A bioinformatics analysis of the arabidopsis thaliana epigenome

Ikhlak Ahmed

2011

PhD Thesis

Ikhlak Ahmed

Laboratoire: CNRS UMR8197 - INSERM U1024 Institut de Biologie de l'ENS(IBENS)

Ecole Doctorale : Sciences Du Végétal

A Bioinformatics analysis of the Arabidopsis thaliana Epigenome

Supervisors: Dr. Chris Bowler and Dr. Vincent Colot

Jury

Prof. Dao-Xiu Zhou	President
Prof. Christian Fankhauser	Rapporteur
Dr. Raphaël Margueron	Rapporteur
Dr. Chris Bowler	Examiner
Dr. Vincent Colot	Examiner
Dr. Allison Mallory	Examiner
Dr. Michaël Weber	Examiner

Acknowledgement

I express my deepest gratitude to my supervisors Dr. Chris Bowler and Dr. Vincent Colot for the support, advice and freedom that I enjoyed working with them. I am very thankful to Chris Bowler who trusted in me and gave me the opportunity to come here and work in the excellent possible conditions. I have no words to express my thankfulness to Vincent Colot for his enthusiastic supervision and the incredibly valuable sessions we spent together that shaped my understanding of DNA methylation.

I am very thankful to my friends Uma Maheswari, Alexis Sarazin, Clara Bourbousse, Stéphanie Drevense who were always there to solve my day-to-day problems. I would also like to thank all the members of the Bowler & Colot groups for their advices, scientific discussions and the moral support they gave me. I also want to thank Sophie de Peindray for her great help.

During this work I have collaborated with many colleagues for whom I have great regard, and I wish to extend my warmest thanks to Frédy Barneche, François Roudier and Hadi Quesneville who have helped me with the work.

I wish to thank my entire extended family for their love and belief in me especially to my sister, Seema Gull, for her constant efforts to keep us all happy and to take care of the family in my absence.

Lastly, and most importantly, I am forever indebted to my parents, AG and papa, for always being there, to care, protect and raise me, their rightful teachings and unconditional love. To them I dedicate this thesis.

Table of Contents

CHAPTER I

INTRODUCTION TO CHROMATIN AND CHROMATIN-LEVEL GENOME ORGANISATION

Introduction

Deoxyribonucleic acid (DNA) is the genetic material that contains the instructions needed for normal development and functioning of almost all known living organisms. However the linear length of naked DNA far exceeds the microscopic dimensions of a cell nucleus. Eukaryotic genomes are therefore compacted into a condensed form called chromatin that is essential to fit the DNA within the confines of a nucleus. Chromatin is composed of DNA plus the proteins (and RNA) that package DNA. The composition and properties of chromatin can vary immensely, e.g. between different types of cells of an organism, during different phases of the cell cycle, or in response to stimuli perceived from the environment.

Besides efficient packing of the genetic material, other major roles of chromatin are to strengthen the DNA and prevent DNA damage, control DNA replication and to provide regulated access of DNA to the transcriptional machinery. Thus while all cells of an organism carry the same DNA sequence, it is the chromatin-mediated selective read out of the genome that distinguishes one cell type from another. Chromatin-mediated control of genome activity can be either transient or else stably transmitted across multiple cell divisions and in some cases even across generations. The study of the transmission of chromatin states without changes of the underlying DNA sequence is a rather new discipline called epigenetics. Chromatin-based epigenetic processes are crucial for development of an adult organism from a fertilised egg and may also be involved in the inheritance of traits. Chromatin has long been recognized to exist in distinct states, starting with the classical cytological definition of euchromatin and heterochromatin. Typically, euchromatin contains transcriptionally active genes whereas heterochromatin is transcriptionally silent, tightly packed and mainly composed

of repeat sequences including transposable elements. However, heterochromatin is sometimes divided into constitutive or facultative forms (Fransz, Soppe et al. 2003). As its name implies, constitutive heterochromatin is observed in most if not all the cells of an organism. This form of heterochromatin mainly occurs around the centromere and near telomeres and its abnormal juxtaposition with genes can result in their transcriptional silencing, in a process called position effect variegation (Reuter and Spierer 1992). Facultative heterochromatin on the other hand does not show any significant enrichment in repetitive sequences and is formed only under specific circumstances or in specific cell-types (Oberdoerffer and Sinclair 2007). A classical example of facultative heterochromatin is the inactive X-chromosome in female mammals.

The fundamental unit of chromatin is the nucleosome (Kornberg and Thomas 1974) (Figure 1.1.1). It consists of a protein octamer, which contains two molecules of each of the core histone proteins (H2A, H2B, H3, H4;(Kornberg and Thomas 1974)), around which 147 bp of DNA is wrapped. In addition a linker histone H1 sits on top of this structure keeping in place the DNA that is wrapped around the nucleosome. Histones are highly basic proteins folded into a C-terminal globular domain and a flexible relatively unstructured N- tail that protrudes from DNA surface in the nucleosome core particle. These N- tails can be subjected to a diverse set of post-translational modifications that modulate their interaction with other chromatin components and hence change the structural and functional properties of chromatin. Histone modifications are generally represented by following a nomenclature system as below:

i. The name of the histone (e.g., H3)

ii. The single-letter amino acid abbreviation (e.g., K for Lysine) followed by the amino acid position in the protein

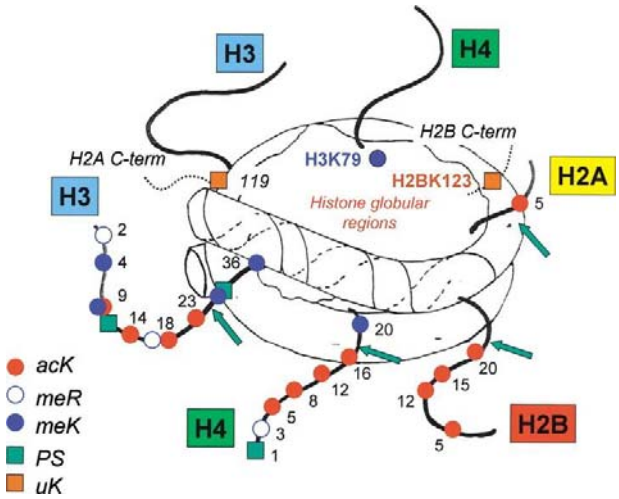
iii. The type of modification (me: methyl, P: phosphate, ac: acetylation, ub: ubiquitin) followed by a numeral that gives the number of groups present.

Using the above nomenclature the di-methylation of histone H3, Lysine 9 would be represented by H3K9me2.

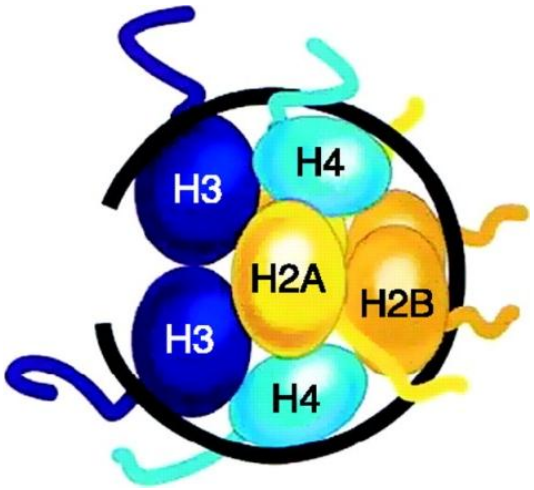
The post-translational modifications of histones most studied are those that lead to gene 'activation' or 'repression', depending on whether the presence of modification correlates with gene activity or silencing. For example, H3K4me3 is associated with gene activation, H3K27me3 is associated with stable repression of genes, while H3K9me2 with long term silencing of mobile repeat elements.

Variation in chromatin structure is also brought about by the incorporation of histone variants. For example, histone H3, is incorporated into chromatin predominantly during replication and is replaced in a replication-independent manner by the histone variant H3.3, which differs from H3 by only four amino acids. H3.3 tends to accumulate in active chromatin (McKittrick, Gafken et al. 2004). Similarly, the histone H2A.Z variant of H2A marks regions around gene promoters (Redon, Pilch et al. 2002; Zhang, Roberts et al. 2005), and CENP-A, which is a more divergent form of histone H3, is located exclusively in centromeric regions, where it plays essential functions including proper segregation of chromosomes (Howman, Fowler et al. 2000; Régnier, Vagnarelli et al. 2005).

Figure 1.1.1: Schematics of nucleosome structure. The nucleosome is the basic unit of chromatin, and consists of a protein octamer containing two molecules each of histones H2A, H2B, H3, and H4, around which 147 base pairs of nuclear DNA is wrapped. Left: the nucleosome core particle is shown with six histone N-termini tails and two C-termini tails. Coloured symbols indicate post-translational modifications and numbers amino acid positions. All DNA-related processes such as transcription, replication and repair must gain access to the DNA, which is mediated by the dynamic nature of the nucleosome structure and regulated by numerous post-translational modifications found at both the N and C-termini of the core histones. Right: Schematic representation of the organization of histones within the nucleosome core particle (DNA in black, N-tails in colour)



(Turner 2002)



(ALLIs CD 2007)

Histone post-translational modifications

A number of histone modifications exist on all the four basic histones that form the nucleosome core and this list is still growing (Bannister and Kouzarides 2011). These modifications can either directly alter the structure of the chromatin by their mere presence on the nucleosomes or can in turn create affinities for other chromatin-associated proteins to cause downstream alterations, e.g. recruitment of Heterochromatin Protein 1 (HP1) by the methylated lysine 9 of histone H3 or Polycomb-group proteins (PcG) by the H3K27me3 to establish a silenced chromatin state. Histone PTMs might also regulate chromatin dynamics by recruiting remodeling enzymes that utilize the energy derived from the hydrolysis of ATP to reposition nucleosomes.

Specific histone modifications are associated with various chromatin dependent processes such as the regulation of gene expression or heterochromatin formation. For example, silent and active genes are typically associated with hypo-acetylated and hyper-acetylated histones respectively. On the other hand, methylation of lysine residues can be associated with either activation or repression depending on the residue, and the degree of methylation or the organism concerned. Thus, H3K9me3 tends to mark active genes in Arabidopsis but is located primarily in heterochromatin in Drosophila and mammals (Berger 2007; Kouzarides 2007; Li, Carey et al. 2007). A number of high-resolution epigenomic maps of histone PTMs have been recently obtained for a range of model organisms, including plants, through a combination of chromatin technologies and genomic tiling microarrays or high-throughput sequencing based approaches. Most studies in plants have focused on the methylation and acetylation of lysine residues on histone H3 (Zhang, Clarenz et al. 2007; Bernatavichute, Zhang et al. 2008; Charron,

He et al. 2009; Zhang, Bernatavichute et al. 2009; Zhou, Wang et al. 2010; Roudier, Ahmed et al. 2011) and have revealed the complex association between different modifications and the underlying chromatin structure. For example, the chromatin domains marked by H3K9me2 are often also marked by the methylation of cytosine residues in DNA (DNA methylation) and this combination is a characteristic feature of repeat and transposable element sequences that form constitutive heterochromatin. H3K4me3, H3K9ac, H3K27ac, and H3K36me3 on the other hand show a strong positive correlation with expression levels and are typically associated with transcriptionally active euchromatic domains (Charron, He et al. 2009; Zhou, Wang et al. 2010). Another histone mark, H3K27me3, is a repressive mark associated with facultative heterochromatin and typically found on silent non-transposable element genes. Polycomb repressive complexe PRC2, a conserved member of the PcG family of proteins is responsible for the deposition of this mark and contributes to chromatin compaction (Margueron and Reinberg 2011) . The marks H3K4me1 and H3K4me2 do not show any significant correlations with gene expression levels and their roles instead depend on the co-occurrence with other chromatic marks such as H3K4me3 or H3K27me3 (Zhang, Clarenz et al. 2007). Thus the chromatin-DNA interactions are often guided by combinations of histone marks that establish and stabilize a given chromatin state. This combinatorial regulation by histone PTMS has been proposed to involve the so-called 'histone code' (Jenuwein and Allis 2001). An overview of known histone modifications along with their associated putative functions is shown in Table 1.1.

Table 1.1: Histone modifications and their proposed functions

Histone	Residue	Type of modification	Proposed Function
H1	S27	Phosphorylation	Transcriptional activation
	K26	Methylation	Transcriptional silencing
H2A	K4 (S.cerevisiae), K5 (mammals), K7 (S.cerevisiae)	Acetylation	Transcriptional activation
	S1,T119 (D.melanogaster)	Phosphorylation	Mitosis
	S122 (S.cerevisiae), S129 (S.cerevisiae), S139 (mammalian H2AX)	Phosphorylation	DNA repair
	K119 (mammals)	Ubiquitination	Transcriptional silencing
	K126 (S.cerevisiae)	Sumoylation	Transcriptional silencing
	K9, K13	Biotinylation	Unknown
H2B	K5, K11(S.cerevisiae), K12 (mammals), K15 (mammals), K16 (S.cerevisiae), K20	Acetylation	Transcriptional activation
	S10 (S.cerevisiae), S14 (Vertebrates)	Phosphorylation	Apoptosis
	S33 (D.melanogaster)	Phosphorylation	Transcriptional activation
	K34 (D.melanogaster), K119 (S. pombe), K120 (mammals), K123 (S.cerevisiae), K143 (Arabidopsis)	Ubiquitination	Transcriptional activation
	K6 or K7 (S. cerevisiae)	Sumoylation	Transcriptional silencing
H3	K4, K9, K14, K18, K23, K27, K56 (S.cerevisiae)	Acetylation	Transcriptional activation
	K4, R17	Methylation	Transcriptional activation
	K36, K79	Methylation	Transcriptional activation (elongation)
	K9, K27, R8	Methylation	Transcriptional silencing
	K9me3 (tri-methylation in Arabidopsis)	Methylation	Transcriptional activation
	T3, S10, T11 (mammals), S28 (mammals)	Phosphorylation	Mitosis
	K4, K9, K18	Biotinylation	Transcriptional activation
H4	K5, K12	Acetylation	Histone deposition
	K8, K16	Acetylation	Transcriptional activation
	K91 (S.cerevisiae)	Acetylation	Chromatin assembly
	R3,	Methylation	Transcriptional activation
	K20	Methylation	Transcriptional silencing
	K59	Methylation	Transcriptional silencing
	S1	Phosphorylation	Mitosis, Chromatin assembly, DNA repair
	K12	Biotinylation	DNA damage response

H1: Linker histone; K: Lysine; R: Arginine; S: Serine; T: Threonine

DNA Methylation

DNA methylation refers to the enzymatic transfer of a methyl group (-CH_3) to the cyclic carbon 5 or nitrogen 4 of cytosines and to the nitrogen 6 of adenines. In eukaryotes, DNA methylation is almost exclusively restricted to the carbons of cytosine residues and acts as a classic epigenetic mark that plays key roles in the control of genome activity. DNA methylation in higher eukaryotes is essential for development and cellular differentiation. However the propagation of this mark dramatically differs between plants and mammals. Thus whereas plants tend to propagate pre-existing DNA methylation states across generations, mammals typically remove this mark during zygote formation and re-establish it through successive cell divisions during development. A recent report, however, indicates that in zygotes hydroxylation of methyl groups occurs rather than the complete removal of methyl groups (Iqbal, Jin et al. 2011).

Methylation of DNA is a hallmark of epigenetic inactivation and heterochromatin in both plants and mammals, typically associated with a silenced chromatin state, and is largely confined to silent repeat elements and transposon sequences. DNA methylation in mammals predominantly occurs on cytosine residues found in a symmetrical sequence context (CG sites) and is estimated to occur at ~70-80% of CG dinucleotides throughout the genome (Ehrlich, Gama-Sosa et al. 1982). Key exceptions to this global methylation of the mammalian genome are CpG islands found as dense clusters of CG dinucleotides near gene promoters. A considerable amount of non-CG methylation is also found in mammalian embryonic stem (ES) cells where one quarter of 5-methyl cytosine was found to occur as CHG or CHH sites (Lister, Pelizzola et al. 2009). In plants, methylation can be observed on any kind of cytosine i.e.,

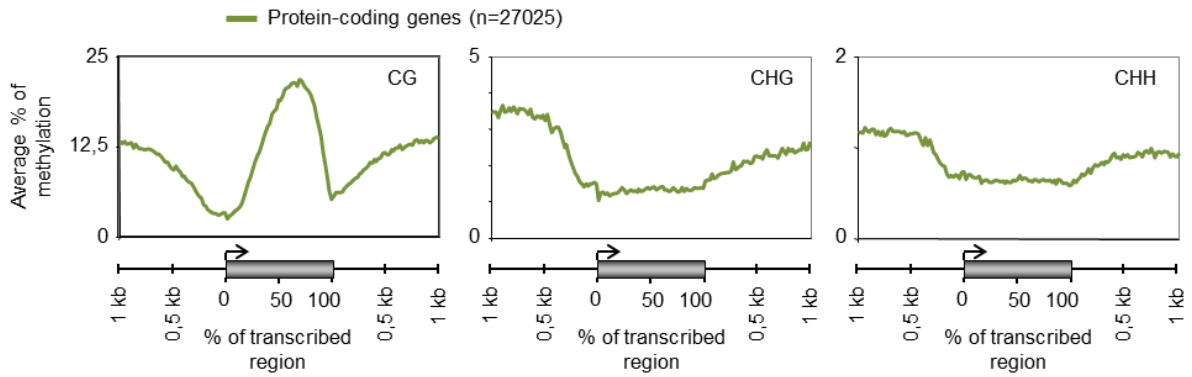
cytosine bases found in a symmetrical sequence context (CG and CHG sites; H is A, T or C) as well as in asymmetrical sequence contexts (CHH sites).

Genomic Context of DNA Methylation

Eukaryotic genomes consist not only of genic sequences but also of transposable elements (TEs), that are capable of moving to new locations and have a potential to increase their number of copies from one generation to another (Orgel and Crick 1980). Genome sequencing has revealed that TEs comprise a large fraction of most eukaryotic genomes, including the human genome (approx. 50% ;(Lander, Linton et al. 2001)) and proliferation of these elements is largely responsible for differences in genome size among eukaryotes (Kidwell 2002). These mobile DNA sequences, often considered as 'selfish' or 'parasitic' elements, are highly mutagenic, and active TEs can disrupt protein coding genes, cause chromosomal breakage, illegitimate recombination or other genome rearrangements. The various classes of transposon sequences employ different mechanisms to proliferate within a genome. For example, Class I elements, termed retrotransposons, use an RNA intermediate for their transposition, Class II or DNA transposons employ a 'cut and paste strategy,' and a third class of transposons called Helitrons are thought to use a 'rolling circle mechanism'. Most TEs of any class are not actively duplicating or transposing mainly due to a mutation or deletion of a part of the TE sequence (non-autonomous elements). However, full-length autonomous copies of TEs are present, and these are typically silent (Slotkin and Martienssen 2007; Lisch 2009). Although TE activity is typically deleterious within the lifespan of an organism, their role over evolutionary time-scales is considered to be a major factor that contributes to shaping the functional genome of an organism.

Genome-wide and other local mapping studies of DNA methylation have long indicated both non-autonomous and full-length TEs to be the main targets of cytosine DNA methylation in eukaryotes. Multiple lines of evidence indicate that eukaryotic DNA methylation serves primarily to keep in check these potentially harmful sequences (Suzuki and Bird 2008) and cytosine methylation is therefore generally seen as a classical silencing epigenetic mark that is associated with the repression of TEs and other repetitive sequences. However, gene-body methylation has been proposed as an ancient property of eukaryotic genomes with preference for exons in most organisms including Arabidopsis thaliana where it is associated with active genes (Zhang, Yazaki et al. 2006; Zilberman, Gehring et al. 2007; Cokus, Feng et al. 2008; Lister, O'Malley et al. 2008; Feng, Jacobsen et al. 2010; Zemach, McDaniel et al. 2010). Unlike in transposons, where cytosine methylation is distributed throughout the length of the TE sequence, methylation within genes occurs predominantly at CG sites only, is confined to the transcriptional part of the gene, and is depleted at both the 5' and 3' ends of coding sequence (Figure 1.2.1). This suggests that methylation at the 5' and 3' ends of genes could be inhibitory to transcription, potentially interfering with initiation or termination. Indeed, methylation of promoter sequences and 5' coding sequences is strongly negatively correlated with the expression of the downstream gene.

Figure 1.2.1: Analysis of CG, CHG and CHH methylation levels in wild type for the entire set of annotated protein coding genes in the Arabidopsis genome (27,025 genes). DNA methylation data are from (Cokus, Feng et al. 2008).



It has also been proposed that moderately transcribed genes are more likely to be methylated than those with low or high expression (Zhang, Yazaki et al. 2006; Zilberman, Gehring et al. 2007), suggesting a scenario where the transcription process itself could contribute to maintaining or enhancing DNA methylation levels over the transcriptional unit. CG methylation in gene bodies of certain mammalian genes has also been observed and is found to be positively correlated with levels of transcription(Jones 1999), and there is now growing evidence that this may be a general phenomenon (Ball, Li et al. 2009). Gene-body methylation has been hypothesized to suppress spurious initiation of transcription within active genes in Arabidopsis and a similar function may exist in mammals (Suzuki and Bird 2008; Feng, Jacobsen et al. 2010; Zemach, McDaniel et al. 2010).

DNA methylation can suppress transcriptional activity or lead to silent chromatin in two ways. First, the methylation of DNA itself may physically impede the binding of transcription factors to

the gene and make the gene inaccessible to the transcription machinery. Second, in a more likely scenario, methylated DNA may be bound by methyl-cytosine binding proteins. These proteins can then recruit additional proteins to the locus, such as histone deacetylases and other chromatin remodelling proteins that can modify histones or remodel the chromatin, thereby forming compact, inactive and silent chromatin.

Establishment and maintenance of DNA methylation

The DNA methyltransferase (DNA MTase) family of enzymes catalyzes the transfer of a methyl group to cytosine bases in DNA and all known DNA MTases use S-adenosyl-L-methionine (SAM or AdoMet) as a methyl group donor. In Arabidopsis, DOMAINS REARRANGED METHYLTRANSFERASE 2 (DRM2), a homologue of the mammalian de novo DNA methyltransferase DNMT3, is primarily responsible for catalyzing de novo DNA methylation (Cao and Jacobsen 2002). DNA METHYLTRANSFERASE 1 (MET1), which is the homologue of mammalian DNMT1, and the plant-specific methyltransferase (CMT3 in Arabidopsis) , are CG-specific and CHG-specific maintenance methyltransferases, respectively. Although, a general perception is that distinct cytosine DNA methyltransferases are responsible for either de novo or maintenance methylation, an emerging view is that different DNA MTases frequently cooperate to catalyse both steps.

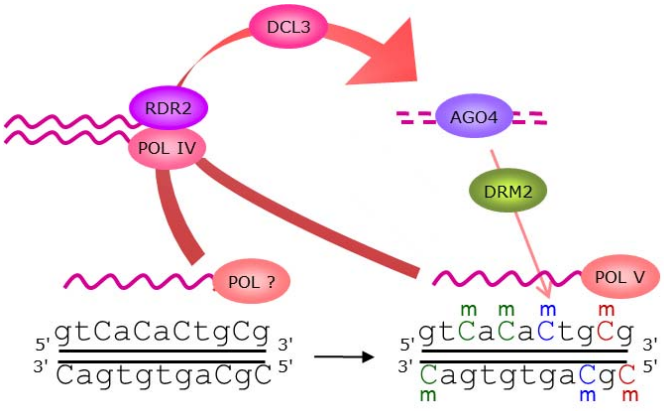
After each round of DNA replication, DNA methylation of the newly synthesized strand is either guided by the parental strand (maintenance) or established de novo through RNA directed DNA methylation (RdDM). RdDM refers to the RNA interference (RNAi) mediated cellular response in plants to the presence of double stranded RNA (dsRNA) in the cell, in which the dsRNA is

processed by the RNase III-like endonuclease, Dicer, to produce small RNAs (siRNAs), which then are loaded on an Argonaut complex to guide DNA MTases to the homologous DNA sequences:. This process induces de novo DNA methylation of cytosine bases in all sequence contexts (CG, CHG, CHH) at the region of siRNA-DNA sequence homology. A prerequisite for siRNA biogenesis is the presence of dsRNA precursors which can form either by bidirectional transcription, transcription through inverted repeats or conversion of transcripts into dsRNA by RNA-dependent RNA polymerases (RDRs). Small RNAs are then incorporated into multiprotein silencing effector complexes to direct either mRNA degradation and repress translation via post transcriptional gene silencing (PTGS), or target DNA methylation and associated repressive chromatin modifications and lead to transcriptional gene silencing (TGS) in a sequence-dependent manner. A vital component of these silencing effector complexes is an argonaute (AGO) protein, which can bind small RNAs through its PAZ domain. Specific members of the argonaute protein family confer functional specificity to different silencing pathways which can be distinguished by either source of dsRNA, size-class of small RNA, or nature of target sequence. Thus, in Arabidopsis whereas the argonaute protein, AGO4 is a member of the RNA-induced transcriptional silencing complex (RITS) involved in TGS and associates with 24-nt siRNAs, AGO1 incorporates into RNA induced silencing complex (RISC) and leads to 21-nt siRNA or miRNA guided cleavage of the target mRNAs (PTGS).

Since cytosines at CG and CHG sites are in a symmetrical sequence context, methylation at these sites has been largely thought to be dependent on maintenance mechanisms (Figure 1.2.2 b; (Teixeira and Colot 2010)). While this holds true for the CG sites, where methylation is mainly under the control of METHYLTRANSFERASE1 in plants (homolog of mammalian

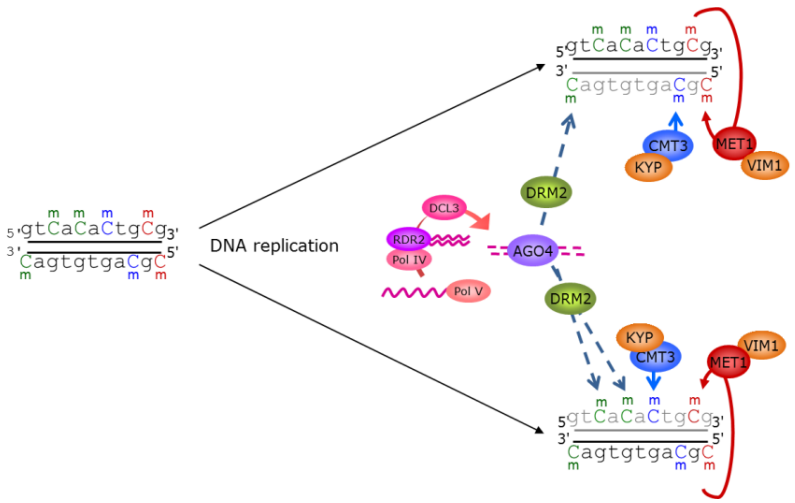
methyltransferase Dnmt1) and Dnmt1 in vitro has a higher affinity for hemi-methylated than for unmethylated CGs (Goll and Bestor 2005). However, the maintenance of methylation at CHG sites does not solely seem to depend on the palindromic symmetry of the sequence, as, methylation maintenance at these sites is mainly carried out by CHROMOMETHYLASE3 (CMT3), a chromodomain containing plant-specific methyl-transferase, and SUVH4, the main histone MTase for histone H3K9 dimethylation (Sharif, Muto et al. 2007; Ooi, O'Donnell et al. 2009). The chromodomain of CMT3 can recognize dimethylated H3K9 and the SRA domain of H3K9 methyl-transferase can bind to methylated CHG sites (Johnson, Bostick et al. 2007).

Figure 1.2.2 a: Proposed model for de novo methylation by the RdDM pathway in Arabidopsis. Primary RNA transcripts are thought to be produced by RNA polymerase Pol IV (or perhaps Pol II) and converted by the RNA-dependent RNA polymerase RDR2 into long dsRNAs. Intra- or intermolecular long dsRNAs could also be produced from inverted repeats or as a result of sense/antisense transcription. The long dsRNAs are then processed by the RNase III enzyme DCL3 into 24-nt siRNAs, which are loaded into a silencing complex containing AGO4. Formation of the siRNA-loaded AGO4 complex, in concert with transcription of the target locus by the RNA polymerase Pol V, would lead to the recruitment of the DNA MTase DRM2 to mediate de novo DNA methylation of the target locus in all sequence contexts. Transcripts produced by either Pol IV or Pol V and corresponding to the methylated region would then be used to amplify the production of siRNAs, creating a reinforcing loop. All cytosines on the two DNA strands are shown as methylated for simplicity. However, not all cytosines within a target sequence are expected to become methylated at once. Black and coloured Cs represent unmethylated and methylated (m) sites, respectively (CG – red; CHG – blue; CHH – green; Taken from Teixeira & Colot 2010).



Teixeira & Colot 2010

Figure 1.2.2 b: Proposed model for maintenance of DNA methylation in Arabidopsis. After DNA replication, the newly synthezised strand (grey) is unmethylated. The SRA- and RING-domain-containing protein VIM1 is thought to recognize hemi-methylated CG sites and help recruit the DNA Mtase MET1 to these sites. Maintenance of CHG methylation is thought to involve a reinforcing loop between the plant-specific DNA MTase CMT3 and the H3K9me2 methyltransferase SUVH4/KYP. CHH methylation is propagated in a locus-specific manner by the constant action of RdDM (represented only by AGO4 and DRM2 for simplicity) and/or by CMT3 and MET1 (not shown). Colours are as in Figure 1.2.2 a. (Figure adopted from Teixeira & Colot 2010)



Teixeira & Colot 2010

Regulation of DNA methylation

Epigenetic regulation implies that marks corresponding to active and inactive chromatin states are not only heritable but also potentially reversible. Cytosine DNA methylation indeed can be lost through passive or active means. Passive loss occurs when methylated cytosine bases fail to maintain their methylated status during multiple rounds of DNA replication. In contrast, active demethylation can occur in non-dividing cells and require the activities of demethylating enzymes known as DNA glycosylases, which function through a base excision repair mechanism, i.e., they remove and replace a methylated cytosine base from the DNA double helix with a normal cytosine. Active DNA demethylation occurs genome-wide in both mammals and plants during pre- and post-zygotic reproductive stages. In mouse embryos, immediately after fertilization the paternal genome is demethylated by an active mechanism and the maternal genome is demethylated by a passive mechanism that depends on DNA replication. Around the time of implantation both are remethylated to different extents in the embryonic and extraembryonic lineages (Reik, Dean et al. 2001; Feng, Cokus et al. 2010). Similarly, in angiosperms like Arabidopsis, genome-wide loss of DNA methylation occurs in the endosperm by simultaneous loss of DNA MTase MET1 and induction of DEMETER (DME), a DNA glycosylase that removes cytosine methylation in all sequence contexts (Gehring, Bubb et al. 2009; Hsieh, Ibarra et al. 2009). Both plants and mammals use DNA methylation for gene imprinting, which is the parent-specific differential expression of alleles of the same gene. Imprinting in mammals is established in the germline cells, passed to the zygote and maintained throughout development of the embryo and adult life and is then erased in primordial germ cells (PGCs) before new imprints are set. In plants imprinting occurs in the endosperm, the structure that nurtures the

developing embryo. Thus, whereas imprinting in mammals is the consequence of differential methylation of specific sequences in gametes (Munshi and Duvvuri 2007), plants such as Arabidopsis implement a more drastic form of imprinting in which the genome-wide loss of methylation in endosperm, mediated by DME, results in a methylation asymmetry between embryo and endosperm (Gehring, Bubb et al. 2009; Hsieh, Ibarra et al. 2009) that also demethylates maternal alleles of imprinted genes and causes endosperm-preferred expression for these genes.

DNA methylation plays an essential role in safeguarding the integrity of the genome, particularly in germ cells, against the activities of parasitic elements such as transposons. A major safeguarding mechanism which ensures the silencing of these elements is via transcriptional (TGS) or posttranscriptional gene silencing (PTGS) mediated by RNAi. TGS at transposons and other repeats in Arabidopsis is strictly enforced via methylation at all three types of sites but largely at CG sites that involves the maintenance activities of MET1 and DDM1, a SWI-SNF chromatin-remodelling factor with intrinsic affinity for transposons and/or repeats. In contrast, genes show a MET1 dependent and DDM1 independent type of methylation (Lippman, Gendrel et al. 2004; Teixeira, Heredia et al. 2009). However, it remains to be seen whether RNAi has any role in gene methylation, though it has been postulated that miRNAs (endogenous ~21-nt small RNAs) may guide DNA methylation of genes in some circumstances like mRNA cleavage of the PHABULOSA gene mediated by miRNA 165 and 166 in *Arabidopsis* and subsequent methylation of the gene downstream of miRNA-DNA sequence homology (Bao, Lye et al. 2004). Nonetheless, most miRNA target genes are not methylated.

TGS directed at transposons and other repeat elements can also epigenetically regulate the expression of neighbouring genes as in the case of the imprinted homeobox gene *FWA (Fujimoto, Kinoshita et al. 2008)*. Thus transposons can play a major role in shaping the functional genome. Transposons, because of their inherent capability to move from one position to another in a genome, are potentially deleterious, and are therefore the main targets of DNA methylation-mediated TGS silencing. Hence, transposons rarely move during the life span of an organism. When a transposon lands closer to a gene sequence, it may affect gene activity by either disrupting the promoter structure in the first place and rendering an active promoter as non-functional, or the new insertion may occur close to the gene regulatory sequences without having a direct impact on promoter activity. However, the subsequent transcription from the new insertion may recruit silencing marks like DNA methylation to the locus and could even additionally spread this methylation via the so-called secondary RdDM (Daxinger, Kanno et al. 2009) or DNA methylation spread (Ahmed, Sarazin et al. 2011) to the promoter regions of neighboring genes.

