and right SV coordinates. For each insertion, the software retrieved from the alignment file of each variety the total number of reads supporting the insertions - **positive reads** - (reads pairs for which one read was mapped within 500 bp spanning the insertion breakpoint, while the mate was aligned to another genomic position and mapped within 500 bp at the 5’ or 3’ end of the inserted transposable element) and the total number of not supporting reads - **negative reads** - (reads supporting the reference genotype, mapping with an insert size similar to the expected one and spanning the SV breakpoint interval). The frequency of the variant was obtained as the ratio of the number of positive reads (those supporting the SV) divided by the total number of reads (sum of positive and negative reads). Based on the frequency, a genotype was assigned to the insertions of each variety: the genotype was assigned only to the samples with at least five reads (both positive and negative) in the regions flanking the SV. Positions were classified as homozygous for reference, heterozygous or homozygous for alternate allele if the ratio was lower than 0.25, between 0.25 and 0.75, or over 0.75, respectively.

As previously explained, the individual used for assembling the reference genome of grapevine was obtained through successive cycles of selfing of a

Pinot Noir seed parent. During the process of self-fertilizations the seed parent was accidentally pollinated in an early generation by Helfensteiner (a variety obtained in Germany from a cross between Pinot Noir and Schiava Grossa). Thus, only part of the reference genome corresponds to one of the haplotypes of Pinot Noir. Based on the SNPs information we classified the regions of the reference genome as either donated by Pinot Noir or donated by Schiava Grossa. In the regions where Pinot Noir carried only heterozygous SNPs, one haplotype was shared between Pinot Noir and PN40024.

In order to validate the call of the genotype for the insertions, we compared SNP genotypes with the genotypes of the insertions in Pinot Noir, within the regions where Pinot Noir shared at least one haplotype with the reference genome. SVs in these regions are expected to be heterozygous. We observed that several insertions were incorrectly classified as homozygous (approximately 30%), due to an excess of positive reads compared to negative reads. Therefore, we applied a further quality control by integrating the negative reads count with the coverage information of the regions flanking the SV breakpoints. We calculated a parameter by dividing the count of negative reads by the mean coverage of the flanking regions. The threshold for this ratio was defined based on the PCR-validation results. Insertions with a ratio higher than 0.3 were classified as heterozygous.

3.11 Gene annotation analysis

In order to characterise the genes interrupted by the SVs, the primary transcripts of the genes of the V2.1 annotation (Vitulo N et al., 2014) were functionally annotated, using the sequences as query for a blastx analysis against the Viridiplantae (taxid: 33090) non redundant protein (nr) database. The *blastx* results were then imported into the Blast2GO 3.1 interface (Conesa A et al., 2005). By retrieving the GO terms associated with each *blastx* hit,

Blast2GO reported a GO annotation for the input gene sequences through the annotation and mapping steps. A summary of the annotation was obtained mapping the results to the Plant GO-Slim, a reduced version of the Gene Ontology with only selected relevant nodes for plants. Over-representation of gene categories affected by SVs, compared to the rest of the genome, was statistically tested using the *Fisher's exact test* integrated in Blast2GO. A false discovery rate correction for multiple testing (Benjamini Y & Hochberg Y, 1995) was used and only gene categories with a corrected p-value<0.05 were considered.

3.12 Validation of structural variants

The identified deletions and insertions were validated experimentally through a PCR-based assay. For each variant, four different primers were designed and combined in three different pairs (Figure 11).

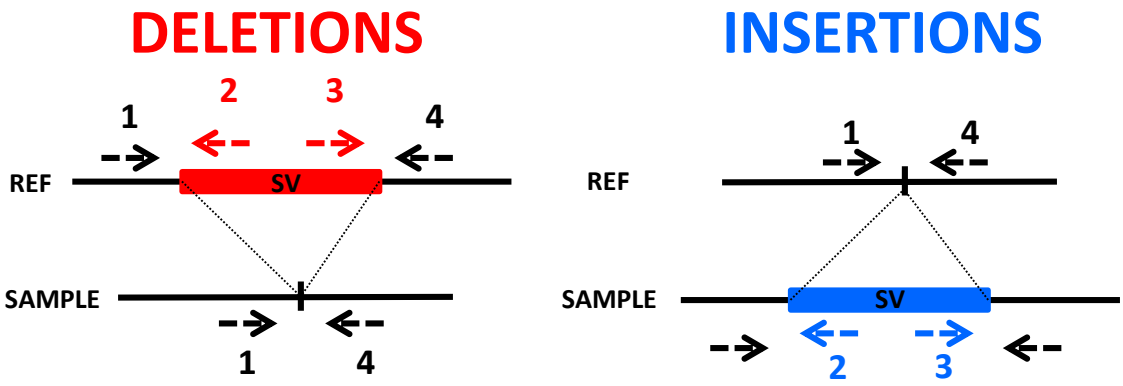


Figure 11: Primer design for validation of structural variants.

Both in the insertions as in the deletions, the primer pairs 1-2 and 3-4 amplified the left and right junction between the reference genome sequence and the SV sequence, respectively. The primer pair 1-4 connected the reference genome sequences flanking the SV. While deletions were validated if the external pair

1-4 amplified, insertions were confirmed by the amplification of the primer pairs 1-2 and 3-4 (Figure 11, Table 1). Deletions were validated only if the sizes of the PCR products were in agreement with the expected sizes, since in smaller deletions (of approximately 1/1.5 Kb), the primer pair 1-4 may amplify the entire sequence also in absence of deletion.

Table 1: Primer pairs required for SV validation. A deletion was validated if primer pair 1-4 amplified, while an insertion was validated if primer pairs 1-2 and 3-4 amplified.

	Primer pairs		
	1+2	**3+4**	**1+4**
DEL	-	-	+
INS	+	+	-

Primers were designed for a random set of the identified SVs, using BatchPrimer3p (You FM et al., 2008), and then manually inspected for further selection. A randomly chosen set of 50 deletions and 50 insertions underwent validation in four different grapevine varieties. Each variety could be either positive or negative for the validated SVs: in total we expected 200 validation events (positives and negatives) for each structural variant category. Events were discarded if no primer pair amplified or a clear genotype could not be determined. The PCR reaction occurred in a 2720 Thermal Cycler (Applied Biosystems, Foster City, CA) under the following conditions: 1 minute at 95°C; 16 cycles of 10 seconds at 95°C, 10 seconds at 67°C (decreasing the temperature by 0.5°C at each cycle) and 10 seconds at 72°C; 25 cycles at 95°C for 10 seconds, at 59°C for 15 seconds and at 72°C for 10 seconds, followed by a final extension at 72°C for 7 minutes.

PCR products were loaded onto a 1% (w/v) agarose gel (160 ml TAE 1X, 1,6 g AGAROSE and 8 µl EUROSAFE) in 1x TAE buffer. In order to verify the identity of the PCR fragments, a 1 Kb DNA ladder and a 100 bp DNA ladder (Biolabs, Ipswich, MA) were run on every gel. The samples separated at 120 V for at least

40 minutes. Finally, images were acquired in a gel documentation station using UV light.

3.13 Validation of SVs on *de novo* assembly

The *de novo* assemblies of six grapevine varieties were used to further validate *in silico* the insertions and deletions identified in the corresponding varieties. The assemblies were obtained with ALLPATHS-LG (Gnerre S et al., 2011) for Cabernet Franc, Heunisch Weiss, Kishmish Vatkana, Rkatsiteli, Sangiovese and Savagnin Blanc. To validate deletions and insertions, a region of 500 bp upstream and downstream the SV breakpoints from the *Vitis vinifera* reference genome was aligned to the *de novo* assembly sequence using blastn with the option *E-value* 10^{-20}. The SVs with flanking sequence regions mapping to the same contig and with the expected orientation were selected for the validation. Deletions were validated if the flanking regions aligned on the same contig with a distance lower than 500 bp; insertions were validated if the upstream and downstream regions mapped to a contig with a distance of at least 500 bp greater than the original distance.

3.14 RNA sequencing experiments

The transcriptome of five varieties (Cabernet Franc, Kishmish Vatkana, Rkatsiteli, Sangiovese and Savagnin Blanc) was sequenced. For each variety two biological replicates for three different tissues (leaves, berries and tendrils) were sampled. Each biological replicate was handled separately during all the RNA sequencing (RNA-Seq) experiment steps. 100 mg of collected tissues were ground in liquid nitrogen and stored at -80°C. RNA was extracted with the *Spectrum plant total RNA* kit (SIGMA, St. Louis, MO) and sequencing libraries were prepared following the Low Sample (LS) Illumina protocol, with the *TruSeq*

Stranded mRNA Library Prep kit. Libraries were sequenced with the HiSeq2500 and 100 bp paired-end reads were obtained. Adapters of raw reads were removed, and reads were further filtered for quality and contaminants. Filtered reads were aligned to the reference genome using TopHat2 version 2.0.6 (Kim D et al., 2013) and the V2.1 gene annotation file (Vitulo N et al., 2014). TopHat takes advantage of the short read aligner Bowtie (Langmead B et al., 2009) for a better alignment resolution of short read segments. To estimate the expression levels in all varieties and tissues, Cufflinks version 2.2.0 was used (Trapnell C et al., 2010, 2012). Expression levels were reported as Fragments Per Kilobase of transcript per Million mapped reads (FPKM), to adjust, in each experiment, for length of transcript and for total number of reads aligned to the transcriptome.

4 RESULTS

4.1 Sequencing

The grapevine population explored in the present work was composed of 128 varieties, including PN40024, the genotype used to produce the grape reference genome. Sequencing metrics are reported in Table 2.

Table 2: Sequencing results of the *V. vinifera* population.

Variety [1]	# [2]	Origin [3]	Group [4]	Coverage [5]	% Ref covered [6]	Physical coverage [7]	Read length [8]	Insert size [9]
Aciaruli Tetri	1	Georgia	*Pontica Georgica*	26.26	86.38	37.15	123.56	405
Agadai	2	Dagestan	admixed	24.90	86.76	33.35	123.49	381
Aglianico	3	Italy	*Italica Tirrenica*	44.81	84.46	91.09	97.75	471
Airen	4	Spain	admixed	26.60	86.95	36.22	123.56	387
Alexandroouli	5	Georgia	*Pontica Georgica Caspica*	27.18	86.73	37.24	123.54	390
Ansonica	6	Italy	*Pontica Insularis*	32.49	80.45	47.98	96.07	353
Ararati	7	Armenia	*Orientalis Caspica trans-Caucasica*	25.48	86.49	38.13	123.43	427
Assyrtiko	8	Greece	admixed	32.45	86.39	46.03	123.57	406
Asyl Kara	9	Dagestan	admixed	20.19	87.18	30.06	126.29	431
Autumn Royal	10	breeding	*Orientalis Antasiatica medi-Asiatica*	5.62	64.94	6.51	75.19	268
Barbera	11	Italy	admixed	57.02	85.98	122.82	98.86	495
Bayan Shirei	12	Azerbaijan	admixed	19.09	86.33	25.85	124.03	389
Berzamino	13	Italy	admixed	14.55	82.89	28.46	109.44	517
Bovale	14	Italy	admixed	13.98	82.97	24.64	97.28	413
Cabernet Franc	15	France	*Occidentalis Gallica*	32.92	85.74	39.89	94.29	266
Cabernet Sauvignon	16	France	*Occidentalis Gallica*	27.97	83.38	47.44	97.79	398
Carignan	17	France	admixed	8.19	79.89	16.16	97.31	481
Catarratto B.C.	18	Italy	admixed	32.73	82.31	52.79	95.63	375
Cesanese d'Affile	19	Italy	*Italica Tirrenica*	20.24	83.78	44.68	107.72	568

Variety [1]	# [2]	Origin [3]	Group [4]	Coverage [5]	% Ref covered [6]	Physical coverage [7]	Read length [8]	Insert size [9]
Chaouch blanc	20	Turkey	*Pontica Meridionalis Balcanica*	37.46	88.05	51.83	123.90	389
Charistvala Kolchuri	21	breeding	-	18.57	84.81	26.54	123.22	415
Chasselas Blanc	22	-	*Occidentalis Teutonica*	31.60	88.40	43.95	123.86	390
Clairette Blanche	23	France	admixed	20.87	87.51	27.28	124.08	371
Coarna Alba	24	Turkey/ Moldova	*Pontica Meridionalis Balcanica*	26.39	87.46	38.39	124.03	413
Corvina Veronese	25	Italy	admixed	36.52	84.90	80.42	97.82	507
Daphnia	26	Greece	admixed	21.55	87.52	28.68	124.01	377
Disecka	27	Slovenia	admixed	13.60	80.52	24.00	95.81	420
Enantio	28	Italy	*Occidentalis Raetica*	13.05	80.47	25.22	98.19	472
Falanghina	29	Italy	admixed	29.77	80.78	44.70	95.57	355
Fiano	30	Italy	*Italica Tirrenica*	25.75	84.72	61.10	96.96	543
Fumat	31	Italy	admixed	13.86	83.54	23.66	108.28	442
Garganega	32	Italy	*Pontica Adriatica*	24.96	82.66	55.64	96.84	522
Garnacha	33	Spain	*Pontica Insularis*	24.81	82.37	34.31	78.84	265
Glera	34	Italy	admixed	27.26	85.32	50.09	96.83	417
Gorula	35	Georgia	*Pontica Georgica Caspica*	29.58	87.27	41.60	123.86	399
Grechetto Bianco	36	Italy	admixed	30.12	81.01	61.86	96.11	487
Greco di Tufo	37	Italy	*Italica Tirrenica*	35.42	80.01	42.56	96.66	290
Grignolino	38	Italy	admixed	13.39	84.02	31.00	120.94	666
Gyulyabi Dagestanskii	39	Dagestan	*Orientalis Caspica trans-Caucasica*	18.35	86.50	25.57	123.43	398
Harslevelue	40	Hungary	admixed	22.16	87.62	32.40	126.30	421
Henab Turki	41	Turkey	*Orientalis Antasiatica medi-Asiatica*	28.03	86.73	42.30	123.95	431
Heunisch Weiss	42	-	*Pontica Balcanica*	57.26	86.36	78.48	94.04	299
Italia	43	breeding	admixed	9.23	73.04	12.51	74.64	277
Kadarka	44	Hungary	*Pontica Meridionalis Balcanica*	25.87	86.87	38.30	123.44	421
Katta Kurgan	45	Uzbekistan	admixed	22.06	86.62	27.89	123.53	361
Khop Khalat	46	Dagestan	admixed	21.86	86.31	28.88	123.52	378
Kishmish Vatkana	47	Uzbekistan	*Orientalis Antasiatica medi-Asiatica*	30.73	83.65	40.36	97.20	305

Variety [1]	# [2]	Origin [3]	Group [4]	Coverage [5]	% Ref covered [6]	Physical coverage [7]	Read length [8]	Insert size [9]
Lambrusco di Sorbara	48	Italy	*Occidentalis Raetica*	15.09	81.45	31.65	101.83	524
Lambrusco Grasparossa	49	Italy	*Occidentalis Raetica*	24.39	84.01	41.65	98.40	400
Limnio	50	Greece	*Pontica Meridionalis Balcanica*	21.08	87.40	28.83	123.51	387
Malvasia Bianca	51	-	admixed	31.76	78.36	55.18	96.05	426
Malvasia Bianca Lunga	52	-	admixed	31.29	82.16	56.40	95.80	420
Malvasia di Sardegna	53	-	admixed	30.66	82.17	46.84	95.22	354
Malvasia Istriana	54	Croatia	*Pontica Adriatica*	14.19	83.59	24.43	108.47	447
Marandi Shemakhinskii	55	Azerbaijan	admixed	20.56	85.55	29.36	125.60	419
Mauzac Blanc	56	France	admixed	29.66	88.37	40.98	123.48	386
Mavrodaphni	57	Greece	admixed	17.79	85.96	24.13	123.48	390
Merlot Noir	58	France	*Occidentalis Gallica*	28.34	87.38	75.51	98.63	602
Montepulciano	59	Italy	*Pontica Adriatica*	26.27	83.71	57.91	97.52	514
Moscato di Scanzo	60	breeding	*Occidentalis Teutonica*	13.20	83.50	30.45	120.95	668
Mtsvane Kachuri	61	Georgia	*Pontica Georgica Caspica*	19.93	86.95	29.89	116.09	401
Muscat a Petits Grains B.	62	-	admixed	29.80	84.74	54.40	97.22	419
Narma	63	Dagestan	*Orientalis Caspica trans-Caucasica*	21.66	86.96	32.82	124.11	433
Nasco	64	Italy	*Pontica Insularis*	33.35	81.50	67.55	96.60	480
Nebbiolo	65	Italy	*Occidentalis Raetica*	34.71	87.97	42.99	120.03	338
Negro Amaro	66	Italy	admixed	12.80	82.93	26.38	104.66	520
Nero d'Avola	67	Italy	*Pontica Insularis*	33.08	84.63	66.52	97.69	464
Nieddu Mannu	68	Italy	*Pontica Insularis*	29.44	78.92	31.62	98.71	269
Nosiola	69	Italy	admixed	23.65	84.25	55.27	96.75	537
Ojaleshi	70	Georgia	admixed	18.47	87.26	29.96	112.56	418
Passerina	71	Italy	*Pontica Adriatica*	36.56	79.60	71.30	96.75	474
Pecorino	72	Italy	*Italica Tirrenica*	34.68	85.18	70.03	97.19	461
Petit Rouge	73	Italy	admixed	11.63	81.24	23.25	102.46	504
Picolit	74	Italy	admixed	28.13	86.33	55.39	97.34	444
Pignoletto	75	Italy	*Pontica Adriatica*	37.61	77.67	66.75	96.38	441
Pinela	76	Slovenia	*Pontica Balcanica*	15.05	83.04	29.93	97.16	465
Pinot	77	France	*Occidentalis Teutonica*	71.43	89.26	108.42	97.39	331
Plechistik	78	Russian Federation	admixed	17.59	86.78	26.54	116.48	405

Variety [1]	# [2]	Origin [3]	Group [4]	Coverage [5]	% Ref covered [6]	Physical coverage [7]	Read length [8]	Insert size [9]
PN40024	79	breeding	-	45.24	92.85	108.07	96.44	496
Raboso Piave	80	Italy	*Occidentalis Raetica*	10.91	82.64	19.85	99.07	436
Red Globe	81	breeding	admixed	9.40	76.07	12.86	74.59	268
Refosco P.R.	82	Italy	*Occidentalis Raetica*	21.85	87.99	44.64	98.36	457
Ribolla Gialla	83	Italy	*Pontica Balcanica*	30.13	85.34	48.03	97.42	364
Ribolla Gialla (Slovenia)	84	-	-	14.57	84.55	28.64	114.93	534
Riesling Weiss	85	Germany	admixed	37.62	88.47	55.21	124.28	412
Rkatsiteli	86	Georgia	*Pontica Georgica Caspica*	44.33	86.35	61.54	94.57	304
Sagrantino	87	Italy	admixed	20.75	83.77	50.27	105.33	609
Sahibi Safid	88	Afghanistan	*Orientalis Antasiatica trans-Caucasica*	21.82	86.63	33.42	123.87	438
Sangiovese	89	Italy	*Italica Tirrenica*	89.79	87.32	125.74	95.72	307
Sauvignon Blanc	90	France	admixed	8.56	84.71	12.57	98.58	342
Savagnin Blanc	91	-	*Occidentalis Teutonica*	59.49	87.39	93.12	96.26	345
Schiava Gentile	92	Italy	*Occidentalis Raetica*	23.83	87.38	39.42	109.92	416
Schiava Grossa	93	Italy	*Occidentalis Raetica*	13.89	86.78	15.61	94.12	244
Schioppettino	94	Italy	admixed	14.32	84.02	27.32	97.61	443
Sciavtsitska	95	-	*Pontica Georgica Caspica*	19.44	86.79	27.01	126.35	405
Shafei	96	Azerbaijan	admixed	22.11	86.44	36.01	118.98	448
Sirgula	97	Georgia	*Pontica Georgica Caspica*	23.44	84.97	35.15	124.13	438
Sultanina	98	Turkey	*Orientalis Antasiatica medi-Asiatica*	82.81	86.77	124.71	132.77	461
Tagobi	99	Tajikistan	*Orientalis Antasiatica trans-Caucasica*	23.66	86.85	40.17	116.86	457
Taifi Rozovyi	100	Uzbekistan	admixed	21.74	87.60	36.57	116.71	448
Tannat	101	France	admixed	22.12	87.63	26.42	99.74	272
Tavkveri	102	Georgia	*Orientalis Caspica trans-Caucasica*	19.60	86.90	32.17	116.54	440
Terbash	103	Turkmenistan	*Orientalis Antasiatica medi-Asiatica*	27.79	84.91	41.01	123.98	431
Terrano	104	Italy	admixed	36.79	85.38	73.62	96.70	453
Thompson Seedless	105	-	-	10.06	80.91	15.91	74.53	291
Tibouren	106	France	admixed	19.28	84.40	39.66	98.37	480

Variety [1]	# [2]	Origin [3]	Group [4]	Coverage [5]	% Ref covered [6]	Physical coverage [7]	Read length [8]	Insert size [9]
Tocai Friulano	107	Italy	admixed	36.96	83.37	60.24	97.22	380
Trebbiano Toscano	108	Italy	*Pontica Adriatica*	29.91	85.40	50.39	95.04	375
Tschvediansis Tetra	109	-	*Pontica Georgica Caspica*	31.07	87.43	49.85	118.26	434
Uva di Troia	110	Italy	*Pontica Adriatica*	41.59	89.48	68.15	97.11	356
V267	111	-	admixed	25.30	87.09	42.44	116.63	449
V278	112	-	*Pontica Georgica*	20.77	86.15	36.14	116.66	471
V292	113	-	*Orientalis Antasiatica medi-Asiatica*	22.06	86.95	35.35	116.15	428
V294	114	-	*Orientalis Antasiatica trans-Caucasica*	24.91	86.80	41.26	116.46	444
V385	115	-	*Orientalis Antasiatica trans-Caucasica*	22.10	86.83	34.89	124.02	451
V389	116	-	admixed	21.15	87.38	36.20	116.54	456
V395	117	-	*Pontica Georgica*	29.79	85.78	44.88	124.11	436
V400	118	-	-	21.27	87.81	36.62	116.68	458
V410	119	-	*Orientalis Antasiatica trans-Caucasica*	23.08	87.55	39.09	119.25	461
V411	120	-	admixed	20.78	86.06	32.60	121.88	444
Verdicchio Bianco	121	Italy	admixed	26.83	84.54	56.94	97.02	487
Verduzzo Friulano	122	Italy	admixed	12.96	79.97	24.08	111.75	519
Vermentino	123	Italy	admixed	9.65	80.11	15.20	91.82	361
Vernaccia S.G.	124	Italy	admixed	37.85	79.14	63.98	96.05	410
Welschriesling	125	-	admixed	35.31	83.02	78.06	96.94	516
Zametovka	126	-	admixed	12.21	82.73	27.78	101.58	559
Zelen	127	Slovenia	admixed	23.22	86.05	46.83	119.83	562
Zinfandel	128	Croatia	admixed	10.36	81.21	17.68	98.08	412

[1] Prime name (short format) of the *V. vinifera* variety. [2] Variety reference number used in plots. [3] Country of origin, if available. [4] Subgroups membership defined with ADMIXTURE K=13. [5] Mean sequence coverage of the uniquely aligned reads. [6] % *V. vinifera* reference genome covered by unique aligned reads. [7] Mean physical coverage of the unique mapped reads. [8] Mean read length of quality trimmed reads. [9] Mean library insert size (bp).

The coverage of the uniquely aligned reads varied among the 128 varieties from 89.8X (Sangiovese) to 5.6X (Autumn Royal), with a mean value of approximately 26X and a standard deviation of 13. Uva di Troia was the cultivar covering the

highest proportion of *V. vinifera* reference genome (89.5%) while Autumn Royal was the one covering the lowest proportion (64.9%). In line with the coverage information, Sangiovese was the variety with the highest physical coverage (125.7X) and Autumn Royal the one with the lowest (6.5X). The read length of the varieties varied between 74.5 bp (Thompson Seedless) and 132.8 bp (Sultanina), with a mean length of 107.5 bp and a standard deviation of 13.7 bp. The insert size had a mean value of 424.3 bp (standard deviation of 79.3 bp), ranging between 244 bp in Schiava Grossa and 668 base pairs in Moscato di Scanzo.

4.2 Single Nucleotide Polymorphism (SNP) analysis

Single nucleotide polymorphisms (SNPs) were called with GATK (Van der Auwera GA et al., 2002; DePristo MA et al., 2011) in 128 grapevine cultivars. Raw SNPs were filtered with a custom developed Perl script (see methods, paragraph 3.3) and 18,296,434 SNPs were retained. Furthermore, SNPs in repeated regions were removed and a total of 9,476,368 SNPs were retained for further analyses.

In order to compute and plot genetic parameters on a whole-genome scale, polymorphic sites were analysed in windows of variable size, containing always 100 Kb of positions not masked by repeat sequence annotators. In each window and in each variety, not all non-repetitive 100 Kb are informative, since a fraction of bases has a coverage below 0.5 or above 2.5 times the modal coverage value. Thus, the number of informative positions is variable in each window. A total of 2,367 windows were obtained over the 19 grapevine chromosomes.

SNPs identified in grapevine were validated with three different approaches.

1. As previously explained, part of the reference genome sequence corresponds to one haplotype derived from Pinot Noir. We selected all genomic regions where Pinot Noir and the reference genome share one haplotype. Only stretches of at least three consecutive windows, discarding the first and the last window (where spurious signals may arise), were selected for validation. In these regions only heterozygous SNPs are expected, while no homozygous SNPs should be called. A total of 175.1 Mb were used to validate the SNPs in Pinot Noir. We recovered 545,646 SNPs: 545,519 SNPs (99.98%) were classified as heterozygous, while 127 as homozygous (0.02%). Out of 82.52 Mb nucleotides used for SNP calling, one homozygous SNP was wrongly detected every 649.74 Kb. We also tested regions in which Pinot Noir shares both haplotypes with the reference sequence and no SNPs are expected. We retained for analysis only stretches of at least three windows, after discarding the first and the last, and we identified a total of 242 SNPs in 2.26 Mb informative bases, corresponding to a false positive SNP every 9.33 Kb.
2. Pinot Noir and Savagnin Blanc are related by a parent-offspring relationship. Thus, across the entire genome the varieties should share always at least one haplotype. Based on the haplotype sharing measure, we estimated that Pinot Noir and Savagnin Blanc share one haplotype across the 19 chromosomes for a total of 286.4 Mb, while for 116.6 Mb the varieties share both haplotypes. We discarded 86 windows with IBS=0, scattered across the genome and amounting to a total of 17.5 Mb. We hypothesize that these windows contain hemizygous DNA, and the parent-offspring duo shared by descent the haplotype carrying the deletion. First, we extracted all the regions where one haplotype is shared between two varieties. In these regions no homozygous SNPs should be identified between the varieties, while heterozygous SNPs are

expected. A total of 1,223,389 SNPs were called in the two varieties with respect to the reference genome. 99.37% of the SNPs were genotyped correctly between the two varieties, while only 0.63% of the SNPs resulted homozygous in one variety compared to the other, producing thus false positive calls. Out of 120.58 Mb nucleotides included in the analysis, homozygous SNPs were wrongly called with a FDR of one false positive SNP every 15.7 Kb. In some specific regions, in addition, both varieties share both haplotypes, as for example on almost the entire chromosome 4 and partially on chromosome 15. We further compared in those regions the SNP calls of the two varieties, in order to estimate the rate of false positive calls: no differences should be observed between the two varieties. 22.67 Mb were extracted from chromosome 4, and the SNP genotypes of Pinot Noir and Savagnin Blanc were compared. A total of 76,897 SNPs were identified, 98.95% of them called with the same genotype in both varieties, while 1.05% were wrongly genotyped in one of the two varieties. A total of 806 SNPs were wrongly called out of 9.49 Mb informative sites, leading to an estimated FDR of one false positive SNP every 11.78 Kb. Two other smaller regions were surveyed on chromosome 15: the beginning of chromosome 15 (from 1 to 11.48 Mb) and a second portion, between 12.21 and 17.96 Mb. SNPs were validated with a precision of 98.4%, and a FDR of 1.6%. Out of 5.47 Mb tested nucleotides, 602 SNPs resulted wrongly genotyped, leading to a FDR of one false positive SNP every 9.09 Kb.

3. In addition, we experimentally validated a set of predicted SNPs in Sangiovese with reads data obtained by our research group with the Single Primer Enrichment Technology (SPET) (NuGEN, San Carlos, CA, USA). Sequencing data were obtained via targeted resequencing of 736 random regions of Sangiovese. After filtering for the probe regions and selecting only regions with at least 100X coverage of SPET reads, a total

of 669 predicted positions were retained. Of the predicted homozygous SNPs (180), 97.78% were confirmed by SPET, while the remaining 2.22% of SNPs were called in heterozygous condition by SPET. Similarly, 94.48% of the predicted heterozygous SNPs (489) were confirmed by SPET, while 2.04% of variant sites were confirmed, although the genotype was called homozygous for the alternate allele by SPET. Only 3.48% of the genome-wide heterozygous SNPs resulted in false positive calls.

To explore the grapevine population structure, and estimate thus the degree to which grape samples can be differentiated into a number K of distinct (ancestral) populations, the software ADMIXTURE (Alexander DH et al., 2009) was used. Different numbers of K populations were tested over a set of 123 varieties. Five of 128 cultivars were discarded from the analysis: the reference genotype, PN40024; Ribolla Gialla (Slovenia) (clone of Ribolla Gialla) and V400 (clone of Rkatsiteli); Charistvala Kolchuri (inter-specific hybrid); and Thompson Seedless (genetically very close to Sultanina).

We first evaluated for different K values the Cross-Validation (CV) error estimate with 10-fold cross-validations (Figure 12).

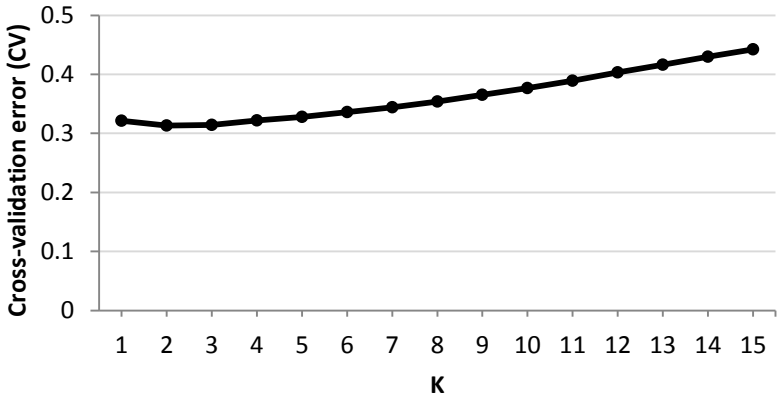


Figure 12: Cross-validation error plot. The mean CV-error value was measured over 20 independent runs, by gradually increasing the K value.

The mean cross-validation error value was calculated over 20 runs of ADMIXTURE, each run performed with a random generated seed number. Good value of K should exhibit low CV error compared to other K values. As depicted in Figure 12, our population showed the lowest value of CV at K=2 and K=3, while the CV values increased with higher K. We further evaluated the true number of K with the four steps method proposed by Evanno G and colleagues (Figure 13).

