Genetic and epigenetic landscapes of spontaneus genome rearrangements in grapevine

Noršić, Dominik

Master's thesis / Diplomski rad

2022

Degree Grantor / Ustanova koja je dodijelila akademski / stručni stupanj: **University of Zagreb, Faculty of Science / Sveučilište u Zagrebu, Prirodoslovno-matematički fakultet**

Permanent link / Trajna poveznica: https://urn.nsk.hr/urn:nbn:hr:217:976183

University of Zagreb
Faculty of Science
Department of Biology

Dominik Noršić

Genetic and epigenetic landscapes of spontaneous genome rearrangements in grapevine

Master thesis

Zagreb, 2022.

Sveučilište u Zagrebu
Prirodoslovno-matematički fakultet
Biološki odsjek

Dominik Noršić

Genetička i epigenetička obilježja spontanih rearanžmana genoma vinove loze

Diplomski rad

Zagreb, 2022.

Ovaj rad je izrađen na odsjeku za molekularnu biologiju Max Planck Instituta za Razvojnu Biologiju pod voditeljstvom Prof. dr. Detlefa Weigela i suvoditeljstvom Prof. dr. Dunje Leljak-Levanić. Rad je predan na ocjenu Biološkom odsjeku Prirodoslovno-matematičkog fakulteta Sveučilišta u Zagrebu radi stjecanja zvanja magistra molekularne biologije.

ACKNOWLEDGEMENTS

I thank my mentor, Dr. Pablo Carbonell-Bejerano, for all of the help provided during the course of my stay in Tübingen, and for all of the advice and suggestions during the course of my work.

I also thank my supervisor, Prof. Dr. Detlef Weigel, for providing me with the amazing opportunity to intern at the Max Planck Institute for Developmental Biology, to hone my skills as a scientist, and for providing great work environment and conditions.

I especially thank all of my friends, family, and colleges, for the provided support during my studies, always being there for me, and making my student years a wonderful experience.

TEMELJNA DOKUMENTACIJSKA KARTICA

Sveučilište u Zagrebu
Prirodoslovno-matematički fakultet
Biološki odsjek Diplomski rad

Genetička i epigenetička obilježja spontanih rearanžmana genoma vinove loze

Dominik Noršić

Rooseveltov trg 6, 10000 Zagreb, Hrvatska

Kultivari vinove loze (*Vitis vinifera* ssp. *vinifera*), ovisno o boji bobica, mogu se podijeliti na crne, bijele i crvene. Bijelim sortama nedostaju funkcionalne kopije gena *VviMybA1* i *VviMybA2* koji reguliraju nakupljanje antocijana u kožici grožđa, čija razina određuje boju bobica. Fokus ovog istraživanja bio je karakterizirati genetičke i epigenetičke procese koji rezultiraju prirodnim somatskim varijantama boje grožđa. Dva događaja strukturne varijacije genoma prethodno su bioinformatički identificirana kao uzrok fenotipa bijele bobice u sorti 'Garnacha Blanca'. Ova dva događaja strukturne varijacije ovdje su potvrđena metodom lančane reakcije polimeraze i potvrđena su kao vjerojatno jedina dva nezavisna podrijetla za ovu sortu nakon genotipizacije zbirke klonskih akcesija. Uzorci sorte 'Garnacha Tinta' s crnom bobicom također su genotipizirani za iste strukturne varijacije, identificirajući jednu kimeru vjerojatnu mutantnu u sloju stanica kože. Kako bi se okarakteriziralo genetsko podrijetlo mutanata s promijenjenom bojom bobica, testirana je metoda sekvenciranja Oxford Nanopore pomoću programa UNCALLED koja omogućuje selektivno ciljanje i obogaćenje specifičnih regija genoma. Između ostalih, regija u kojoj se nalazi gen *VviMybA1* bila je ciljana u somatskoj varijanti sa povraćenom bojom bobice kultivara 'Alvarinho', potvrđujući da je izrezivanje transpozabilnog elementa *Gret1* iz *VviMybA1* promotora uzrok oporavka boje.

(53 stranica, 21 slika, 2 tablica, 63 literaturnih navoda, jezik izvornika: engleski)
Rad je pohranjen u Središnjoj biološkoj knjižnici
Ključne riječi: strukturna varijacija, *Vitis vinifera*, geni *MYBA*, antocijani, boja bobica grožđa nanopore, transpozabilni elementi
Voditelj: Prof. dr. Detlef Weigel
Suvoditelj: Prof. dr. Dunja Leljak-Levanić

Ocjenitelji:
Prof. dr. Dunja Leljak-Levanić

Rad prihvaćen:

BASIC DOCUMENTATION CARD

University of Zagreb
Faculty of Science
Department of Biology

Master Thesis

Genetic and epigenetic landscapes of spontaneous genome rearrangements in grapevine

Dominik Noršić

Rooseveltov trg 6, 10000 Zagreb, Hrvatska

Grapevine cultivars (*Vitis vinifera* ssp. *vinifera*), depending on the color of their berries, can be divided into black, white and red. White cultivars lack functional copies of the *VviMybA1* and *VviMybA2* genes that regulate the accumulation of anthocyanins in grape skin, whose level determines the color of the berry. The focus of this study was to characterize genetic and epigenetic processes that result in natural grape color somatic variants. Two genome structural variation events were bioinformatically identified previously as the cause of the white berry phenotype in 'Garnacha Blanca' derivative cultivar. These two structural variation events were validated here by polymerase chain reaction and were confirmed as the likely only two independent origins for this cultivar after genotyping a collection of clonal accessions. Samples of the ancestral black-berried 'Garnacha Tinta' cultivar were also genotyped for the same structural variations, identifying one likely epidermal cell monolayer chimera mutant. To characterize the genetic origin of mutants with altered berry color, Oxford Nanopore target enrichment sequencing using UNCALLED method was tested. Among others, the region harbouring *VviMybA1* gene was targeted in a color re-gain somatic variant of the white-berried cultivar Alvarinho, confirming that a *Gret1* transposable element excision from *VviMybA1* promoter is the cause of color recovery.

(53 pages, 21 figures, 2 tables, 63 references, original in: english)
Thesis is deposited in Central Biological Library.
Keywords: structural variation, *Vitis vinifera*, *MYBA* genes, anthocyanins, grapevine berry color, nanopore, transposable elements
Supervisor: Prof. dr. Detlef Weigel
Co-supervisor: Prof. dr. Dunja Leljak-Levanić

Reviewers:
Prof. dr. Dunja Leljak-Levanić

Thesis accepted:

CONTENTS

Abbreviations:

UFGT - UDP glucose flavonoid:3-O-glucosyltransferase

MYBA1 - *VviMYBA1*

MYBA2 - *VviMYBA2*

Gret1 - Grapevine retrotransposon 1

LTR - long terminal repeat

CNV - copy number variant

PAV - presence/absence variant

SV - structural variation

SNP - single nucleotide polymorphism

TE - transposable element

LOH - loss of heterozygosity

ONT - Oxford Nanopore Technologies

UNCALLED - Utility for Nanopore Current ALignment to Large Expanses of DNA

IGV - Integrative Genomics Viewer

1. INTRODUCTION

The Grapevine (*Vitis vinifera* ssp. *vinifera*) is one of the most important and widely grown crops in the world (Myles et al. 2011). It was domesticated at least 7.000 years ago, and its diversity has been expanded to 5,000-10,000 different cultivars (This et al. 2007). Despite the huge number of cultivars to choose from, over 50% of all vineyard surface is covered with only 16 cultivars (Carbonell-Bejerano et al. 2019). This is because viticulturists opt for varieties based on their recognizable desired characteristics, such as the size of the fruits, the reproduction method, the contents of certain compounds, the flavor variety, or the color (This et al. 2007). This leads to a bias towards some of the traditional elite cultivars as they possess desirable traits or cultivars that are simply widely present in their region (Carbonell-Bejerano et al. 2019). During the process of domestication, as traits are being selected, changes accumulate at the genetic level, occasionally giving rise to physiologically different variants (Hancock 1992).

The most drastic change domesticated grapevine went through was the emergence of hermaphroditism. Most of the cultivars used in modern vineyards are offspring from individuals that probably germinated hundreds of years ago (McGovern et al. 1996; Carbonell-Bejerano et al. 2019). Grapevine cultivars are propagated vegetatively to prevent segregation and to allow for selection of superior individuals. In contrast, *Vitis vinifera* ssp. *sylvestris*, the wild ancestor of domesticated grapevines, reproduces sexually through seeds. The dioecy of wild grapevines obligates for out-crossing, which leads to high heterozygosity of individual genomes (Zhou et al. 2019). These reproductive features maintain variability in wild populations and lead to a large number of genetic polymorphisms, giving rise to offspring with a wide variety of characteristics, such as berries with different sizes, shapes, sweetness, juiciness, and colors (McGovern et al. 1996). Progeny with segregating inherited parental traits is not desired in modern vineyards as it is important to conserve harvests of consistent quality and uniformity. Therefore, because domesticated cultivars have inherited the high heterozygosity of their wild ancestors, they have to be vegetatively propagated to keep their genotype and attributes.

1.1. Grapevine berry color

One of the most important grapevine traits is the color of its berries. Grapevine cultivars are commonly divided into three groups based on their berry color, those often being black, red, and white. Those three groups additionally have a wide spectrum of differently pigmented variants (Fig. 1). High color variation within grapevine cultivars is attributed to human selection (This et al. 2007). The wild grapevine ancestor of modern cultivated grape is believed to have had berries with black skin (McGovern et al. 1996). Color diversity greatly expanded its spectrum during domestication and color mutants exist for many of today's cultivars (Galet 2000).

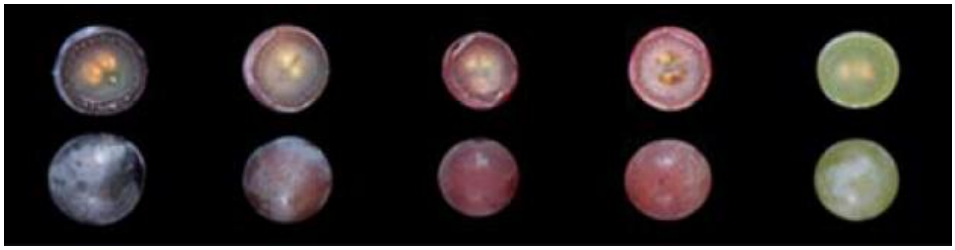


Figure 1. Range of differently colored grapevine (*Vitis vinifera ssp. vinifera*) berries as a result of variation occuring during cultivation, from black (leftmost berry) to white (rightmost berry). Picture was edited and taken from This et al. (2007).

Berry color of a cultivar to be grown is chosen depending on the purpose of cultivation. This choice determines the class of a product to be presented to the market and has had a cultural significance throughout human history. For instance, Ancient Egyptians made black and white wines as they believed those were important for the afterlife (Guasch-Jané et al. 2006). However, no evidence supporting the presence of white grape berries or white wines made from them have been found from these ancient times. Origin of the fruits with white berries is unknown, although they have been mentioned by Pliny the Elder (AD 77) (Walker et al. 2007).

1.1.1. Grapevine color pigment

The main difference between differently colored grapevine berries is the level and composition of the expressed color pigments, anthocyanins, usually in berry skin cell layers (Kobayashi et al. 2001; Walker et al. 2007), but exceptionally also in berry flesh cells of teinturier

cultivars (Röckel et al. 2020). In higher plants, the apical meristems are stratified and are made up of layers of dividing cells. As cell layers develop independently of each other, different plant tissues develop within an organ (Neilson-Jones 1969). The shoot apical meristem of grapevine contains two different layers of cells, with L1 forming the epidermis of all the plants organs and L2 forming most of the internal tissues (Thompson and Olmo 1963; Carbonell-Bejerano et al. 2019). The grape very skin if formed by an epidermal layer derived from meristem L1 and by several subepidermal cell layers derived from L2, while all berry flesh cells derive from the L2. Depending on distribution of anthocyanin in L1 and L2 of the grapevine berry because of different genotype between the layers due to chimeric somatic mutations, it will have a specific coloration (Fig. 2). Exchange of cells between layers, while possible, is a rare occurrence, so each of them keeps the specific expressed phenotype (Carbonell-Bejerano et al. 2019).

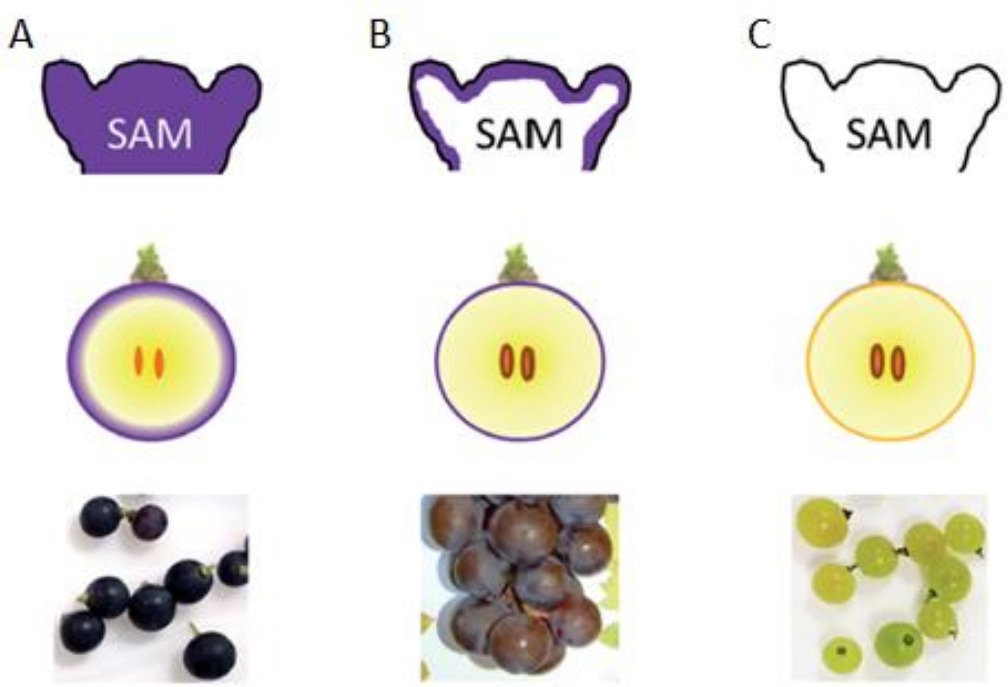


Figure 2. Genetic influence of cell layers of shoot apical meristem (SAM) on berry coloration. Presence of purple color indicates that anthocyanins can be expressed in that layer. Anthocyanins can be expressed in both L1 and L2 (A), only in one (B) or neither (C), resulting in fruits of different color. Image was edited and taken from Carbonell-Bejerano et al. (2019).

1.1.2. Anthocyanin biosynthesis

The crucial step in the anthocyanin biosynthesis pathway was proven to be one of the last ones in anthocyanin biosynthesis pathway, regulated by an enzyme called UDP glucose flavonoid:3-O-glucosyltransferase (UFGT) (Kobayashi et al. 2001), catalyzing the synthesis of red to blue color anthocyanins from colorless anthocyanidin precursors (Ferreira et al. 2018). During the original research performed by Boss et al. (1996), seven genes involved in anthocyanin biosynthesis pathway were analyzed. All of them were expressed in most tissues, except *UFGT*. It was only detected in berry skin, where anthocyanins can be found. In addition, it was observed that *UFGT* expression is nonexistent in all of the white cultivars observed, and the opposite was found for the black and red cultivars (Boss et al. 1996). It was determined, however, that there were no significant differences between *UFGT* genes from black and white grape cultivars (Kobayashi et al. 2001), meaning that the presence of UFGT is regulated on transcription level by a *trans* factor that should be active only in colored genotypes. For the UFGT to be present in a cell, its expression needs to be triggered with two transcription factors, *MYBA1* (*VviMYBA1*) and *MYBA2* (*VviMYBA2*) (Kobayashi et al. 2002). *UFGT* gene is likely induced only by *MYBA* transcription factors, and in a direct manner as they interact with the promotor region of the gene and not trough other factors. (Poudel et al. 2021).

1.1.3. MYBA transcription factors

MYBA1 and *MYBA2* are a part of the berry color locus on chromosome 2 in grapevine (Walker et al. 2007). They are located within a cluster of *MYB* genes, which includes other genes related to anthocyanin biosynthesis control (Fig. 3A) (Matus et al. 2017). This locus is responsible for a wide range of berry color found in grapevine cultivars and that variation is attributed to the combined additive effects of *MYBA* alleles (Fournier-Level et al. 2009). *MYB* genes probably appeared as a result of a rapid amplification, early in higher plant evolution (Rabinowicz et al. 1999). It is suspected that this duplication occurred recently, in evolutionary sense, as there are very few differences in sequence between them, with the purpose of secondary metabolism processes regulation (Walker et al. 2007). While black-berried and red-berried cultivars have at

least one functional copy of *MYBA1* or *MYBA2*, white-berried cultivars lack functional copies and do not accumulate anthocyanins (Kobayashi et al. 2004), consistent with the findings of Boss et al. (1996). At which point in time white cultivars appeared is still unknown, but the events that led to the rise of a white variant have been described.

Absence of anthocyanins in white cultivars are explained by a knock-out mutation of the two *MYBA* genes, which appears in all white grapes (Fig. 3B). Walker et al. (2007) identified a point mutation and a deletion of CA dinucleotide in *MYBA2* coding sequence resulting in inactive gene as the transcription yields a shortened protein due to frameshift (Fig. 3C).

