Corso di dottorato di ricerca in:

"Scienze e Biotecnologie Agrarie"

in convenzione con IGA Technology Services srl

Ciclo 38°

"Structural and functional annotation of genomes for the analysis of centromeric structures in *Vitis* species"

Dottorando:
Mario Liva

Supervisore:
Prof. Fabio Marroni

Co-supervisore:
Prof. Michele Morgante

Anno 2026

Finanziato dall'Unione europea- Next Generation EU, Missione 4 Componente 2 CUP G23D22000780005 (D.M. 352/2022)

UNIVERSITY OF UDINE

Doctoral thesis in Agricultural Sciences and Biotechnology

XXXVIII cycle

PhD course coordinator: Prof. Stefano Bovolenta

Structural and functional annotation of genomes for the analysis of centromeric structures in *Vitis* species

PhD candidate: Mario Liva

Doctorate tutor: Prof. Fabio Marroni

Supervisor: Prof. Michele Morgante

Co-supervisor: Gabriele Magris, PhD

Table of contents:

1 List of abbreviations

BAC clones	Bacterial Artificial Chromosome clones
ChIP-seq	Chromatin Immunoprecipitation Sequencing
FISH	Fluorescence *in situ* hybridization
Gb	Giga bases
Hi-C	High Chromosome Conformation Capture
HOR	Higher Order Repeats
Kbp	Kilo base pairs
Mbp	Mega base pairs
Mya	Million years ago
NGS	Next Generation Sequencing
ONT	Oxford Nanopore Technologies
PacBio	Pacific Biosciences
PCR	Polymerase Chain Reaction
rDNA genes	Ribosomal RNA genes
rRNA	Ribosomal RNA
T2T	Telomere to Telomere
TSD	Target Site Duplication
UPGMA	Unweighted Pair Group Method with Arithmetic Mean

2 Abstract

Third generation sequencing enables fast, accurate reconstruction of reference genomes, producing high-quality assemblies even for non-model species. In this thesis, we present a case study of a structural and functional genome annotation for six grapevine accessions, including a quasi-homozygous line, three cultivars, one wild accession, and one interspecific hybrid, assembled using PacBio HiFi and ONT long reads. The resulting haplotype-resolved assemblies are highly complete, often spanning entire chromosomes from telomere to telomere. Their quality allowed detailed analysis of centromeric regions, revealing multiple tandem repeat families, sometimes forming megabase scaled arrays, interspersed with chromodomain-containing transposable elements, particularly belonging to the Athila family. Additionally, comparative analyses uncovered two major centromere architectures sharing a core set of repeat families. The comparison between the various genotypes highlighted the forces shaping these regions, like repeat expansion, structural rearrangements, and TEs activity. Moreover, the information of cytosine methylation present in the ONT datasets allowed the characterization of the epigenetic state of these regions, revealing both broad patterns and fine-scale regulatory variation. Overall, this work provides a case study demonstrating the power of long-read sequencing for structural and functional genome annotation in understudied species.

3 Introduction

3.1 Importance of high-quality genome assemblies for functional genomics and breeding in grapevine (*Vitis vinifera*)

Reference genomes are essential today for genetic and biological studies, since they are used in several types of bioinformatic analyses (e.g., chromatin structure and accessibility, transcriptome, population genomics, epigenetic modifications, etc). Therefore, obtaining a complete and accurate reference genome is a goal that has been pursued since the emergence of genome sequencing. In the last few years, several high-quality reference genomes belonging to different living organisms have been released, from plants to animals. In multiple cases, genomes were reconstructed from telomere to telomere, giving the opportunity to study regions of the genome that were previously unexplored (Nurk *et al.*, 2022; Garg *et al.*, 2024).

Among plant species, the grapevine (*Vitis vinifera ssp. vinifera*) offers a peculiar case of study. The genome is relatively small (approximately 500 Mbp, 2n=38) (Shi *et al.*, 2023), but high sequence diversity between the two haplotypes present in cultivated varieties (Minio *et al.*, 2017) has made the generation of a haplotype-specific reference genome hard to accomplish. Moreover, high variability in structural variation has been noted between haplotypes of the same individual, with several cases of inversions and duplications, and hemizygosity presence also in coding regions, with genes missing in at least one of the two haplotypes (Zhou *et al.*, 2019; Cochetel *et al.*, 2023).

To overcome this problem, the first attempts of genome assembly have focused on a quasi-homozygous line (approximately 93% of homozygosity) identified as PN40024, obtained by several generations of selfing of the cultivar Helfensteiner. The first few versions of the draft genome, obtained with Sanger sequencing, were highly fragmented and incomplete, presenting big gaps, unanchored contigs, and a high number of contigs per pseudomolecule (19 contigs per pseudomolecule on average) (Canaguier *et al.*, 2017). In the last few years, thanks to the use of long-reads technologies, haplotype-specific references have been assembled for several cultivated varieties, such as Malbec (Calderón *et al.*, 2024), Pinot Noir (Xiao *et al.*, 2025), PN40024 (Shi *et al.*, 2023) and many others (https://grapegenomics.com/pages/all.php).

3.2 Long read sequencing techniques

In the last decade, long read techniques have shaped the world of genome assembly and genomics in general, also thanks to the improvements in bioinformatic pipelines that are able to process and deal with such data (Kolmogorov *et al.*, 2019; Rhie *et al.*, 2021). Nowadays, two main long read technologies are widely used: PacBio HiFi and ONT long reads.

PacBio HiFi is a technology provided by Pacific Biosciences and has the advantage of generating the most accurate long read sequences (Sahu and Liu, 2023). This technology is based on the generation of a SMRTbell template (**Figure 1**): two hairpin adapters are attached to each fragment, which usually has a length between 10kbp and 30kbp, in order to generate a circular molecule. The circular template is then sequenced: this process, which happens inside a small well containing just one template, is based on the sequencing by synthesis technique. The incorporation of the nucleotide by the polymerase causes the emission of a fluorescent signal, which is then detected and associated to the specific base. The sequencing of the same template happens multiple times due to the circular structure of the molecule, generating multiple sub-reads belonging to the same fragment of origin. Subreads are then aligned, adapters are removed and a consensus sequence is obtained using the CCS algorithm. Reads obtained from this procedure are highly accurate, reaching in some cases even 99.9% of accuracy (Logsdon, Vollger and Eichler, 2020), and 99.5% of accuracy of homopolymers up to 5 bases long (Wenger *et al.*, 2019). The underlying principle that makes this strategy so effective is the largely stochastic nature of sequencing errors present within each read, that makes the consensus sequence obtained from multiple reads so accurate.

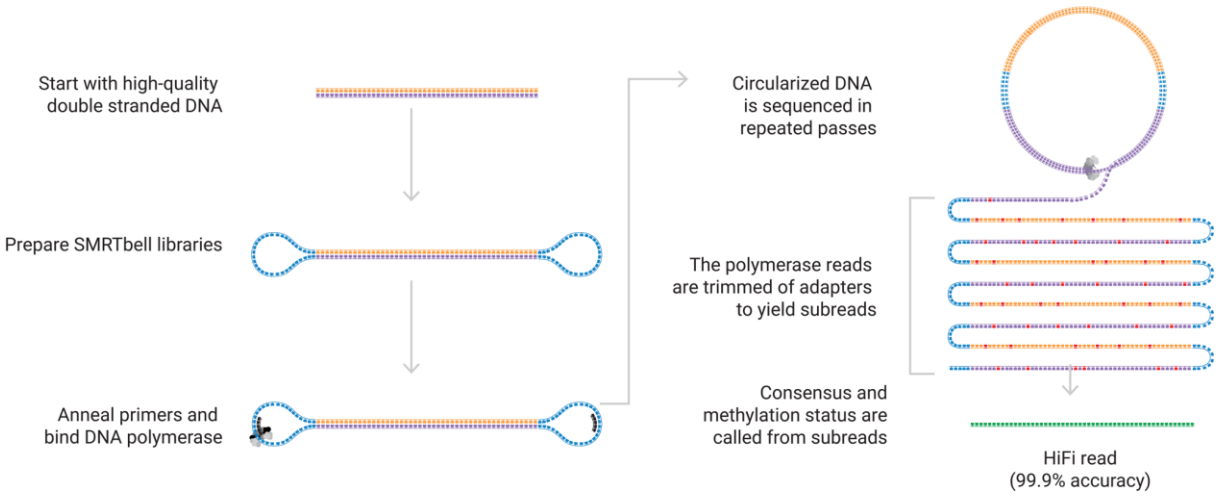


***Figure 1**: Main steps of the PacBio HiFi protocol: generation of the SMRTbell template, sequencing step and generation of the consensus sequence (image taken from www.pacb.com).*

The technology offered by Oxford Nanopore Technologies, on the other hand, has a lower accuracy, but is able to generate longer reads. This technology is based on the sequencing of DNA molecules using a nanopore attached to a membrane (**Figure 2 A**). High molecular weight DNA fragments are used to generate sequencing libraries by ligating adaptors with a motor protein (helicase) already attached to them. The motor protein will then be recognised by the nanopore complex, and will unfold the double helix of DNA, letting one of the strands pass through the pore at a specific speed. The passage of the different bases through the nanopore causes a variation in the ion flux going from one side of the membrane to the other through the pore, leading to a variation of the current: each nucleotide has a specific footprint in the ion current levels. The basecalling step is then performed by converting the signals and variations of the ion current into the sequence of nucleotides. Since the change in ion flux through the pore is generated by a different occupancy of the pore by the nucleotide, different measurements are expected also between modified and unmodified bases, allowing the detection of DNA modifications such as 5-methyl cytosines (**Figure 2 B**) (Liu *et al.*, 2021). Several studies have highlighted the power of this technology in generating reads that can reach lengths of hundreds of Kbp and have more than 99% accuracy (Rang, Kloosterman and de Ridder, 2018; Logsdon, Vollger and Eichler, 2020). Post-sequencing methods have also been developed in the last few years to improve accuracy of reads to values >Q20, at the expenses of reducing the length of final sequences (Stanojević *et al.*, 2024). Deep learning-based base calling and error correction methods have proven very effective in greatly increasing the accuracy of ONT reads, even in the absence of improvements in the biochemistry of the sequencing reactions.

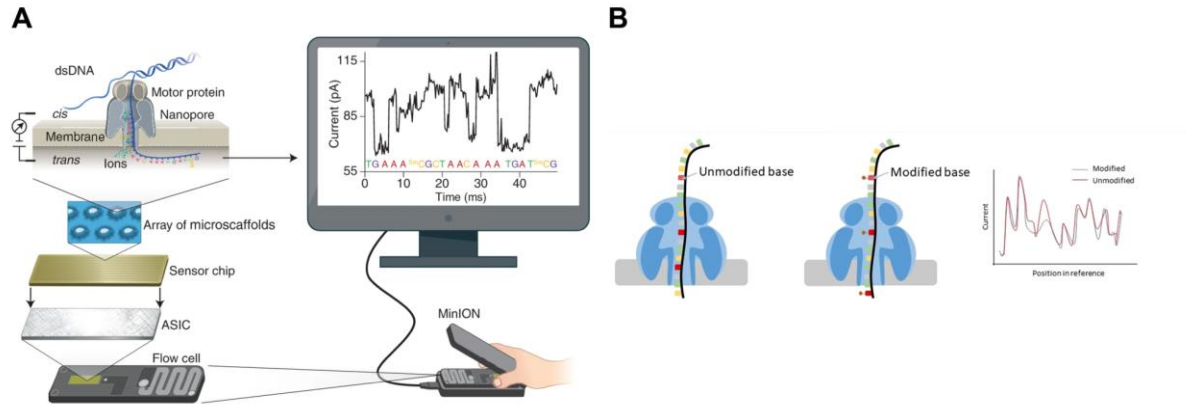


***Figure 2**: Nanopore sequencing method. **A)** Scheme of the flowcell of ONT sequencers and detection of the change of ion current through the pore; **B)** Variation in the ion current in two sequences with/without methylated cytosines. Image modified from Wang et al., 2021 and Xu and Seki, 2020.*

3.3 De-novo assembly techniques

To reconstruct the genome of a specific organism several approaches can be used, based on the technology that has been adopted to generate raw reads and the availability of previous versions of the genome. The most comprehensive and unbiased way of reconstructing genomes is to assemble them de-novo, even though several limitations can hinder the path to reach that goal.

With Sanger sequencing and NGS sequencing, two main approaches were used to perform genome assembly. The method applied to the first technology is relying on overlap-layout-consensus (OLC) paradigm (Pop, 2009), which implies the search of overlaps by performing an all-vs-all comparison of the sequences. Once all the overlaps of the sequences are found, the consensus sequence is built, merging sequences of overlapping reads. Despite the ability of this method of handling also relatively long reads, such as the ones generated with Sanger sequencing (approximately 700-800 bp), and span repeated sequences contained inside the reads, this method has a high computational cost. After NGS sequencing arrival, the shorter reads and higher throughput have led to a change in the genome assembly approach. Instead of using the OLC algorithms, in fact, de Bruijn graph (DBG) based methods have allowed a faster and more efficient way of assembling sequences (Mahadik *et al.*, 2019). In DBG approaches, reads are first divided in shorter k-mers, and then the graph is built based on the k-mers sequences. Despite the much faster computational time and better handling of vast amounts of sequences, this method is more susceptible to sequencing errors and repeats that are longer than the k-mer length.

Both techniques have strong limitations in getting long range information, which are necessary to reach completely assembled genomes. Because of this, the most efficient way of assembling genomes and to scaffold reconstructed contigs was to rely on multiple technologies, in order to exploit the advantages of each method. Hybrid methods often rely on whole genome sequencing methods (like NGS or Sanger) coupled with methods that can get long range interactions, to be able to scaffold and order contigs. Among these methods, BAC clones, mate-pairs and Hi-C are the most widely used.

In the first method, long sequences of the genome of interest (inserts that are 100-300kbp long) are inserted into bacterial single-copy plasmids. BAC templates are then inserted into bacteria: each bacterium takes up to 1 BAC molecule, generating then clones of the same molecule of origin. BAC clones are then fragmented and sequenced with NGS methods: reads coming from the same clone can give the information of the relative position of regions that are close to each other in the genome. Information from

different BAC clones is then integrated by using the known BAC overlap map (Venter, Smith and Hood, 1996).

In the second method, genomic sequences are fragmented to obtain sequences that are 2-5kb long. Fragments are then treated to biotinylate the ends and circularize the fragment. In this way, regions that are few kilobases far from each other are brought close to each other in the circular template. Subsequently, the template is fragmented into sequences that can be sequenced with standard NGS techniques, and biotinylated fragments (i.e. fragments that contain the two ends of the initial genomic fragment) are enriched. Adapters are then added, and the standard NGS sequencing protocol is followed. In this way, alignments against the previously assembled contigs can provide information about linked regions that are distant few kilobases to each other (Wetzel, Kingsford and Pop, 2011).

In the last method, interactions between regions that are spatially close to each other are fixed by crosslinking. Then, DNA is fragmented and marked with biotin. Subsequently, a ligation step is performed to merge fragments that were crosslinked and close to each other. Crosslinks are then removed, and the newly created fragments are enriched by selecting the ones that are biotinylated and sequenced with standard NGS protocols. By mapping the obtained reads, information about reads coming from physically close regions of the genome (using the distance-dependent decay principle, which implies a reduction of the measured interactions as genomic distance increases) can be used to order and orient previously assembled contigs (Ghurye *et al.*, 2019).

As mentioned earlier, none of the methods cited above are able to reconstruct an entire chromosome reference sequence from telomere to telomere (also called T2T). Moreover, being forced to use multiple experiments can lead to an increase in cost and longer times in processing all different types of data. With third generation sequencing technologies, genome assembly has become easier and faster. Several algorithms are able to generate telomere to telomere references by using just one type of technology (either PacBio HiFi or ONT long reads), even though hybrid methods are still needed sometimes. The most widely used assemblers rely mainly on OLC algorithms, optimized for very long reads. Examples of these assemblers are Falcon and Flye (Chin *et al.*, 2016; Kolmogorov *et al.*, 2019). Another group, more efficiently rely on optimized OLC and graphs. Among these we can find HiCanu, Verkko and Hifiasm (Nurk *et al.*, 2020; Cheng *et al.*, 2021; Rautiainen *et al.*, 2023).

Table 1: *Summary of the three main assembly methods.*

	OLC (classic)	DBG	String graph (Hifiasm)
Input reads	Long (any)	Short	Long (very accurate)
Basic unit	Whole reads	k-mers	Whole reads
Graph type	Overlap graph	De Bruijn graph	String graph (pruned OLC)
Overlap finding	All-vs-all alignment	Implicit in k-mers	Minimap2-style overlaps
Computational cost	Very high	Low	Low
Handles repeats	Moderately (with long reads)	Poor if repeat > k-mer	Well (uses phasing + long reads)
Output	Consensus (chimeric) assembly	Consensus (chimeric) assembly	Primary + haplotype assemblies

In particular, Hifiasm is one of the most used assemblers, especially if the main dataset is based on PacBio HiFi technology. Hifiasm is a graph-based assembler which is able to reconstruct haplotype-specific references starting from long reads datasets (**Figure 3**). The Hifiasm algorithm starts with an optimized version of an all-vs-all alignment of the reads in order to group them based on the haplotype of origin. Then, based on the grouping, sequencing errors present in the haplotype-specific reads are corrected. The error correction step is performed multiple times, in order to obtain high quality raw data to reconstruct the contig sequence. Corrected reads are then used to generate a string graph, in which each node corresponds to an oriented read, and each edge corresponds to an overlap between two adjacent reads. Giving the high quality of the reads, string graphs paths are easy to resolve, compared to the ones obtained with old OLC approaches. Moreover, string graphs obtained from Hifiasm contain bubbles corresponding to heterozygous regions, creating a lossless graph, able to store the information of both haplotypes. The generation of the contigs is then performed by following paths on the graph. Additionally, complementary information can be submitted to the software (such as ultra long reads, Hi-C reads or reads coming from Trios): by integrating one or more of these technologies, Hifiasm is able to reconstruct longer contigs and eventually attribute them to the parent of origin.

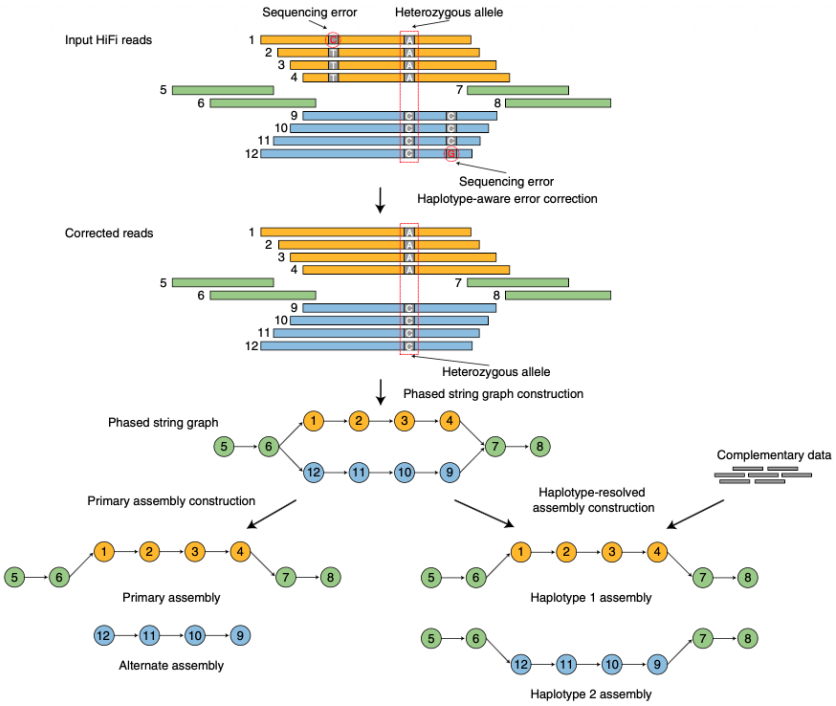


***Figure 3**: Hifiasm assembly approach to generate haplotype specific references. Initially, sequences are binned based on the presence of SNPs. Secondly, several stepst of read correction are performed, in order to remove sequencing errors in the reads. Lastly, a graph is generated, leading to a lossless reconstruction of the primary and alternative haplotype. Complementary data can also be added to phase the sequences in the two haplotypes. Image taken from (Cheng* et al.*, 2021).*

Regardless of the software and the technology used to assemble the sequences, the availability of a previous version of the reference genome can simplify and help the generation of a new assembly. Being able to map contigs on a previous reference can be crucial to orient and position them if Hi-C experiments or mate-pair information is not available. Additionally, having a previous version of the reference genome can allow the identification and assignment of unnamed contigs to specific chromosomes. Moreover, several studies have improved de novo assembly by using reference genomes of related species as a guide (Lischer and Shimizu, 2017).

3.4 Advances in long read sequencing for solving repetitive, structurally complex, and heterozygous regions

To reach the goal of assembling a complete, haplotype resolved genome, several factors can make the assembly process more difficult. Some of them will be described in this chapter, due to their implications in the correct understanding of assembly results.

The most common, and probably mostly widespread, hard-to-assemble regions are the repetitive ones. These can be more generally classified into two main categories: interspersed repeats and tandem repeats. The first ones are usually spread all over the genome and are exemplified by transposable elements (TEs). Transposable elements can be divided in two families, based on their intermediate: class I transposable elements (also called retrotransposons) have an RNA intermediate, which allows a copy-and-paste mechanism, generating an increase of the number of copies of these elements when active; these elements can then be divided based on the presence of the long terminal repeat (LTR) into LTR elements and non-LTR elements. The second family, called class II elements, has a DNA intermediate; this leads to the movement of the same transposable element with a cut-and-paste mechanism (also called jumping genes, due to the fact that they 'move' themselves from one part of the genome to another instead of copying themselves in several parts of the genome) (Wicker *et al.*, 2007). The second type of repeats (tandem repeats) can be exemplified by telomeric repeats, centromeric repeats, gene clusters (such as genes coding for rRNA) and microsatellites. Microsatellites are short tandem repeats (1-6 bp) generated by replication slippage. These regions are usually present in different parts of the genome in non-coding regions and are often used as genetic markers due to their high variability in populations and between individuals (Ellegren, 2004). On the other hand, tandem repeats such as the telomeric repeats and centromeric repeats are located in specific parts of the genome, respectively at the ends of each chromosome and in centromeres. Both telomeric and centromeric repeats can be composed of long stretches of strongly conserved and identical repeats, ranging from several kilobases up to few megabases long arrays (Naish *et al.*, 2021; Minton, 2024; Stephens and Kocher, 2024; Tao *et al.*, 2024). Other tandem repeats can be composed of longer sequences, such as genes families with hundreds of copies of the same gene. Examples of this situation are the gene clusters coding for ribosomal RNAs. These clusters, which can be present in multiple chromosomes in the same genome, are composed of several copies of the same gene, often present in almost identical copies, forming tandem arrays that can reach a total length of several megabases in the whole genome (Nurk *et al.*, 2022).

Another factor that can affect the accomplishment of generating a complete and correct reference genome is the presence of structurally complex regions, such as the ones composed of structural variants (SVs) like inversions, duplications, deletions/insertions and translocations, and all their combinations. Being able to reconstruct these regions is essential to understand and find the correct structure of genome features, since SVs are often associated with characters of interest, like a big inversion causing a change in fruit shape in specific peach cultivars (Guan *et al.*, 2021), a *Gret1* retrotransposon insertion in a promoter, causing a variation in berry colour in grapevine (Kobayashi, Goto-Yamamoto and Hirochika, 2004), or a complex translocated and inverted sequence which is responsible of the change in colour pattern of some insects (Gompert *et al.*, 2025).

Repetitive regions and structural variants are often hard to assemble due to the lack of capacity to include these regions in one whole sequence (i.e. variants contained inside a single read or gapless sequence). Additionally, repeated regions in the genome might induce assemblers to generate errors derived from mis-joining of sequences that are not adjacent or close to each other. In fact, having just part of a repeated region present in one sequence leads to the uncertainty of the exact location of origin, forcing to discard that information in order to avoid a wrong placement of the fragment in the reconstructed assembly or the wrong alignment of the read in the reference sequence. In the last decade, long read sequencing have allowed to mitigate this problem. Long reads such as PacBio HiFi reads and Oxford Nanopore long reads have allowed the spanning of such complex regions (Kovaka *et al.*, 2023): long reads are either able to contain completely some complex structures or repetitive regions, or phase more easily the small variants that might be present in these highly similar regions, by containing anchoring points for the repetitive parts (Mahmoud *et al.*, 2024; Stevanovski *et al.*, 2025).

Being able to phase variants is useful not only to solve complex regions, but also to correctly separate the information from the different haplotypes. In fact, a low phasing power does not only translate into a low ability of reconstructing complex structures, but also in not being able to determine differences in structures between two haplotypes. Some studies have already showed the potential of long reads to detect errors in reference genomes and separate heterozygous variants (Li *et al.*, 2023).

3.5 Characterization of newly assembled genomic regions: centromeric regions

The ability to reconstruct more complete, telomere to telomere assemblies is not only fundamental for having more reliable results from analyses that rely on a reference genome, but has also allowed the study of regions that were previously unexplored due to their complexity, such as centromeric regions (Altemose *et al.*, 2025). In fact, centromeric regions are one of the hardest parts of the genome to assemble and study, due to the presence of large tandem repeat arrays and transposable elements (Naish *et al.*, 2025).

Centromeres are essential for chromosome segregation. They are involved in the interaction and assembly of the kinetochore complex: a protein structure that interacts with the microtubules during cell division (McKinley and Cheeseman, 2016). The positioning of the kinetochore on the chromosome is determined by the presence of a specific histone variant, called CENH3 (CENP-A in mammals) (Earnshaw and Rothfield, 1985; Talbert *et al.*, 2002). CENH3 histones are involved in the recruitment of centromere-associated network (CCAN), which constitutes the inner kinetochore. This complex then interacts with the outer kinetochore (KMN proteins) that then come into contact with microtubules (**Figure 4 A**) (Yu *et al.*, 2000; Yatskevich *et al.*, 2024).

Even though their function is highly conserved across kingdoms, centromeres can be very different both in terms of structure and size (Melters *et al.*, 2013). One way of classifying centromeres can be based on the number of kinetochores and their distribution along the chromosomes.

- With this type of classification, one class of centromeres are those that form monocentric chromosomes. This class of centromeres is the one present, for example, in humans (Altemose *et al.*, 2025), and some plant species such as *Arabidosis*, Maize and grapevine (Chen *et al.*, 2023; Shi *et al.*, 2023; Naish *et al.*, 2025). This architecture is characterized by the presence of satellite arrays localised in a single chromosomal region (**Figure 4 B**), resulting in one chromosome restriction during metaphase (Melters *et al.*, 2013). Centromeric regions, in this case, are often composed of large arrays of tandem repeats, forming higher order repeats. These regions can also be populated by centrophilic TEs, which increase the structural complexity generating hard to assemble patterns of tandem repeats intermixed with transposable elements (Wlodzimierz *et al.*, 2023). Specific methylation patterns have also been charaterized in these regions, with higher level of methylation in centromeric compartments compared to the chromosome's arms, even though they can differ in patterns between

species. In primates, for example, centromeric regions tend to be highly methylated in the CG context, with relatively small windows of unmethylated cytosines, correpsonding to the binding site of the kinetochore (Logsdon *et al.*, 2024). On the other hand, *Arabidopsis* centromeres tend to be highly methylated in all 3 contexts (CG, CHG, CHH) with a depletion of methylation in the active site more emphasised for the CHG context compared to the CG one (Wlodzimierz *et al.*, 2023; Naish *et al.*, 2025).