Methods to identify DNA methylation

Hybridisation-based methods have long been used to characterize methylated regions in DNA. However, these methods cannot be directly used to identify DNA methylation states because the methyl group is located in the major groove of DNA and does not impose a detectable effect on the hydrogen bonding properties of methyl-cytosine (Laird 2010). Therefore, methylation dependent pre-treatments of genomic DNA are employed to reveal the presence or absence of the methyl group at cytosine residues. Currently, three main approaches are used to identify DNA methylation (Lister and Ecker 2009; Laird 2010):

(i) Endonuclease digestion, based on treatment of DNA with methylation-sensitive restriction enzymes followed by hybridization to high-density oligonucleotide arrays. Being dependent on the presence of specific recognition sites for the restriction enzyme, this method can only identify a subset of all methylation sites. However, a mixture of different restriction enzymes is typically used which greatly improves the detection capabilities of this technique. The other major issue with this technique is that precise location of the methyl-cytosines and therefore the context of the methylation is not identifiable.

(ii) Affinity enrichment of methylated regions using antibodies specific for 5mC or using methyl-binding proteins. These techniques rely on the capture of methylated regions by immuno-precipitation of denatured genomic DNA with an antibody specific for methylated cytosine, or methyl-binding proteins, followed by hybridization to either a tiling array {MeDIP-chip} or sequencing {MeDIP-seq}. Affinity-based methods allow for rapid and efficient genome-wide assessment of DNA methylation. However, these techniques are subject to several limitations including low resolution of detection, cross hybridisation, inability to determine individual cytosine context, requirement for a dedicated array, and bias toward CG-rich sequences and low sensitivity for CG-poor regions. Therefore affinity based methods require substantial experimental or bioinformatic adjustment for various kinds of biases which also includes adjustments for the CpG density at different genomic regions.

(iii) Bisulphite conversion, which provides high-resolution detection of DNA methylation, is regarded as the gold-standard methodology for identifying cytosine methylation. Genomic DNA is treated with sodium bisulfite under denaturing conditions which converts cytosines, but not

methyl-cytosines, into uracil via a sulfonation, deamination, desulfonation reaction. Subsequent synthesis of the complementary strand and sequencing allows determination of the methylation status of cytosines on each strand of the genomic DNA simply by observing whether the sequenced base at a cytosine position is a thymine (unmethylated) or a cytosine (methylated).

Thanks to the dramatic advances being made in high-throughput DNA sequencing it is now possible to map the sites of DNA methylation at single-base resolution throughout an entire genome. Autocorrelation analysis has revealed a significant correlation between methylation states of cytosines within 1000 bases (Cokus, Feng et al. 2008). This has prompted the question of whether it is necessary to identify sites of DNA methylation at the single-base resolution and this idea is reinforced by a revised model that points to the possibility of overall methylation levels of a region being important and copied from parent to daughter strands rather than the exact methylation status of each individual cytosine (Jones and Liang 2009). However, numerous studies have demonstrated the critical importance of knowing the methylation status of individual cytosine sites, including a recent report that shows RNA-directed DNA methylation of a single CpG located within a putative conserved element of the Petunia floral homeotic gene pMADS3 that causes ectopic expression of pMADS3 (Shibuya, Fukushima et al. 2009). Further, 5-methylcytosine could play a role like single nucleotide polymorphisms (SNPs), and contribute towards genotype (epigenotype in this case) to phenotype differences. This makes it particularly important to understand and analyse the genome-wide distribution of this mark at single-base resolution.

Bisulphite-based methods tend to be fairly accurate and reproducible. The major sources of bias and measurement error are incomplete bisulphite conversion and differential PCR efficiency for methylated versus unmethylated versions of the same sequence. Bisulphite conversion destroys the self-complementarity of DNA, and therefore, PCR amplification of the bisulfite treated DNA results in four distinct strands (Xi and Li 2009): bisulfite Watson (BSW), bisulfite Crick (BSC), reverse complement of BSW (BSWR), and reverse complement of BSC (BSCR). Furthermore, a T in the bisulfite read could be mapped to either a C or T in the reference sequence but not vice versa. Such asymmetric C/T matching is critical for mapping high-throughput bisulfite reads to the reference genome. Hence aligning millions of bisulfite-treated short reads (BS reads) onto a reference genome remains a challenge and most short read alignment tools, such as BLAT (Kent 2002), SOAP (Li, Li et al. 2008), and Bowtie (Langmead, Trapnell et al. 2009) do not explicitly enable bisulfite mapping. However there are some newly developed software tools like BSMAP (Xi and Li 2009), RMAP (Smith, Chung et al. 2009), MAQ (Li, Ruan et al. 2008) and BS Seeker (Chen, Cokus et al. 2010) that can be used for the alignment of bisulphite-treated short reads to the reference genome. The first whole-genome shotgun bishlphite sequencing library, reported by Cokus et al. for Arabidopsis used an aligner algorithm called CokusAlignment (Cokus, Feng et al. 2008). The CokusAlignment applied several computational filters to the bisulphite-treated short reads including removing sequences that likely mapped to multiple positions and potentially unconverted reads that contained at least three consecutive cytosines in the CHH context. Using reads of length ~30 bases, 2.6 Gb of sequence reads were retained post-filtering, covering; 85% of the 43 million cytosines in the 119 Mb of *Arabidopsis* genome with an average coverage of 20X (Cokus, Feng et al. 2008).

The bioinformatics of epigenome data analysis

The analytical methods used to explore epigenomics data need to be adapted to the approaches used to generate the data in the first place. Chromatin immunoprecipitation followed by hybridisation to tiling microarrays (ChIP-chip) or massively parallel sequencing (ChIP-seq) are currently the most widely used approaches for identification of chromatin modifications. Although microarrays offer lower cost and higher throughput for certain applications, sequencing-based methods are increasingly being used as they offer many advantages, such as increased resolution and higher sampling. A comparison of the ChIP-chip and ChIP-seq technologies is given in Table 1.3.1

High throughput sequencing technologies currently comprise Illumina's Genome Analyzer and HiSeq (http://www.illumina.com/), Roche's 454 (http://www.454.com/), Helicos 'Helioscope and Applied BioSystems' SOLiD (https://products.appliedbiosystems.com) as well as emerging platforms such as Pacific Biosciences (http://www.pacificbiosciences.com/) and Intelligent Bio Systems (http://intelligentbiosystems.com/). These technologies generate millions of short reads that can be aligned to a reference genome. Various software applications have been written that use different algorithmic approaches and flexibility constraints for the efficient mapping of short reads. A summary of some of the widely used tools is given in table 1.3.2.

Table 1.3.1 Comparison of ChIP-chip and ChIP-seq methodologies.

	ChIP-chip	ChIP-seq
coverage	limited by array (genome size, repeats)	whole genome
resolution	25-160 bp (array specific)	1 bp
sample amount	>2 micrograms	10-50 ng
cost	$400-800 per array	$1000 per lane
noise	cross-hybridisation	sequencing errors and biases
data	assumed to be continuous intensities	represented as counts
distribution	log data assumed to be distributed normally	distributional assumptions made for microarray data no longer apply
data analysis	linear models are an established technique	currently available methods assume a negative binomial distribution

Table 1.3.2: Summary of short-read sequence alignment tools

Name	URL	Description
BFAST	http://bfast.sourceforge.net/	Short read mapping, supported by indexing the reference sequences. Can handle billions of short reads and accommodate insertions, deletions, SNPs, and colour errors (can map ABI SOLiD colour space reads). Performs a full Smith Waterman alignment.
Bowtie	http://bowtie-bio.sourceforge.net/	Uses a Burrows-Wheeler transform to create a permanent, reusable index of the genome. Aligns more than 25 million Illumina reads in 1 CPU hour. Supports Maq-like and SOAP-like alignment policies.
BRAT	http://compbio.cs.ucr.edu/brat/	Bisulphite (BS-seq) read mapping
BSMAP	http://code.google.com/p/bsmap/	Bisulphite read mapping

Name	URL	Description
BS Seeker	http://pellegrini.mcdb.ucla.edu/BS_Seeker/BS_Seeker.html	Performs accurate and fast mapping of bisulfite-treated short reads. It aligns Bisulfite-treated reads generated from the Cokus et al's library protocol (with tags) or the Lister et al's library protocol (with no tags)
BWA	http://bio-bwa.sourceforge.net/	Uses a Burrows-Wheeler transform to create an index of the genome. It's a bit slower than bowtie but allows indels in alignment.
CokusAlignment	http://epigenomics.mcdb.ucla.edu/BS-Seq/download.html	Bisulphite read mapping
ELAND	ELAND module within the Illumina Genome Analyzer Pipeline Software(Illumina, Inc., San Diego, CA)	Implemented by Illumina. Includes ungapped alignment with a finite read length.
GMAP and GSNAP	http://research-pub.gene.com/gmap/	Robust, fast, short-read alignment. GMAP: longer reads, with multiple indels and splices; GSNAP: shorter reads, with a single indel or up to two splices per read. Useful for digital gene expression, SNP and indel genotyping.
GNUMAP	http://gemlibrary.sourceforge.net/	Accurately performs gapped alignment of sequence data obtained from next-generation sequencing machines (specifically that of Solexa/Illumina) back to a genome of any size. Includes adaptor trimming, SNP calling and Bisulfite sequence analysis.
MAQ	http://maq.sourceforge.net/	Ungapped alignment that takes into account quality scores for each base.
mrsFAST	http://mrsfast.sourceforge.net/	Short read mapping (supports BS-seq)
Novoalign	http://www.novocraft.com/	Gapped alignment of single end and paired end Illumina GA I & II reads and reads from the new Helicos Heliscope Genome Analyzer. High sensitivity and specificity, using base qualities at all steps in the alignment. Includes adapter trimming, base quality calibration, BS-seq alignment, and option to report multiple alignments per read.
RMAP	http://www.cmb.usc.edu/people/andrewds/rmap/	Short read mapping from 20bp to at most 64bp (supports BS-seq)
SAMTools	http://samtools.sourceforge.net/	Read alignment manipulation, including allele specific expression
SeqMap	http://biogibbs.stanford.edu/~jiangh/SeqMap/	Supports up to 5 mixed substitutions and insertions/deletions. Various tuning options and input/output formats.
SOAP	http://soap.genomics.org.cn/	Robust with a small (1-3) number of gaps and mismatches, uses a 12 letter hash table. SOAP2 is faster than SOAP1

The millions of short reads generated by the sequencing machines are usually presented in a format that associates a quality score to each of the bases. FASTQ is a file format initially developed by the Wellcome Trust Sanger Institute to bundle a FASTA sequence and its quality data. While FASTQ has quickly become the standard for storing short reads, each platform may have native formats that can be converted into FASTQ. For example, Illumina software modules generate data in QSEQ format, which can be directly converted to FASTQ. The ABI SOLiD platform uses a 2 base encoding for their nucleotide sequences and hence uses a “Color Space” FASTA file (CSFASTA).

A FASTQ file normally uses four lines per sequence. Line 1 begins with a '@' character and is followed by a sequence identifier and an optional description (like a FASTA title line). Line 2 is the raw sequence letters. Line 3 begins with a '+' character and is optionally followed by the same sequence identifier (and any description) again. Line 4 encodes the quality values for the sequence in Line 2, and must contain the same number of symbols as letters in the sequence. A minimal FASTQ representation is shown below:

- `@HWI-ST169_0186:4:1:1369:1932#0/1`
- `NCCTAACGACGTTTGGTCAGTTCCATCAACATCATAGCC...`
- `+HWI-ST169_0186:4:1:1369:1932#0/1`
- `BUWUX[WVWVccccc_\ccccccccccacccc_cc^V^^V[[][[^^X^B...`

The quality assurance step processes the FASTQ files by applying some quality checks and clean-up, thereby producing a “clean” FASTQ file per sample. Cleaned-up FASTQ files can be used as input for many of the sequence alignment programs that have been developed to map

short reads to a reference genome. This produces a Sequence Alignment/Map (SAM) file per sample (as well as a file of reads that were not aligned). SAM is a very versatile and near-standardized format for storing many aligned nucleotide sequences. BAM is simply a more compact binary form of the SAM format and the two formats have a back and forth compatibility by using utilities from the SAMtools package (http://samtools.sourceforge.net/). The SAMtools utilities provide all kinds of support to manipulate and operate SAM/BAM files such as sorting, indexing, merging, and building reports of per-base-pair alignment data. SAM files can be sorted or unsorted, but because most tools require the data to be sorted so as to be able to operate within genomic intervals of interest and because sorting doesn't lose any information, it has become a standard practice to work on sorted SAM files.

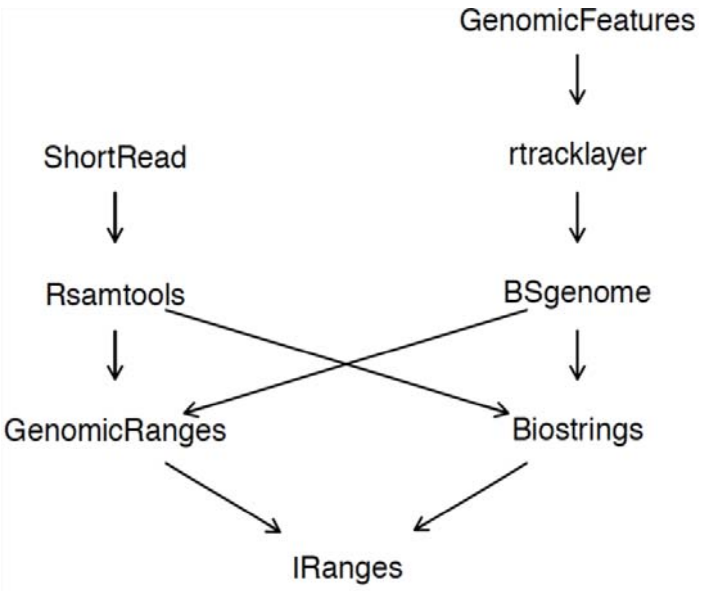
A measure of how many reads cover a given genomic locus is given by the "read depth" parameter. The value simply indicates the height of a pile up of reads that have been assembled or mapped to a given genomic locus. In the case of sequence alignment to a reference genome, higher read depth means more certainty in the "consensus" sequence of the sample and more accuracy in detecting variants from the reference. For *de novo* assembly of a genome, a higher read depth is usually needed, so that large contigs can be formed that are then the building blocks for a draft genome.

R and Bioconductor support

R (http://www.r-project.org/) is a programming language and software environment for statistical computing and graphics. The R language has become a de facto standard among statisticians for developing softwares. Bioconductor (http://www.bioconductor.org/) is a

resource of open software development projects for computational biology and bioinformatics and provides tools for the analysis of high throughput genomic data. Bioconductor uses the R statistical programming language, and is open source and open development. A number of Bioconductor packages have already been developed for analysing next-generation sequencing data and can be used to import fasta, fastq, ELAND, MAQ, BWA, Bowtie, BAM, gff, bed, wig, and other sequence formats, or, trim, transform, align, and manipulate sequences, perform quality assessment for ChIP-seq, RNA-seq, differential expression and other workflows (Figure 1.3.1).

Figure 1.3.1: A selection of Bioconductor packages to analyse high throughput sequencing data. Packages can inherit the functionalities from other packages. Dependencies are shown here with arrows.



The GenomicFeatures package provides a set of tools and methods for making and manipulating transcript centric annotations. With these tools one can easily download the

genomic locations of the transcripts, exons and cds of a given organism, from either the UCSC Genome Browser or a BioMart database. The rtracklayer (Lawrence, Gentleman et al. 2009) package provides an interface between R and genome browsers and lets one manipulate annotation tracks in various formats (currently GFF, BED, bedGraph, BED15, and WIG built-in). This package includes functions that import/export tracks to/from the supported browsers, as well as query and modify the browser state, such as the current viewport. The chipseq (Kharchenko, Tolstorukov et al. 2008) package provides useful tools for design and analysis of ChIP-seq experiments and detection of protein binding positions with high accuracy. These tools include functions that improve tag alignment and correct for background signals. The as-yet-unreleased version of the chipseq package is also aimed at providing convenient interfaces to other powerful packages such as ShortRead and IRanges. The Biostrings package contains memory efficient algorithms for string matching of biological sequences or fast manipulations of larger sets of sequences. The ShortRead package (Morgan, Anders et al. 2009) provides useful tools for analysing high throughput short read sequencing data. These tools include input and output, quality assessment, and downstream analysis functions. The IRanges package provides Infrastructure for manipulating intervals on sequences including functions for representation, manipulation, and analysis of large sequences and subsequences. The package provides efficient low-level and highly reusable classes for storing ranges of integers, Run-Length Encoding vectors, and, more generally, data that can be organized sequentially like Vector objects, as well as views on these Vector objects. The BSgenome package is a container for complete genome sequence of an organism and allows for accessing, analysing, creating, or modifying the data. The biomaRt package (Durinck, Moreau et al. 2005) is an interface to

BioMart databases (e.g., Ensembl, COSMIC, Wormbase and Gramene) and allows users to connect to and search these databases and integrate them with other Bioconductor objects. The Rsamtools package provides an interface to the 'samtools' utilities for manipulating SAM/BAM format files within R environment. The ChIPpeakAnno package (Zhu, Gazin et al. 2010) provides users with facilitation tools for the batch annotation of the peaks identified from either ChIP-chip or ChIP-seq experiments. These tools include functions that find the nearest gene, exon, miRNA or transcription factor binding sites as well as identify Gene Ontology (GO) terms followed by GO enrichment tests. Besides these, there are several packages/tools for visualizing next generation sequencing data, like HilbertVis package(Anders 2009) that provides several functions for visualizing long vectors of integer data by means of Hilbert curves.

Analysis of microarray ChIP-on-chip data

ChIP-on-chip or ChIP-chip combines chromatin immunoprecipitation (ChIP) of DNA fragments associated with a given chromatin modification with hybridization to tiling microarrays in order to verify these DNA fragments. In ChIP-chip, chromatin proteins are covalently cross-linked to the DNA by formaldehyde. The chromatin is then extracted and sheared into fragments typically ~500 bp in length, which sets the limit on the resolution of this technique. The fragmented DNA is fluorescently labelled and hybridized to the tiling microarray, which consists of millions of short (25-70 bp) probes that cover or “tile” the genome at a constant spacing (4 to 100s of nucleotides). The data generated by one experiment consists of an intensity value for each DNA probe. These values measure the relative quantity of DNA at the probe's genomic position in the immunoprecipitated material. In a variant of ChIP-chip, called methyl-DNA

immunoprecipitation (MeDIP), purified DNA is immunoprecipitated with an antibody against methylated cytosine, giving rise to genomic maps of DNA methylation.

In a ChIP-chip or MeDIP experiment, biases can be introduced at various steps that include crosslinking of the DNA to proteins, fragmentation of the chromatin, immunoprecipitation, PCR amplification, and hybridization to the array. In addition, when using two-colour arrays, the "dye-swap" experiments are needed to correct for dye-dependent biases.

Flawed array measurements can be identified by plotting log of the intensity values (log2) of pairs of biological replicates for each probe in an X-Y scatter plot. The data points on these scatter plots should lie scattered close to a straight line. A large difference in the scales of two replicates or a bend in the scatter plot regression curve would indicate problems in the underlying experimental protocol that has led to nonlinearity between the measurements. If the log intensity values for two replicates are distributed approximately along a straight line, a scale normalization (Smyth and Speed 2003) can be used to bring them onto the diagonal. In scale normalization, the log intensity values of each replicate are divided by their median absolute deviations (MAD). The MAD estimates the spread of a distribution, similar to that of a standard deviation, but it is preferable because of its much better robustness with respect to outliers. MAD is the median of the absolute deviations of all measurements from the median. If `median{x1,...,xm}` denotes the median of m measurements x1,..,xm. Then, the absolute deviation of x1 from the median is `|x1-median{x1,..,xm}|`. The MAD of measurements x1,...,xm is defined as

```
MAD = median {|x1-median{x1,..,xm}|,...,|xm-median{x1,..,xm}|}
```

In case of a nonlinear relationship between the replicates, quantile normalization (Bolstad, Irizarry et al. 2003) can be used to map them onto the diagonal and can be performed with several packages available from Bioconductor (e.g. RINGO, vsn, affy). This normalization will make the replicate measurements comparable and allow averaging over them. Outliers in the scatter plot indicate corrupted probes in one of the arrays. Their values should be set to the geometric average of the two neighbouring probes or to NA ("not available"). Many outliers indicate a corrupted array.

Normalisation with the genomic input DNA (called reference) is the most effective way of minimising the strong sequence-dependent probe hybridization biases, chromatin-dependent cross-linking and fragmentation, and many of the other biases. This is achieved by subtracting from each probe the averaged log reference intensity from the averaged log signal intensity. This is equivalent to taking the log of the ratio of the mean signal intensities over the mean reference intensities:

```
log enrichment = arithmetic mean of log (signal) – arithmetic mean of log (reference)
               = log(mean of signal / mean of reference)
```

In addition to the RINGO package that allows to perform many of the normalisation procedures, Bioconductor offers several other packages for the analysis of microarray datasets from different platforms including Affymetrix, Illumina, Nimblegen, Agilent, and other one- and two-color technologies. Major workflows in Bioconductor include pre-processing, quality assessment, differential expression, clustering and classification, gene set enrichment analysis etc.

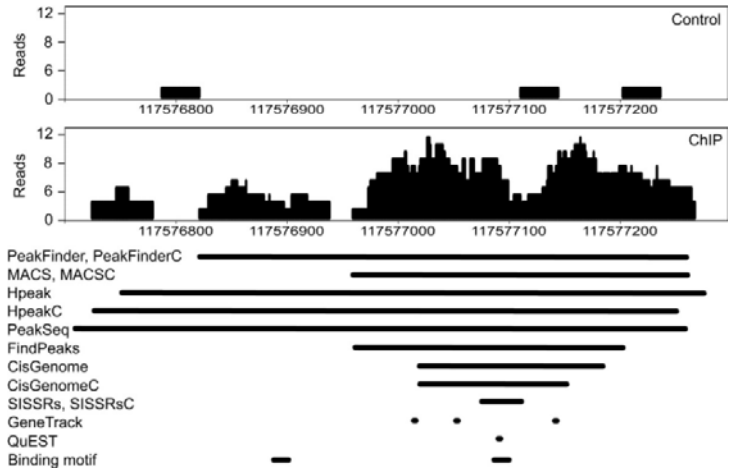
Peak Finding

One of the important goals of high throughput ChIP techniques is to find regions in the genome where more signal or ChIP-seq reads are found than one would expect to see by chance. Such regions are usually characterised by a bell shape centred on a local maximum signal, whose height or depth is significantly greater than the background noise, and defines the DNA regions preferentially associated with the modification or transcription factor under study.

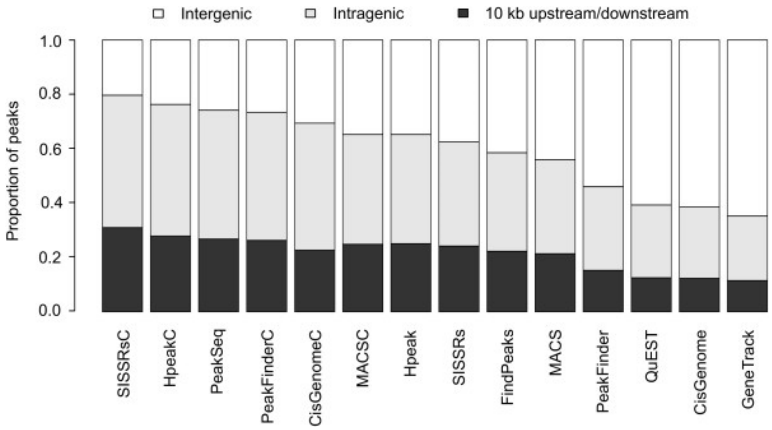
A wealth of algorithmic approaches has been developed for peak finding from high throughput datasets. These can be roughly classified into window-based, overlap-based or hidden Markov models (HMM)-based approaches. In window based methods, one first defines the boundaries of a candidate region and then counts the number of reads or enriched probes within the region. In overlap-based approaches a peak is first identified that corresponds to local maxima of read counts or probe intensity values and then the boundaries of the peak are fitted according to some distribution model. In HMM-based approaches, enrichment along the genome is modelled as a distribution of ChIP-enriched states and background states. As shown in Figure 1.3.2, different mapping strategies may identify mutually exclusive regions as peaks and therefore potential binding regions (Laajala, Raghav et al. 2009; Wilbanks and Facciotti 2010), which could lead to a change in the biological significance of the conclusions.

Figure 1.3.2: A comparison of available peak finding programs. (a) Laajala et al. used 14 peak detection programs to identify binding regions. Same region detected as binding site in all the 14 algorithms is seen to display different peak widths according to the method used. (b) Based on the physical distribution of binding sites detected as peaks, the biological conclusions may change depending upon the algorithm used. For example, in the Laajala et al. study, GeneTrack, QuEST and CisGenome suggested that only less than 40% of the binding sites reside within 10 kb of a gene or are intragenic, whereas with PeakFinderC, SISSRsC, PeakSeq and HpeakC the corresponding estimate was over 70%.

(a)



(b)



Laajala et al. 2009

Furthermore, while the peaks detected for transcription factors tend to be sharp, peaks for histone modifications or RNA Pol II occupancy are typically broader. Thus depending upon the biological study, the choice of algorithm should be changed accordingly. However, the topmost peaks are usually consistent between different algorithms and the difference arises only when looking for marginal peaks. Thus it is always better to use a combination of algorithms and take the union if one is interested in finding all the candidate regions or intersection if only the best of the candidates are to be identified. Further, candidate binding regions can be prioritized using the peak magnitude scores or their p-values, provided by the peak detection programs.

Aims and organisation of the thesis

The main objective of this thesis was to understand the spatial and temporal dynamics of chromatin states in A*rabidopsis* by investigating on a genome-wide scale, patterns of DNA methylation and eleven well-characterized histone post-translational modifications. The results of the analysis is organised into three chapters. The first two result chapters (Chapter II & III) reflect a rather static view of the chromatin and provide details about DNA methylation states and indexing of chromatin by epigenomic marks. Chapter II reports DNA methylation status of transposable element sequences based on a high quality TE annotation provided by our collaborators and publicly available bisulphite single base-resolution DNA methylomes. Chapter III reports results from the analyses of 11 histone modifications along with DNA methylation and provides a combinatorial perspective of epigenoms in establishing a given chromatin state. Chapter IV on the other hand focuses on a specifically chosen histone modification (H2Bub) and uses photomorphogenesis as a biological system to provide a dynamic view of this mark and its effects on the transcriptional state of the underlying chromatin in response to light driven developmental adaptation.

Finally a last Discussion section is provided that summarises the main results and achievements of the thesis.

References:

Ahmed, I., A. Sarazin, et al. (2011). "Genome-wide evidence for local DNA methylation spreading from small RNA-targeted sequences in Arabidopsis." Nucleic Acids Res.

Allemeersch, J., S. Durinck, et al. (2005). "Benchmarking the CATMA Microarray. A Novel Tool forArabidopsis Transcriptome Analysis." Plant Physiology 137(2): 588 -601.

ALLIs CD, J. T. R. D. (2007). Overview and concepts. Epigenetics. New York, Cold Spring Harbour Laboratory Press.

Anders, S. (2009). "Visualization of genomic data with the Hilbert curve." Bioinformatics 25(10): 1231-5.

Ball, M. P., J. B. Li, et al. (2009). "Targeted and genome-scale strategies reveal gene-body methylation signatures in human cells." Nat Biotechnol 27(4): 361-8.

Bannister, A. J. and T. Kouzarides (2011). "Regulation of chromatin by histone modifications." Cell Research 21(3): 381-395.

Bao, N., K. W. Lye, et al. (2004). "MicroRNA binding sites in Arabidopsis class III HD-ZIP mRNAs are required for methylation of the template chromosome." Dev Cell 7(5): 653-62.

Belotserkovskaya, R., S. Oh, et al. (2003). "FACT facilitates transcription-dependent nucleosome alteration." Science (New York, N.Y.) 301(5636): 1090-1093.

Belyayev, A., R. Kalendar, et al. (2010). "Transposable elements in a marginal plant population: temporal fluctuations provide new insights into genome evolution of wild diploid wheat." Mobile DNA 1(1): 6.

Benhamed, M., C. Bertrand, et al. (2006). "Arabidopsis GCN5, HD1, and TAF1/HAF2 interact to regulate histone acetylation required for light-responsive gene expression." The Plant Cell 18(11): 2893-2903.

Benvenuto, G., F. Formiggini, et al. (2002). "The photomorphogenesis regulator DET1 binds the amino-terminal tail of histone H2B in a nucleosome context." Current Biology: CB 12(17): 1529-1534.

Berger, S. L. (2007). "The complex language of chromatin regulation during transcription." Nature 447(7143): 407-412.

Bernatavichute, Y. V., X. Zhang, et al. (2008). "Genome-wide association of histone H3 lysine nine methylation with CHG DNA methylation in Arabidopsis thaliana." PloS One 3(9): e3156.

Betz, J. L., M. Chang, et al. (2002). "Phenotypic analysis of Paf1/RNA polymerase II complex mutations reveals connections to cell cycle regulation, protein synthesis, and lipid and nucleic acid metabolism." Molecular Genetics and Genomics: MGG 268(2): 272-285.

Bolstad, B. M., R. A. Irizarry, et al. (2003). "A comparison of normalization methods for high density oligonucleotide array data based on variance and bias." Bioinformatics 19(2): 185-93.

Bratzel, F., G. López-Torrejón, et al. (2010). "Keeping cell identity in Arabidopsis requires PRC1 RING-finger homologs that catalyze H2A monoubiquitination." Current Biology: CB 20(20): 1853-1859.

Bray, S., H. Musisi, et al. (2005). "Bre1 is required for Notch signaling and histone modification." Developmental Cell 8(2): 279-286.

Cano, F., D. Miranda-Saavedra, et al. (2010). "RNA-binding E3 ubiquitin ligases: novel players in nucleic acid regulation." Biochemical Society Transactions 38(6): 1621-1626.

Cao, X. and S. E. Jacobsen (2002). "Role of the arabidopsis DRM methyltransferases in de novo DNA methylation and gene silencing." Curr Biol 12(13): 1138-44.

Cao, Y., Y. Dai, et al. (2008). "Histone H2B monoubiquitination in the chromatin of FLOWERING LOCUS C regulates flowering time in Arabidopsis." The Plant Cell 20(10): 2586-2602.

Carrozza, M. J., B. Li, et al. (2005). "Histone H3 methylation by Set2 directs deacetylation of coding regions by Rpd3S to suppress spurious intragenic transcription." Cell 123(4): 581-592.

Charron, J.-B. F., H. He, et al. (2009). "Dynamic landscapes of four histone modifications during deetiolation in Arabidopsis." The Plant Cell 21(12): 3732-3748.

Chen, P. Y., S. J. Cokus, et al. (2010). "BS Seeker: precise mapping for bisulfite sequencing." BMC Bioinformatics 11: 203.

Chory, J., C. Peto, et al. (1989). "Arabidopsis thaliana mutant that develops as a light-grown plant in the absence of light." Cell 58(5): 991-999.

Cokus, S. J., S. Feng, et al. (2008). "Shotgun bisulphite sequencing of the Arabidopsis genome reveals DNA methylation patterning." Nature 452(7184): 215-9.

Crevillén, P. and C. Dean (2011). "Regulation of the floral repressor gene FLC: the complexity of transcription in a chromatin context." Current Opinion in Plant Biology 14(1): 38-44.

Culhane, A. C., J. Thioulouse, et al. (2005). "MADE4: an R package for multivariate analysis of gene expression data." Bioinformatics (Oxford, England) 21(11): 2789-2790.

Daxinger, L., T. Kanno, et al. (2009). "A stepwise pathway for biogenesis of 24-nt secondary siRNAs and spreading of DNA methylation." EMBO J 28(1): 48-57.

Deng, X. W. and P. H. Quail (1999). "Signalling in light-controlled development." Seminars in Cell & Developmental Biology 10(2): 121-129.

Durinck, S., Y. Moreau, et al. (2005). "BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis." Bioinformatics 21(16): 3439-40.

Eden, E., R. Navon, et al. (2009). "GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists." BMC Bioinformatics 10(1): 48.

Ehrlich, M., M. A. Gama-Sosa, et al. (1982). "Amount and distribution of 5-methylcytosine in human DNA from different types of tissues of cells." Nucleic Acids Res 10(8): 2709-21.

Ernst, J. and M. Kellis (2010). "Discovery and characterization of chromatin states for systematic annotation of the human genome." Nature Biotechnology 28(8): 817-825.

Feng, S., S. J. Cokus, et al. (2010). "Conservation and divergence of methylation patterning in plants and animals." Proc Natl Acad Sci U S A 107(19): 8689-94.

Feng, S., S. E. Jacobsen, et al. (2010). "Epigenetic reprogramming in plant and animal development." Science (New York, N.Y.) 330(6004): 622-627.

Fleury, D., K. Himanen, et al. (2007). "The Arabidopsis thaliana homolog of yeast BRE1 has a function in cell cycle regulation during early leaf and root growth." The Plant Cell 19(2): 417-432.

