UNIVERSITÀ DEGLI STUDI DI UDINE

Dottorato di Ricerca in Scienze e Biotecnologie Agrarie
Ciclo XXXVIII
Coordinatore: Prof. Mauro Spanghero

TESI DI DOTTORATO DI RICERCA

ANALYSIS OF EPIGENOMIC VARIABILITY IN GRAPEVINE AND ITS RELATION WITH STRUCTURAL VARIATION

DOTTORANDO
Dott. Mirko Celii

SUPERVISORE
Prof. Michele Morgante

CO-SUPERVISORE
Dott. Emanuele De Paoli

ANNO ACCADEMICO 2014/2015

Index

SUMMARY

Epigenetics encompasses a series of chromatin modifications that are potentially inheritable and can result into a change of gene expression without involving a change in the underlying DNA sequence. Epigenetics is involved in several fundamental mechanisms that regulate cell cycle in all eukaryotes including X chromosome inactivation, gene silencing, paramutation, parental imprinting, chromatin position effect, plant gametogenesis, flowering time, stress responses and light signaling.

Within epigenetic modifications, DNA methylation is predominant and widespread in all eukaryotic kingdoms and consists in a reversible reaction which transfers a methyl group on a cytosine. In mammals methylation occurs in CG-rich regions known as CpG islands, whereas in plant methylation may occurs in CG, CHG and CHH contexts, where H may be A,C or T, with different mechanisms.

Depending on the location, DNA methylation may have opposite effects: in heterochromatin it is generally associated to transcriptional inactivity but in the transcribed region of genes, methylation in the CG context is associated to medium-to-high transcriptional level. The silencing effect of DNA methylation represents a useful defense weapon against both retrovirus infection and transposable element (TE) insertions: indeed such sequence elements generally become highly methylated as a plant response to prevent further mobilization. Despite this mechanism, during evolution TEs colonized eukaryotic genomes, up to represent 75% of the total genome in some plant species. . TEs are a major factor underlying the tremendous intra-species genome variability that has been revealed thanks to the introduction of next generation sequencing (NGS) technology and the resequencing of several individuals of the same species. This led scientists to introduce, initially only for bacteria then for any organism, the concept of pan-genome, composed by a common genome shared by all the individuals of the species and a dispensable genome which is not essential for survival, but is the foundation for phenotypic variability.

The dispensable genome consists of the entire set of structural variations (SVs) observed among individuals and is mainly represented by TEs.

TE sequences are generally methylated and their methylation may spread in their flanking regions. Thus, when TEs accidentally insert nearby genes or regulating sequences, they may alter their epigenetic status creating epigenetic variants called epialleles.

Specific protocols of NGS, involving the use of bisulfite, which in the overall process converts unmethylated cytosines to thymines, allow to map all methylcytosines of a genome with single-base resolution. The aim of this work was to analyse the relationship between structural variations, mainly represented by TEs, and epigenetic variations in plants. Grapevine is a suitable model for this study because it is a perennial species and generally, it is vegetatively propagated in agriculture. This technique preserves the genome from recombination, thus it allows maintaining the genotypes stable across clonal generations and focusing on mere epigenetic variation.

Moreover for grapevine a highly homozygous reference genome is already available as well as a set of grapevine-specific TEs annotated. Three biological replicates of leaf nuclear DNA of the cultivar Pinot Noir, which shares one haplotype with the sequenced genome reference, have been sequenced and analysed. To evaluate the spreading of internal TE methylation on the flanking regions, we considered hemizygous TEs in order to compare the same regions on homologous chromosome in presence or absence of TEs. Consistently with other species, grapevine TEs show high methylation in their sequence in both CG and CHG context whereas CHH context is extremely low methylated. Internal TE methylation is generally spread on their flanking regions. Within TEs, retrotranspons show a stronger impact on flanking regions compared to DNA-transposons, with different behaviors according to the differential genomic distribution of TE-groups: Ty3-Gypsy usually insert in highly methylated regions of pericentromeric chromatin, LINEs element are frequently found in highly CG methylated gene bodies, Ty1-Copia display more variable locations. Generally, where not saturated, retrotransposon insertions provoke an increase of methylation in both CG and CHG contexts, supported by statistical analyses. DNA methylation is also present in transcribed regions of grapevine genes, in particular in the CG context. A set of about 19000 genes was utilized to analyze gene body methylation (GBM) in grapevine. Similarly to other species, grapevine GBM displays an asymmetrical bell-shape profile, in which the 5’ is much less methylated than 3’. Surprisingly introns appear more methylated than exons, in contrast with other species such as Arabidopsis, humans and honey bee.

Grapevine introns occupy a large part of the genome (36.7%) and are quite rich in TEs that represent 12.4% of their sequence. The moderate TE content of introns may partially explain their higher methylation compared to flanking exons. However, when excluding genes carrying TE from the analysis, methylation in both exons and introns is reduced but still present, confirming that GBM methylation is independent from TEs, although their insertion may increase it.

Analysis of gene expression showed that genes located in highly methylated regions, especially in the CHG context, show on average a lower expression rate and furthermore their expression tend to be more conserved within varieties. On the contrary, when methylation occurs in gene bodies, transcriptional activity is not reduced and it may be even higher.

Gene expression may also be modulated by TEs: when these are located in gene flanking regions, gene expression rate is significantly lower than unaffected genes, whereas genes whose introns are enriched in TEs display significantly higher methylation and expression rates. Lastly, allele-specific expression analyses indicate that hemizygous TEs may affect the contribution of the two alleles to the expression rate of the gene.

Taken together these data confirm that DNA methylation occurs in grapevine with patterns comparable with other plant species, but with the peculiarity of highly methylated introns whose methylation is generally associated to moderate TE content and medium-to-high expression levels.

INTRODUCTION

Epigenetics in plants

In a modern formulation, the study of epigenetics, (from the ancient greek ***επί***, [epì] meaning "above" genetics) encompasses a series of chromatin modifications that are potentially inheritable and can result into a change of gene expression without involving a change in the underlying DNA sequence.

Epigenetic modifications may be either temporary or inheritable through cell division and gametogenesis and may occur both on DNA and on chromatin proteins.

Many biological processes are modulated by epigenetic regulation, such as X chromosome inactivation, gene silencing, paramutation, parental imprinting and chromatin position effect.

So far, several mechanisms that underlie epigenetic phenomena have been identified and studied, including:

- Histone tail modifications
- Changing of nucleosome positioning
- Covalent modifications of cytosines
- Small non-conding RNAs pathways.

Epigenetic modifications are massive also in humans and other eukaryotic phyla. DNA is an extremely long molecule (up to 1.8 m in human) and thus it needs to be packed in order to fit in the cell nucleus. DNA packing is made possible by protein-DNA interactions; which form the complex called chromatin.

During cell division, DNA reaches the highest packing order in the form of chromosomes, while during the rest of cell cycle, DNA sequence is relaxed in the form of chromatin, which is composed by euchromatin with high transcriptional activity and heterochromatin which is more compact and usually shows a very low transcriptional activity. Heterochromatin may be either constitutive or facultative. Different cells have different needs and thus require the expression of cell-specific genes but not others; in fact it would be extremely expensive in terms of energy cost for a cell to keep all the possible pathways constantly active.

Within a cell, epigenetic modifications contribute on one hand to maintain accessible for the transcription machinery DNA portions carrying specific genes, on the other to condense into heterochromatic regions DNA portions which are not essential for the cell.

Histone tails post-translational modifications and nucleosome positioning

Histones are alkaline proteins with high affinity for DNA, which may be packed by winding around them thanks to the several hydrogen bonds. DNA affinity to histones is modulated by post-translational modifications that occur on histone tails, such as methylation, acetylation, phosphorylation, citrullination, sumoylation, biotinylation, ADP-ribosilation and ubiquitination. Methylation and acetylation are predominant and they may compete for the same lysine residues. Acetylation is generally positively correlated with gene expression, while the actual effect of methylation depends on the residue position. Differences between organisms have also been observed. Moreover, up to three methyl groups may added to a single lysine residue and the methyl groups number may cause opposite effects on the basis of both residue and organism. There are 5 different families of histones, H1/H5, H2A, H2B, H3, H4. Dimers of H2A, H2B, H3, and H4 form an octamer called nucleosome, where a 147 bp DNA stretch can bend 1.67 times. The unbent DNA between two nucleosomes is called linker DNA and may be long up to 80 bp. Because the linker but not the bent DNA is accessible for the transcription machinery, histone tail modifications and nucleosome positioning represent very fine epigenetic regulators of gene expression.

Linker histones H1/H5 provide an additional packing by clipping nucleosomes together in a structure called "30 nm fiber", which on its turn is bent around specific scaffold proteins both in the chromosomal and heterochromatic structures.

Histone proteins are highly conserved within eukaryotes and similar proteins are present also in prokaryotes.

DNA Methylation

DNA covalent modifications occur on cytosines, and include several reactions which, on one hand, do not alter cytosine biological property during replication and transcription and do not affect the pairing with a guanine on the complementary strand, but, on the other hand, may add additional epigenetics information which may contribute to the expression of a certain phenotype.

So far several cytosine modifications have been detected including methylation, hydroxymethylation and formylation; however methylation is predominant and has been deeply investigated.

DNA methylation is a reversible enzymatic reaction, catalysed by DNA-methytransferases which transfer a methyl group from the substrate S-adenosylmethionine to the carbon 5 of the cytosine base. This reaction may be potentially reverted by demethylating enzymes.

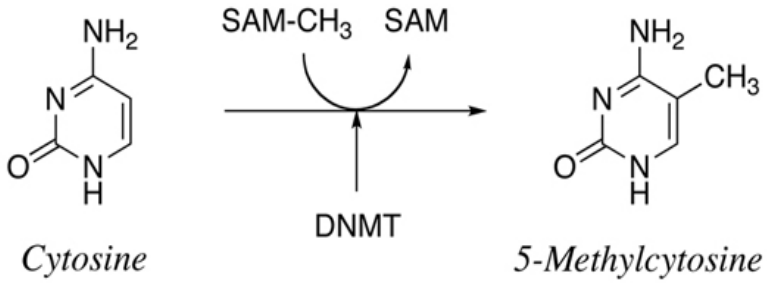


Figure 1 | Schema of cytosine methylation

Cytosine methylation is widespread between organisms; in fact it is in common to both complex organisms and prokaryotes but it is interestingly absent in few model organisms such as *Drosophila* and *Caenorhabditis elegans*. Homology between bacterial and eukaryotic methytransferases suggests a very ancient origin and then a subsequent loss in the systems named above.

The methyl group is located in the major groove of the DNA double helix and does not affect hydrogen bonds involved in base pairing.

5m-cytosine may convert to thymine by spontaneous deamination, which is not repaired by the repair mechanism as the unmethylated cytosine that converts to uracil. Hence in the organism where methylation is present, there is an increase of thymine content during evolution.

An *in vitro* demination reaction may be performed in presence of sodium bisulfite; this reaction is at the basis of the modern Bisulfite Sequencing technology which allows us to map at single base resolution methylated and unmethylated cytosines based on which of the cytosines are converted to uracil and which are not.

Depending on the organisms, DNA-methylation may occur in different contexts: in humans it is predominant in CG rich regions known as CpG islands in which ^{5m}C may represent up to 80% of the total cytosines; in plants 3 different contexts with different methylation mechanisms are known, corresponding to CG, CHG and CHH where H is any base but a G.

Human CpG islands are usually located within few kilobases from gene transcription start site (TSS), and thus their methylation level may alter the expression of flanking genes.

Plant CG and CHG contexts are symmetrical in the Watson and Crick's strands, while the CHH is asymmetrical. However only the CG context is characterized by a symmetrical methylation in both strands. Within plants, CG, CHG and CHH contribution to the total cytosines is variable as well as the methylated fraction of each context (Figure 2)

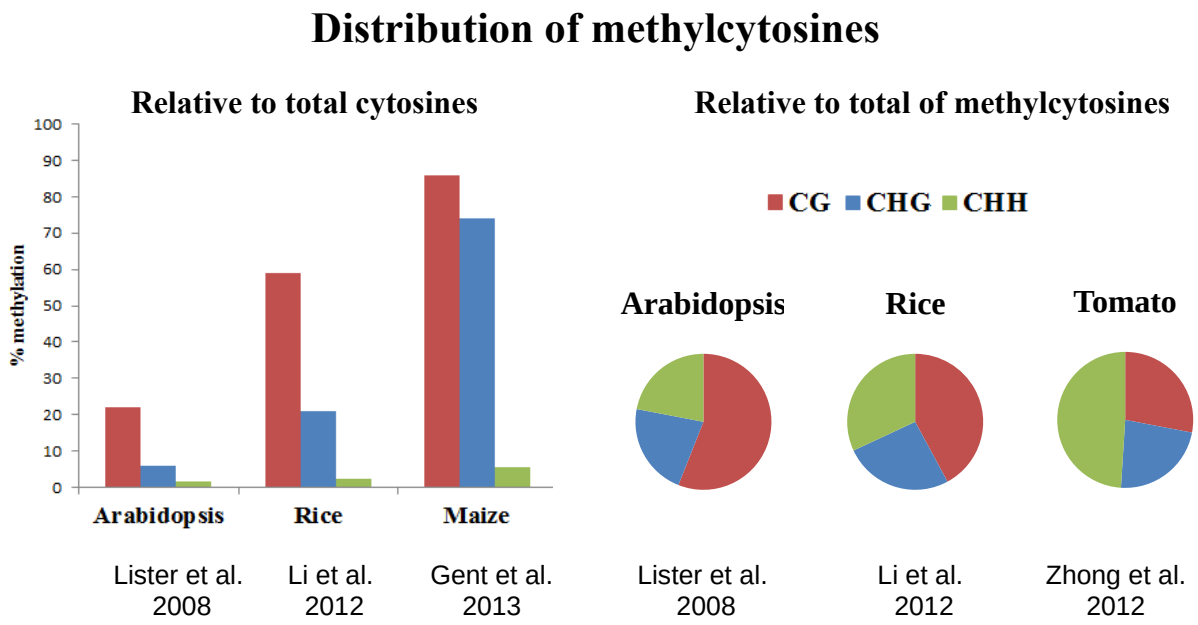


Figure 2 | Patterns of DNA Methylation in Plants

DNA methylation mechanisms may be distinguished in:

- ***de novo*** mechanism, which consists in the methylation of previously unmethylated cytosines.

- **maintenance** mechanism, which acts during cell division and is designated to replicate the methylation profile of the pre-existent strand in the new one.

Different contexts have different mechanisms for both *de novo* and maintenance methylation.

DNA methylation in different sequence contexts

De novo CG methylation in mammals is performed by the DNA methyltransferase class (Dnmt3), which binds H3 tail unmethylated at K4. The same family in plant is called DOMAIN REARRANGEMENT METHYLTRANSFERASE 2 (DMR2) but acts with a very different mechanism which involves siRNAs in a RNA-directed DNA methylation. DMR2 may methylate cytosines also in CHG and CHH contexts (Cao & Jacobsen, 2002).

CG methylation, however, is also maintained during cell replication through a highly conserved process within eukaryotes and it is based on the DNA symmetry of CG: the hemimethylation binding proteins VIM (Hufr1 in human) targets ^{5m}C in CG dinucleotides in the pre-existent DNA strand allowing DNA-methylatransferase MET1 (Dnmt3 in human) to methylate the complementary CG dinucleotide in the newly synthetized strand during replication. This is the only mechanism for the maintenance of methylation in parallel with replication described so far whereas other maintenance mechanisms involve *de novo* methylation of the new strand.

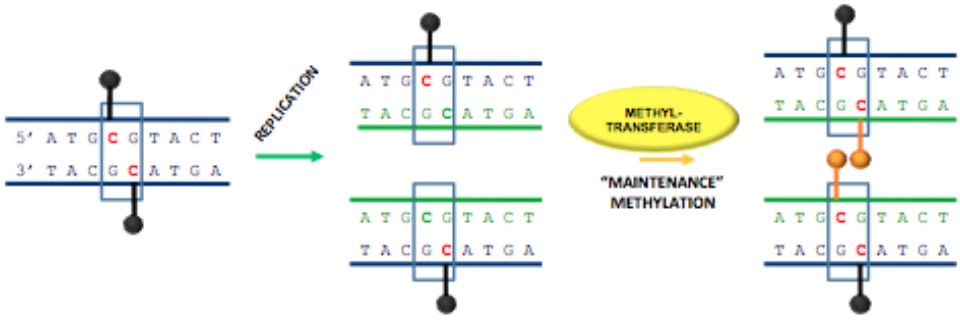


Figure 3 | Maintenance of CG Methylation during DNA replication

The CHG methylation levels in mammals are much lower than in plants, thus different mechanisms are involved for its maintenance. Plants maintain high level of CHG methylation thanks to a self-reinforcing feed-loop of CMT3 DNA-methyltransferase whose action is guided by the methylation of H3K9. In humans the methylation level at CHG sites as well as at CHH ones is very low, and Dnmt3 seems to be involved also in its maintenance.

Because of the asymmetric nature of CHH sites, the maintenance of CHH methylation is thought to happen through a *de novo* mechanism after cell replication. A very sophisticated *de novo* mechanism has been characterized in *Arabidopsis* for CHH methylation and involves the two plant specific RNA polymerases IV and V. Pol IV and Pol V transcription are independent from each other although the two enzymes are both necessary for the silencing of specific loci. The Pol IV transcripts originated from these loci are made double stranded by RDR2 (RNA-dependent RNA polymerase 2) then the resulting dsRNAs are cleaved by DICER-like protein 3 (DCL3) generating short interfering RNAs (siRNAs) that are capable of forming functional complexes with the AGO4 (ARGONAUTE 4) effector protein. Pol V nascent transcripts are targeted by the siRNAs/AGO4 complex and, at the same time, the Pol V C-terminal domain interacts with AGO4 and serves as scaffold for the assembly of a

complex that includes RDM1, an ssDNA-binding protein with a preference for methylated DNA (Gao et al., 2010; Law et al., 2010) and the methyltransferase DMR2. DMR2 is able to methylate in all the three contexts (Cao & Jacobsen, 2002), although in this specific mechanism only CHH methylation has been observed.

Pol V often immunolocalizes to TE-rich loci, suggesting that this pathway may be specifically required to reinforce the silencing of TEs, whose CG and CHG methylation might be already present because of their maintenance mechanisms, by also promoting CHH (*de novo*) methylation. Depletion of Pol V does not affect CHH methylation in pericentromeric regions, suggesting that other mechanisms may be involved in overall CHH methylation. (Wierzbicki et al., 2012; Wierzbicki, Haag, & Pikaard, 2008; Wierzbicki, Ream, Haag, & Pikaard, 2009).

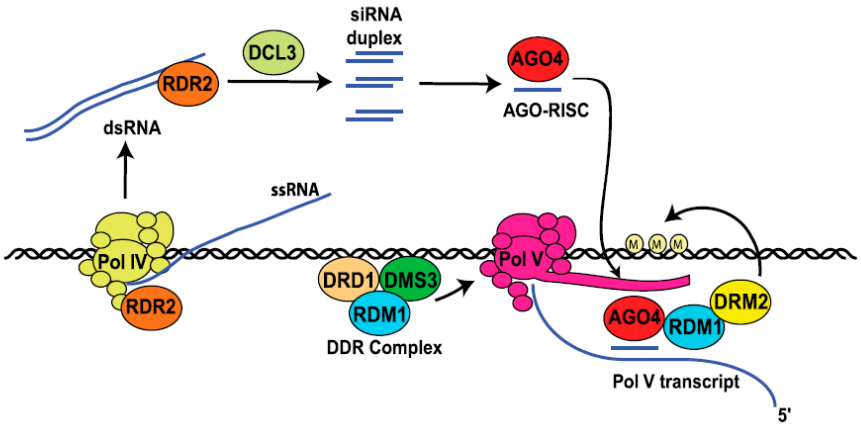


Figure 4 | De novo CHH Methylation pathway (Wierzbicki et al., 2012)

In maize, CHH methylation seems to have a separate localization from that of CG and CHG methylation. Indeed, both CG and CHG but not CHH are highly methylated in intergenic heterochromatic regions, but within 1 kb from the transcription start site (TSS) of active genes, CHH methylation levels show a peak that is proportional to the expression rate of the flanking gene. These "CHH islands" may act as insulators for the heterochromatin propagation and protect genes from epigenetic silencing (Gent et al., 2013).

Functions of DNA methylation

DNA methylation is usually correlated with histone modifications, such as H3K9 dimethylation, that condense chromatin thereby leading to a low transcriptional activity of the region (see Figure 5). Differential methylation between different cells may reflect their

differential needs in terms of expression of specific genes rather than others. This transcriptional inactivating ability of DNA methylation is also useful as a defense weapon against retrovirus infection. Indeed, after their integration in the host genome their sequence is methylated in order to prevent the transcription of their genes and the propagation of the infection (Wassenegger, *Heimes*, Riedel, & Sänger, 1994). Similarly, transposable element sequences are usually methylated to prevent excessive mobilization which may provoke gene disruption or mis-regulation.

In contrast with the general silencing effect, high levels of methylation, exclusively in the CG context, were found in transcribed region of genes (gene bodies) of Arabidopsis and soybean and other species. In Arabidopsis roughly 1/3 of genes bodies present CG methylation and interestingly the highest methylation level is detected in genes with medium to high levels of transcription (Cokus et al., 2008; Hsieh et al., 2009; Lister et al., 2008; Tran et al., 2005; Zhang et al., 2006; Zilberman, Gehring, Tran, Ballinger, & Henikoff, 2007).

GBM is particularly enriched in exons and may be involved in exon definition and splicing regulation (Maor et al., 2015; Chodavarapu et al., 2010; Laurent et al., 2010).

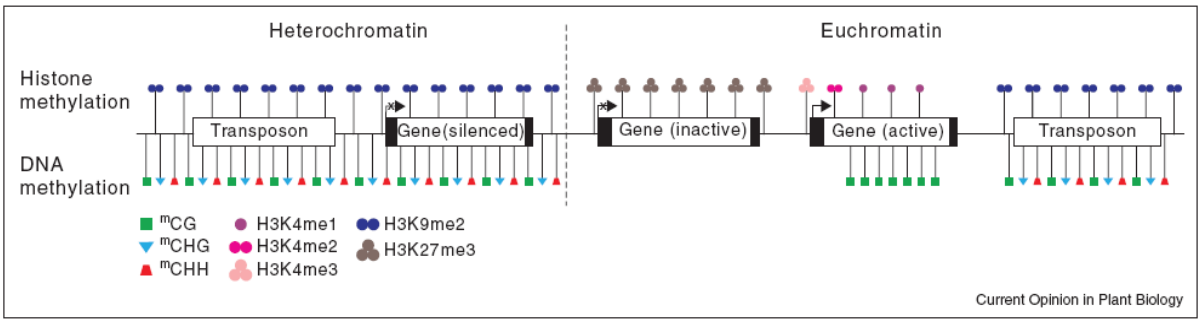


Figure 5 | Summary of epigenetic marks in euchromatin and heterochromatin (Feng & Jacobsen, 2011)

Transposable Elements

DNA methylation is generally associated to chromatin packing and thus to transcriptional inactivity, acting manly (but not exclusively) as a silencing phenomenon. DNA methylation, as well as other epigenetic modifications such as histones post-transcriptional modifications, may also occur as a consequence of Transposable Element (TE) insertions.

As their name suggests, TEs are DNA sequences capable of transposing to other locations of the genome. They were discovered in maize in the 50s by the 1983 Nobel Prize winner Barbara McClintock and so far they have been found in all the eukaryotes species sequenced (e.g. Drosophila, Adams et al., 2000; cereals, Flavell, Rimpau, & Smith, 1977; rice Sequencing Project International Rice Genome, 2005 and human, Lander et al., 2001)

TEs generally encode enzymes able to integrate the TE sequences elsewhere in the genome, however a large number of non-autonomous TEs has been described and generally they require a trans-acting enzyme produced by autonomous TEs.

The insertion event usually produces target site duplication (TSD) externally to both TE termini, with different length and sequence according to the TE family. TSDs are a useful feature to recognize and distinguish TE insertion.

If on one hand TE insertions may promote genetic variability on the other hand, their movement may dramatically affect gene expression by disrupting open reading frames or regulatory sequences such as promoter and enhancers. Genome reacts to TE movement by methylating their sequence in order to silence them and prevent further mobilization which may be lethal for the cell. TE-induced methylation may also be extended to TE flanking regions, potentially affecting the expression of neighboring genes and thus the phenotype. These modifications may be inherited to the offspring and represent an example of genetically transmitted epiallele.

Despite the existence of mechanisms that repress their activity, TEs have been able to successfully colonize eukaryotic genomes so that they can represent more than 75% of the total genomic sequence in some plant species (Baucom et al., 2009; Eichten et al., 2012). Similarly to TEs, genomic retrovirus integrations are generally followed by the methylation of the integrated sequence in order to silence retrovirus genes transcription and stop the propagation of the infection (Wassenegger et al.,1994).

According to Finnegan (1989) and Wicker et al., (2007) TEs may be divided in two principal classes, based on their transposition intermediate:

- class I elements, characterized by an RNA intermediate and a “copy and paste” mechanism;
- class II elements, characterized by a DNA intermediate and by either “cut and paste” or “copy and paste” mechanism.

Below the class, hierarchical classification proposed by Wicker et al. includes subclass, order, superfamily and family.

Class I elements

Class I elements encode a Reverse Transcriptase (RT) and thus are also known as retrotransposons. Their “copy and paste” transposition mechanism involving an RNA intermediate creates a new copy for each transposition event, hence it is not surprising that retroelements usually represent the largest fraction of the repetitive component in eukaryotic genomes. Class I includes several orders. The most abundant in the plant kingdom are LTR elements, while elements of the non-LTR order, LINE and SINE, are more abundant in mammals, including humans.

Long Terminal Repeat (LTR) elements

LTR retroelements are characterized by the presence of long terminal repeats (LTR) with the same orientation at both TE termini, flanked by a 4-6 bp TSD with a variable sequence.

Their size covers a wide range between few kbp up to 25 kbp, each single LTR may reach 5 kb and usually show both a 5’-TG-3’ start and a 5’-CA-3’ end. However, non-autonomous LTR retroelements, lacking the RT genes, may be among the shortest ones and can even shorter than 300 bp in some cases (Gao et al., 2012).

LTR retroelements sequence encodes the retrovirus-like ORFs GAG and POL. The *Pol* ORF encodes the Reverse trascriptase (RT), a DDE integrase (INT), an RNase H (RH) and an aspartic proteinase (AP). The presence of the GAG and POL ORFs suggests a common ancestor for retroelement and retroviruses (Frankel & Young, 1998; Seelamgari et al., 2004)

In plant two main superfamilies are known: ***Gypsy*** and ***Copia***, which differ in the order of the protein domains within the *Pol* ORF.

Long Interspersed Nuclear Elements (LINEs)

LINE retroelements are found in all eukaryotic kingdoms and encode RT and a nuclease necessary for the transposition. Despite the presence of long TSDs and frequently of a poliA /

A-rich tail at the 3’ end, their identification is made difficult by the lack of terminal repeats and by the frequent truncation at the 5’end. The truncation is probably due to a premature retrotranscription termination.
The LINE order includes 5 superfamilies, but only the L1 and RTE superfamilies are present in plants.

Short Interspersed Nuclear Elements (SINEs)

SINEs are short non-autonomous elements whose length may span between 80 and 550 bp. Unlike other retrotransposons, they seem to be originated by an accidental retrotransposition of the Pol III transcripts such as tRNA, 7SL and 5S rRNA. They generally exhibit a Pol III promoter at the 5’ end, which allows them to be expressed, and occasionally a polyT, A-rich, AT- rich or tandem repeat sequences at the 3’ end, whose origin is still unknown. Although they have a different origin compared to other retroelements, SINE elements may be cross-activated by autonomous LINEs and their integration generates a TSD of variable length (5-15bp) .

Class II elements

Class II elements, also known as DNA-transposons, encode the Transposase enzyme and use a DNA intermediate for transposition. On the basis of the transposition mechanism, two subclasses are distinguishable:
- subclass I, characterized by a double stranded DNA intermediate and a “cut and paste” mechanism,
- subclass II, characterized by a single stranded DNA intermediate and a “copy and paste” mechanism.

Terminal Inverted Repeat (TIR) elements

Subclass I DNA-TE comprise only the TIR order, whose elements are characterized by the presence of Terminal Inverted Repeats (TIR) of variable length which are recognized and cut by the transposase enzyme.
The transposition events do not create a new copy of the TE, thus is not surprising that DNA-TE represent a smaller fraction of the genomes than class I element.

However, other complex mechanisms may occasionally contribute to copy number expansion, such as unequal recombination, transposition during replication from a replicated to a non-replicated (yet) locus, and excision repair in the donor site.

TIR elements may be autonomous or not and thus their length is extremely variable. Non-autonomous TIR elements are generally named Miniature Inverted repeat Transposable Elements (MITEs). Although nom-autonomous, mobilization, MITEs may be cross-activated by autonomous TIR- elements.

Several TIR superfamilies have been classified according to TIR and TSD sequence and size. Five of these are present in plants: *Tc1-Mariner*, *hAT*, *Mutator*, *PIF-Harbinger* and *CACTA*. Superfamily behavior is highly specific: in rice *PIF/Harbinger* elements often insert close to genes, affecting expression either positively or negatively (Naito et al., 2009), in maize *Mutator* elements are frequently found within genes (for review, see Lisch, 2002).

Helitron elements

The subclass II of DNA-TEs is represented in plants by the Helitron order, which encodes a Y2-type Tyrosine recombinase, does not display terminal repeats but frequently a TC and a CTRR motif and an hairpin structure at 3' end.

Helitron mobilization occurs through a single strand cleavage that produces a single strand DNA intermediate which undergoes a "copy and paste" rolling-circle mechanism and integrates in the genome without TSDs but frequently in a A / T motif.

Interestingly, Helitron transposition may accidentally include gene fragments to form non autonomous elements, and in maize chimeric genes have been assembled by successive transposition events (Morgante et al., 2005).

Classification		Structure	TSD	Code	Occurrence
Order	Superfamily				
Class I (retrotransposons)					
LTR	*Copia*	GAG AP INT RT RH	4–6	RLC	P, M, F, O
	Gypsy	GAG AP RT RH INT	4–6	RLG	P, M, F, O
	Bel–Pao	GAG AP RT RH INT	4–6	RLB	M
	Retrovirus	GAG AP RT RH INT ENV	4–6	RLR	M
	ERV	GAG AP RT RH INT ENV	4–6	RLE	M
DIRS	*DIRS*	GAG AP RT RH YR	0	RYD	P, M, F, O
	Ngaro	GAG AP RT RH YR	0	RYN	M, F
	VIPER	GAG AP RT RH YR	0	RYV	O
PLE	*Penelope*	RT EN	Variable	RPP	P, M, F, O
LINE	*R2*	RT EN	Variable	RIR	M
	RTE	APE RT	Variable	RIT	M
	Jockey	ORF1 APE RT	Variable	RIJ	M
	L1	ORF1 APE RT	Variable	RIL	P, M, F, O
	I	ORF1 APE RT RH	Variable	RII	P, M, F
SINE	*tRNA*		Variable	RST	P, M, F
	7SL		Variable	RSL	P, M, F
	5S		Variable	RSS	M, O
Class II (DNA transposons) - Subclass 1					
TIR	*Tc1–Mariner*	Tase*	TA	DTT	P, M, F, O
	hAT	Tase*	8	DTA	P, M, F, O
	Mutator	Tase*	9–11	DTM	P, M, F, O
	Merlin	Tase*	8–9	DTE	M, O
	Transib	Tase*	5	DTR	M, F
	P	Tase	8	DTP	P, M
	PiggyBac	Tase	TTAA	DTB	M, O
	PIF–Harbinger	Tase* ORF2	3	DTH	P, M, F, O
	CACTA	Tase ORF2	2–3	DTC	P, M, F
Crypton	*Crypton*	YR	0	DYC	F
Class II (DNA transposons) - Subclass 2					
Helitron	*Helitron*	RPA Y2 HEL	0	DHH	P, M, F
Maverick	*Maverick*	C-INT ATP CYP POL B	6	DMM	M, F, O

Structural features

Long terminal repeats; Terminal inverted repeats; Coding region; Non-coding region; Diagnostic feature in non-coding region; Region that can contain one or more additional ORFs

Protein coding domains

AP, Aspartic proteinase; APE, Apurinic endonuclease; ATP, Packaging ATPase; C-INT, C-integrase; CYP, Cysteine protease; EN, Endonuclease; ENV, Envelope protein; GAG, Capsid protein; HEL, Helicase; INT, Integrase; ORF, Open reading frame of unknown function; POL B, DNA polymerase B; RH, RNase H; RPA, Replication protein A (found only in plants); RT, Reverse transcriptase; Tase, Transposase (* with DDE motif); YR, Tyrosine recombinase; Y2, YR with YY motif

Species groups

P, Plants; M, Metazoans; F, Fungi; O, Others

Figure 6 | Summary of TE features (Wicker et al., 2007)

Structural Variations in Plant

Transposable Element mobilization can be a source of sequence variation among individuals of the same species if the new insertions have not either been eliminated by negative selection or genetic drift or have not gone to fixation, again as a consequence of positive selection or genetic drift. The extent to which TE movement contributes to sequence variation within a species is therefore dependent upon its timing. More ancient transposition events are much less likely to be polymorphic within a species than more recent ones. For decades it has been commonly accepted that the genome of an entire species could be represented by a single individual genome, and that intraspecific variability could be ascribed to single nucleotide polymorphisms (SNPs) and small insertions/deletions. Nowadays, thanks to next generations sequencing (NGS) and the reduction of sequencing time and costs, it has been possible to sequence several individuals of the same species and compare their genomes. This comparison showed in some species an unexpected intraspecific diversity and led scientists to introduce the concept of pan-genome, originally coined in the context of bacteria genomics; the pan-genome is composed by a core genome, which is shared by all the individuals, and a dispensable genome which is present in some individuals but not others and is not essential for survival (Figure 7). The dispensable genome is made up of structural variations (usually greater than 1kb) that together with SNPs and small insertion/deletions contribute to intraspecies variability. Structural variations (SVs) are often enriched in repetitive sequences and include genomic alteration such as insertions, deletions, duplications, inversions and translocations (Figure 8). TEs in higher plants are responsible for the majority of insertions and deletions events, but they can also accidentally be involved in erroneous recombination events responsible of translocations, inversion and tandem duplications.

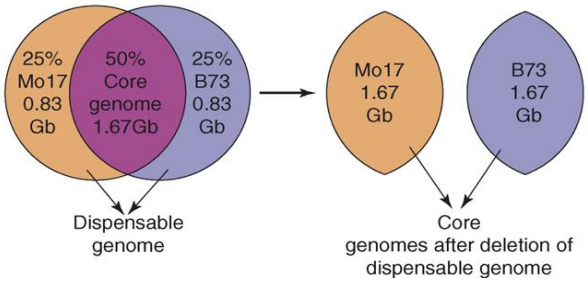


Figure 7 | Maize pan-genome (Morgante, De Paoli, & Radovic, 2007)

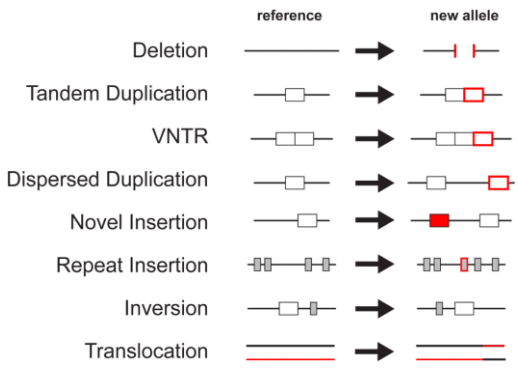


Figure 8 | Structural Variation (Hurles et al., 2008)

VITIS VINIFERA AS A MODEL TO STUDY THE RELATIONSHIP BETWEEN SVS AND EPIGENETIC VARIATION IN PLANT

In this study we propose grapevine, *Vitis vinifera*, as a model for the investigation of the epigenetic impact of structural variations for a series of advantageous features. First, the genome of a highly homozygous grapevine plant, sequenced by the International French-Italian public consortium, is available (Jaillon et al., 2007). The sequencing revealed a 486 Mb genome that is occupied by TEs in at least 41.4% of its length. The relatively small genome size and the moderate contribution of TEs to the whole sequence make grapevine an affordable and manageable genomic system for the detection and analysis of differential SVs. Indeed, thanks to Next Generation Sequencing technology, it has been possible to re-sequence dozens of varieties and their comparison revealed a large number of SVs suitable for evaluating the impact of differential TE content on the epigenome.

Grapevine is also a perennial plant generally reproduced by humans via vegetative propagation, a technique that preserves the distinct genotypes from meiotic recombination and genetic rearrangement. With this respect, grapevine provides the opportunity to investigate epigenetic determinants in a reproducible genomic environment as much as Arabidopsis and maize inbred lines do, with the additional advantage that immortalized heterozygous genomes can be analysed as well. Indeed, heterozygosity of cultivated varieties such as Pinot Noir, the one investigated in the present study, allow for the investigation of genomic structural variation and related epigenetic features within the same individual plant, thereby reducing genotype-dependent effects.

Last but not less important from an epigenetic perspective, vegetative reproduction of grapevine extends over generations, without the caveat of epigenetic resetting during gametogenesis and embryo development, the possibility of accumulation of epigenetics marks as a result of long-term interaction between genome and environment as well as between the genome and its epigenetically active repetitive component.

AIM OF WORK:
Analysis of epigenomic variability in grapevine and its relation with structural variation

Since the introduction of Next Generation Sequencing (NGS) and the consequent reduction of cost and time of DNA sequencing, it has been possible to re-sequence several individuals within the same species highlighting an unexpected genomic variability within the species. This variability not only involves SNPs but also larger DNA elements that severely impact the sequence structure of a genome and for this reason are often regarded as structural variations. Structural variations in higher plants are largely represented by TEs whose mobilization is generally repressed by specific pathways, including DNA methylation, that induce their silencing. Thanks to the implementation of the bisulfite conversion protocol combined with NGS, it is now possible to obtain whole genome DNA methylation maps at single-base resolution.

In plant methylation may occur in the CG, CHG and CHH contexts, where H may be A, C or T, with three different mechanisms. Depending on locations, DNA methylation may have opposite effect: in heterochromatin it is generally associated to transcriptional inactivity whereas in the transcribed region of genes, methylation in the CG context is associated to medium-to-high transcriptional levels.

In this study DNA methylation in grapevine will be analysed in relationship to gene and TE density using expression data that were already available in our research group. The silencing effect of DNA methylation represents a useful defense weapon against TE insertions: indeed their integrated sequence is generally highly methylated in order to prevent their transcription. Internal DNA methylation of TE sequences is often spread into the flanking regions and it may accidentally overlap with gene or regulatory regions and potentially interfere with their expression. An in-depth study of the major TE groups in grapevine has been carried out in order to evaluate both their internal methylation pattern and the potential effect on their flanking regions. Particular efforts have been made to investigate DNA methylation at hemizygous TE insertion sites, which provide a unique system to evaluate the potential epigenetic crosstalk between homologous chromosomes.