- The class of centromeres wich is opposed to the monocentric one, is the one found in holocentric chromosomes. Holocentric chromosomes are characterized by multiple points of kinetochore attachement along the whole chromosome (Melters *et al.*, 2012). Examples of this class can be found in some insects (Drinnenberg *et al.*, 2014) and plant species such as beak-sedge and *Chionographis japonica* (Hofstatter *et al.*, 2022; Kuo *et al.*, 2023). The DNA sequence of holocentromeric chromosomes has several small arrays of centromeric repeats evenly spaced along the chromosome (**Figure 4 C**), with CENH3 enrichment in the core of each array (as happens in *Rhynchospora* species with the *Tyba* arrays and specific TE families, such as *Ty3/gypsy* LTR retrotrasposon CRRh (Marques *et al.*, 2015)). Contrarily to what happens in the monocentric chromosomes, there is no compartimentalization of the centromere from an epigenetic point of view. Nonetheless, fine scale regulation of the methylation has been observed, with high levels of methylation in the CG context corresponding to each repeat array (Hofstatter *et al.*, 2022).
- Another class is also present, with an intermediate structure of chromosomes between the monocentric and the holocentric ones. Metapolycentric chromosomes are in fact characterized by a single widely expanded chromosome constriction during metaphase, composed of multiple distinct repeat arrays, in which just a subset is occupied by the centromeric histon specific variant, and forms one big kinetochore locus (**Figure 4 D**). This type of structure has been observed in leguminous plants, such as the ones belonging to the genus *Pisum* (Macas *et al.*, 2023). Interestingly, in the huge centromeric region, actively transcribed genes were found in between the tandem repeats arrays. Methylation patterns remain pretty similar between centromeric regions and the chromosome arms, with a uniform depletion in methylation levels of CHG and CHH contexts in regions corresponding to specific satellite DNA arrays families (Macas *et al.*, 2023).

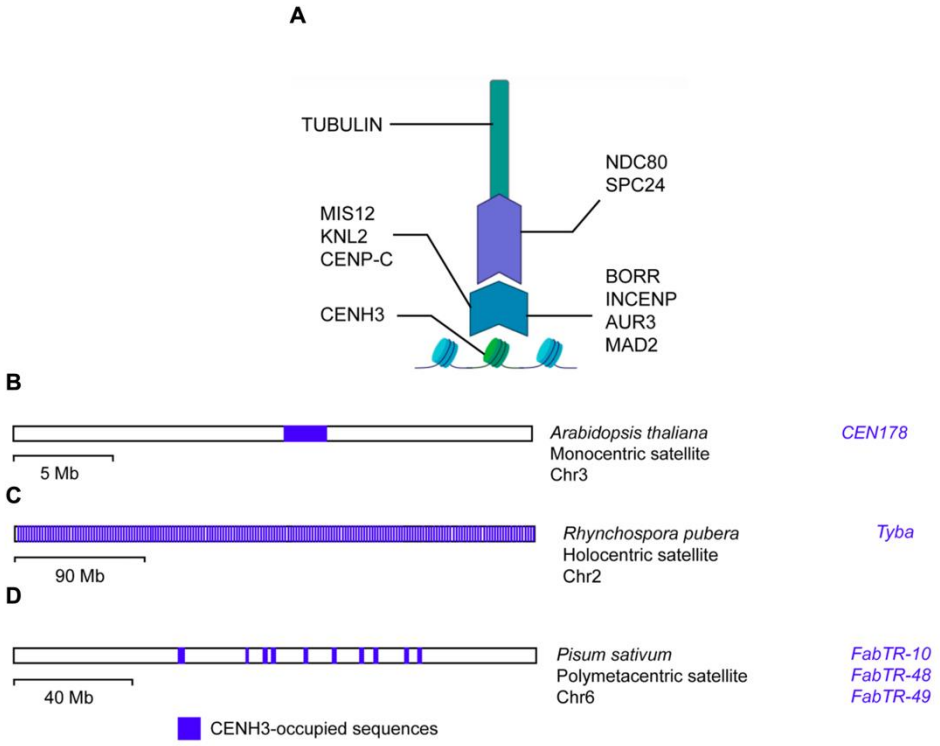


***Figure 4**: Centromere kinetochore in plant species. **A)** Kinetochore complex, interacting with histone specific variants of the centromere (CENH3) and tubulin fibres; **B)** monocentric structure in in Arabidopsis; **C)** holocentric structure in Rhynchospora; **D)** metapolycentric structure in Pisum. Images taken and modified from Naish and Henderson, 2024.*

In grapevine, some studies have tried to identify the composition of centromeric regions. Preliminary studies have focused on ChIP-seq and FISH experiments, relying mostly on these datasets rather than on the full telomere to telomere sequence of chromosomes (Pei *et al.*, 2024). The sequence of centromeric repeat families and their distribution along the chromosome have then been confirmed with the first T2T references and pangenomes (Shi *et al.*, 2023; Liu *et al.*, 2024). Several repeat families have been identified: the most widespread is the 107 bp monomer, followed by multiples of the monomers. The 107-bp repeats were organized in big arrays, ranging from approximately 1kbp up to several megabases, and were the most abundant in all chromosomes but chromosome 3, which was populated by a 135-bp repeat highly similar

to the 107-bp repeat, chromosome 14, which was dominated by 187-bp repeats, and chromosome 18, in which the 66-bp repeat family was the predominant one. In all chromosomes, multiples of the base units were present. Moreover, several TEs were intermixed to centromeric tandem repeats, such as TEs belonging to the *Gyspy* and *Copia* families (**Figure 5**) (Shi *et al.*, 2023).

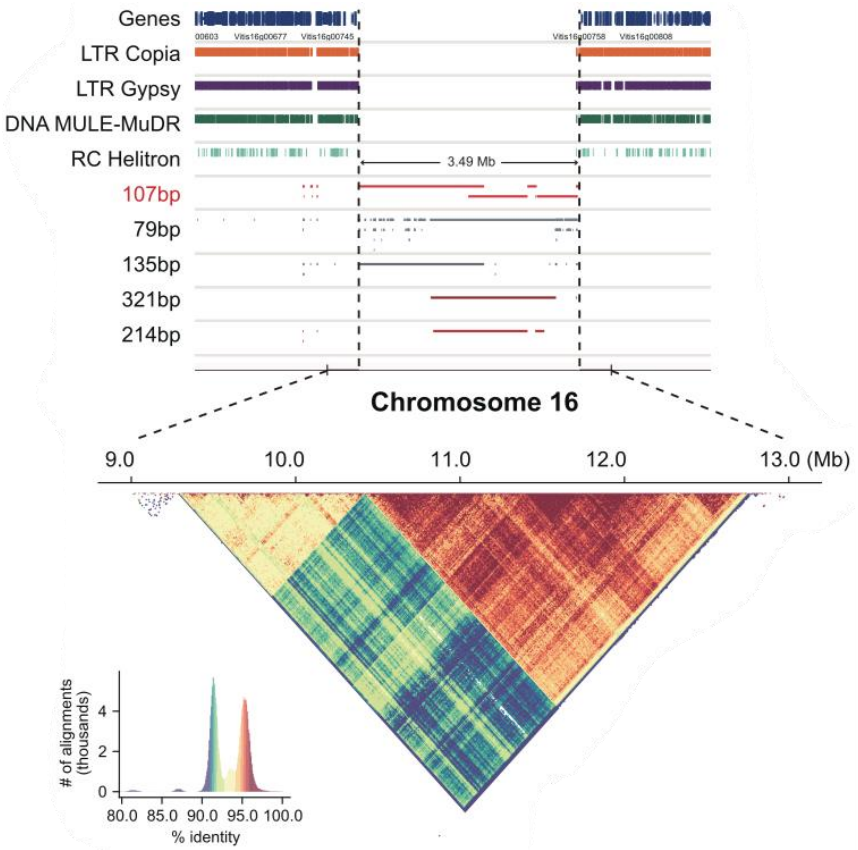


***Figure 5**: Composition and structure of grapevine centromeric region. Tracks describing the centromeric landscape are present in the top part of the image, including Genes, different TEs families and centromeric tandem repeats. In the bottom part, self-identity plot of the centromeric region. Image taken from (Shi* et al.*, 2023)*

3.6 Overview of the selected grapevine varieties and their relevance

In this study, 6 different accessions of grapevine have been analysed in order to get a sufficiently complete and comprehensive view of the differences that can be found in the genomes of cultivated and non-cultivated varieties.

PN40024 is a grapevine accession derived from several steps of selfing of the Helfensteiner cultivar. The high levels of homozygosity of this accession have made it the most widely used for the first genomics studies on the grapevine genome. Because of this, several versions of the reference genome and gene annotations are publicly available for PN40024 (https://grapedia.org/files-download/). Despite its wide presence in genomics studies, PN40024 cannot be used as a case example for the diversity and the actual variety of structures that can be found in grapevine genomes of cultivated varieties.

Consequently, we added two cultivars widely used for wine making: Cabernet Franc and Rkatsiteli (Tabidze *et al.*, 2017; Minio, Cochetel, Vondras, *et al.*, 2022). Being able to reconstruct haplotype specific assemblies in such highly heterozygous samples is essential for understanding the actual variation that can be present in grapevine accessions, and to phase it in the two distinct haplotypes.

Grapevine is widely used not only for wine making: a vast portion of grapevine cultivars are used for making table grapes, which can differ from the "wine" ones for several characteristics, such as berry size, absence of seeds and other sensory characteristics (Di Genova *et al.*, 2014). Among these cultivars, we selected the Sultanina variety, providing us information from the intra-specific differentiated group of table grapes, belonging to *Vitis vinifera L. ssp. vinifera* (Aradhya *et al.*, 2003).

To complete the coverage of all branches of the *Vitis vinifera* species, we selected one cultivar belonging to the *sylvestris* subspecies (hereafter *V. sylvestris*). *V. vinifera*, which contains all the cultivated varieties, has diverged from its wild progenitor *V. sylvestris* ~8000 years ago, during the domestication process (Myles *et al.*, 2011; Magris *et al.*, 2021).

Lastly, interspecific hybrids are widely used as rootstocks in grafted grapevine plants, mostly being generated by crosses with North American *Vitis* species. These species in fact have co-evolved with phylloxera (a pathogen affecting grapevine roots) gaining resistance against this disease (Dodson Peterson *et al.*, 2019). In our study we considered an interspecific hybrid between *V. berlandieri* x *V. riparia*, called Kober 5BB (Minio, Cochetel, Massonnet, *et al.*, 2022).

The different accessions that have been included in this project gives us the opportunity to get a view on the differences and similarities that are present between these 3 grapevine species.

4 Objectives of the PhD project

The PhD project aims to generate a case study of structural and functional annotation of genomes, using long-read technologies. Long-read technologies have allowed to easily study and uncover regions that were inaccessible due to limits in read sequences technologies and algorithms. In this project, we used cutting-edge long read sequences to assemble genomes of 6 grapevine accessions at a haplotype specific resolution. We reconstructed most of the haplotypes from telomere to telomere, generating highly complete reference genomes. Moreover, long reads containing the information of the methylation state of cytosines have given us the opportunity to study haplotype specific methylation patterns.

Additionally, the high completeness of the assembly has also allowed to uncover, investigate and describe regions that were previously unexplored, revealing highly repetitive regions, embedded in highly structured patterns present in centromeric regions. The comparison of these regions between cultivated and wild accessions has also enabled the study of evolution of these elusive regions, bringing out similarities and differences present both within individuals and between haplotypes coming from different accessions.

Moreover, to get functional annotation of the newly assembled genomes, gene prediction has allowed the generation of haplotype specific gene models, allowing the comparison of functional elements between haplotypes belonging to the same individual.

By integrating genomic data coming from different long reads technologies, the PhD project has allowed to get a more complete and comprehensive view both on the structures of grapevine genome and its functional elements, shedding a light on several aspects, ranging from the actual DNA sequence to the epigenetic level (methylation).

This study represents one of the first examples of a complete workflow based on third generation sequences technologies to study genomic variation and variability both within and across individuals of closely related species, at structural and functional levels.

5 Materials and methods

5.1 Sample preparation and sequencing

PacBio HiFi sequencing of PN40024, Cabernet Franc and Rkatsiteli DNA was outsourced to SciLifeLab, National Genomics Infrastructure (NGI), Uppsala, Sweden. For PN40024, Cabernet Franc and Rkatsiteli, 18 Gb, 26Gb and 33Gb of sequence were generated respectively (see **Table 2** for details). Sultanina PacBio HiFi reads were downloaded from project CNP0004225 present on China National GeneBank DataBase (https://db.cngb.org), while VS-1 PacBio HiFi reads were downloaded from DVIT2426 and Kober 5BB reads were downloaded from NCBI (Run ID: SRR20810425) (see **Table 2** for details).

Long reads obtained with Oxford Nanopore Technologies were generated for PN40024, Rkatsiteli and Cabernet Franc samples at Institute of Applied Genomics, Udine, Italy (see **Table 3** for details). For PN40024, 3.6Gb of sequence were generated, with an average size of 5,6Kb but with a minority of reads exceeding 30Kb. For Cabernet Franc, ultra-long reads with 5mCG context call were generated, leading to 30Gb of sequence. For Rkatsiteli, sequencing of ultra-long reads with 5mC (both mCG, mCHG and mCHH) led to the generation of 27Gb of sequence, with an average read length of 26Kb and 20% of the cumulative yield composed of reads with length > 80Kb and up to 200Kb.

***Table 2**: Summary of sequencing statistic of the PacBio HiFi reads.*

	#of reads	Cumulative length (Gb)	Coverage (coverage per haplotype)	Avg. size (kb)
PN40024	1178383	~18	35X	15.4
Cabernet Franc	1841792	~26	49.4X (24.7X)	14
Rkatsiteli	2058922	~33	65X (32.5X)	16
Sultanina	3459832	~60	116.4X (58.2X)	17
VS-1	2161533	~27	54.4X (26.2X)	12
Kober 5BB	1992558	~23	44.8X (22.4X)	11

***Table 3**: Summary of sequencing statistic of the ONT reads.*

	#of reads	Cumulative length (Gb)	Coverage (coverage per haplotype)	Avg. size (kb)
PN40024	646013	~3.6	14X (7X)	5.6
Cabernet Franc	1610000	~30	56X (28X)	18
Rkatsiteli	1067865	~27	54X (27X)	26

5.2 Genome assembly and scaffolding

PacBio HiFi reads were then used to perform genome assembly with Hifiasm assembler (version 0.16.1-r375) (Cheng *et al.*, 2021) with default settings, generating haplotype specific contigs for each sample. Contigs from PN40024 were then aligned to the a previous version of the reference genome (12Xv0; GCF_000003745.3) using NUCmer software (MUMmer4 (Marçais *et al.*, 2018)) to order and orient the contigs into the 19 chromosomes. Gaps between chromosomes were then filled after re-aligning PacBio HiFi reads against the new draft version of the genome (check next section "DNA reads alignment"), and eventually filling the gaps between adjacent contigs with read sequence. For the remaining accessions, contigs belonging to the two haplotypes were aligned independently against the newly assembled genome of PN40024 using NUCmer (MUMmer4 (Marçais *et al.*, 2018)), to order and orient them into the 19 chromosomes version of the assembly. Supplementary contigs were included in the final assembly versions but were not used to perform further analyses.

5.3 Evaluation of completeness of the assembly (statistics, telomere presence and rDNA clusters)

Statistics of assembly completeness were obtained using QUAST version 5.3 (Gurevich *et al.*, 2013). Telomere presence was checked using tidk version 0.2.31 (Brown, Manuel Gonzalez de La Rosa and Blaxter, 2025), looking for grapevine telomeric repeats in the last and first 10000bp of each pseudomolecule (-w 10000 and -c Rosales). rDNA gene clusters were firstly identified in PN40024 using BLAST version 2.2.28 (Altschul *et al.*, 1990) and the sequence of *Arabidospis thaliana* 5.8S and 25S subunits (NCBI sequence

X52320.1) as query. After the identification of the cluster's locations in PN40024, one monomer for each subunit was extracted and used as template to find the corresponding cluster in the remaining accessions.

5.4 Comparison of the chromosome version of the newly assembled haplotypes

Comparison of the newly assembled haplotype-specific references (chromosome level) have been performed with SyRI version 1.7.1 (Goel *et al.*, 2019). Alignments obtained with minimap2 version 2.30-r1287 (Li, 2018) with asm5 preset have been used as input for syri command, to detect syntenic regions, as well as translocations and inversions. The output was then used by Plotsr version 1.1.1 (Goel and Schneeberger, 2022) to check regions annotated as syntenic, translocated or inverted, that are longer than 25Kb (-s 25000) in a graphical way (--itx and -R were also used as additional parameters).

5.5 Transposable elements annotation

Transposable element annotation was performed using EDTA version 2.2.2 (Ou *et al.*, 2019). A curated library of TE was given to EDTA to include grapevine specific TEs, as well as fasta file containing coding regions. Additional parameters --sensitive 1 --species others were also used. To estimate TEs abundances in the genome, EDTA output was used to calculate fraction of bases masked as TEs in 5kbp windows using BedTools version v2.31.1 (Quinlan and Hall, 2010) with the command bedtools intersect. Additionally, masking of Athila elements as well as TE families containing chromodomains in the centromeric regions was performed using Repeatmasker tool, version 4.0.6, with a curated dataset of TEs.

5.6 DNA reads alignment

PacBio HiFi reads and ONT long reads were aligned using minimap2 version 2.28-r1209 (Li, 2018). For PacBio HiFi reads the preset map-pb was used to optimize alignments score to the specific long read technology. Similarly, for ONT reads, the preset that was used was map-ont, together with -y to pass additional tags containing methylation information. For Cabernet Franc and Rkatsiteli, ONT reads were aligned independently to the two haplotypes (sequential alignment), filtered for alignment quality -q 60 and secondary or supplementary alignments (nor primary) -F 256 using SAMtools version

1.15.1 (Danecek *et al.*, 2021). Filtered alignments were then phased and attributed to the corresponding haplotype after SNP calling and phasing with Duet, version 0.6 (Zhou *et al.*, 2022) (clair3 for SNP calling) and haplotype attribution with WhatsHap (Martin *et al.*, 2016). Reads coming from homozygous regions were attributed to both haplotypes.

5.7 DNA methylation levels

Methylation levels were calculating starting from haplotype specific read alignments of ONT long-reads. Alignments from each haplotype were treated independently. Modkit tool, version 0.4.0 (https://github.com/nanoporetech/modkit), was used to call methylation levels for each cytosine. For the CG context, filtering threshold for the methylation calling was set to --filter-threshold 0.75; usage of both strands was activated with --combine-strands as well as the --combine-mods parameter, to combine signals from 5mC and 5hmC. Values for positions with less than 4 reads of coverage were discarded. For CHG and CHH contexts, similar parameters were used: --filter-threshold 0.75 –combine-mods. Values for positions with less than 10 reads of coverage were discarded. Methylation levels of single cytosines were then grouped using BedTools version v2.31.1 (Quinlan and Hall, 2010) to calculate mean levels for 5 kbp windows.

5.8 Definition and characterization of centromeric regions

Centromeric regions in PN40024 were first localized with the use of a putative 107 bp monomer, previously identified (Di Gaspero, G. & Foria, 2015), using BLAST tool (Altschul *et al.*, 1990). After the identification of tandem repeat arrays of the major repeat family, a more comprehensive classification and characterization of the surrounding area were performed manually, adding 13 additional different tandem repeat families, using dotplots generated with Dotter (Sonnhammer and Durbin, 1995). Masking of the newly defined tandem repeat families, together with the TE families of chromodomain-containing TEs (check previous chapter) was performed with RepeatMasker version 4.0.6 (Smit, Hubley and Green, 2015) for all accessions. To define the putative centromeric regions, the first match and last match of the 5 main repeat families (107 bp, 66 bp, 28 bp, 58 bp, Gyp-28) and Athila elements were used, integrating also the evidence of methylation levels. To obtain a highly reliable results, centromeric regions were then analysed and considered just for haplotypes of chromosomes containing the putative regions in a single contig (no contig interruption), while all chromosomes were used for the PN40024 accession, despite interruptions in the continuity of centromeric sequences. Characterization of the defined regions from a

structural point of view was performed using ModDotPlot tool, version 0.9.8 (Sweeten, Schatz and Phillippy, 2024). Both self-comparisons as well as comparisons between different haplotypes were performed using windows of 1 kbp (-w 1000) or 2 kbp (-w 2000) and plotting identity values >70% (-id 70).

5.9 Monomer composition and higher order repeats (HORs) presence in 107bp repeat arrays

HOR identification has been investigated using CENdetectHOR pipeline (Daponte *et al.*, 2025). Centromeric regions of the 19 chromosomes of PN40024 have been used as input to CENdetectHOR software. Several combinations of most of the options have been tried out to find the optimal combination for the detection of the HOR present in the 107 bp arrays. The final setting was the following: consFile option was set to "no" to investigate in an unbiased way the composition of the various repeat families of the centromeres; CHRcons was set using the centromeric sequence of chromosome 01 to extract the consensus sequences; windowSize was set to 350 to get a better resolution on the composition of the windows; kmerSize was set to 8; maxHORgap was set to 30; maxHORlength was set to 10; minHORrep was set to 4 and discreteTree was set to "True". After the identification of the putative monomer length, regions having the 107 bp monomer (or its HORs) as the predominant repeat were selected, in order to consider just the windows composed of the 107 bp repeat family. The file named "centroseq.FULLchr_mons.fasta" containing the sequences of the monomers extracted from all the chromosomes was then used to generate a multi-alignment using the online version of Mafft software (Katoh, Rozewicki and Yamada, 2019), setting the following parameters: --adjustdirectionaccurately to orient sequences based on the best alignment; --allowshift and --unalignlevel 0.1 to allow blindly the presence gaps in the alignemnts; the parameters --maxiterate 2 --retree 1 –globalpair were used to optimise iterative refinement of the alignments. Subsequently, trees were built with Neighbor-Joining (NJ) algorithm based on pairwise distances. The tree file was then plotted using the online version of iTOL (Letunic and Bork, 2024). A similar pipeline was used for Rkatsiteli hap1 centromeric regions. The same parameters used for PN40024 samples for the CENdetectHOR pipeline were used also for the centromeric sequences of Rkatsiteli, with the only exception of the file used to extract the sequences of the monomers. In this case, the file that was used to extract the sequences (consFile) was the one generated by the CENdetectHOR pipeline ran on PN40024, in order to extract the 107 bp monomers starting with the same phase (i.e. same initial sequence). To generate the comparison of all the sequences extracted from both PN40024 and Rkatsiteli hap1,

the two files containing the sequences of the monomers were merged and given to Mafft for the multi-alignment: the parameters modified for the alignment were the same as the ones used previously, for the PN40024 sample alone. The tree file was then generated calculating the alignment score, calculating the rough distance and using the average linkage (UPGMA), due to high number of sequences analyzed. Also in this case, plots were then generated using iTOL.

5.10 Centromeric evolution analyses

To study centromeric evolution, the presence of intact transposable elements has been used. A list of intact TEs containing LTRs was taken from EDTA annotation. Additional intact LTR/TEs that were not annotated automatically were manually selected by performing a BLAST search, taking the first of the two LTR sequences of the annotated TEs as query sequence. The matches were then filtered by taking alignments that were matching the first 100 bases (start base = 1; end base ≥100) with an identity > 91%. An additional filtering was then performed by checking the correspondence of the 2 TSDs sequences of each TE. Moreover, TEs included in regions with structural variations (duplications/inversions) were compared to identify TEs that originated from the same insertion event: to do so, both LTRs similarity as well as complete correspondence of TSDs sequences was required to annotate TEs as elements derived from segmental duplications/inversions and originated from a unique insertion event. The expanded list of intact TEs was then used to date the insertion time. Pairs of LTRs coming from the same TE were used as molecular clock: comparison of the LTRs sequences led to the identification of the number of differences between LTRs, and a constant substitution rate of 1.3E-08 was used to estimate the time of TE insertion (when both LTRs were supposedly identical) as in Magris, 2016. Search of soloLTRs was also performed by comparing TSDs concordance in sequences similar to LTRs. Two soloLTRs were considered to originate from the same excision event when their TSDs were identical. In this case, their presence in multiple genomic locations is interpreted as the result of subsequent duplication or inversion events, rather than independent soloLTR formation events.

5.11 CENH3 data analyses

To investigate the functional part of the centromeres, ChIP-seq data targeting CENH3 variant were aligned against PN40024 genome. Publicly available data targeting CENH3 coming from leaf samples of Pinot Noir (NCBI project ID: PRJNA897873) and Chardonnay (NCBI project ID: PRJNA1021789) varieties were used to check CENH3 distribution. Both enriched sample and Input (control) data were first trimmed and filtered with Fastp, version 0.23.2 (Chen *et al.*, 2018), using the following parameters: -q 20 --length_required 50). Reads were then aligned using Bowtie2 aligner, version 2.4.1 (Langmead *et al.*, 2009), with default settings. Alignments were then processed as in (Logsdon *et al.*, 2024) by filtering with SAMtool version 1.7 (Danecek *et al.*, 2021) for unique alignments (-F 2308) in order keep just one alignment per read (reads mapping to multiple places are randomly assigned to one location). To calculate enrichment of the sequencing experiment with CENH3 compared to the Input, bamCompare command (deepTools2 version 3.1.1 (Ramírez *et al.*, 2016)) was used setting the following parameters: --operation ratio --binSize 1000 --minMappingQuality 1.

5.12 Gene prediction

Gene models were generated for Rkatsiteli and Cabernet Franc haplotypes using BRAKER3 version 3.0.8 (Stanke *et al.*, 2008). Paired-end data from RNA-seq experiments on tendrils, leaves and berries (NCBI project ID: PRJNA373967) were given as input to the software, together with the soft masked version of the reference genome with TEs annotation generated by EDTA in previous steps. Braker was then ran with default parameters and setting --species=vitis_vinifera --prot_seq=/orthodb/Viridiplantae.fa and --useexisting. Additionally, the fraction of bases covered as coding sequences (CDS) was calculated by fist filtering for "CDS" label in the output GTF file from Braker, and then using BedTools version v2.31.1 (Quinlan and Hall, 2010) to intersect the data and calculate the cumulative length of bases annotated as CDS in 5kbp windows.

6 Results

6.1 Raw assembly statistics and completeness of the assembled regions

The genome assembly step performed with PacBio HiFi reads using Hifiasm generated haplotype-specific contigs for all the analysed samples. Raw assembly statistics are present in **Table 4**. On average, for each haplotype 512493137 bp were assembled, with the largest cumulative assembled sequence corresponding to haplotype 1 of Sultanina (536124753 bp) and the shortest one corresponding to haplotype 2 of Rkatsiteli (493356887 bp).

Haplotigs were then ordered, oriented and merged using the old version of the reference, 12xV0 (GCF_000003745.3) as a guide. Gaps of 500bp filled with "N" were inserted as spacers between adjacent haplotigs. The scaffolding resulted in pseudomolecules corresponding to chromosomes, specific for each haplotype of the 6 analysed varieties. For 76 haplotypes, out of 209 (19 chromosomes with 2 haplotypes for 5 accessions and one haplotype for PN40024), the whole chromosome was covered by a single haplotig, from telomere to telomere (**Table 4**), while in 158 haplotypes, both telomeres were anchored to the scaffolds composing the chromosome. In more than 85% of the haplotypes (182/209), chromosomes were assembled in 3 haplotigs or less. The most fragmented assemblies were the two haplotypes of VS-1 (*Vitis sylvestris*) which required 74 and 71 haplotigs to assemble hap1 and hap2, and had L50=13 and L50=15 respectively.