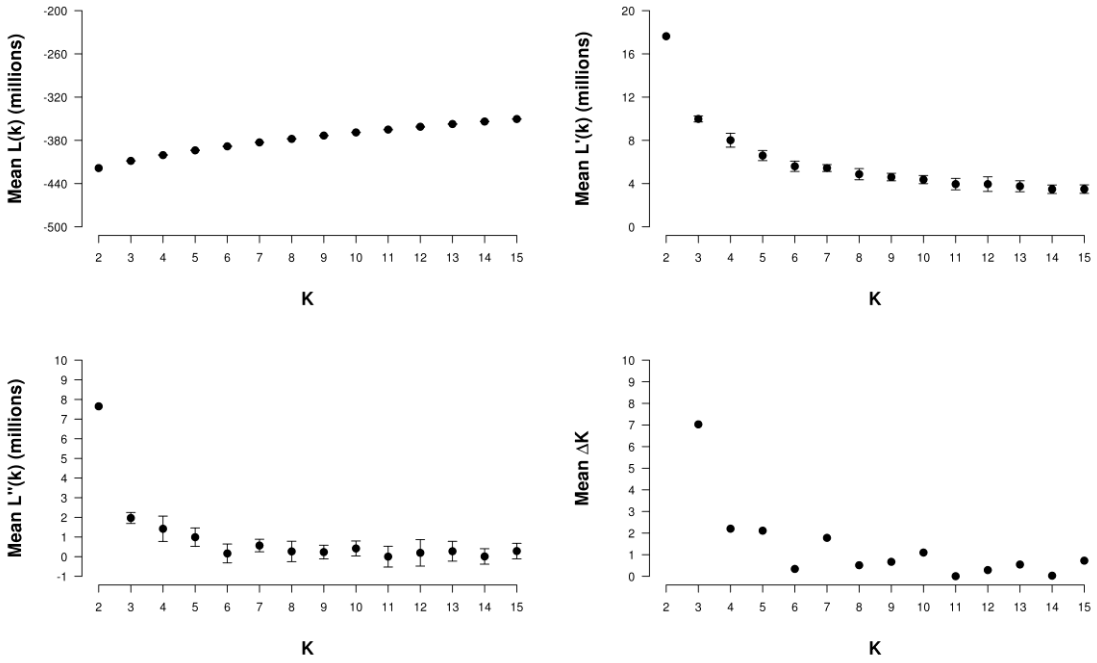


Figure 13: Four step graphical method for the detection of the true number of K (Evanno G et al., 2005).

The authors stated that the modal value of the distribution of ΔK should be located at the real K. As depicted in the last plot of Figure 13, K=3 had the highest mean ΔK value. At K=3 we observed a subdivision of the grapevine population into three main groups, in line with the observations of Negrul AM (1946). Our population was divided in *Proles occidentalis* (green), *Proles orientalis* (red) and *Proles pontica* (blue) (Figure 14).

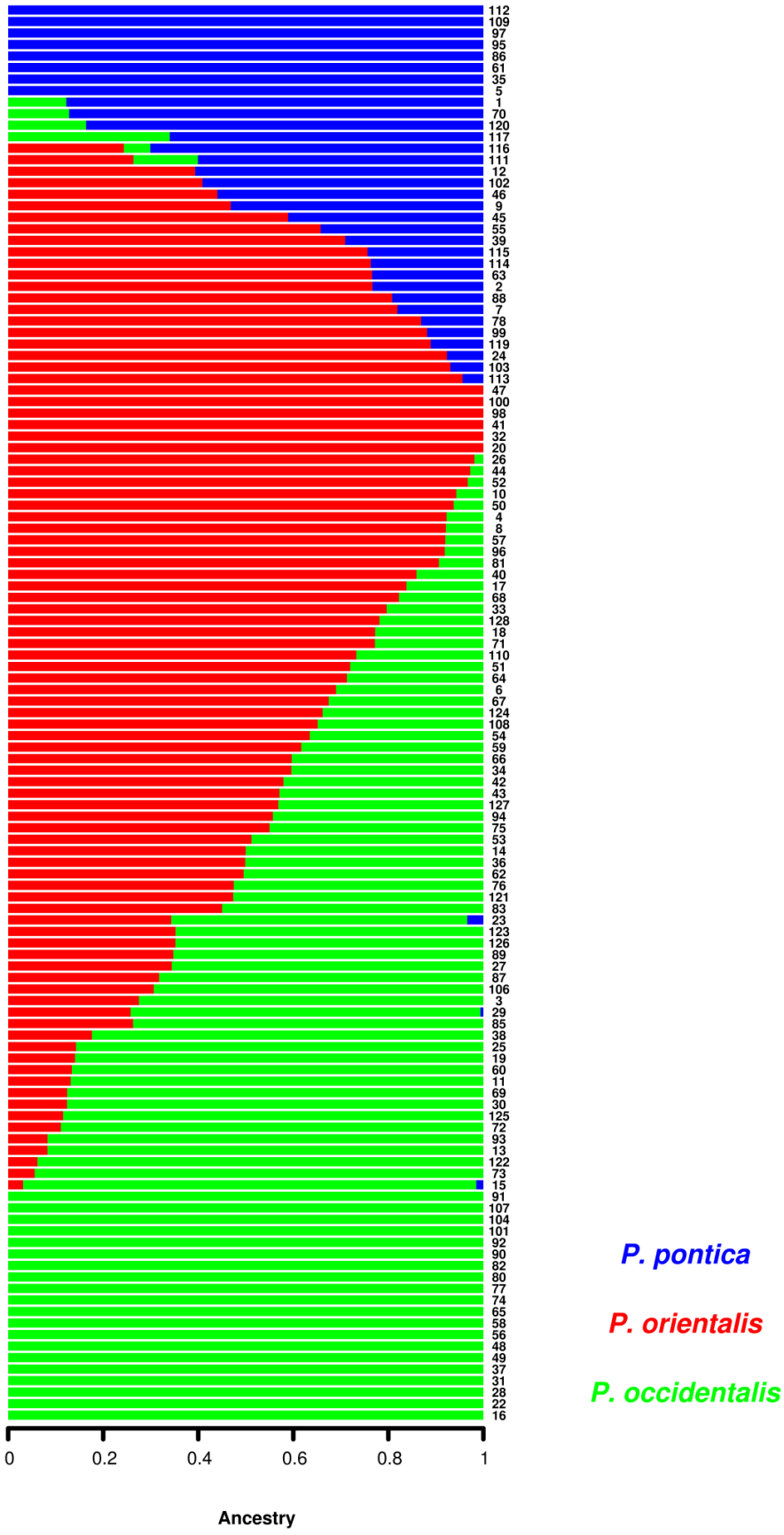


Figure 14: Grapevine population structure with K=3. Numbers refer to the varieties (see Table 2).

Based on the results obtained with ADMIXTURE, our population was composed of a relatively small number of ancient varieties and a majority of recently admixed cultivars, leading to a relatively complex population structure. As depicted in Figure 14 a great proportion of the varieties (73%) showed an admixture between at least two populations. Conversely, approximately 27% of the varieties were ascribed without admixture to a single group: 16% to the *P. occidentalis*, 4.88% to *P. orientalis* and 6.5% to *P. pontica*.

The true number of K describes the primary population structure, but it may miss fine population substructure. Therefore, we gradually increased the number of K, in order to explore the distribution of the varieties within subgroups and investigate evidences of relationships between the individuals and the subgroups. At K=13, nearly 50% of the varieties originated without admixture from distinct groups, while the remaining samples resulted from the mixture of at least two different populations (Figure 15).

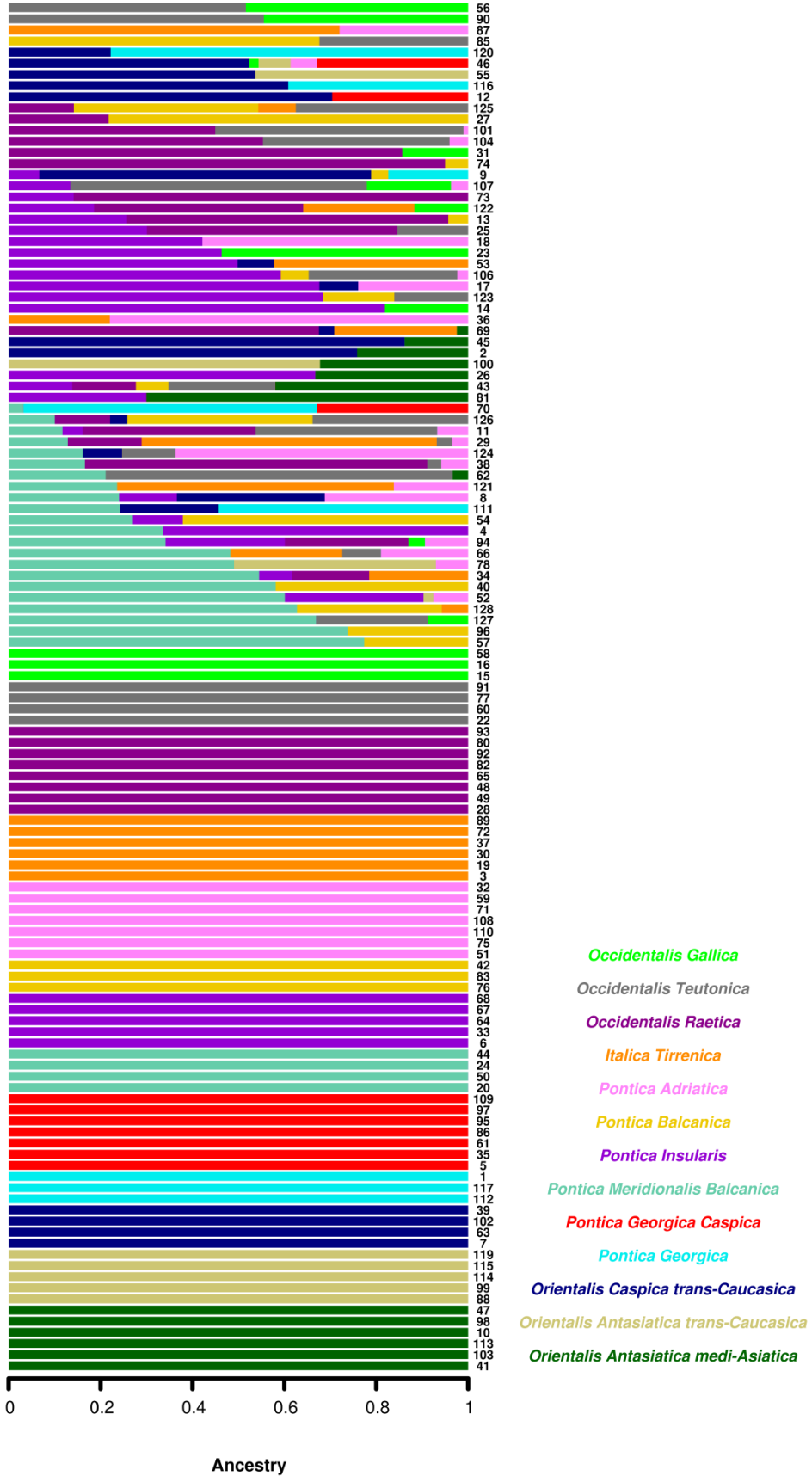


Figure 15: Grapevine population structure with K=13. Numbers refer to the varieties (see Table 2).

In line with the biogeographical groups defined by Troshin LP and colleagues (Troshin LP et al., 1990), we named and described the 13 different groups identified with K=13 as follows:

- *Orientalis Antasiatica medi-Asiatica*: group of *V. vinifera* varieties originating from Central Asia, mostly composed of table cultivars (Autumn Royal, Henab Turki, Kishmish Vatkana, Sultanina, Terbash and V292).
- *Orientalis Antasiatica trans-Caucasica*: group composed of table and double-use varieties of the East and Middle-East, encompassing cultivars originating from Tajikistan, Afghanistan and regions located in the west side of the Caspian Sea (Sahibi Safid, Tagobi, V294, V385 and V410).
- *Orientalis Caspica trans-Caucasica*: group which encompasses table and double attitude cultivars originating from Dagestan, Armenia and other states located between the East side of the Caspian Sea and the West side of the Black Sea (Ararati, Gyulyabi Dagestanskii, Narma and Tavkveri).
- *Pontica Georgica*: group composed of varieties originating from the Georgian Black Sea Basin (Aciaruli Tetri, V278 and V395).
- *Pontica Georgica Caspica*: grapevine varieties originating from the Georgian Black Sea Basin, but in the northern part, stretching from the North of the Black Sea to as far East as the Caspian Sea (Alexandroouli, Gorula, Mtsvane Kachuri, Rkatsiteli, Sciavtsitska, Sirgula and Tschvediansis Tetra).
- *Pontica Balcanica*: group which includes wine cultivars originating from the Northern part of the Balkans (Heunisch Weiss, Pinela and Ribolla Gialla).

- *Pontica Meridionalis Balcanica*: group encompassing wine varieties originating from Turkey, Greece and Southern Balkans (Chaouch Blanc, Coarna Alba, Kadarka and Limnio).
- *Pontica Insularis*: set of varieties of insular origin, for example Sardinia and Sicily (Ansonica, Garnacha, Nasco, Nero d'Avola and Nieddu Mannu).
- *Pontica Adriatica*: group composed of Italian varieties spread over the Adriatic area, showing affinity with *Pontica* individuals (Garganega, Malvasia Istriana, Montepulciano, Passserina, Pignoletto, Trebbiano Toscano and Uva di Troia).
- *Italica Tirrenica*: group of Italian varieties distributed over the Tyrrhenian side (Aglianico, Cesanese d'Affile, Fiano, Greco di Tufo, Pecorino and Sangiovese).
- *Occidentalis Teutonica*: group made up of *V. vinifera* wine cultivars distributed in the Northern Alp region (Chasselas Blanc, Moscato di Scanzo, Pinot and Savagnin Blanc).
- *Occidentalis Raetica*: set of varieties represented by *V. vinifera* wine samples spread in the Southern Alp region (Enantio, Lambrusco di Sorbara, Lambrusco Grasparossa, Nebbiolo, Raboso Piave, Refosco P.R., Schiava Gentile and Schiava Grossa).
- *Occidentalis Gallica*: group composed of grapevines today grown worldwide that originated in the Western part of France (Cabernet Franc, Cabernet Sauvignon and Merlot Noir).

Based on the population partition observed with ADMIXTURE at K=13, we explored the genetic structure of the cultivars via the Principal Coordinates Analysis (PCoA) method (Price AL et al., 2006). We used SNP genotype information and measured inter-individual genotypic distance as follows: for each polymorphic position, the distance between two individuals was set to 1,

0.5 or 0 if they shared zero, one or two alleles, respectively. We then decomposed the principal coordinates of the genotypic distance (Figure 16).

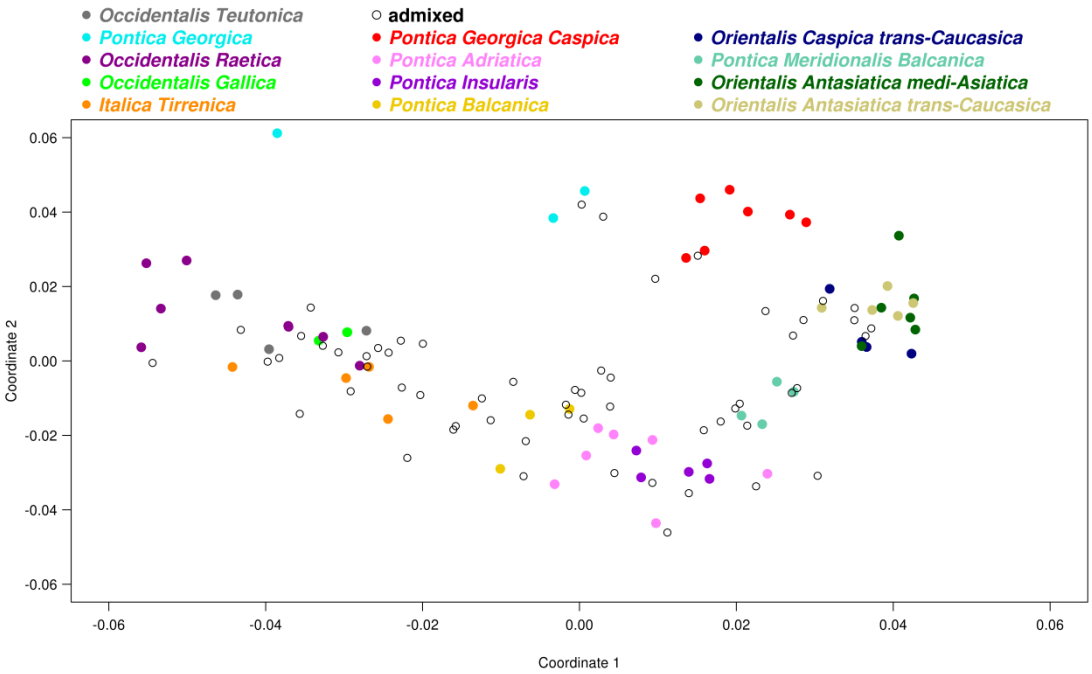


Figure 16: Principal Coordinates Analysis of the SNPs data. Genetic relationships between the 13 grapevine groups. The colours of the groups correspond to the colours of ADMIXTURE. Not coloured dots correspond to admixed varieties.

As depicted in Figure 16 the distribution of the 13 groups over the Cartesian axes approximately reflected the geographical origin of the varieties. The first axis separated the groups geographically from East (right) to West. In the upper part of the plot, the *Pontica Georgica* subgroup resulted the most distant, followed by the *Pontica Georgica Caspica* varieties (upper side of the plot, turquoise and red dots, respectively). Both groups were well separated from the remaining groups, while in the right part of the plot an overlap of the *Orientalis* cultivars was appreciated. The three *Orientalis* subgroups showed a considerable overlap in terms of genotypic distance. Furthermore, the varieties belonging to other groups (*Pontica* and *Occidentalis*) separated clearly over the

first axis. Beside the *Pontica Georgica* subgroups, the *Pontica* groups separated with a gradient ranging between the *Pontica Meridionalis Balcanica* and the *Pontica Balcanica* group. The *Pontica* cluster separated well in turn from the *Occidentalis* varieties, a cluster of groups encompassing cultivars spread from the Tyrrhenian Sea side to North-Europe.

We then explored the haplotype diversity, the nucleotide diversity, the linkage disequilibrium (LD) and the homozygosity of our grapevine population. We selected a subsample of 114 varieties, removing, in addition to the five varieties previously discarded, nine accessions of *V. vinifera* for which accurate identity was unknown. The haplotype diversity of the core dataset of varieties was estimated on blocks of five consecutive markers. The estimate was then averaged over 50 consecutive segments (Figure 17). In order to measure the nucleotide diversity and the linkage disequilibrium, we summarized the genotype SNPs information in windows of 100,000 base-pairs, considering only base-pairs not marked as repetitive sequence. The mean nucleotide diversity and the median linkage disequilibrium r^2 value of each window were plotted over the 19 chromosomes, as depicted in Figure 17. Lastly, based on the SNP genotypes we estimated the degree of homozygosity inside our population. First, based on the window segmentation, we defined for each individual the regions of homozygosity: a window was considered homozygous if the ratio of the heterozygous SNPs over the mean number of base pairs passing quality control for SNP calling was lower than 0.0083 (i.e. 50 heterozygous SNPs in 60,000 positions). We then merged the information of the single varieties together, and measured the total core dataset homozygosity as ratio (Figure 17).

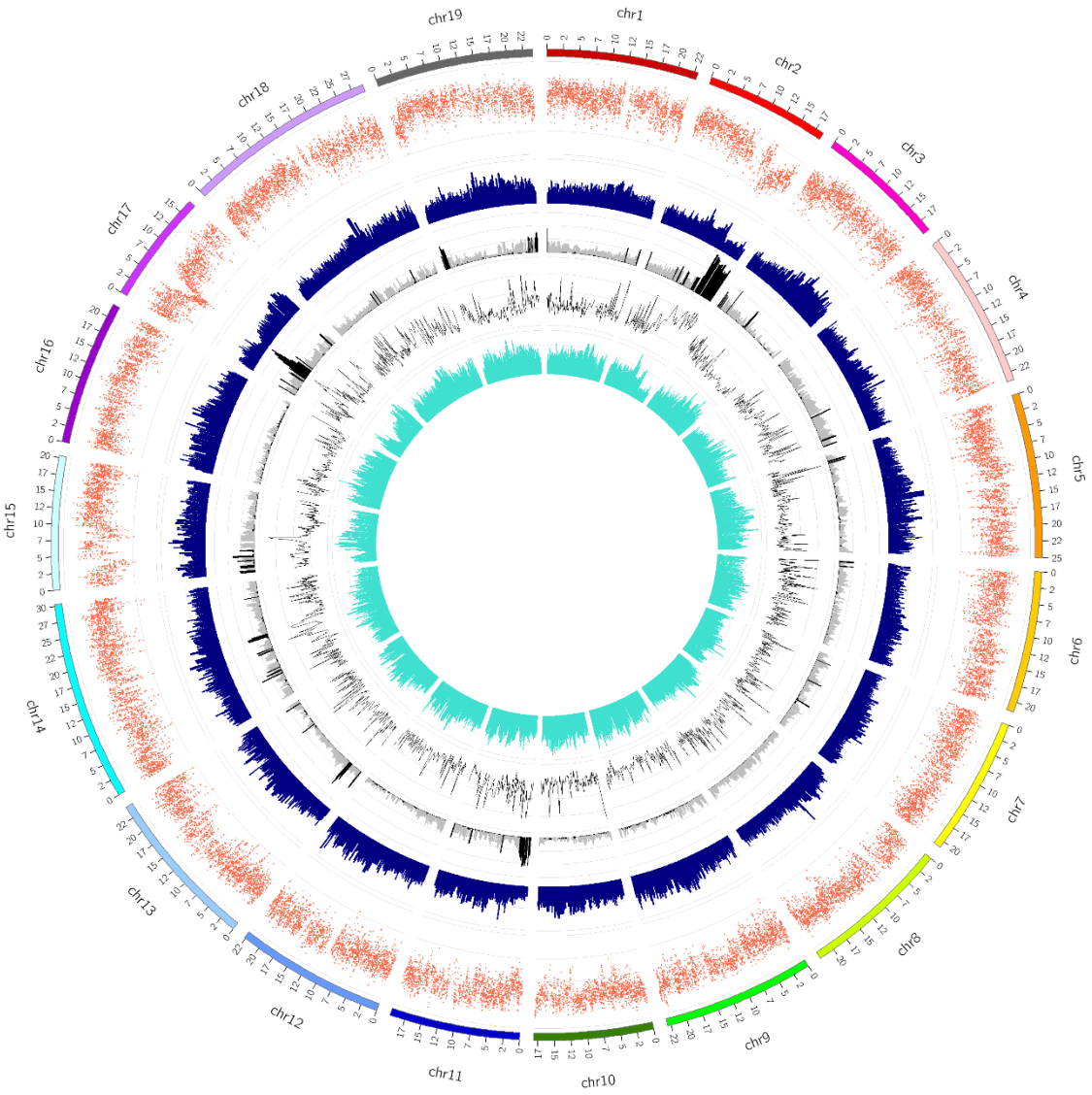


Figure 17: Chromosome distribution of haplotype diversity, nucleotide diversity, homozygosity ratio, linkage disequilibrium and SNP frequency. From outer to inner: haplotype diversity (red, y upper limit 0.75); nucleotide diversity (blue, y upper limit 0.018); homozygosity (grey, y upper limit 0.51; in black were reported the windows with a FDR corrected p-value below 0.025 or above 0.975); linkage disequilibrium (r^2) (black, y upper limit 1); SNP frequency (turquoise, y upper limit 0.020).

In the outer part of the Circos plot (Krzywinski M et al., 2009) the haplotype diversity measure over the 19 *V. vinifera* chromosomes was reported as scatterplot. At first glance, the haplotype diversity values lied around 0.5. Interestingly, some regions with lower haplotype diversity were identified. The

end of chromosome 2 showed a strong decrease of the diversity measure consistently with the observations that white varieties are homozygous at the end of chromosome 2 (Kobayashi S et al., 2004). Furthermore, the region ranging between 5.9 and 6.1 Mb of chromosome 17 showed a strong decrease of the haplotype diversity values, in agreement with the results obtained by Myles S and colleagues in 2011, who identified at this position a selective sweep region. Other smaller regions, such as the beginning of chromosome 11 and chromosome 19, and the region between 3 and 4.5 Mb on chromosome 15 showed a decrease in haplotype diversity. In the same regions described above, a decrease of the nucleotide diversity was observed (blue profile). Moreover, in those regions an increase of the homozygosity ratio was visible (grey profile of the Circos plot). In black we reported windows that violated the null hypothesis of a *Poisson* distribution of homozygous SNPs, after false discovery rate correction. The end of chromosome 2 (from approximately 13.5 Mb to the end of the chromosome) showed a very high homozygosity ratio (approximately 0.5), caused by the white varieties, as previously described. Furthermore, a pronounced level of homozygosity was observed in the putative selective sweep region of chromosome 17. Also the beginning of chromosome 11 and 19 showed a significant increase of the homozygosity ratio values. Furthermore, other smaller regions spread over the 19 chromosomes showed a significant increase or decrease in the homozygosity measure. LD varied with patterns similar to those of haplotype diversity, nucleotide diversity and overall homozygosity. In the previously described regions, where haplotype/nucleotide diversity decreased and overall homozygosity increased, the LD levels increased compared to the overall chromosome distribution. Notably high r^2 values were recorded at the end of chromosome 2 (r^2 around 0.7) and in the selective sweep region of chromosome 17 (r^2 at approximately 1). Finally, we reported the SNP frequency (turquoise) in the innermost part of the circle. At the end of chromosome 2, in the selective sweep region of chromosome 17, as in other

smaller regions, a reduction of the SNPs frequency was observed, in line with the reduction of the haplotype and nucleotide diversity measures and the increase of the homozygosity. Surprisingly, at the beginning of chromosome 11, a different pattern was observed: haplotype and nucleotide diversity decreased, the overall homozygosity was high, while the SNP frequency didn't show a decrease, since several varieties carried the alternative allele in homozygous state.

4.3 SV simulation

To evaluate the performances of various SV discovery tools we simulated 1000 insertions and deletions (see methods, paragraph 3.4) in two reference genomes: *Populus trichocarpa* (Tuskan GA et al., 2006) and *Vitis vinifera* (Jaillon O et al., 2007). Three different sets of reads were used for the identification of the simulated SVs. Two read datasets were represented by the short reads obtained through the next-generation sequencing of the *V. vinifera* and *P. trichocarpa* reference genomes. These real datasets enabled the identification of deletions in presence of possible sequencing bias, but were less informative about the accurate estimation of the false positives. Therefore, in order to precisely estimate the false positives we created a simulated set of reads from the *Populus trichocarpa* genome. Using wgsim (https://github.com/lh3/wgsim), 100 bp long reads were simulated with a mean insert size of 420 bp, a base error rate of 0.01 and a 0 rate of mutations (Figure 18 and Table 3).

The simulated and the real set of reads were aligned to the respective *simulated SV genomes* using BWA (Li H & Durbin R, 2009), with the default parameters. Alignment statistics were reported in Table 3. The poplar real reads showed the smallest mean insert size (224.36 bp). Both *P. trichocarpa* and *V. vinifera* real reads had a very tight insert size peak with a standard deviation of 25.12 and

75.14 bp respectively (Figure 18). The coverage of the alignments ranged between 31.2X for the simulated reads of *P. trichocarpa* and 59.79X for the grapevine real reads dataset.

Table 3: Summary of the read datasets. Alignment statistics of the different read datasets.

Species	Reads type	Read length (bp)	Mean insert size (bp)	Standard deviation (bp)	Mean coverage (X)
P. trichocarpa	**simulated reads**	100	420.13	129.46	31.2
	real reads	75	224.36	25.12	41.2
V. vinifera	**real reads**	100	496.07	75.14	59.79

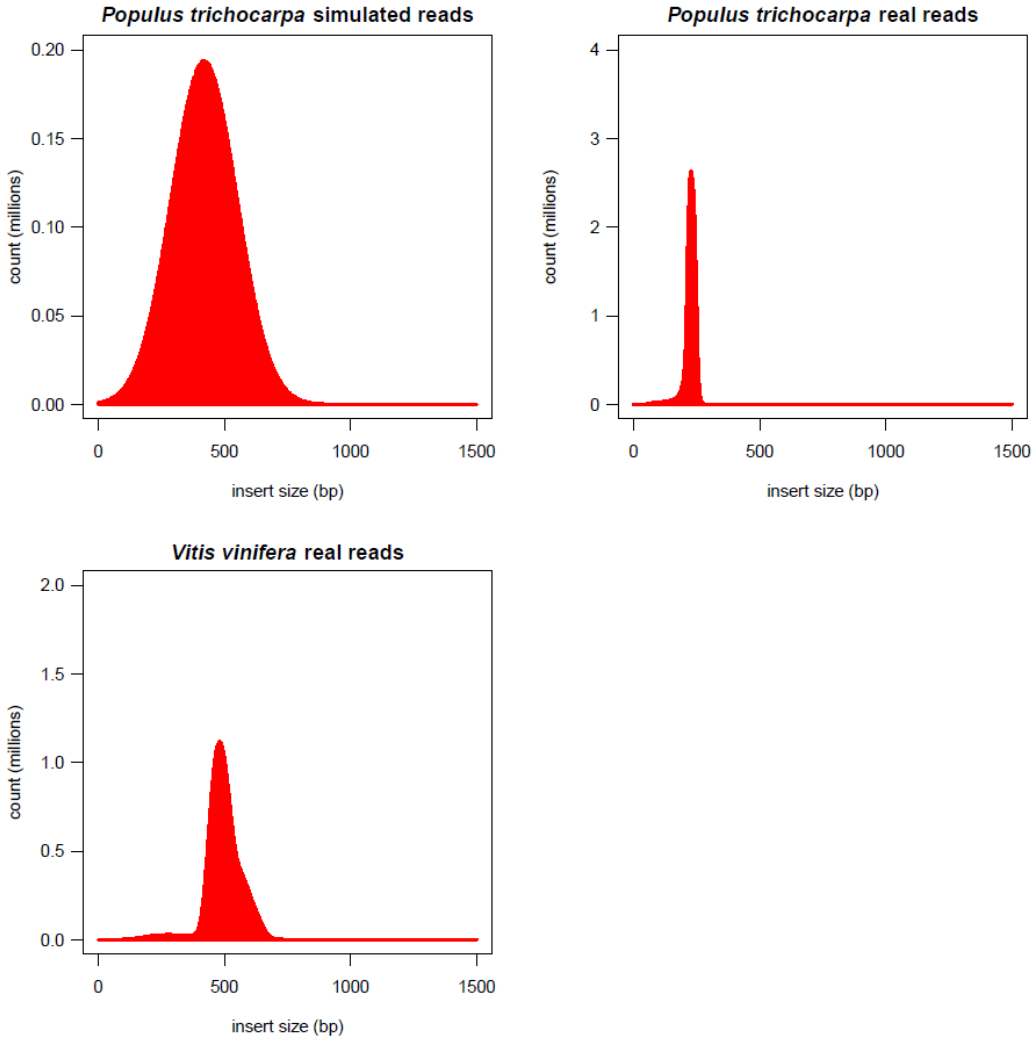


Figure 18: Insert size distribution of the three read datasets. Top left: *Populus trichocarpa* simulated reads; **top right:** *Populus trichocarpa* real reads; **bottom left:** *Vitis vinifera* real reads.

4.4 Simulation of deletions

We investigated the performance of four different tools in the detection of deletions. The deletions obtained with each software were classified as *true positives* if the difference between the breakpoint coordinates and the simulated coordinates (both at the 5' as at the 3' end of the SV) was lower than 250 bp. A deletion not satisfying this condition was classified as *false positive*, while simulated deletions not identified at all by the tools were classified as *false negatives*. The Positive Predictive Value (PPV, or precision) was calculated as the number of *true positives* over all the SVs detected. The *number of exact predictions* was estimated comparing the coordinates and size of the predicted deletions with the simulated SVs breakpoints, while the *mean breakpoint distance* was calculated as the mean distance of the predicted breakpoints from the simulated coordinates (Table 4, Table 5 and Table 6).

A very important threshold for the detection of SV is the number of paired reads supporting it. The simulated set of reads was used first to assay the performances of the four tools (used separately or coupled) when varying the threshold for the detection of SVs. DELLY and all the three pairwise combinations of individual softwares performed better than GASV, CLEVER and Pindel alone, with a slight decay of the total true positives along with the increasing number of PE required (Figure 19).

