JOINT PhD PROGRAMME IN MOLECULAR BIOLOGY

Coordinator: Prof. Giuseppe Legname

CICLE XXX

PhD Thesis

Chromatin Conformation Analysis in *Vitis Vinifera*

SUPERVISOR

Prof. Michele Morgante

PhD STUDENT

Aldo Tocci

CO-SUPERVISORS

Fabio Marroni, PhD

Rachel M. Schwope, PhD

ACADEMIC YEAR 2017/2018

Contents

1 INTRODUCTION 6
1.1 A three-dimensional genome 6
1.1.1 The central biological role of chromatin organization 6
1.1.1 Chromosome territories 10
1.1.2 Structural Domains 13
1.1.3 Self-interacting domains 16
1.1.4 Chromatin loops 19
1.2 The study of the three-dimensional genome in plants 21
1.3 Vitis vinifera genome 26
1.4 The Plant Pan-genome and the NOVABREED project 28
2 AIM OF THE THESIS 33
3 METHODS 35
3.1 Hi-C methods and materials 35
3.1.1 I. Nuclei Preparation 35
3.1.2 II. Digestion 36
3.1.3 III. Biotinylation and Ligation 36
3.1.4 IV. Phenol Chloroform Extraction 37
3.1.5 V. Sonication 37
3.1.6 VI. Biotin Pull-down 38
3.1.7 VII. End Repair and Adapter Ligation 38
3.1.8 VIII. PCR amplification of library and sequencing 39
3.2 Samples characteristics 40
3.3 Hi-C data analysis 41
3.3.1 Homer 41
3.3.2 HiC-Pro 43
3.3.3 Chromosome neighbourhood analysis 45
3.3.4 Distance Dependent Decay function 45
3.4 A/B Compartments analysis 46

3.4.1 Identification of A/B compartments via PCA 46
3.4.2 Functional properties of A/B compartments 47
3.5 Sub-compartment domains analysis 48
3.5.1 Identification and characterization of sub-compartment domains 48
3.5.2 Domain borders analysis 49
3.6 Chromatin loops 50
3.6.1 Effect of SVs on loops detection 51
3.6.2 Characterization of loop interactions 52
3.7 SVs and grapevine chromatin conformation 53
3.7.1 Directionality Index 53
3.7.2 Simulation of large deletions, insertions and inversions 54
3.7.3 Allele-specific Hi-C maps 55
3.7.4 Analysis of chromatin contacts across CNV borders 56
4 RESULTS AND DISCUSSION 59
4.1 Vitis vinifera 3D genome 59
4.1.1 *Vitis vinifera* chromatin organization in the interphasic nucleus. 59
4.1.2 A/B compartments 69
4.1.3 Sub-compartment domains 83
4.1.4 Chromatin loops 92
4.2 SVs effect on grapevine chromatin conformation 101
4.2.1 Simulation of SV presence in Hi-C contact maps 101
4.2.2 Allele-specific Hi-C maps 106
4.3 Spatial proximity is a necessary condition for SV occurrence 112
5 CONCLUSIONS 117
6 BIBLIOGRAPHY 120

1 INTRODUCTION

1.1 A THREE-DIMENSIONAL GENOME

1.1.1 The central biological role of chromatin organization

Genomes are organized in complex structures inside the three-dimensional space of the cell nucleus (Dekker and Misteli, 2015). This physical organization must be non-random, since the information stored into the DNA molecule has to be accessible to the diverse mechanisms of reading, interpretation and propagation (Ramani, Shendure and Duan, 2016). This is not a trivial task, since the length of the DNA molecule can be several orders of magnitude greater than the nuclear space, particularly for eukaryotes. For example, the 23 chromosomes of the human genome account for a total linear length of 2 meters, but such length is compressed into a nucleus with a diameter of 10 micrometres (Felsenfeld and Groudine, 2003).

The solution that ensures compression and functionality at the same time consists in the organization of the DNA molecule into a hierarchy of structural levels (Figure 1).

At the base of this hierarchy there is the DNA molecule packed into chromatin fibers. Chromatin fibers are composed of genomic DNA and histone proteins, and the basic unit of these structures is the nucleosome, made of 147 DNA base pairs (bp) wrapped around a histone octamer (Kornberg, 1974).

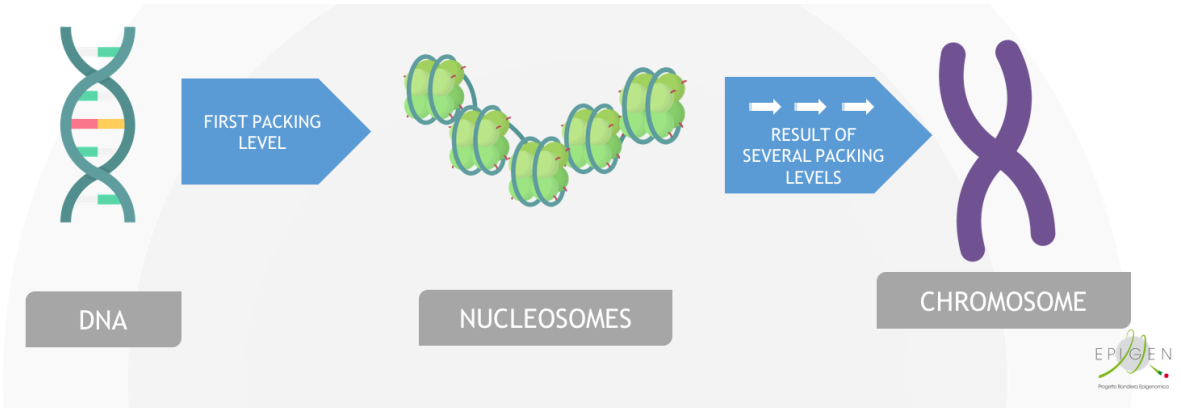


Figure 1 Cartoon representing the main step in DNA structural organization, from basic chromatin fiber formation to fully compressed chromosome (adapted from EPIGEN).

The top level of DNA compression is represented by the metaphase fully compacted chromosomes, formed during the cell division, in which DNA is almost inaccessible to any regulatory signal (Felsenfeld and Groudine, 2003). However, the most complex and poorly understood DNA organization is assumed during interphase. In this context the compromise between convenient three-dimensional structure and efficient function constitutes the central biological role of the chromatin organization (Dekker and Misteli, 2015).

The dynamic structure of the chromatin can affect the functions of the genome from DNA replication and DNA repair to transcription and gene regulation (Cavalli and Misteli, 2013), promoting the interaction between sites at a short distance range; at a long range between distant sites of the same chromosome, and between sites from different chromosomes.

Chromatin structure can be investigated using novel methods designed to reveal physical contacts between loci in regions of interest and across the genome (Lajoie, Dekker and Kaplan, 2015). These methods are based on the chromosome conformation capture (3C) technology

(Dekker *et al.*, 2002), which allows for the capture of chromatin contacts from selected regions at a time.

In order to extend the possibilities of the study from a targeted experiment to a genome wide and high throughput-oriented approach, several methods based on the original 3C have been developed. They include 4C (Chromosome Conformation Capture-on-Chip; Simonis et al. 2006); 5C (Chromosome Conformation Capture Carbon Copy) which results in a genome-wide contacts interrogation for a given locus (Dostie *et al.*, 2006); Hi-C (Lieberman-Aiden *et al.*, 2009) and 3C-seq (Duan *et al.*, 2012) which is able to reconstruct the complete set of interactions for all the loci in the genome. These interactions are represented as a NxN (where N is the number of *loci* captured in the experiment) matrix (contact map), which is commonly represented as a heatmap with colour intensity representing the frequency of contact between any two *loci* (Figure 3).

The above mentioned methods enhanced the study of the 3D organization of chromatin in the interphase nucleus; nonetheless, the research in this field is constantly evolving with new data from single-cell based assays (Nagano *et al.*, 2013; Stevens *et al.*, 2017) and information on the modulation in time of the nucleus architecture, progressing from a 3D to a 4D perspective (Dekker *et al.*, 2017).

Finally, an increasing number of works is pointing the attention to a new way of considering the chromatin inside the nucleus. In particular, it has been proposed that the nucleus is formed by droplets of locally condensed DNA-binding proteins (Figure 2), giving rise to a liquid-liquid phase separation of membrane-less organelles (Plys and Kingston, 2018), such as the nucleolus. Observations both in human cell lines and in Drosophila suggested that the heterochromatin protein 1, which is responsible for the compaction of chromatin and gene silencing, is also able

to form phase-separated droplets in which chromatin is compacted and physically constrained (Larson *et al.*, 2017; Strom *et al.*, 2017).

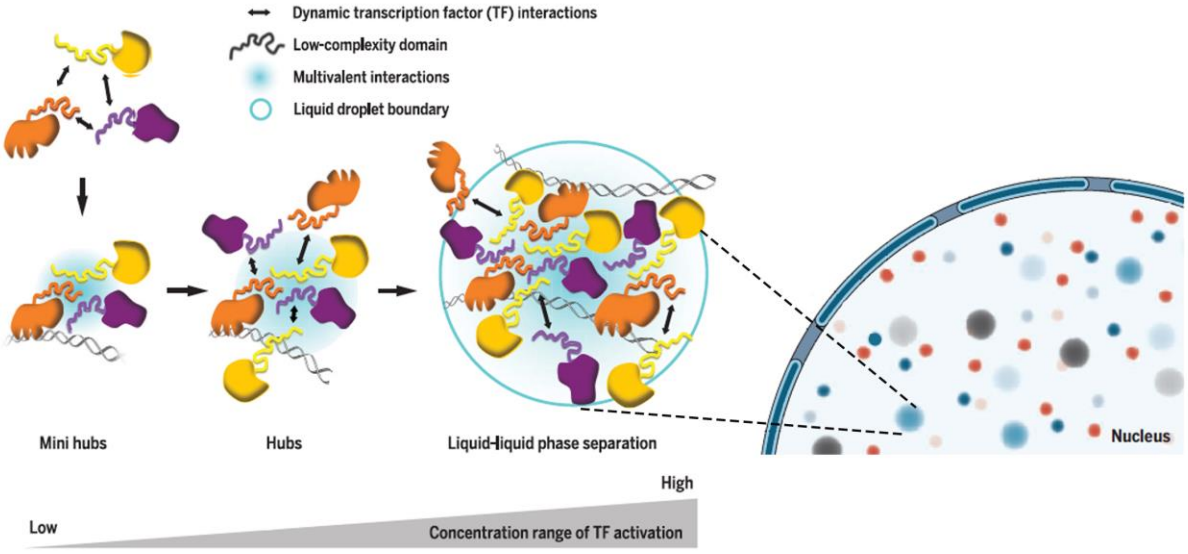


Figure 2: Representation of the nucleus as a liquid-liquid phase separation model. Transcription factors and coactivators condense into high-concentration hubs in the nucleus. Condensation is mediated by low-complexity domains in these proteins (adapted from Chong *et al.*, 2018; Plys and Kingston, 2018).

Liquid-phase separation can be due to concentrated hubs of transcription factors (Figure 2) that interact via their low complexity domains (Chong *et al.*, 2018b). The recruitment of clusters of transcription factors (coactivators) at the super enhancers loci can form phase-separated condensates, facilitating compartmentalization of transcription for specific genes essential for cell-identity maintenance (Sabari *et al.*, 2018). It has been also reported that in human and mouse genome, CpG islands rich and poor regions segregate respectively in different liquid phases (Liu *et al.*, 2018a). This observation suggested a sequence-based separation model that puts in relation the different chromatin structures with the DNA sequence and the thermodynamic factors acting inside the nucleus.

The study of the chromatin structure in a liquid-phase separation perspective could help to gain knowledge on how the chromatin domains are formed and maintained. For instance, topologically associated domains (described below) partially overlapped with the droplets domains found in (Liu *et al.*, 2018b). The formation of membrane-less compartments could explain the modalities of the diffusion or the confinement of the factors regulating a domain, without affecting nearby domains.

The use of the Hi-C analysis in this new perspective and the integration of multidisciplinary approaches from physics to chemistry to biology can address the questions on the biological machinery which links nuclear organization and regulation of gene expression, namely the processes at the base of life in cells and organisms.

1.1.1 Chromosome territories

A chromosome territory (CT) describes the physical space occupied by a chromosome inside the cell nucleus during interphase (Cremer *et al.*, 1982; Lanctôt *et al.*, 2007). In a Hi-C contact map (Figure 3), the CTs are visible as signal-dense blocks along the diagonal (Lieberman-Aiden *et al.*, 2009). CTs are, historically, one of the first structural features of the nucleus described in several microscopic studies since the late 19th century (Cremer and Cremer, 2010). In 1885, Carl Rabl, an Austrian anatomist, proposed a mode of organization for chromosomes in animal interphase nuclei (Rabl, 1885). In his model, Rabl hypothesized that centromeres and telomeres were at the opposite poles of the nucleus, a pattern that has been confirmed by later microscopic and molecular studies in yeast and plants (Cowan, Carlton and Cande, 2001; Duan *et al.*, 2010; Mascher *et al.*, 2017) and is still valid and known as the Rabl configuration.

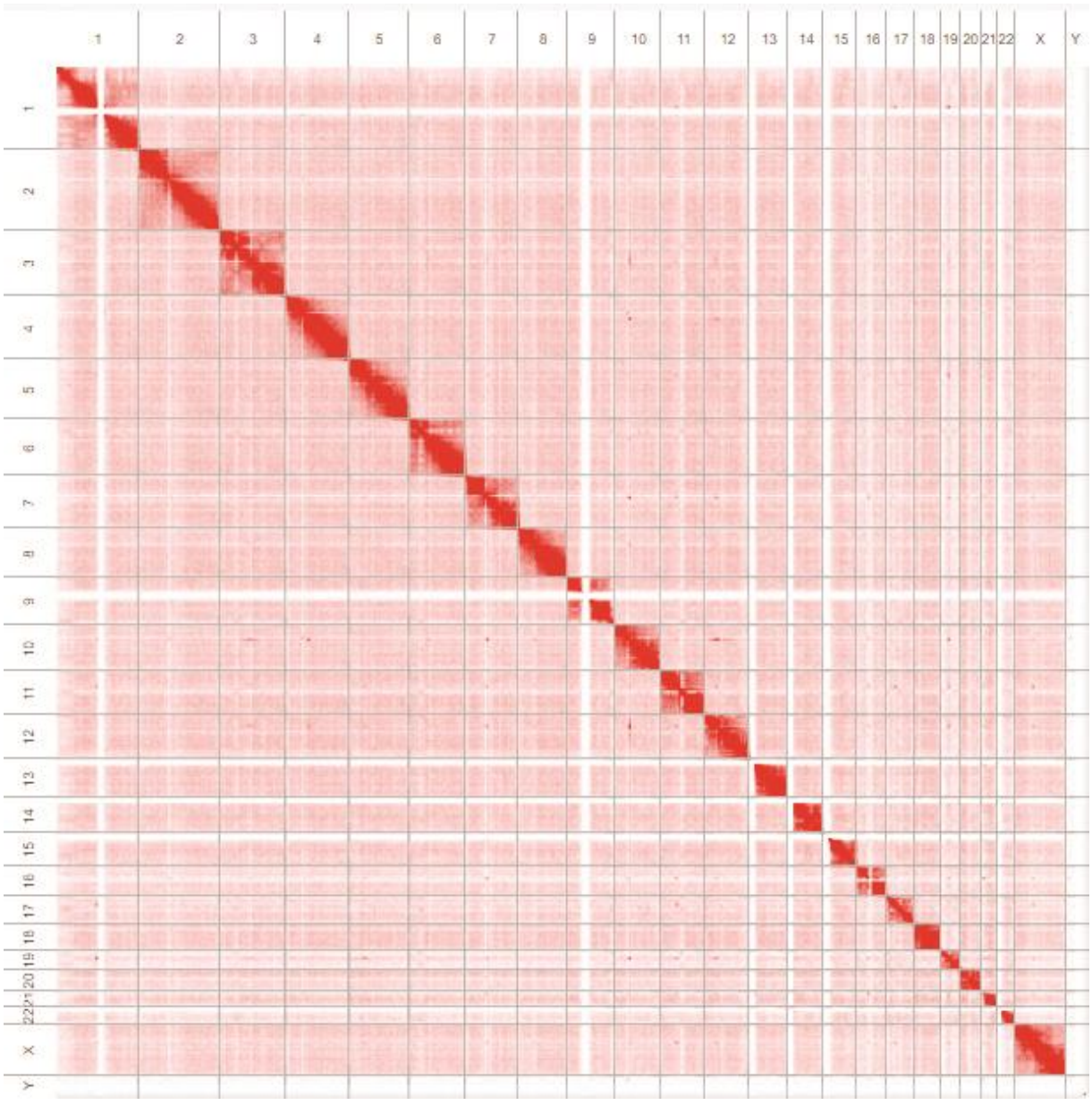


Figure 3 Genome wide Hi-C contact map of human GM06990 cells (Lieberman-Aiden *et al.*, 2009)

The term CT was already introduced by the German biologist Theodor Boveri in 1909. From his studies on the *Ascaris* (horse roundworm) life cycle, he observed that each chromosome visible

during mitosis keeps its individuality also during the interphase and occupies a certain part of the nuclear space (Boveri 1909). Since then, the CT concept has been accepted and rejected several times during the past century and currently CTs are fully readopted (for a detailed review see the "Rise, fall and resurrection of chromosome territories: a historical perspective"-part I and II by Cremer and Cremer 2006).

The presence of CTs in yeasts and some plants is still under debate, but it has been proven that other species, particularly mammals, conserve such nuclear structures (Dixon *et al.*, 2012).

Different studies have reconstructed the major features of CTs, with the main findings indicating that: chromatin of the same CT is mostly in contact with itself, making contact with other CTs only in some restricted regions (Cremer and Cremer 2010); the position of a particular CT is not the same in every cell, but some CTs have a nuclear preferential positioning, which may correlate with genomic properties and functions (Croft *et al.*, 1999; Boyle, 2001; Kosak and Groudine, 2004; Grasser *et al.*, 2008; Takizawa, Meaburn and Misteli, 2008); in human cells, large and gene-poor chromosomes tend to locate at the nuclear periphery, while small and gene-rich chromosomes are grouped at the nuclear core (Croft *et al.*, 1999); homologous chromosomes in diploid interphase cells locate in CTs far apart from each other, which has been observed in human and murine cells but it is not clear if it is only due to a physical constraint (Heride *et al.*, 2010); relative position between CTs is maintained from G1 to G2 cell cycle phases, but it is unknown whether mitosis could cause any rearrangement (Gerlich *et al.*, 2003; Walter *et al.*, 2003); finally, the spatial configuration of CTs is tissue-specific and may even be evolutionarily-conserved (Tanabe *et al.*, 2002; Parada, McQueen and Misteli, 2004). In fact, comparing seven primate species, it was found that the relative positioning of chromosome 18 and 19 was conserved despite the major

rearrangements in karyotype that occurred during higher-primate genome evolution (Tanabe *et al.*, 2002).

CT topology, although varying between cells, is a property of the nucleus emerging from the statistical distribution in a population of cells. Such spatial arrangement sets a non-random organization for chromosomes and genes inside the nucleus (Cavalli and Misteli, 2013), constituting the scaffolding for DNA regulation. Low probability of association exists between the central and peripheral regions of the nucleus (Meaburn and Misteli, 2007). This segregation produces a differentiated microenvironment between core and periphery, which could give rise to difference in regulation such as in the examples of the activating signals in the nucleolus and the repressive features in regions associated with the nuclear lamina (Parada *et al.*, 2002; Finlan *et al.*, 2008). Although genes of different functional status appear to associate with distinct nuclear features (nucleolus, lamina, domains of heterochromatin), the position of a gene alone is not a predictor of its activity. In fact, the expression of genes and general DNA regulation results from the complex interplay between the sequence and the other levels in the hierarchy of the 3D genome (Dekker and Misteli, 2015).

1.1.2 Structural Domains

Within CTs, chromosomes are partitioned into large compartments at the multi-megabase scale known as Structural Domains (SD). These domains are classified into A and B compartments (Figure 4), which in general are considered as indicators of open/closed chromatin. The A/B

compartments correlate with the genetic and epigenetic landscape in a continuous way rather than with a biphasic signal of active/inactive chromatin state (Dekker, 2013).

In particular, the A compartments contain high GC-content regions, are gene rich, and are generally highly transcribed. They are enriched in DNase I hypersensitivity sites and histone modifications marking active (H3K36me3) and poised chromatin (H3K27me3). In contrast, B compartments are gene-poor, transcriptionally less active, and enriched in high levels of the silencing H3K9me3 modification (Dekker, Marti-Renom, and Mirny 2013; Dixon et al. 2012; Jin et al. 2013; Lieberman-Aiden et al. 2009; Rao et al. 2014). The A compartments preferentially cluster with other A compartments in the nucleus, while B compartments associate with B compartments. B compartments are also highly correlated with late replication timing and LADs (Lamina-Associated Domains), suggesting that their nuclear position might be close to the nuclear periphery (Ryba *et al.*, 2010). It has been also shown, in human and mouse cell lines, that the two compartments can be further subdivided into six sub-compartments (A1, A2, and B1-B4) (Rao *et al.*, 2014)

A1 and A2 reflect actively transcribed chromatin, with high gene density, high expression levels and active chromatin marks (H3K36me3, H3K79me2, H3K27ac, and H3K4me1), with A2 more associated with H3K9me3 than A1 and having lower GC content and almost 3 times longer genes. B1 reflects the features of facultative heterochromatin such as low levels of H3K36me3 but higher levels of H3K27me3, while B2 is characteristic of pericentromeric heterochromatin, chromatin interacting with nuclear lamina and nucleolus associated domains (NADs). The subcompartment B3 is also enriched at the nuclear lamina chromatin, but is not associated with NADs. Finally, B4 is positively correlated with regions containing the KRAB-ZNF superfamily genes

(see Huntley, et al 2006 for a detailed description), but was only characterized in the human chromosome 19 and represents only 0.3% of the genome (Rao *et al.*, 2014).

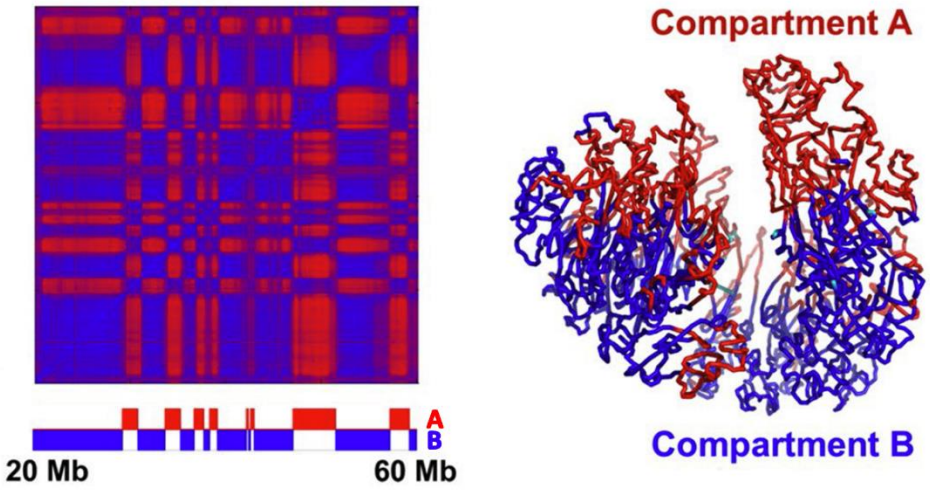


Figure 4 Chromosome partition into A/B compartments on the Hi-C contact map (left) and in a 3D model (right); adapted from (Xie *et al.*, 2017).

The compartmentalization of CTs into distinct A/B compartments and sub-compartments is directly correlated with the cell type-specific gene expression and chromatin status of the genome. For example, during the differentiation of human embryonic stem cells into mesenchymal stem cells or into fibroblasts, the chromatin is reshaped by strong repressive heterochromatin modifications (Xie *et al.*, 2013). Correspondingly, the genome is spatially reorganized, with genes no longer expressed switching from A to B compartments, and genes that need to be expressed switching from B to A accordingly (Dixon *et al.*, 2015). Finally, a recent meta-analysis work carried out on 13 human cell lines (Nurick, Shamir and Elkon, 2018) confirmed the association between A/B compartments and differential gene expression and transcription factors (TF) binding events. Moreover, the effect of A/B compartmentalization on gene regulation

established under basal conditions is still effective even when cells reshape their transcriptional program under treatment (Nurick, Shamir and Elkon, 2018).

1.1.3 Self-interacting domains

Inside CTs and SDs, the next level of the hierarchical 3D structure of the chromatin are the self-interacting domains. They have been identified in the genomes of a wide range of species, from bacteria to human (Dixon *et al.*, 2012; Hou *et al.*, 2012; Nora *et al.*, 2012; Ciabrelli and Cavalli, 2015) and appear as regions in which adjacent loci tend to interact more frequently than with other neighboring domains. Self-interacting domains size ranges from hundreds of kilobases to megabase scale, with each domain separated from another by sharp boundaries. The frequency of interaction across these boundaries suddenly drops (Figure 5), resulting in a structural insulation between adjacent domains (Dixon *et al.*, 2012; Nora *et al.*, 2012; Crane *et al.*, 2015).

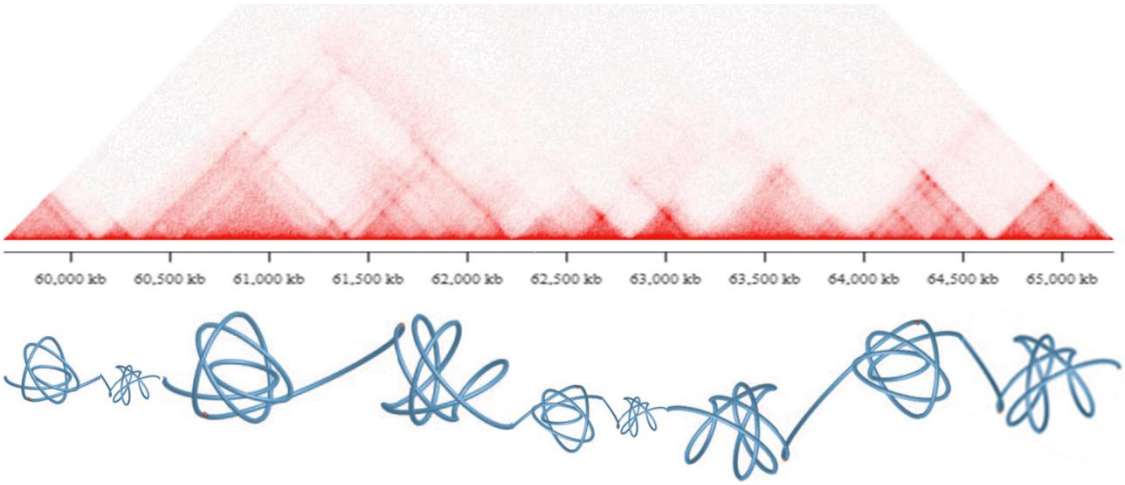


Figure 5 Interaction map of a 5Mb long region showing several TADs (above), and its corresponding chromatin model (bottom); image adapted using data from Robinson *et al.*, 2018.

These chromatin-folding modules are called "physical domains" in *Drosophila* (Sexton *et al.*, 2012) or Topologically Associating Domains (TADs) in mammalian cells (Dixon *et al.*, 2012). Notably, similar structures have been observed in bacteria and yeasts, where these domains are typically referred to as chromosomal interacting domains (CIDs) (Le *et al.*, 2013; Hsieh *et al.*, 2015). In mammals, the domain boundary regions are generally enriched in transcription start sites, active transcription, active chromatin marks, housekeeping genes, tRNA genes, and short interspersed nuclear elements (SINEs), as well as binding sites for architectural proteins like CCCTC binding factor (CTCF) and cohesin (Dixon *et al.*, 2012). In fact, in mammals the depletion of CTCF leads to a loss of TAD structures (Nora *et al.*, 2017), potentially resulting in developmental abnormalities, as seen in mouse embryonic cells (Lupiáñez *et al.*, 2015).

TAD formation results from the combined effect of several architectural proteins that can be explained via the "loop extrusion model" (Sanborn *et al.*, 2015a). Chromatin is looped by a ring of cohesin. The ring progresses on the chromatin fiber until is halted by a block of CTCF bound to

the chromatin in specific orientation. Multiple loop-extruding complexes can give rise to a TAD which borders are sealed by closely spaced CTCF (Sanborn *et al.*, 2015b; Fudenberg *et al.*, 2016). CTCF and cohesin are not the only factors involved in building TADs. Recently, several proteins, called “insulator proteins” and their binding motifs called “insulator motifs” were characterized to mimic the function of CTCF in *D. melanogaster* (Ramírez *et al.*, 2018).

The size, origin and the structure of self-interacting domains vary with the species, but are maintained as features of a wide set of genomes from fungi to mammals (Dekker and Heard, 2015).

The definition of such domains has been a great step forward for the understanding of chromatin organization in the interphase nucleus, and great efforts are being made towards the assessment of the functionality of these domains and their formation mechanisms (Ramani, Shendure and Duan, 2016).

Although TADs are considered as the building blocks of chromosomes (Dixon *et al.*, 2012) from a structural point of view, their functional characterization remains unclear. Some studies suggested that TADs constitute a functional key point in DNA regulation, since groups of genes within the same TAD showed highly correlated expression levels (Nora *et al.*, 2012); or also that TADs represent constrains for gene regulation (Zhan *et al.*, 2017) since they are defining the space of action for enhancers (Lupiáñez *et al.*, 2015; Bonev *et al.*, 2017). Other studies, showed that regulation of gene expression was not significantly affected by TADs disruption upon cohesin removal in human and mouse cell lines (Rao *et al.*, 2017; Schwarzer *et al.*, 2017).

The identification and definition of TADs in plants led to controversial results: studies performed on *A. thaliana* suggest that TADs are not an obvious features of plant genomes (Grob and

Grossniklaus, 2017); in contrast, a recent work on rice found prominent TADs differentiating the chromatin packing (Liu *et al.*, 2017). The presence of TADs in plants will be discussed more in detail in the next sections.

1.1.4 Chromatin loops

In a structural perspective, looping represents the most basic and fundamental step in chromatin folding (Fraser *et al.*, 2015). From a functional point of view, looping is the solution that enables long-range interactions (Figure 6), which can be key effectors in gene expression (Griffith, Hochschild and Ptashne, 1986).

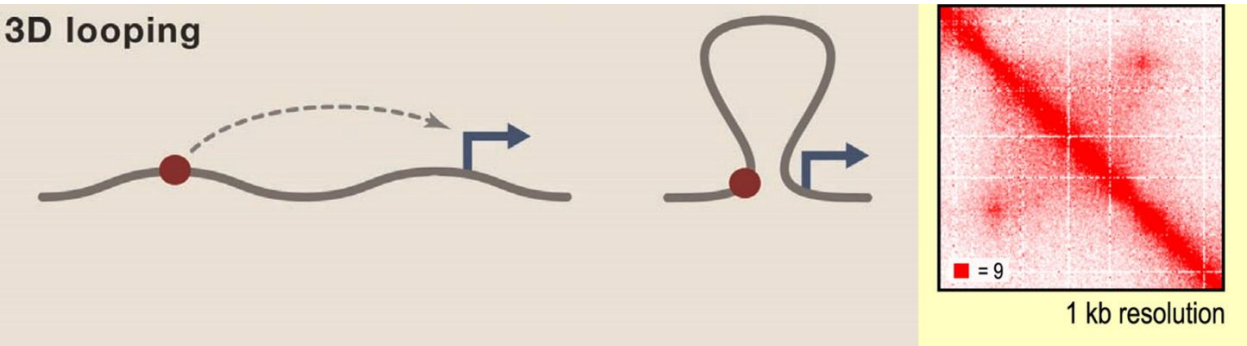


Figure 6 Looping formation between a gene and a regulatory element far away on the linear distance but closer in the 3D space and its correspondent pattern in the Hi-C contact map; adapted from(Rao *et al.*, 2014; Dekker and Mirny, 2016).

Chromatin loops have diverse functional effects on transcription and gene regulation such as: bringing distant enhancers and promoters in contact; allowing the recycling of the RNA polymerase II from its termination site back onto the promoter (Hou and Corces, 2012); enhance transcription directionality of protein-coding genes (Tan-Wong *et al.*, 2012); Polycomb-mediated

repression (Grossniklaus and Paro, 2014); insulator-mediated activation or repression of gene domains (Yang and Corces, 2012).

The best described looping interactions are between genes and their regulatory elements. In this category, a long-range physical contact is formed between control element such as enhancers and a target promoter (Dekker and Misteli, 2015). An extreme example is the *sonic hedgehog* gene that is regulated by an enhancer 1 Mb away from the gene (Lettice *et al.*, 2003). Another well studied case is the α- and β-globin genes regulation. These are clusters of genes, each of which needs to be expressed in a specific developmental stage of mammalian organisms. The element that ensures the correct regulation of such gene clusters is the Locus Control Region (LCR), which is a loop of 40-80 kb (Tolhuis *et al.*, 2002) that interacts sequentially with the appropriate gene in the appropriate developmental stage and only in cells that express the gene (Palstra *et al.*, 2003).

It has also been reported that loop formation can override the gene expression program. For example, the induction of loops between the fetal γ-globin promoter and the LCR in adult human erythroblasts forces the reactivation of the developmentally silenced fetal globin gene, with a reduction of adult β-globin expression; this mechanism could be applied to other genes with loop-dependent expression for therapeutic purposes (Deng *et al.*, 2014).

The formation of chromatin loops is not always related to gene activation. In fact, during the repression mediated by Polycomb complexes, gene silencing elements are recruited to the compacted chromatin via looping events (Grossniklaus and Paro, 2014).