Formosa, T. (2008). "FACT and the reorganized nucleosome." Molecular bioSystems 4(11): 1085-1093.

Fransz, P., W. Soppe, et al. (2003). "Heterochromatin in interphase nuclei of Arabidopsis thaliana." Chromosome Research: An International Journal on the Molecular, Supramolecular and Evolutionary Aspects of Chromosome Biology 11(3): 227-240.

Frappier, L. and C. P. Verrijzer (2011). "Gene expression control by protein deubiquitinases." Current Opinion in Genetics & Development 21(2): 207-213.

Fujimoto, R., Y. Kinoshita, et al. (2008). "Evolution and Control of Imprinted FWA Genes in the Genus Arabidopsis." PLoS Genet 4(4): e1000048.

Gabriel, A., J. Dapprich, et al. (2006). "Global Mapping of Transposon Location." PLoS Genet 2(12): e212.

Gehring, M., K. L. Bubb, et al. (2009). "Extensive demethylation of repetitive elements during seed development underlies gene imprinting." Science 324(5933): 1447-51.

Gerstein, M. B., Z. J. Lu, et al. (2010). "Integrative analysis of the Caenorhabditis elegans genome by the modENCODE project." Science (New York, N.Y.) 330(6012): 1775-1787.

Goll, M. G. and T. H. Bestor (2005). "Eukaryotic cytosine methyltransferases." Annu Rev Biochem **74**: 481-514.

Gu, X., D. Jiang, et al. (2009). "Repression of the floral transition via histone H2B monoubiquitination." The Plant Journal: For Cell and Molecular Biology 57(3): 522-533.

He, Y. (2009). "Control of the transition to flowering by chromatin modifications." Molecular Plant 2(4): 554-564.

Henry, K. W., A. Wyce, et al. (2003). "Transcriptional activation via sequential histone H2B ubiquitylation and deubiquitylation, mediated by SAGA-associated Ubp8." Genes & Development 17(21): 2648-2663.

Hon, G., W. Wang, et al. (2009). "Discovery and annotation of functional chromatin signatures in the human genome." PLoS Computational Biology 5(11): e1000566.

Howman, E. V., K. J. Fowler, et al. (2000). "Early disruption of centromeric chromatin organization in centromere protein A (Cenpa) null mice." Proceedings of the National Academy of Sciences of the United States of America 97(3): 1148-1153.

Hsieh, T. F., C. A. Ibarra, et al. (2009). "Genome-wide demethylation of Arabidopsis endosperm." Science 324(5933): 1451-4.

Husband, B. C. (2004). "Chromosomal variation in plant evolution." American Journal of Botany 91(4): 621 -625.

Hwang, W. W., S. Venkatasubrahmanyam, et al. (2003). "A conserved RING finger protein required for histone H2B monoubiquitination and cell size control." Molecular Cell 11(1): 261-266.

Iqbal, K., S. G. Jin, et al. (2011). "Reprogramming of the paternal genome upon fertilization involves genome-wide oxidation of 5-methylcytosine." Proc Natl Acad Sci U S A 108(9): 3642-7.

Jamai, A., A. Puglisi, et al. (2009). "Histone chaperone spt16 promotes redeposition of the original h3-h4 histones evicted by elongating RNA polymerase." Molecular Cell 35(3): 377-383.

Jenuwein, T. and C. D. Allis (2001). "Translating the Histone Code." Science 293(5532): 1074 -1080.

Ji, H., H. Jiang, et al. (2008). "An integrated software system for analyzing ChIP-chip and ChIP-seq data." Nature Biotechnology 26(11): 1293-1300.

Ji, H. and W. H. Wong (2005). "TileMap: create chromosomal map of tiling array hybridizations." Bioinformatics (Oxford, England) 21(18): 3629-3636.

Jiang, D., X. Gu, et al. (2009). "Establishment of the winter-annual growth habit via FRIGIDA-mediated histone methylation at FLOWERING LOCUS C in Arabidopsis." The Plant Cell 21(6): 1733-1746.

Jiang, D., N. C. Kong, et al. (2011). "Arabidopsis COMPASS-like complexes mediate histone H3 lysine-4 trimethylation to control floral transition and plant development." PLoS Genetics 7(3): e1001330.

Jiang, H. and W. H. Wong (2008). "SeqMap: mapping massive amount of oligonucleotides to the genome." Bioinformatics (Oxford, England) 24(20): 2395-2396.

Jiao, Y., L. Ma, et al. (2005). "Conservation and divergence of light-regulated genome expression patterns during seedling development in rice and Arabidopsis." The Plant Cell 17(12): 3239-3256.

Johnson, L. M., M. Bostick, et al. (2007). "The SRA methyl-cytosine-binding domain links DNA and histone methylation." Curr Biol 17(4): 379-84.

Jones, P. A. (1999). "The DNA methylation paradox." Trends Genet 15(1): 34-7.

Jones, P. A. and G. Liang (2009). "Rethinking how DNA methylation patterns are maintained." Nat Rev Genet 10(11): 805-11.

Kao, C.-F., C. Hillyer, et al. (2004). "Rad6 plays a role in transcriptional activation through ubiquitylation of histone H2B." Genes & Development 18(2): 184-195.

Kato, M., K. Takashima, et al. (2004). "Epigenetic control of CACTA transposon mobility in Arabidopsis thaliana." Genetics 168(2): 961-969.

Kent, W. J. (2002). "BLAT--the BLAST-like alignment tool." Genome Res 12(4): 656-64.

Kharchenko, P. V., A. A. Alekseyenko, et al. (2011). "Comprehensive analysis of the chromatin landscape in Drosophila melanogaster." Nature 471(7339): 480-5.

Kharchenko, P. V., M. Y. Tolstorukov, et al. (2008). "Design and analysis of ChIP-seq experiments for DNA-binding proteins." Nat Biotechnol 26(12): 1351-9.

Kidwell, M. G. (2002). "Transposable elements and the evolution of genome size in eukaryotes." Genetica 115(1): 49-63.

Korlach, J., K. P. Bjornson, et al. (2010). "Real-time DNA sequencing from single polymerase molecules." Methods in Enzymology 472: 431-455.

Kornberg, R. D. and J. O. Thomas (1974). "Chromatin structure; oligomers of the histones." Science (New York, N.Y.) 184(139): 865-868.

Kouzarides, T. (2007). "Chromatin modifications and their function." Cell 128(4): 693-705.

Laajala, T. D., S. Raghav, et al. (2009). "A practical comparison of methods for detecting transcription factor binding sites in ChIP-seq experiments." BMC Genomics 10: 618.

Laird, P. W. (2010). "Principles and challenges of genomewide DNA methylation analysis." Nat Rev Genet 11(3): 191-203.

Lander, E. S., L. M. Linton, et al. (2001). "Initial sequencing and analysis of the human genome." Nature 409(6822): 860-921.

Langmead, B., C. Trapnell, et al. (2009). "Ultrafast and memory-efficient alignment of short DNA sequences to the human genome." Genome Biol 10(3): R25.

Lawrence, M., R. Gentleman, et al. (2009). "rtracklayer: an R package for interfacing with genome browsers." Bioinformatics 25(14): 1841-2.

Li, B., M. Carey, et al. (2007). "The role of chromatin during transcription." Cell 128(4): 707-719.

Li, H., J. Ruan, et al. (2008). "Mapping short DNA sequencing reads and calling variants using mapping quality scores." Genome Res 18(11): 1851-8.

Li, R., Y. Li, et al. (2008). "SOAP: short oligonucleotide alignment program." Bioinformatics 24(5): 713-4.

Lippman, Z., A. V. Gendrel, et al. (2004). "Role of transposable elements in heterochromatin and epigenetic control." Nature 430(6998): 471-6.

Lisch, D. (2009). "Epigenetic regulation of transposable elements in plants." Annu Rev Plant Biol 60: 43-66.

Lister, R. and J. R. Ecker (2009). "Finding the fifth base: genome-wide sequencing of cytosine methylation." Genome Res 19(6): 959-66.

Lister, R., R. C. O'Malley, et al. (2008). "Highly integrated single-base resolution maps of the epigenome in Arabidopsis." Cell 133(3): 523-36.

Lister, R., M. Pelizzola, et al. (2009). "Human DNA methylomes at base resolution show widespread epigenomic differences." Nature 462(7271): 315-22.

Liu, C. L., T. Kaplan, et al. (2005). "Single-nucleosome mapping of histone modifications in S. cerevisiae." PLoS Biology 3(10): e328.

Liu, T., A. Rechtsteiner, et al. (2011). "Broad chromosomal domains of histone modification patterns in C. elegans." Genome Res 21(2): 227-36.

Liu, Y., M. Koornneef, et al. (2007). "The absence of histone H2B monoubiquitination in the Arabidopsis hub1 (rdo4) mutant reveals a role for chromatin remodeling in seed dormancy." The Plant Cell 19(2): 433-444.

Lolas, I. B., K. Himanen, et al. (2010). "The transcript elongation factor FACT affects Arabidopsis vegetative and reproductive development and genetically interacts with HUB1/2." The Plant Journal: For Cell and Molecular Biology 61(4): 686-697.

Loudet, O., T. P. Michael, et al. (2008). "A zinc knuckle protein that negatively controls morning-specific growth in Arabidopsis thaliana." Proceedings of the National Academy of Sciences of the United States of America 105(44): 17193-17198.

Luo, C. and E. Lam (2010). "ANCORP: a high-resolution approach that generates distinct chromatin state models from multiple genome-wide datasets." The Plant Journal: For Cell and Molecular Biology 63(2): 339-351.

Ma, L., J. Li, et al. (2001). "Light control of Arabidopsis development entails coordinated regulation of genome expression and cellular pathways." The Plant Cell 13(12): 2589-2607.

Margueron, R. and D. Reinberg (2011). "The Polycomb complex PRC2 and its mark in life." Nature 469(7330): 343-9.

Martin-Magniette, M.-L., T. Mary-Huard, et al. (2008). "ChIPmix: mixture model of regressions for two-color ChIP-chip analysis." Bioinformatics (Oxford, England) 24(16): i181-186.

Mason, P. B. and K. Struhl (2003). "The FACT complex travels with elongating RNA polymerase II and is important for the fidelity of transcriptional initiation in vivo." Molecular and Cellular Biology 23(22): 8323-8333.

McKittrick, E., P. R. Gafken, et al. (2004). "Histone H3.3 is enriched in covalent modifications associated with active chromatin." Proceedings of the National Academy of Sciences of the United States of America 101(6): 1525-1530.

Mirouze, M., J. Reinders, et al. (2009). "Selective epigenetic control of retrotransposition in Arabidopsis." Nature 461(7262): 427-430.

Miura, A., S. Yonebayashi, et al. (2001). "Mobilization of transposons by a mutation abolishing full DNA methylation in Arabidopsis." Nature 411(6834): 212-214.

Morgan, M., S. Anders, et al. (2009). "ShortRead: a bioconductor package for input, quality assessment and exploration of high-throughput sequence data." Bioinformatics 25(19): 2607-8.

Munshi, A. and S. Duvvuri (2007). "Genomic imprinting - the story of the other half and the conflicts of silencing." J Genet Genomics 34(2): 93-103.

Neff, M. M., C. Fankhauser, et al. (2000). "Light: an indicator of time and place." Genes & Development 14(3): 257-271.

Ng, H. H., F. Robert, et al. (2003). "Targeted recruitment of Set1 histone methylase by elongating Pol II provides a localized mark and memory of recent transcriptional activity." Molecular Cell 11(3): 709-719.

Oberdoerffer, P. and D. A. Sinclair (2007). "The role of nuclear architecture in genomic instability and ageing." Nat Rev Mol Cell Biol 8(9): 692-702.

Oh, S., S. Park, et al. (2008). "Genic and global functions for Paf1C in chromatin modification and gene expression in Arabidopsis." PLoS Genetics 4(8): e1000077.

Ooi, S. K., A. H. O'Donnell, et al. (2009). "Mammalian cytosine methylation at a glance." J Cell Sci 122(Pt 16): 2787-91.

Orgel, L. E. and F. H. Crick (1980). "Selfish DNA: the ultimate parasite." Nature 284(5757): 604-7.

Pavri, R., B. Zhu, et al. (2006). "Histone H2B monoubiquitination functions cooperatively with FACT to regulate elongation by RNA polymerase II." Cell 125(4): 703-717.

Pickart, C. M. (2001). "Mechanisms underlying ubiquitination." Annual Review of Biochemistry 70: 503-533.

Pirngruber, J., A. Shchebet, et al. (2009). "CDK9 directs H2B monoubiquitination and controls replication-dependent histone mRNA 3'-end processing." EMBO Reports 10(8): 894-900.

Pushkarev, D., N. F. Neff, et al. (2009). "Single-molecule sequencing of an individual human genome." Nature Biotechnology 27(9): 847-850.

Redon, C., D. Pilch, et al. (2002). "Histone H2A variants H2AX and H2AZ." Current Opinion in Genetics & Development 12(2): 162-169.

Régnier, V., P. Vagnarelli, et al. (2005). "CENP-A is required for accurate chromosome segregation and sustained kinetochore association of BubR1." Molecular and Cellular Biology 25(10): 3967-3981.

Reik, W., W. Dean, et al. (2001). "Epigenetic reprogramming in mammalian development." Science 293(5532): 1089-93.

Reinberg, D. and R. J. Sims, 3rd (2006). "de FACTo nucleosome dynamics." The Journal of Biological Chemistry 281(33): 23297-23301.

Reuter, G. and P. Spierer (1992). "Position effect variegation and chromatin proteins." BioEssays: News and Reviews in Molecular, Cellular and Developmental Biology 14(9): 605-612.

Riddle, N. C., A. Minoda, et al. (2011). "Plasticity in patterns of histone modifications and chromosomal proteins in Drosophila heterochromatin." Genome Research 21(2): 147-163.

Robzyk, K., J. Recht, et al. (2000). "Rad6-dependent ubiquitination of histone H2B in yeast." Science (New York, N.Y.) 287(5452): 501-504.

Roudier, F., I. Ahmed, et al. (2011). "Integrative epigenomic mapping defines four main chromatin states in Arabidopsis." The EMBO Journal 30(10): 1928-1938.

Roudier, F., I. Ahmed, et al. (2011). "Integrative epigenomic mapping defines four main chromatin states in Arabidopsis." EMBO J 30(10): 1928-38.

Roy, S., J. Ernst, et al. (2010). "Identification of functional elements and regulatory circuits by Drosophila modENCODE." Science (New York, N.Y.) 330(6012): 1787-1797.

Schmitz, R. J., Y. Tamada, et al. (2009). "Histone H2B deubiquitination is required for transcriptional activation of FLOWERING LOCUS C and for proper control of flowering in Arabidopsis." Plant Physiology 149(2): 1196-1204.

Schroeder, D. F., M. Gahrtz, et al. (2002). "De-etiolated 1 and damaged DNA binding protein 1 interact to regulate Arabidopsis photomorphogenesis." Current Biology: CB 12(17): 1462-1472.

Sharif, J., M. Muto, et al. (2007). "The SRA protein Np95 mediates epigenetic inheritance by recruiting Dnmt1 to methylated DNA." Nature 450(7171): 908-12.

Shibuya, K., S. Fukushima, et al. (2009). "RNA-directed DNA methylation induces transcriptional activation in plants." Proc Natl Acad Sci U S A 106(5): 1660-5.

Slotkin, R. K. and R. Martienssen (2007). "Transposable elements and the epigenetic regulation of the genome." Nat Rev Genet 8(4): 272-85.

Smith, A. D., W. Y. Chung, et al. (2009). "Updates to the RMAP short-read mapping software." Bioinformatics 25(21): 2841-2.

Smyth, G. K. (2005). Limma: Linear Models for microarray data. New York, Springer.

Smyth, G. K. and T. Speed (2003). "Normalization of cDNA microarray data." Methods 31(4): 265-73.

Sridhar, V. V., A. Kapoor, et al. (2007). "Control of DNA methylation and heterochromatic silencing by histone H2B deubiquitination." Nature 447(7145): 735-738.

Suzuki, M. M. and A. Bird (2008). "DNA methylation landscapes: provocative insights from epigenomics." Nat Rev Genet 9(6): 465-76.

Tamayo, P., D. Slonim, et al. (1999). "Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation." Proceedings of the National Academy of Sciences of the United States of America 96(6): 2907-2912.

Teixeira, F. K. and V. Colot (2010). "Repeat elements and the Arabidopsis DNA methylation landscape." Heredity 105(1): 14-23.

Teixeira, F. K., F. Heredia, et al. (2009). "A role for RNAi in the selective correction of DNA methylation defects." Science 323(5921): 1600-4.

Turck, F., F. o. Roudier, et al. (2007). "Arabidopsis TFL2/LHP1 specifically associates with genes marked by trimethylation of histone H3 lysine 27." PLoS Genetics 3(6): e86.

Turner, B. M. (2002). "Cellular Memory and the Histone Code." Cell 111(3): 285-291.

Wang, Z., C. Zang, et al. (2008). "Combinatorial patterns of histone acetylations and methylations in the human genome." Nat Genet 40(7): 897-903.

Wessler, S. R. (1996). "Turned on by stress. Plant retrotransposons." Current Biology: CB 6(8): 959-961.

Wheelan, S. J., L. Z. Scheifele, et al. (2006). "Transposon insertion site profiling chip (TIP-chip)." Proceedings of the National Academy of Sciences 103(47): 17632 -17637.

Wilbanks, E. G. and M. T. Facciotti (2010). "Evaluation of algorithm performance in ChIP-seq peak detection." PLoS One 5(7): e11471.

Wood, A., N. J. Krogan, et al. (2003). "Bre1, an E3 ubiquitin ligase required for recruitment and substrate selection of Rad6 at a promoter." Molecular Cell 11(1): 267-274.

Xi, Y. and W. Li (2009). "BSMAP: whole genome bisulfite sequence MAPping program." BMC Bioinformatics 10: 232.

Xiao, T., C.-F. Kao, et al. (2005). "Histone H2B ubiquitylation is associated with elongating RNA polymerase II." Molecular and Cellular Biology 25(2): 637-651.

Xin, H., S. Takahata, et al. (2009). "yFACT induces global accessibility of nucleosomal DNA without H2A-H2B displacement." Molecular Cell 35(3): 365-376.

Xu, L., R. Ménard, et al. (2009). "The E2 ubiquitin-conjugating enzymes, AtUBC1 and AtUBC2, play redundant roles and are involved in activation of FLC expression and repression of flowering in Arabidopsis thaliana." The Plant Journal: For Cell and Molecular Biology 57(2): 279-288.

Zemach, A., I. E. McDaniel, et al. (2010). "Genome-wide evolutionary analysis of eukaryotic DNA methylation." Science (New York, N.Y.) 328(5980): 916-919.

Zhang, H., D. N. Roberts, et al. (2005). "Genome-wide dynamics of Htz1, a histone H2A variant that poises repressed/basal promoters for activation through histone loss." Cell 123(2): 219-231.

Zhang, X., Y. V. Bernatavichute, et al. (2009). "Genome-wide analysis of mono-, di- and trimethylation of histone H3 lysine 4 in Arabidopsis thaliana." Genome Biology 10(6): R62.

Zhang, X., O. Clarenz, et al. (2007). "Whole-genome analysis of histone H3 lysine 27 trimethylation in Arabidopsis." PLoS Biol 5(5): e129.

Zhang, X., O. Clarenz, et al. (2007). "Whole-genome analysis of histone H3 lysine 27 trimethylation in Arabidopsis." PLoS Biology 5(5): e129.

Zhang, X., J. Yazaki, et al. (2006). "Genome-wide high-resolution mapping and functional analysis of DNA methylation in arabidopsis." Cell 126(6): 1189-201.

Zhou, J., X. Wang, et al. (2010). "Genome-wide profiling of histone H3 lysine 9 acetylation and dimethylation in Arabidopsis reveals correlation between multiple histone marks and gene expression." Plant Molecular Biology 72(6): 585-595.

Zhou, V. W., A. Goren, et al. (2011). "Charting histone modifications and the functional organization of mammalian genomes." Nat Rev Genet 12(1): 7-18.

Zhu, B., Y. Zheng, et al. (2005). "Monoubiquitination of human histone H2B: the factors involved and their roles in HOX gene regulation." Molecular Cell 20(4): 601-611.

Zhu, L. J., C. Gazin, et al. (2010). "ChIPpeakAnno: a Bioconductor package to annotate ChIP-seq and ChIP-chip data." BMC Bioinformatics 11: 237.

Zilberman, D., M. Gehring, et al. (2007). "Genome-wide analysis of Arabidopsis thaliana DNA methylation uncovers an interdependence between methylation and transcription." Nat Genet 39(1): 61-9.

Zilberman, D., M. Gehring, et al. (2007). "Genome-wide analysis of Arabidopsis thaliana DNA methylation uncovers an interdependence between methylation and transcription." Nature Genetics 39(1): 61-69.

CHAPTER II

GENOME-WIDE EVIDENCE FOR LOCAL DNA METHYLATION SPREADING FROM SMALL RNA TARGETED SEQUENCES IN ARABIDOPSIS

Introduction

Unlike mammals, where cytosine bases are predominantly methylated in a symmetrical CG sequence context, methyl-cytosines in plants can occur at CG, CHG, and CHH sites, where H represents any nucleotide but guanine. Genome-wide analysis of DNA methylation in Arabidopsis using either immunoprecipitation of methylated DNA followed by hybridization to tiling microarrays ((MeDIP) (Zhang, Yazaki et al. 2006; Zilberman, Gehring et al. 2007) or single base-resolution methylomes (Cokus, Feng et al. 2008; Lister, O'Malley et al. 2008) have shown that most DNA methylation aligns with repeat and TE sequences. These studies further showed that TEs are typically methylated at CG, CHG and CHH sites, but display an enrichment of non-CG methylation considered to be a signature of RdDM. However these reports described only the general trends in CG, CHG and CHH methylation without investigating individual TE sequence. The study of DNA methylation reported here is based on publicly available bisulphite single base-resolution DNA methylomes (Cokus, Feng et al. 2008; Lister, O'Malley et al. 2008) and a TE sequence dataset annotated to high quality standards, which for the first time takes into consideration each of the individual TE sequences separately. I show that CG sites are either almost all methylated or all unmethylated within any given TE sequence and illustrate very clearly the major differences between CG, CHG and CHH methylation over TE sequences. Before this work the general perception about the dense DNA methylation seen over TE sequences was that it involves the joint action of RdDM, which explains CHH methylation, and so-called maintenance DNA methylation, which explains the persistence of some CHG and most CG methylation in RdDM-defective mutant backgrounds. The fact that a substantial number of densely methylated TE sequences have no matching siRNAs is often overlooked or simply

ignored. Our analysis corrects this and provides in addition strong evidence that DNA methylation is established or maintained over some TE sequences through spreading from siRNA-targeted flanking regions. Besides myself, the work presented here also involved other people, including Alexis Sarazin who analysed small RNA sequence libraries, and Hadi Quesneville who developed the annotation pipeline and provided TE sequence annotation. The manuscript was mainly written by me and Vincent Colot.

Nucleic Acids Research Advance Access published May 17, 2011

Nucleic Acids Research, 2011, 1–13
doi:10.1093/nar/gkr324

Genome-wide evidence for local DNA methylation spreading from small RNA-targeted sequences in Arabidopsis

Ikhlak Ahmed[1], Alexis Sarazin[1], Chris Bowler[1], Vincent Colot[1,*] and Hadi Quesneville[2]

[1]Institut de Biologie de l'Ecole Normale Supérieure (IBENS), Centre National de la Recherche Scientifique (CNRS) UMR8197 - Institut National de la Santé et de la Recherche Médicale (INSERM) U1024, 46 rue d'Ulm, 75230 Paris cedex 05 and [2]Unité de Recherches en Génomique-Info, Institut National de la Recherche Agronomique (INRA) UR1164, Centre de recherche de Versailles, Route de Saint Cyr, 78026 Versailles cedex, France

Received December 22, 2010; Revised March 31, 2011; Accepted April 20, 2011

ABSTRACT

Transposable elements (TEs) and their relics play major roles in genome evolution. However, mobilization of TEs is usually deleterious and strongly repressed. In plants and mammals, this repression is typically associated with DNA methylation, but the relationship between this epigenetic mark and TE sequences has not been investigated systematically. Here, we present an improved annotation of TE sequences and use it to analyze genome-wide DNA methylation maps obtained at single-nucleotide resolution in Arabidopsis. We show that although the majority of TE sequences are methylated, ~26% are not. Moreover, a significant fraction of TE sequences densely methylated at CG, CHG and CHH sites (where H = A, T or C) have no or few matching small interfering RNA (siRNAs) and are therefore unlikely to be targeted by the RNA-directed DNA methylation (RdDM) machinery. We provide evidence that these TE sequences acquire DNA methylation through spreading from adjacent siRNA-targeted regions. Further, we show that although both methylated and unmethylated TE sequences located in euchromatin tend to be more abundant closer to genes, this trend is least pronounced for methylated, siRNA-targeted TE sequences located 5′ to genes. Based on these and other findings, we propose that spreading of DNA methylation through promoter regions explains at least in part the negative impact of siRNA-targeted TE sequences on neighboring gene expression.

INTRODUCTION

Transposable elements (TEs) are ubiquitous components of genomes and their differential accumulation is responsible for most of the large variations in genome size seen among eukaryotes. However, mobilization of TEs is inherently mutagenic and is therefore a rare event. Repression of transposition involves a variety of mechanisms, including DNA methylation in plants and mammals (1,2). Moreover, TEs are among the fastest evolving sequences, leading over time to the accumulation of degenerate, non-mobile relics.

In plants, TE sequences are typically methylated at CG, CHG and CHH sites (where H = A, T or C) in a process that requires numerous factors, including small interfering RNAs (siRNAs) to guide methylation of homologous DNA sequences, and so called *de novo* and maintenance DNA methyltransferases (2,3). The model plant Arabidopsis offers several advantages for the detailed exploration of the relationship between DNA methylation and TE sequences, such as a small, almost fully sequenced genome (4) and a large collection of mutants affected in the establishment, maintenance or removal of DNA methylation (3). However, despite the fact that DNA methylation in Arabidopsis has been studied genome wide using a variety of approaches, including bisulphite treatment of genomic DNA followed by whole genome sequencing (5,6), patterns of DNA methylation associated with Arabidopsis TE sequences have not been investigated systematically so far.

We previously described the development of a highly sensitive TE annotation pipeline that doubled the fraction of the Arabidopsis genome detected as TE sequences compared to the initial annotation (7). In the present study, we have refined this pipeline further and have used the resulting set of annotated TE sequences,

*To whom correspondence should be addressed. Tel: +33 1 44 32 35 38; Fax: +33 1 44 32 39 35; Email: colot@biologie.ens.fr

which now cover 21% of the genome sequence, to re-analyze publicly available genome-wide DNA methylation and siRNA datasets. Our analysis indicates that although the majority of TE sequences are densely methylated, >25% are unmethylated at most or all sites, or show significant DNA methylation only over one or two of the three types of sites (CG, CHG and CHH). Furthermore, methylated TE sequences are less often characterized by an abundance of matching siRNAs when located in heterochromatin than in euchromatin. These methylated TE sequences with no or few matching siRNAs tend to show higher levels of DNA methylation towards their extremities and are typically flanked on both sides by methylated TE sequences that are targeted by siRNAs. These observations suggest the existence of local spreading of DNA methylation from siRNA-targeted TE sequences. Further, we show that in euchromatin, both methylated and unmethylated TE sequences are most abundant close to genes. However, this preference is less pronounced for methylated, siRNA-targeted TE sequences upstream of genes. Based on these findings, we propose that the negative impact of siRNA-targeted TE sequences on the expression of neighboring genes which has been observed in *Arabidopsis thaliana* and *Arabidopsis lyrata* (8) results at least in part from local spreading of DNA methylation into promoter regions.

MATERIALS AND METHODS

Sequences

The *A. thaliana* Release 5 genomic sequence was downloaded from TIGR web site (http://ftp.tigr.org/pub/data/a_thaliana/ath1/). Annotations Release 7 was obtained from TAIR as a dump of their database. The three TE reference sequence sets (Opt, Maxsize and OptCoding) used in addition to Repbase Update (RU) were described previously (7).

TE detection pipeline

TE sequence models were detected using the following combination of softwares: BLASTER (9,10), RepeatMasker (11), Censor (12,13). Satellite repeats were detected using RepeatMasker, Tandem Repeat Finder [TRF; (14) and mreps (15)]. The Torque resource manager was used to provide control over batch jobs and distributed compute nodes (http://www.clusterresources.com/pages/products/torque-resource-manager.php). Results were stored in a MySQL database (http://www.mysql.com/).

Each program was run independently. Parameters were chosen to make detection as sensitive as possible. The rate of false positives was minimized by running the TE detection softwares on 200-kb fragments of genomic sequence shuffled by di-nucleotides using the program shuffle [HMMer Package; (16)]. For each of the programs BLASTER, RepeatMasker and Censor, the highest score obtained for these di-nucleotide shuffled chunks was used as a threshold to filter out the results obtained on the true genome chunks. Simple repeats were removed from the TE annotation. TE models <20 bp were discarded.

RepeatMasker was run with the MaskerAid (17) search engine with sensitive parameters ('–cutoff 200 –w –s –gccalc –nolow –no_is'). We found MaskerAid to be more sensitive and much faster than Cross-match, under sensitive parameters. Censor was used at high sensitivity with parameter '–s –ns'. BLASTER now uses WU-BLAST as a search engine, and has also been set to more sensitive parameters, (inspired from MaskerAid settings). We used Blaster with parameters '–W –S 4'. The RMBLR procedure (9) has been replaced by a new procedure, called 'combinedBLR', which now combines the results obtained from BLASTER, RepeatMasker and Censor and gives them to MATCHER for chaining. To do this, we normalized alignment scores to be the hit length times the identity percentage. The MATCHER program has been developed to map match results onto query sequences by first filtering overlapping hits. When two matches overlap on the genomic (query) sequence, the one with the best alignment score is kept, the other is truncated so that only non-overlapping regions remain on the match. As a result of this procedure a match is totally removed only if it is included in a longer one with a best score.

Long insertions or deletions in the query or subject could result in two matches, instead of one with a long gap. Thus the remaining matches are chained by dynamic programming. A score is calculated by summing match scores and subtracting a gap penalty (0.05 times the gap length) as well as a mismatch penalty (0.2 times the mismatch length region), as described previously (18).

The chaining algorithm [(19), pp. 325–329] is modified to produce local alignments. A match is associated with a chain of other matches only if this results in a higher score. The best-scoring chain is kept and the search is repeated minus this chain until no more chain is found. This algorithm is run independently for matches on strand +/+, +/− and −/+. A maximum of 20 bp of overlap is allowed between matches. The chaining algorithm enables the recovery of TE sequences containing long insertions.

Although BLASTER, RepeatMasker and Censor are front ends of the same WU-BLAST program, they cover respectively 21, 18 and 19 Mb of the genome sequence, when the RU TE reference set is used (Supplementary Table S3). Note that without any score threshold, they appear to have a high false positive rate (cover 90, 18 and 23 Mb, respectively). To reduce the false positive rates we rely on a statistical procedure to set their parameters at very high-sensitive values. Supplementary Table S4 shows TE-detection overlaps between the three softwares. BLASTER appears to be the most sensitive, followed by Censor and then RepeatMasker. This is a consequence of the different BLAST parameters used by these programs. When results obtained by the three programs are combined, TE coverage (excluding satellite) is increased to 21.7 Mb.

The Arabidopsis genome contains several regions where TE sequences cluster, often as a result of nested insertions. These are particularly challenging to detect and classic annotation algorithms may fail to connect fragments of split TEs. This prompted us to implement a 'long join' procedure, which is based on age estimates of TE

fragments. Fragments to be joined must be co-linear and have the same age, as estimated using the percentage of identity with the TE reference sequence (20). In the 'nest join' version of the procedure, the inner TE sequence cover >95% of the region between the two fragments to be joined and be younger. TE sequences can also be split as a result of large non-TE sequence insertions. To account for this possibility, TE fragments that have the same age, are separated by an insert of <5 kb and align within 500 bp of each other on the TE reference sequence (Figure 1) are joined in a version of the 'long join' procedure referred to as 'simple join'.

TE fragments already connected in a previous step by MATCHER are split if inner TE fragments are younger than outer joined fragments. Although the 'long join' procedure outperforms MATCHER, which relies on dynamic programming and a scoring scheme that are ill adapted for extreme situations, it did produce only few 'simple join' and no 'nest join'. Indeed, many 'long join' were denied because fragments are too dissimilar in age (>2% difference in age) or too far apart (>100 kb). Results are summarized in Supplementary Table S5.

Assigning confidence scores to TE sequence models

The pipeline provides TE sequence models composed of four lines of evidence, one for each TE reference sequence set. The longest evidence (maximum length) for each TE sequence model is recognized as 'best evidence' and is used to determine the precise genomic coordinates. A score is assigned to the model based on the origin of the best evidence supporting the model. Indeed, because small insertions, unrelated to the TE sequence, may be present in the genomic copies that were used for building the OptCoding, MaxSize and Opt TE reference sequence sets (7), evidence obtained with these different sets may not be all equally reliable. In fact, the coding constraint imposed on the OptCoding design makes this TE reference sequence set more reliable than Opt and MaxSize. A score of 3 (best) is attributed to models supported by at least RU or OptCoding, a score of 2 when support comes from Opt only and a score of 1 when support come from MaxSize alone. In cases where the longest evidence has a lower score than shorter evidence (>100 bp) and does not expand it by >50 bp, the higher score is assigned to the longest evidence.