RESULTS

BS sequencing statistics

Genome-wide DNA methylation analysis was performed by bisulfite sequencing (BS-seq) through the Illumina platform using genomic DNA extracted from young leaf nuclei of the cultivated Pinot Noir variety.

The analysis included three biological replicates of the VCR18 clone, provided by Vivai Cooperativi di Rauscedo (VCR), for which additional genomic and transcriptomic information was already available. Replicate 1 was analysed using a different approach for BS-seq library construction relative to replicate 2 and 3 as a new protocol for bisulfite sequencing was introduced by Illumina later on during the course of this study. The former strategy, based on a consolidated protocol described for the first time in Arabidopsis by (Lister et al., 2008)) and recently reviewed by Urich et al. (2015), involves DNA mechanical fragmentation and bisulfite treatment after adaptor ligation. In contrast, in the new Illumina protocol utilized for replicates 2 and 3, DNA is immediately treated with sodium bisulfite that also contributes to its fragmentation and then random priming with tailed primers followed by tagged adapter pairing is used as a substitute for adaptor ligation. The latter approach is more straightforward and supposed to prevent bisulfite conversion biases ascribed to adapter interference.

BS-seq raw sequencing data were aligned to the genome sequence of the highly homozygous PN40024 genotype sequenced by the international French-Italian public consortium (Jaillon et al., 2007). Mapping was performed using the **ERNE-BS5** software package (Extended Randomized Numerical alignEr – Bisulfite 5, see Materials and methods), an in-house-developed aligning program suitable for efficiently mapping BS-treated reads against large genomes. Further development of ERNE-BS5 functionalities made the algorithm capable of exploiting single nucleotide polymorphism data to assign sequencing reads to specific haplotypes and this aspect was determinant for the choice of this tool among others available.

Mapping efficiency ranged between 71% and 74% between replicates.

Although the total amount of reads produced was comparable among the three replicates, replicates 2 and 3 showed a higher number of uniquely aligned reads to the reference genome

after removal of PCR duplicates. This result was consistent with the reduced number of PCR cycles required for the construction of those two libraries (see Materials and Methods).

ALIGNMENT DATA	Nugen Library		Illumina Libraries			
	rep1	%	rep2	%	rep3	%
TOTAL READS	323694704	100.0	315302298	100.0	316758960	100.0
ALIGNED READS	238708123	73.7	227357200	72.1	224746565	71.0
-> UNIQUE READS	192240859	59.4	159744532	50.7	161422010	51.0
-> DEDUPLICATED UNIQUE READS	31201395	9.6	88605783	28.1	80159607	25.3
LAMBDA CONVERSION (%)	98.38		98.60		98.60	
MEAN C COVERAGE (≥ 4X)	30x		35x		40x	
MEAN CG COVERAGE (≥ 4X)	9x		19x		21x	
MEAN CHG COVERAGE (≥ 10X)	20x		24x		26x	
MEAN CHH COVERAGE (≥ 10X)	19x		22x		25x	

Table 1 | BS-seq sequencing statistics

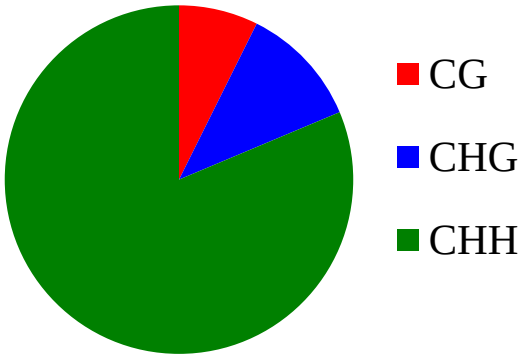
Since a greater amount of reads in both replicate 2 and 3 was available for the analysis of DNA methylation, and the two corresponding libraries had been prepared in parallel with the same protocol, we considered the possibility of merging their data in order to increase both the coverage and the number of cytosines suitable for epigenetic analyses. This approach was put in place after completing the separate analyses for each replicate and verifying the consistency of results. Hereafter, we will present results related to the merged data of replicates 2 and 3, which will be collectively indicated as “replicate 2+3”.

Proper bisulfite conversion represents a critical prerequisite for unbiased quantification of DNA methylation as failure in C-to-T conversion would lead to overestimation of cytosine methylation. To estimate bisulfite conversion efficiency in these preparations, a spike-in of unmethylated lambda phage DNA has been added to each sample. Following mapping of sequencing reads on the lambda genome reference, C-to-T conversion rates were calculated for each cytosine and the global average is reported in Table 1.

In all the three replicates, conversion efficiency was higher than 98.3% and deemed suitable for methylation analyses.

Cytosine methylation may occur in three different sequence contexts: CG, CHG and CHH, where H is any base other than G. While symmetric contexts of the CG or CHG type are maintained in a methylated state and reinforced by specific pathways (Cao & Jacobsen, 2002), CHH contexts include any generic non-symmetric cytosine and methylation of these sites is

thought to be preserved by repeated *de novo* methylation events, which may be tissue- or stage-specific and rarely tracked in a multicellular organism. Indeed, assuming a homogeneous nucleotide distribution, among the three contexts CHH is expected to be the most represented in the genome because of the degeneration of its code. In most plant genomes investigated thus far it is poorly methylated relative to its abundance but it still contributes significantly to total methylcytosines. In contrast, the CG dinucleotide is much less common in the genome (Figure 9) but it is highly methylated (Figure 11) and is the biggest contributor to total methylcytosines (Figure 12).



context	Number of Cs	fraction
CG	11953950	0.07
CHG	18382992	0.11
CHH	132102485	0.81

Figure 9 | Grapevine Genomic Composition of C contexts

The methylation value of a certain position is given by the ratio of the reads supporting the methylation (which show a non–converted cytosine) to the total number of the reads covering this position (showing either cytosine or thymine). Hence, methylation at a position may range between 0 or 1, these values representing completely unmethylated or methylated states respectively. Considering that a single cytosine residue can be either methylated or unmethylated, intermediate methylation levels arise from the fact that during the sequencing of multicellular samples, chromosomes from several cells and tissues are sequenced all together. Thus, the methylation value of each cytosine does not reflect the behavior of a single molecule as much as the average of many molecules, and may be expressed by a fraction as a result of differences in methylation between homologous haplotypes, somatic variation of DNA methylation or both.

When the statistical distribution of these methylation levels is considered, the three contexts show distinct profiles: the CG context shows a bimodal distribution with the two comparable modes positioned at 0 and 1; the CHG context shows an asymmetrical bimodal distribution where the mode located at 0 is much more frequent than the mode at 1; finally, CHH sites

show a very tight unimodal distribution with mode at 0 and a distribution skewed toward low values. Therefore, the majority of cytosines, especially of the CG and CHG types, show either absent or complete methylation, suggesting that the largest part of the methylome is not haplotype- or cell-specific. However, a not negligible fraction of cytosines show an intermediate methylation level, suggesting the presence of regional specificities in different haplotypes or cells.

The distribution of methylation levels affects the power of detecting differential methylation when comparing alternative conditions or haplotypes by sequencing and computational analysis. Depth of sequencing and *in silico* coverage filtering are critical parameters to consider for proper analysis. Ziller et al. (2014) suggests a minimum coverage of 5x when expecting methylation differences greater than 20% between two conditions and a minimum of 10× coverage if 10% methylation differences are expected. In most CG contexts expected differences should be close to 100% because of the strong and symmetric distribution, hence a 4x coverage threshold is considered a good compromise to preserve sensitivity while detecting differential methylation, minimize errors and at the same time discard as few cytosines as possible. In contrast, since the majority of cytosines of both CHG and CHH sites is either unmethylated or lowly methylated, then the expected difference in methylation may be very small and a 10x coverage is recommended. Hereafter, 4x and 10x minimal coverage will be required in each analysis for CG and CHG/CHH sites respectively.

Consistently with the higher number of single-match and deduplicated reads, replicates 2 and 3 showed in all the three contexts a higher number of cytosines covered with the minimal required coverage , compared to replicate 1 (Figure 10).

Moreover, replicates 2 and 3 exhibited higher methylation levels in the CG and CHG contexts, but not in CHH, relative to replicate 1 (Figure 11). This increase of methylation is not proportional and thus results into a different contribution of the three contexts to total methylcytosines (Figure 12b, which is a consequence of the increased number of CHG and CHH cytosines reaching the 10x minimal coverage in replicates 2 and 3.

The discrepancies in the global level of DNA methylation observed between replicates seemed to be ascribed to the use of two different protocols and in particular to the different impact of PCR duplicates, which resulted into very poor coverage in the case of rep1. We hypothesize that the different coverage of highly methylated repetitive regions provided by the two protocols may have contributed to the observed differences.

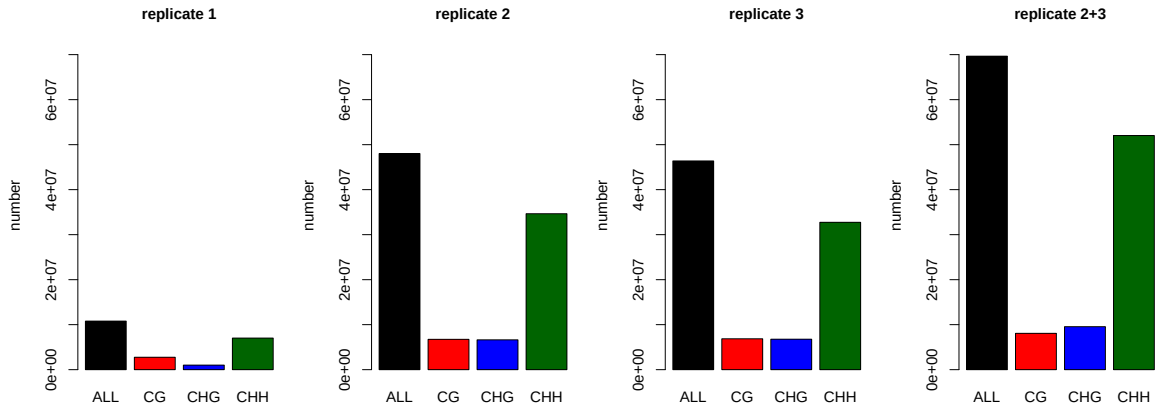


Figure 10 | Number of cytosines reaching the minimal coverage

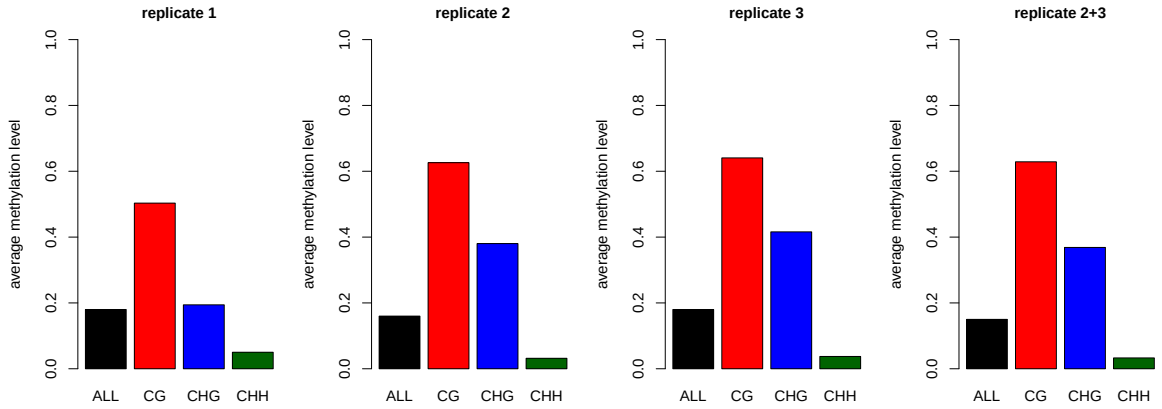


Figure 11 | Grapevine Distribution of methylcytosines in the three different contexts

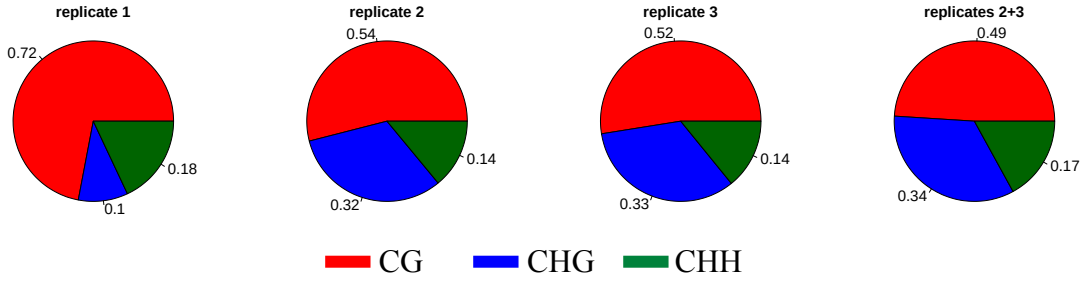


Figure 12 | Genomic Composition of C contexts utilized for methylation analysis in the three replicates

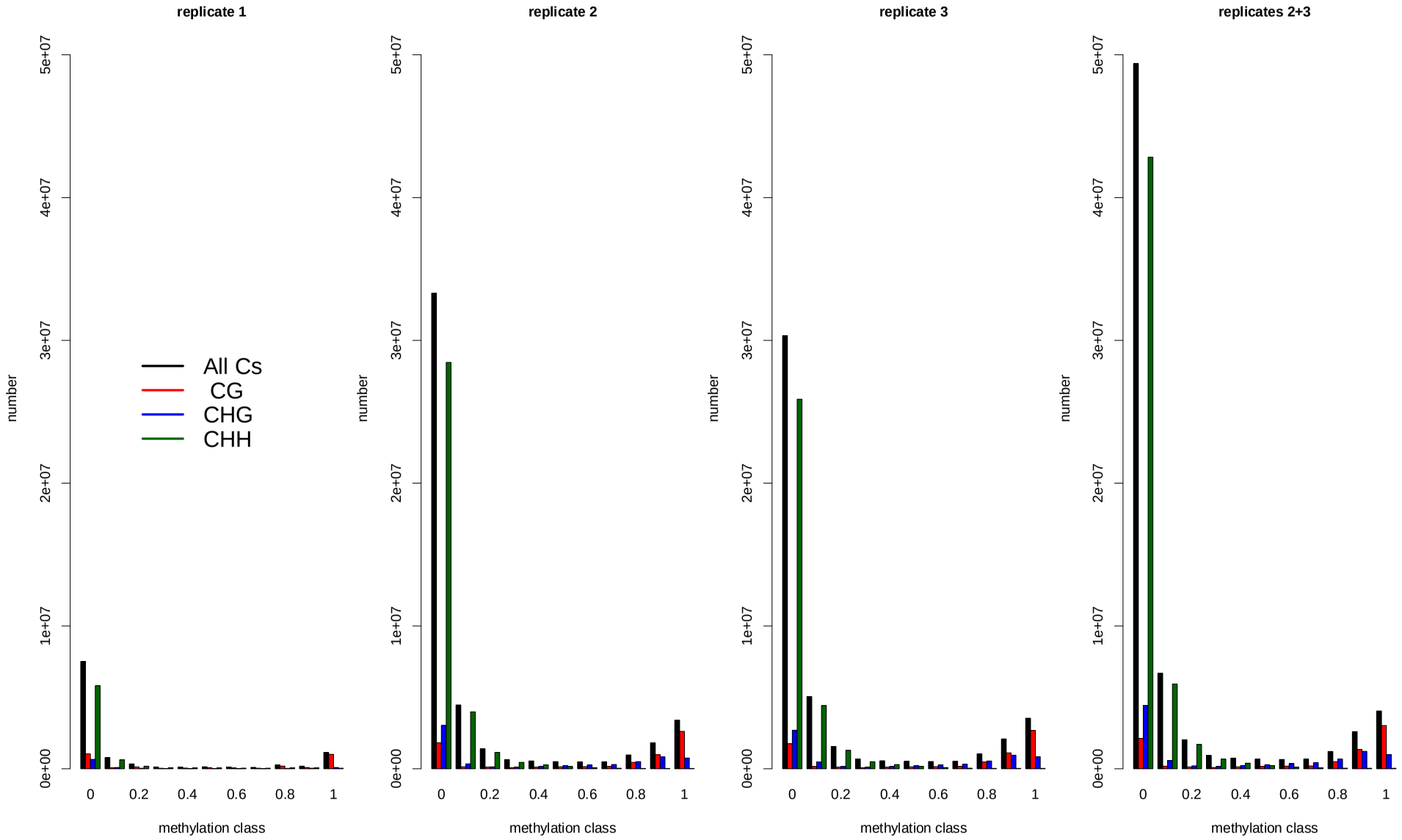


Figure 13 | Distribution of methylation values of all cytosines and for the three contexts separately

All the analyses shown in this work have been performed for all the three replicates separately and also for the merged replicates 2+3. Figures 10-13 show that replicates 2 and 3 are highly similar and when merging them, the number of cytosines with a minimal coverage increase without affecting the estimate of their average methylation level.

Hence hereafter only replicate 2+3 analyses will be shown in the main text whereas the replicate 1 will be shown in supplementary data.

Genome-wide DNA methylation analysis

Genomic landscape of DNA methylation and gene expression

As a repressive epigenetic mark, DNA methylation tends to be associated with heterochromatin. In small plant genomes with low repetitive sequence content (e.g. *Arabidopsis thaliana*) DNA methylation dominates centromeric and pericentromeric regions, although it may be also observed at much lower density across the chromosome arms, often in association with individual transposable element insertions (Lister et al., 2008). The scenario may be different in middle- or large-size genomes where the amount of repetitive DNA is higher and organized in a more complex way. The investigation of the grapevine DNA methylome started with a macroscopic examination of DNA methylation distribution along the chromosomes. All the 19 chromosomes of the genome were divided in 200 kbp windows and for each window the density of genes and TEs was calculated, as well as the average methylation level of CG, CHG and CHH contexts. G+C content in percentage was also used to provide information on base composition changes such as those determined by repetitive DNA in centromeres and pericentromeric regions. The results relative to replicates 2+3 are reported in a circos graph in Figure 14. At first glance, the macroscopic distribution of ^{5m}C is generally positively correlated with TE density and negatively correlated with gene density, in agreement with the expected heterochromatic localization of methylated cytosines. ^{5m}CG and ^{5m}CHG are more abundant in heterochromatic pericentromeric regions and so are TEs, consistently with previous studies (Lister et al., 2008 , Cokus et al., 2008). CHH methylation is generally extremely low and shows a limited increment of methylation level in correspondence of ^{5m}CG and ^{5m}CHG peaks in proximity of centromeres. Sharp transitions are often observed at this scale between highly heterochromatic TE-rich domains and gene-rich regions (see for instance chromosome 2, 15, 18, 19). However, this scenario can be hardly compared to the *Arabidopsis* landscape, where the methylome structure is shaped around the clear distinction between pericentromeric regions and chromosome arms (Lister et al., 2008), whereas the grapevine genome also exhibits patterns of moderate to high DNA methylation across entire chromosome arms.

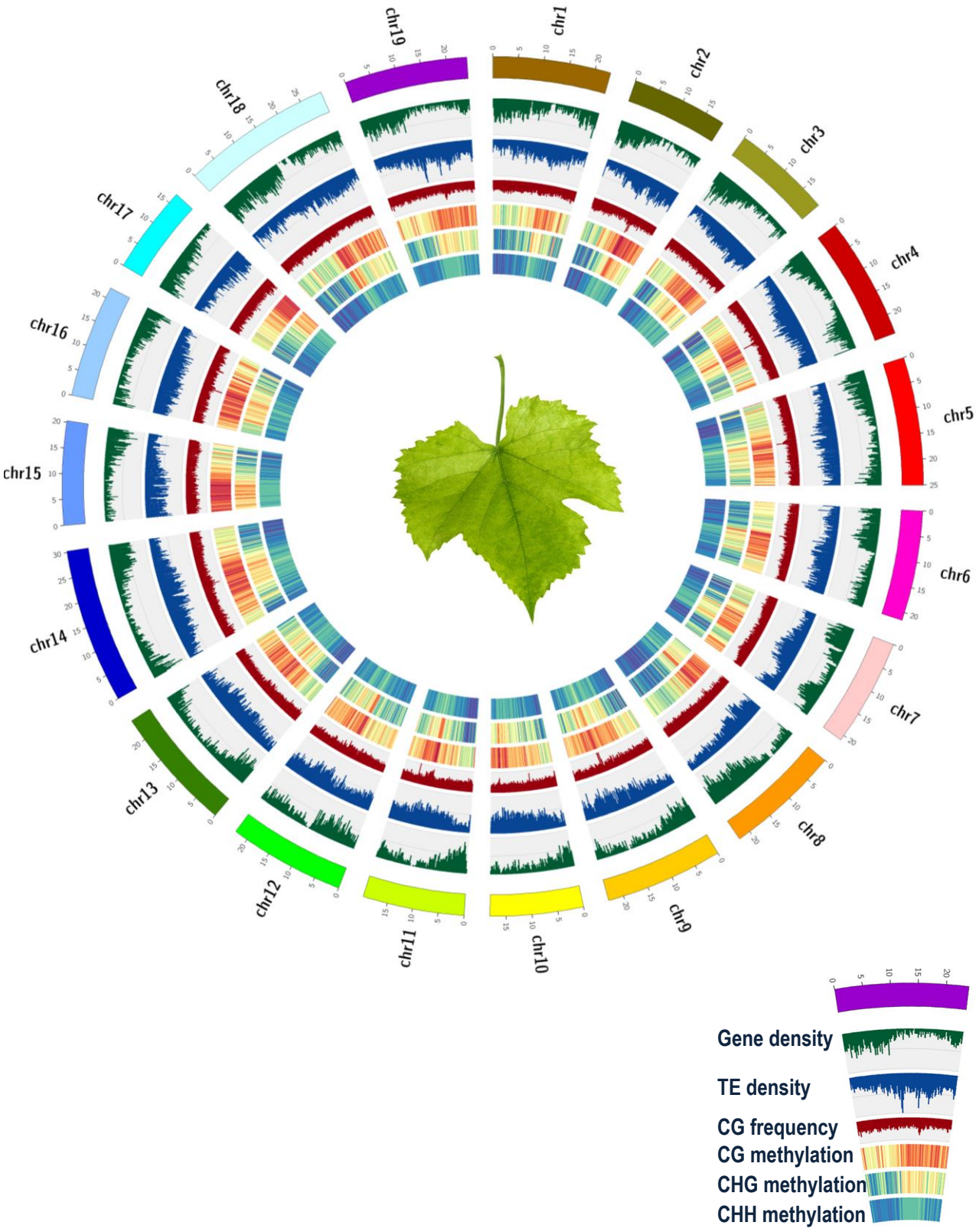


Figure 14 | Circos graph of Grapevine Genome and Methylome. Gene density and TE density, CG frequency and CG, CHG CHH average methylation level are relative to 200 kbp regions. Methylation is expressed in the form of heat map.

While the evolutionary origin of these domains remains to be clarified, one may ask whether such regional epigenetic patterns have an impact on the general level of gene expression at the coding loci involved. To address this question, the 200 kbp windows where grouped in 10 progressive classes according to their average CG methylation level and 10 additional classes based on average CHG methylation level. Coding genes resulted to be distributed across a wide range of regional methylation but tended to be enriched in a sub-compartment of the genome with intermediate CG methylation (20-80%) and low CHG methylation (20-50%) (Figure 15b). According to the general silencing effect of DNA methylation, genes located in more methylated regions tend to be less expressed than those located in low and intermediate methylated regions.(Figure 15a). Nevertheless, gene expression seem to be marginally affected by regional CG methylation, possibly as a result of the confounding effect of gene body methylation, which has a distinct functional role, opposed to gene silencing, and will be discussed later on in the present study. In contrast, gene expression appeared more sensitive to CHG methylation, with higher and more predictable expression levels in poorly CHG-methylated regions that are negatively correlated with TE density.

CHH methylation is extremely low in grapevine and all the genes belong to the lowest methylation class, however when considering centesimal classes between 0 and 0.1, CHH methylation display a similar pattern to CG and CHG.

A comparison of gene expression in leaves between two genetically close varieties such as Pinot Noir and Traminer (first degree relationship, Regner, Stadlbauer, Eisenheld, & Kaserer, 2000)) showed that the fraction of differentially expressed genes is significantly higher in lowly methylated regions, suggesting that epigenetic relaxation could be a prerequisite for transcriptome differentiation in a common genetic background(Figure 15c-d).

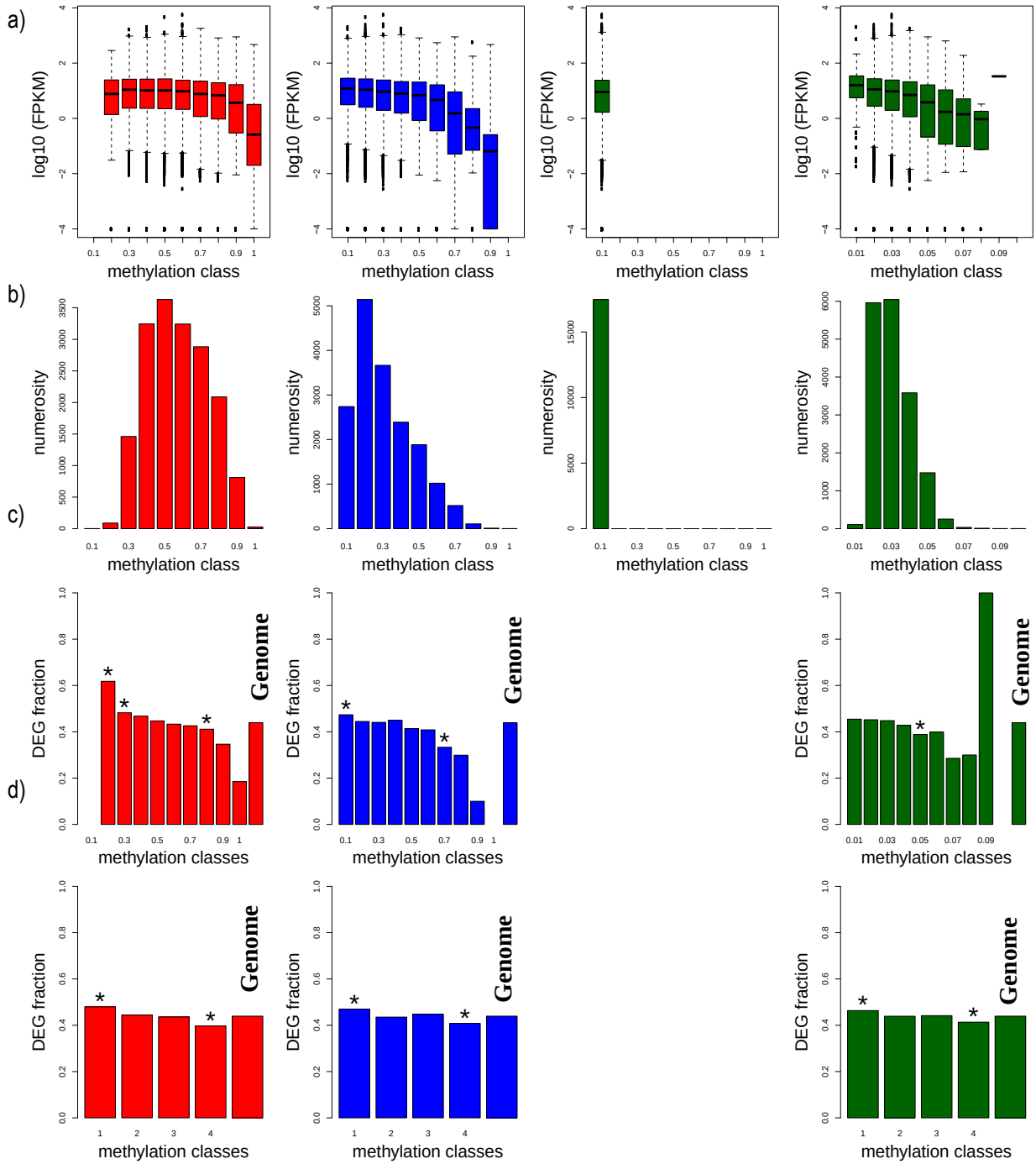


Figure 15
a) Gene expression rate on the basis of the regional methylation
b) Number of occunences pereach class
c) Frequency of significant differential expression between Pinot Noir and Traminer
d) Frequency of significant differential expression between Pinot Noir and Traminer in quantile classes. Lower numbers indicates lower methylation classes.
Significant differences with the DEG fraction of the whole genome are marked with * (Chi-squared test, p-value<0.05)

Methylation profile of transposable elements

Although the movement of transposable elements is a source of genetic and epigenetic variability, which is considered crucial both for evolution and adaptation to environmental conditions, uncontrolled TE mobilization may produce negative effects on organism fitness. As a result of genomic responses to TE proliferation, TE insertion sequences tend to be highly methylated in an attempt by the host to prevent their transcription and mobilization. To verify whether DNA methylation in grapevine TE bodies recapitulates the properties observed in other known species such as Arabidopsis (Cokus et al., 2008), soybean (Schmitz et al., 2013) and maize (Emberton, Ma, Yuan, SanMiguel, & Bennetzen, 2005; Palmer et al., 2003; Rabinowicz et al., 1999; Whitelaw et al., 2003), a meta-analysis of the most abundant TE groups has been performed.

Transposable elements represent 41.4% of the grapevine genome sequence according to an initial estimation (Jaillon et al., 2007). However, the TE annotation currently available was not accurate enough to identify full-length elements and their precise termini. Therefore, a new search was performed across the whole genome using Blast and Repeat masker and more stringent criteria, including 80% of nucleotide identity and a comparable length (between 80% and 125%) with a set of 202 TE sequences obtained from RepBase (Jurka et al., 2005), representing a non-redundant set of *Vitits vinifera* transposable elements, and an internal database of 467 *V. vinifera* TEs.

To enrich the LTR retrotransposon fraction, the tool LTR-finder was launched on the more recent TE annotation database. 4431 TEs fulfilling these requirements were identified and are reported in Figure 16. This set of TEs may not be an unbiased sample of all TEs present in the grapevine genome but may be biased towards the most recently inserted elements that have had less time to accumulate mutations and diverge from the original sequence.

The four different TE-groups analysed (Ty1-Copia LTR-retrotransposons, Ty3-Gypsy LTR-retrotransposons, non-LTR LINE retrotransposons, TIR DNA transposons) show differential enrichment in different genomic localizations: Ty3-Gypsy and TIR elements are more frequently located in intergenic regions, whereas LINE elements are frequently located both in intergenic regions and introns.Ty1-Copia show a more variable genomic localization. Insertions in the exonic compartment of predicted coding sequences are extremely rare, consistently with the expected selection against inactivating insertions, but still present. On

the other hand, almost one third of annotated full length TEs are located in introns, which represent 32% of the genomic sequences of grapevine (Jaillon et al., 2007).

Only a single Helitron element was found using the above mentioned criteria. Since a single element is not suitable for a meta-analysis, it has not been taken in consideration for the following analyses. SINE elements instead were completely absent from both the Repbase and internal databases.

Genomic localization of Transposable Elements					
Annotation	total	intergenic	exonic	intronic	mixed
Ty3-Gypsy	1412	1164	19	222	7
Ty1-Copia	927	867	7	48	5
LINE-1	526	209	0	316	1
TIR	1445	1247	17	156	25
total 4 groups	4310	3487	43	742	38
All TEs	4431	3604	43	746	38

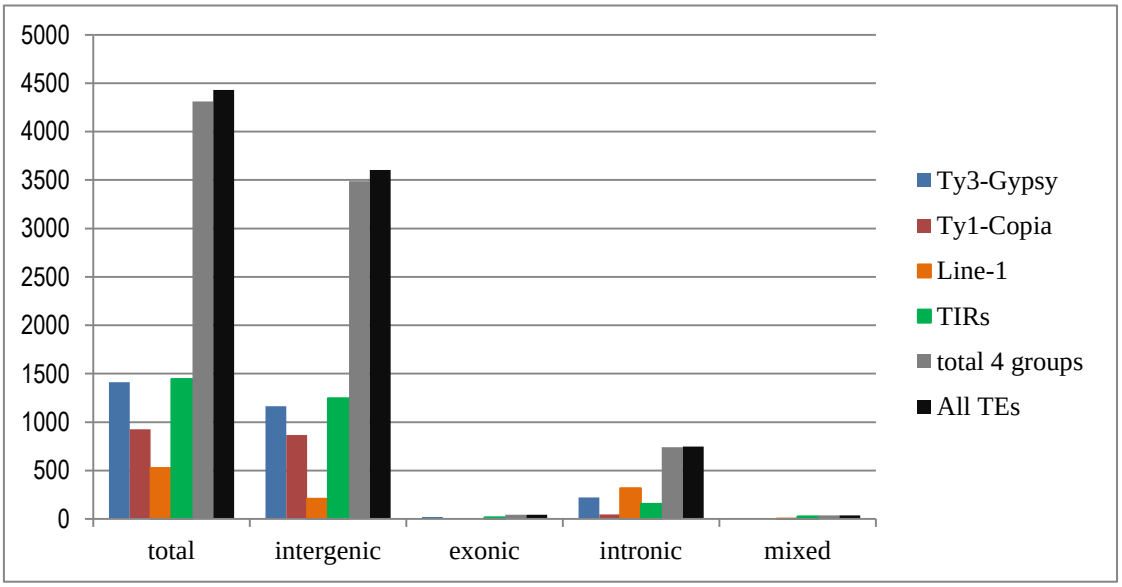


Figure 16 | Full-length TE distribution in grapevine genome in intergenic, exonic and intronic sequences.

The density distribution of TEs across the 200 kbp windows described above showed that the major contribution to centromeric TEs is given by Ty3-Gypsy while the other TE groups, including the intergenic located TIRs, did not show a preferential chromosomal distribution.

A meta-analysis of DNA methylation was carried out among the selected TEs to estimate the average methylation level within the TE body as well as in the regions flanking the insertions up to a distance of 2500 bp. To ameliorate TE termini annotations, TSD and terminal repeats

were identified wherever possible, using information described in Wicker et al., (2007). The LTR-finder software (see Material and methods) was used to predict both TSD and terminal repeats length and positions in the LTR elements. For TIR elements a dedicated R script was designed for the same purpose. In addition to the three compartments (TE body and both flanking regions) considered for all the TE groups, terminal repeats (IRs and LTRs in TIRs and LTR-retrotransposons respectively) were also represented as separate compartments when present. In each compartment CG, CHG and CHH methylation percentiles were computed independently and collapsed in a TE group-specific plot.

Consistently with previous studies that included similar analyses (e.g. in *Glycine max*, Schmitz et al., 2013), transposon bodies appeared highly methylated in both the CG and CHG context; CHH sites were instead extremely lowly methylated in all grapevine TEs (Figure 17).

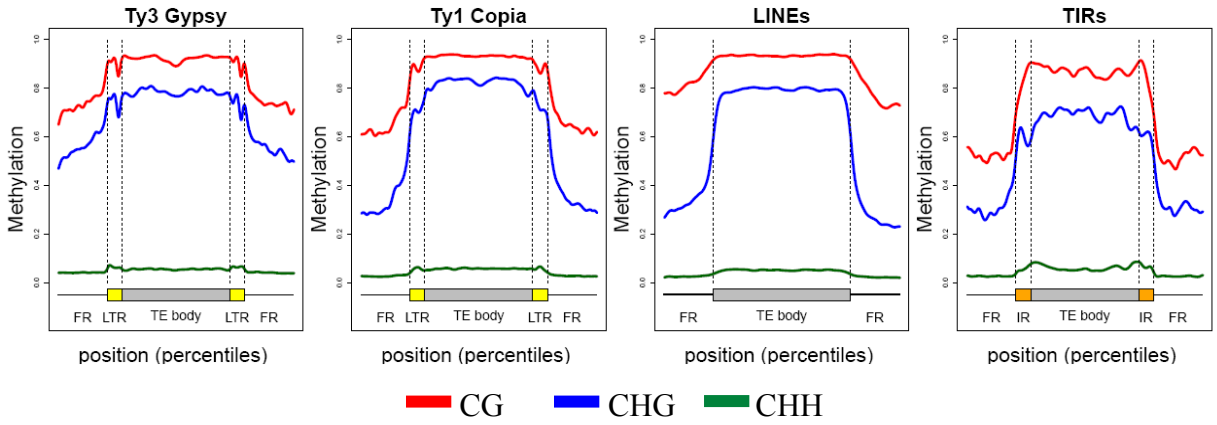


Figure 17 | Methylation profile in Transposon bodies of Ty1-Copia, Ty3 Gypsy, LINE and TIR elements. Methylation is expressed in percentiles in both TE bodies and 2.500 bp flanking regions. Where terminal repeats are present (Yellow in LTR-RTs and Orange in TIRs) their methylation profile is computed independently from the rest of TE body

Although different TE groups showed a comparable TE body methylation profile, the patterns revealed in the flanking regions seemed to be group-specific and suggested different epigenetic characteristics for the preferred genomic targets: class I elements are frequently located in region with high CG methylation and variable CHG methylation (higher for Ty3-Gypsy elements, lower for Ty1-Copia and LINE elements), while TIR elements are located in intermediated CG and moderate CHG methylation levels.

The transition from the TE termini to their flanking regions revealed a fast but progressive decay of methylation level, hence one can speculate that this pattern may be the effect of the propagation of the internal methylation into TE flanking regions for a few hundred bps.

Analysis of allele specific methylation

The analysis of the methylation profile in transposon bodies suggested the spreading of their associated methylation into their flanking regions. In light of the epigenetic mechanisms involved into TE silencing and considering in particular the leading role of heterochromatic small RNAs (hsmRNAs) in guiding the DNA methylation machinery toward the appropriate genomic targets, a question that may be raised is whether the diffusible nature of hsmRNAs and cofactors could offer a simple mechanism to propagate the epigenetic silencing *in trans* and in particular between homologous chromosomes, not dissimilar from what happens in the phenomenon known as paramutation in maize (Arteaga-Vazquez & Chandler, 2010; Patterson, Thorpe, & Chandler, 1993). The grapevine genomic system provides almost unique tools in order to clarify these aspects by comparing in the same individual plant the methylation profile of homologous regions and evaluating the epigenetic crosstalk between homologous chromosomes. Genomic sites characterized by hemizygous TE insertions allow for the investigation of methylation spreading in the same genetic local background in presence and absence of a given TE insertion.

The choice to perform this study in Pinot Noir was motivated by the fact that this variety shares one haplotype with the PN40024 reference, hence by analyzing the phase of Pinot Noir SNPs, it has been possible to define, by subtraction, the alternative haplotype of Pinot Noir and then proceed with the characterization of haplotype-specific DNA methylation.

The logics behind these analyses and the methods utilized are described below here and in the next chapters. In summary, this specific study involved the following steps:

1) Identification of Pinot Noir-derived genomic regions in the sequenced reference
2) Identification of Pinot Noir TE-dependent hemizygous structural variation
3) Estimate of haplotype-specific DNA methylation at TE hemizygous loci
4) Analysis of differential DNA methylation at TE-associated haplotypes

Identification of PN40024 regions derived from Pinot Noir

Accidentally during the process of self-fertilization of Pinot Noir, required to obtain a highly homozygous individual, a cross-fertilization by the cultivar Helfesteiner (obtained from the cross between Pinot Noir and Schiava Grossa in 1931, (Jaillon et al., 2007)) occurred, and thus, only a fraction of the reference genome is derived from Pinot Noir and useful for this analysis.