Loops can connect several interactors, such as promoter-promoter, enhancer-enhancer and multiple promoters and/or multiple enhancers co-localizing from distal loci (Li *et al.*, 2012; Sanyal *et al.*, 2012; Zhang *et al.*, 2013; Ma *et al.*, 2014).

The formation of the loops is mediated by specific proteins and protein complexes made up by transcription factors, cofactors, and DNA binding enzymes. Each loop is characterized by a specific protein combination dependent on the interspecific matching with the binding factors (Dekker and Misteli, 2015). Some protein factors are common to most loops and contribute to the loop establishment, since they put in direct contact the elements they bind. Some examples of common factors are the mediator complex (Kagey *et al.*, 2010), cohesins (Hadjur *et al.*, 2009; Young, 2011) and CTCF protein (Phillips and Corces, 2009).

1.2 THE STUDY OF THE THREE-DIMENSIONAL GENOME IN PLANTS

Spatial genome organization in plants has been analysed with Hi-C, first in the model plant *A. thaliana* (Feng *et al.*, 2014; Grob, Schmid and Grossniklaus, 2014), and then in several species including barley (Mascher *et al.*, 2017), rice (Dong et al. 2017; Liu et al. 2017), foxtail millet, sorghum, tomato and maize (Dong et al. 2017). Plants 3D genomes show differences from species to species, one of the main ones being the way in which chromosomes are arranged in the nucleus (Tiang, He and Pawlowski, 2012). At a global level, three main configurations have been proposed for the chromosomes in the interphase nucleus of plants: the Rabl configuration (mentioned previously, see Figure 7 (a)); the “rosette-like” configuration(Figure 7(b)) (Armstrong,

Franklin and Jones, 2001; Fransz *et al.*, 2002); the "telomere bouquet" configuration (Figure 7 (c)) that seems more related to early meiotic stages (Harper, 2004).

From cytological studies it is known that in some plant species with long chromosomes (hundreds of Mb) such as barley, all the cells of the plant have nuclei with the Rabl configuration (Anamthawat-Jónsson *et al.*, 1990), and this has been recently confirmed by Hi-C experiments (Mascher *et al.*, 2017).

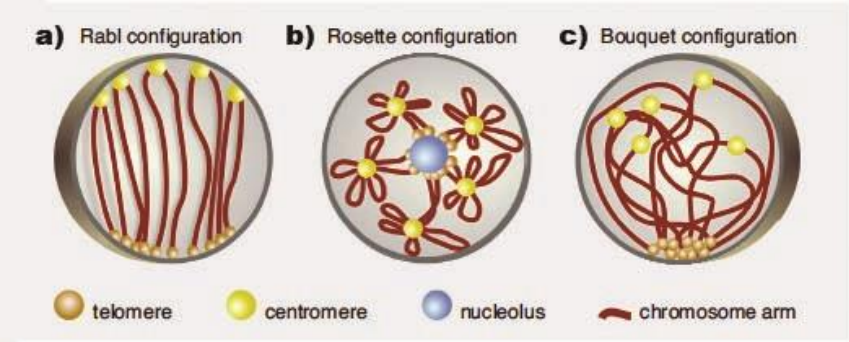


Figure 7 Graphic representation of chromosome configurations in the nucleus (adapted from Grob and Grossniklaus 2017).

In species like rice, instead, this configuration is observed only in certain tissues like xylem and roots (Prieto *et al.*, 2004); a recent high-resolution Hi-C study partially confirmed that telomeres tend to cluster in the rice nucleus, suggesting that a Rabl conformation is possible, although not in a constitutive manner (Liu *et al.*, 2017). However, this is not a fixed feature, since other plant species, such as maize and sorghum, do not show Rabl configuration, even though they have large chromosomes (Dong and Jiang, 1998). In particular, Hi-C experiments confirmed "non-Rabl" configuration for sorghum and maize, with the latter showing a pattern of chromatin interactions halfway between that observed in sorghum and barley (Dong et al. 2017). Moreover, *A.thaliana*, which has relatively short-chromosomes, is purported to have nuclei with an alternative

configuration known as the “rosette configuration” in which centromeres form distinct bodies and chromosome arms loop around such bodies, while telomeres tend to co-localize at the centre of the nucleus (Fransz *et al.*, 2002). This finding has not been yet confirmed by any 3C-based technology analysis, since none of the studies observed a signal in the interaction frequency compatible with the rosette configuration (Grob and Grossniklaus, 2017).

The presence of SDs seems to be a conserved characteristic also in plant genomes. In fact, it has been observed that the chromosome arms (excluding centromeric and pericentromeric regions) of the *Arabidopsis* genome are segmented into Loose Structural Domains (LSDs) and Compacted Structural Domains (CSDs). These domains resemble the A/B compartments in human nuclei (Grob, Schmid and Grossniklaus, 2014). LSDs are open chromatin domains, capable of interactions with distal regions of the genome, while in CSDs the chromatin is more densely packed and is not in contact with regions outside the domain. Briefly, CSDs are more correlated with inactive euchromatin features such as inactive epigenetic marks (high DNA methylation and H3K27me3 enrichment) and abundance of TEs; conversely, LSDs are characterized by active histone modifications (for example H3K4me3) and higher transcription levels (Grob, Schmid and Grossniklaus, 2014).

Also in rice, the chromatin can be partitioned into megabase-scale A/B compartments that tend to localize into distinct domains of a CT. These regions reflect the differential epigenomic landscape of active/inactive chromatin (Liu *et al.*, 2017). At the moment, no recognizable SD has been observed in barley, although the analysis of the first two components allowed the distinction between long arm, short arm and the centromere (Mascher *et al.*, 2017). The lack of prominent A/B compartments in barley may be due to a particular organization of the chromatin.

In fact, most of the chromosome length is occupied by highly repetitive and inactive chromatin, while the actively transcribed chromatin is restricted near the telomeres instead of being distributed along the chromosome arms.

The sub-megabase scale organization of plant genomes is currently under debate, since TADs are not a predominant feature in *A. thaliana* chromatin structure. One reason is attributed to the absence of CTCF or other insulator protein homologs in plants. These are known to be physically associated with the boundaries of animal TADs and act as molecular locks of chromatin interactions (Feng *et al.*, 2014; Wang *et al.*, 2015). Another hypothesis is that the plants used to build all the analysed Arabidopsis Hi-C datasets were 10 to 15 days old: at this stage the plant is growing rapidly, so most of the cells are in mitosis and this could be a complication for TADs detection (Grob and Grossniklaus, 2017). A third possible motivation is that TADs formation is not required or even not allowed in gene-dense genomes like Arabidopsis, and the lack of self-interacting domains should be a common feature of such genomes (Hsieh *et al.*, 2015; Rowley and Corces, 2016).

Recent high-resolution Hi-C studies in *A. thaliana* revealed some portions of the genome that resemble mammalian TADs for organization and dimensions. In particular the existence of “insulator-like” regions (regions with weak interactions with their flanking regions), “TAD-boundary-like” regions (regions interacting preferentially with downstream or upstream other regions, resembling TADs starting or ending point respectively) and “TAD-interior-like” regions in the middle of two adjacent “TAD-boundary-like” regions has been proposed (Wang *et al.*, 2015). By contrast, the rice genome shows prominent local packed chromatin structures described as TADs. They occupy 25% of the genome with a median size of 45kb. Similarly to mammalians, also

in rice the gene expression levels are positively correlated with TAD boundaries regions (Liu *et al.*, 2017).

In general, for plants the concept of TADs needs to be adapted in a "non-canonical" sense. In fact the Hi-C analysis on five plant species revealed a widespread presence of TAD-like domains (Dong et al. 2017). These domains can be divided in four types, since each type is associated to a different epigenetic signature: repressive domain (high DNA methylation); active domain (open chromatin); polycomb domain (high in H3K27me3 mark); intermediate domain (lack of features). Plant "non-canonical" TADs characterization retraces the domains found in *D. melanogaster* (Sexton *et al.*, 2012), while differs from mammalians description since these domains are formed in absence of CTCF and are strongly associated to the A/B compartment status. This fact is similar to what happens in *D. melanogaster* TADs formation, in which, besides the CTCF binding, also A/B chromatin status defines the domain structure (Rowley *et al.*, 2017).

The chromatin in plant genomes can form loops. One of the first examples was observed in maize, where the transcription of the two epi-alleles of the gene *b1* is regulated by the occurrence (in the active allele *B-I*) and by the absence (in the silenced allele *B'*) of looping structures (Louwers, Bader, *et al.*, 2009). Loops seem to be a featured characteristic in *A. thaliana* genome, with more than 20,000 loops identified in a recent genome-wide study (Liu *et al.*, 2016). As in mammalian genomes, Arabidopsis loops have a role in promoting gene expression (Singh and Hampsey, 2007). Nonetheless, loops are also found in correlation with low-expressed or silenced genes, raising the question of whether these genes have different silencing mechanisms from genes without loops (Liu *et al.*, 2016).

Notably, a unique structure of the *A. thaliana* genome is the *KNOT* formation (Feng *et al.*, 2014; Grob, Schmid and Grossniklaus, 2014), resulting from the constitutive contact between ten different chromosomal locations called *KNOT* engaged elements (KEEs) in the Grob, Schmid, and Grossniklaus study, or interactive heterochromatic islands (IHIs) in Feng et al. study. All five Arabidopsis chromosomes are in contact in the *KNOT*, but it is difficult to assess a common epigenetic or genetic landscape for the various KEEs or IHIs regions involved (Grob and Grossniklaus, 2017). However, these regions show significant enrichment for TEs insertions, suggesting that the *KNOT* could act as a trap for TEs (Grob, Schmid and Grossniklaus, 2014). This hypothesis takes strength from the structural analogy of the *KNOT* with the *flamenco* locus of *D. melanogaster*, consisting of several piRNA clusters (Iwasaki, Siomi and Siomi, 2015), described also as TE traps and regulators (Zanni *et al.*, 2013).

1.3 VITIS VINIFERA GENOME

Vitis vinifera is a perennial dicotyledonous species whose genome is composed of 19 chromosomes, for a total length of approximately 485 Mb. Modern grapevines are the result of a domestication path that started 6-8 thousand years ago, when the *Vitis vinifera ssp. sativa* was obtained by breeding and selection from its wild ancestor *Vitis vinifera ssp. sylvestris* (Myles et al. 2011).

Vitis vinifera was the first fruit crop to be fully sequenced, and its genome was assembled in 2007 by the French-Italian Public Consortium for Grapevine Genome Characterization (Jaillon *et al.*,

2007). The reference genome for *Vitis vinifera* was obtained by building a high-quality assembly of the PN40024 line, a nearly-homozygous genotype (with an estimated homozygosity around 93%), obtained by reiterated self-pollination of the Helfensteiner variety.

A large proportion (41.4%) of the grapevine genome is characterized by the presence of transposable elements (Jaillon *et al.*, 2007), while in the transcribed part of the genome 31,827 genes were annotated using different analysis approaches (Vitulo *et al.*, 2014).

Vitis vinifera is a highly heterozygous organism, showing a high genetic diversity (Myles *et al.*, 2011). A recent study on 128 *V. vinifera* varieties identified a total of 9,476,335 single nucleotide polimorphisms (SNPs) and 860,191 INDELs found in the population. Structural variants (SVs; described below) were detected in a subset of 50 grapevine varieties, selected from the 128 varieties population. A total of 18,090 deletions and 45,273 insertions were reported from the SVs analysis (Gabriele Magris, PhD thesis, 2016).

A large proportion of the above-mentioned structural variants are due to transposable elements (TE). TEs are an important constituent of *V. vinifera* genome and can have functional effects. For example, at a macroscopic level, TE density was found to correlate positively with cytosine methylation, both in the CG and in the CHG contexts (Mirko Celii, PhD thesis, 2016). A significant fraction of the highly transcribed genes show high gene body methylation, especially in the CG context. This methylation level is not uniform across the whole gene; in particular, intronic regions appear more methylated than exonic regions. This observation is in contrast with findings in other species such as Arabidopsis and humans, suggesting epigenetic silencing of TEs in *Vitis vinifera* introns (Mirko Celii, PhD thesis, 2016).

1.4 THE PLANT PAN-GENOME AND THE NOVABREED PROJECT

In the last decades, the analysis of variation in plants has revealed high levels of structural diversity among the individuals of a species (Morgante, De Paoli and Radovic, 2007). This observation led to the realization that, in order to obtain a complete description of the genomic variation and composition of a plant species, more than one individual must be analysed. The "pan-genome" was originally defined in bacteria as the complete collection of a species' genetic material (Tettelin *et al.*, 2005). In this seminal study, the pan-genome was defined as the combination of a "core genome" and a "dispensable genome". The former contains sequences shared by all the individuals of a same species, while the latter is made of the variable part, present only in some of the individuals.

The first plant pan-genome was described for maize, for which the comparison of four orthologous *loci* from two inbred lines of maize (B73 and Mo17), revealed that, on average, only 50% of the analysed sequence was shared (Brunner, et al 2005). The remaining 50% of sequence was instead equally divided into B73-private and Mo17-private sequence, respectively (Morgante, De Paoli and Radovic, 2007). This evidence is in contrast with the assumption that individuals belonging to the same species have the same genomic sequence content (collinearity), except for small variations such as SNPs, insertions or deletions (indels), and other small rearrangements (The Arabidopsis Genome Initiative, 2000; Goff *et al.*, 2002; Rafalski, 2002; Yu *et al.*, 2002).

In the maize pan-genome, the "core" fraction contains the majority of genes, and a minority of TEs present in all the individuals at the same genomic locations. The "dispensable genome" instead, contains different types of TEs found at different locations in the two inbred lines, plus

a “gene-like” fraction (Morgante, De Paoli and Radovic, 2007). The genes present in this fraction of the dispensable genome are altered in their structure or in their number, like the *MATE1* gene that is triplicated in aluminium-tolerant individuals (Maron *et al.*, 2013).

The core genome, being shared by all the individuals, is likely essential for vital functions of the organism; conversely, the dispensable genome has been considered for a long time to be non-essential for survival, although it might have consequences on the evolution of the species (Marroni, Pinosio and Morgante, 2014).

In general, the “dispensable genome” of a species is defined by the presence of SVs (Mills *et al.*, 2011). SVs are large (≥1kb) genomic alterations, such as insertions or deletions, translocations, inversions, or duplications (Feuk, Carson and Scherer, 2006), but recently also smaller variants with a minimum length of 50 bp are considered as SVs (Alkan, Coe and Eichler, 2011).

SVs can be categorized as either balanced and unbalanced alterations. Translocations and inversions are examples of balanced SVs, while deletions, insertions and duplications are unbalanced SVs since they alter the DNA copy number (Hurles, Dermitzakis and Tyler-Smith, 2008).

The most common type of SVs in plant dispensable genomes are copy number variants (CNVs) and presence-absence variants (PAVs). In particular, CNVs are sequences present in all the individuals of the same species but in different copy numbers, while PAVs are a particular case of CNV in which a certain sequence is present only in some individuals but totally absent in others (Marroni, Pinosio and Morgante, 2014).

Among the mechanisms capable of generating SVs, non-allelic homologous recombination (NAHR; Hastings et al. 2009), and double strand break (DSB) with single strand annealing (SSA)

are noteworthy (Muñoz-Amatriaín *et al.*, 2013), but the most common event is represented by the recent movement of TEs (Brunner, 2005; Eichten *et al.*, 2011).

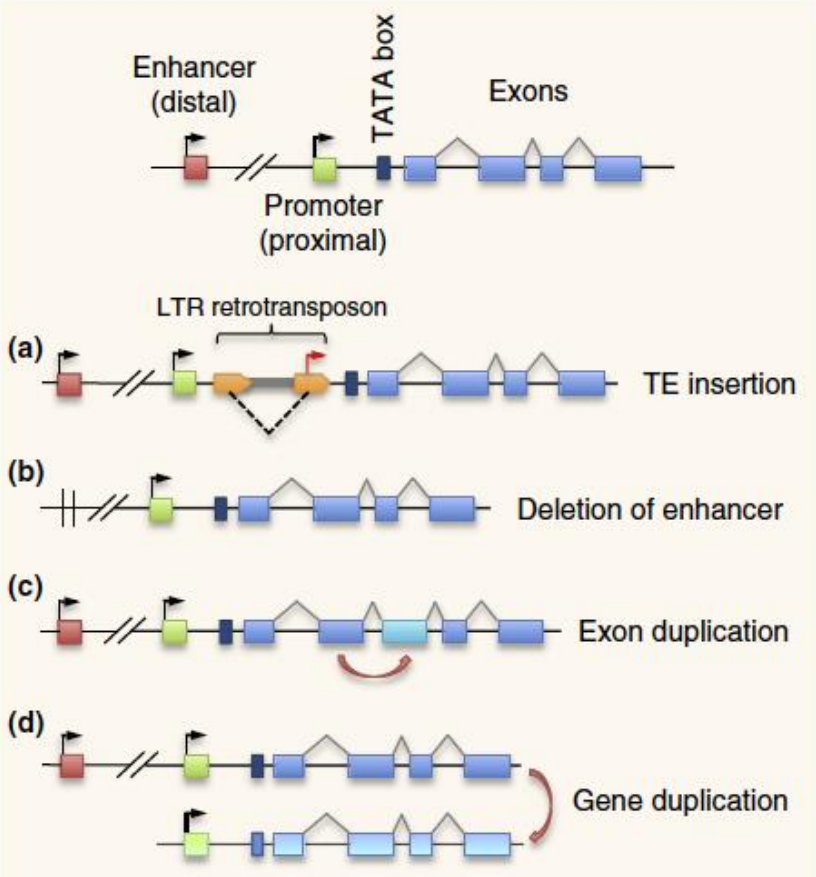


Figure 8 Representation of SVs affecting gene regulation through different mechanisms (Current Opinion in Plant Biology, 2014).

The consequences of SVs are not always obvious when they are in low gene content regions, but they have a different impact when in proximity of genes (Tenaillon, Hollister and Gaut, 2010). Structural variants can affect gene regulation by modifying or destroying regulatory elements (for example enhancers or promoters; Figure 8(a) and (b)), altering the structure of the gene (Figure 8 (c)) or changing the gene copy number (Figure 8(d)) (Marroni, Pinosio and Morgante, 2014). These changes can result in either a disadvantageous or a favourable trait for the plant.

For example, in sweet oranges the insertion of a retrotransposon at the *Myb* gene (coding for a transcription factor) leads to a high anthocyanin content that gives rise to the "blood orange" fruit in cold-stress conditions (Butelli *et al.*, 2012). In addition, the weed pest *Amaranthus palmeri* constitutes an example of how SVs can affect the evolution of a species. In this species, some individuals developed glyphosate resistance due to a CNV of the *5-enolpyruvylshikimate-3-phosphate synthase* gene, with resistant plants holding up to 100 copies of the gene more than not-resistant ones (Gaines *et al.*, 2010).

The mutations caused by TE movement constitute an element of interest in the breeding process, thus defining an important role for the pan-genome concept also in applied plant science. This is the case for grape, in which a PAV affects the anthocyanin synthesis regulator *vvMybA1* gene. Here, the insertion of a TE called *GRET1* (a retrotransposon) in the promoter of the *vvMybA1* gene blocks the transcription of the gene itself, leading to a low amount of anthocyanin in the grape berries. Grapes with white berries are homozygous for *GRET1*, while in grapes with red berries the TE insertion is present in heterozygosis or is totally absent (Kobayashi, 2004).

The present PhD work is part of the ERC-funded project NOVABREED (Novel variation in plant breeding and the plant pan-genomes), the aim of which is to study the composition of the pan-genome of two plant species, *Vitis vinifera* and *Zea mays*. In the last decades, the increased number of sequenced genomes has allowed a deeper inspection of the diversity within a species. The result is a new perspective, in which there is a high level of genetic variation among the individuals of the same species. While the human and mammalian genomes have been intensely investigated, fewer studies have been focused on plants. The NOVABREED project aimed to fill

this gap through extensive genome-wide analysis of grapevine and maize, gathering new knowledge about the genetic diversity that shapes the genomes of both species.

2 AIM OF THE THESIS

As of today, no study investigated the 3D structure of *Vitis vinifera* genome. We set out to perform the first characterization of *V. vinifera* 3D genome structure, investigating its biological function in DNA regulation.

The main objectives of the thesis were:

1) To determine the 3-dimensional chromatin organization in the grapevine genome and assess its stability, both at large (chromosome territories, A/B compartments) and small scale (chromatin loops).
2) To investigate the functional role of chromatin organization in the *V. vinifera* genome. To this aim, we integrated the structural information from the Hi-C data with the genomic and epigenomic features of the grapevine genome.
3) To assess the presence of sub-compartment domains in grapevine 3D chromatin organization, since the existence of such domains is under debate for plant genomes.
4) To investigate the effect of heterozygous structural variation on chromatin conformation. To this aim, we resolved individual haplotypes and examined allele-specific maps of chromatin interaction.
5) To investigate the signatures of SV on Hi-C maps. We thus simulated the effect of SVs on the 3D conformation of the genome

6) To investigate the relationship between SV and 3D structure. To this aim we assessed the role of chromatin 3D structure into the process of SVs formation.

In this work we present the first characterization of the *V. vinifera* 3D genome, describing the multi-scale levels of the chromatin organization and their functional implications. We also show that variation in the DNA sequence can have effect on the 3D conformation of chromatin and vice versa, specific chromatin interactions can be the prerequisite for variation occurrence.

3 METHODS

3.1 HI-C METHODS AND MATERIALS

3.1.1 I. Nuclei Preparation

Based on (Louwers, Splinter, *et al.*, 2009).

Young grapevine leaves were collected from either Azienda Agraria A. Servadei (Udine, UD, Italy) or Vivai Cooperativi Rauscedo (Rauscedo, PN, Italy). For each Hi-C experiment, approximately 2.0 grams of aerial tissue were fixed with 2% formaldehyde in 0.5x Nuclei Isolation Buffer (10mM HEPES, pH 8.0, 125 mM sucrose, 0.5 mM $MgCl_2$, 2.5 mM KCl, 20% (v/v) glycerol, 0.125% Triton X-100, 0.5% (v/v) β-mercaptoethanol. Leaves were fixed for one hour at room temperature under a vacuum. Fixation was quenched with the addition of glycine to 125 mM and an additional 5-minute incubation at room temperature.

Leaves were washed three times with ddH2O, flash-frozen in liquid nitrogen, and ground to a fine powder with a mortar and pestle. Cells were lysed with the addition of 10 mL of ice-cold 1x Nuclei Isolation Buffer (20 mM HEPES, pH 8.0, 250 mM sucrose, 1.0 mM $MgCl_2$, 5.0 mM KCl, 40% (v/v) glycerol, 0.25% Triton X-100, 1.0% (v/v) β-mercaptoethanol) plus 50 uL of protease inhibitor cocktail for plants (Sigma-Aldrich, P9599) and the liquefied sample was filtered through 3 layers of Miracloth. The nuclei suspension was centrifuged at 4°C for 15 minutes at 3000 x g. The supernatant was discarded and the pellet was resuspended in 1 mL of 1x Nuclei Isolation Buffer containing 5 uL protease inhibitor cocktail. Resuspended nuclei

were transferred to a 2 mL Eppendorf tube and centrifuged at 4°C for 5 minutes at 1900 x g. The supernatant was discarded and this wash was repeated. A final wash was performed with the same centrifuge conditions with 1 mL of 1x NEBuffer2 (New England Biolabs, B7002).

3.1.2 II. Digestion

Based on (Rao *et al.*, 2014).

Nuclei were resuspended in 100 uL of 0.5% SDS and incubated at 62°C for 10 minutes. Following this, 145 uL of ddH2O and 50 uL of 10% Triton X-100 were added and the sample was gently mixed and incubated at 37°C for 15 minutes. Next, 25 uL of 10x NEBuffer 2 and either 100 U of MboI (NEB, R0147) or 400 U of HindIII (NEB, R0104) restriction enzyme were added to digest chromatin. Samples were incubated overnight while slowly rotating at 37°C.

3.1.3 III. Biotinylation and Ligation

Digestion reactions were incubated at 62°C for 20 minutes and cooled to room temperature. To each tube was added 50 uL of the biotinylation mixture (37.5 uL of 0.4 mM biotin-14-dCTP (ThermoFisher Scientific, 19518018), 1.5 uL each of 10 mM dATP, dGTP, and dTTP (Euroclone, EMR27X025), and 8 uL of 5U/uL DNA Polymerase I, Large (Klenow) Fragment (NEB, M0210). Reactions were incubated at 37°C for approximately one hour with slow rotation. Next, to each tube was added 900 uL of ligation mix (663 uL ddH2O, 120 uL of 10x NEB T4 DNA ligase buffer (NEB, B0202), 100 uL of 10% Triton X-100, 12 uL of 10 mg/mL Bovine Serum Albumin (NEB, B9000) and 5 uL of 400 U/uL T4 DNA Ligase (NEB, M0202)). Tubes were mixed by inversion and incubated at room temperature for 4 hours with slow rotation. Following ligation, nuclei were pelleted at 1900 x g at room temperature for five

minutes and resuspended in 450 uL of 1x TE. To degrade proteins, 50 uL of 20 mg/mL Proteinase K (NEB, P8107) and 40 uL of 10% SDS were added, and samples were incubated at 65°C overnight.

3.1.4 IV. Phenol Chloroform Extraction

An additional 50 uL of 20 mg/mL Proteinase K was added to the samples, and tubes were incubated for 90 minutes at 65°C. DNA was extracted with the addition of 500 uL of a 25:24:1 phenol:chloroform:isoamyl alcohol mixture (Sigma-Aldrich, P2069), vortexing for three seconds, and centrifugation at 16,000 x g for 5 minutes. The aqueous layer was transferred to a new tube and the extraction was repeated. To the extracted aqueous layer was added 1/10 volume of 3.0 M sodium acetate, 2 uL of 20 mg/mL glycogen, and 2.5 volumes of 100% ethanol. Tubes were incubated at -80°C for one hour and then -20°C for one hour, followed by centrifugation at 16,000 x g at 4°C for 20 minutes. Pellets were washed once with 70% ethanol and dried at 65°C for two minutes, then resuspended for 30 minutes at 37°C in 45 uL of 10 mM Tris buffer and 5 uL of 1mg/mL RNaseA (Sigma-Aldrich, R6513). DNA concentrations were determined using a Qubit Fluorometer (ThermoFisher Scientific) and approximately 5 ug of DNA was used for sonication.

3.1.5 V. Sonication

To bring the sample volume to 100 uL, 10 mM Tris buffer ph 8.0 was added to the 5 ug of resuspended DNA and the sample was transferred to a 0.5 mL sonication tube (Diagenode). Samples were sonicated in a Diagenode Bioruptor for five cycles of 15 seconds on, 90 seconds off on High. Approximately 200 ng each of pre- and post-sonication DNA aliquots were loaded on a 1.4% agarose gel to confirm DNA quality and sonication efficiency.

3.1.6 VI. Biotin Pull-down

For each sample, 150 uL of 10 mg/mL Dynabeads MyOne Streptavidin C1 (ThermoFisher Scientific, 65001) were washed with 400 uL of Tween Wash Buffer (5 mM Tris buffer, pH 7.5, 0.5 mM EDTA, 1 M NaCl, 0.05% Tween 20). Beads were resuspended in 100 uL of 2x binding buffer (10 mM Tris buffer pH 7.5, 1.0 mM EDTA, 2.0 M NaCl) and the sonicated Hi-C DNA was added to the beads. Tubes were slowly rotated at room temperature for 15 minutes. Beads were separated on a magnet and the supernatant was discarded. Beads were washed 2x with 600 uL of Tween Wash Buffer and shaking at 55°C for 2 minutes at 300 rpm in a Thermomixer (Eppendorf).

3.1.7 VII. End Repair and Adapter Ligation

Beads were resuspended in 100 uL 1x NEB T4 ligase buffer and transferred to a new tube, and were then collected again on a magnet. The supernatant was discarded and the beads were resuspended in 100 uL of end-repair mix (88 uL of 1x NEB T4 DNA ligase buffer supplemented with 10 mM ATP, 4 uL of 10 mM dNTP mix, 5 uL of 10 U/uL T4 Polynucleotide Kinase (NEB M0201), 4 uL of 3U/uL T4 DNA polymerase I (NEB, M0203), and 1 uL of 5U/uL DNA Polymerase I, Large (Klenow) fragment (NEB, M0210). Reactions were incubated at room temperature for 30 minutes, then placed on a magnet and the supernatant was removed. Beads were washed 2x with 600 uL of 1x Tween Wash Buffer for 2 minutes at 55°C at 300 rpm. Beads were resuspended in 100 uL of 1x NEBuffer 2 and transferred to a new tube and were placed on a magnet and the supernatant was discarded. Beads were then resuspended in 100 uL of A-tailing mixture (90 uL of 1x NEBuffer 2, 5 uL of 10 mM dATP, and 5 uL of 5 U/uL Klenow exo- enzyme (NEB, M0212). Tubes were incubated at 37°C for 30 minutes, then placed on a magnet and the solution was discarded. Beads were washed 2x

with 600 uL of 1x Tween Wash Buffer for 2 minutes at 55°C at 300 rpm. Beads were then resuspended in 100 uL of 1x T4 DNA ligase buffer (NEB, B0202), then placed on a magnet and the supernatant was discarded. Beads were resuspended in the adapter annealing mix (39 uL 1x T4 DNA ligase buffer supplemented with 5.0% PEG 8000 (Sigma-Aldrich, P5413), 1 uL H2O, 1 uL 10x T4 DNA ligase buffer, 1 uL of 50% PEG 8000, 5 uL T4 DNA ligase, 5 uL of an Illumina Truseq adapter) and incubated for 2 hours at room temperature. Beads were then washed 2x with 600 uL of 1x Tween Wash Buffer for 2 minutes at 55°C at 300 rpm, and 1x with 200 uL NEBuffer 2. Finally, beads were resuspended in 50 uL of NEBuffer 2.

3.1.8 VIII. PCR amplification of library and sequencing

Bead-bound Hi-C DNA was amplified using Q5 polymerase (NEB, M0491) and Illumina TruSeq primer cocktail under the following conditions: **1x** 98°C 1 min; **12x** 98°C 10 sec, 65°C 30 sec, 72°C 30 sec; **1x** 72°C 3 min. Reactions were pooled and separated from the C1 beads, then purified by adding 0.7x volume of AmpureXP beads (Beckman Coulter, A63880) and incubating for 10 minutes at room temperature. Beads were placed on a magnet and washed twice with 70% ethanol, then dried 10 minutes. Beads were resuspended in 20 uL of 10 mM Tris, and the supernatant was removed to a new tube. Size and molarity of fragments was determined via Bioanalyzer (Agilent) or Caliper (Perkin-Elmer), and samples were sequenced for paired-end, 125 bp reads on an Illumina HiSeq 2500 sequencer by IGA Technology Services (Udine, Italy).

3.2 SAMPLES CHARACTERISTICS

Plant material from three grapevine varieties (Pinot noir, Rkatsiteli and Chardonnay) was processed by *in situ* Hi-c. The Rkatsiteli dataset was composed by two biological replicates for the leaf tissue, plus a library of sequences extracted from the shoot apical meristem (SAM). The Chardonnay dataset was composed by a library from a parental individual ("Chardonnay parent") and a library from its self-crossed progeny ("Chardonnay selfed"), both from leaf tissue. A total of 538,104,956 reads were sequenced for Pinot noir; 339,810,300 and 238,304,950 reads for the two Rkatsiteli replicas respectively; 249,414,834 reads for Rkatsiteli SAM; 64,822,168 and 54,482,308 reads for Chardonnay parent and selfed, respectively (*Table 1*).

Table 1 Summary of sequenced and uniquely aligned reads amount for each library. The total number of contacts is reported for both the Hi-C data analysis pipelines used and the amount of PCR duplicates for each dataset.

			Aligned reads		Total contacts		
Dataset ID	**Tissue**	**Sequenced reads**	**HOMER**	**HiC-Pro**	**HOMER**	**HiC-Pro**	**PCR duplicates**
Pinot noir	Leaf	538.104.956	335,817,425	186,838,030	67,973,874	74,832,478	1%
Rkatsiteli-1	Leaf	339.810.300	220,231,875	138,846,094	50,147,564	64,238,238	5%
Rkatsiteli-2	Leaf	238.304.950	147,792,538	93,507,162	24,735,423	43,178,415	2%
Rkatsiteli-3	SAM	249.414.834	110,059,545	69,277,006	1,054,479	2,265,174	92%
Chardonnay parent	Leaf	64.822.168	41,934,548	29,398,212	10,435,441	14,153,928	4%
Chardonnay selfed	Leaf	54.482.308	35,647,516	25,165,888	8,875,518	11,949,901	3%

3.3 HI-C DATA ANALYSIS

Adapters were removed from the reads using *cutadapt* version 1.5 (Martin, 2011) and low quality bases were trimmed and contaminant sequenced were filtered out by *Erne* version 1.4 (Del Fabbro *et al.*, 2013). The clean and trimmed reads were processed with two different pipelines for Hi-C data analysis: *HOMER* version 4.9 (Heinz *et al.*, 2010) and *HiC-Pro* version 2.9.0 (Servant *et al.*, 2015), as detailed below.

3.3.1 Homer

Since some Hi-C ligation products can give rise to chimeric reads (Figure 9), in order to optimize the mapping step, these chimeric sequences must be identified and removed before aligning.

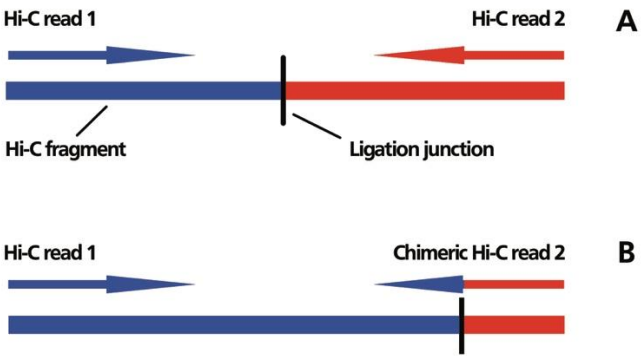


Figure 9 Schematic representation of Hi-C ligation products and relative sequenced reads in case of junction centred on the fragment (A), or junction towards one end of the fragment (B).