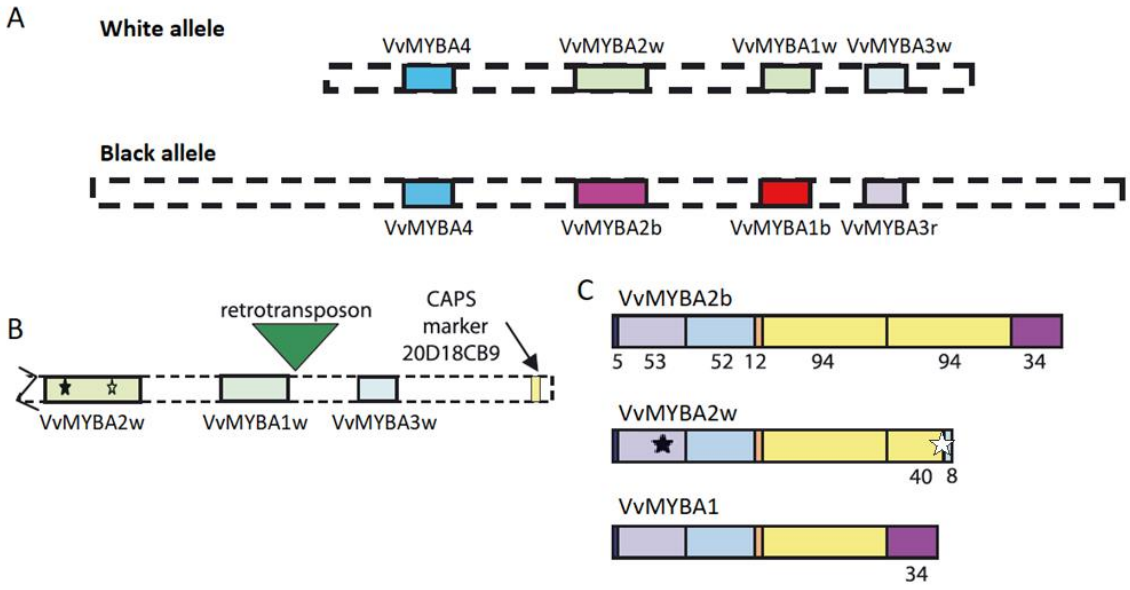


Figure 3. Schematic depicting grapevine (*Vitis vinifera* ssp. *vinifera*) berry color locus, mutation events in its genes and predicted MYBA protein sequences. Not to scale. (A) Relative positions of *MYBA* genes on chromosome 2 for the white and black allele. Differences in color for the same gene indicate a polymorphic sequence. (B) Part of the grapevine berry color locus with indicated mutations that led to the formation of canonical white allele. The stars indicate mutation events leading to the formation of *MYBA2* white variant. Location of the *Gret1* insertion is indicated with a green triangle. (C) Predicted protein sequences for *MYBA1* gene and a black (*VvMYBA2b*) and white (*VvMYBA2w*) variant of *MYBA2*. Boxes of the same color indicate identical sequence. Numbers beneath the boxes indicate the number of amino acids within each. Black star at *VvMYB2w* indicates point mutation and the white star indicates the point of dinucleotide deletion, resulting with a shorter protein. Image was edited and taken from Walker et al. (2007).

What led to the inactivation of *MYBA1* was the insertion of Grapevine retrotransposon 1 (*Gret1*) upstream of the coding sequence, which inhibits its expression, as it was discovered by Kobayashi et al. (2004). *Gret1* additionally can excise itself from the genome, leaving behind a single copy of long terminal repeat (LTR) sequence leading to color recovery in red somatic variants from white-berried cultivars (Kobayashi et al. 2004, Azuma et al. 2009).

All these events led to a formation of a canonical null allele for the grapevine berry color locus. These mutations are present in all the white cultivars observed, suggesting that they had a single origin in a common ancestor. Currently there is no evidence to suggest which mutation occurred first as there is no individual in existence with only one of the *MYBA* genes inactivated except in the case of revertant somatic mutants. The white allele with both *MYBAs* knocked out should have emerged in a heterozygous state. Subsequent segregation of black and white color locus alleles during sexual reproduction, would have generated the first white grapes in individuals homozygous for the white allele as the KO mutations are recessive (Walker et al. 2007).

1.2. Structural variation

Structural variations (SVs), such as copy number variants (CNVs) or presence/absence variants (PAVs), are a genetic difference between individuals, which can result in gene loss, duplications, and the generation of new genes, leading to a variation in phenotype within a species. SVs have been defined as changes in length, copy number, orientation, or chromosomal location of regions of DNA (Escaramis et al. 2015; Yuan et al. 2021). They are frequently the result of mistakes occurring during replication or DNA breaking during mitosis, which are then repaired illegitimately (Carbonell-Bejerano et al. 2019). Polymorphic SVs are most often found in intergenic regions of the plant genome, affecting chromatin loops (Yang et al. 2019), however, they can also be found in proximity of genes, as it is the case with *Gret1* insertion in the regulatory region of *MYBA1* (Kobayashi et al. 2004). There, they can alter the activity of those genes or cause CNVs which can change the number of gene copies within the genome and possibly lead to further functional innovations (Marroni et al. 2014).

For a long time, single nucleotide polymorphisms (SNPs) were thought to be the driving force behind genetic variation. SVs were thought to be a part of "dispensable genome" that is not needed for survival (Morgante et al. 2007), but, in recent years, their central role in shaping the genome landscape became apparent (Feuk et al. 2006; Marroni et al. 2014; Zhou et al. 2019; Vondras et al. 2019). They differ from SNPs as they are considered to span longer parts of the genome and thus, potentially have much bigger effects on the expression of genes and protein function (Chiang et al. 2017). SVs used to be specified as inversions, deletions, duplications, and insertions that spanned at least 1kb of DNA (Feuk et al. 2006), but as technology advanced, many shorter alterations became apparent, so the original definition was modified to incorporate smaller variants as well (Alkan et al. 2011). Eventually obtained capability to generate complex reference genomes, combined with the decreased costs for *de novo* genome assembly, as well as resequencing, have accelerated the study of SVs (Voss-Fels and Snowdon 2016; Gabur et al. 2018) and our ability to define them.

1.2.1. Transposable elements and structural variation

Transposable elements (TEs) are the most prevalent cause for somatic polymorphisms in plants (Mercenaro et al. 2017) and can have a high impact on a phenotype (Carbonell-Bejerano et al. 2017). Their relatively recent activity has been observed in all of angiosperms, so they could presumably be connected to many SVs present (Marroni et al. 2014). SVs can be caused by TE insertion as a result of their unorthodox transposition or homologous recombination, which is often correlated with the repetitive sequences that TEs do possess. Gypsy-like transposable elements, which *Gret1* is a part of, were shown to be most polymorphic, generating insertion polymorphisms in grapevine genomes (Carrier et al. 2012; Mercenaro et al. 2017).

The presence of TEs was found to be highly correlated with CNVs. When annotated, they were proven to overlap with a big percentage of CNVs, as well as insertions/deletions (INDELs) (Mercenaro et al. 2017). TEs are able to cause relatively large INDELS, up to 10 kpb in length, and can further induce recombination events leading to even larger changes in genetic landscape (Lisch 2013; Marroni et al. 2014; Carbonell-Bejerano et al. 2017). Many of TEs have been found

around deleted polymorphic regions, suggesting their important role as deletion mediators (Morgante et al. 2007; Marroni et al. 2014).

Genes that are in close proximity to TEs were shown to have lower expression level on average compared to the rest of the genome. In general, the increase in distance between them was proven to have an increasingly positive effect on the expression (Wang et al. 2013). This could be due to the change in chromatin conformation caused by DNA methylation (Bird 2002) associated with transposable elements as DNA methylation is one of the ways TEs are able to prevent their transposition. Higher rates of methylation further contribute to the rate of mutations, which can be observed between the clones originating from the same initial plant. Exon SVs were shown to be deleterious in higher rate when present in only few individuals, suggesting that mutation accumulate increasingly with every new clone (Vondras et al. 2019).

1.2.2. Grapevine somatic variation

Like it was mentioned previously, grapevine cultivars in modern vineyards are propagated vegetatively as it is desired to preserve some of the key characteristics they possess. During this process, preservation of genotype is not possible in its entirety because of the accumulation of somatic mutations throughout long cycles of vegetative propagation. As the mutations accumulate over time, individuals will increasingly differ from the ancestral seedling that inaugurated the cultivar (McKey et al. 2010; Carbonell-Bejerano et al. 2017). In addition, clonal propagation allows for functional or deleterious variants to hide as heterozygous recessives (McKey et al. 2010).

Mutations an individual to be propagated accumulates in its somatic cells are passed on to its progeny since germline is not fully segregated in plants (Watson et al. 2016). As the mutations accumulate over time, individuals will increasingly differ from the original parent plant over time (McKey et al. 2010; Carbonell-Bejerano et al. 2017).

1.2.3 Somatic structural variation in grapevine berry color diversity

SVs had a big role in the rise of white berry colored grapevine. The most notable event was probably with the aforementioned *Gret1* insertion into the *MYBA1* promotor region, rendering the expression of the gene insufficient to maintain anthocyanin biosynthesis. In combination with the frameshift mutation in *MYBA1*, these two gene variants form a canonical null allele that was found to actually be heterozygous in many of the black-berried grape cultivars (Fournier-Level et al. 2009). Those heterozygous, black-berried cultivars indeed can occasionally generate individuals with white berry color through somatic mutation. Because of that, in some cases, white cultivars do not emerge from segregation of the canonical white allele, but as a result of somatic SV.

White phenotype can be related to loss of heterozygosity (LOH), where a spontaneous somatic deletion event for the functional black allele occurs in a meristem cell line. If the other allele is a null variant, as the plant would be hemizygous for it, the resulting phenotype is a white berry, since anthocyanin biosynthesis cannot take place (Fig. 4) (Walker et al. 2006). The deletion is possibly caused by a recombination event between two sequences flanking the color locus (Schuermann et al. 2005; Walker et al. 2006). It was later shown that the deletion can be caused in a much more complex manner via unbalanced genome rearrangements that emerge during somatic growth (Carbonell-Bejerano et al. 2017). Initially, a resulting individual would be a chimera, as the mutation would happen in only L1 or L2 meristem cell layer. The exchange of cells between layers is possible if some form of tissue damage would occur and the invading cell could take on the role specific for the invaded layer of cells (Kidner et al. 2000; Walker et al. 2006), resulting in a berry that is completely white.

Smaller structural variations for grapevine have been described as the possible cause for the different ripening times between grape individuals (Xu et al. 2016), and some gene families were shown to be affected by duplications, such as already mentioned *MYBA* genes and gene families related to berry size, maturation, and seed formation (Cardone et al. 2016). It was also shown that it is possible for a white-berried cultivar to regain the ability to produce anthocyanins trough SV, resulting in a colored berry mutant phenotype (Azuma et al. 2009).

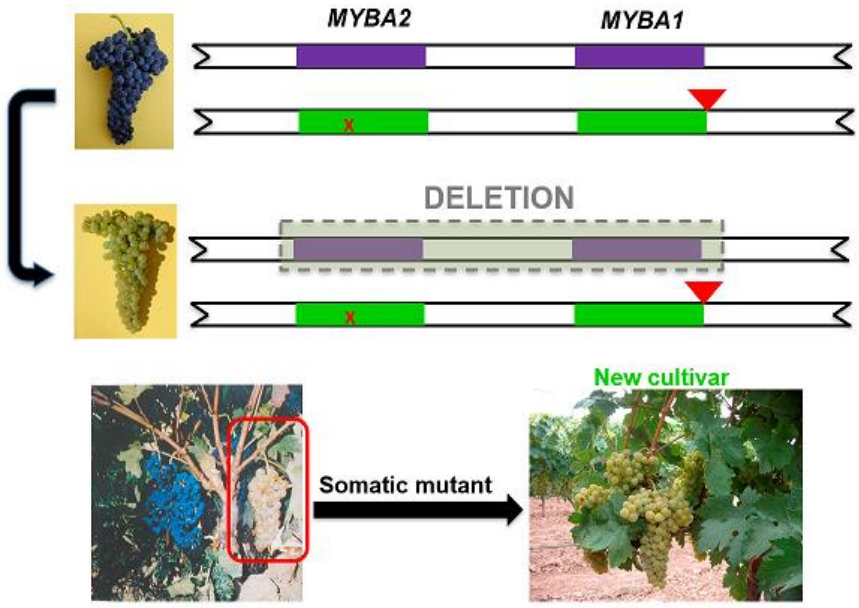


Figure 4. Genetic background for the rise of a somatic mutant trough somatic variation. Purple squares indicate wild type alleles (found in black-berried cultivars), and the green indicate inactivated ones (found in white-berried cultivars). Red X marks the dinucleotide deletion resulting in a shortened *MYBA2* protein and the red triangle indicates *Gret1* insertion. If a plant is heterozygous in the color locus, a deletion of the allele that can express black phenotype will leave the individual with only unfunctional genes specific for the white grape cultivars. Depending on the layer in which the mutation occurs, there will be a change in grapevine berry color in various degrees. Schematic provided by Dr. Pablo Carbonell-Bejerano.

1.3. Nanopore target enrichment sequencing

Second generation sequencing methods, such as Sanger sequencing, greatly improved study of genetic variation, however, sequences gained using these methods were only up to 500 bp long and thus inefficient in solving complex genomic structures, such as repetitive sequences, CNVs or SVs (Magi et al. 2018). To study those, more complex, changes in the genome, methods that enable procurement of longer sequences had to be developed. To that end, Oxford Nanopore Technologies (ONT) has developed a DNA and RNA sequencing technology designed around the use of nanopores. Protein nanopores are embedded into an electro-resistant synthetic

membrane located inside of the flow cell. After the sequencing adapter, that will ligate to the end of molecules, and motor protein that binds to it are added to the DNA/RNA library, it is loaded onto a flow cell, and motor protein leads the molecule to one of the unused nanopores on the membrane (Fig. 5). Ionic current is passing through the pores for the duration of the sequencing process, and as molecules, DNA or RNA, pass through the pore, they cause a disruption in the current (www.nanoporetech.com).

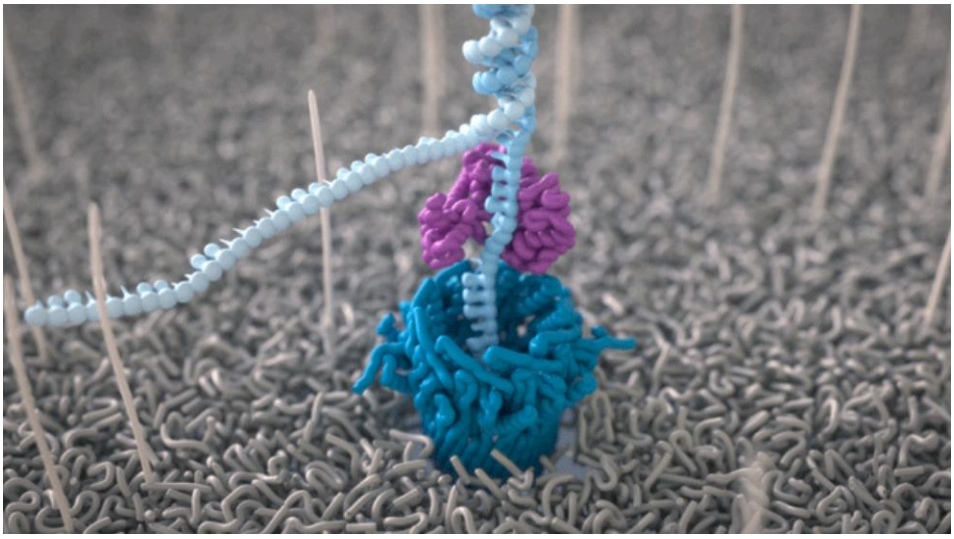


Figure 5. A single protein nanopore (blue) embedded within a synthetic membrane (grey) during the ONT sequencing process. The motor protein (purple) binds the adapter ligated on DNA or RNA molecule and leads it to the pore where the strand is passed through and sequenced. Image was taken from www.nanoporetech.com (14.01.2022.)

Each pore within the flow cell has an electrode connected to a channel, and a sensor, which measures the disruption caused by the molecules. This disruption is recorded as signal that can be further analyzed in real-time, using basecalling algorithms to determine the sequences of the DNA or RNA strands that have passed through. The great advantage that this methodology offers is the ability to directly analyze long fragments of nucleic acids without the need for reassembling the sequence. The only limitation is the length of the molecules themselves. Therefore, this method enables the analysis of structural variants and even repetitive regions.

DNA and RNA are analyzed in their native state, which means that base modifications, such as methylation, remain intact and can also be detected from changes in the electrical signal (www.nanoporetech.com).