In all the accessions, the upper end of the pseudomolecule belonging to chromosome 15 and the bottom end of the pseudomolecule belonging to chromosome 17 were interrupted by the presence of 35S rDNA arrays, which caused the termination of the haplotigs due to the repetitiveness of the sequence.

***Table 4**: Table containing raw assembly statistics obtained with Quast. Each column corresponds to one of the haplotypes reconstructed for each accession.*

Assembly	PN40024	Cabernet Franc - hap1	Cabernet Franc - hap2	Rkatsiteli - hap1	Rkatsiteli - hap2	Sultanina - hap1	Sultanina - hap2	VS-1 - hap1	VS-1 - hap2	Kober 5BB - hap1	Kober 5BB - hap2
# contigs	345	804	378	398	155	709	346	1185	1112	260	137
Total length (bp)	506468382	525806295	523130503	516639849	493356887	536124753	498535378	513694318	504748965	509433796	509485377
Largest contig (bp)	36581781	31333014	29307966	37749015	38871647	39359192	30531888	30326283	23119584	32020157	30370246
GC (%)	35.34	35.56	35.46	35.32	35.13	35.34	35.24	35.14	35.16	35.33	35.04
N50	26359317	14870658	16058559	24768353	24754510	25311536	21773119	10014007	9828651	16893797	19396665
N90	20048913	2463467	3683947	7694154	13686810	19116095	8755454	1671837	1534255	5269626	6798536
L50	9	13	12	10	9	10	10	15	17	12	11
L90	18	39	31	22	19	19	21	56	60	32	28
Chr with 2 anchored telomeres	17	14	14	16	15	16	15	12	17	10	12
T2T in one contig	14	4	3	14	10	14	10	2	1	3	1

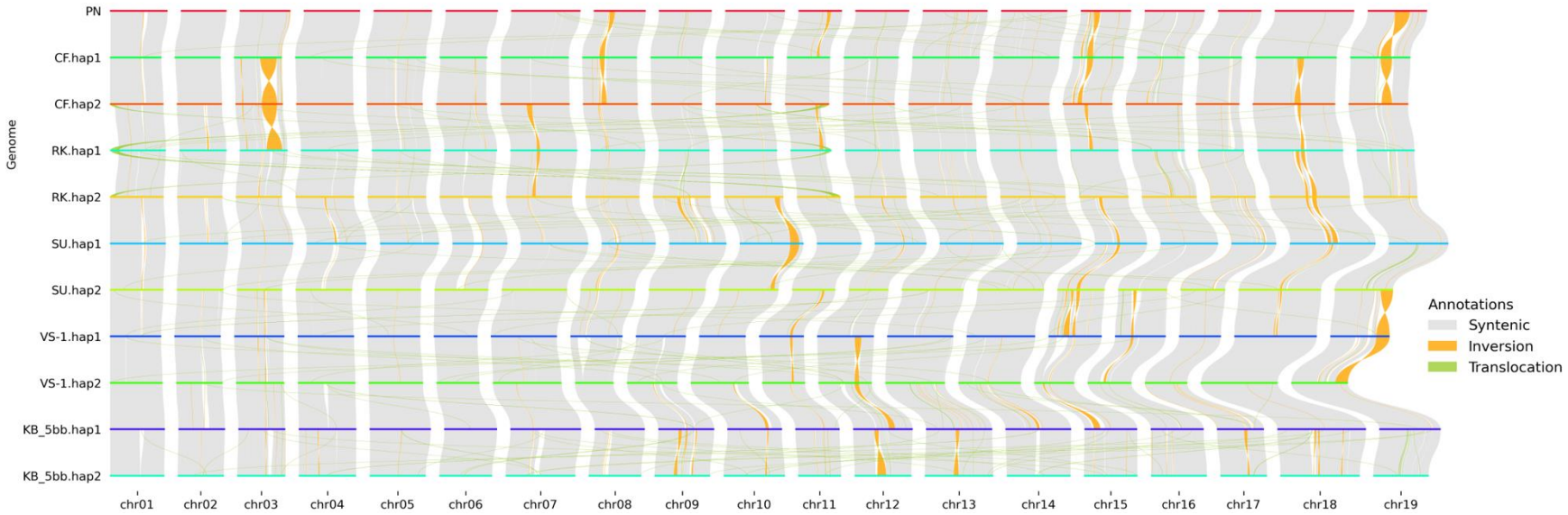


***Figure 6**: Synteny plot of the haplotype-specific references of the 6 accessions. Several structural variants are visible at a chromosomic scale, such as inversions on chr03, chr10 and chr19, and translocation between chr01 and chr11 of Rkatsiteli. Structural variants have been filtered for length > 25 kbp. PN=PN40024; CF=Cabernet Franc; RK=Rkatsiteli; SU=Sultanina; VS-1=Vitis sylvestris VS-1; KB_5bb=Kober 5BB.*

The high completeness of the assembly and its contiguity have allowed the comparison, at a chromosome level, of the pseudomolecules. Synteny plots between all the haplotypes have highlighted large structural variations between the accessions (**Figure 6**). Several inversions that are few Mbp long are present on chromosome 03 of Cabernet Franc hap2, chromosome 19 of Cabernet Franc hap1 and VS-1 hap1. Moreover, the comparison of the chromosomes has shown up a translocation event present in the Rkatsiteli accession between the beginning of chromosome 01 and the end of chromosome 11.

Several smaller inversions as well as interruptions in the synteny between sequences are mostly occurring in centromeric regions. On chromosome 15, inversions are present between Sultanina and VS-1, between Cabernet Franc and PN40024 and structural variation is emphasized by the presence of a big gap in the synteny between the two haplotypes of VS-1. On the other hand, the synteny in other chromosomes is highly conserved between several haplotypes, such as on chromosome 05. Big regions of synteny are also present between a limited number of genotypes, while the rest of the haplotypes present SVs. Example of this case is present on chromosome 14, where the synteny is conserved throughout the haplotypes of PN40024, Cabrent Franc, Rkatsiteli, Sultanina and hap1 of VS-1, while Kover 5BB and VS-1 hap2 tend to be less syntenic.

6.2 Annotation of TEs

Based on curated libraries of TEs and de-novo annotation, TEs annotation has been performed with EDTA (**Supplementary Table 1**). As can be seen from **Table 5**, an average of 48.62% of the chromosomes sequence has been annotated as repetitive/TEs, being slightly higher than previously reported values (41.4% in the 12xV0 version of the genome) (Jaillon *et al.*, 2007), probably due to the improvements in the new genome assembly methods and their ability to better assemble repetitive regions. 28.20% of the TEs were annotated as LTR elements, mainly belonging to the Gypsy family (13.12%). TIR elements accounted for 12.52%, with Mutator being the most abundant family, with 6.22%.

Table 5: *Summary of TE annotation using EDTA. Values are average of the TEs fraction in each haplotype of each accession.*

		All var	
DNA_transposon			0.06%
LTR	**Copia**	9.85%	28.20%
	Gypsy	13.12%	
	Retrovirus	0.60%	
	unknown	4.62%	
TIR	**CACTA**	3.25%	12.52%
	Mutator	6.22%	
	PIF_Harbinger	1.12%	
	Tc1_Mariner	0.31%	
	hAT	1.61%	
nonLTR	**LINE_element**		2.41%
nonTIR	**helitron**		1.80%
repeat_region			3.64%
Total			48.62%

6.3 Identification and annotation of centromeric regions

The high completeness and contiguity of the assemblies have enabled the study and identification of the centromeric regions. In 145 chromosome haplotypes (69% of the cases), centromeric regions were assembled entirely in a single contig. To avoid misinterpretation of the results due to parts missing from the sequence, only chromosomes with centromeres contained entirely in one contig were selected for further analyses.

Several tandem repeat families and TEs families have been identified and characterized in the centromeric regions, with size ranging from 6 bp to 1385 bp. The most common tandem repeats were the 107 bp repeat, the 59 bp repeat family, the 28 bp repeat family, the Gyp-28 repeat family and the 66 bp repeat family (**Figure 7**). The 107 bp repeat family is the one identified in previous studies present in literature (Shi *et al.*, 2023). Self-comparison of the sequence revealed a structure composed of a base unit of 28 bp in length. Moreover, a variant of this monomer was the 135 bp repeat, formed by the union of a whole monomer (107 bp) and part of it (28 bp). Among the other repeats, one of the most interesting ones is the Gyp-28 family, which correspond to repeats that originated from a Gypsy-28 soloLTR element, with matches in the centromeres often containing just part of it, present as isolated repeated regions rather than tandemly repeated arrays.

Apart from the most common repeat families, nine additional repeat families have been characterized, being often present in centromeric or pericentromeric regions (**Supplementary Figure 1**). Among these, the 354 bp repeat family and 342 bp repeat family showed high similarity in their sequences, masking most of the times the same regions, but being separated in other cases.

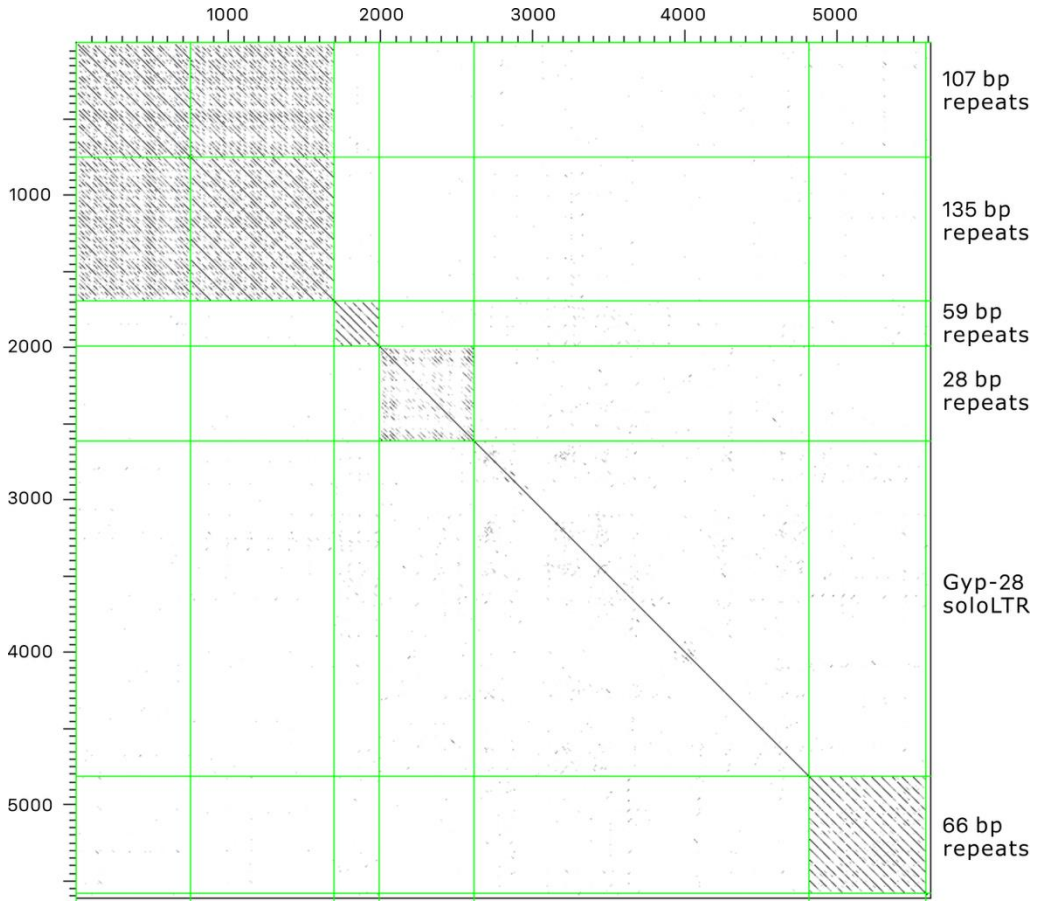


***Figure** 7: Self dotplot (Dotter) of the 6 most characterizing centromeric repeat families. For all repeat families but Gyp-28 soloLTR, several consecutive monomers have been plotted.*

The combination of the repeat families and their abundance varied between haplotypes (**Figure 8**). The most abundant repeat family that has been identified is the 107 bp monomer, covering overall the 33.2% of the regions annotated as centromeres. This monomer is present in almost all chromosomes of all accessions, often forming big arrays of tandem repeats (**Figure 9 A**). Contrarily to the main trend, in some cases such as chromosome 18 of all accessions, chromosome 16 of Cabernet Franc hap2, chromosome 14 of Rkatsiteli hap1 and chromosome 02 of Kober 5BB hap 1 the 107 bp

monomer represents less than 0,1% of the centromeric region. Moreover, one peculiar case was present on chromosome 03 of all accessions, where the big arrays were not composed of 107 bp monomers, but were formed instead by a pentamer of the 28 bp unit, resulting in a 135 bp monomer with high identity with the 107 bp monomer family (from now on, considered as 107 bp monomer family) that represents a tetramer of the same base unit. The second most abundant repeat was the 66 bp family, accounting for 7.90% of the cumulative length of centromeric regions. In chromosomes were the 107 bp monomer was almost lacking, the 66 bp monomer family was the predominant one, accounting for more than 17% of the centromeric region in these haplotypes. Another common repeat family, even though less widespread in terms of cumulative length (1.51%), was the one composed by 59 bp monomers, which was often associated and intermixed with small arrays of the 28 bp monomer family (0.62%) and repeats composed of part of a solo LTR Gypsy element called Gypsy-28 (1.07%). Many other repeat families of various length, accounting globally for 3.6%, were present in low abundance (often < 2%) and were usually not involved in the centromere core structure, even though in some cases were accounting for almost 10% of the centromeric sequence (such as on chromosome 14 of most of the varieties) and were integrated in structures formed by other repeats. Apart from these tandemly arranged repeat families, TE families were also enriched in centromeric regions. The most abundant one was the Athila family, covering 13.37% of the total length of centromeres. Other TE retrotransposons containing a chromodomain were present in these regions, such as Galadriel and CRM which accounted respectively for 5.07% and 4.04% of the centromeric regions. Moreover, also Tekay and Reina families were present in centromeric regions and accounted for 8.54% and 2.15% respectively.

The size of centromeric regions was also highly variable, both between different chromosomes (e.g. chromosome 06 and chromosome 14), as well as between different haplotypes of the same chromosome (e.g. chromosome 12 of Kober 5BB hap2 and chromosome 12 of Rkatsiteli hap1) (**Supplementary Table 2**). Some chromosomes were overall more similar in terms of centromeres size, with at most three haplotypes that exhibit very different dimensions, such as in chromosome 02, 04, 07, 11, 13, 18. On the other hand, chromosome 01, 15, 16 and 19 had more variation in terms of centromere length, having also more intermediate levels between the biggest and smallest haplotypes. The longest centromeric region was several megabases long (6139485 bp) and was present in chromosome 16 of PN40024, while the smallest was less than 400kbp long (380567 bp) and was present in chromosome 15 of Rkatsiteli hap1.

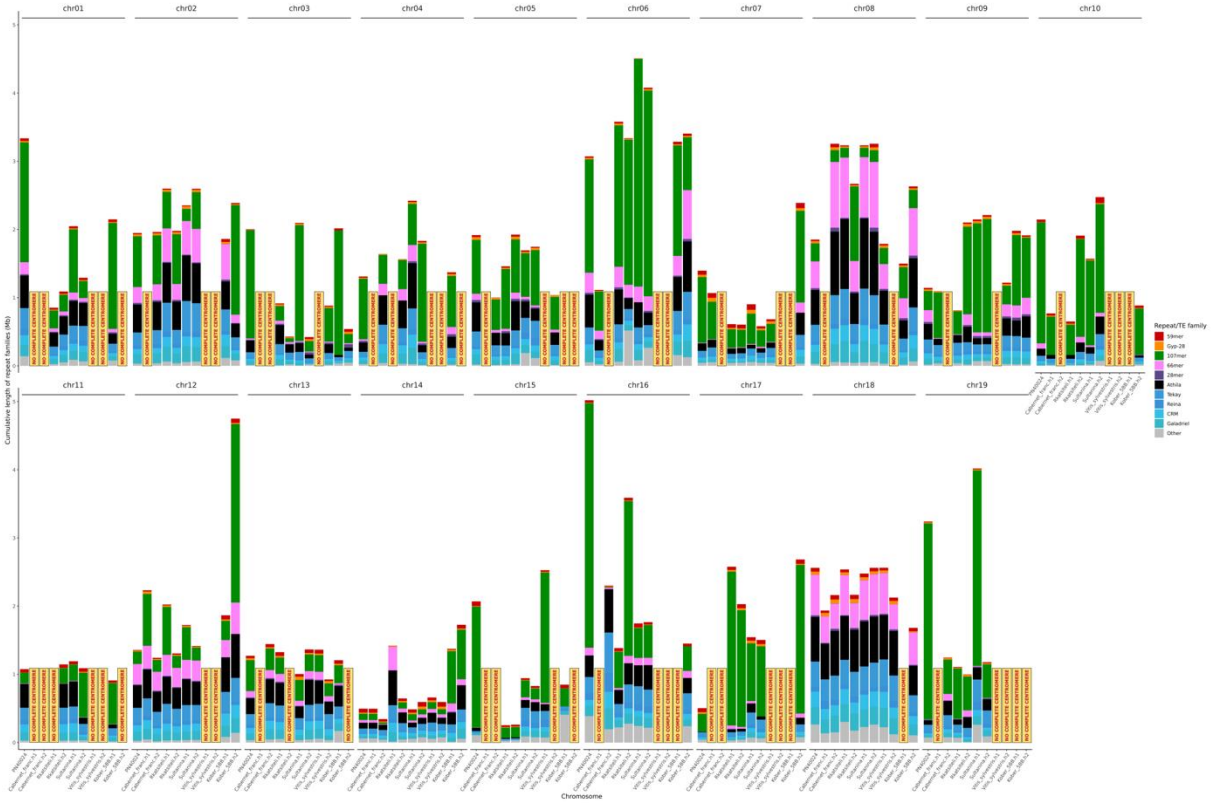


Figure 8: *Cumulative length of bases masked as Repeat/TE families in centromeric regions, by chromosome. Centromeres with "No complete centromere" correspond to haplotypes with centromeres assembled in multiple contigs.*

6.4 Structural variability of grapevine centromeres

Based on the composition and the structure of centromeric regions, two main categories of centromeric regions have been identified (from now on called Type A and Type B).

Type A structure is characterized by large arrays of the 107 bp monomer, in some cases spanning several megabases, surrounding a central core composed of 59 bp monomer, 28 bp monomer and Gyp-28 repeats families, often present in a highly organized and symmetrical structure (**Figure 9 A**). Example of this type of centromeric structure are present in all the accessions that have been analysed, such as in Sultanina chromosome 02 hap2, Cabernet franc chromosome 02 hap1, Kober 5BB chromosome 01 hap1, PN40024 chromosome 06, Rkatsiteli chromosome 08 hap2 and VS-1 chromosome 08 hap1. Variations of this structure were also present among the assembled centromeres: in some cases, one of the two arrays of 107 bp repeats was missing, such as on VS-1 chromosome 15 hap1; in other cases, such as in Rkatsiteli chromosome 10 hap1, two big arrays of 107 bp repeats were present, but the central core comprising 28 bp, Gyp-28 and 59 bp repeats was not organized in a symmetrical structure.

The second main category of centromere structures (Type B) is the one characterized by the almost completely absence of the 107 bp repeat, and the presence of scattered small arrays of the 66 bp repeat family throughout the whole centromeric region, intermixed with Athila elements (**Figure 9 B**). Moreover, repeats of the 6 bp repeat family were present towards the central core. This type of structure was the less represented, being present just in chromosome 18 (all haplotypes) and in some haplotypes of other chromosomes (Kober 5BB chromosome 02 hap1 and Rkatsiteli chromosome 14 hap1). Despite its diversity in the overall structure and presence/absence of big arrays of tandem repeats, a core composed of 59 bp monomer and Gyp-28 repeats families is always present, with the 59 bp repeat being the only part that is in common to all types of centromeric structures.

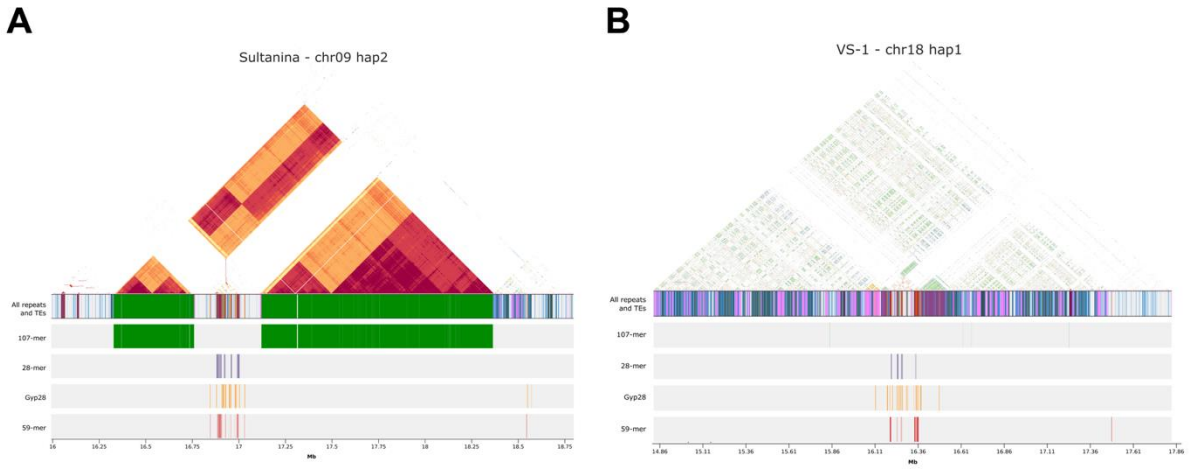


Figure 9: *Centromere structures of grapevine centromeres.* ***A)*** *Type A structure, dominated by tandem repeats arrays of 107 bp monomer family. On top, heatmap of self-identity plot of the centromeric region in chromosome 09 of Sultanina hap2. Tracks highlighting regions masked as tandem repeats or TEs ("All repeats and TEs"), and the four main monomer families: 107 bp, 28 bp, Gyp-28 and 59 bp tandem repeats.* ***B)*** *Type B structure, dominated by absence of big arrays of 107 bp monomer family, but maintaining the 28 bp, Gyp-28 and 59 bp repeat families. On top, heatmap of self-identity plot of the centromeric region in chromosome 09 of Sultanina hap2. Tracks highlighting regions masked as tandem repeats or TEs ("All repeats and TEs"), and the four main monomer families: 107 bp, 28 bp, Gyp-28 and 59 bp tandem repeats.*

Even though most of the centromeric structures can be attributed to either one of the two categories, several minor variations were present in the pool of centromeric haplotypes, especially for the ones with arrays of the 107 bp monomer, belonging to the Type A category. Most of the centromeres were characterized by two relatively big arrays, such as the ones present on chromosome 09 of Sultanina hap2 (**Figure 9 A**) with a central core composed of the 59 bp repeat family, the 28 bp repeat family and the Gyp-28 repeats. A variant of this structure is the one lacking one or sometimes two of the repeats composing the core (**Supplementary Figure 2**). An example of this structure

is chromosome 03 of Kober 5BB hap1, which was lacking the Gyp-28 repeat. In a similar way, chromosome 04 of Cabernet Franc in both haplotypes, chromosome 12 of Kober 5BB in both haplotypes, chromosome 17 of Rkatsiteli hap2, Sultanina hap2 and Kober 5BB hap2, all reconstructed haplotypes of chromosome 14 (apart from Rkatsiteli hap1 and Kober 5BB hap1) and all reconstructed haplotypes of chromosome 07 were lacking the 28 bp repeat family. Another variant was the one lacking 2 of the repeats composing the core, such as in chromosome 15 of Kober 5BB hap1, in which just the 59 bp repeat was present.

In other cases, the main difference from the classical Type A structure was characterized by a core composed of extremely small repeats belonging to the 59 bp, 28 bp and Gyp-28 repeat families, surrounded by the 107 bp big arrays (**Supplementary Figure 3**), such as chromosome 06 of Sultanina hap1, chromosome 09 of Cabernet Franc hap1 and Rkatsiteli hap1.

More variants of these structures were characterized by the presence of just one big array of the 107 bp monomer, having the 28 bp, 59 bp and Gyp-28 repeats present on one side of the array (**Supplementary Figure 4**). Examples of these structures are present in chromosome 03 of PN40024, Rkatsiteli in both haplotypes, Sultanina hap1 and VS-1 hap2, chromosome 06 of Cabernet Franc hap1 and Rkatsiteli hap2, chromosome 08 of Rkatsiteli hap1 and Sultanina hap1, chromosome 10 of all haplotypes (since one of the two arrays of the 107 bp had a very short dimension, often smaller than the actual core), chromosome 11 of all haplotypes, chromosome 14 of Cabernet Franc hap2, all haplotypes of chromosome 15 apart from Kober 5BB hap 1, all haplotypes of chromosome 16 and chromosome 19 of PN40024, Rkatsiteli hap1 and Sultanina in both haps.

An additional and more altered structure was the one having few small clusters of the repeats found usually in one core (28 bp, 59 bp and Gyp-28) in multiple clusters, often before and after a big 107 bp array of tandem repeats (**Supplementary Figure 5**). Examples of these structures can be found in chromosome 02 of Rkatsiteli hap1 and Sultanina in both haplotypes, chromosome 03 of Sultanina hap2 and Kober 5BB hap2, chromosome 08 of Rkatsiteli hap2, Sultanina hap2, Kober 5BB hap2 and VS-1 hap1, chromosome 12 of Cabernet Franc in both haplotypes, chromosome 12 of Rkatsiteli hap2, and chromosome 13 in all haplotypes.

Other variations to the Type A and B structures were characterized by repeat families that were present almost only on specific chromosomes or haplotypes (**Supplementary Figure 6**). Examples are the presence of the 6 bp repeat family in some haplotypes of chromosome 13, the 354 bp and 342 bp repeats families in pericentromeric regions of chromosome 03, 10, 11 and 17, the 187 bp repeat family in some haplotypes of chromosome 05, the 123 bp repeat family in most of the haplotypes

of chromosome 06, the 51 bp repeat family in some haplotypes of chromosome 12 and all of the haplotypes of chromosome 14, the 191 bp repeat family in most of the haplotypes of chromosome 13 and chromosome 18, and the 1385 bp repeat family in all the pericentromeric regions of chromosome 15 and 17.

6.5 Characterization of the functional core of the centromeres

In order to get information about the functional part of the centromere, publicly available data of ChIP-seq experiments targeting the centromere specific histone variant (CENH3) were used to identify the functional domain of the extremely variable centromeric structures. The ratio of the coverage values between the samples containing the antibody targeting CENH3 vs no-antibody sample was used to calculate read enrichment in the samples compared to the input, using 1000bp windows, in identical-by-descent haplotypes.

As shown in **Figure 10**, regions that are enriched in CENH3 are masked mostly by 59 bp monomer, 28 bp monomer and Gyp-28 repeat families, reaching a median enrichment more than 7-fold higher than that of other regions (either other repeat families or regions that are not masked by any TE or repeat family) for the Gyp-28 and more than 10-fold higher for 59 bp repeat family and 28 bp repeat family. Moreover, by taking a closer look at each chromosome of PN40024, regions masked as 59 bp repeats, 28 bp repeats and Gyp-28 monomers were almost always associated with peaks of CENH3 with enrichment values > 30 (**Figure 10 C**). The example of chromosome 01 highlights how 100% of the bases masked as 59 bp repeat had values of enrichment > 30, while the same threshold was achieved in more than 92.5% and 95.2% of the bases masked as Gyp-28 and 28 bp repeats respectively. In some cases, such as in chromosome 03, 08, 11 and 15, repeat families that were not belonging to the 3 functional categories mentioned above were poorly associated with CENH3 enrichment. On the other hand, other centromeres had a higher fraction of repeats belonging to families that are not the 59 bp repeats, Gyp-28 and 28 bp repeats. In chromosome 17, more than 50% of sequences annotated as CRM or Tekay were associated with CENH3 enrichment > 30, while the rest of the repeats and TE families stayed below 35%. Similar patterns were present on chromosome 14, where CRM elements and several tandem repeat families (marked as “Others”) had more than 40% of their regions associated with values of CENH3 enrichment > 30, while the other repeat families remained below 25%. Lastly, chromosome 04 was also characterized by a high enrichment of CENH3 ratios in several TE families (Tekay, CRM, Galadriel, Reina and Athila), with a percentage of masked bases with CENH3 > 30 between 22% and 50%.