We had the availability of SNPs maps of both Pinot Noir and Schiava grossa obtained from whole genome resequencing (data not shown). These data were used to distinguish regions derived from the Pinot Noir cultivar (in which the presence of a haplotype shared with PN40024 determines a distinctive lack of homozygous SNPs between sequencing data from Pinot Noir and the PN40024 genome sequence), from Schiava (in which there is a lack of homozygous SNPs between PN40024 and Schiava) and the regions where the reference haplotype is shared with both varieties and thus not cannot be assigned unambiguously to either one (Figure 18). Hereafter, we will only consider the regions of the PN40024 reference derived from Pinot Noir and the regions shared by the two cultivars. As previously mentioned, the cultivar Pinot Noir and the PN40024 reference genome share one haplotype, here after called reference haplotype, while the non-shared haplotype in Pinot Noir is named alternative haplotype. Being the PN40024 sequence identical by descent or state to one of the two haplotypes present in Pinot Noir, the comparison of their genomes will reveal hemizygous structural variations in Pinot Noir. Only the PN40024 reference genome has been fully sequenced and assembled, hence two different strategies are required to find hemizygous SVs in the two different haplotypes. Both strategies involve the prediction of structural variations among the two haplotypes using bioinformatics tools and then the confirmation of the potential presence of the TE by sequence homology analysis.

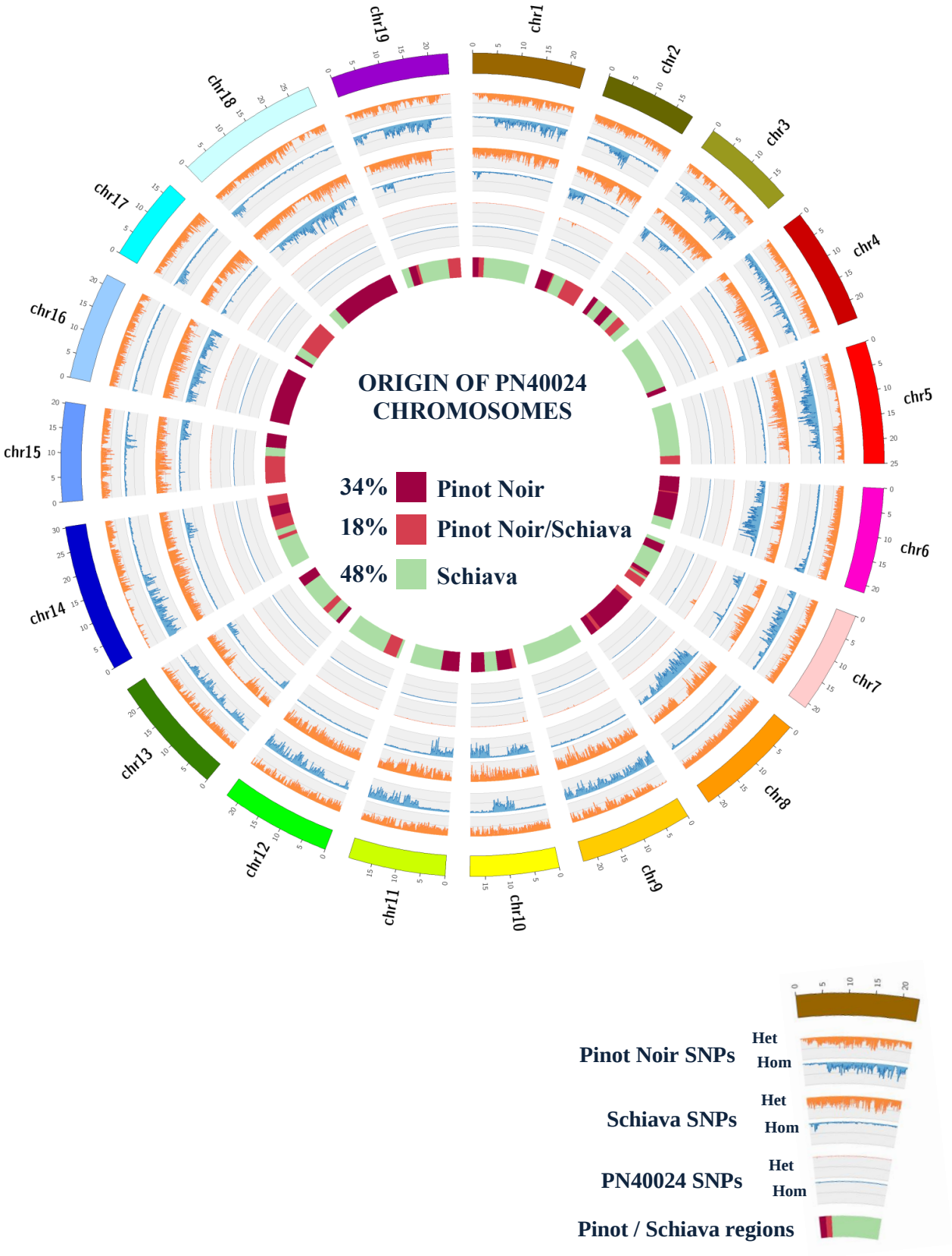


Figure 18 | Circos graph of Pinot Noir and Schiava's SNPs. frequency in 200 kbp windows

Identification of hemizygous structural variation in the Pinot Noir genome

Several tools are available for structural variation predictions. After comparing many of them (including DELLY, GASV, Pindel), DELLY and GASV were chosen because they minimize both false positives and false negatives (Gabriele Magris, PhD Tesis).

These programs were used to detect hemizygous SVs that are present in the reference haplotype but absent from the alternative one that correspond to heterozygous deletions in Pinot Noir when compared to PN40024. When aligning the sample reads on the reference, the insert size of the paired reads that contain the SVs will appear longer than the one expected based on the library insert size. Since the SVs are represented in the reference haplotype, the full sequence of the region involved in the SV is known and through an internal pipeline that includes several tools such as Blast, Blastx, Repeat Masker, LTR-finder it is possible to annotate the possible presence of TEs and also the type of TE involved. By combining the results of DELLY and GASV we found 2023 hemizygous deletions in the alternative haplotype (Table 2) that correspond to an equal number of genomic locations where the reference haplotype contains a fragment of DNA that is not present in the alternative haplotype. 68% of these events could be classified as full-length TEs based on the annotation pipeline mentioned above.

The SVs that are present in the alternative haplotype are absent from the reference genome sequence and correspond to heterozygous insertions in Pinot Noir when compared to PN40024. Unfortunately the above mentioned tools are not efficient in detecting insertion events involving large SVs such as those caused by TEs and we had to utilize a different approach. In this particular case, the bioinformatic evidence of a hemizygous SVs present only in the alternative haplotype consists in a certain number of orphan reads that co-localize around the putative insertion point (Figure 19). An internal pipeline developed by Sara Pinosio (unpublished data) in our research group verifies if the non-mapping reads of each pair can map on a database of TE termini and thus it can associate the SV to a particular TE group. With this method, we found 4127 hemizygous insertions in the alternative haplotype that correspond to an equal number of genomic locations where the alternative haplotype contains a fragment of DNA that is not present in the reference haplotype (Table 2). 90% of these events could be classified as potential full-length TEs based on the annotation pipeline mentioned above.

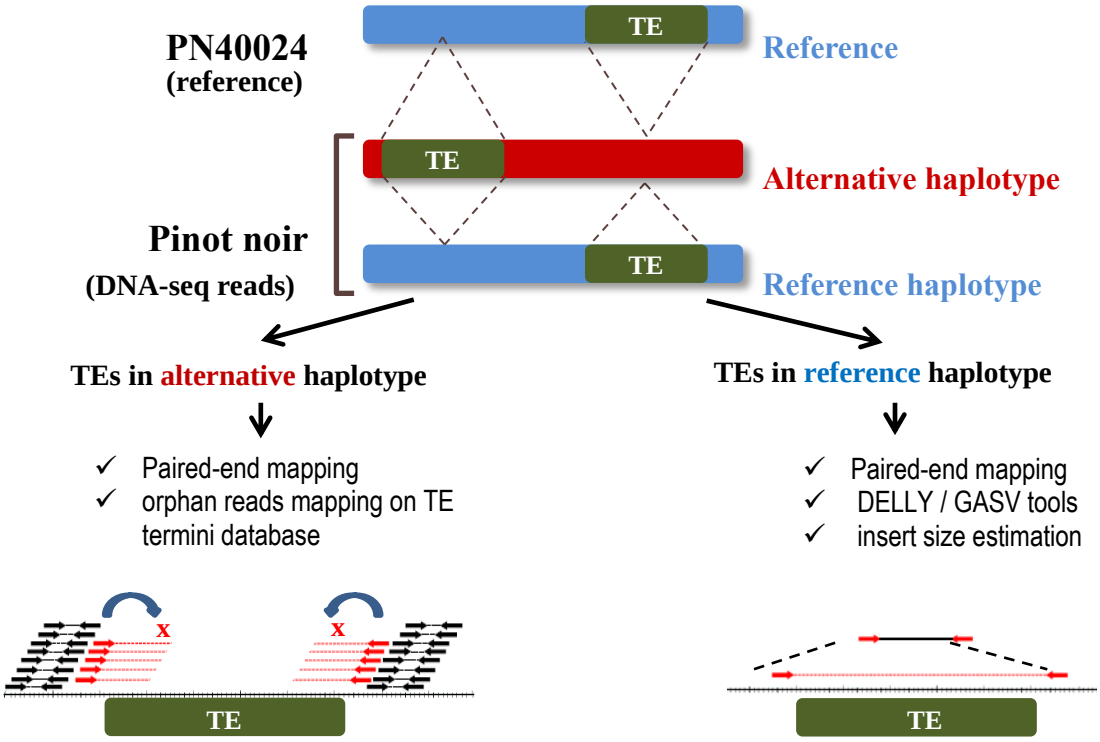


Figure 19 | Flowchart of hemizygous TE identification pipelines

A certain number of DNA elements responsible for structural variation in grapevine and present exclusively in the reference haplotype, could not be identified as known TEs and were flagged as "na*" (not annotated). Since there is no evidence of a TE being inserted, those sequences present solely in the reference haplotype may represent either a deletion in the alternative haplotype that is unrelated to TE activity or simply a falsely predicted SV. However in both cases, they may be used as negative controls (i.e., non-TE structural variants) for the analyses above described.

Not annotated SVs in the alternative haplotype (na*), whose sequences are known only at their termini, may also represent the insertion of a unclassified TE in the alternative haplotype. Hence, they cannot be used as negative controls as they might not represent *bona fide* non-TE structural variants . Few hundreds of SVs have been classified as solo-LTR. These elements originate by non-homologous intramolecular recombination between the two LTRs of a single element, which causes the loss of the internal TE sequence and one of the two LTRs leaving the other LTR on the genome flanked by TSDs. (Ma, Devos, & Bennetzen, 2004; Vitte, Panaud, & Quesneville, 2007)

Hemizygous solo-LTRs generally represent hence TE insertion events followed by deletion caused by the intramolecular recombination event and it will be possible to investigate

whether their effects on the methylation of flanking regions will be similar to those observed for complete elements or not.

Unfortunately solo-LTRs, in the alternative haplotype (**) are indistinguishable by full-length LTRs because only their extremities are known, hence they cannot not be analysed separately from complete elements in the following analyses. For solo-LTRs in the reference haplotype(**), their sequences have been blasted against the genome to verify the absence of a similar LTR within a 25 kb distance from their termini, in order to discard erroneous annotation of full-length LTR-retrotransposons.

Only two solo-LTR were discarded because of the presence of a similar LTR (80% of identity and of length) within 25 Kbs. Because of the low number of solo-LTRs belonging specifically to the RLC or RLX orders, all solo-LTRs have been analysed as a single group.

	Hemizygous TEs in the Reference haplotype		Hemizygous TEs in the alternative haplotype	
Annotation	Class	Frequency	Class	Frequency
Ty3-Gypsy	RLG	376	RLG	1585
Ty1-Copia	RLC	405	RLC	848
LINE-1	RIL	316	RIL	408
Mutator	DTM	79	DTM	243
hAT	DTA	72	DTA	129
CACTA	DTC	62	DTC	97
PIF-Harbinger	DTH	32	DTH	105
retrovirus	RLR	2	RLR	18
not annotated event	na*	269	na*	115
unidentified transposon	XXX	3	XXX	8
unidentified retrotransposon	RXX	7	RXX	19
unidentified LTR retrotransposon	RLX	15	RLX	93
incomplete Ty3-Gypsy	RLG_partial	88	RLG_partial	103
incomplete Ty1-Copia	RLC_partial	5	RLC_partial	1
unidentified solo LTR	RLX_solo_LTR**	16	RLX_solo_LTR**	0
solo LTR of Ty1-Copia	RLC_solo_LTR**	18	RLC_solo_LTR**	0
solo LTR of Ty3-Gypsy	RLG_solo_LTR**	215	RLG_solo_LTR**	0
unidentified class II transposon	DXX	12	DXX	26
unidentified TIR transposon	DTX	0	DTX	4
excessive N content (>80%)	N>80%	3	N>80%	2
Tandem Repeat content >80%	tandem_repeat	25	tandem_repeat	167
mitochondrial DNA	mitochondrion	3	mitochondrion	3
	total	2023	total	3974

Table 2 | Summary of all hemizygous SVs identified in reference and alternative haplotype respectively and their annotation into TE superfamilies where present, according to Wicker et al., 2007)

Localization of hemizygous TEs

Hemizygous TE loci specifically identified with the approaches above described (Figure19), recapitulate the genomic distribution observed for TEs mapped in the reference genome by sequence homology (Figure 16). Ty3-Gypsy elements are most frequently located in intergenic regions, consistently with the pericentromeric enrichment shown in Figure 21, Ty1-Copia are frequently located in intergenic regions but also in gene bodies, LINEs are predominantly found in introns, TIRs are mainly found in intergenic regions, but with no evident preference for pericentromeric regions (Figure 14).

Genomic localization of main hemizygous SV groups						
Annotation	total	intergenic	exonic	intronic	mixed	%
Ty3-Gypsy	1961	1692	47	156	66	32.7
Ty1-Copia	1253	674	71	427	81	20.9
LINE-1	724	113	24	547	40	12.1
TIR	861	644	31	121	65	14.4
not annotated	269	134	16	36	83	4.5
solo-LTR	249	223	2	17	7	4.2
total 6 groups	4975	3480	191	1304	342	83.0
All SVs	5997	4038	199	1370	390	100.0

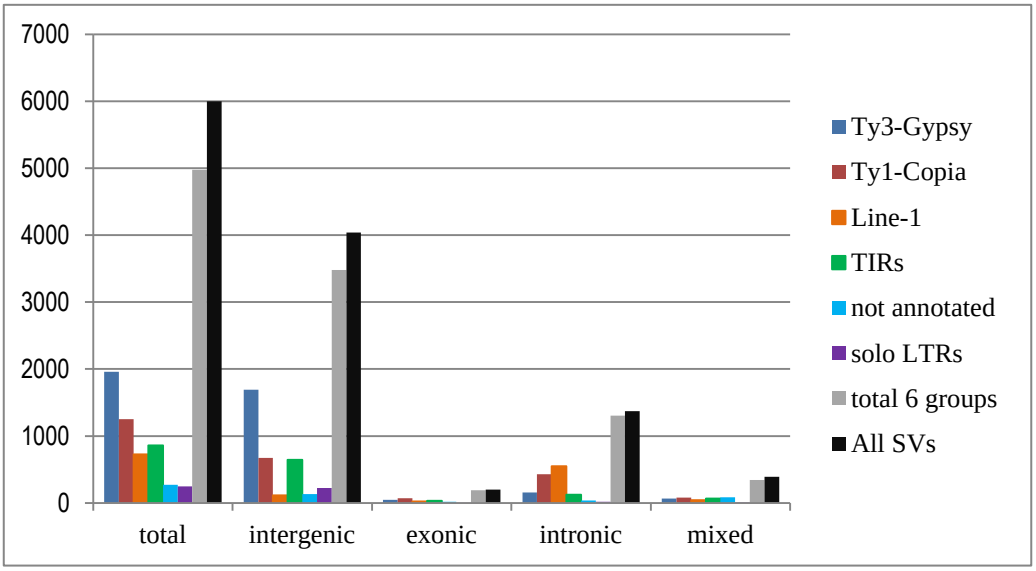


Figure 20 | Hemizygous SVs identified in grapevine across the 19 chromosomes. Unknown and random chromosomes were not considered in this analysis.

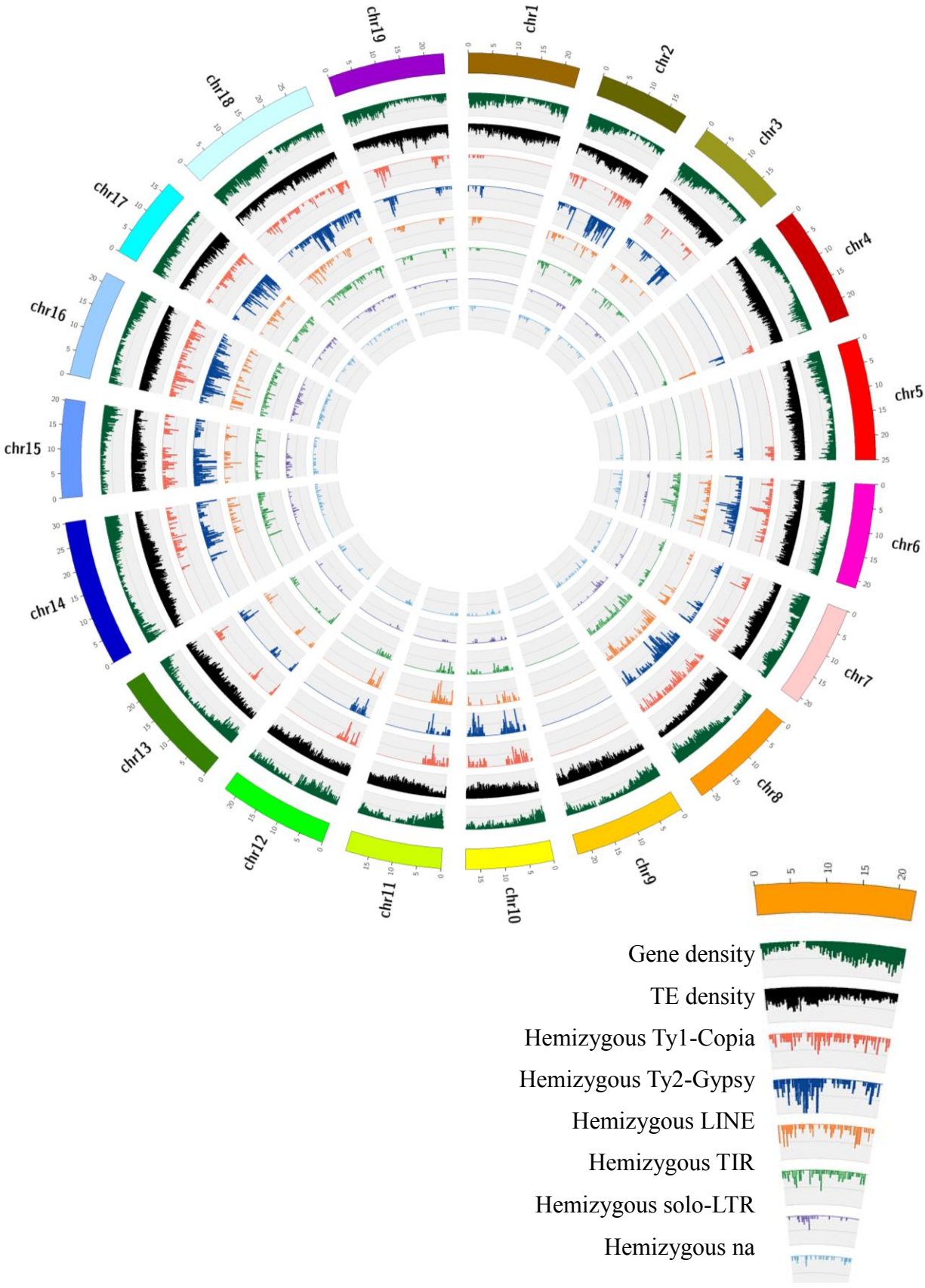


Figure 22 | Circos graph of TEs, gene density and of hemizygous TEs in pinot regions in the reference.

TIRs contribution to the total hemizygous TE is much lower than their contribution to the set of genomic TEs shown in Figure16. This discrepancy can be ascribed to the SV detection criteria, which were designed to accept elements longer than 1000 bp,, thereby excluding the fraction of small and non-autonomous elements such as MITEs.

Allele-specific analysis of DNA methylation spreading

Reads can be unambiguously assigned to a single haplotype as long as they carry at least one SNP that is heterozygous in the individual with the exception of C/T polymorphisms that may be confused with haplotype specific converted cytosines; hence in allele/haplotype-specific analyses of DNA methylation there is a general loss of information which makes a single TE analysis poorly informative per se. Moreover coverage is not always symmetrical in the two haplotypes as shown in a few specific TE examples shown in Figure 22, and thus even within the same 500 bp window, the contribution of cytosines in different positions may produce a biased result. To overcome these limitations a meta-analysis of the main TE groups was performed by comparing flanking regions of the haplotype that carries the TEs with the homologous regions on the other haplotype that is lacking the TE. For each TE, 2500 bp upstream and 2500 bp downstream from the insertion point were considered. Figure 23 displays the methylation profile of the TE-flanking regions in both the haplotype containing the TE and in that without the TE. Ty3-Gypsy flanking regions are usually highly methylated in the CG and CHG contexts both in the TE-carrying haplotype and in the one devoid of the TE, consistently with the frequent pericentromeric heterochromatic location shown in Figure 21 However a moderate increase in CG methylation and a stronger increase in CHG methylation are visible in the TE-carrying haplotype. Ty1-Copia elements instead tend to be found in regions with an intermediate methylation level on average and the haplotype carrying the insertion shows an important increase both in CG and CHG methylation. LINEs are found in highly methylated CG regions which are at the same time lowly methylated in the CHG context. Being CG almost saturated, the increase of methylation is only appreciable in the CHG context. TIRs element are found mainly in intergenic regions usually with a low methylation level but the increase due to the TE insertion in flanking regions is weaker compared to the effect observed in the other TE classes. Interestingly, CG methylation profile

of LINE elements in the unaffected haplotype is not flat as in the other TE groups, suggesting that they preferentially target regions with a high but not extended CG methylation peak rather than wide heterochromatic regions highly methylated in both CG and CHG like Ty3-Gypsy retrotransposons do. It should be noted that LINEs are preferentially located in introns and their behavior will be discussed in light of this observation in the next chapter. Taken together, these data suggest that when local methylation is not saturated, TE insertions generally induce an increase in the methylation level of both CG and CHG context.

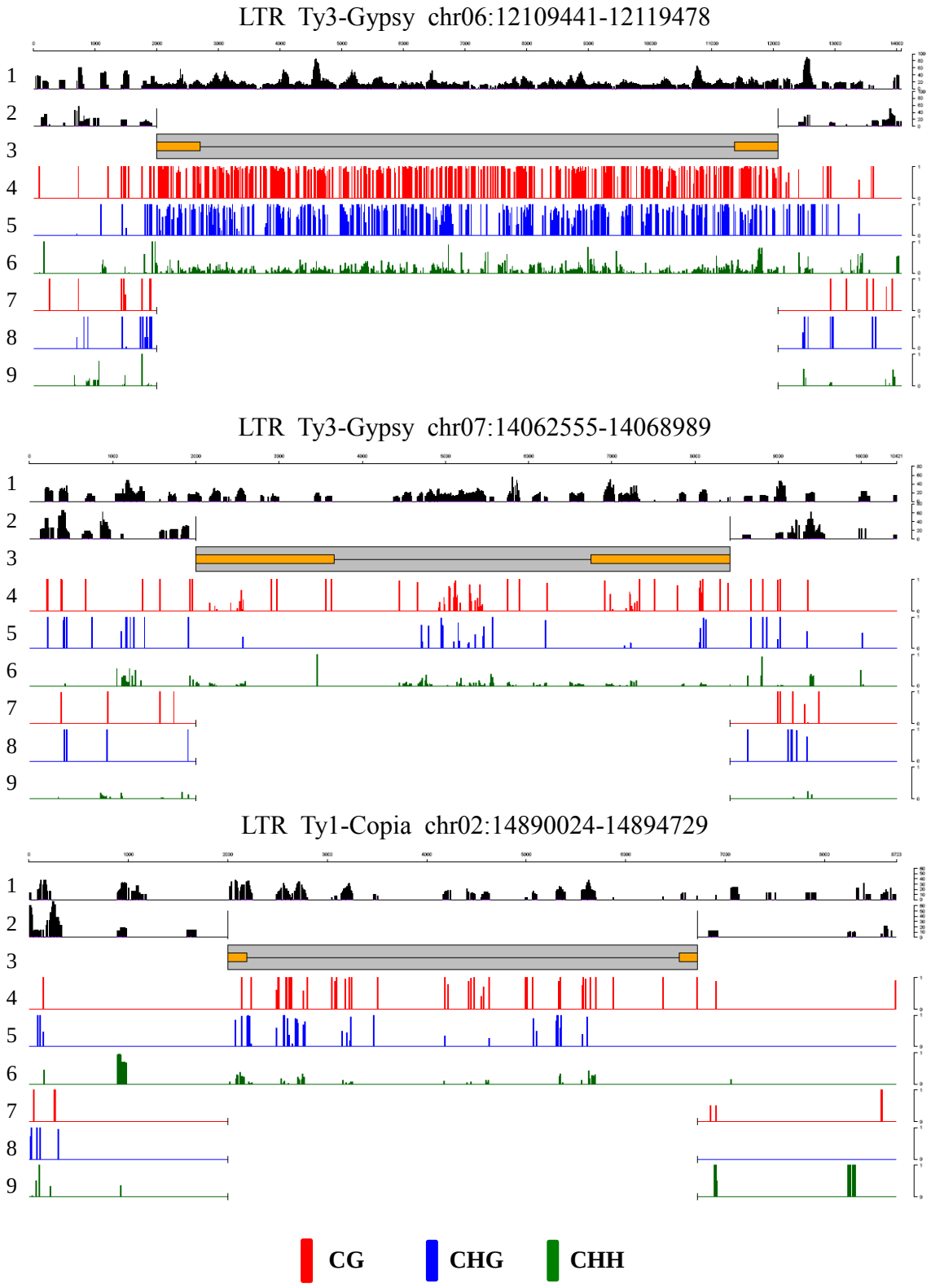


Figure 22 | Three examples of schematic representation of both hemizygous TE bodies and Flanking regions (2000 bp on either side) considering 10bps windows

Lines 1 -2) Scheme of coverage in TE and non-TE haplotype respectively

Line 3) Schematic representation of TE body including LTRs (orange)

Lines 4-5-6) Representation of CG, CHG and CHH methylation profile on TE-haplotype

Lines 7-8-9) Representation of CG, CHG and CHH methylation profile on non TE-haplotype

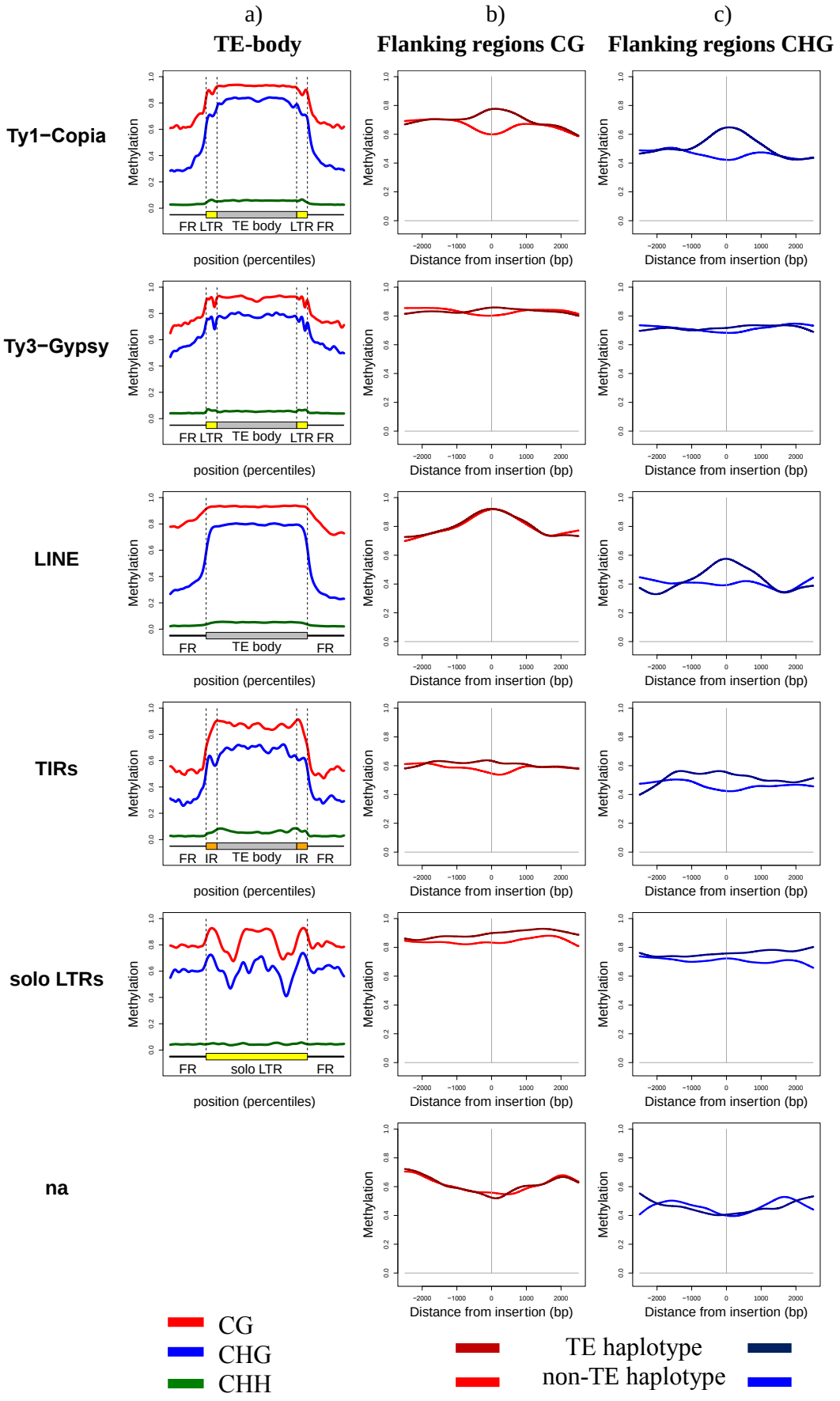


Figure 23 | DNA Methylation profile of both TE bodies and hemizygous TE flanking regions.

a) TE body methylation profile (see Figure 16)
b) Average CG methylation profile of TE flanking regions in bp from insertion point
c) Average CHG methylation profile of TE flanking regions in bp from insertion point

The effect generally decays within 1000 bp, and is observed exclusively in the TE-containing allele, in Ty1-Copia, Ty3-Gypsy and TIR elements suggesting a *cis* but not *trans* effect on DNA methylation of flanking regions for these TEs. LINEs show high CG methylation in both haplotypes and their behavior will be discussed in the next chapter.

Solo-LTRs display an internal methylation profile similar to that of Ty3-Gypsy terminal repeats. However, contrary to the LTRs of full-length elements, they do not seem to spread methylation into their flanking regions. Indeed, when considering their flanking region profile, the typical symmetrical bell-shape profile in the SV-carrying haplotype is absent. However, the SV-carrying haplotype seems to be constitutively more methylated than the unaffected haplotype all over the region with no variation relative to the TE insertion point .

Similarly, na sequences do not display spreading of methylation and, unexpectedly, they show a slightly higher methylation level in the SV carrying haplotype,

Since both solo-LTR and na are present exclusively in the reference haplotype, we hypothesized the existence of a haplotype-dependent bias in the methylation estimates. To confirm this hypothesis, the methylation profiles of hemizygous TE flanking regions were re-analysed for the major TE groups by separating those located in the reference haplotype from those located in the alternative haplotype, i.e. those detected as deletions from those detected as insertions (Figure 24a-b). Similarly to what was observed for the solo-LTR and na profile, the background methylation level of the reference haplotype, irrespective of whether it carries the TE (Figure 24a) or not (Figure 24b), is constitutively higher than the background methylation level in the alternative haplotype. This strand bias may be due to the fact that whereas the reference has been fully sequenced and assembled, the alternative haplotype have been reconstructed by replacing the nucleotides of the PN40024 reference that carry heterozygous SNPs in Pinot Noir. Small indels have not been utilized to discriminate the two haplotypes and, moreover it may be possible that not all Pinot Noir SNPs have been detected. Hence, there is a potential loss of alternative haplotype specific reads and consequently a bias on the measure of the methylation. To verify that the bias is equally distributed on the whole genome, 1000 5 kb regions have been analysed in both haplotypes and their values are reported in form of Boxplot in the figure 24c. This analysis confirms that the reference haplotype is constitutively more methylated than the alternative haplotype independently of TE presence/absence.

To obtain a more reliable haplotype specific methylation profile, the plots of the TEs in the reference and alternative haplotypes were computed separately and then combined 1:1, as shown in the Figure 24c.

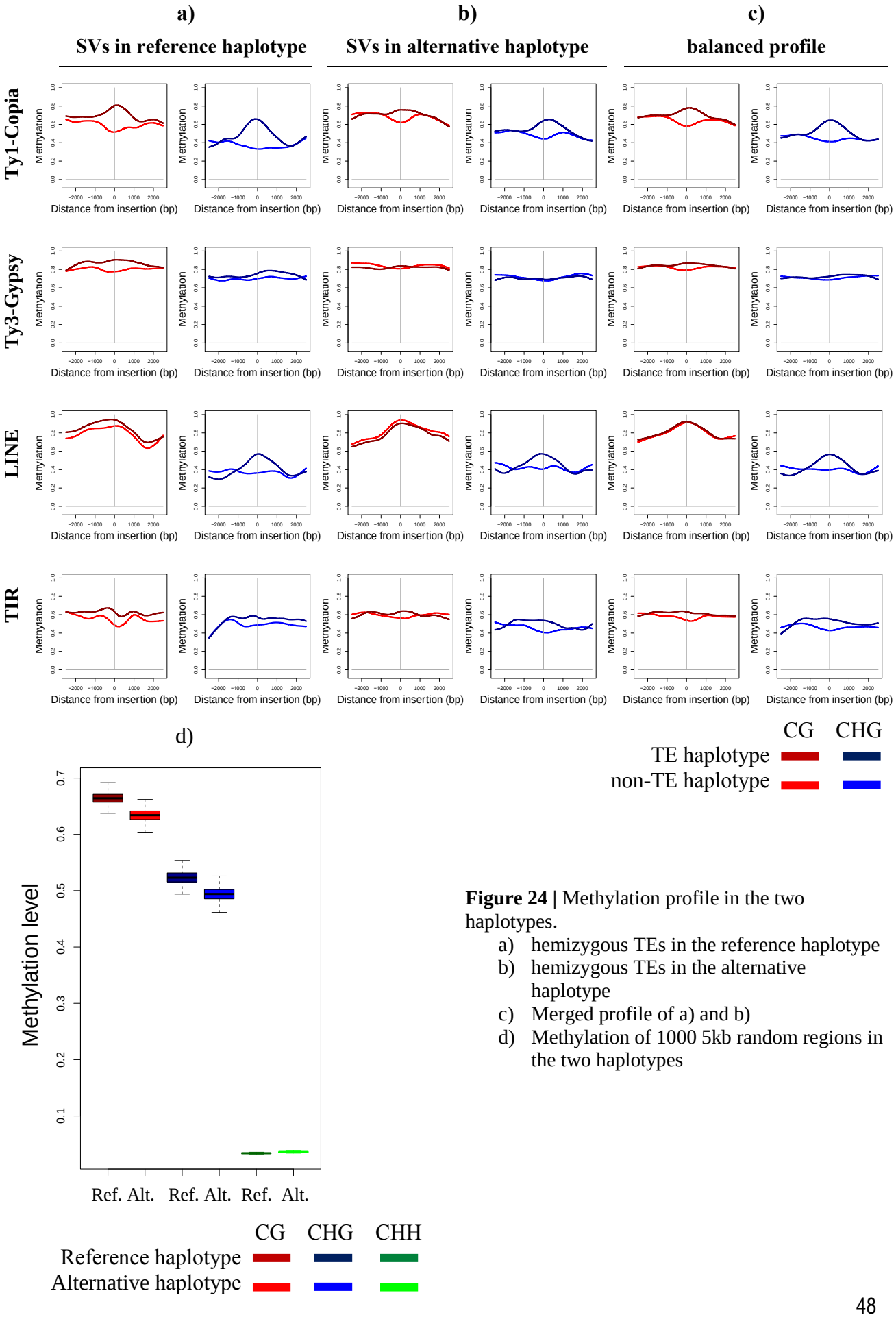


Figure 24 | Methylation profile in the two haplotypes.

a) hemizygous TEs in the reference haplotype
b) hemizygous TEs in the alternative haplotype
c) Merged profile of a) and b)
d) Methylation of 1000 5kb random regions in the two haplotypes

The CHH context has a very low methylation profile in both haplotypes and for all the TE classes there are no evidence of correlation with the TE presence/absence (data not shown).

To provide statistical support to data of differential DNA methylation revealed between homologous haplotypes, Fisher's Exact test, Chi-squared test and Wilcoxon Mann-Whitney test were performed. It should be noticed that in the four TE-groups analysed, the frequency of hemizygous TEs is higher in the alternative haplotype than in the reference haplotype (Table 2). Hence TEs are more often located in the haplotype whose methylation level is underestimated due to the coverage bias described above. As a result, this bias was expected to underestimate the degree of methylation increase observed in their flanking regions.

TE flanking regions were divided in ten 500 bp windows, 5 on each side of the putative insertion sites. Figures 25c-d reports the fraction of more methylated cytosines in the haplotype carrying the transposable element. Then for each cytosine with a sufficient coverage in both haplotypes a Fisher Exact test was performed to verify the hypothesis of a differential methylation. Figures 25e-f reports the fraction of significantly more methylated cytosines in the haplotype carrying the transposable element for each window (p-value <=0.01) If the null hypothesis was correct, for each window an equal distribution of more methylated cytosines in the two haplotypes would be expected, and consequently a fraction of approximately 0.5. On the contrary, the data show that at least within 500 bp from the insertion, the distribution of more methylated cytosines is in favor of the haplotype carrying the TE for both CG and CHG methylation in both class I retrostransposons and class II TIR DNA transposons.

Moreover, for each window the total number of cytosines showing higher methylation in one haplotype with respect to the other (whose ratio is shown in Figures 25c-d) was compared with the null expectation of equal methylation in the two haplotypes using a chi-square test. Figure 25g-h displays the –log (p-value) of such test.