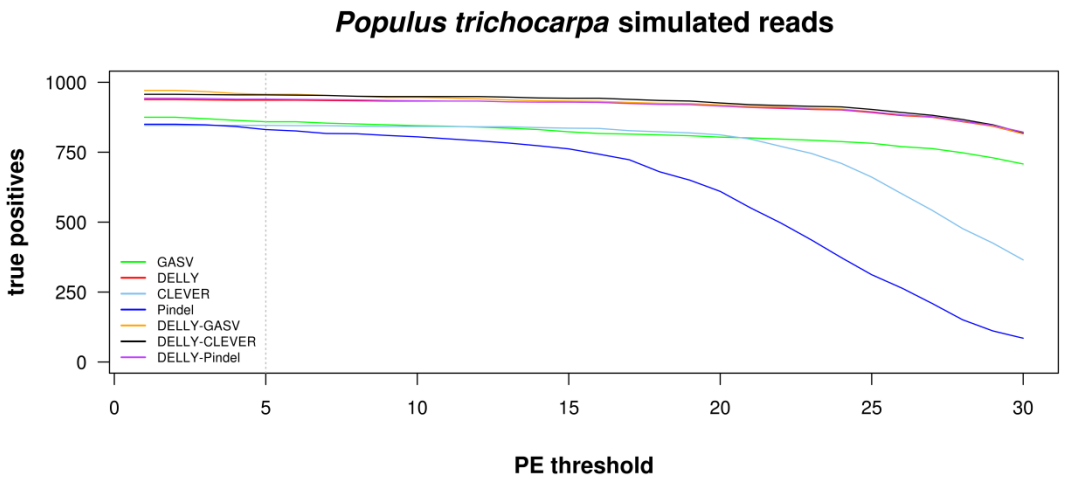


Figure 19: Number of *true positive* deletions identified using simulated reads aligned to the *simulated SV genome* of *Populus trichocarpa*, as a function of the number of supporting paired reads required to call a SV.

Furthermore, by means of the simulated dataset we measured the false positives identified by the four methods, by progressively increasing the PE threshold. All the software tools tested identified very few false positives, with the exception of Pindel (Figure 20).

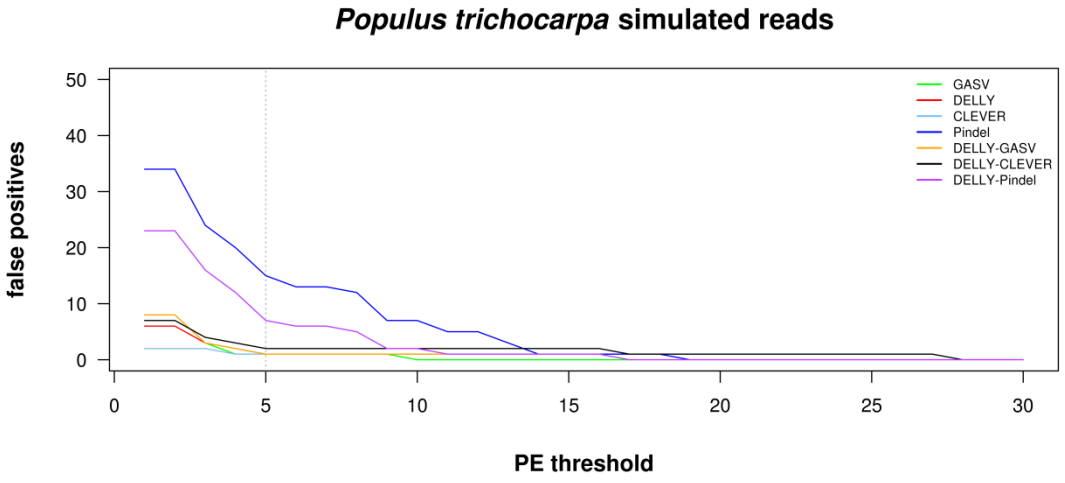


Figure 20: Number of *false positive* deletions identified using simulated reads aligned to the *simulated SV genome* of *Populus trichocarpa*, as a function of the number of supporting paired reads required to call a SV.

Based on the results observed with the simulated reads aligned to the *simulated SV genome* of *P. trichocarpa*, we used a threshold of five paired-end reads for further statistical analyses (Table 4).

Table 4: Deletion statistics with simulated reads. Report obtained through the alignment of simulated reads to the *simulated SV genome* of *P. trichocarpa* and selection of deletions with a support of 5 PE.

	CLEVER	DELLY	GASV	Pindel	DELLY-GASV	DELLY-CLEVER	DELLY-Pindel
Total deletions	847	937	860	846	958	957	947
# true positives	846	936	859	831	957	955	940
# false negatives	154	64	141	169	43	45	60
# false positives	1	1	1	15	1	2	7
Mean breakpoint distance (bp)	12.90	2.82	45.36	1.76	5.21	3.51	3.02
# of exact predictions	67	889	20	829	889	889	893
Sensitivity (%)	84.60	93.60	85.90	83.10	95.70	95.50	94.00
PPV (%)	99.88	99.89	99.88	98.23	99.90	99.79	99.26
FDR (%)	0.12	0.11	0.12	1.77	0.10	0.21	0.74
F1 score (%)	91.61	96.64	92.37	90.03	97.75	97.60	96.56

In terms of true positives, false negatives, false positives and number of exact predictions, DELLY performed better than the other three tools. This software discovered the greatest number of true positives (936), with a *sensitivity* of 93.6%, and the lowest number of false negatives (64). Similarly to CLEVER and GASV, DELLY produced a single false positive, predicting deletions with a Positive Predictive Value (PPV) of 99.89% and a False Discovery Rate (FDR) of 0.11%. Concerning the breakpoint resolution, Pindel resulted the best software with a mean breakpoint size of 1.76 bp. In order to improve the results obtained, we combined the results observed with DELLY (the best performing tool) with any of the other tested packages. Deletions identified by two tools with breakpoint coordinates overlapping within an interval of 250 bp were considered as the same SV event. Comparing the results obtained with the

three possible combinations (DELLY-GASV, DELLY-CLEVER and DELLY-Pindel), the pair DELLY-GASV performed better than the others, with a sensitivity of 95.7%, a PPV of 99.9% and a false discovery rate of 0.1%. Compared to the results obtained with the single tools, the combination of DELLY and GASV enabled overall a better resolution in the deletion discovery with an F1 score of 97.75%. In terms of SV breakpoint resolution, the pair DELLY-GASV identified 92.89% of the deletions with single base resolution and a total mean breakpoint distance resolution of 5.21 bp. The length of the predicted deletions (with DELLY-GASV) and the size of the simulated deletions were strongly correlated (Figure 21), with a Pearson's correlation coefficient of 0.99.

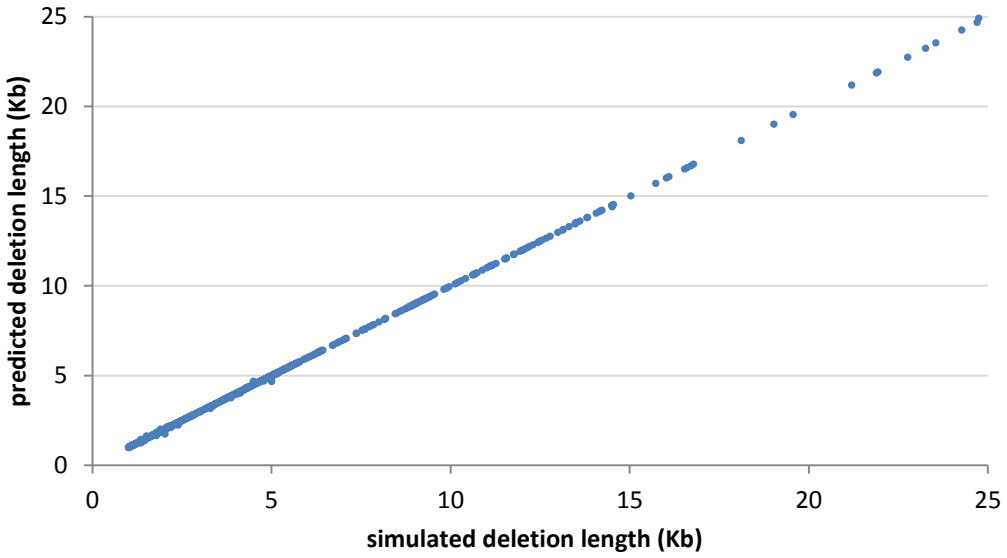


Figure 21: Correlation between the length of predicted and simulated deletions with simulated reads in *P. trichocarpa*.

In addition, we evaluated the performances of the four tools with real reads aligned to the *simulated SV genomes* of *V. vinifera* and *P. trichocarpa*. First, we explored the performances of the tools in terms of detected true positives, by progressively increasing the paired-end threshold (Figure 22).

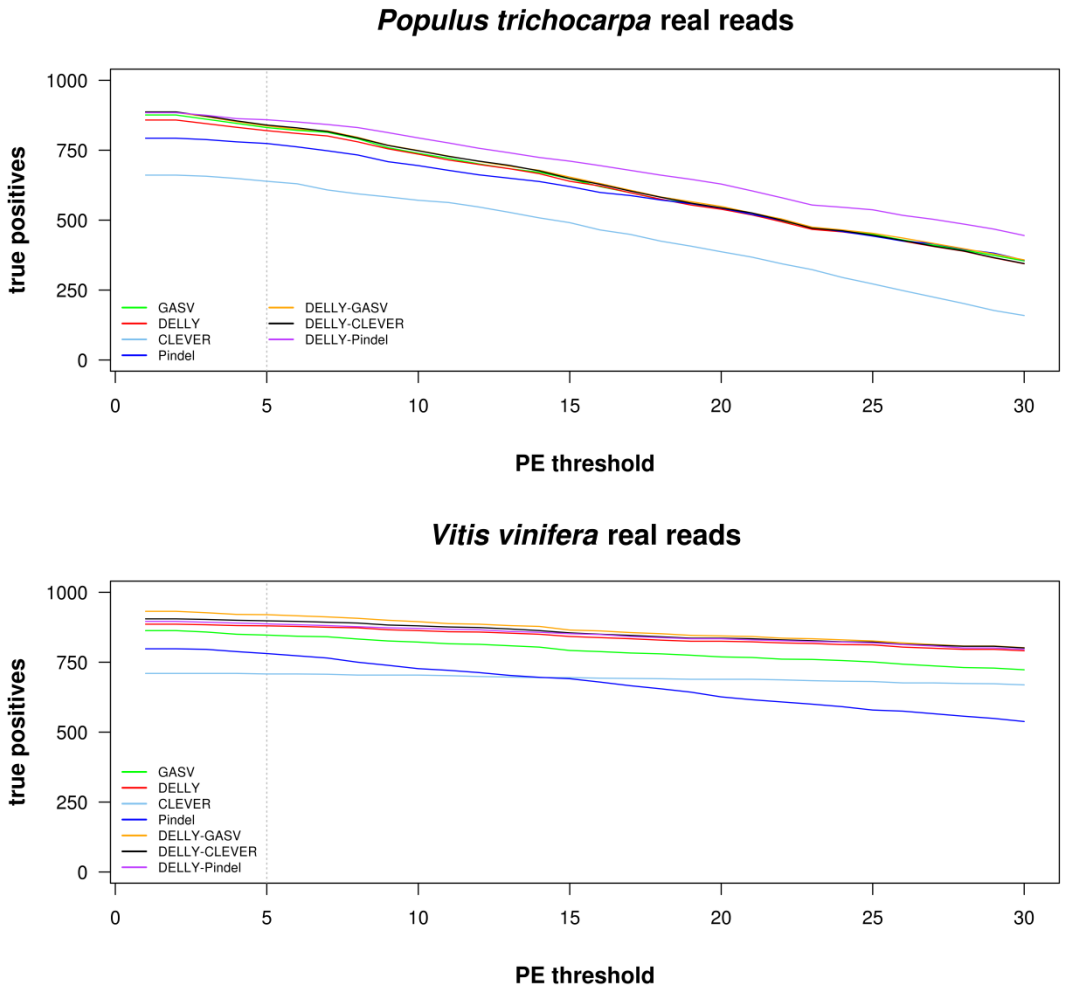


Figure 22: Number of *true positive* deletions identified using real reads aligned to the *simulated SV genomes* of *P. trichocarpa* and *V. vinifera*, as a function of the number of supporting paired reads required to call a SV.

In both species, all the tools or combinations had similar performance, with CLEVER and Pindel being the worst performing tools. By increasing the paired-end threshold, the number of true positives tended to decrease with a higher slope (estimated by linear regression) in poplar (mean tools slope -18.12) compared to grapevine (mean tools slope -4.52).

We evaluated the performances of the tools also in terms of non-simulated positive calls with both real read datasets (Figure 23). Non-simulated positive

calls are the sum of false positives and of real deletions in the reads compared to the assembly.

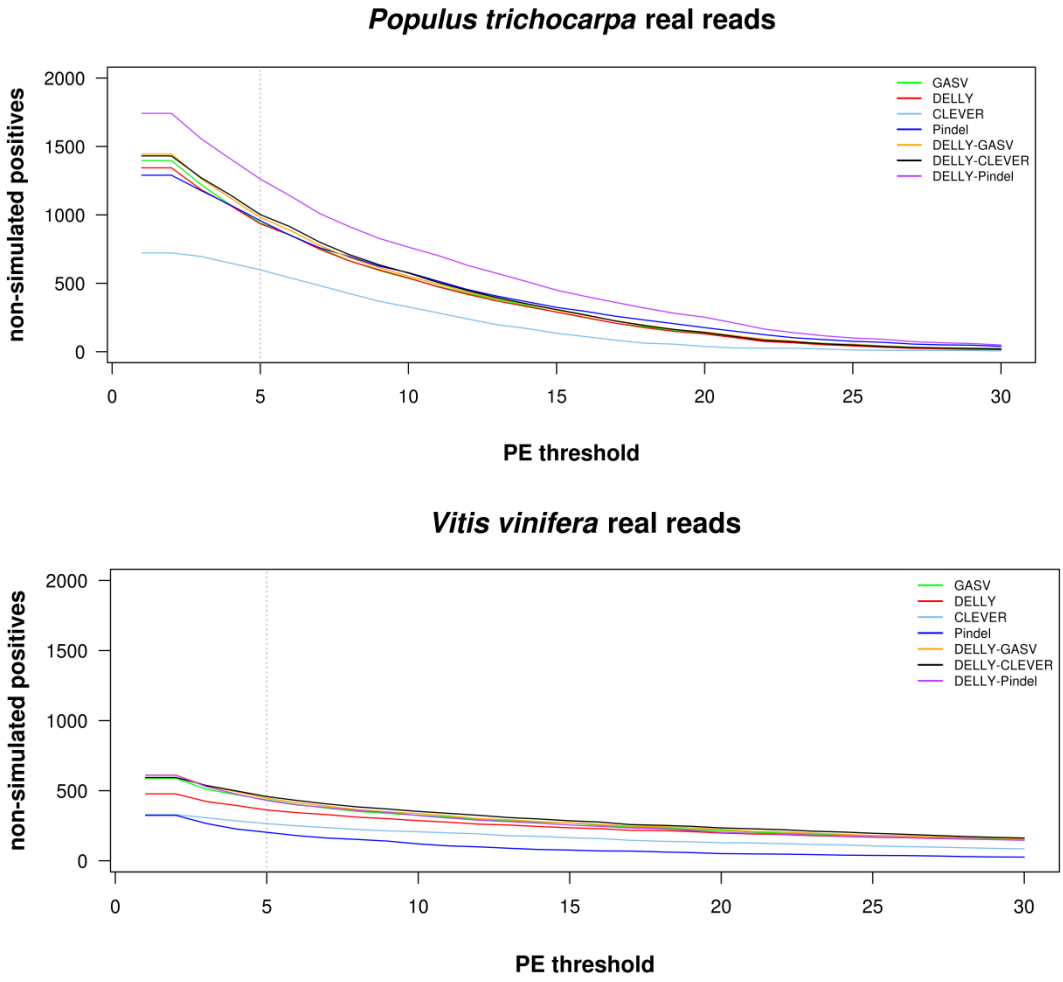


Figure 23: Number of *non-simulated positive* deletions identified using real reads aligned to the *simulated SV genomes* of *P. trichocarpa* and *V. vinifera*, as a function of the number of supporting paired reads required to call a SV.

In poplar all tools discovered a higher number of non-simulated deletions compared to grapevine. By progressively increasing the coverage of read pairs required, a decrease of non-simulated calls was observed with higher incidence in poplar (mean tools slope -44.41) than in *V. vinifera* (mean tools slope -11.82).

Based on the observed results, the deletions identified with at least five paired-end reads were selected for further statistical analyses. We evaluated

the performances of the different tools as reported in Table 5 and Table 6. Where needed, the non-simulated positives were used as an approximation of false positives.

Table 5: *Populus trichocarpa* real reads deletions statistics. Deletions report of real poplar reads aligned to the poplar *simulated SV genome*.

	CLEVER	DELLY	GASV	Pindel	DELLY-GASV	DELLY-CLEVER	DELLY-Pindel
Total deletions	1238	1755	1773	1732	1819	1842	2121
# true positives	639	820	831	774	837	840	859
# false negatives	361	180	169	226	163	160	141
# non-simulated positives	599	935	942	958	982	1002	1262
Mean breakpoint distance (bp)	8.04	3.00	7.87	1.94	3.34	3.54	2.96
# of exact predictions	90	767	83	759	769	770	803
Sensitivity (%)	63.90	82.00	83.10	77.40	83.70	84.00	85.90
PPV (%)	51.62	46.72	46.87	44.69	46.01	45.60	40.50
F1 score (%)	57.10	59.53	59.94	56.66	59.38	59.11	55.05

Table 6: *Vitis vinifera* real reads deletions statistics. Deletions report of real grapevine reads aligned to the modified grapevine *simulated SV genome*.

	CLEVER	DELLY	GASV	Pindel	DELLY-GASV	DELLY-CLEVER	DELLY-Pindel
Total deletions	973	1242	1286	984	1365	1356	1317
# true positives	708	880	847	781	920	898	887
# false negatives	292	120	153	219	80	102	113
# non-simulated positives	265	362	439	203	445	458	430
Mean breakpoint distance (bp)	10.99	4.86	23.39	1.48	6.48	5.94	4.81
# of exact predictions	72	839	42	779	842	841	846
Sensitivity (%)	70.8	88	84.7	78.1	92	89.8	88.7
PPV (%)	72.76	70.85	65.86	79.37	67.40	66.22	67.35
F1 score (%)	71.77	78.50	74.10	78.73	77.80	76.23	76.56

Based on the results of the SVs identified through the alignment of real reads to the simulated genomes, the pair DELLY-GASV performed again better than all the other combinations with an F1 score of 59.38% and 77.8%, respectively in poplar and grapevine. Compared to the other pairs, DELLY-GASV maximised the total number of true positives identified, with a precision of 67.4% and a sensitivity of 92% in grapevine. In poplar the pair DELLY-Pindel performed better in terms of sensitivity, but again deletions obtained with DELLY-GASV showed the highest precision (PPV, 46.01%). On the other hand, the number of non-simulated positives observed (here used as a proxy for false positives) was significantly higher than the number of false positives identified using the simulated reads. This is due to the fact that the non-simulated positives are the sum of two unknown quantities: a) the number of false positives; and b) real heterozygous deletions in the sequenced sample compared to the reference sample. Concerning the single base SV breakpoint resolution, in both species Pindel predicted the highest number of exact breakpoints (98.06% in *P. trichocarpa* and 99.74% in *V. vinifera*), followed by DELLY (93.54% and 95.34%, respectively). In Figure 24 we reported the correlation between the length of the predicted deletions and the size of the simulated SVs identified by the best performing pair DELLY-GASV. Both distributions had a very high Pearson's correlation coefficient (r=0.99 in both species).

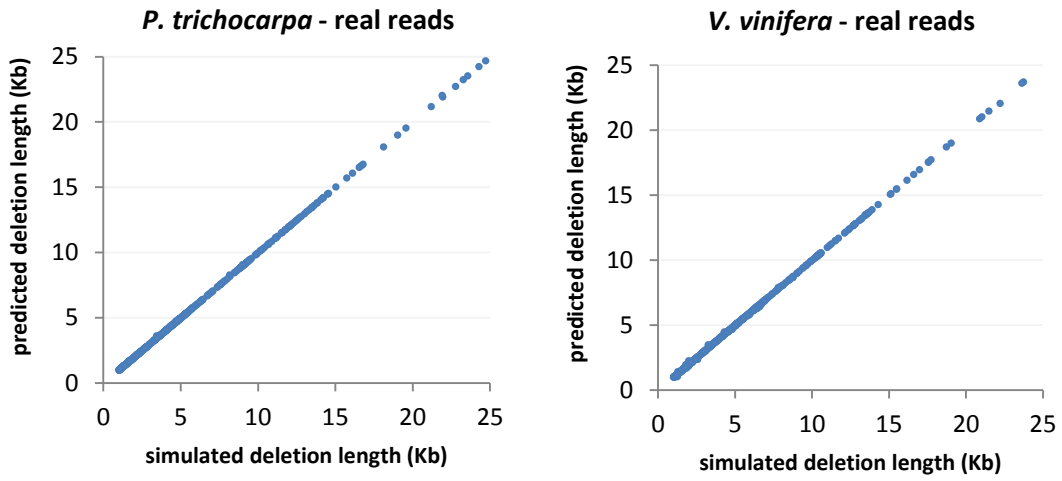


Figure 24: Correlation between the length of predicted and simulated deletions with real reads in *P. trichocarpa* and *V. vinifera*.

4.5 Simulation of insertions

Based on the same simulation data used for the deletions, we evaluated the performance of the insertion discovery pipeline. The same read datasets were used for the analysis: the real and simulated reads of *P. trichocarpa* and the real reads of *V. vinifera* (Table 7).

Table 7: Insertion simulation results.

	Populus trichocarpa		*Vitis vinifera*
	Simulated reads	**Real reads**	**Real reads**
# predictions	871	1230	936
# true positives	860	766	805
# false negatives	140	234	195
# non-simulated positives	0	464	131
Sensitivity (%)	86.00	76.60	80.50
PPV (%)	100	62.28	86.00
FDR (%)	0	-	-
F1 score (%)	92.47	68.70	83.16

With the simulated reads of *P. trichocarpa* the tool achieved an F1 score of 92.47%, identifying 860 true positives and no false positive (PPV of 100%). With real reads the software performed better in grapevine (with an F1 score of 83.16%) compared to poplar. In the latter the pipeline discovered a greater number of non-simulated positives (PPV of 62.28%), which could be either false positives or TE variants missing in the reference genome.

4.6 Structural variants detection in *V. vinifera*

For the SV detection in *Vitis vinifera* we selected a subsample of 50 varieties fulfilling the following criteria: a) at least 20X coverage of the uniquely aligned reads; and b) a k-mer and coverage profile compatible with an unbiased representation of the genome in the reads (determined by visual inspection of profile graphs). The k-mer distribution of a high quality sequencing library enables the distinction of the homozygous genome fraction from the heterozygous and/or hemizygous portion, with two clearly separated peaks. On the other hand, the sequencing coverage, based on the alignement of sequenced reads to the reference genome, enables the distinction between regions covered by reads originated by both alleles, and regions covered only by reads originating from a single allele of the studied genome, corresponding to the hemizygous genome portion.

We selected individuals with a k-mer profile trend ranging between the profile of Ansonica and Barbera (Figure 25). In the latter variety two different well separated peaks were observed: the first corresponded to the heterozygous and/or hemizygous portion of the genome, while the second peak at approximately 110 amounted to the homozygous portion of the genome. On the other hand, in Ansonica the two peaks were barely distinguishable. This particular profile might be influenced by the type of library sequenced and by the sequencing coverage.

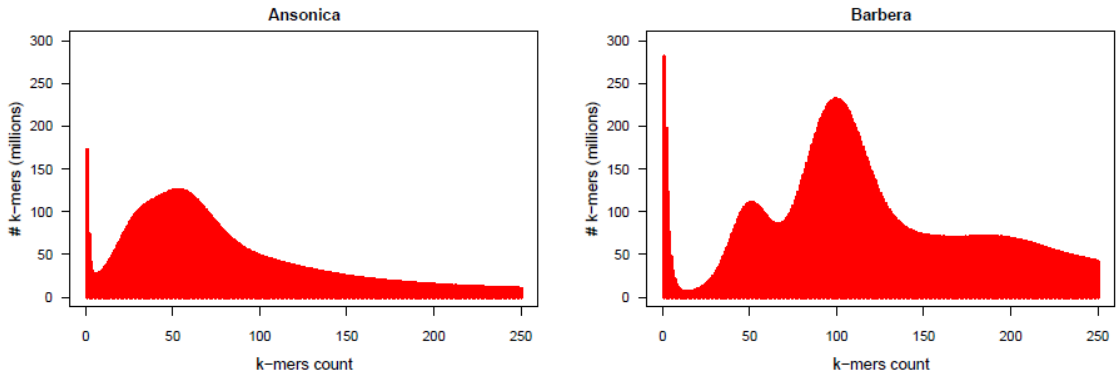


Figure 25: K-mer profile of Ansonica and Barbera. K-mer distribution obtained with a K value of 16.

The 50 varieties were selected also based on the sequencing coverage profile of the uniquely aligned reads. The latter varied with a trend ranging between the Ansonica and the Barbera one's (Figure 26). As above, in the latter variety two peaks were observed at approximately 37X and 73X (corresponding to the hemizygous and homozygous fractions, respectively), while in Ansonica an unique hump was observed.

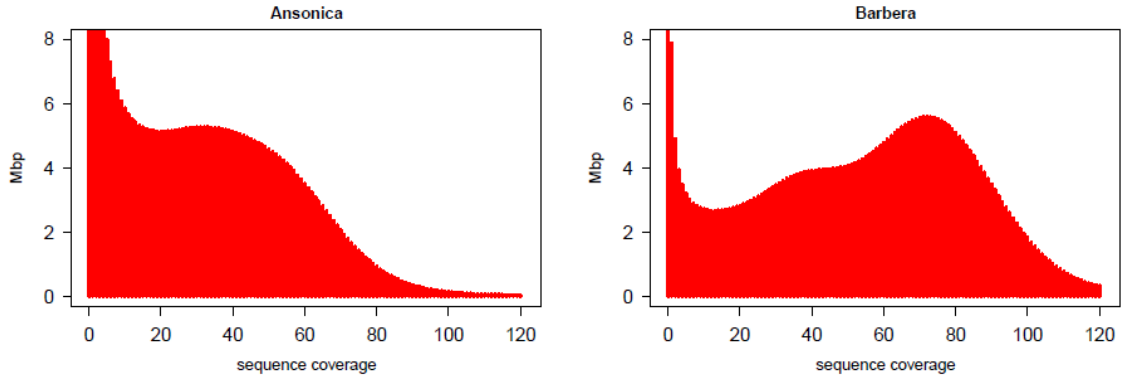


Figure 26: Coverage profile of Ansonica and Barbera. Coverage distribution obtained with uniquely aligned reads.

Concerning the population genetic structure, the varieties selected for the SV analysis covered the entire spectrum of variation detected in our grapevine

germplasm (Figure 27). With reference to the ADMIXTURE analysis using K=13, individuals from 10 groups out of 13 were selected. *Pontica Meridionalis Balcanica*, *Orientalis Antasiatica trans-Caucasica* and *Orientalis Caspica trans-Caucasica* were the only groups excluded from the analysis. 60% of the varieties were not identified as admixed, while 40% resulted from the admixture of two or more ancient populations.

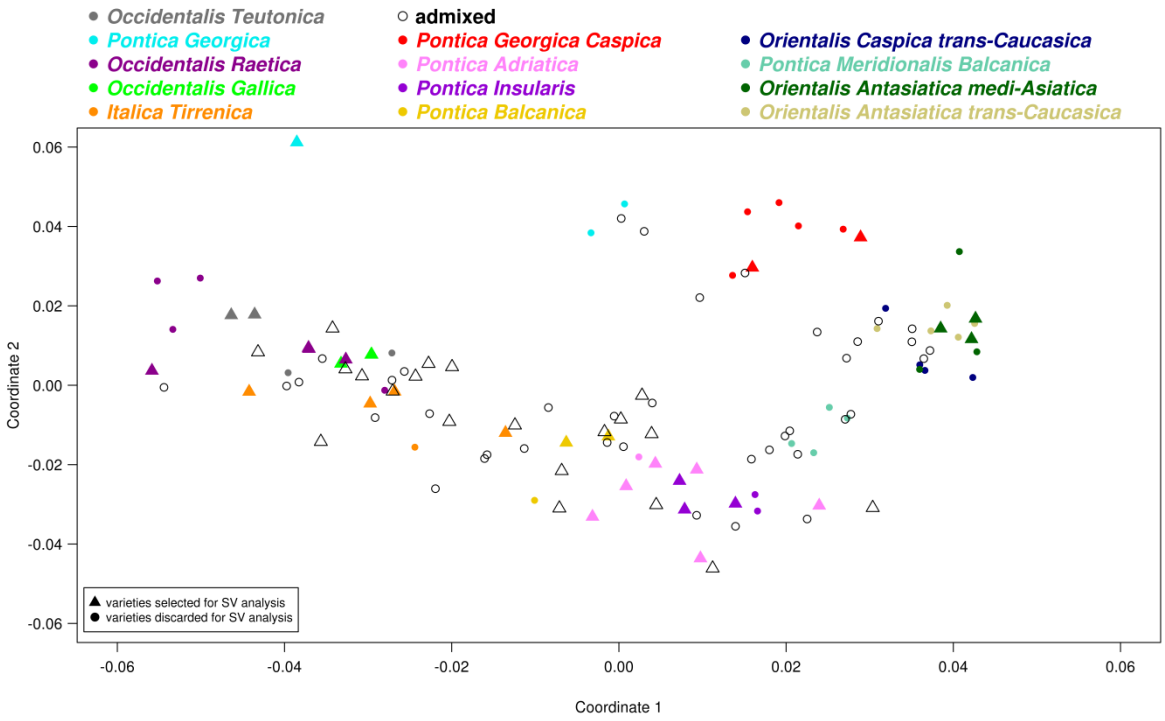


Figure 27: Principal Coordinates Analysis (PCoA) of the 50 varieties selected for the SV analysis. Genetic relationships between the 50 grapevine varieties. The colours of the groups correspond to the colours of ADMIXTURE. Varieties selected for SV analysis are depicted as **triangle**, while varieties discarded are reported as **dot**. Triangles and dots not coloured represent admixed varieties.

4.7 Identification of deletions

According to the simulation results, the combination of DELLY (Rausch T et al., 2012) and GASV (Sindi S et al., 2009) gave the best results in the detection of deletions against the reference sequence. This is the approach we selected for the analysis of our study population.

A total of 18,551 deletions were identified by merging the results of 50 grapevine varieties. Results are summarized in Table 8.

Table 8: Summary of the deletions identified in the grapevine population for each variety.