One limitation standard ONT sequencing has is the lack of the ability to provide results of higher coverage that are needed in some cases, making targeted sequencing necessary. Some enrichment methods designed to alleviate the problem were developed, however, they required longer preparation and special reagents (Kovaka et al. 2021). Kovaka et al. (2021) developed a new Utility for Nanopore Current ALignment to Large Expanses of DNA (UNCALLED). Purpose of this software is to map a streaming raw signal to a DNA sequence reference in combination with ReadUntil, an option provided by ONT devices that can selectively eject a molecule that is being red during the sequencing process. Polarity of the individual pores can be reversed for an approximately 0.01s to eject the molecule passing through the pore. This enables a new strand of DNA/RNA that could match the target reference to be sequenced sooner, as individual pores became available more rapidly. While sequencing with UNCALLED, reads that are unwanted are identified and ejected using ReadUntil, which leads to the enrichment of the targeted regions (Kovaka et al. 2021).

Target regions are provided as a reference before the sequencing starts. Reads that are being sequenced are compared to the reference sequences in real-time by means of identifying k-mers from the electrical signal, and any reads that do not map to reference are ejected. The cumulative effect is that there are more reads of interest being sequenced compared to the rest of the loaded DNA/RNA library, unlike in the regular nanopore run (Kovaka et al. 2021).

Since it takes some time to compare the signal to the reference, shorter strands may pass through the pore without having a chance of being ejected. Some other reported problems associated with the method are common pore blockage, delayed ejections, and lower yield at the end of sequencing. The software's accuracy also worsens the larger and more repetitive the reference is, as the chance to identify similar sequences increases (Kovaka et al. 2021).

UNCALLED real-time sequencing was tested on human genome, where 148 human genes were enriched with 5.5X overall increase in coverage compared to the control flow cell, where the UNCALLED was not activated, as well as on microbial communities (Kovaka et al. 2021).

1.4. Preliminary studies

Using ONT whole-genome sequencing, previous work at Max Planck Institute for Developmental Biology noted three new SVs composing two SV patterns in the grapevine berry color locus, while studying samples of 'Garnacha Blanca' (GB), a cultivar originating from somatic mutation of the black-berried cultivar 'Garnacha Tinta' (GB). These SVs were dubbed DEL, INV1 and INV2. DEL, standing for deletion, was characterized as an excision event for the continuous segment of DNA, spanning 1,445,748 bp (chr02:15,479,868-16,925,616 in the Garnacha genome primary *de novo* assembly) (Fig. 6A). INV1 and INV2, standing for inversion 1 and inversion 2, were characterized as a single SV pattern, consisting of two discontinuous segments. Two inversion SVs were noted, of 826,345 bp for the INV1 (chr02:15,552,954-16,379,299 in the Garnacha genome assembly), while for the INV2, 933,456 bp (chr02:16,034,813-16,968,269 in the Garnacha genome assembly). The events leading to the origin of INV SV pattern were proposed. An inversion of 1,415,315 bp (chr02:15,552,954-16,968,269) occurred first, with a subsequent deletion of approximately 1 Mb, leaving only one segment of DNA between the breakpoints (Fig. 6B).

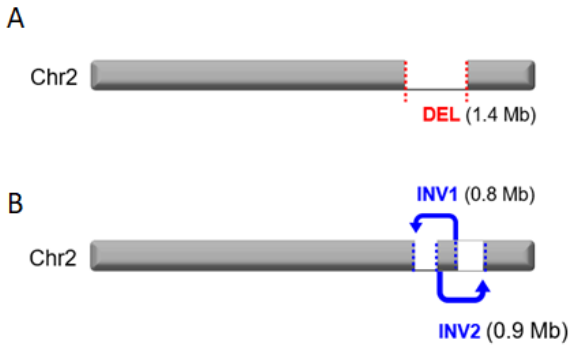


Figure 6. Chromosome 2 schematic depicting the breakpoints of structural variations after they happened and the sequence that remained after deletions. (A) Chromosome 2 after the deletion event. (B) Chromosome 2 after the inversion and subsequent deletion events. Schematic provided by Dr. Pablo Carbonell-Bejerano.

‘Garnacha Tinta’ cultivar is heterozygous for the grape color locus at chromosome 2. One of the alleles for the color locus they have corresponds to the white null allele, while the other one has functional *MYBA* genes. What led to the appearance of white phenotype in ‘Garnacha Blanca’ somatic variants was loss of heterozygosity. Deleted regions associated to DEL and INV1+INV2 SV mutations spanned the region of the genome where *MYBA* genes are located. As they were excised, only unfunctional copies of *MYBA1* and *MYBA2* genes remained, and thus the grape berry coloration was lost. Independent appearance for the INV1+INV2 variation was recorded for white variants, which is not the case for DEL variants. This opens a possibility that all individuals possessing DEL mutation all had a common ancestor in which this single somatic mutation event occured.

A sample of Garnacha cultivar, GT77, was sequenced for the whole genome with Illumina short-reads. What was seen from the obtained sequences is that even though it expresses a black phenotype (Fig. 7), a decrease in coverage is apparent at locations of breakpoint sites characteristic for INV mutants, meaning that it possibly has a predisposition for the appearance of a white mutant if it is a chimeric plant. Indeed, a bud sport (plant part morphologicaly different from the rest of the plant) of GT77, GB77, has white berries, supporting this hypothesis. Some other black-berried individuals may harbor newly detected SV patterns that go unseen in the phenotype. GB77 was additionally sequenced with nanopore.

Figure 7. Garnacha cultivar GT77 berry cluster. Berries are noticeably black, even though the individual possesses an inversion structural variation pattern as it was discovered from illumine short-read sequencing. Picture provided by Dr. Pablo Carbonell-Bejerano.

Single nucleotide polymorphism (SNP) LOH analysis along chromosome 2 was conducted for a collection of 'Garnacha Blanca' accessions. Using this SNPs data, presence of three new noted SVs (DEL, INV1, INV2) was predicted for each sample in the collection.

Spontaneous appearance of 'Red Alvarinho', a red-berried revertant somatic variant was recorded from the ancestral white-berried 'Alvarinho' cultivar (Fig. 8), and it was subjugated to the RNA sequencing analysis (unpublished work from Grape and Wine Research Institute (ICVV, La Rioja, Spain)). Differences in individual gene expression were identified, compared to the white-berried ancestor, notably the ones involved in flavonoid biosynthesis pathway. For the 'Red Alvarinho', considering the over-expression of *MYBA1* detected in the RNA-sequencing, it was hypothesized that the cause of an intense red color reappearing was a recombination event between LTRs of *Gret1* causing it to excise itself from the genome and leading to a partial recovery of *MYBA1* gene expression.

Figure 8. Difference of coloration between White (left) and Red (right) 'Alvarinho' grapes after 'Red Alvarinho' had its ability for anthocyanin production regained, caused by a supposed *Gret1* excision from *MYBA1* promotor. Picture provided by Dr. Pablo Carbonell-Bejerano.

2. GOALS

The general goal of this research is to characterize the genomic landscape of somatic variation for fruit color emerging during the vegetative propagation of grapevine cultivars. To that purpose, the following specific aims were established:

1. To validate SV events as the presumable origin of ‘Garnacha Blanca’ cultivar (GB) and to gain insight into how many independent events there may be in the origin of GB.
2. To assess the hypothesis that new white-berried GB clonal lines may emerge from black-berried ‘Garnacha Tinta’ (GT) plants chimeric for the presence of causal SVs.
3. To determine if SVs in the origin of GB are deleterious for carrier gametes or if these alleles could lead to new white-berried cultivars through sexual reproduction.
4. To assess the hypothesis that a transposable element (TE) movement is in the origin of a ‘Red Alvarinho’ color recovery somatic variant. The potential of a new real-time target enrichment method for ONT sequencing to that aim will be tested.

3. MATERIALS AND METHODS

3.1. Materials

3.1.1. Equipment

Commercial kits:

DNeasy Plant Mini Kit (QIAGEN, Hilden, Germany); Short Read Eliminator Kit (<25kb) (Circulomics, Baltimore, Maryland, USA); ONT Ligation Sequencing Kits SQK – LSK110 and LSK109 (Oxford Nanopore Technologies, Oxford, UK); ONT Wash Kit (EXP-WSH004) (Oxford Nanopore Technologies, Oxford, UK)

Buffers:

Q5 Reaction Buffer Pack (New England Biolabs, Ipswich, Massachusetts, USA), 1X Tris-acetate-EDTA (TAE) buffer

Enzymes:

Q5 High-Fidelity DNA Polymerase (New England Biolabs, Ipswich, Massachusetts, USA)

Dyes:

Orange G 10X Loading dye (0.1 g orange G + 40 mL Glycerol)

Other:

Generuler 10kb DNA Ladder Mix (Fermentas, Waltham, Massachusetts, USA)

Devices:

PCR thermal cycler, gel electrophoresis apparatus, Nanodrop Spectrophotometer, Qubit Fluorometer, MinION sequencing device

Software:

Primer designing tool from NCBI, UNCALLED, Guppy v5, Minimap2, Medaka, Megalodon, Integrative genomics viewer

3.1.2. DNA samples

Grapevine (*Vitis vinifera* ssp. *vinifera*) isolated DNA samples of a collection of clonal accessions of white-berried somatic variant 'Garnacha Blanca' (GB) were provided by the Institute for Grape and Wine Sciences (ICVV, La Rioja, Spain). Clonal GB accessions were collected mainly in the Ebro Valley region of Spain. Most of which come from plants found in the collection of clones and biotypes of EVENA (Navarre Government, Spain), corresponding to samples from old vineyards in Navarre and Álava provinces, or in the ICVV germplasm bank at Finca de la Grajera (samples from La Rioja). From 74 'Garnacha' samples, 5 of them are black-berried 'Garnacha Tinta' (GT) of different origins, 67 samples were white-berried 'Garnacha Blanca' (GB) from La Rioja, Navarre, Álava, Catalonia and France and two samples of red-berried Garnacha Gris from Navarre. Concentrations and purity of the DNAs was measured using Nanodrop and Qubit devices and 44 samples were used for SV genotyping, that had adequate quality (Supplementary table 1). Additional self-cross S1 progeny samples of GB77 and GB78 accessions of GB were also provided from ICVV. For the SV segregation analysis, ten samples each were used of GB77 and GB78 offspring (Supplementary table 2). All of the mentioned DNA samples were stored at -20^{o}C.

Leaf samples of clonal accessions of black-berried 'Garnacha Tinta' ancestor cultivar that were used to search for possible ancestral chimera SVs, were provided from EVENA and ICVV collections, most of them belonging to accessions collected in La Rioja and Navarre regions of Spain, but also from other countries (Supplementary table 3) and their DNAs were extracted. Additionally, samples of 'Red Verdejo', whose DNA was extracted alongside wild-type 'Garnacha' samples, as well as 'Red Alvarinho', whose DNA was sent from ICVV, were used in genotypeing as well.

The grapevine DNA samples used for ONT sequencing target enrichment were provided by ICVV. GB78 sample was taken from GB sample collection, described in section 3.2.1. (Supplementary table 1). Red and White 'Alvarinho' DNA samples, both extracted from leaves, were obtained from ICVV (Supplementary table 5).

3.2. Methods

3.2.1. SV validation

PCR primers were designed using Primer designing tool from NCBI (https://www.ncbi.nlm.nih.gov/) to validate SVs detected by preliminary studies (Supplementary table 4) using 'Garnacha' *de novo* genome assembly (assembly built from GT1608 accession at Max Planck Institute for Biology, unpublished data) as a reference genome. One set of primers was designed to amplify any grapevine sample DNA (wt primers), regardless of genotype relative to the supposed SVs, and thus serve as a positive control, and one primer pair for each mutation (DEL, INV1 and INV2) was designed to amplify DNA only if they are present in a sample. Reverse DEL primer was designed to anneal at the position of the breakpoint, as no specific primers could be found (Table 1, Fig. 9). Sequences for the forward primer used in wt and DEL primer pairs (G-47678), as well as reverse primer for the wt primer pair (G-47686) were provided by Dr. Pablo Carbonell-Bejerano, as they are known to work for GT and GB samples and were used in preliminary research.

Mixtures to load for PCRs were prepared by mixing 1 µL of 1:10 dilution of individual DNA samples with the master mix (per sample: 5 µL Q5 reaction buffer, 5 µL high GC enhancer, 2.5 µL dNTPs, 1.25 µL of each primer in a pair, 0.25 µL Q5 Polymerase). For the PCRs, conditions for each primer pair were used as described in Table 1.

Table 1. List of primer pairs used to amplify DNA containing DEL, INV1 and INV2 mutations and any grapevine DNA (wt), with the conditions for each individual PCR, as well as the expected amplification product length.

MUTATION	ALIAS	PRODUCT LENGTH (bp)	Tm (°C)	ELONGATION (sec)	CYCLES (#)	PCR mix Volume (µL)
wt	Fw:G-47678 Rev:G-47686	731	56	90	32	25
DEL	Fw:G-47678 Rev:G-47702	477	61	60	34	25
INV1	Fw:G-47666 Rev:G-47667	433	57	60	32	25
INV2	Fw:G-47682 Rev:G-47683	566	54.5	60	40	25

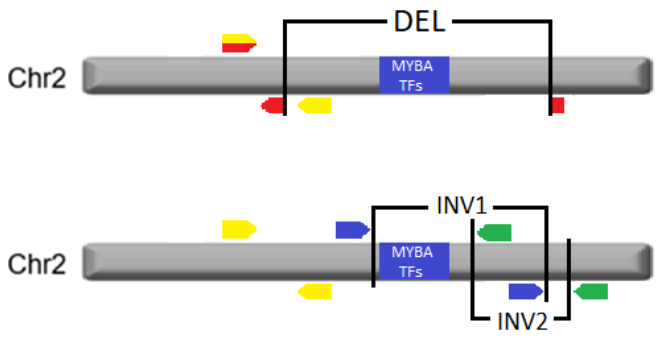


Figure 9. Schematic of relative positions for designed primers, *MYBA* transcription factor genes and breakpoints sites for all three SV breakpoints. Primers depicted above the correspond to forward primers, while the ones depicted bellow the chromosome correspond to reverse primers. Not to scale. ● – wt primers, ● – DEL primers, ● – INV1 primers, ● – INV2 primers

After PCR amplification, amplicon presence and size were checked using electrophoresis. 1% agarose gel was made with 1g of agarose powder, 100 mL TAE buffer and 3µL of ethidium bromide (EtBr). Loading mixture was prepared by mixing 1 µL of amplified DNA with 8 µL of autoclaved dH_2O and 1 µL orange G. Generuler 10kb DNA Ladder Mix was used as a ladder (Fig. 6E). Electrophoresis was run at 120V for 20 min. Finished gel was observed under UV light.

3.2.2. GB collection genotyping

1:10 dilutions were made for Garnacha DNA samples that had concentrations higher than 100 ng/µL. Samples selected for genotyping were prepared for PCR and electrophoresis, using the same procedure as described in section 3.2.2.

3.2.3. Testing wild-type samples for possible chimeras

Wild-type Garnacha samples were prepared for PCR and electrophoresis, using the same procedure as described in section 3.2.2. Only primers designed for INV1 mutation were used. DNAs from younger leaves were extracted using DNeasy Plant Mini Kit from QIAGEN. To that end, frozen plant material was ground to a fine powder under liquid nitrogen, using mortar and pestle.

The required amount of tissue powder was estimated the by filling the eppendorf tube (1.5 mL) half-way. Plant material was kept frozen under liquid nitrogen at all times between steps, as well as all of the dishes used. To isolate DNA from the obtained powder, DNeasy Plant Mini Kit manual from QIAGEN was followed. DNA samples were stored at 4^{o}C.

3.2.4. Structural variation segregation analysis

GB77 and GB78 S1 self-cross progeny DNA samples were diluted 10 times. They were prepared for PCR and electrophoresis, using the same procedure as described in section 3.2.2. For GB77, primers for INV1 were used, and for GB78, DEL primer pair.

3.2.5. ONT sequencing target enrichment

Files to be used when running UNCALLED were prepared according to the provided instructions on the UNCALLED GitHub page (https://github.com/skovaka/UNCALLED). Target sequence reference file had to be prepared for the UNCALLED command to be run. Garnacha genome assembly (unpublished genome assembly, produced with Hifiasm tool after sequencing of Garnacha gDNA with PacBio HiFi, was used as a reference genome for GB78, while for the 'Alvarinho' samples PN40024 12Xv2 reference genome (https://urgi.versailles.inra.fr) was used (Jaillon et al. 2007; Canaguier et al. 2017). A list of 47 genes was made, mostly related to the anthocyanin biosynthesis pathway. A corresponding file containing their sequences was made by extracting them from the corresponding genome assembly using Linux command line, to be provided as target sequences to be enriched during sequencing (Supplementary table 6).

In case of GB78, DEL, INV1 and INV2 breakpoint sites were chosen as targets as well. To start, masking was performed using UNCALLED dedicated scripts available from their GitHub website (https://github.com/skovaka/UNCALLED), as it is recommended by the authors for eukaryotic sequences. As inputs, fasta files of the corresponding reference genome, depending on the sample, was provided, as well as the fasta file containing all of the target sequences. For GB78, 10 kb of sequence upstream and downstream of target genes was provided, and the 30 kb

of upstream and downstream sequence from DEL, INV1 and INV2 breakpoint sites. For 'Alvarinho' samples, length of sequence surrounding the target genes was increased to 30 kb upstream and downstream, while for the *MYBA1* region, provided sequence was 80 kb upstream and downstream of the gene.