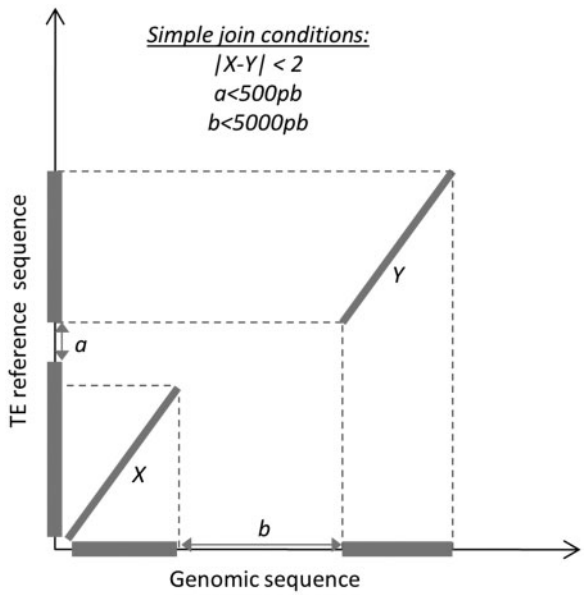


Figure 1. Schematic dot plot representation of 'simple join' conditions. Matching regions between genomic and TE reference sequence are represented by diagonals. Note that these regions might be fragments already connected by MATCHER. *X* and *Y* indicate percentage of identity to the TE reference sequence. *a* and *b* refer to the length of non-matching DNA on the TE reference and genomic sequences, respectively.

Satellites are comprised of highly embedded tandem repeats, and the long-join procedure does not seem to work well when the TE reference sequence sets OptCoding, Opt or MaxSize are used. Indeed these TE-reference-sequence sets tend to merge many satellite units into one big unit. In contrast, RU tends to keep unit boundaries for tandem repeats. Satellites were thus annotated solely based on RU evidence and with no score attached.

RU provides best evidence for the largest set 12 922 (40%) of annotated TE sequence models, followed by Opt 12 214 (38%), MaxSize 5758 (18%) and OptCoding 981 (4%). Of the 31 245 annotated TE models, 13 752 (44%) have a score of 3; 11 773 (37.7%) a score of 2 and 5720 (18.3%) a score of 1. In addition, 3342 sequences were annotated as satellites. Note that because of the stringent statistical threshold used to detect TE sequences with high confidence, some old TE insertions are likely missed, such as those proposed to be responsible for the epigenetic regulation of the imprinted gene *FWA* (21).

DNA methylation analysis

Single-nucleotide resolution DNA methylation data were used from Cokus *et al.* (5). DNA methylation analysis was carried out separately for CG, CHG and CHH sites (Supplementary Figure S6). Sites were considered as methylated if at least 10% of the reads (CG sites) or at least one read (CHG and CHH sites) were indicative of methylation. However, because CG methylation was found to be symmetrical, as expected, methylation status was copied to the opposite strand in cases of no or insufficient coverage for that strand. Although CHG sites are symmetrical, CHG methylation was found to depart significantly from symmetry in a large number of cases, and thus methylation status of CHG sites was not copied to the opposite strand in cases of insufficient coverage for one strand. Once the methylation status of all available sites was established, DNA methylation for TE sequences or any other annotated feature is computed as a fraction of methylated Cs to the total number of covered Cs, for each of the three types of sites.

High-density tiling microarray datasets (22) were downloaded from the Gene Expression Omnibus (GSE5974). Potentially cross-hybridizing probes were identified by aligning them on the genomic sequence with nucmer from the MUMmer v3 package (23) with parameters: –maxmatch –minmatch = 10 –mincluster = 50 –nosimplify. A total of 36 993 probes (out of 382 178) were removed from the analysis because they had multiple matches with

85% identity or more. For each annotated feature, the cumulated sequence length of probes identifying a positive methylation signal was normalized by total length of probes covering the feature.

siRNA density

Small RNAs deep-sequencing data obtained from Arabidopsis Whole-aerial tissues were downloaded from GEO (accession: GSE14696) (24) and used to calculate the 24-nt siRNA density for all TE sequences with defined DNA methylation patterns. As there was a high correlation between different replicates of this library, we merged them together to achieve a ~6 million read library. Reads were mapped to the *A. thaliana* genome using MUMmer v3 and 24-nt siRNA density was calculated as follows:

$$\mathrm{ND} = \frac{\sum \frac{\mathrm{NR}_i}{\mathrm{NM}_i}}{\mathrm{TNR} \times \mathrm{Region\ length}} \times 10^8.$$

Where NR_i is the number of reads corresponding to match M_i and NM_i is the total number of matches for the sequence across the genome. TNR is the total number of reads in library and Region length is the length of TE sequence for which density is being calculated. Densities are expressed as number of reads per kilobite of the sequence per hundred thousand of the library reads.

The R package (http://cran.r-project.org) and Perl (http://www.perl.org) were used for the statistical and DNA methylation analyses, respectively.

RESULTS

Improved annotation of TE sequences

Identification of TE sequences within genomes by homology searches is challenging because many of these sequences are highly degenerate derivatives of functional TEs or occur as nested insertions. Previously, we described a method that substantially improves TE sequence detection (7). This method relies on the use of multiple sets of TE reference sequences, specifically designed to reflect diverse aspects of TE structure and evolution on the one hand, and on a TE annotation pipeline which combines several sequence-similarity search programs on the other (7). The annotation pipeline has been further refined, in particular to allow for the detection of nested insertions through a 'long join procedure'. Briefly, two or more TE sequences separated by <5 kb in the genome are joined together in the final annotation if they align in the same order and orientation within <500 bp from each other on the corresponding TE reference sequence and if they diverge from it to a similar extent (Figure 1; 'Materials and Methods' section).

Using this improved version of the TE-annotation pipeline, we now identify a total of 31 245 TE sequences, which cover 25 Mb (21%) of the 119-Mb genome sequence available. As initially reported (4), retroelements, which transpose through an RNA intermediate, represent the largest fraction of TE sequences (10 Mb), followed by helitrons (8 Mb) and DNA transposons (7 Mb), which transpose through rolling circle and cut and paste processes, respectively. A detailed description of the detected TE sequence models is provided in File 1 in Supplementary Data and the new annotation can be found at TAIR, starting with release 8. Of note, 85% and 2.5%, respectively, of sequences annotated in the TAIR release 7 as pseudogenes (3315/3897) and genes (790/31 726) show at least 75% overlap with our TE annotation (File 2 in Supplementary Data), indicating that they are in all likelihood TEs.

Defining a robust DNA methylation dataset

Two studies have combined bisulphite treatment of genomic DNA, which converts unmethylated cytosines to uracils but leaves methylated cytosines intact, with next-generation sequencing to provide single-nucleotide resolution DNA methylation maps of the Arabidopsis genome (5,6). The two studies produced essentially identical results and although no extensive analysis of TE sequences was carried out, it was concluded in both cases that repeat elements including TEs are typically methylated at CG, CHG and CHH sites. Moreover, these two studies reported that ~30% of genes are methylated, but almost exclusively at CG sites and within part of the transcribed region only.

In order to explore the genome-wide patterns of DNA methylation associated with TE sequences more systematically and in greater detail, the DNA methylation data of Cokus *et al.* (5) were first reassessed. Although the average sequencing coverage was ~20-fold (5), large variations were observed, with 66.7% and 1.3% of sequenced cytosines covered by >10 or <50 reads, respectively (Figure 2). To avoid any potential problems resulting from this uneven coverage, only those cytosines with read depths between 10 and 50 were considered for our analysis, which amounted to 32% of uniquely mapped cytosines.

Although CG sites are usually either unmethylated or methylated on over 80% of the molecules sequenced (5,6), a significant number (15% of total) have intermediate methylation levels (File 3 in Supplementary Data). Thus, among CG sites with at least one read indicative of methylation, only 53% in genes and 66% in TE sequences have methylation levels above 80% (Supplementary Figure S1). Given the possible involvement of transcription in gene body methylation (22,25), lower levels of methylation of CG sites within genes could reflect tissue-specific differences in expression. In the case of TE sequences, which are not transcribed in most cell types, intermediate levels of methylation at CG sites may rather indicate active demethylation or preferential action of the *de novo*, RNA-directed DNA methylation (RdDM) pathway over the so-called maintenance DNA methylation at these sites (3). Our analysis also confirmed that in contrast to CG methylation, CHG and CHH methylation, which are restricted almost exclusively to TE sequences and other repeat elements, rarely reach levels greater than 80% (5,6).

Based on these observations, sites were declared as methylated if at least 10% of the reads for CG sites or at least one read for CHG and CHH sites indicate methylation. Using these criteria, 82, 74 and 31% of CG, CHG

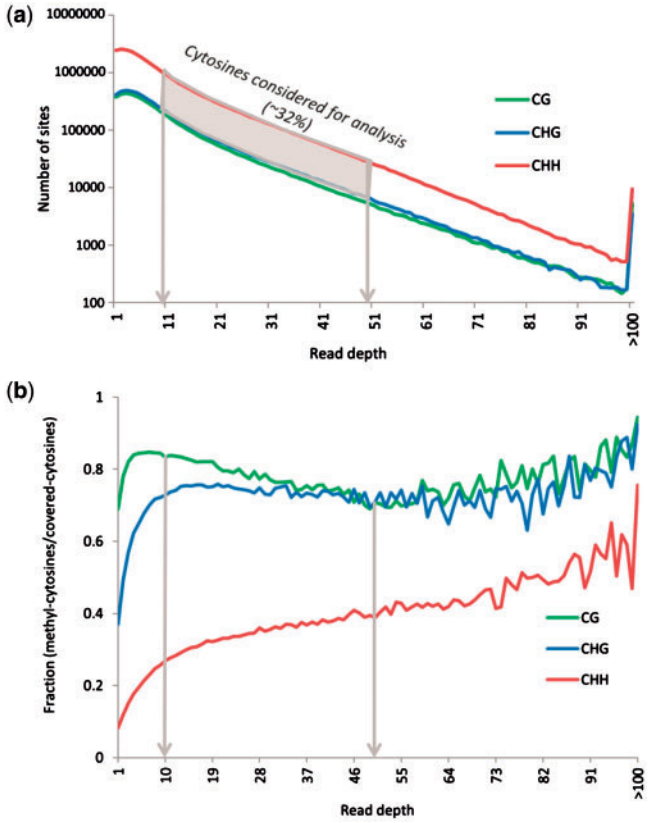


Figure 2. Read-depth coverage map of whole-genome bisulphite sequencing dataset (5). (**a**) The *x*-axis shows the number of bisulphite sequencing reads at a given cytosine and the *y*-axis represents number of sites. Most cytosines in all three sequence contexts are covered by <10 reads. For our analysis, only those cytosines were considered for which read depths were between 10 and 50. This proportion, shown in grey, represents ~32% of the original data in all three cytosine contexts. (**b**) Fraction of methyl-cytosines detected at a given sequencing coverage (Read depth). Read depths below 10 lead to an underestimation of methylated CHG and CHH sites, while read depths above 50 tend to be more often associated with methylated cytosines at all three types of sites.

and CHH sites within TE sequences are methylated, respectively, which is much higher than for genes (26, 4 and 3% for CG, CHG and CHH sites, respectively; File 3 in Supplementary Data). Furthermore, the frequency of methylated sites within the first 500 bp beyond genes decreases ~4-fold for CG sites (7%) and remains consistently low for CHG and CHH sites. In contrast, the frequency of methylated CG, CHG and CHH sites is only reduced 2-fold within the first 500 bp outside of TE sequences (42, 36 and 16%, respectively). This lower reduction in the frequency of methylated sites outside of TE sequences compared to genes could indicate that our annotation pipeline does not precisely define TE sequence boundaries or else that DNA methylation can spread from TE sequences into flanking regions (see below). We also note that among CHG sites declared as methylated, a large proportion (>30%) have statistically significant discordant methylation levels between the two strands ($P < 0.05$ in Chi-square goodness of fit test; Supplementary Figure S2). Furthermore, almost all of the latter are devoid of matching siRNAs (data not shown). These findings indicate that in at least 30% of cases, neither sequence symmetry nor siRNAs play any role in maintaining CHG methylation, which is consistent with this process relying predominantly on a reinforcing loop with dimethylation of lysine 9 of histone H3 (2).

Methylation status of TE sequences

We next determined the DNA methylation status of all individual TE sequences with sufficient information for CG, CHG and CHH sites. Given the repeated nature of TEs and the fact that only sequence reads that map to unique genomic locations with very high confidence are considered (5), cytosine coverage is reduced for TE sequences (66%) relative to the whole genome (85.6%). Nonetheless, a quarter of cytosines within TE sequences have read depths between 10 and 50, a fraction similar to that for the whole genome (27.2%) and almost identical for CG, CHG and CHH sites. Based on these observations, we only considered the 13 667 TE sequences (43.7% of total) for which information is available for >25% of each of the three distinct types of sites and the 3418 TE sequences (10.9% of total) containing only one (CHH) or two types (CHH and CG or CHG) of sites and still fulfilling the >25% coverage criterion for these sites. Two main categories of TE sequences are thus excluded from our analysis, those corresponding to recent insertions and for which reads could not be assigned unambiguously because of two or more possible matches in the genome, and those for which technical or other biases lead to <25% coverage for CG, CHG or CHH sites.

For each of the 17 085 TE sequences retained for analysis, we determined the methylation status separately for CG, CHG and CHH sites. A sequence was deemed methylated at a given type of site if at least 5% of the sites of this type had reads indicative of methylation, a value that is above the level of non-conversion of unmethylated cytosines [2–4%; (5)]. As illustrated in Figure 3, the fraction of methylated sites within individual TE sequences differs dramatically between CG, CHG and CHH sites. Thus, whereas CG sites are typically all unmethylated or all methylated within a given TE sequence, the fraction of methylated CHG sites varies almost linearly between 0% and 100% (with the exception of a peak at 98–100%) and that of CHH sites rarely exceeds 50%. To simplify the analysis, and because the reason(s) for such wide variations in the frequency of methylated CHG and CHH sites remain to be determined, the methylation status of individual TE sequences was simply summarized as either methylated (M) or un-methylated (U) for each of the three types of sites based on the 5% threshold defined above. This convention leads therefore to eight possible DNA methylation patterns (Table 1), or 10 in the case of TE sequences that are devoid of CG and/or CHG sites (Supplementary Table S1). Although all 18 patterns are observed, few predominate.

Thus, among the 13 667 TE sequences with >25% of informative CG, CHG and CHH sites, 58% have an MMM pattern (methylated in all three types of sites) and another 20% have a UUU pattern (unmethylated;

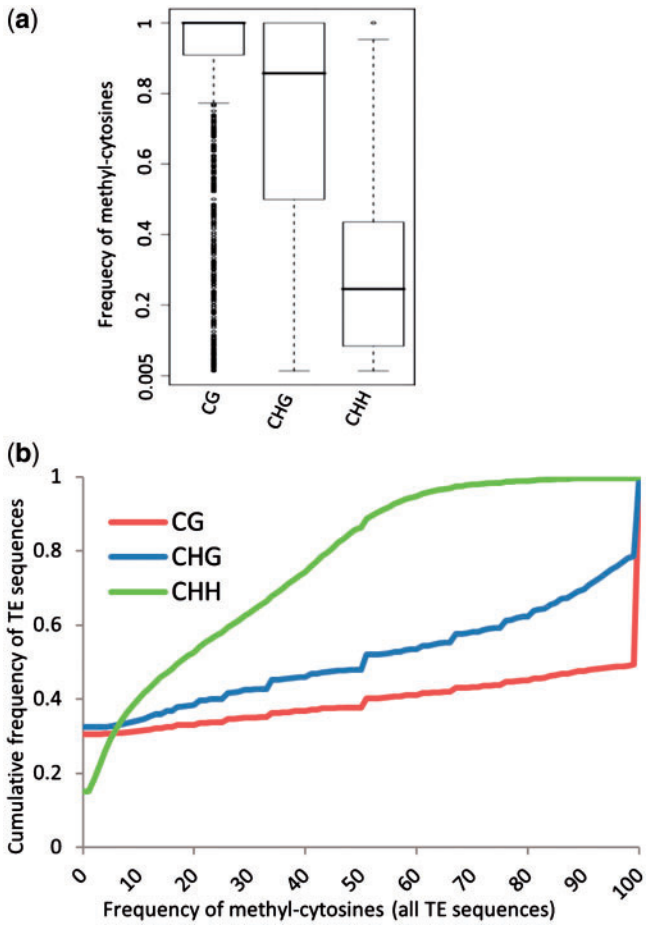


Figure 3. Frequency of methylated CG, CHG and CHH sites in TE sequences. (**a**) Boxplots showing frequency distribution of methyl-cytosines for TE sequences methylated for at least one type of site. Most of these TE sequences have almost all of their CG sites and a majority of their CHG sites methylated. (**b**) Frequency distribution of all TE sequences in relation to percentage of methylated sites, for each of the three types of sites.

Table 1). Moreover, the next most abundant pattern is UUM (6%), which is characterized by a particularly low median level of methylation for methylated CHH sites and a low-median frequency of such sites (1% and 8%, respectively, compared to 6% and 36% for the MMM pattern). Thus, whereas a majority of TE sequences for which information is available for CG, CHG and CHH sites are densely methylated, a large proportion (26%) are not significantly methylated at any of the three types of sites. Another 10% of TE sequences are methylated mainly at CG sites (MUM, MMU and MUU patterns) and have low median levels of methylation and frequency of methylated CG sites (12–40% and 38–75%, respectively, compared to 82% and 100% for the MMM pattern). These TE sequences therefore have methylation patterns resembling those of methylated genes. Finally, a small proportion of TE sequences exhibit non-CG methylation only (UMU and UMM patterns, 5% of total). Methylation data obtained by immunoprecipitation of DNA with an anti-methylcytosine antibody followed by hybridization to a high-density genome-tiling microarray [MeDIP chip; (22)] were used to validate the pertinence of the main and most contrasted patterns MMM and UUU+UUM. Out of the 382 178 probes on the array, 14 482 were extracted that covered, with little risk of cross-hybridization (see 'Materials and Methods' section), 4272 MMM and 1433 UUU + UUM TE sequences. Overall, 75 and 98% of probes corresponding to MMM and UUU + UUM patterns were declared as methylated and unmethylated, respectively, validating our classification and indicating a higher sensitivity of bisulphite sequencing over MeDIP in detecting methylated DNA, as previously reported (26).

Among the 3418 TE sequences devoid of CG and/or CHG sites and with >25% of informative sites of the other type(s), 42% have patterns (−U, -UU and U-U) clearly indicative of no, or very low, methylation (Supplementary Table S1). The -UM and U-M patterns (8% of total) also indicate very low methylation as they

Table 1. Methylation patterns for 13367 TE sequences with CG, CHG and CHH sites

	MMM	UUU	UUM	MUM	MUU	UMU	MMU	UMM
No. TEs	7983	2687	768	554	469	453	449	304
Percentage TEs	0.58	0.20	0.06	0.04	0.03	0.03	0.03	0.02
Nb TEs in Heterochromatin	5802	591	205	279	175	113	210	82
Nb TEs in Euchromatin	2181	2096	563	275	294	340	239	222
Average size (bp)	887	443	396	443	591	635	679	488
Average size in Heterochromatin	929	398	332	453	589	583	612	440
Average size in Euchromatin	777	456	419	433	593	652	738	506
Average distance*	0.24	0.27	0.27	0.26	0.26	0.27	0.26	0.27
Median Frequency of methylated sites								
CG	1.00	0.00	0.00	0.75	0.38	0.00	0.60	0.00
CHG	0.90	0.00	0.00	0.00	0.00	0.17	0.29	0.25
CHH	0.36	0.00	0.08	0.12	0.02	0.02	0.03	0.10
Median level of methylation (methylated reads/total reads)								
CG	0.82	0.00	0.00	0.40	0.12	0.00	0.37	0.00
CHG	0.28	0.00	0.00	0.00	0.00	0.01	0.03	0.02
CHH	0.06	0.00	0.01	0.01	0.00	0.00	0.00	0.01

*Jukes–Cantor distance from reference sequence.

are characterized by low median frequencies of methylated CHH sites (12–13%). In contrast, the −M pattern (20% of total) is characterized by a much higher median frequency of methylated sites (33%), close to that of the MMM pattern (36%). Similarly, the other four patterns (-MM, -MU, M-M, M-U, 30% of total) have median frequencies of methylated sites comparable to those of the MMM pattern. Thus, whereas half of the 3418 TE sequences with no CG or CHG sites are densely methylated, the other half have no or very low methylation, which is twice the fraction of TE sequences with no or very low methylation among those containing all three types of sites. This latter result indicates therefore a critical role for CG and CHG sites in dictating methylation of TE sequences.

Arabidopsis TE sequences can be classified into 13 superfamilies, four corresponding to retroelements (Copia, Gyspy, LINE and SINE), five to well-defined DNA transposons (En-Spm, Harbinger, HAT, MuDR and Pogo) and one each to Helitrons, TEs of a composite nature, Tc1/mariner and other DNA transposons. As shown in Figure 4, the Gypsy and /En-Spm superfamilies have the highest proportion (~90%) of methylated TE sequences, and RC/Helitrons and Tc1/mariner superfamilies the lowest such fraction (40–50%).

Genomic distribution of TE sequences

The 119 Mb of available Arabidopsis genome sequence can be divided into gene-rich/TE-poor and gene-poor/TE-rich regions that form the euchromatic arms of chromosomes and pericentromeric heterochromatin plus interstitial heterochromatic knobs, respectively (4,21,27). Our analysis indicates that more than two thirds of densely methylated TE sequences (MMM, -MM, M-M and −M) and a similar proportion of unmethylated or poorly methylated TE sequences are located within pericentromeric heterochromatin and euchromatin, respectively (Table 1 and Supplementary Table S1). Furthermore, the last two categories of TE sequences correspond mainly to TE relics depleted in CpGs, as indicated by their shorter size compared to their densely methylated counterparts and their higher divergence from the cognate reference TE sequence (Figure 5). Thus, the dense DNA methylation characteristic of heterochromatin results not only from the much higher density of TE sequences compared to euchromatin, but also from the larger ratio of methylated to unmethylated TE sequences within heterochromatin and the longer length of methylated TE sequences on average.

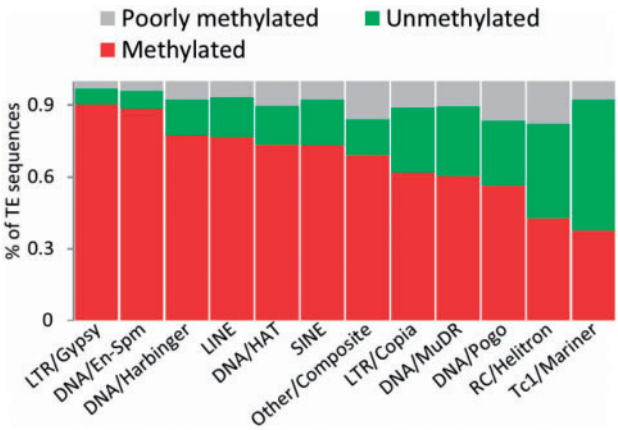


Figure 4. DNA methylation patterns within TE superfamilies. Unmethylated TE sequences are found across all classes but >90% of the sequences for LTR/Gypsy and DNA/En-Spm superfamilies are methylated. The RC/Helitron and Tc1/mariner superfamilies comprise the largest fraction (50–60%) of unmethylated TE sequences.

Although the vast majority of TE sequences are located outside of genes and cluster within heterochromatin, 17% of euchromatic TE sequences intersect with gene annotations (Table 2). Furthermore, most of these TE sequences overlay with exons, suggesting a high incidence of 'exonization' of TE sequences in Arabidopsis. Finally, whereas 53% of these exonic TE sequences are unmethylated (UUU + UUM), 24% are highly methylated (MMM), suggesting a recent origin.

TE sequences and siRNAs

Deep sequencing of small RNAs has revealed that a large fraction of methylated repeat elements present in the Arabidopsis genome are characterized by an abundance of matching 24-nt siRNAs throughout development (6). To investigate more precisely the association of different DNA methylation patterns of TE sequences with endogenous 24-nt-long siRNAs, small RNA deep sequencing data obtained from Arabidopsis whole-aerial tissues were downloaded from GEO (accession: GSE14696) (24) and used to calculate the 24-nt siRNA density for all TE sequences with defined DNA-methylation patterns (see 'Materials and Methods' section). As expected, almost all (95%) unmethylated TE sequences and many (48–58%) of those poorly methylated are devoid of matching 24-nt siRNAs, irrespective of their location in euchromatin or heterochromatin. On the other hand, 89 and 82% of densely methylated TE sequences located respectively in euchromatin and heterochromatin have matching 24-nt siRNAs (Figure 6a and b). This result confirms that most densely methylated TE sequences are associated not only with siRNAs, but also indicates a lower proportion of such sequences in heterochromatin. Indeed, differences between euchromatic and heterochromatic methylated TE sequences were even more pronounced when considering the proportion of those having an abundance of matching siRNAs (density >0.25 reads/kb/10^5 library reads), which is 61% in euchromatin but only 21% in heterochromatin.

Given the high density of TE sequences in heterochromatin, we explored the possibility that methylation of heterochromatic TE sequences with no matching 24-nt siRNAs could occur through local spreading from flanking siRNA-targeted sequences. To this end we considered the set of 749 heterochromatic MMM TE sequences longer than 200 bp and with no matching 24-nt siRNAs but flanked within 1 kb on one or both sides by sequences associated with siRNAs. Each TE sequence was split in equal halves and DNA methylation densities were calculated in non-overlapping 100-bp windows for

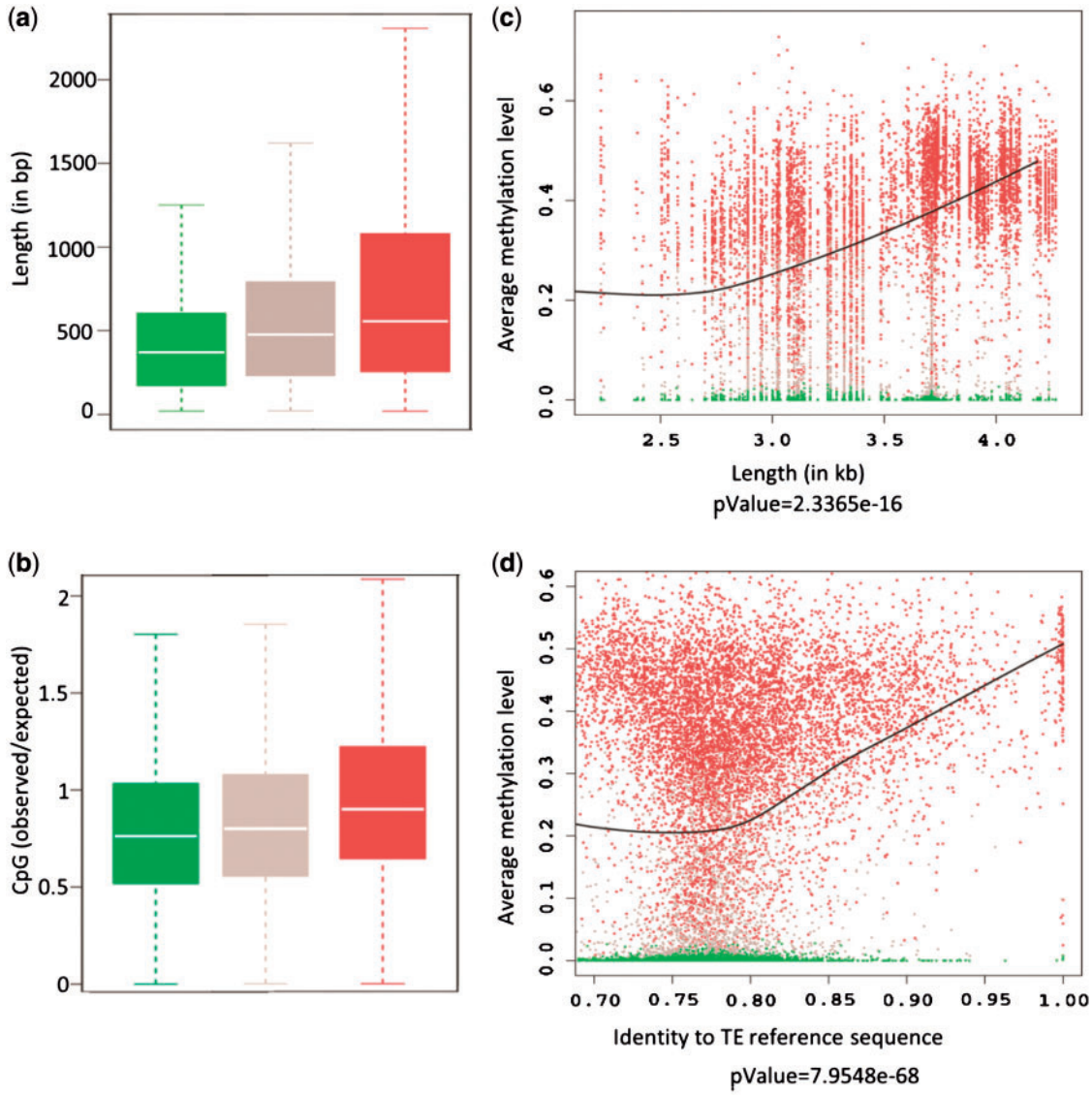


Figure 5. Relationships between DNA methylation, size, CpG content and divergence of TE sequences. Color code is as in Figure 4. (**a**) Unmethylated TE sequences tend to be smaller than their methylated counterparts. (**b**) Boxplots showing observed versus expected CpGs for the three DNA methylation patterns considered. Unmethylated TE sequences are depleted in CpGs compared to poorly methylated TEs (P-value = 0.004793, Wilcoxon rank-sum test) or methylated TEs (P-value $< 1e-10$). Poorly methylated TE sequences also have a lower CG content compared to methylated TE sequences (P-value = $5.369e-13$). (**c** and **d**) Average methylation levels of TE sequences are plotted according to length or percentage of identity with the TE reference sequence. Significant positive correlation (black curve) is observed in each case.