Chimeric reads contain the sequence of the duplicated restriction site (i.e. CAGTCAGT for MboI), thus they could be detected searching for this unique feature. We used the *homerTools trim* utility to find the chimeric reads and trim each read from the duplicated restriction site to the 3' end, keeping only trimming products longer than 20 bp.

For each library, the trimmed read1 and read2 were independently aligned to the *Vitis vinifera* reference (PN40024, version 3, http://services.appliedgenomics.org/pub/grape-assembly/vitis_12xV3.fasta) using *bwa-mem* version 0.7.10 (Li and Durbin, 2009) with default parameters and reads mapping with low quality (MAPQ<10) were filtered out using *samtools* version 0.1.19 (Li *et al.*, 2009). The aligned reads were then fed to HOMER software for the creation of the *unfiltered Tag directory*, specifying that reads were generated from the Illumina platform, with the option *–illuminaPE*, and *–tbp* (maximum tags per base pair) set to 1. A tag directory is a structure used by the Homer software to store the aligned reads. For each variety dataset, single library and pooled libraries *Tag directories* were produced. Each resulting *Tag directory* was then filtered removing the proper paired end reads (both reads on the same chromosome within 1.5x of the estimate fragment length); removing the self-ligation products (both reads near the same restriction site) and the reads starting on a restriction fragment; finally, reads from regions with 5 times more reads than the average (spikes) were removed. Whole genome contact maps were generated at several resolutions (from 1 Mb to 50 Kb) with raw interaction counts (option *–raw*) and with two normalization strategies: normalizing the counts accounting for coverage (*-simpleNorm*) or jointly accounting for coverage and distance (*-norm*). Single chromosome contact maps were generated at higher resolution (from 25 Kb to 5 Kb). The contact maps built by HOMER were

visualized using the graphic renderer *Java Treeview* version 1.1.6 (Saldanha, 2004).

3.3.2 HiC-Pro

Since HiC-Pro is designed to perform an iterative mapping step, pre-mapping trimming of chimeric reads was not required. Instead, non-mapping chimeric reads are trimmed and then realigned. Also, the mapping step is part of the HiC-Pro pipeline, so the software requires as input only the read files and the reference, both must be declared into a configuration file. The HiC-Pro process is divided in several steps: the alignment performed by *bowtie2* version 2.0.2 (Langmead *et al.*, 2009); the Hi-C filtering in which not aligned reads and improper read-pairs are removed; a quality control step in which statistics about the valid ligation products and the PCR duplicates are computed; the contact map construction and finally the ICE map normalization (Imakaev *et al.*, 2012). For each library, raw and ICE-normalized maps were generated at several resolutions (1 Mb, 500 Kb, 150 Kb, 40 Kb and 20 Kb). After converting the generated contact data into the proper format, the Juicebox software version 1.8.8 (Durand, Robinson, *et al.*, 2016) was used to visualize the maps.

The results of the two processes are summarized in (*Figure 10*).

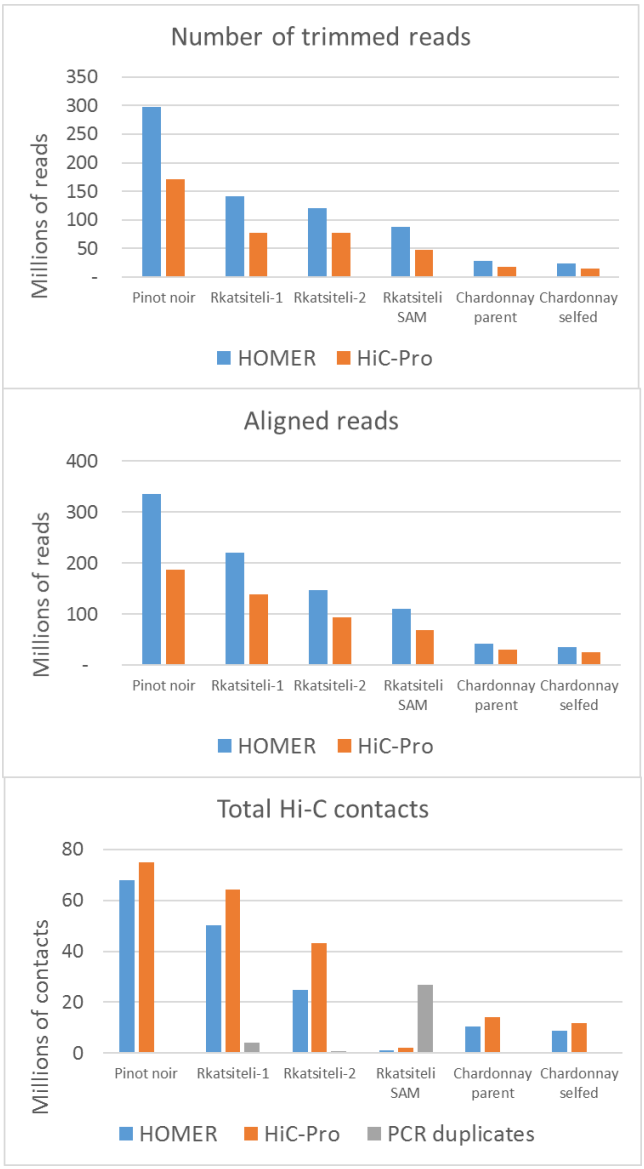


Figure 10 Results of HOMER and HiC-Pro pipelines

The Rkatsiteli SAM library was the one with the lowest yield of contacts in both pipelines used. This was due to a low complexity issue in the library, which caused a high level of duplicated reads (>90%). Duplicated reads are checked by the PCR duplicate filter, which looks for regions with 5 times more coverage than the genome average.

3.3.3 Chromosome neighbourhood analysis

The chromosome neighbourhood heat maps of the grapevine genome were built calculating the relative frequency of interaction between whole chromosomes as described in (Zhang *et al.*, 2012). For each couple of chromosomes, the $\log_2$ ratio of the observed to the expected value was calculated as follows:

$$\log_2\left(\frac{Int_{1,2}}{\left(\left(\frac{Trans_1}{Trans_{TOT}}\right)\left(\frac{Trans_2}{Trans_{TOT}-Trans_1}\right)+\left(\frac{Trans_2}{Trans_{TOT}}\right)\left(\frac{Trans_1}{Trans_{TOT}-Trans_2}\right)\right)*\left(\frac{Trans_{TOT}}{2}\right)}\right)$$

Where "$Int_{1,2}$" is the number of interactions shared between the two chromosomes; "$Trans_1$" and "$Trans_2$" are the number of interactions between one chromosome and the rest of the genome; "$Trans_{TOT}$" is the total number of interchromosomal interactions of the dataset. The heatmaps were generated using the R function *heatmap.2* and data were clustered with *hclust* using the "complete" method.

3.3.4 Distance Dependent Decay function

Interaction frequency is inversely proportional to the distance of the two interacting loci, and the Distance Dependent Decay (DDD) function describes the rate of interaction frequency decay for each Hi-C experiment (Fudenberg and Mirny, 2012). As a first level of control on the reliability of the maps obtained, the DDD function was computed for each dataset using *HiCdat* version 0.99 (Schmid, Grob and Grossniklaus, 2015). Interaction Decay Exponents (IDEs) distributions for each variety dataset was compared and tested via Wilcoxon test for significant differences. In order to identify trends of variation in chromosome conformation across varieties, Pearson's correlation coefficient was computed between Pinot noir, Rkatsiteli and Chardonnay single chromosome IDEs.

3.4 A/B COMPARTMENTS ANALYSIS

3.4.1 Identification of A/B compartments via PCA

A and B compartments were classified according to the sign of the first component (PC1) values, where positive values identified A compartments and negative values identified B compartments. PC1 values resulted from a PCA performed on each grapevine Hi-C dataset on whole genome maps at 50Kb resolution using the HOMER utility *runHiCpca.pl*. Since the PC1 eigenvectors sign may be inconsistent across chromosomes, a manual correction of the signs was carried by direct inspection of the contact map.

In order to assess the global stability of A/B compartments across varieties and organs, the *getHiCcorrDiff.pl* tool included into HOMER was used for a direct comparison of interaction patterns.

The frequency of interaction between the different compartments was obtained by intersecting the compartments coordinates with the interaction frequencies from the map, allowing the distinction between AA, BB and AB interaction contexts. For each context, the proportion of interaction frequency was computed in order to assess which of the three interaction schemes was the most frequent in the grapevine genome.

3.4.2 Functional properties of A/B compartments

The A/B compartment status was correlated with the following genomic data: density of genes, density of TE, DNA methylation level, gene expression levels and histone modification. For *Vitis vinifera* DNA methylation data, the 5mC density for the CG, CHG and CHH contexts was obtained from bisulfite sequencing experiments (Mirko Celii, PhD thesis, 2016). For the genes and TE data, the coordinates obtained from the version 2.1 of the annotated grapevine genome (Vitulo *et al.*, 2014) were used to compute density in the A/B compartments. The expression analysis was carried out using the fragments per kilobase million (FPKM) for each annotated gene as a measure of the expression level. FPKM values were obtained from previous RNA-seq experiments; distribution of not expressed genes was obtained considering the bases covered by genes with FPKM=0.

Finally, histone modification marks and chromatin accessibility data, were obtained respectively from chip-seq and ATAC-seq experiments; in both cases the frequency of peaks found in the different compartments was considered in the analysis.

Each of the above-mentioned datasets was binned into 50 Kb windows and intersected with the set of A/B coordinates. The difference of distribution between A and B compartments was tested using Wilcoxon and chi-squared statistical tests, choosing the proper one according to the type of distribution of the data. In this step of analysis, the data were processed using bedtools version 2.26 (Quinlan and Hall, 2010) for binning and integration between genomic and PCA data; and R version 3.3.3 (R Core Team, 2017) for statistics and plotting.

Differences in the expression pattern between A and B compartments were assessed

comparing the coefficient of variation of expression data obtained from berries of ten grapevine varieties at four different developmental stages (Magris *et al*., paper submitted). Wilcoxon test was used to reveal significant differences between the distributions of variation coefficient in the two compartments.

The genes in the A/B compartments were characterised according to their gene ontology (GO) terms in the "biological process", "cell component" and "molecular function" categories. GO slim annotation were retrieved using biomaRt (BioMart Project, RRID:SCR_002987) (Ensembl Plants Genes 39 version) (Durinck *et al.*, 2005, 2009). The GO term analysis was performed with an in-house script based on the *topGO* R library (Alexa and Rahnenfuhrer, 2016).

3.5 SUB-COMPARTMENT DOMAINS ANALYSIS

3.5.1 Identification and characterization of sub-compartment domains

The annotation of domains inside the A/B compartments was performed using the *Arrowhead* algorithm, available as a tool of the Juicer pipeline (Durand, Shamim, *et al.*, 2016). *Arrowhead* was used to identify sub-compartments domains at 25 Kb, 10 Kb and 5 Kb resolution. The algorithm returns in output a list of coordinate intervals together with a "corner score" which indicates the likelihood for each predicted interval to be a domain. Higher values of the corner score represent more significant results.

Since *Arrowhead* requires a high amount of sequence data, the identification of sub-

compartments was performed merging all the available sequencing data.

The relative distribution of the sub-compartment domains inside the A/B compartments was assessed by computing for each domain, the proportion of length falling in each compartment.

3.5.2 Domain borders analysis

In order to describe the chromatin state inside the above-mentioned domains, a border analysis was performed. Similar to what was reported for other plant species (Dong *et al.*, 2017), in this analysis the chromatin context outside the annotated domains is compared with the one inside the domains.

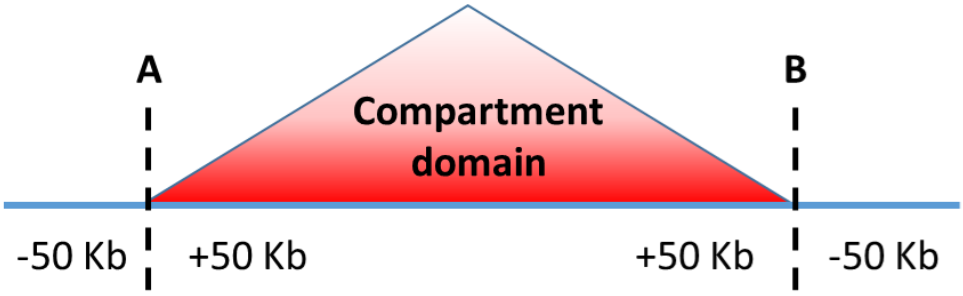


Figure 11 Scheme of the regions around and inside the sub-compartment domain used to perform the domain border analysis.

Taking in account the length distribution of the domains, four regions around each domain were chosen (Figure 11): 50 Kb upstream the A border (outside-left); 50 Kb downstream the A border (inside-left); 50 Kb upstream the B border (inside-right); 50 Kb downstream the B border (outside-right). Each of the 50 Kb regions was binned in 5 Kb windows using *bedtools makewindows*. For domains shorter than 50 Kb, we considered the entire length of the

domain as the inner region to analyse. Gene density, TE density, ATAC peaks, H3K4me3 peaks, gene body methylation, FPKM, CG, CHG and CHH methylation data were intersected with the bins created. For gene and TE density, ATAC and the three methylation contexts the average value for each bin was plotted. For gene body methylation, FPKM and H3K4me3, the median for each bin was plotted. To obtain the outer windows of the plot (white half of *Figure 24*) and the inner windows (red half of *Figure 24*) all the "-50 Kb" and the "+50 Kb" bins were merged respectively.

In order to assess the level of co-regulation between genes inside the sub-compartment domains, Pearson's and Spearman's correlation coefficients were computed from expression data obtained from berries of ten grapevine varieties at four different developmental stages.

The distribution of the correlation values of the domain's genes was compared 1,000 times with the distribution of correlation values of randomly selected genes residing outside the domain. A false discovery rate correction for multiple testing (Y.Benjamini and Y.Hochberg, 1995) was applied to the p-value calculated.

3.6 CHROMATIN LOOPS

Chromatin loops detection was performed using the *–interaction* option of the HOMER software, which searches for pairs of loci sharing greater number of Hi-C contacts than any other two loci at the same distance chosen by chance. The interactions between the

identified loci were referred to as “significant interactions” (Heinz *et al.*, 2010). Significant interactions for grapevine varieties with high number of reads (Pinot noir and Rkatsiteli) were searched in the Hi-C dataset at a resolution of 1Kb with a sliding window of 5Kb, limiting the search space to a maximum distance of 100 Kb, since we expect loops to occur at lower distances. The resulting interactions were filtered by minimum interaction distance of 1 Kb in order to remove all the interactions occurring beyond the resolution power of the analysis. Only interactions with FDR <0.05 were retained.

3.6.1 Effect of SVs on loops detection

In order to test the effect of the presence of SVs on the loop detection, the frequency of SVs and loops were compared across the genome for Pinot noir and Rkatsiteli. The genome was divided into windows that, according to the genotype, were classified as follows: heterozygous windows in which both alleles were different from the reference (sharing 0); homozygous windows in which both alleles were different from the reference (h_sharing 0); heterozygous windows in which only one allele was identical to the reference (sharing 1); and homozygous regions in which both alleles were identical to the reference (sharing 2).

Each significant interaction dataset was intersected with the sharing regions coordinates using *bedtools intersect* and the frequency of loops per 1Mb was computed for each variety. The SVs datasets for the two varieties were processed in the same way. In order to remove any effect of heterozygosis on the analysis, we compared only regions in which varieties were homozygous to the reference, namely h_sharing 0 and sharing 2.

Significant difference between loops and SVs frequencies was assessed using chi-square

test.

3.6.2 Characterization of loop interactions

Loop datasets for Pinot noir and Rkatsiteli were intersected with 2.1 version of *V. vinifera* annotated genes. We distinguished between a) loops coupling a gene and a non-gene region ("gene-other" loops); b) loops between two gene regions ("gene-gene" loops); and c) loops between two non-gene windows ("other-other" loops). The "gene-gene" loops were divided in loops occurring in the same gene and loops occurring between different genes, according to the gene ID in each of the two interacting windows. We tested if the number of "gene-gene" loops, "gene-other" loops and" other-other" loops was significantly different from what expected by chance via a resampling test. To do so, we divided the whole genome in 5 Kb windows. Each window was classified as "G" or "N" depending on the presence or absence of genes. The loops were simulated randomly sampling 6,355 and 4,910 pairs of windows for Pinot noir and Rkatsiteli respectively. This sampling was iterated 100 times and each time the number of G-G ("gene-gene"), G-N ("gene-other") and N-N ("other-other") windows pairs was computed.

Analysis of the GO term for the "same-gene" loops was performed as described in 3.4.2.

To obtain the proportion of loops involving a gene and a putative enhancer, we intersected the "gene-other" loop dataset for Pinot noir and Rkatsiteli with the set of intergenic ATAC peaks obtained in our research group. The dataset of "gene-intergenic ATAC" loops was tested against the simulated set of loops in order to find significant differences from what

expected by chance. For each of the simulated loops set described above, we further classified the “N” windows depending on the presence of intergenic ATAC peaks “A”. We then distinguished a subset of G-A (“gene-intergenic ATAC”) loops from the simulated “gene-other” loops.

In both Pinot noir and Rkatsiteli we classify genes into three categories: “intergenic ATAC” interacting genes (genes involved in “gene-intergenic ATAC” loops), “self-loop” genes (genes involved in “same-gene” loops) and “control” genes (the remainder of the genes). We compared the expression level in each category expressed as FPKM and we identified significant differences using a Wilcoxon rank sum test between all pairs of categories.

3.7 SVs AND GRAPEVINE CHROMATIN CONFORMATION

3.7.1 Directionality Index

Directionality Index (DI) is a method proposed for the identification of topological domains in mammalian genomes (Dixon *et al.*, 2012). The method allows the identification of biases in the direction of interaction frequency in the genome, which means to identify regions in which interactions are highly biased in occurring with downstream or upstream portions of the genome. The direction bias at any given genomic bin is determined by the DI, which is calculated as:

$$DI = \left(\frac{B - A}{|B - A|}\right)\left(\frac{(A - E)^2}{E} + \frac{(B - E)^2}{E}\right)$$

where: A is the number of Hi-C interactions between a given bin and the upstream 2Mb region; B is the number of Hi-C interactions between a given bin and the downstream 2Mb regions; E is the number of expected interactions under the hypothesis to observe equal number of interactions of a given bin in both downstream and upstream regions $(A+B)/2$.

The magnitude of the DI value is proportional to the degree of bias for that given bin; with positive values for downstream bias and negative values for upstream bias.

DI is commonly used for the identification of topological domains. In this study, DI was used to investigate interaction pattern variation between any two Hi-C datasets aligned to the same reference genome and to characterize how chromatin contact patterns give rise to biased interactions in presence of SVs.

3.7.2 Simulation of large deletions, insertions and inversions

SVs presence in Hi-C data was simulated by editing the PN40024 reference and aligning the Pinot noir Hi-C reads on such simulated reference. Deletions in the Pinot noir sample were simulated by adding segments from the *hg19* human genome reference to the PN40024 *V. vinifera* reference. Insertions in the Pinot noir sample were simulated by removing segments of the PN40024 reference. In order to avoid any real SVs presence effect on the simulation, both deletions and insertions were simulated in regions where Pinot noir presented both copies of the genome identical to the PN40024 sequence.

Inversions were simulated by direct editing of the PN40024 reference, inverting the sequence in the selected region.

For deletions, insertions and inversions, one event per chromosome was simulated; each event was characterized by a unique length, ranging from 5 Kb to 2 Mb.

Pinot noir Hi-C reads were aligned to the simulated references, and Hi-C contact maps were obtained using the HiC-Pro pipeline version 2.9.0 (Servant *et al.*, 2015).

Hi-C graphical output for the simulated data was obtained using the HiCPlotter software version 0.7.3 (Akdemir and Chin, 2015), adding to each map the histogram track of the DI.

3.7.3 Allele-specific Hi-C maps

Allele-specific Hi-C maps were obtained for Pinot noir and Rkastiteli, for which haplotypes have been determined by our research group. The resolved Rkatsiteli haplotypes were used as input in the HiC-Pro pipeline (Servant *et al.*, 2015) which allows to build allele-specific Hi-C maps from a set of phased SNPs and a masked reference at the SNPs locations. Each Hi-C read was assigned to one allele according to the SNP carried by the read itself. This approach allowed a good reconstruction of the chromatin conformation of the two alleles for each variety, except for regions with very low number of SNPs.

Allele-specific chromatin interaction patterns were compared using HiCPlotter (Akdemir and Chin, 2015) searching for SVs events. In order to obtain quantitative measurement of differences in the interaction patterns, DI was computed for each map and compared across alleles.

3.7.4 Analysis of chromatin contacts across CNV borders

Eighty-one CNV (corresponding to homozygous deletions either in Pinot noir or Rkatsiteli varieties compared to the reference) were selected from a set of predictions made by depth of coverage analysis (Gabriele Magris, PhD thesis, 2016). Each CNV region was then manually curated, in order to verify the exact location of the CNV borders and the exact prediction of the number of copies in the CNV region. This was done by visually inspecting the sequencing reads of Rkatsiteli and Pinot noir aligned to the PN40024 reference using Tablet (Milne *et al.*, 2013).

The manually annotated set was composed of 73 regions (size range: 4,000-600,000 bp) where one variety was homozygous for the deletion (CNV-present) and the other homozygous for the reference allele (CNV-absent).

A set of control regions (CTR) was built by using 100 randomly chosen regions with the same size distribution as the CNV set. CNV regions in both varieties are excluded from CTR. The random sampling was performed using *bedtools shuffle*, specifying the regions to exclude with the set of CNV coordinates via the *–excl* option.

For each CNV-present, CNV-absent and CTR region, a 5 Kb window was drawn around each border. We called such windows as “Flanking Region” 1 and 2 (FR1, FR2), respectively, indicating the window at the 5’ and at the 3’ end of the region (Figure 12). The number of Hi-C interactions between FR1, FR2 and the rest of the genome were computed using the *make_viewpoints.py* utility (Servant *et al.*, 2015). Since we were interested only in interactions occurring across the CNV borders, we restricted the

distance range for the analysis keeping only the interactions occurring in the range of 15 Kb upstream and downstream each border (Figure 12).

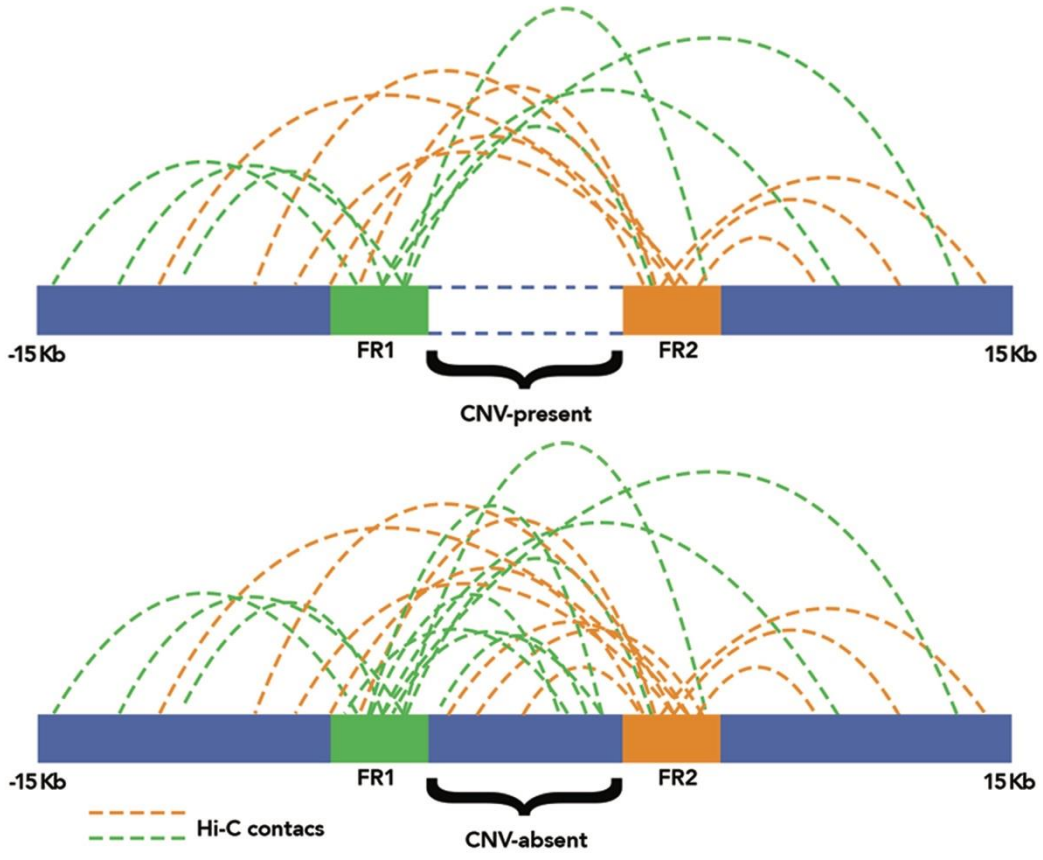


Figure 12 Scheme of the interaction analysis. For each region of interest, the interaction count across FR1-FR2 border and in a range of 15 Kb was evaluated.

In order to compare the interaction counts distributions of CNV-present, CNV-absent and CTR, the two 15 Kb windows were divided into 1 Kb bins using *bedtools makewindows* and the distribution of interactions between the two Flanking Regions (FRs) and each bin was determined. The interaction count in each of the 1Kb bins was normalized by the coverage in that bin using *samtools bedcov*.

In order to avoid the effect of mapping artifacts due to sequence homology between the borders of the analysed regions, a k-mer analysis of the CNV and CTR datasets was performed using the "Tallymer" software (Kurtz *et al.*, 2008). We chose a subset of regions from CNV and CTR showing no significant difference in 10-mers homology (chi-square test), but still preserving statistical power for the analysis.

The CNV-present, CNV-absent and CTR interaction distributions were compared imposing to 0 the distance between the borders of the analysed regions and significant differences were identified performing Wilcoxon rank sum test between all pairs of distributions.

4 RESULTS AND DISCUSSION

4.1 VITIS VINIFERA 3D GENOME

4.1.1 *Vitis vinifera* chromatin organization in the interphasic nucleus.

The nuclear architecture of eukaryotes is the result of a hierarchy of structures in which chromatin is organised. In order to define the chromatin organization of grapevine genome, we used Hi-C reads obtained from tissue-specific libraries of young leaves of three *V. vinifera* varieties: Pinot noir, Rkatsiteli and Chardonnay.

4.1.1.1 Chromosome territories

At the top level of the nuclear structure hierarchy are the chromosome territories (CTs), which define discrete areas in the interphasic nucleus. We sought to identify CTs in grapevine genome and to investigate how different chromosomes relate to each other. We reconstructed a genome-wide Hi-C map for each of the aforementioned grapevine varieties.

Each contact map in

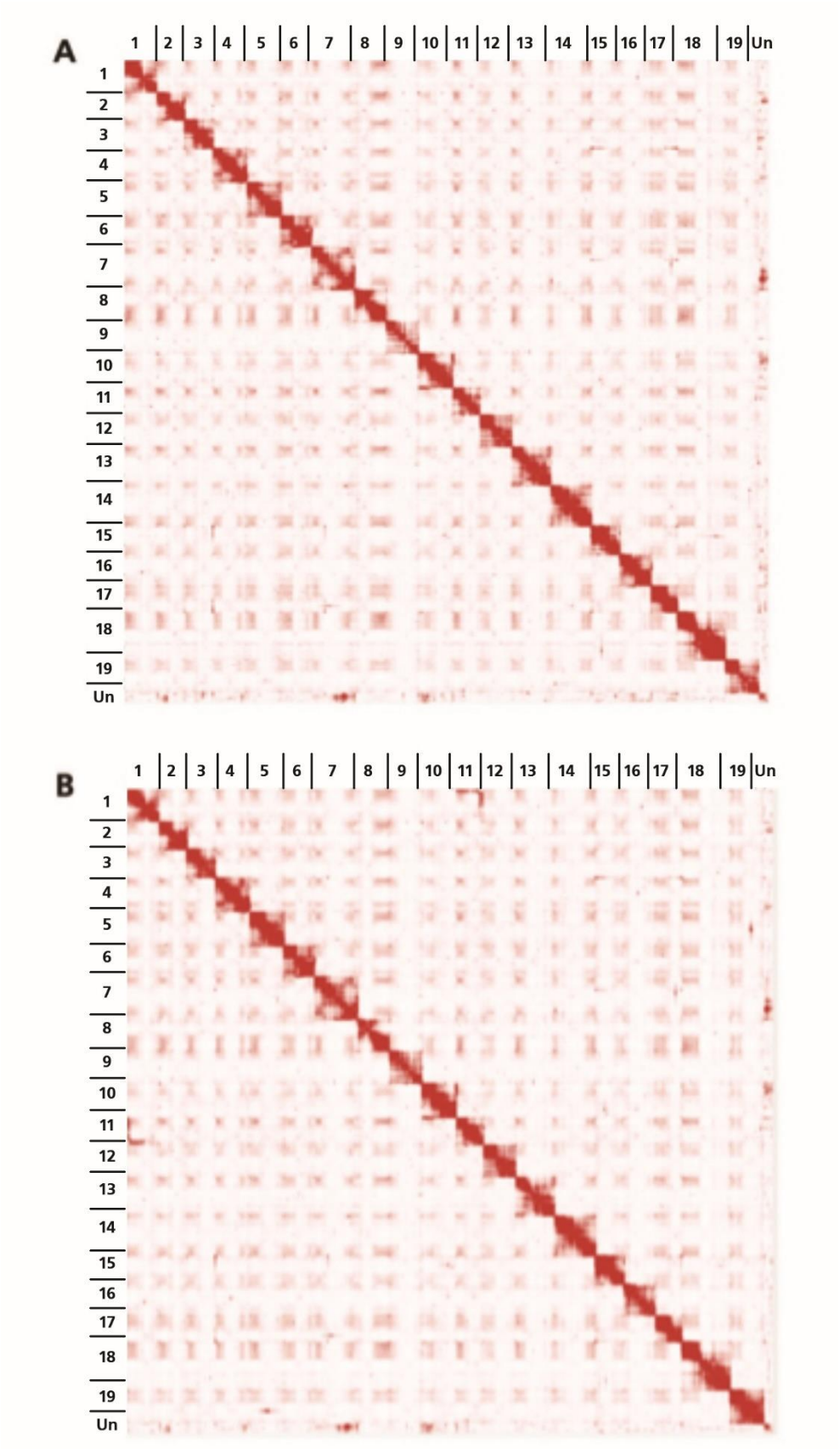
Figure 13 reports the interaction frequency between any two 1Mb size bins across the genome. The contacts occurring between two loci of the same chromosome are defined

"intra-chromosomal" or *cis* interactions; contacts occurring between loci from different chromosomes are defined "inter-chromosomal" or *trans* interactions. Intra-chromosomal interactions were more frequent than inter-, showing high colour intensity in the maps. These findings are in agreement with data from other plant species (Grob, Schmid and Grossniklaus, 2014; Dong *et al.*, 2017; Liu *et al.*, 2017) as well as with data from non-plant organisms (Lieberman-Aiden *et al.*, 2009; Zhang *et al.*, 2012; Rao *et al.*, 2014; Hsieh *et al.*, 2015; Schwartz and Cavalli, 2017), suggesting that CTs are a recognizable structure in grapevine genome.

For each of the three Hi-C maps (*Figure 13*), the interactions occurring outside the main diagonal defined a regular pattern of blocks of signal between several regions of different chromosomes. This kind of signal may be due to physical proximity of *loci* from different chromosomes, representing the set of contacts between pairs of chromosomes.

Finally, at the bottom-right corner of each of the three maps (

Figure 13), the extra chromosome named "unknown" is visible: this is a set of assembled sequences still not anchored onto the *V.vinifera* reference pseudochromosome molecules.


A
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 Un
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 Un
B
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 Un
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 Un

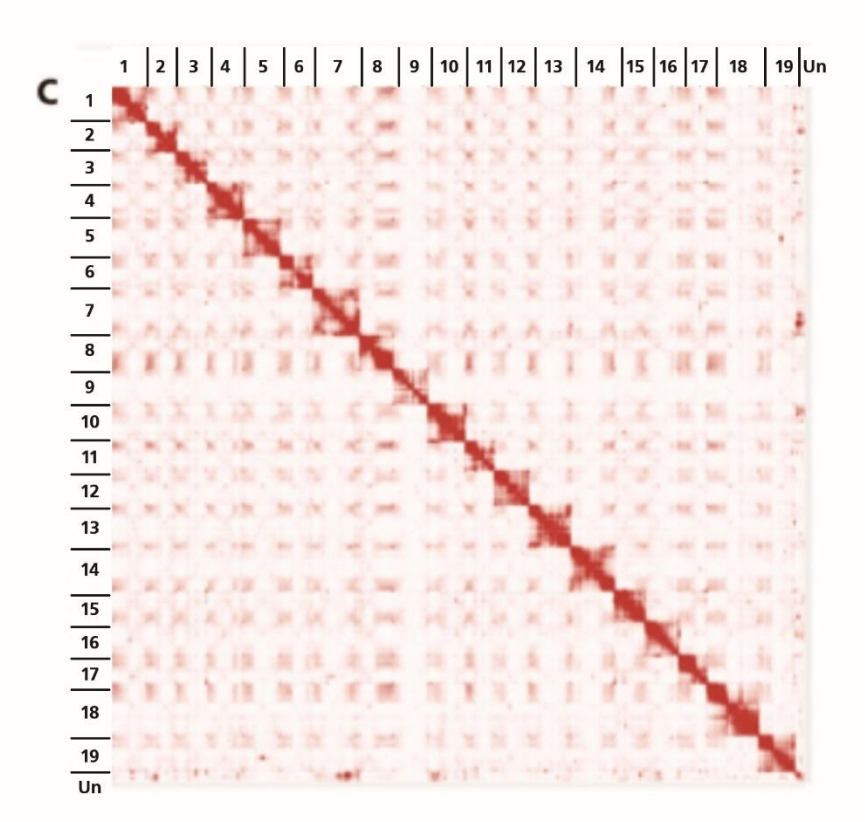


Figure 13 Contact maps for the three varieties Pinot noir (A), Rkatsiteli (B) and Chardonnay (C). Color intensity is proportional to the interaction frequency between 1Mb wide bins.

