UNIVERSITÀ DEGLI STUDI DI UDINE

Dottorato di Ricerca in Scienze e Biotecnologie Agrarie

Ciclo XXIX

Coordinatore: Prof. Giuseppe Firrao

TESI DI DOTTORATO DI RICERCA

Identification and Mapping of Loci Controlling Viability in *Vitis vinifera* crosses

DOTTORANDA

Alice Fornasiero

SUPERVISORE

Prof. Michele Morgante

CO-SUPERVISORE

Dott. Fabio Marroni

ANNO ACCADEMICO 2015 - 2016

Summary

The present research work is part of the NOVABREED project, funded by the European Research Council (ERC). The focus of NOVABREED is the concept of pan-genome as a comprehensive representation of a species genome: its main aim is to characterise the dispensable portion that arises as a consequence of structural variation, and its contribution to the intra-specific genetic and phenotypic variation (Morgante, De Paoli, & Radovic, 2007). The focus of the present work is on the *Vitis vinifera* genome. The goal of the present research is to understand the genetic bases of inbreeding depression in grapevine through the identification of loci controlling viability and survival and their relationship with structural variation.

Vitis vinifera is a very variable species with high levels of both nucleotide diversity and structural variation (Gabriele Magris, PhD thesis, 2016). Cultivated varieties are highly heterozygous (Jaillon et al., 2007) and are expected to carry a large mutational load that is reflected in rather severe inbreeding depression observed upon crossing of related individuals or selfing. We set out to identify loci that are responsible for inbreeding depression both individually as well as a consequence of epistatic interactions. Through the analysis of segregation distortion, defined as deviation of segregation ratio from the expected Mendelian ratio, we explored progenies segregation pattern with the goal of isolating causative mutations of the distortion.

In order to characterize segregation distortion in *Vitis vinifera*, progenies obtained from self-crosses of six varieties and from one out-cross were genotyped using a Genotyping-by-Sequencing approach. The technique used, known as double digest Restriction Site Associated DNA Sequencing (ddRAD-seq) (Peterson, Weber, Kay, Fisher, & Hoekstra, 2012), subsamples the genome at homologous locations among individuals by coupling double restriction enzyme digestion to a selection of genomic fragments by size, allowing fine-scale control of the fraction of regions represented in the final library. SNP genotyping, by means of the software Stacks (Catchen, Hohenlohe, Bassham, Amores, & Cresko, 2013), allowed the identification of several regions of single locus distortion in each progeny assessed.

Progenies deriving from self-fertilization of the six cultivars Cabernet franc, Pinot Noir, Primitivo, Rkatsiteli, Sangiovese and Schiava Grossa, showed distortion in 12 different loci. Progeny of the out-cross between Schiava Grossa and Rkatsiteli did not show any locus of segregation distortion. Overall, ten loci of distortion revealed lethal effects, whereas two loci revealed severe deleterious effects. Among loci with lethal effect, seven showed the complete absence of a homozygous genotype, corresponding to the action of recessive alleles; two showed the action of partially dominant alleles and one showed nearly complete dominance of the lethal allele. Deleterious loci exhibited recessive and overdominant allelic effect on distortion, respectively. In three different varieties, chromosome 8 was revealed to harbour non overlapping loci of distortion with different allelic effects. Analysis of selected progenies over two vegetative seasons showed that five loci caused embryo or seedling lethality in early stages of growth, while two loci exerted their lethal effects on growth only after the first wintering. Fine mapping of the distorted regions allowed to narrow down the causal locus to less than 34 Kbp, in the best case.

Genotype data of the progeny of Rkatsiteli self-cross and of the progeny of Schiava Grossa and Rkatsiteli out-cross was used to build genetic linkage maps of the parental varieties. The three genetic maps were used to analyse recombination frequency along the genome of *Vitis vinifera*.

To identify structural variants (SVs), single nucleotide polymorphisms (SNPs) and small insertion-deletion polymorphisms (INDELs) contributing to the genetic load in the grapevine genome and leading to segregation distortion in the progenies of selfing, candidate loci were screened for mutations affecting genes. Haplotype phasing of alleles allowed to identify mutations belonging to the haplotype that generated segregation distortion in the progenies of selfing. Furthermore, expression of genes affected by such variation was evaluated in a panel of six varieties in three different tissues (leaf, berry and tendril).

The analysis of two-loci segregation distortion was also performed in order to identify epistatic interactions. Fisher's test of independence revealed one strong signal of interaction between loci on chromosome 1 and chromosome 11 in the variety Rkatsiteli, evidenced also by the pseudo-linkage signal in the genetic map. Further evidences showed that the interaction was actually due to a balanced translocation, which was validated through a PCR-based assay. In order to evaluate gene expression differences between the haplotype carrying the translocation and the normal

haplotype in Rkatsiteli, we performed allele specific expression (ASE) analysis for each of the three tissues. Interestingly, the analysis revealed no significant difference in the allele-specific expression profile in the tissues.

In order to detect the balanced translocation in other grapevine varieties, a panel of 196 cultivars was tested, revealing that three varieties - Alexandroouli, Mtsvane Kachuri and Gorula - carried the rearrangement. All four varieties originated from Georgia and belonged to the *Proles pontica*, although not all the varieties of this *Proles* showed the translocation. This suggests that Georgian varieties are distinct and genetically variable from western one and that translocation could be an ancient event never imported to the West and Central Europe varieties (*Proles occidentalis*).

Our study showed that self-fertilization of *Vitis vinifera* cultivars lead to high levels of segregation distortion in the progenies due to the presence of unfavourable alleles in genes. Future studies will be oriented to the characterization of the identified genes.

Furthermore, we generated a fine-scale map of recombination frequency along the genome of *Vitis vinifera*.

Lastly, we detected and validated a balanced translocation involving chromosome 1 and chromosome 11 in the variety Rkatsiteli and we found the chromosomal rearrangement in other three grapevine varieties.

Chapter 1 Introduction

1.1 The pan-genome concept and its application in plants

The key idea at the heart of NOVABREED is the concept of pan-genome. The term was firstly adopted by Tettelin and colleagues (2005) to describe the entire genomic complement of bacterial species, such as *Streptococcus agalactiae* (Tettelin et al., 2005). The genetic content of a single individual was no longer sufficient to represent the entire genetic variability of a species, owing to the presence of an extensive variation among individuals of the same species. Later, the concept of pan-genome was extended to plant species (Morgante, 2006). One striking evidence of extensive plant genetic variation was found in maize: the comparison between genomic regions in the maize inbred lines B73 and Mo17 revealed that approximately half of the sequences were not shared between the two lines (Brunner, Fengler, Morgante, Tingey, & Rafalski, 2005). The differences between the two genomes consisted in a surprisingly high level of structural diversity, mostly due to transposition of retro elements, and resulting in extensive non-homologies (Morgante, De Paoli, & Radovic, 2007).

The pan-genome is composed of a *core* portion, which includes the common genomic features of all individuals of a species and a *dispensable* portion, composed by partially shared or non-shared sequences among individuals. The intra-specific variation observed in plants and described by the dispensable genome is largely due to the movement of transposable elements and to other mechanisms - e.g. different types of errors occurring during recombination (F. Lu et al., 2015; Pinosio et al., 2016).

The major aim of the NOVABREED project is to define pan-genome composition and extent in two economically relevant plant species, grapevine and maize, and to gain new insights into the underlying genetic and phenotypic variation mostly ascribed to structural variation.

1.2 Structural variants as a source of intra-species genetic variation

For a long time, it was assumed that single nucleotide polymorphisms (SNPs) and small insertion-deletion polymorphisms (INDELs) accounted for the majority of intra-genomic variation of a species (Sachidanandam et al., 2001). However, structural variation has been acknowledged as a significant source of intra-specific variation in several species, including human (Lupski & Stankiewicz, 2010) and plant species, such as maize (Swanson-Wagner et al., 2010), barley (Muñoz-Amatriaín et al., 2013), rice (Hurwitz et al., 2010) and many others. Structural variants (SVs) typically refer to events spanning at least 1 Kbp in length and include rearrangements, such as inversion and translocations, and copy number variations. Copy number variants (CNVs) are defined as unbalanced changes resulting from gains or losses of DNA sequences among individual genomes (Żmieńko, Samelak, Kozłowski, & Figlerowicz, 2014). They include deletions, insertions, and duplications. While CNVs are sequences present in all genomes being compared, albeit in different copy number, presence-absence variants (PAVs) constitute an extreme form of CNV, in which a particular sequence is present in some individuals and missing in others (Swanson-Wagner et al., 2010). CNVs can be caused by different mechanisms, including recombination at non-allelic regions of extensive sequence similarity (non-allelic homologous recombination, NAHR); DNA repair at regions with very limited or no homology (e.g. non-homologous end-joining, NHEJ; and micro homology-mediated end-joining, MMEJ); replication-error mechanisms (e.g. fork stalling and template switching, FoSTeS) and the movement of transposable elements (TEs) (Muñoz-Amatriaín et al., 2013). Transposons indeed are at the origin of an important fraction of CNVs due to their capacity of mobilizing gene sequences within the genome (Morgante et al., 2007).

In the human genome, it has been shown that SVs encompass more total nucleotides and arise more frequently than SNPs (Conrad et al., 2010, 2011; Iafrate et al., 2004; Scherer et al., 2007). Furthermore, CNVs have been shown to play a role in defining genetic diversity between human populations (Sudmant et al., 2010; Sudmant et al., 2015) and to have a functional significance for many human disease (Lupski & Stankiewicz, 2010), such as HIV infection susceptibility (Pelak et al., 2011), cancer risk (Willis et al., 2014), and autism development (Girirajan et al., 2013).

In plants, CNVs were reported to overlap multigene families more often than isolated genes (Swanson-Wagner et al., 2010; Zheng et al., 2011) and this is consistent with recombination-based mechanisms of CNV formation. For instance, defence-related gene families, such as nucleotide-binding leucin-rich repeat (NB-LRR) domain containing genes, have been shown to be overrepresented within CNV regions (Żmieńko et al., 2014). Paralogous genes belonging to clusters of multigene families are often functionally redundant, triggering quantitative rather than qualitative changes in phenotype, but still contributing to genetic variation and controlling adaptive traits. In wheat, several major genes controlling flowering time, such as *Vrn-A1* and *Ppd-B1*, are characterized by altered copy number. Plants having an increased copy number of *Vrn-A1* showed longer vernalisation period and exhibited late flowering. Increased copy number of *Ppd-B1*, controlling photoperiod sensitivity, caused early flowering and day-neutral phenotype (Díaz, Zikhali, Turner, Isaac, & Laurie, 2012). Other instances involve insertion of TEs that modify the level and/or pattern of expression of genes: *Hopscotch* TE insertion in a regulatory region of the maize domestication gene teosinte branched1 (*tb1*) causes enhanced expression of the gene (Studer, Zhao, Ross-Ibarra, & Doebley, 2011). In barley, a small CACTA-like transposon inserted upstream the aluminium tolerance gene *HcAACT1* both enhances expression and alters the tissue localization of the gene (Fujii et al., 2012).

1.3 Transposable elements shape plant genomes

Transposable elements (TEs) are segments of DNA that move throughout the genome by their replication and integration into new locations, and represent an extremely variable component of the genome (Kidwell & Lisch, 1997). TEs can affect the overall architecture of the host genome, representing an endogenous source of variation: their activity contributes to the evolution of most eukaryotes, and they have been shown to be very active in the plant kingdom (Baidouri & Panaud, 2013). In maize, 85% of the total DNA is composed of TEs (Schnable et al., 2009): variation in the genome content due to TE movement in maize inbred lines leads to differences in transcript content and likely contributes to phenotypic diversity and heterosis (Springer et al., 2009). In

angiosperms, TEs were found to be more dynamic and labile with respect to the more stable mammalian genomes (Kejnovsky, Leitch, & Leitch, 2009).

TEs are divided in two main classes: retrotransposons (class I) and DNA transposons (class II) (Habibi, Pedram, AmirPhirozy, & Bonyadi, 2015). In plants, the most abundant elements among retrotransposons are Long Terminal Repeats (LTR) retrotransposons (Wicker et al., 2007). Class II elements in plants are prevalently composed of members of the hAT (*hobo*, *Activator* and *Tam3*) and Mutator-like element (MULE) superfamilies (Atkinson, 2015). By copy-and-paste mechanism, retrotransposons generate insertions of new copies along the genome involving an RNA intermediate, whereas cut-and-paste replication of DNA transposons causes excision of the TE from the source DNA and re-integration in the new location (Hickman & Dyda, 2015). Also present in many plant genomes, *Helitrons* are a particular class of DNA transposons which transposes by a copy-and-paste strategy that putatively involves rolling-circle replication (Morgante et al., 2007). Through their mechanism of transposition, TEs are responsible for the majority of SVs ranging between 1 and 25-30 Kbp in size. However, due to their repetitive nature, TEs can also mediate ectopic recombination events leading to larger SVs (Marroni, Pinosio, & Morgante, 2014). Activity of TEs shapes the pattern of gene content and structure and gene expression in several ways. Loss-of-function mutations can be generated when TEs interrupt exons or regulatory elements; gain-of-function mutations can arise when TEs movement introduces new regulatory elements or promoters, or creates new splice sites (Lisch, 2012; Zhao, Ferguson, & Jiang, 2016). Two main mechanisms are responsible for the evolution of new genes through exon shuffling: retroposition and transduplication. Retroposition is typical of retrotransposons and causes reverse transcription of host mRNA by a TE-encoded reverse transcriptase and its integration at a new position. Transduplication mediated by *Helitrons* produces duplication and rearrangement of whole exons-introns structure of host genes (Pritham & Thomas, 2015). Lastly, TE insertion may alter epigenetic context of genes and affect their expression (Lisch, 2012; Morgante et al., 2007).

1.4 Genetic load and the effect of inbreeding and domestication on the mutation load

The genetic load is defined as the reduction in the mean fitness of a population relative to a population composed of individuals having an optimal fitness (i.e. optimal genotypes). The genetic load can be caused by different genetic processes: recurrent deleterious mutations (the mutation load), genetic drift (the drift load), recombination affecting beneficial combinations of alleles, segregation reducing the frequency of fit heterozygotes (the segregation load), migration from other populations bringing less fit alleles into the local population (Whitlock & Davis, 2011).

Deleterious alleles present in a population are caused by the mutation process and compose the mutation load carried by a population. The frequency of such deleterious variants is continuously shaped by the action of drift and selection (Henn, Botigué, Bustamante, Clark, & Gravel, 2015). Frequency of deleterious mutations under mutation-selection balance model is low in sexually reproducing species, because gamete formation and genetic recombination during meiosis may create unfit genotypes that are rapidly eliminated by purifying selection (Kondrashov, 1988). Instead, weakly deleterious alleles may be effectively neutral and subject to the effect of genetic drift (Kimura, 1991). The beneficial effects of sexual reproduction and recombination can be lowered under certain circumstances: for example, selection against deleterious mutations is less effective in regions with low levels of recombination. Population demography and inbreeding are other factors which allow deleterious mutations to rise in frequency (Bersabé, Caballero, Pérez-Figueroa, & García-Dorado, 2015; Price & Arkin, 2015; Zeng & Charlesworth, 2010).

Inbreeding is the mating of individuals that are closely related genetically, and is opposed to outbreeding, which is the mating of unrelated organisms. The phenomenon of inbreeding depression refers to the reduction in fitness of the offspring resulting from the mating between related individuals (Cheptou & Donohue, 2011). Selfing, the most extreme form of inbreeding, reduces heterozygosity by 50% at each generation (Carr & Dudash, 2003) and expose lethal mutations to selection faster in the resulting progenies than in an outcrossing population (Glémin, Ronfort, & Bataillon, 2003). Self-pollination in outcrossing plant species unmasks recessive deleterious alleles by increasing homozygosity of individuals (D. Charlesworth & Willis, 2009),

leading to decreased percentage of germination, higher seedling mortality, lower vigour of the survivors and lower fertility or even sterility (Levadoux, 1956). Inbreeding depression is classically measured as $(w_x - w_s) / w_x$, where w_x and w_s are the fitness in the outbred and in the inbred progeny, respectively. In self-compatible organisms as a particular case, inbreeding depression can be measured as the relative decrease in survival and/or reproduction of individuals after a single generation of self-fertilization (D. Charlesworth & Willis, 2009; Cheptou & Donohue, 2011). Deleterious mutations can accumulate through the process of domestication, during which populations undergo multiple bottlenecks followed by expansions, accompanied by strong artificial selection on numerous traits. During the process of artificial selection on a desired trait, damaging alleles lying in linkage disequilibrium with the beneficial genotype can be accidentally selected as well. This phenomenon, called hitchhiking, can hinder the ability of selection in fixing the beneficial mutation and in weeding the deleterious one (Maynard-Smith & Haigh, 1974). As a consequence, artificial selection can increase the rate at which deleterious mutations accumulate, particularly when the effect of the advantageous mutation outweighs the effects of linked deleterious mutation (Fay & Wu, 2000). By comparing two subspecies of the Asian rice (*Oryza sativa* ssp. *indica* and ssp. *Japonica*) to the ancestral species (*Oryza rufipogon*), Lu J and colleagues (2006) found an increased rate in nonsynonymous substitutions in the cultivars with respect to the wild species, and proposed that amino acid polymorphisms can be accounted for by the accumulation of deleterious mutations during rice domestication. They hypothesized that artificial selection would have increased the frequency of deleterious mutations in the absence of effective recombination, suggesting that hitchhiking had occurred (Lu J et al., 2006).

Selection is less effective in low-recombining regions, where deleterious mutations tend to accumulate (Mezmouk & Ross-Ibarra, 2014; M. Zhang, Zhou, Bawa, Suren, & Holliday, 2016). Renaut S and colleagues (2015) compared the load of deleterious mutations in 21 accessions from natural populations of sunflower (*Helianthus annuus*) and 19 domesticated accessions of the common sunflower, using whole-transcriptome SNP data. Among the detected single-nucleotide polymorphisms, they observed a great disproportion of nonsynonymous substitutions in the domesticated lines compared with wild relatives. Furthermore, they identified similar patterns in other two domesticated species of the sunflower family (i.e. in globe artichoke and in cardoon). They identified regions in the *H. annuus* genome showing a significant excess of deleterious

mutations. These regions had a lower mean recombination rate than the balance of the genome. On the base of their evidences, they concluded by suggesting that deleterious mutations accumulate in low recombining regions of the genome, as a consequence of the reduced efficacy of purifying selection (Renaut & Rieseberg, 2015).

1.5 Consequences of inbreeding on fitness

The consequences of the inbreeding process consist in fitness effects on genotypes and changes of genotype frequencies in progenies. In general, inbreeding lowers fitness-related characters, such as survival, growth and fertility (Theodorou & Couvet, 2006). In plants, consequences of self-fertilization include mutant phenotypes that are lethal in early stages of life, such as chlorophyll-deficient albino seedlings (D. Charlesworth & Willis, 2009). Inbreeding depression is caused by the increased homozygosity in the progenies as a result of inbreeding: two classical theories explain how increased homozygosity can decrease fitness (Figure 1). The **dominance hypothesis** posits that lowered fitness is due to increased homozygosity of recessive deleterious alleles in the progenies. The **overdominance hypothesis**, instead, suggests that the heterozygous combination of alleles at a locus results in higher fitness to either of the homozygous combinations for that locus. Genetic linkage causes major difficulties in resolving dominant from overdominant effects of loci. A locus may display apparent overdominant action that is actually the result of **pseudo-overdominance**, which occurs when two linked loci, each having a deleterious recessive allele in repulsion, both contribute to the phenotype through complementation (Springer & Stupar, 2007). Thus, both dominant and overdominant hypothesis assume that inbreeding depression is caused by an increased homozygosity of individuals, even if they use genetically distinct models to explain how increased homozygosity can lower fitness. Under the dominant hypothesis, deleterious alleles at a locus are generally present at low frequency in populations, thanks to the counteracting action of selection. In a random-mating population at selection–mutation balance, the expected frequency of a deleterious allele depends on the mutation rate, its effect on fitness and on the dominance of the allele. The lower the mutation rate, the more severe the effect of

the allele, and the more the deleterious effect is expressed in the heterozygous genotype, the lower will be the expected frequency of the allele at equilibrium (D. Charlesworth & Willis, 2009). On the contrary, genetic variation at overdominant loci is maintained at intermediate frequencies by the action of balancing selection (e.g. heterozygote advantage): the high fitness of the heterozygote favours the persistence of allelic polymorphisms in the population. The term segregation load refers to the low-fitness homozygous genotypes produced by the breeding of heterozygotes (Carr & Dudash, 2003; Henn et al., 2015).

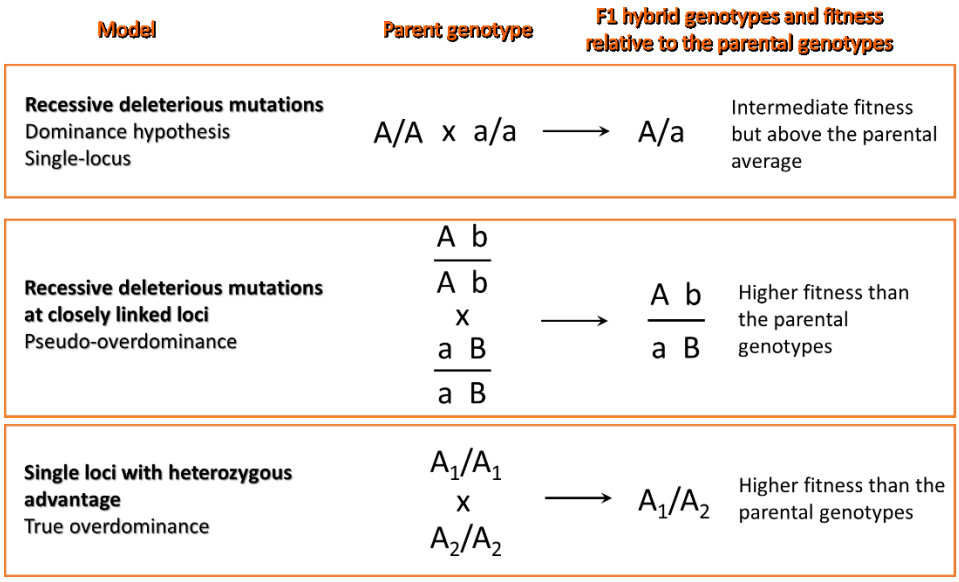


Figure 1. **Main genetic hypothesis for inbreeding depression.** In the **dominance** model, inbreeding depression results from the increased homozygosity of recessive deleterious alleles, the effects of which are masked by dominant alleles in heterozygotes. In the **overdominance** model, heterozygotes at a given locus have an advantage over homozygotes and the loss of heterozygosity in inbred progeny results in inbreeding depression. **Pseudo-overdominance** occurs when repulsion-phase linked loci both contribute via complementing action of dominant alleles to the fitness of heterozygotes, while deleterious recessive alleles at each of the two loci cause decreased fitness of homozygotes.

1.6 Segregation distortion

Segregation distortion (SD) is the deviation of the observed genotypic frequencies in the progenies of a cross from the expected Mendelian frequencies. Imbalanced representation of parental alleles in a segregating progenies can result from different mechanisms: differential inclusion of alleles in the products of meiosis, differential survival or fertilization success of gametes, or differential survival of the zygote. Non-random inclusion of alleles in the gametes during meiosis is defined as meiotic drive. Meiotic drive is thus a pre-meiotic mechanism that implies post-meiotic effects as preferential fertilization or abortion of the gametes or the zygote (Buckler IV et al., 1999; Fishman & Willis, 2005). Post-meiotic transmission distortion systems can be described as reproductive barriers, which act through pre- or post-zygotic mechanisms of selection. Competition among gametes for preferential fertilization (e.g. pollen tube competition) and pollen-pistil incompatibilities are two mechanisms generating differential selection at the gametic level. Several systems of self-incompatibility are used by plants to prevent self-pollination and consequent inbreeding by the recognition of "self" pollen and the inhibition of incompatible pollen tube growth (Franklin-Tong & Franklin, 2003). Hybrid sterility or hybrid inviability genes represent post-zygotic barriers that cause abortion of the zygote in inter-specific cross (Bodenes, Chancerel, Ehrenmann, Kremer, & Plomion, 2016; Fans, Laddomada, & Gill, 1998). Chromosomal rearrangements, which are frequent in inter-specific cross, may also result in abnormal meiotic products with negative effects on hybrid fitness (Myburg, Vogl, Griffin, Sederoff, & Whetten, 2004). The underlying reason of segregation distortion in progenies of crosses is the genetic load: skewed genotypic ratios may arise in progenies for the accumulation of deleterious mutations that reach homozygosity and decrease viability, as described in paragraphs 1.4 and 1.5. Small sample size and genotyping errors are instead non-biological factors that can alter measurement of segregation frequencies in a population (Alheit et al., 2011).

If the fitness of a genotype at a given locus decreases, frequency of the causative damaging allele is expected to lower in the segregating population: since the locus is in linkage with the neighbouring ones, distortion will extend to a cluster of markers surrounding the causative locus (Alheit et al., 2011). Thus, population size matters when analysing segregation distortion: analysis

on large size populations allows the detection of finer causative loci, thanks to a higher probability of recombination events.

Segregation distortion can be assessed not only by analysing single independent loci, but also by looking at pairs of loci, seeking epistatic interactions. Epistasis was firstly described by William Bateson in 1909 to indicate the discrepancy between the prediction of segregation ratios based on the action of individual genes and the actual outcome of a dihybrid cross (P. C. Phillips, 2008). A locus is said to be epistatic to another when its effect masks, prevents or alters the effect of the allele at the second locus (Cordell, 2002). Indeed, when epistatic interaction between loci occurs, not all possible phenotypic, and thus, genotypic classes are observed in dihybrid progenies and some gene combinations result in novel phenotypes (Carr & Dudash, 2003).

1.7 The *Vitis vinifera* genome and inbreeding depression in grapevine

Vitis vinifera is a dicotyledonous, perennial, mainly vegetatively propagated species whose genome is highly heterozygous and small in size (487 Mbp). The reference genome was sequenced and assembled in 2007 by the French-Italian Public Consortium for Grapevine Genome Characterization (Jaillon et al., 2007) from the highly inbred strain of PN40024, obtained from repeated selfing of Pinot Noir and an accidental out-cross with Helfensteiner (itself deriving from a cross between Pinot Noir and Schiava Grossa). Grapevine has an ancient history of domestication, which started 8-10 thousand years ago in the Transcaucasia area and gave rise to the domesticated *Vitis vinifera ssp. sativa* from its wild ancestor *Vitis vinifera ssp. sylvestris* (Myles et al., 2011). From the primary cradle of domestication, cultivated forms would have spread by humans southwards in the Middle East and in North Africa and eventually westwards in Central and in West Europe. These areas may have constituted secondary centres of grapevine domestication, thanks to the genetic contribution from local *sylvestris* populations (Arroyo-Garcia et al., 2006). As a consequence of domestication, some changes in the morphology of cultivated forms have emerged, including shifting from dioecious flower, typical of wild grapevines, to perfect flowers (Cipriani et al., 2010; Myles et al., 2011). Generation and propagation of varieties

has been carried out largely vegetatively, or by crosses (Myles et al., 2011), giving rise to the wide range of cultivars existing today. The majority of present cultivated grapevines with hermaphrodite flowers are also self-compatible. Seed viability and germination rates are reduced upon self-pollination with respect to cross-pollination, even if the two systems are equally efficient in setting seeds (Sabir, 2011). As a result, the development of highly homozygous grapevine varieties is quite difficult due to severe inbreeding depression (Bronner & Oliveira, 1990). Formal experiments assessing inbreeding depression in grapevine are scarcely described in literature and knowledge on the effects of inbreeding mainly derives from the empiric experience of breeders. Progenies of selfed varieties grow weakly and the survival of juvenile plants is reduced compared to plants resulting from the outcrossing. The reduced fitness (low vigour and low fertility) and the premature mortality of seedlings originating from selfing have discouraged their maintenance in cultivation (Cattonaro et al., 2013). Present cultivars are therefore characterized by high levels of heterozygosity, as reported by early studies on the extent of phenotypic variation in progenies of selfed varieties (Snyder & Harmon, 1939) and by more recent studies on the genomic characterization of domesticated varieties and wild species by means of SNP characterization (Myles et al., 2010). The high heterozygosity of the grapevine genome suggests an association with a genetic load of recessive deleterious alleles, which effect is manifested upon self-fertilization (Cattonaro et al., 2013).

The line produced experimentally from the self-pollination of Pinot Noir, and which genome was chosen as the reference for *Vitis vinifera* species, is one documented example of a nearly homozygous grapevine. Nevertheless, effects of inbreeding were observed in the progenies resulting from successive generations of selfing, with progressively increasing level of mortality, reduced vigour and increased sterility. However, the elimination of seedlings that were weak, sterile, and/or with female flowers at each cycle (Bronner & Oliveira, 1990) allowed to obtain the line PN40024, which homozygosity was estimated around 93% (Jaillon et al., 2007). This process of selection should have removed part of the genetic load allowing the reiteration in selfing (Cattonaro et al., 2013).

1.8 Recombination rate in the chromatin context: repeat density, gene density and epigenetic marks

In all eukaryotes, meiotic recombination is an extremely conserved process that has a deep role in shaping genetic variation. Recombination frequency along chromosomes is highly variable: crossovers are less frequently observed in heterochromatic regions, while are more frequent in euchromatic regions (Salomé et al., 2011).

Repetitive DNA sequences compose the constitutive heterochromatin, being localized in the centromeres and in the telomeres, and represent a feature of transposable elements (Bierhoff, Postepska-Igielska, & Grummt, 2014; Termolino, Cremona, Consiglio, & Conicella, 2016). In general, the repeat-containing regions are characterized by a low rate of meiotic recombination (Chen et al., 2008; Pan et al., 2011): crossing over within repeats can promote non-allelic homologous recombination (NAHR) resulting in deleterious genomic rearrangements, such as CNVs involved in neurological disorders in humans (Sasaki, Lange, & Keeney, 2010). Centromeric regions are characterized by lower recombination rate compared to the genome average, as observed in *Arabidopsis* (N. E. Yelina et al., 2012), maize (Shi et al., 2010), and rice (Si et al., 2015). Plant genomes show strong correlations between gene density (which is high in euchromatin) and crossover frequency, especially species with large size genomes, such as maize (Chia et al., 2012), barley (Klaus F X Mayer et al., 2012), wheat (K. F. X. Mayer et al., 2014), and tomato (Sato et al., 2012). All these species show increased crossover frequency and gene density in chromosome arms towards the telomeres (Termolino et al., 2016), distal to the recombinationally suppressed pericentromeric heterochromatin.

Since suppressive epigenetic marks are primarily directed to silence TEs, euchromatin and heterochromatin, which differ in TE content, also differ in their epigenetic signatures (Mirouze et al., 2012). DNA methylation occurs most densely at the centromeric regions, where it is required for transcriptional suppression of repeated sequences (N. E. Yelina et al., 2012). In most eukaryotes, pericentromeric heterochromatin flanking centromere is also enriched in DNA

methylation and histone modifications, repressive for transcription (Simon, Voisin, Tatout, & Probst, 2015).

Epigenetic information is known to influence patterns of meiotic recombination; anyway, the relation between DNA methylation and meiotic recombination in plants is complex. Mirouze et al. (2012) observed that hypomethylation of euchromatic DNA in *Arabidopsis* mutants increased the meiotic recombination rate, suggesting that the further accessibility for the recombination machinery to the already decondensed chromatin could explain the observed increase (Mirouze et al., 2012). Studies in euchromatic regions of hypomethylated mutants of *Arabidopsis* carried out by Melamed-Bessaudo and Levy (2012) led to analogous results. However, the research group did not found any evident change in the recombination rate of heterochromatic regions (Melamed-Bessudo & Levy, 2012). Studies from Yelina and colleagues (2012) instead revealed that mutants of *Arabidopsis* loosing DNA methylation showed an epigenetic remodeling of crossover frequencies, with an increased CO rate in the centromeric regions and compensatory levels in the euchromatic chromosome arms, revealing the presence of homeostatic mechanisms acting on CO distribution and number (N. E. Yelina et al., 2012).

The analysis of the *Vitis vinifera* genome revealed that transposable elements (TEs) and repetitive regions compose more than 40% of the genome. Both class I (retrotransposons) and class II (DNA transposons) elements are present, with a prevalence of class I over class II (Jaillon et al., 2007). Gene density in *Vitis vinifera* is strongly inversely correlated to TE density; however, introns have revealed to be quite rich in TEs (especially LINEs) and repeats, and this may have contributed specifically to the increase in intron size observed in grapevine (Jaillon et al., 2007).

At a macroscopic level, the distribution of methylated cytosines in *Vitis vinifera* was found to correlate positively with TE density, and negatively with gene density (Mirko Celii, PhD thesis, 2016). TE sequences show high levels of methylation both in the CG and the CHG contexts. Methylation of TE sequences in the CHH context is instead extremely low. Regarding gene methylation, transcribed regions of active genes show methylation especially in the CG context. In particular, intronic regions appear more methylated than exonic regions, in contrast with other species such as *Arabidopsis* and humans, suggesting epigenetic silencing of TEs in introns (Mirko Celii, PhD thesis, 2016).

Chapter 2 Objectives

The main objective of the present research work was the identification and characterization of loci controlling viability and survival in the *Vitis vinifera* genome, by means of the analysis of segregation distortion in progenies deriving from self- and out-crosses.

Vitis vinifera species is characterized by small-size, highly heterozygous genome (Jaillon et al., 2007). Present cultivated varieties are characterized by a high degree of genetic diversity, in terms of structural and single-nucleotide variation, and progenies of selfed varieties show severe inbreeding depression (Cattonaro et al., 2013).

Previous studies in the framework of the NOVABREED project provided a detailed description of the high variability of the *Vitis vinifera* species. Through the analysis of a large set of genotypes that are representative of the cultivated varieties present today, grapevine genetic variability was described in terms of single nucleotide polymorphisms (SNPs), small insertion deletion polymorphisms (INDELs) and structural variation (SV). The highly heterozygous genomes of grapevine varieties were shown to harbour a burden of deleterious alleles, whose negative effects may be compensated by the presence of the non-affected copy of the allele.

In order to detect the effects of lethal and highly deleterious mutations on the viability/survival of individuals, segregation was analysed in progenies deriving from the selfing of six varieties and in one progeny of an out-cross. In order to detect segregation distortion, genotyping of the progenies was obtained by means of double digest DNA associated DNA sequencing (ddRAD-seq) technique. Single-locus and two-loci segregation distortion in the progenies was measured, focusing on unfavourable combinations of alleles with severe effects on gene function. Selection and dominance coefficients were measured to analyse mutation severity and effect on fitness.

Furthermore, a fine-scale map of the recombination frequency of the *Vitis vinifera* genome was produced. The estimation of recombination frequency along the genome allowed further characterization of the detected loci. The fine-mapping of candidate loci was carried out at haplotype level in a set of *Vitis vinifera* varieties. Fine-mapped loci were screened for the presence

of mutations affecting gene function, in terms of SVs, INDELs, non-sense and non-synonymous SNPs.

The analysis of epistatic interaction in mapping populations allowed the discovery of a strong signal between chromosome 1 and chromosome 11 in the variety Rkatsiteli. This signal was later identified as a physical interaction between the two chromosomes, as a result of a balanced translocation. To further explore the hypothesis of the balanced translocation, several analyses were performed. The genetic map of Rkatsiteli, revealing the pseudo-linkage signal, and the skewed distribution of the genotypic classes in the progenies of selfing both supported the hypothesis. Further evidences were found through an *in silico* analysis of the balanced translocation: different types of sequencing reads (i.e. previously obtained mate-pair reads and SMRT-seq reads for Rkatsiteli) were found to span the translocation breakpoints on chromosome 1 and on chromosome 11. The analysis of sequencing reads allowed to define both the translocation breakpoints at the base-pair level. Based on the breakpoint coordinates, a PCR assay was carried out in Rkatsiteli in order to validate experimentally the translocation. A screening in a panel of 196 grapevine varieties revealed other three varieties to carry the chromosomal rearrangement. To assess the effect of genes re-location in a new chromosomal context on their expression, the allele specific expression (ASE) analysis was performed. The analysis allowed to compare gene expression in the haplotype carrying the translocation and in the normal haplotype in three tissues of Rkatsiteli.

Chapter 3 Materials and Methods

3.1 In silico identification of deleterious mutations in the grapevine population

The analysis of the genetic variability of the grapevine genome, in terms of both single nucleotide variation and structural variation, was performed previously in the NOVABREED project. The analysis of SNPs and INDELs was carried out in a population of 137 varieties (128 *Vitis vinifera* cultivars and 9 introgression lines, listed in Appendix 1) and produced high-quality SNP/INDEL datasets that were validated with different approaches (Gabriele Magris, PhD thesis, 2016). Based on the primary transcripts of genes of the V2.1 annotation (Vitulo et al., 2014), SNPs and INDELs were functionally annotated using the 2013-08-23 version of Annovar (Wang, Li, & Hakonarson, 2010). Both SNPs and INDELs were classified as intergenic, intronic, splicing, stopgain, stoploss and UTR. Furthermore, SNPs were divided in synonymous and non-synonymous, while INDELs were divided in frameshift and non-frameshift insertions and deletions.

The detection of structural variants (SVs) was also performed previously in the NOVABREED project (Gabriele Magris, PhD thesis, 2016), and was carried out on the 50 variaties of the population with the highest library quality (identified by an asterisk in Appendix 1). Deletions were detected by integrating results obtained by DELLY (Rausch et al., 2012) and GASV (Sindi, Helman, Bashir, & Raphael, 2009); insertions were defined using a pipeline that rely on a database of transposable elements (Pinosio et al., 2016). Deletions occurring outside genes were classified as intergenic. Deletions affecting genes were further divided in three categories: a) exonic, if they involved only exon regions; b) intronic, if they involved only intron regions; c) mixed, if they spanned intronic and exonic regions. Similarly, insertions were classified as intergenic, exonic and intronic.

Full details about the Material and Methods of SNP/INDEL/SV discovery and validation in the grapevine population are described in the PhD thesis of Gabriele Magris.

3.1.1 Identification of mutations affecting gene function

The previously obtained data of genetic variation was the base to generate a list of mutations in genes for each variety of the grapevine population. Both non-sense mutations (stopgain and stoploss SNPs and frameshift INDELs), missense mutations (nonsynonymous SNPs with deleterious effect, described in the next section) and structural variation in genes (insertions and deletions in exonic/intronic/mixed regions) were included in the analysis. Each gene (i.e. the primary transcript from the V2.1 annotation) affected by one or more mutations in at least one of the varieties was included in the list. Furthermore, transcriptome analysis (described in detail in the section *RNA sequencing experiments* of par. 3.2.5) in a panel of six *Vitis vinifera* varieties was used to integrate the information of mutations in genes with the information of gene expression.

3.1.2 Prediction of the effect of missense mutations

The functional effect of amino acid substitution was predicted for the subset of nonsynonymous SNPs previously detected and annotated in the grapevine population.

Version 1.1.5 of the software PROVEAN (Protein Variation Effect Analyzer, Wang, Li, & Hakonarson, 2010) was used to predict effects of deleterious mutations among nonsynonymous sense SNPs (Magris G, unpublished results). The analysis was carried out in the population of 137 grapevine varieties and two outgroup species, *Vitis rupestris Du Lot* and *Vitis armata*, were also added. PROVEAN is based on the alignment of protein sequences through position-specific iterated (psi) BLAST to identify closely related protein homologs. Then, the software compares alignment scores between the query protein sequence and its homolog sequences before and after the introduction of the amino acid substitution. Prediction is based on the fact that if a

particular amino acid change is infrequently or never seen in a protein alignment with related species, it is more likely to be deleterious. For the present work, protein homologs sequences search was carried out against the NCBI non-redundant database, by restricting the search to Viridiplantae. Functional effect of nonsynonymous substitutions may be biased by the ancestral/derived state of the reference allele, and variants are less likely to be classified as deleterious when the reference allele is derived with respect to the outgroup (Simons, Turchin, Pritchard, & Sella, 2014). In order to correct for this bias in the current approach, positions where the reference allele was derived were identified thanks to the information given by the two outgroup species. Threshold value for discovering deleterious mutations was set to default (significant score must be lower than -2.5).

3.1.3 Frequency spectrum and Tajima's D test

The distribution of the SNP allele frequencies in the population of 137 grapevine varieties was summarized through the frequency spectrum. All the categories of SNPs, based on the classification of Annovar and PROVEAN, were represented in the frequency spectrum. The distributions of the categories of SNPs were pairwise compared using the Wilcoxon Mann-Whitney test (p value < 0.001).

Tajima's D test was performed for the different categories of SNPs, deletions and insertions (as previously described) in order to test the null hypothesis of mutation-drift equilibrium and constant population size. The test was carried out in the subset of 50 varieties chosen for the detection of SVs. Tajima's D test was computed according to the equation described by Fumio Tajima in 1989 (Tajima, 1989).

In order to compute the significance of Tajima's D values, the reference genome was divided in 1,320 windows containing 200 Kb of non-repetitive nucleotides (i.e. sequences not masked by repeat sequence annotators). The *D* value was computed for each class of mutation in each window. Since the grapevine population is not at equilibrium, significance of Tajima's D values should not be tested relative to zero. Rather, the distribution of *D* values was statistically

compared in pairs of categories (e.g. the Tajima's D distribution of deleterious mutations was compared to that of synonymous mutations) with the Wilcoxon Mann-Whitney test (p value < 0.001).

3.2 A segregation distortion analysis in grapevine crosses to identify deleterious mutations

3.2.1 Progenies and phenotype scoring

Crosses and plant material

Vitis vinifera crosses were carried out using field-grown plants at the Azienda Agraria Universitaria A. Servadei in Udine (Italy). Self-fertilization and outcrossing were performed under natural conditions and a single individual per variety was used as the parent variety to obtain progenies. Six varieties were self-fertilized: Cabernet Franc, Pinot Noir, Primitivo, Rkatsiteli, Schiava Grossa, and Sangiovese. In order to carry out self-crosses, inflorescences of each parental variety were enclosed in paper bags before blooming to prevent fertilization by other pollen sources and to favour self-pollination. One out-cross was performed using Rkatsiteli as the male parent and Schiava Grossa as the female parent. In order to accomplish the outcrossing, emasculation of hermaphrodite flowers of the seed parent was performed to prevent self-pollination. For each cluster, calyptra and anthers of each flower were removed two days before the first onset of blooming. Each emasculated cluster was then enclosed in a paper bag until blooming. Then, pollen was collected at noon from multiple individuals of the variety selected as the male parent and was repeatedly applied on stigmas of the emasculated clusters, until the disappearance of stigmas exudate. Pollinated clusters were bagged again.

Primitivo and Rkatsiteli self-crosses were performed on 2012, before the beginning of the current PhD project. Cabernet Franc and Sangiovese self-crosses were performed on 2013. Self-crosses were made for Pinot Noir and for Schiava Grossa on 2014; on the same year, Rkatsiteli variety was

self-crossed to generate a second progeny of selfing. Lastly, the out-cross between Schiava Grossa and Rkatsiteli also was made on 2014.

Seeds were collected from overripe berries between September and October of the respective year; then, they were rinsed with 1.5 % hydrogen peroxide for 24 hours. Cold stratification in cotton wool at 4°C lasted three months. On March of the respective following year, seeds were transplanted into cold greenhouse at the Azienda Agraria Universitaria A. Servadei for germination under natural light. Regarding crosses made on 2012, the plant material of the progenies was collected on 2014 (i.e. leaf tissue was collected in the field during the second vegetative season of the survived progenies). For all the other progenies, apical leaflets (when available) or cotyledons were collected within two months after seed germination (between April and May of the respective seedling years, 2014 and 2015). At this timing, seedlings were growing in the greenhouse. The following number of progenies was sampled from each cross:

- Cabernet Franc self-cross: 68 progenies
- Pinot Noir self-cross: 85 progenies
- Primitivo self-cross: 62 progenies
- Rkatsiteli self-cross 1: 86 progenies
- Rkatsiteli self-cross 2: 152 progenies
- Sangiovese self-cross: 87 progenies
- Schiava Grossa self-cross: 91 progenies
- Schiava Grossa X Rkatsiteli: 192 progenies

Marker segregation in progenies of selfing followed an F2-like design (codominant markers *ab* x *ab* that segregate in a 1:2:1 ratio); whereas segregation of codominant markers in the progenies of the outcross was analysed according to a two-way pseudo testcross design (further described in par 3.2.3; see Figure 2 for the explanation of the strategy using dominant markers).