For each sample, size selection was performed by following the Short Read Eliminator Kit (<25kb) handbook (v2.0) from Circulomics. The following change was made, after adding 51 µL of Buffer EB, the tube was not incubated at room temperature for 20 minutes. Instead, as it was suggested for longer DNAs in the protocol, the samples were incubated at 50^{o}C, and additionally they were incubated for 30 minutes. Then, Genomic DNA by Ligation (SQK-LSK110 for 'Alvarinho samples, SQK-LSK110 for GB78) library preparation protocol provided by ONT was followed, as to prepare samples for loading into a flow cell.

Flow cell was inserted into the minION device after DNA was loaded as described in the protocol. The sequencing process was started and the UNCALLED command was run in Linux command line: "uncalled realtime <masked reference> --port 8000 -t 16 --enrich -c 3 > <output name>.PAF", where <masked reference> was substituted with the name of the file created as described in section 3.3.2. for each individual sample, that was sequenced at the time. Number of threads was set to 16 ("-t"), argument "–enrich" was set as it enables to keep reads that map to the reference and the decision time of 3 seconds was given ("-c") to check if the individual reads being sequenced map to a reference or not. Finally, generated mapping summary file was saved under the name <output name>.PAF. Sequencing process was stopped whenever number of sequencing pores sequencing was very low, approximately 24 hours. 400 ng of GB78 DNA library and 330 ng of 'Red Alvarinho' DNA library were loaded once. 600 ng of 'White Alvarinho' DNA library was loaded, and Wash Kit was used after the first run as described in the corresponding protocol from ONT, to prepare the flow cell for a second loading. 600 ng of DNA library was loaded again, and sequencing started one more time.

Data processing was done in Linux command line. After the raw sequencing data files (fast5) were created from sequencing MinKNOW software, basecalling with Guppy v5 was performed (https://github.com/LernerLab/GuPPy) to generate fastq files of nucleotide sequence.

They were mapped to a corresponding reference (GB78 to a Garnacha genome assembly, 'Alvarinho' samples to 12Xv2 PN40024 reference genome) with Minimap2 (https://github.com/lh3/minimap2), producing bam files, and afterwards, vcf files from variant calling of the bam files comparing to the corresponding reference genome assembly were generated with Medaka (medaka_variant command), (https://github.com/nanoporetech/medaka) as they were needed to run the Megalodon program (https://github.com/nanoporetech/megalodon). Megalodon was run using the original fast5 files obtained from sequencing, vcf files created with Medaka and the corresponding reference genome assembly as inputs. As an output a final, bam file, was generated by Megalodon alongside with cytosine-methylation labelled reads, and they were indexed using samtools index.

Generated files were used to visualize sequenced reads against the corresponding reference genome in Integrative Genomics Viewer (IGV) program. While running Megalodon, an option for methylation calling was activated for the Red and White 'Alvarinho' samples. Bam file with methylation data was generated and observed in IGV. A second version of Megalodon was run for 'Red Alvarinho', where the 12Xv2 PN40024 reference modified by Dr. Pablo Carbonell-Bejerano was used. It was modified by removing 8762 bp of Gret1 TE sequence from the promoter of *MYBA1*, leaving the presumably remaining LTR region sequence. The rest of the process was done the same as in all other cases.

4. RESULTS

4.1. Structural variation validation

All three structural variation (SV) breakpoints bioinformatically previously identified form ONT reads were confirmed by PCR amplification to be present in the white mutants of Garnacha cultivar (Fig. 10). All DNAs used were amplified when using wt positive control primer pair, showing that DNA is suitable for PCR amplification. When analyzed with DEL primers, a clear signal was only detected for GB1662 sample, which is indeed the GB accession in which the DEL SV was identified by ONT whole-genome sequencing. When using primers designed to amplify inversion mutations, GB55, GB77 and GT77 DNA samples amplified in both cases, while GT1608 (wt) and GB1662 did not, confirming again previous ONT sequencing results.

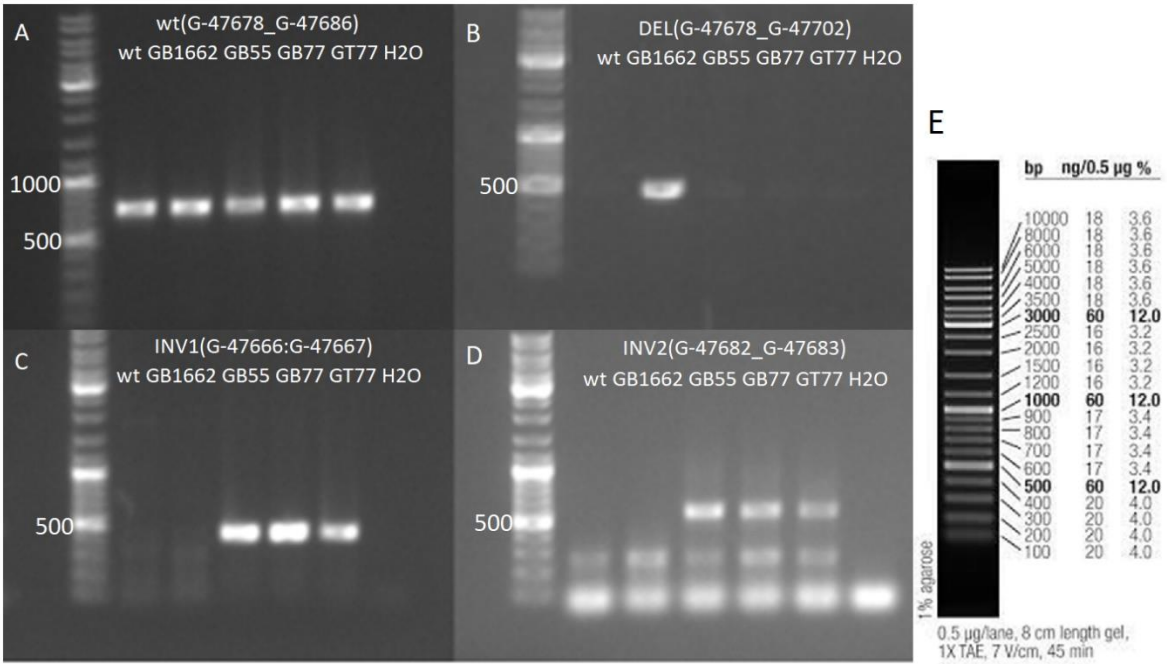


Figure 10. PCR check for working DNA samples (A). All DNA samples show a band when used in combination with wild-type primers. Validations of the proposed DEL (B), INV1 (C) and INV2 (D) breakpoint sites. All samples show bands as predicted by nanopore sequencing (performed by Dr. Pablo Carbonell) for each primer pair. (E) Generuler 10kb DNA Ladder used for electrophoresis. Sample numbers were asigned randomly and have no biological meaning. wt – GT1608, GB – Garnacha Blanca, GT – Garnacha Tinta

Primers used proved to be unspecific, as they produced unwanted bands in the gel in all three cases. Unwanted amplicons for INV2 were much shorter than the desired PCR product and can easily be distinguished. When testing for DEL and INV1, there were unwanted signals roughly the same length of the desired product, however, the signal is much weaker in those cases.

4.2. Garnacha Blanca sample collection genotyping

To gain insight into how many independent mutation events there may be in the origin of 'Garnacha Blanca' cultivar, genotyping of newly detected breakpoints in a collection of clonal accessions of GB was conducted. Genotyping results for the individual Garnacha samples are displayed in Table 2. Most of the samples displayed the SV pattern expected from previous LOH genotyping.

Table 2. Genotyping results for individual garnacha samples tested with DEL, INV1 and INV2 primer combinations. Sample numbers were asigned randomly and have no biological meaning. 🔴 – negative for a mutation, 🟢 – positive for a mutation, 🟡 – unclear results. GB – Garnacha Blanca, R - root

SAMPLE	DEL	INV1	INV2	SAMPLE	DEL	INV1	INV2
GB1-2	🔴	🟢	🟢	GB17-2	🔴	🟢	🟢
GB4-2	🔴	🟢	🟢	GB31-1	🟡	🟢	🟢
GB11-1	🟡	🟢	🟢	GB50-1	🔴	🟢	🟢
GB25-2	🔴	🟢	🟢	GB54-2	🟡	🟢	🟢
GB27-1	🔴	🟢	🟢	GB56-2	🔴	🟢	🟢
GB29-2	🔴	🟢	🟢	GB61-2	🔴	🟢	🟢
GB30-2	🔴	🟢	🟢	GB63-2	🟢	🔴	🔴
GB32-2	🔴	🟢	🟢	GB76-1	🔴	🟢	🟢
GB57-2	🔴	🟢	🟢	GB1608	🔴	🟢	🟢
GB34-2	🔴	🟢	🟢	GB1662	🟢	🔴	🔴
GB35-1	🔴	🟢	🟢	GB1657	🟢	🔴	🔴
GB45-2	🔴	🟢	🟢	GB428	🟢	🔴	🔴
GB49-2	🔴	🟢	🟢	GB177	🔴	🔴	🔴
GB55-2	🔴	🟢	🟢	GB75-1	🔴	🟢	🟢
GB60-1	🔴	🟢	🟢	GB51-1	🔴	🟢	🟢
GB64-1	🟢	🔴	🔴	GB84-2	🔴	🟢	🟢
GB73-2	🔴	🟢	🟢	GB5-2	🟡	🟢	🟢
GB80-1	🟢	🔴	🔴	GB36-1	🔴	🟢	🟢
GB77-2	🔴	🟢	🟢	GB71	🟢	🔴	🔴
GB53-2	🔴	🟢	🟢	GB78-2	🟢	🔴	🔴
GB6-1	🔴	🟢	🟢	GB3-1	🔴	🟢	🟢
GB15-2	🔴	🟢	🟢	GB84-2R	🔴	🟢	🟢

Seven samples were positive for DEL genotype, INV1 and INV2 mutations were always occurring together, and 32 samples were shown to have them. For four samples the results were unclear, and one was proven to be negative when genotyped for any of the SVs (Fig. 11). With unclear results, samples GB11-1, GB31-1, GB54-2 and GB5-2, when tested for INV1 and INV2 mutations, their DNAs amplified in both cases. When tested for DEL mutation, it was unclear whether the specific amplification occurred or not. GB177 did not have its DNA amplify in any of the cases.

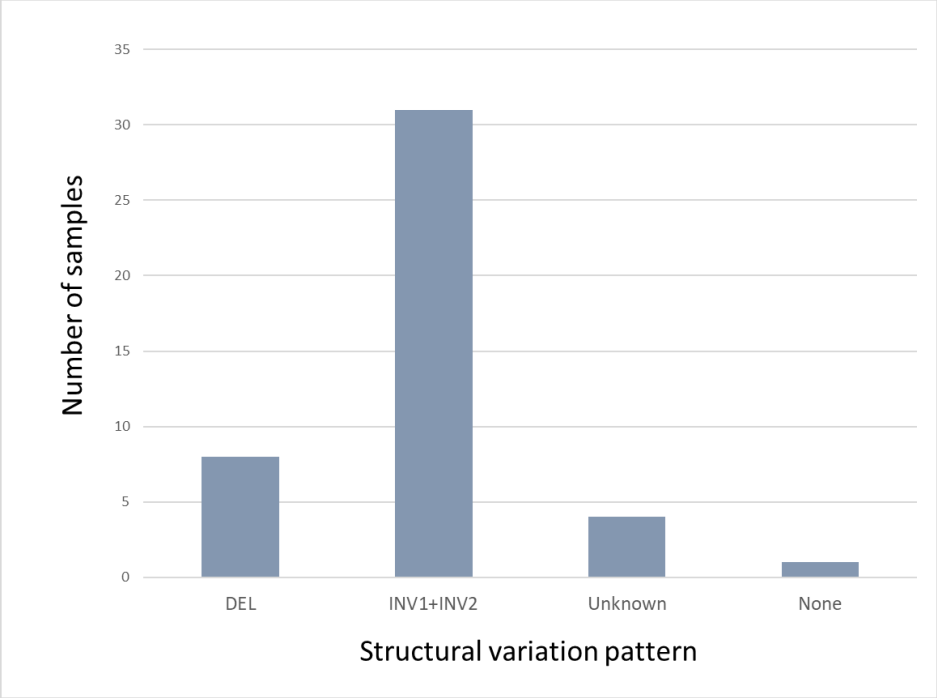


Figure 11. Number of samples possessing the specific individual structural variation pattern (DEL, INV1+INV2) out of 44 total samples. Made according to table 2.

4.3. Testing 'Garnacha Tinta' samples for possible chimeras

To assess the hypothesis that independent new white-berried GB clonal lines may emerge from ancestor black-berried 'Garnacha Tinta' cultivar, 17 GT samples were genotyped for INV1 SV. Except for GT77, none of tested black-berried GT samples show a band in the gel after PCR amplification of INV1 breakpoint (Fig. 12). Only faint, unspecific bands can be seen in the gel for all samples other than GB77. GB55 was confirmed as a positive control in which IN1 breakpoint was identified from ONT sequencing and PCR validation.

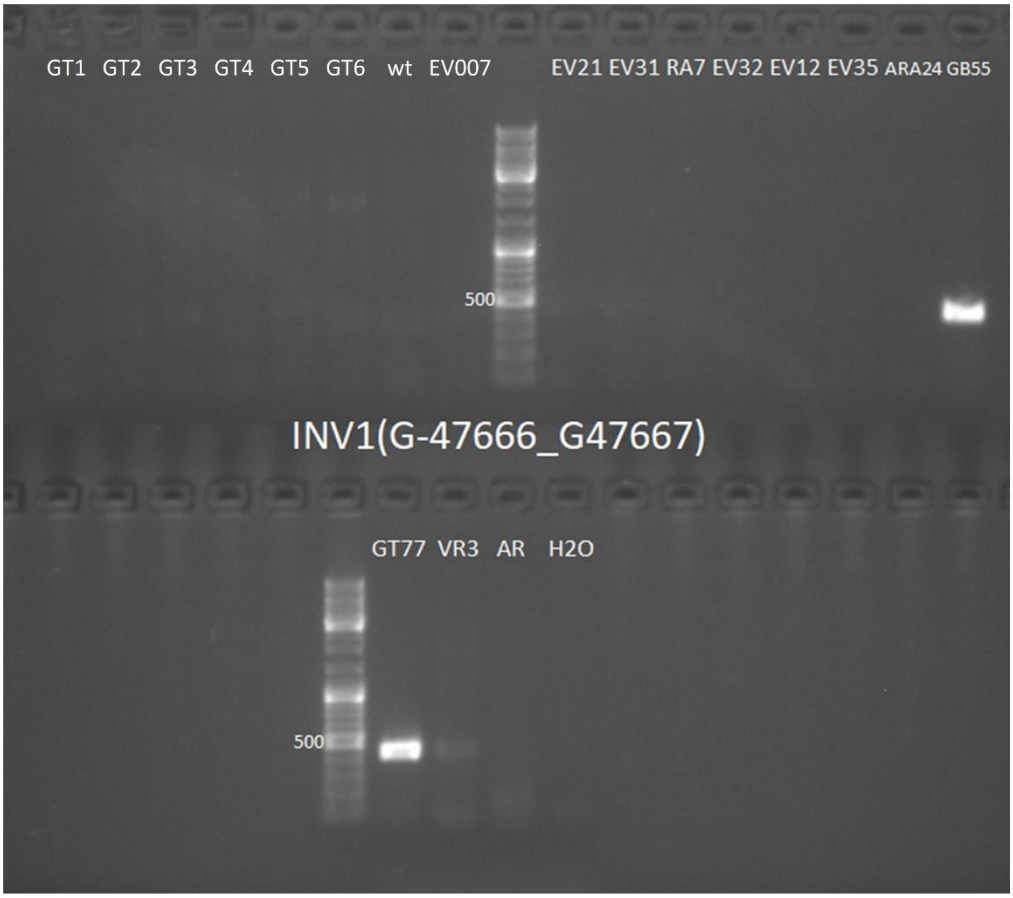


Figure 12. Collection of black-berried phenotype samples of 'Garnacha Tinta' cultivar, as well as 'Red Alvarinho' and 'Verdejo Rubi' cultivar samples, tested for a chimeric genotype. None of the samples carry the INV1 mutation, as their DNA did not specficly amplify during PCR. Sample numbers were asigned randomly and have no biological meaning. wt – GT1608, GB55 and GT77 – positive control, GT – 'Garnacha Tinta', EV – EVENA, VR – 'Verdejo Rubi' (Red Verdejo), AR – 'Albarino Rubi' (Red Alvarinho)

4.4. Structural variation segregation analysis

To determine if SVs in the origin of GB are deleterious for the presumably hemizygous gametes and to see if these mutations can segregate to produce novel alleles to give rise to new white-berried cultivars through sexual reproduction, segregation of DEL and INV1 SV breakpoints was analyzed in offspring of GB78 (Fig. 13A) and GB77 (Fig.13B) accessions, respectively. 70% of GB78 S1 self-cross samples carry the DEL breakpoint, while for GB77 90% of the S1 self-cross samples carry the INV1 mutation.