Lastly, for each window of each TE the average methylation level for both CG and CHG context was calculated for each context independently, the list of values of the two haplotypes were compared trough Wilcoxon Mann-Whitney Test to test for deviations from the expectation of having two identical distributions of methylation levels. Figures 25e-f show the –log (p-value) of such test, in which significant p-values, (lower than 0.01) are above the horizontal black line representing a p-value=0,01. These data, consistently with the Chi-square and Fisher tests, confirm that the haplotypes carrying the TE are significantly more methylated at least within 500bp from the insertion point in both class I LTR retrotransposons and class II TIR DNA transposons, whilst in LINEs is significant only in the CHG context.

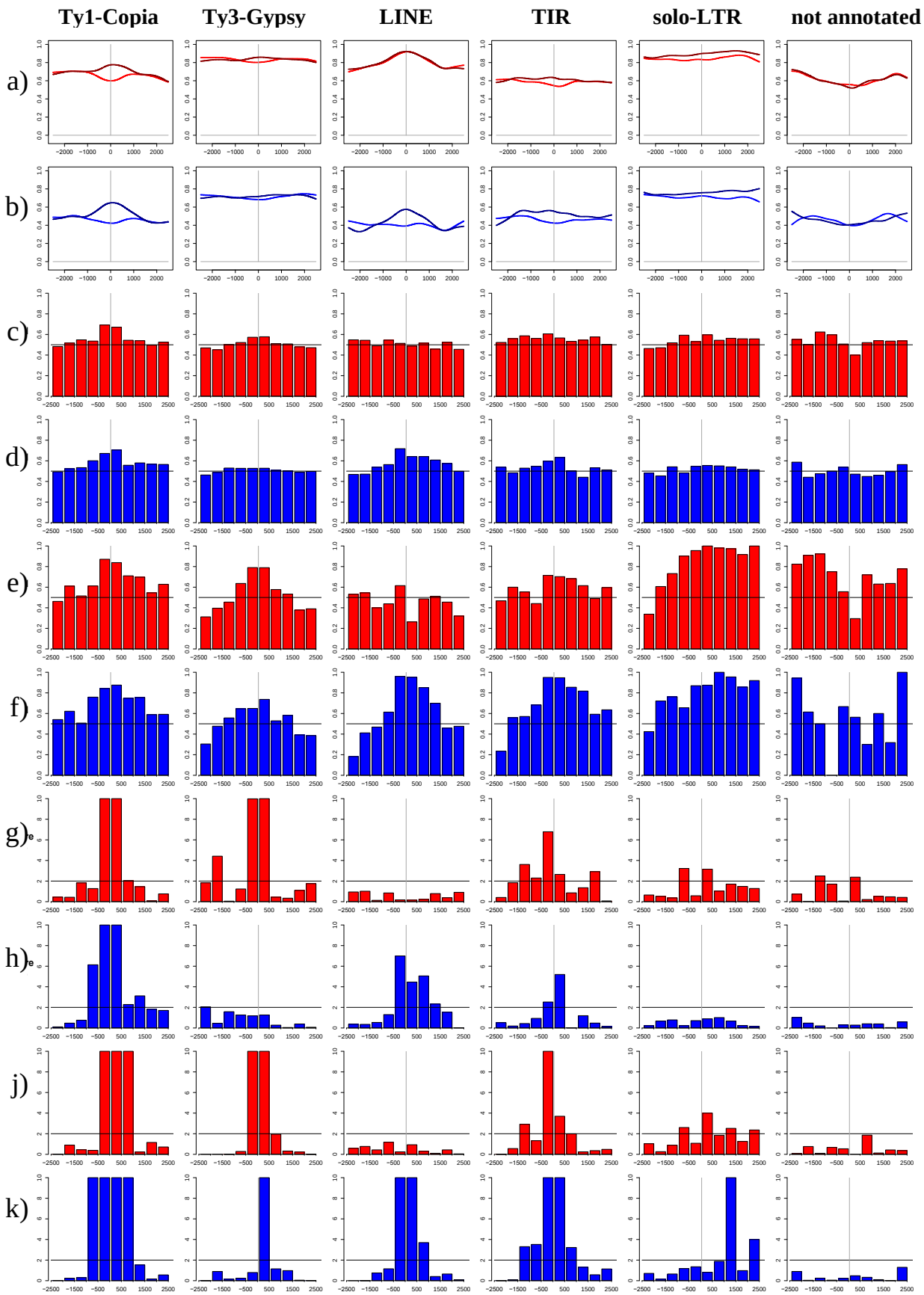


Figure 25 | See next page for the complete description

Figure 25 |
Statistical analysis of between-haplotype differential DNA methylation in the flanking regions of hemizygous TEs

a-b) Average DNA methylation levels (red: CG; blue: CHG) in the TE flanking regions: TE-carrying haplotypes (dark colour) and haplotypes devoid of TEs (light colour) are represented separately (see also Figure 23 for details).

c-d) Fraction of total Cs that are more methylated in the TE-haplotype. Values are reported for each 500 bp bin of distance from the insertion site within a +/- 2500 bp range; c: CG context; d: CHG context.

e-f) Fraction of total Cs that are significantly deviating from the null expectation of equal methylation in the two haplotypes (Fisher's Exact Test, p-value < 0.01). Values are reported for each 500 bp bin of distance from the insertion site within a +/- 2500 bp range; e: CG context; f: CHG context.

g-h) Deviation from the null expectation of equal methylation in the two haplotypes. Chi square Test log P values are reported for each 500 bp bin of distance from the insertion site within a +/- 2500 bp range; g: CG context; h: CHG context.

j-k) Deviation from the null expectation of equal methylation in the two haplotypes. Wilcoxon Mann-Whitney Test log P values are reported for each 500 bp bin of distance from the insertion site; g: CG context; h: CHG context.

Beyond the meta-analysis approach, additional methods were adopted for the representation of DNA methylation behavior within TE flanking regions. One of these strategies sought to disentangle and quantify the contribution of each single hemizygous insertion to the average tendencies revealed by the meta-analysis.

To minimize artefacts in the quantification of between-haplotype differential methylation introduced by differences in sequencing coverage, the average methylation value of a 2kb-wide region around the insertion point of both haplotypes was considered to evaluate at single-TE resolution the effect of the insertion on the flanking regions. The choice to restrict the analysis to a 2kb window was supported by the previous observation that TE-induced methylation on flanking regions is generally negligible after 1000 bp on either side (Figure 2-25). The distribution of methylation differences between the haplotype with and the one without the TE at each insertion site was represented by a histogram where the differences are sorted in classes and positive classes indicate higher levels of methylation in the TE-carrying haplotype. The histogram reports occurrences of insertion sites belonging to each class so that deviations from a symmetrical distribution are symptomatic of differential methylation (Figure 26-31, column a). In addition, as histograms of methylation differences do not show absolute methylation levels, haplotype methylation values for each individual insertion site

were plotted in a dotplot (Figure 2 26-31, column b), with the x and y axes representing the methylation levels of the TE-carrying and the unaffected haplotype, respectively. Points underneath the bisector line indicate higher methylation levels in the TE-haplotype. To provide more robust evidence, the Wilcoxon Mann-Whitney test was performed to test the hypothesis of a differential distribution of methylcytosines among the two haplotypes within the 2 kb region. The same analyses and graphical representations mentioned above were replicated for the subset of TEs showing a significant methylation difference according to a p-value <=0.01 (Figures 26-31 columns b and d).

Ty1-Copia

In the case of hemizygous Ty1-copia insertions the histogram of methylation differences (Figure 26a) shows a tendency toward positive values meaning higher methylation in the TE haplotype, consistently with data shown in Figure 23.

However, a considerable number of TEs exhibits a similar methylation level in the CG context. Figure 26c shows that when this happens the two haplotypes are often both saturated in methylation whereas regions showing different and not saturated CG methylation are often more methylated in the TE-haplotype. CHG methylation is also increased in the TE haplotype in most of the cases, albeit generally low in magnitude. The subset of P value-filtered insertion sites confirms that wherever a methylation difference is significant, the haplotype carrying the TE is the most methylated. By considering CG and CHG together, the increase of methylation in the TE-haplotype becomes more evident. Differences in CHH methylation are minimal and are equally distributed among the two haplotypes and thus no methylation increase can be associated to TE insertion. These small differences may reach significance because of their higher numerosity compared to CG and CHG, but are always equally distributed among the two haplotypes. The variable pre-existent CG level showed in the dotplot is consistent with the variable genomic location of Ty1-Copia elements shown in Figure 23, and also confirms that Ty1-Copia insertions generally elicit an increase of methylation in both the CG and CHG contexts where not already saturated

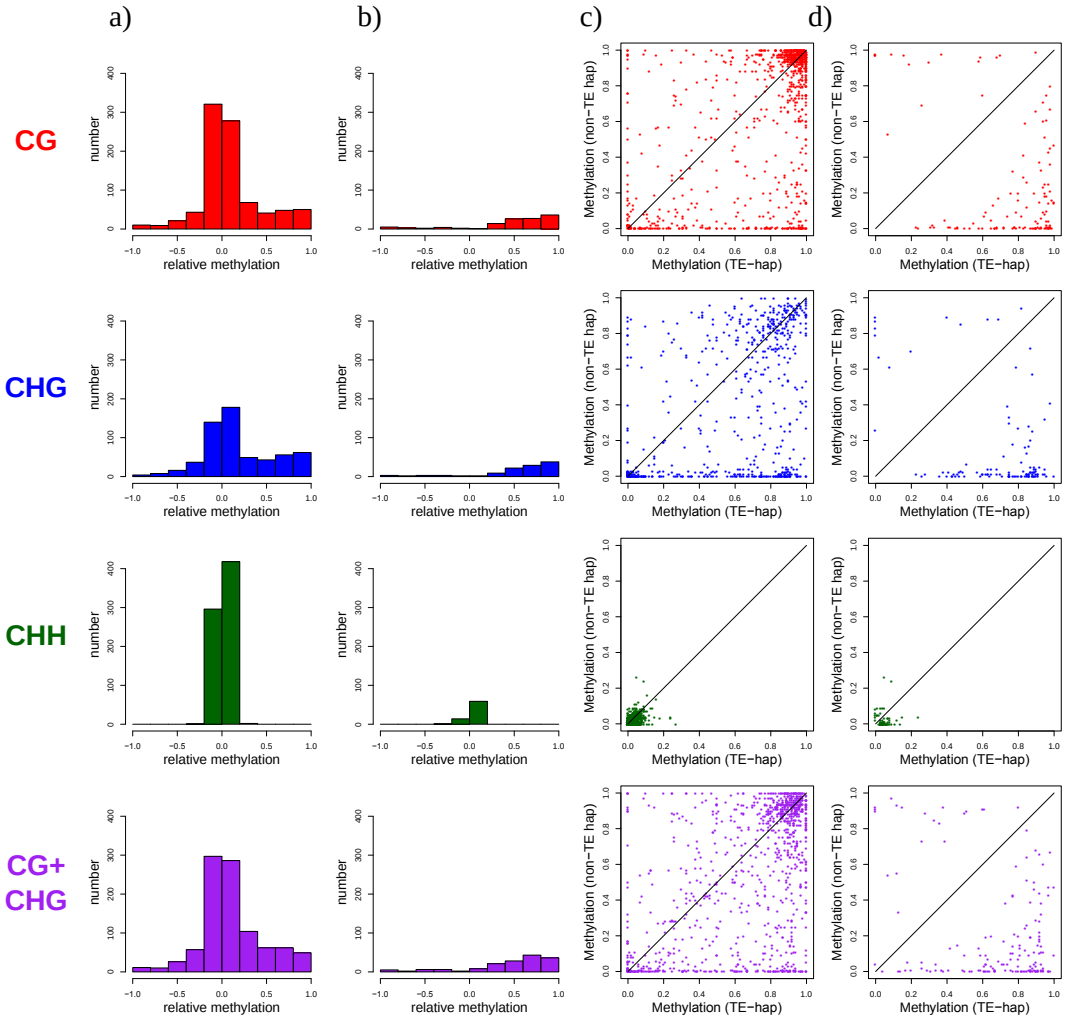


Figure 26 | Individual Ty1-Copia flanking regions analyses. Methylation is calculated over a region of 2kb around the insertion point in both haplotypes
a) Distribution of the difference of methylation values between TE haplotype and non-TE haplotype
b) Subset of Wilcoxon Mann-Whitney tests p-value positives (<0.01) of a)
c) Dotplot
d) Subset of Wilcoxon Mann-Whitney tests p-value positives (<0.01) of c)

Ty3-Gypsy

The majority of insertion sites involving Ty3-Gypsy retrotransposons, exhibits negligible methylation differences between the two haplotypes in all the contexts. Considering the patterns observed for Ty1-copia elements, this result is particularly manifest for the CG and CHG contexts. As suggested by the high density of dots in the top-right corner of the dotplot (Figure 25c), which represent highly methylated regions in both haplotypes, the excess of sites with no apparent TE effect could be a consequence of the preferential pericentromeric

and hererochromatic localization of Ty3-Gypsy elements, which would integrate in already methylated regions. The combination of CG and CHG contexts consolidates the result obtained for the two separate contexts, whereas CHH methylation, generally present at very low level, does not show any correlation with presence/absence of TE.

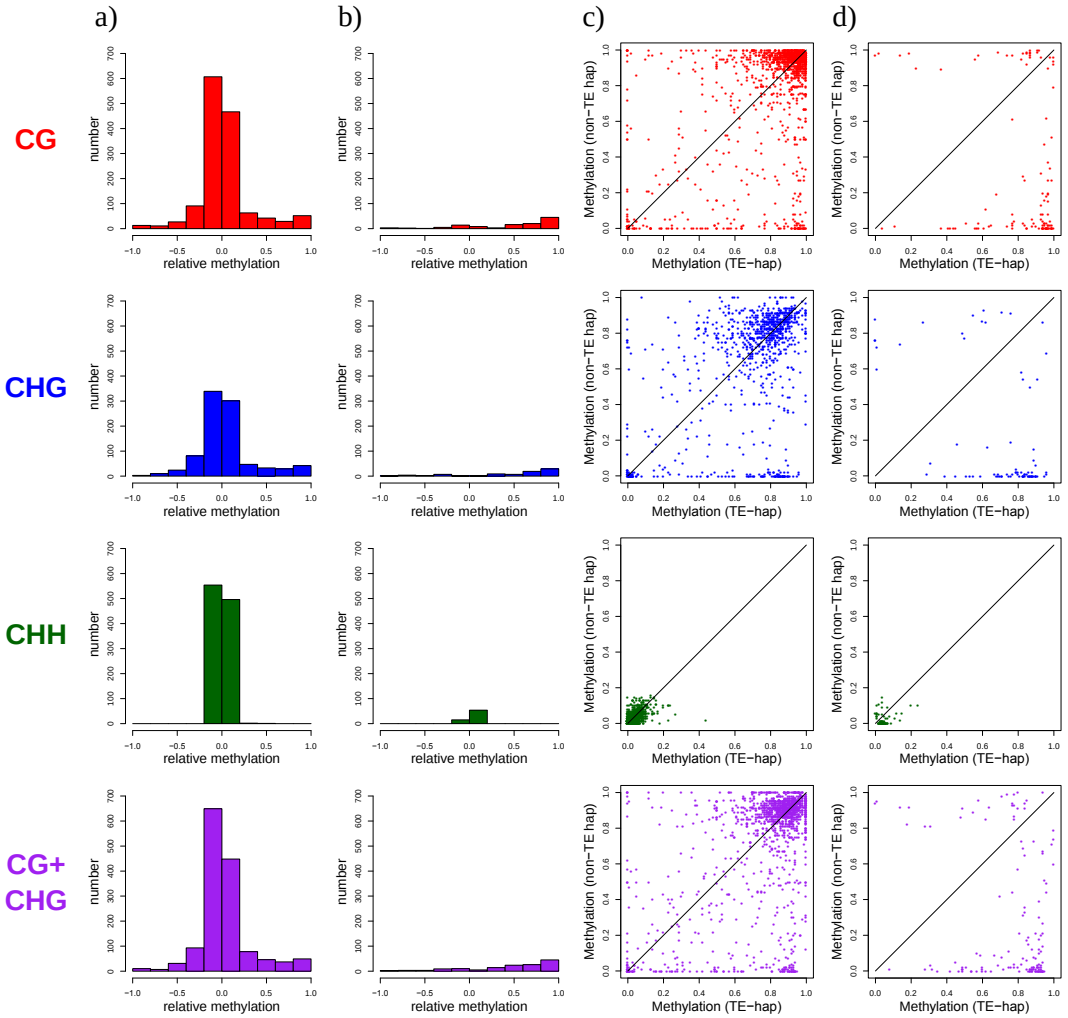


Figure 27 | Individual Ty3-Gypsy flanking regions analyses. Methylation is calculated over a region of 2kb around the insertion point in both haplotypes
a) Distribution of the difference of methylation values between TE haplotype and non-TE haplotype
b) Subset of Wilcoxon Mann-Whitney tests p-value positives (<0.01) of a)
c) Dotplot
d) Subset of Wilcoxon Mann-Whitney tests p-value positives (<0.01) of c)

LINEs

LINE elements show a peculiar enrichment in regions that are saturated in the CG context and lowly methylated in the CHG context.. The majority of these elements show an increase of

CHG methylation in the TE haplotype, according to Figure 23, while the CG methylation is high in both haplotypes and only a small increase in methylation seems to be caused by the TE insertion, as shown in the Figure 26a. These data are consistent with their general intronic localization, which is compatible with a high level of CG methylation that does not negatively affect gene expression, as discussed in the next chapter.

CHH methylation does not seem to be affected by TE-insertions, as observed for LTR-retrotransposons.

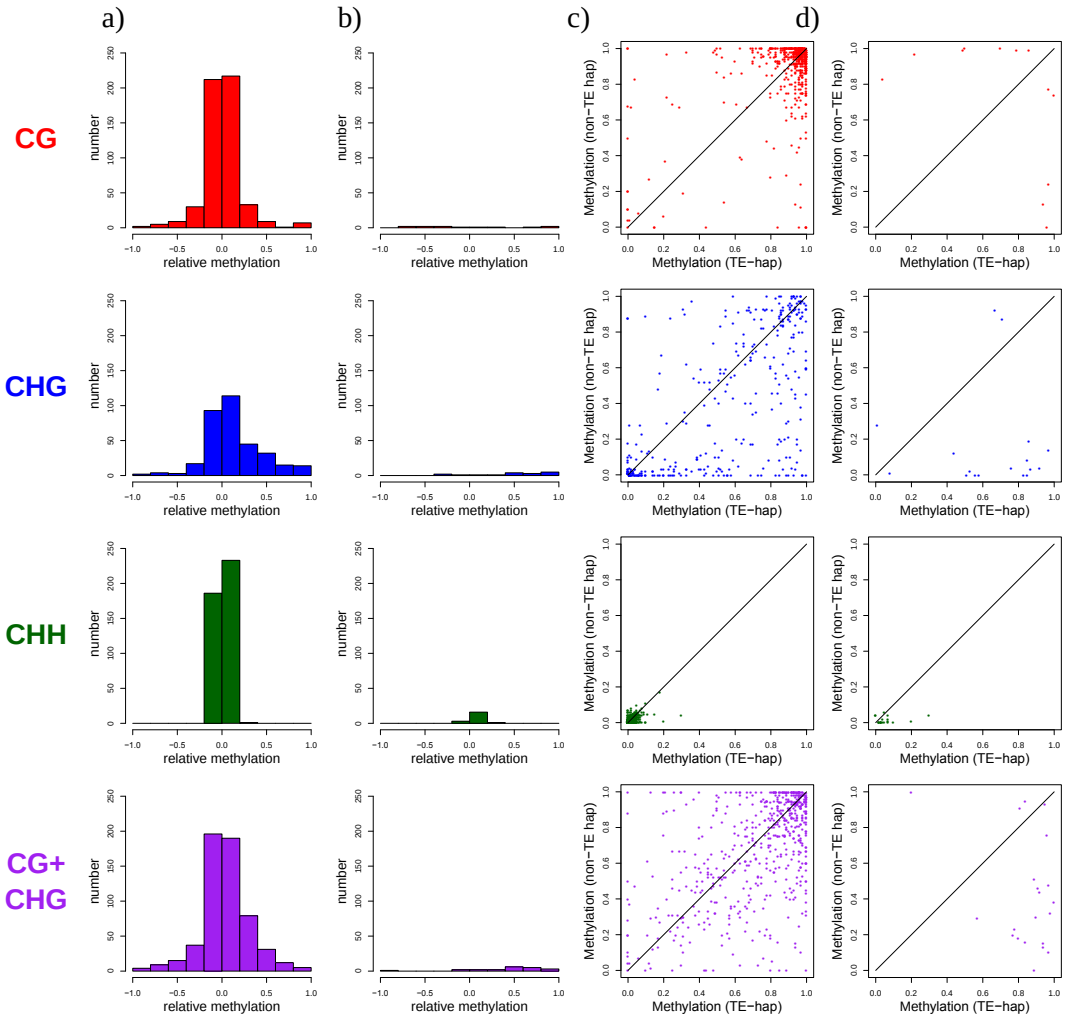


Figure 28 | Individual LINE flanking regions analyses. Methylation is calculated over a region of 2kb around the insertion point in both haplotypes
a) Distribution of the difference of methylation values between TE haplotype and non-TE haplotype
b) Subset of Wilcoxon Mann-Whitney tests p-value positives (<0.01) of a)
c) Dotplot
d) Subset of Wilcoxon Mann-Whitney tests p-value positives (<0.01) of c)

TIRs

TIR elements, in accordance with the observed intergenic but not pericentromeric preferential insertion (Figures 19 and 20), show a wide range of pre-insertion methylation levels, which may be increased by the TE insertion if not already saturated, in particular in the CHG context or considering CG and CHG together. However, TIR elements show a weaker effect in terms of methylation increase than retrotransposons.

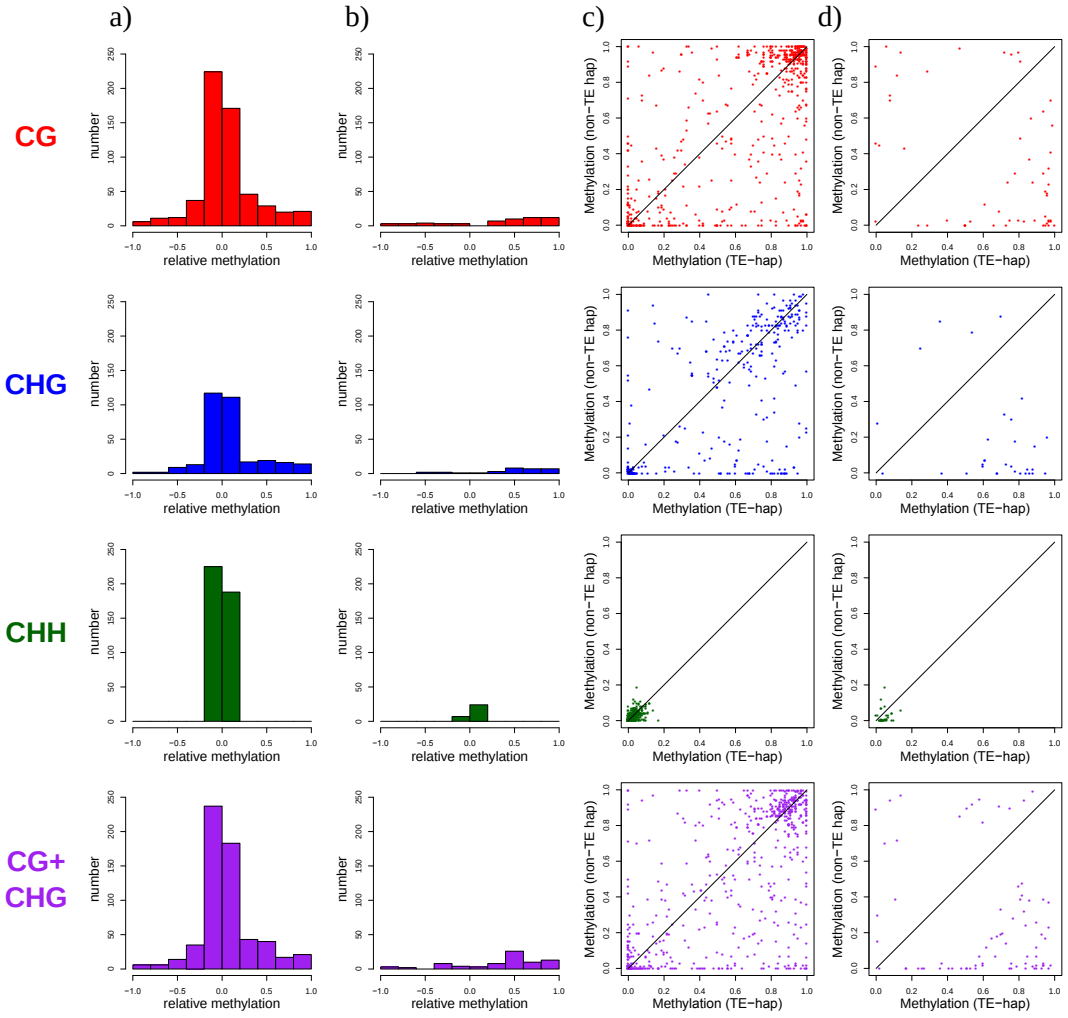


Figure 29 | Individual TIRs flanking regions analyses. Methylation is calculated over a region of 2kb around the insertion point in both haplotypes
a) Distribution of the difference of methylation values between TE haplotype and non-TE haplotype
b) Subset of Wilcoxon Mann-Whitney tests p-value positives (<0.01) of a)
c) Dotplot
d) Subset of Wilcoxon Mann-Whitney tests p-value positives (<0.01) of c)

Solo LTRs

The set of Solo LTRs used in this analysis is mainly belonging to the Ty3-Gypsy superfamily, hence is not surprising that since this type of TE is mainly located in heterochromatic regions, both haplotypes carrying solo-LTRs show an high level of CG and CHG methylations.

Since only solo-LTRs present in the reference haplotype and absent from the alternative one were taken in exam, it has to be considered that the haplotype bias tend to amplify the read differences between haplotypes

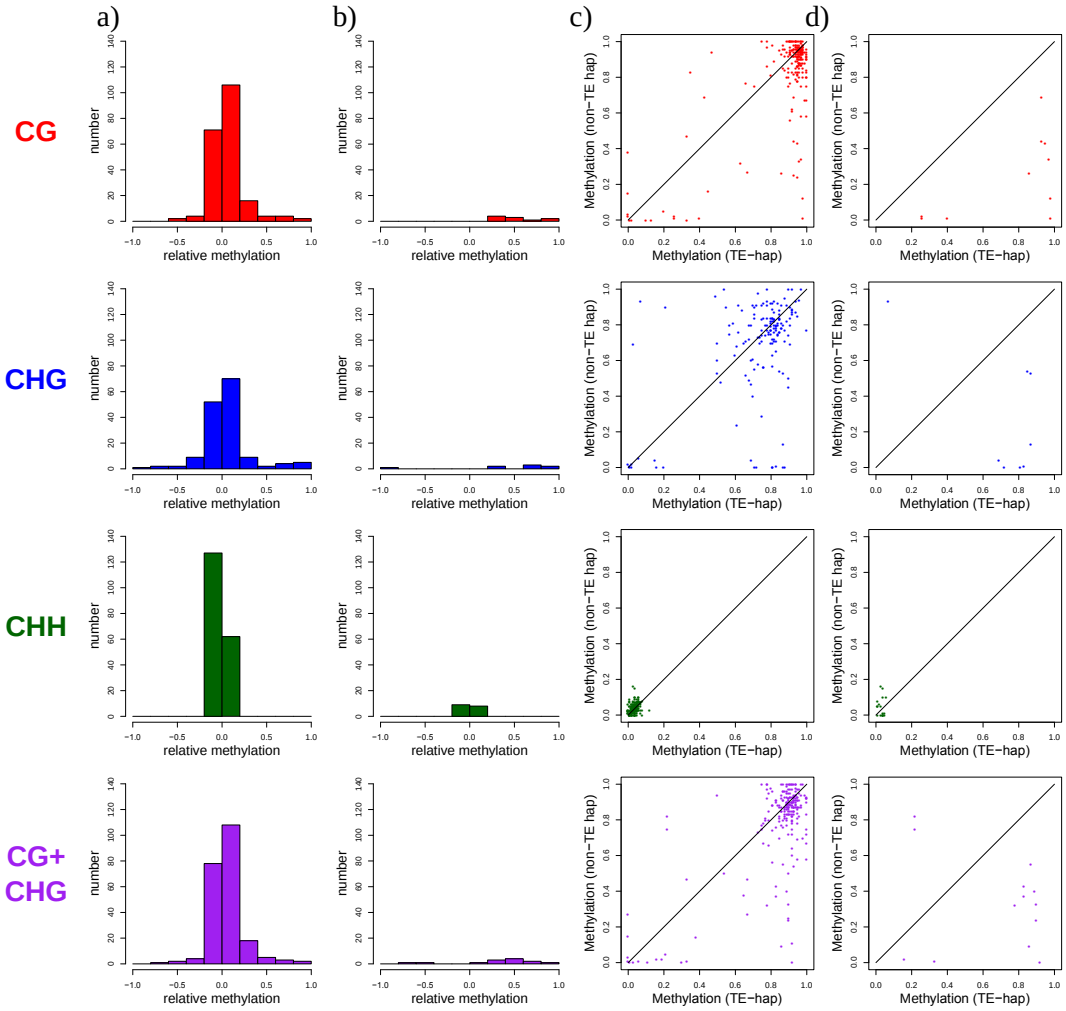


Figure 30 | Individual solo LTR flanking regions analyses. Methylation is calculated over a region of 2kb around the insertion point in both haplotypes
a) Distribution of the difference of methylation values between TE haplotype and non-TE haplotype
b) Subset of Wilcoxon Mann-Whitney tests p-value positives (<0.01) of a)
c) Dotplot
d) Subset of Wilcoxon Mann-Whitney tests p-value positives (<0.01) of c)

Taken together these data show that TE insertions are generally associated with an increase of methylation in their flanking regions within 1000 bp provided that the pre-existent methylation level is not saturated. However, different TE-groups show specific behaviours: Ty1-Gypsy retrotransposons are often localized in pericentromeric heterochromatic regions, hence their pre-existent methylation level in both CG and CHG contexts is high and the TE–induced increase in methylation can only be revealed in non-saturated regions. LINE elements

are mainly found in regions with a high pre-existent CG methylation and thus the increase of methylation in their flanking regions is restricted to the CHG context. Ty1-Copia and TIRs show high heterogeneity in their genomic localization, resulting into variable pre-existent methylation levels in the CG context in particular that will affect the effects of the TE insertion on methylation. No differential methylation is detectable for ”na” elements either in the meta-analyses of flanking regions (Figure 23) or in the single-TE analyses (Figure 31). Moreover the CHH behaviour emerged from the single TE-analysis of “na” SVs is comparable to the same analysis in the four different macro-groups of TE, providing more robust evidence of its negligible or undetectable role in TE-induced propagation of methylation in flanking regions in grapevine.

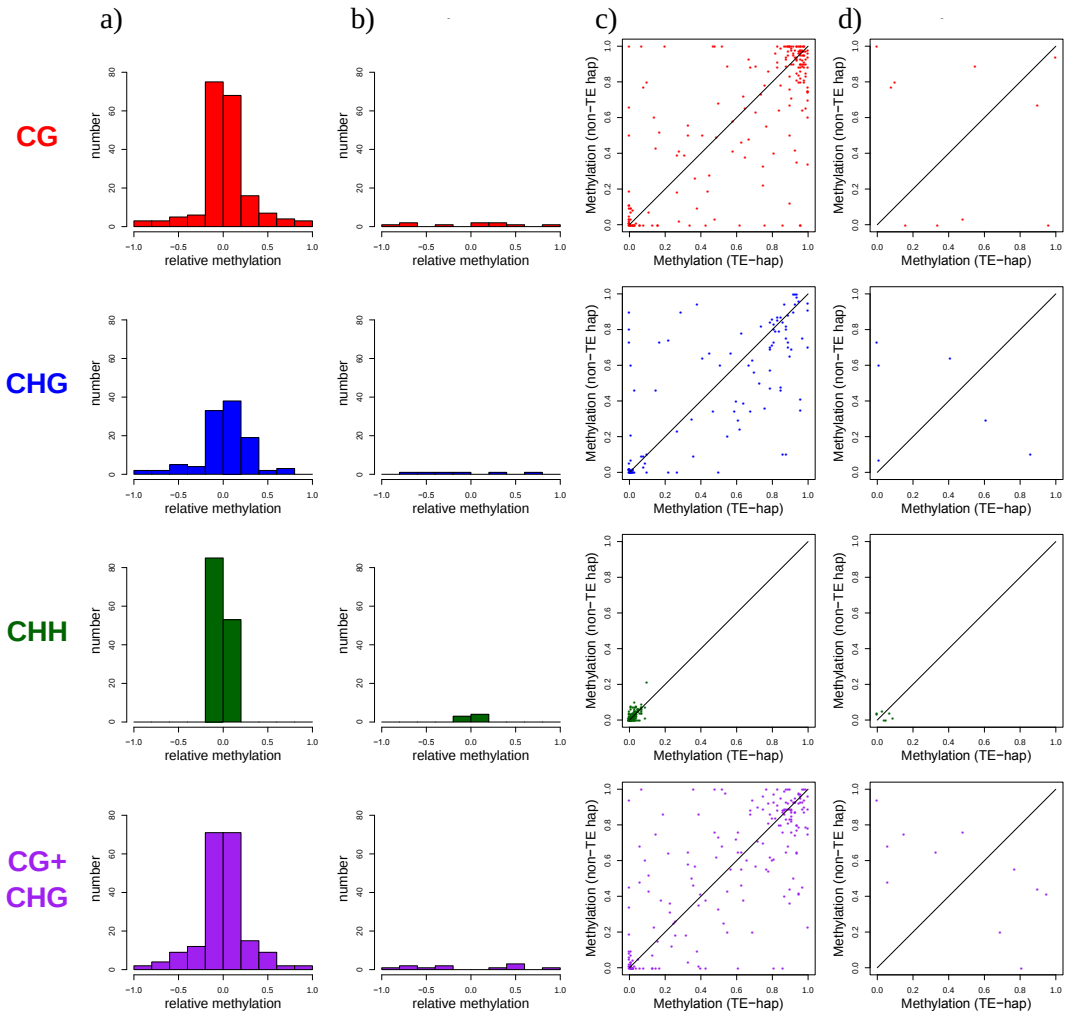


Figure 31 | Individual NA-insertions flanking regions analyses. Methylation is calculated over a region of 2kb around the insertion point in both haplotypes
a) Distribution of the difference of methylation values between TE haplotype and non-TE haplotype
b) Subset of Wilcoxon Mann-Whitney tests p-value positives (<0.01) of a)
c) Dotplot
d) Subset of Wilcoxon Mann-Whitney tests p-value positives (<0.01) of c)

Features of DNA methylation in gene bodies

DNA methylation is generally associated to chromatin packing and thus to transcriptional inactivity. However, both in plant and animals, methylation may be found in the transcribed regions of active genes, exclusively in the CG context (Feng et al., 2010). Such methylation, called Gene Body Methylation (GBM), is not involved in gene silencing and is rather required for efficient transcription and splicing regulation (Maor et al., 2015). Variation in GBM does not seem to quantitatively affect gene expression, as in *Brachypodium* and rice the methylation level is a long-term property of conserved genes (Takuno & Gaut, 2013). Nevertheless, in some species GBM has been found to be positively correlated with gene expression (e.g. in soybean; Schmitz et al, 2013). To investigate the relationship between gene body methylation and gene expression in grapevine, a genome-wide meta-analysis was performed.

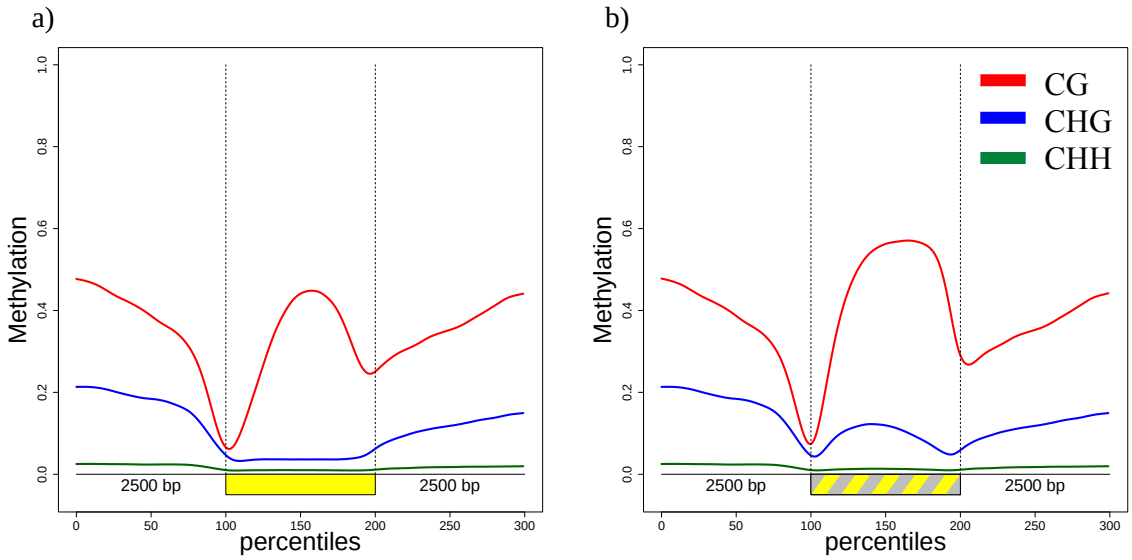


Figure 32 | Gene body methylation in exonic (a) and exonic and intronic (b) sequences in percentiles.

In order to properly identify and compare methylation changes at transcription start and termination sites as well as at exon/intron boundaries, the analysis was restricted to a set of 19896 genes where both 5'- and 3'-UTR were unambiguously annotated in the PN40024 genome reference.

The analysis showed that in grapevine exonic CG methylation exhibits the typical asymmetrical bell-like profile observed in other plant species, in which the 5' upstream region is much less methylated than the 3' downstream one (Figure 32), suggesting that the

transcription start site (TSS) is generally lowly methylated. When including the intronic sequences into the analysis, CG methylation levels strongly increase and also a CHG bell-like profile becomes visible. The distribution of methylation at CHG sites appears more symmetrical across the gene length than does GC methylation (Figure 32b), presumably because of the symmetric distribution of introns in the gene bodies. CHH methylation instead tends to be extremely low both in the gene bodies and in the flanking regions.

TEs may be frequently localized in introns (Figures 16 and 20), whose methylation level may be affected by highly methylated TE sequences. Hence, for the investigation of gene body methylation as a property of the transcribed sequence per se, only exons were considered in some analyses. To evaluate a potential quantitative effect of GBM, genes were grouped by methylation classes based on the average CG methylation of the curve peak observed at the 50^{th}-65^{th} percentile interval, and their gene body methylation profile was plotted separately in Figure 33.