	Total deletions	Private deletions	Hetero[¥]	Homo[§]	Hetero (Mb)	Homo (Mb)
Ansonica	4375	24	1965	2410	10.86	12.55
Barbera	5231	33	2239	2992	11.89	15.91
Cabernet Franc	5038	51	2445	2593	13.51	13.77
Cabernet Sauvignon	4863	19	2538	2325	13.98	11.96
Catarratto B.C.	4763	16	1986	2777	10.53	14.44
Corvina Veronese	4811	17	2619	2192	13.67	11.40
Falanghina	4257	23	2099	2158	11.40	11.19
Fiano	5757	67	2628	3129	12.84	17.19
Garganega	5174	38	2260	2914	11.74	15.18
Glera	5697	32	2963	2734	16.27	13.83
Grechetto Bianco	4805	23	1999	2806	10.17	14.74
Greco di Tufo	4458	50	1829	2629	10.40	13.67
Heunisch Weiss	6345	75	3267	3078	17.86	16.18
Kishmish Vatkana	5964	102	2406	3558	13.07	18.90
Lambrusco Grasparossa	4479	16	2081	2398	10.94	12.36
Malvasia Bianca Lunga	4887	44	1997	2890	10.13	15.37
Malvasia di Sardegna	4512	29	2180	2332	11.46	12.32
Merlot Noir	6209	159	3004	3205	16.49	16.68
Montepulciano	4667	24	2265	2402	11.52	12.59
Muscat a Petits Grains B.	5337	47	2585	2752	13.95	14.60
Nasco	4570	26	2150	2420	11.46	12.61
Nebbiolo	5567	64	3219	2348	17.84	12.04
Nero d'Avola	5683	41	2587	3096	13.67	16.31
Nosiola	5222	32	2508	2714	12.94	14.22
Passerina	4558	27	1836	2722	9.70	14.34
Pecorino	5443	35	2748	2695	14.44	14.40
Picolit	6597	81	3560	3037	19.30	15.92
Pignoletto	4232	28	1709	2523	9.05	13.07
Pinot	4948	14	3295	1653	17.63	8.73
Refosco P.R.	5410	26	2948	2462	16.17	12.43
Ribolla Gialla	5591	27	2950	2641	16.18	13.63
Riesling Weiss	5973	100	2971	3002	16.28	15.75
Rkatsiteli	5509	100	2759	2750	15.03	14.48
Sangiovese	6081	70	3192	2889	17.89	14.74
Savagnin Blanc	4994	19	2818	2176	15.09	11.29
Schiava Gentile	4792	40	2998	1794	15.16	9.78

	Total deletions	Private deletions	Hetero[¥]	Homo[§]	Hetero (Mb)	Homo (Mb)
Sirgula	6183	168	2678	3505	14.49	18.62
Sultanina	5316	99	2506	2810	13.86	14.87
Tannat	5176	106	2315	2861	13.20	15.38
Terbash	6370	170	2772	3598	15.50	18.96
Terrano	5502	38	2837	2665	14.59	14.29
Tibouren	5771	54	3090	2681	16.64	13.84
Tocai Friulano	4826	22	2363	2463	13.06	12.50
Trebbiano Toscano	5429	31	2823	2606	14.94	13.88
Uva di Troia	5515	37	2748	2767	14.88	14.36
V395	6248	243	2205	4043	12.50	21.78
Verdicchio Bianco	5579	42	2633	2946	14.00	15.35
Vernaccia S.G.	4390	37	1791	2599	9.65	13.68
Welschriesling	5242	18	2235	3007	11.65	16.08
Zelen	5700	81	2879	2821	14.83	15.21

[¥] Number of heterozygous deletions; [§] number of homozygous deletions.

Picolit was the variety with the highest number of deletions (6597), accounting for 35.22 Mb in total. Pignoletto, with 4232 deletions (accounting for 22.12 Mb), resulted the variety with the lowest number of deletions. Picolit and Pignoletto showed the highest and the lowest number of heterozygous deletions respectively. V395 showed the highest number of homozygous deletions (64.71% of the deletions identified in V395). Pinot and Schiava Gentile had the lowest proportion of homozygous deletions (33.41% and 37.44%, respectively), proving their closeness to the reference genome. Focusing on the private deletions, V395 resulted the variety with the highest number of private deletions (243), in line with the results observed with the genotypic distances measured pairwise between the varieties using the SNP data: V395 resulted to be a fairly distant variety from the others. Also phenotypically the V395 plant exhibited a habit more similar to the wild *vinifera* than to the domesticated grapes, displaying non-domesticated characters.

We annotated the identified deletions with our custom developed pipeline (Figure 28) and we found that 65% of the deletions were annotated as a transposable element, thus confirming the important contribution of

transposable elements to genetic variation in plants (Kidwell MG & Lisch D, 1997). Only 33% of the deletions were not associated to a transposon superfamily, while 3% of the deletions were represented mostly by tandem repeats sequences (approximately 80%) and other type of sequences.

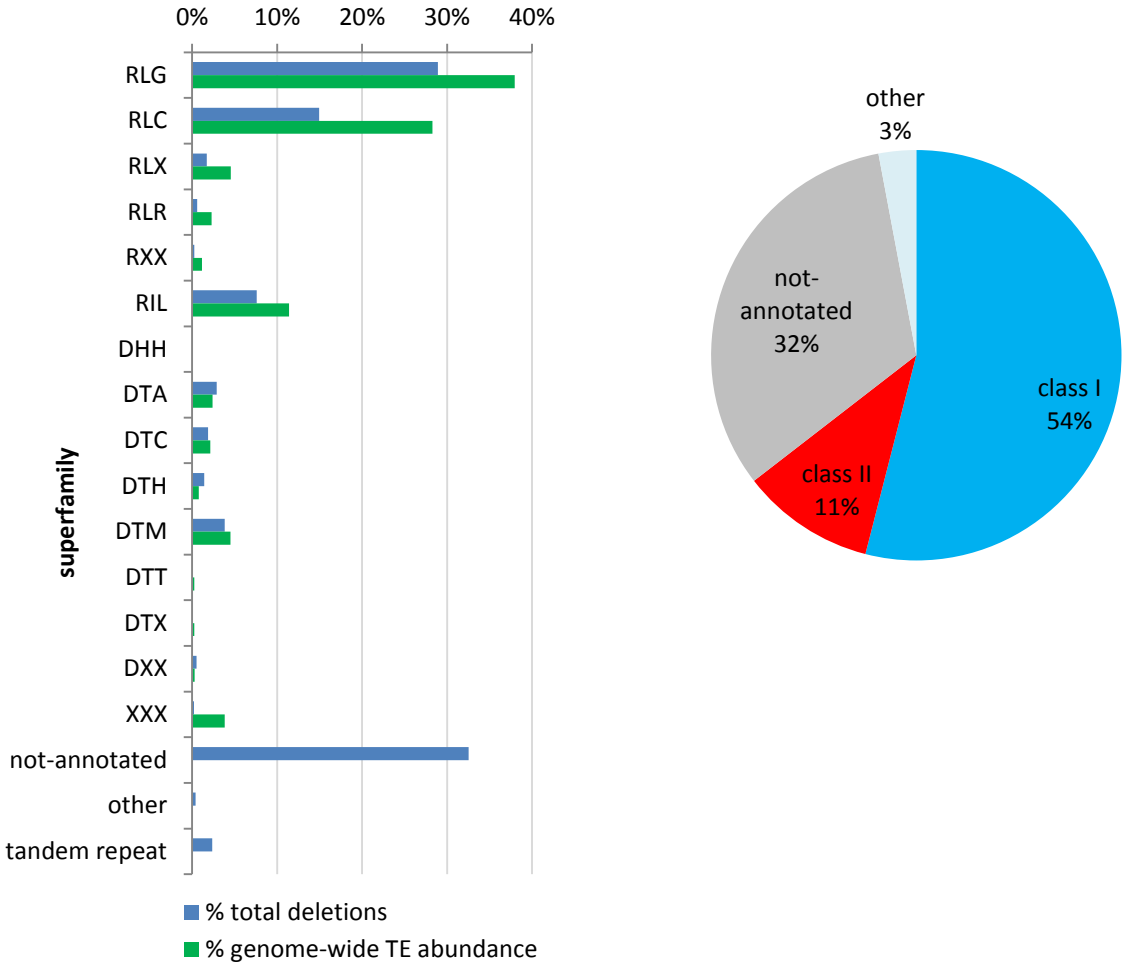


Figure 28: Annotation of the deletions identified in the grapevine population. Left three-letter code classified superfamily (Wicker T et al., 2007), as percentage of the total deletions identified in the *V. vinifera* population (blue) and genome-wide TE superfamily abundance, as percentage of the total amount of base-pairs masked as TE (green). **Right** summarized classification of SVs in classes.

Approximately 54% of the deletions were classified as transposable elements of class I, while 11% were associated to DNA transposons of class II. Among the retrotransposon elements, the LTR Gypsy (RLG) superfamily accounted for the vast majority of the classified TEs (28.93%), followed by Copia (RLC) elements (14.95%). LINE (RIL) represented 7.6% of the deletions. A total of 1,913 class I elements (10.3% of deletions) were masked (with RepeatMasker) by a single LTR for more than 90% of their sequence length, and were thus classified as solo-LTR. Concerning the class II elements, Mutator (DTM) was the most represented superfamily (3.84%), followed by hAT (DTA) elements (2.87%). As depicted in Figure 28, the annotation results reflected the genome-wide abundance of TEs. It is important to point out that the definition of deletion has a technical meaning and not necessarily a biological meaning. It refers to the fact that the sequence of the feature is present in the reference genome and is absent from another variety, without referring to the direction of the SV event. It is possible that most of the SV events defined as deletions with respect to the reference sequence are actually insertions in the genome of the reference individual.

To explore the genome-wide distribution of the deletions in the *V. vinifera* genome, we divided the genome in windows of constant size in terms of mappable reads. A total of 1862 windows were obtained with a mean size of 227.57 Kb and a standard deviation of 20.11 Kb. We looked at the distribution of the deletions across the windows. By comparing the distribution of the deletions with the null hypothesis of *Poisson* distribution and correcting for multiple testing, approximately 5% of the windows (96 out of 1862) showed significantly higher or lower density of deletions. Then, for each window, we plotted across the 19 grapevine chromosomes the number of transposable elements belonging to the class I superfamilies.

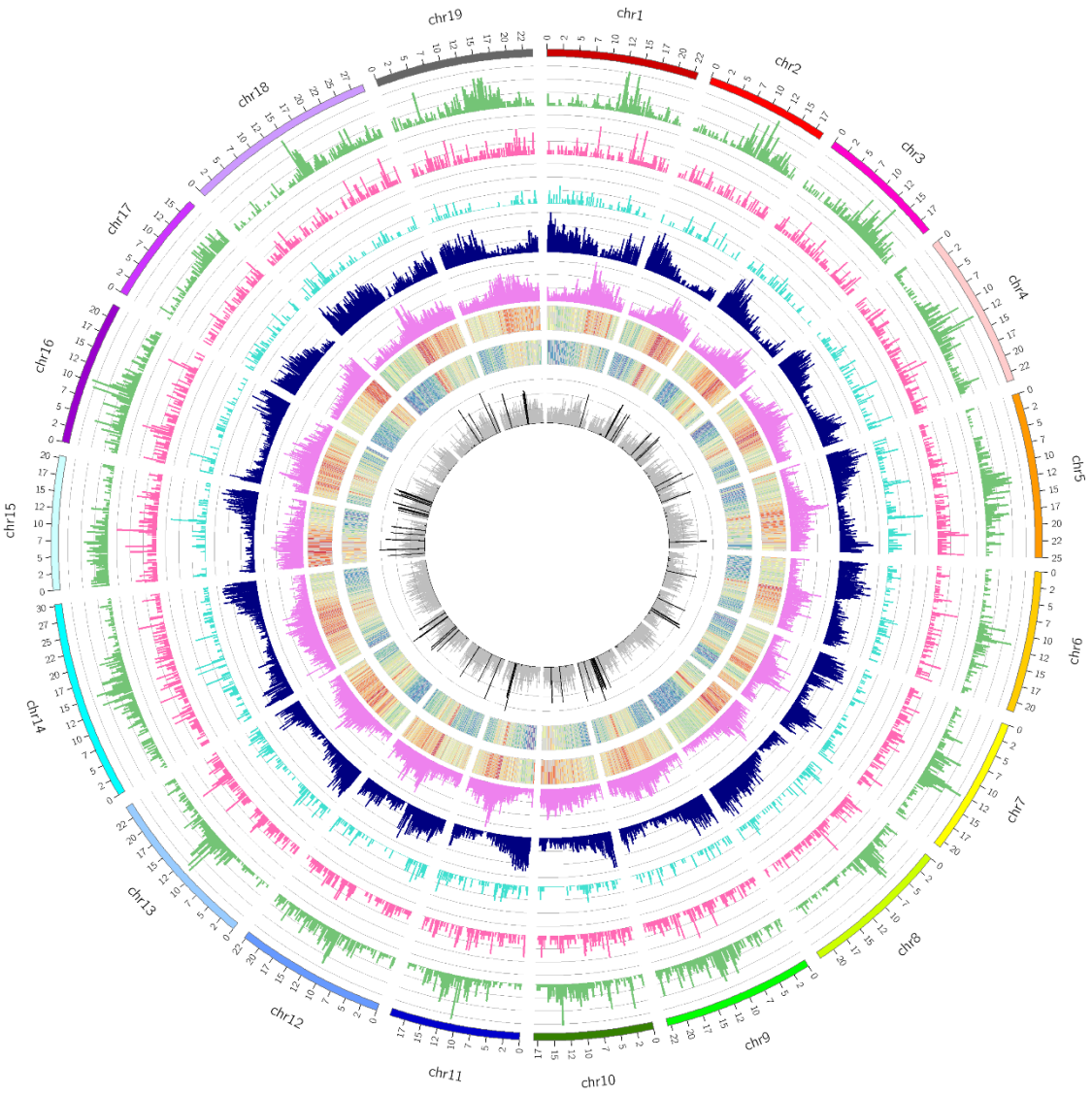


Figure 29: Chromosome distribution of the deletions involving class I transposable elements. From the outside in: Gypsy elements (green, y upper limit 15), Copia elements (pink, y upper limit 10), LINE elements (turquoise, y upper limit 10), gene density (blue, y upper limit 51 Kb/window), repeat density (violet, y upper limit 220 Kb/window), CG methylation context (heat map), CHG methylation context (heat map) and lastly the total deletions distribution (grey, y upper limit 30). In the innermost circle, black bars represent the windows with significant FDR corrected *Poisson* distribution.

As depicted in Figure 29, deletions involving Gypsy elements tend to accumulate in regions poor of genes but rich in repeats. We fitted the 'Gypsy-repeat' and 'Gypsy-gene' distributions with a linear model and the p-values measured for

both comparisons were highly significant (p-values<$2E^{-16}$). Per window, we estimated that the increase of one Gypsy deletion led to a decrease of approximately 2 Kb of coding sequences and to an increase of approximately 9 Kb of repetitive sequences. As a consequence, RLG occurred especially in pericentromeric regions where the gene density decreased deeply, while the repeats fraction increased considerably. We further compared the distribution of the deletions classified as Gypsy with the methylation profiles of the CG and CHG contexts. Methylation data were obtained from our group of research performing methylation analyses on Pinot Noir (Mirko Celii, PhD thesis). For each window and for each context, the mean methylation value was calculated. An increase in the methylation levels in the CG and in the CHG contexts was observed in the regions where Gypsy tended to accumulate: a more pronounced methylation level increase was observed in the CG context compared to the CHG context.

When a new copy of an LTR retrotransposon is created, the two LTR sequences are identical, while over time they tend to accumulate mutation. As proposed by SanMiguel and collaborators, LTR insertion times may be estimated by comparing the two LTR sequences (SanMiguel P et al., 1998). We compared the time of insertion of the LTR-retros identified in deletions (SV-LTRs) with the age of LTR elements not affected by SV in our grapevine population (noSV-LTRs) (Figure 30). A total of 2,871 SV-LTRs and 1,187 noSV-LTRs were investigated.

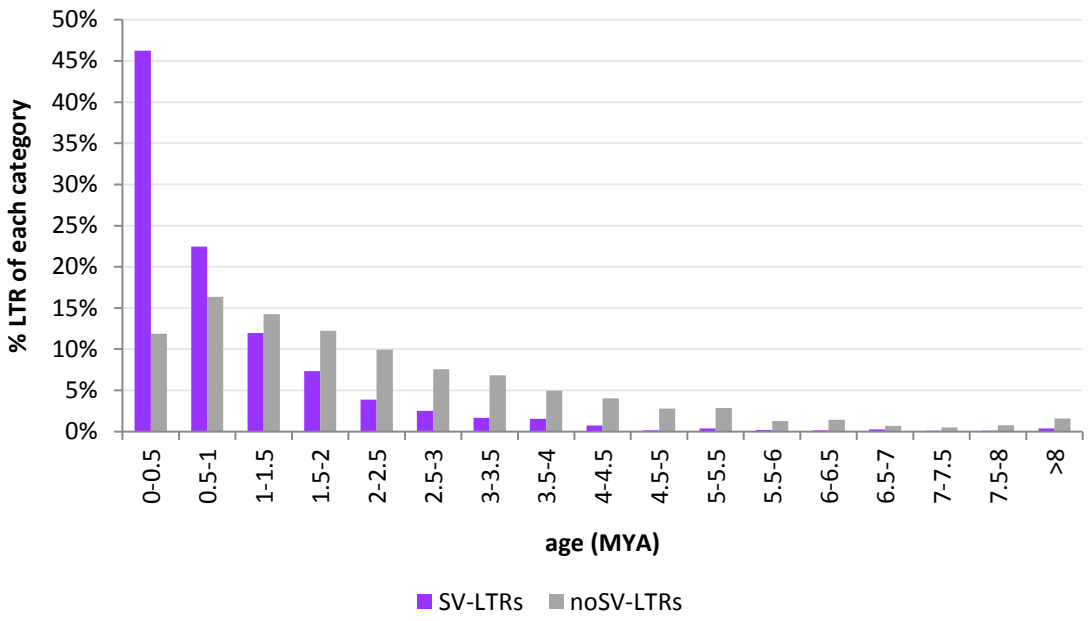


Figure 30: Insertion age of LTR retrotransposons involved in structural variation (SV-LTRs) and of LTR elements shared in the grapevine population (noSV-LTRs).

As depicted in Figure 30, LTR retrotransposons involved in SV tend to be younger than the LTR-retros fixed in the population of grapevine. More than 45% of the SV-LTR elements moved recently, showing an insertion time below 0.5 million of years (MYA). We performed the non-parametric two-sample Kolmogorov-Smirnov test and rejected the null hypothesis of no difference between the two distributions, since the distributions resulted significantly different (p-value<0.05): the SV-LTR elements are significantly younger than the shared LTR retrotransposons.

The length of the transposable elements can vary drastically from few hundred base pairs to more than 10,000 bp (Bennetzen JL, 2000). We evaluated the length distribution of the deletions classified as transposable elements in our dataset (Figure 31).

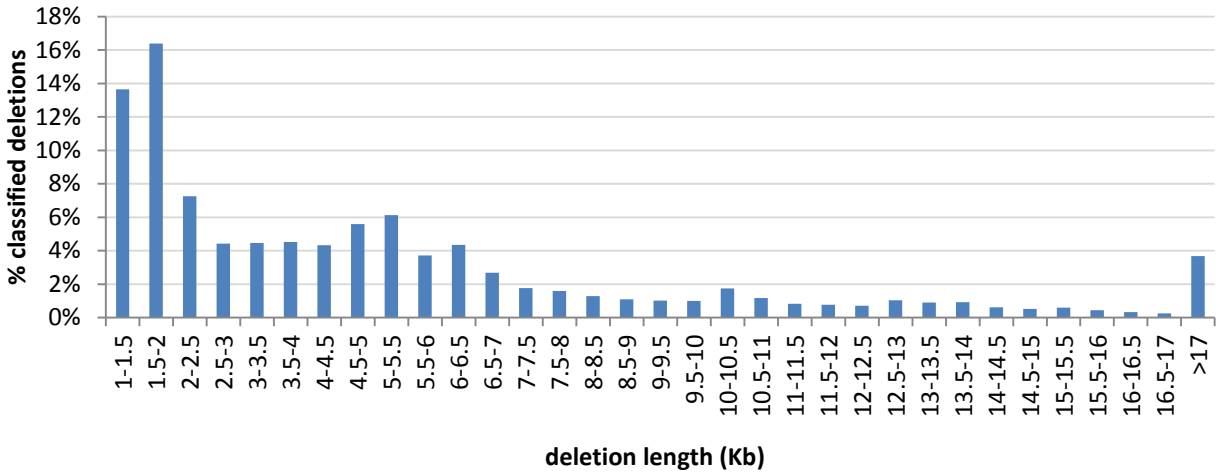


Figure 31: Length distribution of deletions classified as transposable elements (calculated as percentage of the classified deletions).

Three peaks were observed at 1.5-2 Kb, at 5-5.5 Kb and at 10-10.5 Kb, corresponding respectively to incomplete retrotransposable elements or DNA elements, to Copia TEs and to Gypsy transposable elements.

The length of the unclassified deletions showed a completely different distribution (Figure 32). Short deletions were the majority, and variants shorter than 4 Kb accounted for 50% of the unclassified deletions.

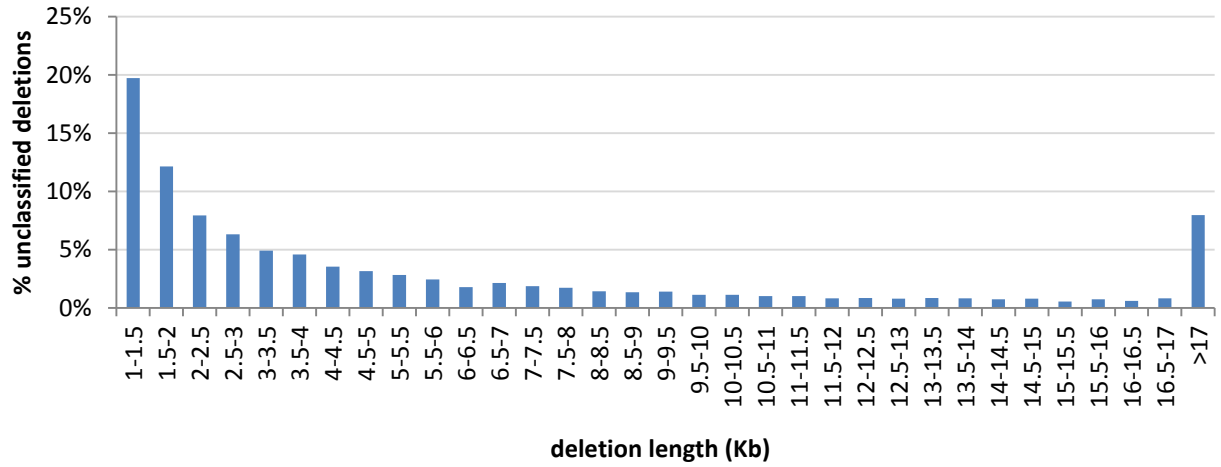


Figure 32: Length distribution of the unclassified deletions (calculated as percentage of the unclassified deletions).

In order to study the intraspecific distribution of the deletions, we investigated the frequency of the classified deletions across the 50 grapevine individuals (Figure 33). The majority of the classified deletions were observed at low frequency, but 42.34% had a frequency in the population greater than 0.5. The discovery of deletions is subjected to an ascertainment bias, since the analytical procedure identifies deletions with respect to a single haplotype captured in the reference genome. One of the consequences is also the fact that SVs classified as deletions may have biologically originated by insertions in the reference individual and other varieties that share the same haplotype.

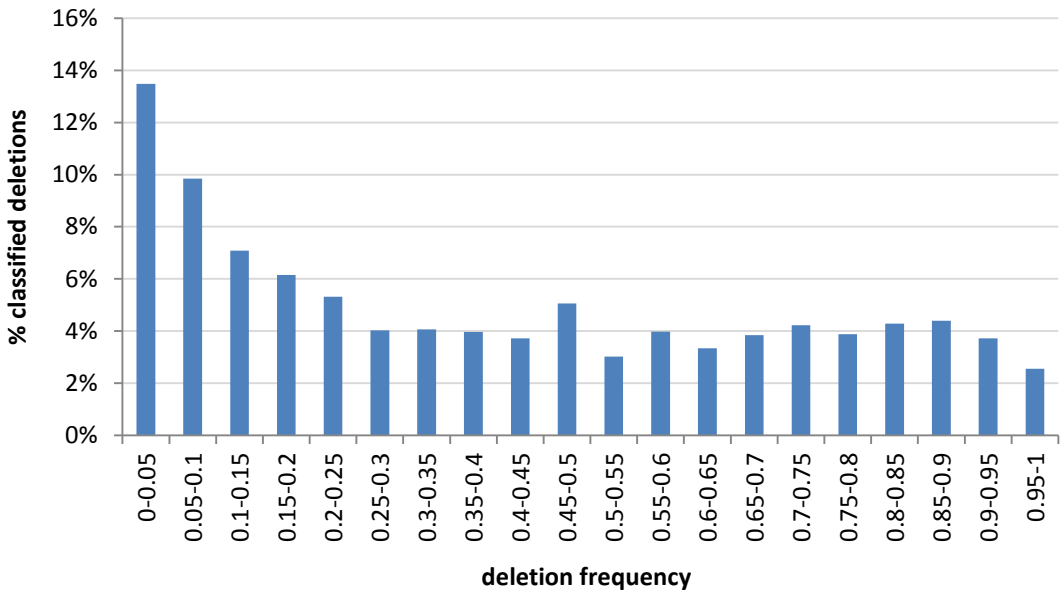


Figure 33: Frequency distribution of deletions classified as transposable elements (calculated as percentage of the classified deletions).

More than 2% of the classified deletions were called in homozygous state between the 50 varieties. These calls are likely caused by assembly errors in the reference sequence, since both Pinot Noir and Schiava Grossa (parental varieties of the reference genome) carried the called SV in homozygous condition.

The frequency spectrum of unclassified deletions showed a completely different pattern (Figure 34). The majority of the SVs were rare (mostly private): more than 35% of the unclassified deletions had a frequency lower than 0.05, corresponding to heterozygous deletions carried by only one variety.

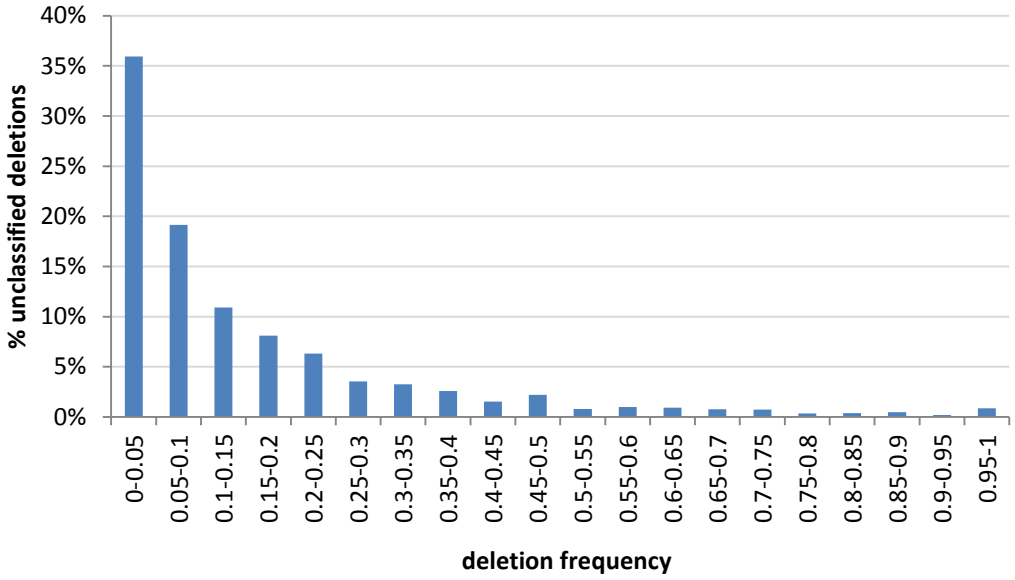


Figure 34: Frequency distribution of unclassified deletions (calculated as percentage of the unclassified deletions).

We then investigated the *V. vinifera* genes affected by deletions. Out of the 101,935,035 bp involved in deletions, 21,673,623 bp corresponded to genic regions. As much as 13% of the gene space was influenced by deletions. Introns cover 115.5 Mb of the reference genome and deletions affected 18.4 Mb of those regions (15.9% of the total intron length); CDS cover 33.8 Mb of the reference genome and deletions involved 2.5 Mb of those regions (7.4% of total CDS length). Lastly, 5' and 3' UTR regions cover respectively 5.5 and 10.8 Mb of the reference genome and deletions affected 0.33 Mb (5.9%) and 0.47 Mb (4.4%), respectively.

An accurate analysis of the deletions overlapping genes revealed that deletions comprising LINE elements tended to accumulate in genes. Approximately 80% of the deletions classified as LINEs involved gene regions; 93.62% of the LINE deletions located in a gene, affected only the intronic portion. On the other hand, the other gene regions were affected mostly by non-annotated deletions.

Deletions affected a total of 5,679 genes. We functionally annotated the genes with Blast2GO. 68% of the genes involved in deletions were associated with at least one Gene Ontology (GO) annotation. Nine over-represented functional GO categories were detected, involved in biological processes, cellular components or molecular function.

Table 9: Over-represented GO categories influenced by deletions.

GO term	Category[§]	Annotated genes interrupted by deletions (%)	FDR
Nucleotide binding	F	21.67%	$6.20E^{-10}$
Cellular protein modification process	P	14.65%	$1.44E^{-02}$
Reproduction	P	8.63%	$1.72E^{-02}$
DNA metabolic process	P	5.17%	$1.23E^{-02}$
Cell cycle	P	5.48%	$1.11E^{-04}$
Embryo development	P	3.93%	$4.95E^{-02}$
Regulation of gene expression, epigenetic	P	3.13%	$2.26E^{-03}$
Motor activity	F	0.85%	$2.28E^{-02}$
Nuclear envelope	C	0.80%	$2.37E^{-02}$

§ F: molecular function; P: biological process; C: cellular component

The GO category *nucleotide binding* accounted for 21.67% of the annotated genes interrupted by deletions, resulting thus the most influenced category, followed by genes involved in cellular protein modification processes (14.65%). Genes belonging to the former category may be related to disease resistance genes, while the latter encompasses genes involved in post-translational

modification processes, which modulate the activity of proteins (Mann M & Jensen ON, 2003).

4.8 Identification of insertions

We used the custom developed pipeline for the insertion analysis across the 50 grapevine varieties. A total of 54,254 insertions were identified after merging the results (Table 10).

Table 10: Summary of the insertions identified in the grapevine population for each variety.

	Total insertions	Private insertions	Hetero[¥]	Homo[§]	Hetero Mb	Homo Mb
Ansonica	10049	98	7003	3046	45.16	18.49
Barbera	10971	160	8803	2168	53.91	13.30
Cabernet Franc	10071	189	7048	3023	42.26	17.87
Cabernet Sauvignon	10957	87	8608	2349	53.67	13.78
Catarratto B.C.	10493	47	7413	3080	46.51	18.11
Corvina Veronese	9048	85	7314	1734	42.79	10.48
Falanghina	9599	85	6903	2696	42.89	16.41
Fiano	11367	111	8628	2739	51.57	16.53
Garganega	9142	61	6937	2205	40.67	13.33
Glera	9850	67	7563	2287	45.76	13.52
Grechetto Bianco	9782	54	7048	2734	43.15	16.06
Greco di Tufo	9727	84	6442	3285	40.38	19.85
Heunisch Weiss	10727	150	7889	2838	48.21	17.07
Kishmish Vatkana	12203	421	8556	3647	53.53	22.34
Lambrusco Grasparossa	8685	73	6568	2117	39.44	12.86
Malvasia Bianca Lunga	10355	151	7203	3152	44.56	19.33
Malvasia di Sardegna	9432	92	6998	2434	43.26	14.37
Merlot Noir	11592	272	9426	2166	55.82	12.86
Montepulciano	9709	89	7387	2322	44.13	14.25
Muscat a Petits Grains B.	9299	113	6887	2412	41.66	14.15
Nasco	9867	88	7493	2374	45.93	14.28
Nebbiolo	12310	213	9072	3238	57.31	19.42
Nero d'Avola	11212	158	8456	2756	51.39	16.27
Nosiola	9441	78	7299	2142	44.01	12.63
Passerina	9652	58	6694	2958	42.29	17.95
Pecorino	10070	62	8029	2041	48.28	12.05
Picolit	11187	176	8568	2619	52.08	15.21

	Total insertions	Private insertions	Hetero[¥]	Homo[§]	Hetero Mb	Homo Mb
Pignoletto	9065	38	6304	2761	39.16	16.53
Pinot	10446	99	8118	2328	50.14	13.78
Refosco P.R.	10551	71	8720	1831	52.16	10.57
Ribolla Gialla	9834	45	7346	2488	45.28	14.59
Riesling Weiss	13402	321	9710	3692	60.62	22.72
Rkatsiteli	9298	291	6826	2472	40.81	15.12
Sangiovese	10906	103	7989	2917	48.48	17.52
Savagnin Blanc	9177	55	7249	1928	43.71	11.04
Schiava Gentile	10338	108	8020	2318	49.25	14.09
Sirgula	12558	745	9073	3485	56.23	20.97
Sultanina	10464	251	7471	2993	45.96	17.86
Tannat	12375	375	8173	4202	50.59	25.74
Terbash	12655	730	8989	3666	55.15	22.01
Terrano	10518	80	8263	2255	49.25	13.57
Tibouren	10555	154	8398	2157	50.29	12.96
Tocai Friulano	10117	73	7478	2639	46.72	15.99
Trebbiano Toscano	9837	72	7359	2478	44.59	14.40
Uva di Troia	11055	134	8420	2635	52.41	15.39
V395	12025	1011	7426	4599	46.99	28.12
Verdicchio Bianco	10401	83	7766	2635	47.53	15.85
Vernaccia S.G.	9467	104	6548	2919	41.52	17.46
Welschriesling	10374	88	8107	2267	48.00	13.90
Zelen	11339	181	8725	2614	53.05	15.73

[¥] Number of heterozygous insertions; [§] number of homozygous insertions.