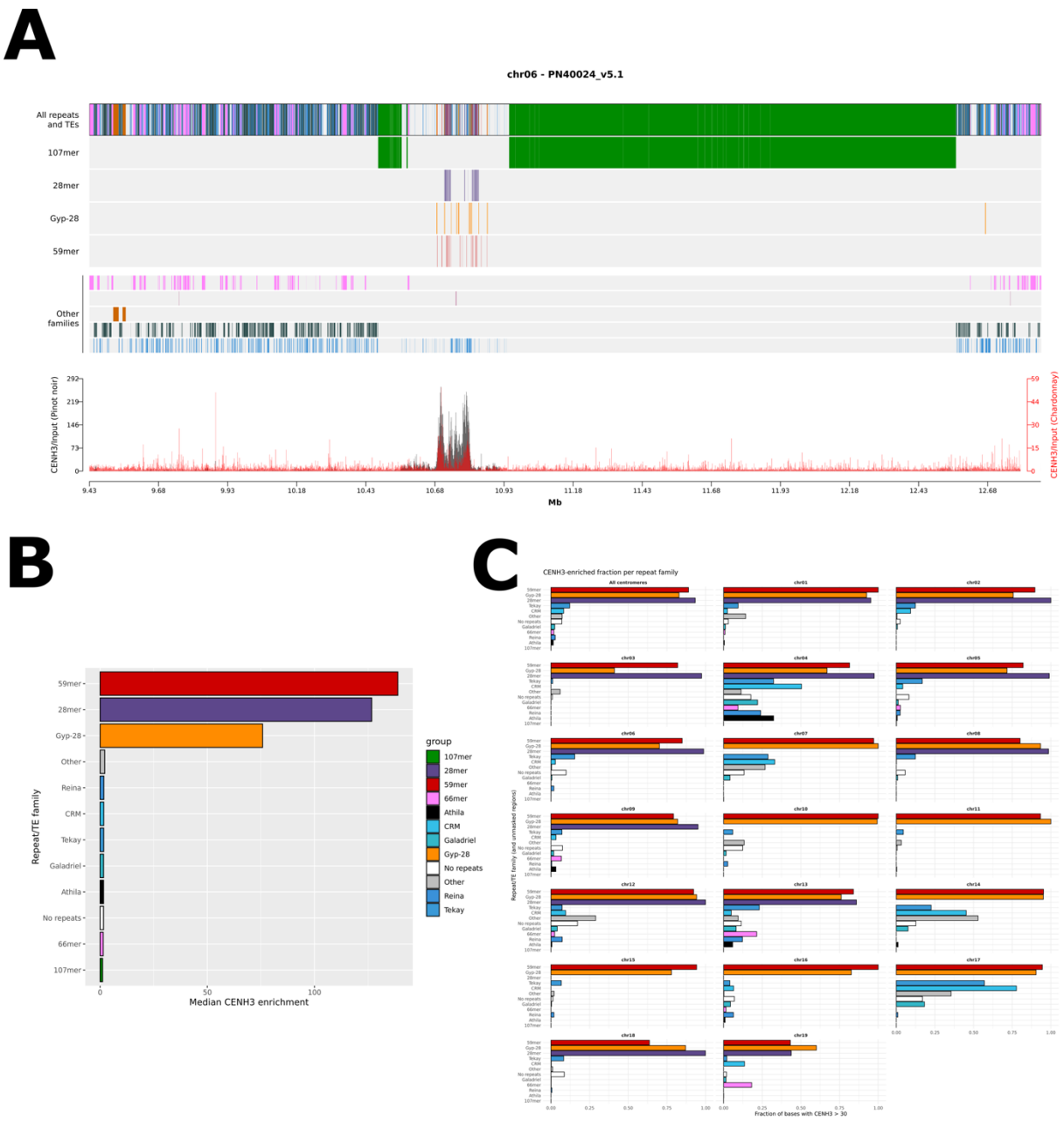


Figure 10: *Centromere functional repeats. As shown by the 3 facets, enrichment of the functional part of the centromere (CENH3) tend to co-localise mostly with the 59 bp repeat family, the 28 bp repeat family and the Gyp-28 repeats, respectively.* ***A)*** *Centromeric region of chromosome 06 in PN40024. Repeat landscape with tracks corresponding to (from the top to the bottom): all repeats and TE families, 107 bp repeat family, 28 bp repeat family, Gyp-28 repeats, 59 bp repeat family, 66 bp repeat family, 191 bp repeat family, 123 bp repeat family, centromeric Athila elements and TE families containing chromodomains (CRM, Galadriel and Tekay). Bottom track shows CENH3 ChIP-seq experiments: enrichment (sample/Input ratio) in 1kbp windows from Pinot noir (black) and Chardonnay (red) samples.* ***B)*** *Histogram with main repeat and TE families in centromeric regions in PN40024. Median CENH3 enrichment levels for each repeat/TE family.* ***C)*** *Fraction of bases masked as specific repeat/TE family with CENH3 enrichment levels > 30 in all centromeric regions of PN40024 and in centromeres of each chromosome of PN40024.*

6.6 Megabase-scale arrays composition and structure

Even though the 107 bp monomer repeats do not appear to be the functional ones, similarity in terms of structure, length and other characteristics with the repeat arrays found in *Arabidopsis* and Human, and their predominance in grapevine centromeric regions has encouraged their study and their characterization. To do so, CENdetectHOR tool was used to both extract monomers composing the different arrays and eventually spot any HOR found in the large arrays.

The longest continuous array was present in chromosome 06 of PN40024, with 3492456 bp between the first and last monomer of the array. Across all the accessions, chromosome 06 was the one with the highest median length of the 107 bp arrays (1058838 bp). Preliminary steps performed on PN40024 genotype on the composition of the arrays showed the prevalence of the monomeric version of the repeat, being the most abundant. Analysis of the frequency of distance of 8 bp k-mers showed the highest peaks for a distance of 107 bp. For chromosome 03, as shown from the previous analyses, the most abundant repeat was 135 bp long, instead of the canonical 107 bp monomer. Interestingly, in two regions of chromosome 07 (from 12603752 to 12660802 and from 12662552 to 13381452), the predominant peak was the dimeric version, of length 214 bp (**Supplementary Figure 7**).

To investigate the similarity of the different monomers, intermediate outputs of the CENdetectHOR pipeline were used to compare the sequences coming from the 19 chromosomes. A total of 1710 sequences extracted from these regions were compared using Mafft software, and an alignment tree was generated to represent visually the alignments. As can be seen in **Figure 11**, in the majority of cases, sequences coming from the same chromosome cluster together, reflecting a higher similarity between sequences coming from the same haplotypic region, rather than a mixed and homogeneous distribution of the variants between all the chromosomes. In particular, chromosomes 02, 06, 08, 10, 15, 16 and 19 were the ones that contained monomers that were not mixed with others. In other cases, such as in chromosomes 11 and 14, monomer families were grouped with monomers coming from other chromosomes. Moreover, comparison of 2900 monomers coming from PN40024 and Rkatsiteli hap1 centromeres has confirmed a similar overall trend, where monomers coming from the same chromosome but different genotype were more similar than monomers coming from different chromosomes within each genotype (**Supplementary Figure 8**). Also in this case, some chromosomes tended to be more clustered and conserved between the two

genotypes than others, such as chromosomes 10, 13 and 16, confirming the trend already seen at within-genotype level.

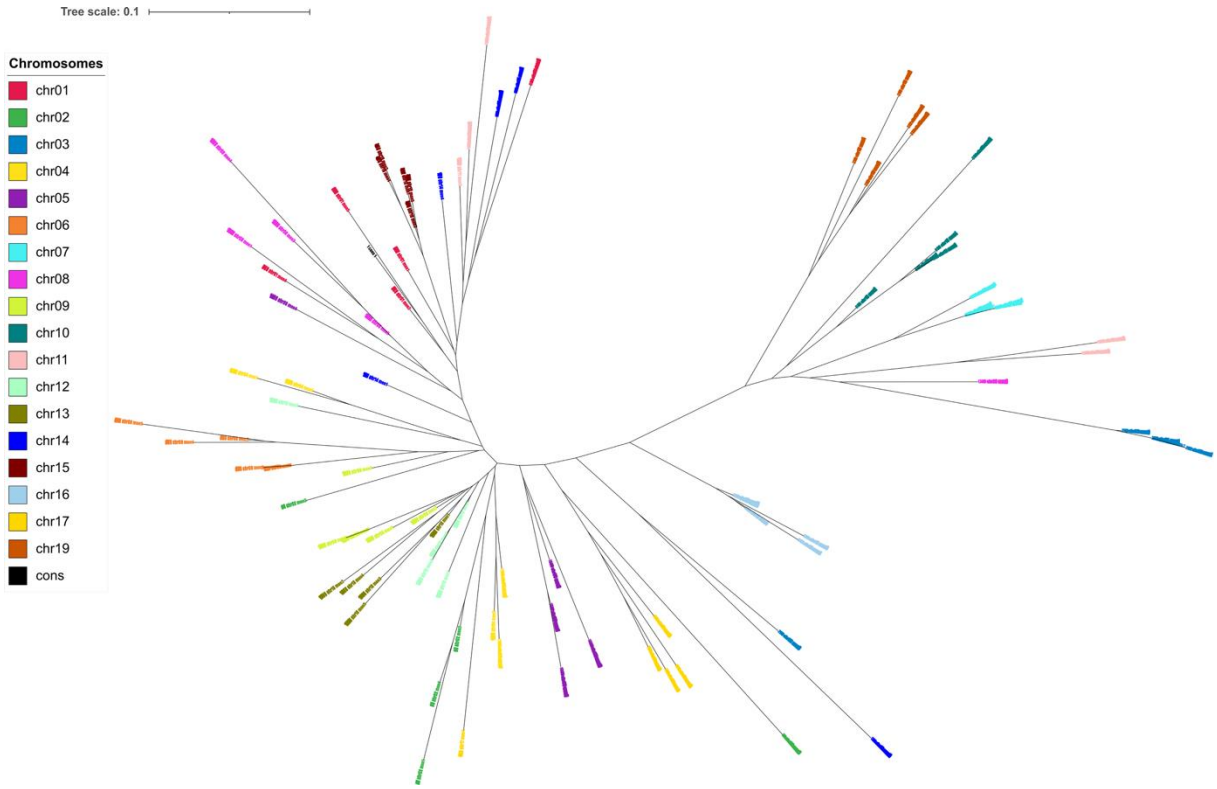


***Figure 11**: Unrooted tree of monomers composing the 107 bp repeat family arrays in PN40024. The tree was generated based on Mafft multi-alignment of the monomers extracted by CENdetectHOR pipeline. Labels are coloured based on the chromosome of origin.*

Identity plots revealed similarity blocks inside the big tandem repeat arrays, allowing a subdivision of the arrays based on the similarity of the monomers. In most of the cases the arrays contained at least two blocks of similarity, reaching even more than 10 distinct blocks in bigger arrays, such as the one in chromosome 04 of PN40024. An opposite pattern was instead present in small arrays, such as the ones present in chromosome 14 of Rkatsiteli hap2.

As shown in **Figure 12** several evolutionary layer could be spotted, based on the identity of the sequences, with a gradient of progressively increasing similarity as we move towards inner and more recent layers. The first evolutionary layer, which is the most ancient one, presents a homogeneous but low identity between 70% and 90%. This layer is present at one corner of the bigger array of 107 bp monomer repeats and on the smaller array, at the opposite side of the central core composed of the functional repeats. A second, and more recent array, is present in the bigger array, developed on one side of the first layer. Going even more towards the center of the array, another evolutionary

layer is present (layer 3), developed from a region which is placed within the previous array. Higher levels of identity of this region suggests a more recent expansion, compared to the previous layers. Lastly, a region with even higher identity is present on the left side of layer 3. This evolutionary layer is characterized by identity levels higher than 95% and presents 3 main subregions (layers 4.1, 4.2 and 4.3), with identity higher than 97%. Both in layer 4 and 3, several sub-regions are recognizable inside the main evolutionary layers, characterized by higher identity, and corresponding to more recent repeat expansions starting from monomers present in the respective layer.

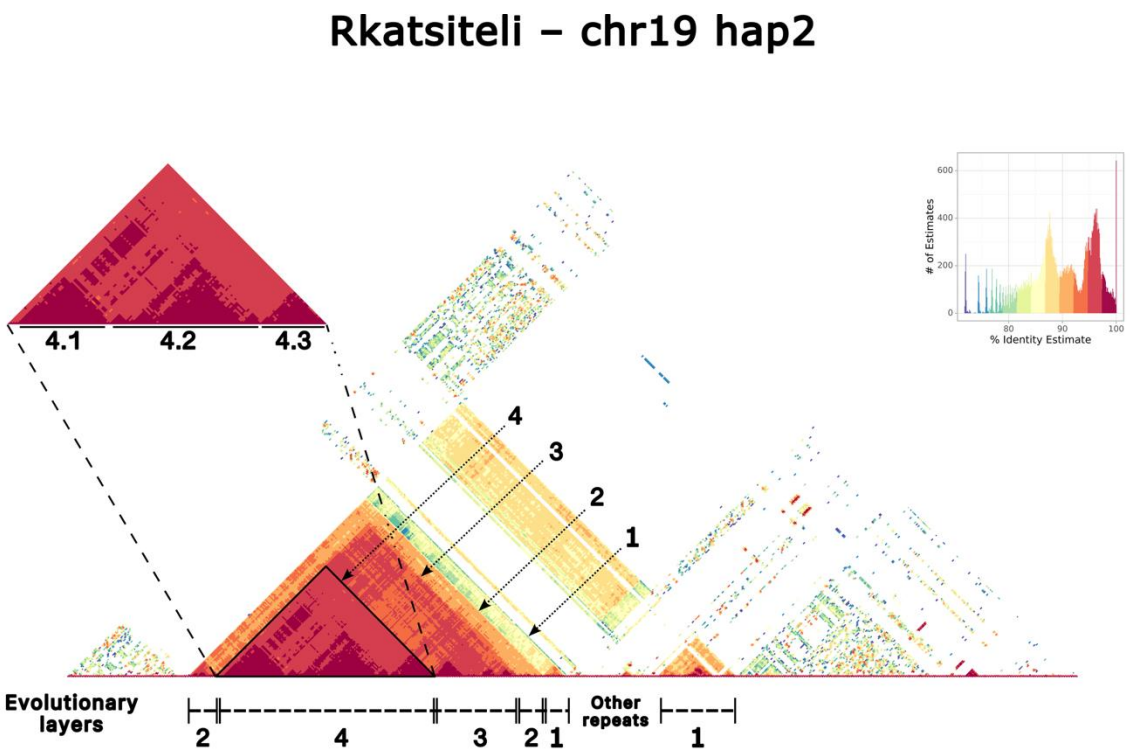


***Figure 12**: Self-identity plot of centromeric region in chromosome 19 of Rkatsiteli hap1. Evolutionary layers present in the 107 bp repeat family array are indicated by numbers, from the oldest (1) to the most recent (4). Zoom on the layer 4 highlights the presence of 3 additionally sub-layers with even higher sequence identity. On the top left, identity box indicated the frequency of windows with a specific identity. Identity plot has been generated with Moddotplot using 2kbp windows and a minimum identity of plotted windows of 70%.*

6.7 Centromere evolution: intra- and inter-haplotype evolution

After highlighting the highly variable structure of centromeres, centromere evolution has been investigated, in order to uncover the processes that generated these structures. Intra-haplotype evolution has been studied in order to describe the dynamics that are shaping centromeric structures, while inter-haplotype evolution have been investigated to trace similarities, differences and mechanisms that have shaped the evolution of centromeres throughout the grapevine lineage.

To be able to reconstruct the evolutionary history of the centromeres, intact TEs belonging to the LTR family have been used to get insights on the estimated time of structural variation events, such as duplications and inversions. On chromosome 09 of PN40024 we were able to find 11 intact TEs throughout the whole centromeric region. Moreover, 6 of these were inserted within the highly structured region composed of duplicated and inverted repeats, surrounded by two arrays of the 107 bp monomer family. The first array spans 141733 bp, containing 1198 intact copies of the 107 bp monomer. The masking with the 107 bp monomer sequence is not contiguous, but it is instead interrupted 53 times by small sequences with a median length of 17 bp, and 2 times by bigger gaps: the first one with a length of 1727 bp, and the second one with a length of 6759 bp, both filled with sequences masked as CRM, Galadriel and Tekay TE families. Similarly, the second array is 158566 bp long, containing 1342 intact 107 bp monomers and being as well interrupted 53 times by sequences that have a median length of 10 bp. The two arrays are separated by a central core, composed of an inverted repeat which is 312010 bp long (from coordinate 15103515 to 154155523 of chromosome 09) which contains in turn a forward duplication that spans for 28367 bp (from coordinate 15215566 to 15243931 of chromosome 09). Functional repeats (i.e. 59 bp monomer and Gyp-28 repeats) are present in the last part of the inverted repeat (closer to the duplication) and inside the forward duplication present in the core.

To study the evolution of the structure of the central core, TSD and LTR have been compared among the intact TEs. Presence of intact LTR has allowed the estimation of the insertion time of the elements. Three groups of identical TEs associated with 3 TE insertion events have been identified in this region.

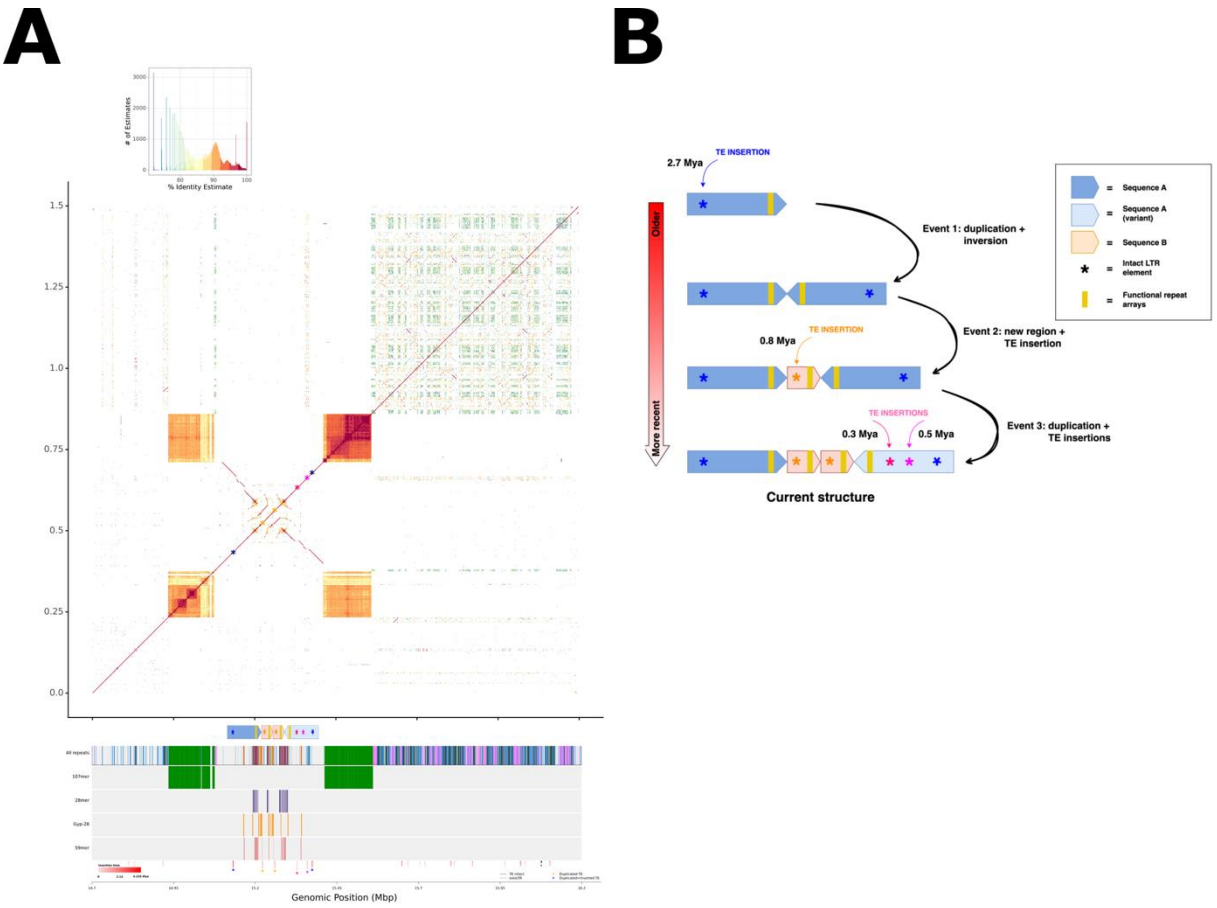


***Figure 13**: Scheme of evolution of chromosome 09 in PN40024. **A)** Self comparison heatmap of centromeric region in chr009 of PN40024. Bottom track refers to All repeats and TEs families, 107 bp repeat family (green), 28 bp repeat family (purple), Gyp-28 repeats (orange) and 59 bp repeat family (red). Arrows indicate intact TE (full line) or soloLTR (dashed). Stars indicte intact TE present in the central core. Color of the arrows refers to estimated insertion time of the TE. **B)** Scheme of evolution of the centromere core (sequence between the two 107bp tandem repeat arrays) in chromosome 09 in PN40024. The centroemre core underwent a first duplication and inversion event, which generated two identycal TEs on the two sides. Subsequently, a central part saw an insertion of a new TE; the central part then was duplicated, generating a forward duplication (and the duplication of the TE). Lastly two more indipendent TEs insertions happened in one of the two inverted repeats, differenciating the two sides of the core.*

- The first group is composed by 2 identical TEs, with the same TSD, present in the inverted region. The presence of two identical TSD in different TEs suggests that the two elements originated by one common insertion event, which happened 2.7 Mya, before the duplication and inversion event.
- The second group is composed by 2 identical TEs (different from the previous group), with the same TSD, present in the forward-duplicated region. Similarly to the previous case, the presence of identical TSD suggest a common insertion event, happened 0.8 Mya, before the duplication event.
- Lastly, the third group is composed by different elements present in just one of the duplicated regions. These elements have different TSD, indicated indipendent events that happened after the formation of the structure in which they are contained. The estimated time of insertion for these elements was 0.3 and 0.5 Mya.

By using this information, the step that led to formation of the centromere of chromosome 09 as it is today have been reconstructed. As shown in **Figure 13**, a first sequence *A* containing the functional repeats and a TE (blue star) underwent a first duplication and inversion event (*Event 1*) less than 2.7 Mya. This led to the formation of an intermediate, which evolved further with a formation of a central spacer (sequence *B*) between the two inverted repeats, also containing the functional repeats. This was then invaded by a new TE insertion event (orange star). This insertion, which happened 0.8 Mya, gives a new boundary to the previous *Event 1*, placing it between 2.7 and 0.8 Mya. The last duplication event happened more recently than 0.8 Mya, when also new independent TE insertions happened in one of the two inverted repeats 0.3 Mya and 0.5Mya (magenta and fuchsia stars).

The presence of intact TEs was widespread among most of the centromeric regions, as well as solo LTR, generated by the excision of TEs (**Figure 14**). In all the centromeric regions, a total of 1890 TEs and 887 solo LTRs were identified. A total of 34 TEs were duplicated or duplicated and inverted, coming from duplications/inversion events rather than independent insertion events. In each haplotype, the number of soloLTR ranged from 0 to 45, while TEs ranged from 0 to 60. In most cases, both soloLTR and intact TEs were present at the same time. Higher abundances of TEs and soloLTRs were present in specific chromosomes (such as on chromosome 18) or in other specific haplotype that were lacking most of the tandem repeat families, like chromosome 16 of Cabernet Franc hap2. From a broader point of view, higher levels of solo LTRs and intact TEs were observed in haplotypes were the composition of centromeres had higher frequencies of TE families (both Athila and the others chromodomain-containing TEs).

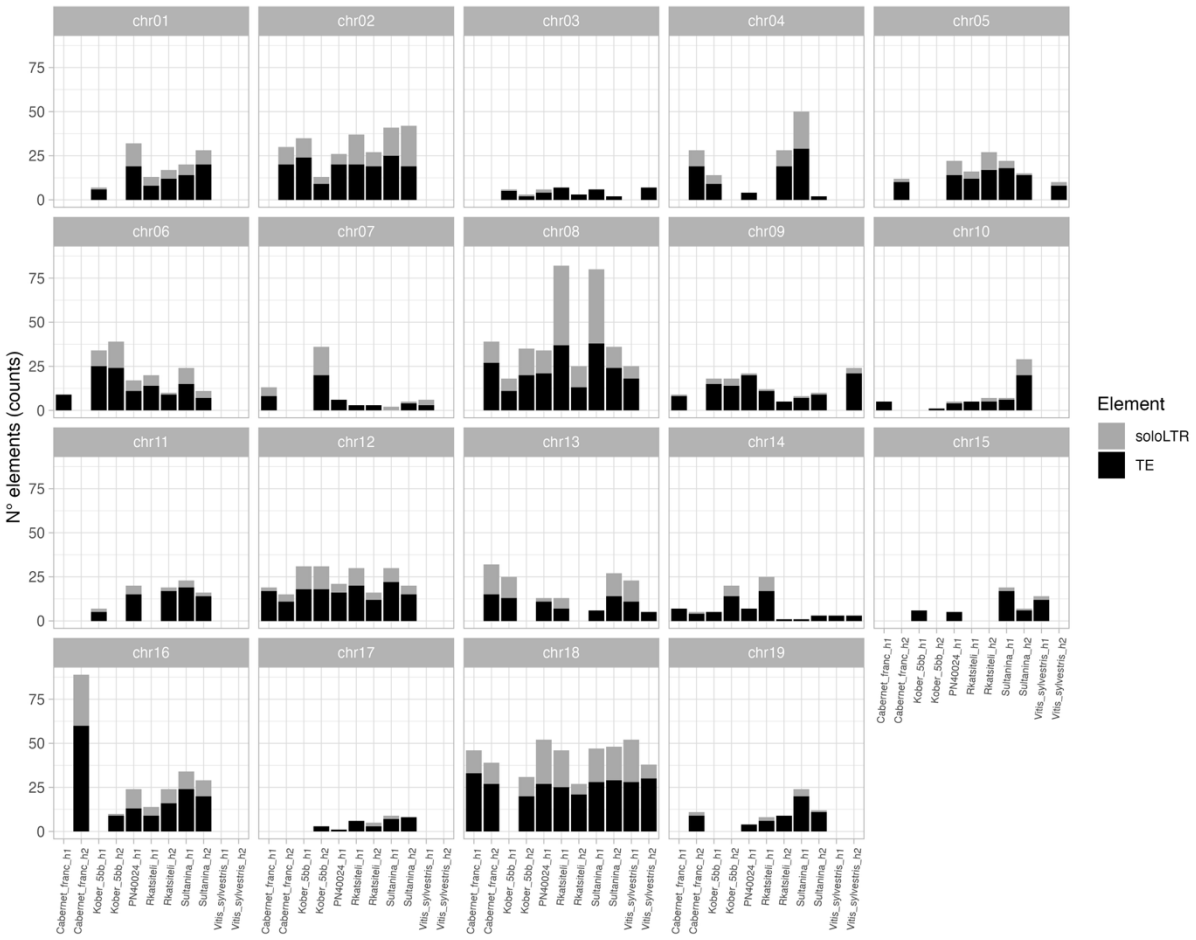


***Figure 14**: Number of intact TEs and soloLTRs for each haplotype.*

To study also the dynamics that affect different genotypes, haplotypes coming from different individuals have also been compared for chromosome 09. Firstly, the comparison between PN40024 and VS-1 hap2 highlighted highly similar structures, indicating a common evolution path (*Events* from *1* to *5* in **Figure 15**). Moreover, recent events of TE insertion/excision were present privately in the two haplotypes, differentiating them, while most of the intact TE and soloLTR were found in both haplotypes. These events were spotted in centromeres' borders, rather than around the functional core of the centromere.