Table 2. TE sequences within genes

Methylation pattern	Number of TEs	Percentage pattern	Intronic	Exonic	Percentage intronic	Percentage exonic
UUU	456	0.44	116	336	0.26	0.74
MMM	249	0.24	59	189	0.24	0.76
UUM	105	0.10	22	82	0.21	0.79
UMU	64	0.06	11	53	0.17	0.83
MUU	54	0.05	12	42	0.22	0.78
MUM	48	0.05	11	37	0.23	0.77
MMU	41	0.04	10	31	0.24	0.76
UMM	30	0.03	5	25	0.17	0.83

each half and its corresponding siRNA-associated flank. As shown in Figure 6c, median DNA methylation densities are uniformly high along the 1 kb flanks, but decrease progressively within the first 500 bp of TE sequences with no matching siRNAs. Correspondingly, the 5371 siRNA-associated MMM TE sequences show higher DNA methylation than their flanks, which may or may not have matching siRNAs (Figure 6d). Taken together, these findings provide strong evidence that DNA methylation can spread over short distances (~500 bp) from siRNA-targeted regions into flanking sequences. Furthermore, analysis of additional methylomes (6) reveals that DNA methylation gradients are abolished in plants defective for the CG maintenance methyltransferase MET1 but are still detectable in plants

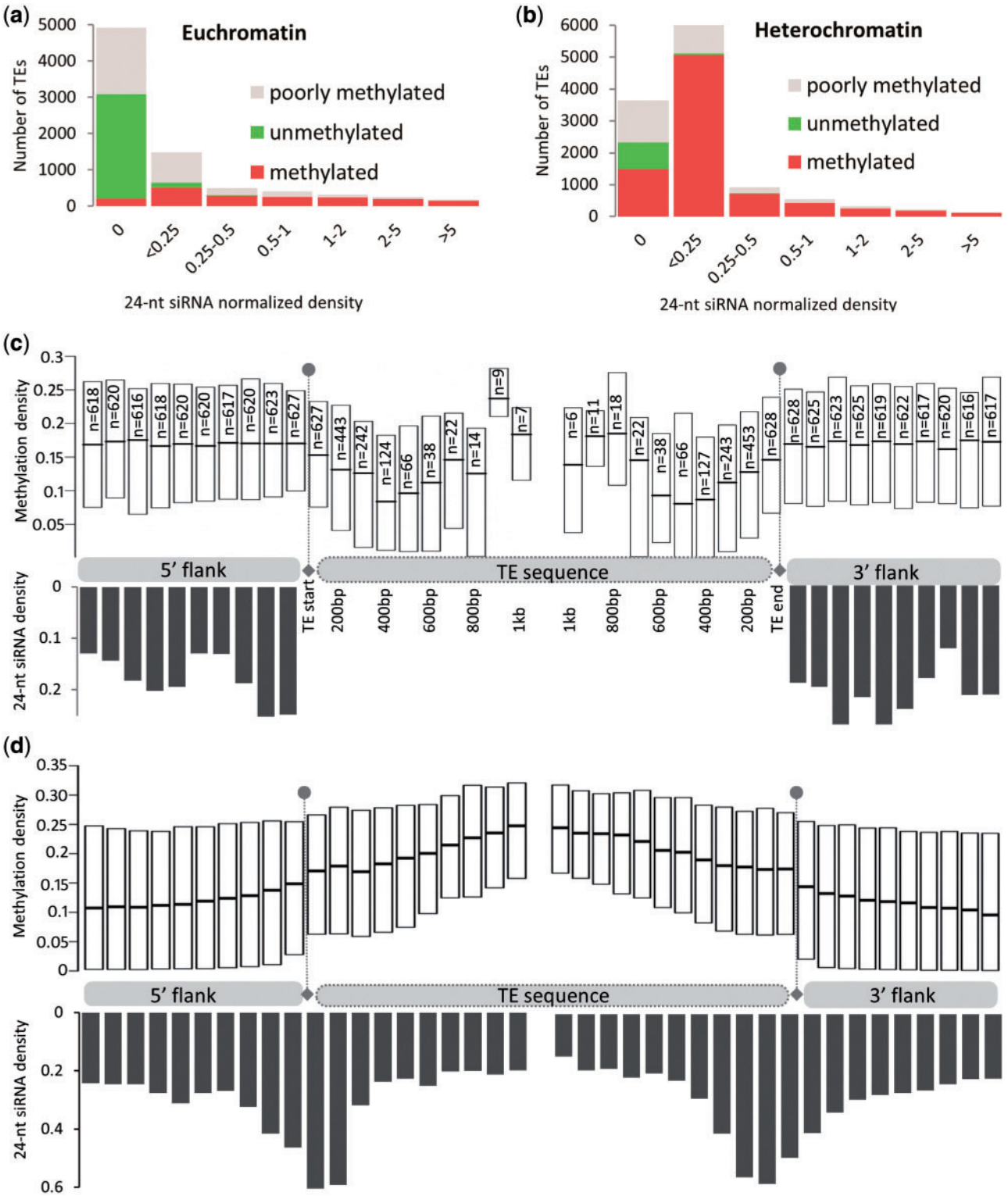


Figure 6. Relationship between methylated TE sequences and 24-nt siRNAs. (**a**) Methylated euchromatic TE sequences are almost always associated with an abundance of siRNAs. (**b**) A significant number of heterochromatic TE sequences are methylated but not associated with siRNAs (**c**) Methylated hetrochromatic TE sequences (size >200 bp) not associated with siRNAs but flanked within 1 kb on one or both sides by sequences associated with siRNAs. These TE sequences were split in half and DNA methylation densities were calculated in 100-bp windows along the two flanks and TE sequence halves by dividing the number of reads indicative of methylation at CG, CHG and CHH sites by the total number of cytosine-covering reads. Results are shown as boxplots of DNA methylation densities. Average normalized siRNA densities are also indicated for each 100-bp window. DNA methylation densities are uniform along the 1 kb flanks, but decrease progressively within the first 500 bp of TE sequences from both sides. (**d**) TE sequences associated with 24-nt siRNAs show increasing methylation from their extremities and decreasing methylation in their flanks.

that are simultaneously defective for the *de novo* DNA methyltransferases DRM1 and DRM2 and the CHG-specific DNA methyltransferase CMT3 [*drm1*, *drm2* and *cmt3* (*ddc*); Supplementary Figure S3a and b]. Nonetheless, the gradient in the *ddc* triple mutant is less steep than in wild type. Finally, plants defective for three of the four known Arabidopsis DNA demethylases [*ros1dml2dml3* or *rdd* triple mutant; (6)] display DNA methylation gradients similar to wild type (Supplementary Figure S3c). Taken together, these results rule out any significant contribution of active DNA demethylation to the gradients observed and suggest a complex set of interactions between different DNA methyltransferases in promoting or limiting DNA methylation spread.

To analyze further the local spreading of DNA methylation from siRNA-targeted TE sequences, methylation densities were plotted separately for CG, CHG and CHH sites (Figure 7). Although gradients are observed in wild type for the three types of sites, slopes are maximal for CHG, suggesting that CHG methylation spreads over shorter distances than CG and CHH methylation. Furthermore, CHG methylation is completely abolished both in the flanks and within TE sequences in the *ddc* triple mutant (Figure 7), suggesting that at least in this background the residual CG and CHH methylation gradients are contributed by DNA methyltransferases other than DRM1, DRM2 and CMT3. These results, together with the absence of any discernible methylation gradient for CHG and CHH in *met1* (Supplementary Figure S4), provide additional evidence that the extent of DNA methylation spread results from complex interactions between different DNA methyltransferases.

Whereas most MMM TE sequences with no matching siRNAs show decreasing DNA methylation towards their middle, uniform DNA methylation across the entire length is observed for some large TE sequences. This suggests either a more extensive spreading of DNA methylation in these cases or the existence of DNA methylation mechanisms not associated, directly or indirectly, with siRNAs. In agreement with the latter hypothesis, the few euchromatic MMM TE sequences

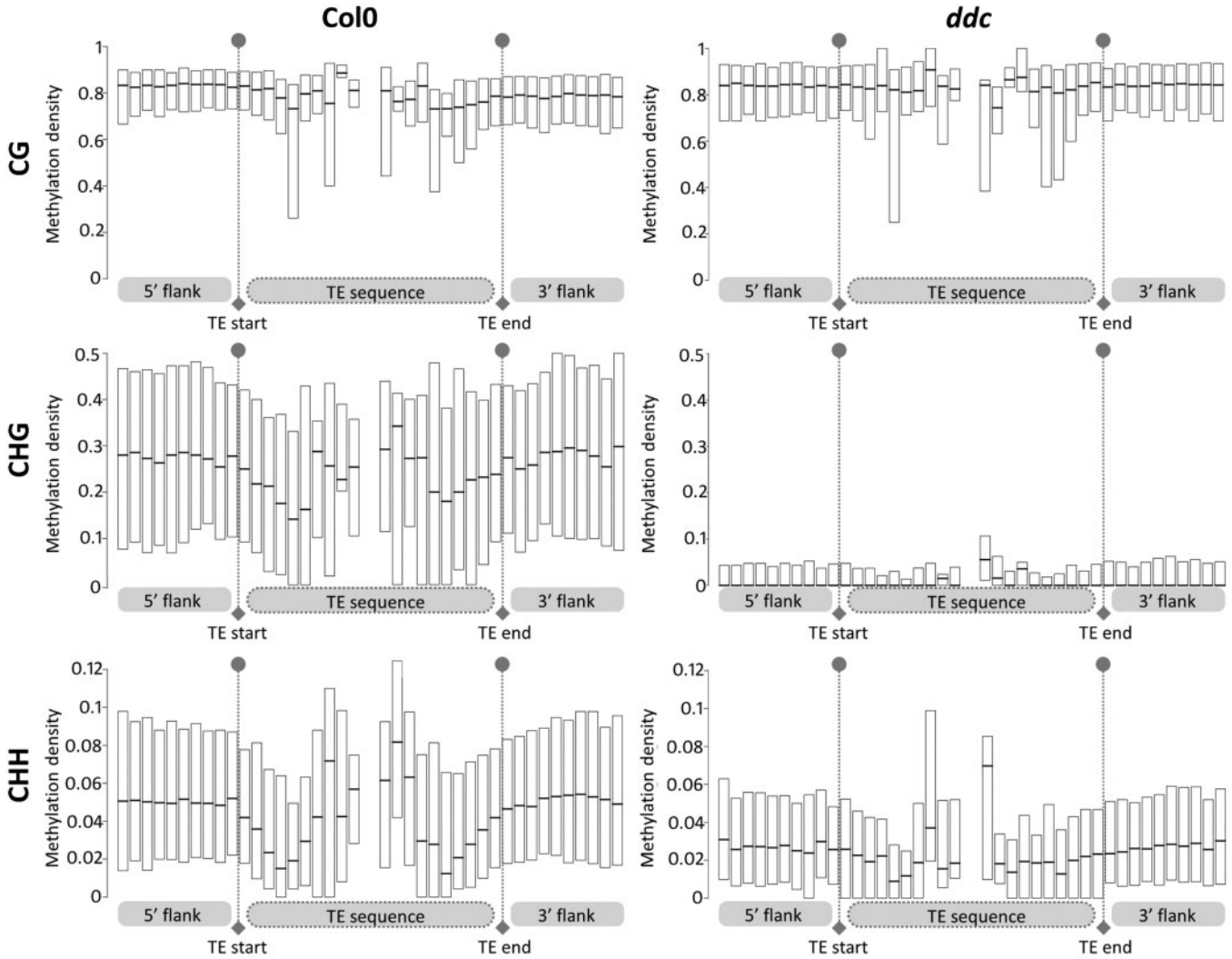


Figure 7. Analysis of methylation spreading for CG, CHG and CHH sites. The figures in the first and second columns correspond to wild type (Col0) and the *drm1*, *drm2*, *cmt3* triple mutant (*ddc*), respectively.

with no matching siRNAs tend to have uniform DNA methylation density throughout their length and are typically flanked by unmethylated sequences (data not shown).

TE sequences and flanking genes

It was previously shown that TE sequences tend to be less methylated when located close to genes, presumably because of deleterious effects of TE methylation on the expression of neighboring genes (28). The proportion of unmethylated TE sequences was reported to drop from ~55% within genes to below 20% for the first 500 bp window away from genes, with little further decrease beyond this point. However, this analysis did not distinguish euchromatic from heterochromatiic genes (28), which are characterized by dramatically distinct intergenic regions (short and TE-poor versus long and TE-rich, respectively). This prompted us to explore further the underrepresentation of methylated TE sequences near genes using our extended dataset and only considering genes within euchromatin. To this end, methylated and unmethylated TE sequences were scored in 100-bp windows for a distance of up to 1 kb upstream and downstream of genes. Our analysis reveals that in euchromatin, both methylated and unmethylated TE sequences tend in fact to over accumulate close to the 5′- and 3′-ends of genes (Figure 8a). Moreover, although the ratio of unmethylated versus methylated TE sequences drops with distance away from genes, as previously reported (28), this drop is rather limited (60% to a minimum of 40%), specific to the 5′-end of genes and less discernible when considering only methylated TE sequences with matching siRNAs, which are the least abundant overall (Figure 8b and Supplementary Table S2). These results suggest therefore that methylated TE sequences have more deleterious effects on transcription initiation than termination and that these effects are more severe when methylated TE sequences have matching siRNAs.

We next tested if spreading of DNA methylation from siRNA-targeted TE sequences could provide a plausible explanation for the deleterious effects of TE methylation on gene expression. For this, DNA methylation densities were calculated in non-overlapping 100-bp windows for all methylated euchromatic TE sequences ($n = 401$) associated with 24-nt siRNAs and flanked by sequences not associated with siRNAs. Although spreading is less pronounced than in heterochromatin, it is nonetheless clearly detectable over ~200 bp beyond siRNA-targeted sequences. Moreover, our analysis suggests that DNA methylation spreads from the center of euchromatic TE sequences towards their extremities, which are often not associated with siRNAs (Supplementary Figure S5) unlike their heterochromatic counterparts (Figure 6d).

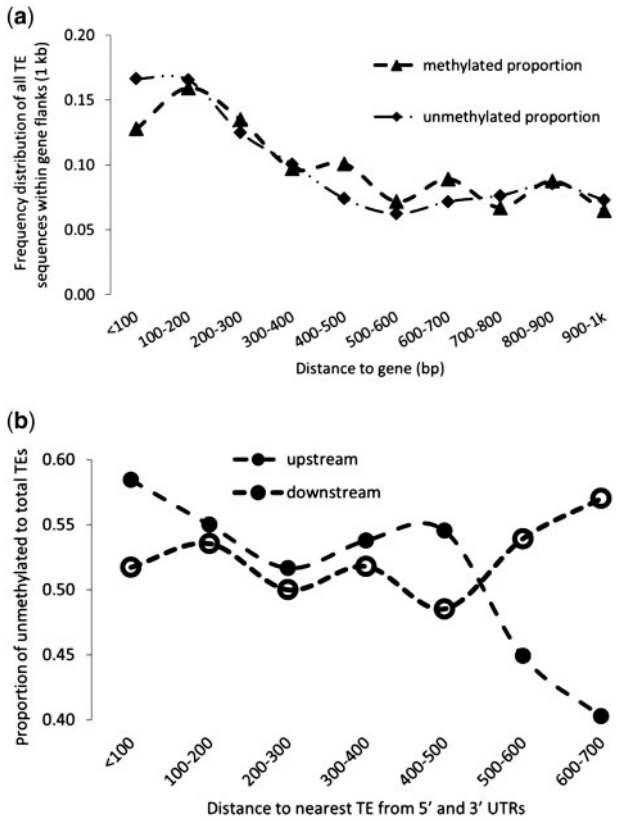


Figure 8. Distance between TE sequences and genes in euchromatin. (**a**) Both methylated and unmethylated TE sequences tend to accumulate close to genes. Note that because results do not substantially differ for the 5′- and 3′-ends of genes, they are not distinguished in the figure. (**b**) The proportion of unmethylated to total TE sequences drops slightly farther away from the 5′-end of genes. No similar drop is observed from the 3′-end of genes. Only the TE sequence closest to the start or stop codon was considered for this analysis.

DISCUSSION

Using a refined version of our previous annotation pipeline, we have obtained the most extensive dataset for TE sequences in the Arabidopsis genome to date. Given the high sensitivity and specificity of this pipeline, this dataset is not expected to evolve substantially in the future. Similarly, the use of stringent criteria for the analysis of DNA methylation of TE sequences makes our conclusions particularly robust, and should facilitate comparison of DNA methylation patterns between different conditions as well as between Arabidopsis accessions. Furthermore, the methods used to determine the DNA methylation status of individual TE sequences based on bisulphite-sequencing data are general and can be implemented to the systematic analysis of the association between DNA methylation and any annotated features of genomes for which single-nucleotide resolution methylomes are available (29).

Based on the 17 085 TE sequences (out of a total of 31 245) for which DNA methylation could be examined with high precision, we have found that 26% are unmethylated and another 15% have methylation patterns that depart significantly from the dense CG, CHG and CHH methylation typically reported. These two categories of TE sequences mainly correspond to short and highly degenerate relics located in euchromatin, many of which are missed by less sensitive detection pipelines. These relics

also tend to be depleted in CG sites, suggesting an important function for these sites in determining DNA methylation density. Thus we can propose a scenario in which TE sequences progressively lose CG sites because of their higher methylation levels compared to CHG and CHH sites and because of the higher mutability of methylcytosines compared to cytosines. This progressive loss would in turn reduce the potential for the affected sequences to perpetuate methylation at CHG and CHH sites, leading ultimately to complete loss of DNA methylation.

While our analysis confirms the association of siRNAs with DNA methylation over TE sequences, an unexpectedly high number of densely methylated TE sequences are characterized by the absence or near absence of matching siRNAs. Such TE sequences are preferentially found in heterochromatin. We have shown that in these cases, DNA methylation most likely results from local spreading (within 500 bp) from flanking, siRNA-targeted sequences (Figure 6c). We have also provided evidence that DNA methylation spread occurs in euchromatin as well, but that the extent of spreading is more limited than in heterochromatin (~200 bp versus ~500 bp; Figure 6c and Supplementary Figure S5). This could reflect either a facilitating role of heterochromatin, an inhibitory effect of euchromatin, or, as reported previously (30), a higher DNA demethylation activity in euchromatin. The spreading phenomenon we have uncovered here appears distinct from so-called secondary RdDM, which is caused by the biogenesis of secondary siRNAs from sequences adjoining those initially targeted by RdDM (31–34). Furthermore, the persistence of DNA methylation gradients for CG and CHH sites in the *ddc* triple mutant background (Figure 7) argues against an important role for RdDM in DNA methylation spreading. However, we cannot rule out that RdDM is involved, notably during the reproductive phase when it is most active (3), and that spreading of DNA methylation is maintained by MET1 and/or other DNA methyltransferases independently of RdDM during plant growth.

Finally, our study indicates that TE sequences present in euchromatin are more abundant closer to genes than away from them. This pattern is observed both upstream and downstream of genes, which could reflect a preference for TEs to insert in 'open' chromatin. Indeed, preferential insertion close to or within genes has been noted for several TE families in maize and rice (35), even though such events are unlikely to be maintained over evolutionary timescales because of their high potential to be deleterious. We have also shown that methylated TE sequences are slightly underrepresented compared to their unmethylated counterparts close to the 5′-end of genes and that methylated TE sequences with matching siRNAs are least abundant and somewhat more uniformly distributed within the 5′-end of genes than methylated sequences with no matching siRNAs (Supplementary Table S2). Given the known inhibitory effect of DNA methylation on promoter activity, it is therefore reasonable to speculate that DNA methylation spread contributes significantly to the negative impact of methylated TE sequences on neighboring gene expression. In support of this view, both in *A. thaliana* and *A. lyrata*, genes that are located <500 bp away from TE sequences tend to be expressed at lower levels than genes further away, and this reduction in gene expression is more pronounced when the TE sequences have matching siRNAs (8).

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank David Swarbreck for valuable insights on TE annotation, Eva Huala for helping us to access The Arabidopsis Information Ressource, François Roudier for critical reading of the manuscript and members of the Colot group for discussions.

FUNDING

Agence Nationale de la Recherche (ANR) 'DDB1 project' (to C.B. and V.C., in part); Centre National de la Recherche Scientifique (CNRS) 'Groupement de Recherche Elements Transposables' (to V.C and H.Q., in part). PhD studentships from ANR and CNRS (I.A. and A.S., respectively). Funding for open access charges: CNRS.

Conflict of interest statement. None declared.

REFERENCES

1. Slotkin,R.K. and Martienssen,R. (2007) Transposable elements and the epigenetic regulation of the genome. *Nat. Rev. Genet.*, **8**, 272–285.
2. Law,J.A. and Jacobsen,S.E. (2010) Establishing, maintaining and modifying DNA methylation patterns in plants and animals. *Nat. Rev. Genet.*, **11**, 204–220.
3. Teixeira,F.K. and Colot,V. (2010) Repeat elements and the Arabidopsis DNA methylation landscape. *Heredity*, **105**, 14–23.
4. The Arabidopsis Genome Initiative. (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature*, **408**, 796–815.
5. Cokus,S.J., Feng,S., Zhang,X., Chen,Z., Merriman,B., Haudenschild,C.D., Pradhan,S., Nelson,S.F., Pellegrini,M. and Jacobsen,S.E. (2008) Shotgun bisulphite sequencing of the Arabidopsis genome reveals DNA methylation patterning. *Nature*, **452**, 215–219.
6. Lister,R., O'Malley,R.C., Tonti-Filippini,J., Gregory,B.D., Berry,C.C., Millar,A.H. and Ecker,J.R. (2008) Highly integrated single-base resolution maps of the epigenome in Arabidopsis. *Cell*, **133**, 523–536.
7. Buisine,N., Quesneville,H. and Colot,V. (2008) Improved detection and annotation of transposable elements in sequenced genomes using multiple reference sequence sets. *Genomics*, **91**, 467–475.
8. Hollister,J.D., Smith,L.M., Guo,Y.L., Ott,F., Weigel,D. and Gaut,B.S. (2011) Transposable elements and small RNAs contribute to gene expression divergence between *Arabidopsis thaliana* and *Arabidopsis lyrata*. *Proc. Natl Acad. Sci. USA*, **108**, 2322–2327.
9. Quesneville,H., Bergman,C.M., Andrieu,O., Autard,D., Nouaud,D., Ashburner,M. and Anxolabehere,D. (2005) Combined evidence annotation of transposable elements in genome sequences. *PLoS Comput. Biol.*, **1**, 166–175.

10. Quesneville,H., Nouaud,D. and Anxolabehere,D. (2003) Detection of new transposable element families in *Drosophila melanogaster* and Anopheles gambiae genomes. *J. Mol. Evol.*, **57(Suppl. 1)**, S50–S59.
11. Smit,A.F.A., Hubley,R. and Green,P. (1996–2004) Institute for Systems Biology.
12. Jurka,J., Klonowski,P., Dagman,V. and Pelton,P. (1996) CENSOR–a program for identification and elimination of repetitive elements from DNA sequences. *Comput. Chem.*, **20**, 119–121.
13. Kohany,O., Gentles,A.J., Hankus,L. and Jurka,J. (2006) Annotation, submission and screening of repetitive elements in Repbase: RepbaseSubmitter and Censor. *BMC Bioinformatics*, **7**, 474.
14. Benson,G. (1999) Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.*, **27**, 573–580.
15. Kolpakov,R., Bana,G. and Kucherov,G. (2003) mreps: Efficient and flexible detection of tandem repeats in DNA. *Nucleic Acids Res.*, **31**, 3672–3678.
16. Eddy,S.R. (1998) Profile hidden Markov models. *Bioinformatics*, **14**, 755–763.
17. Bedell,J.A., Korf,I. and Gish,W. (2000) MaskerAid: a performance enhancement to RepeatMasker. *Bioinformatics*, **16**, 1040–1041.
18. Chao,K.M., Zhang,J., Ostell,J. and Miller,W. (1995) A local alignment tool for very long DNA sequences. *Comput. Appl. Biosci.*, **11**, 147–153.
19. Gusfield,D. (1997) *Algorithms on Strings, Trees, and Sequences: Computer Science and Computational Biology*. Cambridge University Press, New York, NY.
20. Kapitonov,V. and Jurka,J. (1996) The age of Alu subfamilies. *J. Mol. Evol.*, **42**, 59–65.
21. Lippman,Z., Gendrel,A.V., Black,M., Vaughn,M.W., Dedhia,N., McCombie,W.R., Lavine,K., Mittal,V., May,B., Kasschau,K.D. *et al.* (2004) Role of transposable elements in heterochromatin and epigenetic control. *Nature*, **430**, 471–476.
22. Zilberman,D., Gehring,M., Tran,R.K., Ballinger,T. and Henikoff,S. (2007) Genome-wide analysis of *Arabidopsis thaliana* DNA methylation uncovers an interdependence between methylation and transcription. *Nat. Genet.*, **39**, 61–69.
23. Kurtz,S., Phillippy,A., Delcher,A.L., Smoot,M., Shumway,M., Antonescu,C. and Salzberg,S.L. (2004) Versatile and open software for comparing large genomes. *Genome Biol.*, **5**, R12.
24. Fahlgren,N., Sullivan,C.M., Kasschau,K.D., Chapman,E.J., Cumbie,J.S., Montgomery,T.A., Gilbert,S.D., Dasenko,M., Backman,T.W., Givan,S.A. *et al.* (2009) Computational and analytical framework for small RNA profiling by high-throughput sequencing. *RNA*, **15**, 992–1002.
25. Miura,A., Nakamura,M., Inagaki,S., Kobayashi,A., Saze,H. and Kakutani,T. (2009) An Arabidopsis jmjC domain protein protects transcribed genes from DNA methylation at CHG sites. *EMBO J.*, **28**, 1078–1086.
26. Lister,R. and Ecker,J.R. (2009) Finding the fifth base: genome-wide sequencing of cytosine methylation. *Genome Res.*, **19**, 959–966.
27. Bernatavichute,Y.V., Zhang,X., Cokus,S., Pellegrini,M. and Jacobsen,S.E. (2008) Genome-wide association of histone H3 lysine nine methylation with CHG DNA methylation in *Arabidopsis thaliana*. *PLoS One*, **3**, e3156.
28. Hollister,J.D. and Gaut,B.S. (2009) Epigenetic silencing of transposable elements: a trade-off between reduced transposition and deleterious effects on neighboring gene expression. *Genome Res.*, **19**, 1419–1428.
29. Pelizzola,M. and Ecker,J.R. (2010) The DNA methylome. *FEBS Lett.*
30. Penterman,J., Zilberman,D., Huh,J.H., Ballinger,T., Henikoff,S. and Fischer,R.L. (2007) DNA demethylation in the Arabidopsis genome. *Proc. Natl Acad. Sci. USA*, **104**, 6752–6757.
31. Kanno,T., Bucher,E., Daxinger,L., Huettel,B., Bohmdorfer,G., Gregor,W., Kreil,D.P., Matzke,M. and Matzke,A.J. (2008) A structural-maintenance-of-chromosomes hinge domain-containing protein is required for RNA-directed DNA methylation. *Nat. Genet.*, **40**, 670–675.
32. Henderson,I.R. and Jacobsen,S.E. (2008) Tandem repeats upstream of the Arabidopsis endogene SDC recruit non-CG DNA methylation and initiate siRNA spreading. *Genes Dev.*, **22**, 1597–1606.
33. Daxinger,L., Kanno,T., Bucher,E., van der Winden,J., Naumann,U., Matzke,A.J. and Matzke,M. (2009) A stepwise pathway for biogenesis of 24-nt secondary siRNAs and spreading of DNA methylation. *EMBO J.*, **28**, 48–57.
34. Saze,H. and Kakutani,T. (2007) Heritable epigenetic mutation of a transposon-flanked Arabidopsis gene due to lack of the chromatin-remodeling factor DDM1. *EMBO J.*, **26**, 3641–3652.
35. Dooner,H.K. and Weil,C.F. (2007) Give-and-take: interactions between DNA transposons and their host plant genomes. *Curr. Opin. Genet. Dev.*, **17**, 486–492.

Supplementary Information

Supplementary Figure 1: Global CG, CHG and CHH methylation profiles for TE sequences and genes. (a,b,c) The x-axis represents the fraction of reads that are indicative of methylation at a given site and y-axis the fraction of sites with a given methylation level.

Supplementary Figure 2: Asymmetric nature of CHG methylation. (a,b) To test symmetry of CG and CHG methylation, CG and CHG sites that had adequate read depths on both the strands were analyzed for methylation status on the complementary strands. Four different levels were distinguished for CG methylation, i.e., No (0% reads methylated), Low (<10%), Moderate (>10% and <80%) and High (>80%). CHG sites were simply treated as methylated or unmethylated. For CG sites, No or Low methylation on one strand and High methylation on the other strand is considered as inconsistent, while for CHG sites inconsistency is considered when one strand is declared Methylated and the other Unmethylated. No significant inconsistency was observed for CG sites (green arrows). In contrast, a large number of CHG sites showed inconsistency of methylation patterns between the two strands (red arrow). (c) Analysis of methylation asymmetry in CHG sites using whole genome bisulphite sequencing data sets of Cokus et al. 2008 (#1) or Lister, et al. 2008 (#2) and considering four different levels of methylation for CHG sites rather than two {No (0% reads methylated), Low (<10%), Moderate (>10% and <20%) and High (>20%)}. The combinations that would suggest an inconsistency in methylation between the two strands for a CHG site are Low methylation on one strand and High on the other (Low-High), No on one and Moderate on the other (No-Mod) and No on one and High on the other (No-High). While very few Low-High combinations were found, a significant fraction of No-Mod and No-High sites were observed for both the bisulphite sequencing data sets mentioned above (#1 & #2). (d) CHG sites with at least 20 reads on each strand and 6 methylated

reads on one of the strands were taken to test the disproportionate levels of CHG methylation. A chi-square goodness of fit test was conducted to identify any significant differences between the methylation levels of two oppositely stranded cytosines and the p-values were adjusted for multiple testing using false discovery rate (FDR). A total of 30.6% (1062/3471) of the sites show significantly (p-value < 0.05) disproportionate methylation between the two strands. Red colour indicates p-value < 0.05 and green is p-value < 0.01. Furthermore it was also noticed that two or more neighbouring CHG sites tend to be more methylated on the same strand.

Supplementary Figure 3: DNA methylation spreading in met1, ddc and rdd mutant plants. (a) No significant methylation gradients are visible in the *met1* background (b,c) Persistence of DNA methylation spreading in *drm1drm2cmt3* (*ddc*) and *ros1dml2dml3* (*rdd*) triple mutants. Methylation densities were computed as described in Figure 6c.

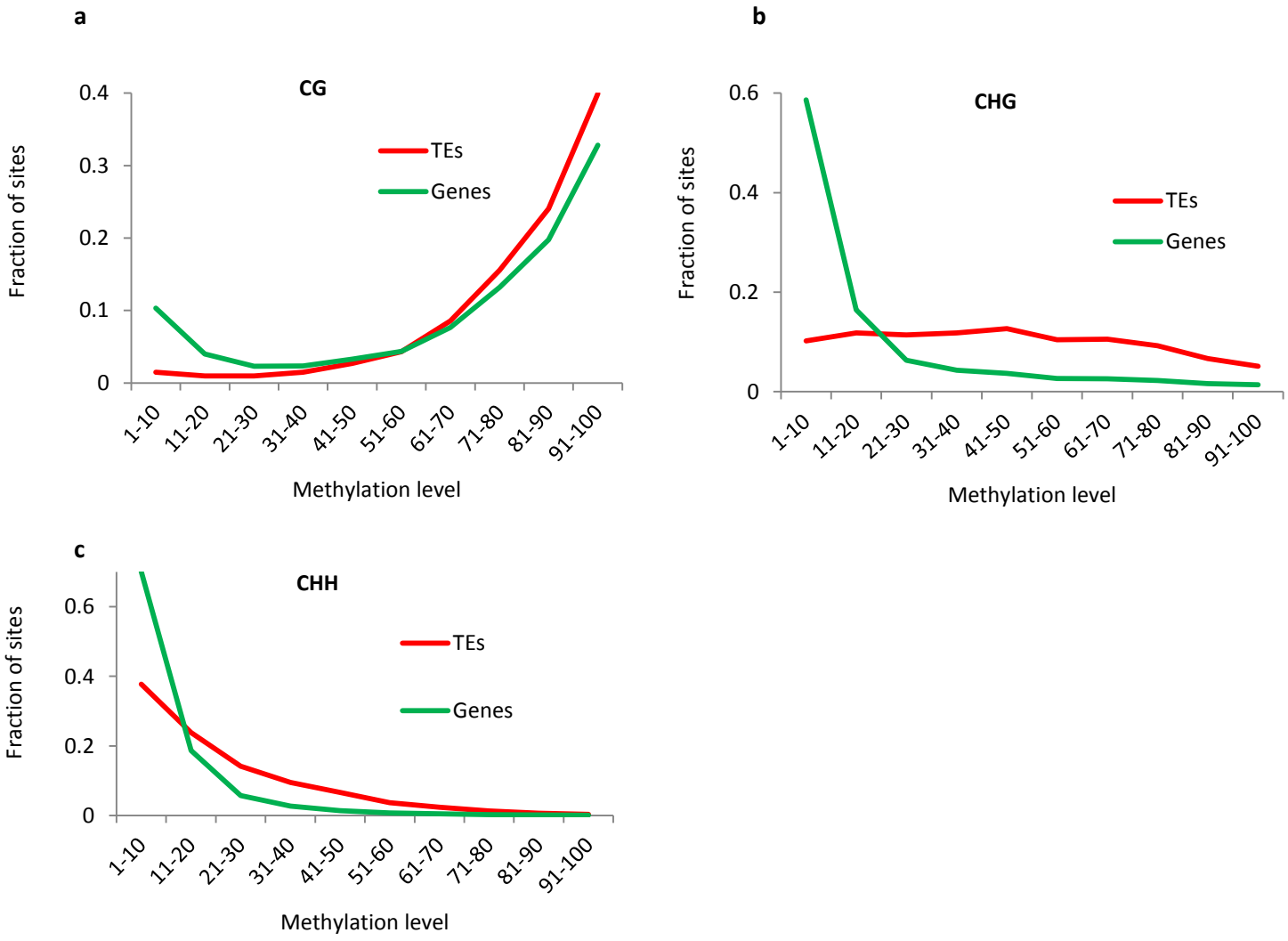
Supplementary Figure 4: Methylation densities for CG, CHG and CHH sites in *met1*. The *met1* mutant has no CG methylation neither in flanks nor in the TE sequence. However in *met1*, lower levels of CHG and CHH methylation in both flanks and TE sequence are still visible. There is no significant drop in CHG methylation from flanks to the TE sequence and only a marginal drop in CHH methylation.