The variety with the highest number of insertions was Riesling Weiss with 13,402 total insertions identified, while Lambrusco Grasparossa showed the lowest number (8,685). V395 presented the highest number of homozygous insertions with 28.12 Mb involved in homozygous insertions. On the contrary, Riesling Weiss had the highest number of heterozygous insertions, involving more than 60 Mb. Concerning the private insertions, as observed for the deletions, V395 showed the highest proportion of private SVs (8.41% of the total insertions identified in the variety), a further evidence supporting differentiation from the *vinifera* population. Besides V395, Terbash and Sirgula (two varieties showing greater haplotype distance from the reference compared to the other cultivars) had the highest proportion of private insertions.

As occurred for deletions, insertions are also caused by the movement of transposable elements. In addition, our analysis method imposed a very strong bias on the detected insertions, since they had to match a previously described transposable element or a previously detected deletion. The distribution among TE superfamilies is, however, not biased by our detection algorithm and is shown in Figure 35.

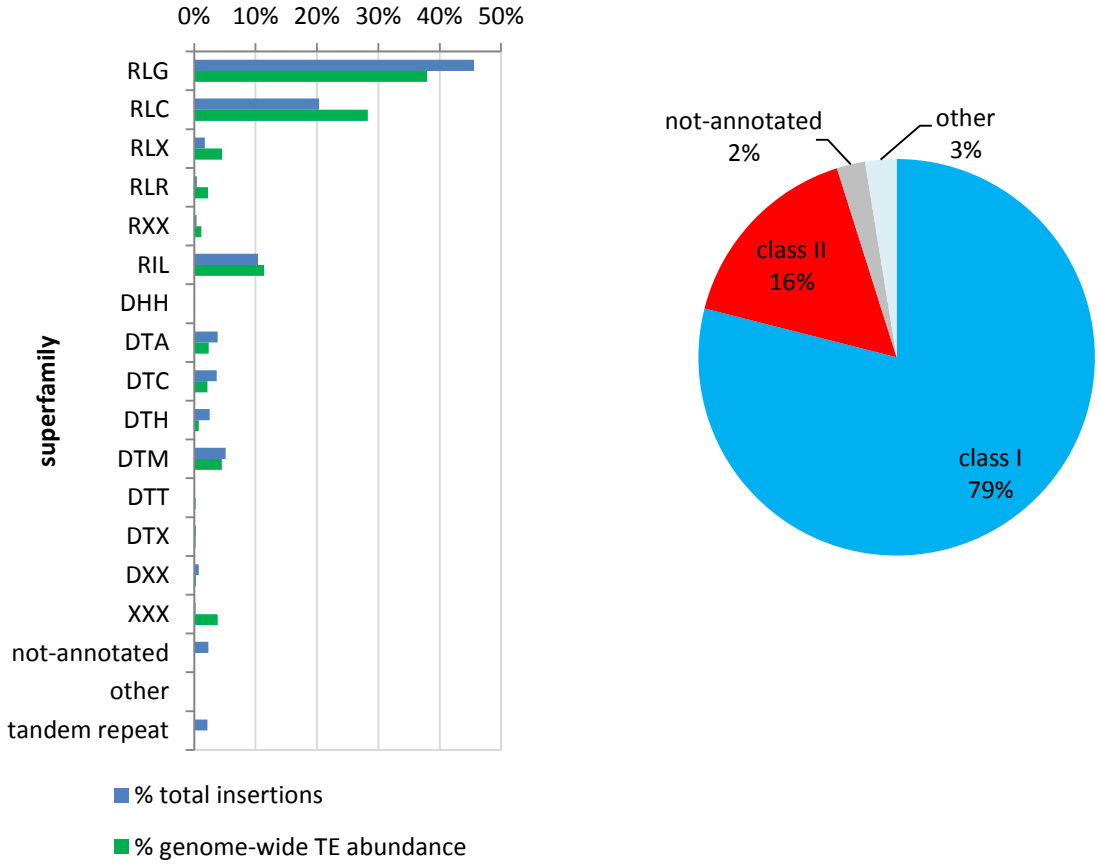


Figure 35: Annotation of the transposable elements involved in insertions. Left three-letter code classified superfamily (Wicker T et al., 2007), as percentage of the total insertions discovered in the grapevine population (blue) and genome-wide TE superfamily abundance, as percentage of the total genome base-pairs masked as TE (green). **Right** summarized classification of the transposable elements in classes.

95% of the total insertions were annotated as TE: 79% were classified as transposable elements of class I (retrotransposons), while 16% as elements of class II (DNA transposons). Only 2% of the insertions mapped to non-annotated sequences, which might correspond to rare transposable elements detected as deletions and unclassified with the annotation pipeline. Among class I elements, Gypsy were the most represented (45.63%), followed by Copia TEs (20.37%). LINEs accounted for 10.38% of the insertions. Concerning the class II elements, the Mutator superfamily was the most represented (5.14%), followed by hAT (DTA) and CACTA (DTC) superfamilies.

As accomplished for the deletions, we measured the genome-wide distribution of the insertions across windows with fixed mappability. By comparing the distribution of the insertions with the null hypothesis of *Poisson* distribution and correcting for multiple testing, approximately 15% of the windows (291 out of 1862) showed significantly higher or lower density of insertions. The distribution of the class I transposable elements was plotted for each window across the 19 *V. vinifera* chromosomes (Figure 36).

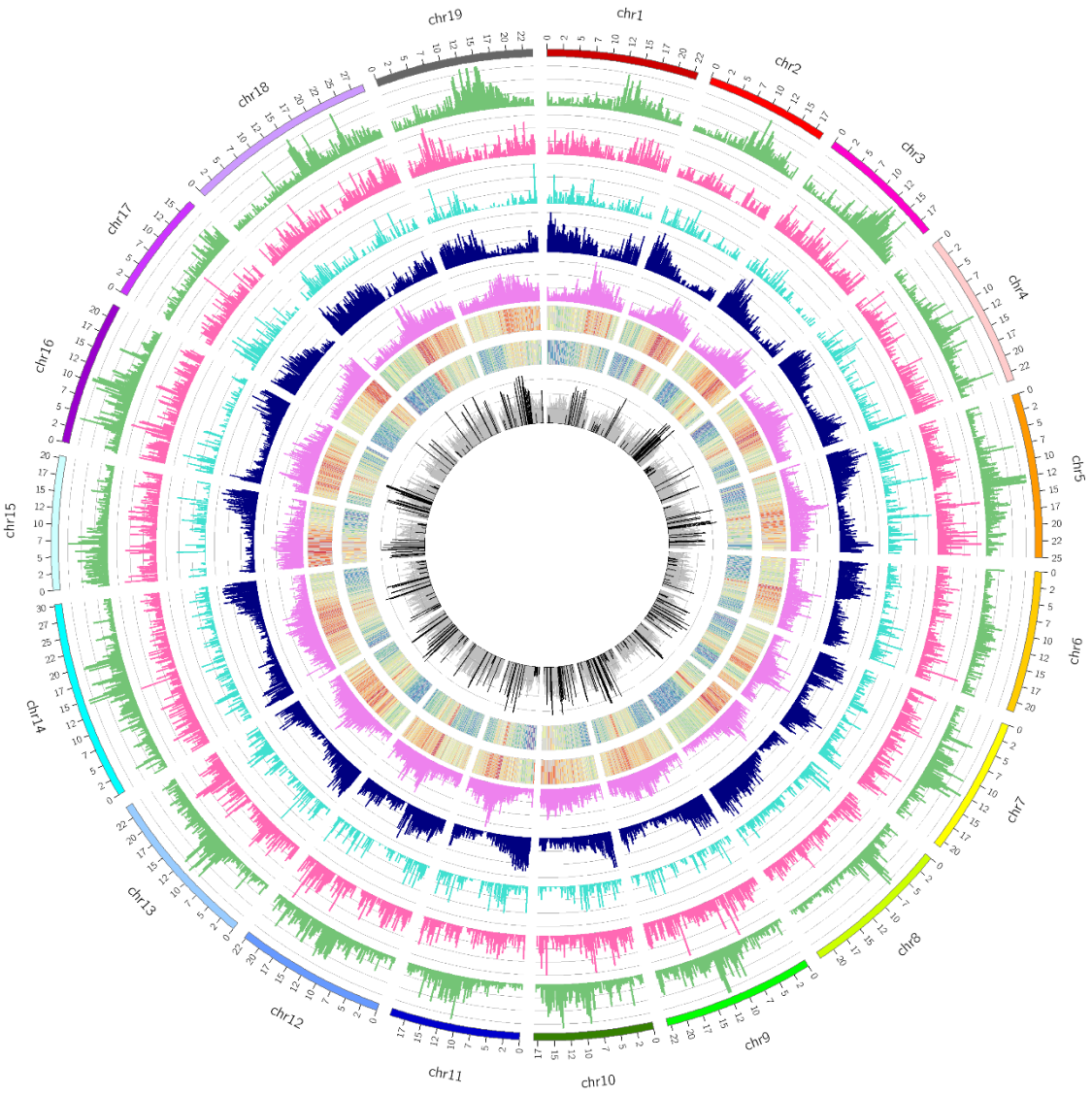


Figure 36: Chromosome distribution of the insertions involving class I transposable elements. From the outside in: Gypsy elements (green, y upper limit 45), Copia elements (pink, y upper limit 20), LINE elements (turquoise, y upper limit 20), gene density (blue, y upper limit 51 Kb/window), repeat density (violet, y upper limit 220 Kb/window), CG methylation context (heat map), CHG methylation context (heat map) and total insertions distribution (grey, y upper limit 60). Black bars in the innermost circle represent the windows with significant FDR corrected *Poisson* distribution.

As observed for the deletions, Gypsy elements accumulated in pericentromeric regions, where the gene density decreased consistently, while on the contrary the repeats density increased drastically. We fitted the 'Gypsy-repeat' and

'Gypsy-gene' distributions with a linear model and again the p-values resulted highly significant (p-values<$2E^{-16}$). An increase of one Gypsy element, per window, led to a decrease on average of approximately 0.6 Kb of coding sequences and to an increase of 2.7 Kb of repetitive sequences. On the other hand, Copia and LINE showed a uniform distribution across most of the 19 chromosomes of *V. vinifera*. On a few chromosomes, such as chromosomes 12, 13 and 19, an opposite distribution trend of the Copia elements was observed compared to Gypsy TEs: an increase of the Copia TEs was appreciated where the number of Gypsy decreased. Concerning the methylation levels of the CG and CHG contexts, an increase in the methylation level was observed in the repeats rich regions. In correspondence of the RLG peaks, a pronounced increase in the methylation level was observed in the CG context, and less pronounced in the CHG context.

We evaluated the length distribution of the transposable elements causing insertions (Figure 37). While the size of the TEs, identified as deletions, was exactly defined by the length of reference genome missing in the variety, the size of the TEs identified as insertions could only be inferred, based on the length of the TEs present in the database that provided the signal for defining the event of insertion. Different peaks were appreciated. The first peak was observed at 1.5-2 Kb and both DNA and RNA transposons fell in this length category. A second peak was observed at 4.5-5.5 Kb, represented by Copia TEs; while a third peak at 6-6.5 Kb was represented by LINE transposable elements. Lastly, at 10-10.5 Kb the peak corresponded to Gypsy transposable elements.

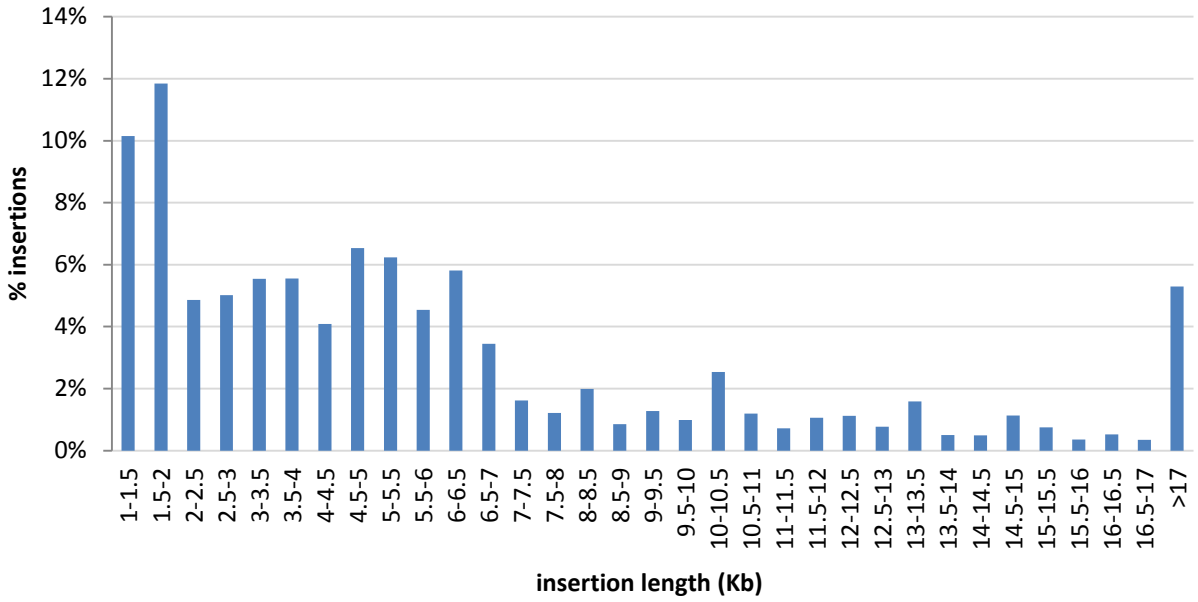


Figure 37: Length distribution of insertions classified as transposable elements.

The frequency distribution of the insertions inside the grapevine population was then measured (Figure 38). The majority of the insertions showed a very low frequency in the population, resulting thus rare events not shared between the varieties. 88.66% of the insertions had an insertion frequency below 0.5.

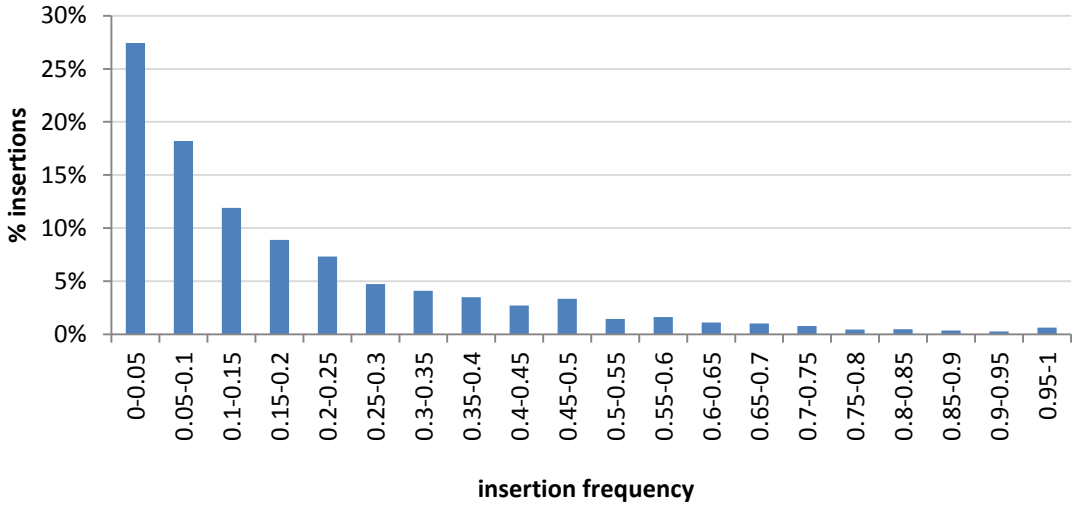


Figure 38: Frequency distribution of insertions in the grapevine population.

Furthermore, we evaluated the functional role of the genes interrupted by insertions. A total of 7,829 genes were interrupted by insertions in at least one variety. Approximately 80% of the insertions disrupting genes affected only the introns. 8.61% of the insertions were predicted in coding regions, while 4.36% and 2.79% insertions affected the 3' and 5' UTR regions, respectively: the remaining insertions influenced more than one gene fraction. A functional annotation of the genes interrupted by transposable elements allowed the annotation of 73% of the genes with at least one Gene Ontology term. Eleven over-represented functional categories were identified.

Table 11: Over-represented GO categories of genes disrupted by insertions.

GO term	Category[§]	Annotated genes interrupted by insertions (%)	FDR
Nucleotide binding	F	20.33%	$6.97E^{-08}$
Cellular protein modification process	P	14.76%	$2.54E^{-05}$
Cytosol	C	13.28%	$2.25E^{-03}$
Reproduction	P	9.42%	$1.91E^{-10}$
DNA metabolic process	P	5.22%	$1.81E^{-05}$
Cell cycle	P	5.59%	$9.50E^{-10}$
Endoplasmic reticulum	C	3.85%	$2.01E^{-02}$
Embryo development	P	4.26%	$1.36E^{-06}$
Regulation of gene expression, epigenetic	P	3.31%	$8.08E^{-09}$
Tropism	P	1.57%	$6.80E^{-09}$
Nuclear envelope	C	0.78%	$5.72E^{-04}$

[§] F: molecular function; P: biological process; C: cellular component

As observed for the deletions, the nucleotide binding category was the most influenced by insertions. 20.33% of the genes annotated with this GO term were interrupted by an insertion, followed by genes involved in the cellular protein modification process (14.76%).

4.9 SV validation through a PCR-based assay

We validated experimentally through a PCR-based assay both deletions and insertions. A randomly chosen set consisting of 50 deletions and 50 insertions was selected for validation. All these SVs but one were called in four selected varieties for validation: Pinot, Refosco P.R., Tibouren and Rkatsiteli. Positions with an unclear genotype, with unexpected PCR product size or without amplification of any primer pair were discarded. Results of the validations are reported cumulatively for the four varieties in Table 12.

Table 12: PCR validation summary results.

	SV type	
	DELETIONS	**INSERTIONS**
True positives (#)	90	61
True negatives (#)	78	32
False positives (#)	8	1
False negatives (#)	6	7
PPV (%)	91.84	98.36
FDR (%)	8.16	1.64
F1 score (%)	92.78	93.75
Accuracy (%)	92.31	92

Deletions were experimentally validated with an accuracy of 92.31%. Insertions were validated with an accuracy of 92%. F1 score of the two SVs categories was similar: 92.78% for the deletions and 93.75% for the insertions. Compared to the deletions, we confirmed the insertions with a higher PPV (98.36%) and lower FDR (1.64%), since only one false positive was detected. Five homozygous deletions and five homozygous insertions events were confirmed as SV, although the experimentally validated genotype was heterozygous.

The validated deletions varied in length between 1 Kb and 20.7 Kb, with a mean length of approximately 5 Kb. The insertions had a mean size of 5.6 Kb, ranging between 1.1 Kb and 10.6 Kb.

We evaluated the transposable elements composition of the PCR validated SVs (Table 13). In both categories, retrotransposons of class I were more frequent than DNA transposons. Among the former, LINE elements in the deletions accounted for 31.37% of the validated events, followed by Copia elements (17.65%). Concerning the insertions, Gypsy TEs were the most frequent superfamily in the PCR validation (36.73%), followed by LINEs (30.61%). Among the deletions, approximately 27% of the PCR validated SVs resulted unclassified, while all insertions were associated with a transposable element.

Table 13: Transposable elements classification of the PCR validated SVs.

	SV validated	
	DELETIONS	**INSERTIONS**
Class I (%)	62.75	89.8
Class II (%)	7.84	10.2
Unclassified (%)	27.45	0

58% of the validated deletions involved genes sequences. More than 50% of gene's deletions were due to LINE, while approximately 27% were due to unclassified deletions. Introns were the gene regions mainly affected by deletions: only five genes coped with deletions in exon sequences. Among the insertions, 22 genes were interrupted by a SV (approximately 45% of the validated insertions), and five of them were disrupted in the exon regions. LINEs resulted the most common TE superfamily (59%) interacting with genes.

4.10 SV validation based on the comparison to the *de novo* assembly

We furthermore validated the deletions and insertions of six grapevine varieties (Cabernet Franc, Heunisch Weiss, Kishmish Vatkana, Rkatsiteli, Sangiovese and Savagnin Blanc) on the respective *de novo* assemblies (Table 14).

Table 14: Validation statistics of SVs on the *de novo* assemblies of six *V. vinifera* varieties.

Variety	DELETIONS				INSERTIONS			
	Homo[§]	Hetero[¥]	% Homo validated	% Hetero validated	Homo[§]	Hetero[¥]	% Homo reconstructed	% Hetero reconstructed
Cabernet Franc	2048	719	98.73%	68.43%	932	2919	32.40%	6.41%
Heunisch Weiss	2341	821	98.85%	80.88%	760	2803	37.37%	5.17%
Kishmish Vatkana	2795	670	99.18%	79.40%	731	2964	46.92%	5.67%
Rkatsiteli	2145	668	99.39%	81.29%	620	2519	30.65%	4.13%
Sangiovese	2066	884	98.60%	81.11%	781	2728	31.50%	3.70%
Savagning Blanc	1685	755	99.47%	86.23%	483	2616	32.51%	3.78%

[§] Number of homozygous SVs with both breakpoint flanking regions mapped to the same contig;
[¥] number of heterozygous SVs with both breakpoint flanking regions mapped to the same contig.

Concerning the deletions, we were able to validate the vast majority of heterozygous and homozygous deletions. We validated between 98.6% and 99.47% of the homozygous deletions that could be placed on contigs, while the heterozygous deletions were validated in lower percentage, with values ranging between 68.43% and 86.23%. On the other hand, validating the insertions on the assemblies was more complicated. In fact, we were able to reconstruct between 30.65% and 46.92% of the homozygous insertions, while only between 3.7% and 6.41% of the heterozygous insertions. The observed percentage of reconstructed insertions is the lower estimate of true positives. The fraction of non-reconstructed insertions includes: false positives, (heterozygous) insertions not reconstructed by the assembler and insertions whose flanking regions are assembled in two different contigs. Hence, the assemblies could be very useful for the deletion discovery but less efficient for the insertion discovery. On the other hand, the homozygous SVs were validated with more efficiency (both deletions and insertions) than the heterozygous one. It is important to point out that the procedure of *de novo* assembly reconstructs arbitrarily only one haplotype of the variety in regions affected by SV, and thus either version of

heterozygous SVs is included in the consensus sequence. With a random probability of selecting one haplotype, we expected to validate 50% of the heterozygous SV events. While heterozygous deletions were validated with higher percentage, heterozygous insertions were validated with very low percentage. In both situations the assembly software tended to assemble and retain the allele without the TE: this is presumably determined by the difficulty in assembling, extending and scaffolding contigs that contain repeated sequences. This indicated that the assembly tended to be interrupted in correspondence of a transposable element insertion, breaking thus the contig continuity, while for a TE deletion it might reconstruct only part of it.

4.11 Transcriptome analysis

We estimated the contribution of structural variants to gene expression alteration in five grapevine varieties. RNA-Seq data were produced for three different tissues (leaves, berries and tendrils) for Cabernet Franc, Kishmish Vatkana, Rkatsiteli, Sangiovese and Savagnin Blanc. We divided the primary transcripts of the V2.1 genes annotation (Vitulo N et al., 2014) in four categories:

- Genes non-affected by SV (NO SV)
- Genes disrupted only in the exon regions (EXON)
- Genes disrupted only in the intronic portion (INTRON)
- Genes characterised by structural variants influencing more than one gene fraction (MIXED)

The expression profiles of the four above described gene categories were explored in two biological replicates with Cufflinks (Trapnell C et al., 2012). Overall, berry genes showed a lower level of expression in all categories compared to leaves and tendrils (Figure 39).

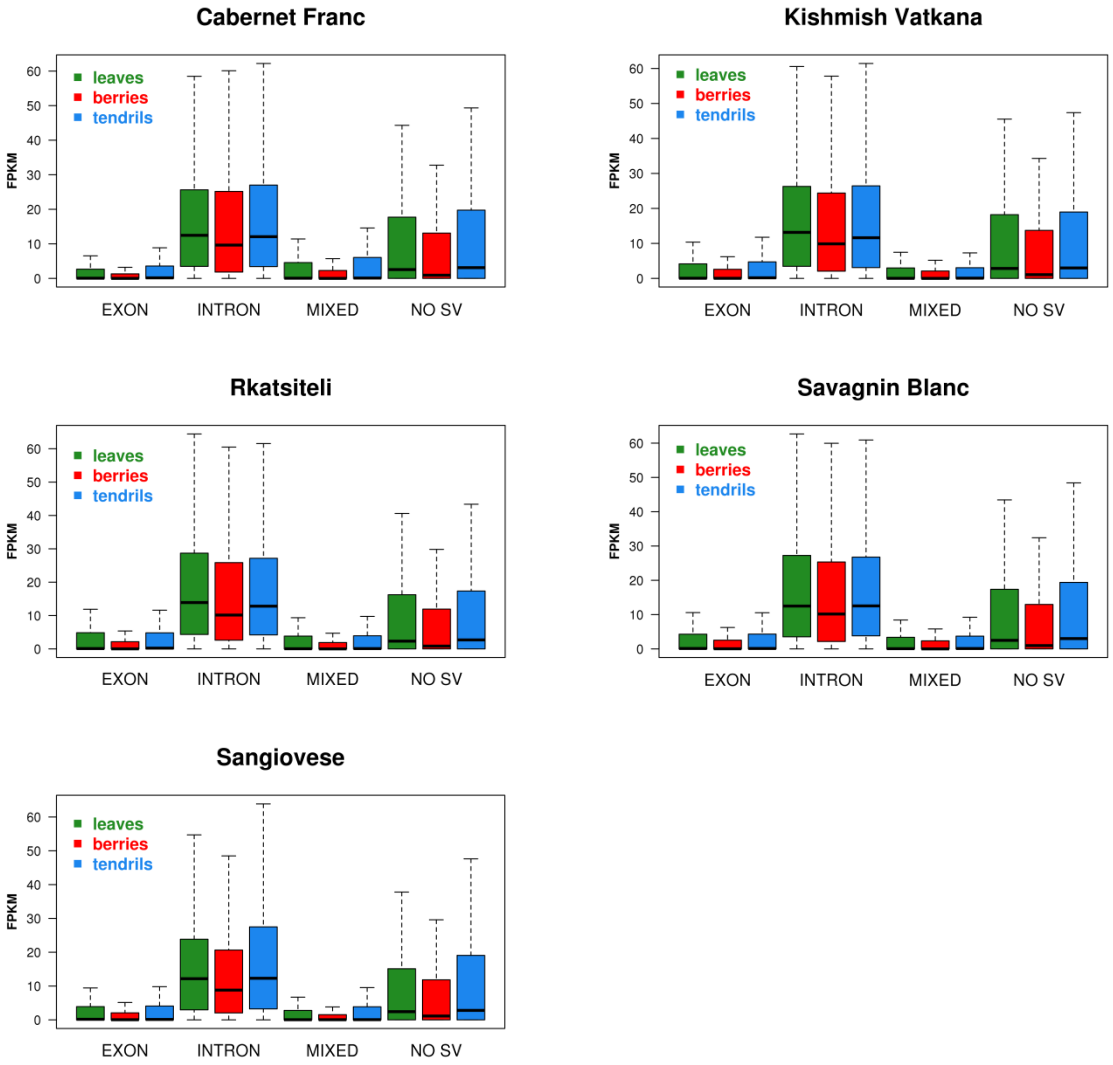


Figure 39: Expression profiles of genes in three tissues of five grapevine varieties. FPKM expression levels of the primary transcript of the V2.1 gene annotation.

In order to statistically compare the expression profiles of the genes disrupted by SVs and the genes non-affected by SVs, the non-parametric Wilcoxon-Mann-Whitney test was used. All expression profiles of the genes interrupted by SVs resulted significantly different from genes without SV. Genes disrupted by SVs in exons or in exon-intron regions showed in all varieties a drastic decrease in the expression profiles compared to the genes not affected by structural variants. The expression levels of the genes carrying a variant only

in the intronic portion increased significantly compared to the genes not affected by SVs in all three tissues. For each tissue and category of genes, we measured the number of genes with no expression (Figure 40).

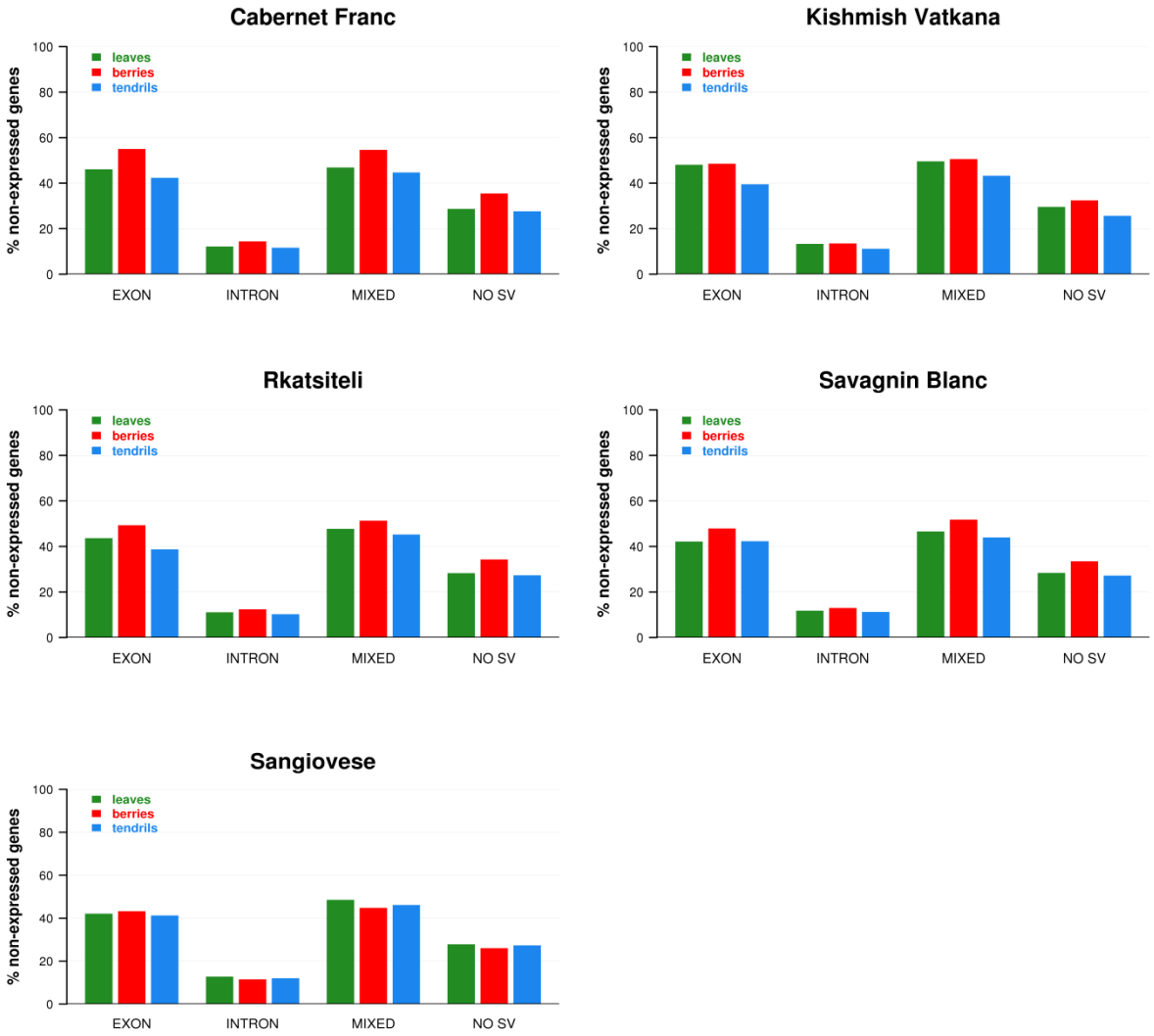


Figure 40: Proportion of non-expressed genes in three tissues of five grapevine varieties. The null hypothesis of no difference in the proportion of non-expressed genes affected by SVs and non-expressed genes not affected by SVs was rejected in all comparisons (chi-square p-value<0.05).