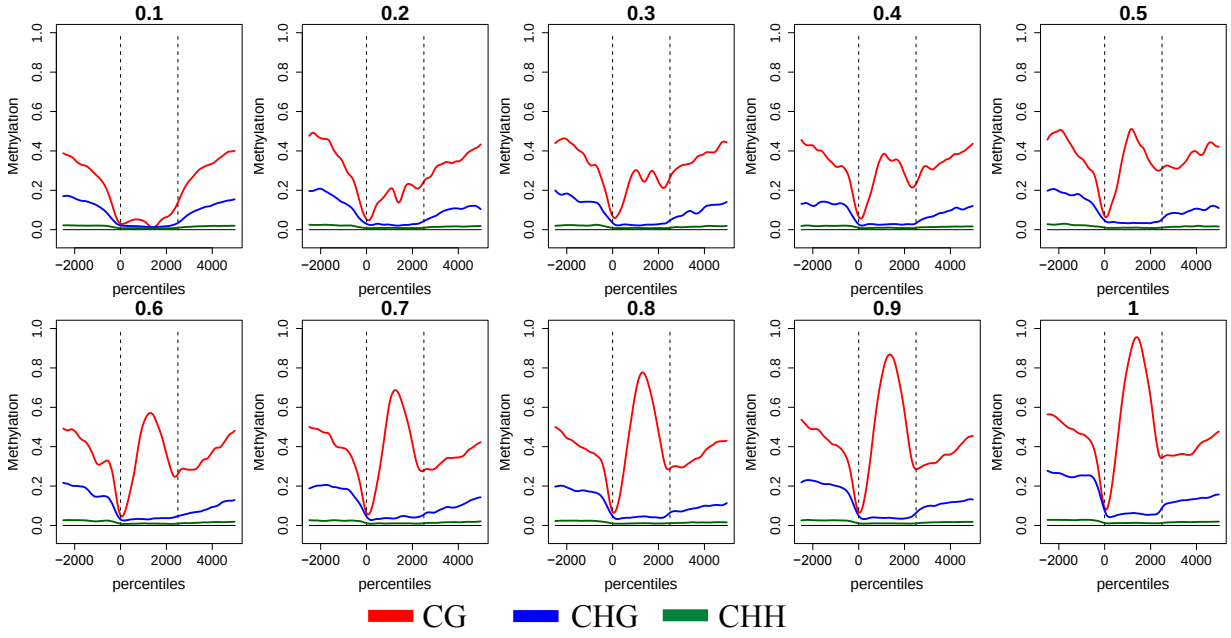


Figure 33 | Gene body methylation classes obtained calculating the average between the 50^{th} and the 65^{th} in the gene body in the gene body

Figure 33 shows that the peak of CG methylation reflects the general magnitude of gene body CG methylation in each class, and hence the level of CG methylation at the peak was used as a key parameter summarizing the amount of gene methylation of each class.

No evident correlation between flanking region methylation and gene body methylation is observable at this stage.

The abundance of each methylation class revealed a bimodal distribution of CG peak methylation (Figure 34a), setting aside a distinct group of genes with 100% methylation. When compared with expression data, a very limited variation in transcriptional levels was observed between the methylation classes, indicating that, consistent with observations made in other plant species, gene body CG methylation in grapevine does not preclude transcriptional activity and indeed on average the highest methylated genes show similar or even higher transcript levels than the lowest methylated ones (Figure 34b) .

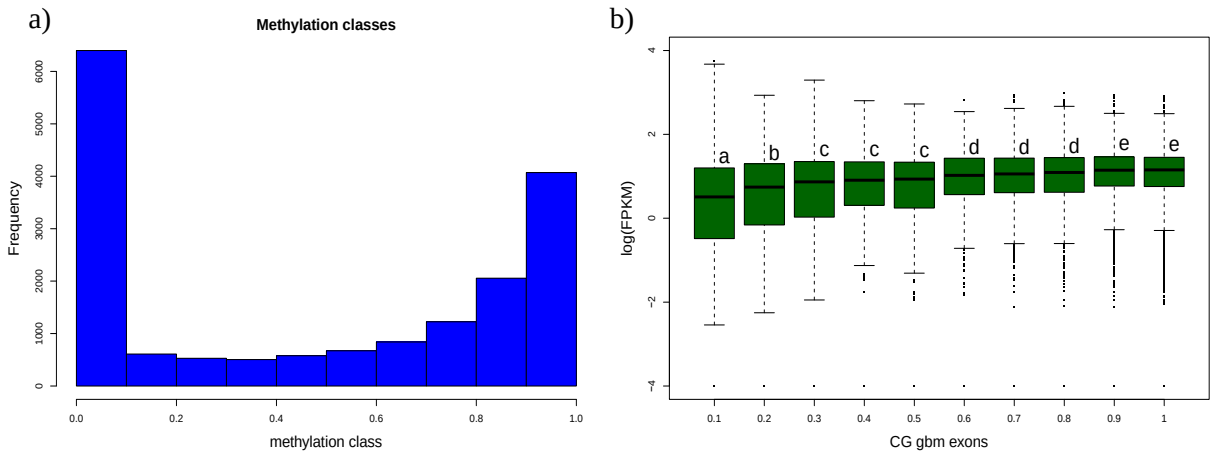


Figure 34 | Frequency of GBM classes (a), Expression rate of gene belonging to GBM classes (b). Clusters with the same letter code are not significantly different (Wilcoxon Mann-Whitney test (p-value <0.05)

Beyond gene expression, other gene features were examined across the different methylation classes, namely gene length, exon number, total exon length and total intron length. Interestingly longer genes tend to be more methylated, as well as genes with higher number of exons/introns and high total intronic sequence (Figure 35).

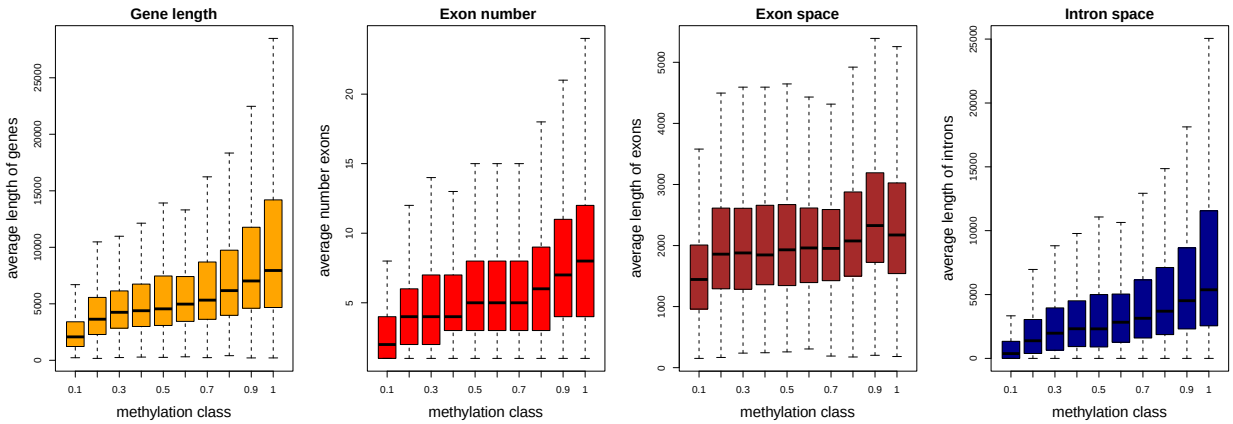


Figure 35 | Gene length, exon number, exon space and intron space in genes belonging to GBM classes

To better investigate this relationship, genes were grouped by exon number classes, and their methylation profile was plotted separately in Figure 36. As a confirmatory result, genes sorted by the total number of exons showed a clear correlation between exon number and the overall magnitude of the GBM profile. CG methylation in the introns always appeared higher than in the exons but, somewhat interestingly, the rate of methylation as a function of the increasing exon number was higher for exons than for introns. This observation suggests the bell-shape profile of gene body methylation may be mainly dictated by the spliced exonic component of the gene and only marginally affected by intron methylation. Adding on top of this speculation, CHG methylation, being restricted to the intronic intervals, was found to be invariable across gene length.

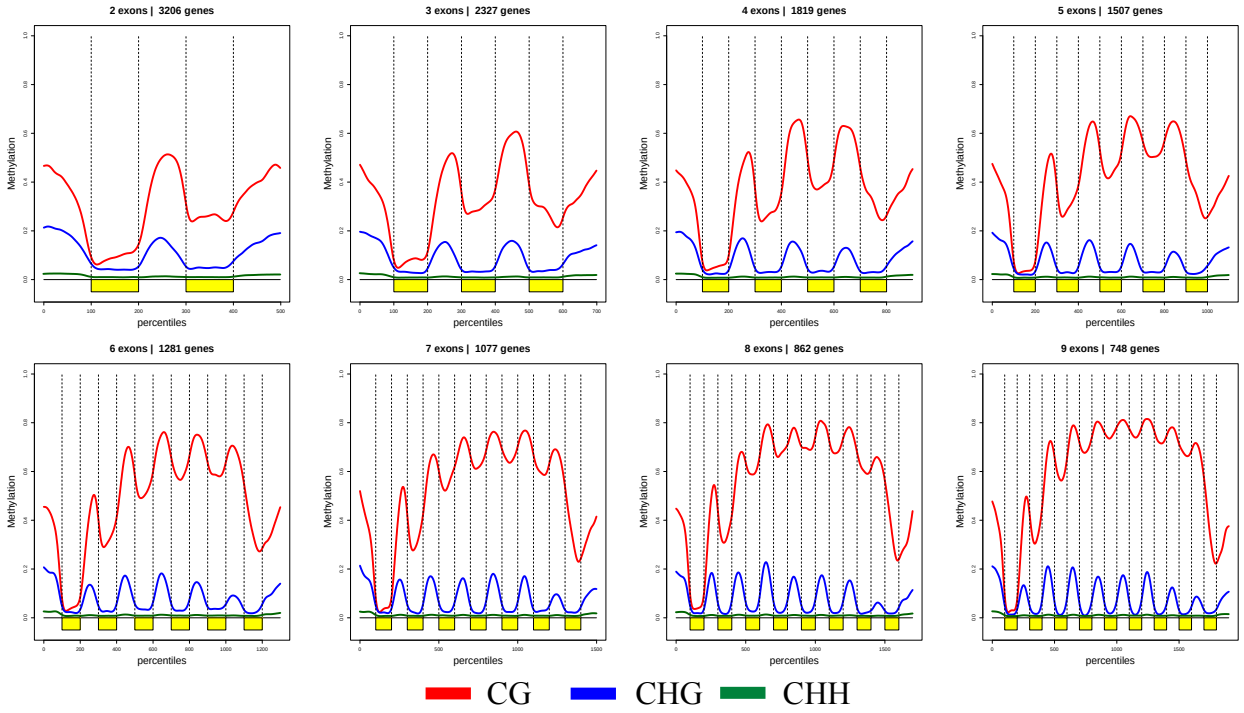


Figure 36 | Gene body Methylation profile of in genes grouped by exon

The finding of highly methylated introns is surprising considering that in other species including Arabidopsis, human and honeybee their methylation level is always lower than their flanking exons, both in the CG and CHG contexts (Chodavarapu et al., 2010; Lyko et al., 2010, Hodges et al., 2009). As mentioned previously, 36.7% of grapevine genome is occupied by introns that in turn are colonized by one third of total TEs mapped in the genome.

Mapped TEs, and in particular retrotransposons, are more frequently located in the intronic sequence of genes belonging to higher classes of gene body methylation (Figure 37). Hence, higher methylation in both the CG and CHG sites of intronic sequences compared to the

flanking exons may be a consequence of the epigenetic silencing of TE and the spreading of TE methylation on their flanking regions shown in the previous chapter. Nevertheless, exon and intron methylation profiles have been also computed after excluding genes containing any annotated TE-related fragment. Interestingly CHG methylation was almost completely undetected also in introns, whereas CG methylation decreases in both exons and introns. Moreover, in genes with high GBM, when excluding those carrying TE-annotation, exons and introns display a similar pattern, suggesting that intronic TEs may be responsible both for the CHG and the CG differential methylation, but no for the fact that grapevine introns are at least as methylated as exons in the CG context.

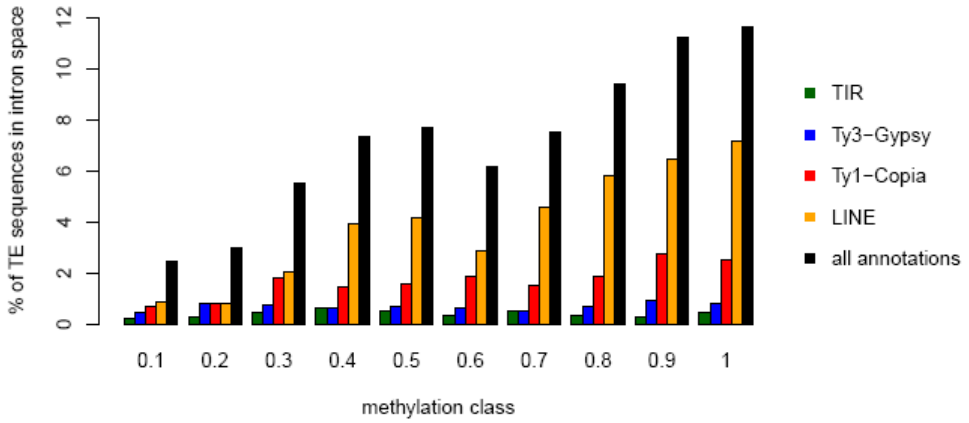


Figure 37 | Intronic TE annotations in genes grouped by GBM classes

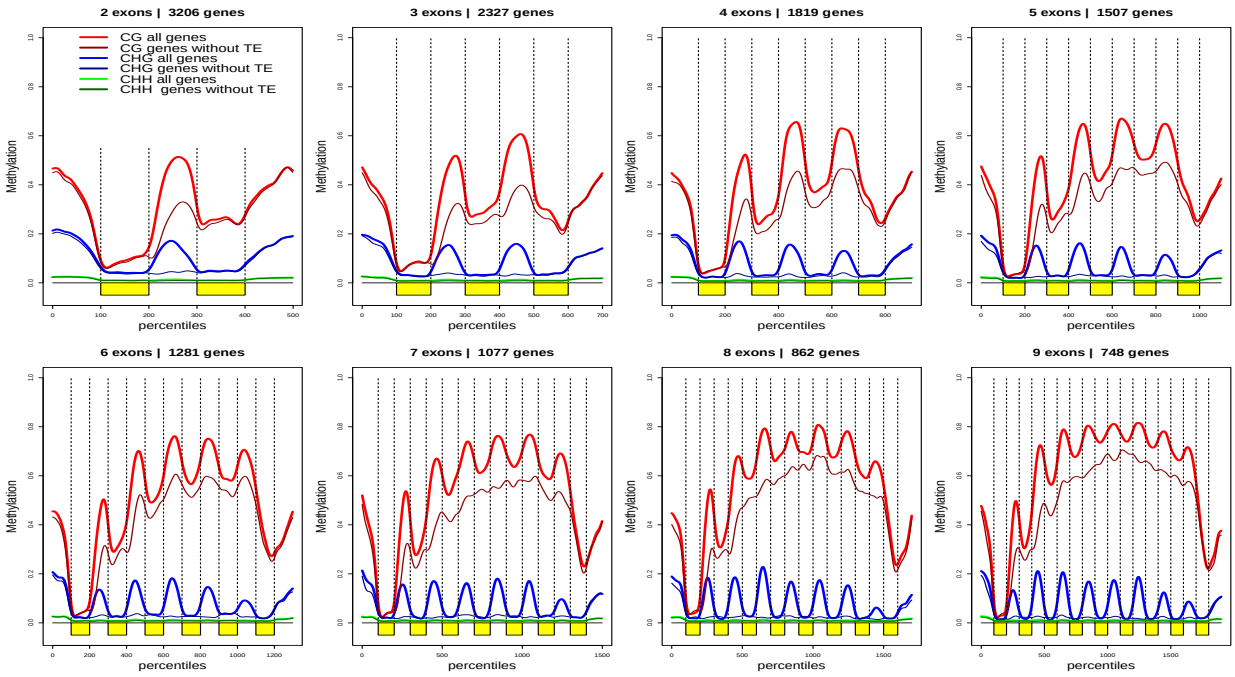


Figure 38 | Gene methylation profile of in genes grouped by exon number, including or excluding genes carrying TE annotations

Methylation in the first exon

The methylation pattern of exons and introns (Figures 36 and 38) shows that exon 1 methylation is generally extremely low, irrespective of either TE content or GBM. This pattern was not unexpected, since also in human the first exon is generally unmethylated in transcribed genes (Brenet et al., 2011; Sengupta & Smith, 1998) and its methylation has a stronger effect on transcription than the methylation in the promoter.
To verify if grapevine first exons display a similar pattern, exon 1 has been analysed separately from the rest of the exonic sequence. Figure 39 shows that exon1 is generally unmethylated, suggesting that also in grapevine absence of methylation may be required for its transcription activity.
The analysis of the expression rate of genes with differential exon1 methylation shows a very different pattern compared to gene body methylation. Indeed, genes with lower exon1 methylation (Figure 40 and Table 3, groups a-b) are significantly more expressed than genes with higher methylation (Figure 40 and Table 3, group f), according to the Wilcoxon Mann Whitney test (p-value < 0.05) whereas for the whole gene body, higher methylation is associated to an higher expression rate.

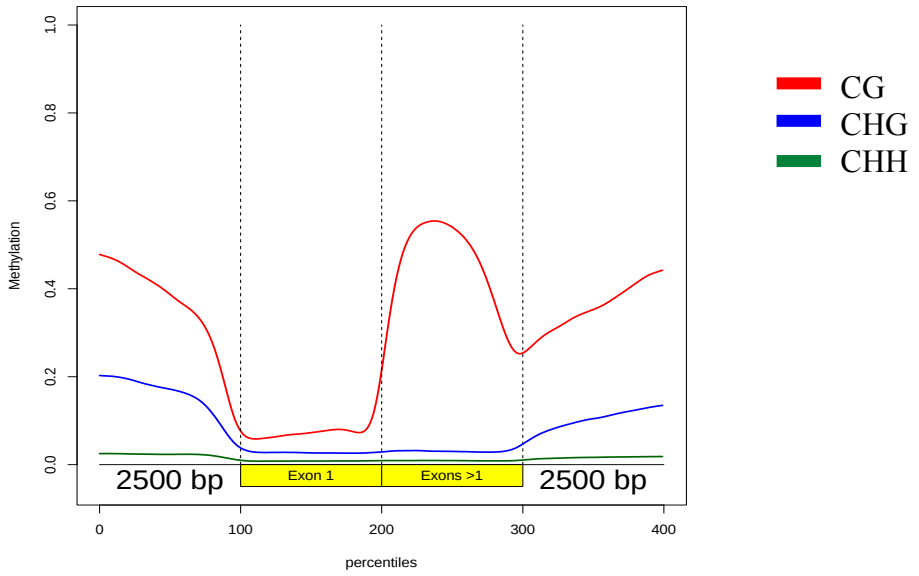


Figure 39 | Gene Body Methylation profile of Exon 1 vs all the other exons

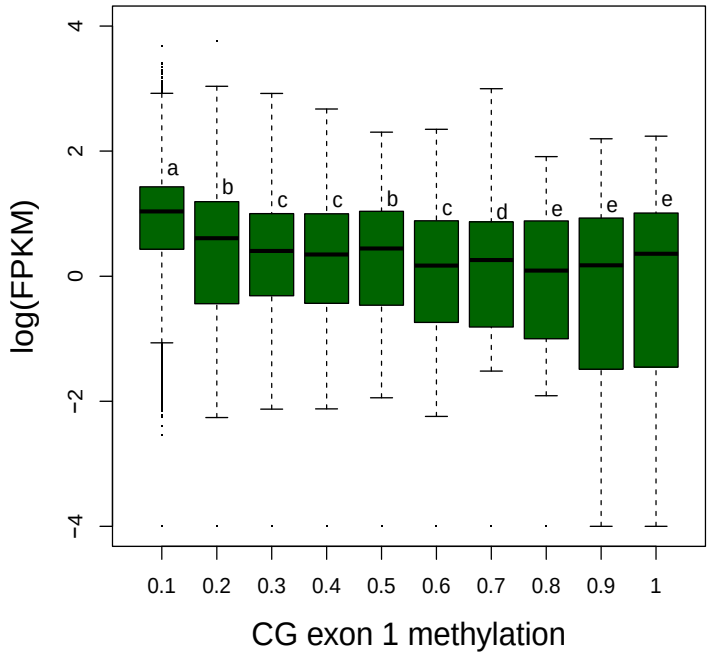


Figure 40 | Gene expression according the average methylation in exon 1.

Clusters of genes according to Wilcoxon Mann Whitney Test (p-value <0.05)										
Exon1 methylation	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
0.1	a									
0.2		b			b					
0.3			c	c	c	c				
0.4				d	d	d	d			
0.5					e					
0.6						f	f	f	f	f
0.7							g	g	g	g
0.8								h	h	h
0.9									i	i
1.0										j

Table 3 | Clusters of group of genes according to their exon1 methylations. Clusters with the same letter code are not significantly different according to Wilcoxon Mann-Whitney P-value <0.05.

Association between LINE insertions and methylation of gene bodies

Based on the above analyses, LTR-retrotransposons and TIR DNA transposons show a clear *cis* effect by increasing the methylation of the TE-carrying haplotype through the spreading of methylation into the flanking regions. *Trans* effects, capable of propagating the epigenetic silencing to the homologous alleles, cannot be ruled out but for LTR and TIR elements a pre-existing heterochromatic landscape appears to be the most simple explanation behind the comparable methylation levels of homologous regions surrounding a hemizygous insertion observed in a number of instances. The reasoning around the possible effects of LINE elements is similar but a few considerations may be added. First of all, the pre-insertion methylation profile of LTR and TIR elements is flat, whereas the LINE profile is curved, probably as a result of LINEs' preferential localization in gene bodies and introns in particular. Indeed, Figure 32, 36 and 38 show a clear curved methylation profile for the CG context in gene bodies, especially when considering also introns .By extending the flanking region analysis up to 10 kb from the insertion point, the curve profile extinguishes within few kilobases, at a distance in which most of cytosines are evidently outside the majority of the genes, which have a median length in grapevine of about 3399 bp (Jaillon et al., 2007) (Figure 41).

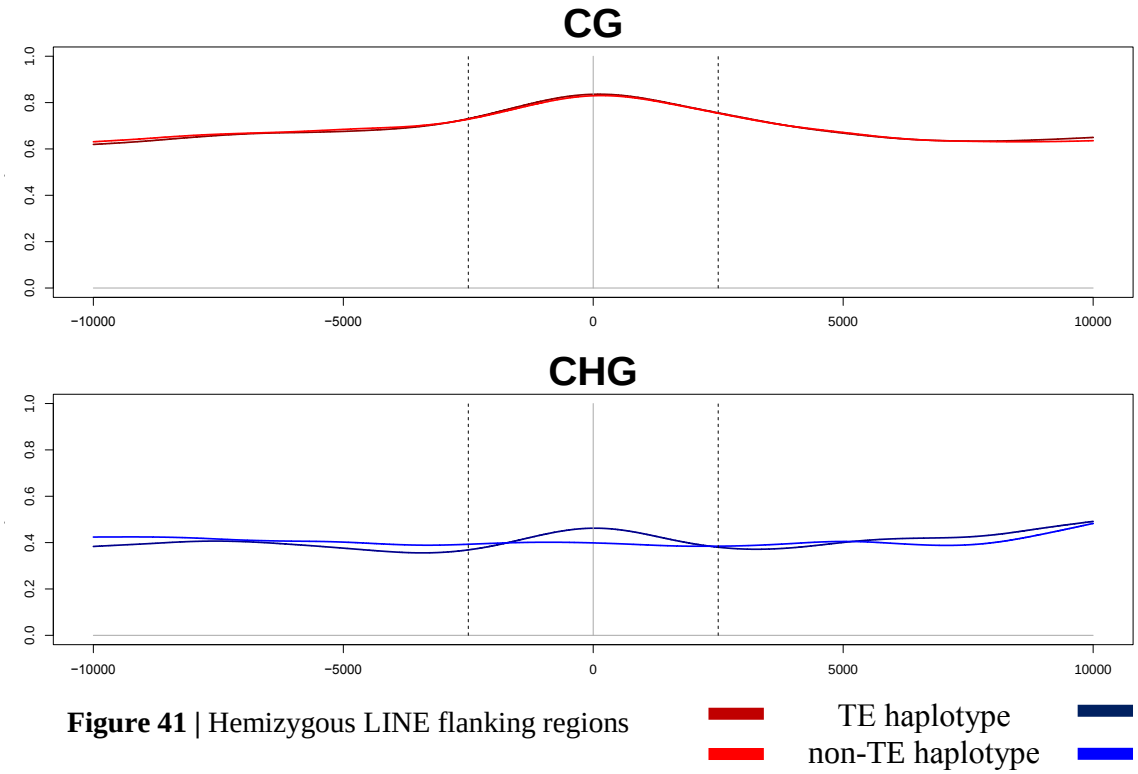


Figure 41 | Hemizygous LINE flanking regions

Second, the CHG methylation increase observed exclusively in the TE haplotype supports the hypothesis of pre-insertion saturated methylation. Thus, LINE elements may be somehow attracted to specific CG-saturated but lowly methylated CHG loci.

To further investigate the LINE behavior, methylation profiles of hemizygous LINEs have been sorted according to the genomic compartment in order to compare both genic and intergenic insertion sites Figure 42). Interestingly, LINEs located in intergenic regions show a pre-and post-insertional methylation state very similar to that observed for elements found within introns. These intergenic LINE profiles may suggest the existence of a certain number of non-annotated genes/pseudogenes and also support the LINE preference for insertions in CG-saturated but low CHG-methylated loci, which may also be intergenic.

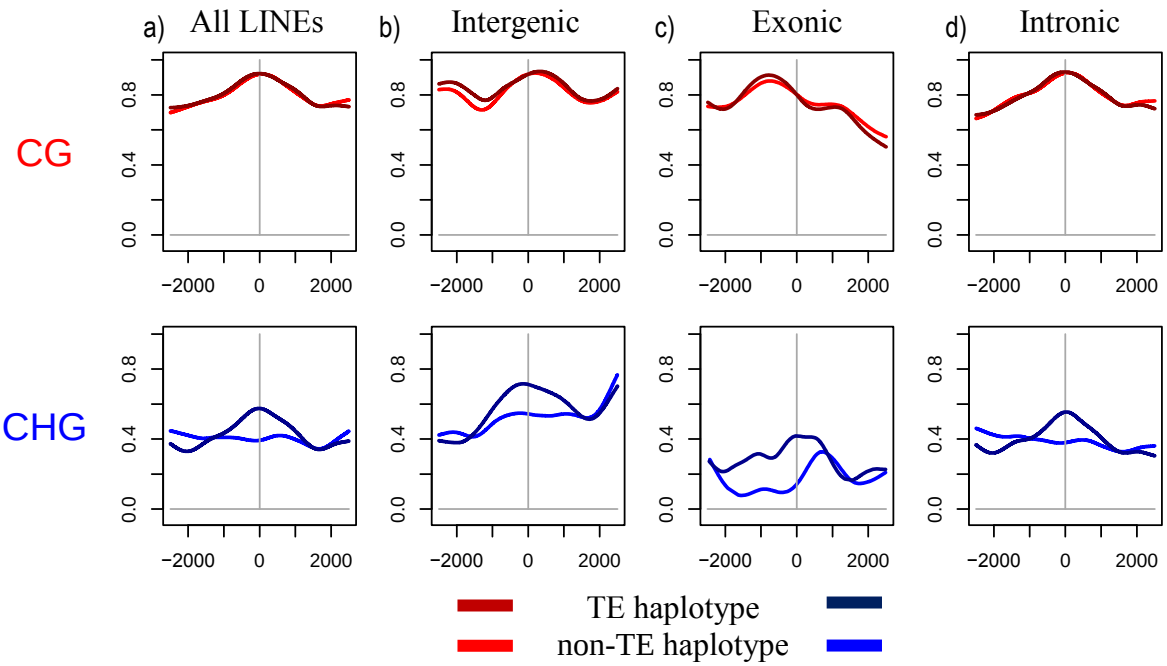


Figure 42 | Hemizygous LINE flanking regions in intergenic, exonic and intronic loci.

Intronic and Exonic Ty1-Copia insertions

Figure 20 shows that hemizygous Ty1-Copia elements have very variable locations: they are prevalently found in intergenic regions (53%), but also in exonic (6%) and intronic (36%) sequences.

Similar to intronic LINEs, intronic Ty1-Copia display a high methylation level in the CG context of both haplotypes and an increase in CHG methylation in the haplotype carrying the TE. Intergenic Ty1-Copia insert in regions with intermediate methylation in CG and low methylation in CHG, inducing an increase of methylation in both contexts. Interestingly, Ty1-Copia elements may be occasionally found also in exons, which are generally lowly methylated in both CG and CHG contexts. The pre-existent methylation profile surrounding the insertion site reflects the intron-exon-intron boundaries in both CG and CHG contexts (as shown in Figures 36 and 38 and, as expected higher methylation is found in the haplotype carrying the TE. The general Ty1-Copia pattern (Figure 23 and 44a) is hence the weighted combination of these three location-specific patterns and the small depression of the CG methylation close to the insertion point in the unaffected allele can be ascribed to the weighted contribution of exonic Ty1-Copia insertions.

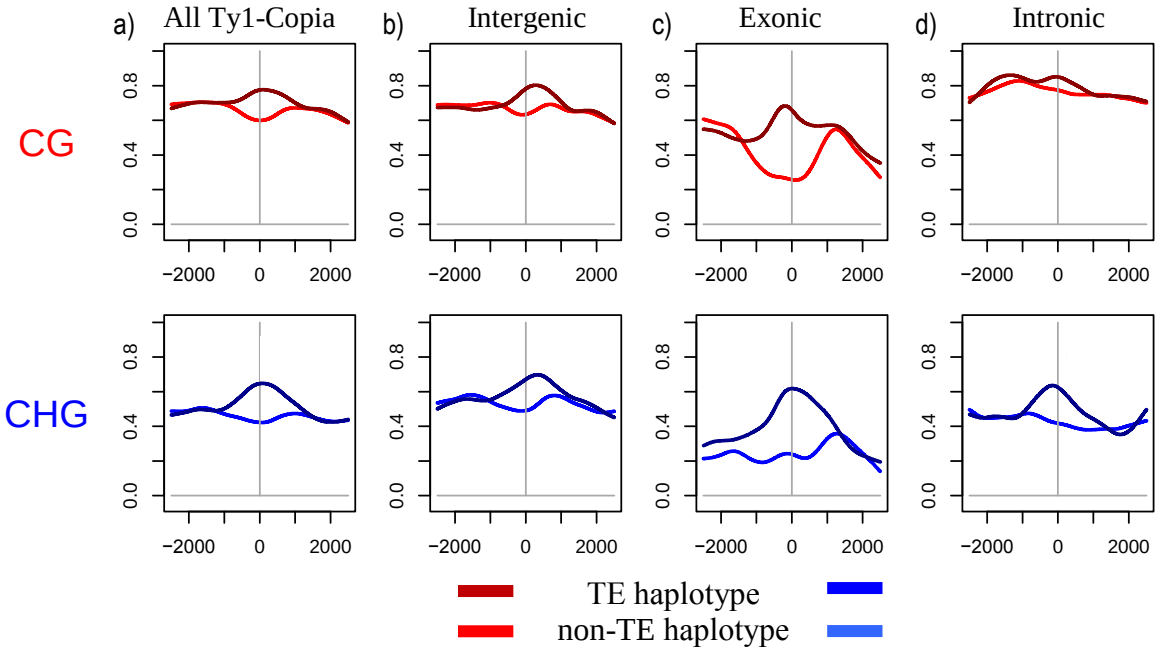


Figure 44 | Hemizygous Ty1-Copia flanking regions in intergenic, exonic and intronic loci.

Transposable elements insertion may modulate gene expression

In the previous chapters we have shown that DNA methylation is positively correlated with gene expression when occurring in the gene body (Figure 34b), and negatively correlated when occurring in the macroregion where the genes are located. Furthermore, DNA methylation occurs in TE loci and it's often spread on TE flanking regions. TEs occupy a big fraction of grapevine genome (41.4%) that interacts with the genic compartment. Part of them are also responsible for the peculiarly large intron space of this genome (36.9%), which significantly contributes to the total sequence space occupied by genes (46.3%) (Jaillon et al., 2007). Several studies have well established that the physical proximity of TEs to gene sequences may affect gene expression, although it is sometimes difficult to distinguish genetic effects (e.g. disruption of pre-existing regulatory sequences) from epigenetic effects (e.g. propagation of silencing epigenetic marks) generated by transposed elements. To explore TE effects on gene expression, we carried out a series of analyses focusing on genes physically associated with TEs.

In the first analysis, we examined the distribution of expression levels (expressed in FPKM units) of genes physically associated to full-length TEs (Figure 16) either in the introns, in the 2500 bp upstream regions or in the 2500bp downstream regions and we compared it with the distributions of genes devoid of TEs.

The majority of TE groups does not seem to cause significant variations in terms of expression when inserted upstream; Ty1-Copia elements are the only group associated with a significant decrease of expression in this case, according to the Wilcoxon Mann-Whitney test (p-value <0.05). In contrast, if taken all together, TEs located downstream to a gene show a significant negative effect on gene expression.

TE insertions in upstream sequences may disrupt promoters and regulatory sequences, provide their own promoter (thereby enhancing transcription), or cause the silencing of the flanking genes by spreading epigenetic marks. Hence, the fact that there are not significant differences in expression between genes carrying or not TEs in the upstream regions may be due to the compensation of positive and negative effects of TEs on the transcription. Conversely, the insertion of TEs in downstream regions is more often associated with lower

expression and we may speculate that in this compartment TEs have fewer chances to positively contribute to gene expression.

Genes carrying Ty1-Copia, LINE and TIRs insertions in the introns are significantly higher expressed than genes devoid of TEs. As genes whose introns are populated by TEs tend to also present higher gene body methylation levels, this is consistent with the positive correlation between gene body methylation and expression (Figure 34b).

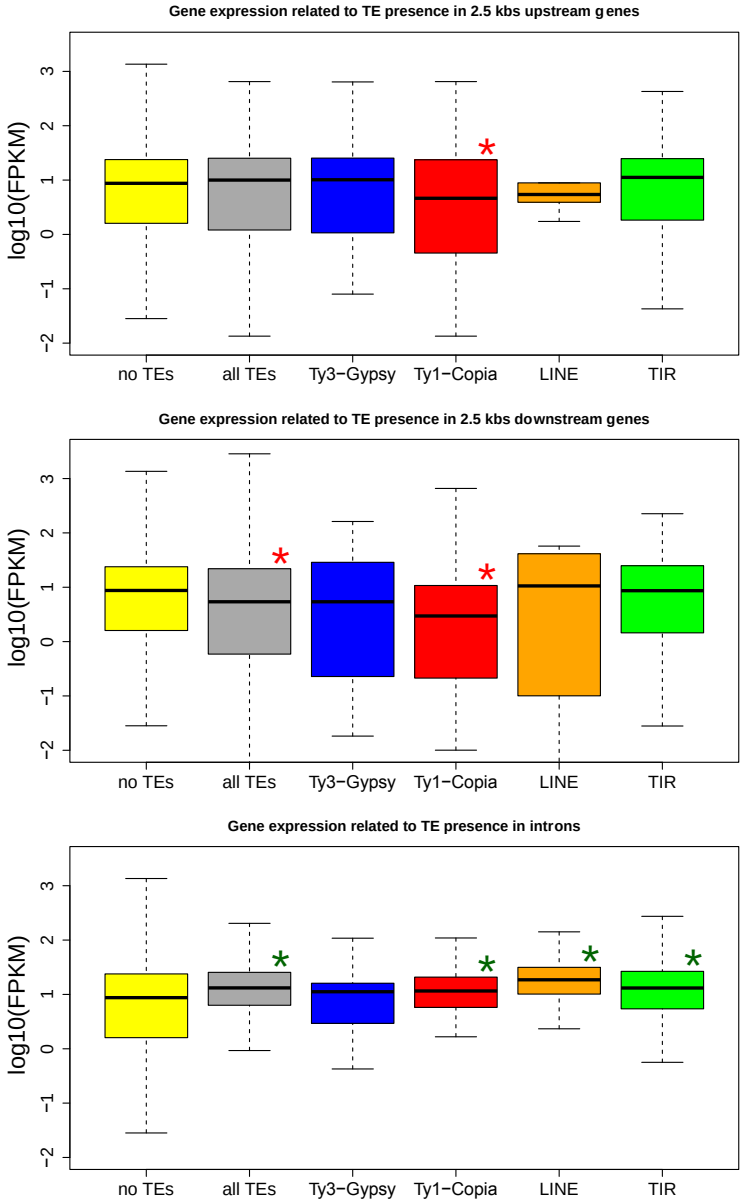


Figure 45 | Log10 of FPKM of genes related or not to TEs. *marked boxplot are significantly different from unaffected genes. Red* are significantly lower expressed than unaffected genes, Green* are significantly more expressed than unaffected genes.

To further investigate the effects of TEs on gene expression we also considered genes associated with hemizygous TEs in order to restrict the analysis to differences in allele expression within the same gene and eliminate the caveats of comparing genes that present different expression levels per se. In our research group allele specific expression data were available (Eleonora Paparelli, PhD), for all the replicates of the same Pinot Noir clone taken in exam for methylation analyses.

The software ALLIM (Allelic imbalance metre, (Pandey, Franssen, Futschik, & Schlötterer, 2013), was utilized for measuring allele specific gene expression (ASE) in the three replicates. This tool is suitable for mapping RNA-seq data on both the reference and the alternative haplotype. ALLIM detected 7393 genes where polymorphisms between the two alleles could make the ASE analysis possible. Out of this group of genes, 6348 intersected with the set of genes used for the Gene body methylation analyses and 1271 were associated to a hemizygous SV in either 2500 bp flanking region or in the gene body. Genes devoid of SVs (5077) and genes associated to SVs in both haplotypes (193) were discarded. SVs consisting of incomplete elements (see Table 2) were aggregated to each corresponding TE group. The remaining 1078 genes were distributed among the TE-groups as indicated in Table 4.

Genomic localization of main groups of hemizygous SVs							
	Ty3-Gypsy	Ty1-Copia	LINE	TIR	soloLTR	na	total
All	160	221	241	225	6	52	1040
up	76	61	13	85	3	11	253
exon	4	6	0	13	0	7	35
intron	30	96	200	40	2	7	439
down	47	51	22	73	1	11	209

Table 4 | Hemizygous SVs associated to genes with 2 alleles identified by ALLIM. RLC_partial and RLG_partial SVs are aggregated to Ty1-Copia and Ty3-Gypsy respectively. SVs overlapping exon-intron boundaries are considered in the exon group. SVs spanning over gene body and flanking regions are only considered only in “”all”

To verify whether the presence of a hemizygous SVs is correlated to a significant change in the expression ratio of the two alleles, we used the Wilcoxon Mann Whitney test to compare the between-allele $\log_2$ratio in genes associated and not associated to SVs.