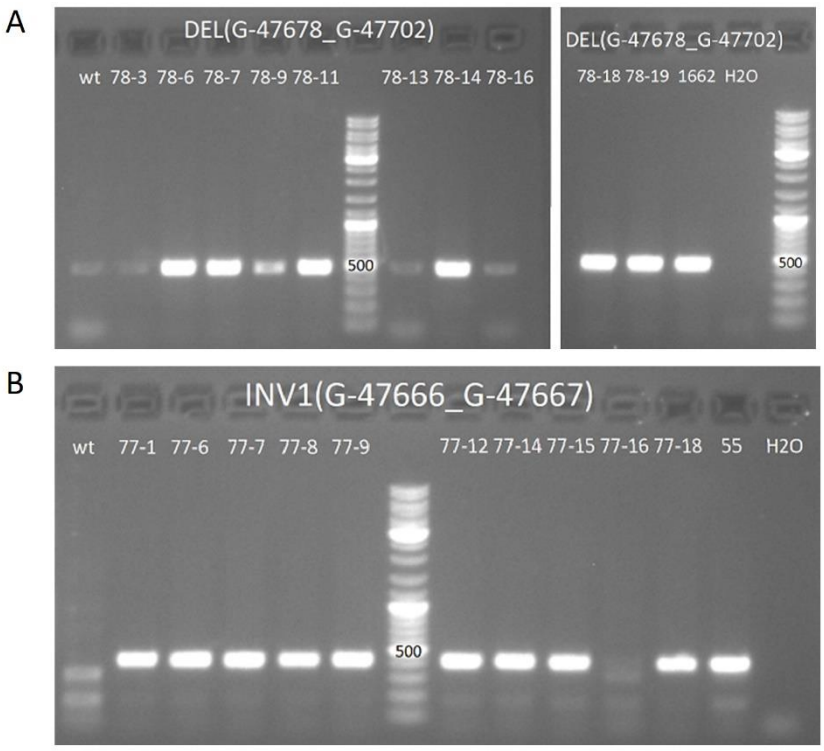


Figure 13. (A) Genotipization of randomly selected GB78 (DEL mutant) self-cross progeny. 30% of GB78 offspring individuals did not inherit DEL mutation, as GB78-3/13/16 did not have their DNA amplify. (B) Genotipization of randomly selected GB77 (INV mutant) self-cross progeny. One out of 10 offspring of GB77, GB77-16, did not inherit INV mutation. All of the samples used belong to the GB cultivar, except for the wt. Sample numbers were asigned randomly and have no biological meaning. wt – GT1608, GB55 – positive control

4.5. Nanopore target enrichment

4.5.1. Method Troubleshooting

To characterize SV in a white mutant of 'Garnacha', GB78, a new real-time target enrichment method for ONT sequencing, using UNCALLED, was tested. When initiating a real-time target enrichment ONT sequencing run, after the UNCALLED software was turned on, there were fewer nanopores sequencing at any given time and that drop can be noticed as soon as the software is turned on (Fig. 14,15). After 7 minutes of sequencing the UNCALLED command was run and a substancial drop in the amount of reads being sequenced at same time can be seen. It fell from 60% pores sequencing to aproximatley 30% in just 2 minutes and an imediate drop in overall number of available, healthy pores can be seen, as indicated by the overall decreese in green color and a coresponding increese in blue color, indicating damaged/blocked pores, on the

graph of sequencing data UI (Fig. 14). The number of pores sequencing is decreased drasticly after 12 hours, however, there was no substencial decrese in the percentage of healthy pores after the four hour mark and it remained mostly consistent trough the rest of the sequencing run (Fig. 16).

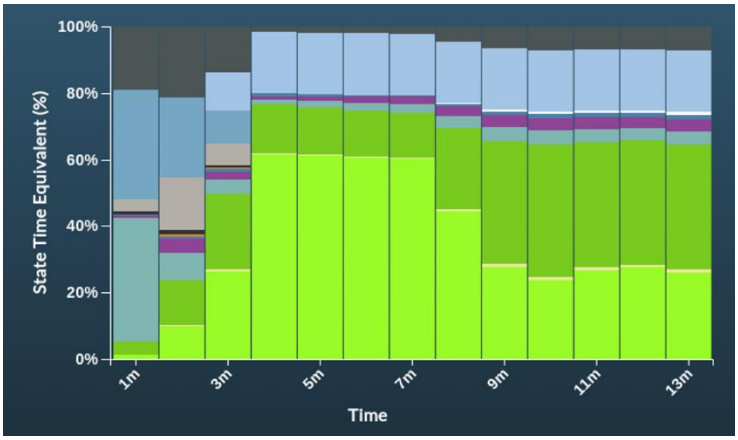


Figure 14. Screenshot of ONT sequencing software sequencing data for the first 14 minutes of sequencing. Light green indicates the percentage of pores that are currently reading the strand, darker green indicates percentage of pores that are waiting for a strand to be introduced to the pore and blue indicates the percentage of pores that are damaged. The UNCALLED software was turned on after seven minutes. Overall decrease in the number of healthy pores and pores that are sequencing can be seen immediately after the seven-minute mark.

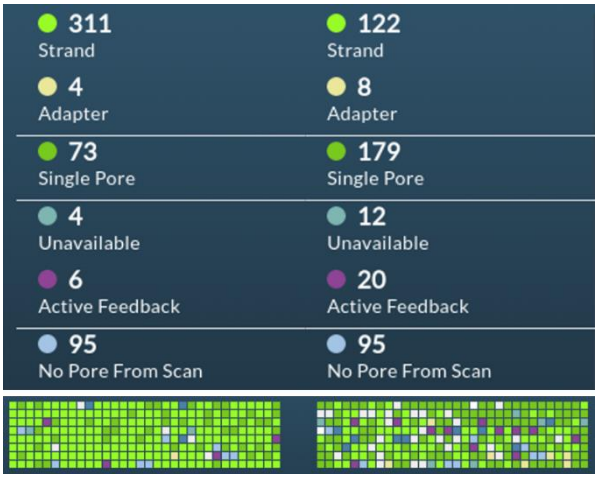


Figure 15. Screenshot of channel state panel from ONT sequencing software during the sequencing run. States of individual pores can be seen before the UNCALLED command is turned on (left side) and after it is turned on (right side). Light green indicates pores that are currently reading the strand, darker green indicates pores that are waiting for a strand to be introduced to the pore and blue indicates pores that are damaged. After the command was run, the number of pores available for sequencing increased by ~2.5X.

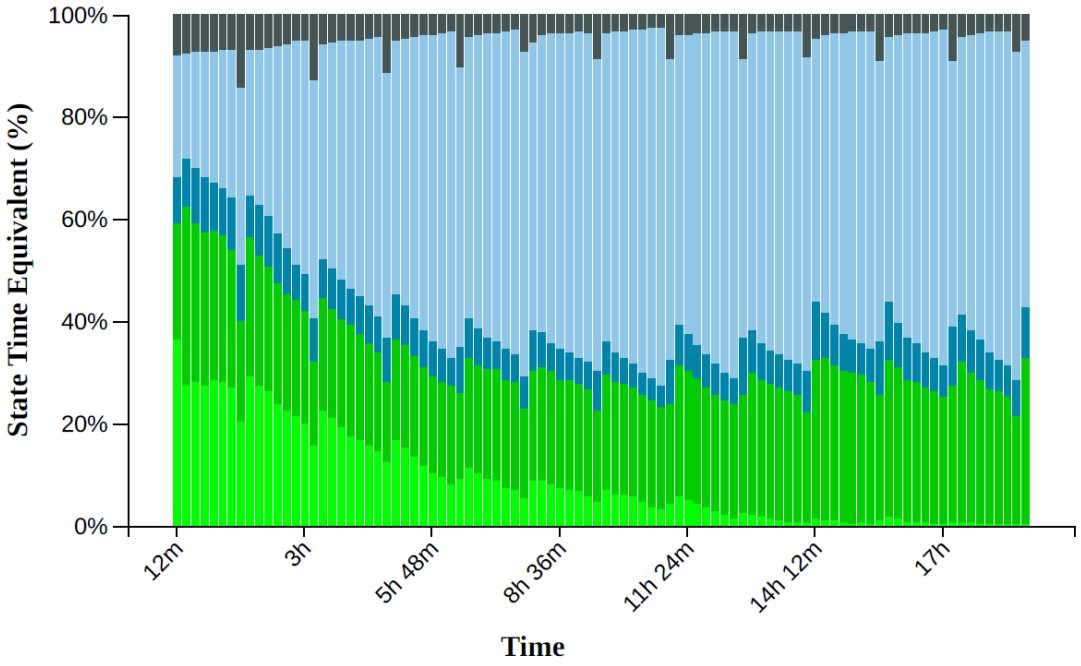


Figure 16. Screenshot of ONT sequencing software sequencing data for the first loading of the flow cell. Light green indicates the percentage of pores that are currently reading the strand, darker green indicates percentage of pores that are waiting for a strand to be introduced to the pore and blue indicates the percentage of pores that are damaged. A rapid overall decrease in the number pores that are sequencing can be seen through first 12 hours of sequencing. After 14 hours of sequencing there is only a negligible percentage of pores sequencing. The percentage of healthy pores remains mostly consistent after four hours of sequencing.

Most of the reads sequenced were not the ones in target. Out of total 1.12 Gb sequenced during the first trial run, that was for GB78 sample, 0.025 Gb of sequence was the one in target, or 2.26% (Fig. 17A). Mapping depth for the off-target sequences was ~2.4X, while for the in-target sequences ~17.9X depth was achieved. The final enrichment was 7.4 times compared to the rest of the genome (Fig. 17B). Average length for the in-target reads sequenced was approximately 29 kb, while the average length of all reads sequenced was around 1.5 kb.

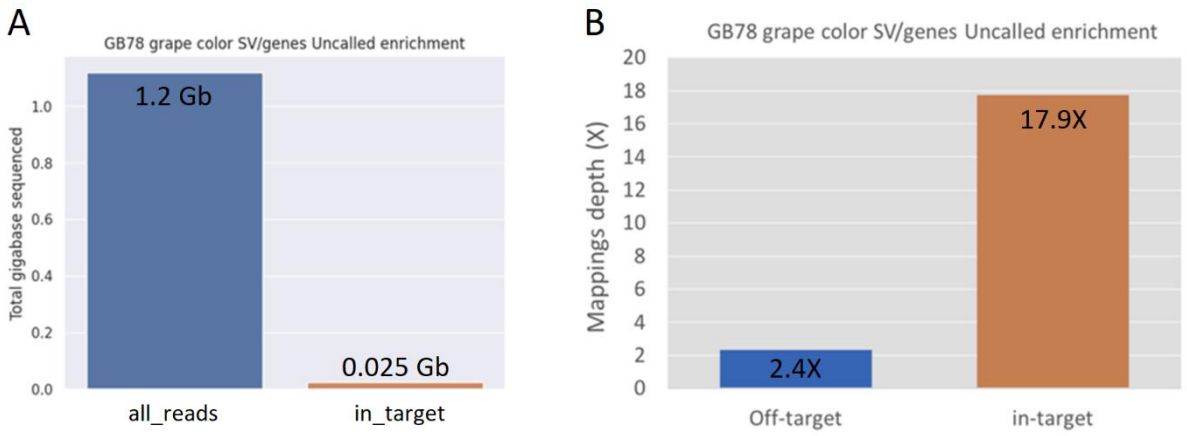


Figure 17. Nanopore target enrichment data at the end of sequencing of GB78 sample. (A) Number of bases sequenced in gigabases compared to the number of bases sequenced that were in-target. Out of 1.2 Gb sequenced, 0.025 Gb were the ones in target (2.26% of the sequenced bases). (B) Mapping depth compared between off-target (2.4X) and in-target (17.9X) sequences. The enrichment of the in-target sequences was 7.4X compared to the rest of the genome.

4.5.2. DEL mutation validation using UNCALLED ONT sequencing

White mutant of 'Garnacha', GB78, was sequenced using a new real-time target enrichment method for ONT sequencing, using UNCALLED, to characterize DEL SV and to confirm its presence with ONT sequencing. Right at the position of the upstream extreme of DEL breakpoint, all the way through to the downstream breakpoint position of GB78 sample, the sequencing depth is half on average, compared to the surrounding sequence (Fig. 18,19), and thus the DEL mutation was confirmed to be present in the sample.

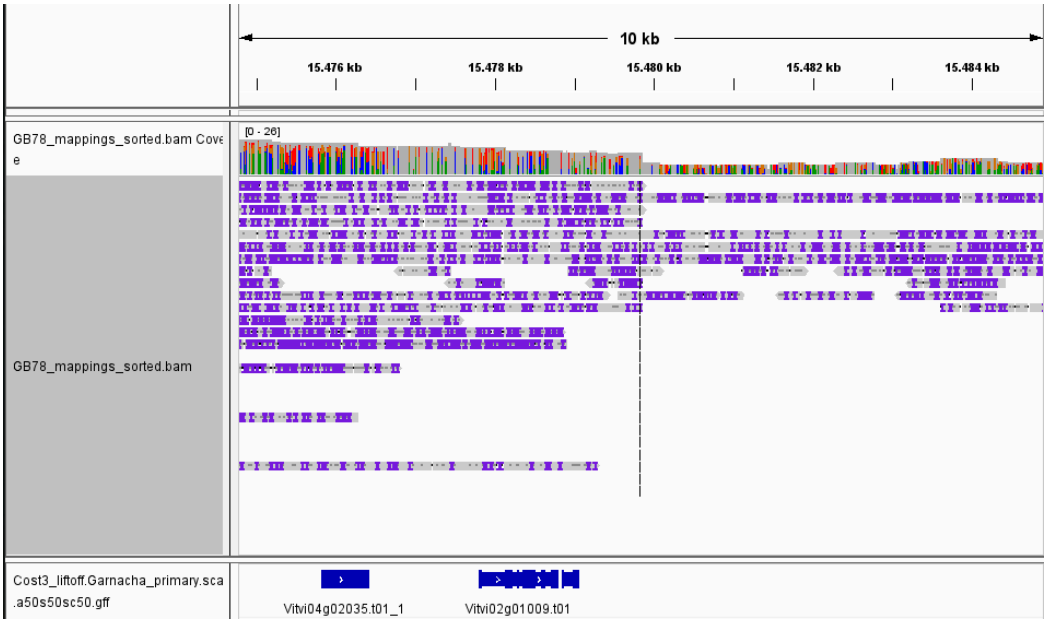


Figure 18. Screenshot of the upstream DEL breakpoint region for the sample GB78, visualized in Integrative genomics viewer program. A 50% drop in coverage can be seen right at the supposed position of the breakpoint.

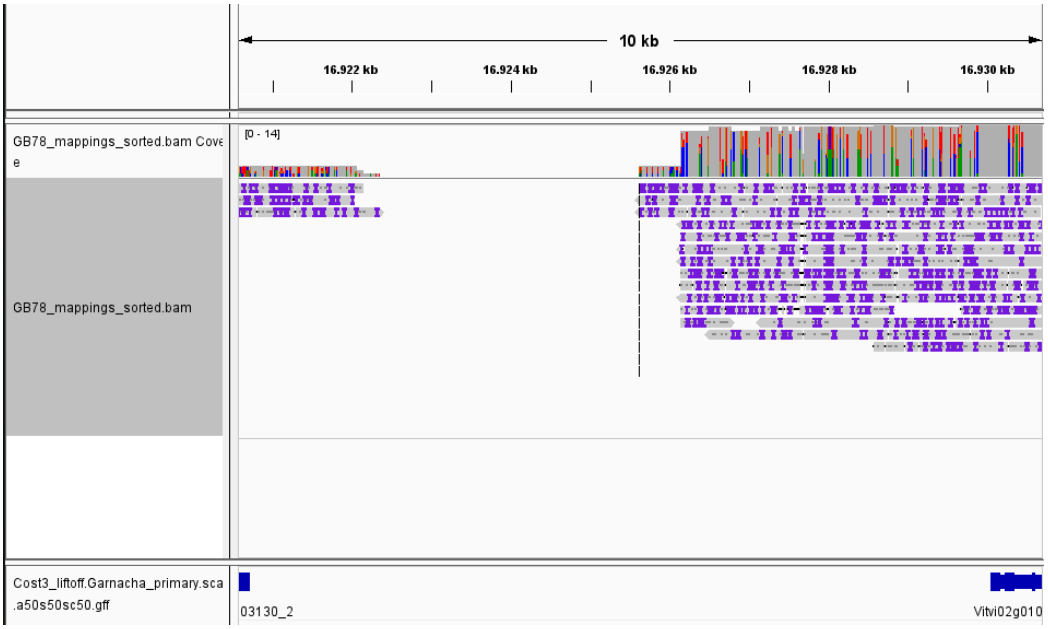


Figure 19. Screenshot of the downstream DEL breakpoint region for the sample GB78, visualized in Integrative genomics viewer program. An immediate 100% drop in coverage can be seen right at the supposed position of the breakpoint, followed by a 50% drop in coverage upstream of the breakpoint.

4.5.3. *Gret1* insertion site analysis using UNCALLED ONT sequencing

Target enrichment method for ONT sequencing, using UNCALLED was also used to assess the hypothesis that the movement of a TE (*Gret1*) is the reason for 'Red Alvarinho' somatic variant color recovery. The potential to study associated DNA methylation variation was also assessed. No ONT reads mapping to *Gret1* sequence that is present in the promotor of *MYBA1* of the 12Xv2 reference genome were detected in the sequenced reads of 'Red Alvarinho', except for the LTR region. For the ancestral, white-berried White Alvarinho, read sequences that correspond to the Gret1 were present (Fig. 20).