Supplementary Figure 5: DNA methylation spread in euchromatin. Methylated euchromatic TEs associated with siRNAs and flanked by sequences not associated with siRNAs were considered. Given that many methylated euchromatic TE sequences have matching siRNAs only towards their middle, TE sequences were classified according to their association (+) or lack of association (-) with siRNAs within 500bp from their 5' and 3' extremities and in 100 bp windows. Each panel shows boxplots of DNA methylation densities along these 500 bp of TE

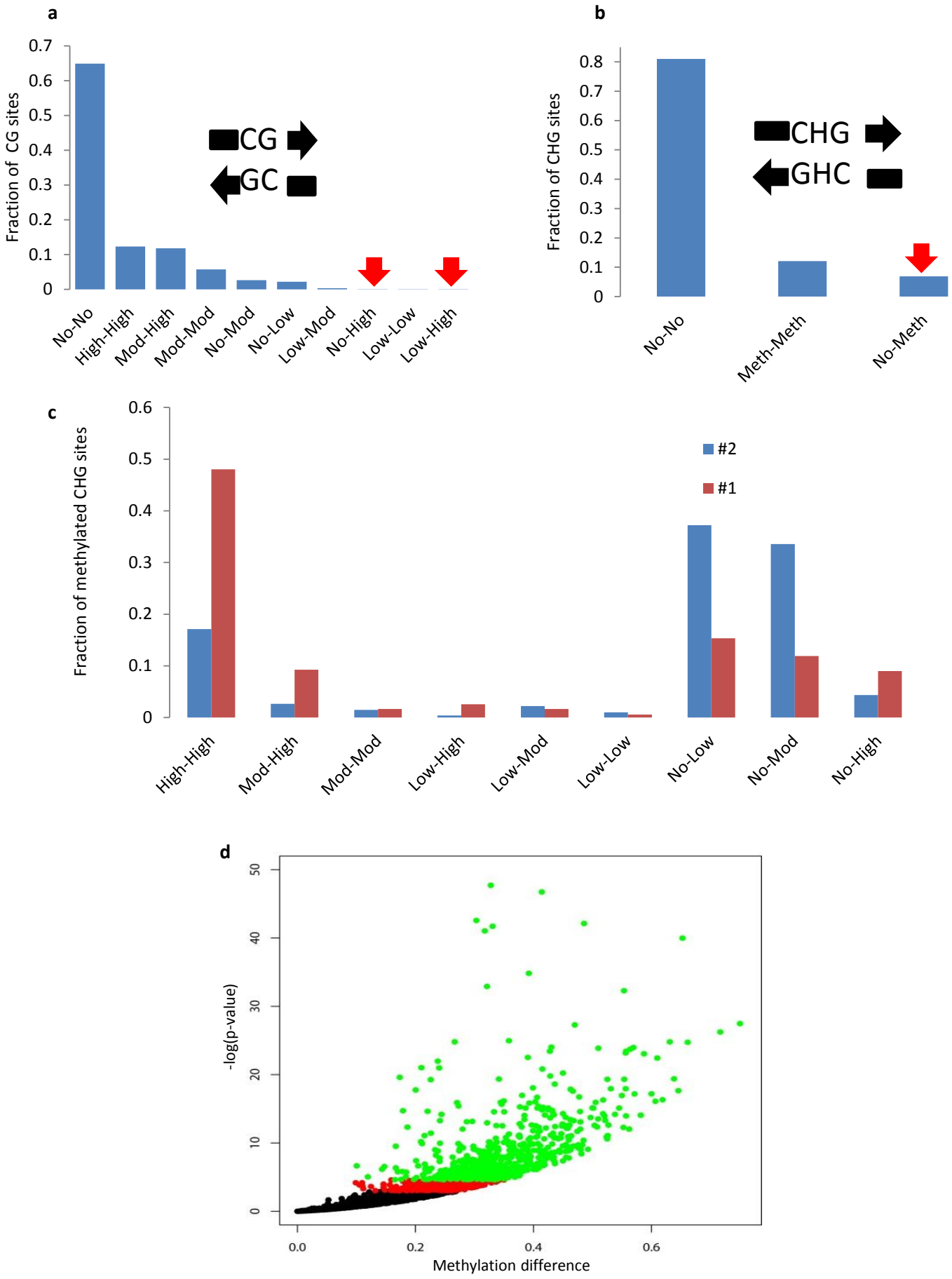
sequence as well for the 5' or 3' 1 kb flanks in 100 bp windows. DNA methylation spread extends on average over 200 bp from the last TE window associated with siRNAs.

Supplementary Figure 6: Quantifying DNA methylation over a given sequence. The Figure is an example of a 3.025 kb long ATHILA0_1 TE of LTR/Gypsy element which is methylated in all three sequence contexts. Each vertical bar in a track shows methylation status at the position. The height of the bar indicates sequence depth and colour shows methylation level with red colour indicating all reads methylated and light green all reads unmethylated for that cytosine. The red crosses in CG Plus and CG Minus tracks indicate positions where the sequence depth is too low to consider them and the arrows indicate CGs recovered from the other strand. Note that 9 CGs in the plus strand have sufficient sequencing depth whereas only two CGs in the minus strand are adequately covered, indicating a possible bias in the bisulphite sequencing technology. The blue horizontal lines in CHG and CHH tracks shows minimum (10) and maximum (50) read depths for a cytosine to be considered in the analysis. For each of the three types of sites, a given TE sequence is called as unmethylated (U) or methylated (M) and a pattern of three letters reflects methylation state for CG, CHG and CHH, respectively.

Supplementary Figure 1

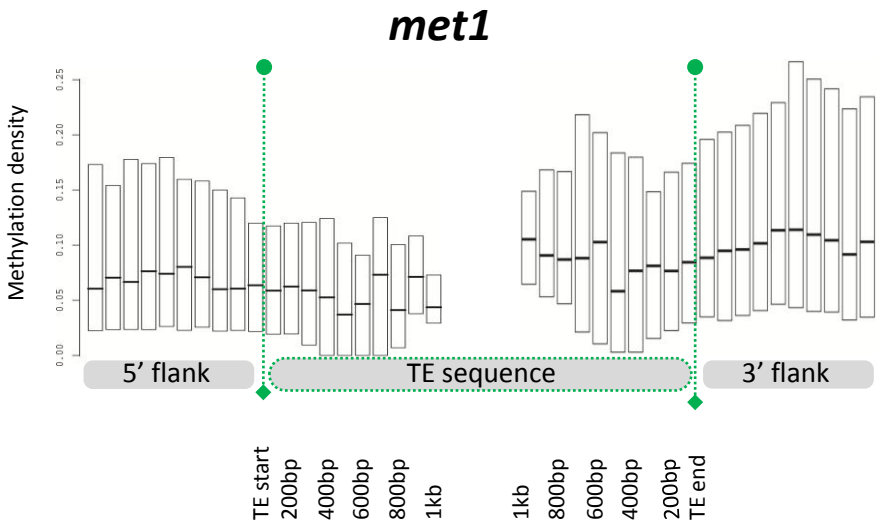


Supplementary Figure 2

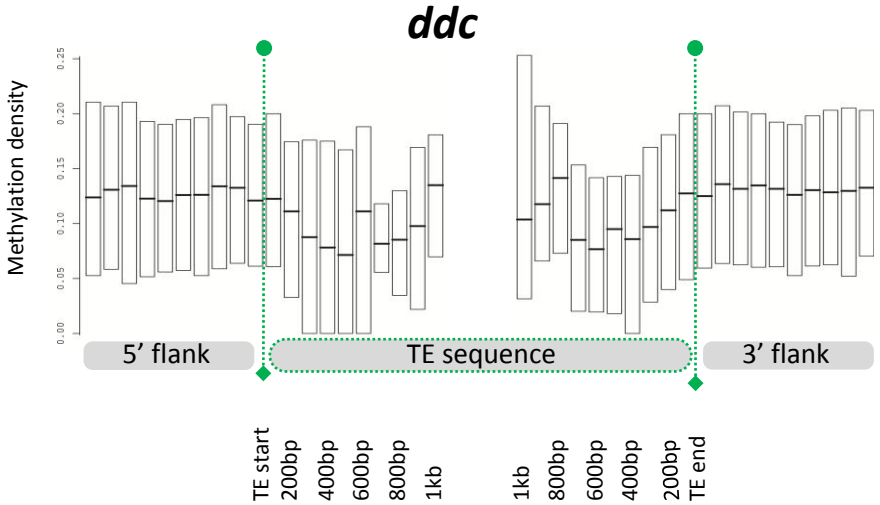


Supplementary Figure 3

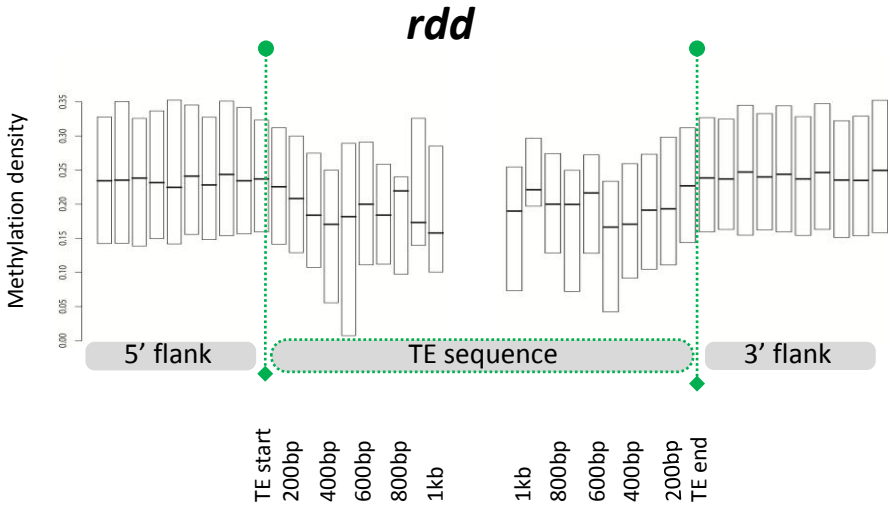
a



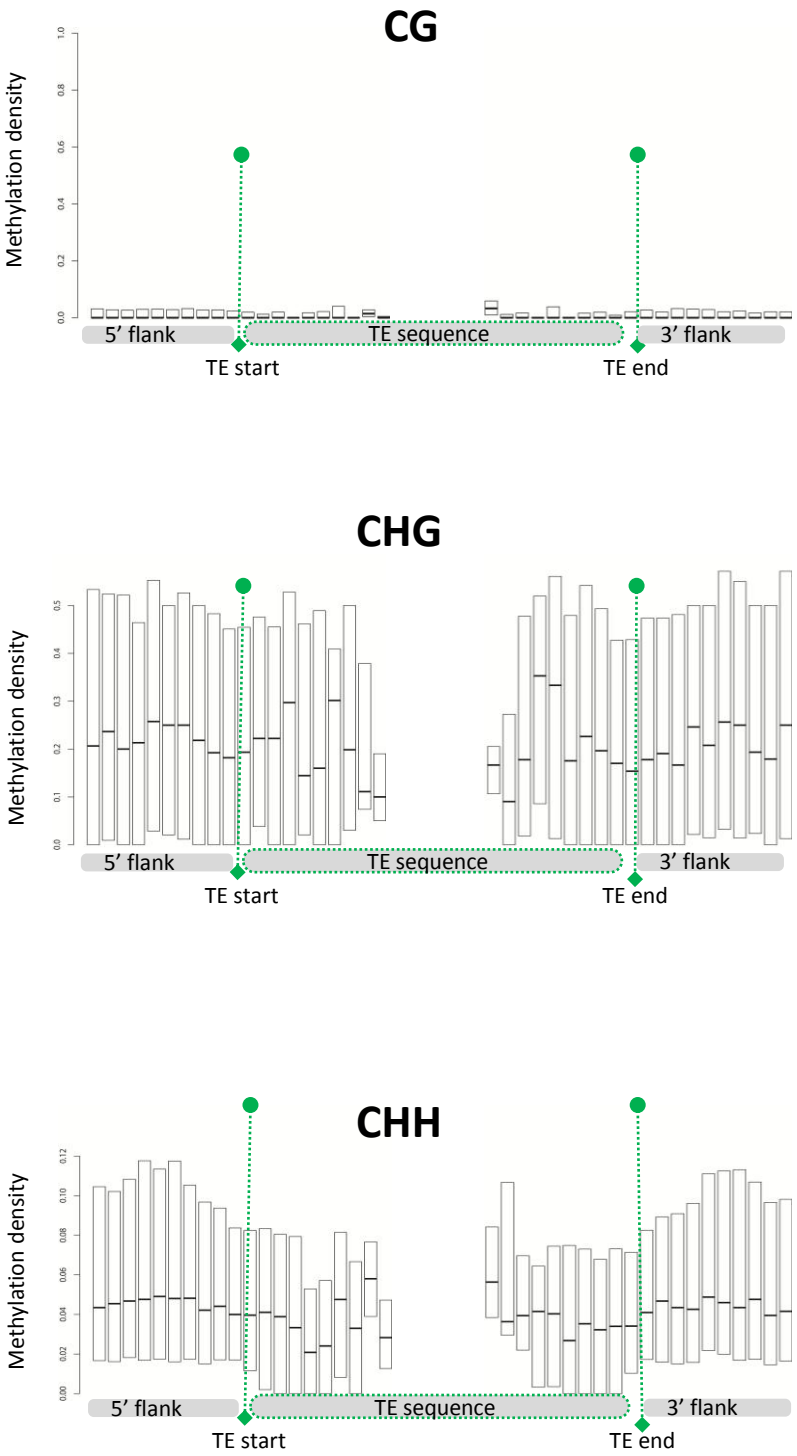
b



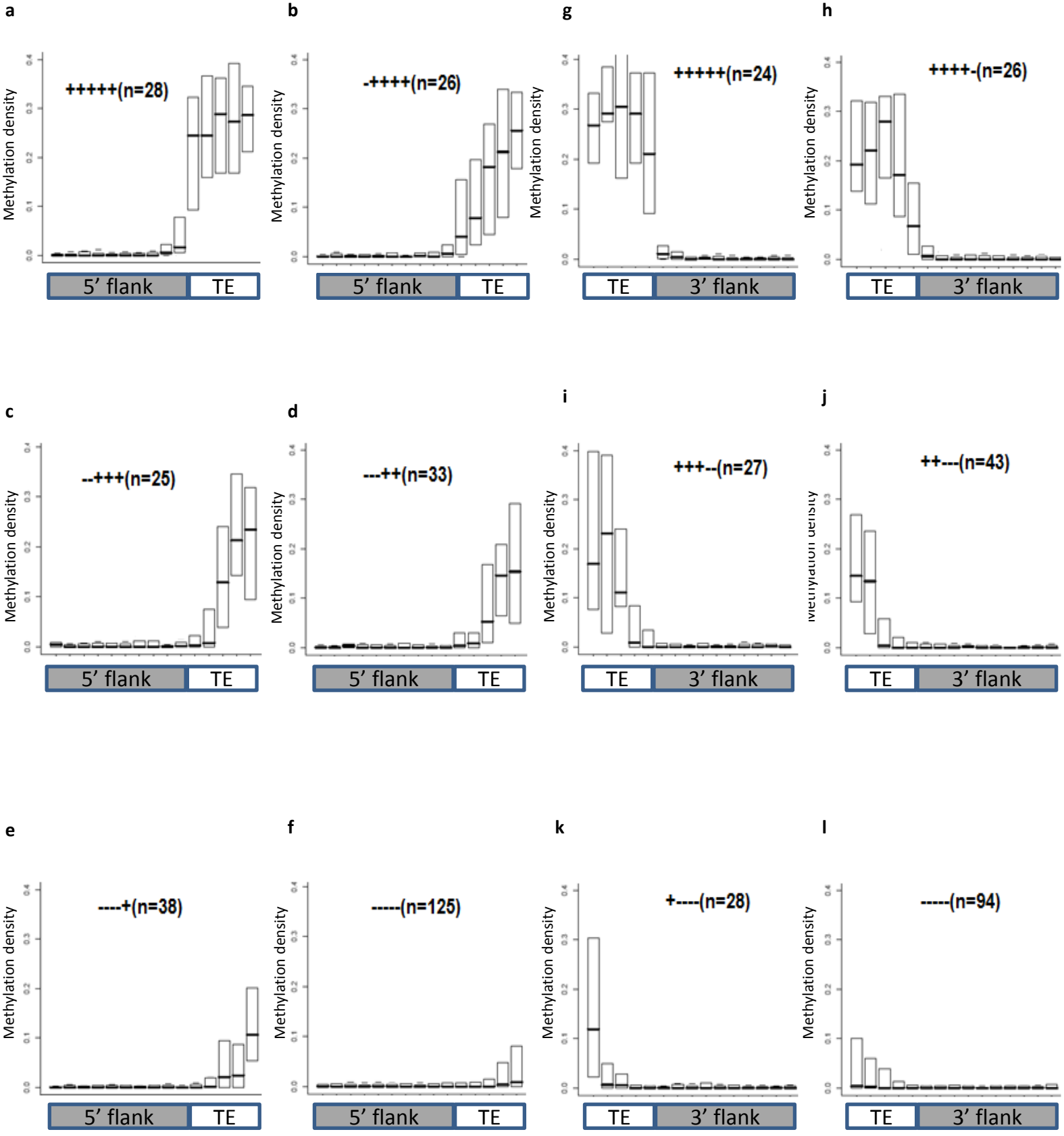
c



Supplementary Figure 4



Supplementary Figure 5



Supplementary Figure 6

$$Density = \frac{\sum C_m}{N}$$

C_m = methyl-cytosine

N = # cytosines in context having sufficient sequencing depths

CG = 100%
CHG = 100%
CHH = 51%
Pattern Assigned = **MMM**
(**M**ethylated CG, **M**ethylated CHG, **M**ethylated CHH)

	-UU	-MM	-UM	-MU	M-M	U-U	U-M	M-U	--M	--U
Nb	438	414	162	80	427	341	111	94	689	662
%	0.40	0.38	0.15	0.07	0.44	0.35	0.11	0.10	0.51	0.49
%Heterochromatin	0.28	0.66	0.38	0.53	0.68	0.31	0.25	0.59	0.62	0.40
%Euchromatin	0.72	0.34	0.62	0.48	0.32	0.69	0.75	0.41	0.38	0.60
Median size	86	77	134	53	76	71	89	77	47	36
Average size	153	130	169	116	120	108	127	116	72	61
	Median Frequency of methylated sites									
CG	NA	NA	NA	NA	1.00	0.00	0.00	1.00	NA	NA
CHG	0.00	1.00	0.00	1.00	NA	NA	NA	NA	NA	NA
CHH	0.00	0.39	0.12	0.00	0.38	0.00	0.13	0.00	0.33	0.00
	Median level of methylation (methylated reads/total reads)									
CG	NA	NA	NA	NA	0.78	0.00	0.00	0.43	NA	NA
CHG	0.00	0.31	0.00	0.08	NA	NA	NA	NA	NA	NA
CHH	0.00	0.06	0.01	0.00	0.05	0.00	0.01	0.00	0.04	0.00

Supplementary Table 1- *Methylation patterns for* 3,418 TE sequences devoid of CG and/or CHG sites

Distance to the nearest TE from gene	<100*	100-200	300-400	400-500	500-600	600-700	700-800	800-900	900-1k
Upstream									
Total	450	431	225	187	147	144	128	131	115
meth_siRNA	85	88	59	52	45	43	31	35	29
meth_no-siRNA	102	106	45	33	36	43	30	33	29
unmethylated	263	237	121	102	66	58	67	63	57
Downstream									
Total	203	183	139	136	115	107	110	97	82
methy_siRNA	54	45	34	37	23	24	24	23	20
meth_no-siRNA	44	40	33	33	30	22	24	31	22
unmethylated	105	98	72	66	62	61	62	43	40

* does not include TE sequences overlapping with gene annotation

Supplementary Table 2 *– Number of nearest TE sequences in 100 bp intervals 1 kb upstream and downstream of euchromatic genes. Methylated and associated with siRNAs (meth_siRNA); Methylated and non-siRNA associated (meth_no-siRNA) ; and unmethylated*

Method	HSP score threshold	Shuffled genome sequences		*A.thaliana* genome
		Number of HSP above the threshold	Fraction of the HSP above the threshold (*p-value*)	TE coverage
BLASTER	*no*	*all*	1	90Mb (~72%)
	47	22214	0.05	53Mb (~42%)
	69	485	0.001	34Mb (~26%)
	142	0	*Na*	21Mb (~17%)
RepeatMasker	*no*	*all*	1	18Mb (~14%)
	226	131	0.05	18Mb (~14%)
	270	0	*Na*	18Mb (~14%)
Censor	*no*	*all*	1	23Mb (~18%)
	250	294	0.05	20Mb(~16%)
	306	0	*Na*	19Mb (~15%)

Supplementary Table 3 *- False positive rate assessment computed from the number of high-scoring segment pairs (HSP) obtained on the shuffled genome sequences with the Repbase Update (RU) TE reference set for different HSP score threshold and the corresponding TE coverage on the A. thaliana genome sequence.*

Query	*BLASTER*	*Censor*	*RepeatMasker*
BLASTER	-	86.4	83
Censor	91.6	-	92.6
RepeatMasker	94.1	99	-

Supplementary Table 4 *- Base-pair percentage that overlaps between predictions of the TE detection programmes*

	RU	OptCoding	MaxSize	Opt
Deny long join	742	763	1012	874
Simple join	170	164	253	223
Too long join	7	11	12	11
Deny nest join	3	2	2	2
Nest join	0	0	0	0
Split	110	122	178	252

Supplementary Table 5 *- "Long join procedure" results.*

Supplementary file 3 -The values indicate fraction of sites showing a given methylation level

CG

Methylation Level (meth reads/unmeth reads)	Genes	Gene 3' 500bp flank	Gene 5' 500bp flank	TEs	Non-TE Insert	TE 3' 500bp flank	TE 5' 500bp flank
0	0.73528	0.91892	0.93589	0.17719	0.26631	0.57668	0.57710
1-10	0.02729	0.02785	0.02578	0.01208	0.01282	0.02573	0.02642
11-20	0.01056	0.00898	0.00708	0.00786	0.00693	0.01050	0.01157
21-30	0.00611	0.00371	0.00302	0.00809	0.00662	0.00731	0.00778
31-40	0.00624	0.00321	0.00234	0.01221	0.00777	0.00866	0.00895
41-50	0.00869	0.00316	0.00235	0.02221	0.01786	0.01257	0.01290
51-60	0.01148	0.00387	0.00217	0.03552	0.02868	0.01861	0.01851
61-70	0.02026	0.00532	0.00338	0.07051	0.06146	0.03566	0.03561
71-80	0.03491	0.00678	0.00427	0.12797	0.10789	0.06330	0.06254
81-90	0.05227	0.00777	0.00564	0.19783	0.18290	0.09343	0.09267
91-100	0.08691	0.01042	0.00806	0.32853	0.30077	0.14753	0.14595

CHG

Methylation Level (meth reads/unmeth reads)	Genes	Gene 3' 500bp flank	Gene 5' 500bp flank	TEs	Non-TE Insert	TE 3' 500bp flank	TE 5' 500bp flank
0	0.95143	0.94795	0.94918	0.25572	0.33612	0.64145	0.64031
1-10	0.02849	0.03074	0.02779	0.07583	0.07053	0.07011	0.07134
11-20	0.00799	0.00929	0.00861	0.08796	0.07618	0.05669	0.05836
21-30	0.00308	0.00344	0.00398	0.08510	0.07943	0.04862	0.04630
31-40	0.00210	0.00224	0.00241	0.08803	0.07828	0.04146	0.04344
41-50	0.00180	0.00191	0.00195	0.09443	0.08393	0.04075	0.03986
51-60	0.00130	0.00113	0.00134	0.07764	0.06865	0.02998	0.02850

61-70	0.00125	0.00116	0.00161	0.07876	0.06833	0.02599	0.02767
71-80	0.00109	0.00097	0.00114	0.06886	0.06195	0.02052	0.02087
81-90	0.00078	0.00059	0.00099	0.04974	0.04259	0.01412	0.01406
91-100	0.00069	0.00056	0.00100	0.03794	0.03401	0.01031	0.00929

CHH

Methylation Level (meth reads/unmeth reads)	Genes	Gene 3' 500bp flank	Gene 5' 500bp flank	TEs	Non-TE Insert	TE 3' 500bp flank	TE 5' 500bp flank
0	0.96328	0.96077	0.95983	0.68562	0.70028	0.83880	0.83477
1-10	0.02574	0.02818	0.02798	0.11871	0.11490	0.07967	0.08125
11-20	0.00684	0.00732	0.00766	0.07495	0.07216	0.03948	0.04055
21-30	0.00209	0.00213	0.00240	0.04439	0.04126	0.01868	0.01933
31-40	0.00098	0.00086	0.00105	0.02986	0.02833	0.01062	0.01086
41-50	0.00051	0.00040	0.00056	0.02086	0.01877	0.00634	0.00672
51-60	0.00025	0.00015	0.00022	0.01138	0.00997	0.00327	0.00320
61-70	0.00016	0.00010	0.00014	0.00741	0.00756	0.00171	0.00177
71-80	0.00008	0.00006	0.00008	0.00412	0.00424	0.00090	0.00093
81-90	0.00004	0.00001	0.00004	0.00184	0.00161	0.00036	0.00041
91-100	0.00002	0.00002	0.00003	0.00086	0.00091	0.00017	0.00021

Additional Methods

The whole genome shotgun bisulphite sequencing from Cokus et al. (Cokus, Feng et al. 2008) was obtained through personal communication. Whole genome shotgun bisulphite sequencing of wild type Arabidopsis plants (Columbia-0), and from *met1*, *drm1 drm2 cmt3* (*ddc*), and *ros1 dml2 dml3* (*rdd*) null mutants corresponding to GEO accession number GSE10966 were downloaded from Gene Expression Omnibus (http://www.ncbi.nlm.nih.gov/geo). The raw data for GSE10966 was downloaded as a tar file and eight text files belonging to plus and minus strands of each chromosome for the following four samples were extracted:

- GSM276809: Col-0_methylC-seq
- GSM276810: met1_methylC-seq
- GSM276811: ddc_methylC-seq
- GSM276812: rdd_methylC-seq

Each file was sorted based on the chromosome and start-position fields and separated into text files for each chromosome using the following commands:

```
#The first two arguments to the sort function (+0 and -1) are for the
#chromosome sort which is in the first column of the file. The third argument
#to the sort function (+2n) is to add second sort on the 3rd field and the
#type of this sort is numerical.
sort +0 -1 +2n *_plus.txt > *_plus1.txt
#remove the file
rm *_plus.txt
#rename file
mv *_plus.txt *_plus.txt
#Separate data for each chromosome into its respective text file
awk '/^1/{print}' ./GSM276809/GSM276809_plus.txt >
./GSM276809/GSM276809_plus_chr1.txt
```

The format of the above generated read sequence files is as follows:

chr	strand	start	end	sequence
1	+	72	120	ATATCTATGAATTTTTAAATATTTAATTTTTTAAATTCGAAATCGGTTT
1	+	109	154	CGAAATCGGTTTTTCTGGTTGAAAATTATTGTGTATATAATGATAA
1	+	116	158	GGTTTTTCTGGTTGTAAATTATTGTGTATATAATGTTAAGTTT
1	+	125	173	GGTTGAAAATTATTGTGTATATAATGATAATTTTATCGTTTTTATGTAA

```
1       +       144     188     TATAATGATAATTTTATCGTTTTTATGTAATTGTTTATTGTTGTG
1       +       149     195     TGATAATTTTATCGTTTTTATGTAATTGTTTATTGTTGTGTGTAGAT
1       +       165     219     TTTATGTAATTGTTTATTGTTGTGTGTAGATTTTTTAAAAATATTATTTGAGGTT
1       +       179     233     TATTGTTGTGTGTAGATTTTTTAAAAATATTATTTGAGGTTAATACACATTTTAT
1       +       187     241     TGTGTAGATTTTTTAAAAATATTATTTGAGGTCNATATAAATTCTATTTCTTGTG
1       +       212     260     TTGAGGTTAATATAAATTTTGTTTTTTGTGGTTTTTTTTTTTTTTTTTA
```

Sodium bisulphite converts unmethylated cytosines to uracil, but 5-methylcytosines remain unconverted. Hence, after amplification by polymerase chain reaction (PCR), unmethylated cytosines appear as thymines and methylated cytosines appear as cytosines. Lister et al. used ELAND module within the Illumina Genome Analyzer Pipeline Software to map the reads to the reference *A. thaliana* genome. ELAND aligns only 32 bases or shorter reads, allowing up to two mismatches to the reference sequence. For reads longer than 32 bases, the authors considered only the first 32 bases for the alignment, while the remaining sequence was still appended to the read regardless of similarity to the reference genome. For this reason, we only considered the first 32 bases in each read. The above files were converted into a General Genome Feature (GFF) format in a two-step process using two different custom designed Perl scripts. The first Perl programme “format_raw_seq.pl” takes as input the read sequence files as generated above for each chromosome and the chromosome fasta sequence file and generates forward and reverse strand distribution files (chr*_distribution_[forward[reverse]].txt). These distribution files are used by another Perl programme “gen_gff.pl” to convert them into GFF format files. These distribution files are then used by a range of other programmes in the downstream analysis.

The format_raw_seq.pl is run as follows and the parameters to the programme can be supplied in a POSIX style:

```
perl format_raw_seq.pl --seq_reads_plus read_sequence_file --seq_reads_minus
read_sequence_file --chr_num chr_number --chr_fasta chr_fasta_sequence_file -
-dir directory --help|-?
```

The command line parameters here are:

`--seq_reads_plus`	read sequence file for plus strand
`--seq_reads_minus`	read sequence file for minus strand
`--chr_num`	chromosome number (integer)
`--dir`	directory path (can be also. or ..)
`--help or -?`	Help message

```
#!c:\perl\bin\perl.exe
# format_raw_seq.pl
use Getopt::Long;
use Cwd;
my $dir= "./";
my($file_reads_plus,$file_reads_minus,$chr_fasta,$chr);
my ($help);
usage() && exit if (@ARGV < 1);
#-- prints usage if no command line parameters are passed or there is an unknown
#   parameter or help option is passed
GetOptions('help|?' => sub { usage(); exit;}, 'seq_reads_plus=s' => \$file_reads_plus,
'seq_reads_minus=s' => \$file_reads_minus,'chr_num=i' => \$chr,'chr_fasta=s' =>
\$chr_fasta,'dir=s' => \$dir);
#get the absolute path of current working directory
$dir= getcwd() if $dir=~/^\./;
#replace "\" in the windows path with "/"
$dir=~s/\\/\//g;
#add a "/" at the end of directory path if it is not there
$dir.="/" if ($dir!~/\/$/);
#declare global variables
my $start_searchp=0;
my $start_searchm=0;
my $flag=1;
my($frp,$frm);
open($frp,'<',"$dir$file_reads_plus") or (usage() && die "\n\n$!: Can't open file
$dir$file_reads_plus");
open($frm,'<',"$dir$file_reads_minus")or (usage() && die "\n\n$!: Can't open file
$dir$file_reads_minus");
open(FH,"$dir$chr_fasta") or (usage() && die "\n\n$!: Can't open file $dir$chr_fasta");
my @arr=<FH>;
close FH;
shift @arr if ($arr[0]=~/^>/);
chomp @arr;
open ($fh, '+>',"$dir$chr_fasta.tmp") or die "cant open temporary file $dir$chr_fasta.tmp";
print $fh join("",@arr);
seek  ($fh,0,0);
open (FHP,">$dir" . "Chr" . $chr . "_distribution_forward.txt");
open (FHM,">$dir" . "Chr" . $chr . "_distribution_reverse.txt");
```

```
my $pos=0;
my ($data,$reads,$context,$str);
while(read ($fh, $data,1))
{
        $reads="";
        $context="";
        $str="";
        if(uc($data) eq "C")
        {
          $str="plus";
          $context=getcontext($pos+1,$str,$fh);
          $reads=get_reads($str,$chr,$pos+1,$frp);
          if(($$reads[0] + $$reads[1])> 0)
                {
                $$reads[0]=0 unless($$reads[0]>0);
                $$reads[1]=0 unless($$reads[1]>0);
                print FHP ($pos+1,"\t$$reads[0]\t$$reads[1]\t",($$reads[0] +
                $$reads[1]),"\t$$context[0]\t$$context[1]\n") ;
                }
        }
        elsif(uc($data) eq "G")
        {
                $str="minus";
                $context=getcontext($pos+1,$str,$fh);
                $reads=get_reads($str,$chr,$pos+1,$frm);
                if(($$reads[0] + $$reads[1])> 0)
                {
                $$reads[0]=0 unless($$reads[0]>0);
                $$reads[1]=0 unless($$reads[1]>0);
                print FHM ($pos+1,"\t$$reads[0]\t$$reads[1]\t",($$reads[0] +
                $$reads[1]),"\t$$context[0]\t$$context[1]\n");
                }
        }
        #Display the progress after every 100,000 bases are processed
        if ($pos % 100000==0)
        {
        print ("Done for chr$chr " , $pos+1,"\n");
        }
        $pos++;
}
close $fh;
close FHP;
close FHM;
unlink("$dir$chr_fasta.tmp");
print "completed for chr$chr\n";


############################SUBROUTINES####################################
#This subroutine displays programme usage
sub usage
{
  print "\n\nusage: perl format_raw_seq.pl [--seq_reads_plus read_sequence_file] [--
  seq_reads_minus read_sequence_file] [--chr_num chr_number] [--chr_fasta
  chr_fasta_sequence_file] [--dir directory] [--help|-?]\n\n\n";
  format STDOUT=
        ********************************************************************************
         compare_lists.pl: This programme takes aligned bisulphite reads for
         plus and minus strands of a chromosome along with the chromosome sequence
         in fasta format and generates distribution files for each cytosine that is
         covered by at least one read.

         Written by    ::Ikhlak Ahmed
        ********************************************************************************
        The Input file is a tab delimited text file with format as under:-

        chr     strand  start   end     sequence
        1       +       72      120     ATATCTATGAATTTTTAAATATTTAATTTTTTAAATTCGAAATCGG
        1       +       109     154     CGAAATCGGTTTTTCTGGTTGAAAATTATTGTGTATATAATGATAA
        1       +       116     158     GGTTTTTCTGGTTGTAAATTATTGTGTATATAATGTTAAGTTT
.
  write;}
```

```
#This subroutine takes position of the cytosine, its strand and file handler of the chromosome
#sequence and returns the context of the cytosine
sub getcontext()
{
        my $pos= shift;
        my $str= shift;
        my $fh=shift;
        my $curr_pos=tell($fh);
        my (@data);
        if($str eq 'plus')
        {
                seek $fh,$pos-1,0;
                read ($fh, $data[0],3);
                if ($data[0]=~/^CG/)
                {
                $data[1]="CG" ;
                }
                elsif($data[0]=~/C[ACT]G/)
                {
                $data[1]="CHG";
                }
                elsif($data[0]=~/C[ACT]{2}/)
                {
                $data[1]="CHH";
                }
                else
                {
                $data[1]="ND";
                }
        }
        elsif($str eq 'minus')
        {
                seek($fh,$pos-3,0);
                read ($fh, $data[0],3);
                @tarr=split(//,$data[0]);
                @tarr=reverse(@tarr);
                $data[0]=join("",@tarr);
                $data[0]=~tr/ATCG/TAGC/;
                if ($data[0]=~/^CG/)
                {
                $data[1]="CG" ;
                }
                elsif($data[0]=~/C[ACT]G/)
                {
                $data[1]="CHG";
                }
                elsif($data[0]=~/C[ACT]{2}/)
                {
                $data[1]="CHH";
                }
                else
                {
                $data[1]="ND";
                }
        }
        seek ($fh,$curr_pos,0);
        return (\@data);
}
#This subroutine takes chromosome, position, strand of the cytosine and file handler of the
sequence_read file  and returns number of reads manifesting methylation and number of reads
manifesting unmethylation for that cytosine within the first 32 bp of the read
sub get_reads
{
        my $strand= shift;
        my $chr= shift;
        my $pos=shift;
        my $fh=shift;
        my @meth_unmeth;
        my @arr="";
        $start_search=$start_searchp if($strand eq "plus");
        $start_search=$start_searchm if($strand eq "minus");
```

```
        seek ($fh, $start_search,0);
        $flag=1;
        while(chomp($line=<$fh>))
        {
                @arr="";
                @arr= split(/\t/,$line);
                #To search entire read regardless of the length of the read uncomment the below if
                #condition and comment the next one
                #if($arr[0]==$chr and ($pos >= $arr[2] and $pos <= $arr[3]))
                #only search the 1st 32 bases of the read
                if($arr[0]==$chr and ($pos >= $arr[2] and $pos <= ($arr[2]+31)))
                {
                        if($flag==1)
                        {
                        $start_searchp=tell($fh)-(length($line)+1)if ($strand eq 'plus');
                        $start_searchm=tell($fh)-(length($line)+1)if ($strand eq 'minus');
                        $flag=0;
                        }
                        #print "$chr\t$strand\t$pos";;
                        #print"\t$arr[0]\t$arr[2]\t$arr[3]";
                        $residue=substr($arr[4],$pos-$arr[2],1);
                        #print "\t$residue\n";
                        if($residue eq 'C')
                        {
                        $meth_unmeth[0]++;
                        }
                        elsif($residue eq 'T')
                        {
                        $meth_unmeth[1]++;
                        }
                }
                last if ( $arr[0]==$chr and $pos < $arr[2]);
        }
        return \@meth_unmeth;
}
```