Comparison of these regions with Sultanina hap2 have added information on the evolution of chromosome 09 centromeres in the grapevine genealogy. In this case, most of the intact TEs and soloLTRs were present in one of the two haplotypes. On the other hand, the only TEs that were in common between the haplotypes were the ones involved

in the highly structured core, contained inverted and duplicated sequences (Sequence *A*). In fact, two inverted TEs, with the same LTR and TSDs as the ones in PN40024 were found also in Sultanina, attributing them to the same insertion event (blue star in **Figure 15**). The presence of another common TEs, with LTRs and TSDs identical to the ones present on the tandem duplication in PN40024 (Sequence *B*) also suggest that the TEs insertion event happened around 0.8 Mya was the same (orange star). Moreover, the presence of just one copy of this TEs hints at an independent evolutionary trajectory of the Sultanina haplotype, originated before the tandem duplication event that then led to the formation of the rest of the structures (*Event 6* in **Figure 15**).

Further improvement in the understanding of the evolution of centromeres in chromosome 09 was achieved adding haplotypes with more diverse structures. Haplotype 2 of Sultanina showed similarities with hap1 of the same individual, with almost all the same intact TEs and soloLTR present in both haplotypes. Both haplotypes were characterized by a big inversion, compared to the PN40024 haplotype and the VS-1 hap2 haplotype, that involved the vast majority of the centromeric region. One difference in the central core was the absence of the intact TE in the non-duplicated region (Sequence *B*), probably due to the diversification of that region, that led to the missed identification of the TE in one of the two haplotypes (*Event 7* in **Figure 15**). Another difference that was present in both of the haplotypes of Sultanina, compared to PN40024, was a SNP (C→T) in the TSD of one of the two TEs involved in the duplicated and inverted repeats of the centromere core (cyan star). The presence of this SNP, shared between the two haplotypes of the same individual, suggests a more recent and common ancestry.

Going back in the evolutionary history, Rkatsiteli hap2 showed even fewer similarities with the previous haplotypes. Even in this case, some TEs are private to this haplotype, mainly in the regions outside the two big 107 bp tandem repeat arrays. The central core contains the duplicated and inverted regions, with the previously identified intact TEs (blue and fuchsia star). The main difference of this haplotype is the lack of the TE between the two inverted repeats, in the region that underwent the direct duplication event (Sequence *B*). This region in fact, is just partially present in the Rkatsiteli hap2 haplotype, with several chunks of sequence that are missing (as proven by the absence of the TEs (orange star) that was duplicated in the other haplotypes) probably due to accumulation of mutations and minor SVs (*Event 8* in **Figure 15**).

Similarly to Sultanina haplotypes and Rkatsiteli hap2, the two Kober 5BB haplotypes have probably separated by the PN40024 evolutionary branch after the generation of the Sequence *B*, even though the high fragmentation of the original sequence, as well as the diversity accumulated during the diversification makes hard to tell from which stage

of evolution it derives, placing the *Event 9* from **Figure 15** in between *Event 2* and *Event 4*. Some residual similarity with Sequence *A* is still present, even though a rearrangement similar to a translocation in one of the two duplicated and inverted segments has led to the movement of part of the sequence outside of the central core present between the two 107 bp tandem repeat arrays; the other region, similar to Sequence *A* has also undergone big rearrangements, with most of the sequence being missing or mutated, even though the duplication and inversion of the sequence is still visible (**Supplementary Figure 9**). Moreover, further diversification in haplotype 2 in the sequence more similar to Sequence *B* has led to the formation of a different version of the centromeric sequence (*Event 10*).

Lastly, haplotype 1 of Cabernet Franc and Rkatsiteli hap1 likely originated from an older separation with the other haplotypes, presenting just one of the two big duplicated and inverted sequences (Sequence *A*). The separation (*Event 11*) happened after the insertion of the TE in that region, and can be placed after the *Event 1* in **Figure 15**. These two haplotypes show a high similarity in the sequence, even though they share just few intact TEs, highlighting a higher similarity among them, compared to the rest of the haplotypes. Among the small differences between the two haplotypes, the intact TE present in Sequence *A* (blue star) was missing from Rkatsiteli hap1, probably due to a diversification process of accumulation of small mutations and minor SVs (*Event 12*). As expected, these regions present a high identity with the first part of Sequence *A*, and almost no similarity with the directly duplicated core composed of Sequence *B* and the surrounding areas, even though functional repeats are already present.

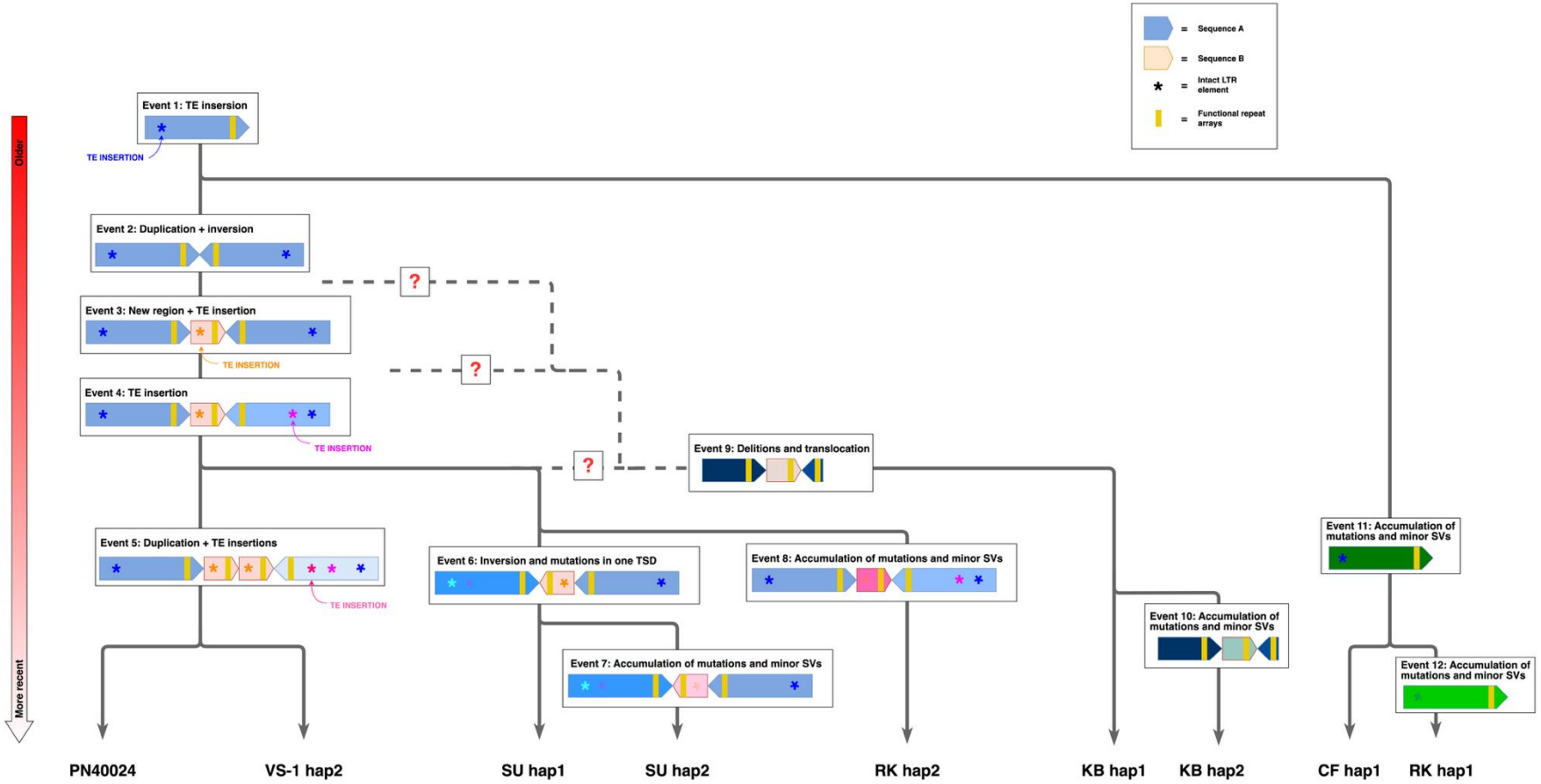


***Figure 15**: Evolution of centromeres core in chromosome 09 throughout the grapevine genealogy. An initial sequence containing the functional repeats and an intact TE took 2 main evolutionary trajectories: one led to the formation of haplotype 1 of Rkatsiteli and Cabernet Franc, while the other underwent multiple main structural variations, such as duplications, inversion, deletions and translocations, generating several variants of a structure that is more similar to the one that we can find in PN40024.*

6.8 Epigenetic landscape of centromeric regions

Since centromeric regions are highly influenced and characterized by their epigenetic DNA state, we investigated the DNA methylation levels in Cabernet Franc and Rkatsiteli genotypes.

Haplotype specific alignments of ONT reads led to a median coverage of 29X for Cabernet Franc hap1 and hap2, and 30X for Rkatsiteli hap1 and hap2. Alignments with methylation levels information were then processed to generate levels of methylation for each cytosine, which were then filtered for a minimum coverage of 4X for the CG context and 10X for CHG and CHH (**Table 6**). For the subsequent analyses, an average of 6678515 CG positions per haplotype, 13807780 of CHG positions per haplotype and 96520629 CHH position per haplotype were used.

Table 6: *Summary of number of C positions analysed. For each haplotype, cytosines were divided in the 3 contexts and then filtered for minimum coverage (4X for CG and 10X for CHG and CHH).*

	Cabernet Franc		Rkatsiteli					
	Hap1	Hap2	Hap1			Hap2		
	CG	CG	CG	CHG	CHH	CG	CHG	CHH
N° of called positions (Modkit output)	6724468	6677323	6934507	19711076	135238764	7023066	19755445	135496633
N° of called positions (filtering for coverage)	6528047	6492319	6793250	13854936	96775741	6900445	13760624	96265517

From a broader point of view, looking at the methylation landscape of the entire chromosomes, a bimodal compartmentalization can be seen in almost each chromosome (**Figure 16 A**). The first type of compartment, localized on the chromosome arms, is characterized by lower levels of methylation in all contexts, and a higher frequency of coding regions. The lower levels of methylation and higher levels of coding sequences are also associated to lower levels of TEs presence. On the other hand, the second type of compartment presents opposite characteristics, having higher levels of TEs and methylation in all 3 contexts, and lower frequencies of coding sequences. This compartment is usually localized in centromeric and peri-centromeric regions. Moreover, the transition between the compartments is not discrete, but tends to be gradual. Despite the high conservation in the localization of these compartments, these rules tend to be followed when centromeric regions are present in the middle of the chromosome (metacentric chromosomes). Chromosomes with centromeres localized more to one end or the other (submetacentric or acrocentric) such as in chromosome 11 of Rkatsiteli hap2 or chromosome 08 of Cabernet Franc hap2, instead tend to have the

smaller arm with characteristics that are more similar to the second type of compartment, with still quite high levels of methylation, TE abundance, and reduced levels of coding regions (**Supplementary Figure 10**, **Supplementary Figure 11**).

To get a better view of which components are more or less methylated in centromeric regions, we took a closer look at centromeres of haplotypes shared with PN40024, in order to analyse also the epigenetic state of cytosines present in functional regions (i.e. CENH3 enriched). Taking a closer look at chromosome 02 of Cabernet Franc hap2 (**Figure 16 B**), high levels of CG methylation were present throughout the whole region. The highest levels were present in regions masked as big arrays of the 107 bp repeat family. These regions present consistent and continuous high levels of CG methylation. In the other regions, the landscape was more variable, with bigger fluctuations in the methylation levels. In these regions in fact, high methylation levels were peaking in correspondence to regions masked as centromeric TEs, while subtle but evident dips in the methylation levels were present in regions in between repetitive elements. Interestingly, lower levels of methylation were present in correspondence of some predicted genes structures, also when intermixed with repeat families (such as Gyp-28 and 59 bp repeats), suggesting a fine scale regulation of the epigenetic state in this region, despite the overall strong trend. Intriguingly, just few predicted genes were associated with strong depletion in methylation levels, while most of them showed just a moderate decrease in the methylation state.

Very similar patterns were present in the almost identical haplotypes of Cabernet Franc hap2 and Rkatsiteli hap2 (**Figure 16 C**). CG methylation presented identical patterns of methylation as before, in all the previously mentioned regions (tandem repeat arrays of 107bp, centromeric TEs and predicted genes). Slightly different patterns were present for the CHG context. For these cytosines, in fact, high but fluctuating methylation levels were present for regions masked both as tandem repeat families, as well as centromeric TEs, lacking the more consistent values observed for the CG context. Despite this small difference, big depletions in methylation levels were observed also in this case, in correspondence of specific predicted genes. For the CHH context, patterns similar to the ones present in the CHG context were present throughout the centromeric region, even though the overall lower levels of methylation (levels always below 50%) made the very low levels in correspondence of predicted genes even more subtle and less striking.

Moving to the regions annotated as functional, due to the presence of high enrichment levels of CENH3 variant, methylation levels appeared to be in line with the rest of the repetitive regions of the centromere, presenting high levels of methylation in all 3 contexts. Surprisingly, as in the case of chromosome 02 in Cabernet Franc hap2 and Rkatsiteli hap2, the presence of the functional part of the centromere did not influence

the almost absence of methylation of very close genes, with very high levels of methylation and CENH3 enrichment, and very low levels of methylation and predicted genes being adjacent to each other.

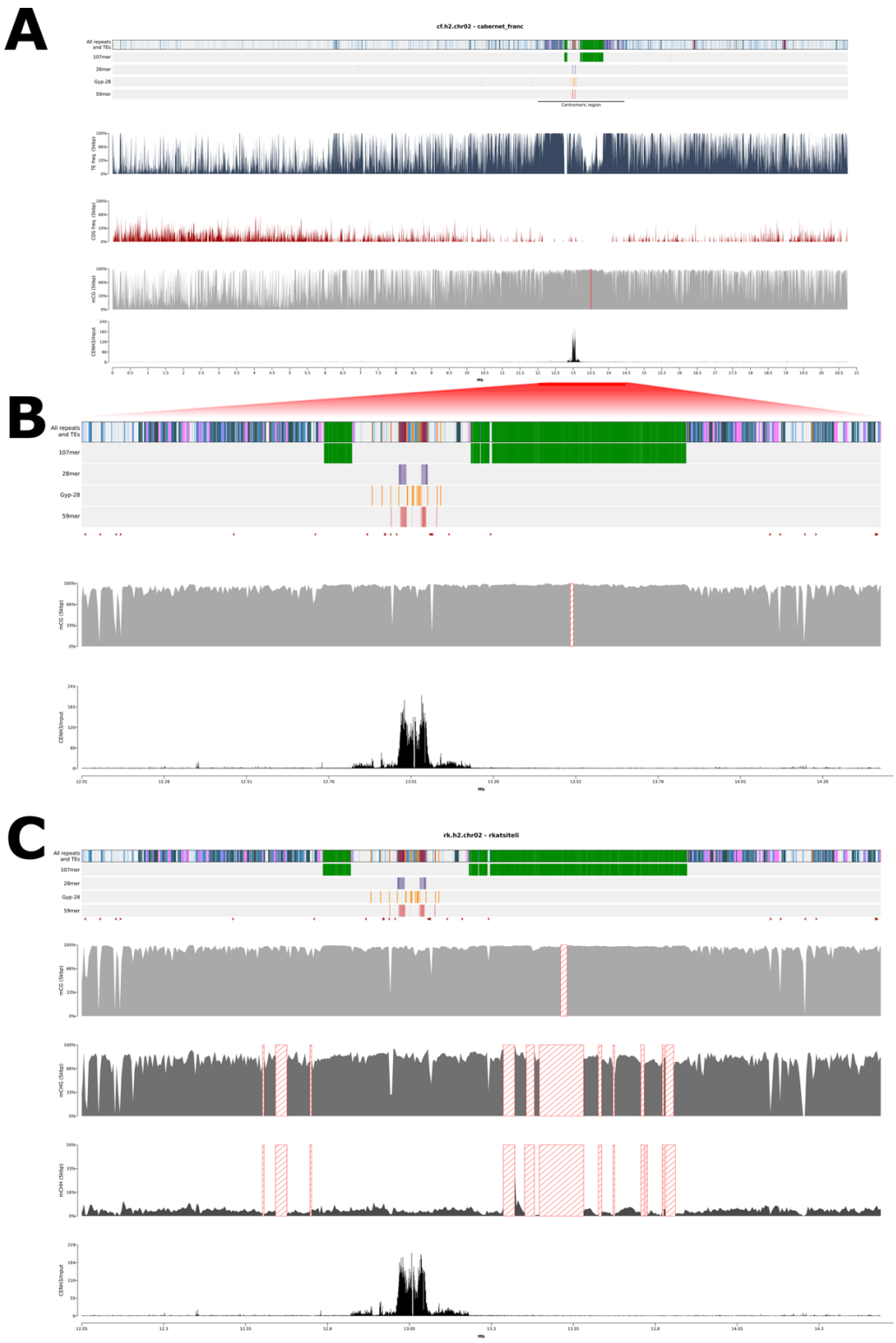


***Figure 16**: Methylation landscape of chromosome 02. **A)** Chromosome-level methylation landscape in chromosome 02 of Cabernet Franc hap2; from the top: Repeat and TEs families; percentage of bp annotated as TEs in 5kbp windows; percentage of bp annotated as CDS in 5kbp windows; average methylation levels of CG context in 5kbp windows; enrichment levels calculated as ratio of coverage of CENH3/Input in 5kbp windows. **B)** Zoom on centromeric region of chromosome 02 of Cabernet Franc hap2. **C)** Chromosome-level methylation landscape in chromosome 02 of Rkatsiteli hap2; from the top: Repeat and TEs families; average methylation levels of CG context in 5kbp windows; average methylation levels of CHG context in 5kbp windows; average methylation levels of CHH context in 5kbp windows; enrichment levels calculated as ratio of coverage of CENH3/Input in 1kbp windows.*

7 Discussion

The advent of third generation sequencing techniques has enabled a more complete understanding of genomic sequences. In the last few years, several genomes have been assembled from telomere to telomere, allowing not only a more comprehensive and complete reconstruction of the genomic sequence, but also enabling the study of highly complex regions, such as the centromeric ones. Moreover, being able to distinguish and phase the haplotypes of a diploid or polyploid species is essential to spot haplotype specific variation and to study evolutionary processes that are ongoing in the different haplotypes. Being able to phase the information in the two haplotypes, moreover, is essential in diploid species with high heterozygosity levels, such as the different cultivars of grapevine.

In this context, we used third generation sequencing techniques to assemble T2T-haplotype specific references of several grapevine cultivars and interspecific hybrids. The high completeness and contiguity in the sequence have allowed us to study highly repetitive regions and to characterize them, both structurally and functionally.

Moreover, the availability of data at haplotype specific level, has allowed us to investigate and study the dynamics that shape the evolution of these regions within single haplotypes. Additionally, the comparison of these regions in different genotypes has also allowed to shed light on the processes and evolutionary steps that have affected the centromeric regions.

PacBio HiFi reads and Hifiasm software have been used to generate haplotype specific assemblies of one almost homozygous accession, 3 cultivated varieties, one wild accession and one interspecific hybrid of grapevines. Additional data coming from ONT long reads have allowed to assemble easily and in a very fast way all the 19 chromosomes of grapevine, at a single haplotype resolution. The high potential of these techniques in assembling highly complete genomes was proven by the presence of several chromosomal haplotypes being reconstructed in one single contig, from telomere to telomere. Moreover, the comparison of the newly assembled sequences with SyRI has confirmed the presence of high levels of variation between the different haplotypes. A big translocation between chromosome 01 and chromosome 11 of Rkatsiteli was visible in these results, confirming previous studies based on other techniques, such as mate pairs sequencing and PCR based experiments (Fornasiero, 2017). Many other main rearrangements were also present between different haplotypes, such as big inversions in chromosomes' arms, and several regions lacking syntheny, often present in centromeric regions. In some cases, long stretches of syntenic regions, sometimes

spanning almost all the chromosome suggest the presence of shared haplotypes, inherited by descent, like on chromosome 04 between PN40024 and Cabernet Franc hap1 and between Cabernet Franc hap2 and Rkatsiteli hap1.

The high quality of the assembled genomes has then allowed the in-depth characterization of centromeric regions. Similarly to what was already found from preliminary studies on grapevine centromeres, several repeat families (such as the 107 bp repeat family) and centrophilic TEs (such as Athila elements and TE containing chromodomains) were found in these regions. The combination of all these elements led to a high variability in terms of composition and dimension of centromeric regions. Moreover, the variability was observed not only between different chromosomes, but also between different haplotypes of the same chromosome, sometimes even within the same individual, with no differences in diversity levels between cultivated and wild accessions. Despite the high variability observed in all chromosomes, some were characterized by more distinctive features that made them consistent across all haplotypes. Chromosome 03, for example, was the only one characterized by arrays of a variant of the 107 bp repeat family, which was 135 bp long, and was composed of a combination of a whole 107 bp monomer attached to just part of it (28 bp coming from the 107 bp monomer). Also in chromosome 18, the composition and structure of the centromere was conserved across all haplotypes, lacking completely the arrays of 107 bp monomer family, which were present in almost all the haplotypes of all centromeres.

The variation in terms of sequence composition was also observed in structure of the centromeres. Two main structure types were identified in all the centromeric regions. The first one (*type A*) was characterized by big tandem repeats arrays of the 107bp monomers spanning in some cases several hundreds of kbs, which often surround a central core composed of other repeat families (59 bp, 28 bp and Gyp-28) that present a highly symmetrical and ordered structure. In addition to the big arrays of tandem repeats, other repeat families as well as TEs were present in the more outer regions, such as 66bp repeats, intermixed with repeats annotated as Athila retrotransposons and other TE families containing chromodomains, such as Galadriel and Tekay. These structures resemble characteristics discovered in other species, such as the big arrays composed of one tandem repeat families found in Human, where the 171bp α-satellite forms megabase-scale arrays, and characteristics found in *Arabidopsis*, where both large tandem repeat arrays of the 178 bp repeat as well as Athila TEs are dominating the centromeric regions. The second one (*type B*) on the other hand, is lacking completely the 107 bp repeat arrays, presenting a more homogeneous but less organized structure. This type of centromeres is characterized by intermixed sequences of 66 bp repeats, retrotransposons of the Athila family and other chromodomain containing TEs, which

compose the vast majority of the centromeric region. Also in this case, 59 bp, 28 bp and Gyp-28 repeats were localized in a core, despite the absence of a symmetrical structure as in *type A* centromeres. Moreover, in centromeres with *type B* structure, higher frequencies of intact TEs and solo-LTR were observed, suggesting a higher activity of TEs compared to the ones dominated by large tandem repeat arrays.

The high variability observed in centromeric regions of grapevine is in line with the theory known as the "centromere paradox" (Henikoff, Ahmad and Malik, 2001), in which high variability in these regions is contraposed to highly conserved functions and proteins interacting with the genetic sequence. This was highlighted not only by the wide range in diversity in composition of centromeric region, but also in structure of the centromeres. Despite the high variability, common features were present in all assembled centromeres: repeats of the 59 bp family were in fact the only conserved repeat present in all types of structures and centromeric repeat composition. This peculiar characteristic suggested its implication in the interaction with the functional proteins of the centromeres, despite their small relative abundance in the centromere. Moreover, analogies between the big 107 bp arrays and the arrays of the functional centromeric repeats in Human and *Arabidopsis* raised the necessity of evidence coming from complementary sources that would indicate the actual functional region of the centromere.

The presence of publicly available data coming from ChIP-seq experiments targeting the centromeric histone variant CENH3 in Pinot Noir and Chardonnay allowed us to investigate which parts of the centromeric regions were the functional ones. Mapping of reads coming from these experiments in identical-by-descent haplotypes led to the identification of the core enriched in CENH3, which confirmed the previous hypothesis of the 59 bp repeat family as the functional one. Even though the higher enrichment levels were observed in the regions masked with this repeat family, neighbouring regions saw some low yet relevant levels of enrichment of CENH3, probably due to limitations in the resolution of ChIP-seq experiments, and the difficulties in mapping short-read sequences against highly repetitive regions. Nevertheless, differently from what has been proposed in the past, the largest and most predominant arrays of the 107 bp repeat monomer do not seem to be involved in the functional component of the centromere (also emphasized by their absence on chromosome 18). This highlights the peculiar situation found in grapevine centromeres, which is very different from what is widely described in Arabidopsis and humans, where the functional repeat families are also the ones that form the largest tandemly repeated arrays.

To get a more comprehensive view of the characteristics of the centromeric regions, thanks to the availability of the methylation state of cytosine with ONT reads, the

epigenetic landscape of these regions has been investigated. From a chromosome scale point of view, two main compartments have been highlighted, characterized by different levels of methylation and different composition in term of TEs and gene density. As happens in most of the monocentric chromosomes of several species, 2 main compartments have been observed in grapevine chromosomes: higher methylation levels in all 3 contexts have been measures in centromeric regions, consistently with a higher abundance in TEs and a depletion in coding sequences; in contrast, lower levels of methylation as well as lower TEs abundance and higher gene density were present in the two chromosome arms. Focusing on centromeric regions, both similarities and differences with other monocentric plant species were present. Similarly to what happens in *Arabidopsis*, centromeres were highly methylated in the CG context, especially in the large arrays composed of the 107 bp tandem repeats. High methylation levels were also observed for the CHG and CHH contexts. High levels of methylation were also present in regions annotated as TEs, in line with common patterns that are happening even outside of centromeres, where transposons are highly methylated in order to silence and repress their activity (Kim and Zilberman, 2014). Despite the overall trend of methylation pattern in centromeric regions due to the high abundance of tandem repeats and TEs, depletion in DNA methylation was observed in correspondence of some of the predicted coding sequences. These regions were present mostly in correspondence of gaps in the regions densely covered by tandem repeats or TEs. Interestingly, this variation in the methylation levels, even inside centromeric regions which tend to be highly methylated, suggests a fine scale regulation of DNA methylation levels. Similar patterns in the regulation of methylation were also observed in other plant species, usually characterized by other types of centromere structures. Fine scale regulation of methylation was in fact observed in holocentric centromeres of *Rhynchospora pubera* where centromeric repeats and expressed genes are present in adjacent regions (Hofstatter *et al.*, 2022).

The variability in structure and composition that has been observed in these regions has then prompted us to investigate the mechanisms that drive the evolution of these highly ordered and structured regions. To do so, information contained in TEs have been used as molecular clock to estimate insertion time and set boundaries to the duplication and inversion events that affected the centromeric regions. Due to the insertion mechanism based on the generation of a cDNA intermediate, at the insertion time, LTR retrotransposons will have 2 main characteristics:

1. They will contain identical LTRs at the two ends. During the generation of the cDNA intermediate, in fact, one of the two initial LTR is used as template for the generation of the other LTR present at the end of the cDNA molecule.