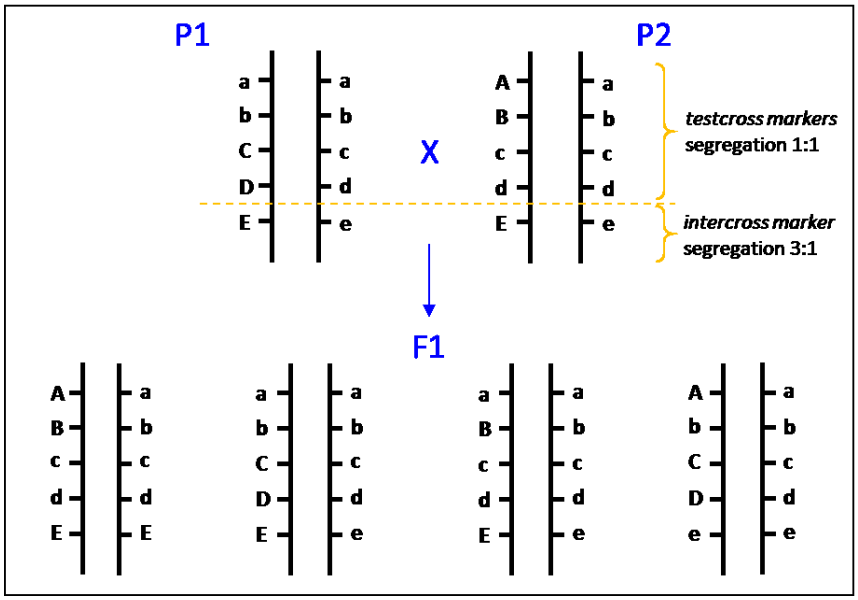


Figure 2. **The two-ways pseudo testcross approach with dominant markers.** Parental varieties P1 and P2 are highly heterozygous: P1 is heterozygous for the testcross markers *C* and *D*; P2 for the testcross markers *A* and *B*. The reciprocal parent is homozygous for a "null" allele. In the F1 progenies, testcross markers segregate in a 1:1 ratio. The intercross marker *E* is heterozygous in both parental varieties and segregates in a 3:1 ratio in the progenies.

Seedling sampling and phenotyping

Stratified seeds were maintained in the greenhouse to induce germination. At the respective seedling year of each population, leaves or cotyledons were harvested from seedlings within two months after the beginning of germination.

Phenotyping was performed for the progenies of selfing and they were scored for survival after germination. Sampled seedlings, that were successively genotyped, constituted the fraction of germinated seeds which survived within two months after germination (this timing was called T_0, Figure 3). Ideally, a single subapical leaf was removed from seedlings at the stage of two to three true leaves. In case of seedlings with weak or delayed growth, one or both cotyledons were collected, saving leaves. In extreme cases of seedlings which growth was dramatically impaired, sampling of the entire epicotyl allowed to obtain the genotypic information at the expense of the

phenotypic observation. These individuals were classified as “no phenotype”, because survival could not be assessed past two months after germination.

Phenotyping carried out on May 2015 (for the progenies of selfing performed on 2013) consisted in scoring survival at the beginning of the second vegetative season, after the first overwintering, and this time was called T_2. A refinement of phenotype scoring was made for progenies which second vegetative year was set on 2016. It consisted in scoring survived individuals at the end of the first vegetative season, before winter season, at a time called T_1 (between September and October 2015). Scoring was performed also at T_2 in these progenies. In addition to survival, seedlings were scored for vigour. Parameters of stem lignification and diameter, and buds sprouting were taken into account to score vigour (Figure 3). According to these parameters, seedlings were categorized into three classes:

- Normal phenotype: lignified stem, stem diameter higher than 3 mm and many buds giving rise to shoots. This phenotype was related to a plant height of approximately 15-20 cm, based on shoot length.
- Weak phenotype: lignified or partially lignified stem, stem diameter lower than 3 mm and one or few buds giving rise to short shoots. Plantlets showing this phenotype were shorter than plantlets from previous category (shoot length not exceeding 10 cm).
- Very weak phenotype: not lignified stem, stem diameter lower than 1 mm and no living buds, no resumption of vegetative growth after winter. Plantlets were classified as dead at T_2.

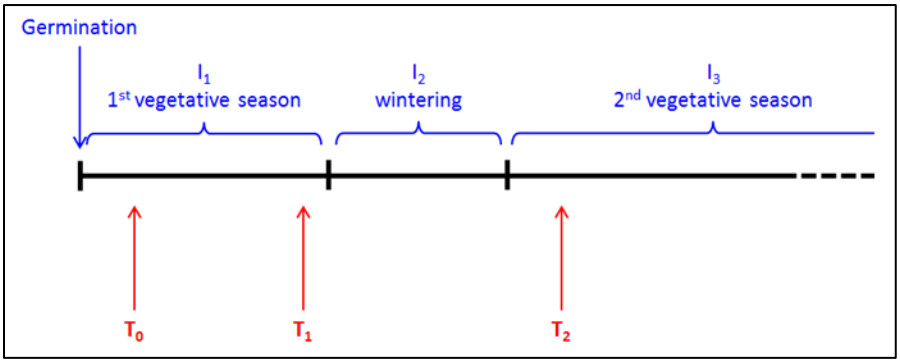


Figure 3**. Timing of phenotypic scoring.** Among seedlings survived to germination at T_0 (genotyped within 2 months after germination), seedlings phenotypes were scored at T_1 (end of the first vegetative season –

I_1, approximately five months after germination) and at T_2 (after wintering - I_2, approximately thirteen months after germination).

3.2.2 ddRAD-seq and DNA-seq markers

DNA extraction

After labelling seedlings, apical leaflets - when available - or cotyledons were harvested and stored at -80°C. Leaf tissue, kept at low temperature in dry ice, was ground with stainless steel beads using TissueLyser (Qiagen, Hilden, DE), set at 20 Hz for 1 minute.

Genomic DNA extraction protocol was modified from Doyle JJ & Doyle JL (1990). Lysis buffer (2.5% w/v CTAB, 8% w/v NaCl, 125 mM Tris-HCl pH 8.0, 25 mM EDTA pH 8.0, 0.1% v/v B-mercaptoethanol, final concentration) was added to ~100 µg of ground tissue supplemented with 1-2% w/w polyvynilpolypyrrolidone (PVPP) powder. Samples were left in water bath for 20 minutes at 65°C and mixed frequently. After cooling down samples at room temperature (RT), chloroform and isoamyl alcohol (24 : 1 v/v) were added and samples were shaken for ten minutes. Samples were then centrifuged at 16,000 g for 25 minutes at RT to obtain phase separation. DNA containing aqueous phase was transferred from tube to a fresh plate. 0.1 M sodium acetate, 0.4 M NaCl and 40% v/v cold isopropanol (final concentration) were added to DNA and samples were stored O/N at -20°C. After centrifuging plate at 2000 g for 30 minutes at 4°C and discarding surnatant, pelleted DNA was washed twice by adding 70% v/v cold ethanol and centrifuging at 2000 g for 20 minutes at 4°C. After removing any trace of ethanol, high-salt TE 1X (10 mM Tris-HCl pH 8.0, 1 mM EDTA pH 8.0 and 0.7 M NaCl, final concentration) supplemented with 1 µg RNAse A (Sigma-Aldrich, St. Louis, MO) was added to samples. DNA was then re-suspended at 37°C for 15 minutes. 0.7 M ammonium acetate (final concentration) and 70% v/v cold ethanol were added to re-suspended DNA and samples were stored at -80°C for 5 minutes. After centrifuging plate at 2000 g for 30 minutes at 4°C, DNA was washed twice with 70% v/v cold ethanol (as mentioned before). After drying samples, DNA was re-suspended in sterile milliQ water. DNA quality (260/280 and 260/230 ratios) was checked using Nanodrop (Thermo Scientific,

Waltham, MA), while concentration was measured using Qubit 2.0 Fluorometer (Invitrogen, Carlsbad, CA).

ddRAD-seq libraries

Double digest restriction site associated DNA sequencing (ddRAD-seq) is part of the family of Genotyping-By-Sequencing (GBS) techniques (Elshire et al., 2011), whose aim is to subsample the genome at homologous locations to identify and type SNPs evenly throughout the genome. The ddRAD-seq method allows to sequence a high number of samples by selecting a precise fraction of the genome in all samples by means of two restriction enzymes. This method allows to obtain a flexible number of markers, based on the choice of restriction enzymes and the fraction of the genome to size-select. Taking advantage of two-tiers indexing, ddRAD-seq procedure allows to obtain a high level of multiplexing and, thus, to reduce sequencing costs.

ddRAD-seq libraries were prepared for progenies deriving from self- and out-crosses of *Vitis vinifera* varieties (number of progenies is listed in par. 3.2.1), as well as for the parent varieties of the crosses. Library preparation procedure was modified from Peterson BK et al. (2012) and was performed as follows (Peterson et al., 2012; see also Figure 4).

Double-Digestion: 250 ng of genomic DNA were digested using 2.4 U of rare-cutter SphI-HF and 2.4 U of common-cutter MboI restriction enzymes (New England Biolabs, Ipswich, MA). Reaction was carried out for 90 minutes at 37°C in a total volume of 30 µL. Restriction enzymes were then inactivated for 20 minutes at 65°C. Digested DNA was purified using 1.5 volumes of Agencourt AMPure XP beads (Beckman Coulter, Brea, CA).

Adapter Ligation: a set of 24 different *adapters P1* designed to bind to SphI-HF restriction site was used. *Adapters P1* are characterized by inline barcodes that varies in length and in sequence to obtain samples multiplexing (intra-pool multiplexing). *Adapter P2*, binding to MboI restriction site, was common to all samples. *Adapter P2* contained a Y-shaped structure that prevented amplification of common fragments MboI-MboI and in turn facilitates positive selection of fragments generated by cuts with both enzymes. Digested DNA was ligated to 1 pmole *adapter P1* and to 3 pmole *adapter P2* using 160 U of T4 DNA ligase (New England Biolabs, Ipswich, MA).

Reaction was carried out for 60 minutes at 23°C and for 60 minutes at 20°C in a total volume of 30 µL. T4 DNA ligase was then inactivated for 20 minutes at 65°C.

Pooling and Size Selection: after ligation, sets of 24 samples were created by pooling ~150 ng of DNA from samples individually barcoded with a unique P1-adapter. Pools were cleaned with 1.5 volumes of AMPure XP beads and size selected manually, by selecting 300-450 bp fragments range (adapters included) from 1.5% agarose gel. Gel extraction was carried out using MinElute Gel Extraction Kit (Qiagen, Hilden, DE).

Amplification and Sequencing: pools of samples were amplified through PCR reaction. The following primers pairs were used: *primer PCR1* annealed to *adapter P1* tail, while *primer PCR2* sequence, containing TruSeq index for Illumina sequencing, was complementary to *adapter P2* tail. The addition of unique sequencing indexes to each pool of 24 differentially labelled samples allowed to obtain an inter-pool multiplexing level. ~20 ng of DNA were amplified using 5 pmole of common *primer PCR1* and 5 pmole of barcoded *primer PCR2*. Reaction was carried out in a total volume of 20 uL, with 10 amplification cycles, using the following programme: denaturation for 30 seconds at 95°C, annealing for 30 seconds at 60°C and extension for 45 seconds at 72°C. Amplified DNA was purified using 1.2 volumes of AMPure XP beads and checked for size selection at Agilent 2100 Bioanalyzer using DNA 1000 Assay (Agilent Technologies, Santa Clara, CA).

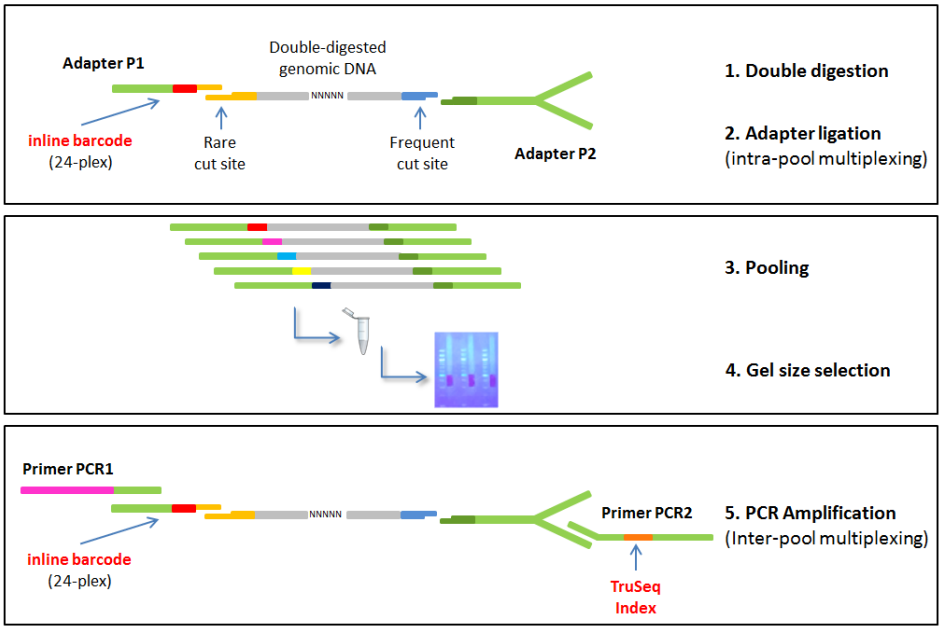


Figure 4. **ddRAD-seq workflow**. Genomic DNA was digested using two different restriction enzymes, a rare and a frequent cutter (1). Ligation of fragmented DNA to adapters containing different inline barcodes labelled uniquely each sample (2). Differently labelled samples were pooled together (3) and size selected (4). Each pool was then amplified using primers containing Illumina sequencing indexes (5).

DNA-seq libraries

DNA-seq technique was used to construct libraries from highly homozygous individuals deriving from the selfing of four *Vitis vinifera* varieties (Cabernet Franc, Primitivo, Rkatsiteli, and Sangiovese) with the goal of reconstructing chromosome-level haplotypes. Four or five individuals for each progenies under study were chosen in order to reach the highest percentage of homozygosity along the genome. Whole-genome libraries were sequenced at low-coverage (6-9 X) and data were used to extrapolate high-density haplotype information (see the section: *High-density haplotype phasing: SNPs, deletions, insertions and INDELs* of par 3.2.5).

DNA-seq procedure was carried out as follows. ~300 ng of genomic DNA were fragmented using Bioruptor sonicator (Diagenode, Liege, BE) and purified with 0.7 volumes of AMPure XP beads to obtain a fragment size range greater than 300 bp. 40 ng of fragmented DNA were used as input for library construction. DNA-seq library preparation was performed following RUBICON -

ThruPLEX DNA-seq Kit standard procedure (Rubicon Genomics, Ann Arbor, MI) and library insert size was checked at Caliper GX using High Sensitivity Assay (PerkinElmer, Waltham, MA).

Both ddRAD-seq pools and DNA-seq libraries were sequenced at the Institute of Applied Genomics (Udine, Italy) on Illumina HiSeq2500 platform using *HiSeq PE Cluster Kit v4 cBot* or *TruSeq Rapid Cluster Kit PE* (Illumina, San Diego, CA), to obtain 125 bp and 100 bp paired-end reads, respectively. The CASAVA 1.8.2 version of the Illumina pipeline was used to process raw data.

ddRAD-seq data processing and genotype calling

For the analysis of RAD-type short read sequences, Stacks software package version 1.35 was used. This software was created by Catchen J et al. (2013) to work with data obtained from reduced representation genotyping techniques to identify loci, either de novo or from a reference genome, and to call genotypes using a maximum likelihood statistical model (Catchen et al., 2013). First, reads were processed in order to demultiplex samples according to Illumina TruSeq indexes (inter-pool labels) and *adapter P1* inline barcodes (intra-pool labels). Through the *process_radtags* programme of Stacks pipeline, *adapters* were removed and reads were trimmed to 85% of their length. Reads were then aligned to the grapevine reference genome (obtained from the highly inbred strain PN40024 by the French-Italian Public Consortium for Grapevine Genome Characterization, Jaillon et al., 2007) using Bowtie2 software package version 2.0.2 with default settings (Langmead & Salzberg, 2012). Running the *pstacks* programme, set of aligned reads were used to assemble loci. For each individual of the progenies and for the parent(s), stacks of reads were grouped into bi-allelic loci. *pstacks* used information on small deletions and insertions (INDELs) in the reads from the CIGAR string of each alignment to keep read length constant. A deletion in the read with respect to the reference was filled in by Ns to regain phase with the reference and the end of the read was trimmed. Conversely, an insertion in the read relative to the reference was trimmed out and the end of the read was padded by Ns. In this way, all reads had the same length and could stack properly. Alleles were retained using a minimum stack depth of three reads. Then, irrespective of the reference sequence, polymorphic nucleotide sites were identified within each stack. SNP calling of Stacks pipeline worked by estimating the maximum-

likelihood value of the sequencing error rate at each nucleotide position, in order to discard false negatives due to sequencing errors and retain true positives. On the base of bi-allelic loci identified in the parents, *cstacks* programme generated a *Catalogue* of segregating loci. Since more than one SNP site can be present in a RAD locus, the *Catalogue* was built by means of called haplotypes in the parents, based on genomic coordinates. Segregating loci in the progenies were obtained by matching samples haplotypes against the *Catalogue*, requiring at least 70% of successfully genotyped samples (*sstacks* programme). A minimum coverage of six reads was applied to retain homozygous calls. Lastly, allelic states were converted into a set of mappable genotypes using *genotypes* programme of Stacks pipeline. An accurate data cleaning was applied on datasets. First, ddRAD loci located in regions characterized by repeats and microsatellites were removed from the analysis. Repetitive regions were previously defined at the Institute of Applied Genomics using ReAS (R. Li et al., 2005) and Sputnik (Abajian, 1994) annotations and an internal database of transposable elements (Dario Copetti, PhD thesis). Second, SNP calls in regions where parental varieties are known to be homozygous were discarded. Regions of homozygosity were defined in each variety by visual inspection of graphics reporting the distribution of the number of heterozygous SNPs along the genome and by defining an empirical threshold based on coverage. The examination of heterozygote SNPs in grapevine varieties was performed by counting occurrences within 2,367 windows containing 100 Kb positions not masked by repeat sequence annotators (Gabriele Magris, PhD thesis, 2016). Since bi-allelic loci were evaluated, interspersed single monomorphic loci (i.e. non-contiguous loci showing mono-allelic state) represented another source of error. Third, to discard poorly genotyped data, dataset was cleaned from samples having more than 20% of missing ddRAD loci and from ddRAD loci missed in more than 20% of the samples (function *clean.geno* in phd_fornasiero.r available on https://github.com/Novabreed/alicefornasiero-phd-code). Functions collected in phd_fornasiero.r were created using the version 3.2.3 of the R language (R Core Team, 2015). Thresholds used for the datacleaning were chosen after comparing several thresholds combinations to optimize balance between data quality and data amount. Lastly, ddRAD loci were refined manually: consecutive loci showing phasing incoherence (i.e. having phase opposite to that of adjacent genotype blocks, see paragraph "*Low-density haplotype phasing*"), possibly because of misalignments, were removed from analysis.

Low-density haplotype phasing

Stacks software assigned arbitrarily the haplotype phase to loci called in the progenies. This means that, at each locus, phase assignment was independent from the phase of neighbouring loci. Genotyped individuals deriving from the selfing were expected to inherit large haplotype blocks, since it is unlikely that a crossing-over event occurred between two adjacent markers. Hence, it was expected that blocks of contiguous loci had the same haplotypic phase. ddRAD-seq haplotypes were phased using an internally developed function in R (function *phase.geno* in phd_fornasiero.r available on https://github.com/Novabreed/alicefornasiero-phd-code). In the used approach (Figure 5), each locus was compared to the previous one and to the following one: three consecutive ddRAD loci were considered at once among all samples of the progenies. Within each triplet, haplotype cis/trans status of loci was counted across all samples for each possible pair (pair of loci 1+2, 1+3 and 2+3). If pairs of loci having the same haplotypic phase (*cis* occurrences) exceeded a defined threshold (default being set at 0.2) with respect to pairs of loci having opposed haplotypic phases (*trans* occurrences), that pair of loci actually had the same haplotypic phase. Conversely, if *trans* occurrences exceeded the threshold, the two loci had opposite phase relative to one another. The use of triplets helped when pair 1+2 could not be resolved due to lack of information. In that case, information on pairs 2+3 and 1+3 was used to increase ability of phasing pair 1+2 as well.

Before phasing

Chr	Position	Sample 1	Sample 2	Sample 3	Sample 4	Sample 5	Sample 6
chr1	402889	kk	kk	hh	kk	hh	kk
chr1	476550	kk	kk	hh	kk	hh	kk
chr1	476850	hh	hh	kk	hh	kk	hh
chr1	503938	kk	kk	hh	kk	hh	kk
chr1	585801	hh	hh	kk	hh	kk	hh
chr1	660862	kk	kk	hh	kk	hh	kk
chr1	712836	kk	kk	hh	kk	hh	kk

After phasing

Chr	Position	Sample 1	Sample 2	Sample 3	Sample 4	Sample 5	Sample 6
chr1	402889	kk	kk	hh	kk	hh	kk
chr1	476550	kk	kk	hh	kk	hh	kk
chr1	476850	kk	kk	hh	kk	hh	kk
chr1	503938	kk	kk	hh	kk	hh	kk
chr1	585801	kk	kk	hh	kk	hh	kk
chr1	660862	kk	kk	hh	kk	hh	kk
chr1	712836	kk	kk	hh	kk	hh	kk

Figure 5. **Schematic illustration of low-density haplotype phasing**. *h* and *k* represent two allelic states. *hh* and *kk* represent homozygous genotypes for *h* and *k* alleles, respectively. Samples 1, 2, 4, and 6 all have the same phase, opposed to the phase of samples 3 and 5. However, before phasing, phase assignment at each locus is arbitrary with respect to the others in the haplotypic block. After phasing, samples 1, 2, 4, and 6 are homozygous for *k* allele in the reconstructed haplotype block; while samples 3 and 5 are homozygous for allele *h*.

DNA-seq data processing and SNP/INDEL calling

Raw paired end reads obtained by low-coverage sequencing of DNA-seq libraries were processed as follows. Adapters were masked using Cutadapt version 1.4.1 (Martin, 2011) in order to check for contaminations using Check Contaminants pipeline version 1.3 (internally developed at Institute of Applied Genomics). Raw sequences were then quality trimmed using erne-filter version 1.4.6 (Del Fabbro, Scalabrin, Morgante, & Giorgi, 2013) and reads that matched chloroplast contaminants were filtered out using BBTools programme BBDuck2 (Bushnell B, Joint Genome Institute, CA). Short read sequences were aligned against the *Vitis vinifera* genome sequence through the software package Burrows-Wheeler Aligner (BWA) version 0.7.10 (H. Li & Durbin, 2010). Alignments were carried out using the *mem* command (which performs local

alignments using the maximal exact matches algorithm) with default settings (minimum seed length 19, mismatch penalty 4, gap open penalties for deletions and insertions 6, gap extension penalty 1, Picard tools compatibility). Alignment output in Sequence Alignment/Map (SAM) format was transformed into Binary Alignment/Map (BAM) format and alignments were sorted using the software package SAMtools version 0.1.19 (H. Li & Durbin, 2009). A series of tools of the Picard suite version 1.78 (http://broadinstitute.github.io/picard) was then utilized for manipulating SAM/BAM data: *CleanSam* to clean alignment data by soft-clipping beyond-end-of-reference alignments and marking unmapped reads, *CollectAlignmentSummaryMetrics* to get some quality control metrics, *CollectInsertSizeMetrics* to get the insert size distribution of paired-end libraries. For each sample, mean coverage was calculated by dividing the total number of uniquely aligned bases by the number of covered positions. Reads with a mapping quality below 10 were discarded using SAMtools, and quality-filtered reads were checked for mate-pairing using Picard *FixMateInformation* tool. Both PCR and optical duplicated reads were identified and discarded through Picard *MarkDuplicatesWithMateCigar* tool. Single nucleotide polymorphisms (SNPs) and small insertions and deletions (INDELs) were called using *UnifiedGenotyper* tool from GATK (DePristo et al., 2011; Van der Auwera et al., 2013) with default parameters. Subsequently, SNP/INDEL calls were filtered by discarding those falling in regions rich in repeats and microsatellites (as previously mentioned).

3.2.3 Genetic linkage maps construction

Segregating populations deriving from the selfing and the outcrossing of grapevine varieties were used to generate genetic linkage maps. Genetic maps of Rkatsiteli and Schiava Grossa varieties were constructed using genotypic data obtained from ddRAD loci segregation. Segregation of codominant markers *ab x ab* (1:2:1 segregation ratio) in the progenies of selfing was used to generate the map of Rkatsiteli. A two-way pseudo testcross approach was used to map testcross markers in the progenies of the out-cross Schiava Grossa x Rkatsiteli. The two-way pseudo testcross strategy (Grattapaglia & Sederoff, 1994) consists in crossing two highly heterozygous individuals and analysing the parental origin and the genetic segregation of markers in the

progenies, allowing the construction of a genetic map for each parent. In the progenies of the out-cross, two separate data sets of testcross markers (1:1 segregation ratio) were obtained, one for each parent (*ab* x *aa* and *aa* x *ab*; see Figure 2) and recombination was calculated separately for the female and for the male parent. Furthermore, *ab x ab* markers were informative in half of the cases, because only *aa* and *bb* genotypes could be used to reconstruct segregation.

R/qtl package (Broman, 2010) was used in the R language environment version 3.2.3 to group markers into linkage groups. The 19 linkage groups corresponding to the chromosomes of the *Vitis vinfera* genome were constructed using a maximum recombination fraction of 0.15 and a minimum LOD score of 18. RECORD (REcombination Counting and ORDering) software package (Van Os, H., P. Stam, R. G. F. Visser and H. J. van Eck, 2005) was utilized to find the relative order of markers within each linkage group for the self-cross data sets. MSTmap (Minimum Spanning Tree) software package (Wu, Bhat, Close, & Lonardi, 2008) was used for ordering markers of the out-cross dataset. In both cases, Kosambi's mapping function was used to convert recombination frequencies into map distances in centiMorgan (cM). Two rounds of markers ordering were performed. An implementation of the SMOOTH software package (described in Van Os, H., P. Stam, R. G. F. Visser and H. J. van Eck, 2005) developed at the Institute of Applied Genomics (Scaglione D, unpublished methods) was used after the first round of ordering to identify and correct genotyping errors.

3.2.4 Identification and characterization of single-locus segregation distortion in progenies of crosses

Single-locus segregation distortion

To detect single-locus segregation distortion (SD) in the progenies of selfing, marker segregation was tested for significant deviation from the expected ratio of 1:2:1. Deviation of the observed genotype frequencies from the expected frequencies was tested through the Chi square Goodness of Fit test, with two degrees of freedom. In the progenies of the out-cross, testcross markers *ab* x *aa* or *aa* x *ab* were tested for deviation from the 1:1 segregation ratio, intercross

markers *ab x ab* were tested for deviation from the 1:2:1 ratio, *ab* x *bc* and *ab* x *cd* markers were tested for deviation from the 1:1:1:1 ratio. Chi square Goodness of Fit was used to test for significance, using a number of degrees of freedom according to the marker type. Considering, on average, two recombination events for each chromosome (being 19 the number of chromosomes in the *Vitis vinifera* genome), 38 independent genomic regions were expected (Bodenes et al., 2016). Therefore, according to Bonferroni correction (Bonferroni, 1936), a significance level of ≈0.0013 (i.e. 0.05/38) was required to obtain a genome-wide error rate of $\alpha = 0.05$. However, a more stringent threshold ($\alpha = 0.001$) was applied.

Estimation of genotype fitness

Estimation of genotype fitness was performed as described in the work of Morton N E et al. (1956). Fitness was calculated by measuring the coefficients of selection (*s*) and dominance (*h*) for the putative lethal/deleterious allele in each region of single-locus SD (Newton E Morton, Crow, & Muller, 1956). For each locus of SD, expected genotypic frequencies were calculated based on the observed frequencies.

The model used to estimate genotype fitness is reported in Table 1: *s* defines the probability of death of genotypes homozygous for the lethal allele and *h* is the measure of dominance, being 0 for a completely recessive locus and 1 for a lethal allele with complete dominance. Consider A is the allele carrying a lethal/deleterious mutation and B is the wild-type allele. In the general model illustrated in Table 1, fitness of genotype BB is 1; fitness of AA is 1-*s*; and fitness of AB is 1-*sh*. Coefficients *s* and *h* vary depending on the segregation mode of the locus. *t* is the selection coefficient against the other allele (B in our example) in overdominant loci, where the genotype with the highest fitness is AB.

- For a completely recessive locus, the homozygote for the lethal allele has fitness 0, since the selection coefficient *s* is 1. The value for the dominance coefficient *h* is 0; thus, the heterozygote and the homozygote for the wild type allele have fitness 1.
- For a completely additive locus, homozygote for the lethal allele has fitness 0, and *h* is exactly 0.5: this means the allele has additive effect and fitness of the heterozygote is 0.5.

- Partially dominant locus is an intermediate case between a complete dominant locus (where both *s* and *h* are 1 as consequence of complete penetrance of a dominant lethal allele) and an additive locus where *h* is 0.5. In this case, *h* gets a value ranging from 0.5 to 1 (extremes excepted); while *s* is 1.
- For a complete overdominant locus, the two classes of homozygotes have fitness 0, while fitness of heterozygote is 1, thanks to the complementing action of the two alleles at the locus. In this case, selection acts against both alleles: *s* and *t* are the coefficients that measure selection against each allele, respectively.
- For a partially overdominant locus where distortion has a deleterious effect, both homozygotes have lower fitness (but higher than 0) compared to the heterozygotes, which has the highest fitness (but lower than 1).

genotype	AA	AB	BB
fitness	1-s	1-sh	1

Table 1. **Estimation of genotype fitness**. Model used to calculate fitness of genotypes, where A is the lethal allele and B is the wild-type allele at the locus. *s* is the coefficient of selection against the allele A, while *h* is the coefficient of dominance.

Fine-mapping of candidate loci

In order to fine-map regions of single-locus segregation distortion, the haplotypes of the six parent varieties under study were compared to the haplotypes of the 128 *Vitis vinifera* varieties (listed in Appendix 1), by means of short-block haplotype comparison. Haplotype phase was carried out for each variety using Beagle software package, version 3.2.2 (Browning & Browning, 2007). The haplotype phasing of the varieties used as parents for the crosses was computed with the aim to obtain the maximum possible accuracy within the limit of reasonable computational time (number of iterations was set to 1,000). Phasing of the other varieties of the grapevine population was previously obtained with the same software using 20 iteration cycles. Each parent haplotype was compared to the haplotypes in the population by means of sliding blocks of three SNPs. The fine mapping was performed based on the following assumption: if the parent

haplotype contained a recessive allele with lethal effect, this should not be found in homozygosis in the population. Consequently, fine mapping consisted in determining regions where short-block haplotypes of parental varieties were never present in homozygosis in the population.

Estimation of the recombination frequency and its correlation with other genomic features

Genetic linkage map information was used to estimate the recombination frequency in the *Vitis vinifera* genome. A pseudo-marker of physical distance, that was named "centroid", was estimated for each genetic bin in the genetic maps. The physical position of a centroid was defined as the mean physical distance between the proximal and the distal marker within a genetic bin. Centroid positions were used to measure the physical distance between two consecutive genetic bins. The ratio between the genetic distance (in cM) and the physical distance (in Mb) of consecutive markers was calculated for each genetic map separately. The average ratio between genetic and physical distance (cM/Mb) was calculated in 1,871 sliding windows of 1 Mb (with sliding steps of 200 Kb) in Rkatsiteli and Schiava Grossa.

The mean recombination frequency between two windows randomly sampled 1,000 times gave the null distribution of the recombination frequency along the genome. The average recombination rate in each locus of SD was then compared to the null distribution.

The genome was divided in regions of low, intermediate, and high recombination frequency. This was achieved by classifying windows having values of recombination frequency less than the 33rd percentile as "low frequency", between the 33rd and the 66th percentile as "intermediate frequency" and above the 66th percentile as "high frequency". A comparable number of windows, namely 620, 615 and 636 windows, fell into each of the three categories, respectively.
Significance of the relationship between each genomic feature and the recombination frequency in the three categories was then tested with the Wilcoxon Mann-Whitney test (p value < 0.001).
The genomic features considered for the analysis were the following:

a) the ratio of deleterious over tolerated nonsynonymous SNPs in expressed genes (see the section: *RNA sequencing experiments* of par 3.2.5 for gene expression analysis);
b) the CG and the CHG methylation profiles, which were obtained in previous research work in the NOVABREED project (Mirko Celii, PhD thesis, 2016). The average level of DNA methylation was measured in windows of 200 Kbp in the variety Pinot Noir;
c) the gene density, measured as the number of bp annotated as primary transcripts of genes in the V2.1 annotation (Vitulo et al., 2014); and
d) the repeat density, measured as the number of bp annotated as repetitive elements on the base of the ReAS annotation (R. Li et al., 2005).

3.2.5 Characterization of loci of segregation distortion

High-density haplotype phasing: SNPs, deletions, insertions and INDELs

Vitis vinifera varieties are characterized by highly heterozygous genomes and, depending on the degree of relationship, they can share none, one or both variable-length stretches of homologous chromosomes. In general, the rate of haplotype sharing between two varieties depends on the degree of their genetic relation, determined by the number of generations which separate them from the common ancestor. The looser the relation between two varieties, the higher the number of generations separating them, and the more fragmented the portions of chromosomes shared, as a consequence of recombination events along generations.
Thus, none of the haplotypes of heterozygous samples are necessarily described by the haplotype of the reference sequence. This can be a drawback when knowing haplotype structure is important. Haplotype phasing was performed for four grapevine varieties (Cabernet Franc, Primitivo, Rkatsiteli, and Sangiovese). Haplotype information was important to detect which mutations belonged to the haplotype carrying the distortion in order to obtain a list of putative candidates. Furthermore, information on haplotype phasing in Rkatsiteli was used for the ASE analysis on Rkatsiteli (see par 3.3.5 and 4.3.5).
Haplotype phasing was accomplished based on known information of heterozygous SNPs in the parental varieties (par 3.1) and the respective information of homozygous SNPs in the progenies

of selfing (see section: *DNA-seq data processing and SNP/INDEL calling* in par 3.2.2). Indeed, at heterozygous SNP positions of the parent variety, homozygous SNPs in the progenies made possible to distinguish haplotypes. Progenies of selfing were screened for their percentage of homozygosity along the genome according to the genotyping data of the ddRAD-seq, and four to five individuals were chosen for low-coverage whole-genome DNA-seq. The individuals were selected in order to obtain the highest rate of homozygosity with the information of at least one homozygous genotype. Every homozygous individual provided (ideally) long range information on one of the two haplotypes of the heterozygous parent. The other haplotype was obtained by subtraction from the parental multilocus genotype. To reduce the possibility of errors and to increase the fraction of the genome in which homozygous individuals were sequenced, several subjects were used.

An internally developed approach in R language was adopted to separate haplotypes (function *define.haplo* in phd_fornasiero.r available on https://github.com/Novabreed/alicefornasiero-phd-code). The strategy consisted in comparing homozygous SNP patterns across samples and calculating the best estimate of the two alternative haplotypes. This was done by comparing the SNP homozygote state of each sample relative to one another, based on the fact that two different homozygote states are expected (i.e. homozygote for the reference allele or homozygote for the alternative allele). By analysing consecutive windows of 1,000 SNP positions (averaging 400 Kbp and thus having a low probability of being split by a recombination event), the most frequent pattern of homozygotes across all samples was obtained in each window, thus defining a "consensus". At each position within a window, one haplotype could be distinguished from the other based on the consensus pattern. When one of the two homozygous states was missing, the information of the relative haplotype could be retrieved knowing the other one.

Figure 6 illustrates regions classified as homozygous and heterozygous based on density of homozygous and heterozygous SNPs in windows of 100 Kb non-repetitive DNA, in Rkatsiteli and in the progenies selected for the haplotype phasing. The same approach was adopted for the four grapevine varieties. In the top panel, the parental variety Rkatsiteli is shown: more than 91% of the genome is classified as heterozygous (blue), while less than 9% is classified as homozygous (pink; either homozygous for the reference allele or for the alternative allele). In the middle panel,

one progeny of Rkatsiteli selfing is illustrated: 62% of the genome is classified as homozygous (pink), while 34% as heterozygous (blue). For a small fraction of the genome (4%), the heterozygous or homozygous state was not determined (grey). The bottom panel shows the genome-wide homozygosity (93%) obtained by merging the four progenies of Rkatsiteli selfing.

By adopting the same strategy, heterozygous deletions, insertions, and INDELs previously detected in the parental varieties were integrated within SNP-defined haplotypes. Deletions and insertions which were heterozygous in the parental variety were quantified in progenies of selfing, through a pipeline developed at the Institute of Applied Genomics (Scaglione D, unpublished pipeline). SV quantification was performed using non-quality-filtered alignment data. In this way, reads aligning on repetitive regions and transposable elements, usually discarded by applying quality filters, could be used for the SV quantification. After integrating all variation information deriving from SNPs, SVs and INDELs, the strategy previously described was used to compute haplotype phasing.

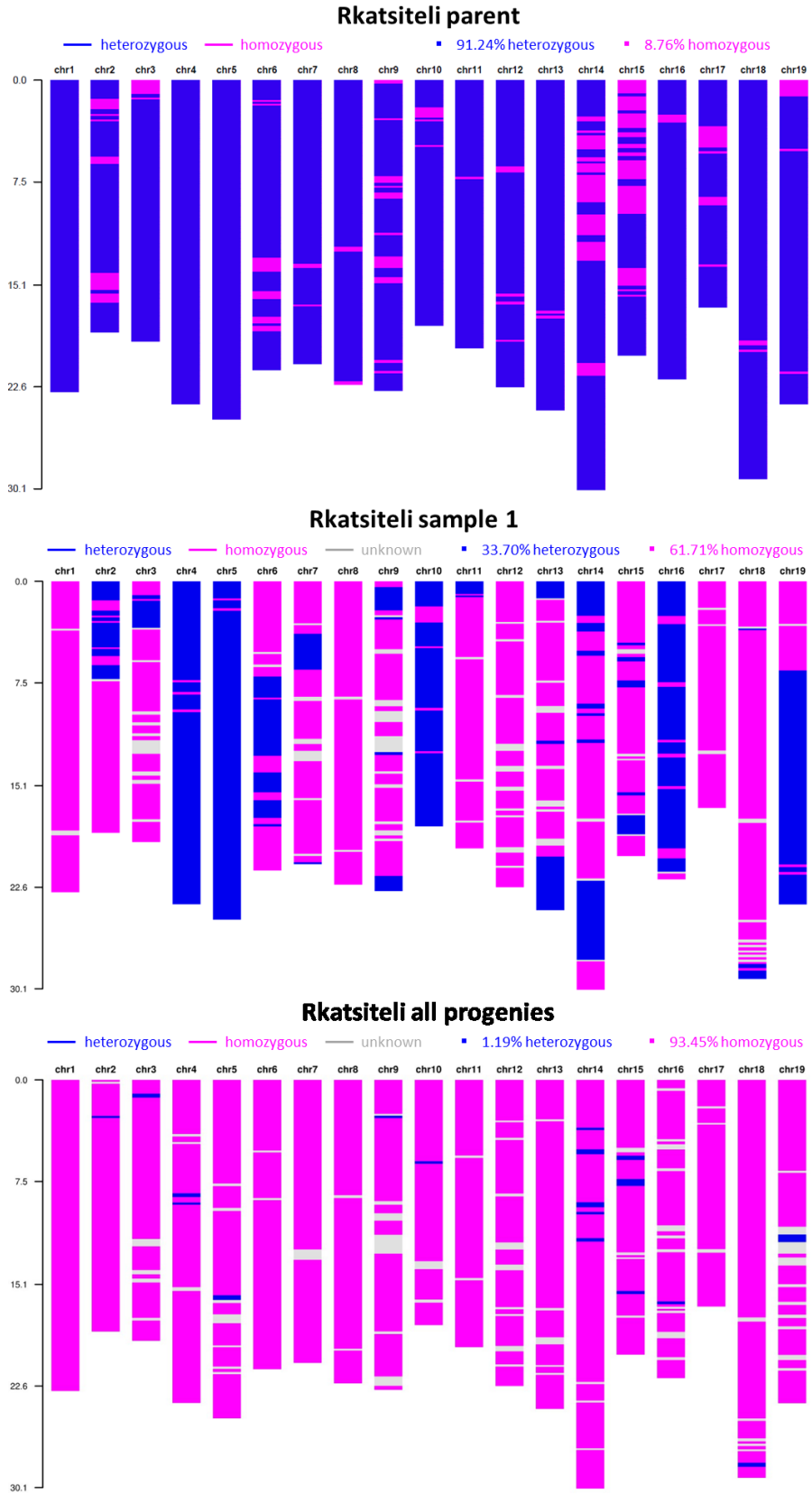

Rkatsiteli parent
heterozygous
homozygous
91.24% heterozygous
8.76% homozygous
chr1 chr2 chr3 chr4 chr5 chr6 chr7 chr8 chr9 chr10 chr11 chr12 chr13 chr14 chr15 chr16 chr17 chr18 chr19
0.0
7.5
15.1
22.6
30.1
Rkatsiteli sample 1
heterozygous
homozygous
unknown
33.70% heterozygous
61.71% homozygous
chr1 chr2 chr3 chr4 chr5 chr6 chr7 chr8 chr9 chr10 chr11 chr12 chr13 chr14 chr15 chr16 chr17 chr18 chr19
0.0
7.5
15.1
22.6
30.1
Rkatsiteli all progenies
heterozygous
homozygous
unknown
1.19% heterozygous
93.45% homozygous
chr1 chr2 chr3 chr4 chr5 chr6 chr7 chr8 chr9 chr10 chr11 chr12 chr13 chr14 chr15 chr16 chr17 chr18 chr19
0.0
7.5
15.1
22.6
30.1

Figure 6**. Homozygous and heterozygous regions in Rkatsiteli parental variety and in the progenies of selfing.** Percentage of homozygosity/heterozygosity is measured in Rkatsiteli (top), in one progeny of selfing (middle), and by merging the four progenies used for the haplotype phasing (bottom). Full details on how homozygosity was measured in the windows are described in the PhD thesis of Gabriele Magris (Gabriele Magris, PhD thesis, 2016). Pink: homozygous windows; blue: heterozygous windows; grey: undetermined windows. In the bottom panel, homozygosity refers to windows where at least one individual is homozygous.

RNA sequencing experiments

Transcriptome analysis was performed previously in the NOVABREED project through RNA-seq experiment on five grapevine varieties - Cabernet franc, Kishmish vatkana, Rkatsiteli, Sangiovese and Traminer (Paparelli E, unpublished results). Transcriptome data was used in the current work to identify which candidate genes in loci of SD were expressed in the parent variety and in the other available varieties.

The experiment was carried out on two biological replicates of three tissues (leaf, berry and tendril). Each biological replicate was handled independently during the RNA-seq library preparation procedure. For RNA extraction, 100 mg of tissue were ground in liquid nitrogen and Spectrum plant total RNA kit (SIGMA, St. Louis, MO) was used. Libraries were prepared using the Low Sample (LS) protocol of Illumina *TruSeq Stranded mRNA Library Prep Kit* (Illumina, San Diego, CA), following standard procedure. Sequencing was performed on Illumina HiSeq2500 sequencer to obtain 100 bp paired-end reads. After adapter removal and quality and contamination filtering, reads were aligned against the reference genome using TopHat2 version 2.0.6 (Kim & Salzberg, 2011). The V2.1 gene annotation of the reference genome (Vitulo et al., 2014) was used for gene annotation. Cufflinks version 2.2.0 (Trapnell et al., 2010, 2012) was used to estimate expression levels in each tissue of the five varieties. To adjust for transcript length and the total number of reads aligned to the transcriptome, expression level was reported as Fragments Per Kilobase of transcript per Million mapped reads (FPKM). RNA-seq data on leaf tissue of Pinot Blanc, Pinot Meunier, Pinot Gris, and Pinot Noir clones was available from previous work (Miculan M, unpublished results) and was added to the NOVABREED transcriptome analysis.