In order to perform this analysis, a possible origin of bias was taken into account. Indeed, the reference alleles are on average more expressed than the allele belonging to the alternative

haplotypes (Figure 46), presumably because RNA-seq reads align more efficiently on the reference genome rather than on the reconstructed alternative haplotype.

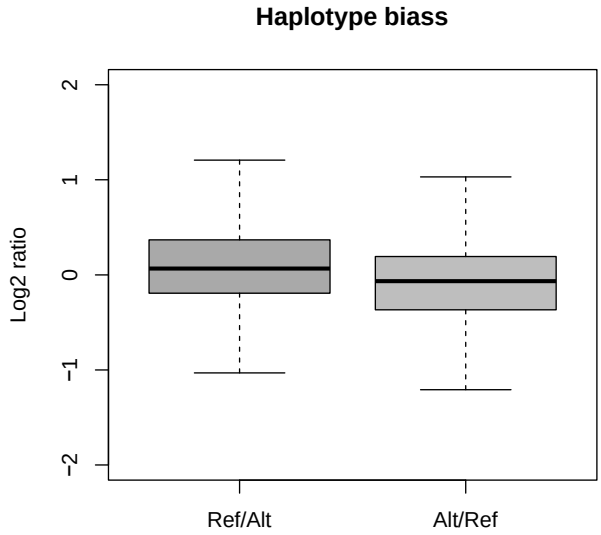


Figure 46 | Log2ratio of genes unrelated to SVs, expressed as Reference allele / alternative allele and vice-versa.

Since SVs occur in the two haplotypes with different frequency (432 in the reference haplotype, 646 in the alternative haplotype), the expression ratio between the TE-allele and the TE-devoid allele is overestimated when the TE is on the reference haplotype (40% of the SVs) and underestimated when the TE is carried by the alternative haplotype (60% of SVs).
Hence, in order to perform a suitable comparisons of ASE between the set of genes associated with SVs and the set of control genes, devoid of SVs, the latter consisted of 40% random genes for which the between-allele log_2ratio was calculated as reference_allele / alternative_allele and 60% random genes for which the between-allele log_2ratio was calculated as alternative_allele / reference_allele (Figure 47).
The result of this analysis indicated that alleles carrying SVs tended to show a lower log_2ratio, indicating a reduction in the contribution of the TE-allele to the overall gene expression level. Focusing on single TE groups, these differences were always significant when involving Ty1-Copia elements in both flanking regions (either upstream or downstream) and introns. For the case of Ty3-Gypsy superfamily they were significant only when occurring upstream and for TIRs only when occurring in introns or downstream.
We cannot rule out that the significance of these tests may also be affected by the reduced numerosity of the set of genes taken in exam.

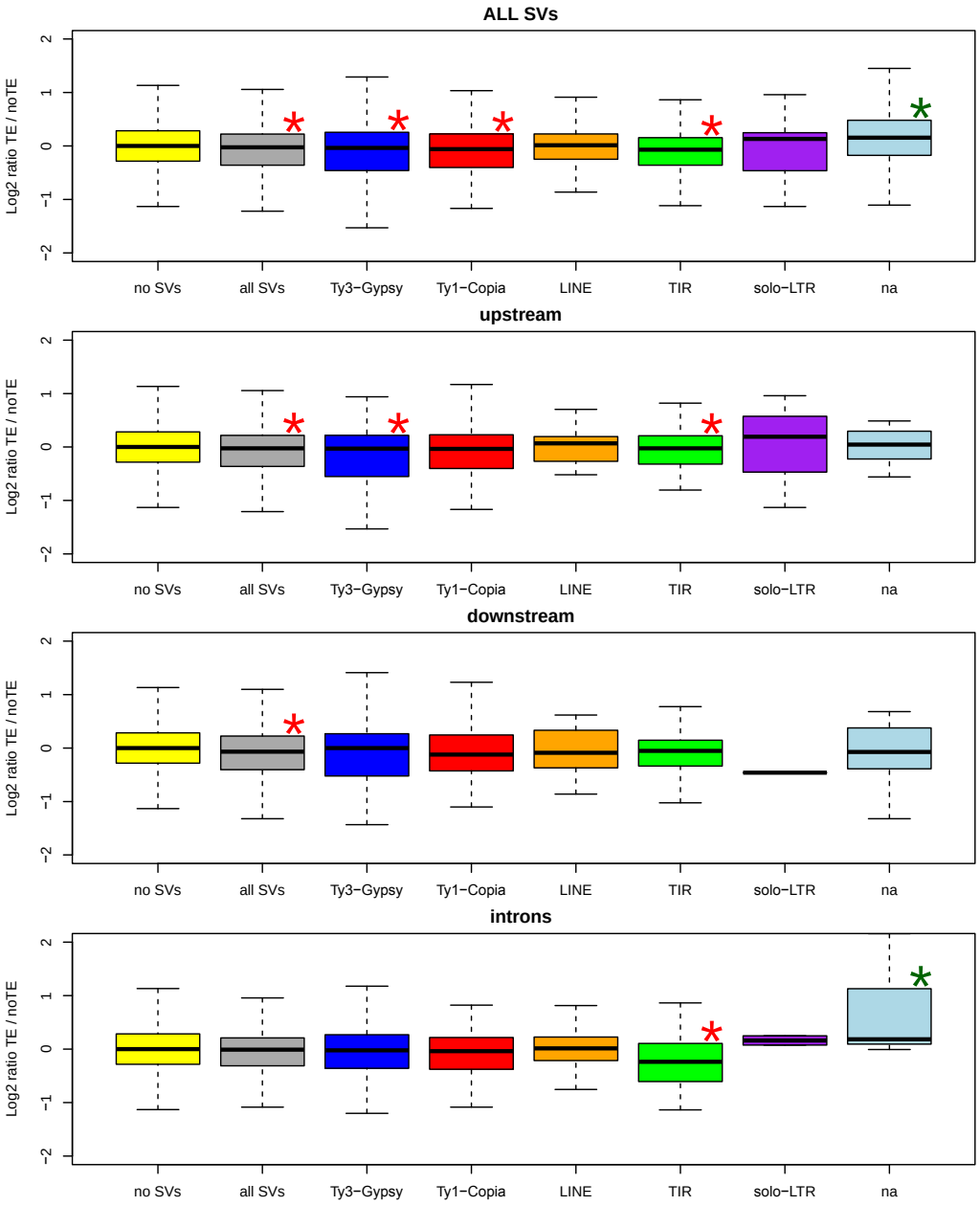


Figure 47 | Distribution of Log2ratio in SV-related and non-related genes. Significant values for Wilcoxon Mann-Whitney test (p-value<0.05) are marked with*, red stars indicate a lower log2ratio, green star a higher log2ratio.

DISCUSSION

DNA methylation is an epigenetic mark widespread among eukaryotes and its contrasting roles in transcriptional silencing of heterochromatic regions as well as transcriptional stability of gene sequences has been described in both plants and mammals.

DNA methylation cannot be directly detected through DNA sequencing as the C:::G pairing is not affected by cytosine methylation. However, through specific protocols that involve a sodium bisulfite treatment, unmethylated cytosines are converted to thymines and then by comparing a treated sample with an untreated reference, it is possible to obtain a full DNA methylome with a single-base resolution. Since a new Illumina protocol for BS-seq was introduced during the course of this study, replicates 2 and 3 were constructed with this new protocol, in which the DNA sample is immediately treated with bisulfite, which also contributes to its fragmentation, and then is ligated to Illumina adaptors. In contrast, in the traditional protocol utilized for replicate 1, the DNA sample is initially fragmented, then ligated to the adaptors and finally treated with bisulfite prior to amplification and sequencing. These protocol differences may partially explain the higher methylcytosine content estimated in replicates 2 and 3 than in replicate 1 (Figures 10,11,12,13) and the discrepancies in the distribution of methylcytosines between the three contexts CG, CHG and CHH. Despite these differences, locus-specific analyses in both gene bodies and TE flanking regions revealed similar patterns among the three replicates and yielded the same substantial conclusions.

The silencing effect of DNA methylation represents a useful defense against both retrovirus infection and TE insertions. TE sequences integrated in the genomes are generally *de novo* methylated in all contexts through the RdDM pathway in order to prevent further mobilization; indeed TE sequences are predicted targets of smRNAs in several plant species, e.g. Arabidopsis (Cokus et al., 2008), soybean (Schmitz et al., 2013), maize (Gent et al., 2013) and tomato (Zhong et al., 2013).

As expected, in grapevine TE bodies are highly methylated in both CG and CHG contexts, consistently with previous studies in *Glycine max* (Schmitz et al., 2013), in Arabidopsis (Cokus et al., 2008), tomato (Zhong et al., 2013) and maize (Emberton et., 2005; Palmer et al., 2003; Rabinowicz et al., 1999; Whitelaw et al., 2003) but interestingly no significant CHH methylation increase is detectable in grapevine whereas in soybean it is present in both LINE and TIR elements, even though at low level. So far no maintenance mechanisms have been

characterized for CHH methylation other than repeated *de novo* methylation via RdDM (Wierzbicki et al., 2012). Methylation at these sites is prominent during the early stages of embryo development (Jullien et al., 2012) thus a compelling hypothesis that has been proposed claims that vegetative propagation by bypassing embryogenesis might preclude a major event of *de novo* CHH methylation, which indeed tends to be lower in vegetatively propagated relative to sexually propagated species (Robert J. Schmitz, oral presentation). Grapevine would not be an exception under these assumptions, which might be verified taking in exams TE bodies in F1 seedlings of the Pinot Noir cultivar.

However, the CHH context is generally low methylated in all plant species so far analysed, e.g., *Glycine max* (Schmitz et al., 2013), Arabidopsis (Cokus et al., 2008), tomato (Zhong et al., 2013). Interestingly, in maize CHH islands, whose methylation level is much lower than in CG and CHG sequences, have been found nearby genes and may act as epigenetic insulator that protect genes by the spreading of epigenetic marks (Gent et al., 2013).

Methylation in TE bodies is often spread into the flanking regions, but despite a similar internal profile among the TE groups analysed, in the flanking regions methylation patterns display a high variability that reflects both the different pre-existent methylation profile of the target loci and the specific TE behavior. To better evaluate the effect of TE insertions on their flanking regions, sequences as closest as possible, one carrying a TE and one unaffected must be taken in exam.

Eichten et al., (2012) compared orthologous regions in two maize varieties, Mo17 and B73, in which TEs are uniquely present in the B73 reference but not in Mo17. A similar approach has been proposed for this study, with the advantage that PN40024 reference and Pinot Noir share the reference haplotype, thereby private PN40024 TEs must be hemizygous TEs in Pinot Noir. Considering hemizygous TEs, their flanking regions are compared to their homologous regions, in absence of TE, on the other chromosome, making these analyses not sensitive to environmental or developmental influences since the two haplotypes that are compared are observed in the same cells. Moreover, an in-house pipeline implemented in the team where the present study was performed allowed for the identification of hemizygous TEs uniquely present in the alternative haplotype. The seminal work by Eichten and coworkers showed that in maize DNA methylation and heterochromatic histone marks such as H3K9 dimethylation present in the internal TE sequence are spread on the flanking regions in some retrotransposon families (49 out of 144). Spreading families show higher internal CG and CHG methylation; in contrast, non-spreading families show higher CHH methylation. All these spreading families belong to the LTR retrotransposon order, in particular to the Ty3-Gypsy superfamily,

and the dating of their insertion, based on LTR divergence, revealed that they are between the youngest families of the maize genome (< 3 million of years).

Observations in maize are in general compatible with the retrotransposon behavior investigated in grapevine. Indeed, in grapevine both Ty3-Gypsy and Ty1-Copia elements show an increase of methylation in their flanking regions. Moreover, grapevine shows significant spreading of methylation also in LINE elements. Hemizygous TEs analysis in grapevine (Figures 23-31) shows that also in grapevine TE internal methylation is often spread on flanking regions, in particular in the CHG context that is generally far to be saturated. The strongest effect is observed in Class I retrotransposons. In particular Ty3-Gypsy elements generally insert in pericentromeric/heterochromatic regions and thus in loci already highly methylated in both CG and CHG contexts prior to the insertion and thus an increment of DNA methylation is appreciable only where the pre-existent level is not saturated; LINE elements are most frequently found in introns of transcribed genes which are as well highly methylated but only in the CG context and for the same reason a significant increment is found in particular in the CHG; Ty1-Copia elements are instead found both in intergenic sequences and in gene bodies and they often induce a methylation increase in both CG and CHG contexts. However, within the Ty1-Copia superfamily a high variability in the methylation level of the unaffected haplotype is observed, reflecting possible location-specific effects.

Unfortunately a more detailed taxonomic classification of repetitive elements is not available for grapevine yet and a possible future perspective may be the characterization of retrotransposon families according to Wicker et al., (2007) recommendations and then an analysis of DNA methylation based on family-specific patterns.

In this study also class II TIR elements have been analysed. They are frequently locate in intergenic regions but with no preference for heterochromatic loci, similarly to Ty3-Gypsy, and cause a significant increase of methylation on their flanking regions, although weaker than retrotransposons. This may be due both to a lower internal methylation level (Figure 17) and a bias in the hemizygous TE calling, because on theory they may be the effect of either a de novo insertion in one chromosome or an excision in the other. For what concerns Class I retroelements, their “copy and paste” mechanism suggests that the majority of hemizygous TE represents de novo insertions. On the contrary, the “cut and paste” mechanism of Class II TIR elements may potentially generate two hemizygous TE loci for each transposition event, one corresponding to the insertion in the new locus, and the other representing its deletion in the previous locus.

No mechanisms of demethylation following a TE excision are known; hence it would not be surprising that the flanking regions of an excised TE showed a similar pattern to the flaking region of the homologous TE on the other haplotype.

So far it has not been possible to distinguish whether an hemizygous TIR element has been generated by an insertion or a deletion, thereby in this work deletions involving TIR element may have been potentially taken in exam. This may partially explain the weaker effect of TIRs on the spreading of methylation in flanking regions (Figure 23 and 25).

The moderate TE content and the interspersed pattern of TEs across grapevine genome represent an ideal compromise for the evaluation of TE insertion effect on nearby genes.

For example in Arabidopsis, TEs are particularly enriched in pericentromeric region and thus their potential effect on nearby genes is hard to be evaluated. The maize genome is instead extremely enriched in TEs and TE insertion effects in their flanking regions may overlap with precedent insertions and rearrangement patterns increasing the noise and the possibility of evaluating insertion-specific effects. Despite the moderate TE content, the re-sequencing of dozen of grapevine varieties revealed a conspicuous number of SVs, that represent the most recent transposition events and thereby the most mobile families.

Moreover the high heterozygosity of grapevine allowed us to perform haplotype specific analyses with the advantage of comparing homologous loci within the same individual rather than inbred lines , as in Arabidopsis and maize. Spreading of heterochromatic marks that may alter gene expression has been well characterized also in other eukaryotic species, (e.g. Arabidopsis, Liu et al., 2004; Saze & Kakutani, 2007; Soppe et al., 2000), generally with a negative effect on transcription. In maize, instead, specific TE insertions have been correlated to abiotic stress response, and contribute to the activation of several genes when inserting in their upstream regions by acting as local enhancers (Makarevitch et al., 2015), suggesting the existence of a case-by-case pattern.

Preliminary analysis of gene expression in this study showed that genes located in highly methylated regions, especially in the CHG context, show lower expression on average (Figure 15a) and furthermore their expression tends to be more conserved within varieties than the expression of genes located in lower methylated regions (Figure 15c-d).

Genes with TEs located in their flaking regions generally display a lower expression rate than unaffected genes, whereas when they are located in introns gene expression rate is significantly higher (Figure 44).Furthermore, the allele-specific expression analysis shows that hemizygous SVs may modulate the contribution of the two alleles on gene expression (Figure 46).

Beyond playing a prominent role in gene silencing, DNA methylation may also be compatible with gene transcription. Indeed, in many eukaryotic species so far sequenced, a conspicuous fraction of transcriptionally active gene bodies show intermediate to high levels of methylation, usually restricted to the CG context. Grapevine is not an exception and several of its gene sequences show a bell-shaped methylation profile spanning the entire transcribed region from the transcription start site to the transcription termination site and involving CG methylation but very low CHG and CHH methylation levels. However, CG methylation is generally absent in the first exons of transcribed genes, consistently with previous analyses in humans (Jain et al., 2015; Lee, Evans, Kim, Chae, & Kim, 2014).

When sorting genes according to their bell-peak methylation, they distribute across a wide range of levels from absent to almost saturated methylation; however, the majority of genes show either a very low (<0.1) or a high (>0.7) average methylation.

Recent studies in humans (Hodges et al., 2009), honeybees (Lyko et al., 2010) and *Arabidopsis* (Chodavarapu et al., 2010) showed that exons display higher methylation than their flanking introns. This fact, together with the higher occupancy of nucleosome in exons, suggests that gene body methylation may be involved in exon definition and alternative splicing regulation (Maor et al., 2015) Surprisingly, several independent observations made in this work, provide evidence for an opposite scenario in grapevine, where introns are much more methylated than their flanking exons both in CG and in CHG contexts (Figures 23, 25, 28, 32, 35, 36, 37 ,38, 41,42 and 43). .

The higher CG and CHG methylation present in the introns (Figures 32, 36, and 38) may lead one to speculate that it might be the effect of highly methylated TEs, which indeed occupy 12.4% of the grapevine intronic sequence (Jaillion et al., 2007). To verify this hypothesis, gene body methylation profiles have been re-examined excluding all genes carrying TE annotations. In absence of TEs, methylation is globally reduced in both exons and introns. However, the most prominent reduction occurs in the introns and in particular in the CHG contexts, whose methylation drop to negligible levels. CG methylation declines as well in the intron albeit never below the exon methylation levels (Figure 38). Interestingly, in genes devoid of TEs but presenting high GBM, exon and intron methylation profiles converge whereas in low-GBM genes, introns remain more methylated.

These data suggest that intronic CG methylation even in absence of TEs is at least comparable with the exonic methylation, and thus it cannot be totally ascribed to the frequent localization of TEs in grapevine introns. Moreover, introns with high CG methylation may be somehow

attractive for TEs, in particular LINEs, that locate prevalently in introns characterized by a saturated CG methylation level and low CHG methylation (Figures 16, 20 and 42).

Despite the predominance of LTR elements within plants retrotransposons, non-LTR LINE elements are the major contributor of the repetitive fraction of human genome, and they are estimated to account for up to 400 million SV events within the human population (Xing et al., 2009). Somatic LINEs mobilization is associated to more than 70 human diseases, including cancer and neurodegenerative diseases. LINEs somatic insertion may be not casual and implied in early development of brain cell (Coufal et al., 2009).

In grapevine all the identified LINE elements belong to the L1 superfamily and are found prevalently in introns (Figures 16 and 20). When considering hemizygous LINEs in gene bodies, interestingly both the TE containing and the unaffected haplotype are highly methylated in the CG context. This may be due either to a pre-existent high methylation level in the intron or to an inter-chromosomal cross-talk that may occur via smRNAs, which indeed have been reported to targets LINE sequences in at least one species (soybean, Schmitz et al., 2013). However, the frequent high methylation in gene bodies, independent of TEs (Figure 38), and the *cis* propagation of CHG methylation into TE flanking regions but not in the other haplotype suggest that the cross-talk hypothesis should be rejected.

Preferential intronic insertion site in LINEs is evident (>75% within hemizygous TEs and >68% within homozygous TEs, Figures 16 and 20), hence there must be either a somehow positive role of LINE insertions in introns that allows this enrichment or a negative selection against intergenic LINE enrichment. Since intergenic DNA is frequently enriched in TEs, the second hypothesis appears unsustainable. Although the mechanisms that drive this intronic enrichment compels further investigation, the particular methylation pattern of saturated CG and low CHG levels that seems to be attractive for LINE insertions, represent an excellent starting point for future analysis. LINE family characterization and single-gene investigation may be helpful to provide more insight into this phenomenon.

MATERIALS AND METHODS

Plant Material

The samples selected are from the Pinot Noir clone VCR18, provided by Vivai Cooperativi di Rauscedo (http://www.vivairauscedo.com/). Each of the three replicates is a pool of genomic DNA, extracted from leaf nuclei of three different plants. The three groups of plants used for the three pools were grown in different rows .

WGBS library preparation

Replicate 1 BS-seq library was constructed with the Nugen Kit Ovation® Ultralow Methyl-Seq Library Systems (http://www.nugen.com/products/ngs/ovation-ultralow-methyl-seq-library-systems). Firstly 100 ng of DNA sample were fragmented by sonication with 3 cycles 15-90 in order to enrich the sample in the fraction of 300-600 bps fragments. Fragmented DNA was then processed as indicated in Nugen protocol.

In addition, gel size selection was performed after adaptors ligation step, considering the 120bp adaptors length, a gel slice corresponding to the 400-700 bp range was excited and the DNA purified with the QIAquick® Gel Extraction Kit cat. nos. 28704. As indicated by manufactures' guideline, bisulfite conversion was performed after Final Repair step with the suggested Quiagen EpiTect® Fast Bisulfite Handbook kit #59824, lastly the library was amplified through 15 cycles of PCR

Replicate 2 and 3 libraries were constructed with the Illumina TruSeq DNA Methylation Kit. https://support.illumina.com/downloads/truseq-dna-methylation-library-prep-guide.html

According to this protocol, no fragmentation and gel size selection are required, hence 100 ng of DNA sample were immediately processed as indicated in the Illumina protocol, including the bisulfite conversion as a first step through the EZ DNA Methylation-Gold™ Kit https://www.zymoresearch.com/epigenetics/dna-methylation/bisulfite-conversion/ez-dna-methylation-gold-kit which provides also for the fragmentation of the sample.

The library was finally amplified through 10 cycles of PCR.

Sequencing

Replicate 1 was sequenced with the Illumina HiSeq™ 2000 sequencer (http://www.illumina.com/documents/products/datasheets/datasheet_hiseq2000.pdf) whereas replicates 2 and 3 were sequenced with the more advanced Illumina HiSeq™ 2500 sequencer (http://www.illumina.com/systems/hiseq_2500_1500.html), in both cases according to the manufacturer's instructions.

Alignment of bisulfite converted reads

BS-seq reads were aligned with the in-house developed aligning program ERNE v.1.4.5 (http://erne.sf.net), suitable for efficiently mapping BS-treated reads.

Estimation of bisulfite conversion efficiency

To estimate bisulfite conversion efficiency a spike-in of unmethylated lambda phage DNA has been added to each sample.

Fastq files of read 1 and 2 were aligned on lamba genome with erne-bs5 with the following command

```
erne-bs5 --query1 FILE_Read1.fastq --query2 FILE_Read2.fastq --output OUTPUT.bam --reference GENOME.ehm --threads N
```

Conversion efficiency was calculated with ns-methylation-statistics, a tool included in the in-house developed package NGS-SUITE v1.3 (http://ngs-suite.sf.net) with the following command

```
ns-methylation-statistics --input OUTPUT.bam --reference LAMBDA.fasta --output OUTPUT.stats --conversion-check
```

Alignment of Pinot Noir reads on PN40024 reference genome

Fastq files of read 1 and 2 were aligned on PN40024 genome with erne-bs5 with the following command

erne-bs5 --query1 FILE_Read1.fastq --query2 FILE_Read2.fastq --output OUTPUT.bam --reference GENOME.ehm --threads N

Finally the methylome is produced with erne-meth with the following command

erne-meth --fasta GENOME.fasta --input OUTPUT.bam --output-prefix OUTPUT --annotations-erne –deduplicate

Alignment of haplotype specific bisulfite converted reads

Pinot Noir and the PN40024 reference share one haplotype, thereby to align reads on the alternative haplotype it's necessary to construct its sequence in fasta format.

In Prof. Morgante's research team a SNP map of Pinot Noir was available (data not shown), hence taking advantage of GATK (https://www.broadinstitute.org/gatk/) it has been possible to create the alternative reference in which all the nucleotides of the reference genome carrying a SNP were substituted with their homologous nucleotide on the alternative haplotype with the following command.

```
packages/sw/bio/gatk/2.1-13/GenomeAnalysisTK.jar \
-R /MY_reference.fasta \
-T FastaAlternateReferenceMaker \
-o /MY_alternative_reference.fasta \
--variant SNP_map.vcf
```

With the in-house developed ns-disaplotipization from NGS-SUITE, all the paired-reads aligning unambiguously on one of the two haplotypes were selected and stored in separated files with the following command.

ns-disaplotipization --first-bam OUTPUT_REF.bam --second-bam OUTPUT_ALT.bam --first-fasta GENOME_REF.fasta --second-fasta GENOME_ALT.fasta --bs-seq --prefix DISAPLOTIPIZATION

For each haplotype, BAM files were aligned separately in order to create two separated methylomes with erne-meth for the reference and the alternative haplotypes with the following commands.

erne-meth --fasta GENOME_REF.fasta --input DISAPLOTIPIZATION_REF.bam --output-prefix OUTPUT --annotations-erne –deduplicate

erne-meth --fasta GENOME_ALT.fasta --input DISAPLOTIPIZATION_ALT.bam --output-prefix OUTPUT --annotations-erne –deduplicate

Structural Variants Prediction

Prediction of TEs solely present in the reference haplotype of Pinot Noir

To detect SVs present solely in the reference haplotype, a combined approach, which includes DELLY (version 0.3.3), GASV (version 2.0), and an internal pipeline was utilized.

SVs present in the reference genome and absent in the sample are called “deletions” in the reference whereas SVs present only in the sample are called “insertions” in the reference (Hurles at al., 2008, Figure 8)

DELLY and GASV are two free tools designed to find deletions in a reference genome whereby insert size of paired-end reads is greater than expected.

Being the PN40024 reference equal to one haplotype of Pinot Noir, DELLY and GASV deletions will thereby detect hemizygous SVs present solely in the reference haplotype. DELLY was launched with the default parameters, and the output file was filtered for a size of included in 1kb – 25 kbs range and with at least 2 paired end reads supporting the deletion events. Consistently with the parameters set in DELLY, GASV was launched with the option *minClusterSize* =2 which requires at least 2 paired reads for the deletions prediction and was filter for the same size range of DELLY.

DELLY and GASV outputs were integrated as long as their coordinates differed for less than 250 bp and when not coincident, DELLY coordinates were considered.

An internal Python Script (Pinosio, Personal communication) was developed to evaluate the ratio between the number of reads supporting the deletion and the total number of reads supporting either the deletion event or the reference genotype.

The analysis was computed on regions of 500 bp flanking both sided of SVs coordinates. SVs with a ratio <0.25 were considered false positives and then discarded.

Since deletion sequence is present in the reference, an internal pipeline has been developed in order to annotate the potential presence of TEs in each deletion.

This pipeline takes advantage of a grapevine specific set of 202 TEs obtained from RepBase

(Jurka et al., 2005), and an internal database of 467 TEs and includes the usage of Tandem Repeat Finder (Benson, 1999), RepeatMasker (Smit AFA, Hubley R & Green P, *RepeatMasker Open-3.0*). LTR_finder (Xu & Wang, 2007), REPET (Flutre et al., 2011) to provide for the annotation of the superfamilies of TE involved in the SVs, where present.

Prediction of TEs solely present in the alternative haplotype of Pinot Noir

Since DELLY and GASV are not efficient in detecting insertion events involving large SVs such as those caused by TEs, for SVs solely present in the alternative haplotype, generally known as insertions, an internal python script has been developed for their detection. (Sara Pinosio, personal communication).

Whereby there is an insertion in the reference genome, within the paired-reads spanning the insertions site, only one of the two will map on the genome whereas the other will map on a database of TE termini (Figure 19). Database of TEs used for the insertion detection was enriched with the deletion set produced in the previous step.

Reads mapping the flanking regions of each putative insertion point where assembled though CAP3 (Huang & Madan, 1999), creating a consensus sequences that was aligned with blastn (Altschul, Gish, Miller, Myers, & Lipman, 1990) on PN40024 reference genome. If the reconstructed consensus sequences mapped with opposite orientation and at distance lower than the mean sequenced library insert size, a putative insertion site was identified. Orphans reads were aligned with CAP3 and the consensus created was aligned with blastn on the 500 termini of TE database (with the only exception of LINE in which the whole sequence were considered for the 5' terminal because of the often truncated 5' end) in order to annotate the superfamily of TE involved, where present.

Consistently with deletions, the ratio between the number of reads supporting the insertion and the total number of reads supporting either the insertion event or the reference genotype was computed in order to discard insertions with a lower than 0.25 ratio, presumably false positives. Whereby insertion site prediction is not represented by a single nucleotide, the mean position of the interval was considered the insertion point.

Haplotype specific homozygous TE data of both deletions and insertions were obtained from our group of research (Gabriele Magris, PhD thesis).

Genomic landscape analyses

Genome was divided in 200 kbp regions, for each region the average methylation level of CG, CHG, CHH contexts have been calculated, as well as the number of genes and full-length TEs; the frequency of CG dinucleotides; the number of hemizygous TEs; and the frequency of Pinot Noir, Schiava grossa and PN40024 SNPs.

All the analyses in this work were computed considering only Cytosines with a minimal coverage of at least 4xfor CG context and 10xfor the CHG and CHH contexts.

Circos graphs

Figures 14, 18 and 21 were produced with the Circos software (http://circos.ca/software/, Krzywinski et al., 2009).

Correlation between regional methylation and gene expression

Expression data of several Grapevine varieties were already available in Prof. Morgante's team, including Pinor Noir. Genes were grouped in 10 progressive classes according to the average methylation level of the 200 bp window in which they locate, for both CG and CHG independently.

Log_{10} of FPKM (Fragments Per Kilobase Of Exon Per Million Fragments Mapped) of the genes belonging to each class were plotted in a Boxplot trough R function **boxplot()** in Figure 15a whereas the numerosity of each class was plotted in Figure15b.

Cuffdiff 2 (Trapnell et al., 2013) was used to calculate the log2 ratio between Pinot Noir and Traminer FPKM in Figure 15c.

Significant differences between the sets of gene grouped by regional methylation class with the genome fraction of DEGs according to Chi-squared test (p-value <=0.05) are marked with a *.

Identification of Pinot Noir derived regions in the PN40024 reference

SNP maps of Pinot Noir, Schiava grossa and PN40024 were available in Prof. Morgante's research group.

Pinot Noir regions in the PN40024 reference were assigned where there was a lack of homozygous SNPs between Pinot Noir and PN40024, similarly Schiava grossa regions were assigned whereby there was a lack of homozygous SNPs between Schiava grossa and PN40024. Regions in which there is a lack of both Pinot Noir and Schiava grossa

homozygous SNPs, are considered equal in the two varieties and then are not assignable unambiguously to one of the two.

Transposon body methylation profile

Full-length TEs were searched across the genome through Blastn and Repeat Masker using as a query the sets of grapevine-specific TEs previously mentioned.

Full-length TEs were accepted whether they show at least 80% of nucleotide identity and a length included in the size range between 80% and 125% of the query length.

Coordinates of LTR and TIR elements were extended of 40 bp on both termini in order to verify the presence of the TSDs and Terminal Repeats. For LTR elements LTR-finder (Xu & Wang, 2007) was used, and all elements not showing LTRs were discarded.

For TIRs element, an internal script provided to confirm the presence of both TIRs and TSDs according to Wicker et al. (2007) indications. (Figure 6)

LINE element termini don't show repetitive sequence, their often truncated 5' and the variable TSD length lead us to consider blast and repeat masker coordinates the most reliable.

Solo-LTR termini were predicted by Delly and Gasv.

TE coordinates were then extended of 2500 bp in order to include a part of flanking regions in the analyses. LINE element regions were divided in 3 parts: upstream, LINE body and downstream whereas LTR and TIR element regions were divided in 5 parts (upstream, terminal repeat 1, TE body, terminal repeat 2, downstream) and their methylation percentile in all the three contexts were calculated separately.

For each TE group data were collapsed and their methylation profiles of the three contexts were computed independently with the R function **smoothingSpline()** with the parameter spar=0.5 and plotted with the **plot()** function in the Figure 17.

Analysis of hemizygous TE flanking regions

Individual hemizygous TE representation

For the single TEs representation in Figure 22, only reference specific hemizygous TEs, whose internal sequences is available in the reference genome, were considered. For each TE, both TE sequences and 2000 flanking regions from each side were divided in 10 bp windows, and for each window the mean coverage, and the mean methylation levels of the three contexts in both haplotypes were computed and plotted in separated lines. TE scheme was drawn according to LTRs coordinates predicted by LTR-Finder.

Hemizygous TE methylation profile

For each TE only flanking regions were considered and the data belonging to the TE-carrying or unaffected haplotype were collapsed separately.

The R function **smoothingSpline()** (spar=1.0) was used to create the methylation profile of the two haplotypes that were plotted in the same figure, in which the 0 represents the insertion point. (Figure 23b-c).

In Figure 24a-b, the methylation profile was computed in independently for reference and alternative haplotypes specific TEs. In Figure 24c, the average methylation profiles of Figures24a-b is reported. Figure 24d display the analysis of 1000 regions of 3 kb length chosen randomly in the genome for the two haplotypes and for the three contexts separately, whose average values are reported in boxplot.

Figure 41 display the methylation patter of LINEs 10kb-flanking regions, whereas in Figures 42 and 43 is shown the methylation pattern of LINEs and Ty1-Copia respectively in Intergenic, exonic and intronic regions separately according to GFF annotations (see next Chapter).

Fisher's Exact Test

To provide statistical support to the methylation profile data, a Fisher's exact test was performed for each Cytosins with a sufficient coverage in both haplotype, to verify the hypothesis of a differential methylation of a single cytosine in the two alleles.

Cytosines whose contexts differ in the two haplotypes because of the presence of a SNP were not considered. The Fisher's Exact test was performed with the R function **fisher.test()** with all the default parameter (including the alternative hypothesis = "two-sided").
The **fisher.test()** function requires a 2x2 input matrix as shown in the following table.

	allele with TE	uneffected allele
reads suppoting ^{5m}C	a	b
reads suppoting C	c	d

Flanking regions were divided in 500 bp windows, and for each window the number of differentially methylated cytosines (according to a p-value <=0.01) was computed, in both the TE and unaffected haplotypes.
Figure 25c-d shows the fraction of the significant more methylated Cytosines in the TE haplotype for each 500 bp window, whereas Figure 25 e-f shows the same fraction considering only cytosines with a significant differential methylation according to the Fisher test

Chi-squared test

To test whether the differential distribution of more methylated cytosines in the two haplotypes was significant, the Chi-Squared test was performed to verify the null hypothesis of a 50:50 distribution of more methylated cytosines in the two haplotypes with the R function **chisq.test()** Figure 26g-h reports the –log(p-value) of such test.

Wilcoxon Mann Whitney test

Flanking regions of each TE were divided in 500 bp windows and for each window the average methylation level in both haplotypes and both for CG and CHG contexts was calculated.
Successively, for each window and for each context, the entire list of values of all TEs of the two haplotype were tested with the R function **wilcox.test()** with the parameter exact=FALSE, alternative="two-sided".
In the Figure 25j-k is shown the –log(p-value) of such test for each window, in order to have greater values than y=2 when p-value is lower than 0.01.

Single-TE analyses

The average methylation value of a 2kb-wide region around the insertion point in both haplotypes was considered to evaluate at single-TE resolution the effect of the insertion on the flanking regions

For each TE the difference of average methylation level between the TE-carrying and unaffected haplotype was computed for CG, CHG, CHH and CG + CHG contexts independently and shown in the form of histogram in the column a) of Figures 26-31

The values of the two haplotypes were represented in a dotplot in the column c) of Figures 26-31, with TE-carrying haplotype on the x axis and the unaffected haplotype on the y axis, in order to have points underneath the bisector line when higher methylation level is present in the TE-carrying haplotype.

To provide statistical support to this analysis, for each hemizygous TE the Wilcoxon Mann-Whitney test was performed with the R function **wilcox.test()** with the parameter exact=FALSE and alternative="two-sided", giving as input the methylation values of all the cytosines in the 2kb region of the two haplotype respectively

Columns b) and d) of Figures 26-31 represent the subset of significant differences of methylation with a p-value<=0.01 of columns a) and c) respectively.

Gene body methylation

Gene prediction of the Grape genome database of university of Padua (http://genomes.cribi.unipd.it/DATA/V2/V2.1/V2.1.gff3, Vitulo et al., 2014), was utilized for the whole gene body analysis.

18986 genes showing both 5' and 3' UTRs were selected, and 2500 bp of flanking regions on both sides were included in all the analyses.

For each gene, methylation in the exons, upstream and downstream regions were expressed in percentiles for the three context independently, then for each context all data were collapsed and the methylation profile were computed with the R function **smoothingSpline().**

In Figure 32a is reported the gene body methylation profile of the exons whereas in 32b introns were exceptionally included.

In order to group genes on the basis of their internal methylation, the average CG methylation level of the bell-peak was computed for each gene between the 50^{th} and 65^{th} percentiles. Genes were grouped in 10 progressive classes according to their CG methylation level at the

bell-peak and the methylation profile of each class was computed with the R function **smoothingSpline()** (spar=1.0) and plotted in Figure 33.

Figure 34a reports the frequency of each class whereas Figure 34b shows in the form of boxplot the $\log_{10}$(FPKM) of the genes belonging to each class.

All the classes were compared trough the Wilcoxon Mann-Whitney test and a letter code was assigned in order to have non-significant differences with the same letter code. Gene length, exon number, total exon space and total intron space were reported according to the methylation class respectively in Figures 35.

Moreover, for each class the number of TE annotations obtained by an internal database was plotted in the Figure 37.

Finally genes were grouped according to their exon number and the methylation profile of each group was computed for all exons and introns separately with the R function **smoothingSpline()** (spar=1.0) and plotted in Figure 36.

To evaluate TE effect on GBM, the exon-intron profile was computed after excluding carrying TE annotation in their introns (Figure 38)

Similarly to Figure 33a, Figure 39 reports the gene body methylation profile with the exon 1 separated from the other exons. Figure 41 shows in the form of boxplot the $\log_{10}$(FPKM) of the genes according on their average methylation in exon 1

A Wilcoxon Mann-Whitney test was performed among all the 10 classes of GBM in exon1 and the results is reported in the table 4 where same letter code correspond to non-significant differences.

Correlation between TE presence and gene expression

Figure 45 displays Boxplot of Log_{10} of FPKM of the genes associated to a TE in their intron, 2500 bp upstream or 2500 bp downstream. Significantly differences against the set of unaffected genes according the Wilcoxon Mann-Whitney test are marked with *.

Haplotype specific expression

RNA-seq data were available in professor Morgante's group (Eleonora Paparelli, PhD) for all the replicates taken in analysis.

Allim (Allelic imbalance meter, Pandey et al., 2013), was used to measure allele specific gene expression (ASE) in the two Pinot Noir haplotypes.

Figure 45 show the log2 ratio of genes unrelated to SVs computed as reference_allele / alternative_allele and vice-versa respectively.
In Figure 47, the log2 ratio in SV-related genes, is always calculated in the ratio SV-allele / unaffected allele.
Since SVs are belonging for the 60% of the occurrences to the alternative haplotype and for the 40% to the reference haplotype, the log2 ratio of genes unrelated to SVs is calculated for the 40% of the genes, chosen randomly, as "reference_allele / alternative_allele" and in the remaining 60% of the genes it calculated as "alternative allele / reference allele".
Boxplot marked with a red * show a significant lower log2ratio of SV-genes (one-tailed Wilcoxon Mann-Whitney test; p-value <0.005), whereas green * indicate significant higher values with the same test.