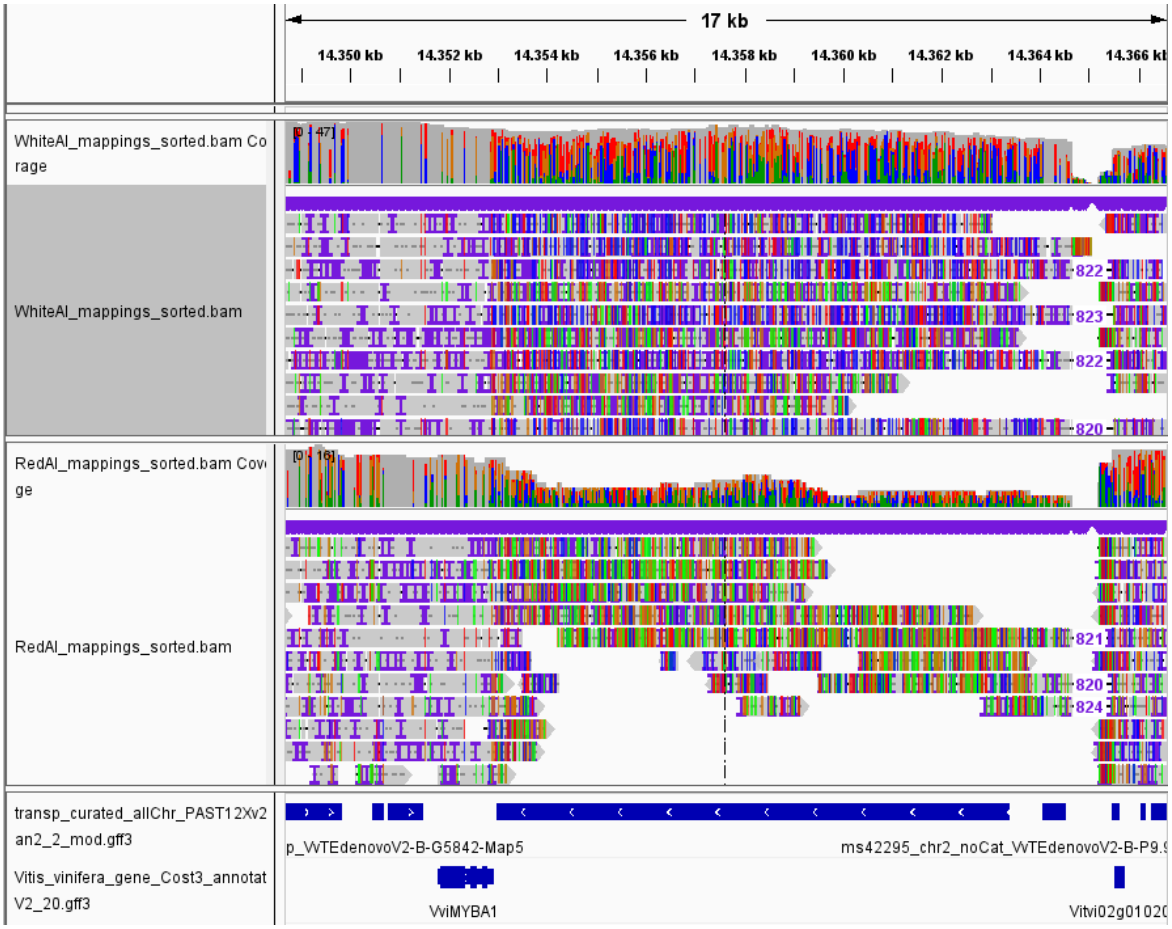


Figure 20. Screenshot of the *MYBA1* promotor region of 'White Alvarinho' (above) and 'Red Alvarinho' (bellow), visualized in Integrative genomics viewer program. *Gret1* jump is confirmed for the red cultivar as there is a noticeable difference in the number of sequenced reads that contain *Gret1* sequence. LTR sequence is still present in the 'Red Alvarinho'.

Thus, Gret1 jump was confirmed using ONT sequencing as the TE was not detected in the color recovery somatic mutant, but it was present in the ancestral, white-berried cultivar. No clipping reads, continuous ones, spanning through the entire *Gret1* region that would correspond to the sequence where *Gret1* excision is present, can be observed amongst the sequenced fragments. The 12Xv2 reference was modified to match the expected sequence more closely, to possibly prevent Megalodon from filtering out needed reads to study methylation. When using the modified reference, no additional information was present after data processing (Fig. 21).

Figure 21. Screenshot of the *MYBA1* promotor region of 'Red Alvarinho', processed and visualized in Integrative genomics viewer program using the modified reference, where the LTR region was removed. Annotated TE map and genes are offset from the start of the LTR by 850 bp (Size of the removed *Gret1* sequence).

5. DISCUSSION

5.1. Structural variation validation

Structural variations detected using ONT sequencing were validated using the designed primers. All samples had their DNAs amplified as predicted by previous nanopore sequencing. GB1662 was confirmed to possess DEL mutation. GB55 and GB77 carry both INV1 and INV2 mutations. GT77, a black-berried individual and the parent plant of GB77 white berry bud sport mutant, was also shown to possess INV mutations. This is discussed in detail in section 5.3. Since the deleted sections of the genome caused by DEL and INV1+INV2 mutations spanned the region where *MYBA* TFs are located from the haplophase carrying the only copy of functional alleles for *MYBA1* and *MYBA2* present in 'Garnacha Tinta' ancestral cultivar (Migliaro et al. 2014), it is apparent that those SVs were the cause of the appearance of white phenotype in GBs. Indeed, deletions spanning the region of chromosome 2 where *MYBA* genes are located were noted to be appearing for many cultivars, and with high variation in deletion size, spanning from 0.08 Mb ('Garnacha Blanca' cultivar) to 4.3 Mb ('Canaiolo Rosa' cultivar) (Migliaro et al. 2017). With a removal of the only functional copies of *MYBA1* and *MYBA2* genes, that are present in GT individuals, the GB plants have lost the ability to produce anthocyanins, and thus their berry color.

Positive control primers (wt) were designed so that they would amplify all gDNA from any grapevine sample (Fig. 8), except in the cases where DNA is not suitable for PCR amplification. There was a risk that individual DNA samples may not have worked as they were extracted over 3 years ago, even though the concentrations and quality were checked. Aged DNA is often highly fragmented due to autolysis and other spontaneous events, reducing the capability of PCR to amplify the template strands (Golenberg et al. 1996). Considering all DNA samples amplified when using positive control primers, they are expected to work with the rest of designed primers if they possess the corresponding mutation.

Designed primers showed some residual unspecific amplification for all three SV breakpoints tested. DEL primers were designed in a way that they would bind to a sequence that is present only if DEL mutation is present. That was achieved by constructing a reverse primer

that has its 3’ end bind to the last six nucleotides before the DEL upstream breakpoint, and 5’ end bind to the first 20 nucleotides after the downstream breakpoint. The primer could bind fully only if the sequence between the breakpoints is missing. However, the primer could still bind partially with its 3’ end to the complementary sequence at the upstream DEL breakpoint, that all samples possess. Even though the binding is weaker, as there is a big part of the primer that is not annealed, some amplification is still possible and that is probably why there is always a faint signal of the same size as the expected product in the gel for all the samples, even the ones that do not carry a DEL mutation. For both INV mutations there seems to be some unspecific binding, as there are many fragments of different sizes always present. Those bands are quite weaker and so can be easily distinguished from the wanted signals, except in a case for INV2, where there is an additional strong unspecific signal. It can be discarded as false positive as it is a too short of a fragment, distinguishable from the expected amplicon for the confirmed presence of the breakpoint. Contamination is not the source of abundant unspecific bands as those do not appear when the DNA sample is replaced by water, so the unwanted amplification must be a result of unspecific binding, which can be expected considering that breakpoint sites lie within repetitive TE sequences (Migliaro et al. 2017).

The problem could be alleviated with the use of more specific primers. However, since the region of interest has high occurrence of TEs, and therefore repetitive regions (Gypsy-like TE are much more abundant in the regions of the genome bordering deletion event sites compared to the rest of the genome (Migliaro et al. 2017)), this goal would be hard to achieve. Additional primers were designed for the purpose of this research, and none were specific enough for the wanted bands to even be visible in the gel after amplification, or simply did not work. More stringent PCR conditions were also tested, but either they still yielded unspecific bands or were not able to amplify in samples carrying the SV breakpoints.

5.2. Garnacha Blanca sample collection genotyping

Samples that carry a deletion-type SV did not carry inversion-type SV and both inversion breakpoints were always appearing together for positive samples. This is expected, as the sequence that these two different SV patterns span overlap in most of the sequence. These mutations would not be able to occur in a same plant, on the same chromosome, as parts of a sequence needed for the mutations to happen (breakpoint sites) are eliminated with the first mutation event, be it DEL or INV. Either DEL or INV1+INV2 SVs are the only two independent origin for GBs in the genotyped collection.

For clonal samples 11-1, 31-1, 54-2 and 5-2 it was not clear whether they possess DEL mutation or not. It is due to the fact that the primer pair designed to amplify samples with DEL mutation were unspecific, as described in section 5.1. Presence or absence of specific amplification had to be estimated from the strength of the signal in the gel after electrophoresis. If the signal was faint, those samples were considered to be negative for DEL mutation, and if the signal was much stronger, they were considered positive, as it can be seen in Figure 6B. Samples 11-1, 31-1, 54-2 and 5-2 had an intermediate signal strength, and therefore could not be classified as either positive or negative. PCR and electrophoresis were done three times to eliminate the possibility of a preparation error, however the results remained consistent. Since DEL SV could not be present for the aforementioned reasons, further study of these samples would be required to find out how does their genotype corelate with the results described here.

The sample GB177 that did not show a band in any of PCR reactions, but still expresses a white phenotype. This result was explained with the recent analysis of cultivar genotyping microsatellite markers, and it was concluded that GB177 does not actually belong to a ‘Garnacha Blanca’ cultivar (unpublished work). With that in mind, the sample worked as it should have.

Overall, PCR reactions were generally reliable, despite the fact that primer combinations used were not fully specific and the DNAs used were 3 years old. Multiple repeats of PCRs were required to get the results predicted by SNP analysis for most samples. Even though it is unlikely, as the results discussed here match the SNP LOH analysis completely, it is possible that there was unspecific amplification happening during PCRs. What could be done further to confirm the

results is sequencing of the amplicons detected on the gel, to confirm they indeed match the targeted sequences. A more detailed and precise analysis of the GB11-1, 31-1, 54-2 and 5-2 samples is required to draw any definite conclusions regarding their genotype.

5.3. Testing 'Garnacha Tinta' samples for possible chimeras

Garnacha Tinta samples, those corresponding to the ancestral clonal lineage of Garnacha showing black-berried phenotype, were genotyped for the presence of the two newly validated SV patterns, as it was observed GB variants were appearing independently in somatic variants of 'Garnacha Tinta' cultivar. That could be explained if only one of the cell layers carries the mutation, so that the deletion of this allele does not affect all cell layers. This mutation could go unnoticed if the anthocyanin production has stopped in only a thin layer off cells, presumably L1 as it forms the monolayer epidermis of a berry (Walker et al. 2006). Since anthocyanins are still synthesized in the L2, as all but the most external cell layer in the berry skin originate from the L2 (Thompson and Olmo 1963; Carbonell-Bejerano et al. 2019), dark phenotype of the pulp tissue underneath the epidermis would overpower the white phenotype of epidermis, resulting in a seemingly completely black berry. If the opposite case would be true, where L2 would carry the mutation wile L1 still produces anthocyanins, expected color of the berry would be red (Barceló et al. 1994; Walker et al. 2006), as the thin layer of epidermal cells would not contain enough anthocyanin to bring out black color. This case of chimerism can be observed in red cultivars 'Malian' and 'Pinot Gris', that both emerged as a result of somatic variation (Walker et al. 2006).

The existence of a black-berried 'Garnacha Tinta' chimeric for loss of *MYBA* genes SV was confirmed using PCR, as INV1 and INV2 mutations were proven to indeed be present in the GT77 sample in spite of its black-berried phenotype. Since GB77, a white variant bud sport, emerged from the black-berried GT77 plant (Fig. 7), I conclude that the predisposition for that berry color existed in its parental plant. What was further needed for a completely white berry phenotype to appear was an additional cell migration event, caused by some form of damage done to meristem tissue (Kidner et al. 2000; Walker et al. 2006). 'Tempranillo Blanco', a white-berried cultivar, originated trough LOH in L1 with subsequent colonization of mutated cells into L2 (Martínez et al.

2007), and same events could describe the emergence of GB77. Original ONT sequencing that discovered the candidate INV1+INV2 SV events in GB77 also show that the coverage in the deleted regions, delimited by SV breakpoints, reach a drop of 50% of read depth, which means that the mutation should be present in all cell layers of GB77. This indicates that, instead of an independent KO mutation in the L2, colonization of the L2 meristem cell layer by INV1+INV2 mutant cells that were already present in the L1 of GT77 should be in the origin of the white berry phenotype of GB77.

GT77 was the only black-berried phenotype sample that was proven to be a chimera, as other wild-type samples do not show a band when genotyped with primers for INV1. As they were negative for the amplification, they do not harbor undiscovered independent INV type SV in the grape color locus, thus no white-berried mutants can be expected to originate from them in the manner GB77 did from GT77. In contrast, additional white-berried bud sports could emerge from the chimeric clonal line to which GT77 accession belongs to.

5.4. Structural variation segregation analysis

Self-cross progeny of GB78 and GB77 was genotyped by PCR to check how DEL, INV1 and INV2 mutations segregate and to see if they are deleterious or lethal for sexual phase. Only INV1 breakpoint was checked for the INV mutants, as due to their proximity, it is likely that all the samples that possess it will also possess the linked INV2 breakpoint. For Mendelian segregation, it would be expected that three of every four self-cross S1 offspring individuals would carry the mutation, or 75%. As 70% of self-crosses of GB78 carry the DEL mutation, and 90% of GB77 self-crosses carry the INV1 mutation, I conclude that all detected SVs segregate and are not deleterious for meiosis or lethal for haploid gametes. This also means that the white-berried alleles of the color locus emerged by somatic SV in ‘Garnacha’ could generate new white-berried cultivars through segregation after sexual crossing.

70% of positive samples is very close to the expected rate of 75%, however, 90% of self-crosses carrying the mutation is relatively high. Nevertheless, this higher number should not matter because if the mutations were lethal for meiosis or haploid gametes, the percentage of

progeny individuals with the mutation would actually fall as the gametes carrying the mutation would not be able undergo fertilization to produce an offspring.

These conclusions could be confirmed with more certainty if more self-cross samples would be genotyped at the same time. The sample size used here was relatively small, as a single sample equates to the 10% of the cases in the analysis. If the analysis would be repeated, I would recommend a bigger sample size to be used for genotyping.

5.5. Nanopore target enrichment

5.5.1. Method Troubleshooting

After the UNCALLED command for real-time target enrichment is started during a ONT MinION run, pores were getting damaged/unavailable more rapidly, presumably due to the more frequent initiation of sequencing, increasing the chance of blockage, or pore clogging caused by ssDNA self-binding. Long term lifetime of individual pores does not decrease when UNCALLED Realtime is activated, and so the inactivity of the pores is most likely temporary (Kovaka et al. 2021). Nuclease flush solves this problem (Kovaka et al. 2021), however, more DNA is needed as the library has to be reloaded onto a flow cell, and the problem will reoccur eventually, as the sequencing process is progressing.

The number of available pores increased significantly throughout the entirety of sequencing run as there was a lot less sequencing done overall. This is because most of the pores are ejecting unwanted fragments, those that do not map to the provided reference, and are waiting for a new strand of DNA to be brought up to the pore, and this increase the time each individual pore is empty (Kovaka et al. 2021). At any given moment, there were fewer pores sequencing than it would be expected, because the requirement that the read must map to a reference is very strict. Only a small percentage of the genome was given as a target, and thus, the chance for a targeted sequence to be brought up to a pore is lower compared to the entirety of the genome.

These effects are cumulative and can be seen in greater extent after the sequencing has been running for a couple of hours. Nanopore flow cells were designed, in theory, to perform sequencing runs lasting up to 72 hours, in optimal conditions (www.nanoporetech.com). Here however, there is practically no sequencing being done after only about 12 hours, even when there are plenty of healthy pores available. Possible reason is that the loaded library is being exhausted of target sequences and the unwanted reads highly outnumber the ones that are left.

At the end, most of the reads sequenced were not the ones in target. This is because any of the off-target sequence that reaches a pore is still sequenced on its first 450-1,500 bp. They are passing through the pore with no chance of being ejected, considering that ~450 bp per second are sequenced in a nanopore in the MinION sequencer and thus the window of 1-3 seconds that UNCALLED needs for the comparison with the reference is long enough for them to pass through the pore with their entire length. The final enrichment achieved was about 7.4X, however, the total sequence saved at the end is about half than what would be expected in a regular ONT run of the same duration, as the number of pores sequencing at any time drops to approximately 50%, when the UNCALLED Realtime command is activated. This makes the effective enrichment approximately 4X. Compared to the 5.5X enrichment achieved by Kovaka et al. (2021) using human genome, effective enrichment described here is lower. In their original comparison between human and bacterial genome, where enrichment was worse for the human genome target enrichment, they hypothesized that the possible causes are a more frequent occurrence of repetitive regions and higher complexity of the genome. As the occurrence of repetitive regions and complexity increase further in plants, this could be the reason for the lower enrichment noted here. The features of the selected targets could also contribute to lower enrichment as the SV targets were in TE-rich regions and 10-30 kb upstream and downstream of selected target genes might also involve repetitive elements.