Transcriptome data of Rkatsiteli was also used to perform allele specific expression (ASE) analysis in this variety (see par 3.3.5).

Finding candidate genes for segregation distortion

The grapevine population of 128 varieties was screened for mutations destroying gene function (i.e. non-sense SNP gain and loss, frameshift INDELs, insertions and deletions in gene sequences) and for mutations altering amino acid sites (i.e. nonsynonymous deleterious SNPs) in regions corresponding to loci of segregation distortion in the progenies of selfing. The mutation had to be heterozygous in the parental variety, since lethality is manifested in the progeny of selfing. To restrict the search to alleles with potential lethal effect, mutations should not be present in homozygosis in the grapevine population. Genes affected by one or more mutations with these characteristics were considered as candidate genes for segregation distortion. Among genes screened for mutations, those affected by mutations destroying gene function (as defined above) were selected as more interesting putative candidates. For those loci of SD in which lethal mutations could not be found, genes affected by nonsynonymous SNPs with deleterious effect were considered as putative candidates. The information of high-density haplotype phase was used, when available, to further restrict the list of candidate genes: putative causative mutations have to be present on the haplotype causing distortion.

Candidate gene expression was analysed both in the transcriptome panel of the six varieties and, when available, in the parental variety of the progenies. Threshold for gene expression was set to a value of at least 1 FPKM, in at least one tissue in one variety.

To further characterize putative candidates, an analysis on gene sequence similarity was performed to detect multiple-copy genes (i.e. genes that were duplicated). Translated gene sequences were aligned against the *Vitis vinifera* proteome using BLASTp. A minimum e-value of 1×10^{-50}, a minimum identity of 70% and a minimum alignment length of 70% was set out to consider a gene to be duplicated.

3.3 Identification and validation of a reciprocal translocation

3.3.1 Analysis of epistatic interaction

Segregation distortion was assessed between pairs of independent loci (i.e. located on different chromosomes) searching for epistatic interactions in the progenies of selfing. To detect two-loci interaction, loci were tested for significant deviation from the segregation ratio expected in a dihybrid cross. Fisher's Exact test was applied on a matrix of loci for pairwise comparisons of chromosomes without reciprocal. The *false discovery rate* (fdr) control (Benjamini, Yoav; Hochberg, 1995) was adopted as the correction method for multiple testing and a significance level α of 0.05 was used. To detect significant decrease/lack of the class of double homozygotes, Pearson's Chi square test was used. Expected frequencies of two loci genotypes were calculated as the product of the frequencies of each homozygous genotype at single locus, thus eliminating bias due to single locus segregation distortion. Independent tests were carried out for each of the four classes of double homozygotes. P values were corrected according to Bonferroni, taking into account two crossing-overs per chromosome, all pairs of chromosomes compared pairwise without reciprocal and the number of tests for each pair of loci: $\frac{38*37*4}{2}$.

3.3.2 Pollen germination analysis

Pollen semisterility is a diagnostic to detect a balanced translocation inherited in heterozygosis (Burnham, 1930). To test whether Rkatsiteli variety was heterozygous for the translocation, pollen viability was assessed by means of pollen tube germination. Pollen sampling was carried out at the Azienda Agraria A. Servadei. Grapevine blooming interval time on vintage 2015 lasted eight days. Experiment was carried out by sampling two replicates of pollen for each clone and varieties on three days: June, 1st (batch1), June, 3rd (batch 2), and June, 4th (batch 3). In this way, varieties having different bloom timings could be sampled at their maximum peak of blooming. Nine different clones of Rkatsiteli variety were tested. Cabernet Franc, Kishmish Vatkana, Pinot Noir,

Sangiovese and Traminer were used as control varieties. Two replicates for each clone and for each variety were taken at each day. Pollen grains were grown for two days on 20% sucrose, 2% agar, 100 g/L boric acid medium growth. Measurement of pollen viability was carried out by counting pollen grains developing pollen tube versus pollen grains which did not germinated. Replicates showing clumps of pollen, for which pollen grains could not be isolated from one another, could not be measured.

3.3.3 Structural analysis of a balanced translocation

The translocation breakpoints were reconstructed on the base of paired-end reads obtained from three progenies of Rkatsiteli selfing, homozygous for translocation. Since these individuals were sequenced at low-coverage, reads from each individual were merged to increase the overall coverage. Furthermore, previously sequenced reads of Rkatsiteli parental variety were also used for the breakpoint detection. Both long-insert paired end reads (mate pair) and partially overlapping reads were utilized for junction reconstruction. The larger insert size of mate pair sequencing has the advantage to pair mates across larger distances than conventional paired-end reads. Partially overlapping reads have an insert size which is shorter than the length of both reads, thus extending for a longer range compared to previous read types. Through the contribution of all reads type, both junctions of the reciprocal translocation could be detected and reconstructed. Translocation in Rkatsiteli is carried in heterozygosis; thus, only reads aligning on the haplotype carrying translocation were informative. Lastly, breakpoint reconstruction was confirmed through the long-range single-molecule real-time sequencing technology (SMRT sequencing - Pacific Biosciences Company; Menlo Park, CA). Sequencing of Rkatsiteli variety using this technology was obtained at the Institute of Applied Genomics. Available single-molecule sequencing reads (with a mean length of 4.5 Kbp after the trimming) allowed the reconstruction of long-range phased alleles carrying the reciprocal translocation. Furthermore, these reads were used to resolve mis-alignment in repetitive regions, allowing to characterize the haplotypes with structural variants.

3.3.4 Validation of the balanced translocation through a PCR-based assay

The translocation identified in Rkatsiteli was validated experimentally. The nine clones of Rkatsiteli, four homozygous progenies of selfing, and a panel of grapevine varieties (196 *Vitis vinifera* varieties and *Vitis labrusca* species, listed in Appendix 2) were tested for the translocation using a PCR-based assay. For each variety assessed, four primers were combined in different ways to amplify both normal and translocated chromosomes (Figure 7 and Table 2). Primer pair 1-2, and primer pair 3-4 targeted normal chromosome 1 and normal chromosome 11, respectively. Pair 3-1 targeted region of chromosome 1 (primer 1) translocated on chromosome 11 (primer 3), which was named translocation breakpoint 2. Pair 4-2 targeted region of chromosome 11 (primer 4) translocated on chromosome 1 (primer 2), the translocation breakpoint 1. Primer design was based on the sequence of previously obtained mate pair reads of Rkatsiteli variety, and was performed using BatchPrimer3 (You et al., 2008). Primer pairs were constructed on the sequence of the reads that spanned the two translocation breakpoints and on the neighbouring sequences. Subsequently, primers were evaluated manually to optimize CG content, length, and annealing temperature.

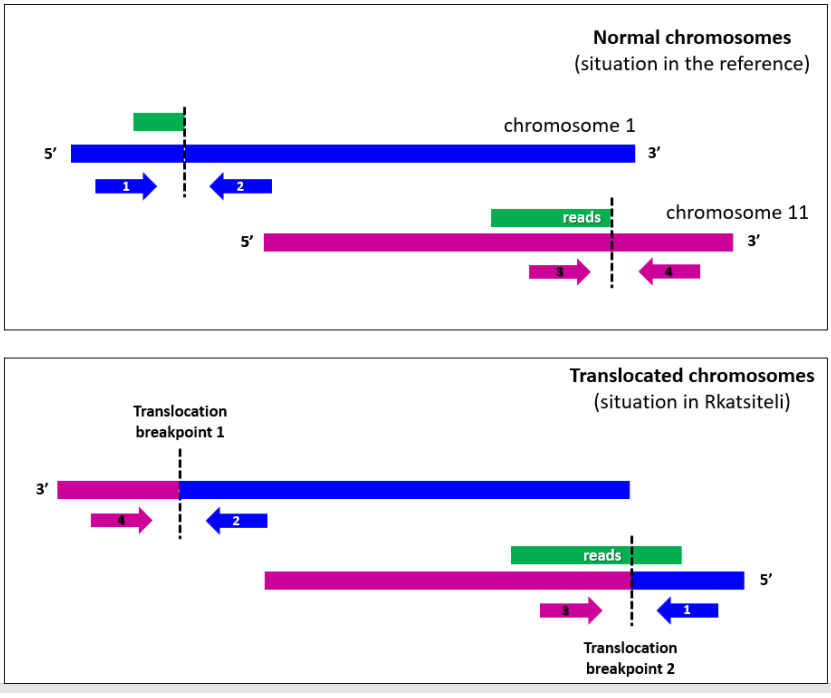


Figure 7. **Primers design to validate translocation.** In the upper panel, chromosome 1 and chromosome 11 are depicted as present in the reference genome (normal chromosomes). Dashed vertical lines on the two chromosomes represent translocation breakpoints. In the lower panel, reciprocal translocation in Rkatsiteli is depicted (translocation involves the initial portion of chromosome 1 and the final portion of chromosome 11). The green box represents sequencing reads spanning the translocation breakpoint 2. Numerated arrows indicated PCR primer pairs used to validate translocation.

Amplification pattern to validate heterozygous translocation				
	Primers pair (forward + reverse)			
	1 + 2	**3 + 4**	**3 + 1**	**4 + 2**
chr 1 native	+	-	-	-
chr 11 native	-	+	-	-
chr 1 translocated on chr 11	-	-	+	-
chr 11 translocated on chr 1	-	-	-	+

Table 2. **Primer pairs required to validate translocation**. Primers pairs 1+2 and 3+4 amplified chromosome 1 and chromosome 11, respectively. Translocation was validated if, at least, primers pair 3+1 amplified.

Regarding amplification reaction, ~20 ng of DNA were amplified using 1X KAPA 2G Fast HotStart DNA Polymerase (KAPABIOSYSTEMS, Woburn, MA) and 0.5 µM of each forward and reverse primers. Reaction was carried out in 10 µL volume. Amplification programme was performed using a touch-down step under the following conditions: 60 seconds at 95°C; 10 cycles of 15 seconds at 95°C, 15 seconds at 65°C (-1°C per cycle) and 60 seconds at 72°C; 20 cycles of 15 seconds at 95°C, 15 seconds at 55°C and of 60 seconds at 72°C, followed by a final extension of 7 minutes at 72°C. Amplified products were loaded onto a 1% (w/v) agarose gel in 1X TAE buffer and run was carried out at 110 V for 60 minutes.

3.3.5 Allele specific expression analysis

The relocation of a chromosomal portion into a new position of the genome, due to chromosomal rearrangements such as translocations, can have effects on the expression of the relocated genes. In Rkatsiteli, haplotype phase for SNPs (described in section: *High-density haplotype phasing: SNPs, deletions, insertions and INDELs* of par. 3.2.5) was used to perform the ASE analysis. Since

Rkatsiteli is heterozygous for the balanced translocation, the analysis was performed by comparing ASE level in the "translocated" haplotype and in the "normal" haplotype on chromosome 1 and on chromosome 11 relative to the ASE in the rest of the genome.

In order to perform ASE on Rkatsiteli, Allim software package was used (Pandey, Franssen, Futschik, & Schlötterer, 2013). This software required the two FASTA format sequences of Rkatsiteli haplotypes, the gene annotation of the reference genome, and the transcriptome data in each tissue. A typical bias arising in ASE analysis derives from the mapping of both alleles against a common reference. Indeed, success rate of read mapping is biased if one allele is more similar to the reference than to the alternative allele. In order to reduce mapping bias, Allim used a polymorphism-aware reference genome which accounted for sequence variation between the alleles and estimated the residual mapping bias using a sequence-specific simulation tool. Polymorphism-aware reference was obtained using GATK *alternaterferencemaker* (Mirko Celii, PhD thesis, 2016). Lastly, Allim quantified ASE in simulated as well as in experimental data by determining the number of reads unambiguously assigned to one of the haplotypes and provided the *fdr*-corrected p values for statistical significance of allelic imbalance. The number of reads aligning to each gene sequence in each haplotype was considered as unit of gene expression. The analysis was performed for each tissue dataset separately and considering, for each haplotype, the sum of the number of reads in the two replicates. For each tissue dataset, ASE in translocated regions was measured as the $\log_2$ ratio of reads belonging to the normal haplotype (hapN) to reads belonging to the haplotype with translocation (hapT). Analogously, ASE level in the genome (excluded regions belonging to the translocated portions) was measured through the $\log_2$ ratio of reads belonging to each haplotype. The measure of ASE in the genome was used to calculate the expected ASE values in the translocated regions. Chi square test was applied to assess for significant deviation of ASE in the translocated regions relative to the expected, given the ASE level of the genome (α level of significance of 0.05).

Chapter 4 Results and Discussion

4.1 An *in silico* analysis of deleterious mutations in grapevine

Previous work on both nucleotide diversity and structural variation has shown the high genetic variability of the *Vitis vinifera* species. Nucleotide diversity, in terms of SNPs and INDELs, was detected by analysing a set of 128 *Vitis vinifera* varieties and 9 introgression lines selected for downy mildew resistance (Appendix 1). Furthermore, two outgroup species, *Vitis rupestris Du Lot* and *Vitis armata*, were added to the analysis. Structural variants (SVs) were detected in a subset of 50 varieties selected from the population of 128 grapevine varieties (varieties are indicated by an asterisk in Appendix 1).

A total of 9,476,335 SNPs and 860,191 INDELs were found in the population. Regarding SVs, a total of 18,090 deletions and of 45,273 insertions was detected in the subset of 50 varieties (Gabriele Magris, PhD thesis, 2016).

4.1.1 Analysis of mutations affecting gene function

Missense deleterious mutations

Prediction of the effect of nonsynonymous mutations in the grapevine population was carried out to detect SNPs with potential deleterious effect.

Among a total of 439,632 nonsynonymous SNPs found in the population of 137 varieties, 419,210 were scored by PROVEAN. Of this fraction, 112,492 (26.83 %) were found to have deleterious effects (i.e. the amino-acid substitution is not conserved in the population and the effect is predicted to alter protein function; PROVEAN score < -2.5). The remaining 306,718 (73.17 %) nonsynonymous SNPs were scored as neutral (i.e. the amino acid substitution is conserved, and the predicted effect is tolerated).

Figure 8 shows the ratio of mutations with deleterious effect to mutations with tolerated effect, for each variety of the population. Variation in the proportion of deleterious mutations over tolerated mutations is not marked among individuals in the whole population, with values ranging from 0.20 to 0.25. Variation is even less marked if considering a subset of the original dataset. If we exclude the 9 introgression lines carrying non-*vinifera* chromosome segments; the reference genotype, (PN40024); Ribolla Gialla (Slovenia) (clone of Ribolla Gialla); V400 (clone of Rkatsiteli); Charistvala Kolchuri (inter-specific hybrid); and Thompson Seedless (genetically very close to Sultanina), the values shift to a range of 0.23 - 0.25.

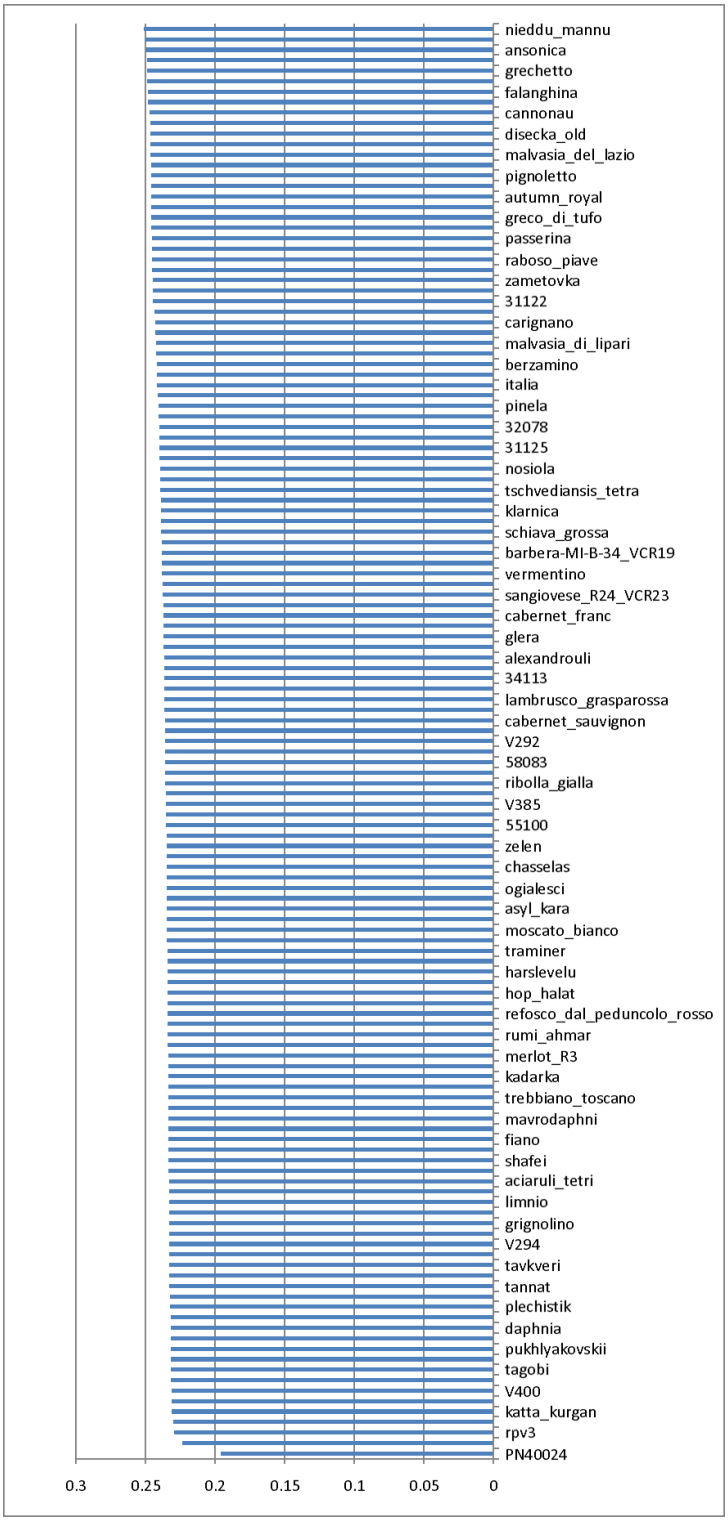

nieddu_mannu
ansonica
grechetto
falanghina
cannonau
disecka_old
malvasia_del_lazio
pignoletto
autumn_royal
greco_di_tufo
passerina
raboso_piave
zametovka
31122
carignano
malvasia_di_lipari
berzamino
italia
pinela
32078
31125
nosiola
tschvediansis_tetra
klarnica
schiava_grossa
barbera-MI-B-34_VCR19
vermentino
sangiovese_R24_VCR23
cabernet_franc
glera
alexandrouli
34113
lambrusco_grasparossa
cabernet_sauvignon
V292
58083
ribolla_gialla
V385
55100
zelen
chasselas
ogialesci
asyl_kara
moscato_bianco
traminer
harslevelu
hop_halat
refosco_dal_peduncolo_rosso
rumi_ahmar
merlot_R3
kadarka
trebbiano_toscano
mavrodaphni
fiano
shafei
aciaruli_tetri
limnio
grignolino
V294
tavkveri
tannat
plechistik
daphnia
pukhlyakovskii
tagobi
V400
katta_kurgan
rpv3
PN40024
0.3
0.25
0.2
0.15
0.1
0.05
0

Figure 8. **Analysis of nonsynonymous mutation**. Each bar represents the ratio of the number of nonsynonymous deleterious SNPs to the number of nonsynonymous tolerated SNPs per individual in the population of 137 varieties.

Structural variants affecting genes

Among the total set of structural variants detected in the subset of 50 grapevine varieties, 7,978 deletions and 13,913 insertions were found to affect 5,673 and 7,132 genes, respectively (Table 3). 22% of deletions and 16% of insertions were found in exonic regions; whereas 35% of deletions and 81% of insertions located in introns. The remaining 43% of deletions and 3% of insertions were placed in the mixed regions, since these structural variants encompassed more than one gene fraction (Figure 9).

Structural variants	Total	N. genes	Exonic	N. genes	Intronic	N. genes	Mixed	N. genes
N. deletions	7,978	5,673	1,759	1,510	2,777	1,912	3,442	2,621
N. insertions	13,913	7,132	2,213	2,007	11,200	5,184	500	487

Table 3. SV-disrupted genes: deletions and insertions are subdivided in exon-only, intron-only, and "mixed" (i.e. spanning both exons and introns).

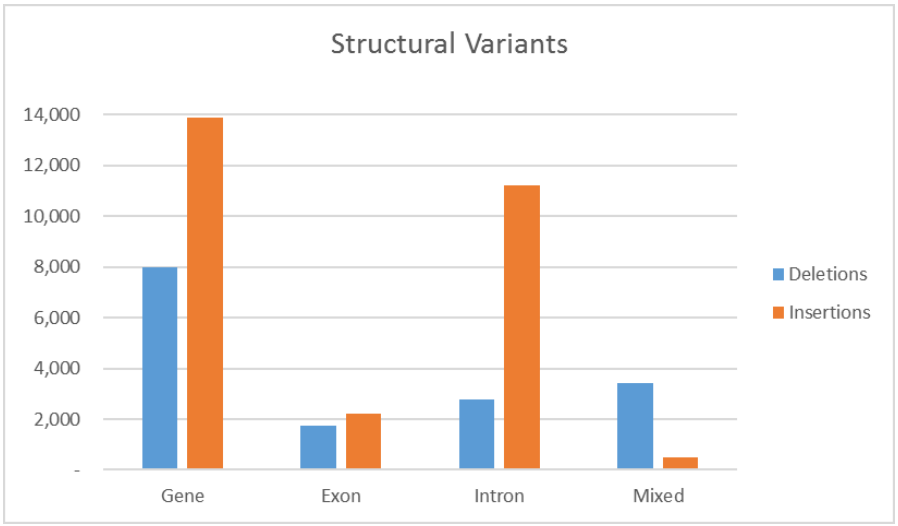


Figure 9. **Distribution structural variants**: deletions and insertions in genes, exonic regions, intronic regions and "mixed" regions (i.e. spanning both exonic and intronic fractions of genes).

We observed a disproportion between the fraction of TEs inserted in intronic regions and the fraction located in exons. This finding may be explained by the preferential insertion of LINE (RIL) and Copia (RLC) retrotransposons in introns and by the fact that insertions in exons are more likely to affect gene function and less likely to be conserved.

4.1.2 Identification of putative loss of function mutations

We investigated how many genes of the *Vitis vinifera* genome were affected by the different categories of neutral and deleterious variation. In the population under study, we found that 484,558 mutations affected 18,274 out of 31,845 genes (i.e. the 57.4% of genes) composing the grapevine genome (Vitulo et al., 2014). Among the entire variation complement found in genes, we detected: a) mutations with non functional effect, namely nonsense SNPs (9,833 stopgain and 1,251 stoploss SNPs), frameshift INDELs (7,304 frameshift deletions and 4,647 frameshift insertions) and structural variation in exonic regions (2,213 insertions and 1,759 deletions); b) mutations with altering effect, that is nonsynonymous SNPs with deleterious effect (112,492 SNPs); and c) mutations with neutral effect on function (306,718 nonsynonymous SNPs with tolerated effect). The great majority of polymorphic mutation with functional effect is attributable to nonsynonymous deleterious SNPs (83%), which corresponded to the 25.6% of the total nonsynonymous SNPs; non-sense polymorphisms creating a stop codon represented the 7.3%, while stop loss mutations reached the 0.92%; mutations shifting the reading frame corresponded to the 8.8%. Regarding the effect of mutations on gene function, and consequently on fitness, we focused particular attention to the fraction of variants resulting in non functional or truncated protein products (i.e. non-sense mutations, frameshift INDELs, and SVs in genes). Regarding this type of mutations, the majority of genes (6,136) were affected by stopgain SNPs, followed by genes affected by frameshift deletions (5,134). 2,007 genes were interrupted by insertions in exonic regions, while 1,510 genes were interrupted by deletions in exons. Deletions spanning both exonic and intronic regions involved 2,621 genes. Table 4 shows the number of genes affected by the various classes of nucleotide variation (SNPs and INDELs; one gene can be affected by more than one mutation). Number of genes affected by structural variants is reported in Table 3.

	N. hit	N. genes
Stop gain	9,833	6,163
Stop loss	1,251	1,145
Frameshift deletions	7,304	5,134
Frameshift insertions	4,647	3,610
Nonsynonymous deleterious SNPs	112,492	21,521
Nonsynonymous tolerated SNPs	306,718	25,247

Table 4. Distribution of nucleotide variation in the grapevine population. Among the mutations significantly affecting the gene function there are non-sense SNPs (stopgain and stoploss), frameshift INDELs and nonsynonymous SNPs with deleterious effect. Nonsynonymous SNPs with tolerated effect represent a class of mutations that alters amino acid sequence, but which effect is tolerated in the gene product.

4.1.3 Analysis of the frequency spectrum and the Tajima's D test

Under the mutation-selection balance, highly deleterious mutations are not fixed in the population. They are rather kept at very low frequency by purifying selection. On the contrary, weakly deleterious mutations may be approximately neutral and subject to genetic drift, rather than to selection (Mezmouk & Ross-Ibarra,2014). In a population at equilibrium, it is expected that mutations negatively affecting the function of crucial genes are maintained at very low frequencies; while silent mutations can occur in a range of frequencies (from low to high).

The allele frequency spectrum of the 137 varieties under study illustrated in Figure 10 shows the distribution of allelic frequencies for the classes of nucleotide mutation. Among low frequency derived-alleles (0.05 or lower), the fraction of highly deleterious variants (62.3% stopgain codons, 59.1% deleterious SNPs and 54.7% mutations causing splice sites) is significantly higher than the fraction of mutations with weak or neutral effect (48.5% tolerated SNPs and 47.1% synonymous SNPs) and the fraction of silent mutations (51.3% intergenic mutations and 49.4% intronic mutations) (p value < 0.001 based on a Wilcoxon Mann-Whitney test for each comparison). A higher proportion of low-frequency segregating SNPs at deleterious and nonsense sites relative

to synonymous sites (p value < 0.001, Wilcoxon Mann-Whitney test) suggests the action of purifying selection acting at nonsynonymous sites (Branca et al., 2011).

Thus, the frequency spectrum suggests that the grapevine population is not at the equilibrium (i.e. it is not in a standard neutral model), but rather, that selection has occurred on the fraction of highly deleterious mutations lowering their frequency of segregation in the population.

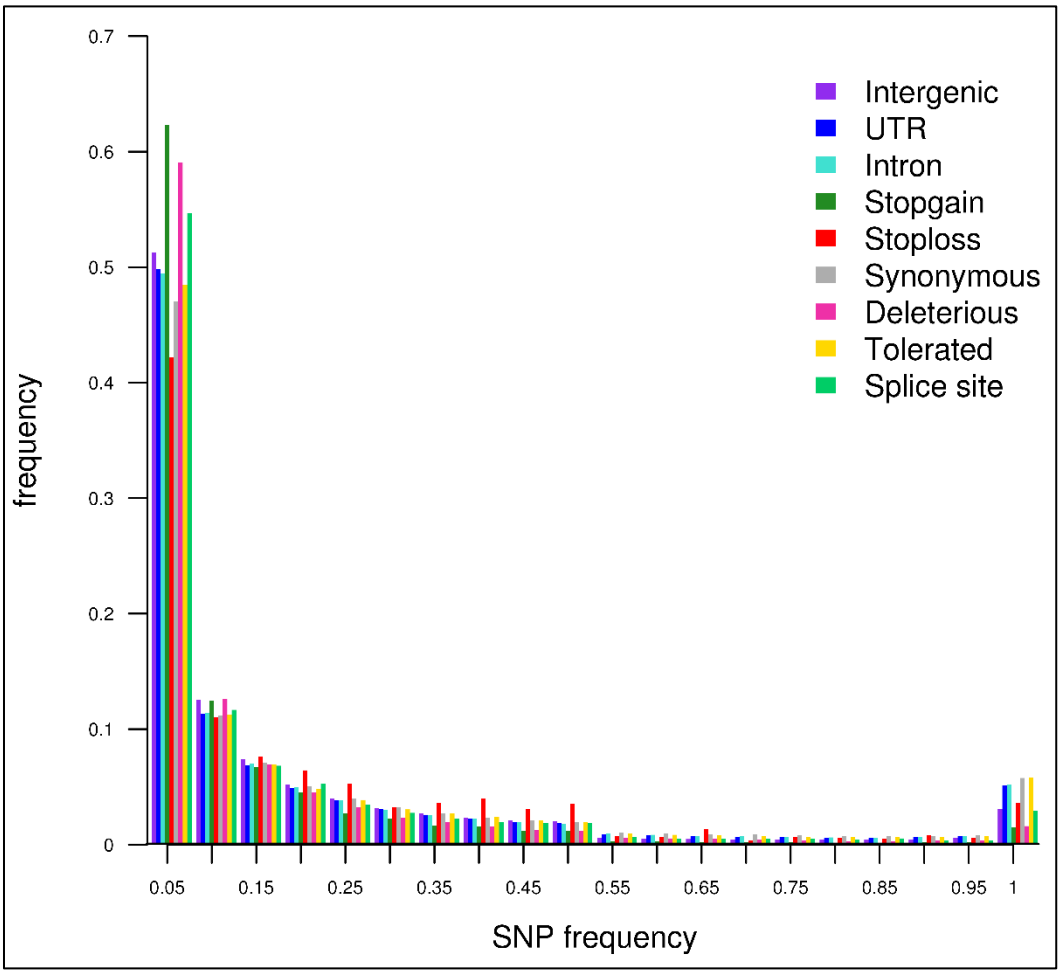


Figure 10. **Spectrum of allele frequency in the grapevine population of 137 varieties.** Derived SNP frequency is reported on the x-axis, proportion of each class of SNPs is reported on the y-axis.

According to the neutral theory, most of genetic variability at the molecular level is selectively neutral, i.e. it does not affect the fitness of the organism. Changes in allele frequencies are caused

by the balance between mutation rate and genetic drift (Kimura, 1991). Thus, frequency fluctuation for neutral mutations is random, rather than directional (as for the case of selection), and it depends on population size and on the number of segregating sites. A statistic used to quantitatively describe the frequency spectrum is Tajima's D. The purpose of Tajima's test is to identify sequences that do not fit the neutral theory model between mutation and genetic drift at equilibrium (Tajima, 1989).

The grapevine population under study it is not at the equilibrium, and *D* values of Tajima's test are shifted towards positive values. To assess their significance, *D* values were not compared to zero; instead, *D* values for each class of mutation was pairwise tested using the Wilcoxon Mann Whitney test (all pairwise comparisons showed a p value < 0.001, with the exception of the comparison between intronic SNPs and UTR SNPs, that was not significant).

Values of the Tajima's D for the different classes of mutation are shown in Figure 11.

Mutations with deleterious effect, such as nonsense and deleterious SNPs, and deletions and insertions in exons showed negative values of Tajima's D. On the contrary, mutations with neutral or nearly-neutral effect, such as intergenic and intronic mutations, either SNPs or SVs, and synonymous and nonsynonymous tolerated SNPs showed positive values.

Positive values of Tajima's D measured for neutral variants in grapevine is coherent with the scenario of a bottleneck generated during the domestication process that caused the loss of rare alleles and the increase of common alleles. Evidences suggest that *Vitis vinifera* is still recovering from the effect of the bottleneck and the expected *D* values for the evolving population is thus positive, as grapevine has not yet reached the equilibrium. Recovery from bottleneck may have occurred slower in grapevine compared to other cultivated species (see Meyer, DuVal, & Jensen, 2012 for a review on the comparison between the domestication process in perennial and in annual food crops). This lagging could be due to a longer juvenile phase of grapevine, typical of perennial species as compared to annual species, and to the reduced number of sexual generations past domestication, in favour of clonal propagation (Gaut, Díez, & Morrell, 2015; Miller & Gross, 2011; Myles et al., 2011).

Significant difference in the distribution of *D* values for mutations with deleterious effect (e.g. negative *D* for deleterious SNPs) relative to the distribution for mutations with neutral effect (e.g.

positive *D* for synonymous SNPs; p value of the comparison < 2e-16) indicate that there is an excess of low-frequency polymorphisms and suggests that purifying selection has occurred. These evidences may indicate that grapevine population is undergoing size expansion, following the bottleneck, determining an excess of rare alleles as a substrate for selection. In particular, negative values were measured for stopgain and deleterious SNPs and SV variants in exonic regions, thus suggesting that selection is acting on rare alleles with deleterious effect on gene function. Since the population is not at the equilibrium, non-negative *D* values of weakly-deleterious mutations (e.g. splice site SNPs), but significantly lower than *D* values of neutral mutations (e.g. synonymous SNPs; p value of the comparison < 2e-16), could be yet due to negative selection acting on them.

Stop loss mutations were removed from the analysis. Indeed, stop loss mutations detected in the varieties may represent stopgain mutations in the reference that we were not able to distinguish from true stop loss events. Deletions in the present analysis represent only a fraction of the total amount of deletion events. In fact, we were able to detect deleted regions with respect to the reference, while we could not detect deletions in the reference with respect to the other varieties. Deletions occurring at high frequency in the population may be present in the reference as well, affecting the detection of these events. As a consequence, detected deletions would represent only the fraction of events with lower frequency. This may explain the great shift of Tajima's D towards negative values observed for the deletions. Tajima's D measurement observed for insertions did not suffer from this bias and suggests the effect of selection on heavy deleterious mutation (negative value for insertions in exons) and on milder deleterious mutation (non-negative value for insertions in introns), with respect to the positive value for insertions in intergenic regions (non-deleterious mutation).

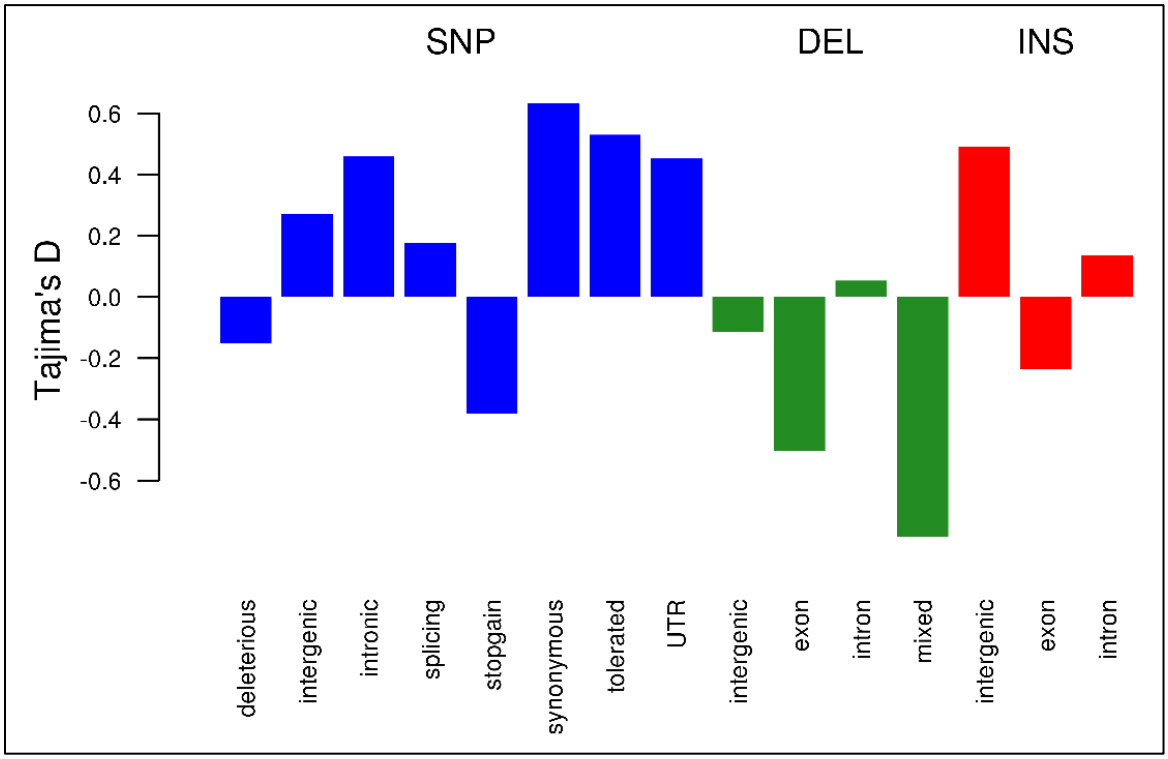


Figure 11. **Tajima's D values for SNPs, deletions and insertions in the subset of 50 varieties of the grapevine population.** SNPs are divided according to their functional annotation and SVs are divided based on their location with respect to intergenic or genic (exonic, intronic and mixed) regions.

4.2 An *in vivo* analysis of deleterious mutations in grapevine

4.2.1 Phenotype scoring in the progenies of selfing

Phenotyping was performed for the six progenies of selfing and we tested for seedling survival at different timings: within two month past germination (T_0) and at the beginning of the second vegetative year (T_2). Progenies which second vegetative year was set on 2016 were also scored for vigour before (T_1) and after (T_2) the first winter season. Phenotyping information was used alongside genotyping information (recorded at T_0 for all progenies, except those indicated in Table 7 that were genotyped at T_2) to analyse the progression of survival in time and to establish the onset of segregation distortion.

We observed asynchrony in germination and variation in vigour within each population. Table 5, Table 6, and Table 7 report viability at different timings and phenotypic scores of progenies of self-crosses for each seeding year.

We analysed two self-cross progenies in 2014, derived from selfing of Cabernet Franc and of Sangiovese, respectively (Table 5). Genotypic frequencies at T_0 represent the fraction of germinated seeds and the seedlings that survived until two months after germination (i.e. no seedling was lost between germination and sampling for genotyping) in Cabernet Franc self-cross progenies. For Sangiovese self-cross progenies, instead, a fraction of seedlings died prematurely before the time of sampling. There was a critical phase of growth early after seed germination which determined very premature lethality of Sangiovese self-cross progenies, which was not observed in the progenies of Cabernet Franc self-cross. Thus, genotypic frequencies in the progenies of Sangiovese self-cross at the time of sampling (T_0) represent the fraction of seedlings that survived until two months after germination. Survival of seedlings was then scored at T_2 after overwintering.

		T_0 (05/2014)	T_2 (05/2015)		
Cross	**Germinated**	**Genotyped**	**Dead**	**Alive**	**No phenotype***
Cabernet franc self	68	68	46	22	0
Sangiovese self	100	87	18	69	0

* Plantlets that were sacrificed

Table 5. Phenotype scoring at T_2 of progenies of Cabernet Franc and Sangiovese self-crosses. In progenies of Sangiovese self-cross, 87 individuals were genotyped out of a total of 100 germinated seedlings. 13 seedlings died in the time window between germination and genotyping at T_0.

We analysed three self-cross progenies in 2015, derived from the selfing of Pinot Noir, Rkatsiteli, and Schiava Grossa, respectively (Table 6). Experiments in 2014 suggested there are two major critical developmental stages which determine seedling survival, the first one early after germination and the second one during the first winter season. For the three self-cross progenies assessed in 2015, we paid particular attention to collect plant material for genotyping as early as possible – sampling all germinated seeds and, if necessary, using the entire epigeal part of extremely weak seedlings for DNA extraction. For this reason, viability at later stages of development could not be assessed for this fraction of the progenies. We also scored survival before and after winter dormancy, comparing survival at the end of the vegetative growth of the first year (T_1) and survival after overwintering (T_2). Furthermore, at T_1 and at T_2, plantlets were scored not only for survival, but also for classes of vigour. Phenotype of progenies described in Table 6 were thus scored at T_0, T_1 and T_2. Before and after winter dormancy, seedlings were classified based on the stem diameter, used as indicator of vigour, and on the length of the shoot. Stem diameter at T_1 was a predictor for winter frost survival. Indeed, we observed a negative relationship between stem thickness/lignification at T_1 and resumption of vegetative growth at T_2. Bud sprouting was used as an indicator for resumption of vegetative growth after wintering. According to these parameters, seedlings were categorized into three classes (described in par 3.2.1): normal phenotype, weak phenotype and very weak phenotype (dead at T_2).

		T_0 (05/2015)	T_1 (10/2015)			T_2 (05/2016)			
Cross	**Germinated**	**Genotyped**	**Dead**	**Weak**	**Normal**	**Dead**	**Weak**	**Normal**	**No phenotype***
Pinot nero self	85	85	9	21	52	22	16	44	3
Rkatsiteli self progeny 2	152	152	24	27	76	77	8	42	25
Schiava grossa self	91	91	0	20	61	57	12	12	10

* Plantlets that were sacrificed

Table 6. Phenotype scoring at T_1 and T_2 of progenies of Rkatsiteli, Pinot Noir and Schiava Grossa self-crosses. In these progenies, phenotype of progenies was classified based on stem diameter, length of the shoot, and number of buds giving rise to shoots. Based on these criteria, seedlings were classified in the categories "normal", "weak" or "dead".

The self-cross progenies of Rkatsiteli showed a different progression of mortality after germination compared to the other progenies. Germination rate in the progenies of Rkatsiteli self-cross was comparable to self-crosses of other varieties. At the end of the first vegetative season, before winter (T_1), mortality was approximately 16%. A fraction of seedlings that survived until T_1 (approx. 18%) showed a weak phenotype. At the beginning of the second vegetative season (T_2) approximately 75 % of the progeny was die. A small fraction of the individuals that were classified as weak at T_1 (7.7%) survived until T_2. At T_2, less than 40% of the initial progeny survived and was scored as normal phenotype. In Figure 12A, survived individuals are shown at spring time at T_2. Among them, it was possible to identify a gradient of weakness (Figure 12B). Proceeding from left to right, the first individuals showed a normal phenotype, with lignified stem with diameter greater than 3 mm, many buds already sprouted and a plant height of 20 cm. The second and third individuals showed intermediate-to-weak phenotypes, with lignified stem with diameter within the lower bound of normal range, but few or no buds, whose sprouting was delayed. Plant height was reduced comparing to the first phenotype, ranging from 5 to 10 cm. Finally, the individual on the right has died, and it was characterized by non-lignified stem with a diameter lower than 1 mm and no shoot emerging from buds. The mortality in progenies of Rkatsiteli self-cross observed at T_2 is associated with low stem lignification and diameter, which appear to be critical factors for surviving winter frost and/or for storing adequate reserves and supporting resumption of vegetative growth in the next season.

Figure 12. **Progeny of Rkatsiteli self-cross**. Rate of survival of the progenies after wintering, at the beginning of the second vegetative season (A). Phenotypes of the progenies (B) going from the normal one (on the left) to the weakest one (on the right): stem growth and bud sprouting were considered when differentiating phenotypes at T_1 and T_2.

The progenies of Schiava Grossa self-cross also showed atypical progression of mortality. At T_1, all seedlings survived and the majority (75%) of progenies showed a normal phenotype. At the beginning of the second vegetative season, 70% of seedlings showed no bud sprouting and were scored as very weak phenotype, while approx. 15% of seedlings showed few buds late in sprouting. This evidence suggests that the critical phase in this progeny is not linked to proper development during the first vegetative season, since seedlings showed normal stem growth at T_1, but rather, the critical factor for survival may occur during winter or at the resumption of vegetative growth.

In addition to progenies that were germinated within the timeframe of this PhD, we also analysed progenies of selfing of Primitivo and Rkatsiteli that were germinated in 2013, and for which we had information on the number of survivors at T_1. We genotyped these progenies at T_2. Genotypic frequencies at T_2 represent the fraction of seedlings that resumed vegetative growth after the first winter (Table 7). We were able to obtain genotypic frequencies not earlier than the stage of emergence of epicotyl and to follow seedling mortality only in subsequent developmental stages.

Sporophytic selection may also have occurred at any earlier stage, from embryo development to emergence of the radicle.

	T_1 (06/2013)	T_2 (05/2014)		
Cross	**Available**	**Dead**	**Genotyped**	**No phenotype***
Primitivo self	80	18	62	0
Rkatsiteli self progeny 1	232	146	86	0

* Plantlets that were sacrificed

Table 7. Phenotype scoring at T_2 of progenies of Primitivo and Rkatsiteli self-crosses. Samples were harvested for genotyping on the 2nd vegetative season.