REFERENCES

Adams, M. D., Celniker, S. E., Holt, R. A., Evans, C. A., Gocayne, J. D., Amanatides, P. G., … Venter, J. C. (2000). The genome sequence of Drosophila melanogaster. *Science (New York, N.Y.)*, *287*(5461), 2185–95. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/10731132

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, *215*(3), 403–10. http://doi.org/10.1016/S0022-2836(05)80360-2

Arteaga-Vazquez, M. A., & Chandler, V. L. (2010). Paramutation in maize: RNA mediated trans-generational gene silencing. *Current Opinion in Genetics & Development*, *20*(2), 156–63. http://doi.org/10.1016/j.gde.2010.01.008

Baucom, R. S., Estill, J. C., Chaparro, C., Upshaw, N., Jogi, A., Deragon, J.-M., … Bennetzen, J. L. (2009). Exceptional Diversity, Non-Random Distribution, and Rapid Evolution of Retroelements in the B73 Maize Genome. *PLoS Genetics*, *5*(11), e1000732. http://doi.org/10.1371/journal.pgen.1000732

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

Brenet, F., Moh, M., Funk, P., Feierstein, E., Viale, A. J., Socci, N. D., & Scandura, J. M. (2011). DNA methylation of the first exon is tightly linked to transcriptional silencing. *PloS One*, *6*(1), e14524. http://doi.org/10.1371/journal.pone.0014524

Cao, X., & Jacobsen, S. E. (2002). Role of the arabidopsis DRM methyltransferases in de novo DNA methylation and gene silencing. *Current Biology : CB*, *12*(13), 1138–44. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/12121623

Chodavarapu, R. K., Feng, S., Bernatavichute, Y. V., Chen, P.-Y., Stroud, H., Yu, Y., … Pellegrini, M. (2010). Relationship between nucleosome positioning and DNA methylation. *Nature*, *466*(7304), 388–392. http://doi.org/10.1038/nature09147

Cokus, S. J., Feng, S., Zhang, X., Chen, Z., Merriman, B., Haudenschild, C. D., … Jacobsen, S. E. (2008). Shotgun bisulphite sequencing of the Arabidopsis genome reveals DNA methylation patterning. *Nature*, *452*(7184), 215–9. http://doi.org/10.1038/nature06745

Eichten, S. R., Ellis, N. a, Makarevitch, I., Yeh, C.-T., Gent, J. I., Guo, L., … Springer, N. M. (2012). Spreading of heterochromatin is limited to specific families of maize retrotransposons. *PLoS Genetics*, *8*(12), e1003127. http://doi.org/10.1371/journal.pgen.1003127

Emberton, J., Ma, J., Yuan, Y., SanMiguel, P., & Bennetzen, J. L. (2005). Gene enrichment in maize with hypomethylated partial restriction (HMPR) libraries. *Genome Research*, *15*(10), 1441–6. http://doi.org/10.1101/gr.3362105

Feng, S., Cokus, S. J., Zhang, X., Chen, P.-Y., Bostick, M., Goll, M. G., … Jacobsen, S. E. (2010). Conservation and divergence of methylation patterning in plants and animals. *Proceedings of the National Academy of Sciences of the United States of America*, *107*(19), 8689–94. http://doi.org/10.1073/pnas.1002720107

Feng, S., & Jacobsen, S. E. (2011). Epigenetic modifications in plants: An evolutionary perspective. *Current Opinion in Plant Biology*, *14*(2), 179–186. http://doi.org/10.1016/j.pbi.2010.12.002

Finnegan, D. J. (1989). Eukaryotic transposable elements and genome evolution. *Trends in Genetics : TIG*, *5*(4), 103–7. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/2543105

Flavell, R. B., Rimpau, J., & Smith, D. B. (1977). Repeated sequence DNA relationships in four cereal genomes. *Chromosoma*, *63*(3), 205–222. http://doi.org/10.1007/BF00327450

Flutre, T., Duprat, E., Feuillet, C., & Quesneville, H. (2011). Considering transposable element diversification in de novo annotation approaches. *PloS One*, *6*(1), e16526. http://doi.org/10.1371/journal.pone.0016526

Frankel, A. D., & Young, J. A. (1998). HIV-1: fifteen proteins and an RNA. *Annual Review of Biochemistry*, *67*, 1–25. http://doi.org/10.1146/annurev.biochem.67.1.1

Gao, D., Chen, J., Chen, M., Meyers, B. C., & Jackson, S. (2012). A highly conserved, small LTR retrotransposon that preferentially targets genes in grass genomes. *PloS One*, *7*(2), e32010. http://doi.org/10.1371/journal.pone.0032010

Gao, Z., Liu, H.-L., Daxinger, L., Pontes, O., He, X., Qian, W., … Zhu, J.-K. (2010). An RNA polymerase II- and AGO4-associated protein acts in RNA-directed DNA methylation. *Nature*, *465*(7294), 106–9. http://doi.org/10.1038/nature09025

Gent, J. I., Ellis, N. A., Guo, L., Harkess, A. E., Yao, Y., Zhang, X., & Dawe, R. K. (2013). CHH islands : de novo DNA methylation in near-gene chromatin regulation in maize. *Genome Research*, *23*, 628–637. http://doi.org/10.1101/gr.146985.112.as

Hodges, C., Bintu, L., Lubkowska, L., Kashlev, M., & Bustamante, C. (2009). Nucleosomal fluctuations govern the transcription dynamics of RNA polymerase II. *Science (New York, N.Y.)*, *325*(5940), 626–8. http://doi.org/10.1126/science.1172926

Hsieh, T.-F., Ibarra, C. A., Silva, P., Zemach, A., Eshed-Williams, L., Fischer, R. L., & Zilberman, D. (2009). Genome-wide demethylation of Arabidopsis endosperm. *Science (New York, N.Y.)*, *324*(5933), 1451–4. http://doi.org/10.1126/science.1172417

Huang, X., & Madan, A. (1999). CAP3: A DNA sequence assembly program. *Genome Research*, *9*(9), 868–77. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=310812&tool=pmcentrez&rendertype=abstract

Hurles, M. E., Dermitzakis, E. T., & Tyler-Smith, C. (2008). The functional impact of structural variation in humans. *Trends in Genetics : TIG*, *24*(5), 238–45. http://doi.org/10.1016/j.tig.2008.03.001

Jaillon, O., Aury, J.-M., Noel, B., Policriti, A., Clepet, C., Casagrande, A., … Wincker, P. (2007). The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature*, *449*(7161), 463–467. http://doi.org/10.1038/nature06148

Jain, S., Xie, L., Boldbaatar, B., Lin, S. Y., Hamilton, J. P., Meltzer, S. J., … Su, Y.-H. (2015). Differential methylation of the promoter and first exon of the RASSF1A gene in hepatocarcinogenesis. *Hepatology Research : The Official Journal of the Japan Society of Hepatology*, *45*(11), 1110–1123. http://doi.org/10.1111/hepr.12449

Jullien, P. E., Susaki, D., Yelagandula, R., Higashiyama, T., & Berger, F. (2012). DNA methylation dynamics during sexual reproduction in Arabidopsis thaliana. *Current Biology : CB*, *22*(19), 1825–30. http://doi.org/10.1016/j.cub.2012.07.061

Jurka, J., Kapitonov, V. V, Pavlicek, A., Klonowski, P., Kohany, O., & Walichiewicz, J. (2005). Repbase Update, a database of eukaryotic repetitive elements. *Cytogenetic and Genome Research*, *110*(1-4), 462–7. http://doi.org/10.1159/000084979

Krzywinski, M., Schein, J., Birol, I., Connors, J., Gascoyne, R., Horsman, D., … Marra, M. A. (2009). Circos: An information aesthetic for comparative genomics. *Genome Research*, *19*(9), 1639–1645. http://doi.org/10.1101/gr.092759.109

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., … Szustakowki, J. (2001). Initial sequencing and analysis of the human genome. *Nature*, *409*(6822), 860–921.

http://doi.org/10.1038/35057062

Laurent, L., Wong, E., Li, G., Huynh, T., Tsirigos, A., Ong, C. T., … Wei, C.-L. (2010). Dynamic changes in the human methylome during differentiation. *Genome Research*, *20*(3), 320–31. http://doi.org/10.1101/gr.101907.109

Law, J. A., Ausin, I., Johnson, L. M., Vashisht, A. A., Zhu, J.-K., Wohlschlegel, J. A., & Jacobsen, S. E. (2010). A protein complex required for polymerase V transcripts and RNA- directed DNA methylation in Arabidopsis. *Current Biology : CB*, *20*(10), 951–6. http://doi.org/10.1016/j.cub.2010.03.062

Lee, C.-J., Evans, J., Kim, K., Chae, H., & Kim, S. (2014). Determining the effect of DNA methylation on gene expression in cancer cells. *Methods in Molecular Biology (Clifton, N.J.)*, *1101*, 161–78. http://doi.org/10.1007/978-1-62703-721-1_9

Lev Maor, G., Yearim, A., & Ast, G. (2015). The alternative role of DNA methylation in splicing regulation. *Trends in Genetics : TIG*, *31*(5), 274–80. http://doi.org/10.1016/j.tig.2015.03.002

Lisch, D. (2002). Mutator transposons. *Trends in Plant Science*, *7*(11), 498–504. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/12417150

Lister, R., O'Malley, R. C., Tonti-Filippini, J., Gregory, B. D., Berry, C. C., Millar, a. H., & Ecker, J. R. (2008). Highly Integrated Single-Base Resolution Maps of the Epigenome in Arabidopsis. *Cell*, *133*(3), 523–536. http://doi.org/10.1016/j.cell.2008.03.029

Liu, J., He, Y., Amasino, R., & Chen, X. (2004). siRNAs targeting an intronic transposon in the regulation of natural flowering behavior in Arabidopsis. *Genes & Development*, *18*(23), 2873–8. http://doi.org/10.1101/gad.1217304

Lyko, F., Foret, S., Kucharski, R., Wolf, S., Falckenhayn, C., & Maleszka, R. (2010). The honey bee epigenomes: differential methylation of brain DNA in queens and workers. *PLoS Biology*, *8*(11), e1000506. http://doi.org/10.1371/journal.pbio.1000506

Ma, J., Devos, K. M., & Bennetzen, J. L. (2004). Analyses of LTR-retrotransposon structures reveal recent and rapid genomic DNA loss in rice. *Genome Research*, *14*(5), 860–9. http://doi.org/10.1101/gr.1466204

Makarevitch, I., Waters, A. J., West, P. T., Stitzer, M., Hirsch, C. N., Ross-Ibarra, J., & Springer, N. M. (2015). Transposable Elements Contribute to Activation of Maize Genes in Response to Abiotic Stress. *PLoS Genetics*, *11*(1), e1004915. http://doi.org/10.1371/journal.pgen.1004915

Morgante, M., Brunner, S., Pea, G., Fengler, K., Zuccolo, A., & Rafalski, A. (2005). Gene duplication and exon shuffling by helitron-like transposons generate intraspecies diversity in maize. *Nature Genetics*, *37*(9), 997–1002. http://doi.org/10.1038/ng1615

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

Naito, K., Zhang, F., Tsukiyama, T., Saito, H., Hancock, C. N., Richardson, A. O., … Wessler, S. R. (2009). Unexpected consequences of a sudden and massive transposon amplification on rice gene expression. *Nature*, *461*(7267), 1130–4. http://doi.org/10.1038/nature08479

Palmer, L. E., Rabinowicz, P. D., O'Shaughnessy, A. L., Balija, V. S., Nascimento, L. U., Dike, S., … McCombie, W. R. (2003). Maize genome sequencing by methylation filtration. *Science (New York, N.Y.)*, *302*(5653), 2115–7. http://doi.org/10.1126/science.1091265

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

Patterson, G. I., Thorpe, C. J., & Chandler, V. L. (1993). Paramutation, an allelic interaction, is associated with a

stable and heritable reduction of transcription of the maize b regulatory gene. *Genetics*, *135*(3), 881–94. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1205727&tool=pmcentrez&rendertype=abstract

Rabinowicz, P. D., Schutz, K., Dedhia, N., Yordan, C., Parnell, L. D., Stein, L., … Martienssen, R. A. (1999). Differential methylation of genes and retrotransposons facilitates shotgun sequencing of the maize genome. *Nature Genetics*, *23*(3), 305–8. http://doi.org/10.1038/15479

Regner, F., Stadlbauer, A., Eisenheld, C., & Kaserer, H. (2000). Genetic Relationships Among Pinots and Related Cultivars. *Am. J. Enol. Vitic.*, *51*(1), 7–14. Retrieved from http://www.ajevonline.org/content/51/1/7.abstract

Saze, H., & Kakutani, T. (2007). Heritable epigenetic mutation of a transposon-flanked Arabidopsis gene due to lack of the chromatin-remodeling factor DDM1. *The EMBO Journal*, *26*(15), 3641–52. http://doi.org/10.1038/sj.emboj.7601788

Schmitz, R. J., He, Y., Valdés-lópez, O., Res, G., Gent, J. I., Ellis, N. a, … Ecker, J. R. (2013). Epigenome-wide inheritance of cytosine methylation variants in a recombinant inbred population Epigenome-wide inheritance of cytosine methylation variants in a recombinant inbred population, 1–13. http://doi.org/10.1101/gr.152538.112

Seelamgari, A., Maddukuri, A., Berro, R., de la Fuente, C., Kehn, K., Deng, L., … Kashanchi, F. (2004). Role of viral regulatory and accessory proteins in HIV-1 replication. *Frontiers in Bioscience : A Journal and Virtual Library*, *9*, 2388–413. Retrieved from https://www.researchgate.net/publication/8361164_Role_of_viral_regulatory_and_accessory_proteins_in_HIV-1_replication

Sengupta, P. K., & Smith, B. D. (1998). Methylation in the initiation region of the first exon suppresses collagen pro-alpha2(I) gene transcription. *Biochimica et Biophysica Acta*, *1443*(1-2), 75–89. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/9838053

Sequencing ProjectInternational Rice Genome. (2005). The map-based sequence of the rice genome. *Nature*, *436*(7052), 793–800. http://doi.org/10.1038/nature03895

Shirasu, K., Schulman, A. H., Lahaye, T., & Schulze-Lefert, P. (2000). A contiguous 66-kb barley DNA sequence provides evidence for reversible genome expansion. *Genome Research*, *10*(7), 908–15. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=310930&tool=pmcentrez&rendertype=abstract

Soppe, W. J., Jacobsen, S. E., Alonso-Blanco, C., Jackson, J. P., Kakutani, T., Koornneef, M., & Peeters, A. J. (2000). The late flowering phenotype of fwa mutants is caused by gain-of-function epigenetic alleles of a homeodomain gene. *Molecular Cell*, *6*(4), 791–802. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/11090618

Takuno, S., & Gaut, B. S. (2013). Gene body methylation is conserved between plant orthologs and is of evolutionary consequence. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(5), 1797–802. http://doi.org/10.1073/pnas.1215380110

Tran, R. K., Henikoff, J. G., Zilberman, D., Ditt, R. F., Jacobsen, S. E., & Henikoff, S. (2005). DNA methylation profiling identifies CG methylation clusters in Arabidopsis genes. *Current Biology : CB*, *15*(2), 154–9. http://doi.org/10.1016/j.cub.2005.01.008

Trapnell, C., Hendrickson, D. G., Sauvageau, M., Goff, L., Rinn, J. L., & Pachter, L. (2013). Differential analysis of gene regulation at transcript resolution with RNA-seq. *Nature Biotechnology*, *31*(1), 46–53. http://doi.org/10.1038/nbt.2450

Urich, M. A., Nery, J. R., Lister, R., Schmitz, R. J., & Ecker, J. R. (2015). MethylC-seq library preparation for base-resolution whole-genome bisulfite sequencing. *Nature Protocols*, *10*(3), 475–83. http://doi.org/10.1038/nprot.2014.114

Vitte, C., Panaud, O., & Quesneville, H. (2007). LTR retrotransposons in rice (Oryza sativa, L.): recent burst amplifications followed by rapid DNA loss. *BMC Genomics*, *8*, 218. http://doi.org/10.1186/1471-2164-8-218

Vitulo, N., Forcato, C., Carpinelli, E. C., Telatin, A., Campagna, D., D'Angelo, M., … Valle, G. (2014). A deep survey of alternative splicing in grape reveals changes in the splicing machinery related to tissue, stress condition and genotype. *BMC Plant Biology*, *14*, 99. http://doi.org/10.1186/1471-2229-14-99

Wassenegger, M., Heimes, S., Riedel, L., & Sänger, H. L. (1994). RNA-directed de novo methylation of genomic sequences in plants. *Cell*, *76*(3), 567–76. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/8313476

Whitelaw, C. A., Barbazuk, W. B., Pertea, G., Chan, A. P., Cheung, F., Lee, Y., … Quackenbush, J. (2003). Enrichment of gene-coding sequences in maize by genome filtration. *Science (New York, N.Y.)*, *302*(5653), 2118–20. http://doi.org/10.1126/science.1090047

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

Wierzbicki, A. T., Cocklin, R., Mayampurath, A., Lister, R., Jordan Rowley, M., Gregory, B. D., … Pikaard, C. S. (2012). Spatial and functional relationships among Pol V-associated loci, Pol IV-dependent siRNAs, and cytosine methylation in the Arabidopsis epigenome. *Genes and Development*, *26*(16), 1825–1836. http://doi.org/10.1101/gad.197772.112

Wierzbicki, A. T., Haag, J. R., & Pikaard, C. S. (2008). Noncoding transcription by RNA polymerase Pol IVb/Pol V mediates transcriptional silencing of overlapping and adjacent genes. *Cell*, *135*(4), 635–48. http://doi.org/10.1016/j.cell.2008.09.035

Wierzbicki, A. T., Ream, T. S., Haag, J. R., & Pikaard, C. S. (2009). RNA polymerase V transcription guides ARGONAUTE4 to chromatin. *Nature Genetics*, *41*(5), 630–4. http://doi.org/10.1038/ng.365

Xing, J., Zhang, Y., Han, K., Salem, A. H., Sen, S. K., Huff, C. D., … Jorde, L. B. (2009). Mobile elements create structural variation: analysis of a complete human genome. *Genome Research*, *19*(9), 1516–26. http://doi.org/10.1101/gr.091827.109

Xu, Z., & Wang, H. (2007). LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Research*, *35*(Web Server issue), W265–8. http://doi.org/10.1093/nar/gkm286

Zhang, X., Yazaki, J., Sundaresan, A., Cokus, S., Chan, S. W.-L., Chen, H., … Ecker, J. R. (2006). Genome-wide high-resolution mapping and functional analysis of DNA methylation in arabidopsis. *Cell*, *126*(6), 1189–201. http://doi.org/10.1016/j.cell.2006.08.003

Zhong, S., Fei, Z., Chen, Y.-R., Zheng, Y., Huang, M., Vrebalov, J., … Giovannoni, J. J. (2013). Single-base resolution methylomes of tomato fruit development reveal epigenome modifications associated with ripening. *Nature Biotechnology*, *31*(2), 154–159. http://doi.org/10.1038/nbt.2462

Zilberman, D., Gehring, M., Tran, R. K., Ballinger, T., & Henikoff, S. (2007). Genome-wide analysis of Arabidopsis thaliana DNA methylation uncovers an interdependence between methylation and transcription. *Nature Genetics*, *39*(1), 61–9. http://doi.org/10.1038/ng1929

Ziller, M. J., Hansen, K. D., Meissner, A., & Aryee, M. J. (2014). Coverage recommendations for methylation analysis by whole-genome bisulfite sequencing. *Nature Methods*, (November), 2–5. http://doi.org/10.1038/nmeth.3152

ACKNOWLEDGMENTS

Firstly, I would like to express my sincere gratitude to my supervisor Prof. Michele Morgante for the continuous support of my Ph.D study and related research, for his patience, motivation, and knowledge. His guidance helped me in all the time of research and writing of this thesis. I could not have imagined having a better advisor and mentor for my Ph.D study.

A very special thanks goes to my co-supervisor Dr. Emanuele De Paoli, who followed me step-by-step for all these three years, for his invaluable teaching, patience and helpfulness.

Besides my supervisors, I would like to thank my thesis reviewers Prof. Emidio Albertini and Prof. Giuseppe Macino, for their insightful comments and suggestions.

My sincere thanks also goes to Dr. Fabio Marroni for his help in statistical analyses; to Dr.Emanuela Aleo for her help in library construction and all the colleagues of Institute of Applied Genomics for their suggestions, support and the friendly atmosphere.

Last but not the least, I would like to thank my family: my parents, my brothers and my girlfriend Valentina for supporting and encouraging me during all these years.

SUPPLEMENTARY DATA

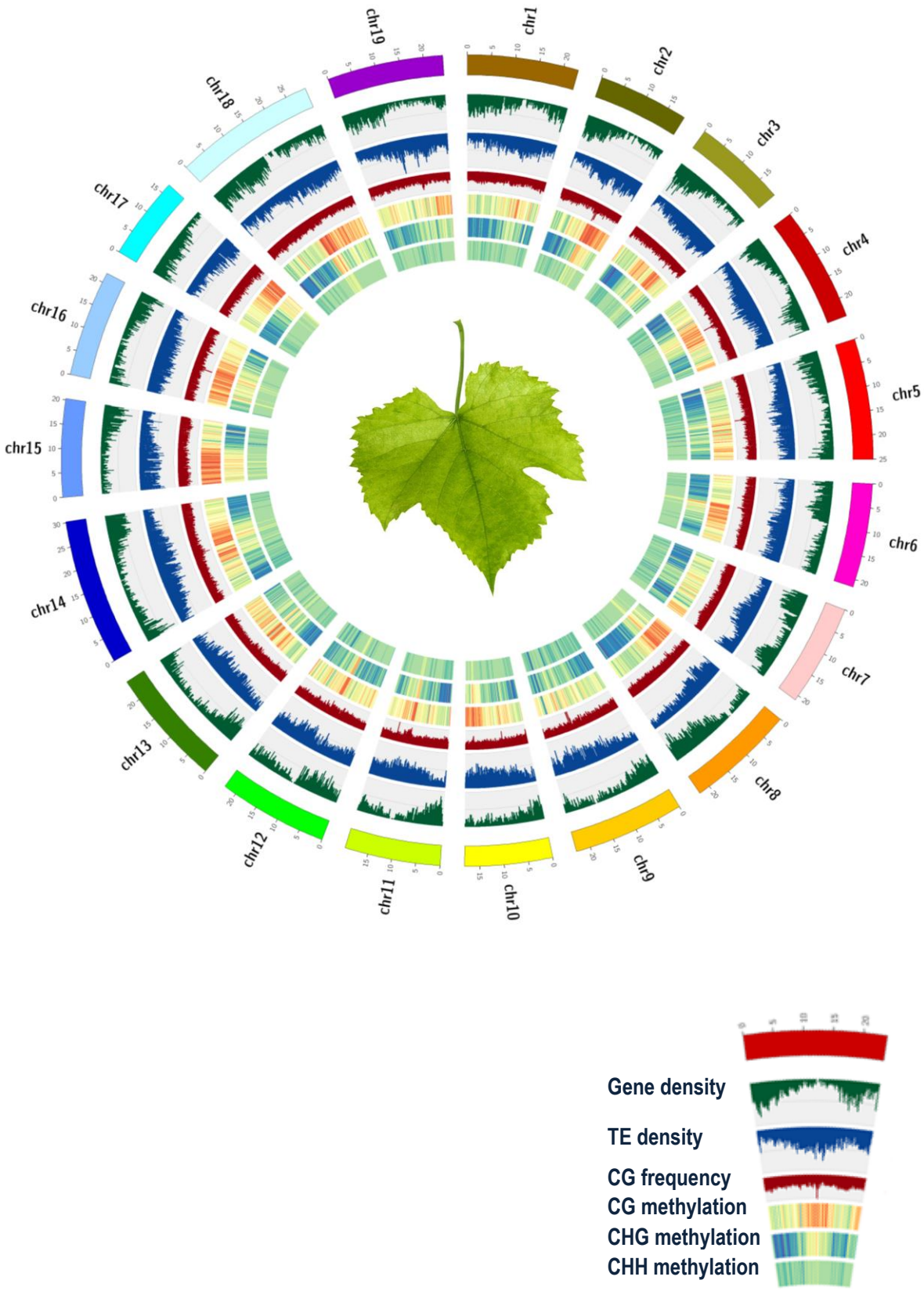


Figure S1 | Circos graph of Grapevine Genome and Methylome (replicate 1). Gene density and TE density, CG frequency and CG, CHG CHH average methylation level are relative to 200 kbp regions. Methylation is expressed in the form of heat map.

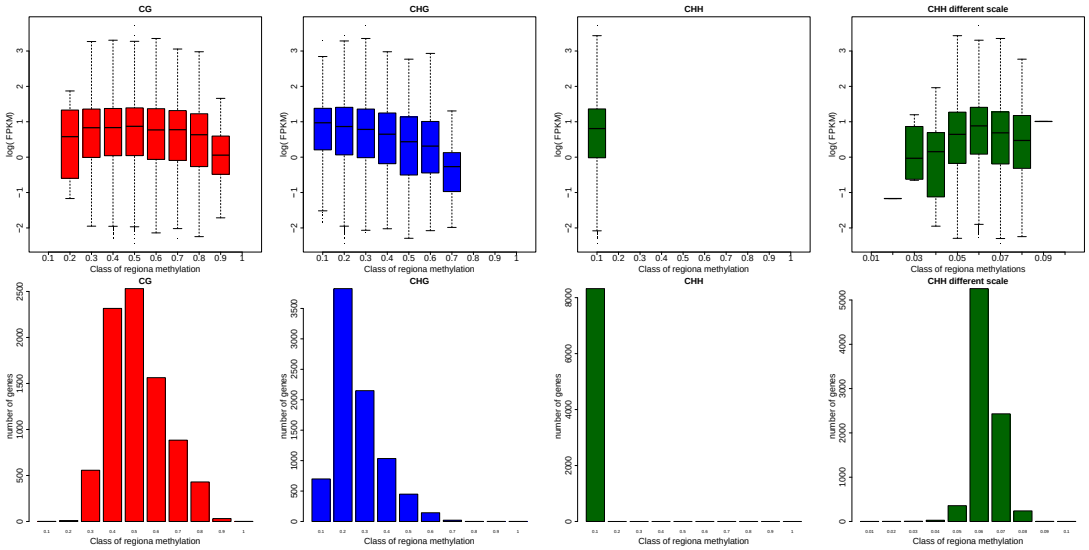


Figure S2
a) Gene expression rate on the basis of the regional methylation for CG and CHG respectively
b) number of occurrences for each class

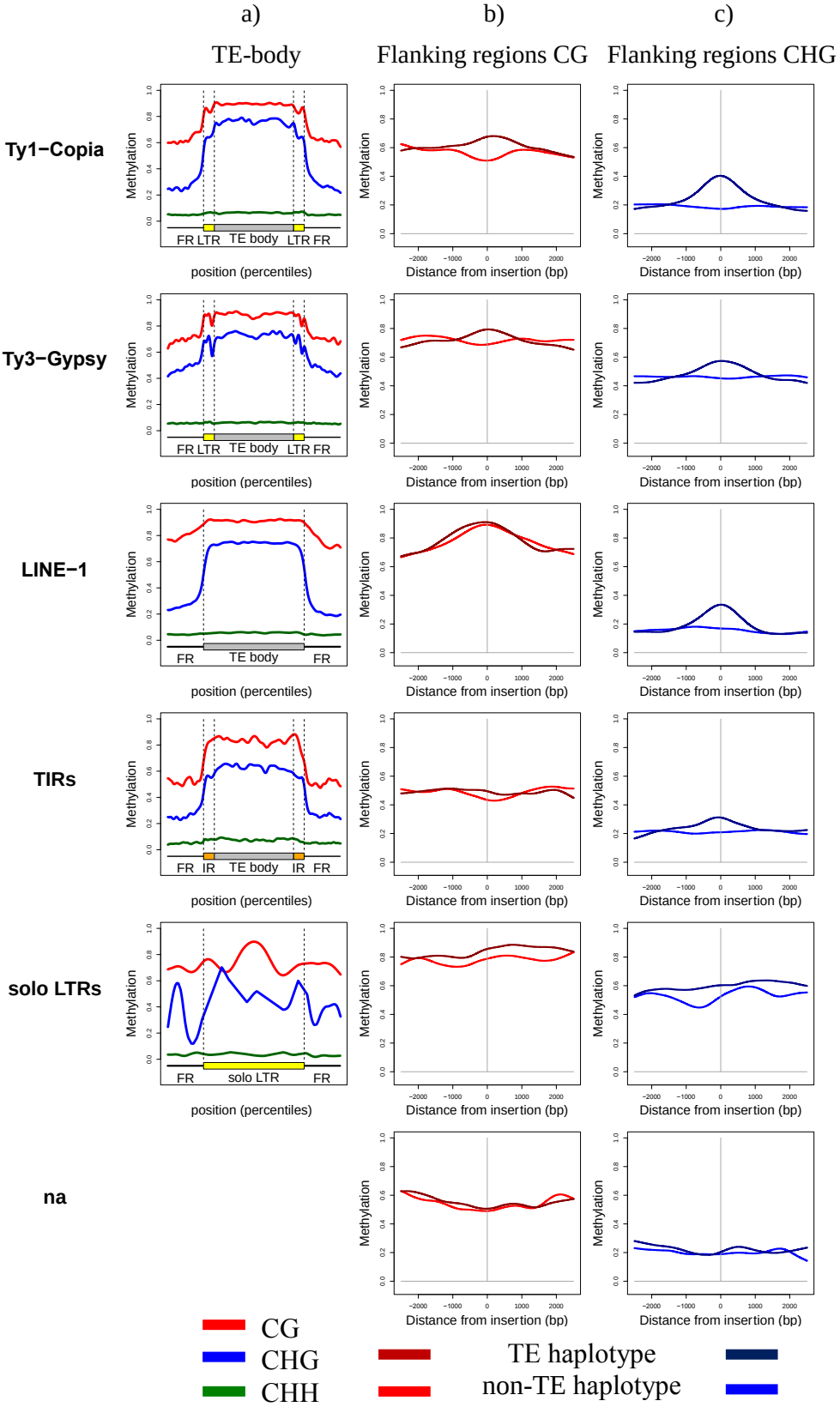
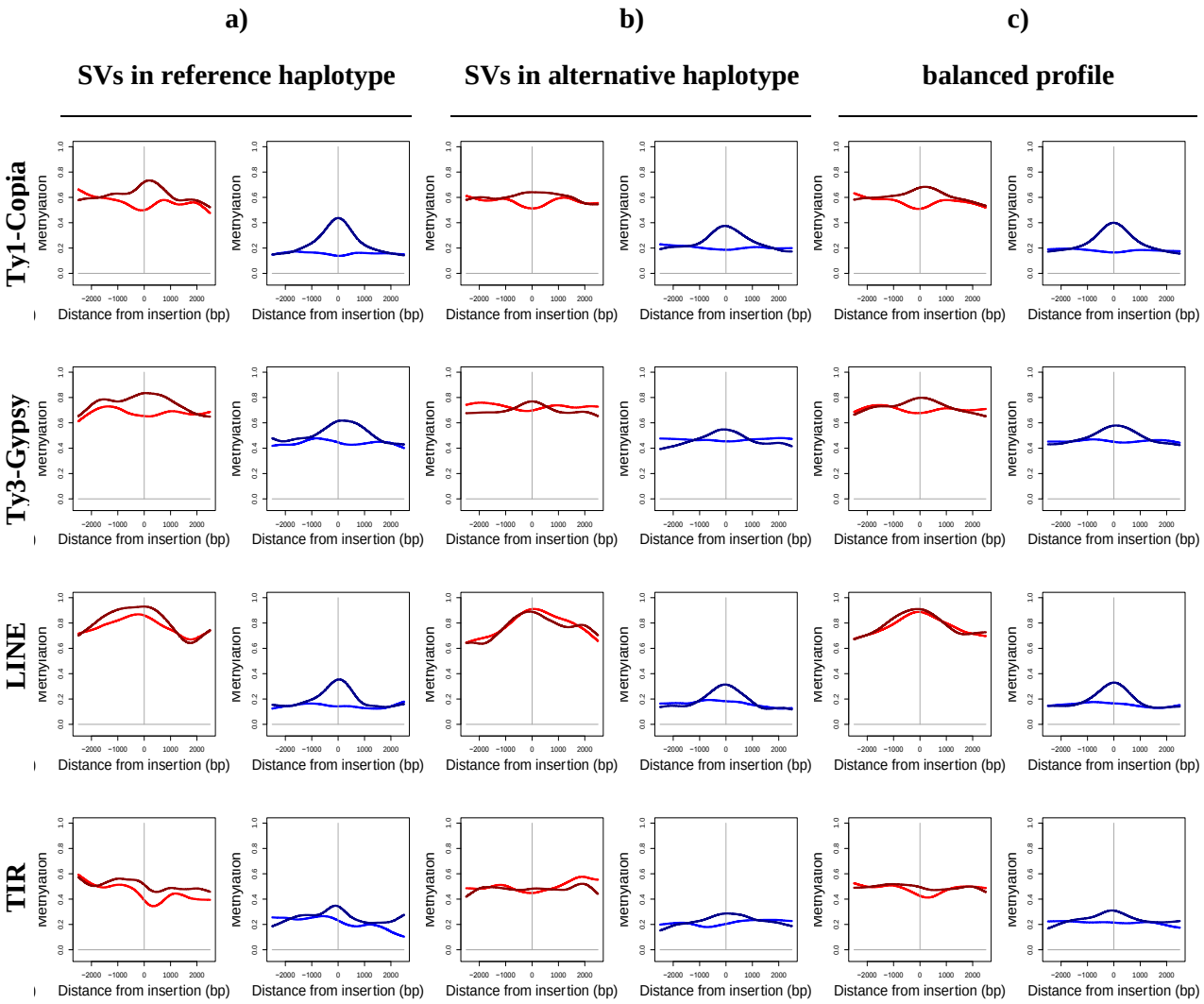


Figure S3 | Methylation profile of both TE bodies and hemizygous TE flanking regions.

d) TE body methylation profile

e) Average CG methylation profile of TE flanking regions in bp from insertion point

f) Average CHG methylation profile of TE flanking regions in bp from insertion point



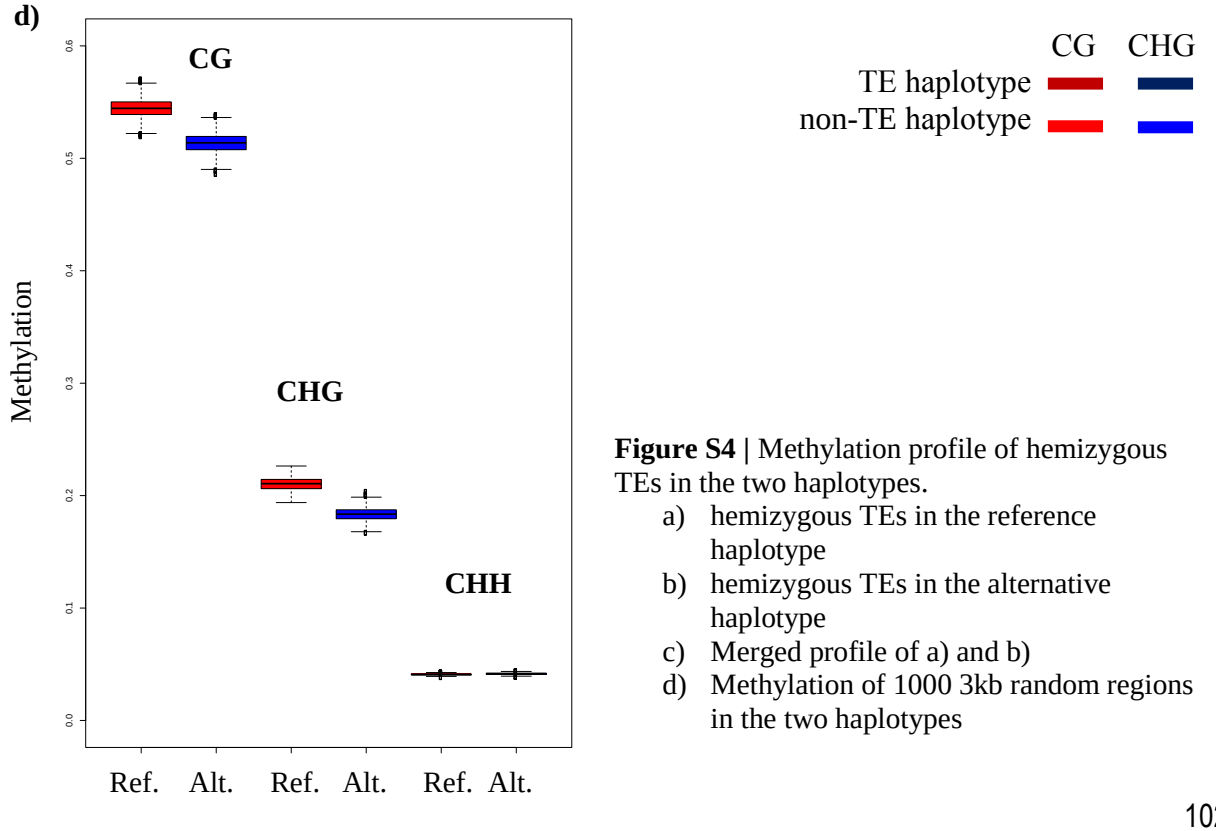


Figure S4 | Methylation profile of hemizygous TEs in the two haplotypes.

a) hemizygous TEs in the reference haplotype
b) hemizygous TEs in the alternative haplotype
c) Merged profile of a) and b)
d) Methylation of 1000 3kb random regions in the two haplotypes