To understand the spatial distribution of CTs inside the nucleus, we calculated the $\log_2$-ratio between the observed and expected Hi-C *trans*-interactions for each pair of chromosomes. We used this as a measure of the spatial proximity between any two chromosomes in order to reconstruct the "neighbourhood" of the grapevine interphase nucleus. In the first place, we observed that the signal in the Rkatsiteli map (*Figure 14* C) was biased by the high frequency of interaction between chromosomes 1 and 11. This increased *trans*-interaction frequency between the two chromosomes was not only due to the physical proximity of chromosomes 1 and 11, but may be the effect of a previously described reciprocal translocation event between those two chromosomes in Rkatsiteli (Alice Fornasiero, PhD thesis 2017).

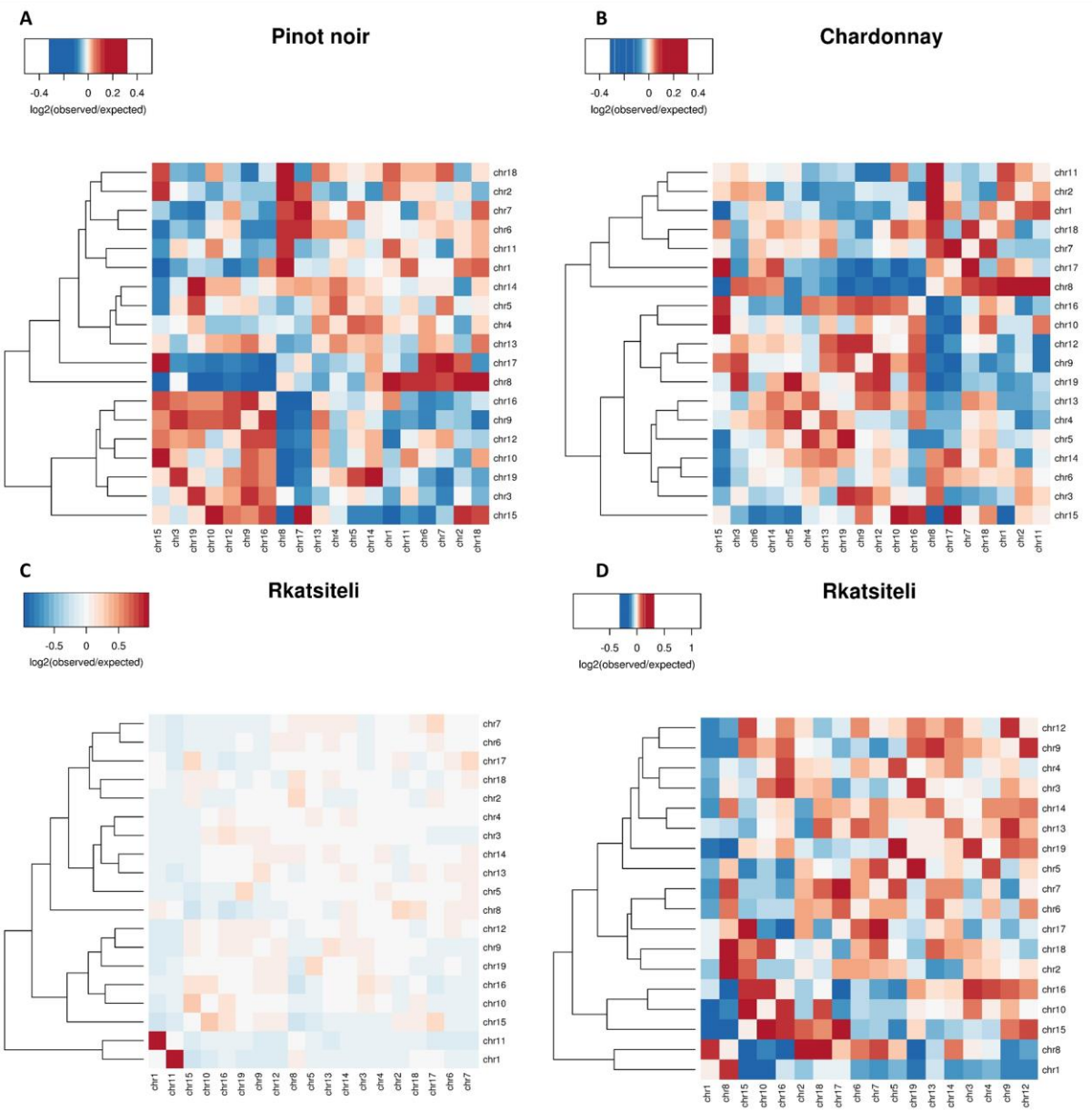


Figure 14 Observed/expected number of contacts between all pairs of whole chromosomes in Pinot noir, Rkatsiteli and Chardonnay varieties. Red indicates enrichment; blue indicates depletion. In the last map, chr11 is excluded from Rkatsiteli due to a translocation between chr1 and chr11.

We then removed chromosome 11 from the Rkatsiteli map (Figure 14 D) and observed a signal of *trans*-chromosome interaction comparable with those of the Pinot noir and

Chardonnay varieties (Figure 14 A,B). Overall, we did not observe clusters of chromosomes that are conserved across all the varieties. Comparing only Pinot noir and Chardonnay (Figure 14 A, B), we found some recurrent higher than expected *trans*-interactions of the chromosome 8 with chromosomes 1, 2, 7, 11 and 18. We did not observe any pattern between the spatial proximity of chromosomes in the nucleus and their physical features such as length, as observed in human and mouse cells, where the shortest, gene-rich chromosomes were grouping together (Lieberman-aiden *et al.*, 2010; Zhang *et al.*, 2012). Differently from what has been found in *A. thaliana*, where all the five chromosomes shared equal interactions (Grob, Schmid and Grossniklaus, 2014), our results suggest that grapevine chromosomes can form clusters, which may be conserved in the cell population and across varieties.

4.1.1.2 Differences in Distance Dependent Decay across grapevine varieties

We computed the distance dependent decay (DDD) of interactions for Pinot noir, Rkatsiteli and Chardonnay in order to compare their pattern of interaction across all the chromosomes.

We obtained three functions (*Figure 15*) in which the $\log_{10}$(contact frequency) decreases for all the chromosomes at a similar rate with the $\log_{10}$(distance). We observed nearly 100% probability of finding contacts between loci 200-300 bp apart and a wider range of contact probabilities (from 4% to 0.3%) between loci at the opposite ends of the chromosomes.

The range of interaction frequencies occurring at long distances (>10 Mb;

$\log_{10}$(distance)>7 in *Figure 15*) may be explained by the tendency of the telomeric regions of a chromosome to be in contact, bringing the ends of the chromosome arms in physical proximity. Such a structure is in agreement with the Rabl organization of the nucleus (Cowan, Carlton and Cande, 2001). Similar examples were found in other plant species like *A.thaliana* (Grob and Grossniklaus 2017) and barley (Mascher *et al.*, 2017), with the latter showing an extreme case in which all the chromosomes showed the same configuration with high contact frequency between loci of the two chromosome arms and tips.

We used the slopes (interaction decay exponents; IDEs) to measure the decay of interaction of each chromosome in each variety. Comparing the IDEs distributions, we did not obtain any significant difference in the overall chromatin organization across the analysed grapevine varieties (Figure 16 A). We then computed the correlation between single chromosome IDEs across the three varieties. We observed that chromosomes 10, 11 and 17 were the ones showing major changes in the correlation coefficients, meaning a different interaction decay relative to other chromosomes. Interestingly, chromosomes 17 and 11 are the shortest chromosomes in grapevine genome (19,560,009 and 20,151,551 bp respectively compared with the average length 24,933,237 bp), and were showing steeper slopes than the average (*Figure 15*).

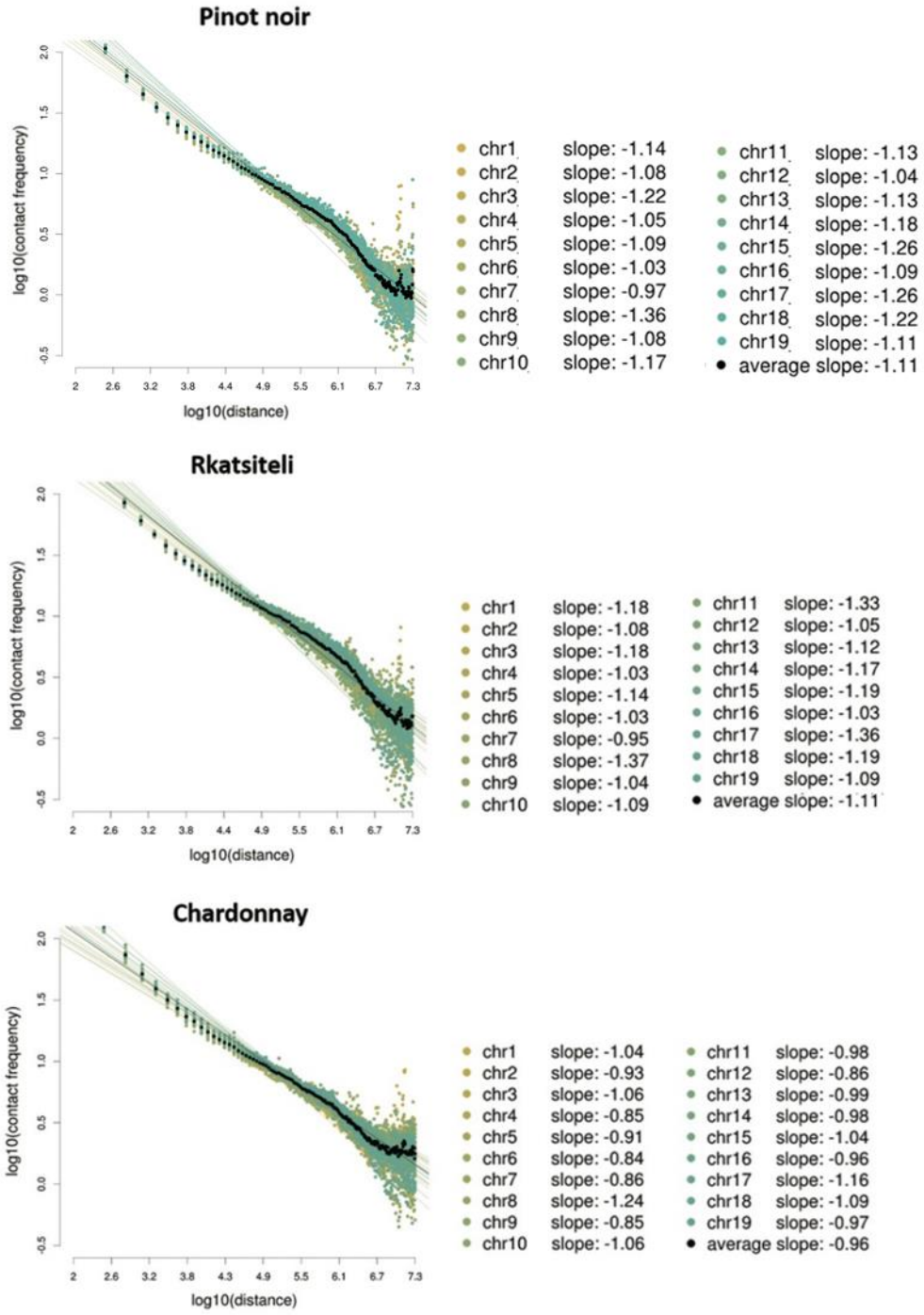


Figure 15 Distance dependent decay plot for the three varieties. For each plot is reported the slope for every chromosome and the average slope which defines the general trend of interaction frequency depending on the distance.

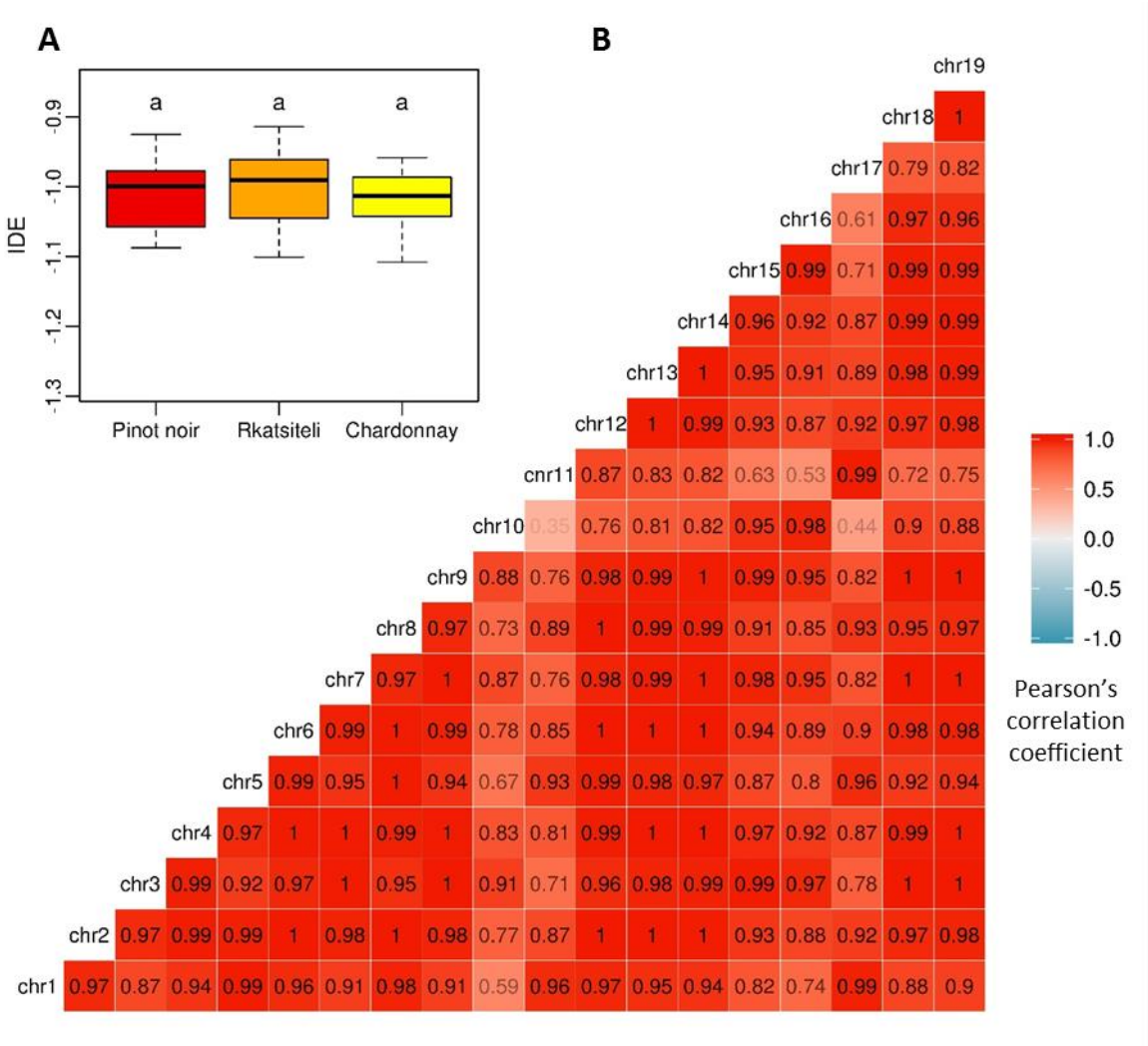


Figure 16 A: Distribution of IDEs of the Pinot noir, Rkatsiteli and Chardonnay full chromosome set. Same letters indicate no significant difference between distributions (Wilcoxon's p-value<0.05). B: Correlation matrix between single chromosome IDEs of each grapevine varieties. Colour intensity is proportional to the correlation value for each comparison.

4.1.1.3 Characterization of discrete *inter*-chromosome contacts

From inspection of the Hi-C contact maps, we observed contacts that were conserved across the three varieties. These contacts appeared as bright foci connecting different chromosomes on the maps (*Figure 17* A). In particular, we could distinguish interactions

between telomeres (white arrow in *Figure 17* A) and contacts between centromeres of different chromosomes (blue arrow in *Figure 17* A). These results suggest that in the grapevine nucleus telomeres and centromeres of different chromosomes co-localize in discrete nuclear portions. This observation could indicate that chromosomes in grapevine nucleus are organized following a Rabl configuration (Rabl, 1885; Cowan, Carlton and Cande, 2001; Cremer and Cremer, 2006) in which telomeres and centromeres are localized at the opposite poles of the nucleus.

Finally, similar patterns of discrete *trans*-interactions were also observed in *A.thaliana* and were due to contacts between genomic regions forming an interacting structure called the KNOT (Grob, Schmid and Grossniklaus, 2014).

We found a unique feature of *trans*-interaction between chromosomes 1 and 11 in the Rkatsiteli map (*Figure 17* B). This seemingly long-range interaction is due to a previously described reciprocal translocation event between those two chromosomes (Alice Fornasiero, PhD thesis 2017) and shows the efficiency of the method in revealing chromosomal rearrangements by building the contact map. Similar patterns revealing both balanced and unbalanced translocation events were recognizable in a Hi-C study on the chromosome rearrangements in human tumours (Harewood *et al.*, 2017).

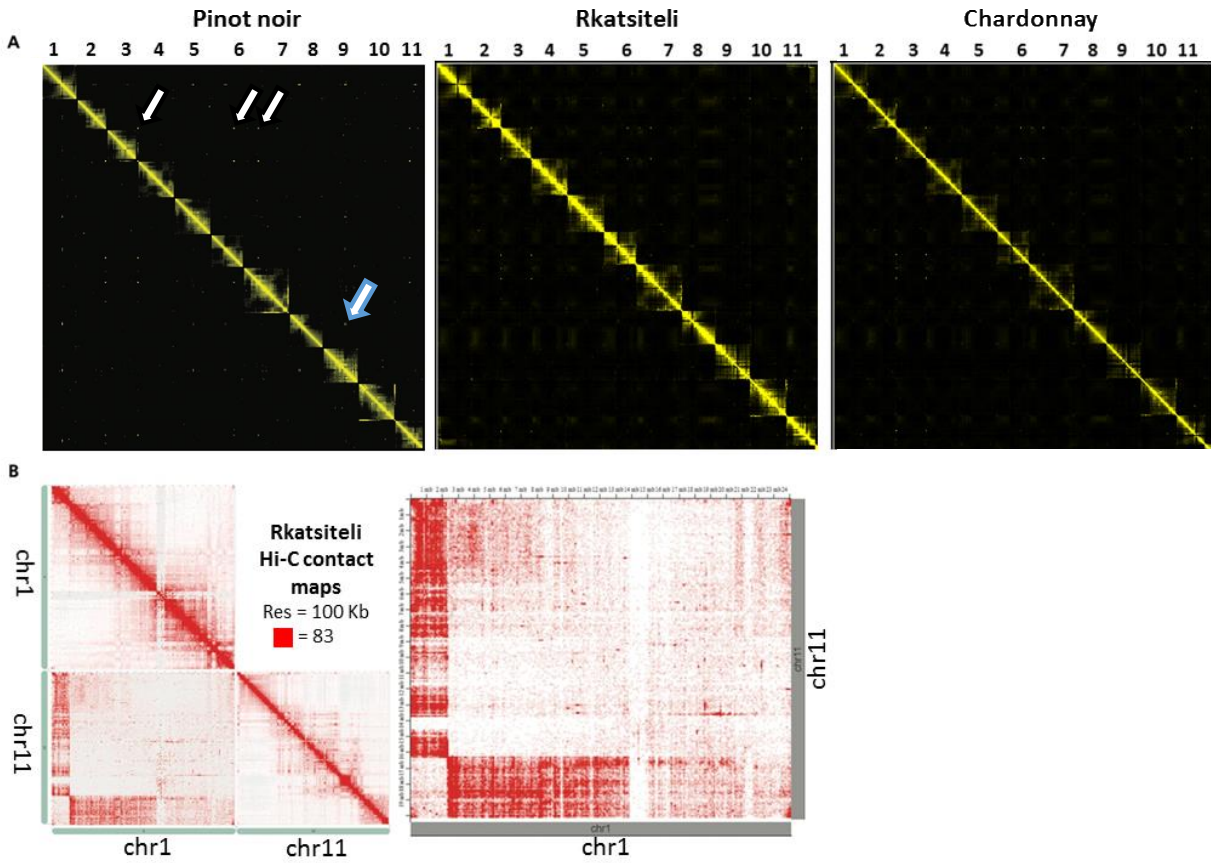


Figure 17 A: Detail of discrete contacts in chr1-chr11 genomic region in the three grapevine varieties. White arrow: interactions between telomeres; blue arrow= contacts between centromeres of different chromosomes. B: Detail of the Rkatsiteli contact map (100 Kb resolution) in the chr1-chr11 region. Chromosome names and coordinates are reported at map axes.

4.1.2 A/B compartments

4.1.2.1 Nuclear compartmentalization of grapevine chromatin

Previous studies have shown that in both mammals (Dixon et al., 2012; Lieberman-aiden et al., 2010; Rao et al., 2014) and in some plants, such as tomato, rice, barley and *A. thaliana* (Grob, Schmid and Grossniklaus, 2014; Dong *et al.*, 2017; Mascher *et al.*, 2017), the nuclear genome can be partitioned into two compartments. Such compartments,

called A and B, are described respectively as the active and inactive parts of the nucleus. A/B compartments can be identified based on a principal component analysis (PCA) of the contact map. We found that such compartmentalization is also a feature of the *V. vinifera* genome structure. In fact, we could divide grapevine chromatin into into A/B compartments via PCA, with the A compartment identified by positive values of the first principal component (PC1) and B compartment by negative values of PC1. Globally, chromosomes showed two main types of compartment composition. Of the 19 chromosomes, 12 were characterized by a "bi-modal" composition, in which each arm had a preferential enrichment for a different compartment (*e.g.* chr3 in *Figure 18* A). In the other 7 cases, chromosomes 4,5,6,7,11,13 and 18 showed a "tri-modal" composition, in which the arm extremities were enriched in positive values (A compartment), while the central region was enriched in negative values (B compartment) (*e.g.* chr4 in Figure 18 A). We assessed the distribution of the A and B compartments inside the grapevine nucleus using the PCA analysis as described in the methods section.

In the normalized genome-wide contact map (*Figure 18* B), we observed a "plaid pattern" made up by alternating blocks of high and low interaction frequency, also seen in other Hi-C analyses (Lieberman-Aiden *et al.*, 2009; Grob, Schmid and Grossniklaus, 2014; Rao *et al.*, 2014; Liu *et al.*, 2016; Dong *et al.*, 2017; Schwartz and Cavalli, 2017). The genome-wide A/B division was consistent with the plaid pattern (*Figure 18* B) and led us to classify different kinds of *inter*-chromosome contacts. We defined three classes of *inter*-chromosomal interactions: those between two A compartments, those between two B compartments, and those between an A and a B compartment. In order to understand how nuclear compartmentalization might promote or restrict chromosome positioning,

we measured the frequency with which each class of interaction occurs.

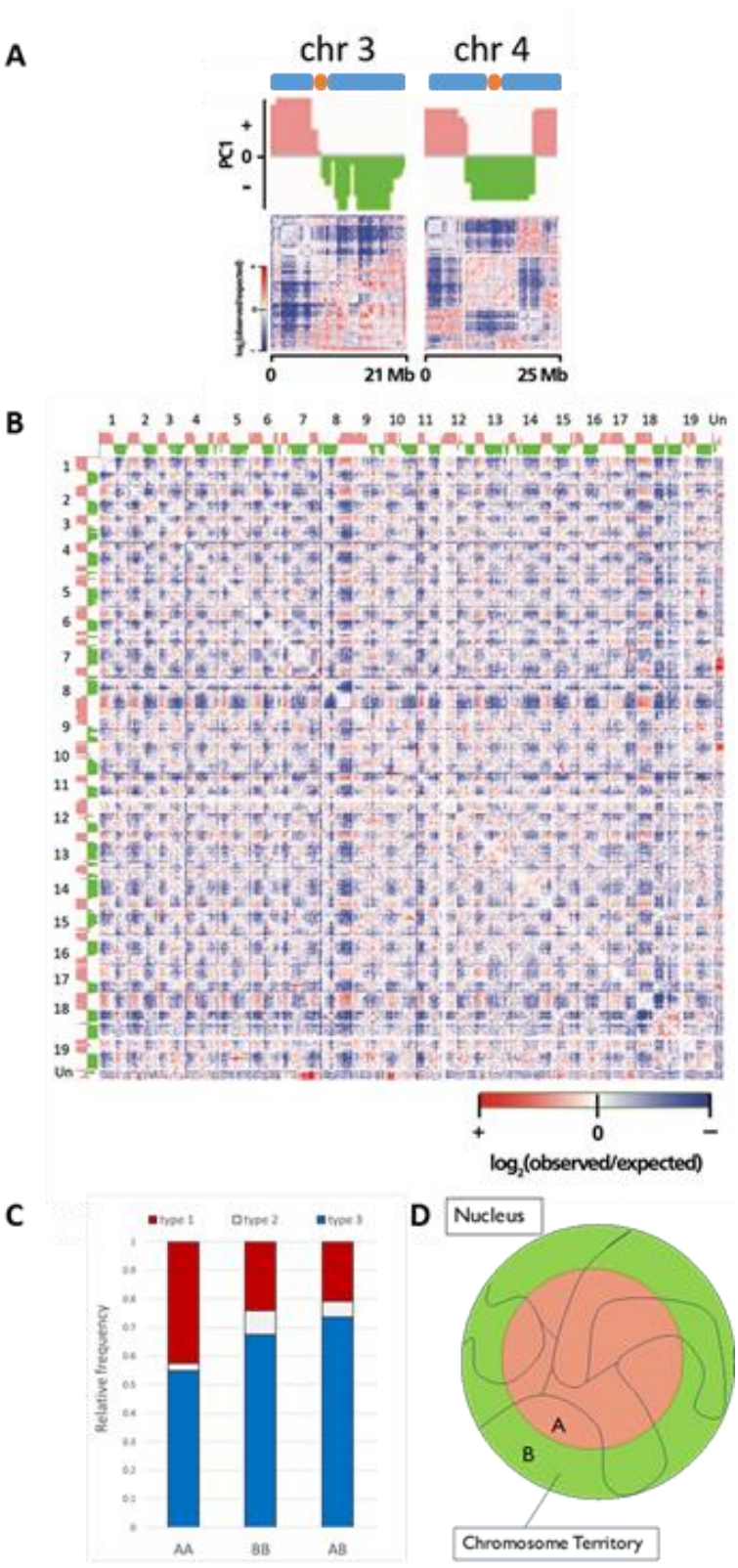


Figure 18 Nuclear compartmentalization of grapevine chromatin.

(*continues from previous page)*

(A) Chromosome 3 showed a bi-modal compartment composition, while chromosome 4 showed a tri-modal one. Compartment A and B are depicted in pink and green respectively above the contact maps.
(B) Genome wide coverage–normalized contact map for Pinot noir. The map reports the interaction frequency in 1Mb windows, colors from blue to red depict loci from lower to higher than expected interaction frequency.
(C) Histogram showing the relative abundance of more than expected, as expected and less than expected (type 1, 2 and 3 respectively) contact frequencies in the three interaction contests.
(D) We have drawn a model representing how A (pink) and B (green) compartments are organized inside the nucleus divided into the nucleus.

We divided the contact matrix (*Figure 18* B) into: more than expected (type 1, red pixels), as expected (type 2, white pixels), and less than expected (type 3, blue pixels) interaction frequencies. Here, the expected frequency of interaction is computed by assuming that each locus has an equal chance of interacting with every other locus in the genome and that loci are expected to interact depending on their linear distance along the chromosome (Heinz *et al.*, 2010).

For each of the AA, BB and AB types of interactions we assessed their interaction frequencies (*Figure 18* C) in relation to the expectations based on the above described model. Interestingly, the AA context was the one with the highest type 1 contacts rate (42%) compared to BB context (24%) and AB context (21%). The BB context presented relative enrichment for type 2 contacts (8% versus 2% of AA and 5% of AB).

The higher probability of finding AA interactions than BB interactions and the low probability of finding AB interactions could reflect the fact that A and B compartments occupy distinct nuclear locations, with B composed by regions at the opposite poles of the nucleus which will rarely interact, and A composed by regions confined at the nuclear core which have higher interaction probability. From the observations gathered, we hypothesize a biphasic nuclear model, in which the A compartment regions of each

chromosome are located toward the core of the nucleus and the B compartment regions are left at the nuclear periphery. We have drawn a graphical representation of such model in *Figure 18* D. These results are in agreement with a Rabl nuclear configuration for *V.vinifera* genome. In fact, both the centromeric and telomeric regions of each chromosome were mostly characterized by B compartment features, meaning that they could occupy opposite poles of the nucleus.

4.1.2.2 A/B compartments are globally conserved across varieties/tissues

To investigate the degree of conservation of A/B compartmentalization across grapevine varieties, we compared the PC1 values among the Pinot noir, Rkatsiteli and Chardonnay Hi-C datasets. This approach did not show any appreciable difference at a global qualitative level (Figure 19 A). We performed the same analysis to assess the stability of the A/B compartment organization in different tissue/cell types. Instead of comparing different varieties, we compared the PC1 from the Hi-C datasets of two different Rkatsiteli organs, namely leaf and the shoot apical meristem (SAM) (Figure 19 B). The two organs represent different stages in development, with the leaf mainly composed of fully mature cells and SAM composed mainly by undifferentiated cells.

In order to avoid any effect of the low complexity issue of the SAM Hi-C library (discussed in the methods section) in the analysis, we subsampled the Rkatsiteli leaf dataset obtaining a dataset which can be compared to the SAM Hi-C data.

The comparison of the PC1 between the two organs showed global consistency in the

A/B pattern.

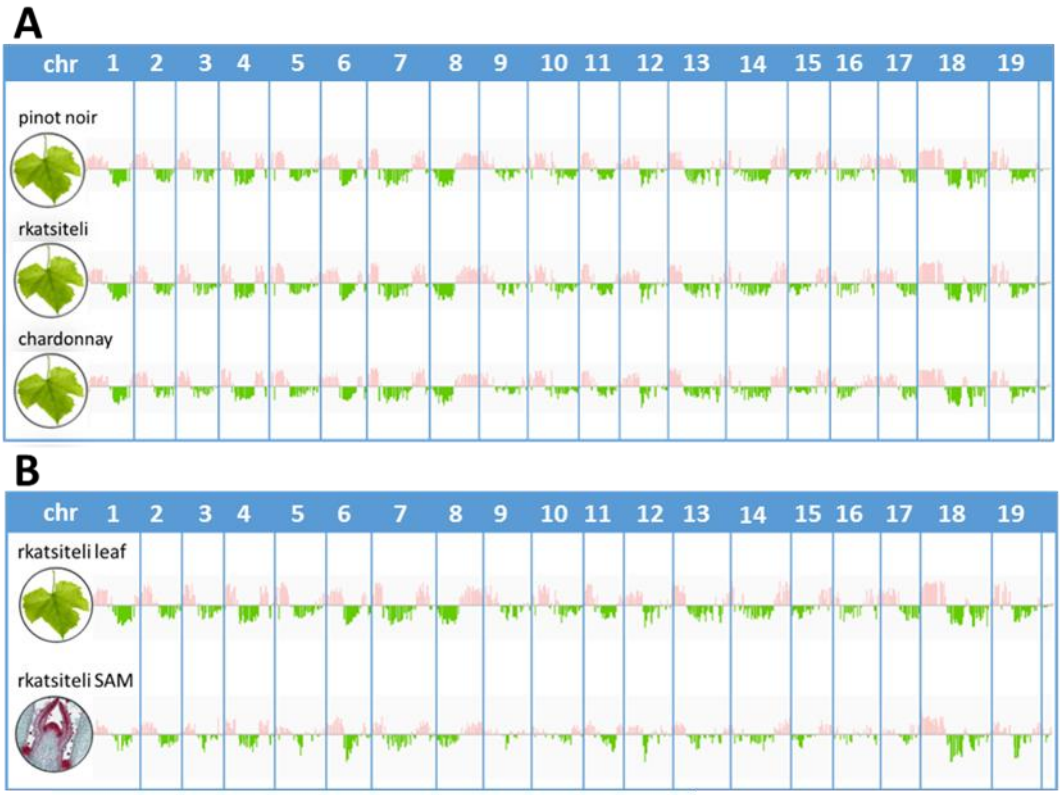


Figure 19 PC1 values comparison across leaf tissue of Pinot noir, Rkatsiteli and Chardonnay grapevine varieties (A) and across Rkatsiteli leaf and SAM (B).

A quantitative evaluation of the chromatin compartments conservation across varieties and organs was obtained using two methods. In the first one (Figure 20 A) we directly compared the PC1 values for each dataset, obtaining a regression line and the respective r^2 value as a measure of the correlation. For all the comparisons we observed high r^2 values (between 0.85 and 0.93), except for the comparison between Rkatsiteli leaf vs Rkatsiteli SAM (r^2 =0.09). In the second method (Figure 20 B) we obtained a correlation score for each pair of varieties and organs datasets using the *getHiCcorrDiff* tool from the HOMER software (Heinz *et al.*, 2010). For all the compared varieties, we observed

high correlation values (medians between 0.94 and 0.97), meaning a high degree of conservation of chromatin compartmentalization across varieties.