We observed that germination rate (scored as emergence of epicotyl) in progenies deriving from self-cross was lower than germination rate of progenies deriving from outcrossing of the parental varieties. In Figure 13, progenies deriving from selfing of Sangiovese (A) and of Cabernet Franc (C) showed a germination rate around 65% compared to 89% of germinated progenies deriving from the outcrossing of the same parental variety (B, out-cross of Sangiovese and Traminer; C, out-cross of Cabernet Franc and Rkatsiteli). Sangiovese progenies showed an even weaker phenotype than Cabernet Franc progenies: at this stage, 67% of germinated seeds in the former progeny had not yet set cotyledons versus 12% of the latter one.

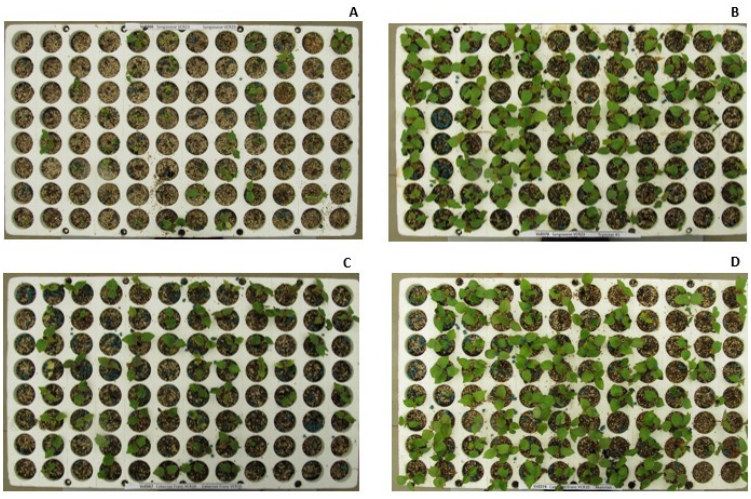


Figure 13. **Germination in progenies of selfing and of outcrossing.** Germination rate comparison between progenies of Sangiovese self-cross (A) and progenies from the out-cross between Sangiovese and Traminer (B). Germination rate comparison between progenies of Cabernet Franc self-cross (C) and progenies from the out-cross between Cabernet Franc and Rkatsiteli (D).

4.2.2 Markers

ddRAD sequencing results

The seven segregating populations analysed in the present work were genotyped by means of a reduced-representation sequencing technique (the ddRAD-seq technique). Sequencing metrics for the six progenies of selfing and the progenies of the outcross are reported in Table 8.

Cross	N. samples sequenced	Average raw reads per sample (M)	Average % of retained reads[$]	Average retained reads per sample[$] (M)	Average median coverage per sample	Average median insert size per sample (bp)*
Cabernet Franc self	68	3.71	98.72	3.68 ± 1.29	10.38 ± 2.47	272.97 ± 9.47
Pinot Nero self	85	4.62	99.81	4.61 ± 1.67	13.41 ± 2.55	275.84 ± 6.26
Primitivo self	62	4.23	98.40	4.13 ± 1.25	10.97 ± 1.60	284.47 ± 2.44
Rkatsiteli self	238	3.17	99.55	3.16 ± 1.36	11.68 ± 3.59	259.78 ± 6.71
Sangiovese self	87	5.01	99.82	5.00 ± 2.63	15.41 ± 6.50	251.53 ± 3.15
Schiava Grossa self	91	4.52	99.63	4.51 ± 1.20	13.45 ± 2.39	268.47 ± 4.68
Schiava Grossa x Rkatsiteli	192	2.86	99.55	2.86 ± 1.08	11.40 ± 3.22	262.16 ± 7.05

* fragment insert size (inline barcode length excluded)

[$] unambiguously demultiplexed reads with expected RE sites

Table 8. ddRAD-seq sequencing metrics. Average number of raw reads per sample and average number of retained reads per sample is given. Retained reads correspond to reads containing the expected restriction sites and the expected inline barcode sequence. Standard deviation values are provided for average number of retained reads per sample, average median coverage per sample, and average median insert size per sample.

We aimed to obtain 3.5 – 4 M reads per sample in self-cross progenies, while a lower data amount was required for the out-cross progenies (≈ 3 M reads per sample). We required at least 20 M reads for each parental variety.

Rate of raw reads containing the expected restriction site overhang on both pairs (SphI on forward read and MboI on reverse read) and univocal inline barcodes was high, ranging from 98.4% to 99.8%. Average number of successfully demultiplexed reads per sample was 2.86 ± 1.08 M in the out-cross progenies. Average number of successfully demultiplexed reads per sample in self-crosses varied from 3.17 M reads for Rkatsiteli progenies (with a standard deviation of ± 1.36 M) to 5 M for Sangiovese progenies (which showed also the highest standard deviation, ± 2.63 M).

Reads from parental varieties that aligned to the reference genome ranged from 90 to 94%. Primitivo self-cross progenies showed the highest homogeneity in terms of retained reads per sample, with an average of 4.13 M reads and a standard deviation of ± 1.67 M reads. According to retained read number, average median coverage per sample varied from a minimum value of 11X in Primitivo progenies and in the out-cross progenies to a maximum value of 15X in Sangiovese progenies. Average median insert size per sample was homogeneous among different progenies, with values ranging from 251 bp to 284 bp.

Table 9 shows the total number of samples sequenced and the total number of loci obtained in each cross progenies. Furthermore, it describes the number of loci excluded from the analysis at each data cleaning step and the number of good quality loci retained for the subsequent analysis of segregation distortion. They represent the 50-60% compared to the raw data set ("Good quality retained loci").

Segregating loci falling in regions characterized by repeats or microsatellites were discarded from the analysis. Through this step, approximately 30% of loci were removed ("N. loci in repetitive regions"). Subsequently, automatic custom data cleaning and manual curation were applied to the dataset of genotyped RAD tag loci ("Manual + automatic curation (discarded loci)"). Phasing of RAD tags was performed using information of the homozygous tags, and assuming that the probability of recombination between two adjacent tags was low (i.e. tag genotypes observed in sequence on a large proportion of subjects are likely part of the same haplotype). After phasing haplotypes, manual data curation was carried out on datasets and evident mis-called loci were removed from analysis (see section: *Low-density haplotype phasing* of par. 3.2.2).

Cross	N. samples sequenced	N. samples with > 20% missing loci	N. samples analysed	Total loci	N. loci in repetitive regions	N. loci missing > 20% samples	Manual + automatic curation (discarded loci)	Good quality retained loci
Cabernet Franc self	68	4	64	16,303	5,345	1,570	760	8,628
Pinot Nero self	85	3	82	21,257	7,032	669	638	12,918
Primitivo self	62	5	57	15,648	5,106	1,131	935	8,476
Rkatsiteli self	238	21	217	9,264	2,648	1,041	387	5,188
Sangiovese self	87	15	72	17,690	5,888	2,328	2,181	7,293
Schiava Grossa self	91	3	88	12,447	3,955	580	279	7,633
Schiava Grossa x Rkatsiteli	192	21	171	11,978	3,370	1,115	-	7,493

Table 9. Total number of samples sequenced and total number of called loci is described for each progeny. In addition, number of samples and number of loci excluded from analysis at each data cleaning step is provided. Good quality retained loci column reports the final number of loci after the data cleaning process, used for downstream analysis.

Whole genome low-coverage DNA sequencing results

Highly homozygous individuals in the progenies of Cabernet Franc, Primitivo, Rkatsiteli, and Sangiovese self-crosses were sequenced at low coverage in order to infer haplotypes, by including all SNP/INDEL and SV heterozygous loci present in each parent. Sequencing metrics are reported in Table 10. We set out to obtain 7-8 X coverage per sample, corresponding to about 35-40 M reads per sample. We obtained a minimum of 6 X coverage in Sangiovese self-cross samples (with about 31.6 M reads per sample) and a maximum of 10 X coverage in Rkatsiteli self-cross samples (with about 48 M reads per sample). The lower coverage obtained in Sangiovese samples compared to the others did not negatively affect downstream analysis. Samples were chosen in order to maximize homozygous genotype information along the genome. In Table 11, percentage of genome covered by at least one or two homozygote individuals is indicated (percentage is calculated on the base of RAD genotyping data). We called from 9.7 to 9.9 M SNPs and from 0.85 to 0.9 M INDELs in each batch of samples ("Total SNPs" and "Total INDELs" in Table 11). After removing source of errors as SNPs/INDELs falling in repetitive regions, number of SNPs retained for downstream analysis was reduced to 4.1-4.2 M and number of INDELs was reduced to 0.53-0.55 M, in each batch ("SNPs retained" and "INDELs retained"). Since our samples were sequenced at low-coverage, we did not require a minimum coverage for SNP/INDEL calling, to avoid excessive data drop. High quality SNP/INDEL datasets for the population of grapevine varieties were available from previous work (Gabriele Magris, PhD thesis, 2016). To reduce false positive calls in the progenies as a result of sequencing errors, only those polymorphic positions corresponding to heterozygous SNPs/INDELs in the parental variety were retained in the low-coverage datasets of the progenies. In addition to haplotype phase inference for SNP/INDEL heterozygous loci in the varieties, we inferred phase for deletions and insertions. Previously identified SVs in the parent varieties were quantified in the progenies. Genotype information of RAD loci in the progenies was used to better define thresholds for SV heterozygous call in the

parent varieties. On the base of the genotype information, we assured that SV quantification in the progenies was coherent with the SV call in the varieties and we did not require a minimum number of reads for quantification. Table 11 provides the number of heterozygous SV loci in parent varieties that was detected in the progenies ("Deletions" and "Insertions").

Cross	Samples sequenced	Average raw reads per sample (M)	Average % aligned reads per sample	Average reads aligned per sample (M)	Average coverage per sample	Median insert size per sample (bp)
Cabernet franc self	5	36.65	96.83	35.49 ± 3.19	6.97 ± 0.60	499.40 ± 31.64
Primitivo self	5	39.85	97.13	38.70 ± 7.35	8.01 ± 1.32	491.40 ± 37.24
Rkatsiteli self	4	48.01	97.77	46.94 ± 3.61	9.89 ± 0.70	481.25 ± 16.79
Sangiovese self	5	31.55	96.87	30.57 ± 0.74	6.28 ± 0.17	478.80 ± 20.12

Table 10. Whole genome low-coverage DNA-seq sequencing metrics. Average number of raw reads per sample and average percent and number of reads after alignment is provided. Standard deviation values are provided for average aligned reads, average coverage, and median insert size for sample.

Cross	Homozygosity[1] (%)	Homozygosity[2] (%)	Total SNPs	SNPs retained	Total INDELs	INDELs retained	Deletions	Insertions
Cabernet franc self	99.64	93.67	9,967,945	4,247,221	872,101	539,697	4,914	6,687
Primitivo self	99.16	87.71	9,795,769	4,226,793	874,769	528,459	6,648	9,293
Rkatsiteli self	99.53	86.18	9,715,466	4,156,005	898,479	549,822	5,356	6,251
Sangiovese self	98.83	86.70	9,746,629	4,207,163	852,120	528,459	5,780	6,965

[1] percentage of genome covered by at least one homozygote individual.
[2] percentage of genome covered by at least two homozygote individuals.

Table 11. Table shows percentage of homozygosity across genome for each cross progeny, considering at least one or two homozygous progenies (based on RAD tags genotyping data). Number of SNPs and INDELs before and after cleaning in repetitive regions and number of SVs is indicated.

4.2.3 Genetic linkage maps

For the construction of the genetic maps, we used the populations with the largest number of individuals that were genotyped by means of ddRAD-seq: the progenies of Rkatsiteli self-cross (238 individuals) and the progenies of Schiava Grossa x Rkatsiteli out-cross (192 individuals). Individuals having more than 20% missing markers were excluded from the analysis ("N.

individuals with >20% missing markers" in Table 12). During the step of generation of linkage groups, markers that did not group in any of the 19 expected linkage groups were removed from the dataset ("Excluded RAD markers"). The number of markers considered in the final analysis is reported in Table 12 ("Mapped RAD markers").

The progenies of Rkatsiteli selfing was considered as an F2-like segregating population. From this population, 5,188 RAD markers were used for the construction of the Rkatsiteli genetic map. Segregating loci in the progenies of the out-cross Schiava Grossa x Rkatsiteli were split in two datasets according to the two-ways pseudo test-cross design (Grattapaglia & Sederoff, 1994). For each parent, we used markers that were heterozygous in Schiava Grossa (parental configuration *ab* x *aa*) or in Rkatsiteli (parental configuration *aa* x *ab*), respectively. More than 3,400 RAD markers were analysed in each parent map. The map of Schiava Grossa reports the recombination rates in the seed parent, the map of Rkatsiteli reports the recombination rates in the pollen parent.

	self-cross	**out-cross**	
	Rkatsiteli	**Rkatsiteli**	**Schiava grossa**
N. of individuals	238	192	192
N. individuals with >20% missing markers	22	21	21
N. individuals considered	216	171	171
Total n. of RAD markers	5,188	3,689	3,639
Excluded RAD markers	190	213	209
Mapped RAD markers	4,998	3,476	3,430

Table 12. Features of genetic linkage maps. Number of individuals and of markers before and after data cleaning is indicated.

For each segregating population, RAD markers were grouped into 19 linkage groups (LGs). Based on the physical locations of markers, we were able to unambiguously assign the 19 LGs to the 19 chromosomes of the grapevine genome. Furthermore, markers belonging to previously unassigned scaffolds on chromosome Unknown could be assigned to the respective chromosomes, summing up more than 13 Mbp to the grapevine assembled genome (Table 13).

	self-cross	out-cross	
	Rkatsiteli	Rkatsiteli	Schiava grossa
N. of assigned scaffolds	15	81	85
bp assigned	2,389,995	11,433,399	13,124,505

Table 13. Scaffolds from chromosome Unknown and number of bp that were assigned to chromosomes through genetic map construction.

Table 14 describes number of loci and number of genetic bins per LG, length of LGs in centiMorgan (cM) and LG resolution as cM per bin for the genetic map of Rkatsiteli – deriving from the progenies of the self- and the out- cross, respectively – and of Schiava Grossa. All the mapped RAD markers were grouped onto the same chromosomes to which they had been physically positioned previously, providing evidence that chromosome assignment was correct.

Figure 14, Figure 15 and Figure 16 show genetic bins location along each chromosome in each of the three maps described in the same order used for Table 14.

The genetic map of Rkatsiteli deriving from the F2-like mapping population spanned a total of 965 cM, with a total of 1,392 genetic bins and a mean resolution of 0.72 cM per bin. A total of 789 genetic bins were mapped on the genetic map of Rkatsiteli deriving from the population of the out-cross. The total length of this map was 1,290 cM and mean resolution was 1.71 cM per bin. The genetic map of Schiava Grossa spanned a total of 1,013 cM, with a total of 644 bins and a mean resolution of 1.66 cM per bin.

LGs	Rkatsiteli from self-cross progeny				Ratsiteli from out-cross progeny				Schiava Grossa from out-cross progeny			
	n. loci	n. bins	LG length (cM)	resolution (cM/n.bins)	n. loci	n. bins	LG length (cM)	resolution (cM/n.bins)	n. loci	n. bins	LG length (cM)	resolution (cM/n.bins)
1	360	101	50.00	0.50	260	60	65.69	1.09	240	51	55.27	1.08
2	167	49	46.63	0.95	115	30	64.81	2.16	164	38	50.19	1.32
3	187	55	36.58	0.67	120	29	48.71	1.68	118	21	45.95	2.19
4	294	79	54.45	0.69	183	41	77.11	1.88	180	29	51.08	1.76
5	353	82	49.23	0.60	216	49	82.12	1.68	148	34	40.98	1.21
6	251	72	58.04	0.81	202	42	79.50	1.89	194	36	64.54	1.79
7	385	97	78.26	0.81	314	64	92.07	1.44	189	37	60.40	1.63
8	348	92	53.66	0.58	249	60	75.70	1.26	164	25	56.33	2.25
9	185	56	39.95	0.71	105	25	53.65	2.15	149	38	46.93	1.23
10	174	38	38.99	1.03	147	40	71.11	1.78	168	34	52.86	1.55
11	270	75	49.11	0.65	175	31	50.04	1.61	187	29	56.79	1.96
12	259	82	47.18	0.58	164	41	54.62	1.33	192	35	50.99	1.46
13	293	87	53.30	0.61	235	53	85.13	1.61	186	39	63.78	1.64
14	247	74	73.04	0.99	157	42	91.17	2.17	297	50	63.86	1.28
15	115	35	31.02	0.89	75	15	35.24	2.35	163	30	53.02	1.77
16	214	57	44.50	0.78	155	37	49.18	1.33	162	36	50.79	1.41
17	198	61	44.40	0.73	123	26	52.95	2.04	156	26	52.24	2.01
18	433	120	73.60	0.61	275	70	107.77	1.54	265	45	70.44	1.57
19	265	80	42.78	0.53	206	34	54.20	1.59	108	11	27.49	2.50
	4998	**1392**	**964.72**	**0.72**	**3476**	**789**	**1290.76**	**1.71**	**3430**	**644**	**1013.91**	**1.66**
	tot. loci	tot. bins	tot. length	mean resolution	tot. loci	tot. bins	tot. length	mean resolution	tot. loci	tot. bins	tot. length	mean resolution

Table 14. Description of the genetic linkage maps of Rkatsiteli (derived from the progenies of selfing and of outcrossing, respectively) and of Schiava Grossa (derived from the progenies of outcrossing).

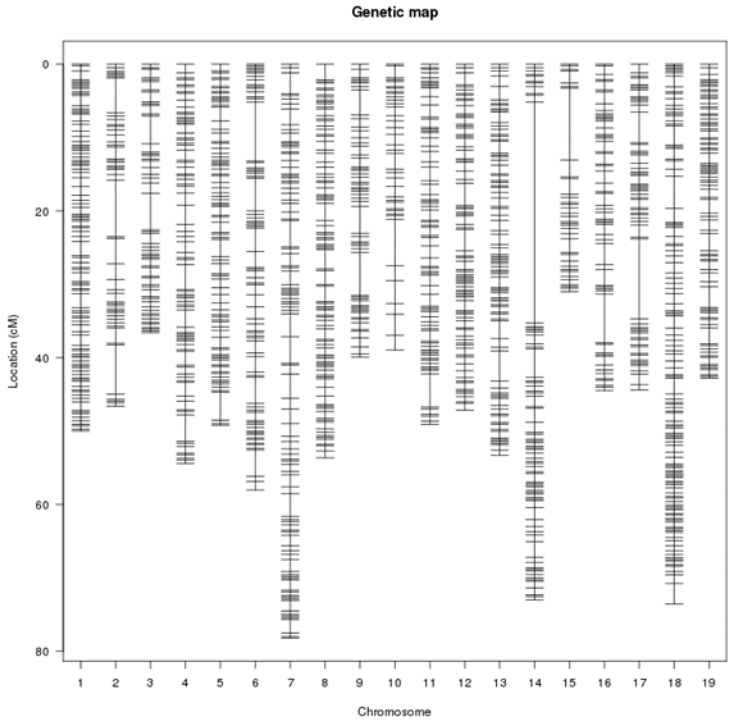


Figure 14. **Rkatsiteli genetic linkage map derived from the F2-like mapping population**. The position of bins is shown in Kosambi cM along the 19 chromosomes. Regions in chromosomes 2, 14, 15, 16 and 17 lacking markers represent homozygous (or hemizygous) regions in Rkatsiteli. A homozygous region of 5 Mbp at the telomeric end of chromosome 15 accounts for the shortened length of this LG.

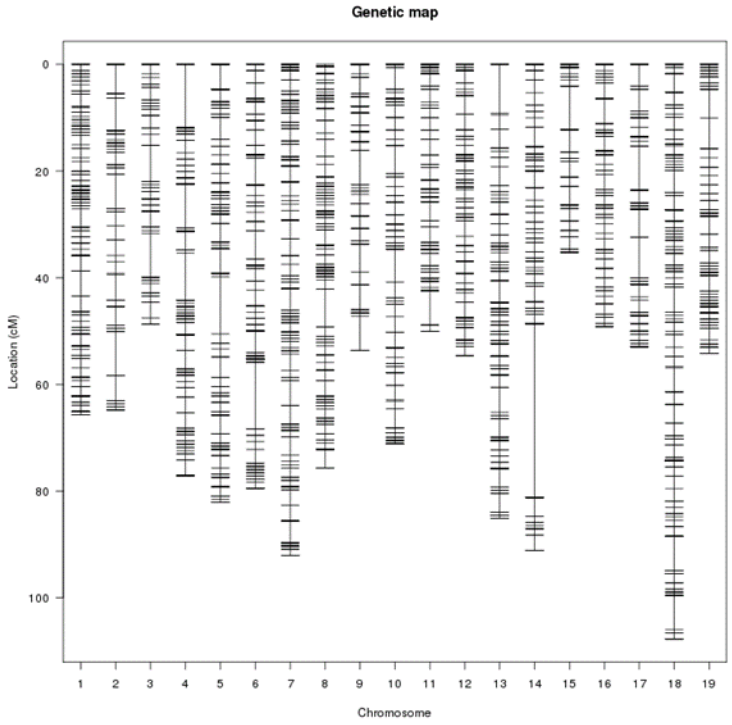


Figure 15. **Rkatsiteli genetic linkage map derived from the progenies of out-cross**. The position of bins is shown in Kosambi cM along the 19 chromosomes.

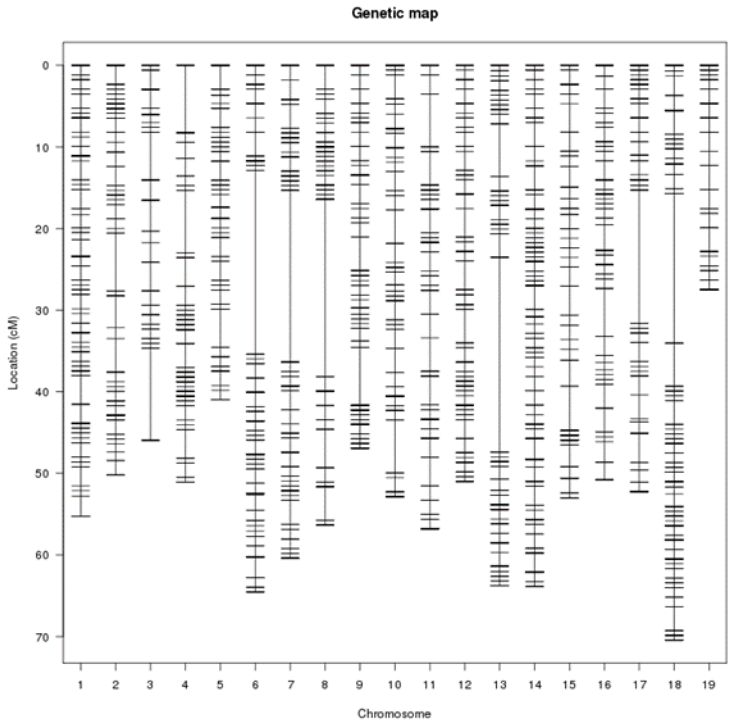


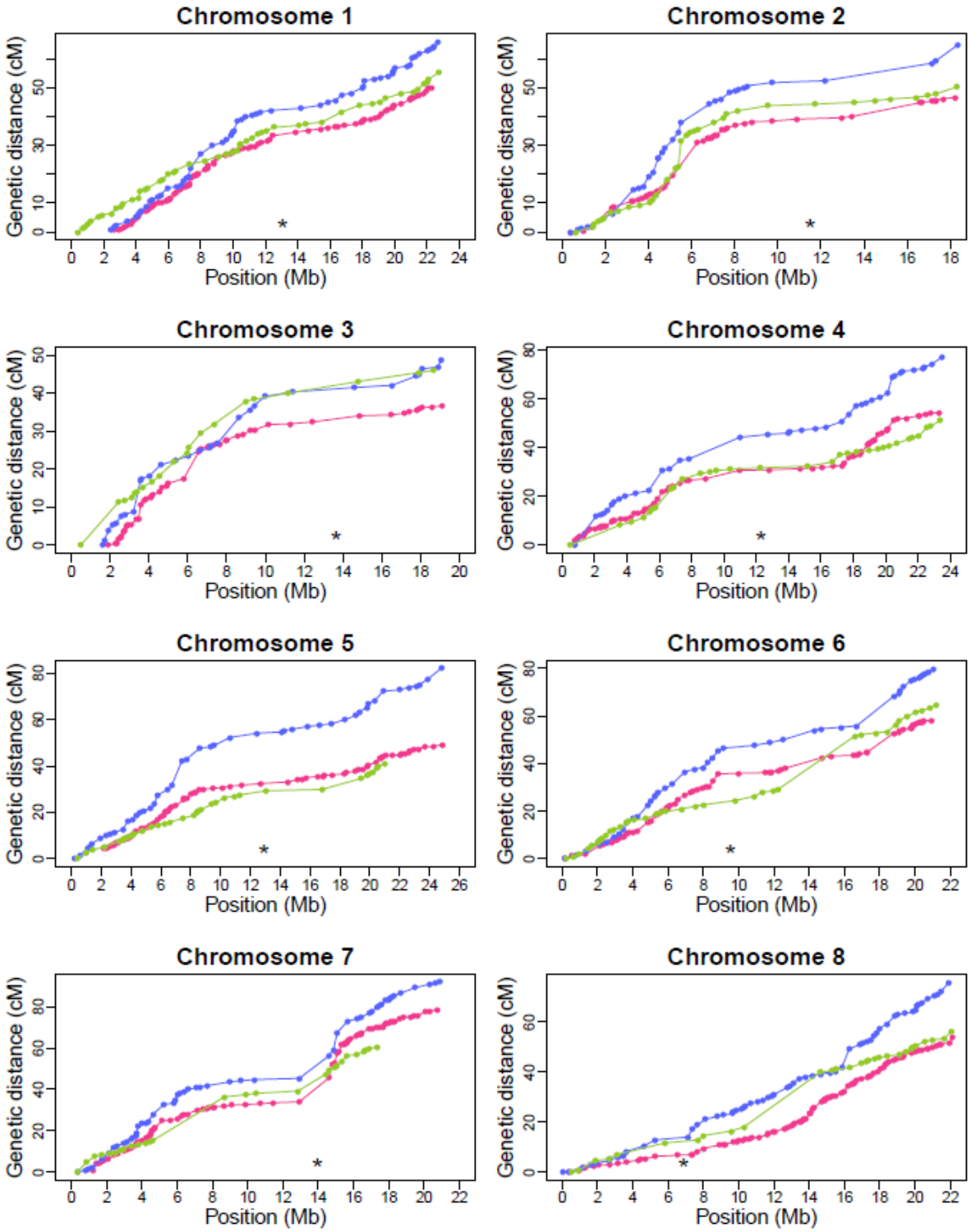
Figure 16. **Schiava Grossa genetic linkage map derived from the progenies of out-cross.** The position of bins is shown in Kosambi cM along the 19 chromosomes. Regions in chromosomes 3, 4, 5, 6, 7, 8, 13, 17, and 18 lacking markers represent homozygous (or hemizygous) regions in the variety. Homozygous regions at the telomeric end of chromosomes 5, 7, and 19 account for the shortened length of the corresponding LGs.

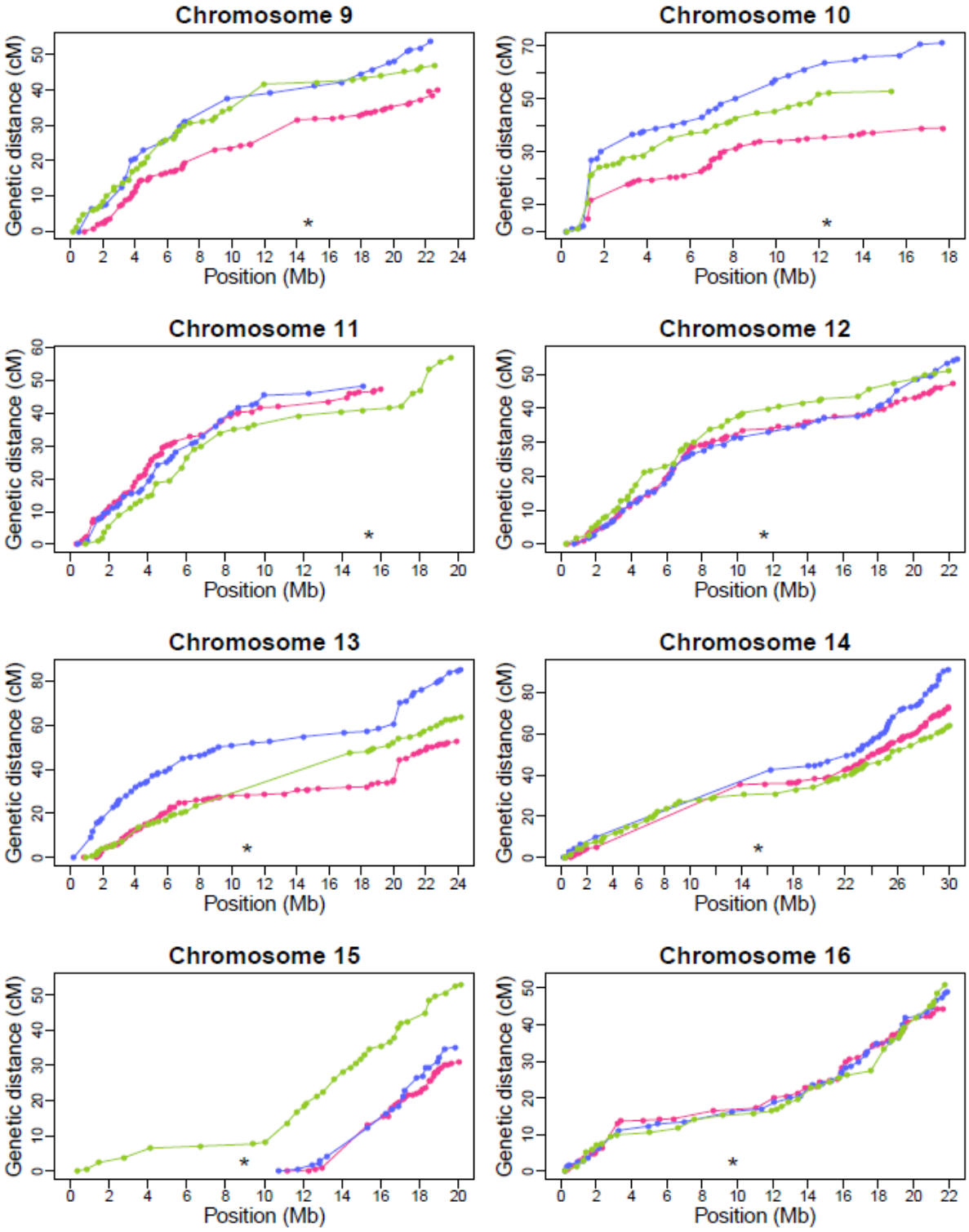
The genetic map of Rkatsiteli based on the F2-like progenies has, on average, shorter LGs compared to the corresponding LGs generated from the progenies of outcross (see also Figure 17). This effect may be due to different recombination rates in the same variety when it is used as female or as male parent. Heterochiasmy, that is the difference in crossing-over (CO) number and position between female and male meiosis, is a widespread phenomenon in animal and plant species (Lenormand & Dutheil, 2005). In hermaphroditic plants lacking sex chromosomes, regions close to the telomeres have a lower recombination in female than in male meiosis (Ne Giraut et al., 2011), and this difference may have reflected in the construction of genetic maps. Differences in genetic maps of reciprocal backcross populations were observed in barley by Phillips et colleagues (2015). By comparing the recombination landscape in male and female meiosis under different temperature regimes, they observed an increase in crossing-over frequency in male meiosis, but not in female meiosis (Phillips et al., 2015). By analysing back-cross populations of *Arabidopsis*, Giraut and co-workers (2011) observed a dramatic difference in the number of crossing-over between male and female meiosis, estimating an average of 11.15 and of 6.6 COs per male and female meiocyte, respectively. This difference affected the length of the genetic maps, with a male to female CO ratio of 1.73 (Ne Giraut et al., 2011).

These evidences may explain the increased size of Rkatsiteli map derived from the progenies of outcross, as this variety was used as the male parent of the cross. In the F2-like progenies of Rkatsiteli, half of the Rkatsiteli gametes donated to the self-cross derived from male gametogenesis. The same observation may explain some differences between Rkatsiteli and Schiava Grossa maps, where the varieties were used as male and female parent, respectively. Some of the difference may also be due to suppression of recombination in regions enriched in structural variants. SVs indeed can generate regions of hemizygosity, interfering with normal homologous pairing. Local variation in the haplotype structure will have a strong influence on estimates of genetic distance (Dooner & He, 2008; Eckardt NA, 2008). Lastly, differences in length for some of the LGs may be due to large regions of homozygosity extending to the telomeres of the chromosomes. For example in Schiava Grossa, chromosome 5, chromosome 7, and chromosome 19 show extended regions of homozygosis towards the telomeric end (spanning 3.5 Mbp, 3 Mbp, and 5 Mbp, respectively). A homozygous region of 5 Mbp is present at the telomeric end of chromosome 15 in Rkatsiteli.

In Figure 17, the order and the distance of markers on the genetic map are shown relative to their placement on the reference sequence. Markers segregating in the Rkatsiteli F2-like mapping population are shown in purple, markers segregating from Rkatsiteli parent in the mapping population of the out-cross are shown in blue, and markers segregating from Schiava Grossa parent in the mapping population of the out-cross are shown in light green. In each chromosome, asterisk indicates the location of the centromere. As expected, recombination rate is not constant along chromosomes and tends to be repressed in centromeric and pericentromeric regions.

The notion that the centromere exerts a negative effect on meiotic recombination both within itself and on proximal chromosomal regions was recognized in the thirties (Beadle & Ker, 1931) and was termed the centromere effect (Choo, 1998). This effect has been documented in many plant species - including *Arabidopsis* (Round, Flowers, & Richards, 1997), *Solanum/Lycopersicon* (Sherman & Stack, 1995), *Zea mays* (Anderson et al., 2003), *Oryza sativa* (Limborg, McKinney, Seeb, & Seeb, 2016) – and in animal species such as *Drosophila* (Mather, 1939), mouse (Billings et al., 2010), and humans (Mahtani & Willard, 1998). The condensed state of centromeric heterochromatin does not fully explain the effect of recombination suppression in the proximal regions. In *Drosophila*, deletions of centromeric heterochromatin resulted in lowered levels of meiotic exchange in centromere-adjacent euchromatin (Yamamoto & Miklos, 1978). It was proposed that crossing over events in regions close to the centromere can be a constraint for the assembly of the kinetochore (Ellermeier et al., 2010) with subsequent negative effects on chromosome segregation (Vincenten et al., 2015).


Chromosome 1
Genetic distance (cM)
Position (Mb)
Chromosome 2
Genetic distance (cM)
Position (Mb)
Chromosome 3
Genetic distance (cM)
Position (Mb)
Chromosome 4
Genetic distance (cM)
Position (Mb)
Chromosome 5
Genetic distance (cM)
Position (Mb)
Chromosome 6
Genetic distance (cM)
Position (Mb)
Chromosome 7
Genetic distance (cM)
Position (Mb)
Chromosome 8
Genetic distance (cM)
Position (Mb)


Chromosome 9
Genetic distance (cM)
Position (Mb)
Chromosome 10
Genetic distance (cM)
Position (Mb)
Chromosome 11
Genetic distance (cM)
Position (Mb)
Chromosome 12
Genetic distance (cM)
Position (Mb)
Chromosome 13
Genetic distance (cM)
Position (Mb)
Chromosome 14
Genetic distance (cM)
Position (Mb)
Chromosome 15
Genetic distance (cM)
Position (Mb)
Chromosome 16
Genetic distance (cM)
Position (Mb)

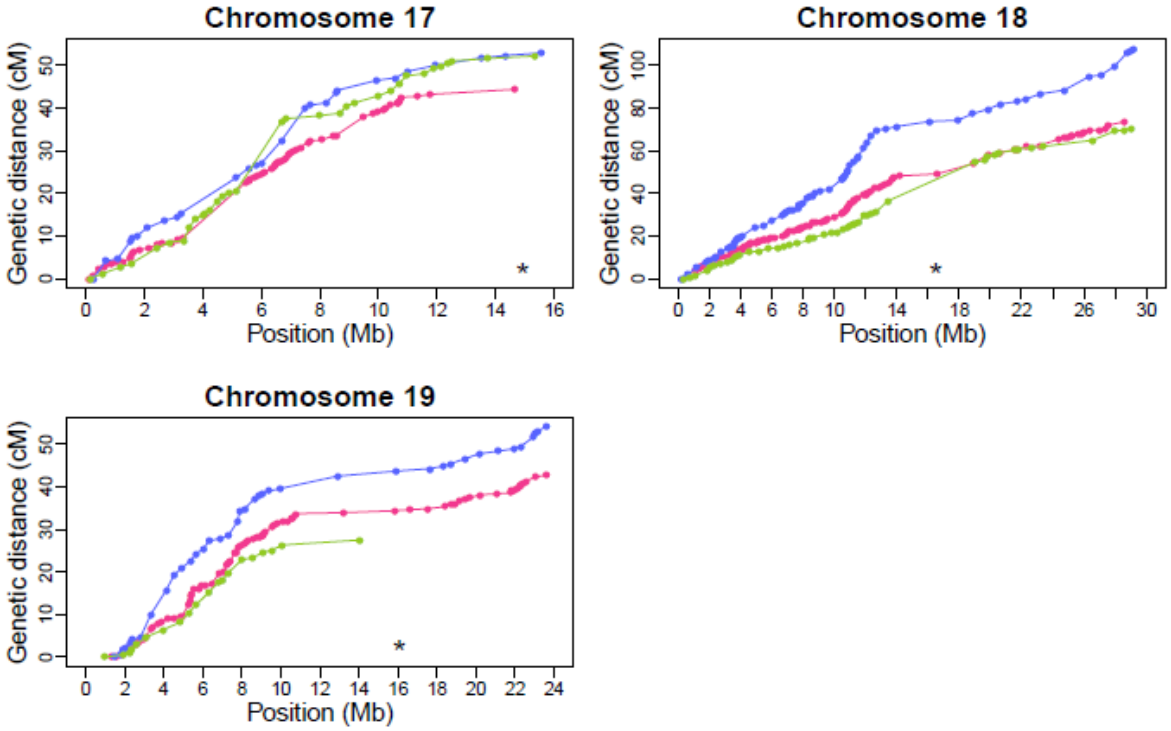


Figure 17. **Genetic position of RAD markers (in cM) relative to their physical position (in Mbp)**. Purple: segregating markers in the progenies of Rkatsiteli selfing; blue: segregating markers in Rkatsiteli from the progenies of out-cross; green: segregating markers in Schiava Grossa from the progenies of out-cross. For each chromosome, the asterisk is located in correspondence of the centromere position. Shortened length of chromosomes 5, 7, and 19 in Schiava Grossa and of chromosomes 15 in Rkatsiteli is due to large regions of homozygosis at the telomeric end of these chromosomes.

4.2.4 Analysis of single-locus segregation distortion (single-locus SD)

Identification of single-locus SD

We used Chi square test for assessing deviations of the observed genotypic frequencies from the expected ratio in the progenies. Twelve regions of single-locus segregation distortion (SD) were identified and all the progenies of selfing showed at least one locus of SD (the lowest p value at each locus of SD in indicated in Table 15). Sangiovese showed the highest number of loci with SD (four in total), followed by Primitivo with three loci, and Rkatsiteli with two loci. The progenies of the out-cross, instead, did not show any locus of SD.

Figure 18 illustrates the twelve single-locus SD regions for chromosome and for variety. The upper panels show the log-transformed p values of the Chi square test for each genotyped RAD-tag locus as a function of the physical position (red dots). Values above threshold of 3 (corresponding to the log-transformed, Bonferroni corrected α level of significance of 0.001) are statistically significant. The lower panels show the count of the genotypic classes for each RAD locus: recessive homozygotes (blue), heterozygous (green) and dominant homozygotes (purple). Recombination in progenies defined boundaries of loci of SD: borders were defined by the least represented genotypic class. Boundaries of recessive and partially dominant loci were defined by the lowest count of homozygous recessive individuals; in dominant loci, by the lowest count of heterozygotes; in the overdominant locus, by the lowest count of both classes of homozygotes. The initial and final chromosomal coordinates of single-locus SD regions are reported in Figure 15. Cabernet Franc, Rkatsiteli and Sangiovese all showed close, but non-overlapping loci of distortion on chromosome 8. Regarding the size of the identified loci, we observed a narrow region of distortion on chromosome 8 in Rkatsiteli and in Sangiovese progenies (330 Kb and 260 Kb, respectively). Locus of chromosome 5 in Sangiovese and of chromosome 15 in Pinot Noir showed instead the widest genetic interval, spanning 8.7 Mbp and 7.8 Mbp, respectively.

The developmental stage in which the genotypic distortion occurred differed among progenies (Table 15). In the progenies of Cabernet Franc, Pinot Noir, Sangiovese, and Schiava Grossa single-locus SD occurred at T_0. In the progenies of Rkatsiteli the onset of SD occurred later, at the T_2

stage. In the case of Primitivo, we have no genotypic information of the progenies before the T_2 stage.

Cross	Chromosome	Start coordinate (Mbp)	End coordinate (Mbp)	Effect at T_0	Effect at T_2	Locus segregation	P-values*
Cabernet franc self	8	19,451,711	22,189,347	L	L	partially dominant	< 0.0001
Pinot nero self	15	4,579,661	12,409,362	L	L	recessive	< 0.0001
Primitivo self	4	3,198,069	5,888,137	-	L	recessive	< 0.0001
Primitivo self	11	1	2,853,881	-	L	partially dominant	< 1.00e-07
Primitivo self	12	6,805,962	8,194,361	-	L	dominant	< 1.00e-15
Rkatsiteli self	8	16,528,331	16,856,387	ns	L	recessive	< 1.00e-07
Rkatsiteli self	18	20,288,310	21,565,864	ns	L	recessive	< 1.00e-08
Sangiovese self	5	8,164,204	16,866,264	L	L	recessive	< 1.00e-04
Sangiovese self	8	14,986,249	15,246,148	L	L	recessive	< 1.00e-05
Sangiovese self	9	8,924,697	14,320,556	L	L	recessive	< 1.00e-04
Sangiovese self	11	17,946,141	18,995,833	D	D	overdominant	< 1.00e-04
Schiava grossa self	4	10,896,476	13,648,030	D	D	recessive	< 0.001
Schiava grossa X Rkatsiteli	none	-	-	-	-	-	-

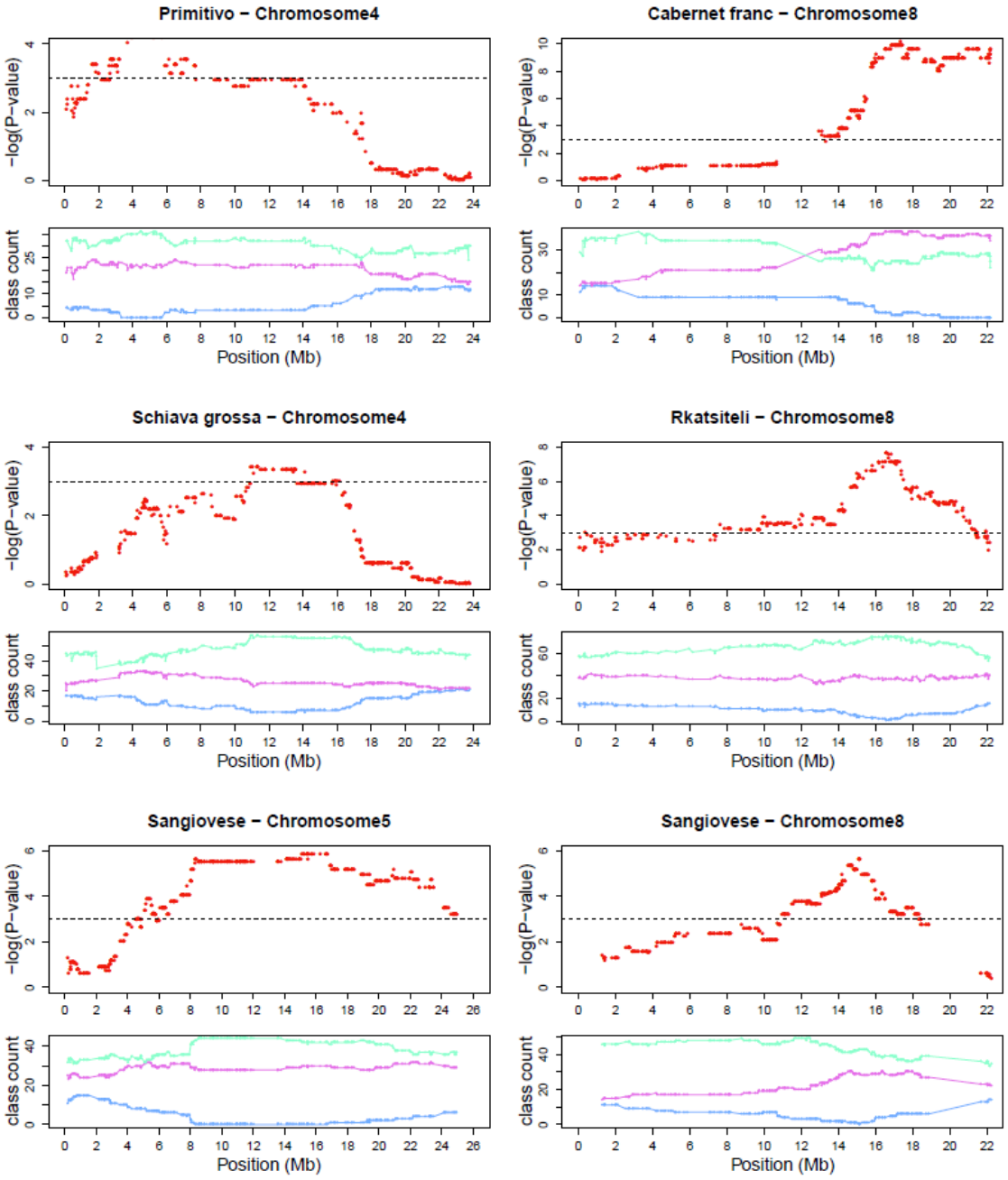
L: lethal effect
D: deleterious effect
ns: not significant
- : data not available
* : p values refer to T_2

Table 15. Chromosome, start and end coordinates of loci showing single-locus segregation distortion. Effect of distortion on locus segregation at T_0 (two months after germination) and at T_2 (at the beginning of the second vegetative season), and segregation mode.