As expected, genes disrupted in the exons or in intron-exons regions showed an increase of the non-expressed genes. In all three tissues, the percentage of genes with no expression increased significantly in the former genes compared to the dataset of genes not influenced by SVs: chi-square test p-values ranged

between $8.23E^{-11}$ (in tendrils of Rkatsiteli) and $3.54E^{-50}$ (in berries of Rkatsiteli). On the other hand, the number of non-expressed genes affected by SV only in the introns decreased drastically compared to the non-disrupted genes. All tissues showed a very similar trend with a significant reduction in number of genes with no expression: chi-square test p-values varied between $5.63E^{-58}$, in tendrils of Kishmish Vatkana, and $1.91E^{-105}$ in berries of Rkatsiteli.

4.12 Grapevine pan-genome

We investigated how the varieties contributed to the coverage of the pan-genome of *V. vinifera*, by assessing the number of new structural variants identified by adding one variety at a time (Figure 41).

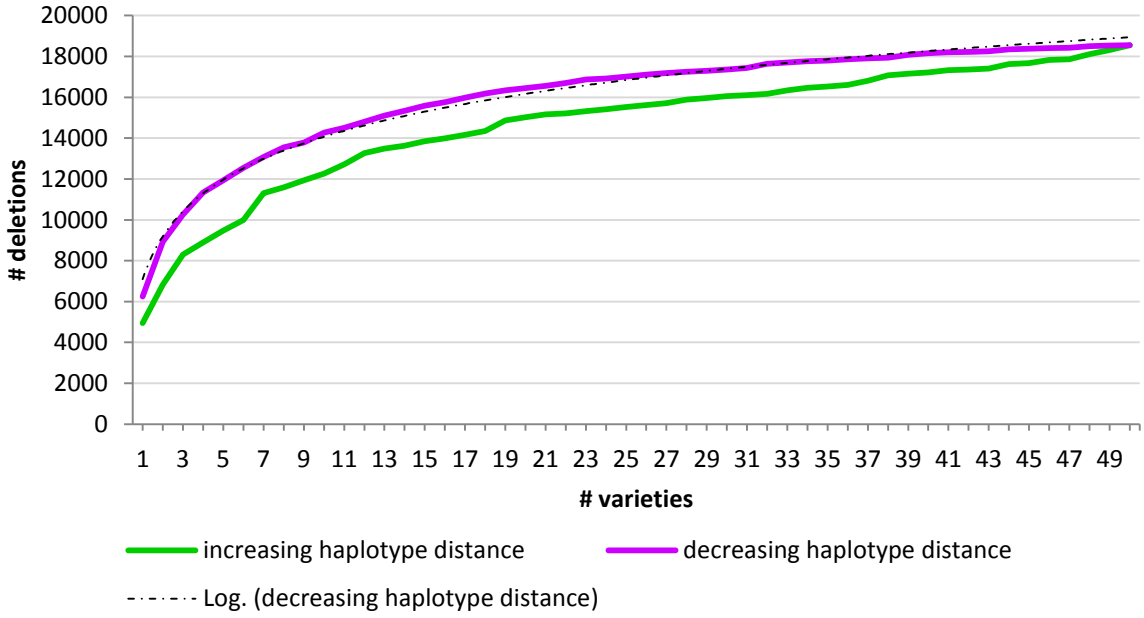


Figure 41: Deletion saturation curve. Number of deletions discovered in the grapevine population by gradually increasing the number of varieties.

As depicted in Figure 41, even after adding all 50 varieties, an increase in the number of deletions was still observed. Based on the genome-wide haplotype

distance measures, it was observed that varieties more distant from the reference genome explained a greater proportion of SVs. Already three varieties explained 50% of the deletions observed in the grapevine population and 32 varieties saturated the curve with 95% of the SVs identified. After a certain number of samples the curve tended to reach a plateau, with slight increases due to the private deletions carried by the single varieties, without ever reaching an asymptote. Our data could be described by a logarithmic curve with the following equation y = 3023 ln(x) + 7100 and an R^2 value of 0.99. By doubling the varieties we estimated an increase of approximately 2,100 private deletions. A very similar trend was observed for the insertions (Figure 42).

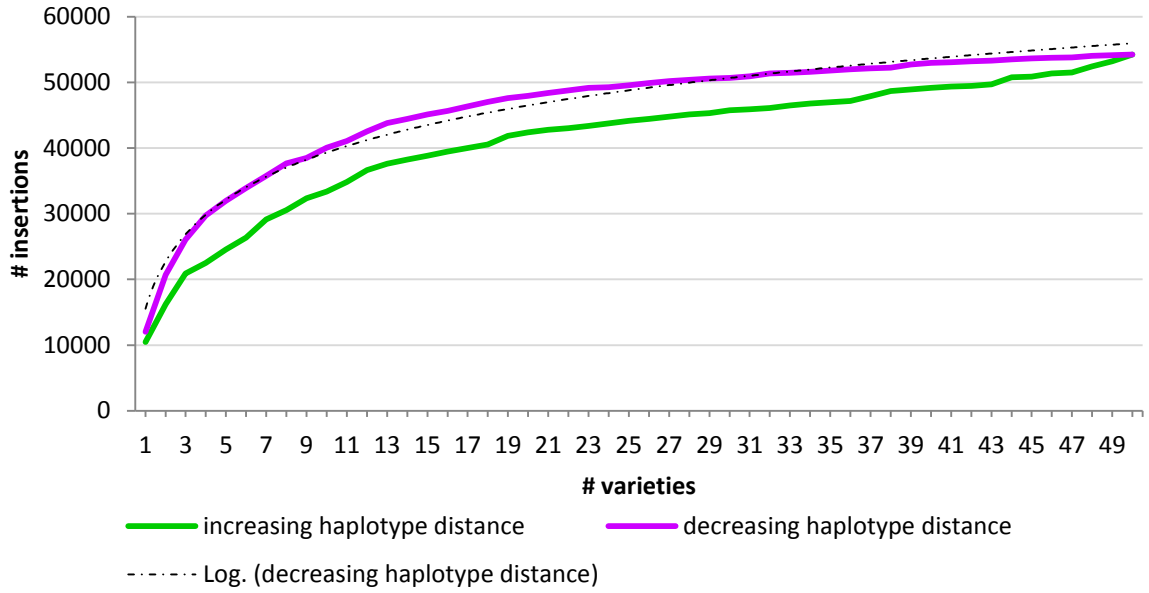


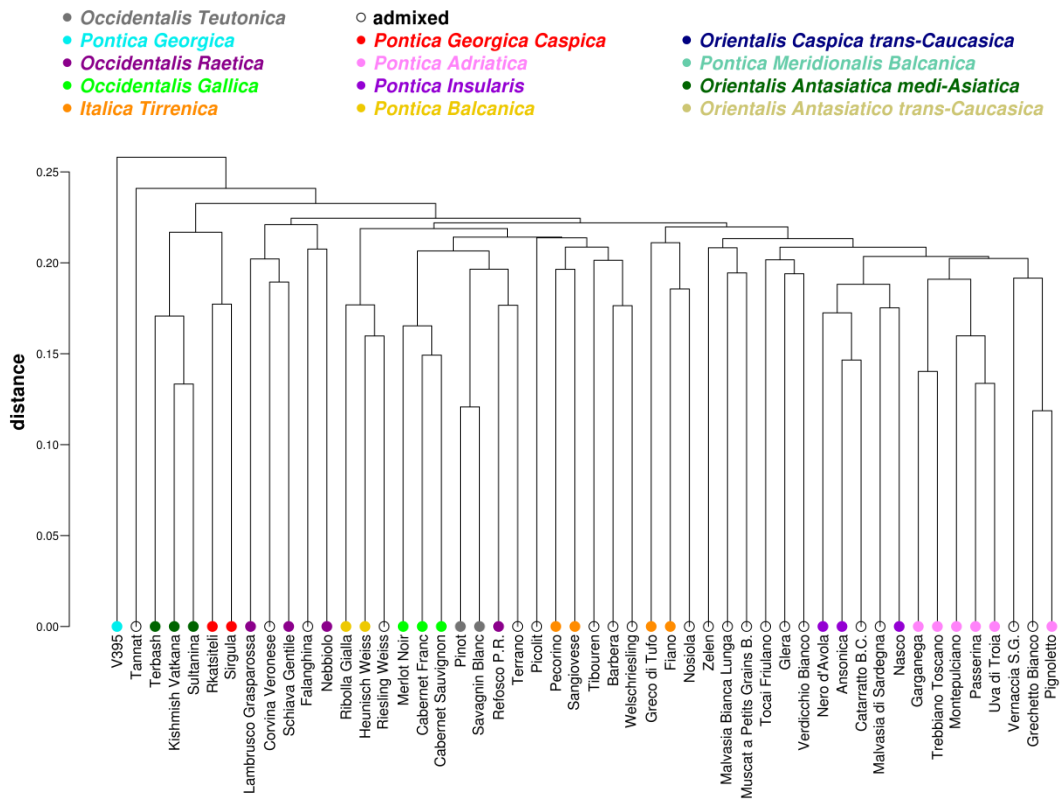
Figure 42: Insertion saturation curve. Number of insertions discovered in the grapevine population by increasing the number of varieties.

By gradually increasing the number of varieties, the curve reached a plateau, with slight increases mediated by private insertions, without ever reaching again an asymptote. The most distant varieties explained the greatest increase of insertions identified. Four varieties explained 50% of the insertions identified in the *V. vinifera* population, while 33 individuals explained 95% of them. As for

the deletions, the pan-genome saturation trend mediated by insertions could be described by a logarithmic curve, with the following equation y = 10332 ln(x) + 15538 and R^2=0.98. We estimated that 100 varieties will contribute to the grapevine dispensable fraction with approximately 7,000 insertions.

4.13 Grapevine population structure based on SV

We calculated the genotypic distances between the 50 grapevine varieties, based on their SV genotypes. The genetic distances were measured pairwise: distance was set to 0, 0.5, or 1, if the varieties shared both haplotypes, one or none, respectively, for each SV. With the R function hclust we performed an UPGMA hierarchical clustering separately for deletions and insertions (Figure 43).



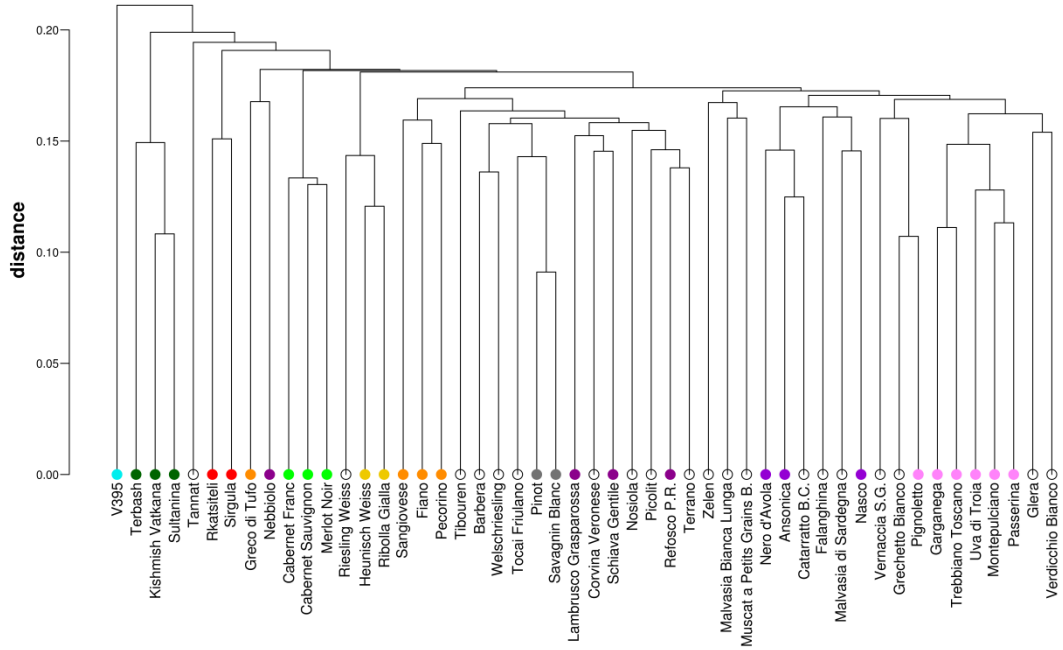


Figure 43: Hierarchical clustering of the grapevine varieties, based on the genotypic distances measured with the deletion genotypes (upper side) and with the insertion genotypes (lower side). The colours of the groups correspond to the colours of ADMIXTURE. Not coloured dots correspond to admixed varieties.

As depicted in Figure 43, the 50 varieties were grouped consistently with the (K=13) ADMIXTURE clusters obtained with SNP data. In both SV categories, V395, the most distant variety, clearly separated from all other varieties. The separation of *Orientalis* varieties (Terbash, Kishimish Vatkana and Sultanina) and *Pontica Georgica* varieties (Rkatisteli and Sirgula) from all other wine grapes was confirmed by the SV-based UPGMA. Furthermore, Pinot Noir and Savagnin Blanc, related by a parent-offspring relationship, were grouped together.

Lastly, in order to validate the genotyping based on structural variants, we compared the matrices of genotypic distances obtained separately for deletions, insertions and SNPs. We performed the Mantel's permutation test (of the APE R package (Paradis E et al., 2004)) with 10,000 random permutations (Mantel N, 1967): all the pairwise comparisons of genetic distance matrices

based on SNPs, insertions, and deletions resulted to be correlated (p-value<0.0001).

5 DISCUSSION

In the present work we performed a genome-wide analysis of the sequence variation for single nucleotide polymorphisms and structural variants on a set of 128 *Vitis vinifera* varieties, representative of the genetic diversity within the species. By means of next-generation sequencing technology we sequenced 122 cultivars, while sequences for six other varieties were retrieved from the Sequence Read Archive (Wheeler DL et al., 2005). In a set of 123 varieties we characterised the population structure with ADMIXTURE (Alexander DH et al., 2009), based on SNP data. According to the ΔK method (Evanno G et al., 2005), the K value of 3 explained the major break in the structure of our population. At K=3 the grapevine population was fragmented in three groups, in accordance with the observations of Negrul AM: *Proles orientalis*, *Proles pontica* and *Proles occidentalis* (Negrul AM, 1946). A great proportion of the varieties (73%) showed admixture between at least two groups. Most of our germplasm was composed of varieties belonging to the *occidentalis* or *orientalis* groups. A few varieties had pure ancestry. Most individuals were admixed. On the other hand, very few individuals presented admixture between the *pontica* and *occidentalis* groups. By increasing gradually the value of K we observed a greater subdivision of the population into smaller groups, in line with known kinships between the varieties. This subdivision was confirmed also by measures and plots of genotypic distance. The PCoA representation showed a clear separation between varieties with an East to West gradient over the first axis, and a separation between cultivated and wild varieties, over the second axis. Georgian wine varieties clearly separated from all others and did not contribute to the ancestry of the Western European varieties. Winegrapes of Southern

Europe have high *orientalis* ancestry, which is phenotypically reflected by many traits shared with table grapes. Based on the PCoA results, we observed that germplasm today cultivated across short geographical ranges, such as the Italian peninsula, includes a very broad range of genetic diversity. These regions may have acted as centres of accumulation of genetic diversity, given their geographical location and historical relevance for inland and maritime trades with the East. To date, several studies explored the grapevine diversity (Myles S et al., 2011; Bacilieri R et al., 2013; Emanuelli F et al., 2013; De Lorenzis G et al., 2015), but never with an accuracy and depth as in the present work.

In a subset of 50 varieties, representative of the genetic diversity of the grapevine population, we investigated at genome-wide level the structural variants (SV) shaping the *V. vinifera* pan-genome. The pan-genome of a species is characterised by a core fraction shared between all the individuals, and a dispensable portion, the absence of which is completely tolerated in a diploid individual (homozygous SV) or compensated for by the presence of homologous DNA sequences (heterozygous DNA) (Tettelin H et al., 2005; Morgante M et al., 2007). Each individual of the species may contribute to the pan-genome with a set of SVs: it will carry private sequences, missing in other varieties, and at the same time will lack other portions of the genome.

Several mutational mechanisms contribute to the generation of structural variants, which extensively affect the genome of any individual. Non-allelic homologous recombination (NAHR) or illegitimate recombination contribute to the genome size contraction, inducing deletion of genome fractions (Devos KM et al., 2002; Stankiewicz P & Lupski JR, 2002), while on the contrary, the movement of transposable elements tends to increase the plant genome size (Vitte C & Panaud O, 2005). On the other hand, unequal homologous recombination contributes to the removal of LTR retrotransposon, leading to the formation of solo-LTRs, as observed, for example, in barley (Shirasu K et al.,

2000) or in rice (Vitte C & Panaud O, 2003; Ma J et al., 2004; Vitte C et al., 2007). The TE movement may be identified both through the detection of deletions and insertions (since the definition of deletion has just an operational meaning). Instead, solo-LTRs may be not distinguished from complete LTR retrotransposons if searching for insertions, while we may be able to identify them, through deletions, only if the fragment is present in the reference genome and absent from the sample (Figure 44).

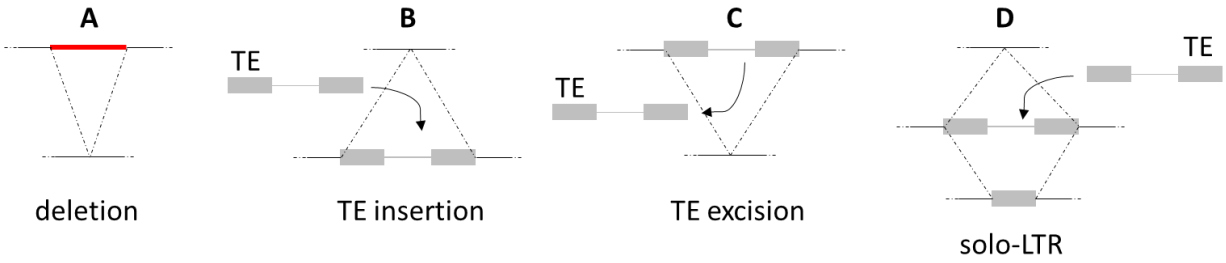


Figure 44: Structural variation formation mechanisms. Examples of SV formation: **A**. removal of sequences mediated by NAHR; **B.** insertion of TEs; **C.** excision of TEs; **D.** generation of solo-LTRs: a LTR transposable element may insert in a genomic region and, at later stage, it may undergo unequal homologous recombination which leads to the formation of solo-LTR in the sequence.

In grapevine, no previous work explored genome-wide structural variation within a population of varieties and with an accuracy and exhaustiveness as in the present work. Di Genova A and colleagues (2014) recently described a catalogue of SVs between two grapevines, the table grape Sultanina and the wine grape PN40024. By combining information of PEM signature and of the *de novo* assembly of Sultanina with the reference genome, the authors discovered a very high number of prevalently small insertions and deletions (ranging in size between 1 and 46.2 Kb and 1 and 9.99 Kb, respectively), in addition to other complex structural variants, such as inversions or inter and intra chromosomal rearrangements. In the present work, we used a similar approach to identify SVs in 50 grape cultivars with the final aim of characterising

the dispensable fraction of grapevine pan-genome. A further work of Giannuzzi and co-workers (2011) revealed, by means of the Whole-genome Shotgun Sequence Detection (WSSD), Segmental Duplications (SDs) within a single genome. SDs are portions of DNA sequence with a length varying between 1 and 200 Kb, sharing a very high sequence identity (greater than 90%) and mapping to different loci in a genome. Based on a combination of molecular cytogenetics and read-depth analysis, the authors observed that recent SDs (with a length >= 10 Kb and a sequence identity > 94%) affected 17.47% of the grapevine genome. These are regions of high instability and are a substrate for non-allelic homologous recombination, which induce the creation of structural variation, contributing thus to the evolution of genes involved in different processes, as for example the NBS-LRR gene family, involved in disease resistance, or genes controlling the berry development and ripening process (Giannuzzi G et al., 2011). The identification of a high number of SD in the grapevine genome confirms the important contribution of SV to shaping genetic architecture of grape.

In the present work we focused on the detection of deletions and insertions, ranging in size between 1 and 25 Kb, by means of the paired-end mapping information of NGS sequenced reads aligned to the *V. vinifera* reference genome (Jaillon O et al., 2007). PEM-based methods do not allow the discovery of all types of SVs, thus, the small SVs discovered in the present work are only a subset of the entire dataset of structural variants that make up the dispensable fraction of the grapevine pan-genome. Copy Number Variants (CNVs), driven by non-homologous recombination, are also a great source of variation both in human and plants. Based on the read depth, the depth of coverage (DOC) method (that we are planning to complete in the near future) may enable the identification of larger deletions or duplications.

While for the detection of deletions several tools are available, the discovery of insertions is more difficult. Available tools are limited to the discovery of small insertions enclosed between two paired reads. Deletions were identified by means of DELLY (Rausch T et al., 2012) and GASV (Sindi S et al., 2009), while for the detection of insertions we employed a pipeline developed by our group, aimed at the detection of insertions resulting from the movement of transposable elements (Sara Pinosio, unpublished results). The performances of the tools in the detection of SVs were assayed by simulating 1000 insertions and deletions in two different species: *Populus trichocarpa* and *Vitis vinifera*. In order to correctly estimate the number of false positives, reads were simulated in *P. trichocarpa* and aligned to the SV simulated reference genome. Compared to the simulated reads, an increase in the number of non-simulated positives was observed with real sequence data, especially in *P. trichocarpa*. Non-simulated positives are the sum of false positives and real heterozygous deletions in the sequenced sample compared to the reference genome.

A total of 18,551 deletions and 54,254 insertions were identified in our grapevine population. Since a great fraction of transposable elements move via the 'copy-and-paste' mechanism, it is important to point out that the definition of deletion used in this thesis has a technical meaning, since the absence of a stretch of reference sequence in a given variety may be actually caused by an insertion of that feature in the reference genome. For each variety we evaluated the extension in terms of amount of base pairs of the structural variants (Figure 45). On average, in each individual, the extension of insertions is approximately twice the extension of deletions: this is in agreement with the consideration that deletions were discovered with an ascertainment bias with respect to a single haplotype represented in the reference genome (assuming that a great fraction of events are in fact insertions), while insertions were discovered from both haplotypes of each heterozygous variety. It is important to point out that the identification of SVs based on a reference genome is not

exhaustive. Portions present in the reference genome and missing in an individual may be identified quite easily. On the other hand, sequences present in an individual, and missing in the reference genome, especially regions of nested transposable elements, are more complicated to be identified. SVs were genotyped in our population as either homozygous or heterozygous. In all varieties we observed a high proportion of base pairs in hemizygous state, i.e. positions characterised by heterozygous SVs. Riesling Weiss was the variety with the highest hemizygous fraction of the genome amounting to 76.9 Mb, while Pignoletto was the variety with the smaller portion. The average length of the hemizygous regions in each variety was estimated to 61.26 Mb; this means that - strikingly - 12.6% of the genome in each variety lacks a counterpart on the homologous chromosome. The extension of hemizygosity estimated in the present work is definitely an underestimate of the real size, since we considered only small SVs, ranging in size between 1 and 25 Kb. *De novo* assemblies obtained for six grapevine varieties are demonstrating high levels of hemizygosity in the grapevine genome, and the discovery of CNVs will enable a better estimate of the total amount of hemizygous DNA.

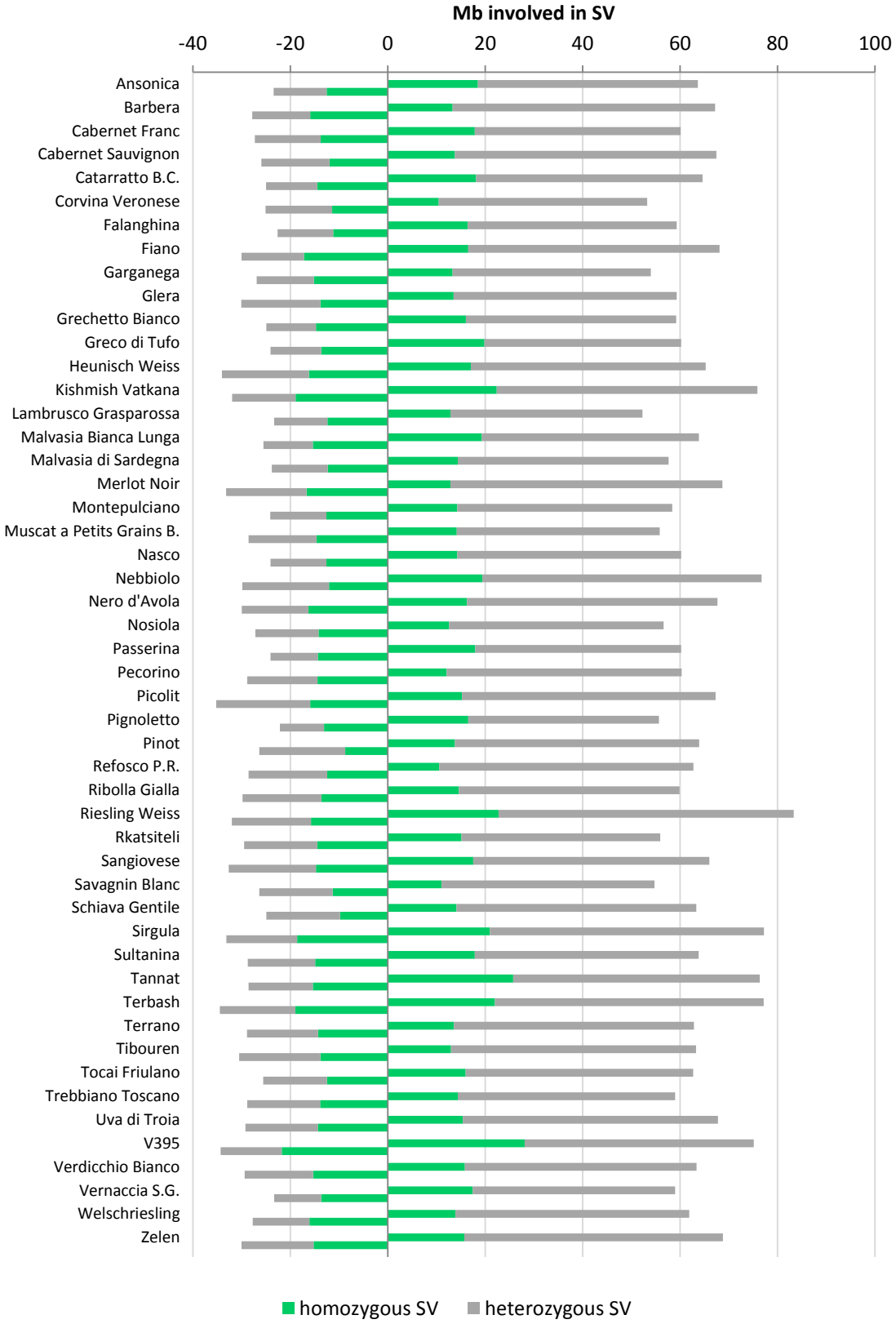


Figure 45: Heterozygous and homozygous SV extension (Mb). Negative values represent the Mb involved in deletions, while positive values the Mb involved in insertions.

We provided a first insight into the pan-genome size considering only the small SVs (ranging between 1 and 25 Kb): the pan-genome is necessarily increased in size with respect to our estimation by the contribution of Copy Number Variants (CNVs) and by SVs missing in the reference genome, that may be identified only by means of the *de novo* assembly. Although SVs longer than 25 Kb and up to hundreds of Kb may rapidly inflate the size of pan-genome, these events are expected to have occurred much less frequently than small SV caused by the movement of TE. A total of 101.94 Mb were deleted in our population, while the cumulative length of insertions was estimated at 329.9 Mb. Based on a reference genome length of approximately 486 Mb and by adding small SVs, the grapevine pan-genome size is now estimated at approximately 816 Mb. Based on our estimates, the core fraction shared between all chromosomes amounted to 384 Mb, while the dispensable fraction accounted for 431 Mb. Grapevine is a diploid species and thus, if we consider the dispensable portion as sequences of DNA missing in at least one individual of the species, the new pan-genome estimates are defined as detailed below. Sixty-five Mb of deletions were observed in homozygous state in at least one of the fifty varieties, and 323.1 Mb of insertions were present at least in heterozygous state in one variety (i.e involved a stretch of dispensable DNA, missing in at least one individual of the species). According to this estimate, the pan-genome has a size of 816 Mb; the core fraction covers 427.8 Mb, while the dispensable portion extends for 388.2 Mb.

Since the production of the first pan-genome for *Streptococcus agalactiae* (Tettelin H et al., 2005), numerous studies have been carried out in bacteria (Donati C et al., 2010; Baddam R et al., 2014; Liu F et al., 2014; Zhou Y et al., 2014). Only recently, pan-genomic data of higher organisms were made available, based on structural variation studies of CNVs and PAVs. Several studies focused on the sequence diversity within genes, which may have an important phenotypic impact. The model plant *Arabidopsis thaliana* was

comprehensively investigated: a first study of Ossowski S and colleagues revealed that 3.4 Mb of sequence were extremely different between two divergent ecotypes (Ossowski S et al., 2008). A wider screening of 18 accession of *A. thaliana* revealed that between 2.1 and 3.7 Mb of the reference sequence was missing in other accessions (Gan X et al., 2011). Rice was the first crop deeply studied. A first study of 2.4 Mb on chromosome four of two rice accessions belonging to different subspecies (ssp *indica* and spp *japonica*) revealed that the homologous regions differed for the presence/absence of 27 genes (Han B & Xue Y, 2003). A CGH study of ssp *japonica* and ssp *indica* enabled the detection of 641 CNVs, amounting to approximately 7.6 Mb (Yu P et al., 2011). A further study of Schatz MC and colleagues on three divergent rice accessions revealed that 92% of genes were present in the core genome, while the remaining 8% were variable for presence/absence (Schatz MC et al., 2014). Compared to our results, in soybean a greater core fraction (80.1%) was estimated from the comparison of seven *Glycine soja* genomes (Li Y et al., 2014). Instead, in maize a smaller core fraction was estimated (Brunner S et al., 2005; Morgante M et al., 2007). A simple comparison of four randomly selected genomic regions, between the reference genome B73 and the line Mo17, revealed that approximately 50% of the DNA was shared, while the remaining DNA was present in either line. A further work of Springer NM and colleagues discovered approximately 2,800 CNVs or PAVs between Mo17 and B73 (Springer NM et al., 2009). On the other hand, the comparison of six inbred lines to the reference enabled the identification of 296 genes missing in one of the six individuals, and of 570 genes missing in B73, but present in one of the six lines (Lai J et al., 2010). Lastly, by means of RNA sequencing, Hirsch CN and colleagues discovered 8,681 high confidence representative transcript assemblies (RTAs), missing in the reference sequence (Hirsch CN et al., 2014).

In the present work, we characterised for the first time the pan-genome of *Vitis vinifera*. We observed that structural variants are a very important source of

genetic variation and contribute to the dispensable portion of the grape pan-genome. As stated by Golicz AA and colleagues, in order to get more realistic estimates of the dispensable fraction, it is very important to choose the most appropriate individuals explaining the greatest diversity within a species (Golicz AA et al., 2015). We observed that by sampling approximately 35 individuals, ordered by decreasing haplotype distance from the reference genotype, we had been able to capture more than 95% of the SVs identified in the entire set of 50 varieties, the remaining 15 individuals contributing to only another 5%. Considering that the set of 50 varieties was itself a subset that maximised the genetic diversity captured by SNPs, we strongly believe to have reached a fairly good saturation of the pan-genome for small SVs.