In conclusion, while the enrichment does work, it is not very effective when used in conjunction with plant samples, probably due to the very repetitive nature of their genomes. The level of enrichment achieved could only be useful in some specific cases, although the method could still be optimized. Additional runs were performed, and as certain factors were changed (loaded DNA concentration, DNase treatment duration, etc.) there was also variation in the

quality of the end result. In particular, loading much higher concentration of DNA seemed to be quite beneficial for one of the runs. This observation cannot be taken with certainty, however, as the quality of the individual runs also varied greatly with the condition of the flow cell used. If individual factors are taken into consideration and are further optimized, there is a possibility that the quality of this method used on plant samples could be increased to a more satisfactory level.

5.5.2. DEL mutation validation using UNCALLED ONT sequencing

GB78 was validated to be a DEL mutant, as it was found that the region of genome spanning between the DEL breakpoints is missing in one of the chromosomes. It is probable that the GB78 plant inherited the DEL mutation by vegetative propagation from a single GT plant in which the deletion of 1.4 Mb originally emerged, as no new independent occurrences of white-berried bud sports with DEL-type SV have been found to this date. White phenotype can be explained if the other allele inherited from the last sexual reproduction was the canonical white allele, as it is the case in 'Garnacha Tinta' ancestral cultivar of 'Garnacha Blanca' derivative cultivar, thus leaving the plant with only unfunctional copies of *MYBA* genes after deletion of the functional copy.

When GB78 sequenced reads were mapped to the GT1608 'Garnacha Tinta' *de novo* genome assembly and visualized in IGV, sequencing depth drops between the breakpoint sites as one of the two chromosome copies has been lost from this region of the genome in GB78. As that part of the sequence is missing in the GB78 genome, when running nanopore, the chance of that region to be sequenced is only half compared to the rest of the target sequences. Other target sequences provided are present in both chromosome copies, and as the result, their coverage is double on average. This fact confirms that GB78 indeed does carry a deletion mutation in one allele, so that it is hemizygous, as it was also proven by genotyping (See section 5.2).

In some locations on the genome, zero coverage can be seen, and it is quite apparent next to a downstream breakpoint of the DEL mutation (Fig. 14). Those results are most likely not a result of mutations in both chromosomes. The emergence of those gaps could have happened as a result off difference between the sample genome and the reference genome provided for the

analysis. One other possible explanation could be that the eventual secondary formations enabled by the presence of specific sequences interfered with the nanopore system and prevented the sequencing process to be performed for the fragments containing those regions.

5.5.3. *Gret1* insertion site analysis using UNCALLED ONT sequencing

Concerning the sequencing of the *Gret1* LTR TE that is found in the promoter of *MYBA1* of the canonical white allele of the frape color locus, in the sequenced reads of 'Red Avarinho', LTR region can be found, but for the rest of the transposon the sequencing coverage is half throughout its entire length, compared to the surrounding sequence. What that means is that the whole 10 kb *Gret1* TE is only present in one of the alleles and thus had 50% less chance of being sequenced, resulting in half of the expected coverage. When observing the sequences belonging to the 'White Alvarinho', this drop in coverage is not present confirming that it is homozygous for *Gret1* TE insertion. The conclusion is that Gret1 sequence is missing in one of the alleles, leaving only one copy of the LTR region behind, and it is how the phenotype was reverted to the red berry color variant in 'Red Alvarinho' compared to the ancestral white-berried 'Alvarinho'. This result confirms the hypothesis that was previously established for the higher *MYBA1* expression in 'Red Alvarinho' compared to 'White Alvarinho' observed in a berry skin RNA-sequencing comparison (ICVV, unpublished data). Color recovery was observed in other somatic variant cultivars, like 'Benitaka' a red-berried derivative cultivar of the white-berried cltivar 'Italia', where it was hypothesized that the color recovery was the result of homologous recombination between *MYBA1* and *MYBA3* genes, as they share identical parts of the promoter region and part of the coding sequence. The expression could be induced by the promoter sequence derived from *MYBA3* (Azuma et al. 2009). This is likely not the case here, as one copy of LTR sequence remains in the promotor region, suggesting the recombination happened between two LTR regions of *Gret1*. Reports of intra-LTR recombination of *Gret1*, resulting in color recovery of grapevine berries, do exist, as it was the case with 'Ruby Okuyama', another red-berried somatic variant derived from 'Italia' (Kobayashi et al. 2004, Kobayashi et al. 2005).

One of the goals was to check how methylation changed between variants, however, to that end, soft-clipping reads, are needed, as those would be the ones containing relevant information at sites of SV events. The soft-clipping reads would be the ones corresponding to the sequences obtained from the chromosome copy in which Gret1 was excised. In that chromosome copy, the sequence upstream of *Gret1* is followed immediately by the LTR sequence and then sequence downstream of the *Gret1*. Library fragments from that region, after they are sequenced and mapped, would not map to a reference completely as one part of the sequence they possess is located ~8.8 kb further away in the reference. The result is a clipping read that annotates only partially.

Soft-clipping reads were not present because the Megalodon tool used performs the methylation calling based on the presence of C nucleotides according to the reference used and thus filters out soft-clipping mapping reads. As the clipping part cannot be analyzed for methylation, no data could be gathered for most reads in the region. Providing Megalodon a modified reference for the analysis, in which the *Gret1* was removed leaving only one LTR mimicking the TE jump, did not result in the appearance of relevant reads. Megalodon filters those reads out, as they existed in the output reads and were useful to detect the deletion of *Gret1* TE before data processing with Megalodon. I conclude that Megalodon may not be a good tool of choice to study methylation when large SVs are present, as the filtering for clipping aligned reads seems to be too strict. Alternative tools would have to tested, such as Nanopolish, used by Kovaka et al. (2021) in their original testings, where methylation was successfully detected, as to enable the study of methylation in the context of plant genomes.

Methylation analysis using identical methodology did succeed in cases where large genome rearrangements were not present. The grapevine sex locus region was one of the provided targets for the enrichment. While the results of the sex locus region analysis are out of scope for this thesis, it displays that variation in methylation status of DNA can be successfully detected with ONT sequencing target enrichment, using UNCALLED Realtime and Megalodon (Supplementary figure 1).

6. CONCLUSION

Research described in this paper showed that structural variations (SVs), proposed during preliminary studies, were present in the individual plants of 'Garnacha Blanca' white-berried grape cultivar derived from somatic variation of 'Garnacha Tinta'. Thus, these SVs are the cause for the white berry color phenotype. Deletion of functional *MYBA* genes, in combination with the presence of a canonical null allele resulted with inability for the cells to synthesize anthocyanins, rendering berry color white. Only two types of SV events were found in a collection of 'Garnacha Blanca' indicating that the cultivar includes accessions from two independent genetic origins. Both SV patterns segregate and are not deleterious for meiosis or lethal for haploid gametes. White-berried individuals of 'Garnacha' could therefore generate new white-berried cultivars through segregation of the SV alleles after sexual reproduction. GT77 was shown to be the only chimeric plant out of all GT samples studied as it harbors an INV SVs in one of its cell layers, despite its black phenotype. It was suspected that mutation occurred in the L1 cell monolayer and thus could have gone unnoticed. The appearance of GB77, bud sport of GT77, should then explained by a cell migration event, where mutated cells colonized the L2 cell layer.

DEL SV was validated in GB78 using ONT target enrichment sequencing. *Gret1* jump was confirmed as the origin for berry color re-gain in 'Red Alvarinho' using the same ONT approach. ONT sequencing target enrichment method, using UNCALLED, was assessed in the context of a plant genome. Effective enrichment achieved was not very effective, proximately 4X. The likely reason is that due to the repetitive nature of a plant genome, which makes recognition of targeted sequences difficult as there is a higher chance they would appear throughout the genome. The DNA methylation analysis pipeline used, involving Megalodon program, was recognized as flawed as by the end of data processing important information was lost. The suspect for this loss is Megalodon program, as its filtering eliminated important reads required for analyses of regions involving SV between the sequenced sample and the mapping reference. Study of methylation was shown to be possible with the methodology and data processing used. Overall, the method could be improved and brought to a more satisfactory level with further optimization and testing of alternative tools.

7. LITERATURE

Alkan C., Coe B., Eichler E. (2011): Genome structural variation discovery and genotyping. Nat Rev Genet 12: 363–376.

Azuma A., Kobayashi S., Goto-Yamamoto N., Shiraishi M., Mitani N., Yakushiji H., Koshita Y. (2009): Color recovery in berries of grape (Vitis vinifera L.) 'Benitaka', a bud sport of 'Italia', is caused by a novel allele at the VvmybA1 locus. Plant Sci 176(4): 470-478.

Barceló A.R., Calderón A.A., Zapata J.M., Muñoz R. (1994): The histochemical localization of anthocyanins in seeded and seedless grapes (Vitis vinifera). Sci Hortic 57(3): 265-268.

Bird A. (2002): DNA methylation patterns and epigenetic memory. Genes Dev 16: 6-21.

Boss P.K., Davies C., Robinson S.P. (1996): Expression of anthocyanin biosynthesis pathway genes in red and white grapes. Plant Mol Biol 32: 565–569.

Canaguier A., Grimplet J., Di Gaspero G., Scalabrin S., Duchêne E., Choisne N., Mohellibi N., Guichard C., Rombauts S., Le Clainche I., Bérard A., Chauveau A., Bounon R., Rustenholz C., Morgante M., Le Paslier M.C., Brunel D., Adam-Blondon A.F. (2017): A new version of the grapevine reference genome assembly (12X.v2) and of its annotation (VCost.v3). Genom Data 14: 56-62.

Carbonell-Bejerano P., Royo C., Torres-Pérez R., Grimplet J., Fernandez L., Franco-Zorrilla J. M., Lijavetzky D., Baroja E., Martínez J., García-Escudero E., Ibáñez J., Martínez-Zapater J. M. (2017): Catastrophic unbalanced genome rearrangements cause somatic loss of berry color in grapevine. Plant Physiol 175(2): 786–801.

Carbonell-Bejerano P., Royo C., Mauri N., Ibáñez J., Martínez Zapater J. M. (2019): Somatic variation and cultivar innovation in grapevine. Advances in Grape and Wine Biotechnology, IntechOpen

Cardone M.F., D'Addabbo P., Alkan C., Bergamini C., Catacchio C.R., Anaclerio F., Chiatante G., Marra A., Giannuzzi G., Perniola R., Ventura M., Antonacci D. (2016): Inter-varietal structural variation in grapevine genomes. Plant J 88(4): 648-661.

Chiang C., Scott A.J., Davis J.R., Tsang E.K., Li X., Kim Y., Hadzic T., Damani F.N., Ganel L. (2017): The impact of structural variation on human gene expression. Nat Genet 49(5): 692-699.

Escaramis G., Docampo E., Rabionet R. (2015): A decade of structural variants: description, history and methods to detect structural variation. Brief Funct Genomics 14: 305–314.

Ferreira V., Pinto-Carnide O., Arroyo-García R., Castro I. (2018) Berry color variation in grapevine as a source of diversity. Plant Physiol Biochem 132: 696-707.

Feuk L., Carson A.R., Scherer S.W. (2006): Structural variation in the human genome. Nat Rev Genet 7: 85-97.

Fournier-Level A., Le Cunff L., Gomez C., Doligez A., Ageorges A., Roux C., Bertrand Y., Souquet J.M., Cheynier V., This P. (2009): Quantitative genetic bases of anthocyanin variation in grape (Vitis vinifera L. ssp. sativa) berry: a quantitative trait locus to quantitative trait nucleotide integrated study. Genetics 183(3): 1127-1139.

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 (Edinb) 104(4): 351-362.

Gabur I., Chawla H.S., Snowdon R.J., Parkin I.A.P. (2019): Connecting genome structural variation with complex traits in crop plants. Theor Appl Genet 132(3): 733-750.

Galet P. (2000): Dictionnaire encyclopedique des cepages. Hachette, Paris.

Golenberg E. M., Bickel A., Weihs P. (1996): Effect of Highly Fragmented DNA on PCR. Nucleic Acids Res 24(24): 5026–5033.

Guasch-Jané M.R., Andrés-Lacueva C., Jáuregui O., Lamuela- Raventós R.M. (2006): First evidence of white wine in ancient Egypt from Tutankhamun's tomb. J Archeol Sci 33: 1075–1080.

Hancock J.F. (1992): Plant evolution and origin of crop species, 1st ed. Prentice-Hall, Englewood Cliffs

Jaillon O., Aury J.M., Noel B., Policriti A., Clepet C., Casagrande A., Choisne N., Aubourg S., Vitulo N., Jubin C., Vezzi A., Legeai F. et al. (2007): French-Italian Public Consortium for Grapevine Genome Characterization. The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. Nature 449(7161): 463-7.

Kidner C., Sundaresan V., Roberts K., Dolan L. (2000): Clonal analysis of the Arabidopsis root confirms that position, not lineage, determines cell fate. Planta 211: 191–199.

Kobayashi S., Ishimaru M., Ding C.K., Yakushiji H., Goto N. (2001): Comparison of UDP-glucose:flavonoid 3-O-glucosyltransferase (UFGT) gene sequences between white grapes (Vitis vinifera) and their sports with red skin. Plant Sci 160(3): 543-550.

Kobayashi S., Ishimaru M., Hiraoka K., Honda C. (2002): Mybrelated genes of the Kyoho grape (Vitis labruscana) regulate anthocyanin biosynthesis. Planta 215: 924–933.

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

Kobayashi S., Goto-Yamamoto N., Hirochika H. (2005): Association of VvmybA1 gene expression with anthocyanin production in grape (Vitis vinifera) skin-color mutants. J Jpn Soc Hort Sci 74(3): 196–203.

Kovaka S., Fan Y., Ni B., Timp W., Schatz M.C. (2021): Targeted nanopore sequencing by real-time mapping of raw electrical signal with UNCALLED. Nat Biotechnol 39(4): 431-441.

Lisch D. (2013): How important are transposons for plant evolution? Nat Rev Genet 14, 49–61.

Magi A., Semeraro R., Mingrino A., Giusti B., D'Aurizio R. (2018): Nanopore sequencing data analysis: state of the art, applications and challenges. Brief Bioinform 19(6): 1256-1272.

Marroni F., Pinosio S., Morgante M. (2014): Structural variation and genome complexity: is dispensable really dispensable? Curr Opin Plant Biol 18: 31-6.

Martínez J., Vicente T., Martínez T., Chavarri J.B., Garcia-Escudero E. (2007): Tempranillo blanco, características de una nueva variedad de vid. Vida Rural 244: 44–48.

Matus J.T., Cavallini E., Loyola R., Höll J., Finezzo L., Dal Santo S., Vialet S., Commisso M., Roman F., Schubert A., Alcalde J.A., Bogs J., Ageorges A., Tornielli G.B., Arce-Johnson P. (2017): A group of grapevine MYBA transcription factors located in chromosome 14 control anthocyanin synthesis in vegetative organs with different specificities compared with the berry color locus. Plant J 91: 220–236.

Migliaro D., Crespan M., Muñoz-Organero G., Velasco R., Moser C., Vezzulli S. (2014): Genetic characterisation of berry colour variants. Aust J Grape Wine Res 20: 485-495.

Migliaro D., Crespan M., Muñoz-Organero G., Velasco R., Moser C., Vezzulli S. (2017): Structural dynamics at the berry colour locus in *Vitis vinifera* L. somatic variants. Acta Hortic 1157: 27-32

Morgante M., De Paoli E., Radovic S. (2007): Transposable elements and the plant pan-genomes. Curr Opin Plant Biol 10: 149–155.

McGovern P.E., Fleming S.J., Katz S.H. (1996): The origins and ancient history of wine: food and nutrition in history and anthropology, 1st ed. Gordon and Breach, New York

McKey D., Elias M., Pujol B., Duputié A. (2010): The evolutionary ecology of clonally propagated domesticated plants. New Phytol 186: 318-332

Mercenaro L., Nieddu G., Porceddu A., Pezzotti M., Camiolo S. (2017): sequence polymorphisms and structural variations among four grapevine (Vitis vinifera L.) cultivars representing sardinian agriculture. Front Plant Sci 8: 1279.

Myles S., Boyko A.R., Owens C.L., Brown P.J., Grassi F., Aradhya M.K., Prins B., Reynolds A., Chia J.M., Ware D., Bustamante C.D., Buckler E.S. (2011): Genetic structure and domestication history of the grape. Proc Natl Acad Sci U S A. 108(9): 3530–3535.

Neilson-Jones, W. (1969): Plant Chimeras, 2nd ed. Methuen, London.

Poudel P., Azuma A., Kobayash, S., Koyama K., Goto-Yamamoto N. (2021): VvMYBAs induce expression of a series of anthocyanin biosynthetic pathway genes in red grapes (Vitis vinifera L.). Sci Hortic 283: 110121.

Rabinowicz P.D., Braun E.L., Wolfe A.D., Bowen B., Grotewold E. (1999): Maize R2R3 Myb genes: sequence analysis reveals amplification in the higher plants. Genetics 153: 427–444.

Schuermann D., Molinier J., Fritsch O., Hohn B. (2005): The dual nature of homologous recombination in plants. Trends Genet 21:172–181.