These results suggest a conservation of the A/B compartments in the chromatin of the three grapevine varieties Pinot noir, Rkatsiteli and Chardonnay. This was an expected result, since a perturbation in the A/B compartmentalization of the nucleus requires strong alterations in the genomic structure, such as during cell differentiation (Dixon *et al.*, 2015).

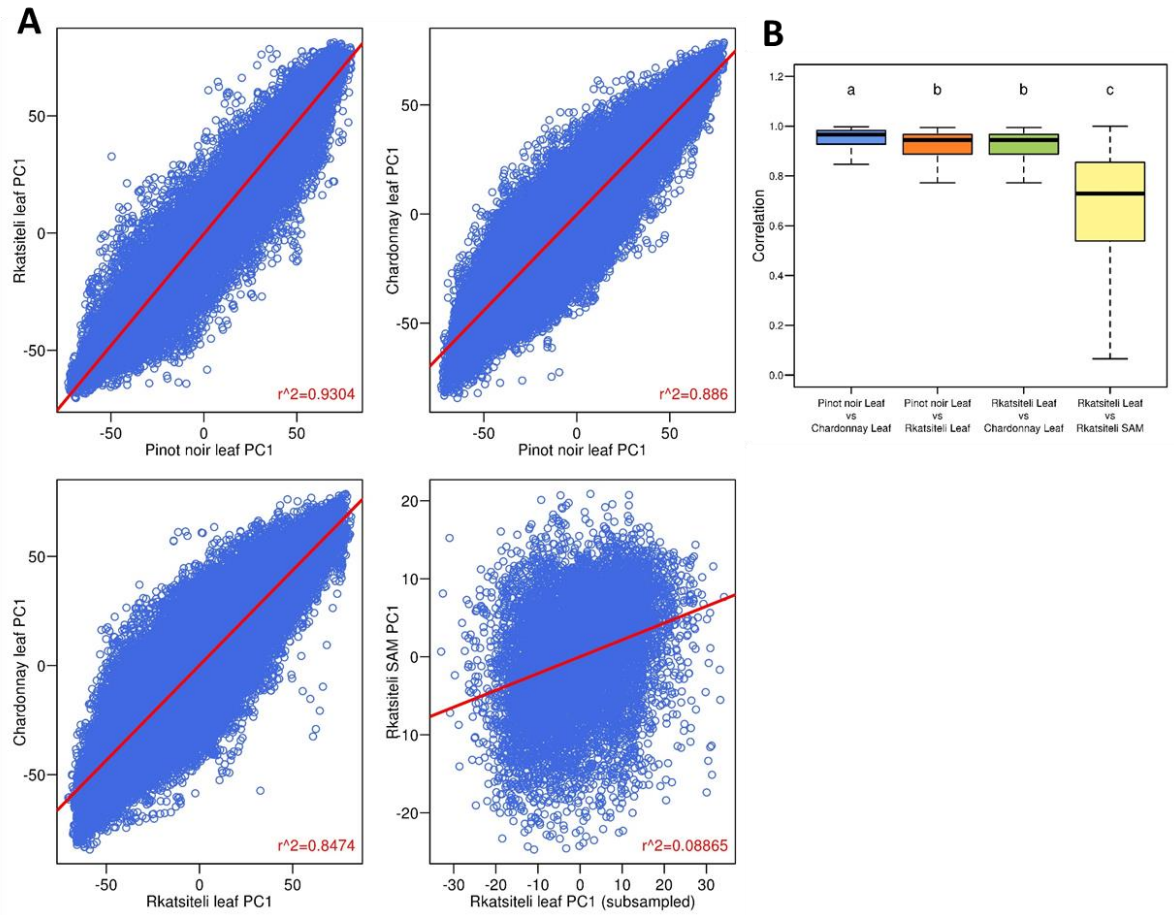


Figure 20 A) PC1 values comparison between leaf samples in the three varieties and in the leaf vs SAM tissue of Rkatsiteli. For each scatterplot is reported in red the regression line and the r^2 value. B) Contact pattern correlation between varieties and organs. Different letters indicate significantly different distributions (pairwise Wilcoxon p <0.01).

We still observed high degree of correlation also between leaf and SAM, but at significantly lower level than the comparisons between varieties (median: 0.64; $p<0.01$). The obtained results suggest that the nuclear organization into A/B blocks could be established in the early stages of cell differentiation, and that differences in chromatin conformation between organs are due to small rearrangements in the A/B compartmentalization. Finally, considering the correlation between Rkatsiteli leaf and SAM in relation with the other comparisons made between varieties (Figure 20), we could conclude that A/B compartments are significantly more conserved across varieties than across different tissues and developmental stages of the same variety.

4.1.2.3 A/B compartments correlate with active/inactive states of chromatin

The characterization of the chromatin structure in mammals, flies and in other plant species found that A compartments are related to active chromatin state and B compartments are associated with inactive chromatin. To verify whether the *V. vinifera* A/B compartments are coupled with functional characteristics of the genome, we associated the identified compartments with known genetic and epigenetic features, namely expression levels, number of genes, methylation, TE content, H3K4me3 and ATAC-seq data. These features were shown to be good descriptors of the chromatin state in terms of activity/inactivity and accessibility to factors regulating the genome functions (Lieberman-Aiden *et al.*, 2009; Grob, Schmid and Grossniklaus, 2014; Dong *et al.*, 2017). The grapevine A compartment showed higher density of genes, H3K4me3 marks and

intergenic ATAC-seq peaks (*Figure 21*). On the other hand, the B compartment showed higher levels of DNA methylation in all the three contexts of CG, CHG and CHH, as well as an enrichment in TE density (Figure 21, plots 1-5). These results are in agreement with the observations already present in the literature, indicating that the A compartment is characterized mainly by markers of active chromatin, while the B compartment is characterized by chromatin inactivation marker modifications (Dekker, Marti-Renom, and Mirny 2013; Dixon et al. 2012; Jin et al. 2013; Lieberman-Aiden et al. 2009; Rao et al. 2014).

Gene density was higher in the A compartment than in B, while TE density was higher in the B compartment than in A (Figure 21, plot 4-5). As shown in other Hi-C studies, this could indicate that chromatin organization can constitute boundaries between genomic regions with different functions or acting as a confinement for regions subjected to high variation (Grob and Grossniklaus, 2017; Xie *et al.*, 2017).

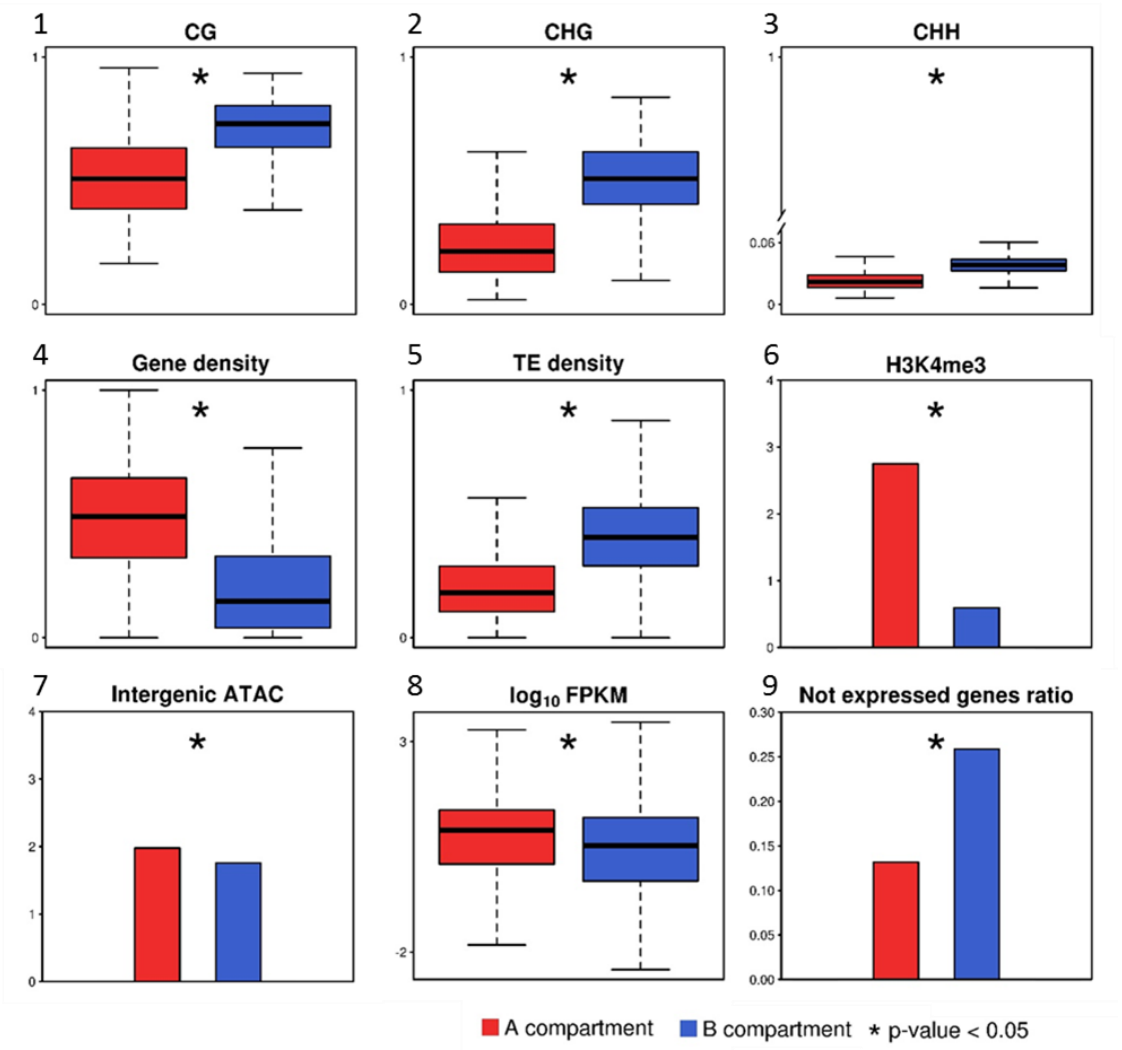


Figure 21 Genomic and epigenomic features of the grapevine genome in A/B compartments.
1-5: In all the plots, the y axis reports the density of the analyzed feature.
6-7: Histone modifications and chromatin accessibility in the grapevine genome A/B compartments. On the y axis, the average number of peaks per 50Kb window are reported.
8-9: Expression rate and not expressed genes distribution in the grapevine genome A/B compartments. In the plot on the left, the y axis reports the value of $\log_{10}$(FPKM); in the plot on the right, the y axis report the ratio of not expressed genes over the total number of genes in A and B compartments respectively.

To understand how the A and B compartments affect gene expression, FPKM for each annotated gene was stratified by compartment. Genes in A compartment showed a significantly higher level of expression compared to genes in B compartment (*Figure 21*, plot 8-9). In addition, the proportion of genes showing no expression in the B

compartment was twice the value observed for the A compartment (p-value <0.05) (*Figure 21*, plot 8-9). These results, together with the gene density result, showed that the A compartment contains the majority of genes, and the most actively expressed. Conversely, the B compartment contains fewer genes, among which a relatively high number is not expressed, or showing lower expression levels compared to genes in A compartment.

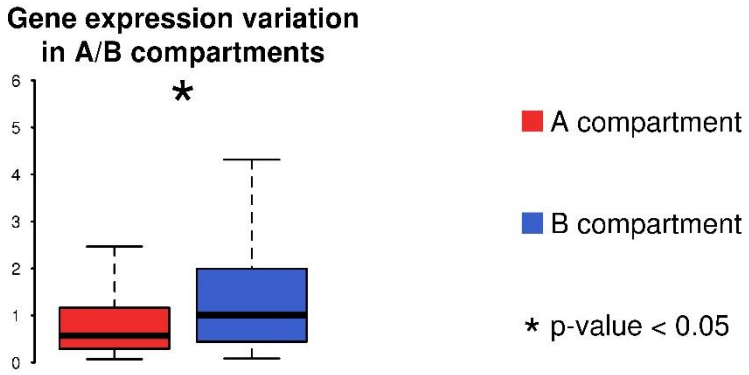


Figure 22 Comparison of variation of gene expression levels in A and B compartments. On the y axis is reported the coefficient of variation; significant difference between distribution assessed with Wilcoxon test.

We also investigated the relative stability of gene expression in the two compartments by comparing the coefficient of variation of the gene expression computed across ten grapevine varieties at four developmental stages. We observed significantly less variation in the expression of genes in the A compartment than in the B. This observation suggested that genes in the A compartment are subjected to regulation pathways which could be highly conserved, opening questions about the compartment's composition in gene classes and gene functions. Therefore, we performed a GO term analysis in order to test whether there was a differential enrichment in classes of genes between the two

compartments. We observed that the A compartment was enriched in GO terms of biological processes which are required for the basic functions of the cell such as growth, differentiation, development, homeostasis, transport, biosynthesis, photosynthesis, etc. (Table 2 A). On the other hand, B compartment showed significant enrichment (p-value <0.05) for GO terms involved in biological processes not strictly required for the cellular growth and maintenance, such as secondary metabolic process, or other biological processes which are activated upon changes in the status of the cell, such as signal transduction, response to stress and DNA metabolic pathways. Finally, we also observed cell-specific biological processes in the B compartment such as those involved in the pollination process (Table 2 A). Interestingly, we observed that the A and B compartments were differently enriched for cellular component terms that belong respectively to the core and the peripheral cellular districts. In fact, the A compartment showed enrichment in GO terms defining all the intracellular components, including the nucleus and the main cellular organs (Table 2 B); the B compartment was enriched in terms describing peripheral components such as the plasma membrane and the cell wall (Table 2 B).

Table 2 Gene Ontology terms enrichment in the A/B compartments for biological process terms (A) and cellular component (B) terms. For each category, only the significantly enriched terms (p-value <0.05) are reported

A

Biological Process

A Compartment

Term	Significant	Expected	classicFisher
biosynthetic process	3122	2921.61	**4.10E-15**
nucleobase-containing compound metabolic...	3003	2812.54	**4.00E-14**
cellular component organization	1408	1272.5	**1.20E-13**
response to endogenous stimulus	369	305.95	**4.80E-12**
response to abiotic stimulus	360	302.52	**2.70E-10**
cellular process	7974	7816.8	**3.50E-09**
post-embryonic development	246	202.37	**4.00E-09**
growth	105	78.89	**5.10E-09**
multicellular organism development	452	392.38	**1.00E-08**
anatomical structure morphogenesis	149	119.36	**1.60E-07**
cell growth	72	54.19	**1.70E-06**
transport	1701	1620.3	**5.20E-05**
cellular homeostasis	224	194.82	**6.60E-05**
flower development	78	62.42	**0.00014**
cell differentiation	205	182.47	**0.00132**
photosynthesis	112	97.41	**0.0042**
carbohydrate metabolic process	765	727.14	**0.00495**
regulation of gene expression, epigeneti...	72	61.05	**0.00657**
response to external stimulus	189	173.55	**0.01921**
translation	468	443.83	**0.01955**
response to biotic stimulus	152	138.57	**0.02256**
lipid metabolic process	551	528.89	**0.04206**

B Compartment

Term	Significant	Expected	classicFisher
secondary metabolic process	202	122.72	**1.00E-16**
cell communication	577	456.73	**2.90E-12**
signal transduction	509	408.17	**6.40E-10**
DNA metabolic process	269	217.18	**1.50E-05**
pollen-pistil interaction	58	36.2	**2.40E-05**
pollination	63	44.73	**1.00E-03**
response to stress	513	467.03	**4.40E-03**
cellular protein modification process	840	791.33	**1.31E-02**

B

Cellular Component

A Compartment

Term	Significant	Expected	classicFisher
intracellular	6646	6281.78	**< 1e-30**
cell	7736	7489.49	**7.70E-22**
nucleus	2384	2186.95	**6.50E-19**
cytoplasm	3892	3677.96	**4.20E-16**
plastid	696	627.76	**7.70E-08**
nucleoplasm	207	178.56	**2.80E-05**
mitochondrion	519	473.1	**3.10E-05**
cytosol	641	595.42	**0.00019**
Golgi apparatus	365	331.1	**0.00023**
thylakoid	185	162.39	**0.00045**
nucleolus	202	179.26	**0.00077**
endosome	116	99.12	**0.00078**
vacuole	370	348.68	**0.01769**
endoplasmic reticulum	341	320.56	**0.01788**
ribosome	298	282.6	**0.04824**

B Compartment

Term	Significant	Expected	classicFisher
membrane	2430	2212.89	**1.70E-15**
plasma membrane	558	489.2	**6.40E-05**
cell wall	108	91.24	**0.023**
external encapsulating structure	108	91.24	**0.023**

In addition to the expression data, also the H3K4me3 histone modification marker confirmed a higher frequency of active transcription sites in A compartments which were almost six times the B compartment amount (*Figure 21*, plot 6). Finally, via the chromatin accessibility assay data (ATAC-seq), we observed a significant (chi-square p-value > 0.05) higher number of intergenic ATAC peaks in the A compartment than in B. Intergenic ATAC peaks are markers of putative enhancers, thus an enrichment for such markers could indicate a region in which the chromatin structure allows dynamic interactions between distant loci of the genome.

A general profile of the chromatin condensation state across the nuclear compartments could be drawn from these results, indicating that chromatin is more open and accessible in A compartment, while it is more condensed in the B compartment (*Figure 21*, plot 7).

The results reported here confirm the observations made in other Hi-C data analyses performed on several plant species. The distribution of genetic and epigenetic features across A/B compartments was analysed in detail for *A.thaliana* (Grob, Schmid and Grossniklaus, 2014), tomato, rice, maize, foxtail millet and sorghum (Dong *et al.*, 2017). In all the cases, the results presented here confirm the findings of the cited works. In *A.thaliana*, the chromatin accessibility assay was not performed; conversely, in this work only the distribution of H3K4me3 marker was analysed, while other histone modifications besides H3K4me3 were taken into account in the other works. Finally, the present work is the only in which besides the expression patterns, also the distribution of not expressed genes in the two compartments was considered.

Taken together, these observations confirmed that also in grapevine genome the A and B compartments reflect the characteristics of active and inactive chromatin, respectively.

Moreover, the definition of these nuclear structural compartments is the result of the interplay of several features, and at the same time, the chromatin state defines and can be defined by the local genomic and epigenomic context.

4.1.3 Sub-compartment domains

We sought for domains of locally compacted chromatin smaller than the A/B compartments, commonly referred to as TADs. Since the TADs definition is not well established in plants, we adopted a general terminology for *V. vinifera*, calling such domains as sub-compartment domains.

From the analysis of the pooled grapevine Hi-C dataset, we identified a total of 747 sub-compartment domains, covering approximatively 21% of the whole genome (*Table 3*). These findings are similar to those from previous work on rice in which 1,763 domains were found, covering 25% of the genome (Liu *et al.*, 2017).

The size of the domains found in grapevine ranged from 60 Kb to 2Mb in length and were sparsely distributed along the chromosomes instead of occurring consecutively as seen in mammalians (Dixon *et al.*, 2012).

Table 3 Summary of the domains annotated in grapevine genome with individual chromosome description and total.

***V. vinifera* sub-compartment domains**			
chr	counts	cumulative length	chr %
chr1	47	5,060,000	20.9
chr2	36	5,800,000	28.3
chr3	34	4,130,000	19.6
chr4	44	5,605,000	22.1
chr5	43	7,010,000	27.3
chr6	55	5,975,000	26.4
chr7	30	4,430,000	14.0
chr8	50	5,555,000	23.6
chr9	14	2,285,000	9.4
chr10	33	5,635,000	22.0
chr11	24	3,920,000	19.5
chr12	28	4,700,000	19.4
chr13	45	6,880,000	23.6
chr14	55	7,585,000	24.8
chr15	39	4,355,000	20.4
chr16	35	3,950,000	16.9
chr17	30	3,640,000	18.6
chr18	57	8,675,000	24.0
chr19	48	5,680,000	22.9
total	747	100,870,000	20.8

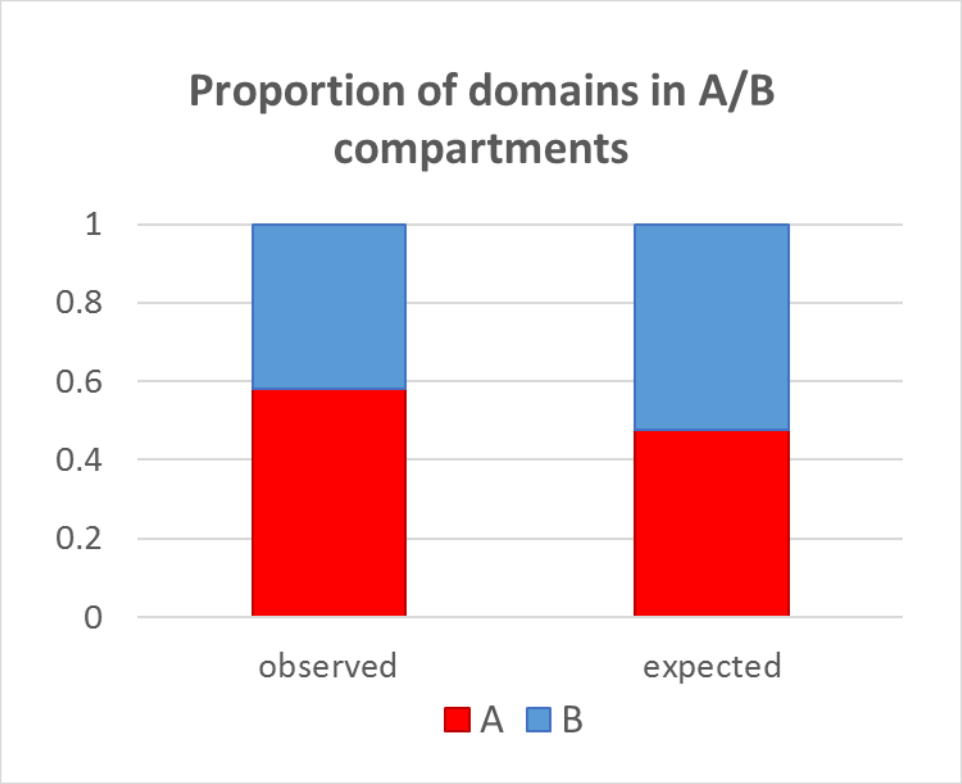


Figure 23 Presence of sub-compartment domains into A compartment was higher than expected (chi-squared p-value <0.01) in *V. vinifera* chromatin.

To understand how the sub-compartment domains relate to the higher order of compartmentalization, we assessed their distribution in the A/B compartments, resulting in: 429 domains out of 747 (57.4%) located in the A compartment, while 309 out of 747 (41.3%) were in the B compartment (*Figure 23*). The remaining 9 domains (1.2% of the total 747) were found to span the A/B boundary, having half of their length (from 45% to 55% of total length) in both compartment A and compartment B. The distribution of the domains and the total genomic proportion of A and B compartment (covering respectively 47.7% and 52.2% of the total genome) were significantly different (p-value <0.01; chi-squared statistical test).

We then asked what is the role of the sub-compartment domains, in terms of chromatin

activity and genomic functions. We analysed genomic features such as the enrichment of gene density, TE and epigenomic features at the domain borders (*Figure 24*) and in the 50 Kb outside (white area) and inside (red area) the domains. The analysis made on the gene density and the TE density reported opposite trends. Gene density was much higher outside the domains (between 420 and 555 bases per 5Kb bin) than inside the domains (270 bases per bin). TE density outside the domains was lower than inside, ranging from an average of 1100 base pairs outside, to an average of 1620 base pairs per bin inside the domain.

Intergenic ATAC peaks did not show any difference in trend outside and inside the domains. This could suggest that sub-compartment domains found in this analysis do not constitute any physical constraint to the distribution of putative enhancers in the genome.

All the three methylation contexts (CG, CHG and CHH) showed a common trend with the minimum at the domain boundary and an increase of methylation level inside the domains (0.59 to 0.62 for CG; 0.29 to 0.39 for CHG; 0.024 to 0.032 for CHH).

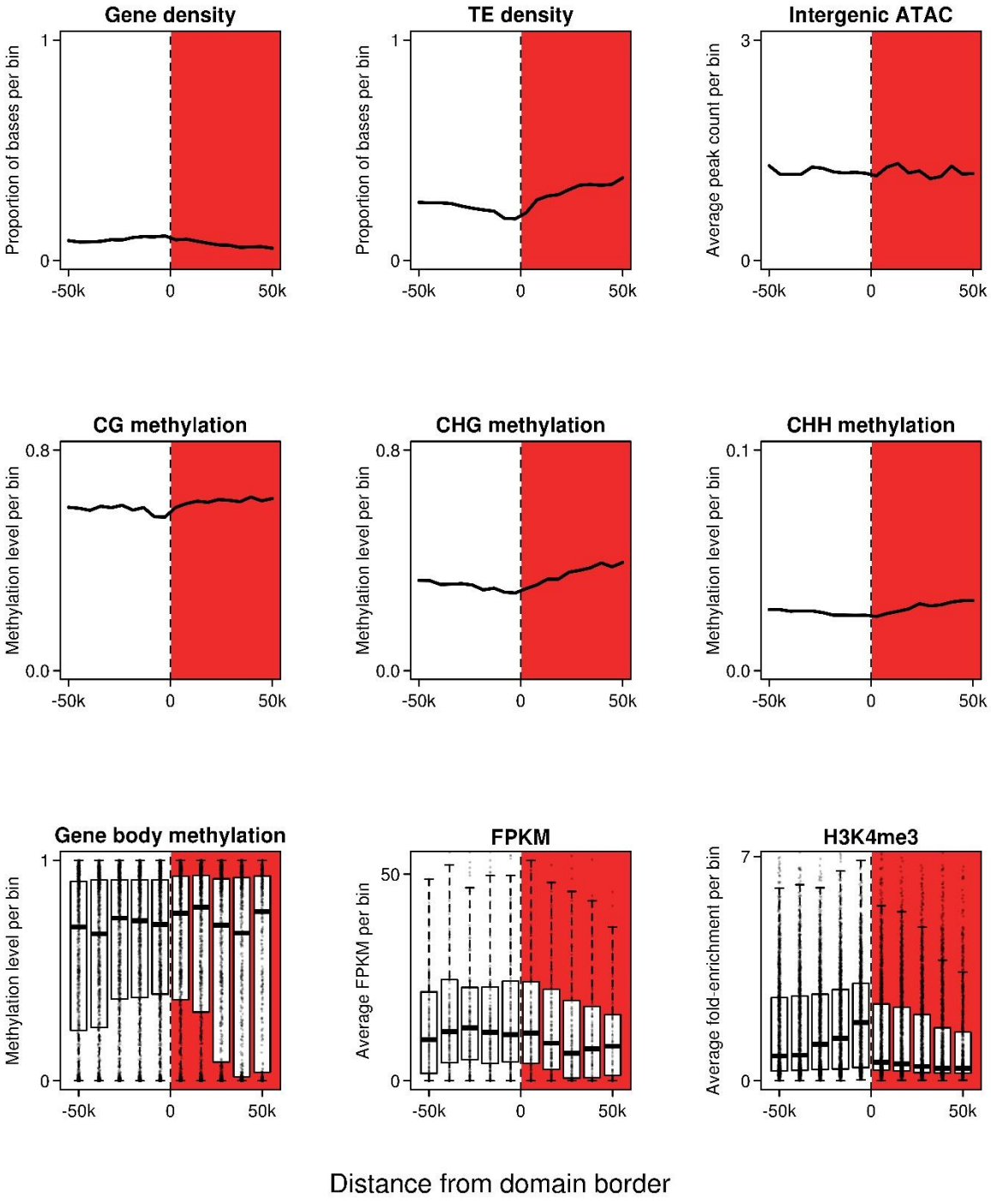


Figure 24 Domain borders analysis of genomic and epigenomic features of the grapevine genome at 50 Kb inside (red area) and outside (white area) the domains. Each value was computed for 5 Kb bins except for the boxplots in the last row, in which data were divided into 10 Kb bins.

Gene body methylation showed a constant level outside the domains, but after the border the trend in methylation level was unclear.

The expression level (measured as FPKM) of genes was lower inside the domain (median between 6 and 11) when compared to the extra-domain area (median between 9 and 13).

The last feature analysed was the enrichment for the H3K4me3 histone modification, which showed a maximum peak at the border (median: 2.03), followed by a dramatic decrease inside the domain (median between 1.13 and 0.3).

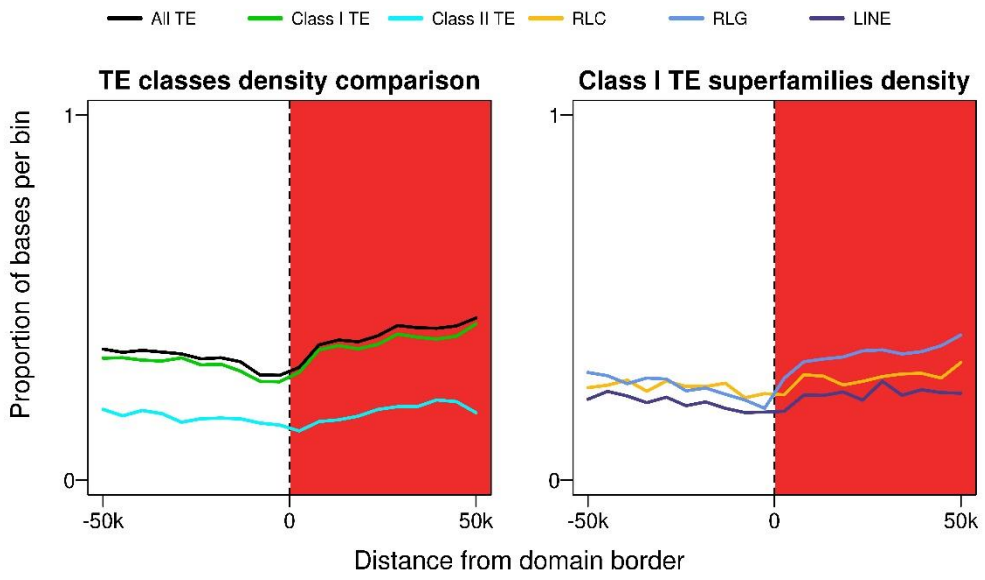


Figure 25 Detailed proportion of TE density at the domain borders every 5 Kb bin. In A, class I and class II elements are reported in relation of the total TE density; in B is reported the density for each superfamily of class I elements.

One of the most prominent features of the sub-compartment domains was the enrichment in TE density inside the domains. We stratified the analysis observing that

there is a significant enrichment in density of class I TE (RNA elements) inside the domains (from 0.23 at the border, to 0.43 at 50 Kb inside the domain). On the other hand, class II TE (DNA elements) showed an overall constant density, both outside and inside the domains (Figure 25). This observation is in agreement with the previously shown results (*Figure 24*), in particular with the gene density plot. Inside the class I TE, we further distinguished between *copia-like* elements (RLC), *gypsy-like* elements (RLG) and long interspersed nuclear elements (LINE). Interestingly, RLG was the most abundant superfamily across the domain border and showed enrichment inside the domain (up to 0.4) when compared to RLC and LINE, which remain at a baseline between 0.15 and 0.18 (Figure 25).

In conclusion, these results suggest that the sub-compartment domains we found could represent regions of the grapevine genome with high levels of chromatin compression, low accessibility, global inactivation of gene transcription, and high presence of repetitive DNA. These sub-compartment domains seem to be mostly defined by an increased level of LTR-retrotransposons of the gypsy superfamily that are usually found in plants and also in grapevine in highly heterochromatic pericentromeric regions (Gabriele Magris, PhD thesis, 2016).

This is in agreement with results found in other Hi-C studies on plants. In particular, the same trend in gene density and H3K4me3 were also reported in rice (Liu *et al.*, 2017). Moreover, our results showed agreement with a similar analysis performed by Dong et al. (2017); hence, the domains found in grapevine could be analogous to the “repressive domains” described in the cited work.

Moreover, the fact that the majority of domains is located in the active grapevine

chromatin A compartment, opens new questions on the origin of such domains, if their structure is a consequence of the DNA sequence content, or if their structure is the framework that influences the DNA regulation.

Then we asked if the sub-compartment domains we found in grapevine genome could have a functional role, influencing the gene expression. In order to assess the level of co-regulation between genes inside the sub-compartment domains, we measured the coefficient of correlation for expression levels inside and outside the domains. Of the 747 sub-compartment domains, 123 (16%) could not be used for the analysis since they were containing less than three genes. Of the reported domains, 134 (18% of the total) showed significantly higher correlation coefficients of gene expression compared to the genes outside the domain.

From a study on variation of expression in ten grapevine varieties (Magris et al., paper submitted), we know that the expression correlation is higher for consecutive genes up to an inter-TSS distance (the distance between the transcription start sites) of 2Kb, than for genes with longer inter-TSS distances. In the same study, has been also observed significant higher correlation for genes up to an inter-TSS distance of 48 Kb in comparison with randomly sampled unlinked genes.

The genes in the majority of our predicted sub-compartment domains (66%) did not show significant higher expression correlation than randomly sampled unlinked genes. This observation suggests that the predicted sub-compartment domains have not the same features in gene regulation as reported in other studies in different organisms, where genes inside the domains showed highly correlated expression levels (Nora *et al.*, 2012; Zhan *et al.*, 2017). Our results indicate that the majority of the domains we

detected are not corresponding to functional units of the genome capable of affecting gene expression. The positive results observed in only the 18% of the cases could be due to conserved locations of the genome, in which sub-compartment domains boundaries are present in a significant portion of the cell population.

These observations reflect the state of the art of the sub-compartment domains or TADs identification in plants, which existence and definition is still not clear (Wang *et al.*, 2015; Liu *et al.*, 2016, 2017; Dong *et al.*, 2017).