2. During the insertion process, two identical sequences called target site duplications (TSDs) are generated from the insertion position, generating a fingerprint of the insertion event.

These two characteristics can then be used to estimate the insertion time by exploiting point 1: assuming a constant mutation rate, time can be inferred by spotting differences accumulated independently in the two LTRs. Moreover point 2 can be used to link multiple TEs that originated from the same insertion event: identical TSDs in different TEs implies a common original insertion event, and a subsequent duplication of the entire region containing the TE, that led to the duplication of the transposon without an actual TE activity. The presence of intact TEs in both duplicated and inverted regions in chromosome 09 of PN40024 has enabled the reconstruction of the events that led to the generation of the highly structured and symmetrical core of the centromere. A comparison of the structure found in chromosome 09 of PN40024 with the other haplotypes assembled in the remaining accessions has given an even wider and more complete overview of how the different mechanisms can lead to the generation of such diverse structures. Interestingly, no clustering in terms of centromere similarity was found between the cultivated and the wild groups. This suggests the presence of multiple mechanisms, acting alone or together, that drive the evolution of the centromere structure present in each haplotype. One possibility might leverage on the independence of the selection acting on these regions compared to the one acting on rest of chromosome's arms and genes of the same haplotype, allowing a rapid evolution of the centromeric regions and generating structures and variants that are markedly more different between haplotypes of the same individual than those coming from more distantly related accessions. By contrast, the presence of important traits in regions close to the centromeres, and used for selection, might influence the stability and conservation of some centromeric structures in the population, leading to very similar structures present in distant accessions. Few examples of these similarities between distant haplotypes can be found in chromosome 09 of PN40024 and VS-1 hap2, where centromeric sequences are almost identical and in chromosome 18, where the composition and main structure of all the assembled haplotypes tend to be highly similar. On the other hand, hap1 and hap2 of chromosome 09 of Rkatsiteli are very different, with different structures that co-exist within the same individual. Even more drastic cases of haplotype diversity in centromeric regions within the same individual, even though being a interspecific hybrid (with haplotypes coming from two different species), was observed in chromosome 02 of Kober 5BB, where the centromeric composition and structure was each associated to one of the two main centromere types (type A for hap2 and type B for hap1).

Moreover, expansions of the tandem repeat arrays add another layer in the whole picture of forces that are shaping the centromeric diversity, with both variation in terms of length of tandem repeated arrays (especially of the 107 bp monomers) as well as similarity between sequences composing the different arrays. Comparison of the monomers composing the arrays in PN40024 and Rkatsiteli hap1 haplotypes highlighted a clusterization of the sequences mostly based on monomers of the same chromosomes rather than the genotype. This suggests that the variation observed in one haplotype of a specific chromosome is strongly influenced by the initial diversity of monomers present in the array in parental haplotypes of the same chromosome. This peculiarity was also confirmed by the prevalence of the HOR of length two (dimeric version of the 107 bp repeat) in chromosome 07, which was conserved across all the accessions, as well as the presence of the 135 bp instead of the canonical 107 bp repeat in chromosome 03 and the almost complete absence of the 107 bp repeat family in chromosome 18 of all the genotypes. Similar and even neater patterns have been already described in *Arabidopsis*, where the higher similarity between the same chromosome in different accessions rather than between different chromosomes of the same accession is even stronger, likely due to the strong selection acting on the big tandem repeat arrays formed by the functional repeat family (Wlodzimierz *et al.*, 2023). Moreover, different evolutionary layers have been spotted in almost all the arrays of the 107 bp repeat family. Several expansion events led to the formation of blocks composed of similar variants, that originated from monomers present in the previous evolutionary layer, or from multiple independent events originated from the same layer.

The dynamicity of these regions and the recent activity of several mechanisms that modify their structures and composition, underlines how different forces can generate or reduce diversity. The alternation of these two forces can lead to subsequent cycles of differentiation and homogenization of centromeres. TE activity for example can be a source of variation, both adding new sequences via insertions or by deleting sequences, with the generation of soloLTRs or via recombination between similar TEs. On the other hand, duplication events of big chunks of sequence or recombination of repeated sequences can lead to the formation of highly organized structures, consequently reducing variation at a small scale. As exposed in Wlodzimierz et al., 2023, in *Arabidopsis*, cycles of TE invasion can lead to speciation and generation of variation, while a co-existing but opposite mechanism can reduce this activity via homologous recombination using sister chromatid as template, eliminating TE inserted in one of the two chromatids. With a similar mechanism, variants of tandem repeats present in one of the two haplotypes can be “seeded” on the sister one, leading to a reduction in TE frequency, but increasing diversity in structures and repeats, leading to speciation of centromeric regions.

Lastly, in an even more complex mechanism where both repeat families and TEs are involved, TE activity can be a source of seeding of tandem repeats too, as shown in studies in *Rhynchospora pubera*. In this holocentric species, in fact, repeats associated with CENH3 variants (*Tyba* repeats) were found in a non-autonomous element of the Helitron family (TCR1) that spread then several *Tyba*-containing loci all over the genome, as a resulted of TCR1 activity.

In conclusion, the evolution of centromeric regions is driven by a complex network of interacting mechanisms involving TEs, tandem repeats and recombination processes. The interplay between these processes generates recurrent cycles of diversification and homogenization, producing centromeres that are simultaneously dynamic and highly organized.

8 Conclusions

To conclude, this PhD project provided an example of application of long reads techniques and their potential to study both structurally and functionally the genome of a non-model plant species.

In the first part of the project, the focus was on the generation of highly complete reference genomes. The combination of PacBio HiFi reads and ONT ultra long reads allowed the reconstruction of a highly accurate and complete reference sequence of the 6 grapevine accessions, while the use of an assembly software based on graphs allowed the generation of haplotype specific references, leading to the generation of 11 haplotype specific references, assembled in most cases from telomere to telomere. The highly informative results formed the bases to the following analysis on the similarities and diversities of these genotypes.

The second part of the PhD project focused mainly on the characterization of unexplored sequences, such as the centromeric ones. Thanks to the completeness of the assemblies, we were able to identify several repeat families composing the grapevine centromeres, which were varying both across chromosomes within the same individual as well as between genotypes. Moreover, a deeper look at the structures found in these regions allowed to uncover highly organized structures and large arrays composed of tandem repeats, shaped by structural variation events as well as high TEs activity.

Moreover, the use of publicly available ChIP-seq data targeting the centromere-specific histone variant CENH3, together with the epigenetic information of the methylation status of cytosines coming from ONT data, allowed the identification of the putative functional part of the centromere and the involved repeat family. The identified repeat family (59bp repeats), which surprisingly differed from the one identified in preliminary studies present in literature, was the only one that was present in all the assembled haplotypes of all varieties and co-localized with CENH3 enrichment peaks. The functional annotation of these regions was then completed by characterizing the methylation levels of cytosines in all three contexts, with a haplotype specific resolution, revealing both the high methylation patterns at a broad scale level, as well as the small regulation of methylation at a fine scale.

Lastly, the comparison of the structures between the assembled haplotypes allowed the study of the forces acting on these regions. In particular, we showed how several duplication and inversion events shaped the structure and evolution of the centromeric regions in chromosome 09, thanks to the presence of intact LTR-TEs that allowed us to date and order the different structural re-arrangements.

The present study demonstrates the potential of long sequencing techniques to generate and study newly assembled genomes. The high quality of the results and the richness of the information that can be obtained with these methods represents an insightful tool to study genomes of poorly known species both from a structural and functional point of view.

9 References

Altemose, N. *et al.* (2025) "Complete genomic and epigenetic maps of human centromeres," *Science*, 376(6588), p. eabl4178. Available at: https://doi.org/10.1126/science.abl4178.

Altschul, S.F. *et al.* (1990) "Basic local alignment search tool," *Journal of Molecular Biology*, 215(3), pp. 403–410. Available at: https://doi.org/https://doi.org/10.1016/S0022-2836(05)80360-2.

Aradhya, M.K. *et al.* (2003) "Genetic structure and differentiation in cultivated grape, Vitis vinifera L.," *Genetical Research*. 2003/06/25, 81(3), pp. 179–192. Available at: https://doi.org/DOI: 10.1017/S0016672303006177.

Brown, M.R., Manuel Gonzalez de La Rosa, P. and Blaxter, M. (2025) "tidk: a toolkit to rapidly identify telomeric repeats from genomic datasets," *Bioinformatics*, 41(2), p. btaf049. Available at: https://doi.org/10.1093/bioinformatics/btaf049.

Calderón, L. *et al.* (2024) "Diploid genome assembly of the Malbec grapevine cultivar enables haplotype-aware analysis of transcriptomic differences underlying clonal phenotypic variation," *Horticulture Research*, 11(5), p. uhae080. Available at: https://doi.org/10.1093/hr/uhae080.

Canaguier, A. *et al.* (2017) "A new version of the grapevine reference genome assembly (12X.v2) and of its annotation (VCost.v3)," *Genomics Data*, 14, pp. 56–62. Available at: https://doi.org/https://doi.org/10.1016/j.gdata.2017.09.002.

Chen, J. *et al.* (2023) "A complete telomere-to-telomere assembly of the maize genome," *Nature Genetics*, 55(7), pp. 1221–1231. Available at: https://doi.org/10.1038/s41588-023-01419-6.

Chen, S. *et al.* (2018) "fastp: an ultra-fast all-in-one FASTQ preprocessor," *Bioinformatics*, 34(17), pp. i884–i890. Available at: https://doi.org/10.1093/bioinformatics/bty560.

Cheng, H. *et al.* (2021) "Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm," *Nature Methods*, 18(2), pp. 170–175. Available at: https://doi.org/10.1038/s41592-020-01056-5.

Chin, C.-S. *et al.* (2016) "Phased diploid genome assembly with single-molecule real-time sequencing," *Nature Methods*, 13(12), pp. 1050–1054. Available at: https://doi.org/10.1038/nmeth.4035.

Cochetel, N. *et al.* (2023) “A super-pangenome of the North American wild grape species,” *Genome Biology*, 24(1), p. 290. Available at: https://doi.org/10.1186/s13059-023-03133-2.

Danecek, P. *et al.* (2021) “Twelve years of SAMtools and BCFtools,” *GigaScience*, 10(2), p. giab008. Available at: https://doi.org/10.1093/gigascience/giab008.

Daponte, A. *et al.* (2025) “CENdetectHOR: a comprehensive tool for CENtromere profiling and HOR detection,” *bioRxiv*, p. 2025.01.07.631657. Available at: https://doi.org/10.1101/2025.01.07.631657.

Dodson Peterson, J.C. *et al.* (2019) “Grape Rootstock Breeding and Their Performance Based on the Wolpert Trials in California,” in D. Cantu and M.A. Walker (eds.) *The Grape Genome*. Cham: Springer International Publishing, pp. 301–318. Available at: https://doi.org/10.1007/978-3-030-18601-2_14.

Drinnenberg, I.A. *et al.* (2014) “Recurrent loss of CenH3 is associated with independent transitions to holocentricity in insects,” *eLife*. Edited by A.A. Hyman, 3, p. e03676. Available at: https://doi.org/10.7554/eLife.03676.

Earnshaw, W.C. and Rothfield, N. (1985) “Identification of a family of human centromere proteins using autoimmune sera from patients with scleroderma,” *Chromosoma*, 91(3), pp. 313–321. Available at: https://doi.org/10.1007/BF00328227.

Ellegren, H. (2004) “Microsatellites: simple sequences with complex evolution,” *Nature Reviews Genetics*, 5(6), pp. 435–445. Available at: https://doi.org/10.1038/nrg1348.

Fornasiero, A. (2017) “Identification and mapping of loci controlling viability in Vitis vinifera crosses.” Available at: https://air.uniud.it/handle/11390/1132182.

Garg, V. *et al.* (2024) “Unlocking plant genetics with telomere-to-telomere genome assemblies,” *Nature Genetics*, 56(9), pp. 1788–1799. Available at: https://doi.org/10.1038/s41588-024-01830-7.

Di Gaspero, G. & Foria, S. (2015) *Grapevine Breeding Programs for the Wine Industry*. 1st ed. Edited by A. Reynolds. Elsevier.

Di Genova, A. *et al.* (2014) “Whole genome comparison between table and wine grapes reveals a comprehensive catalog of structural variants,” *BMC Plant Biology*, 14(1), p. 7. Available at: https://doi.org/10.1186/1471-2229-14-7.

Ghurye, J. *et al.* (2019) “Integrating Hi-C links with assembly graphs for chromosome-scale assembly,” *PLOS Computational Biology*. Edited by I. Ioshikhes, 15(8), p. e1007273. Available at: https://doi.org/10.1371/journal.pcbi.1007273.

Goel, M. *et al.* (2019) “SyRI: finding genomic rearrangements and local sequence differences from whole-genome assemblies,” *Genome Biology*, 20(1), p. 277. Available at: https://doi.org/10.1186/s13059-019-1911-0.

Goel, M. and Schneeberger, K. (2022) “plotsr: visualizing structural similarities and rearrangements between multiple genomes,” *Bioinformatics*, 38(10), pp. 2922–2926. Available at: https://doi.org/10.1093/bioinformatics/btac196.

Gompert, Z. *et al.* (2025) “Adaptation repeatedly uses complex structural genomic variation,” *Science*, 388(6744), p. eadp3745. Available at: https://doi.org/10.1126/science.adp3745.

Guan, J. *et al.* (2021) “Genome structure variation analyses of peach reveal population dynamics and a 1.67 Mb causal inversion for fruit shape,” *Genome Biology*, 22(1), p. 13. Available at: https://doi.org/10.1186/s13059-020-02239-1.

Gurevich, A. *et al.* (2013) “QUAST: quality assessment tool for genome assemblies,” *Bioinformatics*, 29(8), pp. 1072–1075. Available at: https://doi.org/10.1093/bioinformatics/btt086.

Henikoff, S., Ahmad, K. and Malik, H.S. (2001) “The Centromere Paradox: Stable Inheritance with Rapidly Evolving DNA,” *Science*, 293(5532), pp. 1098–1102. Available at: https://doi.org/10.1126/science.1062939.

Hofstatter, P.G. *et al.* (2022) “Repeat-based holocentromeres influence genome architecture and karyotype evolution,” *Cell*, 185(17), pp. 3153-3168.e18. Available at: https://doi.org/10.1016/j.cell.2022.06.045.

Jaillon, O. *et al.* (2007) “The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla,” *Nature*, 449(7161), pp. 463–467. Available at: https://doi.org/10.1038/nature06148.

Katoh, K., Rozewicki, J. and Yamada, K.D. (2019) “MAFFT online service: multiple sequence alignment, interactive sequence choice and visualization,” *Briefings in Bioinformatics*, 20(4), pp. 1160–1166. Available at: https://doi.org/10.1093/bib/bbx108.

Kim, M.Y. and Zilberman, D. (2014) “DNA methylation as a system of plant genomic immunity,” *Trends in Plant Science*, 19(5), pp. 320–326. Available at: https://doi.org/10.1016/j.tplants.2014.01.014.

Kobayashi, S., Goto-Yamamoto, N. and Hirochika, H. (2004) “Retrotransposon-Induced Mutations in Grape Skin Color,” *Science*, 304(5673), p. 982. Available at: https://doi.org/10.1126/science.1095011.

Kolmogorov, M. *et al.* (2019) "Assembly of long, error-prone reads using repeat graphs," *Nature Biotechnology*, 37(5), pp. 540–546. Available at: https://doi.org/10.1038/s41587-019-0072-8.

Kovaka, S. *et al.* (2023) "Approaching complete genomes, transcriptomes and epi-omes with accurate long-read sequencing," *Nature Methods*, 20(1), pp. 12–16. Available at: https://doi.org/10.1038/s41592-022-01716-8.

Kuo, Y.-T. *et al.* (2023) "Holocentromeres can consist of merely a few megabase-sized satellite arrays," *Nature Communications*, 14(1), p. 3502. Available at: https://doi.org/10.1038/s41467-023-38922-7.

Langmead, B. *et al.* (2009) "Ultrafast and memory-efficient alignment of short DNA sequences to the human genome," *Genome Biology*, 10(3), p. R25. Available at: https://doi.org/10.1186/gb-2009-10-3-r25.

Letunic, I. and Bork, P. (2024) "Interactive Tree of Life (iTOL) v6: recent updates to the phylogenetic tree display and annotation tool," *Nucleic Acids Research*, 52(W1), pp. W78–W82. Available at: https://doi.org/10.1093/nar/gkae268.

Li, H. (2018) "Minimap2: pairwise alignment for nucleotide sequences," *Bioinformatics*, 34(18), pp. 3094–3100. Available at: https://doi.org/10.1093/bioinformatics/bty191.

Li, K. *et al.* (2023) "Identification of errors in draft genome assemblies at single-nucleotide resolution for quality assessment and improvement," *Nature Communications*, 14(1), p. 6556. Available at: https://doi.org/10.1038/s41467-023-42336-w.

Lischer, H.E.L. and Shimizu, K.K. (2017) "Reference-guided de novo assembly approach improves genome reconstruction for related species," *BMC Bioinformatics*, 18(1), p. 474. Available at: https://doi.org/10.1186/s12859-017-1911-6.

Liu, Y. *et al.* (2021) "DNA methylation-calling tools for Oxford Nanopore sequencing: a survey and human epigenome-wide evaluation," *Genome Biology*, 22(1), p. 295. Available at: https://doi.org/10.1186/s13059-021-02510-z.

Liu, Zhongjie *et al.* (2024) "Grapevine pangenome facilitates trait genetics and genomic breeding," *Nature Genetics*, 56(12), pp. 2804–2814. Available at: https://doi.org/10.1038/s41588-024-01967-5.

Logsdon, G.A. *et al.* (2024) "The variation and evolution of complete human centromeres," *Nature*, 629(8010), pp. 136–145. Available at: https://doi.org/10.1038/s41586-024-07278-3.

Logsdon, G.A., Vollger, M.R. and Eichler, E.E. (2020) “Long-read human genome sequencing and its applications,” *Nature Reviews Genetics*, 21(10), pp. 597–614. Available at: https://doi.org/10.1038/s41576-020-0236-x.

Macas, J. *et al.* (2023) “Assembly of the 81.6 Mb centromere of pea chromosome 6 elucidates the structure and evolution of metapolycentric chromosomes,” *PLOS Genetics*. Edited by C. Köhler, 19(2), p. e1010633. Available at: https://doi.org/10.1371/journal.pgen.1010633.

Magris, G. (2016) *Characterisation of the pan-genome of Vitis vinifera using Next Generation Sequencing*. University of Udine. Available at: https://hdl.handle.net/11390/1132723 (Accessed: November 27, 2025).

Magris, G. *et al.* (2021) “The genomes of 204 Vitis vinifera accessions reveal the origin of European wine grapes,” *Nature Communications*, 12(1), p. 7240. Available at: https://doi.org/10.1038/s41467-021-27487-y.

Mahadik, K. *et al.* (2019) “Scalable Genome Assembly through Parallel de Bruijn Graph Construction for Multiple k-mers,” *Scientific Reports*, 9(1), p. 14882. Available at: https://doi.org/10.1038/s41598-019-51284-9.

Mahmoud, M. *et al.* (2024) “Utility of long-read sequencing for All of Us,” *Nature Communications*, 15(1), p. 837. Available at: https://doi.org/10.1038/s41467-024-44804-3.

Marçais, G. *et al.* (2018) “MUMmer4: A fast and versatile genome alignment system,” *PLOS Computational Biology*, 14(1), pp. e1005944-. Available at: https://doi.org/10.1371/journal.pcbi.1005944.

Marques, A. *et al.* (2015) “Holocentromeres in Rhynchospora are associated with genome-wide centromere-specific repeat arrays interspersed among euchromatin,” *Proceedings of the National Academy of Sciences*, 112(44), pp. 13633–13638. Available at: https://doi.org/10.1073/pnas.1512255112.

Martin, M. *et al.* (2016) “WhatsHap: fast and accurate read-based phasing,” *bioRxiv*, p. 085050. Available at: https://doi.org/10.1101/085050.

McKinley, K.L. and Cheeseman, I.M. (2016) “The molecular basis for centromere identity and function,” *Nature Reviews Molecular Cell Biology*, 17(1), pp. 16–29. Available at: https://doi.org/10.1038/nrm.2015.5.

Melters, D.P. *et al.* (2012) “Holocentric chromosomes: convergent evolution, meiotic adaptations, and genomic analysis,” *Chromosome Research*, 20(5), pp. 579–593. Available at: https://doi.org/10.1007/s10577-012-9292-1.

Melters, D.P. *et al.* (2013) “Comparative analysis of tandem repeats from hundreds of species reveals unique insights into centromere evolution,” *Genome Biology*, 14(1), p. R10. Available at: https://doi.org/10.1186/gb-2013-14-1-r10.

Minio, A. *et al.* (2017) “How Single Molecule Real-Time Sequencing and Haplotype Phasing Have Enabled Reference-Grade Diploid Genome Assembly of Wine Grapes,” *Frontiers in Plant Science*, Volume 8-2017. Available at: https://doi.org/10.3389/fpls.2017.00826.

Minio, A., Cochetel, N., Vondras, A.M., *et al.* (2022) “Assembly of complete diploid-phased chromosomes from draft genome sequences,” *G3 Genes|Genomes|Genetics*, 12(8), p. jkac143. Available at: https://doi.org/10.1093/g3journal/jkac143.

Minio, A., Cochetel, N., Massonnet, M., *et al.* (2022) “HiFi chromosome-scale diploid assemblies of the grape rootstocks 110R, Kober 5BB, and 101–14 Mgt,” *Scientific Data*, 9(1), p. 660. Available at: https://doi.org/10.1038/s41597-022-01753-0.

Minton, K. (2024) “Tandem repeat variation of human centromeres,” *Nature Reviews Genetics*, 25(7), p. 455. Available at: https://doi.org/10.1038/s41576-024-00741-x.

Myles, S. *et al.* (2011) “Genetic structure and domestication history of the grape,” *Proceedings of the National Academy of Sciences*, 108(9), pp. 3530–3535. Available at: https://doi.org/10.1073/pnas.1009363108.

Naish, M. *et al.* (2021) “The genetic and epigenetic landscape of the Arabidopsis centromeres,” *Science*, 374(6569). Available at: https://doi.org/10.1126/science.abi7489.

Naish, M. *et al.* (2025) “The genetic and epigenetic landscape of the Arabidopsis centromeres,” *Science*, 374(6569), p. eabi7489. Available at: https://doi.org/10.1126/science.abi7489.

Naish, M. and Henderson, I.R. (2024) “The structure, function, and evolution of plant centromeres,” *Genome research*, 34(2), pp. 161–178. Available at: https://doi.org/10.1101/gr.278409.123.

Nurk, S. *et al.* (2020) “HiCanu: accurate assembly of segmental duplications, satellites, and allelic variants from high-fidelity long reads,” *Genome research*, 30(9), pp. 1291–1305. Available at: https://doi.org/10.1101/gr.263566.120.

Nurk, S. *et al.* (2022) “The complete sequence of a human genome,” *Science*, 376(6588), pp. 44–53. Available at: https://doi.org/10.1126/science.abj6987.

Ou, S. *et al.* (2019) "Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline," *Genome Biology*, 20(1), p. 275. Available at: https://doi.org/10.1186/s13059-019-1905-y.

Pei, D. *et al.* (2024) "The evolution and formation of centromeric repeats analysis in Vitis vinifera," *Planta*, 259(5), p. 99. Available at: https://doi.org/10.1007/s00425-024-04374-6.

Pop, M. (2009) "Genome assembly reborn: recent computational challenges," *Briefings in Bioinformatics*, 10(4), pp. 354–366. Available at: https://doi.org/10.1093/bib/bbp026.

Quinlan, A.R. and Hall, I.M. (2010) "BEDTools: a flexible suite of utilities for comparing genomic features," *Bioinformatics*, 26(6), pp. 841–842. Available at: https://doi.org/10.1093/bioinformatics/btq033.

Ramírez, F. *et al.* (2016) "deepTools2: a next generation web server for deep-sequencing data analysis," *Nucleic Acids Research*, 44(W1), pp. W160–W165. Available at: https://doi.org/10.1093/nar/gkw257.

Rang, F.J., Kloosterman, W.P. and de Ridder, J. (2018) "From squiggle to basepair: computational approaches for improving nanopore sequencing read accuracy," *Genome Biology*, 19(1), p. 90. Available at: https://doi.org/10.1186/s13059-018-1462-9.

Rautiainen, M. *et al.* (2023) "Telomere-to-telomere assembly of diploid chromosomes with Verkko," *Nature Biotechnology*, 41(10), pp. 1474–1482. Available at: https://doi.org/10.1038/s41587-023-01662-6.

Rhie, A. *et al.* (2021) "Towards complete and error-free genome assemblies of all vertebrate species," *Nature*, 592(7856), pp. 737–746. Available at: https://doi.org/10.1038/s41586-021-03451-0.

Sahu, S.K. and Liu, H. (2023) "Long-read sequencing (method of the year 2022): The way forward for plant omics research," *Molecular Plant*, 16(5), pp. 791–793. Available at: https://doi.org/https://doi.org/10.1016/j.molp.2023.04.007.

Shi, X. *et al.* (2023) "The complete reference genome for grapevine (Vitis vinifera L.) genetics and breeding," *Horticulture Research*, 10(5), p. uhad061. Available at: https://doi.org/10.1093/hr/uhad061.

Smit, A., Hubley, R. and Green, P. (2015) *RepeatMasker Open-4.0*, *http://www.repeatmasker.org*.

Sonnhammer, E.L.L. and Durbin, R. (1995) "A dot-matrix program with dynamic threshold control suited for genomic DNA and protein sequence analysis," *Gene*, 167(1),

pp. GC1–GC10. Available at: https://doi.org/https://doi.org/10.1016/0378-1119(95)00714-8.

Stanke, M. *et al.* (2008) "Using native and syntenically mapped cDNA alignments to improve de novo gene finding," *Bioinformatics*, 24(5), pp. 637–644. Available at: https://doi.org/10.1093/bioinformatics/btn013.

Stanojević, D. *et al.* (2024) "Telomere-to-Telomere Phased Genome Assembly Using HERRO-Corrected Simplex Nanopore Reads," *bioRxiv*, p. 2024.05.18.594796. Available at: https://doi.org/10.1101/2024.05.18.594796.

Stephens, Z. and Kocher, J.-P. (2024) "Characterization of telomere variant repeats using long reads enables allele-specific telomere length estimation," *BMC Bioinformatics*, 25(1), p. 194. Available at: https://doi.org/10.1186/s12859-024-05807-5.

Stevanovski, I. *et al.* (2025) "Comprehensive genetic diagnosis of tandem repeat expansion disorders with programmable targeted nanopore sequencing," *Science Advances*, 8(9), p. eabm5386. Available at: https://doi.org/10.1126/sciadv.abm5386.

Sweeten, A.P., Schatz, M.C. and Phillippy, A.M. (2024) "ModDotPlot—rapid and interactive visualization of tandem repeats," *Bioinformatics*, 40(8), p. btae493. Available at: https://doi.org/10.1093/bioinformatics/btae493.

Tabidze, V. *et al.* (2017) "Whole genome comparative analysis of four Georgian grape cultivars," *Molecular Genetics and Genomics*, 292(6), pp. 1377–1389. Available at: https://doi.org/10.1007/s00438-017-1353-x.

Talbert, P.B. *et al.* (2002) "Centromeric Localization and Adaptive Evolution of an Arabidopsis Histone H3 Variant," *The Plant Cell*, 14(5), pp. 1053–1066. Available at: https://doi.org/10.1105/tpc.010425.