The vast majority of SVs identified in grapevine resulted from the activity of transposable elements, in accordance with the observations of Kidwell MG and Lisch D (1997). Class I retrotransposable elements move via the 'copy-and-paste' mechanism, leaving a copy of the sequence in the original position, while class II DNA transposons move via the 'cut-and-paste' mechanism, creating at the same time a deletion and an insertion. Class I TEs cover 131.1 Mb of the reference genome and are responsible for 54% and 79% of deletions and insertions, respectively; Class II TEs cover 16.2 Mb of the reference genome and are responsible for 10% and 16% of the deletions and insertions, respectively. Transposable elements had an important role in the eukaryotic genome evolution (Bennetzen JL, 2000; Biémont C & Vieira C, 2006; Feschotte C & Pritham EJ, 2007). The relative recent activity of TEs, usually with bursts of activity associated with stress (Grandbastien M, 1998), hybridization or polyploidy (Voytas DF & Naylor GJP, 1998; Liu B & Wendel JF, 2000), induced high levels of structural variation in several different angiosperms. Very recent peaks of activity (possibly related to plant domestication) were observed for example in soybean (Wawrzynski A et al., 2008) and also in grapevine (Moisy C et al., 2008). Plant genomes underwent genome amplification, induced by

retrotransposition, and contraction, through either illegitimate or homologous recombination (Vitte C & Panaud O, 2005). Following an approach already used in maize (SanMiguel P et al., 1998; Brunner S et al., 2005; Baucom RS et al., 2009), *Arabidopsis* (Devos KM et al., 2002), *Medicago* (Wang H & Liu J-S, 2008), rice (Ma J & Bennetzen JL, 2004; Vitte C et al., 2007; Hurwitz BL et al., 2010) and melon (Garcia-Mas J et al., 2012), we estimated the timing of TE insertional activity. We observed that LTR retrotransposons involved in structural variation moved very recently, with 69% of the surveyed LTR transposable elements with an insertion age of less than one million of years, and with a significantly different trend compared to the insertion time of LTR-retros shared in the grapevine population. The insertion time profile of LTR retrotransposons involved in SV was very similar to the age distribution of complete LTR-retros observed in expanded regions of *O. sativa* (compared to *O. nivara*) by Hurwitz BL and colleagues (2010). On the other hand, the insertion age of grapevine shared LTR retrotransposons showed a similar profile compared to the age of LTR-retros identified in the melon genome (Garcia-Mas J et al., 2012). Finally, Brunner S and co-workers (2005) observed a very similar pattern comparing shared and non-shared LTR retrotransposons in two maize inbred lines. The very recent activity of transposable elements of different classes seems therefore to be largely responsible for the high number of structural variants observed in grapevine, that correspond to insertions that have not had yet either gone to fixation in the population or have not been lost from it.

Furthermore, we observed that structural variants affected the gene space. Deletions affected 5,679 genes, while insertions occurred in 7,828 genes. A total of 2,608 genes resulted affected both by deletions and insertions. The number of genes influenced by both SV categories was significantly lower than what expected by chance. We simulated 100 times 18,551 and 54,254 random deletions and insertions, respectively, in the reference genome and measured the number of genes affected by SV. Simulations predicted that on average

9,217.58 genes should be influenced by deletions, with a standard deviation of 60.15 (confidence intervals 9,099.69 and 9,335.47), while 17,037.59 genes should be affected by insertions, with a standard deviation of 114.41 (confidence intervals 16,813.35 and 17,261.83): a number of genes significantly higher than what we actually observed in our grapevine population. Thus, the genome of each variety tolerates a high number of SVs affecting genes. For example in Merlot Noir we observed a total of 3,734 genes affected by SV. Out of these, 2,293 genes were affected by deletions. Homozygous deletions are tolerated by Merlot Noir, since the variety does not show phenotypic abnormalities. Concerning the heterozygous deletions, we are not able to predict their effect, since the copy of the gene affected by SV is present in the genome along with a wild-type copy on the homologous chromosome. If heterozygous deletions occurred in essential genes, the wild-type copy may have a dominant effect and is able to provide the functional gene product. On the other hand, homozygous deletions that cause a complete loss of the gene function may have a strong negative impact.

In the gene space, structural variants affected mostly non-coding gene regions. Both deletions and insertions affected mainly introns, and only with lower incidence exonic or UTR regions. Genes affected by SVs in the latter regions showed a strong reduction in the expression levels compared to the non-affected genes, while surprisingly, genes with SVs in introns showed higher than average levels of expression. In line with these observations, we discovered a significant increase in the number of non-expressed genes disrupted by SVs in coding regions, while, on the contrary, a significant reduction of non-expressed genes affected by SVs in introns. LINE transposable elements affected prevalently introns of genes that are transcribed. Jaillon O and colleagues (2007) observed similar results: 75% of the transposable elements in introns was represented by LINEs which contributed to the longer size of introns in grapevine, compared to other plant species. As suggested by

Jiang K and colleague, the intron size expansion may be related to the evolutionary history of grapevine domestication (Jiang K & Goertzen LR, 2011). On the other hand, the excess of non-expressed genes showing disruptions of exons might be due to the fact that such SVs might have negative effects and might have undergone negative selection. Alternatively, we cannot exclude that gene models carrying SVs in exons or UTR regions may have been erroneously predicted or may correspond to pseudogenes without evidence of transcription. A functional annotation of the genes disrupted by SVs revealed a total of nine and eleven Gene Ontology categories preferentially affected by deletions and insertions, respectively. The *nucleotide binding* GO term was the gene category mostly influenced by SV. Both insertions and deletions affected a high number of genes with nucleotide binding functions, most of which are genes involved in plant disease resistance. Furthermore, genes related to post-translational protein modification or cytosolic activity were influenced by SV with lower incidence. Proteins with a Nucleotide-Binding Site (NBS) and Leucine-Rich Repeat (LRR) are mainly involved in plant innate immunity. The NBS-LRR class of genes account for hundreds of copies in plant species (Ellis J et al., 2000). Since pathogen populations evolve quickly, genes involved in disease resistance need to generate novel variation in the host plant genomes. Thus, these genes may have high levels of inter- and intraspecific variation (McHale L et al., 2006), induced by different mechanism as unequal crossing-over, sequence exchange or gene conversion and usually reside in regions with high SV (Baumgarten A et al., 2003; Kuang H et al., 2004; Mondragon-Palomino M & Gaut BS, 2005).

Advances in next-generation sequencing improved considerably the identification of structural variation. Both CNVs and TEs have been widely associated with human diseases, while in plants few efforts have been made (Saxena RK et al., 2014). Structural variants have been investigated in different species: in maize (Springer NM et al., 2009; Beló A et al., 2010), in *Arabidopsis* (DeBolt S, 2010), in soybean (McHale LK et al., 2012), in rice (Xu X et al., 2012),

in sorghum (Zheng L-Y et al., 2011), in barley (Muñoz-Amatriaín M et al., 2013), in melon (Sanseverino W et al., 2015) and in grapevine (Giannuzzi G et al., 2011; Di Genova A et al., 2014). SVs in human have been linked to several diseases including schizophrenia (The International Schizophrenia Consortium, 2008) and autism (Marshall CR et al., 2008). Also in plants several studies have been carried out. In grapevine, Kobayashi S and colleagues (2004) discovered that the insertion of Gret1, in the promoter region of the *MybA1* transcription factor gene, inactivated (in white varieties) the anthocyanin biosynthesis pathway. On the other hand, the insertion of a class II transposable element in the *TFL1A* promoter resulted in an upregulation of the gene expression, affecting drastically the size and the branching pattern of the Carignan fruit cluster (Fernandez L et al., 2010). Furthermore, McHale LK and colleagues observed a significant enrichment of CNV in known R gene clusters, genes involved in disease resistance (McHale LK et al., 2012). Ong-Abdullah M and colleagues recently observed in oil palm, that the loss of DNA methylation of the Karma LINE transposon (Komatsu M et al., 2003), within an important flowering transcription factor gene, was responsible for the mantled phenotype, resulting in a drastically decrease of oil yields (Ong-Abdullah M et al., 2015). In the near future, it would be of great interest to determine the phenotypic effects of other SVs.

Finally, in order to validate the genotyping based on structural variants, we correlated, by means of the Mantel test (Mantel N, 1967), the matrices of genetic distances obtained from SV genotypes and SNP genotypes. We observed that the genetic relationships described by structural variants were consistent with those described by SNPs (p-value<0.0001), despite the likely more recent origin of the SVs in comparison to SNPs.

In conclusion, based on the results of the present work, we observed that SVs extensively affect the grapevine pan-genome. Not only intergenic regions were

involved, but also gene space resulted affected. We observed that structural variants affecting exonic regions are associated with reduced gene expression, while SVs in introns, caused predominantly by LINEs, are associated to an increase of expression levels, compared to genes not affected by SV. Finally, concerning the methylation context, SVs mediated by the movement of transposable elements were generally associated with high levels of DNA methylation, which is aimed at suppressing further TE activity (Vaughn MW et al., 2007; Eichten SR et al., 2011; Schmitz RJ et al., 2013).

6 REFERENCES

Abajian C (1994). **Sputnik**.

Alexander DH, Novembre J and Lange K (2009). **Fast model-based estimation of ancestry in unrelated individuals.** *Genome Research*, 19 (9), 1655–64.

Alkan C, Coe BP and Eichler EE (2011). **Genome structural variation discovery and genotyping.** *Nature Reviews Genetics*, 12 (5), 363–376.

Altschul SF, Gish W, Miller W, Myers EW and Lipman DJ (1990). **Basic local alignment search tool.** *Journal of Molecular Biology*, 215 (3), 403–10.

Arroyo-Garcia R, Ruiz-Garcia L, Bolling L, Ocete R, Lopez MA, Arnold C, Ergul A, Soylemezoglu G, Uzun HI, Cabello F, Ibanez J, Aradhya MK, Atanassov A, Atanassov I, Balint S, Cenis JL, Costantini L, Goris-Lavets S, Grando MS, Klein BY, McGovern PE, Merdinoglu D, Pejic I, Pelsy F, Primikirios N, Risovannaya V, Roubelakis-Angelakis K a, Snoussi H, Sotiri P, Tamhankar S, This P, Troshin L, Malpica JM, Lefort F and Martinez-Zapater JM (2006). **Multiple origins of cultivated grapevine (Vitis vinifera L. ssp. sativa) based on chloroplast DNA polymorphisms**. *Molecular Ecology*, 15 (12), 3707–3714.

Bacilieri R, Lacombe T, Le Cunff L, Di Vecchi-Staraz M, Laucou V, Genna B, Péros J-P, This P and Boursiquot J-M (2013). **Genetic structure in cultivated grapevines is linked to geography and human selection.** *BMC Plant Biology*, 13 (1), 25.

Baddam R, Kumar N, Shaik S, Lankapalli AK and Ahmed N (2014). **Genome dynamics and evolution of Salmonella Typhi strains from the typhoid-endemic zones.** *Scientific Reports*, 4, 7457.

Baucom RS, Estill JC, Chaparro C, Upshaw N, Jogi A, Deragon J-M, Westerman RP, SanMiguel PJ and Bennetzen JL (2009). **Exceptional Diversity, Non-Random Distribution, and Rapid Evolution of Retroelements in the B73 Maize Genome**. *PLoS Genetics*, 5 (11), e1000732.

Baumgarten A, Cannon S, Spangler R and May G (2003). **Genome-Level Evolution of Resistance Genes in Arabidopsis thaliana**. *Genetics*, 165 (1), 309–319.

Beló A, Beatty MK, Hondred D, Fengler KA, Li B and Rafalski A (2010). **Allelic genome structural variations in maize detected by array comparative genome hybridization.** *TAG. Theoretical and Applied Genetics. Theoretische Und Angewandte Genetik*, 120 (2), 355–67.

Benjamini Y and Hochberg Y (1995). **Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing**. *Journal of the Royal Statistical Society B*, 57 (1), 289–300.

Bennetzen JL (2005). **Transposable elements, gene creation and genome rearrangement in flowering plants**. *Current Opinion in Genetics & Development*, 15 (6), 621–627.

Bennetzen JL (2000). **Transposable element contributions to plant gene and genome evolution**. *Plant Molecular Biology*, 42 (1), 251–269.

Benson G (1999). **Tandem repeats finder: a program to analyze DNA sequences.** *Nucleic Acids Research*, 27 (2), 573–80.

Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG, Hall KP, Evers DJ, Barnes CL, Bignell HR, Boutell JM, Bryant J, Carter RJ, Keira Cheetham R, Cox AJ, Ellis DJ, Flatbush MR, Gormley NA, Humphray SJ, Irving LJ, Karbelashvili MS, Kirk SM, Li H, Liu X, Maisinger KS, Murray LJ, Obradovic B, Ost T, Parkinson ML, Pratt MR, Rasolonjatovo IMJ, Reed MT, Rigatti R, Rodighiero C, Ross MT, Sabot A, Sankar S V, Scally A, Schroth GP, Smith ME, Smith VP, Spiridou A, Torrance PE, Tzonev SS, Vermaas EH, Walter K, Wu X, Zhang L, Alam MD, Anastasi C, Aniebo IC, Bailey DMD, Bancarz IR, Banerjee S, Barbour SG, Baybayan PA, Benoit VA, Benson KF, Bevis C, Black PJ, Boodhun A, Brennan JS, Bridgham JA, Brown RC, Brown AA, Buermann DH, Bundu AA, Burrows JC, Carter NP, Castillo N, Chiara E Catenazzi M, Chang S, Neil Cooley R, Crake NR, Dada OO, Diakoumakos KD, Dominguez-Fernandez B, Earnshaw DJ, Egbujor UC, Elmore DW, Etchin SS, Ewan MR, Fedurco M, Fraser LJ, Fuentes Fajardo K V, Scott Furey W, George D, Gietzen KJ, Goddard CP, Golda GS, Granieri PA, Green DE, Gustafson DL, Hansen NF, Harnish K, Haudenschild CD, Heyer NI, Hims MM, Ho JT, Horgan AM, Hoschler K, Hurwitz S, Ivanov D V, Johnson MQ, James T, Huw Jones TA, Kang G-D, Kerelska TH, Kersey AD, Khrebtukova I, Kindwall AP, Kingsbury Z, Kokko-Gonzales PI, Kumar A, Laurent MA, Lawley CT, Lee SE, Lee X, Liao AK, Loch JA, Lok M, Luo S, Mammen RM, Martin JW, McCauley PG, McNitt P, Mehta P, Moon KW, Mullens JW, Newington T, Ning Z, Ling Ng B, Novo SM, O'Neill MJ, Osborne MA, Osnowski A, Ostadan O, Paraschos LL, Pickering L, Pike

AC, Pike AC, Chris Pinkard D, Pliskin DP, Podhasky J, Quijano VJ, Raczy C, Rae VH, Rawlings SR, Chiva Rodriguez A, Roe PM, Rogers J, Rogert Bacigalupo MC, Romanov N, Romieu A, Roth RK, Rourke NJ, Ruediger ST, Rusman E, Sanches-Kuiper RM, Schenker MR, Seoane JM, Shaw RJ, Shiver MK, Short SW, Sizto NL, Sluis JP, Smith MA, Ernest Sohna Sohna J, Spence EJ, Stevens K, Sutton N, Szajkowski L, Tregidgo CL, Turcatti G, Vandevondele S, Verhovsky Y, Virk SM, Wakelin S, Walcott GC, Wang J, Worsley GJ, Yan J, Yau L, Zuerlein M, Rogers J, Mullikin JC, Hurles ME, McCooke NJ, West JS, Oaks FL, Lundberg PL, Klenerman D, Durbin R and Smith AJ (2008). **Accurate whole human genome sequencing using reversible terminator chemistry.** *Nature*, 456 (7218), 53–9.

Biémont C and Vieira C (2006). **Junk DNA as an evolutionary force**, 443 (October), 521–524.

Browning BL and Browning SR (2009). **A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals.** *American Journal of Human Genetics*, 84 (2), 210–23.

Brunner S, Fengler K, Morgante M, Tingey S and Rafalski A (2005). **Evolution of DNA sequence nonhomologies among maize inbreds.** *The Plant Cell*, 17 (2), 343–360.

Campbell PJ, Stephens PJ, Pleasance ED, O'Meara S, Li H, Santarius T, Stebbings L a, Leroy C, Edkins S, Hardy C, Teague JW, Menzies A, Goodhead I, Turner DJ, Clee CM, Quail M a, Cox A, Brown C, Durbin R, Hurles ME, Edwards P a W, Bignell GR, Stratton MR and Futreal PA (2008). **Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing.** *Nature Genetics*, 40 (6), 722–729.

Conesa A, Götz S, García-Gómez JM, Terol J, Talón M and Robles M (2005). **Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research.** *Bioinformatics*, 21 (18), 3674–6.

Conrad DF, Pinto D, Redon R, Feuk L, Gokcumen O, Zhang Y, Aerts J, Andrews TD, Barnes C, Campbell P, Fitzgerald T, Hu M, Ihm CH, Kristiansson K, Macarthur DG, Macdonald JR, Onyiah I, Pang AWC, Robson S, Stirrups K, Valsesia A, Walter K, Wei J, Tyler-Smith C, Carter NP, Lee C, Scherer SW and Hurles ME (2010). **Origins and functional impact of copy number variation in the human genome.** *Nature*, 464 (7289), 704–12.

De Lorenzis G, Chipashvili R, Failla O and Maghradze D (2015). **Study of genetic variability in Vitis vinifera L. germplasm by high-throughput Vitis18kSNP array: the case of Georgian genetic resources.** *BMC Plant Biology*, 15 (1), 154.

DeBolt S (2010). **Copy number variation shapes genome diversity in Arabidopsis over immediate family generational scales.** *Genome Biology and Evolution*, 2, 441–53.

Del Fabbro C, Scalabrin S, Morgante M and Giorgi FM (2013). **An extensive evaluation of read trimming effects on Illumina NGS data analysis.** *PloS One*, 8 (12), e85024.

DePristo MA, Banks E, Poplin R, Garimella K V, Maguire JR, Hartl C, Philippakis AA, del Angel G, Rivas MA, Hanna M, McKenna A, Fennell TJ, Kernytsky AM, Sivachenko AY, Cibulskis K, Gabriel SB, Altshuler D and Daly MJ (2011). **A framework for variation discovery and genotyping using next-generation DNA sequencing data.** *Nature Genetics*, 43 (5), 491–8.

Devos KM, Brown JKM and Bennetzen JL (2002). **Genome size reduction through illegitimate recombination counteracts genome expansion in Arabidopsis.** *Genome Research*, 12 (7), 1075–9.

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

Donati C, Hiller NL, Tettelin H, Muzzi A, Croucher NJ, Angiuoli S V, Oggioni M, Dunning Hotopp JC, Hu FZ, Riley DR, Covacci A, Mitchell TJ, Bentley SD, Kilian M, Ehrlich GD, Rappuoli R, Moxon ER and Masignani V (2010). **Structure and dynamics of the pan-genome of Streptococcus pneumoniae and closely related species.** *Genome Biology*, 11 (10), R107.

Edwards D and Batley J (2010). **Plant genome sequencing: applications for crop improvement**. *Plant Biotechnology Journal*, 8 (1), 2–9.

Eichler EE and Sankoff D (2003). **Structural Dynamics of Eukaryotic Chromosome Evolution**. *Science*, 301 (5634), 793–797.

Eichten SR, Swanson-Wagner RA, Schnable JC, Waters AJ, Hermanson PJ, Liu S, Yeh C-T, Jia Y, Gendler K, Freeling M, Schnable PS, Vaughn MW and Springer NM (2011). **Heritable epigenetic variation among maize inbreds.** *PLoS Genetics*, 7 (11), e1002372.

El Baidouri M and Panaud O (2013). **Comparative genomic paleontology across plant kingdom reveals the dynamics of TE-driven genome evolution.** *Genome Biology and Evolution*, 5 (5), 954–65.

Ellis J, Dodds P and Pryor T (2000). **Structure, function and evolution of plant disease resistance genes.** *Current Opinion in Plant Biology*, 3 (4), 278–284.

Emanuelli F, Lorenzi S, Grzeskowiak L, Catalano V, Stefanini M, Troggio M, Myles S, Martinez-Zapater JM, Zyprian E, Moreira FM and Grando MS (2013). **Genetic diversity and population structure assessed by SSR and SNP markers in a large germplasm collection of grape.** *BMC Plant Biology*, 13 (1), 39.

Evanno G, Regnaut S and Goudet J (2005). **Detecting the number of clusters of individuals using the software structure: a simulation study**. *Molecular Ecology*, 14 (8), 2611–2620.

Fernandez L, Torregrosa L, Segura V, Bouquet A and Martinez-Zapater JM (2010). **Transposon-induced gene activation as a mechanism generating cluster shape somatic variation in grapevine**. *Plant Journal*, 61 (4), 545–557.

Feschotte C, Jiang N and Wessler SR (2002). **Plant Transposable Elements: Where Genetics Meets Genomics**. *Nature Reviews Genetics*, 3 (5), 329–341.

Feschotte C and Pritham EJ (2007). **DNA transposons and the evolution of eukaryotic genomes.** *Annual Review of Genetics*, 41, 331–68.

Feuk L (2006). **Structural variants: changing the landscape of chromosomes and design of disease studies**. *Human Molecular Genetics*, 15 (90001), R57–R66.

Flutre T, Duprat E, Feuillet C and Quesneville H (2011). **Considering transposable element diversification in de novo annotation approaches.** *PloS One*, 6 (1), e16526.

Fournier-Level A, Lacombe T, Le Cunff L, Boursiquot J-M and This P (2010). **Evolution of the VvMybA gene family, the major determinant of berry colour in cultivated grapevine (Vitis vinifera L.).** *Heredity*, 104 (4), 351–62.

Freeman JL, Perry GH, Feuk L, Redon R, McCarroll SA, Altshuler DM, Aburatani H, Jones KW, Tyler-Smith C, Hurles ME, Carter NP, Scherer SW and Lee C (2006). **Copy number variation: new insights in genome diversity.** *Genome Research*, 16 (8), 949–61.

Gan X, Stegle O, Behr J, Steffen JG, Drewe P, Hildebrand KL, Lyngsoe R, Schultheiss SJ, Osborne EJ, Sreedharan VT, Kahles A, Bohnert R, Jean G, Derwent P, Kersey P, Belfield EJ, Harberd NP, Kemen E, Toomajian C, Kover PX, Clark RM, Rätsch G and Mott R (2011). **Multiple reference genomes and transcriptomes for Arabidopsis thaliana.** *Nature*, 477 (7365), 419–23.

Garcia-Mas J, Benjak A, Sanseverino W, Bourgeois M, Mir G, González VM, Hénaff E, Câmara F, Cozzuto L, Lowy E, Alioto T, Capella-Gutiérrez S, Blanca J, Cañizares J, Ziarsolo P, Gonzalez-Ibeas D, Rodríguez-Moreno L, Droege M, Du L, Alvarez-Tejado M, Lorente-Galdos B, Melé M, Yang L, Weng Y, Navarro A, Marques-Bonet T, Aranda MA, Nuez F, Picó B, Gabaldón T, Roma G, Guigó R, Casacuberta JM, Arús P and Puigdomènech P (2012). **The genome of melon (Cucumis melo L.).** *Proceedings of the National Academy of Sciences of the United States of America*, 109 (29), 11872–7.

Gaut BS, Morton BR, McCaig BC and Clegg MT (1996). **Substitution rate comparisons between grasses and palms: synonymous rate differences at the nuclear gene Adh parallel rate differences at the plastid gene rbcL.** *Proceedings of the National Academy of Sciences of the United States of America*, 93 (September), 10274–10279.

Giannuzzi G, D'Addabbo P, Gasparro M, Martinelli M, Carelli FN, Antonacci D and Ventura M (2011). **Analysis of high-identity segmental duplications in the grapevine genome.** *BMC Genomics*, 12, 436.

Gnerre S, Maccallum I, Przybylski D, Ribeiro FJ, Burton JN, Walker BJ, Sharpe T, Hall G, Shea TP, Sykes S, Berlin AM, Aird D, Costello M, Daza R, Williams L, Nicol R, Gnirke A, Nusbaum C, Lander ES and Jaffe DB (2011). **High-quality draft assemblies of mammalian genomes from massively parallel sequence data.** *Proceedings of the National Academy of Sciences of the United States of America*, 108 (4), 1513–8.

Golicz AA, Batley J and Edwards D (2015). **Towards plant pangenomics**. *Plant Biotechnology Journal*, 1–7.

Grandbastien M (1998). **Activation of plant retrotransposons under stress conditions**. *Trends in Plant Science*, 3 (5), 181–187.

Grassi F, Labra M, Imazio S, Spada A, Sgorbati S, Scienza A and Sala F (2003). **Evidence of a secondary grapevine domestication centre detected by SSR analysis**. *TAG Theoretical and Applied Genetics*, 107 (7), 1315–1320.

Han B and Xue Y (2003). **Genome-wide intraspecific DNA-sequence variations in rice**. *Current Opinion in Plant Biology*, 6 (2), 134–138.

Harris TD, Buzby PR, Babcock H, Beer E, Bowers J, Braslavsky I, Causey M, Colonell J, Dimeo J, Efcavitch JW, Giladi E, Gill J, Healy J, Jarosz M, Lapen D, Moulton K, Quake SR, Steinmann K, Thayer E, Tyurina A, Ward R, Weiss H and Xie Z (2008). **Single-molecule DNA sequencing of a viral genome.** *Science*, 320 (5872), 106–9.

Hirsch CN, Foerster JM, Johnson JM, Sekhon RS, Muttoni G, Vaillancourt B, Peñagaricano F, Lindquist E, Pedraza MA, Barry K, de Leon N, Kaeppler SM and Buell CR (2014). **Insights into the maize pan-genome and pan-transcriptome.** *The Plant Cell*, 26 (1), 121–35.

Horner DS, Pavesi G, Castrignano T, De Meo PD, Liuni S, Sammeth M, Picardi E and Pesole G (2010). **Bioinformatics approaches for genomics and post genomics applications of next-generation sequencing**. *Briefings in Bioinformatics*, 11 (2), 181–197.

Huang X and Madan A (1999). **CAP3: A DNA Sequence Assembly Program**. *Genome Research*, 9 (9), 868–877.

Hurles ME, Dermitzakis ET and Tyler-Smith C (2008). **The functional impact of structural variation in humans.** *Trends in Genetics : TIG*, 24 (5), 238–45.

Hurwitz BL, Kudrna D, Yu Y, Sebastian A, Zuccolo A, Jackson S a., Ware D, Wing R a. and Stein L (2010). **Rice structural variation: A comparative analysis of structural variation between rice and three of its closest relatives in the genus Oryza**. *Plant Journal*, 63 (6), 990–1003.

Iafrate AJ, Feuk L, Rivera MN, Listewnik ML, Donahoe PK, Qi Y, Scherer SW and Lee C (2004). **Detection of large-scale variation in the human genome.** *Nature Genetics*, 36 (9), 949–51.

Jaillon O, Aury J-M, 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 A-F, Weissenbach J, Quétier F and Wincker P (2007). **The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla.** *Nature*, 449 (7161), 463–7.

Jiang K and Goertzen LR (2011). **Spliceosomal intron size expansion in domesticated grapevine (Vitis vinifera).** *BMC Research Notes*, 4, 52.

Jurka J, Kapitonov V V, Pavlicek A, Klonowski P, Kohany O and Walichiewicz J (2005). **Repbase Update, a database of eukaryotic repetitive elements.** *Cytogenetic and Genome Research*, 110 (1-4), 462–7.

Kallioniemi A, Kallioniemi O, Sudar D, Rutovitz D, Gray J, Waldman F and Pinkel D (1992). **Comparative genomic hybridization for molecular cytogenetic analysis of solid tumors**. *Science*, 258 (5083), 818–821.

Kejnovsky E, Leitch IJ and Leitch AR (2009). **Contrasting evolutionary dynamics between angiosperm and mammalian genomes**. *Trends in Ecology & Evolution*, 24, 572–582.

Kidwell MG and Lisch D (1997). **Transposable elements as sources of variation in animals and plants**. *Proceedings of the National Academy of Sciences*, 94 (15), 7704–7711.

Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R and Salzberg SL (2013). **TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions.** *Genome Biology*, 14 (4), R36.

Kimura M (1980). **A Simple Method for Estimating Evolutionary Rates of Base Substitutions Through Comparative Studies of Nucleotide Sequences**. *Journal of Molecular Evolution*, 16, 111–120.

Kobayashi S, Goto-Yamamoto N and Hirochika H (2004). **Retrotransposon-induced mutations in grape skin color.** *Science*, 304 (5673), 982.

Komatsu M, Shimamoto K and Kyozuka J (2003). **Two-step regulation and continuous retrotransposition of the rice LINE-type retrotransposon Karma.** *The Plant Cell*, 15 (8), 1934–44.

Korbel JO, Urban AE, Affourtit JP, Godwin B, Grubert F, Simons JF, Kim PM, Palejev D, Carriero NJ, Du L, Taillon BE, Chen Z, Tanzer A, Saunders ACE, Chi J, Yang F, Carter NP, Hurles ME, Weissman SM, Harkins TT, Gerstein MB, Egholm M and Snyder M (2007). **Paired-end mapping reveals extensive structural variation in the human genome.** *Science*, 318 (5849), 420–6.

Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D, Jones SJ and Marra MA (2009). **Circos: an information aesthetic for comparative genomics.** *Genome Research*, 19 (9), 1639–45.

Kuang H, Woo S-S, Meyers BC, Nevo E and Michelmore RW (2004). **Multiple genetic processes result in heterogeneous rates of evolution within the major cluster disease resistance genes in lettuce.** *The Plant Cell*, 16 (11), 2870–94.

Kumar A and Bennetzen JL (1999). **Plant retrotransposons**. *Annual Review of Genetics*, 33, 479–532.

Lai J, Li R, Xu X, Jin W, Xu M, Zhao H, Xiang Z, Song W, Ying K, Zhang M, Jiao Y, Ni P, Zhang J, Li D, Guo X, Ye K, Jian M, Wang B, Zheng H, Liang H, Zhang X, Wang S, Chen S, Li J, Fu Y, Springer NM, Yang H, Wang J, Dai J, Schnable PS and Wang J (2010). **Genome-wide patterns of genetic variation among elite maize inbred lines**. *Nature Genetics*, 42 (11), 1027–1030.

Langmead B, Trapnell C, Pop M and Salzberg SL (2009). **Ultrafast and memory-efficient alignment of short DNA sequences to the human genome**. *Genome Biology*, 10 (3), R25.

Levadoux L (1956). **Les populations sauvages et cultivées de Vitis vinifera L**. *Annales de L'amélioration Des Plantes*, (6), 59–118.

Levin HL and Moran J V. (2011). **Dynamic interactions between transposable elements and their hosts**. *Nature Reviews Genetics*, 12 (9), 615–627.

Li H and Durbin R (2009). **Fast and accurate short read alignment with Burrows-Wheeler transform.** *Bioinformatics*, 25 (14), 1754–60.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R and 1000 Genome Project Data Processing Subgroup (2009). **The Sequence Alignment/Map format and SAMtools.** *Bioinformatics*, 25 (16), 2078–9.

Li R, Ye J, Li S, Wang J, Han Y, Ye C, Wang J, Yang H, Yu J, Wong GK-S and Wang J (2005). **ReAS: Recovery of ancestral sequences for transposable elements from the unassembled reads of a whole genome shotgun.** *PLoS Computational Biology*, 1 (4), e43.

Li Y, Zhou G, Ma J, Jiang W, Jin L, Zhang Z, Guo Y, Zhang J, Sui Y, Zheng L, Zhang S, Zuo Q, Shi X, Li Y, Zhang W, Hu Y, Kong G, Hong H, Tan B, Song J, Liu Z, Wang Y, Ruan H, Yeung CKL, Liu J, Wang H, Zhang L, Guan R, Wang K, Li W, Chen S, Chang R, Jiang Z, Jackson S a, Li R and Qiu L (2014). **De novo assembly of soybean wild relatives for pan-genome analysis of diversity and agronomic traits**. *Nature Biotechnology*, 32 (10), 1045–1052.

Lisch D (2012). **How important are transposons for plant evolution?** *Nature Reviews Genetics*, 14 (1), 49–61.

Liu B and Wendel JF (2000). **Retrotransposon activation followed by rapid repression in introgressed rice plants.** *Genome / National Research Council Canada = Genome / Conseil National de Recherches Canada*, 43, 874–880.