Findings in *D. melanogaster* (Rowley *et al.*, 2017) suggested that TADs are a characteristic of all eukaryotes, but only mammalian TADs require CTCF at TADs boundaries. Moreover, a recent single cell Hi-C study even put in doubt the actual existence of TADs as physical organization units of genomes, questioning whether they result from a statistical effect generated merging the interactions from individual cells (Flyamer *et al.*, 2017). Moreover, other studies on human and mouse cell lines showed that depletion of cohesin (which together with CTCF constitutes the protein structure present at the TAD boundaries) causes the disruption of TADs, but not a dramatic effect on gene expression regulation (Rao *et al.*, 2017; Schwarzer *et al.*, 2017). Taking these observations together, the TAD boundaries seen in mammals are conserved locations with CTCF/cohesin binding sites which are maintained across the cell population. Therefore, they constitute foci of chromatin interactions in a Hi-C map which is the average of the interactions across the entire cell population. Cohesin depletion causes the disappearance of such foci in the Hi-C map, and TAD structures are not revealed. Nonetheless, in cohesin depletion state, TADs are still present in cells, but cannot be revealed by Hi-C at cell population level, since the single cell variability in TAD location results into a distribution

of probabilities in which every location along the genome can be a TAD boundary.

The lack of a CTCF/cohesin system in plant genomes constitutes a situation comparable to the mammalian cohesin-depleted genomes. This could explain the fact that in plant genomes the assessment of TAD structures led to contrasting results. In fact, although TADs are not prominent features in *A.thaliana* genome (Wang *et al.*, 2015), in five plant species (maize, sorghum, tomato, foxtail millet and rice) "wide-spread" blocks of local highly condensed chromatin were found at sub-megabase scale (Dong *et al.*, 2017; Liu *et al.*, 2017) and recently TAD structures with active gene enrichment at their boundaries were observed also in cotton (Wang *et al.*, 2018). In conclusion, the lack of a consensus for TAD boundary locations in plant cell population indicates the need for single cell Hi-C studies for a reliable investigation of TAD or sub-compartment domains structures.

4.1.4 Chromatin loops

We analysed the Hi-C interaction data for the Pinot noir and Rkatsiteli leaf samples, searching for significant interactions occurring between loci at distances greater than 1000 bp. A total of 6,355 and 4,910 unique significant long-range interactions were detected (FDR <0.05), respectively. In order to assess to what extent the detection of long distance interactions may have been affected by the frequent presence of SVs that is characteristic of grapevine, we partitioned the genome of Pinot noir and Rkatsiteli according to the haplotype sharing with the PN40024 reference. For each variety, we distinguished between: sharing 0 regions (heterozygous and with both haplotypes

different from reference); h_sharing 0 regions (homozygous and different from reference); sharing 1 regions (heterozygous and with one haplotype identical to the reference); and sharing 2 (homozygous and identical to the reference) regions (described in detail in the methods section). For each sharing region, we assessed the distribution of SVs and loops in Pinot noir and Rkatsiteli (Figure 26). By definition, SVs are absent in sharing 2 regions, while the SVs effect on the loops detection was expected to be influencing the analysis in sharing 0 and sharing 1 regions. We observe a low frequency of SVs in the sharing 2 regions (Pinot noir: 0.013; Rkatsiteli: 0.015) that may be due to errors either in the SVs calling method or in the sharing regions definition.

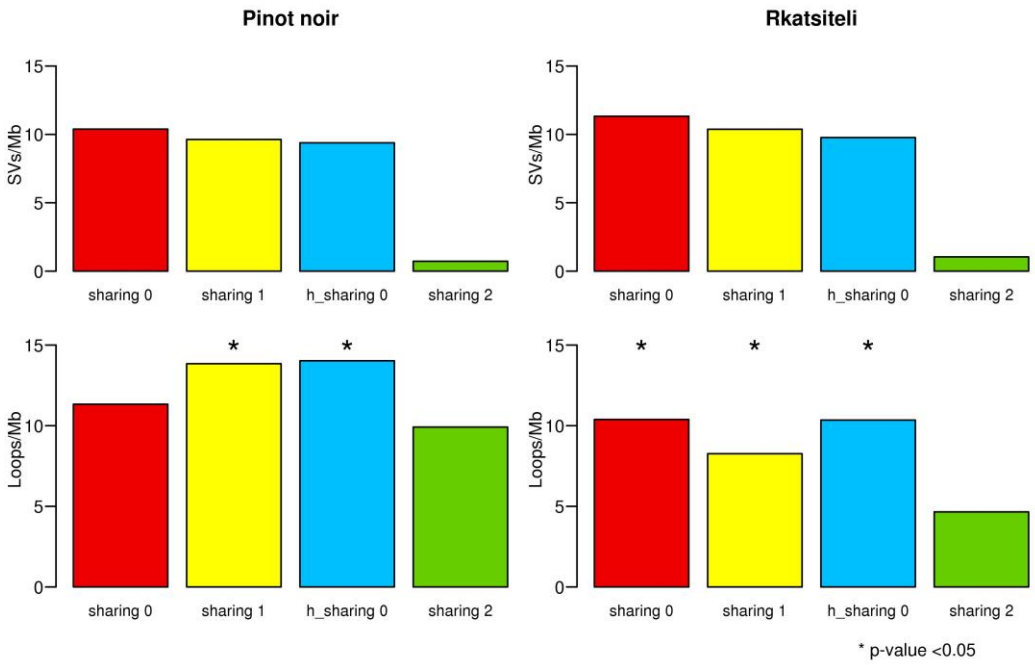


Figure 26 SVs and loops frequency found in Pinot noir and Rkatsiteli according to the level of haplotype sharing with the reference sequence. We tested the difference in loop frequency of each sharing region compared with the sharing 2 region; frequencies significantly higher than the frequency of loops in sharing 2 region (green), indicate an effect of SVs on the detection of loops.

We can consider the sharing 2 region as SV-free, so the detection of loops in such region

is not affected by the presence of SVs. In order to evaluate the effect of the SVs on the detected loops, we compared the frequency of loops in sharing 2 with the frequency of loops in each of the other sharing regions (Figure 26). In Pinot noir we observed loop that sharing 1 and h_sharing 0 regions have loop frequencies significantly higher than the sharing 2. This was an expected outcome, since these regions are more affected by the presence of SVs, so part of the loops observed in sharing 1 and h_sharing 0 could be due to SVs. We did not expect to find no significant difference when testing the loop frequency in sharing 2 against the sharing 0 region. On the other hand, all the comparisons made in Rkatsiteli resulted in a significantly higher frequency of loops than expected in the sharing 0, sharing 1 and h_sharing 0 regions, thus the loop detection in these regions is affected by the presence of SVs.

Table 4 Summary of loops involving genes across *V.vinifera* genome.

class	count	
	Pinot noir	Rkatsiteli
gene-other	874	746
gene-gene	4412	3464
other-other	1069	700
same gene	2402	1790
tot	6355	4910

In order to characterize the interacting partners that are brought in contact by the loops, we considered the subset of Pinot noir and Rkatsiteli loops involving genes. We observed that the majority of the loops (Pinot noir: 83%; Rkatsiteli: 86%) involved at least a gene (*Figure 27* A); with 4,412 loops in Pinot noir and 3,464 loops in Rkatsiteli occurring between two coding regions. Of these “gene-loops” more than half were loops between

different genes ("gene-gene"; Pinot noir: 53%; Rkatsiteli: 60%), while the remainder occurred inside the same gene (*Table 4*). We tested the significance of our results comparing the loops found in Pinot noir and Rkatsiteli to a set of simulated data (described in the methods section). We observed that for all the loop categories described we found results significantly different ($p<0.01$) from the random distribution (*Figure 27* C). Our results are in agreement with an analogous analysis, carried out in maize, which also reported that 74.89% of the total 5,616 loops detected occurred between genes (Dong *et al.*, 2017). As shown in both Pinot noir and Rkatsiteli (*Figure 27* B), looping events inside the same gene were also observed in *A. thaliana* chromatin loops (Liu *et al.*, 2016), where 12% of gene-gene loops were found to form "self-loops", occurring between the 5' and 3' portions of the same coding region.

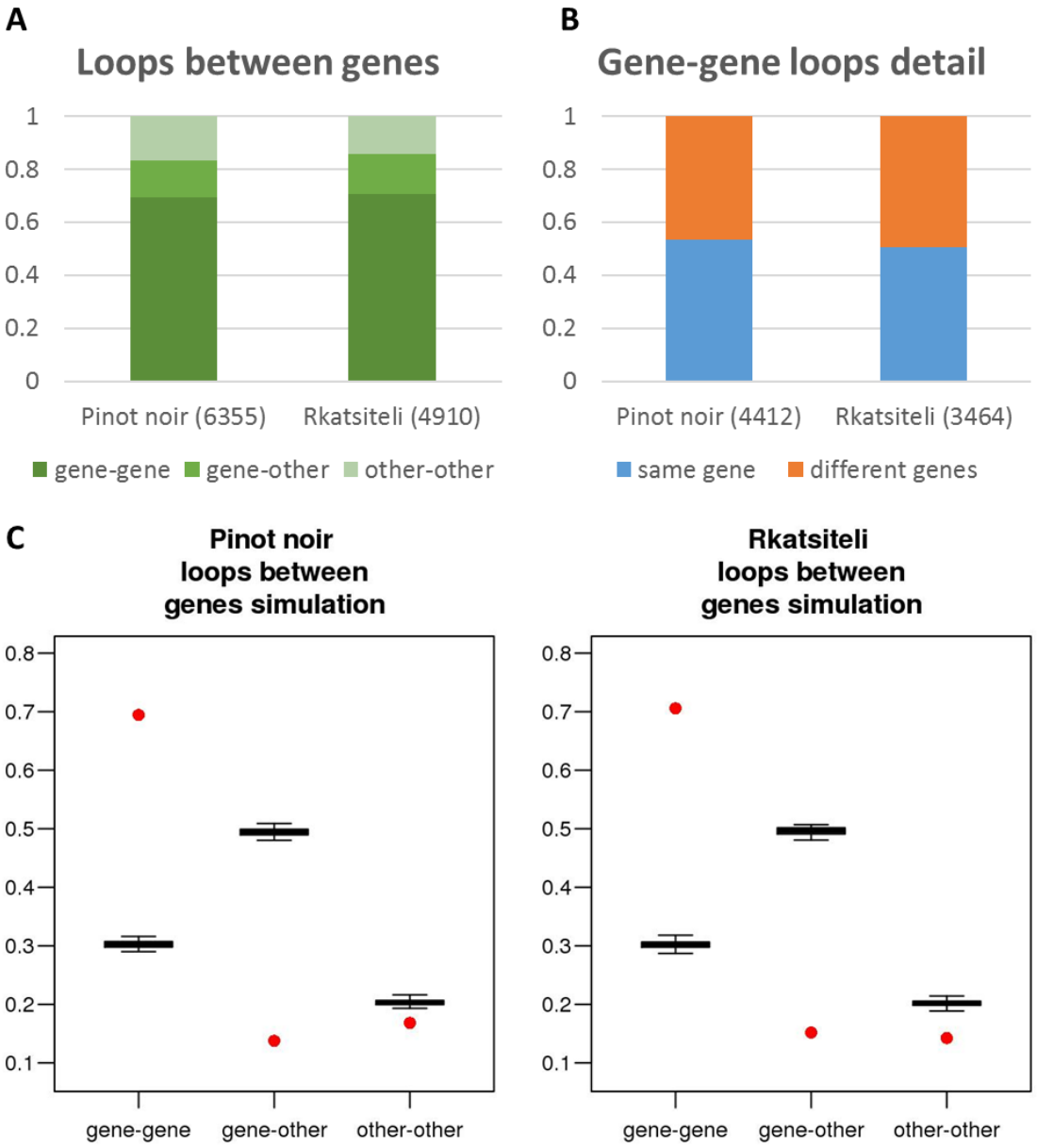


Figure 27: A) Distributions of interactions involving genes in Pinot noir and Rkatsiteli. In both varieties, most of the loops occurred in regions enriched in genes. B) Stratification of the loops involving genic regions in loops occurring between different genes and loops occurring inside the same gene in Pinot noir and Rkatsiteli. C) Test for significant results via comparison of the Pinot noir and Rkatsiteli data (red points) with a simulated loop dataset constituting a random distribution of loops (green boxes) in three categories (gene-gene; gene-other; other-other).

We investigated the biological function of the "self-looped" genes by means of a gene ontology analysis. We found a significant enrichment (Fisher's p-value <0.05) in GO terms mainly describing constitutive biological processes (Table 5). This observation is in

agreement with the self-looped gene model, in which the physical contact between the promoter and the transcription termination site would promote the recycling of the transcription machinery on the gene (Tan-Wong *et al.*, 2012).

Table 5 GO term analysis for "self-looped" genes in the biological process category.

	GO.ID	Term	Annotated	Significant	Expected	classicFisher
Biological Process	GO:0006139	nucleobase-containing compound metabolic...	4100	553	417.91	**2.40E-15**
	GO:0009987	cellular process	11395	1265	1161.49	**1.10E-09**
	GO:0006259	DNA metabolic process	738	124	75.22	**1.10E-08**
	GO:0009628	response to abiotic stimulus	441	72	44.95	**3.70E-05**
	GO:0016043	cellular component organization	1855	237	189.08	**8.40E-05**
	GO:0040029	regulation of gene expression, epigeneti...	89	21	9.07	**0.00019**
	GO:0009605	response to external stimulus	253	43	25.79	**0.00055**
	GO:0006810	transport	2362	282	240.76	**0.0016**
	GO:0009719	response to endogenous stimulus	446	63	45.46	**0.0047**
	GO:0007275	multicellular organism development	572	77	58.3	**0.00664**
	GO:0006091	generation of precursor metabolites and ...	264	40	26.91	**0.00696**
	GO:0006464	cellular protein modification process	2689	310	274.09	**0.0073**
	GO:0006950	response to stress	1587	188	161.76	**0.01343**
	GO:0009791	post-embryonic development	295	42	30.07	**0.01628**
	GO:0009653	anatomical structure morphogenesis	174	27	17.74	**0.01763**
	GO:0009908	flower development	91	16	9.28	**0.02088**
	GO:0040007	growth	115	19	11.72	**0.02333**
	GO:0016049	cell growth	79	14	8.05	**0.02768**
	GO:0007049	cell cycle	410	53	41.79	**0.04168**

We also sought to find which are the regions interacting with genes, in the loops characterized as "gene-other". We found that 10% of the 874 "gene-other" loops of Pinot noir (N=87) occurred between genes and regions marked by intergenic ATAC peaks. We called this category of interactions "gene-intergenic ATAC" loops. We observed comparable results in Rkatsiteli, where the 13% of the 746 "gene-other" loops (N=97) were revealed to be "gene-intergenic ATAC" loops (*Figure 28* A).

We then simulated a set of "gene-other" loops (as described in the methods section) and

applied the same system of classification, computing for each of the 100 iterations how many interactions were classified as “gene-intergenic ATAC” loops. We found that our results were significantly different ($p<0.01$) from the simulated data (*Figure 28* B), in particular, the loops are connecting a gene and an intergenic ATAC region more often than what expected by chance both in Pinot noir and Rkatsiteli.

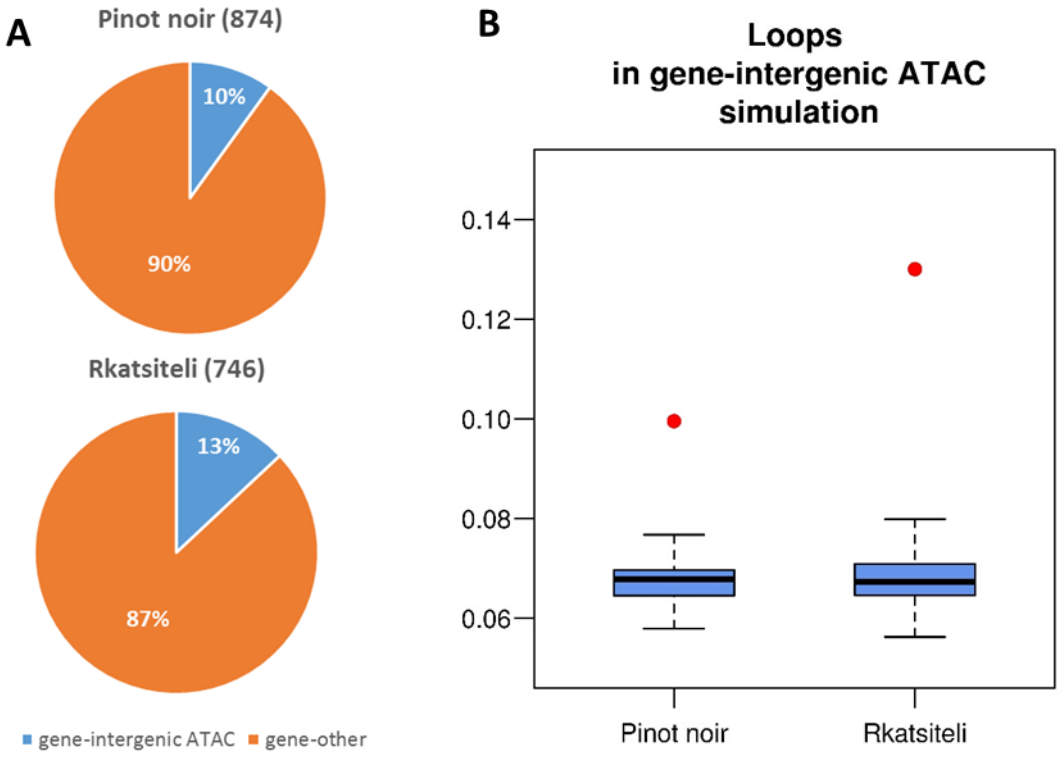


Figure 28 A: In both Pinot noir and Rkatsiteli, a small portion (10% and 13% respectively) of the “gene-other” loops was classified as “gene-intergenic ATAC”. B: Comparison between real data (red points) and a simulated dataset of “gene-intergenic ATAC loops” (blue box) in Pinot noir and Rkatsiteli.

Intergenic ATAC peaks are markers for open chromatin and are enriched at putative enhancer locations of the genome, therefore our results suggest that part of the loops we found in grapevine are involved in regulatory mechanisms of gene expression, bringing in contact genes and enhancers.

We then asked if loops could affect the expression of genes, therefore we took into account two categories of genes involved in loops with putative biological meaning, namely the genes interacting with an intergenic ATAC region and the "self-looped" genes. For Pinot noir and Rkatsiteli we compared the FPKM of the genes in these two categories with a control set composed by all the genes not belonging to either of the two categories (Figure 29). We observed that both the intergenic ATAC-interacting genes and the "self-looped" genes showed significantly higher expression levels than the control.

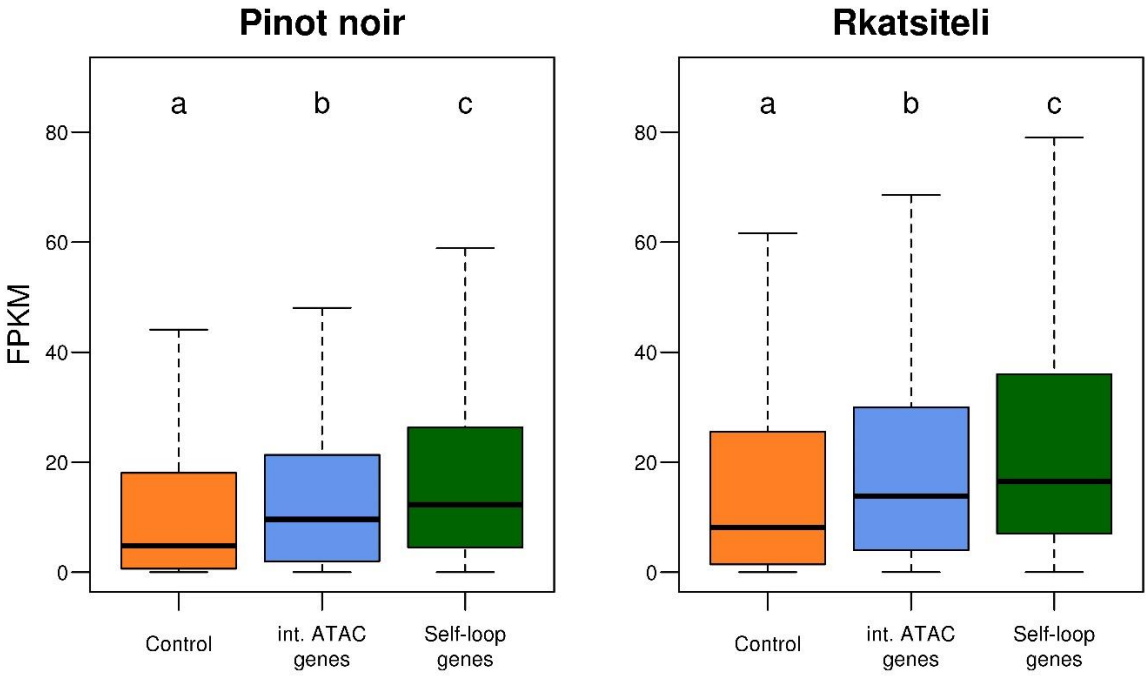


Figure 29 Expression level comparison between genes interacting with an intergenic ATAC region (int. ATAC genes), "self-looped" genes and the control set. The comparison was carried out in Pinot noir and Rkatsiteli and significant difference between the distribution was computed by pairwise Wilcoxon test (same letters indicate no significant difference; p-value <0.05).

In particular, the "self-looped" genes were the most highly expressed in both varieties. Similar results were obtained in a recent work on rice (Dong *et al.*, 2018), where "self-

looped genes" showed high levels of expression compared to a control set of genes. In that work the influence of neighbouring genes on the expression of the "self-looped" genes was also tested, observing that the expression level of these genes was dependent from the presence or absence of highly expressed neighbours (Dong *et al.*, 2018).

Loops are structural elements of the chromatin organization which can have several functional roles (Rao *et al.*, 2014). Loops bring in physical contact loci whose linear distance ranges from thousands of bases up to two megabases. There are stable looping structures, conserved across tissues and species, and loops that can constitute dynamic structures, which can be untied and reformed. Loops can be involved in gene regulation, occurring in most of the cases between a gene promoter and an enhancer, but they can also occur between the borders of sub-compartment domains, as shown for mammals (Rao *et al.*, 2014; Smith *et al.*, 2016). Since loops cause significantly higher interaction frequency than expected in pairs of distant loci, they represent outliers from the DDD function (*Figure 15*). However, artefacts in the alignment could give raise to the same kind of signal in the Hi-C contact map. For example, a deletion brings in physical proximity regions that in the reference sequence are distant apart from one another. Consequently, such regions will appear in the Hi-C matrix as a signal originated by two distant loci interacting more than any two loci at the same distance. So, the loop signal could be misinterpreted in presence of SVs. We performed our analysis focusing on genomic regions containing genes, which are less affected by the presence of SVs that could influence the results. We identified different categories of loops involving genes. Our observations suggested that loops in grapevine genome could have a role in the regulation of gene expression, directly influencing the physical contacts between genes

or different parts of the same gene and allowing the coupling between genes and enhancers.

4.2 SVs EFFECT ON GRAPEVINE CHROMATIN CONFORMATION

4.2.1 Simulation of SV presence in Hi-C contact maps

To investigate the potential of detecting effects of SVs on the chromatin conformation in the *V. vinifera* genome, we simulated large chromosomal rearrangements such as deletions, insertions and inversions. For each simulated event we built one Hi-C contact map with a resolution of 20 Kb. We performed one simulation for each of the 19 chromosome of the *V. vinifera* variety Pinot noir, obtaining 19 cases of deletion, 19 cases of insertion and 19 cases of inversions. Each simulated map was compared with a non-simulated dataset in order to visually identify the heat map pattern produced by the different possible SVs. The simulated SVs ranged in length from 5 Kb to 2 Mb. In addition to the information from the contact map, we also used the directionality index (DI) method (see methods section) to detect specific variations in the signal for each simulated event.

4.2.1.1 Deletions and insertions

In all the 19 simulated deletion events, a main and a secondary feature characterized the interaction map pattern. The main feature was the interruption of the signal on the map. This is a consequence of the lack of reads in the deleted region. If the deletion is carried by both alleles in the sample (homozygous deletion), the region between the deletion borders (D1 and D2) is completely missing in the sample, with no Hi-C reads mapping resulting in a blank space in the contact map (Figure 30 B). The secondary feature is a higher than expected signal (white arrow in Figure 30 B) at the intersection point of the deletion borders (D1-D2). This pattern was originated by two adjacent loci in the sample that are mapping at a distance greater than zero on the reference. For example, in Figure 30 B we reported a simulated deletion of 1 Mb. The points D1 and D2 are distant 1Mb on the Hi-C map, but they are adjacent (distance=0) in the sample. As a result, D1 and D2 will appear in the map as two loci sharing long range interactions.

In 17 cases out of 19 the DI showed an interruption of the signal inside the deleted region as the one reported in Figure 30 B. The DI track did not show significant variations in pattern for two simulated deletions with length 10 Kb and 5 Kb, which were also the smallest ones. We then repeated the simulation using 5 Kb resolution maps (the maximum resolution for our Hi-C dataset) in order to measure the limit for this analysis. However, even at a 5k resolution, the 5kb and 10 Kb simulated deletions could not be revealed, setting to 10 Kb the resolution limit for this analysis.

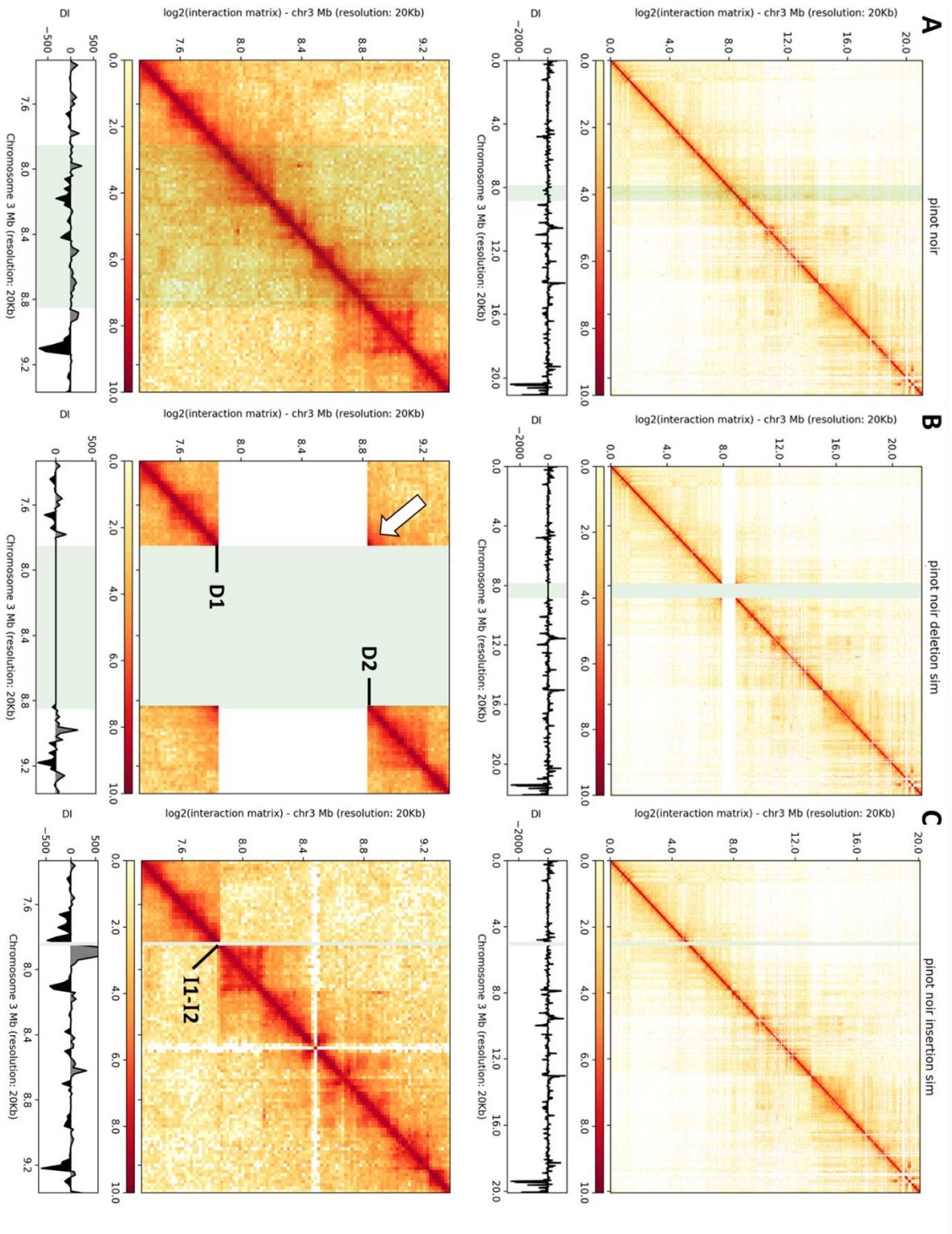


Figure 30 Contact map comparison between Pinot noir chr 3 (A), a simulated deletion (B) and a simulated insertion (C). Detail of the simulated region (chr3:7,851,201-8,851,201) is reported in the low row; the simulated region is highlighted in green.

The 19 simulated insertions showed a common contact pattern characterized by a dramatic decrease in interaction count at the insertion border (Figure 30 C). At this locus, the log2 value of the interaction matrix ranged from zero to two, while in the non-simulated maps the values in the same region ranged from eight to ten. This pattern of interaction is caused by a region that is present in the sample, but is missing in the reference. As an example, we reported a 1 Mb simulated insertion on chromosome three of Pinot noir (Figure 30 C). The borders of the insertion (I1 and I2), are at distance 1 Mb in the sample, but were mapped at distance zero on the reference. We observed less frequent interactions between I1 and I2 (close to zero) than between other adjacent loci on the map (average $\log_2$(frequency) = 9). At the same time, I1 and I2 have comparable frequency of interaction with other loci at 1 Mb distance (Figure 30 C).

For 16 out of 19 simulated insertions, the DI reported significant biased interactions at the insertion borders, consisting of negative DI values in regions immediately upstream of the insertion, followed by positive values in regions immediately downstream the insertion. Negative DI values indicated upstream biased interactions, while positive values indicated downstream bias. We measured the difference in magnitude of DI value (representing the degree of bias) across the insertion borders and found that it did not show significant correlation with the length of the 16 simulated insertions (Pearson's correlation p-value=0.1079). In two of the simulated cases the DI did not show any significant bias across the insertion borders. These were the same cases in which also the simulated deletion was not revealed, confirming the small size of the two simulated events (5 Kb and 10 Kb) was the resolution limit for this analysis.

4.2.1.2 Inversions

Comparing the 19 non-simulated maps with the 19 simulated inversions, two complementary contact patterns emerged. The first pattern was similar to the insertion pattern, consisting of a lower than expected interaction frequency between the borders of the event. Unlike the simulated insertions, the same pattern was found at both extremities of the inverted region. We reported as an example the 150 Kb inversion simulated in Pinot noir chromosome 17 (Figure 31). Due to the inversion, the points A1 and A2 were mapped at distance zero on the reference, while they were at distance 150 Kb on the sample. The same happened to the points B1 and B2. As a result, A1-A2 and B1-B2 no longer shared interactions at distance zero but at distance 150 Kb, and therefore their interaction frequency was lower than expected (between zero and two, instead of ten). The second pattern was complementary to the first one, and was revealed for all the 19 cases as a higher than expected frequency of interaction between distant point on the map. This pattern, evident as an off-diagonal point of frequent interaction, is highlighted in Figure 31. In the inverted samples A1-B1 and A2-B2 were interaction partners, since their distance was zero; while on the reference their distance was greater than zero (in this case 150 Kb), giving rise to this long-range high interaction signal outside the main diagonal (white arrow in Figure 31).

All the 19 simulations showed a characteristic DI pattern at the inversion borders. Similar to what resulted for the insertions, the DI showed a dramatic change in the bias direction at both borders of the event. However, the DI bias degree was different in the two inversion borders.

In particular, at the upstream border the DI was always positive (downstream bias); while at the downstream border the DI was always negative (upstream bias).

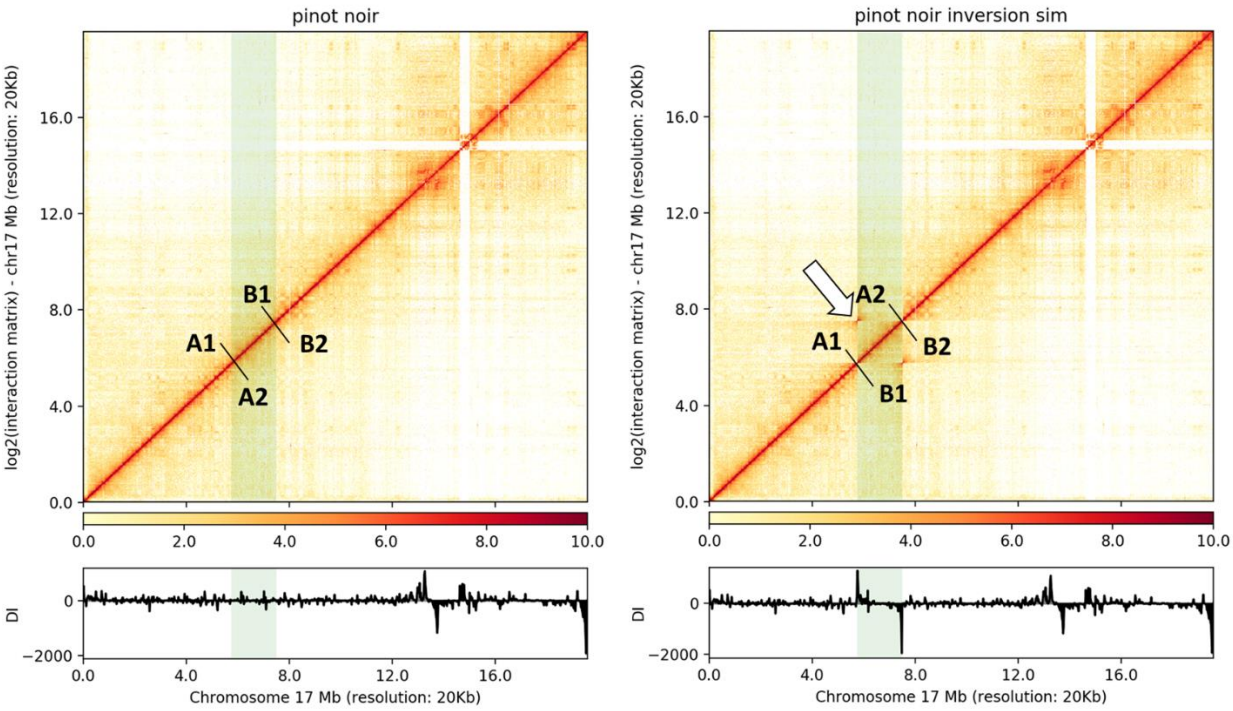


Figure 31 Comparison between normal Pinot noir chromosome 7 contact map (A) and a simulation of inversion. The region of the simulated event (chr7: 21,968,380-22,118,379) is highlighted in green.