This P., Lacombe T., Cadle-Davidson M., Owens C.L. (2007): Wine grape (Vitis vinifera L.) color associates with allelic variation in the domestication gene VvmybA1. Theor Appl Genet 114(4): 723-30.

Thompson M.M., Olmo H.P. (1963): Cytohistological studies of Cytochimeric and tetraploid grapes. Am J Bot 50: 901.

Vondras A.M., Minio A., Blanco-Ulate B., Figueroa-Balderas R., Penn M.A., Zhou Y., Seymour D., Ye Z., Liang D., Espinoza L.K., Anderson M.M., Walker M.A., Gaut B., Cantu D. (2019): The genomic diversification of grapevine clones. BMC Genomics 20(1): 972.

Voss-Fels K., Snowdon R.J. (2016): Understanding and utilizing crop genome diversity via high-resolution genotyping. Plant Biotechnol J 14(4):1086–1094.

Walker A.R., Lee E., Robinson S.P. (2006): Two new grape cultivars, bud sports of cabernet sauvignon bearing pale-coloured berries, are the result of deletion of two regulatory genes of the berry colour locus. Plant Mol Biol 62: 623-635.

Walker A.R., Lee E., Bogs J., McDavid D.A.J., Thomas M.R., Robinson S.P. (2007): White grapes arose through the mutation of two similar and adjacent regulatory genes. Plant J 49: 772-785.

Wang X., Weigel D., Smith L.M. (2013) Transposon variants and their effects on gene expression in Arabidopsis. PLoS Genet 9(2): e1003255.

Watson J. M., Platzer A., Kazda A., Akimcheva S., Valuchova S., Nizhynska V., Nordbor M., Riha K. (2016): Germline replications and somatic mutation accumulation are independent of vegetative life span in Arabidopsis. Proc Natl Acad Sci U S A 113(43): 12226–12231.

Xu Y., Gao Z., Tao J., Jiang W., Zhang S., Wang Q., Qu S. (2016): Genome-wide detection of SNP and SV variations to reveal early ripening-related genes in grape. PLoS One 11(2): e0147749.

Yang N., Liu J., Gao Q., Gui S., Chen L., Yang L., Huang J., Deng T., Luo J., He L., Wang Y., Xu P., Peng Y., Shi Z., Lan L., Ma Z., Yang X., Zhang Q., Bai M., Li S., Li W., Liu L., Jackson D., Yan J. (2019): Genome assembly of a tropical maize inbred line provides insights into structural variation and crop improvement. Nat Genet 51: 1052-1059.

Zhou Y., Minio A., Massonnet M., Solares E., Lv Y., Beridze T., Cantu D., Gaut B.S. (2019): The population genetics of structural variants in grapevine domestication. Nat Plants. 5(9): 965-979.

https://nanoporetech.com/how-it-works (accessed on 14.01.2022.)

https://nanoporetech.com/resource-centre/introduction-nanopore-sequencing (accessed on 14.01.2022.)

https://nanoporetech.com/products/minion (accessed on 23.01.2022.)

https://github.com/skovaka/UNCALLED (accessed on 29.01.2022.)

https://github.com/LernerLab/GuPPy (accessed on 29.01.2022.)

https://github.com/lh3/minimap2 (accessed on 29.01.2022.)

https://github.com/nanoporetech/medaka (accessed on 29.01.2022.)

https://github.com/nanoporetech/megalodon (accessed on 29.01.2022.)

https://urgi.versailles.inra.fr/Species/Vitis/Data-Sequences/Genome-sequences (accessed on 30.01.2022.)

https://www.ncbi.nlm.nih.gov/ (accessed on 06.02.2022.)

8. CURRICULUM VITAE

I was born on 13th of April 1995. In Zagreb, Croatia. In my childhood, I went to the elementary school "OŠ Dragutin Domjanić", after which I became a chef, graduating from University of Tourism and Catering, Zagreb, in 2013. After that I became Technician Nutritionist, graduating from the Public University, Zagreb, in 2015. In 2016. I got accepted to Faculty of Science, Zagreb, to study undergraduate molecular biology and subsequently graduate molecular biology. During my studies I participated in many university related events, mainly "Dan i Noć na PMF-u" for several years. I worked in the laboratory of Archaeobotany and Geobotany for my practice and was an intern at Max Planck Institute for Developmental Biology, where I did my masters thesis work.

9. SUPPLEMENTARY DATA

1. List of samples used for SV validation in ‘Garnacha Blanca’ sample collection

2. List of samples used for SV segregation analysis

3. List of GT samples used for chimera search analysis

4. List of designed primers used for genotyping as well as their information and characteristics

5. ‘Alvarinho’ samples data from nanodrop analysis

6. Gene target list for ONT target enrichment sequencing

7. The grapevine sex locus region variation in methylation status, detected using ONT sequencing target enrichment using UNCALLED in combination with Megalodon program

Supplementary table 1. List of samples used for structural variation validation in 'Garnacha Blanca' sample collection as well as their concentrations and purity measured on nanodrop spectrophotometer.

#	conc. (ng/uL)	260/280	260/230	#	conc. (ng/uL)	260/280	260/230
GB1-2	254.60	1.84	2.02	GB17-2	492.70	1.80	1.87
GB4-2	329.70	1.78	1.93	GB31-1	63.50	1.85	1.65
GB11-1	387.80	1.82	2.14	GB50-1	313.10	1.84	2.13
GB25-2	193.20	1.87	1.96	GB54-2	374.80	1.84	1.99
GB27-1	257.10	1.81	2.09	GB56-2	84.50	1.83	2.17
GB29-2	85.30	1.82	1.74	GB61-2	283.20	1.84	2.20
GB30-2	298.40	1.78	0.72	GB63-2	204.50	1.84	2.18
GB32-2	285.50	1.82	2.05	GB76-1	222.80	1.79	1.82
GB57-2	251.30	1.84	2.03	GB1608	141.60	1.78	1.61
GB34-2	339.20	1.86	2.16	GB1662	180.00	1.79	1.72
GB35-1	64.80	1.81	2.34	GB1657	166.90	1.77	1.62
GB45-2	175.20	1.86	2.11	GB428	150.60	1.79	1.80
GB49-2	309.70	1.84	2.06	GB177	205.70	1.76	1.42
GB55-2	263.10	1.89	1.94	GB75-1	91.50	1.83	1.85
GB60-1	240.10	1.88	2.15	GB51-1	72.60	1.87	2.00
GB64-1	294.20	1.83	2.11	GB84-2	251.90	1.80	1.84
GB73-2	206.80	1.85	2.29	GB5-2	92.00	1.82	1.71
GB80-1	320.10	1.84	2.31	GB36-1	11.70	1.63	0.72
GB77-2	425.00	1.84	2.05	GB71	131.50	1.84	1.92
GB53-2	247.60	1.83	2.05	GB78-2	123.40	1.84	1.97
GB6-1	114.50	1.87	2.33	GB3-1	57.00	1.79	1.48
GB15-2	449.20	1.79	1.93	GB84-2R	32.10	1.80	1.95

Supplementary table 2. List of samples used for structural variation segregation analysis as well as their concentrations and purity measured on nanodrop spectrophotometer.

Sample	conc. (ng/uL)	260/280	260/230
GB77-1	65.8	1.7	1.23
GB77-6	107.6	1.84	2.29
GB77-7	115.1	1.86	3.19
GB77-8	80.5	1.86	4.83
GB77-9	104.8	1.89	3.02
GB77-12	75.7	1.64	1.17
GB77-14	68.4	1.87	3.07
GB77-15	64.4	1.69	0.95
GB77-16	71.9	1.91	2.39
GB77-18	66.1	1.86	3.38
GB78-3	32.5	2.1	3.85
GB78-6	71	1.86	2.22
GB78-7	72.7	1.86	2
GB78-9	94.9	1.91	1.83
GB78-11	44	1.95	7.55
GB78-13	60.3	1.91	4.88
GB78-14	70	1.85	1.82
GB78-16	70	1.8	2.28
GB78-18	70.5	1.86	2.61
GB78-19	35.8	1.85	6.46

Supplementary table 3. List of 'Garnacha Tinta' (GT) samples used for chimera search analysis.

Cultivar	Accession ID	Country and Province of of origin of accession	Collector entity	Navarre area	Town of collection
Garnacha Tinta	GT1	Spain, Navarre	EVENA	Baja Montaña	Aibar
Garnacha Tinta	GT2	Spain, Navarre	EVENA	Baja Montaña	Eslava
Garnacha Tinta	GT3	Spain, Navarre	EVENA	Baja Montaña	San Martin de Unx
Garnacha Tinta	GT4	Spain, Navarre	EVENA	Valdizarbe	OBANOS
Garnacha Tinta	GT5	Spain, Navarre	EVENA	Tierra Estella	Dicastillo
Garnacha Tinta	GT6	Spain, Navarre	EVENA	Ribera Alta	Olite
Garnacha Tinta	EV007	Spain, Navarre	EVENA	Baja Montaña	Aibar
Garnacha Tinta	EVENA-31	Spain, Navarre	EVENA	Ribera Alta	Tafalla
Garnacha Tinta	EVENA-21	Spain, Navarre	EVENA	Ribera Alta	Falces
Garnacha Tinta	RA-7	Spain, Navarre	EVENA	Ribera Alta	Olite
Garnacha Tinta	EVENA-32	Spain, Navarre	EVENA	Ribera Baja	Ablitas
Garnacha Tinta	EVENA-12	Spain, Navarre	EVENA	Baja Montaña	Liedena
Garnacha Tinta	EVENA-35	Spain, Navarre	EVENA	Tierra Estella	Arroniz
Garnacha Tinta	ARA 24	Spain, Aragon	CITA		

Supplementary table 4. List of designed primers used for genotyping as well as their sequence, orientation, length, position in Garnacha genome assembly and GC content.

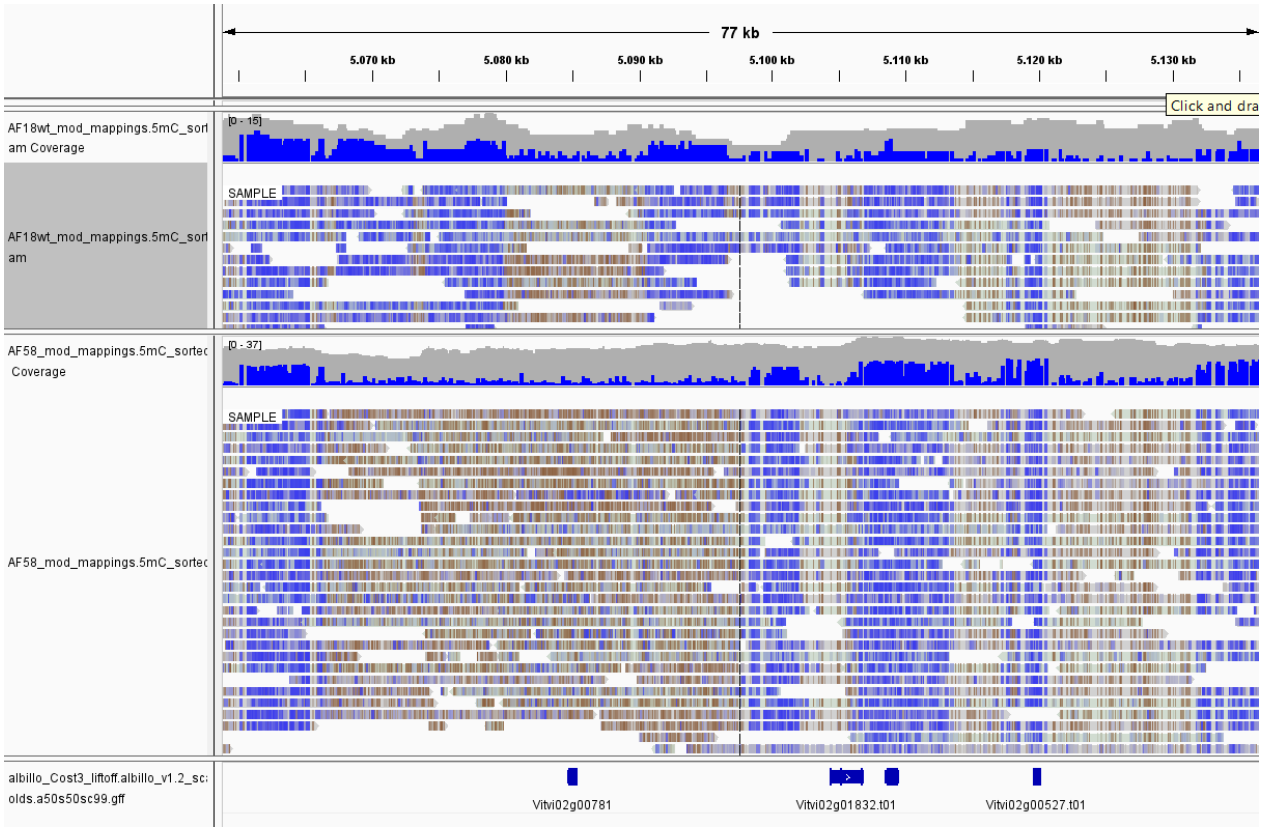
wt	Sequence (5'->3')	Template strand	Length	Primer position	GC%
Forward primer	TAGGCTATGTGCAATAAAGG	Plus	20	chr02:15,479,411-15,479,430	40.00
Reverse primer	CAAACCCTCGTAACATTGT	Minus	19	chr02:15,480,142-15,480,161	42.10
Product length	731				
DEL	Sequence (5'->3')	Template strand	Length	Primer position	GC%
Forward primer	TAGGCTATGTGCAATAAAGG	Plus	20	chr02:15,479,411-15,479,430	40.00
Reverse primer	GATAGTACATCAAACACATCAATTGC	Minus	26	chr02:15,479,862-15,479,868; chr02:16,925,616-16,925,636	34.62
Product length	477				
INV1	Sequence (5'->3')	Template strand	Length	Start	GC%
Forward primer	ACCTGAATTGTGCAACCGAGC	Plus	21	chr02:15,552,813-15,552,834	52.38
Reverse primer	TCATGGGCATCCCTAACTCCG	Minus	21	chr02:16,379,539-16,379,560	57.14
Product length	433				
INV2	Sequence (5'->3')	Template strand	Length	Start	GC%
Forward primer	TAGGATACTACGAGCTTTTCTTGAC	Plus	25	chr02:16,034,693-16,034,718	40.00
Reverse primer	GCAAATGGGATAACACAAGGGG	Minus	22	chr02:16,968,683-16,968,705	50.00
Product length	566				

Supplementary table 5. 'Alvarinho' samples data obtained from nanodrop spectrophotometer analysis.

Sample	Phenotype	Nanodrop ug/uL	260/280	260/230
White Alvarinho	wt white	379	1.83	1.95
Red Alvarinho	red revertant	210	1.73	1.25

Supplementary table 6. Gene target list for ONT target enrichment sequencing.

VIT_216s0039g02230	UFGT1
VIT_206s0009g03010	F3'5'H
VIT_208s0105g00380	LDOX3
VIT_202s0025g04720	LDOX1
VIT_203s0017g00830	LDOX2
VIT_211s0118g00360	LDOX4
VIT_213s0067g01020	LDOX4
VIT_200s0361g00040	ANR1
VIT_204s0079g00690	GST4
VIT_203s0017g00870	3AT
VIT_216s0050g02480	ABCC1
VIT_208s0007g03560	ANM1
VIT_216s0050g00930	AM1
VIT_216s0050g00910	AM2
VIT_216s0050g00900	AM3
VIT_201s0010g03510	AOMT1
VIT_214s0068g00930	CHS1
VIT_201s0011g04760	MYBC2-L1
VIT_214s0006g01620	MYBC2-L3
VIT_208s0007g07230	MYB5A
VIT_206s0004g00570	MYB5B
VIT_215s0046g00170	MYBPA1
VIT_211s0016g01310	MYBPA4
VIT_200s0341g00050	MYBPAL1
VIT_202s0033g00380	MYBA8
VIT_203s0038g02310	MYB4A
VIT_204s0023g03710	MYB4B
VIT_207s0130g00040	MYB3
VIT_207s0005g01210	MYBF1
VIT_205s0077g01360	MYBF2
VIT_218s0001g03470	FLS4
VIT_205s0020g01090	HYH
VIT_204s0008g05210	HY5
VIT_214s0066g01090	MYB24
VIT_210s0003g01170	bHLH_UV
VIT_06s0009g02810	F3'5'Hb
VIT_06s0009g02830	F3'5'Hc
VIT_06s0009g02840	F3'5'Hd
VIT_06s0009g02880	F3'5'Hh
VIT_06s0009g02920	F3'5'Hi
VIT_06s0009g02970	F3'5'Hj
VIT_06s0009g03000	F3'5'Hk
VIT_06s0009g03010	F3'5'Hk2
VIT_06s0009g03140	F3'5'Ho
VIT_16s0022g01500	F3'5'Hq
VIT_16s0022g01510	F3'5'Hr
VIT_16s0022g01540	F3'5'Hs



Supplementary figure 1. The grapevine sex locus region variation in methylation status (CG) of DNA, between the wt (AF18, above) and sex mutant (AF58, bellow) detected using ONT sequencing target enrichment, with UNCALLED Realtime and Megalodon. Methylated cytosines are colored grey, while unmethylated cytosines are colored blue. The change in methylation status is apparent close to the sex locus region, as grey color substituted blue in many cases.