The format of the distribution files as generated by format_raw_seq.pl is as follows:

cytosine_position	meth_reads	unmeth_reads	depth	nucleotide_triplet	context
73	0	1	1	CAT	CHH
76	1	0	1	CCA	CHH
77	0	1	1	CAT	CHH
84	0	1	1	CCC	CHH
85	0	1	1	CCT	CHH
86	0	1	1	CTA	CHH
93	0	1	1	CCT	CHH
94	0	1	1	CTA	CHH
100	0	1	1	CCC	CHH
101	0	1	1	CCT	CHH

To generate gff files from distribution files, gen_gff.pl programme was run with the following parameters. (The order of the parameters is immaterial):

```
perl gen_gff.pl --chr chr_number --strand plus|minus --dir directory --file
distribution_file_name --type type_to_be_appended --source source --help|-?
```

The command line parameters here are:

`--chr`	chromosome number or name
`--strand`	chromosome strand (plus or minus)
`--file`	the name of the distribution file
`--dir`	directory path (can be also. or ..)
`--type`	this parameter is used as an additional tag for type field of the gff file. By default the type is cytosine “context_strand” (e.g. CG_forward)
`--help or -?`	Help message

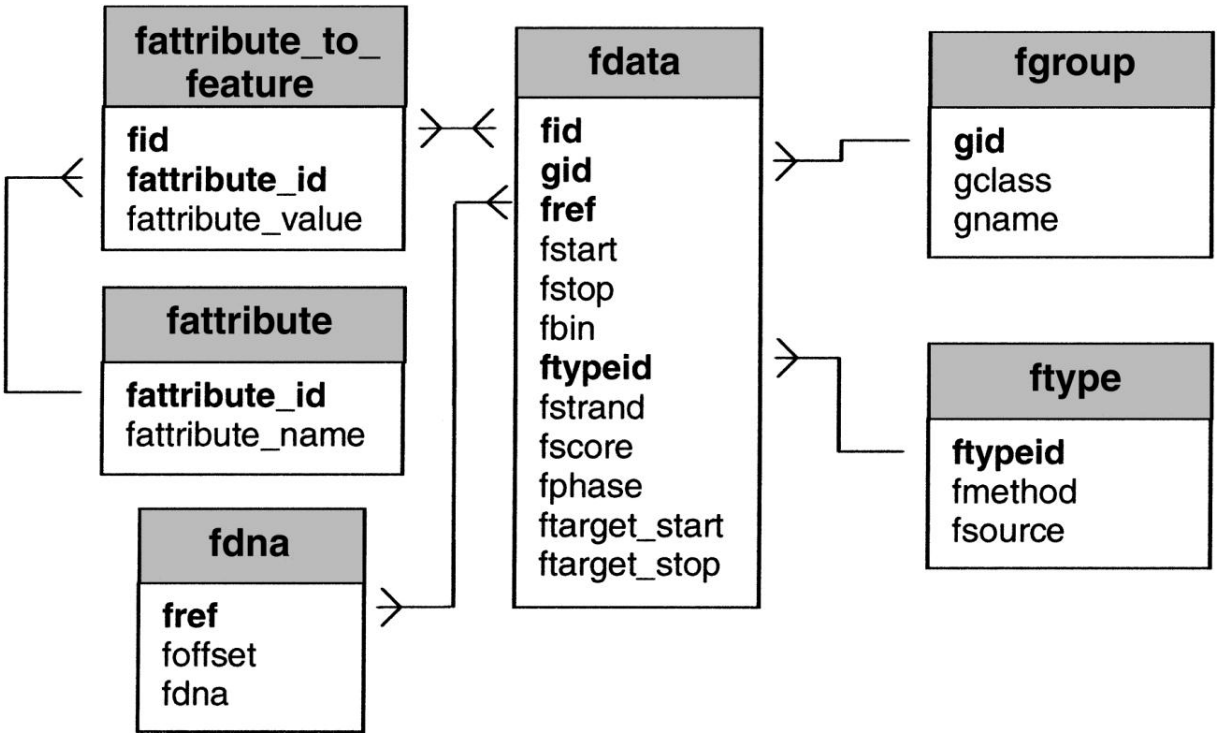
```
#!/usr/bin/perl
#gen_gff.pl
use Getopt::Long;
use Tie::File;
use Cwd;
my ($dir,$chr,$str,$file,$type,$source,$help);
$source="lister";
usage() && exit if (@ARGV < 1);
#-- prints usage if no command line parameters are passed or there is an unknown
#   parameter or help option is passed
GetOptions('help|?' => sub { usage(); exit;}, 'dir=s' => \$dir, 'chr=s' => \$chr,'strand=s' =>
\$str,'file=s' => \$file,'type=s' => \$type,"source=s"  => \$source);
#This subroutine displays programme usage
sub usage
{
  print "\n\nusage: perl gen_gff.pl [--chr chr_number] [--strand plus|minus] [--dir directory] [-
  -file distribution_file_name] [--type type_to_be_appended] [--source source] [--help|-
  ?]\n\n\n";
}
#get the absolute path of current working directory
$dir= getcwd() if $dir=~/^\./;
#replace "\" in the windows path with "/"
$dir=~s/\\/\//g;
#add a "/" at the end of directory path if it is not there
$dir.="/" if ($dir!~/\/$/);
$chr=ucfirst($chr);
open (FHW , ">" . $dir  . $chr ."_". $str ."_annot.gff") or (usage() && die "\n\n$!: Can't open
output file");
print "processing starts for $file.....................\n";
tie @array, 'Tie::File' , "$dir$file" or (usage() && die "\n\n$!: Can't open file $dir$file");
foreach  $id(0..$#array)
{
        @tarr=split(/\s+/,$array[$id]);
        $con=$tarr[5] ."_" . $str ."_" . $type;
        $pos=$tarr[0];
```

```
        $height=$tarr[3];
        #The height attribute is set to the read depth if depth is less than 100 reads for a
        #given cytosine otherwise height is incremented by 10 for every 1000 reads above 100
        $height=100 + sprintf("%.f",(($height-100)/1000)*10) if ($height >100);
        $val=sprintf("%.2f" ,$tarr[1]/$tarr[3]);
        $group="ID  " . $pos ." ". $tarr[1] . "/" . $tarr[3] . " ; Height " . $height . ";
        Meth_Unmeth \"" . $tarr[1] . "/" . $tarr[2] . "\" ; Ratio $val";
        print FHW "$chr\t$source\t$con\t$pos\t$pos\t$val\t+\t.\t$group\n";
}
close FHW;
```

Data visualization and downstream analysis

The GFF format files generated represent a standard file format for storing genomic features in a text file. These files can be uploaded to a genome viewer called Generic Genome Browser (GBrowse) from GMOD (http://www.gmod.org/wiki/GBrowse). GBrowse, which is Perl-based, is a combination of database and interactive web pages for manipulating and displaying annotations on a genome and uses several adaptors (Perl Modules) for interacting with various kinds of databases and different types of data sources. We used a Bio::DB::GFF adaptor and MySQL (http://www.mysql.com/) relational database server for our GBrowse implementation. The Bio::DB::GFF database uses a minimal schema to represent features on sequences. The main tables are fdata, which contains the position and type of each feature, fgroup, which tracks the grouping of subfeatures into features, fdna, which stores the raw DNA sequence, and fattribute_to_feature, which allows attribute information to be attached to features. The fattribute and ftype tables, respectively, hold attribute names and the method and source fields. For retrieval efficiency, the fdna table fragments each DNA into small pieces and stores the beginning of each piece in the foffset field.

Flowchart for Bio::DB::GFF database schema



Stein L D et al. Genome Res. 2002;12:1599-1610

The BioPerl (http://www.bioperl.org) module Bio::DB::GFF provides several methods and objects that besides giving a fast and indexed access to a sequence annotation database also provides methods to load GFF files into a database and multiple schemas are supported through a system of adaptors and aggregators. However, to provide an easy access to these methods, the BioPerl also offers several Perl scripts that make it seamless to work with GFF files and upload them into a relevant database schema. One such BioPerl script, bp_load_gff.pl can both create the database as well as incrementally load new data into a Bio::DB::GFF database. This script was launched with the following command line arguments:

```
bp_load_gff.pl --dsn 'dbi:mysql:database=databasename;host=localhost;
port=3306' --adaptor dbi::mysql --user username --pass password -c chr1.fa
chr2.fa …
```

```
bp_load_gff.pl --dsn 'dbi:mysql:database=databasename;host=localhost;
port=3306' --adaptor dbi::mysql --maxfeature 1000000000 --user username --
pass password chr1forward_annot.gff chr1reverse_annot.gff …
```

The various command-line options that bp_load_gff.pl script can take are:

`--dsn`	Data source (default dbi:mysql:test)
`--adaptor`	Schema adaptor (default dbi::mysqlopt)
`--user`	Username for mysql authentication
`--pass`	Password for mysql authentication
`--fasta`	Fasta file or directory containing fasta files for the DNA
`--create`	Force creation and initialization of database
`--maxfeature`	Set the value of the maximum feature size (default 100 Mb)
`--group`	A list of one or more tag names (comma or space separated) to be used for grouping in the 9th column.
`--upgrade`	Upgrade existing database to current schema
`--gff3_munge`	Activate GFF3 name munging (see Bio::DB::GFF)
`--quiet`	No progress reports

Other BioPerl scripts that can be used with the Bio::DB::GFF database schema are bp_bulk_load_gff.pl and bp_fast_load_gff.pl. To get help documentation for the usage of these scripts, one can use the command line Perl function “perldoc” e.g., perldoc bp_bulk_load_gff.pl. GBrowse can also work with an advanced version of the Bio::DB::GFF database schema called Bio::DB::SeqFeature::Strore that I have used for the GBrowse implementation of the genome-

wide MeDIP and ChIP-chip datasets discussed in Chapters III & IV. The BioPerl script that works with Bio::DB::SeqFeature::Strore module is bp_seqfeature_load.pl and can be launched to upload GFF files to a MySQL database as follows:

```
bp_seqfeature_load.pl --dsn 'dbi:mysql:database=databasename;host=localhost;
port=3306' --adaptor DBI::mysql --user username --password password -c
chr1.fa chr2.fa …
```

```
bp_seqfeature_load.pl --dsn 'dbi:mysql:database=databasename;host=localhost;
port=3306' --adaptor DBI::mysql --user username --password password gff_files
```

The command-line options for bp_seqfeature_load.pl programme are explained as below:

`-d, --dsn`	DBI data source (default dbi:mysql:test)
`-a, --adaptor`	The storage adaptor (class) to use (default DBI::mysql)
`-c, --create`	Create the database and reinitialize it (default false) Note, this will erase previous database contents, if any.
`-u, --user`	User to connect to database as
`-p, --password`	Password to use to connect to database

The APIs provided in the BioPerl modules Bio::DB::GFF and Bio::DB::SeqFeature::Store can be used to retrieve the sequence and annotation data from the database. The Perl code below shows how to interact with the GFF databases via API methods provided by these modules:

Using Bio::DB::GFF Module

```
#connect to a MySQL Bio::DB::GFF database from within a perl programme
use Bio::DB::GFF;
my $db      = Bio::DB::GFF->new( -adaptor => 'dbi::mysql',-dsn =>
              'dbi:mysql:database=databasename;host=localhost',-user
              =>'username',-pass => 'password');
#Define feature types that you are interested in. A feature type is a
combination of third and second fields in a GFF file separated by a colon
my @types= ('CG_forward:source1','CG_reverse:source1','CG_forward:source2',
            'CG_reverse:source2');
```

```
#Define a segment by specifying chromosome and start and stop coordinates.
The segment contains all feature types present within that region
my $segment = $db->segment('Chr1',1=>100000);
#Retreive from the database the feature types that you are interested in.
Here @features is an array containing references of objects of class
Bio::DB::GFF::Feature. A feature object is a stretch of sequence that
corresponding to a single annotation in a GFF database
my @features = $segment->features(-types => \@types);
#Iterate over each of the feature objects and print it's ID, type of the
feature and the attributes (methylated_reads/unmethylated_reads)
foreach my $feature (@features)
 {
      print $feature->display_id ,"\t",$feature->type,"\t",
      $feature->attributes('Meth_Unmeth'),"\n";
 }
#Delete the specified type of features from the database
$success=$db->delete(-type=>[@types]);
```

Using Bio::DB::SeqFeature::Store Module

```
#connect to a MySQL Bio::DB::SeqFeature::Store database from within a perl
programme
use Bio::DB::SeqFeature::Store;
my $db      = Bio::DB::SeqFeature::Store->new( -adaptor => 'DBI::mysql',
              -dsn => 'dbi:mysql:databasename',-user =>'username',-pass =>
               'password');
#Define feature types
my @types= ('CG_forward:source1','CG_reverse:source1','CG_forward:source2',
            'CG_reverse:source2');
#delete these feature types from all chromosomes
for $i(1..5)
{
@features = $db->features(-type => \@types,-seq_id => 'Chr'.$i,-start  => 1,
                          -end => 3043300,);
$success = $db->delete(@features);
print ($success,"\t done for chromosome $i\n\n");
}
```

GBrowse uses a configuration file that usually resides in gbrowse.conf directory and through it one can customize the display of a track, which is defined as a horizontal band containing a number of graphical elements called "glyphs" that correspond to a given sequence feature. A glyph indicates how each feature is to be rendered and a variety of glyphs are available, some of which are described below:

`arrow`	An arrow; can be unidirectional or bidirectional. It is also capable of displaying a scale with major and minor tickmarks, and can be oriented horizontally or vertically.

`box`	A filled rectangle.
`cds`	Draws CDS features, using the phase information to show the reading frame usage. At high magnifications draws the protein translation.
`dna`	At high magnification draws the DNA sequence. At low magnifications draws the GC content.
`dot`	A circle, useful for point features like SNPs, stop codons, or promoter elements.
`line`	A simple line.
`pinsertion`	A triangle designed to look like an insertion location (e.g., a transposon insertion).
`processed_transcript`	multi-purpose representation of a spliced mRNA, including positions of UTRs
`primers`	Two inward pointing arrows connected by a line.
`redgreen_box`	A box that changes from green->yellow->red as the score of the feature increases from 0.0 to 1.0. Useful for representing microarray results.
`xyplot`	Histograms and other graphs plotted against the genome.
`whiskerplot`	Box and whisker plot for statistical data.

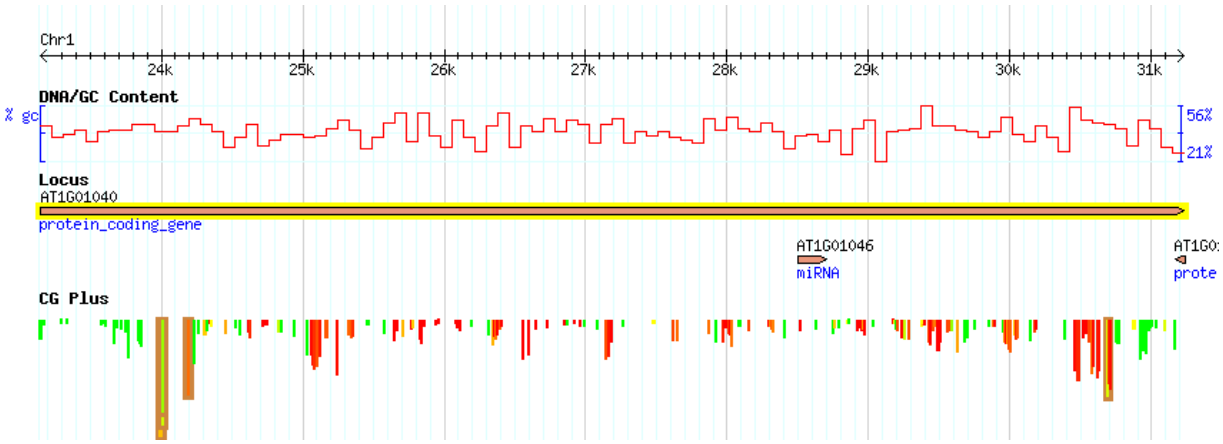
Below is an excerpt from our GBrowse configuration file that draws a track representing CG methylation of the plus strand and in this track each CG site is represented by a vertical bar. The height of the bar indicates read depth and colour shows methylation level (methylated reads/

total reads) with red colour indicating all reads methylated and light green all reads unmethylated for that cytosine.

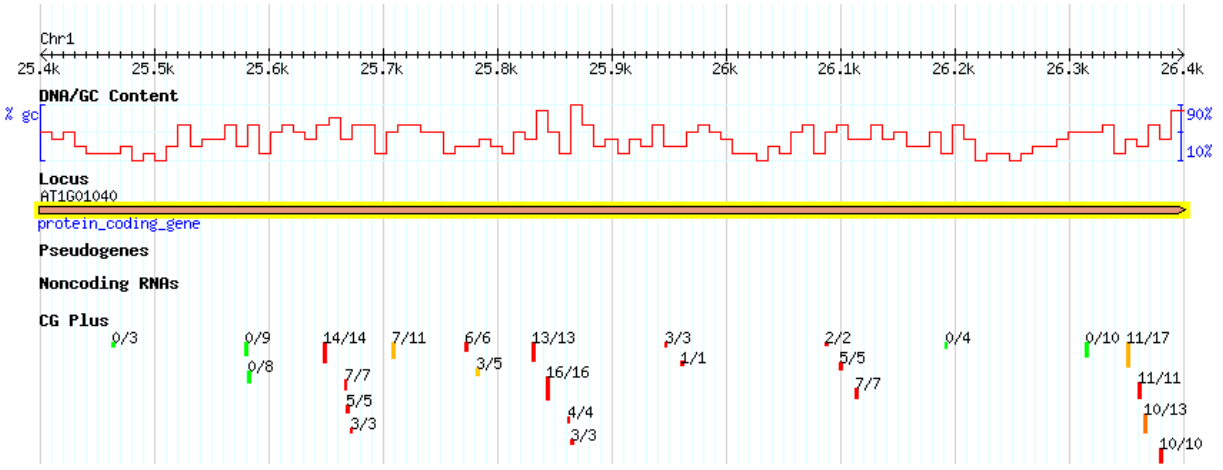
```
[CGPlus]
feature           = CG_forward
glyph             = redgreen_box
pad_top           = 10
color_subparts    = 1
#The height attribute in the GFF file was set to read depth if depth is less
#than 100 reads for a given cytosine otherwise height was incremented by 10
#for every 1000 reads above 100. The following subroutine highlights a CG bar
#if the read depth is greater than 50
hilite                  = sub
                              {
                              my $feature = shift;
                              my $height=$feature->attributes('Height');
                              return 1 if($height > 50);
                              return 0;
                              }
#returns the height for the bar which is estimation for the read depth at CG
#site
height            =   sub
                              {
                              my $feature = shift;
                              my $height=$feature->attributes('Height');
                              return $height;
                              }

linewidth         = 2


#The label at each vertical bar in the track and is given by methylated
#reads/total reads at the cytosine site
label             = sub
                              {
                              my $feature = shift;
                              my $m=$feature->attributes('Meth_Unmeth');
                              my @t=split(/\//,$m);
                              $m=$t[0] . "/" . ($t[0] + $t[1]);
                              return $m ;
                              }
category    = Whole-genome bisulphite
key         = CG Plus
```

A GBrowse screenshot showing CG methylation track as rendered by configuration settings shown above is displayed in Supplementary Figure 7.

Supplementary Figure 7: GBrowse screenshot showing CG methylation.



At low resolution



At high resolution

CHAPTER III

INTEGRATIVE EPIGENOMIC MAPPING DEFINES FOUR MAIN CHROMATIN STATES IN ARABIDOPSIS

Introduction

Chromatin demarcates the functional elements of a genome, such as genes, promoters and enhancers, as well as transposons and repeats through the effects of chemical modifications to DNA and histone proteins. These modifications can in turn either directly alter the chromatin structure and change its organization or can create an affinity for other chromatin associated proteins (e.g., HP1 or PcG proteins) that cause downstream alterations in chromatin structure and establishes a given chromatin state. Different chromatin states are thus marked by unique combinations of modified histones and proteins that organise the genome into domains with distinct functional properties. Chromatin ChIP-chip and ChIP-seq are powerful tools to characterise and determine the functional consequences of variations in chromatin composition along the gene. In the manuscript presented here, we reported the genome-wide mapping of a broad set of 11 histone marks and DNA methylation. Using combinatorial analysis based on these 12 epigenomic marks, we could show that only a limited number of combinations are actually used, in spite of the total combinatorial space potentially available. Further analysis using fuzzy *c-means* clustering partitioned the Arabidopsis epigenome into four major chromatin types with distinct functional properties. These chromatin types form short and highly interspersed domains along chromosome arms and index actively transcribed genes, chromatin repressed by PcG proteins, transposons and silent repeat elements in heterochromatin, and ambiguous chromatin mostly corresponding to intergenic regions. This first comprehensive view of the Arabidopsis epigenome suggests simple principals of organisation, similar to what has been reported in metazoans, and provides a resource to refine our understanding of the control of genome activity at the level of chromatin. My contribution

to this study was the design and development of the computational approaches and bioinformatics analyses required for data analysis and interpretation.

The EMBO Journal (2011), 1–11 | © 2011 European Molecular Biology Organization | Some Rights Reserved 0261-4189/11
www.embojournal.org

THE EMBO JOURNAL

EMBO open

Integrative epigenomic mapping defines four main chromatin states in Arabidopsis

François Roudier[1,*], Ikhlak Ahmed[1], Caroline Bérard[2], Alexis Sarazin[1], Tristan Mary-Huard[2], Sandra Cortijo[1], Daniel Bouyer[3], Erwann Caillieux[1], Evelyne Duvernois-Berthet[1], Liza Al-Shikhley[1], Laurène Giraut[4], Barbara Després[1], Stéphanie Drevensek[1], Frédy Barneche[1], Sandra Dèrozier[4], Véronique Brunaud[4], Sébastien Aubourg[4], Arp Schnittger[3], Chris Bowler[1], Marie-Laure Martin-Magniette[2,4], Stéphane Robin[2], Michel Caboche[4] and Vincent Colot[1,*]

[1]Institut de Biologie de l'Ecole Normale Supérieure, Centre National de la Recherche Scientifique (CNRS) UMR8197, Institut National de la Santé et de la Recherche Médicale (INSERM) U1024, Paris, France, [2]AgroParisTech/Institut National de la Recherhe Agronomique (INRA), UMR518 Mathématiques et Informatiques Appliquées, Paris, France, [3]Institut de Biologie Moléculaire des Plantes, CNRS UPR2357, Université de Strasbourg, Strasbourg, France and [4]Unité de Recherche en Génomique Végétale, UMR8114 INRA/CNRS/Université d'Evry Val d'Essonne (UEVE), Evry, France

Post-translational modification of histones and DNA methylation are important components of chromatin-level control of genome activity in eukaryotes. However, principles governing the combinatorial association of chromatin marks along the genome remain poorly understood. Here, we have generated epigenomic maps for eight histone modifications (H3K4me2 and 3, H3K27me1 and 2, H3K36me3, H3K56ac, H4K20me1 and H2Bub) in the model plant Arabidopsis and we have combined these maps with others, produced under identical conditions, for H3K9me2, H3K9me3, H3K27me3 and DNA methylation. Integrative analysis indicates that these 12 chromatin marks, which collectively cover ~90% of the genome, are present at any given position in a very limited number of combinations. Moreover, we show that the distribution of the 12 marks along the genomic sequence defines four main chromatin states, which preferentially index active genes, repressed genes, silent repeat elements and intergenic regions. Given the compact nature of the Arabidopsis genome, these four indexing states typically translate into short chromatin domains interspersed with each other. This first combinatorial view of the Arabidopsis epigenome points to simple principles of organization as in metazoans and provides a framework for further studies of chromatin-based regulatory mechanisms in plants.

The EMBO Journal advance online publication, 12 April 2011; doi:10.1038/emboj.2011.103
Subject Categories: chromatin and transcription; plant biology
Keywords: Arabidopsis; chromatin; DNA methylation; epigenome; histone modifications

Introduction

Packaging of DNA into chromatin is pivotal for the regulation of genome activity in eukaryotes. The basic unit of chromatin is the nucleosome, which is composed of 147 bp of DNA wrapped around a protein octamer composed of two molecules each of the core histones H2A, H2B, H3 and H4. Covalent modifications of histones, DNA methylation, incorporation of histone variants, and other factors, such as chromatin-remodelling enzymes or small RNAs, all contribute to defining distinct chromatin states that modulate access to DNA (Berger, 2007; Kouzarides, 2007). In particular, different histone modifications are thought to act sequentially or in combination in order to confer distinct transcriptional outcomes (Strahl and Allis, 2000; Jenuwein and Allis, 2001; Berger, 2007; Lee *et al*, 2010a). More generally, it is now well established that the precise composition of chromatin along the genome, which defines the epigenome, participates in the selective readout of the genomic sequence.

Thanks in part to a compact, almost fully sequenced and well-annotated genome, the flowering plant *Arabidopsis thaliana* has become a model of choice for exploring the epigenomes of multicellular organisms and the contribution of chromatin to the regulation of genome activity during development or in response to the environment. Indeed, epigenomic profiling in Arabidopsis has begun to provide insights into the relationship between transcriptional activity and localization of chromatin marks or histone variants (Roudier *et al*, 2009; Feng and Jacobsen, 2011). For instance, H3K4me3 and H3K36me2 are detected at the 5′- and 3′-ends of actively transcribed genes, respectively (Oh *et al*, 2008; Zhang *et al*, 2009), while H3K27me3 broadly marks repressed genes (Turck *et al*, 2007; Zhang *et al*, 2007; Oh *et al*, 2008). In contrast, cytosine methylation (5mC) has a dual localization. It is present predominantly over silent transposable elements (TEs) and other repeats, where it is associated with H3K9me2 and H3K27me1, but also in the body of ~30% of genes, many

*Corresponding authors. F Roudier or V Colot, Institut de Biologie de l'Ecole Normale Supérieure, CNRS UMR8197, INSERM U1024, 46 rue d'Ulm, 75230 Paris Cedex 05, France. Tel.: +33 014 432 3538; Fax: +33 014 432 3935; E-mails: roudier@biologie.ens.fr or colot@biologie.ens.fr

Received: 23 November 2010; accepted: 10 March 2011

of which are characterized by moderate expression levels (Lippman *et al*, 2004; Zhang *et al*, 2006; Zilberman *et al*, 2006; Turck *et al*, 2007; Vaughn *et al*, 2007; Bernatavichute *et al*, 2008; Cokus *et al*, 2008; Lister *et al*, 2008; Jacob *et al*, 2010). Furthermore, the variant histone H2A.Z, which is preferentially deposited near the 5′-end of genes and promotes transcriptional competence, antagonizes DNA methylation and vice versa (Zilberman *et al*, 2008). However, extensive combinatorial analyses of these and other chromatin marks have not been performed so far in Arabidopsis and meta-analysis of published data is complicated by the fact that biological materials and methodologies often differ between studies.

Here, we report the epigenomic profiles of eight histone modifications (H3K4me2, H3K4me3, H3K27me1, H3K27me2, H3K36me3, H3K56ac, H4K20me1 and H2Bub). Integrative analyses of these and other profiles, previously obtained under identical conditions for DNA methylation, H3K9me2, H3K9me3 and H3K27me3 (Turck *et al*, 2007; Vaughn *et al*, 2007), indicate a low combinatorial complexity of chromatin marks in Arabidopsis, as recently reported for metazoans (Wang *et al*, 2008; Hon *et al*, 2009; Ernst and Kellis, 2010; Gerstein *et al*, 2010; Roy *et al*, 2010; Kharchenko *et al*, 2011; Liu *et al*, 2011; Riddle *et al*, 2011; Zhou *et al*, 2011). Furthermore, our study identifies four main chromatin states in Arabidopsis, which have distinct indexing functions and which typically form short domains interspersed with each other. This first comprehensive view of the Arabidopsis epigenome suggests simple principles of organization, as recently proposed for Drosophila (Filion *et al*, 2010), and provides a resource to refine our understanding of the control of genome activity at the level of chromatin.

Results

Epigenomic profiling of 12 chromatin marks

Epigenomic maps were generated for eight histone modifications (H3K4me2, H3K4me3, H3K27me1, H3K27me2, H3K36me3, H3K56ac, H2Bub and H4K20me1) using chromatin extracted from young seedlings and immunoprecipitation followed by hybridization to a tiling microarray that covers the entire chromosome 4 of Arabidopsis at ~900 bp resolution (Turck *et al*, 2007). Data previously obtained for 5mC (Vaughn *et al*, 2007), H3K9me2, H3K9me3 and H3K27me3 (Turck *et al*, 2007) using similar materials and methodologies were also considered. Epigenomic profiling was additionally performed for seven of these marks (H3K4me2, H3K4me3, H3K27me1, H3K27me3, H3K36me3, H2Bub and 5mC) using a tiling microarray covering the whole-genome sequence at 165 bp resolution. Chromosome 4 and whole-genome maps were also obtained for histone H3 to control for nucleosome occupancy. The 12 marks were chosen because they were shown in previous studies to be associated with distinct transcriptional activities or subnuclear localization in Arabidopsis. In addition, our selection was focussed to a large extent on histone lysine methylation, which exists in three forms (mono-, di- and trimethylation) and therefore has a versatile indexing potential (Sims and Reinberg, 2008).

Collectively, the 12 chromatin marks cover almost all of the regions that are detectably associated with histone H3, which amount to ~90% of the total genome sequence (data not shown; Chodavarapu *et al*, 2010). The distribution of each chromatin modification was characterized in detail along chromosome 4. In agreement with previous reports (Lippman *et al*, 2004; Turck *et al*, 2007; Zhang *et al*, 2007, 2009; Oh *et al*, 2008; Tanurdzic *et al*, 2008), H3K4me2, H3K4me3, H3K9me3, H3K27me3 and H3K56ac are mostly found in euchromatin (Figure 1A; Supplementary Figure S1; Supplementary Table I), which reflects the fact that these different modifications are associated almost exclusively with genes (Figure 1B). H2Bub and H3K36me3, for which no epigenomic maps have been reported to date in plants, are also characterized by a predominant distribution over genes. In contrast, H4K20me1 is found in heterochromatin mainly and associates with TE and other repeat element sequences (Figure 1B), like H3K9me2 (Lippman *et al*, 2004; Bernatavichute *et al*, 2008). The present analysis reveals in addition that, like 5mC (Zhang *et al*, 2006; Zilberman *et al*, 2006), H3K27me1 and H3K27me2 are dual marks associated not only with TEs but also with a fraction of genes (Supplementary Tables II–IV).