The information on the effect of SD, whether lethal or deleterious, and on the locus segregation mode was obtained from the distribution of the genotypic classes. In Table 15, the effect of distortion at T_0 and at T_2, and locus segregation mode is described for the twelve regions. A lethal effect implies that a genotypic class is completely missing; a deleterious effect causes a significant reduction in the observed genotypic frequency compared to the expected genotypic frequency, but the genotypic class does not disappear. Distortion in the genotypic frequencies indicated that five loci had lethal effect at T_0, while two loci caused deleterious effect at T_0. At T_2, ten loci showed

lethal effects (including two loci detected in Rkatsiteli progeny that did not show significant distortion at T_0) and two loci caused deleterious effect.

The information about the model of inheritance of alleles involved in SD was inferred from the relative ratio among genotypic classes. Under a fully recessive lethal model, genotypic classes are expected with a ratio of $1/3$ homozygotes to $2/3$ heterozygotes. Eight loci segregated according to a completely recessive model. They were located on chromosome 15 in Pinot Noir progenies; on chromosome 4 in Primitivo progenies; on chromosomes 8 and 18 in Rkatsiteli progenies; on chromosomes 5, 8, and 9 in Sangiovese progenies; and on chromosome 4 in Schiava Grossa progenies. SD loci on chromosome 8 in Cabernet Franc progenies and on chromosome 11 of Primitivo progenies segregated according to a partially dominant model. Indeed, ratio of heterozygotes : homozygotes ranged in between the ratio expected for an additive locus ($1/2$ heterozygotes : $1/2$ non-affected homozygotes) and the ratio expected for a fully lethal dominant locus (0 heterozygotes : 1 non-affected homozygotes). The SD locus on chromosome 12 of Primitivo progenies segregated as a nearly completely dominant model. As many as 95% of individuals that survived until T_2 were homozygotes for the non-deleterious allele, while only 5% of survivors were heterozygous. Lastly, SD locus on chromosome 11 in Sangiovese progenies segregated according to a model of overdominance. Here, we observed a significant decrease in the frequency of both homozygous classes with respect to heterozygous genotypes, although the two homozygous classes did not completely disappear.


Primitivo – Chromosome4
Cabernet franc – Chromosome8
Schiava grossa – Chromosome4
Rkatsiteli – Chromosome8
Sangiovese – Chromosome5
Sangiovese – Chromosome8
−log(P−value)
class count
Position (Mb)

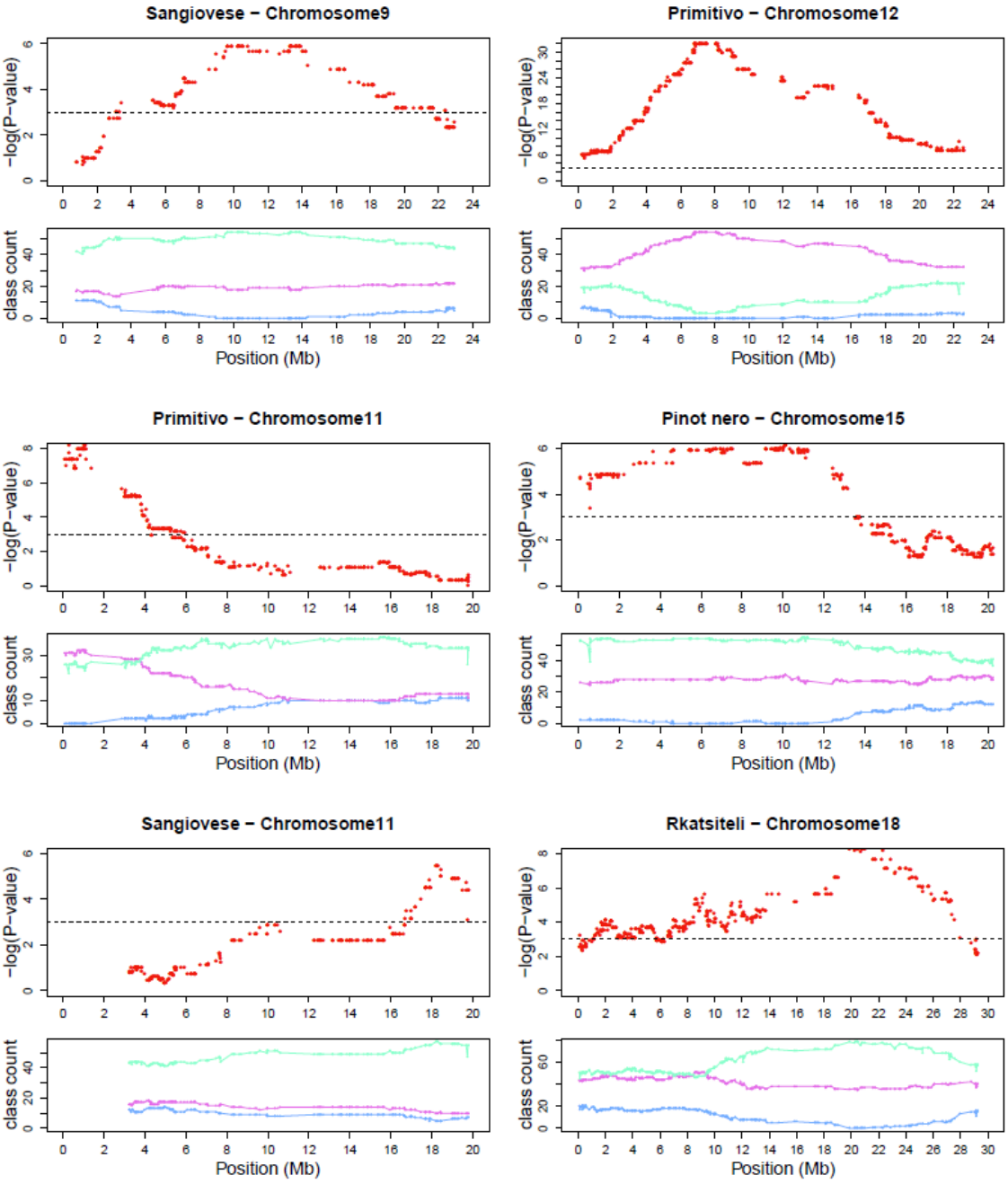


Figure 18. **Analysis of segregation distortion**. The log-transformed p values ($-\log_{10}$(p value)) from Chi square test are plotted along position on chromosomes (upper panel). Values above threshold of 3 (dashed horizontal line) are statistically significant ($\alpha = 0.001$). Genotypic classes count is shown in lower panel: recessive homozygotes (blue), heterozygote (green) and dominant homozygote (purple).

The effect of lethal and deleterious alleles on fitness

In order to quantify the effect of deleterious alleles on the fitness of progenies, we estimated coefficients of selection (*s*) and dominance (*h*) for the putative lethal/deleterious allele at each locus under study (Table 16).

According to the model proposed by Morton NE et al. on 1956, *h* value is 0 for a completely recessive allele and 1 for an allele causing the same probability of death in a heterozygote as in a homozygote. Thus, alleles showing some dominance have a more aggressive effect on fitness compared to completely recessive alleles. Selection against alleles showing dominance is more efficient than selection against completely recessive alleles.

Putative mutant allele on chromosome 15 in Pinot Noir progenies, on chromosome 4 in Primitivo progenies, on chromosomes 8 and 18 in Rkatsiteli progenies, and on chromosomes 5, 8, and 9 in Sangiovese progenies showed $s = 1$ and $h = 0$, meaning complete recessivity. Putative mutant allele on chromosome 4 in Schiava Grossa progenies showed $s < 1$ and $h = 0$. In this case, the effect of distortion is not lethal, but deleterious and few recessive homozygotes can survive. Alleles on chromosome 8 in Cabernet Franc progenies and on chromosome 11 in Primitivo progenies showed $s = 1$ and $h = 0.75$ and 0.61, respectively (estimation refers to T_2). Since the dominance of the mutant allele was higher than for an additive allele (where $h = 0.5$), we observe more than 50% of death among heterozygotes. Allele on chromosome 12 in Primitivo progenies showed nearly complete dominance ($h = 0.97$ and $s = 1$). Locus of SD on chromosome 11 in Sangiovese progenies segregated as overdominant: *s* and *t* estimates of selection against each of the two homozygotes at T_2 were 0.67 and 0.83, respectively. The values of selection against both alleles were lower than 1, meaning that the effect on fitness was deleterious.

cross	Locus on	T0				T2			
		effect	*h*	*s*	*t*	*effect*	*h*	*s*	*t*
Cabernet franc self	chromosome 8	L	0.61	1	-	L	0.75	1	-
Pinot noir self	chromosome 15	L	0	1	-	L	0	1	-
Primitivo self	chromosome 4	na	na	na	na	L	0	1	-
	chromosome 11	na	na	na	na	L	0.61	1	-
	chromosome 12	na	na	na	na	L	0.97	1	-
Rkatsiteli self	chromosome 8	ns	ns	ns	ns	L	0	1	-
	chromosome 18	ns	ns	ns	ns	L	0	1	-
Sangiovese self	chromosome 5	L	0	1	-	L	0	1	-
	chromosome 8	L	0	1	-	L	0	1	-
	chromosome 9	L	0	1	-	L	0	1	-
	chromosome 11	D	-	0.65	0.82	D	-	0.67	0.83
Schiava grossa self	chromosome 4	D	0	0.85	-	D	0	0.87	-

L: lethal effect
D: deleterious effect
ns: not significant
na: not available
h: dominance
s: selection against lethal/deleterious allele
t: selection against alternative allele (in overdominant loci)

Table 16. Estimates of dominance (*h*) and selection (*s* and *t*) of the putative lethal/deleterious allele in each locus of segregation distortion at T_0 and at T_2.

Fine-mapping of candidate loci of SD

Regions of single-locus distortion detected in the six progenies of selfing were fine-mapped through a population-level analysis in the population of 128 *Vitis vinifera* varieties. The analysis consisted in comparing short blocks haplotypes of the parent varieties to the haplotypes in the population. In the population, we calculated the frequency in homozygosis of each short-block haplotype defined in the parent variety. The reasoning we adopted for the fine-mapping was the following: if the parent haplotype contained a lethal allele, this should never be found in homozygosis in the population. Thus, all regions where one haplotype phase was never found in homozygosis in the 128 varieties were defined as putative candidate lethal loci. Through short-

blocks haplotype analysis in the germplasm, we could fine-map candidate regions of segregation distortion to 34 Kbp in the best case (locus on chromosome 8 in Sangiovese, Table 17).

Cross	Locus on	SD Candidate region length (Kbp)	
		Segregation defined	Population defined
Cabernet franc self	Chromosome 8	2,738	620
Pinot nero self	Chromosome 15	7,830	2,484
Primitivo self	Chromosome 4	2,690	1,130
	Chromosome 11	2,854	1,480
	Chromosome 12	1,388	760
Rkatsiteli self	Chromosome 8	328	150
	Chromosome 18	1,278	920
Sangiovese self	Chromosome 5	8,702	2,530
	Chromosome 8	260	34
	Chromosome 9	5,396	2,390
	Chromosome 11	1,050	399
Schiava grossa self	Chromosome 4	2,752	577

Table 17. Fine-mapping of single-locus SD regions. The length of the regions is defined by means of segregation and recombination in progenies deriving from selfing ("Segregation defined"). The fine mapping of the regions is achieved by means of haplotype comparison in the population of grapevines ("Population defined").

4.2.5 Characterization of loci of segregation distortion

Loci of SD in the genomic context: recombination frequency, deleterious SNPs, methylation profiles, gene density and repeat density

Meiotic recombination rate is not constant along chromosomes. Noticeably, it is lower in highly condensed and transcriptionally inert heterochromatin, mainly composed of repetitive sequences, and is higher in gene-rich, actively transcribed and structurally relaxed euchromatin (Eichten et al., 2011; Ellermeier et al., 2010).

Furthermore, low recombining regions of the genome are those expected to accumulate deleterious mutations, as a consequence of reduced efficacy of purifying selection (J. Lu et al., 2006; Rodgers-Melnick et al., 2012).

We investigated whether candidate deleterious loci were located in high- or low- recombining regions and we compared recombination frequency to other genomic features of the genome (Figure 19). The upmost panel of Figure 19 illustrates recombination frequency along the chromosomes carrying loci of SD (represented below the recombination distribution, as black boxes). Colours used range from green (low recombination) to red (high recombination, see also legend of Figure 19). The remaining panels show the remaining features included in the analysis (from top to bottom): a) the ratio of nonsynonymous deleterious SNPs over nonsynonymous tolerated SNPs in expressed genes; b) CG and CHG methylation profiles; c) the distribution of gene density (coding DNA sequence, CDS) and d) the distribution of repeat density.

According to the null distribution of recombination frequency in the genome, windows of 200 Kbp were divided in low (values lower than the 33rd percentile, 2.09 cM/Mbp), intermediate (values between the 33rd and 66th percentile, between 2.09 and 4.31 cM/Mbp) and high (values higher than the 66th percentile, 4.31 cM/Mbp) recombination frequency.
According to the three categories of recombination frequency, loci of SD were distributed as follows:

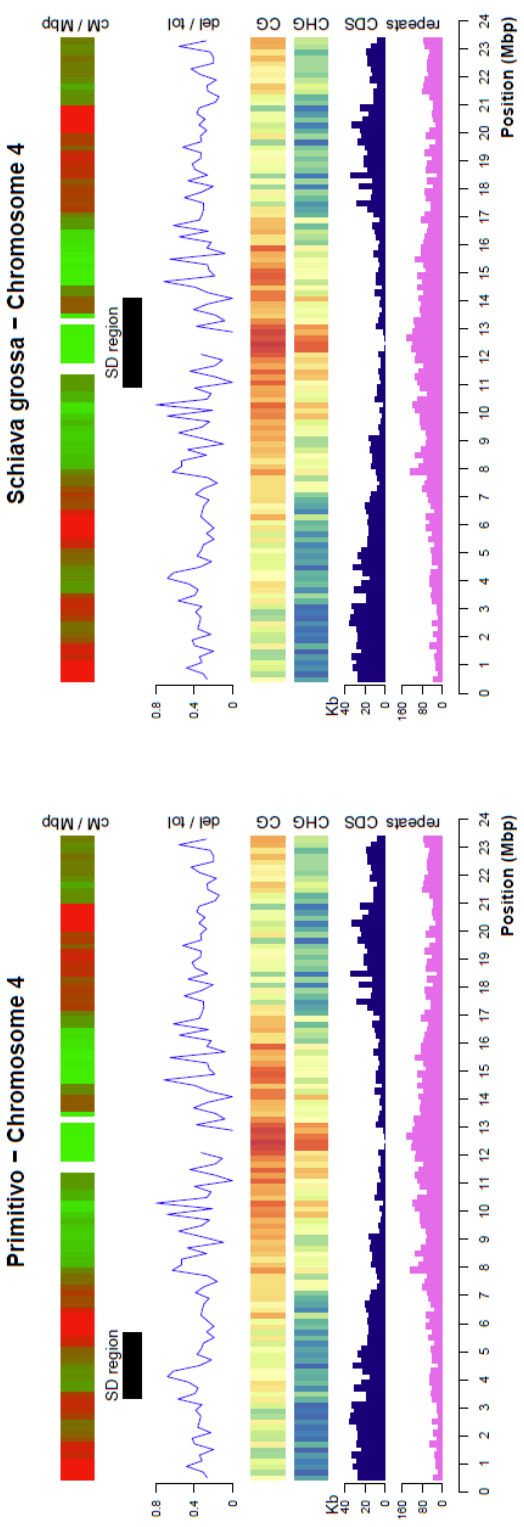
- Four loci were found in regions showing low recombination frequencies: locus on chromosome 4 in Schiava Grossa, on chromosome 5 in Sangiovese, on chromosome 15 in Pinot Noir and on chromosome 18 in Rkatsiteli.
- Loci on chromosomes 8 and 9 in Sangiovese were located on regions of intermediate recombination frequency.
- The remaining six loci were sited in regions showing high recombination frequencies: locus on chromosome 8 in Cabernet Franc, on chromosomes 4 and 12 in Primitivo, on chromosome 11 in both Primitivo and Sangiovese, and on chromosome 8 in Rkatsiteli.

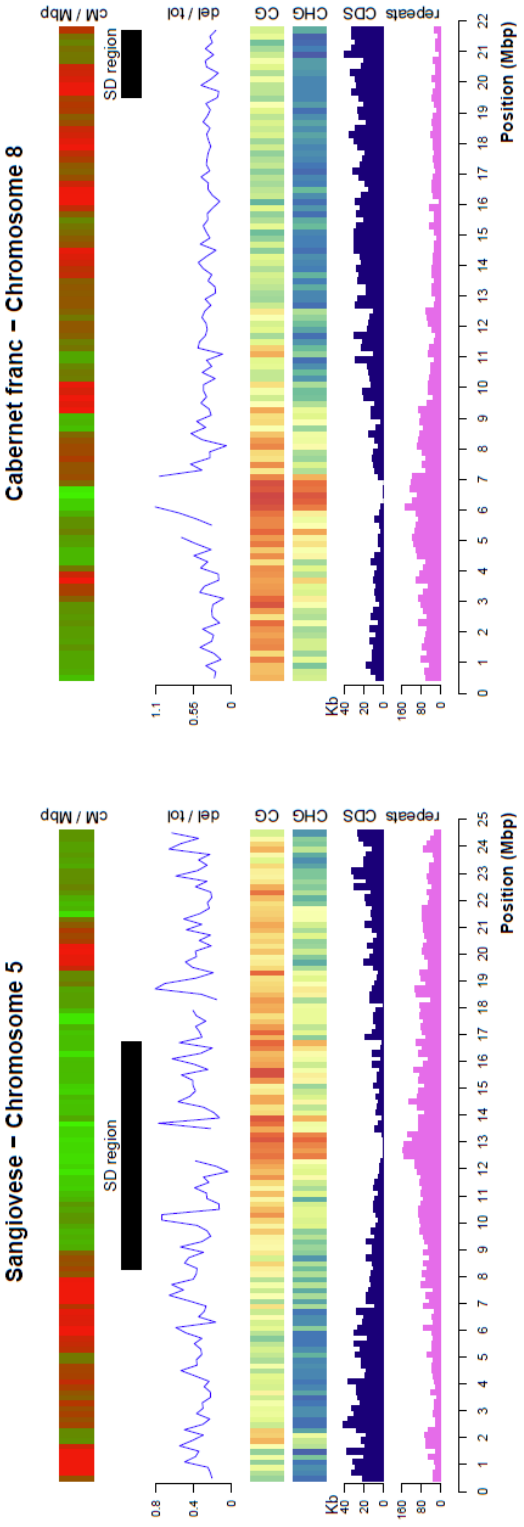
To test the relationship between the recombination frequency and the other genomic features, regions were this measure could not be estimated (often corresponding to centromeric regions) were removed from analysis. Box plots of Figure 21 show the distribution of each genomic feature in low (green), intermediate (yellow) and high (red) recombining regions of the genome. At increasing levels of recombination, gene density increases, while repeat density and methylation profiles in the CG and CHG contexts decrease (Wilcoxon Mann Whitney test, p value < 0.001). No significant trend was observed for the distribution of the ratio of deleterious over tolerated SNPs. Summarizing, low recombining regions showed high levels of repeat density and methylation, ad low gene density, describing a scenario coherent with an inactive transcriptomic context. High-recombining regions correlated positively with gene density and negatively with repeat density and methylation profiles, indicating actively transcribed chromatin regions (Paape et al., 2012; N. Yelina, Diaz, Lambing, & Henderson, 2015; N. E. Yelina et al., 2012).

As a consequence of their location in low-recombining regions of the genome, loci of segregation distortion on chromosome 5 (8.7 Mbp) and on chromosome 9 (5.4 Mbp) in Sangiovese, and on chromosome 15 (7.8 Mbp) in Pinot Noir showed the largest size.

Loci of SD located in highly-recombining regions were smaller in size compared to SD loci in low-recombining regions (size ranged from 0.26 Mbp on chromosome 8 in Sangiovese to 2.8 Mpb on chromosome 11 in Primitivo).

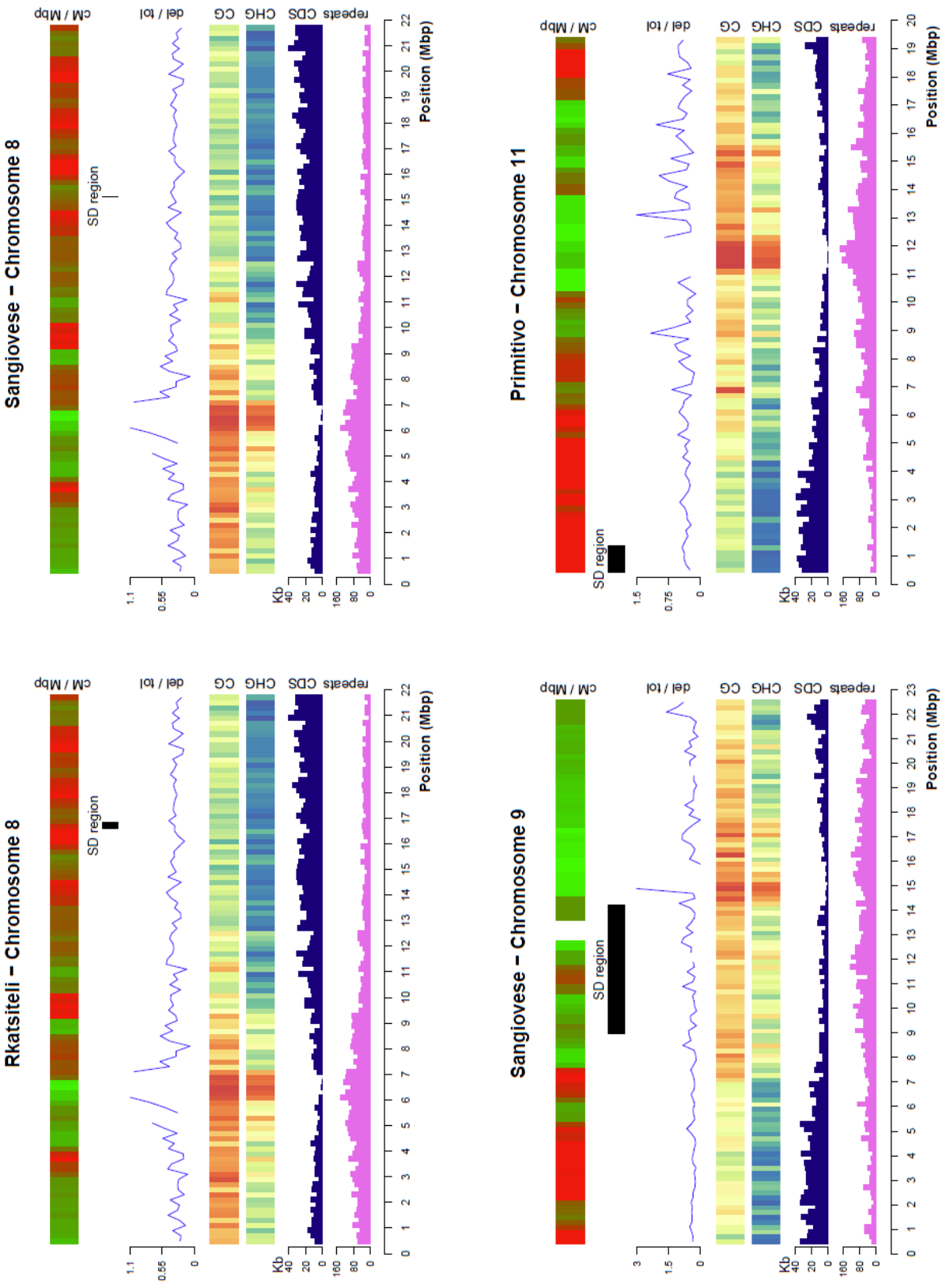
The equal distribution of the loci of SD in the tree categories of recombination frequency shows that their location is independent from the recombination rate. This evidence suggests that deleterious alleles causing segregation distortion in the progenies of a single cross generation may locate in the genome independently of the recombination frequency. From the literature it is known that recombination facilitates the removal of damaging alleles from a population, thanks to a higher efficacy of selection in removing them (B. Charlesworth, 2007; Chun, Fay, & Pritchard, 2011; Rodgers-Melnick et al., 2012). Thus, deleterious alleles in the loci of SD residing in high recombining regions may be removed faster than deleterious alleles in low recombining regions.


Schiava grossa – Chromosome 4
SD region
cM / Mbp
del / tot
CG
CHG
CDS
repeats
Kb
Position (Mbp)


Cabernet franc – Chromosome 8
SD region
cM / Mbp
del / tot
CG
CHG
CDS
repeats
Kb
Position (Mbp)

Primitivo – Chromosome 4
SD region
cM / Mbp
del / tot
CG
CHG
CDS
repeats
Kb
Position (Mbp)

Sangiovese – Chromosome 5
SD region
cM / Mbp
del / tot
CG
CHG
CDS
repeats
Kb
Position (Mbp)


Sangiovese – Chromosome 8
Primitivo – Chromosome 11
Rkatsiteli – Chromosome 8
Sangiovese – Chromosome 9
cM / Mbp
del / tol
CG
CHG
CDS
repeats
SD region
Kb
Position (Mbp)

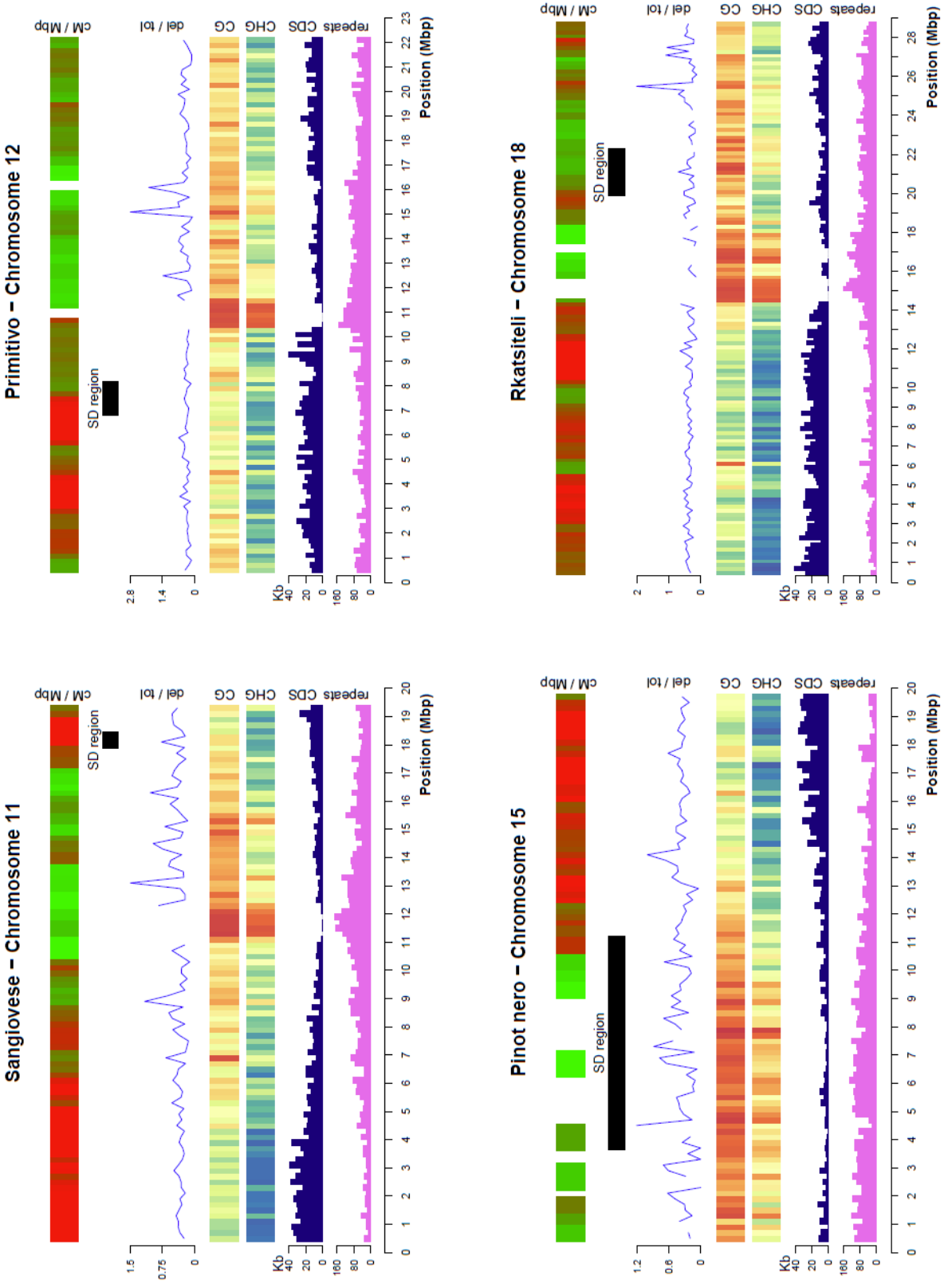
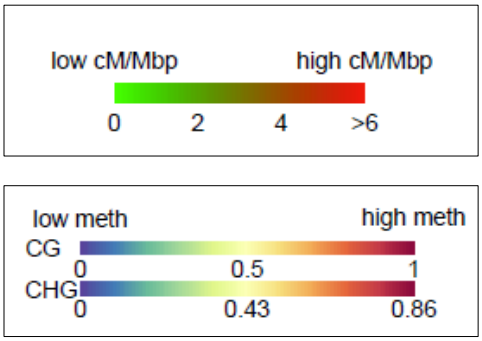

Primitivo – Chromosome 12
Rkatsiteli – Chromosome 18
Sangiovese – Chromosome 11
Pinot nero – Chromosome 15
SD region
cM / Mbp
del / tol
CG
CHG
CDS
repeats
Kb
Position (Mbp)

Figure 19. **Candidate single-locus SD regions in the genomic context**: gene (CDS, coding sequence) density, repeat density, CHG and CG methylation profiles, ratio of nonsynonymous deleterious SNPs to nonsynonymous tolerated SNPs in expressed genes and recombination frequency are plotted for each chromosome showing SD in the progenies of selfing.

Legend of Figure 20.



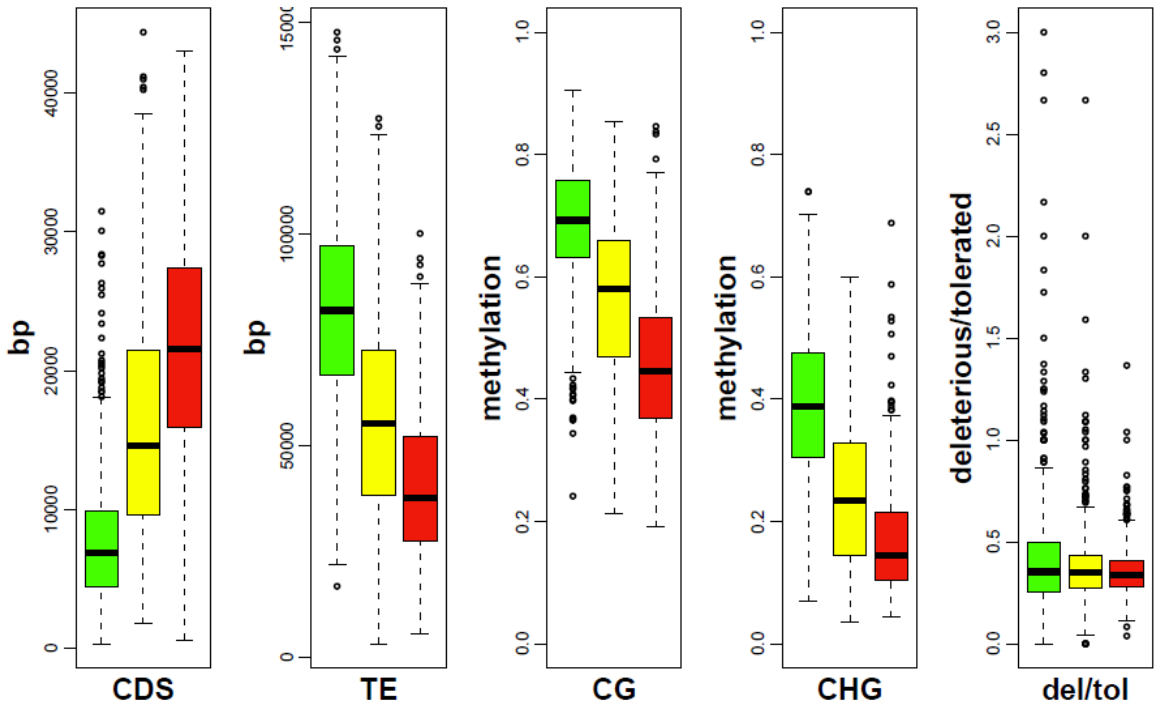


Figure 21. **Box plots of each genomic feature in regions of low (green), intermediate (yellow) and high (red) recombination frequency.** Five genomic features are described: gene density (CDS, coding sequence), repeat density (TE), CG and CHG methylation profiles and the ratio of deleterious SNPs to tolerated SNPs (del/tol). On the y-axis, count of CDS and TE in bp, average methylation level of CG and CHG contexts in 200 Kb windows, ratio of deleterious over tolerated SNPs. Relation between the recombination frequency and the distribution of the related feature was significant for CDS, TE, CG and CHG (Wilcoxon Mann Whitney, p values < 0.001). The distribution of the ratio deleterious SNPs over tolerated SNPs showed no significant relation with the recombination frequency.

Candidate causal mutations

Skewed genotypic frequencies in loci of SD can derive from occurrence in homozygosis (e.g. after selfing) of a lethal/deleterious allele. Damaging consequence of a damaging allele can vary based on the severity of mutation, on the level of expression in the heterozygote (i.e. the degree of dominance), and on its effect on fitness. Once loci of SD were defined in the progenies of selfing, we started searching for candidate causal mutations affecting viability and survival in the parent

varieties. Causal mutations may include small variants such as SNPs and INDELs and structural variants such as deletions and insertions.

In order to restrict the list of candidate mutations, we again fine-mapped loci by taking advantage of the haplotype phase in the population of 128 varieties. Among polymorphic alleles for deleterious mutations in the parent variety, we excluded those found in homozygosis in at least one variety of the population. Ideally, if the effect of the mutation is lethal, it should be never inherited in homozygosis and we should not find it in the grapevine population. We also considered whether the genes were expressed in the varieties (there may be mutations in pseudogenes) and, more importantly, whether the candidate alleles belonged to the distorted haplotype, when this information was available.

We considered as putative candidates for lethal phenotype the mutations falling in the categories of stopgain and stoploss, of frameshift INDELs, and of SVs encompassing genes (summarized in Table 18). We considered as less damaging (and less likely to result in lethal phenotype) those mutations falling in the category of nonsynonymous SNPs with deleterious effect (summarized in Table 19) and we define them as deleterious.

For each type of mutation with lethal effect as described above, the number of mutations and the number of genes involved are described in Table 18. Table 19 shows the number of genes harbouring nonsynonymous SNPs with deleterious or with tolerated effect. Both Table 18 and Table 19 refers to loci of SD after the fine-mapping.

Variety	Locus on chromosome	N. genes hit by lethal mut	N. lethal mutations	Deletions	Insertions	Frameshift deletions	Frameshift insertions	Stopgain SNP	Stoploss SNP
Cabernet franc	8	10	14	7	0	2	2	3	0
Pinot nero	15	6	7	0	5	0	1	1	0
Primitivo	4	6	6	0	4	0	0	2	0
Primitivo	11	3	3	0	1	1	0	1	0
Primitivo	12	4	4	0	2	0	1	1	0
Rkatsiteli	8	0	0	0	0	0	0	0	0
Rkatsiteli	18	2	2	0	1	1	0	0	0
Sangiovese	5	10	12	2	7	1	2	0	0
Sangiovese	8	2	2	0	0	1	0	1	0
Sangiovese	9	11	13	3	1	4	0	4	1
Sangiovese	11	3	3	0	0	0	0	3	0
Schiava grossa	4	1	1	0	0	0	1	0	0

Table 18. Number of genes encompassed by lethal mutations, and categories of candidate mutations in loci after the fine-mapping through the analysis at population level.

Variety	Locus on chromosome	N. genes hit by nonsyn mut	N. deleterious nonsyn SNPs	N. tolerated nonsyn SNPs
Cabernet franc	8	19	39	79
Pinot nero	15	7	10	16
Primitivo	4	16	29	36
Primitivo	11	4	5	13
Primitivo	12	10	11	23
Rkatsiteli	8	3	3	5
Rkatsiteli	18	9	13	25
Sangiovese	5	16	22	41
Sangiovese	8	1	1	6
Sangiovese	9	33	53	90
Sangiovese	11	10	16	26
Schiava grossa	4	0	0	0

Table 19. Number of nonsynonymous SNPs, with deleterious and tolerated effect, and number of involved genes in fine-mapped loci.

We reported a more detailed description of mutations heavily affecting or disrupting gene function in Table 20 (it refers to the list of Table 18).

We did not find any lethal mutations in the locus on chromosome 8 in Rkatsiteli. Thus, in Table 20 we reported details about deleterious SNPs found in the locus. Overall, 58 genes were affected by lethal mutations in 11 loci of SD and three genes were affected by deleterious mutations in the locus on chromosome 8 in Rkatsiteli, for a total of 61 candidate genes for distortion. Gene ID and gene function annotation are referred to the V2.1 annotation of the grapevine genome (Vitulo et al., 2014). Chromosome (Chr V), starting coordinate (Start V) and end coordinate (End V) refer to the coordinates of the candidate mutations. Expression of the candidate genes is also reported: the first value reports gene expression data in the whole panel of varieties analysed through RNA-seq ("Expr. All varieties"); the second value considers expression data of the variety in which SD was assessed (this last information is not available for Primitivo and for Pinot Noir). Association of each deleterious allele to the distorted haplotype was carried out for the four low-coverage

progenies (Cabernet Franc, Primitivo, Rkatsiteli, and Sangiovese), when information on the phase of polymorphic position was available. Locus on chromosome 11 in Sangiovese apparently segregated as overdominant. If it was truly overdominant, we should have a single locus in which the only not-lethal combination of alleles could be found in heterozygosis. However, we could never find a locus in which only heterozygous individuals were observed. This evidence suggests that the region of segregation distortion on chromosome 11 in Sangiovese may be a pseudo-overdominant locus. In this configuration, two linked loci would be involved, and each of them would carry a deleterious allele. Homozygous state of the deleterious allele at each of the two loci would lower fitness in the progenies, while high-fitness combination would be allowed via-complementation of the two loci in heterozygosis. If this hypothesis was true, we would expect that each pair of mutations found in the locus could be a potential candidate for the segregation distortion observed in the progenies of selfing.

For haplotype phasing of deleterious alleles in Pinot Noir and Schiava Grossa varieties, we adopted another strategy. The individual used to assemble the reference genome, PN40024, derives from repeated selfing of a Pinot Noir seed parent. During the process of self-fertilization, the seed parent was accidentally pollinated by Helfensteiner (a variety resulting from a cross between Pinot Noir and Schiava Grossa). As consequence, part of the reference genome corresponds to one of the haplotypes of Pinot Noir and part to one of the haplotypes of Schiava Grossa. Based on SNP information, regions of the reference genome were classified as either derived from Pinot Noir or from Schiava Grossa. Attribution of the putative causal mutation to the haplotype depleted in loci of SD can be performed in Pinot noir and in Schiava Grossa if the region corresponds to the portion of the genome in which the reference haplotype is donated by the studied cultivar (or, relatively common case, the reference haplotype corresponds to one haplotype of both the cultivars). In Pinot Noir, the locus of SD corresponded to a region in which the haplotypes of Schiava Grossa and Pinot Noir could not be distinguished; thus, no information about phase could be inferred on alleles for this locus.