4.2.2 Allele-specific Hi-C maps

SVs effect on chromatin conformation was evident when occurring in homozygosity, as we showed in the simulated Hi-C maps. However, when only one copy of the genome was affected by SV, the signal in the Hi-C contact map could be less obviously interpreted. This is due to the fact that in the Hi-C experiment reads from both alleles are generated; thus, if one allele is affected by SV, its contact pattern will be different from that obtained from the other allele which is not affected by SV. When building the Hi-C contact map, the contribution in

signal from both alleles is merged, and in case of heterozygous SV the variation in the contact pattern will be visible, but only half of the signal will follow such pattern.

In order to investigate the effects of structural variation on the chromatin conformation at single allele level, we sought to generate allele-specific Hi-C maps. In this analysis, we used the Rkatsiteli Hi-C dataset, since the phased haplotype data were readily available from a previous work done in our research group (Alice Fornasiero, PhD thesis, 2017).

The allele-specific Hi-C maps were reconstructed for Rkatsiteli Haplotype A and Haplotype B. In the genome-wide contact maps (Figure 32), the translocation event between chromosomes 1 and 11 (highlighted by a circle in Figure 32) is clearly visible in the Haplotype A map, but not in the Haplotype B map, confirming that the reciprocal translocation is present in heterozygous condition.

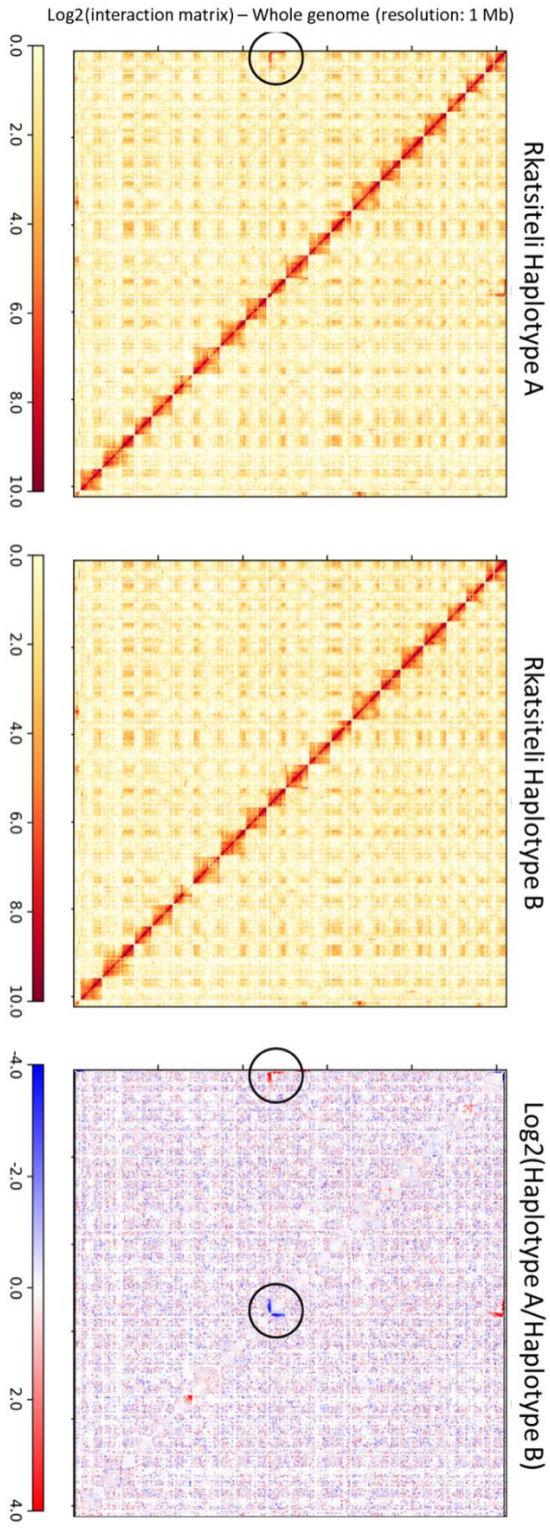


Figure 32 Whole genome maps comparison between Rkatsiteli Haplotype A and B. The main feature emerging is the presence of a translocation between chr1 and chr11 (in the circles). Maps are plotted at 1 Mb resolution.

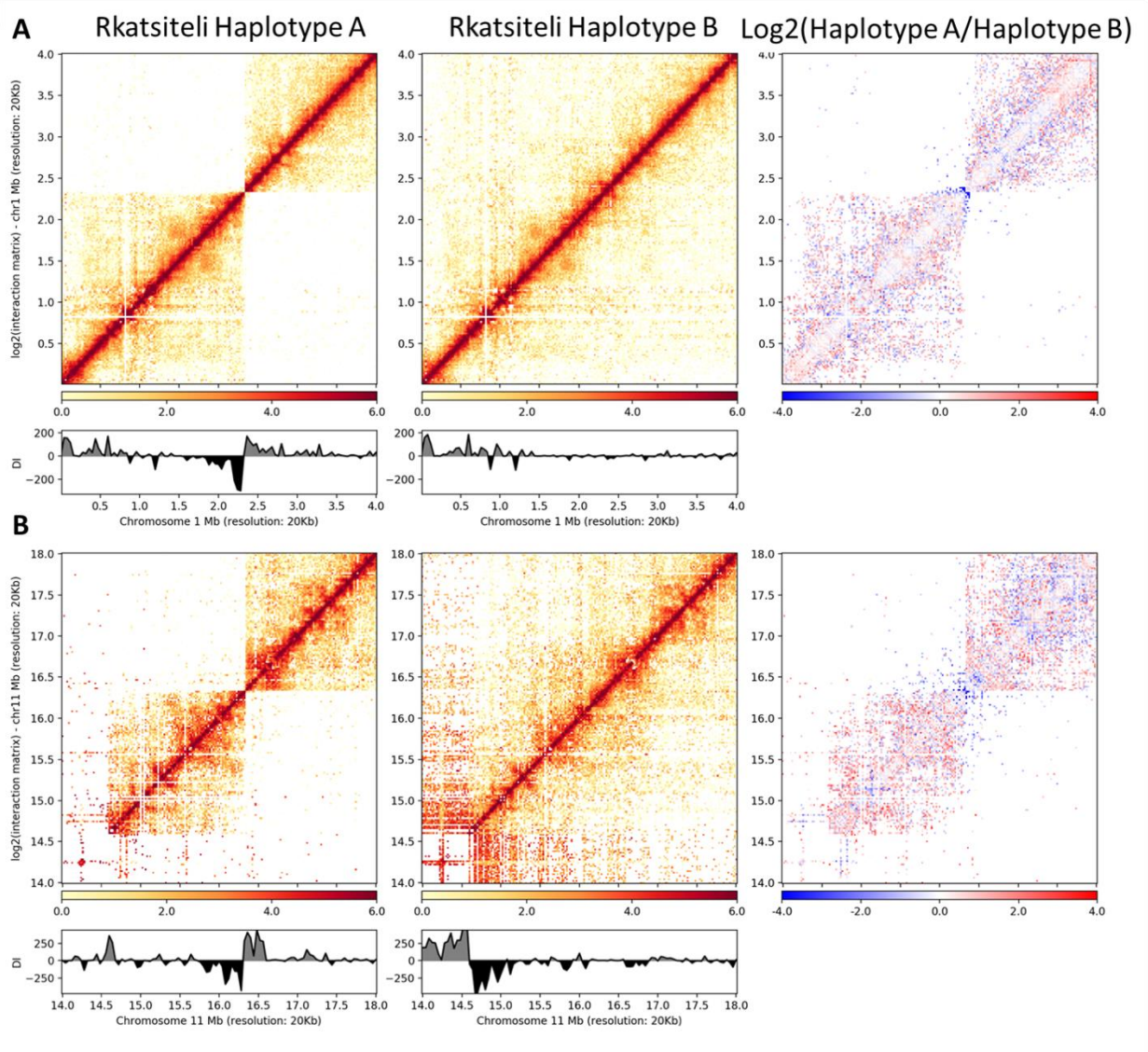


Figure 33 Hi-C contact maps of chromosome 1 (A) and chromosome 11 (B) in the two Rkatsiteli haplotypes. The maps show the detail of the chr1-chr11 translocation borders that are evident in Rkatsiteli Haplotype A, while in Haplotype B both the contact map and the DI track are showing regular patterns of interaction.

A detailed inspection of the region containing the translocation borders in chromosome 1 and 11, revealed the already characterized breakpoints of the translocation at chr1: 2,341,643 and chr11: 16,496,900 (Alice Fornasiero, PhD thesis, 2017) in the Rkatsiteli Haplotype A maps (*Figure 33*). The variation in the contact pattern between the translocated allele and the non-translocated one was also reflected by the DI computed for the two maps. In chr1 of the

Rkatsiteli Haplotype A, the DI track showed a break in the contact pattern around 2.3 Mb. The bins up to the breakpoint showed biased contacts with upstream regions, while bins after the breakpoint showed biased interactions with downstream regions. The DI track of chr1 in Haplotype B did not show any bias in the direction of interactions near the same point of coordinates (*Figure 33* A). This break in Haplotype A was due to the fact that the two regions were mapped at distance 0 on the PN40024 reference, but were on different chromosomes in the Rkatsiteli sample; thus no interaction occurred across the breakpoint region in Rkatsiteli Haplotype A. The same effect was observed on chromosome 11 at around 16.3 Mb (*Figure 33* B). Notably, another DI bias was revealed at 14.6 Mb (*Figure 33* B), although this is likely due to a large stretch of homozygosity. In fact, since the method we applied relies on SNPs presence for the allele-specific assignment of the Hi-C reads, regions with high degree of homozygosity will lack diagnostic SNPs and might appear as regions with absence of mapped reads (deletions).

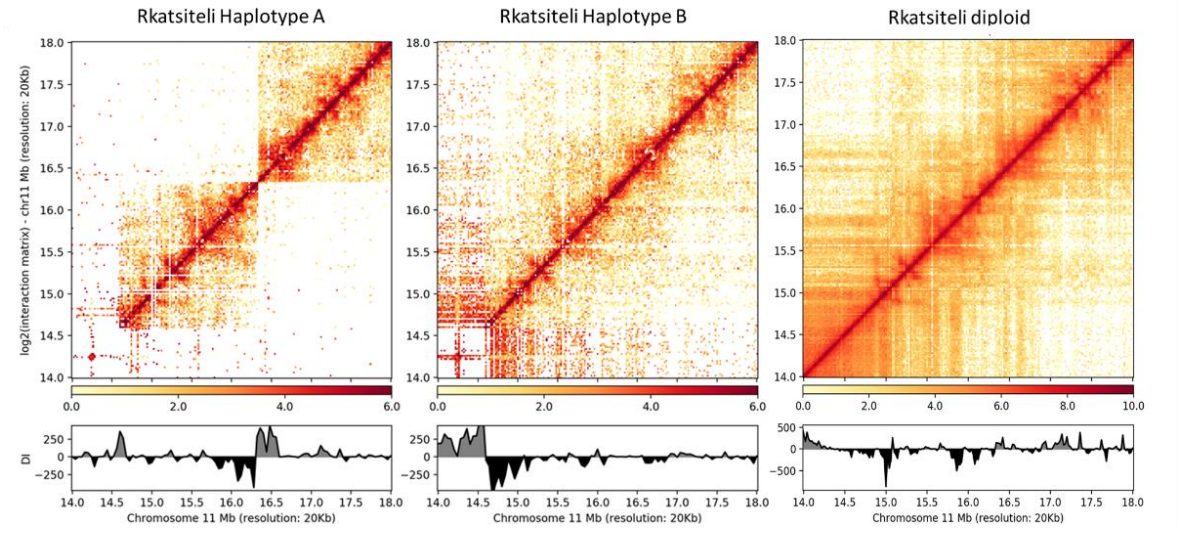


Figure 34 Comparison between the allele-specific maps and the diploid map of Rkatsiteli at the translocation breakpoint in chr11. The signal present in Haplotype A and B at 14.0-14.6 Mb is absent in the diploid map

The discrimination between a stretch of homozygosity and an actual deletion can be easily obtained by integrating the information of the allele-specific Hi-C maps and of the presence of heterozygous SNPs, or by comparing allele-specific Hi-C maps with diploid Hi-C maps (Figure 34). In addition, we reported a novel case of heterozygous SV, consisting of an inversion event at chr7:11.9-12.9 Mb (*Figure 35*). The inversion was seen only in the Rkatsiteli Haplotype A map while no sign of a similar interaction pattern was revealed in the Haplotype B map. The DI track reported biased interactions at the inversion borders with the characteristic profile identified in Figure 31. We also showed the diploid version of the chr7 contact map in the same region. As expected, the inversion contact pattern presented a diluted signal in the diploid map due to the averaged Hi-C data from both alleles (*Figure 35* C). In addition, the haplotype B DI track although showing peaks of interaction bias, was not showing a fully recognizable inversion pattern as for the DI track in Rkatsiteli haplotype A (*Figure 35* A)

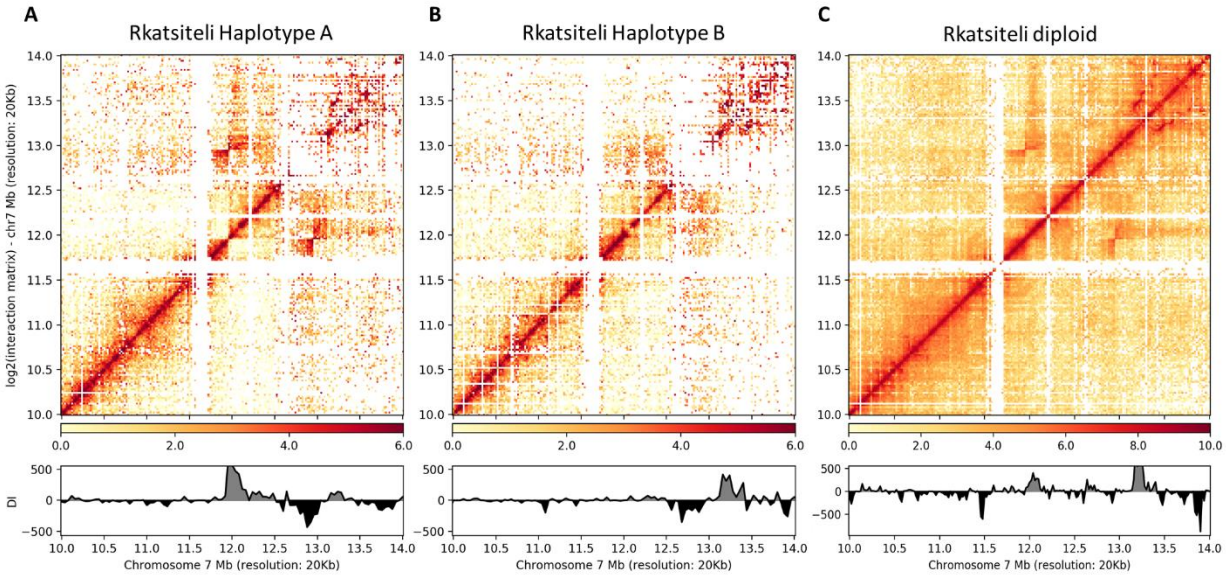


Figure 35 Detail of the heterozygous inversion found in Rkatsiteli chromosome 7 between 11.9 and 12.9 Mb. From the allele-specific maps it is evident that one of the two haplotypes carried the inversion (A and B), while the dilution effect on the inversion signal in the diploid map is evident (C).

4.3 SPATIAL PROXIMITY IS A NECESSARY CONDITION FOR SV OCCURRENCE

We used grapevine (*Vitis vinifera*), a species with high levels of structural variation, to investigate if SV occur in regions that tend to be in physical proximity in the nucleus, meaning that physical contact is a prerequisite for SV occurrence.

CNVs are one of the most common types of SV, in which different individuals possess the same DNA sequence in a different number of copies. In particular, we considered cases in which the CNV resulted in a deletion in the analysed individuals in comparison to the reference sequence, and led to changes in chromosome structure, creating a junction between two formerly separated DNA sequences.

We compared seventy-three manually annotated regions with verified exact borders location and a size range between 4,000 and 600,000 bp, corresponding to homozygous deletions either in Pinot noir or Rkatsiteli varieties compared to the reference. In such set of regions one variety was homozygous for the deletion (CNV-present) and the other was homozygous for the reference allele (CNV-absent).

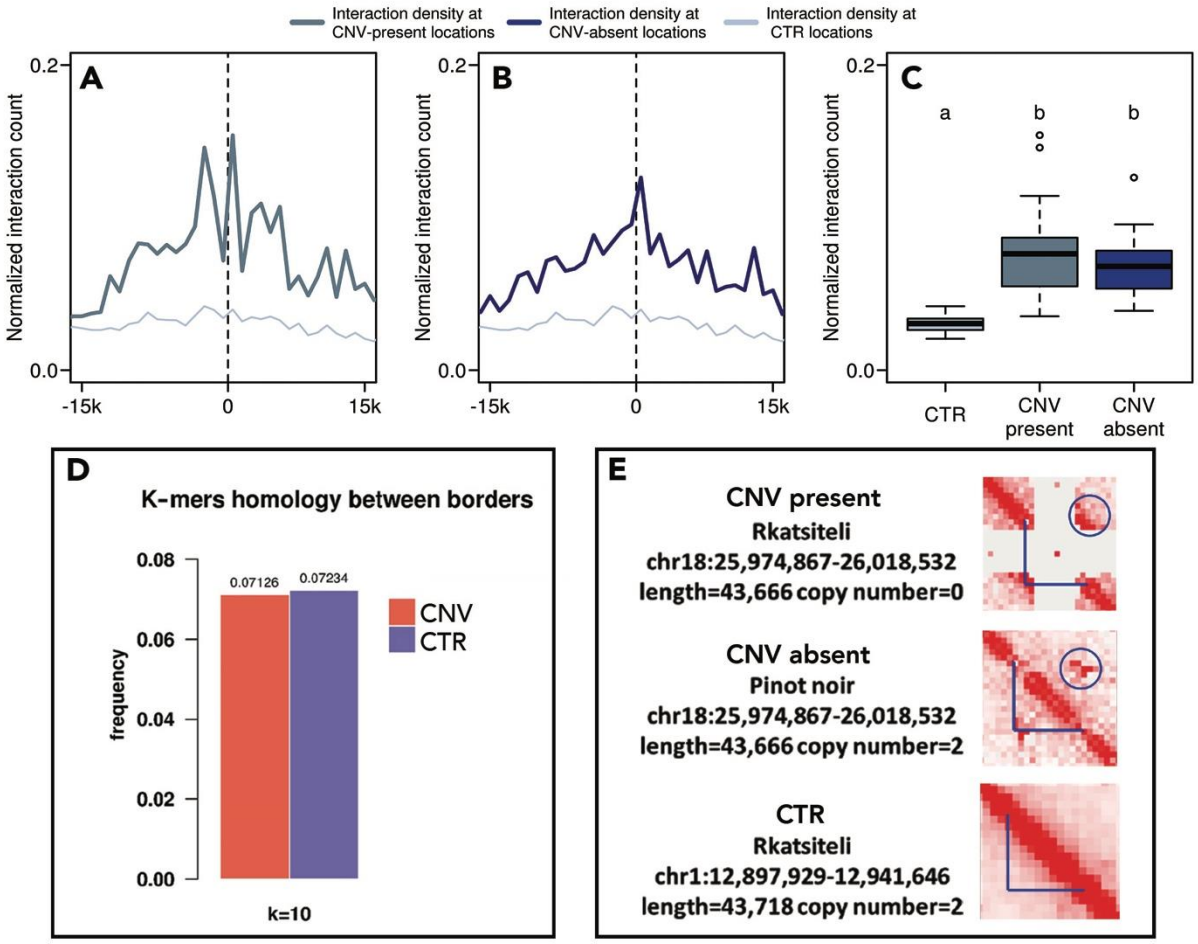


Figure 36 A,B: Interaction count across the border (dashed line) of CNV-present, CNV-absent and CTR regions in the range of 15 Kb upstream and downstream the border. C: Comparison between the distributions of interaction counts for CTR, CNV-present and CNV-absent. Significant differences in distribution assessed via pairwise Wilcoxon test: different letters = significant difference (p-value < 0.05). Interaction counts were normalised by coverage. D: Comparison between the frequency of common 10-mers found across the borders of CNV and CTR regions showed no significant difference (chi-square test). E: Hi-C contact maps for a trio of CNV-present, CNV-absent and CTR regions. Colour intensity is proportional to the interaction frequency between loci; blue lines define the region of interest and the physical contact occurring between each region extremity are indicated by a blue circle.

We compared the interaction frequency across the border of CNV-present and CNV-absent against the interaction frequency across the border of a set of control regions (CTR).

As expected, we observed high interaction counts across the borders for CNV-present regions (Figure 36 A), because the deletion brings in physical proximity regions that in the reference sequence are distant apart from one another. The unexpected result was the observation of

a high interaction frequency across the borders of the same regions even when the CNV is not present (Figure 36 B). Moreover, there was no significant difference (pairwise Wilcoxon's p-value > 0.05) in interaction levels across the borders of CNV-present and CNV-absent regions (Figure 36 C).

A k-mers analysis of both the CNV and CTR borders (Figure 36 D) confirmed that the difference in interaction counts between CTR and CNV was not due to mapping bias, since no significant difference in frequency of identical 10-mers was found (chi-square p-value > 0.01).

We show that the presence of CNV can be visually confirmed in the Hi-C interaction matrix, with increased signal seen across the CNV borders with respect to the surrounding regions. This signal increase is also observed in the CNV-absent variety (Figure 36 E).

Our results indicate that the borders of CNV-absent regions, still having the same distance distribution as the CTR regions, showed similar interaction frequencies as CNV-present, whose borders have distance zero. Thus, CNV-absent borders are distant on the linear DNA strand, but are in physical contact in the 3D chromatin conformation.

Sequence variation among individuals of the same species is partly due to large insertions or deletions (SVs) that can derive from the movement of transposable elements or from defective repair of double strand breaks (DSBs) through Non Homologous End Joining (NHEJ) or unequal Homologous Recombination (HR). In fact, the several DSBs repair mechanisms have different potentials of introducing errors in the DNA code (San Filippo, Sung and Klein, 2008; Shrivastav, De Haro and Nickoloff, 2008; Lieber, 2010). The DSBs repair process can have two outcomes: the restoration of the DNA sequence or can cause genome variability, giving rise to base conversion, inversions, insertions, deletions and translocations (Schubert *et al.*, 2004).

The HR mechanism repairs the DSB during the S and the G2 phase of the cell cycle. It uses the intact sister chromatid as template, resulting in an error-free restoration of the DNA sequence. But in some cases, the HR mechanism may also use homologous sequences as template, deriving from the homologous chromosome or from non-homologous chromosomes, resulting in an error-prone repairing (Puchta, 2005; Heyer, Ehmsen and Liu, 2010; Jasin and Rothstein, 2013). In a particular case, the HR mechanism can result into a non-conservative single strand annealing, which may introduce large deletions in the genome (Jasin and Rothstein, 2013).

The NHEJ mechanism can be activated in every phase of the cell cycle, but mainly occurs during the S1 phase, when the HR is not available (Lieber, 2010). The NHEJ mechanism is less conservative than the HR, but if the ends of the DSB are preserved form nucleotide loss or gain, the NHEJ can restore the pre-break status by direct ligation (Lin, Wilson and Lin, 2013). However, in most of the cases the DSBs is accompanied by loss or gain of nucleotides, and the direct ligation occurring during the NHEJ could give rise to deletions or insertions. This is the case of the “canonical” or “classic” NHEJ (c-NHEJ; (Deriano and Roth, 2013)). In case of microhomology (2-25 bp) between the ends of the DSB, the DNA strands can anneal and then be ligated, giving rise to both insertions or deletions of variable size; such mechanism is called “alternative” NHEJ (a-NHEJ; (Deriano and Roth, 2013; Pannunzio *et al.*, 2014; Pannunzio, Watanabe and Lieber, 2018)). If two unrelated DSB ends are joined together, the a-NHEJ mechanisms could be the source of chromosomal aberrations such as the translocations (Schubert *et al.*, 2004; Deriano and Roth, 2013; Vu *et al.*, 2014).

SV can affect the structure and the regulation of genes, giving rise to either disadvantageous or favorable traits (such as resistance to stress, pathogens or chemicals). In conclusion, our

results point to a physical interaction as a prerequisite for the occurrence of SVs and provide evidence that the three-dimensional organization of a genome can have a dramatic effect not only on the functioning of the genome but also on its structure and variation.

5 CONCLUSIONS

The present work is the first attempt in characterizing *Vitis vinifera* genome investigating its 3D structure. In particular, we focused on the interplay between structure and function in DNA regulation. We used the Hi-C method on three grapevine varieties Pinot noir, Rkatsiteli and Chardonnay to assess the different levels of chromatin organization which characterise the grapevine 3D genome. We observed the stability across varieties of structural features such as distance dependent decay of interactions, relative chromosome positioning inside the nucleus and polarization of telomeres and centromeres as in the Rabl conformation. We also observed that the grapevine genome is organized into physical compartments, namely A and B, which divide the core of the nucleus from the nuclear periphery. Such nuclear compartmentalization is a conserved feature of the grapevine genome and tends to vary more between different tissues than across varieties. A/B compartments are not only structures in which chromatin is organized, but they have functional implications. In fact, the definition of A/B compartments is the result of the interplay between chromatin structure and local genomic and epigenomic features. Chromatin in the A compartment showed enrichment for active transcription of genes, while the chromatin in the B compartment constitutes a more transcription-repressive environment. Moreover, genes in the A compartment showed less variability in expression levels than genes in the B compartment, pointing to a difference in regulatory pathways occurring in the two compartments, offering a point of further investigation on the differential composition in the classes of genes.

We explored the relationship between structure and function at a higher degree of

resolution, investigating the presence of sub-compartmental units of organization. We observed the presence of sub-compartment domains from the analysis of the Hi-C data. However, we were not able to assess a functional characterization of such domains. We did not obtain direct evidence for the existence of such domains in grapevine genome in the way they were characterized in mammalian genomes, in which CTCF/cohesin binding sites constitute a consensus for TAD boundaries across the cell population. Instead, our results were comparable with the observations obtained in other plant genomes, in which the sub-compartment domains are not a prominent structural feature. We hypothesize that the lack of a consensus for TAD boundary locations in plant genomes results into a probability distribution for domain boundaries formation, which cannot be resolved with a Hi-C experiment over a cell population, but needs a single cell Hi-C strategy to be assessed.

We observed a further class of chromatin architectural elements in the grapevine genome, which are long-range interactions, referred to as chromatin loops. Our results suggested that loops can constitute a network of interactions between different genes and can also occur inside the same gene. Moreover, we observed that loops can bring in physical contact genes and enhancers, suggesting that loops in grapevine genome are involved in the regulation of gene expression.

We also investigated the relationship between intra-species variability and chromatin 3D structure. In particular, we simulated the effect of structural variants (deletions, insertions and inversions) on the 3D conformation of the genome, observing changes in the interaction pattern between SV affected and control datasets. We also observed allele-specific effects of heterozygous structural variation, confirming with the Hi-C data

a translocation between chromosome 1 and 11 in one of the two haplotypes of the Rkatsiteli variety; moreover, we reported a novel case of heterozygous inversion on the chromosome 7 of Rkatsiteli. We also investigated the possible effect of chromatin conformation on the occurrence of SVs. We performed a comparative analysis between Pinot noir and Rkatsiteli varieties across a set of regions where one variety was homozygous for the deletion and the other was homozygous for the presence of the deleted segment. Our results pointed to physical interaction as a prerequisite for the occurrence of SVs. In this perspective, chromatin conformation can be a key role player in events from which variation derives, such as movement of transposable elements, defective repair of double strand breaks through Non Homologous End Joining or unequal Homologous Recombination.

6 BIBLIOGRAPHY

Akdemir, K. C. and Chin, L. (2015) 'HiCPlotter integrates genomic data with interaction matrices', *Genome Biology*, 16(1). doi: 10.1186/s13059-015-0767-1.

Alexa, A. and Rahnenfuhrer, J. (2016) *topGO*, *Alexa A and Rahnenfuhrer J (2016). topGO: Enrichment Analysis for Gene Ontology. R package version 2.28.0.* doi: http://dx.doi.org/10.1136/jech-2013-202820.

Alkan, C., Coe, B. P. and Eichler, E. E. (2011) 'Genome structural variation discovery and genotyping', *Nature Reviews Genetics*, 12(5), pp. 363–376. doi: 10.1038/nrg2958.

Anamthawat-Jónsson, K. *et al.* (1990) 'Discrimination between closely related Triticeae species using genomic DNA as a probe', *Theoretical and Applied Genetics*, 79(6), pp. 721–728. doi: 10.1007/BF00224236.

Armstrong, S. J., Franklin, F. C. and Jones, G. H. (2001) 'Nucleolus-associated telomere clustering and pairing precede meiotic chromosome synapsis in Arabidopsis thaliana.', *Journal of Cell Science*, 114(Pt 23), pp. 4207–4217. Available at: http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?dbfrom=pubmed&id=11739653&retmode=ref&cmd=prlinks%5Cnpapers3://publication/uuid/A28AE7C9-4143-4E0D-BF7F-025719432A5B.

Bonev, B. *et al.* (2017) 'Multiscale 3D Genome Rewiring during Mouse Article Multiscale 3D Genome Rewiring during Mouse Neural Development', *Cell*, 171(3), p. 557.e1-557.e24. doi: 10.1016/j.cell.2017.09.043.

Boveri, T. (2008) 'Concerning the Origin of Malignant Tumours by Theodor Boveri. Translated and annotated by Henry Harris', *Journal of Cell Science*, 121(Supplement 1), pp. 1–84. doi: 10.1242/jcs.025742.

Boyle, S. (2001) 'The spatial organization of human chromosomes within the nuclei of

normal and emerin-mutant cells', *Human Molecular Genetics*, 10(3), pp. 211–219. doi: 10.1093/hmg/10.3.211.

Brunner, S. (2005) 'Evolution of DNA Sequence Nonhomologies among Maize Inbreds', *the Plant Cell Online*, 17(2), pp. 343–360. doi: 10.1105/tpc.104.025627.

Butelli, E. *et al.* (2012) 'Retrotransposons Control Fruit-Specific, Cold-Dependent Accumulation of Anthocyanins in Blood Oranges', *The Plant Cell*, 24(3), pp. 1242–1255. doi: 10.1105/tpc.111.095232.

Carty, M. *et al.* (2017) 'An integrated model for detecting significant chromatin interactions from high-resolution Hi-C data', *Nature Communications*, 8(May), p. 15454. doi: 10.1038/ncomms15454.

Cavalli, G. and Misteli, T. (2013) 'Functional implications of genome topology', *Nature Structural & Molecular Biology*. Nature Publishing Group, 20(3), pp. 290–299. doi: 10.1038/nsmb.2474.

Chong, S. *et al.* (2018a) 'Imaging dynamic and selective low-complexity domain interactions that control gene transcription.', *Science (New York, N.Y.)*, 361(6400), p. eaar2555. doi: 10.1126/science.aar2555.

Chong, S. *et al.* (2018b) 'Imaging dynamic and selective low-complexity domain interactions that control gene transcription.', *Science (New York, N.Y.)*, 361(6400), p. eaar2555. doi: 10.1126/science.aar2555.

Ciabrelli, F. and Cavalli, G. (2015) 'Chromatin-driven behavior of topologically associating domains', *Journal of Molecular Biology*. Elsevier B.V., 427(3), pp. 608–625. doi: 10.1016/j.jmb.2014.09.013.

Cowan, C. R., Carlton, P. M. and Cande, W. Z. (2001) 'The Polar Arrangement of Telomeres in Interphase and Meiosis. Rabl Organization and the Bouquet', *Plant Physiol*, (125), pp. 532–538.

Crane, E. *et al.* (2015) 'Condensin-driven remodelling of X chromosome topology during dosage compensation', *Nature*, 523(7559), pp. 240–244. doi: 10.1038/nature14450.

Cremer, T. *et al.* (1982) 'Analysis of chromosome positions in the interphase nucleus of Chinese hamster cells by laser-UV-microirradiation experiments', *Human Genetics*, 62(3), pp. 201–209. doi: 10.1007/BF00333519.

Cremer, T. and Cremer, C. (2006) 'Rise, fall and resurrection of chromosome territories: A historical perspective. Part I. The rise of chromosome territories', *European Journal of Histochemistry*, pp. 161–176. doi: 10.4081/989.