Liu F, Zhu Y, Yi Y, Lu N, Zhu B and Hu Y (2014). **Comparative genomic analysis of Acinetobacter baumannii clinical isolates reveals extensive genomic variation and diverse antibiotic resistance determinants.** *BMC Genomics*, 15, 1163.

Ma J and Bennetzen JL (2004). **Rapid recent growth and divergence of rice nuclear genomes.** *Proceedings of the National Academy of Sciences of the United States of America*, 101 (34), 12404–12410.

Ma J, Devos KM and Bennetzen JL (2004). **Analyses of LTR-retrotransposon structures reveal recent and rapid genomic DNA loss in rice.** *Genome Research*, 14 (5), 860–9.

Magi A, Benelli M, Gozzini A, Girolami F, Torricelli F and Brandi ML (2010). **Bioinformatics for Next Generation Sequencing Data**. *Genes*, 1 (2), 294–307.

Mann M and Jensen ON (2003). **Proteomic analysis of post-translational modifications.** *Nature Biotechnology*, 21 (3), 255–261.

Mantel N (1967). **The detection of disease clustering and a generalized regression approach**. *Cancer Research*, 27 (June), 209–220.

Mardis ER (2008). **The impact of next-generation sequencing technology on genetics.** *Trends in Genetics*, 24 (3), 133–41.

Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen Y-J, Chen Z, Dewell SB, Du L, Fierro JM, Gomes X V, Godwin BC, He W, Helgesen S, Ho CH, Ho CH, Irzyk GP, Jando SC, Alenquer MLI, Jarvie TP, Jirage KB, Kim J-B, Knight JR, Lanza JR, Leamon JH, Lefkowitz SM, Lei M, Li J, Lohman KL, Lu H, Makhijani VB, McDade KE, McKenna MP, Myers EW, Nickerson E, Nobile JR, Plant R, Puc BP, Ronan MT, Roth GT, Sarkis GJ, Simons JF, Simpson JW, Srinivasan M, Tartaro KR, Tomasz A, Vogt KA, Volkmer GA, Wang SH, Wang Y, Weiner MP, Yu P, Begley RF and Rothberg JM (2005). **Genome sequencing in microfabricated high-density picolitre reactors.** *Nature*, 437 (7057), 376–80.

Marroni F, Pinosio S and Morgante M (2014). **Structural variation and genome complexity: is dispensable really dispensable?** *Current Opinion in Plant Biology*, 18, 31–6.

Marschall T, Costa IG, Canzar S, Bauer M, Klau GW, Schliep A and Schönhuth A (2012). **CLEVER: clique-enumerating variant finder.** *Bioinformatics*, 28 (22), 2875–82.

Marshall CR, Noor A, Vincent JB, Lionel AC, Feuk L, Skaug J, Shago M, Moessner R, Pinto D, Ren Y, Thiruvahindrapduram B, Fiebig A, Schreiber S, Friedman J, Ketelaars CEJ, Vos YJ, Ficicioglu C, Kirkpatrick S, Nicolson R, Sloman L, Summers

A, Gibbons CA, Teebi A, Chitayat D, Weksberg R, Thompson A, Vardy C, Crosbie V, Luscombe S, Baatjes R, Zwaigenbaum L, Roberts W, Fernandez B, Szatmari P and Scherer SW (2008). **Structural variation of chromosomes in autism spectrum disorder.** *American Journal of Human Genetics*, 82 (2), 477–88.

Martin M (2011). **Cutadapt removes adapter sequences from high-throughput sequencing reads**. *EMBnet Journal*, 17 (1), 10–12.

McClintock B (1956). **Controlling Elements and the Gene**. *Cold Spring Harbor Symposia on Quantitative Biology*, 21 (0), 197–216.

McGovern PE (2003). **Ancient Wine: The search for the Origins of Viniculture**. In Princeton University Press (pp. 1–15).

McHale L, Tan X, Koehl P and Michelmore RW (2006). **Plant NBS-LRR proteins: adaptable guards.** *Genome Biology*, 7 (4), 212.

McHale LK, Haun WJ, Xu WW, Bhaskar PB, Anderson JE, Hyten DL, Gerhardt DJ, Jeddeloh JA and Stupar RM (2012). **Structural variants in the soybean genome localize to clusters of biotic stress-response genes.** *Plant Physiology*, 159 (4), 1295–308.

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M and DePristo MA (2010). **The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data.** *Genome Research*, 20 (9), 1297–303.

Medvedev P, Stanciu M and Brudno M (2009). **Computational methods for discovering structural variation with next-generation sequencing**. *Nature Methods*, 6 (11), 13–20.

Mills RE, Walter K, Stewart C, Handsaker RE, Chen K, Alkan C, Abyzov A, Yoon SC, Ye K, Cheetham RK, Chinwalla A, Conrad DF, Fu Y, Grubert F, Hajirasouliha I, Hormozdiari F, Iakoucheva LM, Iqbal Z, Kang S, Kidd JM, Konkel MK, Korn J, Khurana E, Kural D, Lam HYK, Leng J, Li R, Li Y, Lin C-Y, Luo R, Mu XJ, Nemesh J, Peckham HE, Rausch T, Scally A, Shi X, Stromberg MP, Stütz AM, Urban AE, Walker JA, Wu J, Zhang Y, Zhang ZD, Batzer MA, Ding L, Marth GT, McVean G, Sebat J, Snyder M, Wang J, Ye K, Eichler EE, Gerstein MB, Hurles ME, Lee C, McCarroll SA and Korbel JO (2011). **Mapping copy number variation by population-scale genome sequencing.** *Nature*, 470 (7332), 59–65.

Moisy C, Garrison KE, Meredith CP and Pelsy F (2008). **Characterization of ten novel Ty1/copia-like retrotransposon families of the grapevine genome.** *BMC Genomics*, 9, 469.

Mondragon-Palomino M and Gaut BS (2005). **Gene conversion and the evolution of three leucine-rich repeat gene families in Arabidopsis thaliana.** *Molecular Biology and Evolution*, 22 (12), 2444–56.

Morgante M, De Paoli E and Radovic S (2007). **Transposable elements and the plant pan-genomes.** *Current Opinion in Plant Biology*, 10 (2), 149–55.

Muñoz-Amatriaín M, Eichten SR, Wicker T, Richmond TA, Mascher M, Steuernagel B, Scholz U, Ariyadasa R, Spannagl M, Nussbaumer T, Mayer KFX, Taudien S, Platzer M, Jeddeloh JA, Springer NM, Muehlbauer GJ and Stein N (2013). **Distribution, functional impact, and origin mechanisms of copy number variation in the barley genome.** *Genome Biology*, 14 (6), R58.

Myles S, Boyko AR, Owens CL, Brown PJ, Grassi F, Aradhya MK, Prins B, Reynolds A, Chia J-M, Ware D, Bustamante CD and Buckler ES (2011). **Genetic structure and domestication history of the grape.** *Proceedings of the National Academy of Sciences of the United States of America*, 108 (9), 3530–5.

Negrul AM (1946). **Origin and classification of cultivated grape.** In Baranov, A., Kai, Y., Lazarevski, M., Palibin, T., & Prosmoserdov, N. (Eds.), (pp. 159–216).

Oliver KR, McComb JA and Greene WK (2013). **Transposable elements: powerful contributors to angiosperm evolution and diversity.** *Genome Biology and Evolution*, 5 (10), 1886–901.

Olson SA (2002). **Emboss opens up sequence analysis**. *Briefings in Bioinformatics*, 3 (1), 87–91.

Ong-Abdullah M, Ordway JM, Jiang N, Ooi S-E, Kok S, Sarpan N, Azimi N, Hashim AT, Ishak Z, Rosli SK, Malike FA, Bakar NAA, Marjuni M, Abdullah N, Yaakub Z, Amiruddin MD, Nookiah R, Singh R, Low EL, Chan K, Azizi N, Smith SW, Bacher B, Budiman MA, Van Brunt A, Wischmeyer C, Beil M, Hogan M, Lakey N, Lim C, Arulandoo X, Wong C-K, Choo C, Wong W, Kwan Y, Alwee SSRS, Sambanthamurthi R and Martienssen RA (2015). **Loss of Karma transposon methylation underlies the mantled somaclonal variant of oil palm**. *Nature*, 525 (7570), 533–537.

Orgel LE and Crick FHC (1980). **Selfish DNA: the ultimate parasite**. *Nature*, 284, 604–607.

Ossowski S, Schneeberger K, Clark RM, Lanz C, Warthmann N and Weigel D (2008). **Sequencing of natural strains of Arabidopsis thaliana with short reads.** *Genome Research*, 18 (12), 2024–33.

Paradis E, Claude J and Strimmer K (2004). **APE: Analyses of Phylogenetics and Evolution in R language**. *Bioinformatics*, 20 (2), 289–290.

Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA and Reich D (2006). **Principal components analysis corrects for stratification in genome-wide association studies.** *Nature Genetics*, 38 (8), 904–9.

R Core Team (2013). **R: A language and environment for statistical computing.** *R Foundation for Statistical Computing, Vienna, Austria.*

Raphael BJ (2012). **Chapter 6: Structural variation and medical genomics.** *PLoS Computational Biology*, 8 (12), e1002821.

Rastogi K, Kim K and Lee S (2013). **Next-Generation Sequencing Technology for Crop Improvement**. *SABRAO Journal of Breeding and Genetics*, 45 (1), 84–99.

Rausch T, Zichner T, Schlattl A, Stütz AM, Benes V and Korbel JO (2012). **DELLY: structural variant discovery by integrated paired-end and split-read analysis.** *Bioinformatics*, 28 (18), i333–i339.

Rice P, Longden I and Bleasby A (2000). **EMBOSS: The European Molecular Biology Open Software Suite**. *Trends in Genetics*, 16 (6), 276–277.

Sachidanandam R, Weissman D, Schmidt SC, Kakol JM, Stein LD, Marth G, Sherry S, Mullikin JC, Mortimore BJ, Willey DL, Hunt SE, Cole CG, Coggill PC, Rice CM, Ning Z, Rogers J, Bentley DR, Kwok PY, Mardis ER, Yeh RT, Schultz B, Cook L, Davenport R, Dante M, Fulton L, Hillier L, Waterston RH, McPherson JD, Gilman B, Schaffner S, Van Etten WJ, Reich D, Higgins J, Daly MJ, Blumenstiel B, Baldwin J, Stange-Thomann N, Zody MC, Linton L, Lander ES and Altshuler D (2001). **A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms.** *Nature*, 409 (6822), 928–33.

SanMiguel P, Gaut BS, Tikhonov A, Nakajima Y and Bennetzen JL (1998). **The paleontology of intergene retrotransposons of maize.** *Nature Genetics*, 20 (1), 43–5.

Sanseverino W, Hénaff E, Vives C, Pinosio S, Burgos-Paz W, Morgante M, Ramos-Onsins SE, Garcia-Mas J and Casacuberta JM (2015). **Transposon Insertions, Structural Variations, and SNPs Contribute to the Evolution of the Melon Genome.** *Molecular Biology and Evolution*, 32 (10), 2760–74.

Saxena RK, Edwards D and Varshney RK (2014). **Structural variations in plant genomes.** *Briefings in Functional Genomics*, 13 (4), 296–307.

Schatz MC, Maron LG, Stein JC, Hernandez Wences A, Gurtowski J, Biggers E, Lee H, Kramer M, Antoniou E, Ghiban E, Wright MH, Chia J, Ware D, McCouch SR and McCombie WR (2014). **Whole genome de novo assemblies of three divergent strains of rice, Oryza sativa, document novel gene space of aus and indica.** *Genome Biology*, 15 (11), 506.

Scherer SW, Lee C, Birney E, Altshuler DM, Eichler EE, Carter NP, Hurles ME and Feuk L (2007). **Challenges and standards in integrating surveys of structural variation.** *Nature Genetics*, 39 (7 Suppl), S7–S15.

Schmitz RJ, He Y, Valdés-lópez O, Res G, Gent JI, Ellis N a, Guo L, Valde O, Stacey G and Ecker JR (2013). **Epigenome-wide inheritance of cytosine methylation variants in a recombinant inbred population Epigenome-wide inheritance of cytosine methylation variants in a recombinant inbred population**, 1–13.

Sebat J, Lakshmi B, Troge J, Alexander J, Young J, Lundin P, Månér S, Massa H, Walker M, Chi M, Navin N, Lucito R, Healy J, Hicks J, Ye K, Reiner A, Gilliam TC, Trask B, Patterson N, Zetterberg A and Wigler M (2004). **Large-scale copy number polymorphism in the human genome.** *Science*, 305 (5683), 525–8.

Shendure J and Ji H (2008). **Next-generation DNA sequencing**. *Nature Biotechnology*, 26 (10), 1135–1145.

Shendure J, Porreca GJ, Reppas NB, Lin X, McCutcheon JP, Rosenbaum AM, Wang MD, Zhang K, Mitra RD and Church GM (2005). **Accurate multiplex polony sequencing of an evolved bacterial genome.** *Science*, 309 (5741), 1728–32.

Shirasu K, Schulman AH, Lahaye T and Schulze-Lefert P (2000). **A contiguous 66-kb barley DNA sequence provides evidence for reversible genome expansion.** *Genome Research*, 10 (7), 908–15.

Sindi S, Helman E, Bashir A and Raphael BJ (2009). **A geometric approach for classification and comparison of structural variants.** *Bioinformatics*, 25 (12), i222–30.

Slotkin RK and Martienssen R (2007). **Transposable elements and the epigenetic regulation of the genome.** *Nature Reviews Genetics*, 8 (4), 272–85.

Springer NM, Ying K, Fu Y, Ji T, Yeh C-T, Jia Y, Wu W, Richmond T, Kitzman J, Rosenbaum H, Iniguez AL, Barbazuk WB, Jeddeloh J a, Nettleton D and Schnable PS (2009). **Maize inbreds exhibit high levels of copy number variation (CNV) and presence/absence variation (PAV) in genome content.** *PLoS Genetics*, 5 (11), e1000734.

Stankiewicz P and Lupski JR (2002). **Genome architecture, rearrangements and genomic disorders**. *Trends in Genetics*, 18 (2), 74–82.

Tettelin H, Masignani V, Cieslewicz MJ, Donati C, Medini D, Ward NL, Angiuoli S V, Crabtree J, Jones AL, Durkin AS, Deboy RT, Davidsen TM, Mora M, Scarselli M, Margarit y Ros I, Peterson JD, Hauser CR, Sundaram JP, Nelson WC, Madupu R, Brinkac LM, Dodson RJ, Rosovitz MJ, Sullivan SA, Daugherty SC, Haft DH, Selengut J, Gwinn ML, Zhou L, Zafar N, Khouri H, Radune D, Dimitrov G, Watkins K, O'Connor KJB, Smith S, Utterback TR, White O, Rubens CE, Grandi G, Madoff LC, Kasper DL, Telford JL, Wessels MR, Rappuoli R and Fraser CM (2005). **Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: implications for the microbial "pan-genome".** *Proceedings of the National Academy of Sciences of the United States of America*, 102 (39), 13950–5.

The International Schizophrenia Consortium (2008). **Rare chromosomal deletions and duplications increase risk of schizophrenia.** *Nature*, 455 (7210), 237–41.

This P, Lacombe T and Thomas MR (2006). **Historical origins and genetic diversity of wine grapes**. *Trends in Genetics*, 22 (9), 511–519.

Trapnell C, Roberts A, Goff L, Petrea G, Kim D, Kelley DR, Pimentel H, Salzberg S, Rinn JL and Pachter L (2012). **Differential gene and transcript expressiont analysis of RNA-seq experiments with TopHat and Cufflinks**. *Natures Protocols*, 7 (3), 562–578.

Trapnell C, Williams B a, Pertea G, Mortazavi A, Kwan G, van Baren MJ, Salzberg SL, Wold BJ and Pachter L (2010). **Transcript assembly and abundance estimation from RNA-Seq reveals thousands of new transcripts and switching among isoforms**. *Nature Biotechnology*, 28 (5), 511–515.

Troshin LP, Nedov P, Litvak I and Guzun N (1990). **Improvement of Vitis vinifera sativa DC. taxonomy**. *Vitis (special Issue) Proceedings of the 5th International Symposium on Grape Breeding, 1989*, 37–43.

Turcatti G, Romieu A, Fedurco M and Tairi A-P (2007). **A new class of cleavable fluorescent nucleotides: synthesis and optimization as reversible terminators for DNA sequencing by synthesis**. *Nucleic Acids Research*, 36 (4), e25.

Tuskan GA, Difazio S, Jansson S, Bohlmann J, Grigoriev I, Hellsten U, Putnam N, Ralph S, Rombauts S, Salamov A, Schein J, Sterck L, Aerts A, Bhalerao RR, Bhalerao RP, Blaudez D, Boerjan W, Brun A, Brunner A, Busov V, Campbell M, Carlson J, Chalot M, Chapman J, Chen G-L, Cooper D, Coutinho PM, Couturier J, Covert S, Cronk Q, Cunningham R, Davis J, Degroeve S, Déjardin A, Depamphilis C, Detter J, Dirks B, Dubchak I, Duplessis S, Ehlting J, Ellis B, Gendler K,

Goodstein D, Gribskov M, Grimwood J, Groover A, Gunter L, Hamberger B, Heinze B, Helariutta Y, Henrissat B, Holligan D, Holt R, Huang W, Islam-Faridi N, Jones S, Jones-Rhoades M, Jorgensen R, Joshi C, Kangasjärvi J, Karlsson J, Kelleher C, Kirkpatrick R, Kirst M, Kohler A, Kalluri U, Larimer F, Leebens-Mack J, Leplé J-C, Locascio P, Lou Y, Lucas S, Martin F, Montanini B, Napoli C, Nelson DR, Nelson C, Nieminen K, Nilsson O, Pereda V, Peter G, Philippe R, Pilate G, Poliakov A, Razumovskaya J, Richardson P, Rinaldi C, Ritland K, Rouzé P, Ryaboy D, Schmutz J, Schrader J, Segerman B, Shin H, Siddiqui A, Sterky F, Terry A, Tsai C-J, Uberbacher E, Unneberg P, Vahala J, Wall K, Wessler S, Yang G, Yin T, Douglas C, Marra M, Sandberg G, Van de Peer Y and Rokhsar D (2006). **The genome of black cottonwood, Populus trichocarpa (Torr. & Gray).** *Science*, 313 (5793), 1596–604.

Van der Auwera GA, Carneiro MO, Hartl C, Poplin R, Del Angel G, Levy-Moonshine A, Jordan T, Shakir K, Roazen D, Thibault J, Banks E, Garimella K V, Altshuler D, Gabriel S and DePristo MA (2002). **From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline**. *Current Protocols in Bioinformatics*, 11 (1110), 11.10.1–11.10.33.

Varshney RK, Nayak SN, May GD and Jackson SA (2009). **Next-generation sequencing technologies and their implications for crop genetics and breeding**. *Trends in Biotechnology*, 27 (9), 522–530.

Vaughn MW, Tanurdzić M, Lippman Z, Jiang H, Carrasquillo R, Rabinowicz PD, Dedhia N, McCombie WR, Agier N, Bulski A, Colot V, Doerge RW and Martienssen RA (2007). **Epigenetic natural variation in Arabidopsis thaliana.** *PLoS Biology*, 5 (7), e174.

Vitte C and Panaud O (2005). **LTR retrotransposons and flowering plant genome size: Emergence of the increase/decrease model**. *Cytogenetic and Genome Research*, 110 (1-4), 91–107.

Vitte C and Panaud O (2003). **Formation of solo-LTRs through unequal homologous recombination counterbalances amplifications of LTR retrotransposons in rice Oryza sativa L.** *Molecular Biology and Evolution*, 20 (4), 528–540.

Vitte C, Panaud O and Quesneville H (2007). **LTR retrotransposons in rice (Oryza sativa, L.): recent burst amplifications followed by rapid DNA loss.** *BMC Genomics*, 8, 218.

Vitulo N, Forcato C, Carpinelli E, Telatin A, Campagna D, D'Angelo M, Zimbello R, Corso M, Vannozzi A, Bonghi C, Lucchin M and Valle G (2014). **A deep survey of alternative splicing in grape reveals changes in the splicing machinery related to tissue, stress condition and genotype**. *BMC Plant Biology*, 14 (1), 99.

Volik S, Zhao S, Chin K, Brebner JH, Herndon DR, Tao Q, Kowbel D, Huang G, Lapuk A, Kuo W-L, Magrane G, De Jong P, Gray JW and Collins C (2003). **End-sequence profiling: sequence-based analysis of aberrant genomes.** *Proceedings of the National Academy of Sciences of the United States of America*, 100 (13), 7696–701.

Voytas DF and Naylor GJP (1998). **Rapid Flux in Plant Genomes**. *Nature Genetics*, 20 (1), 6–7.

Wang H and Liu J-S (2008). **LTR retrotransposon landscape in Medicago truncatula: more rapid removal than in rice.** *BMC Genomics*, 9 (1), 382.

Warnes G, Leisch F, Man M and Warnes G (2011). **Package genetics**.

Wawrzynski A, Ashfield T, Chen NWG, Mammadov J, Nguyen A, Podicheti R, Cannon SB, Thareau V, Ameline-Torregrosa C, Cannon E, Chacko B, Couloux A, Dalwani A, Denny R, Deshpande S, Egan AN, Glover N, Howell S, Ilut D, Lai H, Del Campo SM, Metcalf M, O'Bleness M, Pfeil BE, Ratnaparkhe MB, Samain S, Sanders I, Ségurens B, Sévignac M, Sherman-Broyles S, Tucker DM, Yi J, Doyle JJ, Geffroy V, Roe B a, Maroof M a S, Young ND and Innes RW (2008). **Replication of nonautonomous retroelements in soybean appears to be both recent and common.** *Plant Physiology*, 148 (4), 1760–71.

Wheeler DL, Barrett T, Benson DA, Bryant SH, Canese K, Church DM, DiCuccio M, Edgar R, Federhen S, Helmberg W, Kenton DL, Khovayko O, Lipman DJ, Madden TL, Maglott DR, Ostell J, Pontius JU, Pruitt KD, Schuler GD, Schriml LM, Sequeira E, Sherry ST, Sirotkin K, Starchenko G, Suzek TO, Tatusov R, Tatusova TA, Wagner L and Yaschenko E (2005). **Database resources of the National Center for Biotechnology Information.** *Nucleic Acids Research*, 33 (Database issue), D39–45.

Wicker T, Sabot F, Hua-Van A, Bennetzen JL, Capy P, Chalhoub B, Flavell A, Leroy P, Morgante M, Panaud O, Paux E, SanMiguel P and Schulman AH (2007). **A unified classification system for eukaryotic transposable elements.** *Nature Reviews Genetics*, 8 (12), 973–982.

Wu GA, Prochnik S, Jenkins J, Salse J, Hellsten U, Murat F, Perrier X, Ruiz M, Scalabrin S, Terol J, Takita MA, Labadie K, Poulain J, Couloux A, Jabbari K, Cattonaro F, Del Fabbro C, Pinosio S, Zuccolo A, Chapman J, Grimwood J, Tadeo FR, Estornell LH, Muñoz-Sanz J V, Ibanez V, Herrero-Ortega A, Aleza P, Pérez-

Pérez J, Ramón D, Brunel D, Luro F, Chen C, Farmerie WG, Desany B, Kodira C, Mohiuddin M, Harkins T, Fredrikson K, Burns P, Lomsadze A, Borodovsky M, Reforgiato G, Freitas-Astúa J, Quetier F, Navarro L, Roose M, Wincker P, Schmutz J, Morgante M, Machado MA, Talon M, Jaillon O, Ollitrault P, Gmitter F and Rokhsar D (2014). **Sequencing of diverse mandarin, pummelo and orange genomes reveals complex history of admixture during citrus domestication.** *Nature Biotechnology*, 32 (7), 656–62.

Xu X, Liu X, Ge S, Jensen JD, Hu F, Li X, Dong Y, Gutenkunst RN, Fang L, Huang L, Li J, He W, Zhang G, Zheng X, Zhang F, Li Y, Yu C, Kristiansen K, Zhang X, Wang J, Wright M, McCouch S, Nielsen R, Wang J and Wang W (2012). **Resequencing 50 accessions of cultivated and wild rice yields markers for identifying agronomically important genes.** *Nature Biotechnology*, 30 (1), 105–11.

Xu Z and Wang H (2007). **LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons.** *Nucleic Acids Research*, 35 (Web Server issue), W265–8.

Ye K, Schulz MH, Long Q, Apweiler R and Ning Z (2009). **Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads.** *Bioinformatics*, 25 (21), 2865–71.

You FM, Huo N, Gu YQ, Luo M-C, Ma Y, Hane D, Lazo GR, Dvorak J and Anderson OD (2008). **BatchPrimer3: a high throughput web application for PCR and sequencing primer design.** *BMC Bioinformatics*, 9, 253.

Yu P, Wang C, Xu Q, Feng Y, Yuan X, Yu H, Wang Y, Tang S and Wei X (2011). **Detection of copy number variations in rice using array-based comparative genomic hybridization.** *BMC Genomics*, 12, 372.

Zhang H-B, Zhao X, Ding X, Paterson AH and Wing RA (1995). **Preparation of megabase-size DNA from plant nuclei**. *The Plant Journal*, 7 (1), 175–184.

Zheng L-Y, Guo X-S, He B, Sun L-J, Peng Y, Dong S-S, Liu T-F, Jiang S, Ramachandran S, Liu C-M and Jing H-C (2011). **Genome-wide patterns of genetic variation in sweet and grain sorghum (Sorghum bicolor).** *Genome Biology*, 12 (11), R114.

Zhou Y, Burnham C-AD, Hink T, Chen L, Shaikh N, Wollam A, Sodergren E, Weinstock GM, Tarr PI and Dubberke ER (2014). **Phenotypic and genotypic analysis of Clostridium difficile isolates: a single-center study.** *Journal of Clinical Microbiology*, 52 (12), 4260–6.

7 APPENDIX

Table 15: Common names of the *Vitis vinifera* varieties sampled in the present work.

Prime name (short format)	Common name
Aciaruli Tetri	Aciaruli Tetri
Agadai	Agadai
Aglianico	Aglianico
Airen	Airen
Alexandroouli	Alexandruli
Ansonica	Ansonica
Ararati	Ararati
Assyrtiko	Assyrtiko
Asyl Kara	Asyl Kara
Autumn Royal	Autumn Royal
Barbera	Barbera
Bayan Shirei	Bayan Shirei
Berzamino	Berzamino
Bovale	Bovale
Cabernet Franc	Cabernet Franc
Cabernet Sauvignon	Cabernet Sauvignon
Carignan	Carignano
Catarratto B.C.	Catarratto Bianco Comune
Cesanese d'Affile	Cesanese d'Affile
Chaouch blanc	Chaouch Blanc
Charistvala Kolchuri	Charistvala Kolchuri
Chasselas Blanc	Chasselas
Clairette Blanche	Clairette
Coarna Alba	Pukhlyakovskii
Corvina Veronese	Corvina
Daphnia	Daphnia
Disecka	Disecka
Enantio	Enantio
Falanghina	Falanghina
Fiano	Fiano
Fumat	Fumat
Garganega	Garganega
Garnacha	Cannonau
Glera	Glera

Prime name (short format)	Common name
Gorula	Gorula
Grechetto Bianco	Grechetto
Greco di Tufo	Greco di Tufo
Grignolino	Grignolino
Gyulyabi Dagestanskii	Gyulyabi Dagestanskii
Harslevelue	Harslevelue
Henab Turki	Rumi Ahmar
Heunisch Weiss	Gouais Blanc
Italia	Italia
Kadarka	Kadarka
Katta Kurgan	Katta Kurgan
Khop Khalat	Hop Halat
Kishmish Vatkana	Kishmish Vatkana
Lambrusco di Sorbara	Lambrusco di Sorbara
Lambrusco Grasparossa	Lambrusco Grasparossa
Limnio	Limnio
Malvasia Bianca	Malvasia del Lazio
Malvasia Bianca Lunga	Malvasia Bianca Lunga
Malvasia di Sardegna	Malvasia di Lipari
Malvasia Istriana	Malvasia Istriana
Marandi Shemakhinskii	Marandi Shemakhinskii
Mauzac Blanc	Mauzac
Mavrodaphni	Mavrodaphni
Merlot Noir	Merlot
Montepulciano	Montepulciano
Moscato di Scanzo	Moscato di Scanzo
Mtsvane Kachuri	Mtsvane Kachuri
Muscat a Petits Grains B.	Moscato Bianco
Narma	Narma
Nasco	Nasco
Nebbiolo	Nebbiolo
Negro Amaro	Negro Amaro
Nero d'Avola	Nero d'Avola
Nieddu Mannu	Nieddu Mannu
Nosiola	Nosiola
Ojaleshi	Ojaleshi
Passerina	Passerina
Pecorino	Pecorino
Petit Rouge	Petit Rouge
Picolit	Picolit
Pignoletto	Pignoletto
Pinela	Pinela
Pinot	Pinot

Prime name (short format)	Common name
Plechistik	Plechistik
PN40024	PN40024
Raboso Piave	Raboso Piave
Red Globe	Red Globe
Refosco P.R.	Refosco P.R.
Ribolla Gialla	Ribolla Gialla
Ribolla Gialla (Slovenia)	Ribolla Gialla (Slovenia)
Riesling Weiss	Rhein Riesling
Rkatsiteli	Rkatsiteli
Sagrantino	Sagrantino
Sahibi Safid	Sahibi Safid
Sangiovese	Sangiovese
Sauvignon Blanc	Sauvignon
Savagnin Blanc	Traminer
Schiava Gentile	Schiava Gentile
Schiava Grossa	Schiava Grossa
Schioppettino	Schioppettino
Sciavtsitska	Sciavtsitska
Shafei	Shafei
Sirgula	Sirgula
Sultanina	Sultanina
Tagobi	Tagobi
Taifi Rozovyi	Taifi Rozovyi
Tannat	Tannat
Tavkveri	Tavkveri
Terbash	Terbash
Terrano	Terrano
Thompson Seedless	Thompson Seedless
Tibouren	Rossese
Tocai Friulano	Tocai Friulano
Trebbiano Toscano	Trebbiano Toscano
Tschvediansis Tetra	Tschvediansis Tetra
Uva di Troia	Uva di Troia
V267	V267
V278	V278
V292	V292
V294	V294
V385	V385
V389	V389
V395	V395
V400	V400
V410	V410
V411	V411

Prime name (short format)	Common name
Verdicchio Bianco	Verdicchio
Verduzzo Friulano	Verduzzo
Vermentino	Vermentino
Vernaccia S.G.	Vernaccia S.G.
Welschriesling	Riesling Italico
Zametovka	Zametovka
Zelen	Zelen
Zinfandel	Primitivo

8 ACKNOWLEDGEMENTS

First of all I would like to express my gratitude to my supervisor Prof. Michele Morgante, who provided me with the opportunity to carry out, with his full support, the present work and complete my PhD thesis at the Institute of Applied Genomics.

I would like to warmly thank my co-supervisor Dr. Fabio Marroni, for his helpfulness, his statistical, informatics (R) and scientific support and his friendship.

Furthermore, special thanks go to Dr. Gabriele Di Gaspero, for his scientific support and cooperation.

Additionally, my sincere thanks go to Sara Pinosio for her helpfulness and useful discussion in setting up the SV detection analysis, in addition to her informatics and scientific support.

I would like to thank all my colleagues of the Institute of Applied Genomics (Aldo Tocci, Alessandro Gervaso, Alice Fornasiero, Andrea Zuccolo, Cristian Del Fabbro, Davide Scaglione, Eleonora Paparelli, Ettore Zapparoli, Mara Miculan, Michele Vidotto, Mirko Celii, Rachel Schwope, Simone Scalabrin, Vera Vendramin, Vittorio Zamboni and all the lab people) and of the University of Udine for their friendship, inspiration and full support during the last three years.

Last, but not least, I would like to thank my beautiful family, spread over Europe, for their help, support and unwavering love. In addition, I would like to thank Manno's family, for their warm welcome and support.

Lastly, I'd like to give special thanks to Marta Manno, for her love and full support during at least the last seven years.

The present work has been supported by the European Commission’s European Research Council, within the Seventh Framework Programme for Research (Grant number, 294780).