Locus on chromosome 4 in Primitivo:

Gene ID	Annotation V2.1	Chr V	Start V	End V	Type	Expr. All varieties / Primitivo	On distorted haplotype
04s0008g04010	lupus la	chr4	3349891	3349891	stopgain	yes / na	yes
04s0008g04160	burp domain containing protein	chr4	3483308	3483360	insertion	yes / na	na
04s0008g04520	uncharacterized protein	chr4	3920841	3920905	insertion	yes / na	no
04s0008g04610	pentatricopeptide repeat containing protein	chr4	4058378	4058378	stopgain	yes / na	na
04s0008g05430	rna dependent rna polymerase 6	chr4	4877631	4877643	insertion	yes / na	yes
04s0008g05800	sister chromatid cohesion protein pds5 like protein	chr4	5381661	5381667	insertion	yes / na	no

Locus on chromosome 4 in Schiava Grossa:

Gene ID	Annotation V2.1	Chr V	Start V	End V	Type	Expr. All varieties / S. Grossa	On distorted haplotype
04s0079g00620	rna polymerase beta subunit	chr4	11382735	11382735	frameshift insertion	yes / na	yes

Locus on chromosome 5 in Sangiovese:

Gene ID	Annotation V2.1	Chr V	Start V	End V	Type	Expr. All varieties / Sangiovese	On distorted haplotype
05s0049g01620	cyclopropane fatty acyl phospholipid synthase	chr5	9047438	9047438	frameshift insertion	yes / yes	yes
05s0049g01710	transducin wd 40 repeat containing protein	chr5	9151045	9151199	insertion	no / no	yes
05s0049g01710	transducin wd 40 repeat containing protein	chr5	9151526	9151526	frameshift insertion	no / no	yes
05s0051g00050	myosin xi 2	chr5	10221635	10221727	insertion	yes / yes	yes
05s0051g00730	ring u box domain containing protein	chr5	11703944	11705102	deletion	yes / yes	yes
05s0051g00730	ring u box domain containing protein	chr5	11705897	11705905	insertion	yes / yes	yes
05s0051g00830	dihydroxy acid dehydratase	chr5	11895454	11895466	insertion	yes / yes	yes
05s0136g00140	uncharacterized protein sll0005 like	chr5	13443503	13443514	insertion	yes / yes	yes
05s0029g00656	beta galactosidase 9	chr5	15704808	15704812	insertion	yes / yes	yes
05s0029g00730	methyltransferase like protein 13 like	chr5	15897532	15897533	frameshift deletion	yes / yes	na
05s0029g00830	uncharacterized protein	chr5	16212633	16213005	insertion	yes / yes	yes
05s0029g00850	tmv resistance protein n like	chr5	16228754	16236942	deletion	yes / yes	yes

Locus on chromosome 8 in Cabernet Franc:

Gene ID	Annotation V2.1	Chr V	Start V	End V	Type	Expr. All varieties/ C. franc	On distorted haplotype
08s0007g05540	eukaryotic rpb5 rna polymerase subunit family	chr8	19452130	19461358	deletion	yes / yes	yes
08s0007g05550	protein	chr8	19452130	19461358	deletion	yes / yes	yes
08s0007g05664	protein	chr8	19564291	19564291	frameshift insertion	yes / yes	no
08s0007g06070	no hit	chr8	19879233	19879233	frameshift insertion	no / no	na
08s0007g06180	myb like hth transcriptional regulator family protein	chr8	19994961	19994962	frameshift deletion	yes / yes	yes
08s0007g06570	probable l type lectin domain containing receptor kinase	chr8	20289874	20292707	deletion	no / no	yes
08s0007g06570	probable l type lectin domain containing receptor kinase	chr8	20290486	20290486	stopgain	no / no	na
08s0007g06570	probable l type lectin domain containing receptor kinase	chr8	20291274	20291274	stopgain	no / no	na
08s0007g06570	probable l type lectin domain containing receptor kinase	chr8	20291281	20291283	frameshift deletion	no / no	na
08s0007g06580	probable l type lectin domain containing receptor kinase	chr8	20289874	20292707	deletion	no / no	yes
08s0007g06580	probable l type lectin domain containing receptor kinase	chr8	20293073	20295310	deletion	no / no	na
08s0007g08500	60s ribosomal export protein nmd3 like	chr8	21904754	21904754	stopgain	yes / no	na
08s0007g08600	no hit	chr8	21983299	21989007	deletion	yes / yes	na
08s0007g08610	no hit	chr8	21983299	21989007	deletion	yes / yes	na

Locus on chromosome 8 in Rkatsiteli:

Gene_ID	Annotation V2.1	chr V	start V	end V	type	Provean score	Expr. All varieties/ Rkatsiteli	On distorted haplotype
08s0007g02480	cytokinin riboside 5 monophosphate	chr8	16607301	16607301	del SNP	-6.725	yes / no	yes
08s0007g02610	tld domain containing nucleolar protein	chr8	16723114	16723114	del SNP	-3.22	yes / yes	yes
08s0007g02710	short chain alcohol	chr8	16801600	16801600	del SNP	-3.408	no / no	yes

Locus on chromosome 8 in Sangiovese:

Gene ID	Annotation V2.1	Chr V	Start V	End V	Type	Expr. All varieties / Sangiovese	On distorted haplotype
08s0007g00810	concanavalin a like lectin kinase like protein	chr8	14993602	14993602	stopgain	yes / yes	yes
08s0007g01100	uracil phosphoribosyltransferase	chr8	15218272	15218273	frameshift deletion	yes / yes	yes

Locus on chromosome 9 in Sangiovese:

Gene ID	Annotation V2.1	Chr V	Start V	End V	Type	Expr. All varieties / Sangiovese	On distorted haplotype
09s0002g08300	disease resistance protein rga4 like	chr9	9102428	9102428	stopgain	yes / yes	yes
09s0002g08560	uncharacterized protein	chr9	9646410	9646410	stopgain	yes / no	yes
09s0002g08800	no hit	chr9	10186248	10186256	frameshift deletion	no / no	na
09s0002g08800	no hit	chr9	10186609	10186611	frameshift deletion	no / no	na
09s0002g08940	nuclear pore complex protein nup98 nup96	chr9	10502063	10502063	stopgain	yes / yes	na
09s0002g09250	cytochrome p450 82c4	chr9	11038534	11038534	stoploss	no / no	yes
09s0002g09250	cytochrome p450 82c4	chr9	11040939	11042073	deletion	no / no	na
09s0096g00420	probable disease resistance protein at5g63020 like	chr9	11827141	11827259	insertion	yes / yes	na
09s0096g00490	hydroxycinnamoyl coenzyme a shikimate quinate	chr9	11932706	11932707	frameshift deletion	no / no	yes
09s0070g00150	no hit	chr9	13179122	13179122	stopgain	no / no	na
09s0070g00240	cinnamoyl reductase	chr9	13367322	13367323	frameshift deletion	yes / yes	no
09s0070g00460	no hit	chr9	13674746	13681154	deletion	yes / yes	na
09s0070g00470	zinc finger ccch domain containing protein 16	chr9	13674746	13681154	deletion	yes / yes	na

Locus on chromosome 11 in Primitivo:

Gene ID	Annotation V2.1	Chr V	Start V	End V	Type	Expr. All varieties / Primitivo	On distorted haplotype
11s0016g00460	kinesin like protein	chr11	454293	454601	insertion	yes / na	na
11s0016g01610	uncharacterized protein	chr11	1299819	1299819	stopgain	yes / na	na
11s0016g01620	ankyrin repeat containing protein at2g01680 like	chr11	1309054	1309055	frameshift deletion	no / na	na

Locus on chromosome 11 in Sangiovese:

Gene ID	Annotation V2.1	Chr V	Start V	End V	Type	Expr. All varieties / Sangiovese	On distorted haplotype
11s0052g00760	no hit	chr11	18314330	18314330	stopgain	no / no	*
11s0052g00985	glutamate gated kainate type ion channel receptor subunit 5	chr11	18636798	18636798	stopgain	no / no	*
11s0052g01270	probable xyloglucan endotransglucosylase	chr11	18993031	18993031	stopgain	yes / yes	*

Locus on chromosome 12 in Primitivo:

Gene ID	Annotation V2.1	Chr V	Start V	End V	Type	Expr. All varieties / Primitivo	On distorted haplotype
12s0059g02490	spx and exs domain containing protein 1 like	chr12	7273203	7273211	insertion	yes / na	yes
12s0059g02540	cysteine rich receptor like protein kinase 25 like isoform	chr12	7315218	7315223	insertion	no / na	yes
12s0134g00270	g type lectin s receptor like serine threonine protein	chr12	7812749	7812749	frameshift insertion	yes / na	yes
12s0134g00280	g type lectin s receptor like serine threonine protein	chr12	7818159	7818159	stopgain	yes / na	na

Locus on chromosome 15 in Pinot Noir:

Gene ID	Annotation V2.1	Chr V	Start V	End V	Type	Expr. All varieties / P. Noir	On distorted haplotype
15s0045g00110	60s ribosomal protein l4 l1	chr15	4648068	4648068	frameshift insertion	no / no	na
15s0045g00540	protein transport protein sec23	chr15	5414555	5414560	insertion	yes / yes	na
15s0045g00540	protein transport protein sec23	chr15	5431657	5431667	insertion	yes / yes	na
15s0045g00550	uncharacterized protein	chr15	5491983	5492382	insertion	yes / yes	na
15s0045g01170	3 hydroxyisobutyryl hydrolase 1	chr15	6628291	6628298	insertion	yes / yes	na
15s0107g00150	glutathione s	chr15	8191614	8191614	stopgain	no / no	na
15s0021g01430	aminoacyl t rna synthetase	chr15	11827986	11828141	insertion	yes / yes	na

Locus on chromosome 18 in Rkatsiteli:

Gene ID	Annotation V2.1	Chr V	Start V	End V	Type	Expr. All varieties / Rkatsiteli	On distorted haplotype
18s0072g01080	dnaj heat shock n terminal domain containing	chr18	20632016	20632017	frameshift deletion	yes / yes	yes
18s0075g00130	fanconi anemia group m protein	chr18	21290884	21290889	insertion	yes / yes	yes

Start V: starting coordinate of the mutation
End V: ending coordinate of the mutation
Expression threshold: > 1 FPKM in at least one tissue of one of the six varieties analysed for gene expression and > 1 FPKM in at least one tissue of the variety carrying locus of segregation distortion.
na: not available
del SNP: nonsynonymous SNP with deleterious effect
*: either over-dominant or pseudo-overdominant models may explain the locus segregation mode. If the second model was true, two loci in repulsion phase acting via-complementation would be involved.

Table 20. The list of lethal mutations (stop codons gain or loss, frameshift INDELs, insertions, and deletions) is provided for each SD locus after the fine-mapping. List of deleterious mutations (nonsynonymous deleterious SNPs) is shown for the SD locus on chromosome 8 in Rkatsiteli.

On the base of haplotype phasing data, mutations that did not locate on the haplotype carrying the distortion were discarded from the list showed in Table 20: four out of 61 genes were then discarded from further analysis.

To further explore candidate genes, we asked whether they were single-copy or duplicated, i.e. whether genes had functional redundant copies. Single-copy genes with essential function affected by lethal mutations represent good candidates for further studies, because their function cannot be replaced, if compromised. At this stage, we aimed to give a very preliminary overview of candidate genes and we performed BLASTp alignment of translated candidate gene sequences against the *Vitis vinifera* proteome. Further analysis should be performed to restrict the list of candidates for segregation distortion: for example, structure motifs and patterns should be considered to identify proteins with the same biochemical function. Through the analysis of sequence similarity using BLASTp we detected 36 single-copy genes and 21 duplicated genes. Among the 21 putatively duplicated genes, 14 were found to have at least one expressed copy (Appendix 3). The duplicated copies of three genes not expressed in the corresponding variety also showed no expression. No expression data was available for the four duplicated genes in Primitivo. This analysis is very preliminary, but suggests that at least 14 genes may be excluded from further analysis and that three genes may be pseudogenes.

4.3 Identification and analysis of a reciprocal translocation in Rkatsiteli

4.3.1 Linkage and epistatic interaction

Fisher's test

In order to detect epistatic interactions between independent loci in the progenies of selfing, pairwise comparison of genotyped loci was tested by the Fisher's exact test. We tested whether loci on different chromosomes showed a deviation from the expected segregation ratio for a dihybrid cross. We could detect a signal of two-loci interaction in the progenies of Rakstiteli self-cross. Figure 22 shows the *fdr*-corrected p values of the pairwise comparisons that were significant for an α level of confidence of 0.05. Signal on the diagonal simply represents signal of each chromosome compared to itself. In the area above the diagonal, a strong signal of interaction was detected between markers of chromosome 1 and of chromosome 11.

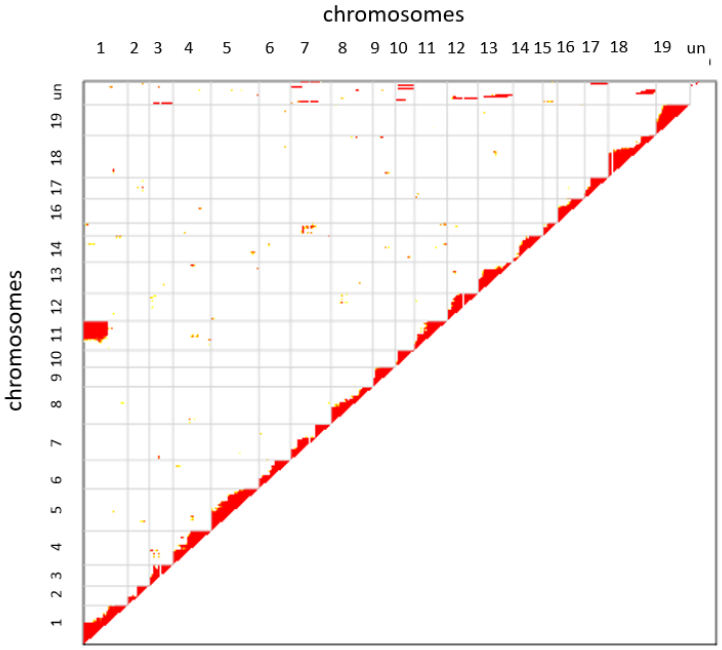


Figure 22. **Fisher's exact test for chromosome pairwise comparisons.**
Chromosome number is indicated on the axes (un: unknown chromosome). P values significant for α = 0.05 are plotted in red. Signal on the diagonal represents the signal of each chromosome compared with itself.
Signal in the area above the diagonal represents significant signal for two-loci comparison.

Pearson's Chi square test

The Pearson's Chi square test was used to detect epistatic alleles exerting lethal/deleterious effects in homozygosis in the progenies of selfing. If damaging alleles at two interacting loci were inherited in homozygosis, we would expect a significant decrease or the absence of one classes of double homozygous in the segregating population. This may be the case for duplicated and functionally redundant genes. We detected the signal between chromosomes 1 and 11 in Rkatsiteli, but we did not find signals of epistatic interaction in the other progenies. The progeny of Rkatsiteli was the largest in size among the six progenies of selfing under study, and the detection of signal in this population confirmed we had the statistical power to perform the test in this progenies. The lack of significant signal may be due to the smaller size of the other progenies. Indeed, we were seeking one of the four classes of double homozygotes, which ideally represent 1/16 of the individuals. Furthermore, if loci were deleterious but not lethal, a significant signal would be even more difficult to detect.

Genetic linkage

The genetic map of Rkatsiteli also showed a strong linkage signal between chromosome 1 and chromosome 11 (interaction between the corresponding LGs resulted in a cross-shaped linkage map, showed in Figure 23). An additional investigation into the type of epistatic interaction revealed that actually this signal was not due to epistasis between the loci (only three of the nine expected genotypes were observed, corresponding to two double homozygous genotypes and the double heterozygous one), but rather to a pseudo-linkage signal indicating a physical interaction between the chromosomes. This pseudo-linkage is explained genetically by the fact that markers at non-homologous chromosomes appear to be linked if the chromosomes take part in a translocation and the loci are close to the translocation breakpoint. The apparent linkage of markers known to be on separate non-homologous chromosomes is a genetic giveaway for the presence of a translocation. All the evidences of interaction strongly favoured the hypothesis of a balanced translocation between portions of the two chromosomes.

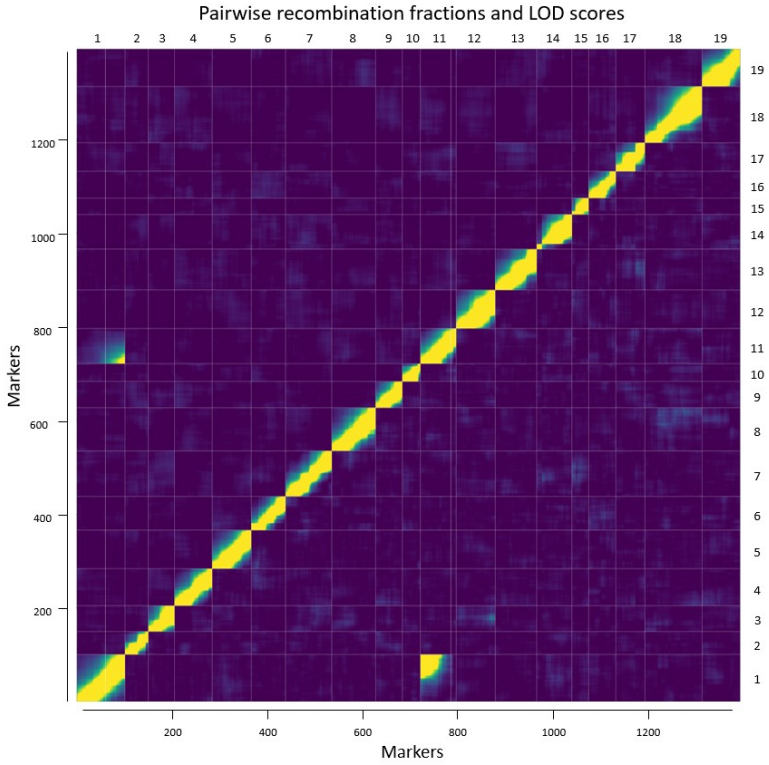
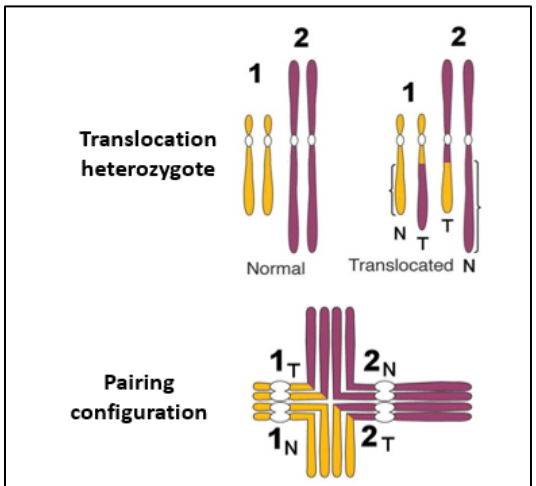


Figure 23. **Heat map of LOD scores and recombination fractions in Rkatsiteli (from the F2-like mapping population).** The 19 LGs (in yellow, along the diagonal) correspond to the 19 chromosomes of the *Vitis vinifera* genome.
The strength of genetic linkage between two markers in the same LG becomes greater both as the recombination frequency decreases and as the LOD score increases.

Homologous regions of reciprocally translocated chromosomes can still pair to form synapses in meiosis. The characteristic configuration deriving from this type of pairing is that of a cross, illustrated in Figure 24. Pairing configuration can be resolved in three different ways to form meiotic products. The segregation of each of the structurally normal chromosomes with one of the translocated chromosomes ($1_T + 2_N$ and $1_N + 2_T$) is called Adjacent-1 segregation. Both meiotic products are duplicated and deleted for different regions. On the other hand, the two normal chromosomes (1_N and 2_N) can segregate together, as well as the two translocated chromosomes, generating $1_N + 2_N$ and $1_T + 2_T$ meiotic products. This type of segregation is called Opposite and meiotic products are complete and viable. Adjacent-1 and Opposite segregation patterns are equally frequent, and thus generate half of the possible alleles, respectively. There is another pattern of segregation, called Adjacent-2, in which homologous centromeres migrate to the same pole, generating meiotic products carrying deletions and duplications. Anyway, this type of segregation pattern is rare. Since Adjacent-1 and Opposite segregation patterns are equally

probable, half the gametes will bear an unbalanced chromosomic content and will be incapable of contributing to the next generation, a condition known as semisterility. The condition of semisterility is an important diagnostic for a balanced translocation in heterozygosis (Griffiths AJF, Miller JH, et al., 2000). An analysis of pollen germination was performed to detect the fraction of male gametes showing sterility (described in par 4.3.2).



Adjacent-1	Products	Effect
$1_T + 2_N$ $1_N + 2_T$	Duplications and deletions	Often not viable
Adjacent-2 (rare)	**Products**	**Effect**
$2_N + 2_T$ $1_N + 1_T$	Duplications and deletions	Often not viable
Opposite	**Products**	**Effect**
$1_T + 2_T$ $1_N + 2_N$	Both complete	Viable

Figure 24. **Chromosome segregation patterns**. Meiotic products resulting from three patterns of chromosome segregation (Adjacent-1, Adjacent-2, and Opposite) in a translocation heterozygote.

Rkatsiteli is heterozygous for the balanced translocation. Looking at the ratio of the genotypic classes in the progenies of selfing, we detected higher than expected proportion of two classes of double homozygotes (corresponding to genotypes homozygous for the balanced translocation and homozygous for the normal chromosomes) and of the class of double heterozygotes (parental genotype) (genotypes in red boxes in the low panel of Figure 25). We also observed the total

absence of the remaining genotypic classes. These corresponded to the genotypic combinations resulting in deletions and duplications of wide genomic regions in one or both chromosomes.
In summary, we observed an excess of genotypic classes deriving from the combination of normal gametes ($1_N + 2_N$) and of gametes carrying the translocation ($1_T + 2_T$); and the lack of genotypic classes deriving from the combinations between the two types of gametes (mainly $1_T + 2_N$ and $1_N + 2_T$). This evidence suggested that genotypes carrying duplications and deletions might result in non-viable genotypes or that gametes carrying unbalanced genetic content could not even form or were not fertile. Analysis on pollen viability was performed in Rkatsiteli in order to measure the rate of pollen germination.

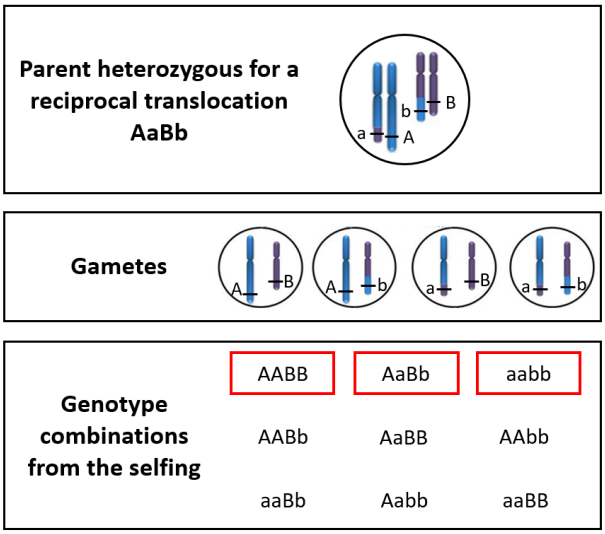


Figure 25. **Gametes generated by a translocation heterozygote and genotypes combination in the progenies of selfing**. Top: genotype of a translocation heterozygote. Middle: genetic content of gametes deriving from the parent. Bottom: genotype combinations in the progenies deriving of selfing. Genotypes showing a complete genomic content are surrounded by red boxes. The remaining genotypes all harbour deletions and/or duplications in their genomic content.

4.3.2 Analysis of pollen germination

Semisterility appears when half the gametes are incapable of contributing to the next generation because of a halved number of viable and, hence, fertile meiotic products. In the thirties, Burnham hypothesized that the reason of 50% of pollen and ovules abortion in partially sterile lines of maize was due to translocation of a portion of a chromosome to a non-homologue one. When semisterile plants were selfed or crossed with plants showing normal fertility, F1 progenies

showed a ratio of 1 normal : 1 semisterile (Burnham, 1930). This condition arises in translocation heterozygotes as a result of the equality of Adjacent-1 and Opposite segregation modalities, because 50% of the spores either would be deficient or would possess a portion in duplicate (see Figure 24). Thus, semisterility can be used as a diagnostic to detect a translocation in heterozygosis. We tested pollen viability by measuring the rate of pollen grains developing pollen tube in nine clones of Rkatsiteli (Table 21) and in replicates of Cabernet Franc, Kishmish Vatkana, Pinot Noir, Sangiovese, and Traminer, used as control varieties (Table 22). Blooming in control varieties is earlier as compared to Rkatsiteli, which is characterized instead by late flowering. Pollen of Pinot Noir and Traminer was not available for sampling on batch 3, because blooming time had already ended at this day time. Consequently, sampling was carried out for batch 1 and 2 for these two varieties. Replicates containing clumps of pollen due to humidity on flower clusters at the time of sampling could not be measured, since pollen grains cannot be isolated for counting. The average percentage of pollen germination for each variety and number of replicates are indicated in Table 22. Provided that pollen germination never reaches exactly 100%, germination percentage varied widely among the varieties in the considered flowering interval of measurement. Maximum germination rate value in control varieties was measured in time interval of batch1 and batch 2 and ranged from 88 to 92%. Expected maximum germination rate in Rkatsiteli was estimated around 44 to 46%, that is to say half of the maximum germination rate observed in control varieties. Indeed, in the hypothesis that gametes with unbalanced genetic content can form and can give rise to pollen grains, and that lethality arises at this latter stage, we expect that 50% of grains cannot develop pollen tube. Maximum germination rate observed in Rkatsiteli reached 72% at day time of batch 3. This value represents the average germination percentage among 17 replicates of batch 3, and standard deviation was calculated in the same way (Table 22). In Table 21, standard deviation (SD) was calculated within replicates of each clone, in batch 3: germination rate in Rkatsiteli is higher than expected (Chi square test, p value < 0.05). On the overall, germination rate measured in Rkatsiteli is lower compared to germination rate measured in control varieties, but it is higher than the expected (Chi square test, p value < 0.05). The observation of pollen tube growth in 72% of pollen grains suggests that either pollen grains deriving from gametes with unbalanced content can only partially germinate, or gametes with unbalanced content cannot form at all.

	batch 1 (6/1/2015)		batch 2 (6/3/2015)		batch 3 (6/4/2015)		
	average %[1]	n. rep[2]	average %[1]	n. rep[2]	average %[1]	SD[3]	n. rep[2]
Rkatsiteli clone 1	67.07	2	55.48	1	75.89	3.10	2
Rkatsiteli clone 2	50.13	2	55.80	2	71.68	5.35	2
Rkatsiteli clone 3	62.44	2	54.84	2	73.01	8.61	2
Rkatsiteli clone 4	66.29	2	66.39	1	72.97	11.14	2
Rkatsiteli clone 5	56.73	2	47.47	2	75.42	-	1
Rkatsiteli clone 6	62.88	2	58.32	2	71.97	13.41	2
Rkatsiteli clone 8	71.08	2	61.29	1	74.06	0.15	2
Rkatsiteli clone 9	68.67	2	67.81	1	59.16	6.86	2
Rkatsiteli clone 10	69.77	2	53.96	2	73.48	5.03	2

[1] average percentage of pollen germination between replicates for each clone.
[2] number of replicates for each clone.
[3] standard deviation measured on replicates within each clone.
na: not available.

Table 21. Germination rate comparison on batch 1, 2, and 3 for nine clones of Rkatsiteli. One or two replicates are sampled for each clone.

	batch 1 (6/1/2015)		batch 2 (6/3/2015)		batch 3 (6/4/2015)	
	average %[1]	n. rep[2]	average %[1]	n. rep2	average %[1]	n. rep[2]
Rkatsiteli*	63.89 ± 8.52	18	57.93 ± 6.28	14	71.96 ± 7.79	17
Cabernet Franc	89.75	1	60.48	2	50.40	2
Kishmish Vatkana	88.64	1	91.63	2	67.24	2
Pinot Nero	89.00	1	85.71	1	na	na
Sangiovese	76.85	2	88.23	2	58.75	2
Traminer	91.60	1	90.76	1	na	na
tot counts (n.)	4,680		3,466		7,447	

[1] average percentage of pollen germination among replicates for each variety. For Rkatsiteli, value of standard deviation among all replicates is provided.
[2] number of replicates for each variety.
* values are averaged among all Rkatsiteli replicates for each batch.
na: not available.

Table 22. Germination rate comparison between Rkatsiteli and control varieties on batch 1, 2, and 3. Germination rate for Rkatsiteli is the average value among all replicates and SD values are also provided.

4.3.3 Structural analysis of the translocation

Both the pseudo-linkage signal in the genetic map of Rkatsiteli and the deviation of two-loci segregation frequency suggested a balanced translocation between portions of chromosome 1 and 11. To confirm our hypothesis, we performed a structural analysis on the alignment of sequencing reads obtained for the parent variety and for progenies homozygous for the haplotype carrying the translocation. We looked at the alignment of mate-pair and partially overlapping reads against the reference genome and against Rkatsiteli assembly and we compared the read alignment configuration in the two cases.

Upper panel of Figure 26 illustrates chromosome 1 (in blue) and 11 (in violet) as found in the reference ("Normal chromosomes"); lower panel illustrates reciprocally translocated chromosomes in Rkatsiteli. In the upper panel, dashed vertical lines indicate breakpoints on chromosomes 1 and 11; in the lower panel, dashed lines define junctions of reciprocal translocation (they are called translocation breakpoints 1 on chromosome 1, and translocation breakpoints 2 on chromosome 11). The green box represents an ideal read of Rkatsiteli belonging to the translocated chromosome and physically spanning the breakpoint 2 (lower panel). When the read is aligned against the reference genome (upper panel), it can extent on chromosome 11 until the breakpoint; the remaining part of the read aligns on chromosome 1. A read sequence which is not aligned from the first residue to the last one is said to be "clipped". Clipped reads stacked together form a "wall", because their alignment ends exactly at the same position. Read alignment configuration with respect to the reference as illustrated in Figure 26 is indicative of a translocation. The finding of the complementary alignment configuration in correspondence of the other breakpoint reveals the presence of a balanced translocation.

Translocation breakpoint 2 was defined at coordinate 2,339,872 bp on chromosome 1 and at coordinate 16,004,969 bp on chromosome 11. The finding of an overlapping read pair spanning the breakpoint 2 showed that chromosome portions are simply juxtaposed. On the contrary, the reciprocal chromosome portions are not adjacent on the breakpoint 1, and we found that they are spaced out by a TE that is not present in the reference sequence. Translocation breakpoint 1 was defined at coordinate 2,341,643 bp on chromosome 1 and at coordinate 16,005,647 bp on

chromosome 11. TE spacing out the two pieces corresponds to a DTM element of the MITE family and it is 1,010 bp long. Considering coordinates of breakpoint 1, it is evident that pieces of chromosomes that were initially broken were not simply re-joint in the new configuration, but rather that chewing activity of DNA polymerase eroded both pieces before they were joined.

Figure 27 illustrates the haplotypes of the heterozygous translocation after the reconstruction based on the structural analysis. For each translocation breakpoint, the illustration depicts a region of 100 Kbp centered on the breakpoint coordinate (upper panel: translocation breakpoint 1; lower panel: translocation breakpoint 2). Genes are represented as green arrows, class I TEs as light-blue boxes and class II TEs as red boxes.

The availability of long-range single-molecule reads for Rkatsitely allowed us to confirm the reconstruction of the haplotypes on the translocation breakpoint 1.

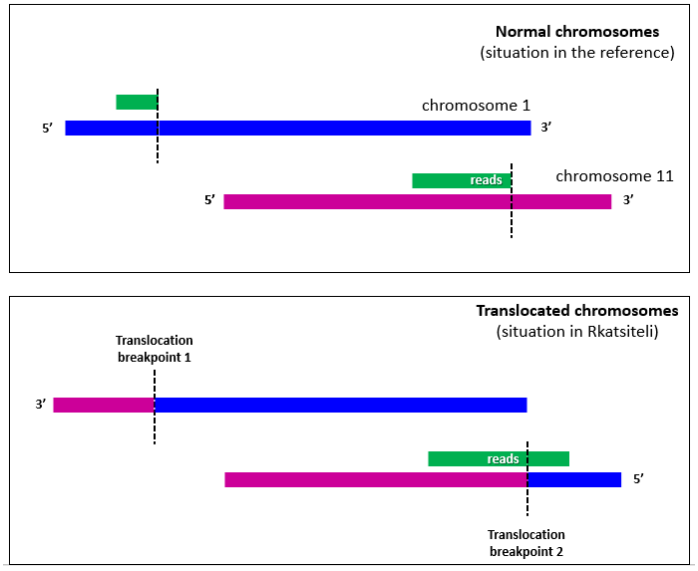


Figure 26. **Balanced translocation in Rkatsiteli**. In the upper panel, chromosome 1 and chromosome 11 are depicted in blue and violet, respectively. They represent “normal” chromosome, as found in the reference genome. In the lower panel, a balanced translocation has occurred and pieces of chromosomes have interchanged. Translocation breakpoints 1 and 2 represent junctions between newly juxtaposed pieces (dashed vertical lines). Green box represents an ideal read spanning translocation breakpoint 2.

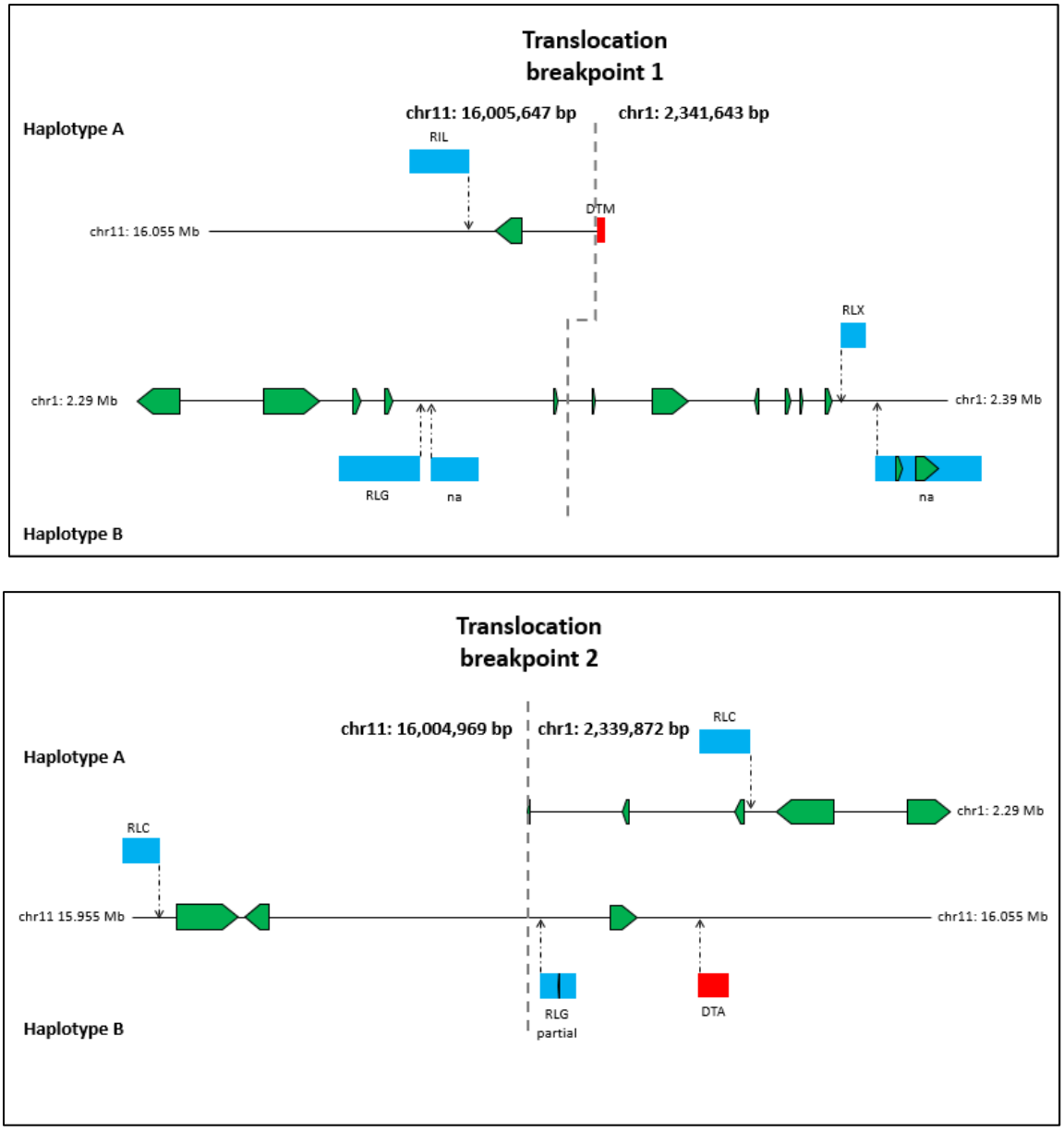


Figure 27. **Haplotypes of the balanced translocation**. Reconstructed haplotypes for 100 Kbp region centred on translocation breakpoint 1 (upper panel) and on translocation breakpoint 2 (lower panel) in Rkatsiteli. Coordinates of breakpoints are indicated. Green arrows: genes; blue boxes: class I TEs; red boxes: class II TEs.

4.3.4 Validation of the balanced translocation and germplasm analysis

In order to validate the translocation in Rkatsiteli, a PCR-based assay was carried out in the nine clones of the parent variety (Figure 28) and in four progeny of selfing (Figure 29). The PCR assay allowed us to validate the translocation breakpoint 2. Figure 28 shows three clones of Rkatsiteli that were tested. Four primer pair combinations were used to validate the heterozygous translocation. On lane 1, primers pair amplified normal chromosome 1; on lane 2, primers pair amplified normal chromosome 11; on lane 3 and 4, two different primers pairs amplified translocation breakpoint 2.

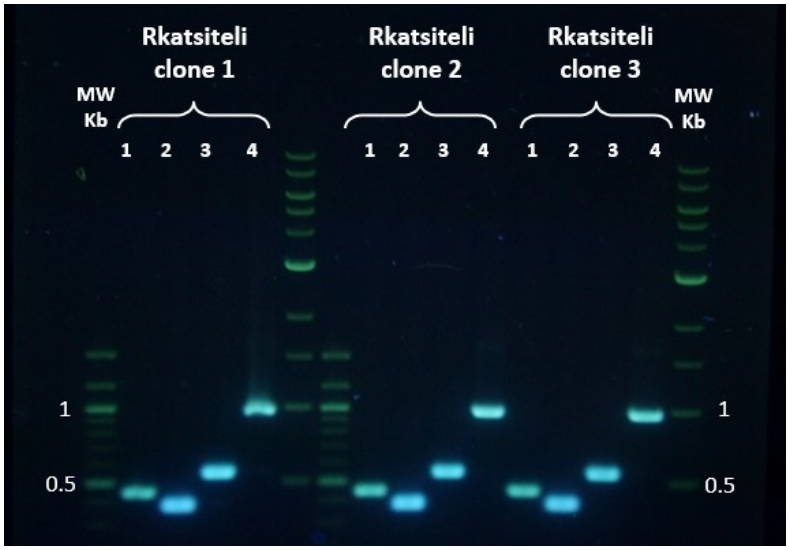


Figure 28. **Validation of the balanced translocation in heterozygosis.** Three clones of Rkatsiteli are shown. The translocation breakpoint 2 is validated by the amplification of two different primer pairs in lane 3 and 4. Amplification products of the normal chromosomes 1 and 11 are shown on lane 1 and 2, respectively, showing that Rkatsiteli is heterozygous for the translocation.

Based on the information of the reconstructed haplotypes, we tested both progenies homozygous for the haplotype carrying translocation and homozygous for the haplotype as found in the reference. In Figure 29, in lanes from 1 to 4, primers pairs amplified products as explained for the previous figure. In samples 1, 2, and 4, amplification products only in lane 3 and 4 confirmed that they were homozygous for translocation. In sample 3, amplification products only in lane 1 and 2 confirmed it was homozygous for normal chromosomes.

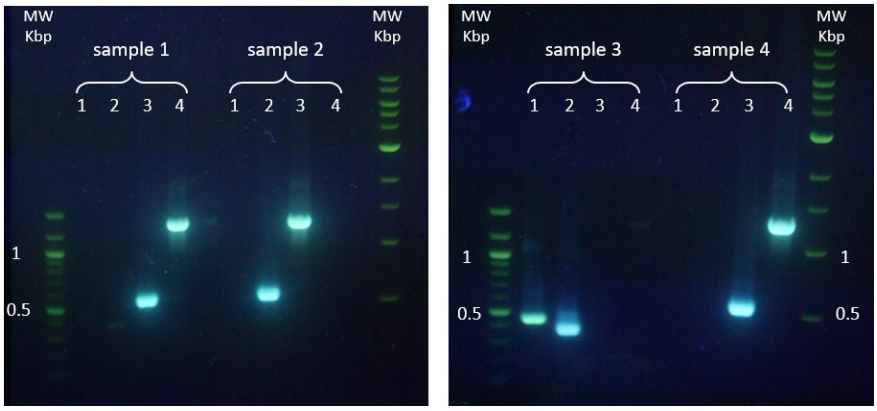


Figure 29. **Four homozygous progenies of Rkatsiteli selfing were tested for translocation.** Samples 1, 2, and 4 were confirmed as homozygotes for the translocation; sample 3 was confirmed as homozygote for normal chromosomes.

196 *Vitis vinifera* varieties and one *Vitis labrusca* species (Appendix 2) were tested for the translocation. Translocation breakpoint 2 was found in heterozygosis in Alexandroouli, Gorula and Mtsvane Kachuri varieties.

All the four grapevine varieties carrying the heterozygous translocation are of Georgian origin, belonging to the *Proles pontica.* According to the classification proposed by Negrul in 1946, grapevine varieties are classified into three main eco-geographical groups - *Proles occidentalis*, *orientalis*, and *pontica* – showing different characteristic, such as in the size of berries and in the resistance to cold temperatures. Finding the evidence that none, among tested varieties belonging to the *Proles occidentalis* and *orientalis*, showed the translocation, suggests this event may be ancient and never brought to the West during grapevine domestication and migration history. Therefore, the evidence is in agreement with the fact that Georgian varieties are genetically distinct and present a different variability compared to varieties belonging to the other two *Proles*.

Translocation breakpoint 1 was tested using different primer combinations and by performing nested-PCR in order to increase primer specificity. Highly repetitive nature of this region and primers' off-target sites made difficult to amplify univocally this region. Further implementation

of PCR strategy and/or primer design will be necessary to amplify insert corresponding to translocation breakpoint 1.

4.3.5 Position effects on gene expression (allele specific expression analysis)

The chromosome repositioning resulting from a translocation may have an effect on the expression of genes that were moved. Indeed, chromosomal rearrangements can interrupt genes located near the breakpoint, or can re-locate genes in a new chromatin environment with effects on their expression, as well as on their epigenetic landscape. This effect is called position-effect variegation (PEV) and the most famous example was found in the *white* locus of *Drosophila* by Muller, in 1930. Through his X-ray experiments and his observations of the polytene chromosome of *Drosophila*, Muller noticed that several inversions and rearrangements placed the euchromatin-located *white* locus in the pericentromeric heterochromatin. The consequence of this rearrangements was an altered packing of genes normally packaged in euchromatic form, suggesting a heterochromatic spreading from the adjacent constitutive heterochromatic region. The spreading of compact packed status of heterochromatin in the adjacent euchromatin caused the transcriptional silencing of the *white* gene in some of the cells where it was normally active (resulting in variegation). Modern studies revealed that spreading of heterochromatin depends on histone methylation and on the activity of enzymes involved in structural protein methylation establishment and maintaining (Elgin & Reuter, 2013).

The re-localization of chromosomal portions in new genomic and chromatin contexts may have effects on gene expression levels. In order to investigate this hypothesis for the balanced translocation, we performed ASE analysis in Rkatsiteli.

The $\log_2$ ratio of the number of reads belonging to the normal haplotype (hapN) to the number of reads belonging to the haplotype carrying translocation (hapT) gave the measure of ASE in the translocated regions. The $\log_2$ ratio of the number of observed reads for each haplotype in the genome represented the overall ASE level in the genome. In Table 23, the observed count of the $\log_2$ ratio for the two haplotypes, in the genome and in each translocated portion, is reported for

each tissue dataset. The expected values for the translocated regions were calculated on the base of observed values in the genome. Chi square was used to test for significant deviation of observed ASE in the translocated regions relative to the expected value, given the ASE level in the genome (p value < 0.05, Table 23).

A slightly significant ASE was found in the region of chromosome 11, in the tendril tissue. In particular, analysis showed that genes located on the haplotype carrying translocation were less expressed than genes located on the normal haplotype.

However, comparisons in the other tissues gave no significant ASE differences. We concluded that no significant difference in the allele-specific expression of genes in translocated portions was detected consequently to the re-positioning of genes in the new genomic context.

leaf	**Obs. ASE hapN**	**Obs. ASE hapT**	**tot genes**	**Exp. ASE hapN**	**Exp. ASE hapT**	**Chi square**	**p value**
whole genome*	7488	8226	15840				
chr1 translocated	73	92	169	80.532	88.468	0.845	0.358
chr11 translocated	61	68	131	62.424	68.576	0.037	0.847
berry	**Obs. ASE hapN**	**Obs. ASE hapT**	**tot genes**	**Exp. ASE hapN**	**Exp. ASE hapT**	**Chi square**	**p value**
whole genome*	6555	6982	13629				
chr1 translocated	72	83	160	77.477	82.523	0.390	0.532
chr11 translocated	65	49	114	55.202	58.798	3.372	0.066
tendril	**Obs. ASE hapN**	**Obs. ASE hapT**	**tot genes**	**Exp. ASE hapN**	**Exp. ASE hapT**	**Chi square**	**p value**
whole genome*	7490	8250	15838				
chr1 translocated	84	88	176	83.751	92.249	0.196	0.658
chr11 translocated	74	57	131	62.337	68.663	4.163	0.041

* $\log_2$ ratio is measured along the genome, excluding the regions interested by the translocation.