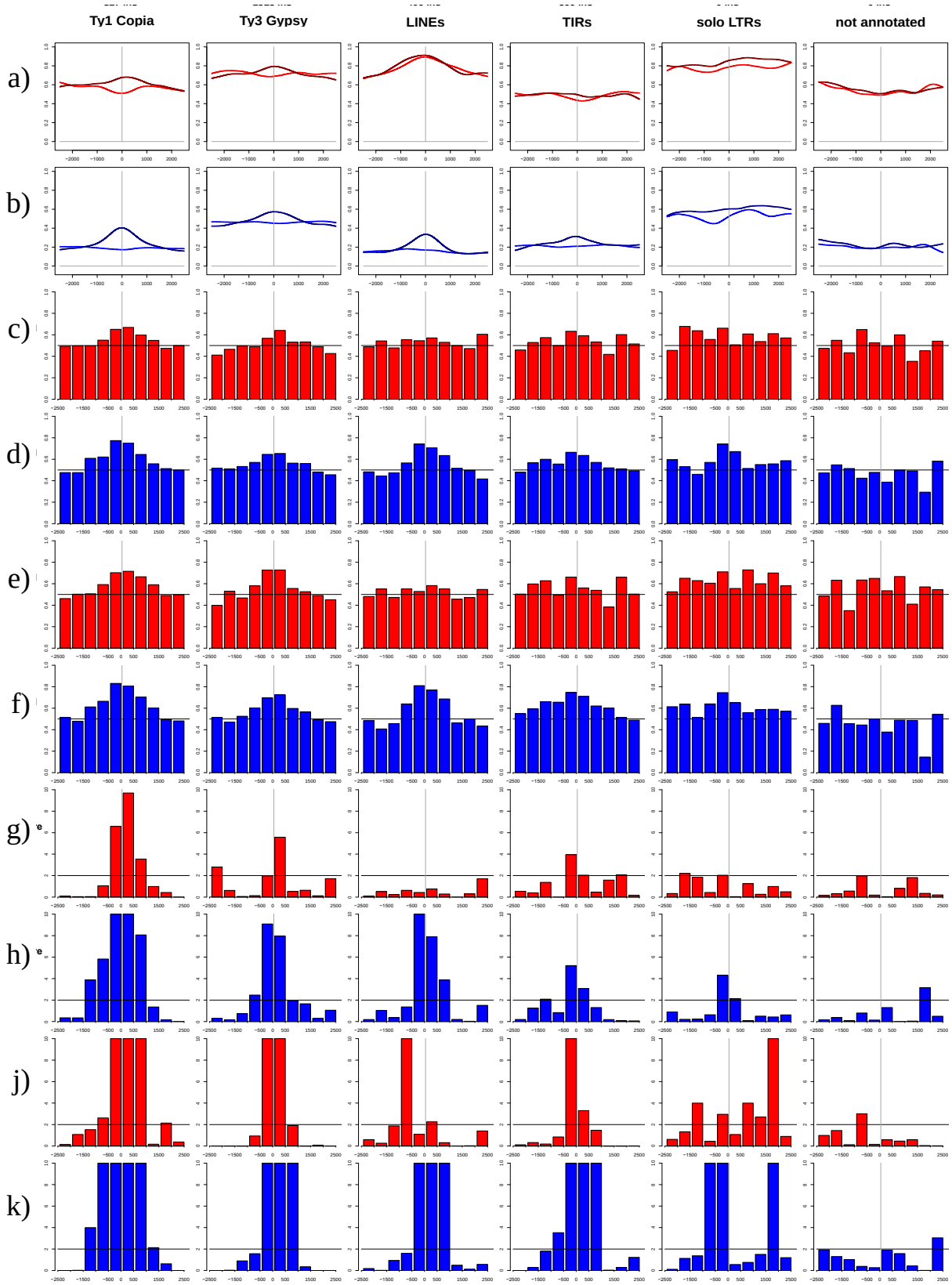


Figure S5 | see next page for the description

Figure S5 | Statistical analysis of between-haplotype differential DNA methylation in the flanking regions of hemizygous TEs

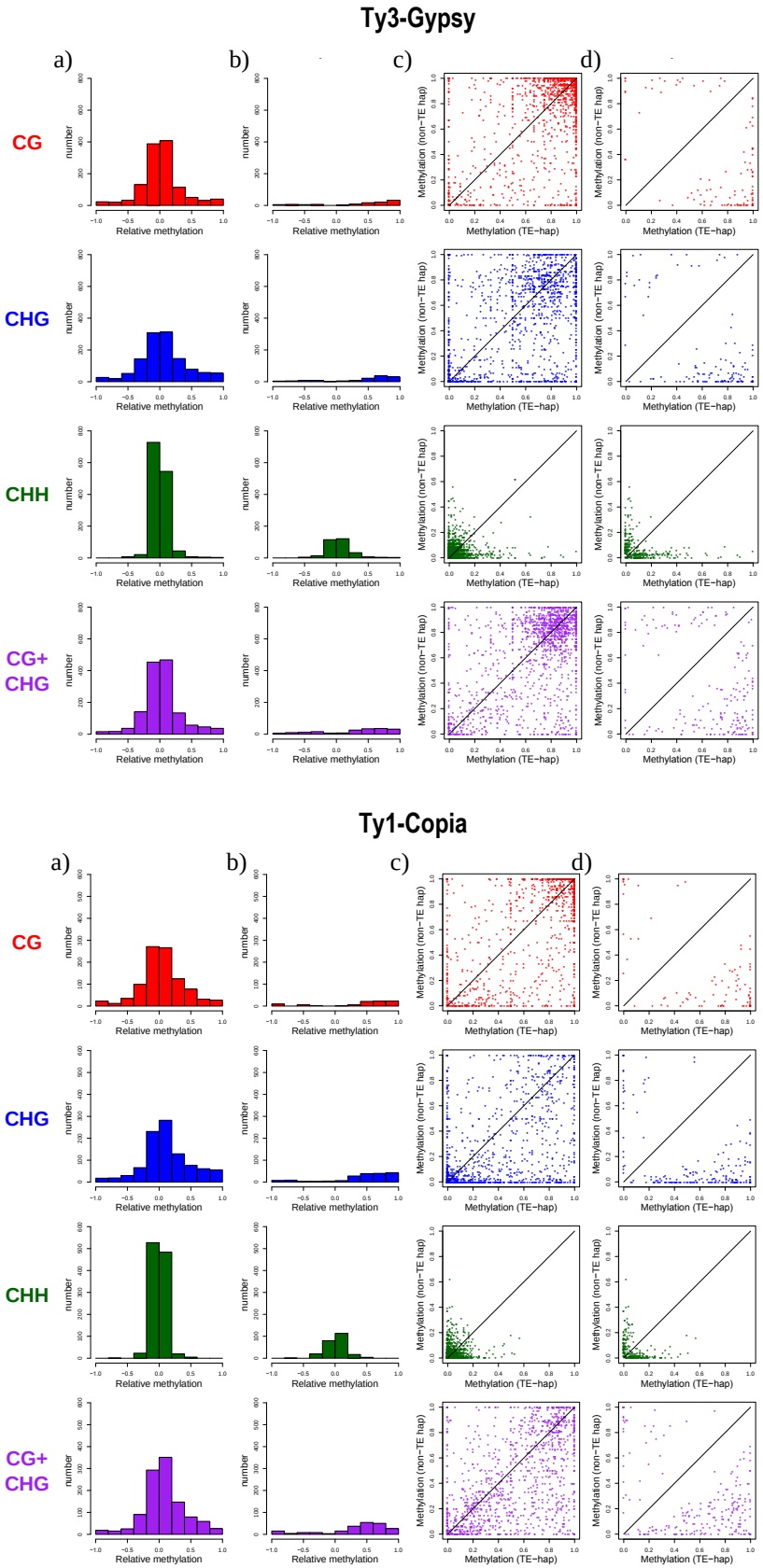
a-b) Average DNA methylation levels (red: CG; blue: CHG) in the TE flanking regions: TE-carrying haplotypes (dark colour) and haplotypes devoid of TEs (light colour) are represented separately (see also Figure S3 for details).

c-d) Fraction of total Cs that are more methylated in the TE-haplotype. Values are reported for each 500 bp bin of distance from the insertion site within a +/- 2500 bp range; c: CG context; d: CHG context.

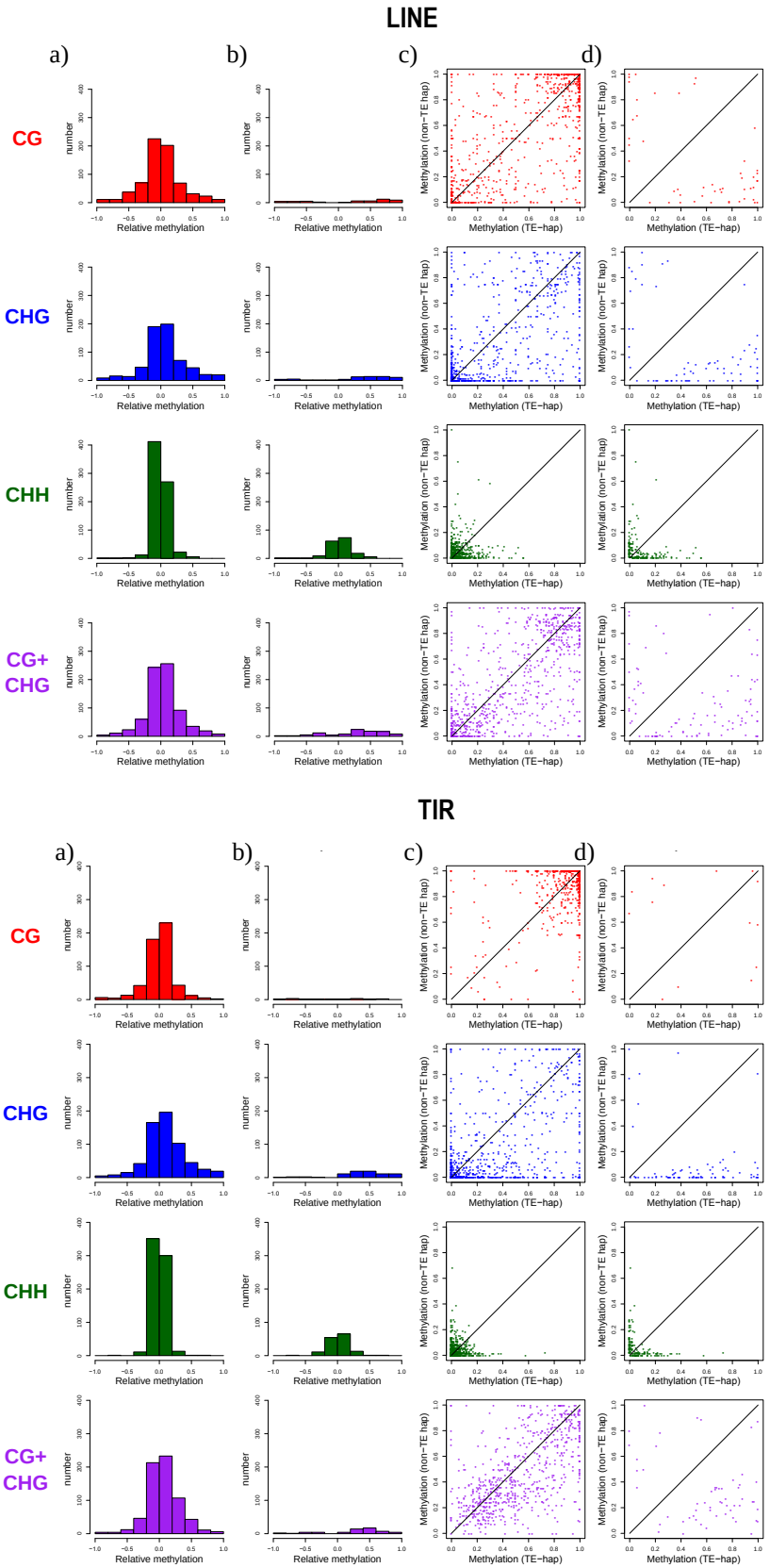
e-f) Fraction of total Cs that are significantly deviating from the null expectation of equal methylation in the two haplotypes (Fisher's Exact Test, p-value < 0.01). Values are reported for each 500 bp bin of distance from the insertion site within a +/- 2500 bp range; e: CG context; f: CHG context.

g-h) Deviation from the null expectation of equal methylation in the two haplotypes. Chi square Test log P values are reported for each 500 bp bin of distance from the insertion site within a +/- 2500 bp range; g: CG context; h: CHG context.

j-k) Deviation from the null expectation of equal methylation in the two haplotypes. Wilcoxon Mann-Whitney Test log P values are reported for each 500 bp bin of distance from the insertion site; g: CG context; h: CHG context.



Figures S6–S7 | Individual TE flanking regions analyses. Methylation is calculated over a region of 2kb around the insertion point in both haplotypes. a) Distribution of the difference of methylation values between TE haplotype and non-TE haplotype; b) Subset of Wilcoxon Mann-Whitney tests p-value positives (<0.01) of a); c) Dotplot ; d) Subset of Wilcoxon Mann-Whitney tests p-value positives (<0.01) of c)



Figures S8 –S9 | Individual TE flanking regions analyses. Methylation is calculated over a region of 2kb around the insertion point in both haplotypes. a) Distribution of the difference of methylation values between TE haplotype and non-TE haplotype; b) Subset of Wilcoxon Mann-Whitney tests p-value positives (<0.01) of a); c) Dotplot ; d) Subset of Wilcoxon Mann-Whitney tests p-value positives (<0.01) of c)

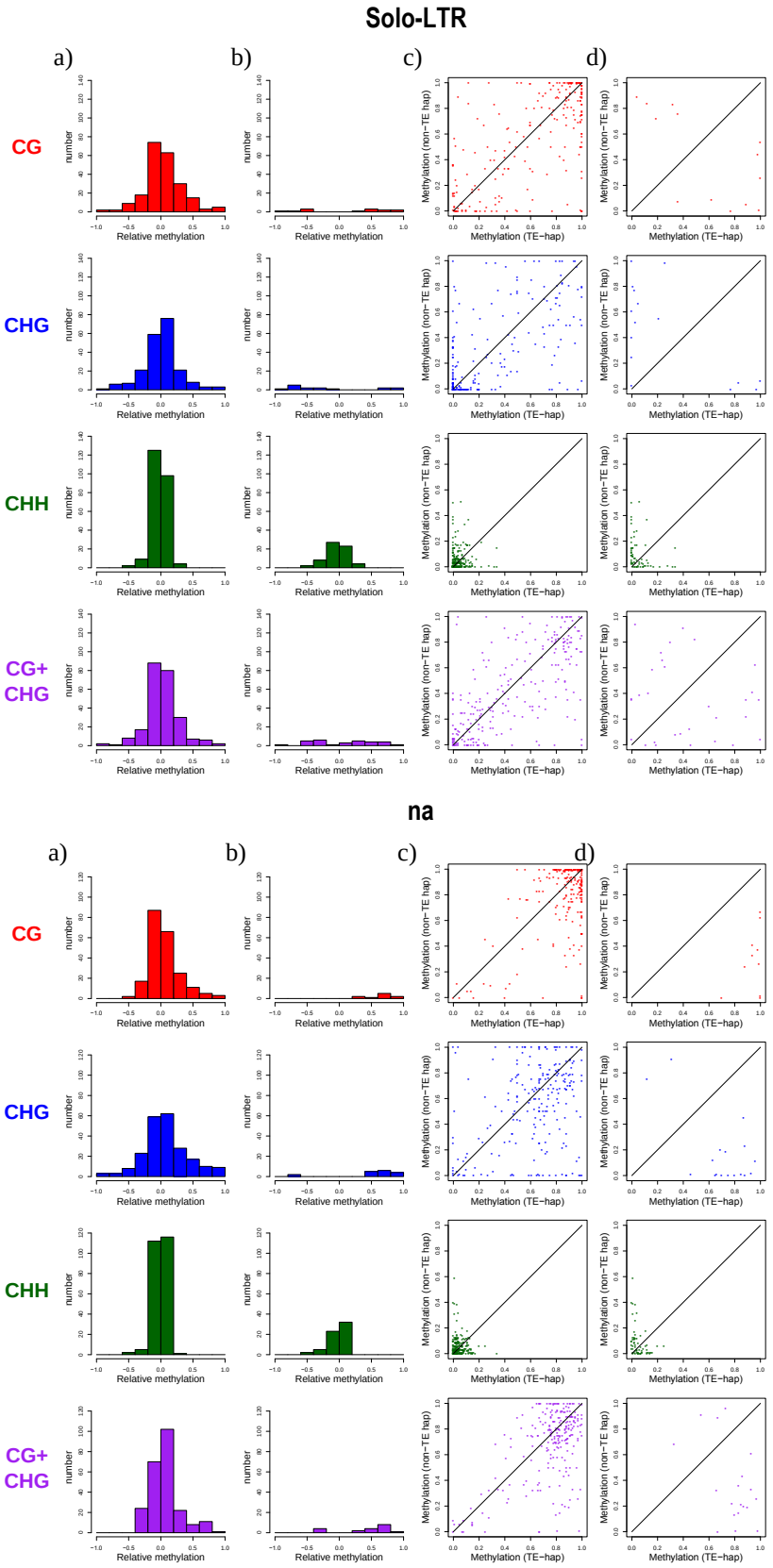


Figure S10 –S11 | Individual TE flanking regions analyses. Methylation is calculated over a region of 2kb around the insertion point in both haplotypes. a) Distribution of the difference of methylation values between TE haplotype and non-TE haplotype; b) Subset of Wilcoxon Mann-Whitney tests p-value positives (<0.01) of a); c) Dotplot ; d) Subset of Wilcoxon Mann-Whitney tests p-value positives (<0.01) of c)

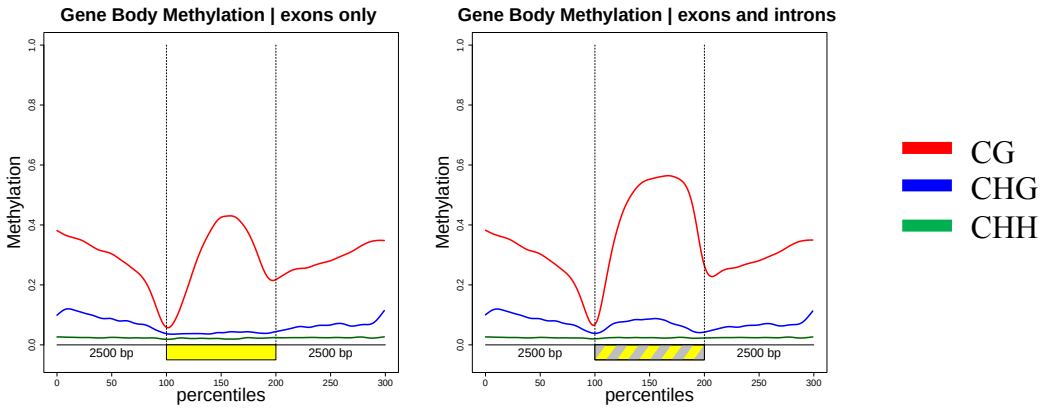


Figure S12 | Gene body methylation in exonic (a) and exonic and intronic (b) sequences in percentiles.

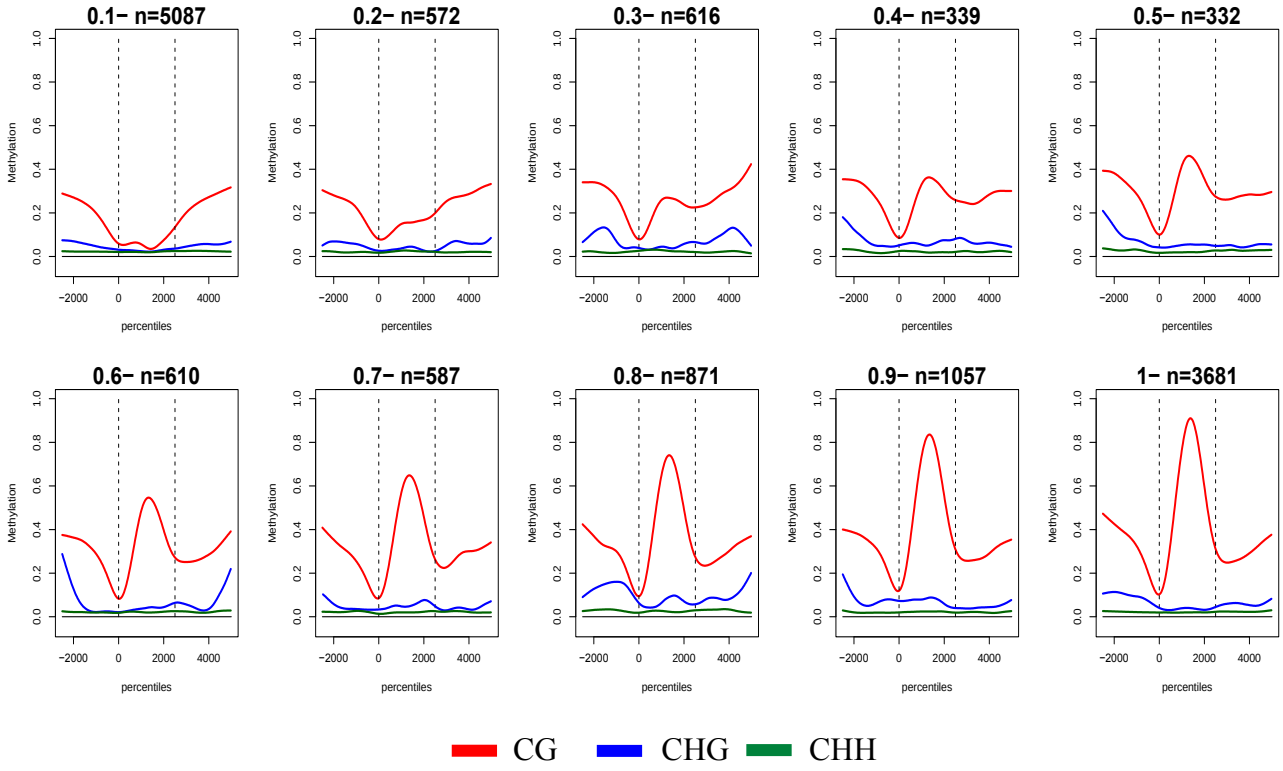


Figure S13 | Gene body methylation classes obtained calculating the average between the 50th and the 65th in the gene body

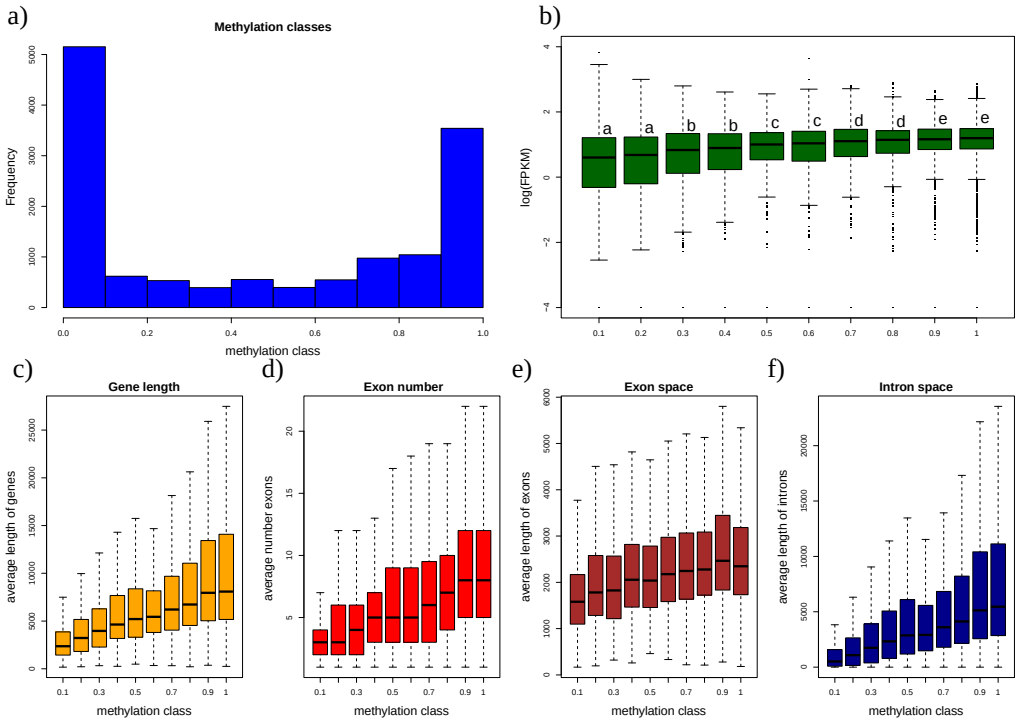


Figure S14 | Frequency of GBM classes (a), Expression rate of gene belonging to GBM classes (b). b) clusters with the same letter code are not significantly different (Wilcoxon Mann-Whitney test (p-value <0.05), Gene length (c), exon number (d),exon space€ and intron space (f) in genes belonging to GBM classes.

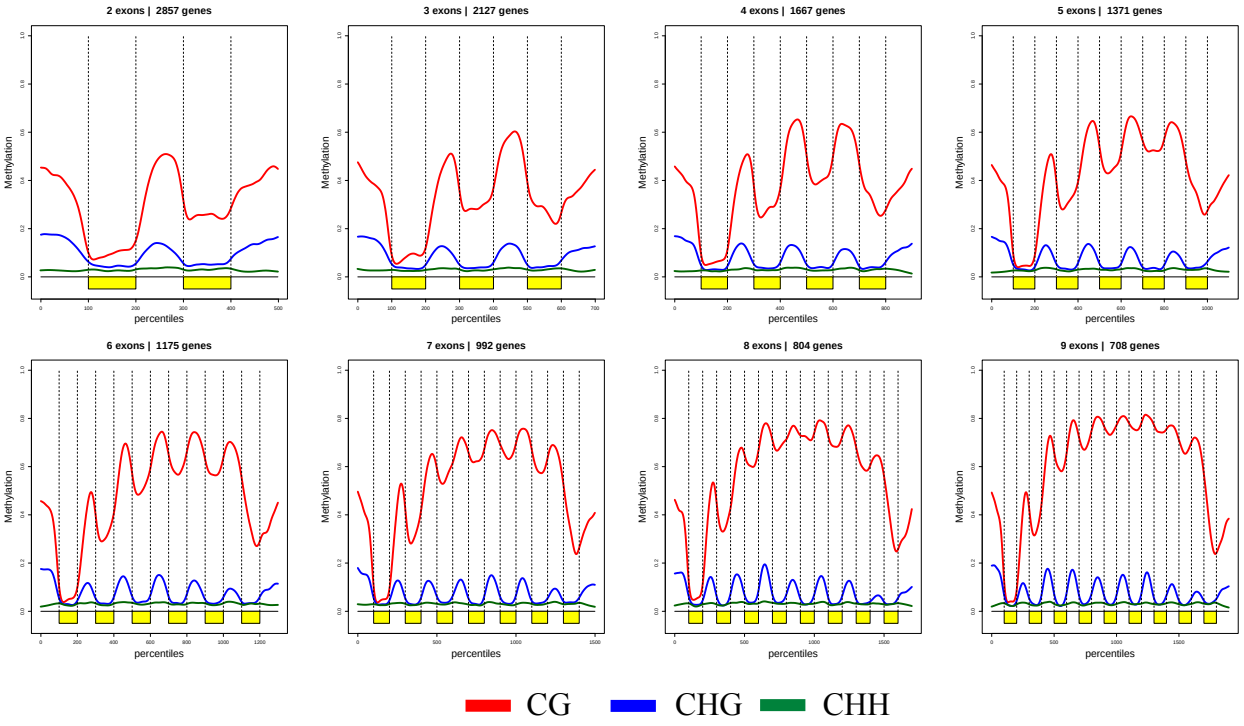


Figure S15 | Gene body Methylation profile of in genes grouped by exon number

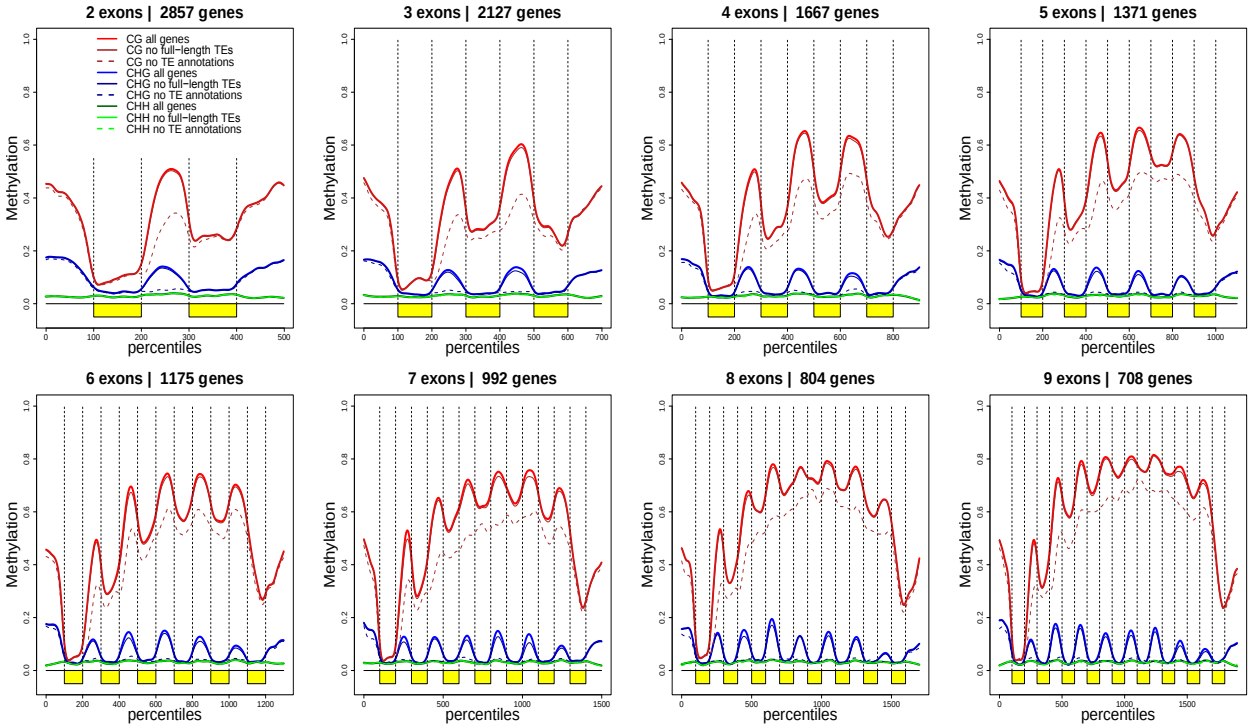


Figure S16| Gene methylation profile of in genes grouped by exon number, including or excluding genes carrying TE annotations

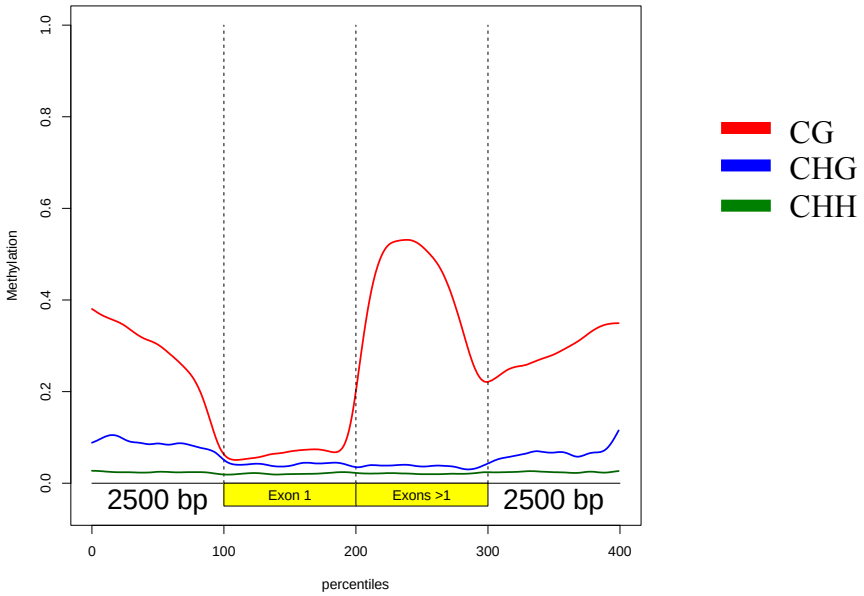


Figure S17 | Gene Body Methylation profile of Exon 1 vs all the other exons

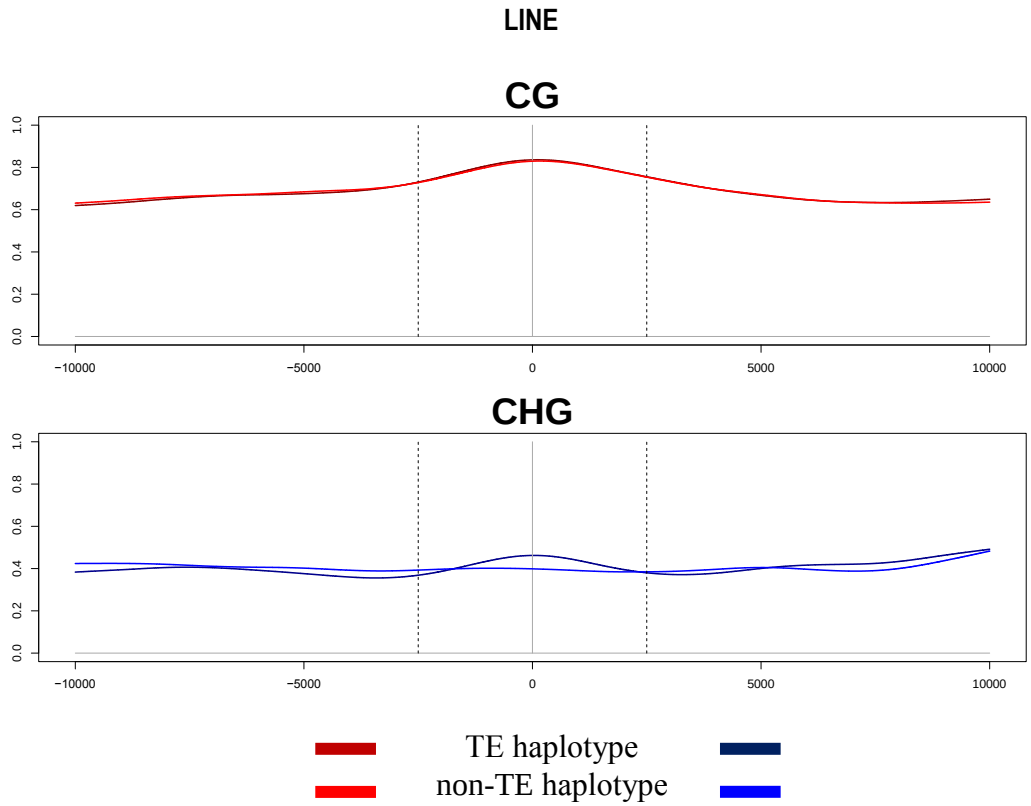


Figure S18 | Hemizygous LINE flanking regions

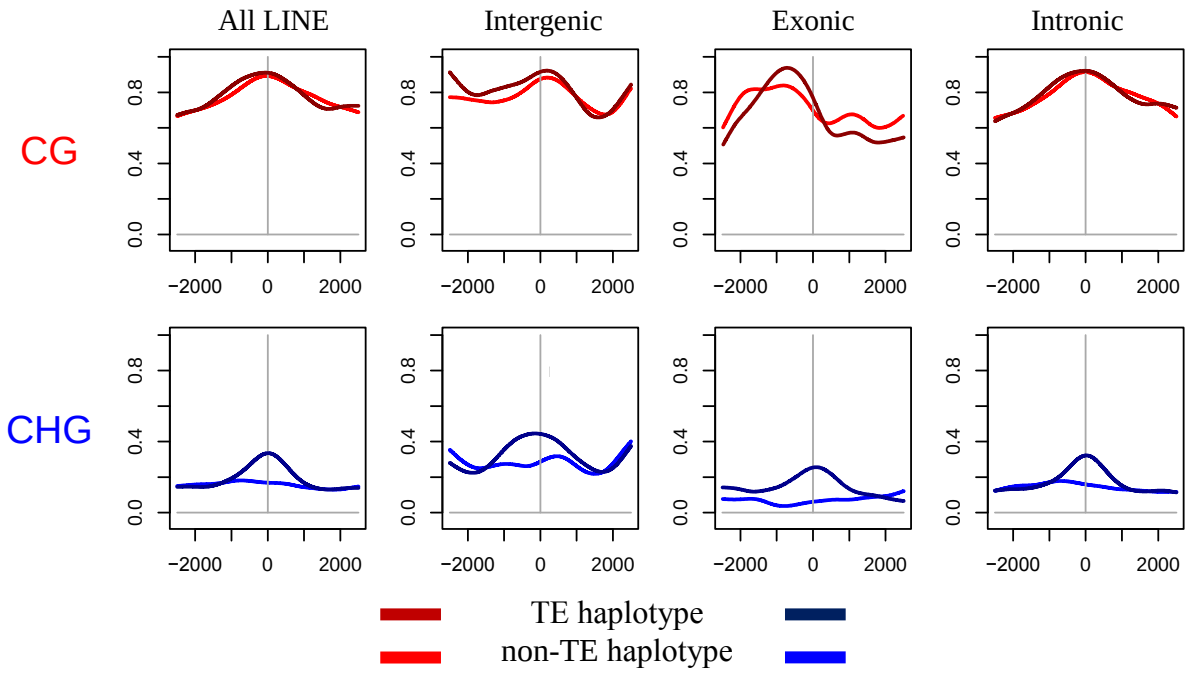


Figure S19 | Hemizygous LINE flanking regions in intergenic, exonic and intronic loci.

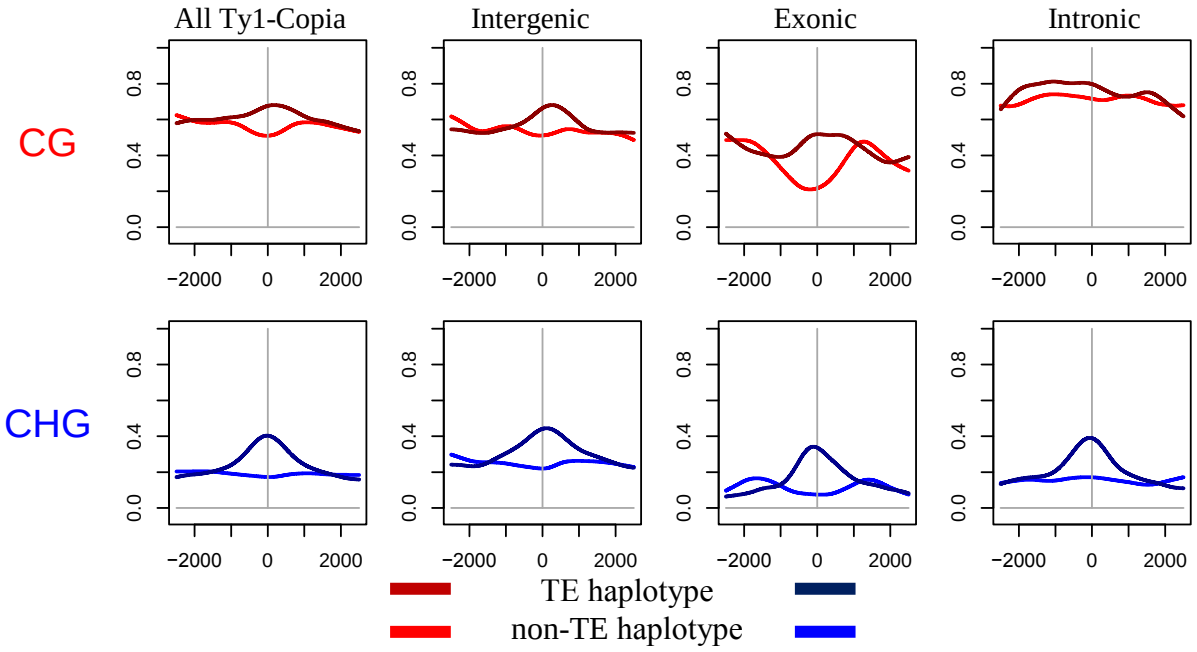


Figure S20 | Hemizygous LINE flanking regions in intergenic, exonic and intronic loci.