Tao, Y. *et al.* (2024) "Atlas of telomeric repeat diversity in Arabidopsis thaliana," *Genome Biology*, 25(1), p. 244. Available at: https://doi.org/10.1186/s13059-024-03388-3.

Venter, J.C., Smith, H.O. and Hood, L. (1996) "A new strategy for genome sequencing," *Nature*, 381(6581), pp. 364–366. Available at: https://doi.org/10.1038/381364a0.

Wang, Yunhao *et al.* (2021) "Nanopore sequencing technology, bioinformatics and applications," *Nature Biotechnology*, 39(11), pp. 1348–1365. Available at: https://doi.org/10.1038/s41587-021-01108-x.

Wenger, A.M. *et al.* (2019) “Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome,” *Nature Biotechnology*, 37(10), pp. 1155–1162. Available at: https://doi.org/10.1038/s41587-019-0217-9.

Wetzel, J., Kingsford, C. and Pop, M. (2011) “Assessing the benefits of using mate-pairs to resolve repeats in de novo short-read prokaryotic assemblies,” *BMC Bioinformatics*, 12(1), p. 95. Available at: https://doi.org/10.1186/1471-2105-12-95.

Wicker, T. *et al.* (2007) “A unified classification system for eukaryotic transposable elements,” *Nature Reviews Genetics*, 8(12), pp. 973–982. Available at: https://doi.org/10.1038/nrg2165.

Wlodzimierz, P. *et al.* (2023) “Cycles of satellite and transposon evolution in Arabidopsis centromeres,” *Nature*, 618(7965), pp. 557–565. Available at: https://doi.org/10.1038/s41586-023-06062-z.

Xiao, H. *et al.* (2025) “Impacts of reproductive systems on grapevine genome and breeding,” *Nature Communications*, 16(1), p. 2031. Available at: https://doi.org/10.1038/s41467-025-56817-7.

Xu, L. and Seki, M. (2020) “Recent advances in the detection of base modifications using the Nanopore sequencer,” *Journal of Human Genetics*, 65(1), pp. 25–33. Available at: https://doi.org/10.1038/s10038-019-0679-0.

Yatskevich, S. *et al.* (2024) “Structure of the human outer kinetochore KMN network complex,” *Nature Structural & Molecular Biology*, 31(6), pp. 874–883. Available at: https://doi.org/10.1038/s41594-024-01249-y.

Yu, H.-G. *et al.* (2000) “The plant kinetochore,” *Trends in Plant Science*, 5(12), pp. 543–547. Available at: https://doi.org/10.1016/S1360-1385(00)01789-1.

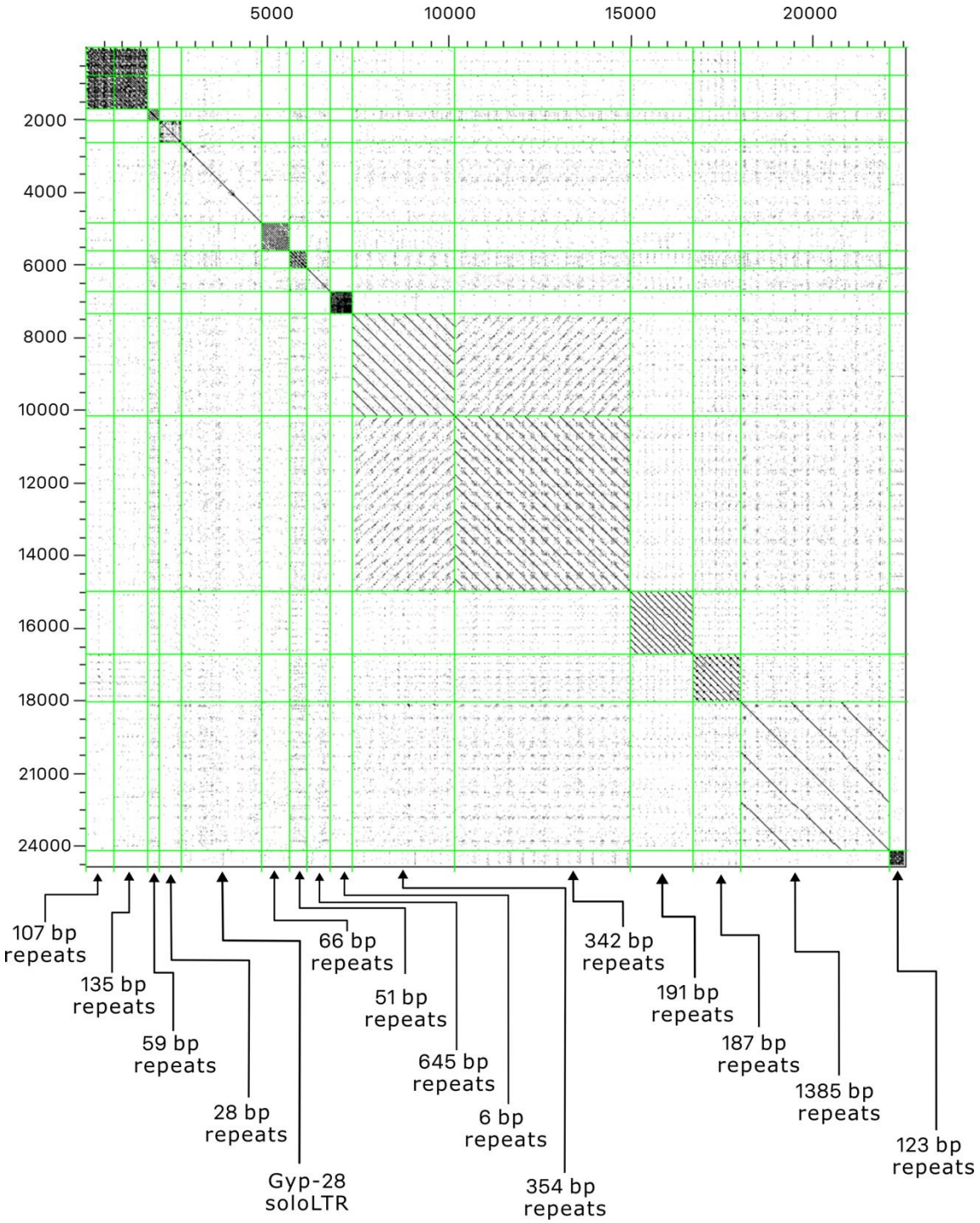
Zhou, Y. *et al.* (2019) “The population genetics of structural variants in grapevine domestication,” *Nature Plants*, 5(9), pp. 965–979. Available at: https://doi.org/10.1038/s41477-019-0507-8.

Zhou, Y. *et al.* (2022) “Duet: SNP-assisted structural variant calling and phasing using Oxford nanopore sequencing,” *BMC Bioinformatics*, 23(1), p. 465. Available at: https://doi.org/10.1186/s12859-022-05025-x.

10 Supplementary material

Supplementary Table 1: *TEs annotation statistics. Percentage of the haplotype-specific reference genome masked as TE family.*

		PN40024	Cabernet Franc hap1	Cabernet Franc hap2	Rkatsiteli hap1	Rkatsiteli hap2	Sultanina hap1	Sultanina hap2	VS-1 hap1	VS-1 hap2	Kober 5BB hap1	Kober 5BB hap2
	DNA transposon	0.06%	0.07%	0.06%	0.06%	0.06%	0.07%	0.06%	0.06%	0.07%	0.06%	0.06%
LTR	Copia	9.31%	9.84%	9.85%	10.73%	10.76%	9.70%	9.77%	9.86%	9.76%	9.39%	9.43%
	Gypsy	13.55%	13.07%	13.17%	14.13%	14.16%	13.83%	13.58%	12.56%	12.37%	11.93%	12.00%
	Retrovirus	0.64%	0.60%	0.59%	0.60%	0.63%	0.57%	0.62%	0.67%	0.63%	0.52%	0.51%
	unknown	4.47%	4.60%	4.61%	4.95%	4.91%	4.79%	4.54%	4.75%	4.58%	4.22%	4.41%
TIR	CACTA	2.99%	3.26%	3.28%	3.38%	3.43%	3.21%	3.23%	3.27%	3.22%	3.20%	3.23%
	Mutator	6.04%	5.96%	6.16%	6.53%	6.45%	6.26%	6.26%	5.97%	5.95%	6.49%	6.40%
	PIF Harbinger	1.02%	1.09%	1.09%	1.28%	1.22%	1.11%	1.10%	1.11%	1.06%	1.17%	1.11%
	Tc1 Mariner	0.28%	0.31%	0.32%	0.35%	0.34%	0.31%	0.31%	0.30%	0.28%	0.30%	0.31%
	hAT	1.56%	1.62%	1.60%	1.62%	1.63%	1.58%	1.57%	1.68%	1.63%	1.63%	1.61%
nonLTR	LINE element	2.25%	2.34%	2.31%	3.01%	2.85%	2.25%	2.31%	2.29%	2.40%	2.24%	2.21%
nonTIR	Helitron	1.62%	1.76%	1.73%	2.09%	2.00%	1.67%	1.74%	1.78%	1.77%	1.85%	1.80%
	Repeat region	5.07%	3.96%	3.71%	2.57%	2.33%	3.87%	3.65%	3.61%	3.35%	4.37%	3.56%
	Total	48.84%	48.49%	48.49%	51.21%	50.77%	49.23%	48.75%	47.92%	47.08%	47.36%	46.63%



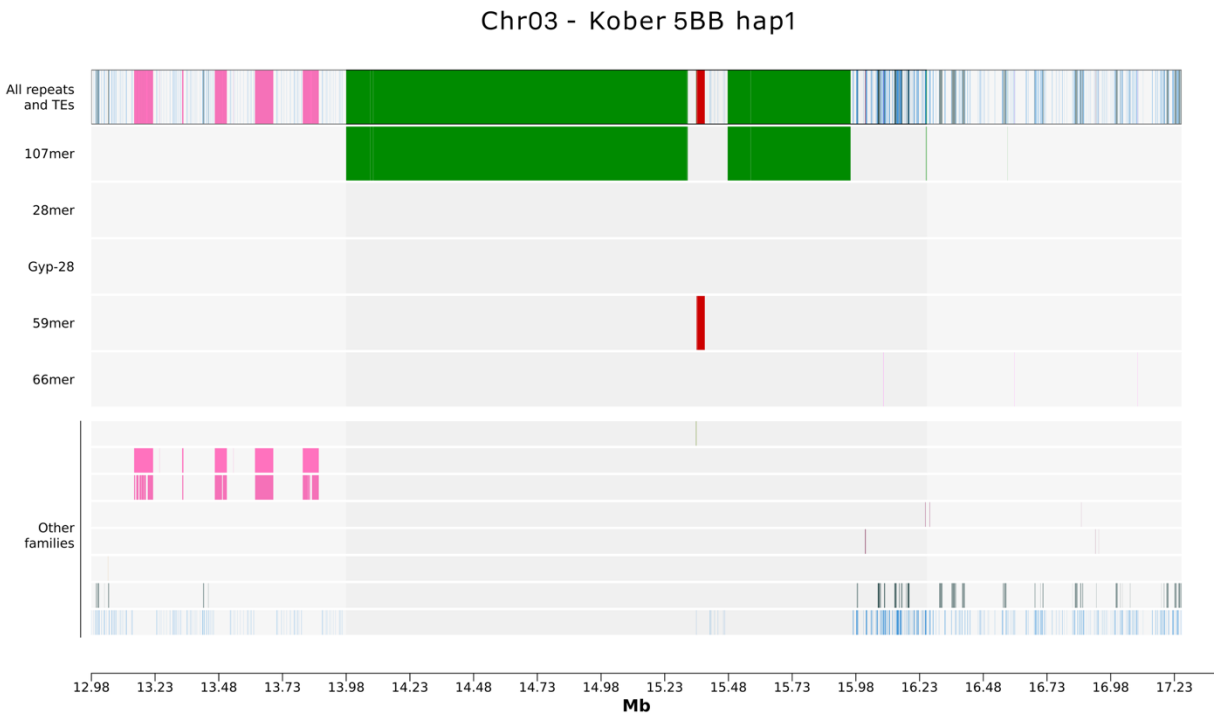
***Supplementary Figure 1**: Self dot plot (Dotter) of all centromeric repeats families. For all the repeat families but Gyp-28 LTR and the 645 bp repeat family, several consecutive monomers have been plotted.*

***Supplementary Table 2**: Table with start and end coordinates of centromeric regions. For each variety: first column corresponds to chromosome name; second column corresponds to start coordinate; third column corresponds to end coordinate. Empty cells correspond to centromeres reconstructed in 2 or more contigs (and therefore not considered).*

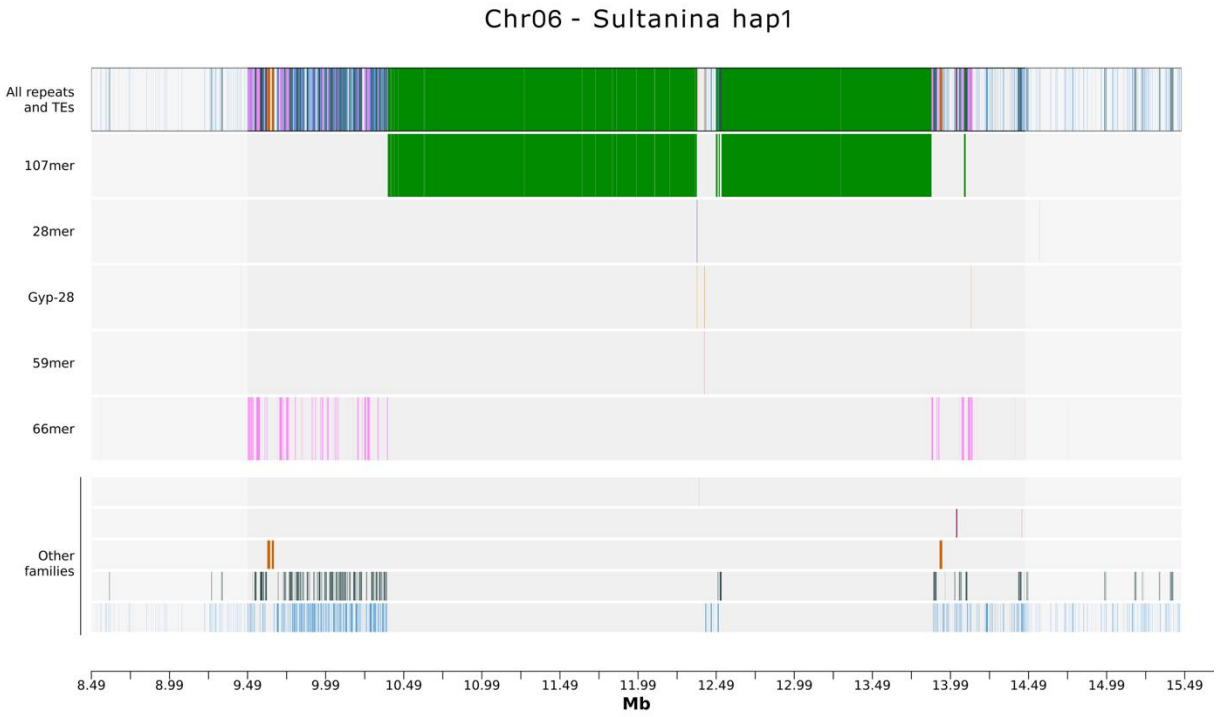
PN40024			Cabernet Franc		
chr01	13855596	17534356	cf.h1.chr01		
			cf.h2.chr01		
chr02	12272821	14574071	cf.h1.chr02		
			cf.h2.chr02	12011768	14437883
chr03	14454735	16588062	cf.h1.chr03		
			cf.h2.chr03		
chr04	12485684	13989815	cf.h1.chr04		
			cf.h2.chr04	11741257	13648923
chr05	12974957	15229600	cf.h1.chr05		
			cf.h2.chr05	12830755	14047976
chr06	9433842	12802725	cf.h1.chr06	9350610	10705731
			cf.h2.chr06		
chr07	12514852	14035635	cf.h1.chr07	12892362	14104948
			cf.h2.chr07		
chr08	6022624	8134871	cf.h1.chr08		
			cf.h2.chr08	6194894	9671392
chr09	14703562	16202296	cf.h1.chr09	15926577	17518979
			cf.h2.chr09		
chr10	19917208	22160292	cf.h1.chr10	19498792	20547619
			cf.h2.chr10		
chr11	13218486	14806960	cf.h1.chr11		
			cf.h2.chr11		
chr12	10166666	11691085	cf.h1.chr12	10122331	12657826
			cf.h2.chr12	9522738	10862023
chr13	11385989	12927925	cf.h1.chr13		
			cf.h2.chr13	11506708	13329769
chr14	14846855	15471482	cf.h1.chr14	14854669	15478970
			cf.h2.chr14	13980261	14431548
chr15	6861638	9216320	cf.h1.chr15		
			cf.h2.chr15		
chr16	8413622	14274318	cf.h1.chr16		
			cf.h2.chr16	8678711	14818195
chr17	15008025	15660532	cf.h1.chr17		
			cf.h2.chr17		
chr18	15015535	17785695	cf.h1.chr18	14928802	17801004
			cf.h2.chr18	14563489	17724058
chr19	15507334	19079192	cf.h1.chr19		
			cf.h2.chr19	18151905	19767470

Rkatsiteli			Sultanina		
rk.h1.chr01	15705283	16766782	su.h1.chr01	14457150	16881072
rk.h2.chr01	13747518	15343127	su.h2.chr01	13764645	15672509
rk.h1.chr02	12421415	15369281	su.h1.chr02	12576524	15265630
rk.h2.chr02	12053720	14491893	su.h2.chr02	12175107	15126815
rk.h1.chr03	14505593	15571726	su.h1.chr03	14216836	16464665
rk.h2.chr03	13614972	14160699	su.h2.chr03	13274029	13923675
rk.h1.chr04			su.h1.chr04	12269320	14983488
rk.h2.chr04	12557956	14289004	su.h2.chr04	11862724	13859788
rk.h1.chr05	12842315	14636685	su.h1.chr05	11513459	14407198
rk.h2.chr05	12914035	15224487	su.h2.chr05	11761383	14910501
rk.h1.chr06	9587680	13541260	su.h1.chr06	9489429	14468908
rk.h2.chr06	9601640	13453896	su.h2.chr06	9439891	13938833
rk.h1.chr07	12342869	13076939	su.h1.chr07	12582056	13645955
rk.h2.chr07	12651342	13382481	su.h2.chr07	12403658	13125858
rk.h1.chr08	6018012	9397548	su.h1.chr08	6052779	9428775
rk.h2.chr08	6098170	8973183	su.h2.chr08	6140654	9618127
rk.h1.chr09	15346942	16768744	su.h1.chr09	16811188	19530500
rk.h2.chr09	16537908	18936299	su.h2.chr09	16002057	18782880
rk.h1.chr10	19031314	19958561	su.h1.chr10	20254398	22057521
rk.h2.chr10	19925455	21977023	su.h2.chr10	19966624	23234294
rk.h1.chr11			su.h1.chr11	13553923	15257120
rk.h2.chr11	13345673	14984097	su.h2.chr11	13635393	15297521
rk.h1.chr12	10090524	12374799	su.h1.chr12	9752418	11750185
rk.h2.chr12	9691376	11187144	su.h2.chr12	4542903	6241805
rk.h1.chr13	11825221	13507597	su.h1.chr13	11818926	13108117
rk.h2.chr13			su.h2.chr13	11678015	13449935
rk.h1.chr14	14271406	15844381	su.h1.chr14	14084157	14697860
rk.h2.chr14	14122831	14939920	su.h2.chr14	14629767	15391632
rk.h1.chr15	7231079	7611645	su.h1.chr15	8192772	10200846
rk.h2.chr15	7504556	7893601	su.h2.chr15	7091347	8789002
rk.h1.chr16	7721696	9870131	su.h1.chr16	9827947	12581048
rk.h2.chr16	8661826	13215365	su.h2.chr16	8013847	10637826
rk.h1.chr17	14129288	16908448	su.h1.chr17	14658823	16874079
rk.h2.chr17	14385381	16562227	su.h2.chr17	14537826	16613961
rk.h1.chr18	14896929	17759689	su.h1.chr18	14932115	17605072
rk.h2.chr18	14623318	18061405	su.h2.chr18	15089649	17864481
rk.h1.chr19	17271094	18709582	su.h1.chr19	15361296	20732674
rk.h2.chr19	16058302	17180164	su.h2.chr19	17873207	19880872

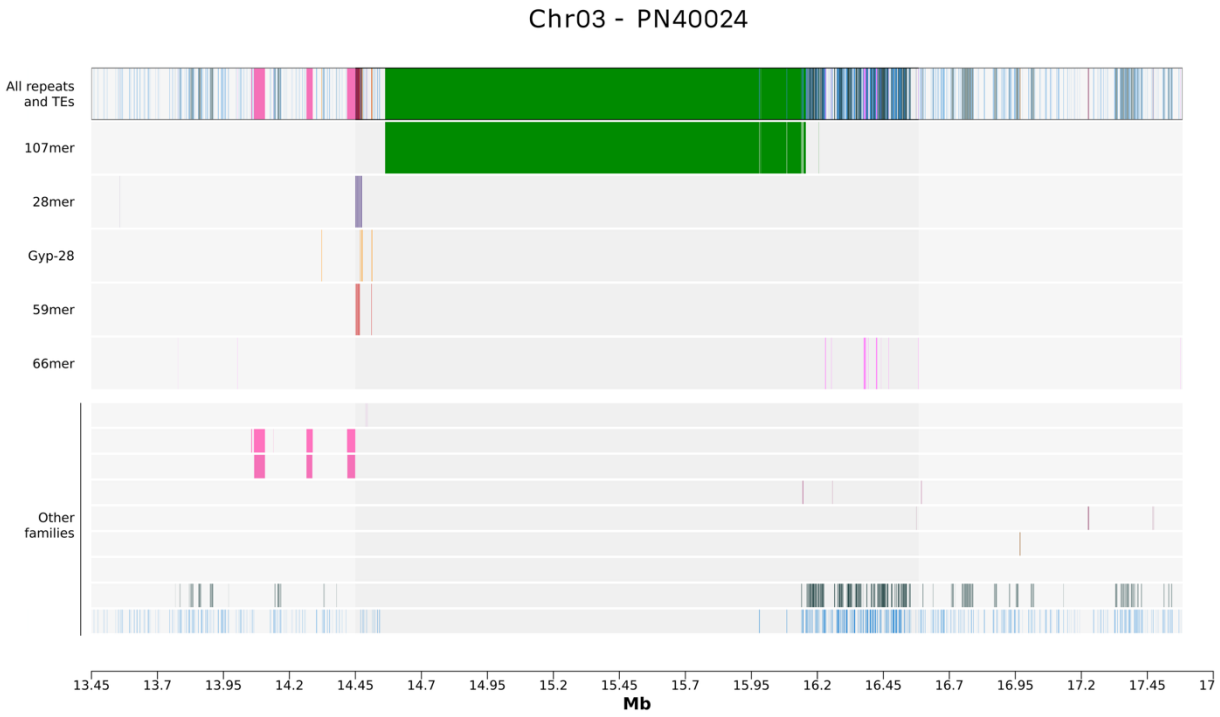
VS-1			Kober 5BB		
vs.h1.chr12			kb.h1.chr12	13331269	15853427
vs.h2.chr12			kb.h2.chr12		
vs.h1.chr18			kb.h1.chr18	12011766	14304177
vs.h2.chr18			kb.h2.chr18	11933006	14635938
vs.h1.chr04			kb.h1.chr04	13978115	16256134
vs.h2.chr04	14379301	15446731	kb.h2.chr04	14893395	15872632
vs.h1.chr08			kb.h1.chr08	11736906	13374314
vs.h2.chr08			kb.h2.chr08		
vs.h1.chr13			kb.h1.chr13		
vs.h2.chr13	12764740	14016671	kb.h2.chr13		
vs.h1.chr16			kb.h1.chr16	9632653	13297595
vs.h2.chr16			kb.h2.chr16	9302325	13218552
vs.h1.chr19	12339195	13489553	kb.h1.chr19		
vs.h2.chr19			kb.h2.chr19	12809281	15292697
vs.h1.chr02	5960179	7964191	kb.h1.chr02	6135972	7869300
vs.h2.chr02			kb.h2.chr02	6099114	9022609
vs.h1.chr04			kb.h1.chr04	14923706	17440343
vs.h2.chr04	14522713	16272846	kb.h2.chr04	14674048	16819697
vs.h1.chr08			kb.h1.chr08		
vs.h2.chr08			kb.h2.chr08	20100702	21280789
vs.h1.chr11			kb.h1.chr11	13427869	14681792
vs.h2.chr11			kb.h2.chr11		
vs.h1.chr13			kb.h1.chr13	9449052	11870745
vs.h2.chr13			kb.h2.chr13	9357415	14510979
vs.h1.chr15	11825285	13543681	kb.h1.chr15	13872046	15392894
vs.h2.chr15	11564239	12914916	kb.h2.chr15		
vs.h1.chr03	15550566	16361651	kb.h1.chr03	14234897	15805135
vs.h2.chr03	13067138	13828975	kb.h2.chr03	13210392	15068361
vs.h1.chr07	6624344	10266121	kb.h1.chr07	8061106	9482485
vs.h2.chr07			kb.h2.chr07		
vs.h1.chr09			kb.h1.chr09		
vs.h2.chr09			kb.h2.chr09	9718678	11987620
vs.h1.chr12			kb.h1.chr12		
vs.h2.chr12			kb.h2.chr12	14088889	16980021
vs.h1.chr16	14827068	17836710	kb.h1.chr16		
vs.h2.chr16	15106526	18126814	kb.h2.chr16	14697877	16927864
vs.h1.chr18			kb.h1.chr18		
vs.h2.chr18			kb.h2.chr18		



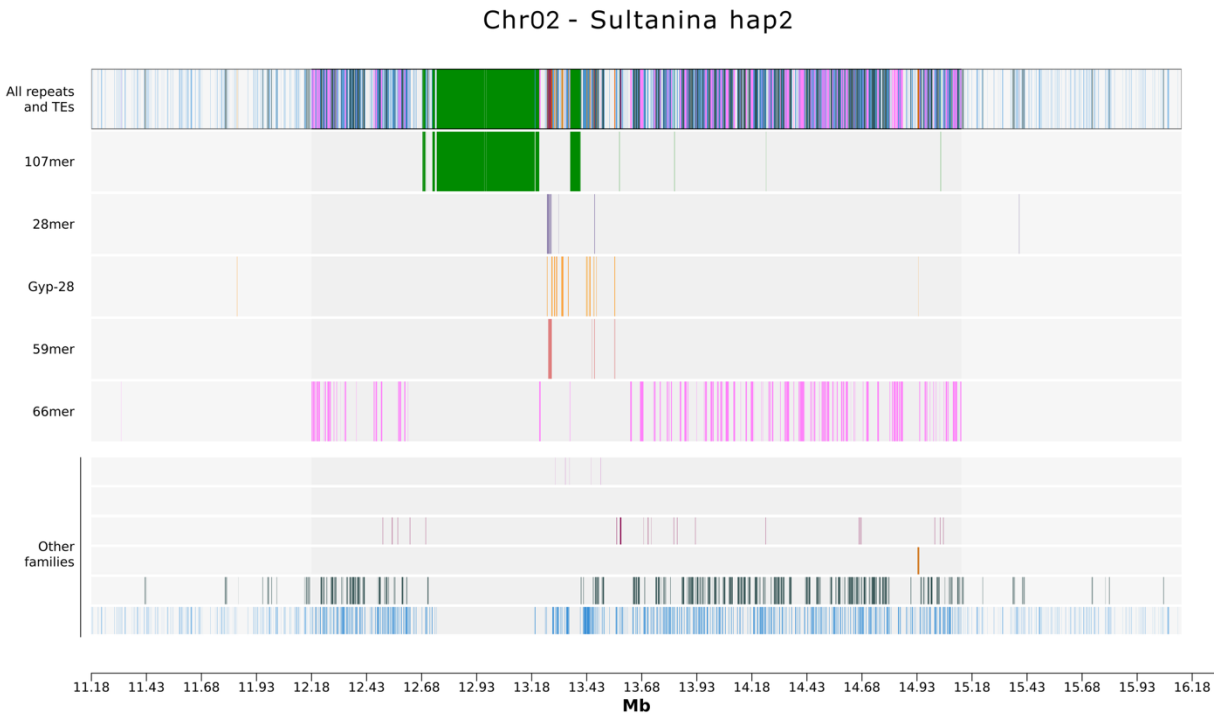
***Supplementary Figure 2**: Repeat landscape of centromeric region of chromosome 03 of Kober 5BB hap1. Peri-centromeric regions (1 Mb upstream and downstream of the centromere coordinates) are plotted with faded colours. From the top track: All centromeric repeats families and TEs families, 107 bp repeat family, 28 bp repeat family, Gyp-28 soloLTR repeats, 59 bp repeat family and 66 bp repeat family. Bottom tracks reffer to minor repeat families and TEs families: 51 bp repeat family, 354 bp repeat family, 342 bp repeat family, 191 bp repeat family, 187 repeat family, 123 bp repeat family, Athila repeats and TE containing chromodomains.*