Table 23. Comparison between ASE in the whole genome and ASE in regions interested by the translocation, in three tissues (leaf, berry, tendril). In the translocated regions, ASE was measured as the $\log_2$ ratio of the number of reads of the normal haplotype (hapN) to the number of reads of the haplotype carrying translocation (hapT). Significant deviation of ASE value in the translocated regions relative to expected levels was tested using Chi square test ($\alpha = 0.05$).

Absence of significant ASE subsequently to a chromosomal rearrangement was also found by Fransz et colleagues (2016), who recently characterized a paracentric inversion in *Arabidopsis thaliana*. The inversion, caused by the activity of a *Vandal* transposon element, is located on chromosome 4 (Fransz et al., 2016). This rearrangement split an F-box and relocated a pericentric

heterochromatic region close to a euchromatic domain. To investigate whether the newly created boundaries at the breakpoints changed the characteristics of the flanking euchromatin, they examined expression and chromatin profiles in the *Col-0* and the *Ler* accessions, respectively with and without inversion. They could not find major differences in the epigenetic profiles in the two accessions, declining the hypothesis of a shift in the heterochromatin-euchromatin transition along the chromosome. Furthermore, they could not find any difference in the gene expression profiles between the two accessions, suggesting that the close proximity of heterochromatin to the genes in euchromatin region did not affect their transcriptional activity.

Chapter 5 Conclusions

Vitis vinifera species is characterized by a highly heterozygous genome (Jaillon et al., 2007) and present cultivated varieties show high levels of genetic diversity (Cattonaro et al., 2013). *Vitis vinifera* has a long history of domestication. Viticulture started during the Neolithic Age, when human populations began collecting and selecting mainly hermaphroditic grapevines amongst the wild dioecious progenitor species *Vitis sylvestris* (Cipriani G at al., 2010). Since then, propagation has been performed through vegetative propagation or by crosses (Myles et al., 2011). *Vitis vinifera* species is self-compatible, but the rate of seed germination and the survival of young seedlings are reduced in progenies deriving from the self-fertilization of cultivars. As a result of preferential clonal propagation and tendency of the progenies of selfing to inbreeding depression, grapevine has maintained high levels of genetic diversity during domestication (Fournier-Level, Lacombe, Le Cunff, Boursiquot, & This, 2010; Barnaud, Lacombe, & Doligez, 2006). The high heterozygosity of the genome underlies a genetic load of lethal and deleterious alleles, which is manifested in homozygosis. Progenies deriving from the self-fertilization of parental varieties show severe inbreeding depression by means of a rapid drop of germination rate, survival, vigour and fertility (Cattonaro et al., 2013).

Different mechanisms have been reported to be responsible for the genetic load of a species: selfing within populations is responsible for faster exposure of lethal mutations to selection than in an outcrossing population (Glémin et al., 2003). Moreover, deleterious mutations can rise in frequency as a consequence of linked selection on beneficial mutations, as in the domestication process (Chun et al., 2011; Mezmouk & Ross-Ibarra, 2014). Lastly, selection against deleterious mutations is less effective in genomic regions with low levels of recombination, where these alleles tend to accumulate.

The aim of the present work was to gain knowledge on the genetic load characterizing the *Vitis vinifera* genome. Genetic load was assessed through the analysis of segregation distortion in progenies of selfing and of outcrossing. Segregation distortion in the progenies of selfing resulted from the increased homozygosity of damaging alleles, harboured in heterozygosis in the parental variety. The lethal or deleterious effect of the causative allele at the homozygous state in the progenies was unmasked, causing a decrease in seedling viability and survival. The mapping of putative candidate loci and the description of mutations affecting gene function in loci of SD gave a first insight into the genetic load borne by the grapevine genome. Knowledge on the lethal effect of mutations in genes can be exploited for more efficient and precise breeding strategies and for crop improvement.

5.1 Segregation distortion: seedlings genotyping and phenotyping

We analysed both loci acting individually (single-locus SD) and pairs of loci acting through epistatic interaction (two-loci SD). The analysis was carried out in progenies deriving from the self-fertilization of six varieties of *Vitis vinifera*: Cabernet Franc, Pinot Noir, Primitivo, Rkatsiteli, Sangiovese, and Schiava Grossa. We also analysed one progenies deriving from the out-cross between Schiava Grossa and Rkatsiteli. Reduced-representation libraries were constructed to obtain a dense panel of genotypic data in the progenies of crosses. This technique allowed to sample homologous fractions of genome across all individuals, in order to maximize the number of loci genotyped in each progenies. On the overall, Chi square test for deviation of observed genotypic frequencies from expected Mendelian ratio revealed 12 regions of single-locus segregation distortion (SD). SD was found in the six progenies of selfing, while the progenies of the out-cross did not show any evidence for distortion. This suggested that level of inbreeding in a cross between two heterozygous grapevine varieties was not deep enough to rise segregation distortion. Interestingly, progenies of three parent varieties, Cabernet Franc, Rkatsiteli and Sangiovese, showed segregation distortion in close but non-overlapping loci on chromosome 8. This evidences that chromosome 8 is frequently involved in segregation distortion, suggesting that lethal alleles in genes required for viability/survival may reside on this chromosome.

Alongside genotyping, progenies of crosses showing segregation distortion were also phenotyped. Phenotyping was carried out over two vegetative seasons: initially, scoring was performed on survived individuals at germination and at the beginning of the second vegetative season. Since we noticed there were two major critical steps during seedling development, i.e. soon after germination and during dormancy, following phenotyping was carried out by considering also seedlings vigour. An additional phenotype scoring timing at the end of the first vegetative season allowed to compare seedlings vigour before and after dormancy. Moreover, we incremented phenotyping by observing stem growth and bud sprouting as indicators of vigour during and after winter, respectively. Soon after the beginning of germination, at the stage of epicotyl emergence, Sangiovese progenies showed a higher degree of lethality compared to the others. The higher lethality in the progenies of selfing in this variety came along with the detection of four loci of single-locus distortion, while the other progenies showed one (progenies of Cabernet franc, Pinot Noir, and Schiava Grossa), two (progenies of Rkatsiteli), or three (progenies of Primitivo) loci of SD. Rkatsiteli progenies was the only one showing late lethality, arising during the winter season. While progenies genotyped at germination and at the end of the first vegetative year did not show any significant deviation of genotype frequencies, segregation distortion at loci on chromosome 8 and 18 were shown for the progeny genotyped at the resumption of the vegetative growth.

5.2 Genetic load and measure of fitness in loci of SD

The burden of deleterious alleles carried by a population was the subject of study of pioneering research work in population genetics and was termed mutation load (Kimura, Maruyama, & Crow, 1963). Mutation load is the component of genetic load that is attributable to the reduction in fitness caused by deleterious mutations. In the middle Fifties, Newton E. Morton proposed the theory that deleterious mutations reduce the fitness of the carrier individual with respect to a genotype with ideal fitness. He proposed that measure of genetic load is given through the coefficient of selection against the deleterious allele and its degree of dominance, and the measure of fitness is calculated with respect to the fittest genotype (Henn et al., 2015). Classical

studies were carried out on children of consanguineous marriages and the genetic load was expressed as mortality, arising when a gamete carrying recessive mutation was doubled to produce a complete homozygote (Morton, 1960; Morton et al., 1956).

In order to characterize the genetic load carried by the segregating populations under study, we estimated the fitness of genotypes at loci of SD.

The effect of dominance on fitness is quantified by the coefficient *h*. *h* = 0 for recessive alleles, *h* = 0.5 for additive alleles, and between 0.5 to 1 for all levels of partial dominant alleles, while it is exactly 1 for dominant alleles. The mutation load is higher under an additive/partial dominance model rather than a recessive model. Among the twelve loci of SD, two of them were found to segregate as partially dominant (i.e. locus on chromosome 8 in Cabernet franc, and locus on chromosome 11 in Primitivo). Eight loci segregated as recessive, while one locus segregated as overdominant. Interestingly, the three non-overlapping loci on chromosome 8 were found to segregate differently. Indeed, in Rkatsiteli and in Sangiovese progenies, respective loci of chromosome 8 were found to act as recessive loci. Further examination of lethal and deleterious alleles in the candidate loci revealed that different mutations were present in Rkatsiteli and Sangiovese. This finding suggested that different causal mutations may be at the base of distortion in close loci. Furthermore, locus on chromosome 12 in Primitivo was found to be nearly dominant, with a penetrance of about 95%. This finding is really interesting: to explain the presence of a nearly dominant mutation in this variety, one would hypothesize that Primitivo was generated from a very rare non-mutant gamete. If the effect on lethality is at gametophytic level, this hypothesis could be assessed by analysing pollen viability and by searching low-frequency lethal alleles in pollen pools.

5.3 Mutations as candidates for inbreeding depression in grapevine

The Darwinist theory stated that natural selection is the driving force of evolutionary change. New mutations with advantageous effects are positively selected and fixed in the population. Selectively neutral mutations are very rare and genetic drift plays no or very little role in evolution. With his "Neutral Theory of Molecular Evolution", the geneticists Motoo Kimura revolutionized the standpoint on evolution. Neutral theory claims that the great majority of evolutionary changes at the molecular level are not due to positive selection on advantageous mutations, but rather are due to random fixation, through genetic drift, of selectively neutral variants under mutation pressure. Thus, neutral or nearly-neutral mutations represent the great majority of variation, and they are maintained by the balance between mutation pressure and genetic drift. However, this theory does not refute the neo-Darwinian one: simply, it states that advantageous mutations fixed by natural selection are extremely rare. Synonymous amino acid substitutions or other substitution in silent sites (e.g. in introns) occur at higher rates than non-synonymous mutations, since they are more likely to be non-deleterious and less likely to be subjected to natural selection (Kimura, 1991).

However, deleterious mutations can rise in frequency in some circumstances, e.g. during artificial selection as a result of hitchhiking effect, as proposed in rice by Lu and colleagues (J. Lu et al., 2006); or in low-recombining regions, as showed by recent research in sunflower (Renaut & Rieseberg, 2015). In a recent population genetics study, Zhang and colleagues used exome capture data to estimate the genome-wide distribution of deleterious alleles in natural populations of black cottonwood (*Populus trichocarpa*) (M. Zhang et al., 2016). Looking at the whole genome, the recombination rate did not show a significant correlation with the proportion of deleterious SNPs, but a significant correlation was found in proximity to putative centromeres. When they looked at the absolute number of deleterious SNPs, they found it was higher in regions of reduced recombination. They found that deleterious alleles were in general present at low frequency, suggesting purifying selection; however, alleles were preferentially enriched both within genomic regions of low-recombination and in regions showing evidence of positive selection. They suggested that demographic history, selection efficiency in small-size populations

and bottlenecks could have contributed to the observed rate of deleterious alleles among populations. Through their results, they concluded that both genomic context and historical demography played a role in shaping the distribution of deleterious alleles in *P. trichocarpa*.

Previous and present work on the analysis of genetic variation in grapevine showed that there is an excess of deleterious mutations (e.g. stopgain and nonsynonymous SNPs with deleterious effect) segregating at lower frequency with respect to either nonsynonymous SNPs with tolerated effect or synonymous SNPs. These findings are consistent with reports in *Arabidopsis*, maize and rice (Günther & Schmid, 2010; Mezmouk & Ross-Ibarra, 2014), and the pattern suggested the action of purifying selection against deleterious alleles. Accordingly, significant distribution of values of Tajima's D were found for mutations highly affecting gene function (such as stopgain SNPs, deleterious SNPs, deletions and insertions in exonic regions) relative to that of neutral mutations (such as synonymous SNPs), suggesting a signature of negative selection.

We also investigated the relation between the candidate loci of SD, found in the progenies of selfing, and the recombination frequency in the regions where they are located. We observed that six loci were located in highly recombining regions, while the remaining six loci were distributed in low- to intermediate- recombining regions. These findings suggested that lethal/deleterious alleles at the base of segregation distortion observed in the progenies of selfing could be present both in low- and in high-recombining regions in the grapevine genome. It is likely that deleterious alleles in loci of SD in high recombining regions will be removed faster by selection than deleterious alleles in low-recombining regions, as reported in a population-based study in poplar (M. Zhang et al., 2016).

In order to identify genes and variants, whether SVs or SNPs, contributing to genetic load and leading to segregation distortion in the progenies of selfing, loci were screened for mutations affecting gene function. We provided a list of 57 candidate genes, among which 33 located on the haplotype carrying distortion based on the high-density haplotype phasing analysis (for the remaining 24 haplotype phase could not be assigned). A preliminary analysis on gene sequence similarity suggested that 36 genes were single-copy and 21 genes were duplicated. We then

looked at the expression of duplicated copies of the candidate gene in the variety carrying the putative lethal allele. Among the 21 putatively duplicated genes, 14 were found to have at least one expressed copy, suggesting a possible functional redundancy of these genes. This analysis is very preliminary and further investigation will be needed to detect the functional redundancy of candidate genes. However, these findings suggest that duplicated expressed genes may be excluded from the list of candidates.

The analysis of genetic variation in a population of 128 grapevine cultivars has revealed that their genomes are characterized by high levels of both SNP and structural variation. Through the analysis of segregation distortion carried out in the present work, we gave further insight on the genetic load carried by six grapevine varieties and assessed in the progenies of selfing. Through the present work, we showed that each variety under study presented different and non-overlapping loci of segregation distortion, we showed the modality of segregation of SD loci and we provided a list of genes and mutations putatively involved. The tracing of grapevine domestication process, an insight of its demographic history (the recent bottleneck and the subsequent population expansion) and the identification of signatures of selection focusing deeply on structural variants represent further interesting topics of investigation to elucidate the genetic load of the *Vitis vinifera* genome.

5.4 Identification and characterization of reciprocal translocation in Rkatsiteli

In order to identify two-loci epistatic interactions, segregation distortion was assessed in pairs of independent loci. Fisher's test could reveal one strong signal of interaction between regions of chromosomes 1 and 11 in Rkatsiteli. A pseudo-linkage signal and significant deviation of observed genotypic frequencies by means of Pearson's Chi square test suggested a physical interaction between the chromosomes. A structural analysis on read alignment and a PCR-based assay validated a balanced translocation in Rkatsiteli. Nearly 200 *Vitis vinifera* varieties across the three grapevine *Proles* were assayed through the PCR assay. Three varieties were found to carry the reciprocal translocation - i.e. Alexandroouli, Mtsvane Kachuri and Gorula - all of them originating from Georgia. The fact that none Western or even Eastern varieties showed the translocation suggested that this rearrangement may be very ancient and that Georgian varieties remained isolated from the more western ones.

Chromosomal rearrangements can impact on the spatial organization of the resulting chromosomes within the nucleus (Croft et al., 1999). Furthermore, gene regulation is dependent not only on regulatory sequences and chromatin context located in *cis*, but also on the specific three-dimensional position in the nucleus of chromosomal compartment that are rich or depleted in genes (Meaburn, Misteli, & Soutoglou, 2007). As a consequence, alteration in local chromatin structure and in the nuclear organisation as a consequence of translocation can result in position effect on gene expression.

To further characterize the translocation, we assessed whether a position effect on gene expression was present as a consequence of chromosomal repositioning in a new location. Thus, we performed allele specific expression analysis (ASE) on the reconstructed high-density haplotypes of Rkatsiteli. Analysis did not reveal any strong difference in the ASE of genes lying on the translocated portions of chromosomes, with respect to the mean ASE along genome.

Figure 30 illustrates a reconstruction of some genomic features along the translocated chromosome. Gene density (CDS), repeat density, CG and CHG methylation profiles, and

conformation of chromatin domains were compared for each of the two translocation breakpoints. Chromatin conformation analysis on HiC data on Pinot Noir variety was previously obtained in the NOVABREED projects (Schwope R, Tocci A, unpublished results). In Figure 30, positive signed values indicated relaxed chromatin status (loose structural domain – LSD, pink), while negative signed values indicated condensed chromatin status (CSD, green). The distribution of the genomic features on reconstructed breakpoints suggested that each translocated portion was re-located on a chromosomal context of destination different for the chromatin conformation status. This is particularly evident for the translocation breakpoint 2.

Further analysis will be made in Rkatsiteli in order to deeply explore allele specific characteristics. For example, changes in the methylation profile affecting allele-specific expression (methyl-ASE) will be an interesting topic of investigation. Correlating structural domains and TEs would allow a better comprehension of the effect of TE movement on gene expression at short and long range. While large-scale differences on chromosome localization inside the nucleus is expected to be highly conserved in varieties of the same species, some differences may arise in the local positioning of chromosomal domains consequently to variation within haplotypes inside genome. For example, the insertion of a transposable elements may be responsible for micro re-adjustment of local chromatin conformation. The analysis of allele-specific structural variation composition of Rkatsiteli could help in understanding the effects of chromatin conformation status due to TE transposition.

In plants, reciprocal translocations have been found in many varieties, as in *Arabidopsis thaliana* (Lysak et al., 2006), in *Oryza sativa L.* (Wang G, Li H, Cheng Z, 2013), and in legume crops as *Vigna angularis* (the azuki bean; Wang L, Isemura T et al., 2015), and *Lathyrus sativus L.* (the grass pea; Talukdar D, 2010). Inversions were found in *Arabidospsis* (Fransz et al., 2016), in *Mimulus guttatus* (Fishman & Willis, 2005), in *Aegilops tauschii* (Z. Zhang, Zhu, Gill, & Li, 2015), among many examples. However, studies on the position effect of reciprocal translocations, and more in general, of balanced rearrangements, have shown significant differences on gene expression especially in mammals, humans included (Harewood et al., 2010).

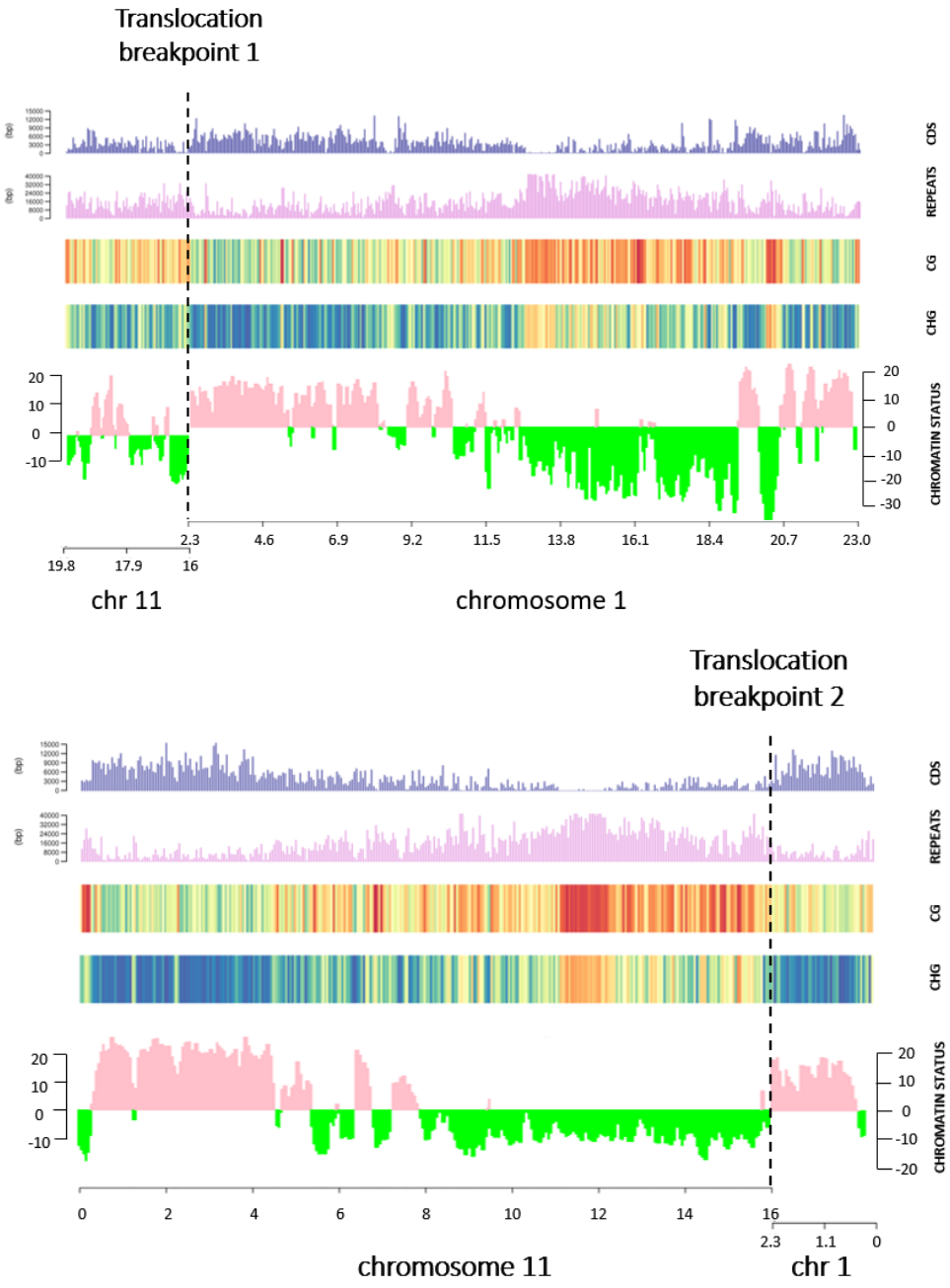


Figure 30. **Reconstruction of genomic features** in the chromosome of translocation breakpoint 1 (top panel) and of translocation breakpoint 2 (bottom panel): gene density (CDS, coding sequence), transposable element density (REPEATS), CG and CHG methylation profiles, and chromatin structure status.

Our study showed that self-fertilization of *Vitis vinifera* cultivars leads to high levels of segregation distortion in the progenies. Segregation distortion in the progenies results from the appearance in homozygosis of unfavourable recessive alleles causing decreased viability and survival of seedlings. Mutations such as structural variants in genes, nonsense and missense SNPs and frameshift INDELs can have strong decreasing effect on fitness, when lethal/deleterious alleles affect genes with crucial function. We provided a list of lethal/deleterious alleles in genes representing putative candidates for the segregation distortion observed in progenies of selfing. Both genes with essential function in the cell (such as genes coding for RNA polymerase subunits) and genes with redundant function (such as genes coding for transferases) were affected by lethal or highly deleterious mutations. Some of these genes represent good candidates for follow-up studies on the regions showing segregation distortion. Perspective of this work is oriented to the further restriction of the number of candidates and to their characterization. The final goal is to provide knowledge on loci controlling viability and survival, which can be used to improve the efficiency and precision of breeding strategies.

The construction of genetic linkage maps of *Vitis vinifera* varieties allowed to generate a fine-scale map of the recombination frequency. Knowledge on the variation of recombination frequency can be exploited by plant breeders, who rely on meiotic cross-overs to fine-map quantitative traits and introgress favourable alleles.

In the present research work we also provided evidences of a balanced translocation detected in the Georgian variety Rkatsiteli. This rearrangement was found also in other three varieties - Alexandroouli, Mtsvane Kachuri and Gorula. We gave a first insight into the characterization of the translocation through ASE analysis. Future experiments will be focused on haplotype differences in the chromatin conformation and in the epigenetic modulation.

References

Abajian, C. (1994). Sputnik.

Alheit, K. V, Reif, J. C., Maurer, H. P., Hahn, V., Weissmann, E. A., Miedaner, T., & Würschum, T. (2011). Detection of segregation distortion loci in triticale (x Triticosecale Wittmack) based on a high-density DArT marker consensus genetic linkage map. *BMC Genomics*, *12*(1), 380. https://doi.org/10.1186/1471-2164-12-380

Anderson, L. K., Doyle, G. G., Brigham, B., Carter, J., Hooker, K. D., Lai, A., … Stack, S. M. (2003). High-Resolution Crossover Maps for Each Bivalent of Zea mays Using Recombination Nodules. *Genetics*, *165*(2).

Arroyo-Garcia, R., Ruiz-Garcia, L., Bolling, L., Ocete, R., Lopez, M. A., Arnold, C., … Martinez-Zapater, J. M. (2006). Multiple origins of cultivated grapevine (Vitis vinifera L. ssp. sativa) based on chloroplast DNA polymorphisms. *Molecular Ecology*, *15*(12), 3707–3714. https://doi.org/10.1111/j.1365-294X.2006.03049.x

Atkinson, P. W. (2015). hAT Transposable Elements. *Microbiology Spectrum*, *3*(4). https://doi.org/10.1128/microbiolspec.MDNA3-0054-2014

Baidouri, M. El, & Panaud, O. (2013). Comparative genomic paleontology across plant kingdom reveals the dynamics of TE-driven genome evolution. *Genome Biology and Evolution*. https://doi.org/10.1093/gbe/evt025

Barnaud, A., Lacombe, T., & Doligez, A. (2006). Linkage disequilibrium in cultivated grapevine, Vitis vinifera L. *Theoretical and Applied Genetics*, *112*(4), 708–716. https://doi.org/10.1007/s00122-005-0174-1

Beadle, G. W., & Ker, W. G. (1931). A possible influence of the spindle fibre on crossing-over in Drosophila. *Proc. Natl. Acad. Sci. USA*, *18*, 160–165.

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

Bersabé, D., Caballero, A., Pérez-Figueroa, A., & García-Dorado, A. (2015). On the Consequences of Purging and Linkage on Fitness and Genetic Diversity. *G3 (Bethesda, Md.)*, *6*(1), 171–81. https://doi.org/10.1534/g3.115.023184

Bierhoff, H., Postepska-Igielska, A., & Grummt, I. (2014). Noisy silence - Non-coding RNA and heterochromatin formation at repetitive elements. *Epigenetics*, *9*(1), 53–61.

https://doi.org/10.4161/epi.26485

Billings, T., Sargent, E. E., Szatkiewicz, J. P., Leahy, N., Kwak, I.-Y., Bektassova, N., … Ukkonen, E. (2010). Patterns of Recombination Activity on Mouse Chromosome 11 Revealed by High Resolution Mapping. *PLoS ONE*, *5*(12), e15340. https://doi.org/10.1371/journal.pone.0015340

Bodenes, C., Chancerel, E., Ehrenmann, F., Kremer, A., & Plomion, C. (2016). High-density linkage mapping and distribution of segregation distortion regions in the oak genome. *DNA Research*, *23*(2), 115–124. https://doi.org/10.1093/dnares/dsw001

Bonferroni, C. E. (1936). *Teoria statistica delle classi e calcolo delle probabilità.* Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze.

Branca, A., Paape, T. D., Zhou, P., Briskine, R., Farmer, A. D., Mudge, J., … Tiffin, P. (2011). Whole-genome nucleotide diversity, recombination, and linkage disequilibrium in the model legume Medicago truncatula. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(42), E864-70. https://doi.org/10.1073/pnas.1104032108

Broman, K. W. (2010). Genetic map construction with R/qtl.

Bronner, A., & Oliveira, J. (1990). Creation and study of the Pinot noir variety lineage. *Vitis Special Issue*, 69–80.

Browning, S. R., & Browning, B. L. (2007). Rapid and Accurate Haplotype Phasing and Missing-Data Inference for Whole-Genome Association Studies By Use of Localized Haplotype Clustering. *The American Journal of Human Genetics Am. J. Hum. Genet*, *8181*, 1084–1097. https://doi.org/10.1086/521987

Brunner, S., Fengler, K., Morgante, M., Tingey, S., & Rafalski, A. (2005). Evolution of DNA sequence nonhomologies among maize inbreds. *The Plant Cell*, *17*(2), 343–60. https://doi.org/10.1105/tpc.104.025627

Buckler IV, E. S., Phelps-Durr, T. L., Buckler, C. S. K., Dawe, R. K., Doebley, J. F., & Holtsford, T. P. (1999). Meiotic drive of chromosomal knobs reshaped the maize genome. *Genetics*.

Burnham, C. R. (1930). Genetical and cytological studies of semisterility and related phenomena in Maize. *Genetics*, *16*, 269–277.

Carr, D. E., & Dudash, M. R. (2003). Recent approaches into the genetic basis of inbreeding depression in plants. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *358*(1434), 1071–84. https://doi.org/10.1098/rstb.2003.1295

Catchen, J., Hohenlohe, P. A., Bassham, S., Amores, A., & Cresko, W. A. (2013). Stacks: An analysis tool set for population genomics. *Molecular Ecology*, *22*(11), 3124–3140. https://doi.org/10.1111/mec.12354

Cattonaro, F., Testolin, R., Scalabrin, S., Morgante, M., & Di Gaspero, G. (2013). Heterozygosity. In *Genetic Diversity in the Grapevine Germplasm*.

Celii, M. (2016). *Analysis of Epigenomic Variability in Grapevine and its Relation with Structural Variation*.

Charlesworth, B. (2007). Mutation-selection balance and the evolutionary advantage of sex and recombination. *Genetical Research*, *89*(5–6), 451. https://doi.org/10.1017/S0016672308009658

Charlesworth, D., & Willis, J. H. (2009). The genetics of inbreeding depression. *Nature Reviews. Genetics*, *10*(11), 783–96. https://doi.org/10.1038/nrg2664

Chen, S. Y., Tsubouchi, T., Rockmill, B., Sandler, J. S., Richards, D. R. D. R., Vader, G., … Kleckner, N. (2008). Global analysis of the meiotic crossover landscape. *Developmental Cell*, *15*(3), 401–15. https://doi.org/10.1016/j.devcel.2008.07.006

Cheptou, P.-O., & Donohue, K. (2011). Environment-dependent inbreeding depression: its ecological and evolutionary significance. *The New Phytologist*, *189*(2), 395–407. https://doi.org/10.1111/j.1469-8137.2010.03541.x

Chia, J. M., Song, C., Bradbury, P. J., Costich, D., de Leon, N., Doebley, J., … Ware, D. (2012). Maize HapMap2 identifies extant variation from a genome in flux. *Nature Genetics*, *44*(7), 803–807. https://doi.org/10.1038/ng.2313

Choo, K. H. (1998). Why is the centromere so cold? *Genome Research*, *8*(2), 81–2. https://doi.org/10.1101/GR.8.2.81

Chun, S., Fay, J. C., & Pritchard, J. K. (2011). Evidence for Hitchhiking of Deleterious Mutations within the Human Genome. *PLoS Genet*, *7*(8).

Cipriani, G., Spadotto, A., Jurman, I., Gaspero, G. Di, Crespan, M., Meneghetti, S., … Testolin, R. (2010). The SSR-based molecular profile of 1005 grapevine (Vitis vinifera L.) accessions uncovers new synonymy and parentages, and reveals a large admixture amongst varieties of different geographic origin. *Theoretical and Applied Genetics*, *121*(8), 1569–1585. https://doi.org/10.1007/s00122-010-1411-9

Conrad, D. F., Keebler, J. E. M., DePristo, M. A., Lindsay, S. J., Zhang, Y., Casals, F., … Awadalla, P. (2011). Variation in genome-wide mutation rates within and between human families. *Nature Genetics*, *43*(7), 712–4. https://doi.org/10.1038/ng.862

Conrad, D. F., Pinto, D., Redon, R., Feuk, L., Gokcumen, O., Zhang, Y., … Hurles, M. E. (2010). Supp 1: Origins and functional impact of copy number variation in the human genome. *Nature*, *464*(7289), 704–12. https://doi.org/10.1038/nature08516

Cordell, H. J. (2002). Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans. *Human Molecular Genetics*, *11*(20), 2463–8.

https://doi.org/10.1093/hmg/11.20.2463

Croft, J. A., Bridger, J. M., Boyle, S., Perry, P., Teague, P., & Bickmore, W. A. (1999). Differences in the Localization and Morphology of Chromosomes in the Human Nucleus. *The Journal of Cell Biology*, *145*(6), 1119–1131. Retrieved from http://www.jcb.org

Del Fabbro, C., Scalabrin, S., Morgante, M., & Giorgi, F. M. (2013). An extensive evaluation of read trimming effects on Illumina NGS data analysis. *PloS One*, *8*(12), e85024. https://doi.org/10.1371/journal.pone.0085024

DePristo, M. A., Banks, E., Poplin, R., Garimella, K. V, Maguire, J. R., Hartl, C., … Daly, M. J. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics*, *43*(5), 491–8. https://doi.org/10.1038/ng.806

Díaz, A., Zikhali, M., Turner, A. S., Isaac, P., & Laurie, D. A. (2012). Copy number variation affecting the photoperiod-B1 and vernalization-A1 genes is associated with altered flowering time in wheat (Triticum aestivum). *PLoS ONE*. https://doi.org/10.1371/journal.pone.0033234

Dooner, H. K., & He, L. (2008). Maize Genome Structure Variation: Interplay between Retrotransposon Polymorphisms and Genic Recombination. *The Plant Cell*, *20*(249:258). https://doi.org/10.1105/tpc.107.057596

Eckardt NA. (2008). Retrotransposon Polymorphisms Affect Genic Recombination in Maize. *The Plant Cell*, *20*(247). https://doi.org/10.1105/tpc.108.200213

Eichten, S. R., Swanson-Wagner, R. A., Schnable, J. C., Waters, A. J., Hermanson, P. J., Liu, S., … Raftery, A. (2011). Heritable Epigenetic Variation among Maize Inbreds. *PLoS Genetics*, *7*(11), e1002372. https://doi.org/10.1371/journal.pgen.1002372

Elgin, S. C. R., & Reuter, G. (2013). Position-effect variegation, heterochromatin formation, and gene silencing in Drosophila. *Cold Spring Harbor Perspectives in Biology*. https://doi.org/10.1101/cshperspect.a017780

Ellermeier, C., Higuchi, E. C., Phadnis, N., Holm, L., Geelhood, J. L., Thon, G., & Smith, G. R. (2010). RNAi and heterochromatin repress centromeric meiotic recombination. *Proceedings of the National Academy of Sciences of the United States of America*, *107*(19), 8701–5. https://doi.org/10.1073/pnas.0914160107

Elshire, R. J., Glaubitz, J. C., Sun, Q., Poland, J. A., Kawamoto, K., Buckler, E. S., & Mitchell, S. E. (2011). A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS ONE*, *6*(5), 1–10. https://doi.org/10.1371/journal.pone.0019379

Fans, J. D., Laddomada, B., & Gill, B. S. (1998). Molecular mapping of segregation distortion loci in Aegilops tauschii. *Genetics*, *149*(1), 319–327.

Fay, J. C., & Wu, C. I. (2000). Hitchhiking under positive Darwinian selection. *Genetics*, *155*(3),

1405–1413.

Fishman, L., & Willis, J. H. (2005). A novel meiotic drive locus almost completely distorts segregation in Mimulus (monkeyflower) hybrids. *Genetics*, *169*(1), 347–353. https://doi.org/10.1534/genetics.104.032789

Fournier-Level, a, Lacombe, T., Le Cunff, L., Boursiquot, J.-M., & This, P. (2010). Evolution of the VvMybA gene family, the major determinant of berry colour in cultivated grapevine (Vitis vinifera L.). *Heredity*, *104*(4), 351–362. https://doi.org/10.1038/hdy.2009.148

Franklin-Tong, N. V. E., & Franklin, F. C. H. (2003). Gametophytic self-incompatibility inhibits pollen tube growth using different mechanisms. *Trends in Plant Science*, *8*(12), 598–605. https://doi.org/10.1016/j.tplants.2003.10.008

Fransz, P., Linc, G., Lee, C.-R., Aflitos, S. A., Lasky, J. R., Toomajian, C., … Schranz, M. E. (2016). Molecular, genetic and evolutionary analysis of a paracentric inversion in Arabidopsis thaliana. *The Plant Journal*, doi: 10.1111/tpj.13262. https://doi.org/10.1111/tpj.13262

Fujii, M., Yokosho, K., Yamaji, N., Saisho, D., Yamane, M., Takahashi, H., … Ma, J. F. (2012). Acquisition of aluminium tolerance by modification of a single gene in barley. *Nature Communications*, *3*, 713. https://doi.org/10.1038/ncomms1726

Gaut, B. S., Díez, C. M., & Morrell, P. L. (2015). Genomics and the Contrasting Dynamics of Annual and Perennial Domestication. *Trends in Genetics*, *31*(12), 709–719. https://doi.org/10.1016/j.tig.2015.10.002

Girirajan, S., Johnson, R. L., Tassone, F., Balciuniene, J., Katiyar, N., Fox, K., … Selleck, S. B. (2013). Global increases in both common and rare copy number load associated with autism. *Human Molecular Genetics*, *22*(14), 2870–80. https://doi.org/10.1093/hmg/ddt136

Glémin, S., Ronfort, J., & Bataillon, T. (2003). Patterns of Inbreeding Depression and Architecture of the Load in Subdivided Populations. *Genetics*, *165*(4).

Grattapaglia, D., & Sederoff, R. (1994). Genetic Linkage Maps of Eucalyptus grandis and Eucalyptus urophylla Using a Pseudo-Testcross: Mapping Strategy and RAPD Markers Dario. *Genetics*, *137*(4), 1121–37. https://doi.org/10.1007/s11033-010-0612-2

Griffiths AJF, Miller JH, S. D. (2000). *An Introduction to Genetic Analysis. 7th edition.* New York: W. H. Freeman.

Günther, T., & Schmid, K. J. (2010). Deleterious amino acid polymorphisms in Arabidopsis thaliana and rice. *Theoretical and Applied Genetics*, *121*(1), 157–168. https://doi.org/10.1007/s00122-010-1299-4

Habibi, L., Pedram, M., AmirPhirozy, A., & Bonyadi, K. (2015). Mobile DNA Elements: The Seeds of Organic Complexity on Earth. *DNA and Cell Biology*, *34*(10), 597–609. https://doi.org/10.1089/dna.2015.2938

Harewood, L., Boyle, S., Perry, P., Delorenzi, M., Bickmore, W. A., & Reymond, A. (2010). The effect of translocation-induced nuclear reorganization on gene expression. *Genome Research*, *20*, 554–564. https://doi.org/10.1101/gr.103622.109.554

Henn, B. M., Botigué, L. R., Bustamante, C. D., Clark, A. G., & Gravel, S. (2015). Estimating the mutation load in human genomes. *Nature Publishing Group*, *16*(6), 1–11. https://doi.org/10.1038/nrg3931

Hickman, A. B., & Dyda, F. (2015). Mechanisms of DNA Transposition. *Microbiology Spectrum*, *3*(2), MDNA3-0034-2014. https://doi.org/10.1128/microbiolspec.MDNA3-0034-2014

Hurwitz, B. L., Kudrna, D., Yu, Y., Sebastian, A., Zuccolo, A., Jackson, S. A., … 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. https://doi.org/10.1111/j.1365-313X.2010.04293.x

Iafrate, A. J., Feuk, L., Rivera, M. N., Listewnik, M. L., Donahoe, P. K., Qi, Y., … Lee, C. (2004). Detection of large-scale variation in the human genome. *Nature Genetics*, *36*(9), 949–51. https://doi.org/10.1038/ng1416

Jaillon, O., Aury, J.-M., Noel, B., Policriti, A., Clepet, C., Casagrande, A., … Wincker, P. (2007). The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature*, *449*(7161), 463–7. https://doi.org/10.1038/nature06148

Kejnovsky, E., Leitch, I. J., & Leitch, A. R. (2009). Contrasting evolutionary dynamics between angiosperm and mammalian genomes. *Trends in Ecology & Evolution*, *24*(10), 572–582. https://doi.org/10.1016/j.tree.2009.04.010

Kidwell, M. G., & Lisch, D. (1997). Transposable elements as sources of variation in animals and plants. *Proc. Natl. Acad. Sci. USA*, *94*(15), 7704–7711. https://doi.org/10.1073/pnas.94.15.7704

Kim, D., & Salzberg, S. L. (2011). TopHat-Fusion: an algorithm for discovery of novel fusion transcripts. *Genome Biology*, *12*(8), 15. https://doi.org/10.1186/gb-2011-12-8-r72

Kimura, M. (1991). The Neutral Theory of Molecular Evolution: a review of recent evidence. *Jpn. J. Genet.*, *66*(367–386), 367–86. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/1954033

Kimura, M., Maruyama, T., & Crow, J. F. (1963). The mutation load in small populations. *Genetics*, *48*(10).

Kondrashov, A. S. (1988). Deleterious mutations and the evolution of sexual reproduction. *Nature*, *336*(6198), 435–440. https://doi.org/10.1038/336435a0

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods*, *9*(4), 357–9. https://doi.org/10.1038/nmeth.1923

Lenormand, T., & Dutheil, J. (n.d.). Recombination Difference between Sexes: A Role for Haploid Selection. https://doi.org/10.1371/journal.pbio.0030063

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

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics (Oxford, England)*, *25*(14), 1754–60. https://doi.org/10.1093/bioinformatics/btp324

Li, H., & Durbin, R. (2010). Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics (Oxford, England)*, *26*(5), 589–95. https://doi.org/10.1093/bioinformatics/btp698

Li, R., Ye, J., Li, S., Wang, J. J. J., Han, Y., Ye, C., … Eichler, E. (2005). ReAS: Recovery of Ancestral Sequences for Transposable Elements from the Unassembled Reads of a Whole Genome Shotgun. *PLoS Computational Biology*, *1*(4), e43. https://doi.org/10.1371/journal.pcbi.0010043

Limborg, M. T., McKinney, G. J., Seeb, L. W., & Seeb, J. E. (2016). Recombination patterns reveal information about centromere location on linkage maps. *Molecular Ecology Resources*, *16*(3), 655–661. https://doi.org/10.1111/1755-0998.12484

Lisch, D. (2012). How important are transposons for plant evolution? *Nat Rev Genet*, *14*(1), 49–61. https://doi.org/10.1038/nrg3374

Lu, F., Romay, M. C., Glaubitz, J. C., Bradbury, P. J., Elshire, R. J., Wang, T., … Buckler, E. S. (2015). High-resolution genetic mapping of maize pan-genome sequence anchors. *Nature Communications*, *6*, 6914. https://doi.org/10.1038/ncomms7914

Lu, J., Tang, T., Tang, H., Huang, J., Shi, S., & Wu, C. I. (2006). The accumulation of deleterious mutations in rice genomes: A hypothesis on the cost of domestication. *Trends in Genetics*, *22*(3), 126–131. https://doi.org/10.1016/j.tig.2006.01.004

Lupski, J. R., Reid, J. G., Gonzaga-Jauregui, C., Rio Deiros, D., Chen, D. C. Y., Nazareth, L., … Gibbs, R. a. (2010). Whole-genome sequencing in a patient with Charcot-Marie-Tooth neuropathy. *The New England Journal of Medicine*, *362*(13), 1181–1191. https://doi.org/10.1056/NEJMoa0908094

Lysak, M. A., Berr, A., Pecinka, A., Schmidt, R., Mcbreen, K., & Schubert, I. (2006). Mechanisms of chromosome number reduction in Arabidopsis thaliana and related Brassicaceae species. *PNAS*, *103*(13), 5224–5229.

Magris, G. (2016). *Characterisation of the pan-genome of Vitis vinifera using Next Generation Sequencing*. University of Udine.

Mahtani, M. M., & Willard, H. F. (1998). Physical and genetic mapping of the human X

chromosome centromere: repression of recombination. *Genome Research*, *8*(2), 100–10. https://doi.org/10.1101/GR.8.2.100

Marroni, F., Pinosio, S., & Morgante, M. (2014). Structural variation and genome complexity: Is dispensable really dispensable? *Current Opinion in Plant Biology*, *18*(1), 31–36. https://doi.org/10.1016/j.pbi.2014.01.003

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

Mather, K. (1939). Crossing over and Heterochromatin in the X Chromosome of Drosophila Melanogaster. *Genetics*, *24*(3), 413–35. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/17246931

Mayer, K. F. X., Rogers, J., Dole el, J., Pozniak, C., Eversole, K., Feuillet, C., … Praud, S. (2014). A chromosome-based draft sequence of the hexaploid bread wheat (Triticum aestivum) genome. *Science*, *345*(6194), 1251788–1251788. https://doi.org/10.1126/science.1251788

Mayer, K. F. X., Waugh, R., Langridge, P., Close, T. J., Wise, R. P., Graner, A., … Fincher, G. B. (2012). A physical, genetic and functional sequence assembly of the barley genome. *Nature*, 1–83. https://doi.org/10.1038/nature11543

Maynard-Smith, J., & Haigh, J. (1974). The hitch-hiking effect of a favourable gen. *Genet. Res., Camb*, *23*, 23–35.