Cremer, T. and Cremer, M. (2010) 'Chromosome territories.', *Cold Spring Harbor perspectives in biology*, 2(3), pp. 1–22. doi: 10.1101/cshperspect.a003889.

Croft, J. A. *et al.* (1999) 'Differences in the localization and morphology of chromosomes in the human nucleus', *Journal of Cell Biology*, 145(6), pp. 1119–1131. doi: 10.1083/jcb.145.6.1119.

Dekker, J. *et al.* (2002) 'Capturing chromosome conformation.', *Science (New York, N.Y.)*, 295(5558), pp. 1306–11. doi: 10.1126/science.1067799.

Dekker, J. *et al.* (2017) 'The 4D nucleome project', *Nature*, pp. 219–226. doi: 10.1038/nature23884.

Dekker, J. and Heard, E. (2015) 'Structural and functional diversity of Topologically Associating Domains', *FEBS Letters*. Federation of European Biochemical Societies, 589(20), pp. 2877–2884. doi: 10.1016/j.febslet.2015.08.044.

Dekker, J., Marti-Renom, M. A. and Mirny, L. A. (2013) 'Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data', *Nature Reviews Genetics*, 14(6), pp. 390–403. doi: 10.1038/nrg3454.

Dekker, J. and Mirny, L. (2016) 'The 3D Genome as Moderator of Chromosomal Communication', *Cell*. Elsevier Ltd, 164(6), pp. 1110–1121. doi: 10.1016/j.cell.2016.02.007.

Dekker, J. and Misteli, T. (2015) 'Long-Range Chromatin Interactions. - PubMed - NCBI'. Available at: http://www.ncbi.nlm.nih.gov/pubmed/26430217.

Deng, W. *et al.* (2014) 'Reactivation of developmentally silenced globin genes by forced

chromatin looping', *Cell*, 158(4), pp. 849–860. doi: 10.1016/j.cell.2014.05.050.

Deriano, L. and Roth, D. B. (2013) 'Modernizing the Nonhomologous End-Joining Repertoire: Alternative and Classical NHEJ Share the Stage', *Annual Review of Genetics*. doi: 10.1146/annurev-genet-110711-155540.

Dixon, J. R. *et al.* (2012) 'Topological domains in mammalian genomes identified by analysis of chromatin interactions', *Nature*. Nature Publishing Group, 485(7398), pp. 376–380. doi: 10.1038/nature11082.

Dixon, J. R. *et al.* (2015) 'Chromatin architecture reorganization during stem cell differentiation', *Nature*, 518(7539), pp. 331–336. doi: 10.1038/nature14222.

Dong, F. and Jiang, J. (1998) 'Non-Rabl patterns of centromere and telomere distribution in the interphase nuclei of plant cells', *Chromosome Research*, 6(7), pp. 551–558. doi: 10.1023/A:1009280425125.

Dong, P. *et al.* (2017) '3D chromatin architecture of large plant genomes determined by local A/B compartments', *Molecular Plant*. Elsevier Ltd, 10(12), pp. 1497–1509. doi: 10.1016/j.molp.2017.11.005.

Dong, Q. *et al.* (2018) 'Genome-wide Hi-C analysis reveals extensive hierarchical chromatin interactions in rice', *Plant Journal*, 94(6), pp. 1141–1156. doi: 10.1111/tpj.13925.

Dostie, J. *et al.* (2006) 'Chromosome Conformation Capture Carbon Copy (5C): A massively parallel solution for mapping interactions between genomic elements', *Genome Research*, 16(10), pp. 1299–1309. doi: 10.1101/gr.5571506.

Duan, Z. *et al.* (2010) 'A three-dimensional model of the yeast genome', *Nature*, 465(7296), pp. 363–367. doi: nature08973 [pii]\n10.1038/nature08973.

Duan, Z. *et al.* (2012) 'A genome-wide 3C-method for characterizing the three-dimensional architectures of genomes', *Methods*, pp. 277–288. doi: 10.1016/j.ymeth.2012.06.018.

Durand, N. C., Robinson, J. T., *et al.* (2016) 'Juicebox Provides a Visualization System for Hi-C Contact Maps with Unlimited Zoom', *Cell Systems*, 3(1), pp. 99–101. doi:

10.1016/j.cels.2015.07.012.

Durand, N. C., Shamim, M. S., *et al.* (2016) 'Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments', *Cell Systems*, 3(1), pp. 95–98. doi: 10.1016/j.cels.2016.07.002.

Durinck, S. *et al.* (2005) 'BioMart and Bioconductor: A powerful link between biological databases and microarray data analysis', *Bioinformatics*, 21(16), pp. 3439–3440. doi: 10.1093/bioinformatics/bti525.

Durinck, S. *et al.* (2009) 'Mapping identifiers for the integration of genomic datasets with the R/ Bioconductor package biomaRt', *Nature Protocols*, 4(8), pp. 1184–1191. doi: 10.1038/nprot.2009.97.

Eichten, S. R. *et al.* (2011) 'B73-Mo17 Near-Isogenic Lines Demonstrate Dispersed Structural Variation in Maize', *PLANT PHYSIOLOGY*, 156(4), pp. 1679–1690. doi: 10.1104/pp.111.174748.

Del Fabbro, C. *et al.* (2013) 'An extensive evaluation of read trimming effects on illumina NGS data analysis', *PLoS ONE*, 8(12). doi: 10.1371/journal.pone.0085024.

Felsenfeld, G. and Groudine, M. (2003) 'Controlling the double helix', *Nature*, 421(6921), pp. 448–453. doi: 10.1038/nature01411.

Feng, S. *et al.* (2014) 'Genome-wide Hi-C Analyses in Wild-Type and Mutants Reveal High-Resolution Chromatin Interactions in Arabidopsis', *Molecular Cell*, 55(5), pp. 694–707. doi: 10.1016/j.molcel.2014.07.008.

Feuk, L., Carson, A. and Scherer, S. (2006) 'Structural variation in the human genome.', *Nat Rev Genet*, 7(2), pp. 85–97. doi: 10.1038/nrg1767.

Finlan, L. E. *et al.* (2008) 'Recruitment to the nuclear periphery can alter expression of genes in human cells', *PLoS Genetics*, 4(3). doi: 10.1371/journal.pgen.1000039.

Flyamer, I. M. *et al.* (2017) 'Single-nucleus Hi-C reveals unique chromatin reorganization at oocyte-to-zygote transition', *Nature*. Nature Publishing Group, 544(7648), pp. 110–114. doi:

10.1038/nature21711.

Fransz, P. *et al.* (2002) 'Interphase chromosomes in Arabidopsis are organized as well defined chromocenters from which euchromatin loops emanate', *Proceedings of the National Academy of Sciences*, 99(22), pp. 14584–14589. doi: 10.1073/pnas.212325299.

Fraser, J. *et al.* (2015) 'An Overview of Genome Organization and How We Got There: from FISH to Hi-C', *Microbiology and Molecular Biology Reviews*, 79(3), pp. 347–372. doi: 10.1128/MMBR.00006-15.

Fudenberg, G. *et al.* (2016) 'Formation of Chromosomal Domains by Loop Extrusion', *Cell Reports*. The Author(s), 15(9), pp. 2038–2049. doi: 10.1016/j.celrep.2016.04.085.

Fudenberg, G. and Mirny, L. A. (2012) 'Higher-order chromatin structure: Bridging physics and biology', *Current Opinion in Genetics and Development*, pp. 115–124. doi: 10.1016/j.gde.2012.01.006.

Gaines, T. A. *et al.* (2010) 'Gene amplification confers glyphosate resistance in Amaranthus palmeri', *Proceedings of the National Academy of Sciences*, 107(3), pp. 1029–1034. doi: 10.1073/pnas.0906649107.

Gerlich, D. *et al.* (2003) 'Global chromosome positions are transmitted through mitosis in mammalian cells', *Cell*, 112(6), pp. 751–764. doi: 10.1016/S0092-8674(03)00189-2.

Goff, S. A. *et al.* (2002) 'A Draft Sequence of the Rice Genome (Oryza sativa L . ssp .)', *Science*, 296(5565), pp. 92–100. doi: 10.1126/science.1068037.

Grasser, F. *et al.* (2008) 'Replication-timing-correlated spatial chromatin arrangements in cancer and in primate interphase nuclei.', *Journal of cell science*, 121(Pt 11), pp. 1876–1886. doi: 10.1242/jcs.026989.

Griffith, J., Hochschild, A. and Ptashne, M. (1986) 'DNA loops induced by cooperative binding of lambda repressor.', *Nature*, 322(6081), pp. 750–2. doi: 10.1038/322750a0.

Grob, S. and Grossniklaus, U. (2017) 'Chromosome conformation capture-based studies reveal novel features of plant nuclear architecture', *Current Opinion in Plant Biology*.

Elsevier Ltd, 36, pp. 149–157. doi: 10.1016/j.pbi.2017.03.004.

Grob, S., Schmid, M. and Grossniklaus, U. (2014) 'Hi-C Analysis in Arabidopsis Identifies the KNOT, a Structure with Similarities to the flamenco Locus of Drosophila', *Molecular Cell*. Elsevier Inc., 55(5), pp. 678–693. doi: 10.1016/j.molcel.2014.07.009.

Grossniklaus, U. and Paro, R. (2014) 'Transcriptional Silencing by Polycomb-Group Proteins', *Cold Spring Harbor Perspectives in Biology*, 6(11), pp. 1–26. doi: 10.1101/cshperspect.a019331.

Hadjur, S. *et al.* (2009) 'Cohesins form chromosomal cis-interactions at the developmentally regulated IFNG locus', *Nature*. doi: 10.1038/nature08079.

Harewood, L. *et al.* (2017) 'Hi-C as a tool for precise detection and characterisation of chromosomal rearrangements and copy number variation in human tumours', *Genome Biology*, 18(1). doi: 10.1186/s13059-017-1253-8.

Harper, L. (2004) 'A bouquet of chromosomes', *Journal of Cell Science*, 117(18), pp. 4025–4032. doi: 10.1242/jcs.01363.

Hastings, P. J. *et al.* (2009) 'Mechanisms of change in gene copy number', *Nature Reviews Genetics*, 10(8), pp. 551–564. doi: 10.1038/nrg2593.

Heinz, S. *et al.* (2010) 'Simple Combinations of Lineage-Determining Transcription Factors Prime cis-Regulatory Elements Required for Macrophage and B Cell Identities', *Molecular Cell*, 38(4), pp. 576–589. doi: 10.1016/j.molcel.2010.05.004.

Heride, C. *et al.* (2010) 'Distance between homologous chromosomes results from chromosome positioning constraints', *Journal of Cell Science*, 123(23), pp. 4063–4075. doi: 10.1242/jcs.066498.

Heyer, W.-D., Ehmsen, K. T. and Liu, J. (2010) 'Regulation of Homologous Recombination in Eukaryotes', *Annual Review of Genetics*. doi: 10.1146/annurev-genet-051710-150955.

Hou, C. *et al.* (2012) 'Gene Density, Transcription, and Insulators Contribute to the Partition of the Drosophila Genome into Physical Domains', *Molecular Cell*, 48(3), pp. 471–484. doi:

10.1016/j.molcel.2012.08.031.

Hou, C. and Corces, V. G. (2012) 'Throwing transcription for a loop: Expression of the genome in the 3D nucleus', *Chromosoma*, pp. 107–116. doi: 10.1007/s00412-011-0352-7.

Hsieh, T. H. S. *et al.* (2015) 'Mapping Nucleosome Resolution Chromosome Folding in Yeast by Micro-C', *Cell*, 162(1), pp. 108–119. doi: 10.1016/j.cell.2015.05.048.

Hurles, M. E., Dermitzakis, E. T. and Tyler-Smith, C. (2008) 'The functional impact of structural variation in humans.', *Trends in genetics : TIG*, 24(5), pp. 238–45. doi: 10.1016/j.tig.2008.03.001.

Imakaev, M. *et al.* (2012) 'Iterative correction of Hi-C data reveals hallmarks of chromosome organization', *Nature Methods*, 9(10), pp. 999–1003. doi: 10.1038/nmeth.2148.

Iwasaki, Y. W., Siomi, M. C. and Siomi, H. (2015) 'PIWI-Interacting RNA: Its Biogenesis and Functions', *Annual Review of Biochemistry*, 84(1), pp. 405–433. doi: 10.1146/annurev-biochem-060614-034258.

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

Jasin, M. and Rothstein, R. (2013) 'Repair of strand breaks by homologous recombination', *Cold Spring Harbor Perspectives in Biology*. doi: 10.1101/cshperspect.a012740.

Jin, F. *et al.* (2013) 'A high-resolution map of the three-dimensional chromatin interactome in human cells', *Nature*. doi: 10.1038/nature12644.

Kagey, M. H. *et al.* (2010) 'Mediator and cohesin connect gene expression and chromatin architecture', *Nature*, 467(7314), pp. 430–435. doi: 10.1038/nature09380.

Kobayashi, S. (2004) 'Retrotransposon-Induced Mutations in Grape Skin Color', *Science*, 304(5673), pp. 982–982. doi: 10.1126/science.1095011.

Kornberg, R. D. (1974) 'Chromatin Structure: A Repeating Unit of Histones and DNA', *Science*, 184(4139), pp. 868–871. doi: 10.1126/science.184.4139.868.

Kosak, S. T. and Groudine, M. (2004) 'Form follows function: The genomic organization of

cellular differentiation', *Genes and Development*, pp. 1371–1384. doi: 10.1101/gad.1209304.

Kurtz, S. *et al.* (2008) 'A new method to compute K-mer frequencies and its application to annotate large repetitive plant genomes.', *BMC genomics*, 9, p. 517. doi: 10.1186/1471-2164-9-517.

Lajoie, B. R., Dekker, J. and Kaplan, N. (2015) 'The Hitchhiker's guide to Hi-C analysis: Practical guidelines', *Methods*, 72(C), pp. 65–75. doi: 10.1016/j.ymeth.2014.10.031.

Lanctôt, C. *et al.* (2007) 'Dynamic genome architecture in the nuclear space: regulation of gene expression in three dimensions', *Nature Reviews Genetics*, 8(2), pp. 104–115. doi: 10.1038/nrg2041.

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

Larson, A. *et al.* (2017) 'Liquid droplet formation by HP1α suggests a role for phase separation in heterochromatin.', *Nature*, 547(7662), pp. 236–240. doi: doi: 10.1038/nature22822.

Le, T. B. K. *et al.* (2013) 'High-resolution mapping of the spatial organization of a bacterial chromosome.', *Science (New York, N.Y.)*, 342(6159), pp. 731–4. doi: 10.1126/science.1242059.

Lettice, L. A. *et al.* (2003) 'A long-range Shh enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly', *Human Molecular Genetics*, 12(14), pp. 1725–1735. doi: 10.1093/hmg/ddg180.

Li, G. *et al.* (2012) 'Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation', *Cell*, 148(1–2), pp. 84–98. doi: 10.1016/j.cell.2011.12.014.

Li, H. *et al.* (2009) 'The Sequence Alignment/Map format and SAMtools', *Bioinformatics*, 25(16), pp. 2078–2079. doi: 10.1093/bioinformatics/btp352.

Li, H. and Durbin, R. (2009) 'Fast and accurate short read alignment with Burrows – Wheeler transform', *Bioinformatics*, 25(14), pp. 1754–1760. doi: 10.1093/bioinformatics/btp324.

Lieber, M. R. (2010) 'The Mechanism of Double-Strand DNA Break Repair by the Nonhomologous DNA End-Joining Pathway', *Annual Review of Biochemistry*. doi: 10.1146/annurev.biochem.052308.093131.

Lieberman-aiden, E. *et al.* (2010) 'Comprehensive mapping of long range interactions reveal folding principles of the human genome', *October*, 326(5950), pp. 289–293. doi: 10.1126/science.1181369.Comprehensive.

Lieberman-Aiden, E. *et al.* (2009) 'Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome', *Science*, 326(5950), pp. 289–293. doi: 10.1126/science.1181369.

Lin, W. Y., Wilson, J. H. and Lin, Y. (2013) 'Repair of chromosomal double-strand breaks by precise ligation in human cells', *DNA Repair*. doi: 10.1016/j.dnarep.2013.04.024.

Liu, C. *et al.* (2016) 'Genome-wide analysis of chromatin packing in Arabidopsis thaliana at single-gene resolution', *Genome Research*, 26(8), pp. 1057–1068. doi: 10.1101/gr.204032.116.

Liu, C. *et al.* (2017) 'Prominent topologically associated domains differentiate global chromatin packing in rice from Arabidopsis', *Nature Plants*. Springer US, pp. 0–1. doi: 10.1038/s41477-017-0005-9.

Liu, S. *et al.* (2018a) 'From 1D sequence to 3D chromatin dynamics and cellular functions: a phase separation perspective', *Nucleic Acids Research*. doi: 10.1093/nar/gky633.

Liu, S. *et al.* (2018b) 'From 1D sequence to 3D chromatin dynamics and cellular functions: a phase separation perspective', *Nucleic Acids Research*. doi: 10.1093/nar/gky633.

Louwers, M., Splinter, E., *et al.* (2009) 'Studying physical chromatin interactions in plants using Chromosome Conformation Capture (3C)', *Nature protocols*, 4(8), pp. 1216–1229.

Louwers, M., Bader, R., *et al.* (2009) 'Tissue- and expression level-specific chromatin looping

at maize b1 epialleles.', *The Plant cell*, 21(3), pp. 832–42. doi: 10.1105/tpc.108.064329.

Lupiáñez, D. G. *et al.* (2015) 'Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions', *Cell*, 161(5), pp. 1012–1025. doi: 10.1016/j.cell.2015.04.004.

Ma, W. *et al.* (2014) 'Fine-scale chromatin interaction maps reveal the cis-regulatory landscape of human lincRNA genes', *Nature Methods*, 12(1), pp. 71–78. doi: 10.1038/nmeth.3205.

Maron, L. G. *et al.* (2013) 'Aluminum tolerance in maize is associated with higher MATE1 gene copy number', *Proceedings of the National Academy of Sciences of the United States of America*, 110, pp. 5241–5246. doi: 10.1073/pnas.1220766110.

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

Martin, M. (2011) 'Cutadapt removes adapter sequences from high-throughput sequencing reads', *EMBnet.journal*, 17(1), p. 10. doi: 10.14806/ej.17.1.200.

Mascher, M. *et al.* (2017) 'A chromosome conformation capture ordered sequence of the barley genome', *Nature*. Nature Publishing Group, 544(7651), pp. 427–433. doi: 10.1038/nature22043.

Meaburn, K. J. and Misteli, T. (2007) 'Cell biology: Chromosome territories', *Nature*, pp. 379–781. doi: 10.1038/445379a.

Mills, R. E. *et al.* (2011) 'Mapping copy number variation by population-scale genome sequencing', *Nature*, 470(7332), pp. 59–65. doi: 10.1038/nature09708.

Milne, I. *et al.* (2013) 'Using tablet for visual exploration of second-generation sequencing data', *Briefings in Bioinformatics*, 14(2), pp. 193–202. doi: 10.1093/bib/bbs012.

Morgante, M., De Paoli, E. and Radovic, S. (2007) 'Transposable elements and the plant pan-genomes', *Current Opinion in Plant Biology*, 10(2), pp. 149–155. doi:

10.1016/j.pbi.2007.02.001.

Muñoz-Amatriaín, M. *et al.* (2013) 'Distribution, functional impact, and origin mechanisms of copy number variation in the barley genome', *Genome Biology*, 14(6), p. R58. doi: 10.1186/gb-2013-14-6-r58.

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. doi: 10.1073/pnas.1009363108.

Nagano, T. *et al.* (2013) 'Single-cell Hi-C reveals cell-to-cell variability in chromosome structure', *Nature*, 502(7469), pp. 59–64. doi: 10.1038/nature12593.

Nora, E. P. *et al.* (2012) 'Spatial partitioning of the regulatory landscape of the X-inactivation centre', *Nature*, 485(7398), pp. 381–385. doi: 10.1038/nature11049.

Nora, E. P. *et al.* (2017) 'Targeted Degradation of CTCF Decouples Local Insulation of Chromosome Domains from Genomic Compartmentalization', *Cell*, 169(5), p. 930–944.e22. doi: 10.1016/j.cell.2017.05.004.

Nurick, I., Shamir, R. and Elkon, R. (2018) 'Genomic meta-analysis of the interplay between 3D chromatin organization and gene expression programs under basal and stress conditions', *bioRxiv*. BioMed Central, pp. 1–14. doi: dx.doi.org/10.1101/337766.

Palstra, R.-J. *et al.* (2003) 'The β-globin nuclear compartment in development and erythroid differentiation', *Nature Genetics*, 35(2), pp. 190–194. doi: 10.1038/ng1244.

Pannunzio, N. R. *et al.* (2014) 'Non-homologous end joining often uses microhomology: Implications for alternative end joining', *DNA Repair*. doi: 10.1016/j.dnarep.2014.02.006.

Pannunzio, N. R., Watanabe, G. and Lieber, M. R. (2018) 'Nonhomologous DNA end-joining for repair of DNA double-strand breaks', *Journal of Biological Chemistry*, 293(27), pp. 10512–10523. doi: 10.1074/jbc.TM117.000374.

Parada, L. A. *et al.* (2002) 'Conservation of relative chromosome positioning in normal and cancer cells', *Current Biology*, 12(19), pp. 1692–1697. doi: 10.1016/S0960-9822(02)01166-1.

Parada, L. A., McQueen, P. G. and Misteli, T. (2004) 'Tissue-specific spatial organization of genomes.', *Genome biology*, 5(7), p. R44. doi: 10.1186/gb-2004-5-7-r44.

Phillips, J. E. and Corces, V. G. (2009) 'CTCF: Master Weaver of the Genome', *Cell*, pp. 1194–1211. doi: 10.1016/j.cell.2009.06.001.

Plys, A. J. and Kingston, R. E. (2018) 'Dynamic condensates activate transcription', *Science*, pp. 329–330. doi: 10.1126/science.aau4795.

Prieto, P. *et al.* (2004) 'Chromosomes associate premeiotically and in xylem vessel cells via their telomeres and centromeres in diploid rice (Oryza sativa)', *Chromosoma*, 112(6), pp. 300–307. doi: 10.1007/s00412-004-0274-8.

Puchta, H. (2005) 'The repair of double-strand breaks in plants: Mechanisms and consequences for genome evolution', *Journal of Experimental Botany*. doi: 10.1093/jxb/eri025.

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

Rabl, C. (1885) 'Über Zelltheilung.', *Morph Jb*, (10), pp. 214–330.

R Core Team (2017) 'R: A Language and Environment for Statistical Computing', *R Foundation for Statistical Computing, Vienna, Austria*, p. {ISBN} 3-900051-07-0. doi: http://www.R-project.org/.

Rafalski, A. (2002) 'Applications of single nucleotide polymorphisms in crop genetics', *Current Opinion in Plant Biology*, pp. 94–100. doi: 10.1016/S1369-5266(02)00240-6.

Ramani, V., Shendure, J. and Duan, Z. (2016) 'Understanding Spatial Genome Organization: Methods and Insights', *Genomics, Proteomics and Bioinformatics*. Beijing Institute of Genomics, Chinese Academy of Sciences and Genetics Society of China, 14(1), pp. 7–20. doi: 10.1016/j.gpb.2016.01.002.

Ramírez, F. *et al.* (2018) 'High-resolution TADs reveal DNA sequences underlying genome organization in flies', *Nature Communications*, 9(1). doi: 10.1038/s41467-017-02525-w.

Rao, S. S. P. *et al.* (2014) 'A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping', *Cell*. Elsevier Inc., 159(7), pp. 1665–1680. doi: 10.1016/j.cell.2014.11.021.

Rao, S. S. P. *et al.* (2017) 'Cohesin Loss Eliminates All Loop Domains', *Cell*, 171(2), p. 305–320.e24. doi: 10.1016/j.cell.2017.09.026.

Robinson, J. T. *et al.* (2018) 'Juicebox.js Provides a Cloud-Based Visualization System for Hi-C Data', *Cell Systems*, 6(2), p. 256–258.e1. doi: 10.1016/j.cels.2018.01.001.

Rowley, M. J. *et al.* (2017) 'Evolutionarily Conserved Principles Predict 3D Chromatin Organization', *Molecular Cell*. Elsevier Inc., 67(5), p. 837–852.e7. doi: 10.1016/j.molcel.2017.07.022.

Rowley, M. J. and Corces, V. G. (2016) 'The three-dimensional genome: Principles and roles of long-distance interactions', *Current Opinion in Cell Biology*, pp. 8–14. doi: 10.1016/j.ceb.2016.01.009.

Ryba, T. *et al.* (2010) 'Evolutionarily conserved replication timing profiles predict long-range chromatin interactions and distinguish closely related cell types', *Genome Research*, 20(6), pp. 761–770. doi: 10.1101/gr.099655.109.

Sabari, B. R. *et al.* (2018) 'Coactivator condensation at super-enhancers links phase separation and gene control', *Science*, 361(6400). doi: 10.1126/science.aap9195.

Saldanha, A. J. (2004) 'Java Treeview--extensible visualization of microarray data.', *Bioinformatics (Oxford, England)*, 20(17), pp. 3246–8. doi: 10.1093/bioinformatics/bth349.

San Filippo, J., Sung, P. and Klein, H. (2008) 'Mechanism of eukaryotic homologous recombination.', *Annual review of biochemistry*. doi: 10.1146/annurev.biochem.77.061306.125255.

Sanborn, A. L. *et al.* (2015a) 'Chromatin extrusion explains key features of loop and domain formation in wild-type and engineered genomes', *Proceedings of the National Academy of Sciences*, 112(47), pp. E6456–E6465. doi: 10.1073/pnas.1518552112.

Sanborn, A. L. *et al.* (2015b) 'Chromatin extrusion explains key features of loop and domain formation in wild-type and engineered genomes', *Proceedings of the National Academy of Sciences*, 112(47), pp. E6456–E6465. doi: 10.1073/pnas.1518552112.

Sanyal, A. *et al.* (2012) 'The long-range interaction landscape of gene promoters', *Nature*, 489(7414), pp. 109–113. doi: 10.1038/nature11279.

Schmid, M. W., Grob, S. and Grossniklaus, U. (2015) 'HiCdat: a fast and easy-to-use Hi-C data analysis tool', *BMC Bioinformatics*. BMC Bioinformatics, 16(1), p. 277. doi: 10.1186/s12859-015-0678-x.

Schubert, I. *et al.* (2004) 'DNA damage processing and aberration formation in plants', in *Cytogenetic and Genome Research*. doi: 10.1159/000077473.

Schwartz, Y. B. and Cavalli, G. (2017) 'Three-dimensional genome organization and function in Drosophila', *Genetics*, 205(1), pp. 5–24. doi: 10.1534/genetics.115.185132.

Schwarzer, W. *et al.* (2017) 'Two independent modes of chromatin organization revealed by cohesin removal', *Nature*, 551(7678), pp. 51–56. doi: 10.1038/nature24281.

Servant, N. *et al.* (2015) 'HiC-Pro: an optimized and flexible pipeline for Hi-C data processing', *Genome Biology*, 16(1), p. 259. doi: 10.1186/s13059-015-0831-x.

Sexton, T. *et al.* (2012) 'Three-dimensional folding and functional organization principles of the Drosophila genome', *Cell*, 148(3), pp. 458–472. doi: 10.1016/j.cell.2012.01.010.

Shrivastav, M., De Haro, L. P. and Nickoloff, J. A. (2008) 'Regulation of DNA double-strand break repair pathway choice', *Cell Research*. doi: 10.1038/cr.2007.111.

Simonis, M. *et al.* (2006) 'Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture-on-chip (4C)', *Nature Genetics*, 38(11), pp. 1348–1354. doi: 10.1038/ng1896.

Singh, B. N. and Hampsey, M. (2007) 'A Transcription-Independent Role for TFIIB in Gene Looping', *Molecular Cell*, 27(5), pp. 806–816. doi: 10.1016/j.molcel.2007.07.013.

Smith, E. M. *et al.* (2016) 'Invariant TAD Boundaries Constrain Cell-Type-Specific Looping

Interactions between Promoters and Distal Elements around the CFTR Locus', *American Journal of Human Genetics*. Elsevier Ltd, 98(1), pp. 185–201. doi: 10.1016/j.ajhg.2015.12.002.

Stevens, T. J. *et al.* (2017) '3D structures of individual mammalian genomes studied by single-cell Hi-C', *Nature*. Nature Publishing Group, 544(7648), pp. 59–64. doi: 10.1038/nature21429.

Strom, A. R. *et al.* (2017) 'Phase separation drives heterochromatin domain formation', *Nature*, 547(7662), pp. 241–245. doi: 10.1038/nature22989.

Takizawa, T., Meaburn, K. J. and Misteli, T. (2008) 'The Meaning of Gene Positioning', *Cell*, pp. 9–13. doi: 10.1016/j.cell.2008.09.026.

Tan-Wong, S. M. *et al.* (2012) 'Gene loops enhance transcriptional directionality', *Science*, 338(6107), pp. 671–675. doi: 10.1126/science.1224350.

Tanabe, H. *et al.* (2002) 'Evolutionary conservation of chromosome territory arrangements in cell nuclei from higher primates', *Proceedings of the National Academy of Sciences*, 99(7), pp. 4424–4429. doi: 10.1073/pnas.072618599.

Tenaillon, M. I., Hollister, J. D. and Gaut, B. S. (2010) 'A triptych of the evolution of plant transposable elements', *Trends in Plant Science*, pp. 471–478. doi: 10.1016/j.tplants.2010.05.003.

Tettelin, H. *et al.* (2005) 'Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: Implications for the microbial "pan-genome"', *Proceedings of the National Academy of Sciences*, 102(39), pp. 13950–13955. doi: 10.1073/pnas.0506758102.

The Arabidopsis Genome Initiative (2000) 'Analysis of the genome sequence of the flowering plant Arabidopsis thaliana.', *Nature*, 408(6814), pp. 796–815. doi: 10.1038/35048692.

Tiang, C.-L., He, Y. and Pawlowski, W. P. (2012) 'Chromosome Organization and Dynamics during Interphase, Mitosis, and Meiosis in Plants', *Plant Physiology*, 158(1), pp. 26–34. doi: 10.1104/pp.111.187161.

Tolhuis, B. *et al.* (2002) 'Looping and interaction between hypersensitive sites in the active ??-globin locus', *Molecular Cell*, 10(6), pp. 1453–1465. doi: 10.1016/S1097-2765(02)00781-5.

Vitulo, N. *et al.* (2014) 'A deep survey of alternative splicing in grape reveals changes in the splicing machinery related to tissue, stress condition and genotype', *BMC Plant Biology*, 14(1). doi: 10.1186/1471-2229-14-99.

Vu, G. T. H. *et al.* (2014) 'Repair of Site-Specific DNA Double-Strand Breaks in Barley Occurs via Diverse Pathways Primarily Involving the Sister Chromatid', *The Plant Cell*. doi: 10.1105/tpc.114.126607.

Walter, J. *et al.* (2003) 'Chromosome order in HeLa cells changes during mitosis and early G1, but is stably maintained during subsequent interphase stages', *Journal of Cell Biology*, 160(5), pp. 685–697. doi: 10.1083/jcb.200211103.

Wang, C. *et al.* (2015) 'Genome-wide analysis of local chromatin packing in Arabidopsis thaliana', *Genome Research*, 25(2), pp. 246–256. doi: 10.1101/gr.170332.113.

Wang, M. *et al.* (2018) 'Evolutionary dynamics of 3D genome architecture following polyploidization in cotton', *Nature Plants*, 4(2), pp. 90–97. doi: 10.1038/s41477-017-0096-3.

Xie, W. *et al.* (2013) 'Epigenomic analysis of multilineage differentiation of human embryonic stem cells', *Cell*, 153(5), pp. 1134–1148. doi: 10.1016/j.cell.2013.04.022.

Xie, W. J. *et al.* (2017) 'Structural modeling of chromatin integrates genome features and reveals chromosome folding principle', *Scientific Reports*. Springer US, 7(1), p. 2818. doi: 10.1038/s41598-017-02923-6.

Y.Benjamini and Y.Hochberg (1995) 'Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing', *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1), pp. 289–300. doi: 10.2307/2346101.

Yang, J. and Corces, V. G. (2012) 'Insulators, long-range interactions, and genome function', *Current Opinion in Genetics and Development*, pp. 86–92. doi: 10.1016/j.gde.2011.12.007.

Young, R. A. (2011) 'Control of the embryonic stem cell state', *Cell*, pp. 940–954. doi: 10.1016/j.cell.2011.01.032.

Yu, J. *et al.* (2002) 'A draft sequence of the rice genome (*Oryza sativa* L. ssp. indica).', *Science (New York, N.Y.)*, 296(5565), pp. 79–92. doi: 10.1126/science.1068037.

Zanni, V. *et al.* (2013) 'Distribution, evolution, and diversity of retrotransposons at the flamenco locus reflect the regulatory properties of piRNA clusters.', *Proceedings of the National Academy of Sciences of the United States of America*, 110(49), pp. 19842–7. doi: 10.1073/pnas.1313677110.

Zhan, Y. *et al.* (2017) 'Reciprocal insulation analysis of Hi-C data shows that TADs represent a functionally but not structurally privileged scale in the hierarchical folding of chromosomes', *Genome Research*, 27(3), pp. 479–490. doi: 10.1101/gr.212803.116.

Zhang, Y. *et al.* (2012) 'Spatial organization of the mouse genome and its role in recurrent chromosomal translocations', *Cell*, 148(5), pp. 908–921. doi: 10.1016/j.cell.2012.02.002.

Zhang, Y. *et al.* (2013) 'Chromatin connectivity maps reveal dynamic promoter–enhancer long-range associations', *Nature*, 504(7479), pp. 306–310. doi: 10.1038/nature12716.