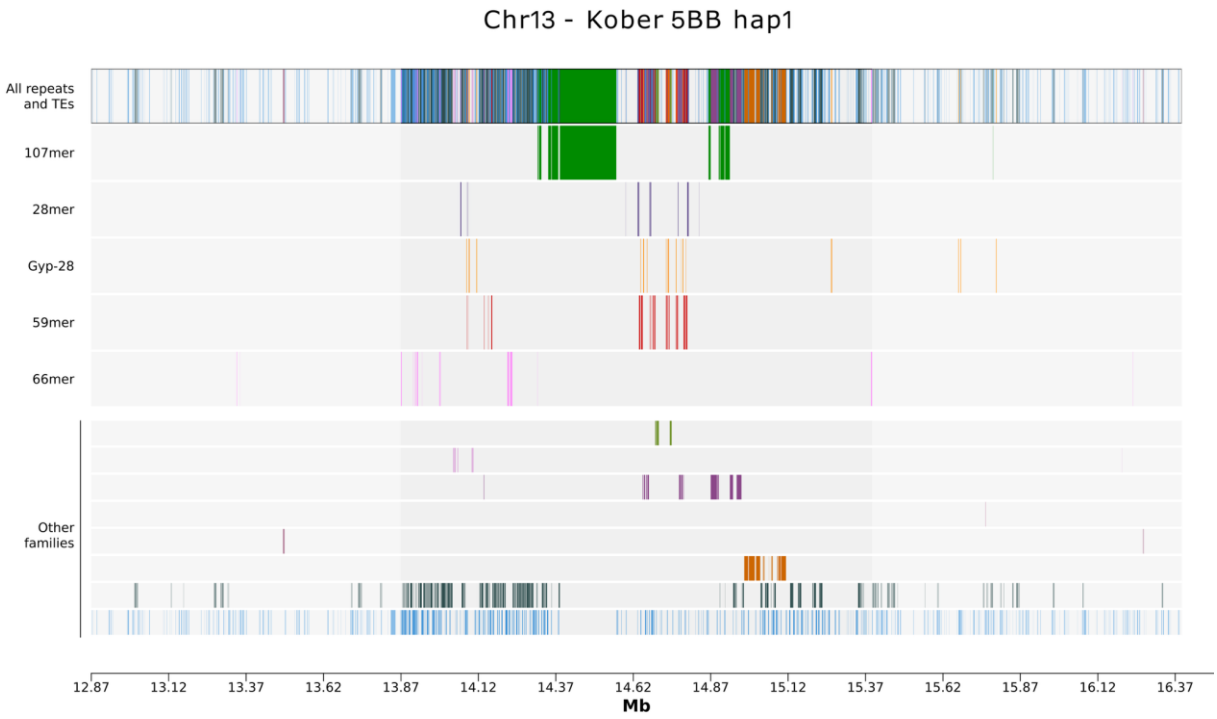
***Supplementary Figure 3**: Repeat landscape of centromeric region of chromosome 06 of Sultanina hap1. Peri-centromeric regions (1 Mb upstream and downstream of the centromere coordinates) are plotted with faded colours. From the top track: All centromeric repeats families and TEs families, 107 bp repeat family, 28 bp repeat family, Gyp-28 soloLTR repeats, 59 bp repeat family and 66 bp repeat family. Bottom tracks reffer to minor repeat families and TEs families: 645 bp repeat family, 191 bp repeat family, 123 bp repeat family, Athila repeats and TE containing chromodomains.*



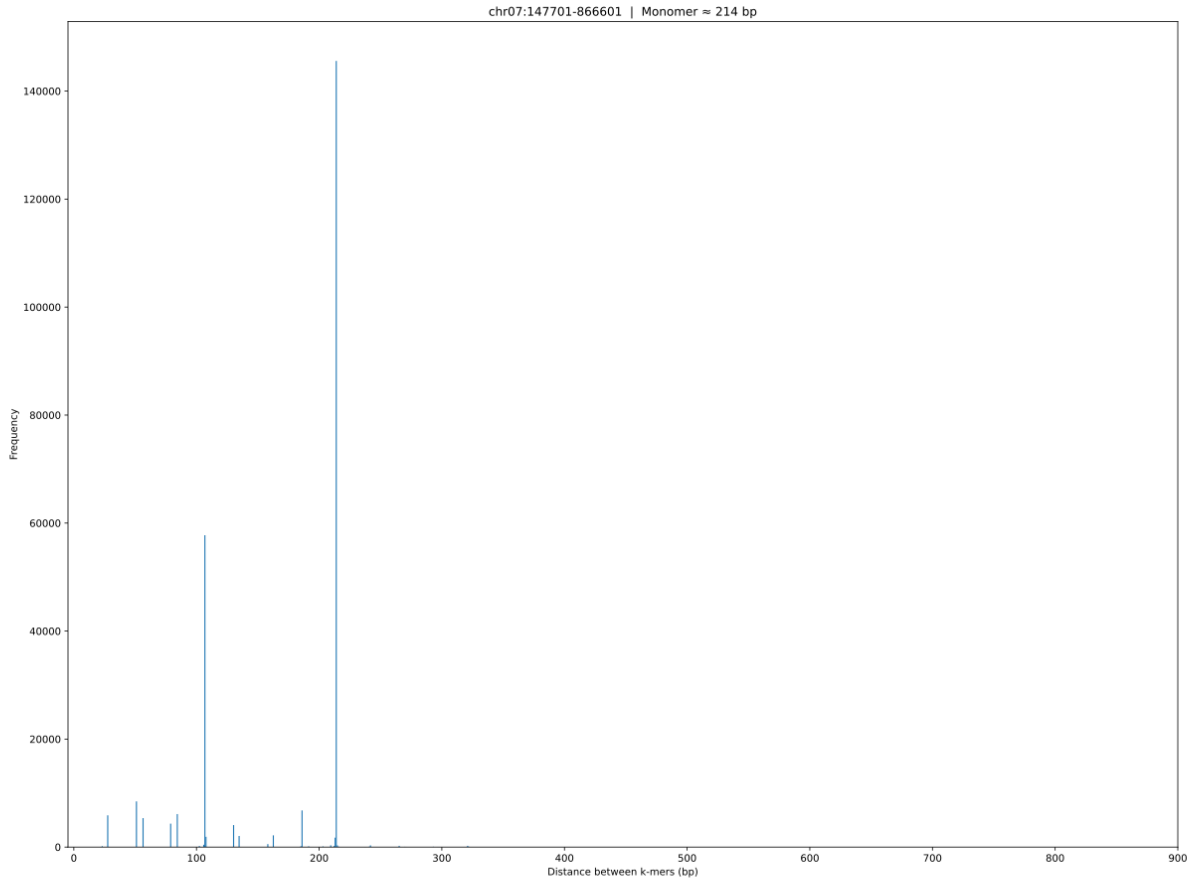
***Supplementary Figure 4**: Repeat landscape of centromeric region of chromosome 03 of PN40024. Peri-centromeric regions (1 Mb upstream and downstream of the centromere coordinates) are plotted with faded colours. From the top track: All centromeric repeats families and TEs families, 107 bp repeat family, 28 bp repeat family, Gyp-28 soloLTR repeats, 59 bp repeat family and 66 bp repeat family. Bottom tracks reffer to minor repeat families and TEs families: 645 bp repeat family, 354 bp repeat family, 342 bp repeat family, 191 bp repeat family, 187 bp repeat family, 1385 bp repeat family, 123 bp repeat family, Athila repeats and TE containing chromodomains.*



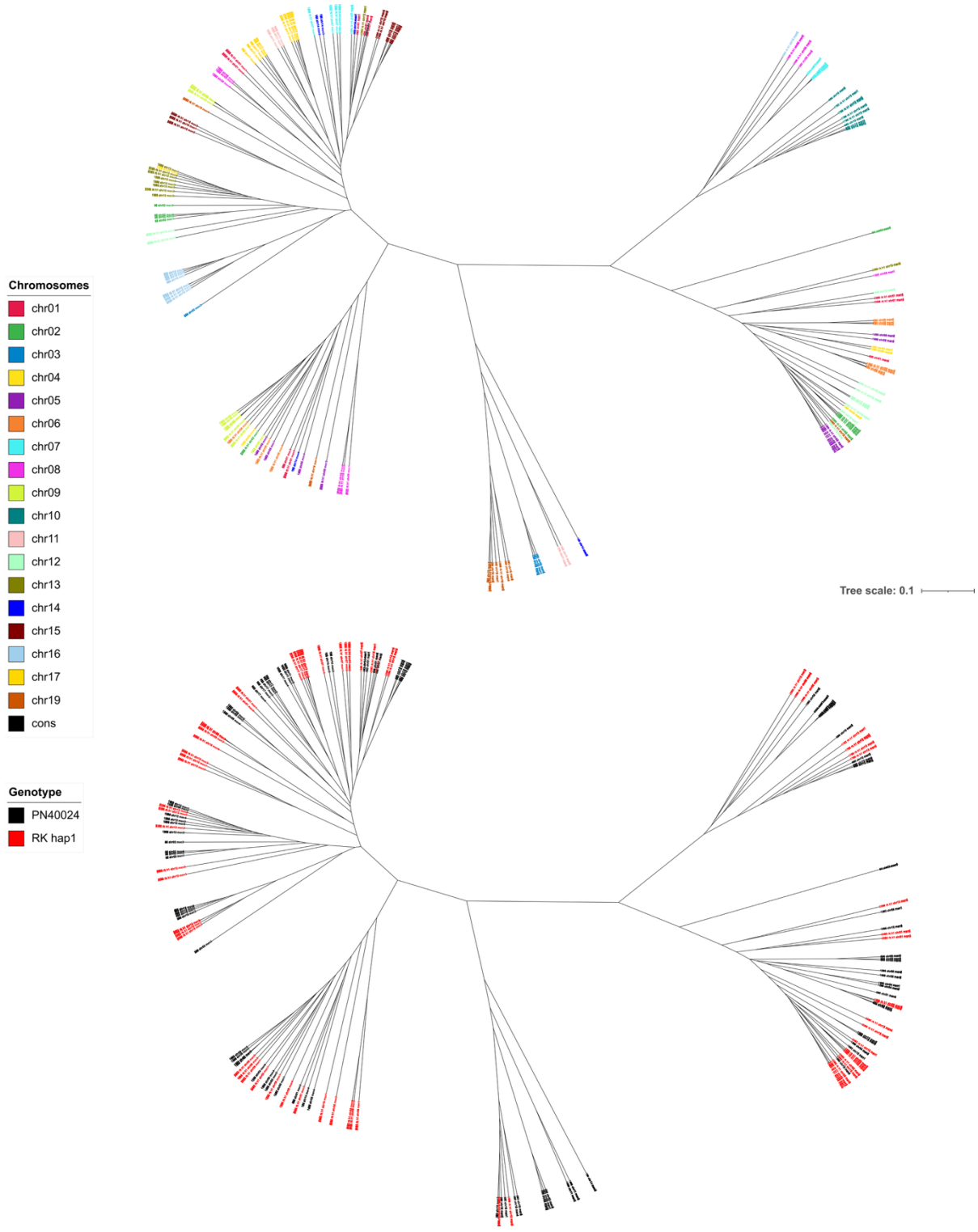
***Supplementary Figure 5**: Repeat landscape of centromeric region of chromosome 02 of Sultanina hap2. Peri-centromeric regions (1 Mb upstream and downstream of the centromere coordinates) are plotted with faded colours. From the top track: All centromeric repeats families and TEs families, 107 bp repeat family, 28 bp repeat family, Gyp-28 soloLTR repeats, 59 bp repeat family and 66 bp repeat family. Bottom tracks reffer to minor repeat families and TEs families: 645 bp repeat family, 6 bp repeat family, 191 bp repeat family, 123 bp repeat family, Athila repeats and TE containing chromodomains.*



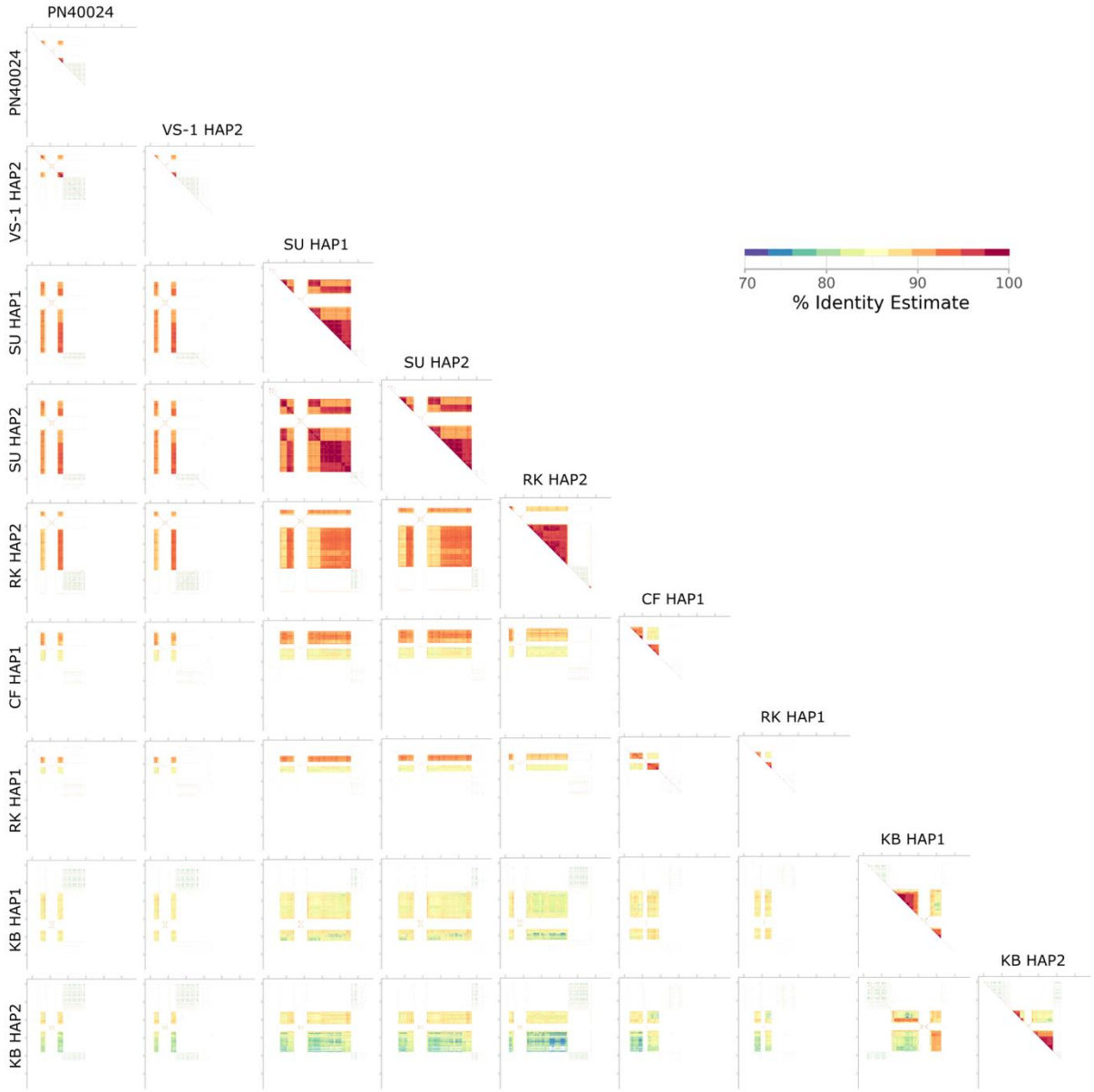
***Supplementary Figure 6**: Repeat landscape of centromeric region of chromosome 13 of Kober 5BB hap1. Peri-centromeric regions (1 Mb upstream and downstream of the centromere coordinates) are plotted with faded colours. From the top track: All centromeric repeats families and TEs families, 107 bp repeat family, 28 bp repeat family, Gyp-28 soloLTR repeats, 59 bp repeat family and 66 bp repeat family. Bottom tracks reffer to minor repeat families and TEs families: 51 bp repeat family, 645 bp repeat family, 6 bp repeat family, 191 bp repeat family, 187 bp repeat family, 123 bp repeat family, Athila repeats and TE containing chromodomains.*



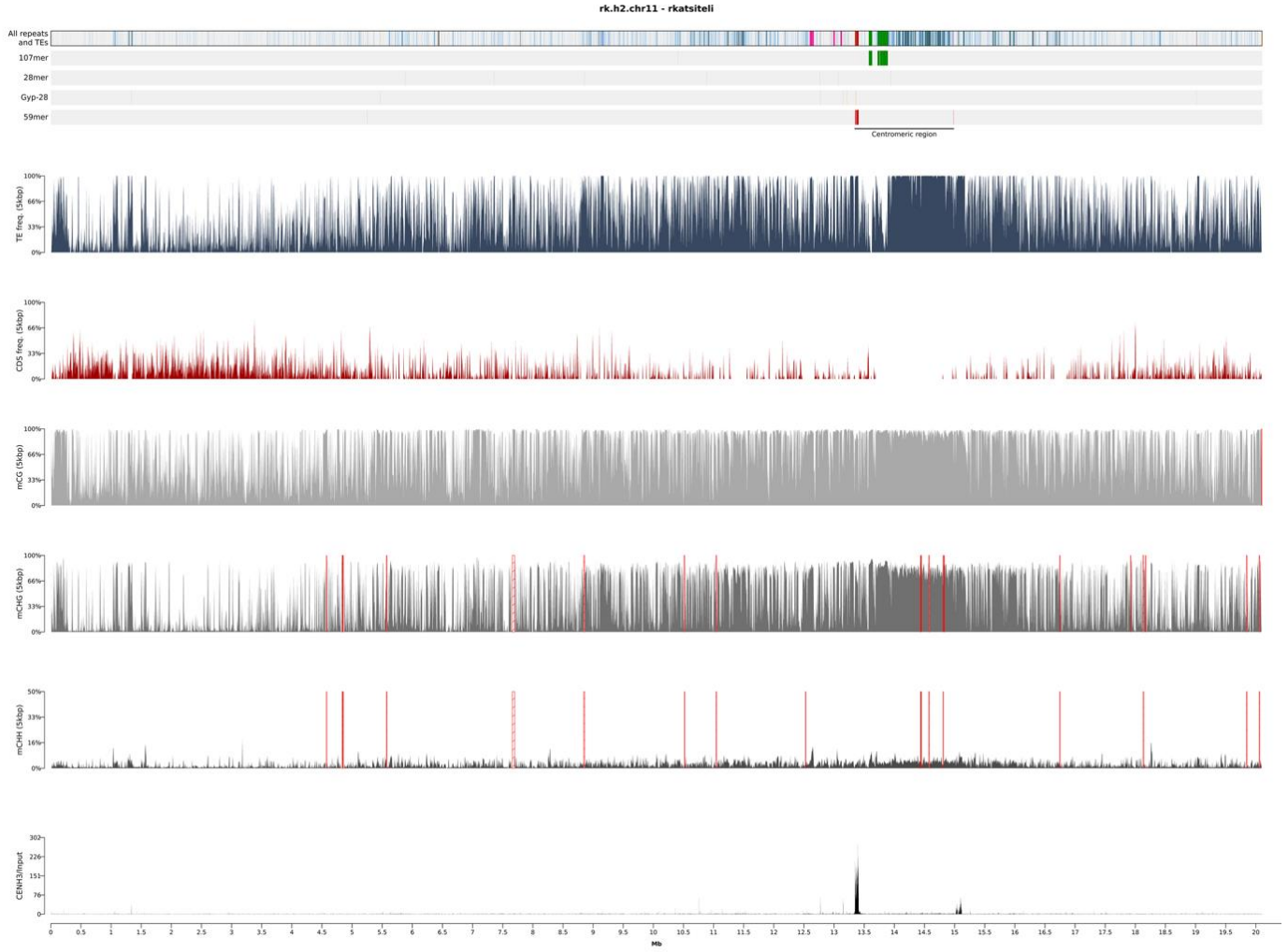
***Supplementary Figure 7**: Histogram plot from CENdetectHOR pipeline showing frequency of k-mer distances inside the 107 bp tandem repeat array. Highest peak at 214 bp suggests a higher abundance of identical k-mers at a distance of 214 bp, corresponding to dimers of the 107 bp monomer.*



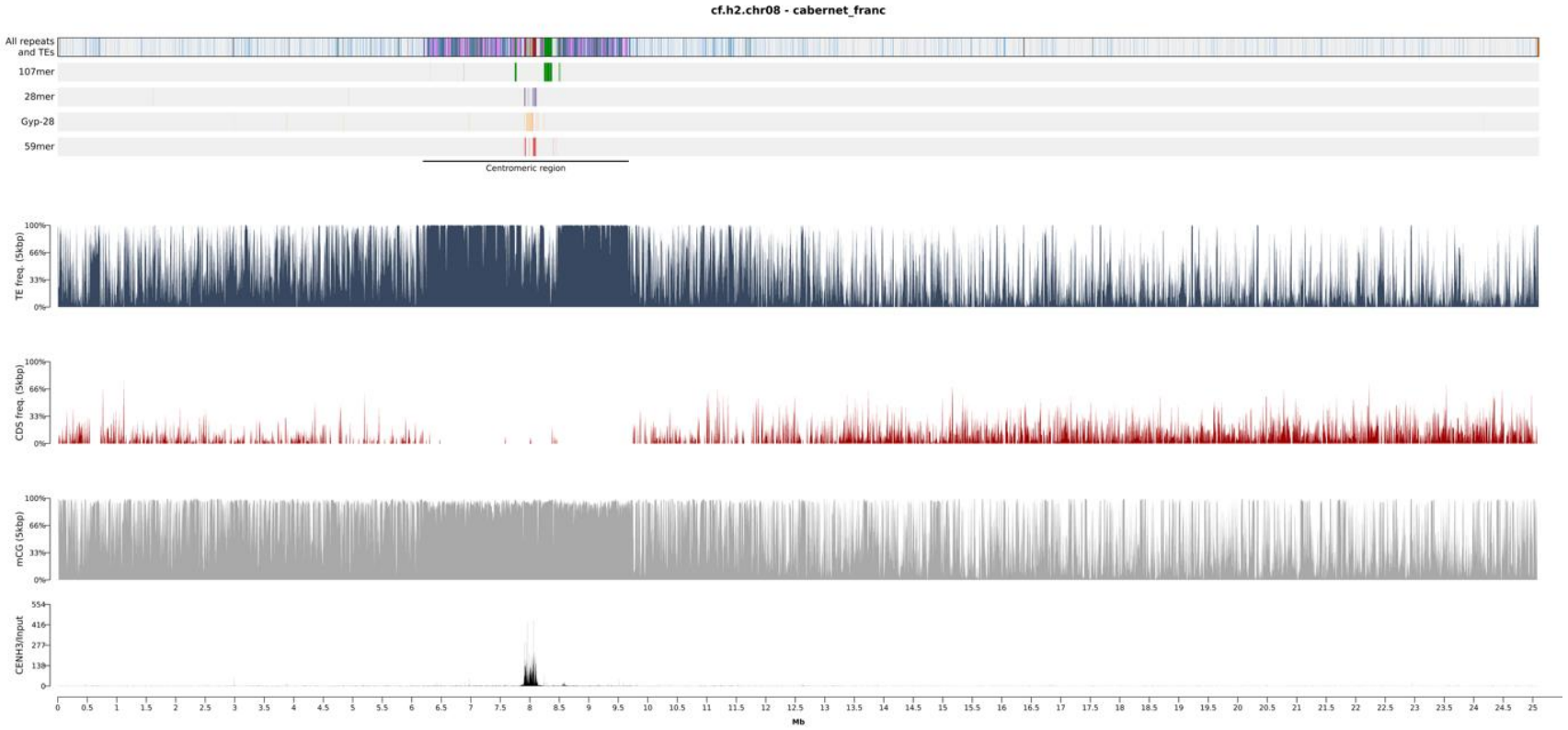
***Supplementary Figure 8**: Unrooted tree of monomers composing the 107 bp repeat family arrays in PN40024 and Rkatsiteli hap1. The tree was generated based on Mafft multi-alignment of the monomers coming from both genotypes and extracted by CENdetectHOR pipeline. In the upper plot, labels are coloured based on the chromosome of belonging. In the bottom plot, labels are coloured based on the genotype of belonging.*



Supplementary Figure 9: *Comparison of all haplotypes reconstructed for chromosome 09 using ModDotPlot. Colour refers to sequence identity in 1kbp windows. Just windows with >70% identity have been plotted. SU = Sultanina; RK = Rkatsiteli; CF= Cabernet Franc; KB = Kober 5BB.*



Supplementary Figure 10: *Methylation landscape at a chromosome level in chromosome 11 of Rkatsiteli hap2. From the top to the bottom section: repeat and TEs families; percentage of bp annotated as TEs in 5kbp windows; percentage of bp annotated as CDS in 5kbp windows; average methylation levels of CG context in 5kbp windows; average methylation levels of CHG context in 5kbp windows; average methylation levels of CHH context in 5kbp windows; enrichment levels calculated as ratio of coverage of CENH3/Input in 5kbp windows. Red boxes correspond to regions with no methylation information, due to insufficient coverage.*



Supplementary Figure 11: *Methylation landscape at a chromosome level in chromosome 08 of Cabernet Franc hap2. From the top to the bottom section: repeat and TEs families; percentage of bp annotated as TEs in 5kbp windows; percentage of bp annotated as CDS in 5kbp windows; average methylation levels of CG context in 5kbp windows; enrichment levels calculated as ratio of coverage of CENH3/Input in 5kbp windows. Red boxes correspond to regions with no methylation information, due to insufficient coverage.*

11 Acknowledgements

I would like to thank Professor Morgante for the guidance and insights throughout this PhD, and Professor Marroni for his support during this work.

I am also grateful to Gabriele and Gabbo for their help, discussions, and for the personal support that went well beyond the academic side.

My thanks extend to all the people at Institute of Applied Genomics and IGATech, whose expertise and collaboration contributed to my growth, and with whom I shared many enjoyable lunch and coffee breaks.

Lastly, I want to thank the dearest ones in my life, whose constant support made this journey possible.