Meaburn, K. J., Misteli, T., & Soutoglou, E. (2007). Spatial genome organization in the formation of chromosomal translocations. *Seminars in Cancer Biology*, *17*(1), 80–90. https://doi.org/10.1016/j.semcancer.2006.10.008

Melamed-Bessudo, C., & Levy, A. A. (2012). Deficiency in DNA methylation increases meiotic crossover rates in euchromatic but not in heterochromatic regions in Arabidopsis. *Proceedings of the National Academy of Sciences of the United States of America*, *109*(16), E981-8. https://doi.org/10.1073/pnas.1120742109

Meyer, R. S., DuVal, A. E., & Jensen, H. R. (2012). Patterns and processes in crop domestication: an historical review and quantitative analysis of 203 global food crops. *New Phytologist*, *196*(1), 29–48. https://doi.org/10.1111/j.1469-8137.2012.04253.x

Mezmouk, S., & Ross-Ibarra, J. (2014). The Pattern and Distribution of Deleterious Mutations in Maize. *G3 (Bethesda, Md.)*, *4*(1), 163–71. https://doi.org/10.1534/g3.113.008870

Miller, A. J., & Gross, B. L. (2011). From forest to field: perennial fruit crop domestication. *American Journal of Botany*, *98*(9), 1389–414. https://doi.org/10.3732/ajb.1000522

Mirouze, M., Lieberman-Lazarovich, M., Aversano, R., Bucher, E., Nicolet, J., Reinders, J., & Paszkowski, J. (2012). Loss of DNA methylation affects the recombination landscape in Arabidopsis. *Proceedings of the National Academy of Sciences of the United States of*

America, *109*(15), 5880–5. https://doi.org/10.1073/pnas.1120841109

Morgante, M. (2006). Plant genome organisation and diversity: the year of the junk! *Current Opinion in Biotechnology*, *17*(2), 168–173. https://doi.org/10.1016/j.copbio.2006.03.001

Morgante, M., De Paoli, E., & Radovic, S. (2007). Transposable elements and the plant pan-genomes. *Current Opinion in Plant Biology*, *10*(2), 149–155. https://doi.org/10.1016/j.pbi.2007.02.001

Morton, N. E. (1960). The mutational load due to detrimental genes in man. *American Journal of Human Genetics*, *12*(764), 348–364.

Morton, N. E., Crow, J. F., & Muller, H. J. (1956). An Estimate of the Mutational Damage in Man From Data on Consanguineous Marriages. *Proceedings of the National Academy of Sciences of the United States of America*, *42*(11), 855–863. https://doi.org/10.1073/pnas.42.11.855

Muñoz-Amatriaín, M., Eichten, S. R., Wicker, T., Richmond, T. A., Mascher, M., Steuernagel, B., … Stein, N. (2013). Distribution, functional impact, and origin mechanisms of copy number variation in the barley genome. *Genome Biology*, *14*, R58. https://doi.org/10.1186/gb-2013-14-6-r58

Myburg, A. A., Vogl, C., Griffin, A. R., Sederoff, R. R., & Whetten, R. W. (2004). Genetics of Postzygotic Isolation in Eucalyptus: Whole-Genome Analysis of Barriers to Introgression in a Wide Interspecific Cross of Eucalyptus grandis and E. globulus. *Genetics*, *166*(3), 1405–1418. https://doi.org/10.1534/genetics.166.3.1405

Myles, S., Boyko, A. R., Owens, C. L., Brown, P. J., Grassi, F., Aradhya, M. K., … Buckler, E. S. (2011). Genetic structure and domestication history of the grape. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(9), 3530–3535. https://doi.org/10.1073/pnas.1009363108

Myles, S., Chia, J.-M. M., Hurwitz, B., Simon, C., Zhong, G. Y., Buckler, E., & Ware, D. (2010). Rapid genomic characterization of the genus Vitis. *PLoS ONE*, *5*(1). https://doi.org/10.1371/journal.pone.0008219

Ne Giraut, L., Falque, M., Drouaud, J., Pereira, L., Martin, O. C., & Mé Zard, C. (2011). Genome-Wide Crossover Distribution in Arabidopsis thaliana Meiosis Reveals Sex-Specific Patterns along Chromosomes. *PLoS Genet*, *7*(11). https://doi.org/10.1371/journal.pgen.1002354

Paape, T., Zhou, P., Branca, A., Briskine, R., Young, N., & Tiffin, P. (2012). Fine-scale population recombination rates, hotspots, and correlates of recombination in the Medicago truncatula genome. *Genome Biology and Evolution*, *4*(5), 726–37. https://doi.org/10.1093/gbe/evs046

Pan, J., Sasaki, M., Kniewel, R., Murakami, H., Blitzblau, H. G., Tischfield, S. E., … Keeney, S. (2011). A Hierarchical Combination of Factors Shapes the Genome-wide Topography of Yeast Meiotic Recombination Initiation. *Cell*, *144*, 719–731.

https://doi.org/10.1016/j.cell.2011.02.009

Pandey, R. V., Franssen, S. U., Futschik, A., & Schlötterer, C. (2013). Allelic imbalance metre (Allim), a new tool for measuring allele-specific gene expression with RNA-seq data. *Molecular Ecology Resources*, *13*(4), 740–745. https://doi.org/10.1111/1755-0998.12110

Pelak, K., Need, A. C., Fellay, J., Shianna, K. V, Feng, S., Urban, T. J., … NIAID Center for HIV/AIDS Vaccine Immunology, on behalf of N. C. for H. V. I. (2011). Copy number variation of KIR genes influences HIV-1 control. *PLoS Biology*, *9*(11), e1001208. https://doi.org/10.1371/journal.pbio.1001208

Peterson, B. K., Weber, J. N., Kay, E. H., Fisher, H. S., & Hoekstra, H. E. (2012). Double digest RADseq: An inexpensive method for de novo SNP discovery and genotyping in model and non-model species. *PLoS ONE*, *7*(5). https://doi.org/10.1371/journal.pone.0037135

Petkov, P. M., Broman, K. W., Szatkiewicz, J. P., & Paigen, K. (2007). Crossover interference underlies sex differences in recombination rates. *Trends in Genetics*, *23*(11), 539–542. https://doi.org/10.1016/j.tig.2007.08.015

Phillips, D., Jenkins, G., Macaulay, M., Nibau, C., Wnetrzak, J., Fallding, D., … Ramsay, L. (2015). The effect of temperature on the male and female recombination landscape of barley. *New Phytologist*. https://doi.org/10.1111/nph.13548

Phillips, P. C. (2008). Epistasis: the essential role of gene interactions in the structure and evolution of genetic systems. *Nat Rev Genet*, *9*(11), 855–867. https://doi.org/10.1038/nrg2452

Pinosio, S., Giacomello, S., Faivre-Rampant, P., Taylor, G., Jorge, V., Le Paslier, M. C., … Morgante, M. (2016). Characterization of the Poplar Pan-Genome by Genome-Wide Identification of Structural Variation. *Molecular Biology and Evolution*, *33*(10), 2706–19. https://doi.org/10.1093/molbev/msw161

Price, M. N., & Arkin, A. P. (2015). Weakly Deleterious Mutations and Low Rates of Recombination Limit the Impact of Natural Selection on Bacterial Genomes. *mBio*, *6*(6), e01302-15. https://doi.org/10.1128/mBio.01302-15

Pritham, E. J., & Thomas, J. (2015). Helitrons, the Eukaryotic Rolling-circle Transposable Elements. *Microbiology Spectrum*, *3*(4). https://doi.org/10.1128/microbiolspec.MDNA3-0049-2014

R Core Team. (2015). R: A Language and Environment for Statistical Computing. Vienna, Austria. Retrieved from https://www.r-project.org/

Rausch, T., Zichner, T., Schlattl, A., Stütz, A. M., Benes, V., & Korbel, J. O. (2012). DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics (Oxford, England)*, *28*(18), i333–i339. https://doi.org/10.1093/bioinformatics/bts378

Renaut, S., & Rieseberg, L. H. (2015). The accumulation of deleterious mutations as a consequence of domestication and improvement in sunflowers and other compositae crops. *Molecular Biology and Evolution*, *32*(9), 2273–2283. https://doi.org/10.1093/molbev/msv106

Rodgers-Melnick, E., Mane, S. P., Dharmawardhana, P., Slavov, G. T., Crasta, O. R., Strauss, S. H., … DiFazio, S. P. (2012). Contrasting patterns of evolution following whole genome versus tandem duplication events in Populus. *Genome Research*, *22*(1), 95–105. https://doi.org/10.1101/gr.125146.111

Round, E. K., Flowers, S. K., & Richards, E. J. (1997). Arabidopsis thaliana centromere regions: genetic map positions and repetitive DNA structure. *Genome Research*, *7*(11), 1045–53. https://doi.org/10.1101/GR.7.11.1045

Sabir, A. (2011). Influences of Self-and Cross-pollinations on Berry Set, Seed Characteristics and Germination Progress of Grape (Vitis vinifera cv. Italia). *ISSN Online Int. J. Agric. Biol*, *13*, 1560–8530. Retrieved from http://www.fspublishers.org

Sachidanandam, R., Weissman, D., Schmidt, S. C., Kakol, J. M., Stein, L. D., Marth, G., … International SNP Map Working Group. (2001). A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature*, *409*(6822), 928–933. https://doi.org/10.1038/35057149

Salomé, P., Bomblies, K., Fitz, J., Laitinen, R., Warthmann, N., Yant, L., & Weigel, D. (2011). The recombination landscape in Arabidopsis thaliana F 2 populations. *Heredity*, *108*, 447–45595. https://doi.org/10.1038/hdy.2011.95

Sasaki, M., Lange, J., & Keeney, S. (2010). Genome destabilization by homologous recombination in the germ line. *Nature Reviews. Molecular Cell Biology*, *11*(3), 182–95. https://doi.org/10.1038/nrm2849

Sato, S., Tabata, S., Hirakawa, H., Asamizu, E., Shirasawa, K., Isobe, S., … Gianese, G. (2012). The tomato genome sequence provides insights into fleshy fruit evolution. *Nature*, *485*(7400), 635–641. https://doi.org/10.1038/nature11119

Scherer, S. W., Lee, C., Birney, E., Altshuler, D. M., Eichler, E. E., Carter, N. P., … Feuk, L. (2007). Challenges and standards in integrating surveys of structural variation. *Nature Genetics*, *39*(7 Suppl), S7–S15. https://doi.org/10.1038/ng2093

Schnable, P. S., Ware, D., Fulton, R. S., Stein, J. C., Wei, F., Pasternak, S., … Wilson, R. K. (2009). The B73 Maize Genome: Complexity, Diversity, and Dynamics. *Science*, *326*(5956), 1112–1115. https://doi.org/10.1126/science.1178534

Sherman, J. D., & Stack, S. M. (1995). Two-Dimensional Spreads of Synaptonemal Complexes from Solanaceous Plants. VI. High-Resolution Recombination Nodule Map for Tomato (Lycofiersicon esculentum). *Genetics*, *141*(2), 683–708.

Shi, J., Wolf, S. E., Burke, J. M., Presting, G. G., Ross-Ibarra, J., Dawe, R. K., … Doebley, J. (2010). Widespread Gene Conversion in Centromere Cores. *PLoS Biology*, *8*(3), e1000327. https://doi.org/10.1371/journal.pbio.1000327

Si, W., Yuan, Y., Huang, J., Zhang, X., Zhang, Y., Zhang, Y., … Yang, S. (2015). Widely distributed hot and cold spots in meiotic recombination as shown by the sequencing of rice F2 plants. *The New Phytologist*, *206*(4), 1491–502. https://doi.org/10.1111/nph.13319

Simon, L., Voisin, M., Tatout, C., & Probst, A. V. (2015). Structure and Function of Centromeric and Pericentromeric Heterochromatin in Arabidopsis thaliana. *Frontiers in Plant Science*, *6*, 1049. https://doi.org/10.3389/fpls.2015.01049

Simons, Y. B., Turchin, M. C., Pritchard, J. K., & Sella, G. (2014). The deleterious mutation load is insensitive to recent population history. *Nat Genet*, *46*(3), 220–224. https://doi.org/10.1038/ng.2896

Sindi, S., Helman, E., Bashir, A., & Raphael, B. J. (2009). A geometric approach for classification and comparison of structural variants. *Bioinformatics (Oxford, England)*, *25*(12), i222-30. https://doi.org/10.1093/bioinformatics/btp208

Snyder, E., & Harmon, F. (1939). Grape Progenies of Self-Pollinated Vinifera Varieties. *American Society for Horticulture Science*, *37*, 625–626.

Springer, N. M., & Stupar, R. M. (2007). Allelic variation and heterosis in maize: How do two halves make more than a whole? *Genome Research*, *17*(3), 264–275. https://doi.org/10.1101/gr.5347007

Springer, N. M., Ying, K., Fu, Y., Ji, T., Yeh, C.-T. T., Jia, Y., … Schnable, P. S. (2009). Maize inbreds exhibit high levels of copy number variation (CNV) and presence/absence variation (PAV) in genome content. *PLoS Genetics*, *5*(11), e1000734. https://doi.org/10.1371/journal.pgen.1000734

Studer, A., Zhao, Q., Ross-Ibarra, J., & Doebley, J. (2011). Identification of a functional transposon insertion in the maize domestication gene tb1. *Nature Genetics*, *43*(11), 1160–3. https://doi.org/10.1038/ng.942

Sudmant, P. H., Kitzman, J. O., Antonacci, F., Alkan, C., Malig, M., Tsalenko, A., … Eichler, E. E. (2010). Diversity of human copy number variation and multicopy genes. *Science (New York, N.Y.)*, *330*(6004), 641–6. https://doi.org/10.1126/science.1197005

Sudmant, P. H., Mallick, S., Nelson, B. J., Hormozdiari, F., Krumm, N., Huddleston, J., … Eichler, E. E. (2015). Global diversity, population stratification, and selection of human copy-number variation. *Science*, *349*(6253).

Swanson-Wagner, R. A., Eichten, S. R., Kumari, S., Tiffin, P., Stein, J. C., Ware, D., & Springer, N. M. (2010). Pervasive gene content variation and copy number variation in maize and its undomesticated progenitor. *Genome Research*, *20*(12), 1689–1699.

https://doi.org/10.1101/gr.109165.110

Tajima, F. (1989). Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics*, *123*(3), 585–595. https://doi.org/PMC1203831

Talukdar, D. (2010). Reciprocal Translocations in Grass Pea (Lathyrus sativus L .): Pattern of Transmission , Detection of Multiple Interchanges and their Independence. *J Hered*, *101*(2), 169–176. https://doi.org/10.1093/jhered/esp106

Termolino, P., Cremona, G., Consiglio, M. F., & Conicella, C. (2016). Insights into epigenetic landscape of recombination-free regions. *Chromosoma*. https://doi.org/10.1007/s00412-016-0574-9

Tettelin, H., Masignani, V., Cieslewicz, M. J., Donati, C., Medini, D., Ward, N. L., … Fraser, C. M. (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. https://doi.org/10.1073/pnas.0506758102

Theodorou, K., & Couvet, D. (2006). On the expected relationship between inbreeding, fitness, and extinction. *Genetics, Selection, Evolution : GSE*, *38*(4), 371–87. https://doi.org/10.1051/gse:2006010

Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D. R., … Pachter, L. (2012). Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nature Protocols*, *7*(3), 562–78. https://doi.org/10.1038/nprot.2012.016

Trapnell, C., Williams, B. A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M. J., … Pachter, L. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature Biotechnology*, *28*(5), 511–515. https://doi.org/10.1038/nbt.1621

Van der Auwera, G. A., Carneiro, M. O., Hartl, C., Poplin, R., Del Angel, G., Levy-Moonshine, A., … DePristo, M. A. (2013). From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Current Protocols in Bioinformatics*, *43*(1110), 11.10.1-33. https://doi.org/10.1002/0471250953.bi1110s43

Van Os, H., P. Stam, R. G. F. Visser and H. J. van Eck. (2005). RECORD: a novel method for ordering loci on a genetic linkage map. *Theor. Appl. Genet.*, *112*(1), 30–40.

Vincenten, N., Kuhl, L.-M., Lam, I., Oke, A., Kerr, A. R., Hochwagen, A., … Keeney, S. (2015). The kinetochore prevents centromere-proximal crossover recombination during meiosis. *eLife*, *4*, 923–937. https://doi.org/10.7554/eLife.10850

Vitulo, N., Forcato, C., Carpinelli, E., Telatin, A., Campagna, D., D'Angelo, M., … 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.

https://doi.org/10.1186/1471-2229-14-99

Wang, K., Li, M., & Hakonarson, H. (2010). ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Research*, *38*(16), e164. https://doi.org/10.1093/nar/gkq603

Wang G, Li H, Cheng Z, J. W. (2013). A novel translocation event leads to a recombinant stable chromosome with interrupted centromeric domains in rice. *Chromosoma*, *122*(4), 295–303. https://doi.org/DOI: 10.1007/s00412-013-0413-1

Wang L, Kikuchi S, Muto C, Naito K, Isemura T, Ishimoto M, Cheng X, Kaga A, T. N. (2015). Reciprocal translocation identified in Vigna angularis dominates the wild population in East Japan. *J Plant Res*, *128*(4), 653–663. https://doi.org/DOI: 10.1007/s10265-015-0720-0

Whitlock, M., & Davis, B. (2011). Genetic Load. In *eLS*. Chichester, UK: John Wiley & Sons, Ltd. https://doi.org/10.1002/9780470015902.a0001787.pub2

Wicker, T., Sabot, F., Hua-Van, A., Bennetzen, J. L., Capy, P., Chalhoub, B., … Schulman, A. H. (2007). A unified classification system for eukaryotic transposable elements. *Nature Reviews Genetics*, *8*(12), 973–982. https://doi.org/10.1038/nrg2165

Willis, J. A., Mukherjee, S., Orlow, I., Viale, A., Offit, K., Kurtz, R. C., … Klein, R. J. (2014). Genome-wide analysis of the role of copy-number variation in pancreatic cancer risk. *Frontiers in Genetics*, *5*, 29. https://doi.org/10.3389/fgene.2014.00029

Wu, Y., Bhat, P. R., Close, T. J., & Lonardi, S. (2008). Efficient and accurate construction of genetic linkage maps from the minimum spanning tree of a graph. *PLoS Genetics*, *4*(10), e1000212. https://doi.org/10.1371/journal.pgen.1000212

Yamamoto, M., & Miklos, G. L. (1978). Genetic studies on heterochromatin in Drosophila melanogaster and their implications for the functions of satellite DNA. *Chromosoma*, *66*(1), 71–98. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/416935

Yelina, N., Diaz, P., Lambing, C., & Henderson, I. R. (2015). Epigenetic control of meiotic recombination in plants. *Science China Life Sciences*. https://doi.org/10.1007/s11427-015-4811-x

Yelina, N. E., Choi, K., Chelysheva, L., Macaulay, M., de Snoo, B., Wijnker, E., … Henderson, I. R. (2012). Epigenetic Remodeling of Meiotic Crossover Frequency in Arabidopsis thaliana DNA Methyltransferase Mutants. *PLoS Genetics*. https://doi.org/10.1371/journal.pgen.1002844

You, F. M., Huo, N., Gu, Y. Q., Luo, M.-C., Ma, Y., Hane, D., … Anderson, O. D. (2008). BatchPrimer3: a high throughput web application for PCR and sequencing primer design. *BMC Bioinformatics*, *9*, 253. https://doi.org/10.1186/1471-2105-9-253

Zeng, K., & Charlesworth, B. (2010). The effects of demography and linkage on the estimation of selection and mutation parameters. *Genetics*, *186*(4), 1411–1424.

https://doi.org/10.1534/genetics.110.122150

Zhang, M., Zhou, L., Bawa, R., Suren, H., & Holliday, J. A. A. (2016). Recombination Rate Variation, Hitchhiking, and Demographic History Shape Deleterious Load in Poplar. *Molecular Biology and Evolution*, *33*(11), 2899–2910. https://doi.org/10.1093/molbev/msw169

Zhang, Z., Zhu, H., Gill, B. S., & Li, W. (2015). Fine mapping of shattering locus Br2 reveals a putative chromosomal inversion polymorphism between the two lineages of Aegilops tauschii. *Theoretical and Applied Genetics*, *128*(4), 745–755. https://doi.org/10.1007/s00122-015-2469-1

Zhao, D., Ferguson, A. A., & Jiang, N. (2016). What makes up plant genomes: The vanishing line between transposable elements and genes. *Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms*, *1859*(2), 366–380. https://doi.org/10.1016/j.bbagrm.2015.12.005

Zheng, L.-Y., Guo, X.-S., He, B., Sun, L.-J., Peng, Y., Dong, S.-S., … Jing, H.-C. (2011). Genome-wide patterns of genetic variation in sweet and grain sorghum (Sorghum bicolor). *Genome Biology*, *12*(11), R114. https://doi.org/10.1186/gb-2011-12-11-r114

Żmieńko, A., Samelak, A., Kozłowski, P., & Figlerowicz, M. (2014). Copy number polymorphism in plant genomes. *TAG. Theoretical and Applied Genetics. Theoretische Und Angewandte Genetik*, *127*(1), 1–18. https://doi.org/10.1007/s00122-013-2177-7

Appendix 1

Population of 137 varieties of *Vitis vinifera* species (128 cultivars and 9 introgression lines) used in the framework of the NOVABREED project for the identification of SNP variation in *Vitis vinifera* species. Varieties indicated by the asterisk were used also for the detection of SVs.

Variety[1]	Origin[2]
31122	breeding
31125	breeding
32078	breeding
34111	breeding
34113	breeding
55100	breeding
58083	breeding
76026	breeding
Aciaruli Tetri	breeding
Agadai	Dagestan
Aglianico	Italy
Airen	Spain
Alexandroouli	Georgia
Ansonica*	Italy
Ararati	Armenia
Assyrtico	Greece
Asyl Kara	Dagestan
Autumn Royal	breeding
Barbera*	Italy
Bayan Shirei	Azerbaijan
Berzamino	Italy
Bovale	Italy
Cabernet Franc*	France
Cabernet Sauvignon*	France
Cannonau	Spain
Carignano	France
Catarratto Bianco Comune*	Italy
Cesanese D'affile	Italy
Chaouch Blanc	Turkey
Charistvala Kolchuri	breeding
Chasselas	-
Clairette Blanche	France
Corvina Veronese*	Italy
Daphnia	Greece

Variety[1]	Origin[2]
Disecka Old	Slovenia
Enantio	Italy
Falanghina*	Italy
Fiano*	Italy
Fumat	Italy
Garganega*	Italy
Glera*	Italy
Gorula	Georgia
Gouais Blanc*	-
Grechetto Bianco*	Italy
Greco di Tufo*	Italy
Grignolino	Italy
Gyulyabi Dagestanskii	Dagestan
Harslevelue	Hungary
Hop Halat	Dagestan
Italia	breeding
Kadarka	Hungary
Katta Kurgan	Uzbekistan
Kishmish Vatkana*	Uzbekistan
Klarnica	-
Lambrusco Grasparossa*	Italy
Lambrusco Sorbara	Italy
Limnio	Greece
Malvasia Bianca Lunga*	-
Malvasia del Lazio	-
Malvasia di Lipari	-
Malvasia Istriana	Croatia
Marandi Shemakhinskii	Azerbaijan
Mauzac	France
Mavrodaphni	Greece
Merlot Noir*	France
Montepulciano*	Italy
Moscato Bianco*	-
Moscato di Scanzo	breeding
Mtsvane Kachuri	Georgia
Narma	Dagestan
Nasco*	Italy
Nebbiolo*	Italy
Negroamaro	Italy
Nero d'Avola*	Italy
Nieddu Mannu	Italy

Variety[1]	Origin[2]
Nosiola*	Italy
Ogialesci	Georgia
Passerina*	Italy
Pecorino*	Italy
Petit Rouge	Italy
Picolit*	Italy
Pignoletto*	Italy
Pinela	Slovenia
Pinot Blanc Meunier*	France
Plechistik	Russian Fed
PN40024*	breeding
Primitivo	Croatia
Pukhlyakovskii	Moldova
Raboso Piave	Italy
Red Globe	breeding
Refosco dal Peduncolo Rosso*	Italy
Rhein Riesling*	Germany
Ribolla Gialla*	Italy
Riesling Italico*	-
Rkatsiteli*	Georgia
Rossese	France
Rpv3	breeding
Rumi Ahmar	Turkey
Sagrantino	Italy
Sahibi Safid	Afghanistan
Sangiovese*	Italy
Sauvignon*	France
Schiava Gentile*	Italy
Schiava Grossa	Italy
Schioppettino	Italy
Sciavtsitska	-
Shafei	Azerbaijan
Sirgula*	Georgia
Sultanina*	Turkey
Tagobi	Tajikistan
Taifi Rozovyi	Uzbekistan
Tannat*	France
Tavkveri	Georgia
Terbash*	Turkmenistan
Terrano*	Italy
Thompson Seedless	-

Variety[1]	Origin[2]
Tocai Friulano*	Italy
Traminer*	-
Trebbiano Toscano*	Italy
Tschvediansis Tetra	-
Uva di Troia*	Italy
V267	-
V278	-
V292	-
V294	-
V385	-
V389	-
V395*	-
V400	-
V410	-
V411	-
Verdicchio Bianco*	Italy
Verduzzo	Italy
Vermentino	Italy
Vernaccia*	Italy
Zametovka	-
Zelen*	Slovenia

[2] Prime name (short format) of the variety.

[3] Country of origin, if available.

Appendix 2

List of 196 varieties of *Vitis vinifera*, and *Vitis lambrusca* species tested for the presence of translocation in the present work.

Species	**Variety[2]**	**Origin[3]**
Vitis labrusca		United states
Vitis vinifera	Abouhou	Morocco
Vitis vinifera	Aciaruli Tetri	Georgia
Vitis vinifera	Ag Emchek	Russian Federation
Vitis vinifera	Ag Emchek - Emchek Izum	Russian Federation
Vitis vinifera	Ag Izyum	Dagestan
Vitis vinifera	Agadai	Dagestan
Vitis vinifera	Aglianico	Italy
Vitis vinifera	Airen	Spain
Vitis vinifera	Aleatico	Italy
Vitis vinifera	Alexandroouli	Georgia
Vitis vinifera	Almeria Negra	France
Vitis vinifera	Almura Shavi	Georgia
Vitis vinifera	Angur Siekh Gissarskii	Tajikistan
Vitis vinifera	Ansonica	Italy
Vitis vinifera	Aramon	France
Vitis vinifera	Ararati	Armenia
Vitis vinifera	Arneis	Italy
Vitis vinifera	Assyrtico	Greece
Vitis vinifera	Asyl kara	Dagestan
Vitis vinifera	Barbera	Italy
Vitis vinifera	Bayan Shirei	Azerbaijan
Vitis vinifera	Beregovsky Rosovy	Russian Federation
Vitis vinifera	Berzamino	Italy
Vitis vinifera	Bobal	Spain
Vitis vinifera	Bovale	Italy
Vitis vinifera	Cabernet Franc	France
Vitis vinifera	Cannonau	Italy
Vitis vinifera	Carignane	Spain
Vitis vinifera	Catarratto Bianco Comune	Italy
Vitis vinifera	Cesanese D'Affile	Italy
Vitis vinifera	Chaouch Blanc	Turkey
Vitis vinifera	Charistvala Kolchuri 631	Georgia
Vitis vinifera	Chasselas	France
Vitis vinifera	Chenin Blanc	France
Vitis vinifera	Cividin	Italy

Species	Variety[2]	Origin[3]
Vitis vinifera	Cjanorie	Italy
Vitis vinifera	Clairette Blanche	France
Vitis vinifera	Colombard	France
Vitis vinifera	Corvina	Italy
Vitis vinifera	Daphnia	Greece
Vitis vinifera	Disecka Old	Slovenia
Vitis vinifera	Dolcetto	Italy
Vitis vinifera	Doroi Krasnyi	Uzbekistan
Vitis vinifera	Efremovskii Vtoroi	Russian Federation
Vitis vinifera	Enantio	Italy
Vitis vinifera	Falanghina	Italy
Vitis vinifera	Feteasca Neagra	Moldova
Vitis vinifera	Fiano	Italy
Vitis vinifera	Folle Blanche	France
Vitis vinifera	Fumat	Italy
Vitis vinifera	Furmint	Hungary
Vitis vinifera	Garganega	Italy
Vitis vinifera	Glera	Italy
Vitis vinifera	Gorula	Georgia
Vitis vinifera	Gouais Blanc	France
Vitis vinifera	Graciano	Spain
Vitis vinifera	Grechetto	Italy
Vitis vinifera	Greco di Tufo	Italy
Vitis vinifera	Grignolino	Italy
Vitis vinifera	Gyulyabi Dagestanskii	Dagestan
Vitis vinifera	Harslevelue	Hungary
Vitis vinifera	Hatal Baar	-
Vitis vinifera	Hoca Cibi	-
Vitis vinifera	Hop Halat	Dagestan
Vitis vinifera	Italia	breeding
Vitis vinifera	Kadarka	Hungary
Vitis vinifera	Kaltak Kara Tagapskii	Russian Federation
Vitis vinifera	Kandahari Siah	Afghanistan
Vitis vinifera	Katta Kurgan	Uzbekistan
Vitis vinifera	Kishmish Vatkana	Uzbekistan
Vitis vinifera	Klarnica	Slovenia
Vitis vinifera	Kosorotovskii	Russian Federation
Vitis vinifera	Krasnostop Anapskii	-
Vitis vinifera	Krasnotop Rostov	-
Vitis vinifera	Kreaca	Serbia
Vitis vinifera	Lacu Kere	-

Species	Variety[2]	Origin[3]
Vitis vinifera	Lambrusco Grasparossa	Italy
Vitis vinifera	Lambrusco Salamino	Italy
Vitis vinifera	Lambrusco Sorbara	Italy
Vitis vinifera	Limnio	Greece
Vitis vinifera	Maccabeo	-
Vitis vinifera	Malbech	-
Vitis vinifera	Malvasia Bianca Lunga	-
Vitis vinifera	Malvasia del Lazio	Italy
Vitis vinifera	Malvasia di Candia Aromatica	Italy
Vitis vinifera	Malvasia di Lipari	Italy
Vitis vinifera	Malvasia Istriana	Croatia
Vitis vinifera	Mammolo	-
Vitis vinifera	Marandi Shemakhinskii 564	Azerbaijan
Vitis vinifera	Mauzac	France
Vitis vinifera	Mavrodaphni	Greece
Vitis vinifera	Merlot Noir	France
Vitis vinifera	Montepulciano	Italy
Vitis vinifera	Moscato Bianco	Italy
Vitis vinifera	Muscat of Alexandria	Egypt
Vitis vinifera	Moscato di Scanzo	breeding
Vitis vinifera	Moscato Rosa	-
Vitis vinifera	Mourvedre	Russian Federation
Vitis vinifera	Mtsvane Kachuri	Georgia
Vitis vinifera	Muchuchar	Dagestan
Vitis vinifera	Narma	Dagestan
Vitis vinifera	Nasco	Italy
Vitis vinifera	Nebbiolo	Italy
Vitis vinifera	Negroamaro	Italy
Vitis vinifera	Nero d'Avola	Italy
Vitis vinifera	Nieddu Mannu	Italy
Vitis vinifera	Nosiola	Italy
Vitis vinifera	Ogialesci	Georgia
Vitis vinifera	Parellada	Spain
Vitis vinifera	Passerina	Italy
Vitis vinifera	Pecorino	Italy
Vitis vinifera	Petit Rouge	Italy
Vitis vinifera	Petit Verdot	France
Vitis vinifera	Picolit	Italy
Vitis vinifera	Pignoletto	Italy
Vitis vinifera	Pinela	Slovenia
Vitis vinifera	Pinot	France

Species	Variety[2]	Origin[3]
Vitis vinifera	Plechistik	Russian Federation
Vitis vinifera	PN40024	Breeding
Vitis vinifera	Primitivo	Italy
Vitis vinifera	Pukhlyakovskii	Russian Federation
Vitis vinifera	Raboso Piave	Italy
Vitis vinifera	Rasheh	-
Vitis vinifera	Rassegui	Tunisia
Vitis vinifera	Refosco dal Peduncolo Rosso	Italy
Vitis vinifera	Regina	Spain
Vitis vinifera	Rezè	Switzerland
Vitis vinifera	Rhein Riesling	Germany
Vitis vinifera	Ribolla Gialla	Italy
Vitis vinifera	Riesling Italico	France
Vitis vinifera	Rkatsiteli	Georgia
Vitis vinifera	Rossese	Italy
Vitis vinifera	Roussaïtis	Greece
Vitis vinifera	Rumi Ahmar	Egypt
Vitis vinifera	Sagrantino	Italy
Vitis vinifera	Sahibi Safid	Afghanistan
Vitis vinifera	Saint Laurent	France
Vitis vinifera	Sangiovese	Italy
Vitis vinifera	Sarakh	Dagestan
Vitis vinifera	Sarfeher	Hungary
Vitis vinifera	Sauvignon	France
Vitis vinifera	Schiava Gentile	Italy
Vitis vinifera	Schiava Grossa	Italy
Vitis vinifera	Schioppettino	Italy
Vitis vinifera	Sciavtsitska	-
Vitis vinifera	Shafei	Azerbaijan
Vitis vinifera	Shahani	Iran
Vitis vinifera	Shavrany	Dagestan
Vitis vinifera	Shilokhvostyi	Russian Federation
Vitis vinifera	Sibircov	-
Vitis vinifera	Sirgula	Georgia
Vitis vinifera	Sultanina	Turkey
Vitis vinifera	Stanichnyi Belyi	Russian Federation
Vitis vinifera	sylvestris v267	
Vitis vinifera	sylvestris Armenia v411	Armenia
Vitis vinifera	sylvestris Azerbaijan v385	Azerbaijan
Vitis vinifera	sylvestris Azerbaijan v395	Azerbaijan
Vitis vinifera	sylvestris Azerbaijan v410	Azerbaijan

Species	Variety[2]	Origin[3]
Vitis vinifera	sylvestris Dagestan, Zangelansk v278	Dagestan
Vitis vinifera	sylvestris Derbent v389	Derbent
Vitis vinifera	sylvestris Turkmenistan v292	Turkmenistan
Vitis vinifera	sylvestris Turkmenistan v294	Turkmenistan
Vitis vinifera	sylvestris v400	-
Vitis vinifera	Syrah	France
Vitis vinifera	Tagobi	Tajikistan
Vitis vinifera	Taifi Rozovyi	Uzbekistan
Vitis vinifera	Tavkveri	Georgia
Vitis vinifera	Tavlinskii Pozdnii	Dagestan
Vitis vinifera	Tazzelenghe	Italy
Vitis vinifera	Tempranillo	Spain
Vitis vinifera	Terbash	Russian Federation
Vitis vinifera	Terra Promessa	-
Vitis vinifera	Terrano	Italy
Vitis vinifera	Tibouren	France
Vitis vinifera	Tinta Cao	Portugal
Vitis vinifera	Tocai Friulano	Italy
Vitis vinifera	Tolstokoryi	Russian Federation
Vitis vinifera	Traminer	Italy
Vitis vinifera	Trebbiano Toscano	Italy
Vitis vinifera	Trnjak	Ex-Yugoslavia
Vitis vinifera	Tschvediansis Tetra	-
Vitis vinifera	Uva di Troia	Italy
Vitis vinifera	Varyoshkin	Russian Federation
Vitis vinifera	Veltliner	Austria
Vitis vinifera	Verdicchio	Italy
Vitis vinifera	Verduzzo	Italy
Vitis vinifera	Vermentino	Italy
Vitis vinifera	Vernaccia	Italy
Vitis vinifera	Viognier	France
Vitis vinifera	Vitouska	Slovenia
Vitis vinifera	Vranac	Montenegro
Vitis vinifera	Zaarma	Iran
Vitis vinifera	Zametovka	-
Vitis vinifera	Zelen	Slovenia
Vitis vinifera	Zemialski Ciorni	-
Vitis vinifera	Zilavka	Bosnia and Herzegovina

[2] Prime name (short format) of the variety.

[3] Country of origin, if available.

Appendix 3

For each candidate gene, table describes the number of duplicated copies according to BLASTp analysis and the information of gene expression in at least one copy.

Candidate gene ID	Annotation V2.1	Variety	Expression[1]	n. copies (Blastp)	Expression in dup. genes[2]
VIT_204s0008g04010.1	lupus la	Primitivo	na	1	no duplicated gene
VIT_204s0008g04160.1	burp domain-containing protein	Primitivo	na	2	na
VIT_204s0008g04610.1	pentatricopeptide repeat-containing protein	Primitivo	na	2	na
VIT_204s0008g05430.1	rna-dependent rna polymerase 6	Primitivo	na	1	no duplicated gene
VIT_204s0079g00620.1	rna polymerase beta subunit	Schiava Grossa	na	1	no duplicated gene
VIT_205s0029g00656.1	beta galactosidase 9	Sangiovese	yes	2	yes
VIT_205s0029g00730.1	methyltransferase-like protein 13-like	Sangiovese	yes	1	no duplicated gene
VIT_205s0029g00830.1	uncharacterized protein	Sangiovese	yes	2	yes
VIT_205s0029g00850.1	tmv resistance protein n-like	Sangiovese	yes	1	no duplicated gene
VIT_205s0049g01620.1	cyclopropane-fatty-acyl-phospholipid synthase	Sangiovese	yes	2	yes
VIT_205s0049g01710.1	transducin wd-40 repeat-containing protein	Sangiovese	no	1	no duplicated gene
VIT_205s0051g00050.1	myosin xi-2	Sangiovese	yes	1	no duplicated gene
VIT_205s0051g00730.1	ring u-box domain-containing protein	Sangiovese	yes	1	no duplicated gene
VIT_205s0051g00830.1	dihydroxy-acid dehydratase	Sangiovese	yes	1	no duplicated gene
VIT_205s0136g00140.1	uncharacterized protein sll0005-like	Sangiovese	yes	1	no duplicated gene
VIT_208s0007g00810.1	concanavalin a-like lectin kinase-like protein	Sangiovese	yes	1	no duplicated gene
VIT_208s0007g01100.1	uracil phosphoribosyltransferase	Sangiovese	yes	1	no duplicated gene
VIT_208s0007g02480.1	cytokinin riboside 5 monophosphate phosphoribohydrolase log7	Rkatsiteli	no	8	yes
VIT_208s0007g02610.1	tld domain containing nucleolar protein	Rkatsiteli	yes	1	no duplicated gene
VIT_208s0007g02710.1	short chain alcohol	Rkatsiteli	no	1	no duplicated gene
VIT_208s0007g05540.1	eukaryotic rpb5 rna polymerase subunit family protein	Cabernet Franc	yes	1	no duplicated gene
VIT_208s0007g05550.1	protein	Cabernet Franc	yes	3	yes
VIT_208s0007g06070.1	NA	Cabernet Franc	no	1	no duplicated gene
VIT_208s0007g06180.1	myb-like hth transcriptional regulator family protein	Cabernet Franc	yes	1	no duplicated gene
VIT_208s0007g06570.1	probable l-type lectin-domain containing receptor kinase -like	Cabernet Franc	no	2	no
VIT_208s0007g06580.1	probable l-type lectin-domain containing receptor kinase -like	Cabernet Franc	no	2	no
VIT_208s0007g08500.1	60s ribosomal export protein nmd3-like	Cabernet Franc	no	1	no duplicated gene
VIT_208s0007g08600.1	NA	Cabernet Franc	yes	1	no duplicated gene
VIT_208s0007g08610.1	NA	Cabernet Franc	yes	1	no duplicated gene

Candidate gene ID	Annotation V2.1	Variety	Expression[1]	n. copies (Blastp)	Expression in dup. genes[2]
VIT_209s0002g08300.1	disease resistance protein rga4-like	Sangiovese	yes	1	no duplicated gene
VIT_209s0002g08560.1	uncharacterized protein	Sangiovese	no	1	no duplicated gene
VIT_209s0002g08800.1	NA	Sangiovese	no	4	yes
VIT_209s0002g08940.1	nuclear pore complex protein nup98-nup96	Sangiovese	yes	1	no duplicated gene
VIT_209s0002g09250.1	cytochrome p450 82c4	Sangiovese	no	6	yes
VIT_209s0070g00150.1	NA	Sangiovese	no	1	no duplicated gene
VIT_209s0070g00460.1	NA	Sangiovese	yes	1	no duplicated gene
VIT_209s0070g00470.1	zinc finger ccch domain-containing protein 16	Sangiovese	yes	1	no duplicated gene
VIT_209s0096g00420.1	probable disease resistance protein at5g63020-like	Sangiovese	yes	5	yes
VIT_209s0096g00490.1	hydroxycinnamoyl-coenzyme a shikimate quinate hydroxycinnamoyltransferase-like	Sangiovese	no	7	yes
VIT_211s0016g00460.1	kinesin-like protein	Primitivo	na	1	no duplicated gene
VIT_211s0016g01610.1	uncharacterized protein	Primitivo	na	1	no duplicated gene
VIT_211s0016g01620.1	ankyrin repeat-containing protein at2g01680-like	Primitivo	na	3	na
VIT_211s0052g00760.1	NA	Sangiovese	no	1	no duplicated gene
VIT_211s0052g00985.1	glutamate-gated kainate-type ion channel receptor subunit 5	Sangiovese	no	2	no
VIT_211s0052g01270.1	probable xyloglucan endotransglucosylase hydrolase protein 23	Sangiovese	yes	20	yes
VIT_212s0059g02490.1	spx and exs domain-containing protein 1-like	Primitivo	na	1	no duplicated gene
VIT_212s0059g02540.1	cysteine-rich receptor-like protein kinase 25-like isoform 3	Primitivo	na	1	no duplicated gene
VIT_212s0134g00270.1	g-type lectin s-receptor-like serine threonine-protein kinase at4g03230-like	Primitivo	na	1	no duplicated gene
VIT_212s0134g00280.1	g-type lectin s-receptor-like serine threonine-protein kinase at4g03230-like	Primitivo	na	2	na
VIT_215s0021g01430.1	aminoacyl-t-rna synthetase	Primitivo	yes	1	no duplicated gene
VIT_215s0045g00110.1	60s ribosomal protein l4 l1	Pinot Noir	no	5	yes
VIT_215s0045g00540.1	protein transport protein sec23	Pinot Noir	yes	1	no duplicated gene
VIT_215s0045g00550.1	uncharacterized protein	Pinot Noir	yes	1	no duplicated gene
VIT_215s0045g01170.1	3-hydroxyisobutyryl- hydrolase 1	Pinot Noir	yes	2	yes
VIT_215s0107g00150.1	glutathione s-	Pinot Noir	no	5	yes
VIT_218s0072g01080.1	dnaj heat shock n-terminal domain-containing	Rkatsiteli	yes	1	no duplicated gene
VIT_218s0075g00130.1	fanconi anemia group m protein	Rkatsiteli	yes	1	no duplicated gene

[1]: expression of the candidate gene in at least one tissue of the variety carrying the putative lethal allele.
[2]: expression of at least one gene among the duplicated genes in the variety carrying the putative lethal allele.

Acknowledgements

I would like to express my gratitude to my supervisor, Professor Michele Morgante, for giving me the opportunity to attend this PhD school, for all the teaching and the knowledge he provided me, for the interesting topics discussed in these three years.

I would like to thank my co-supervisor, Dr. Fabio Marroni, for helping me and supporting me during these three years, for introducing me to the "R world" and for the statistical and scientific support.

Special thanks go to Gabriele Magris and to Gabriele di Gaspero, for their scientific support and their contribution to this work.

I thank my past and present "PhD student colleagues" Aldo Tocci, Ettore Zapparoli, Gabriele Magris, and Mirko Celii, for sharing adventures and misadventures, for being always helpful and supportive, for their friendship.

I thank everyone at the Institute of Applied Genomics (especially Irena Jurman, Federica Magni, Emanuela Aleo and Eleonora di Centa) and at the University of Udine (especially Nicoletta Felice and Giusi Zaina), their technical support and friendship have contributed to the realization of this work.

I thank my family and Alessandro for following me in this path, for the strength and the love the give me every day.

The present work has been supported by the European Commission's European Research Council, within the Seventh Framework Programme for Research.

(Grant number, 294780)