Each chromatin mark defines domains of contiguous tiles and the number of these domains ranges from 306 for H3K9me2 to 1163 for H3K4me3. For H3K4me3, H3K36me3, H3K56ac, H3K9me3, H2Bub or H3K27me3, domains have similar median length between euchromatin and heterochromatin and mostly coincide with single transcription units (Supplementary Table II; Supplementary Figure S2). By contrast, H3K9me2, H4K20me1, H3K27me1, H3K27me2 and 5mC form small domains in euchromatin but large domains in heterochromatin, as a result of the dense clustering of TE and other repeat sequences in the latter (Supplementary Figure S2; Supplementary Table II).

Combinatorial analysis of chromatin marks

As a first step in exploring the combinatorial deposition patterns of chromatin marks, unbiased pairwise association analyses were carried out. A heat map generated from the calculated association values (Supplementary Table V) and organized by hierarchical clustering reveals two clear groups of correlated pairs that distinguish genes from TE sequences (Figure 1C). Next, co-occurrence of marks was registered over each of the ~20 000 tiles of the chromosome 4 array. Of the $2^{12} = 4096$ combinations theoretically possible, only 665 were observed and among these, only 38 concerned at least 100 tiles (Supplementary Figure S3A). This indicates therefore a limited repertoire of chromatin signatures in Arabidopsis, as in other eukaryotes (Ernst and Kellis, 2010; Kharchenko *et al*, 2011; Liu *et al*, 2011). The four prevalent combinations of marks are H3K27me1 + 5mC + H3K9me2 + H4K20me1 + H3K27me2, H3K56Ac + H2Bub + H3K4me3 + H3K4me2 + H3K9me3 + H3K36me3, H3K27me3 + H3K27me2 + H3K4me2 and H3K27me3 + H3K27me2, which cover 10.9, 6.8, 4.7 and 4.6% of the tiling array, respectively. Whereas the first combination is almost exclusively associated with TE sequences, the other three are mainly present over genes (Supplementary Figure S3B). Furthermore, like H3K27me3 + H3K27me2, most of the remaining combinations represented by at least 100 tiles are subcombinations of the three prevalent ones (Supplementary Figure S3B and data not shown).

To complement this tile-centric analysis and to identify the prevalent combinatorial patterns of the 12 chromatin marks,

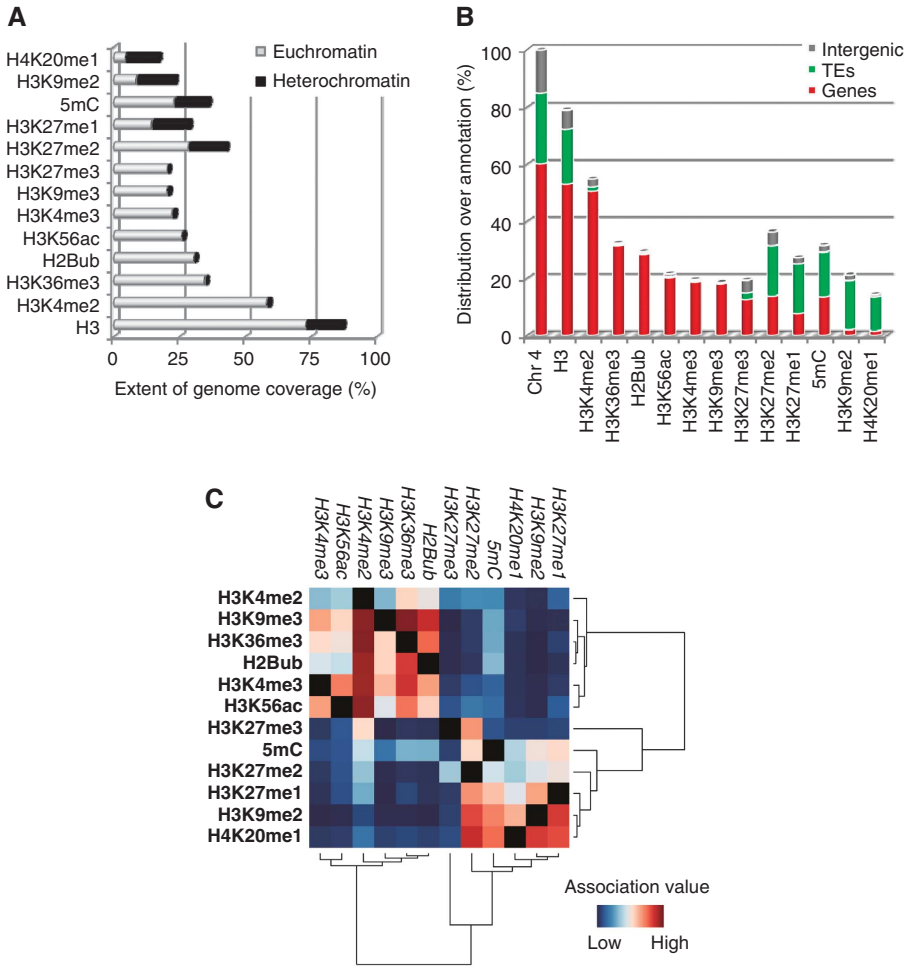


Figure 1 Genomic distribution of chromatin marks. (**A**) Relative coverage of chromatin marks in the euchromatin and heterochromatin of chromosome 4. Coordinates for heterochromatin are 1.61–2.36 Mb (knob) and 2.78–5.15 Mb (pericentromeric regions). (**B**) Chromosome-wide distribution of chromatin marks over annotated features. Tiles that overlap annotated genes or transposable elements (TAIR8) by at least 50 bp were assigned to the corresponding annotation and otherwise called 'intergenic'. (**C**) Pairwise association analysis of the 12 chromatin marks along chromosome 4. Mean association values were calculated for each pair of modifications over all marked tiles and are shown as a directional heat map organized by hierarchical clustering using Pearson's correlation distances.

unsupervised *c*-means clustering was performed. The number of clusters (k) was varied from 2 to 11 and $k=4$ was determined to be optimal in maximizing homogeneity within clusters and heterogeneity between them. The four chromatin states (CS1–CS4) defined by these four clusters are also identified by PCA analysis (data not shown), thus reinforcing their significance. Whereas CS1 regroups ~90% of the tiles associated with H3K4me3, H3K36me3, H3K9me3 and H2Bub as well as the majority of H3K4me2- and H3K56ac-marked sequences, H3K27me3 and H3K27me2 are the most prevalent modifications in CS2 (Figure 2A). As expected from their composition, CS1 and CS2 are mainly associated with genes (Figure 2B) and have antagonistic indexing functions, being prevalent among active and repressed/lowly expressed genes, respectively (Figure 2C). CS3, which is associated predominantly with TE sequences (Figure 2B), regroups most of the tiles marked by H3K9me2, H4K20me1 and H3K27me1 as well as ~50% of those marked by H3K27me2 and 5mC (Figure 2A). In contrast to the other three chromatin states, CS4 is not particularly enriched in any chromatin mark (Figure 2A) and is found mainly outside of genes and TE sequences (Figure 2B). Nonetheless, CS4 also marks ~10% of genes, most of which display low expression (Figure 2C). In keeping with the domain layout of individual marks, CS1–CS4 typically form small domains interspersed with each other, except in cytologically defined heterochromatin, where CS3 forms larger domains as a result of the clustering of TE sequences (Figure 2D; Supplementary Figure S4).

Chromatin signatures of genes

To investigate further the chromatin indexing of genes, pairwise analysis of chromatin modifications was carried out specifically over genic tiles, which revealed a tight association between H3K4me3 and H3K56ac, between H3K36me3, H3K9me3 and H2Bub and between H3K27me2 and H3K27me3 (Figure 3A). Next, average enrichment levels were calculated within and around genes for all marks except H3K9me2 and H4K20me1, which are almost exclusively associated with TE and other repeat sequences. As shown in Figure 3B, values are highest within the transcribed region for the 10 chromatin modifications considered and

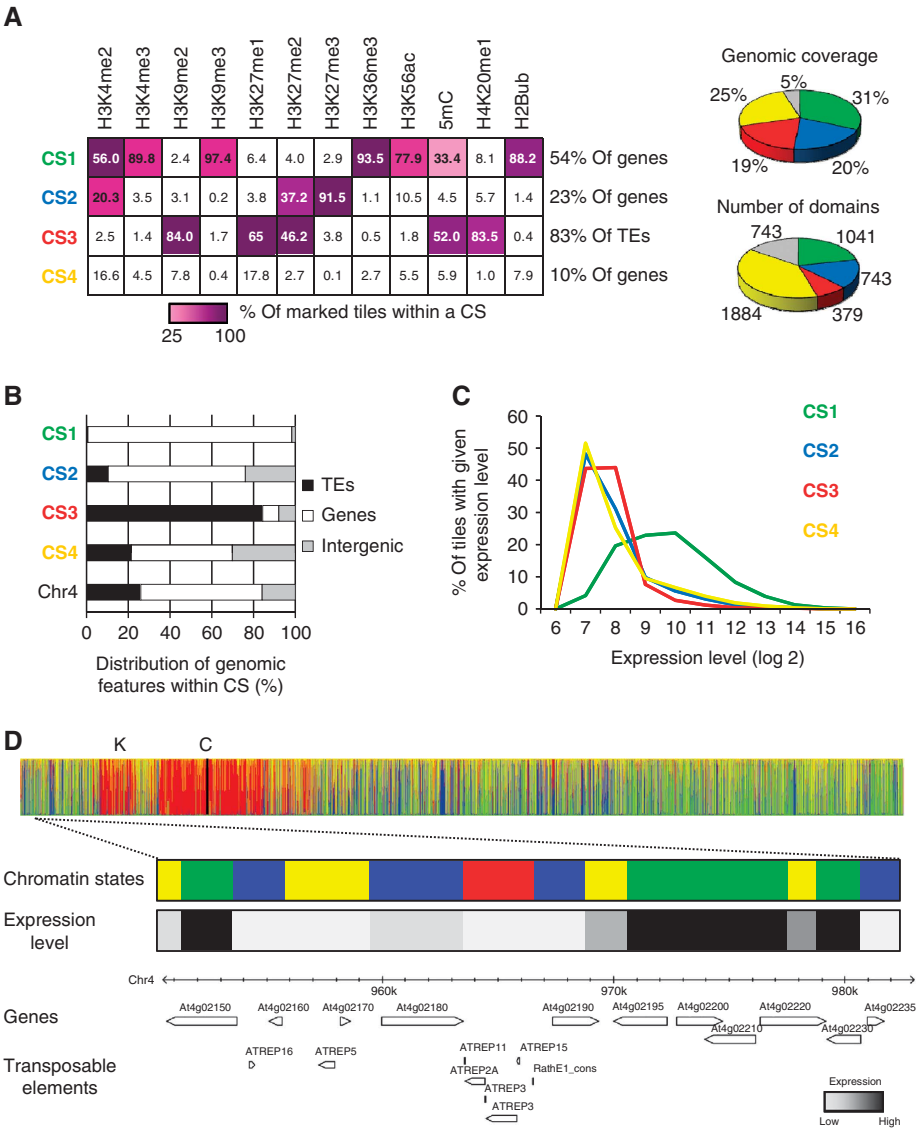


Figure 2 The Arabidopsis epigenome contains four predominant chromatin states. (**A**) The table on the left indicates the composition of the four predominant chromatin states (CS) identified by *c*-means clustering. The distribution of the 12 chromatin marks over the four CS is indicated as a heat map for values ranging from 25% (light purple) to 100% (dark purple). The degree of homogeneity of each CS is indicated by the percentage of tiles assigned to it that are associated with each of the 12 chromatin marks (numbers inside cells). Note that no single mark is present over >20% of the tiles assigned to CS4, in contrast to what is observed for CS1–CS3. The percentage of genes indexed by CS1, CS2 and CS4 and the percentage of TE annotations indexed by CS3 are also shown. Pie charts indicate the relative genomic coverage of the four CS and the number of domains that they each form. Grey colour corresponds to tiles that cannot be unambiguously assigned to any of the four CS (see Materials and methods). (**B**) Relative proportion of genomic features within each CS. Tiles that overlap annotated genes or transposable elements (TAIR8) by at least 50 bp were assigned the corresponding annotation. All other tiles were considered as 'intergenic'. (**C**) Relationship between chromatin states and gene expression level. The percentage of tiles associated with a given CS is represented according to expression level. The dashed line represents the distribution of all annotated genes of chromosome 4. Expression data (Schmid *et al*, 2005) were obtained by averaging appropriate developmental stages. (**D**) Distribution of the four CS along chromosome 4. For each tile, membership to a given CS is colour coded. K: heterochromatic knob. The non-sequenced part of the centromere (C) is represented by the vertical black line. The high interspersion of chromatin states seen outside of heterochromatin is highlighted in a genome browser view of a 30-kb euchromatic region (positions 0.95–0.98 Mb).

are typically lowest upstream or downstream of it. However, distribution patterns vary substantially between marks, as previously established in several instances (Turck *et al*, 2007; Zhang *et al*, 2007; Jacob *et al*, 2010). H3K4me3, H3K56ac, H3K4me2, H3K36me3 and H3K9me3 all peak at the 5′-end of the transcribed region, but the first two marks more sharply than the other three (Figure 3B). In contrast, H2Bub as well as H3K27me1 are highest more centrally, 5mC is most enriched in the 3′-half of the transcribed region and both H3K27me2 and H2K27me3 show an even distribution across transcribed regions. Finally, H3K27me2 differs from all other marks including H3K27me3 in that it remains high in flanking regions, a difference which does not result from the presence of H3K27me2-marked TE sequences adjacent to genes nor from the lower signal to noise ratio measured for this mark (see legend of Figure 3A, data not

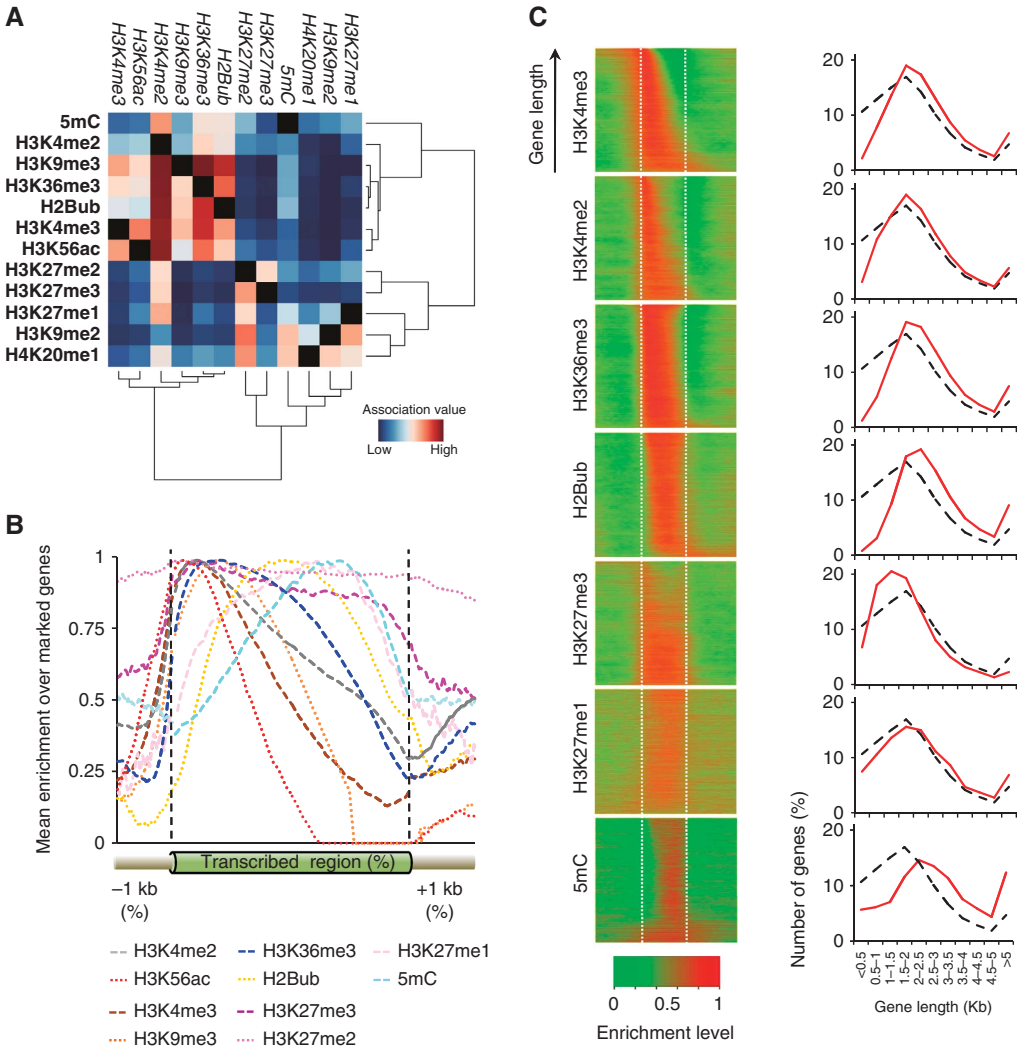


Figure 3 Distribution of chromatin marks over genes. (**A**) Pairwise association analysis of the 12 chromatin marks along chromosome 4. Mean association values were calculated for each pair of modifications over all marked genic tiles and are shown as a directional heat map organized by hierarchical clustering using Pearson's correlation distances metric. (**B**) Mean enrichment levels relative to histone H3 are plotted along marked genes (transcribed region, scaled to accommodate for different gene lengths, bin size of 1%) as well as up to 1 kb of upstream and downstream sequences (bin size of 10 bp). Maximum value for any given mark is arbitrarily set to 1. Data were obtained using the chromosome 4 tiling array. Note that values for H3K27me2 in upstream and downstream regions are significantly higher than for unmarked genes (>0.9 versus ~0.6, not shown). (**C**) Left panels: Enrichment levels relative to histone H3 for marked genes sorted by length. Each line represents a single gene as well as 1 kb of upstream and downstream sequences. Enrichment is indicated as a heat map, with maximal (red) and minimal (green) values set to 1 and 0, respectively. Right panels: Frequency distribution of marked (red line) and all genes (black dashed line) according to their length. Data were obtained using the whole-genome tiling array.

shown). Using the genome-wide profiles obtained for seven chromatin modifications, we could show in addition that contrary to H3K27me3, which preferentially marks small genes as noted before (Luo and Lam, 2010), H2Bub, H3K36me3, 5mC and, to a lesser extent, H3K4me2 as well as H3K4me3 tend to be associated with longer genes (Figure 3C). Unlike these chromatin modifications, H3K27me1 does not exhibit preferential association in relation to gene length (Figure 3C).

It has been established that H3K4me3 and H3K56ac mark genes that are highly and broadly expressed (Oh *et al*, 2008; Tanurdzic *et al*, 2008; Zhang *et al*, 2009). Conversely, H3K27me3 is preferentially associated with genes that are expressed at low levels or in a tissue-specific manner (Turck *et al*, 2007; Zhang *et al*, 2007; Oh *et al*, 2008; Jacob *et al*, 2010) and 5mC tends to mark moderately expressed genes (Zilberman *et al*, 2006; Vaughn *et al*, 2007). Our analysis confirms these results and indicates in addition that H2Bub, H3K36me3 and H3K9me3 tend to mark highly expressed genes, like H3K4me3 and H3K56ac (Figure 4A). On the other hand, H3K4me2 does not appear to index genes in relation to their expression level and H3K27me1 as well as H3K27me2 tend to be associated with genes that are expressed at low level or in a tissue-specific manner, like H3K27me3 (Figure 4A and B). However, H3K27me1 and H3K27me2/3 mark largely non-overlapping sets of genes with different ontologies (Figure 3A; Supplementary Tables III, IV, VI and VII), which suggests the existence of two distinct gene repression systems associated with methylation of H3K27. For most chromatin marks, average enrichment

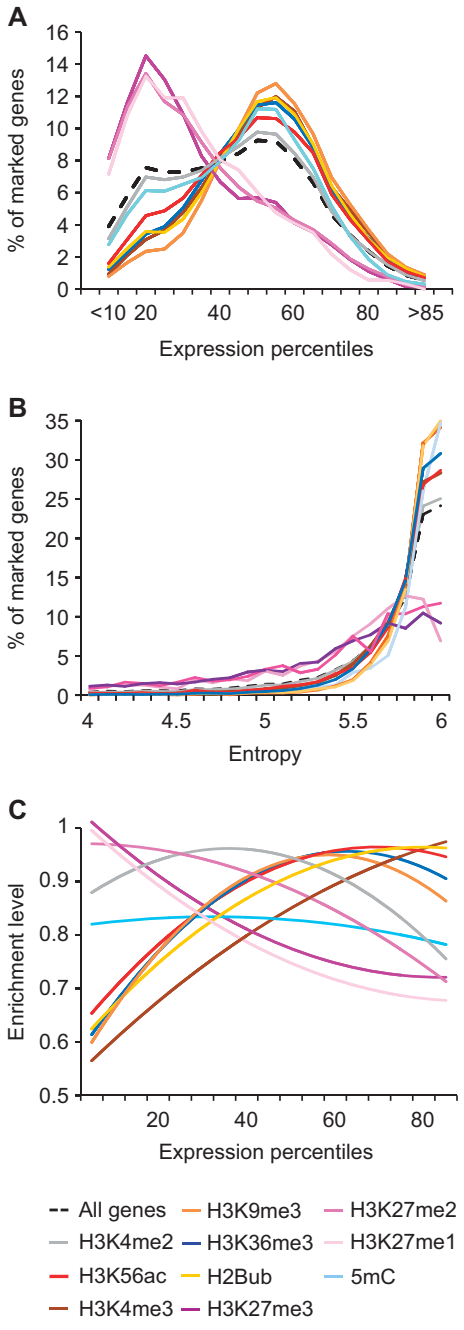


Figure 4 Chromatin indexing in relation to gene expression. (**A**) Distribution density of marked genes according to expression percentiles. Genes were binned according to their absolute expression values in whole seedlings. The dashed line indicates the distribution of all annotated genes on chromosome 4 across all expression percentiles. Expression data (Schmid *et al*, 2005) were obtained by averaging appropriate developmental stages. (**B**) Tissue specificity of marked genes as estimated by Shannon entropy calculation. Low entropy values indicate high tissue specificity. The fraction of marked genes associated with a given entropy value is plotted for each chromatin modification. (**C**) Relationship between gene expression and enrichment level for each chromatin modification. Maximum enrichment level is set to 1 in each case.

levels correlate either positively or negatively with expression levels (Figure 4C). Thus, values for H3K4me3, H3K56ac, H3K36me3, H2Bub and H3K9me3 increase gradually with gene expression, at least up to mid expression levels, whereas values for H3K27me1, H3K27me2 and H3K27me3 show an opposite trend. Whether these correlations reflect expression of genes in a variable number of cells, or true differential enrichment in relation to expression level, remains to be determined.

Collectively, our findings indicate that H3K4me3 and H3K27me3 are diagnostic of two antagonist chromatin states that are associated with most active and repressed genes, respectively. However, ~13% (3433 out of 27 294) of genes marked by H3K4me3 or H3K27me3 in whole seedlings present both marks, in agreement with previous observations (Oh *et al*, 2008; Zhang *et al*, 2009). To explore this further, H3K4me3 and H3K27me3 were mapped genome-wide using chromatin extracted from roots and profiles were compared with those obtained for whole seedlings (this study) or aerial parts only (Oh *et al*, 2008). Out of the 3433 genes with both marks in whole seedlings, 284 genes (8.3%) are only marked by H3K4me3 in roots and by H3K27me3 in aerial parts or vice versa (Figure 5A; Supplementary Table VIII). Correspondingly, a majority of these genes show differential expression between roots and aerial parts (Figure 5B), which is in contrast to genes with persistent co-marking in both plant parts (Figure 5A and C). Thus, it can be concluded that co-marking in whole seedlings results for a number of genes from the mixing of cells with opposite chromatin indexing in the two plant parts. By extension, it is likely that persistent co-marking in one or the other plant parts (Figure 5A) reflects similar mixing of cells with distinct epigenomes, but this time within organs. Co-marking could nevertheless correspond to *bona fide* bivalent marking in some cases, as originally reported in mammals for key regulatory genes poised for activation (Wang *et al*, 2009) and as also described in Arabidopsis for a small number of genes encoding transcription factors (Jiang *et al*, 2008; Berr *et al*, 2010). In this respect, it is noteworthy that ontology analysis of the 224 genes with persistent co-marking in both roots and aerial parts (Figure 5A) indicates significant enrichment for terms associated with regulation of transcription (data not shown).

Discussion

A small number of prevalent chromatin states index the Arabidopsis genome

Using an integrative analysis of the distribution of 12 chromatin marks, we show that the Arabidopsis epigenome is organized around four predominant chromatin states with distinct biochemical, transcriptional and sequence properties. This representation refines the classical segmentation between cytologically defined heterochromatin and euchromatin. A first chromatin state (CS1) corresponds to transcriptionally active genes and is typically enriched in the trimethylated forms of H3K4 and H3K36. Two further states correspond to two distinct types of repressive chromatin. H3K27me3-marked repressive chromatin (CS2) is mainly associated with genes under PRC2-mediated repression (Turck *et al*, 2007; Zhang *et al*, 2007), while H3K9me2- and H4K20me1-marked repressive chromatin (CS3) corresponds to classical heterochromatin and is almost exclusively located over silent TEs (Lippman *et al*, 2004; Bernatavichute *et al*, 2008). A fourth chromatin state (CS4) is characterized by the

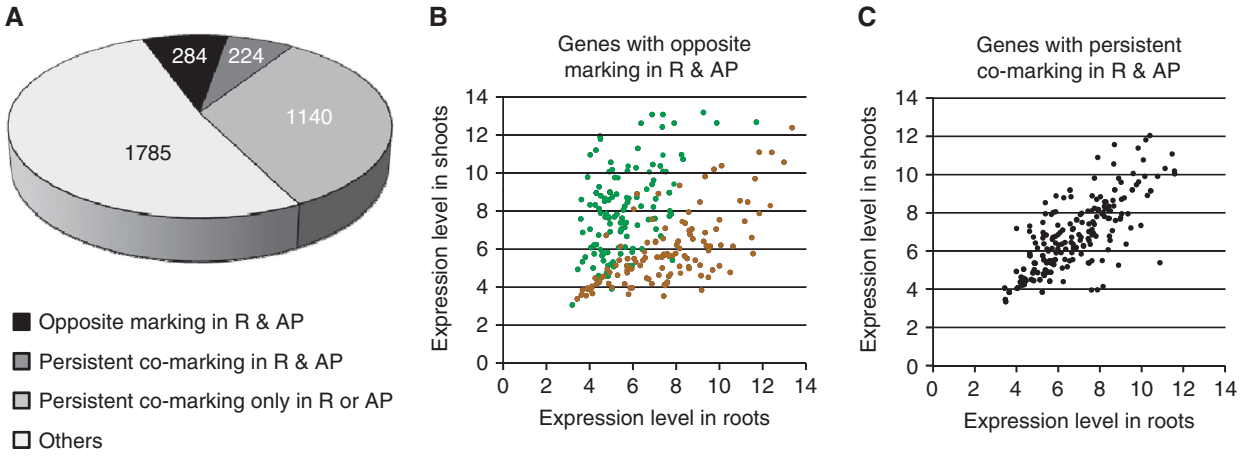


Figure 5 Analysis of genes co-marked with H3K27me3 and H3K4me3 in whole seedlings. (**A**) The 3433 genes co-marked in whole seedlings were split into different classes according to their marking in roots (R; this study) and aerial parts (AP; Oh *et al*, 2008). 'Others' indicate genes with other marking patterns in the two plant parts. This class, which is not expected based on the co-marking observed in whole seedlings, can be explained in part by the fact that the different data sets were not all generated using the same conditions and methodologies. (**B**) Expression analysis in roots and shoots (Schmid *et al*, 2005) for the 284 genes showing opposite marking in roots and aerial parts. Brown dots indicate genes with H3K4me3 in roots and H3K27me3 in aerial parts, green dots indicate genes with the opposite marking pattern. (**C**) Expression analysis in roots and shoots (Schmid *et al*, 2005) for the 224 genes showing persistent co-marking in roots and aerial parts.

absence of any prevalent mark and is associated with weakly expressed genes and intergenic regions.

This rather simple organization of Arabidopsis chromatin into four main states shows similarities with that recently reported for Drosophila cells. Indeed, based on the integration of epigenomic maps obtained for 53 chromatin proteins, it was concluded that the Drosophila epigenome is organized into a mosaic of five principal chromatin types that display distinct functional properties (Filion *et al*, 2010). Specifically, Arabidopsis CS2 and CS3 are similar to Drosophila 'BLUE' and 'GREEN' chromatin types, which correspond to repressive chromatin associated with the Polycomb pathway and classical heterochromatin, respectively. Furthermore, CS4, which has no prevalent chromatin mark and indexes some weakly expressed genes as well as intergenic regions is reminiscent of Drosophila 'BLACK' chromatin, which is relatively gene poor and constitutes a repressive environment distinct from heterochromatin. In contrast, transcriptionally active chromatin is represented by a single chromatin state in Arabidopsis (CS1) but by two distinct types in Drosophila that differ in several ways, including the enrichment of H3K36me3 in 'YELLOW' but not in 'RED' chromatin.

Other large-scale epigenomic studies have been performed in yeast (Liu *et al*, 2005), *C. elegans* (Gerstein *et al*, 2010; Liu *et al*, 2011), Drosophila (Kharchenko *et al*, 2011; Roy *et al*, 2010; Riddle *et al*, 2011) and human cells (Wang *et al*, 2008; Hon *et al*, 2009; Ernst and Kellis, 2010; Zhou *et al*, 2011), which all indicate a relatively low combinatorial complexity of chromatin marks. Furthermore, the two main repressive chromatin states defined in Arabidopsis (CS2 and CS3) have similar counterparts in metazoans, indicating that they are highly conserved between plants and animals. On the other hand, the single predominant chromatin state (CS1) that we have identified for transcriptionally active genes in Arabidopsis has no obvious equivalent in these other organisms. Instead, several chromatin states have been associated with expressed genes in other organisms. This discrepancy likely results from the smaller size of genes and intergenic regions in Arabidopsis (~2 kb each on average), as well as the relatively lower resolution of our data. Indeed, our analysis shows that distribution patterns vary substantially between chromatin marks associated with active genes (Figure 3B), which suggests that CS1 could be further refined into at least two additional chromatin signatures, specific to the promoter and transcribed region of these genes.

Although the number of chromatin states identified via this type of integrative approach may appear surprisingly low, such analyses aim to identify prevalent combinations of chromatin marks or chromatin proteins. Furthermore, the heterogeneity of the biological material used in many of these studies, including ours, likely hampered the detection of certain chromatin states such as those that are specific to rare cell types. Ultimately, only a knowledge of the epigenomes of individual cell types will enable a full understanding of the functional impact of chromatin-level regulation on genome activity.

Chromatin indexing of genes in Arabidopsis

Our work indicates that the Arabidopsis epigenome is mainly organized at the level of single transcription units and that the distribution of chromatin marks along genes is linked to the transcription process (Figures 2 and 3). For example, H3K4me3 peaks around the transcription start site of actively expressed genes, as observed in all other eukaryotes examined to date (Rando and Chang, 2009). Similarly, H3K56ac is specifically located at gene promoters and shows preferential marking of active genes, suggesting that, like in yeast, it could facilitate rapid transcriptional activation (Williams *et al*, 2008). In contrast to H3K4me3, H3K4me2 shows no particular association with highly expressed genes or with specific parts of genes. Rather than being a constitutive mark of transcription, H3K4me2 may be implicated in

fine tuning of tissue-specific expression, as recently reported in mammals (Pekowska *et al*, 2010).

The distribution of H3K36me3, H3K9me3 and H2Bub over the transcribed regions of expressed genes suggests that these modifications are linked with transcriptional elongation. In the case of H2Bub, this is in agreement with the distribution reported in mammals and yeast (Minsky *et al*, 2008; Schulze *et al*, 2009). For H3K9me3, enrichment over the coding region of expressed genes in Arabidopsis (this study; Caro *et al*, 2007; Turck *et al*, 2007; Charron *et al*, 2009) contrasts with the enrichment predominantly over heterochromatin in animals. However, association with the transcribed regions of some active genes has been reported in mammals (Vakoc *et al*, 2005, 2006; Squazzo *et al*, 2006). Whether H3K9me3 could serve different outcomes depending on genomic and or chromatin context and whether it has any role in transcription regulation in plants remains to be determined. Given the discrepancy between the low amounts of H3K9me3 reported in bulk histones (Jackson *et al*, 2004; Johnson *et al*, 2004) and its apparent abundance reported by ChIP-chip, it is also possible that the H3K9me3 antibody we used recognizes another modification in Arabidopsis, which would be H3K36me3 based on our epigenomic analysis. However, *in vitro* competition assays using an H3K36me3 peptide suggest that this is unlikely (Supplementary Figure S5).

H3K36me3 preferentially marks exons of transcribed genes in yeast, *C. elegans* and mammals (Kolasinska-Zwierz *et al*, 2009) and it was shown to be involved in the control of alternative splicing in mammals (Luco *et al*, 2010). In Arabidopsis, however, H3K36me3 peaks in the first half of the coding region, which is in contrast to the 3′-end enrichment reported in other organisms (Wang *et al*, 2009). This preferential enrichment at the 5′-end, which is not dependent on gene length, could indicate that the principles governing H3K36me3 deposition differ between plants and other eukaryotes. In fact, H3K36me3 distribution in Arabidopsis resembles that of H3K79me3 in mammals (Wang *et al*, 2009). As Arabidopsis lacks a clear homologue of the H3K79 methyltransferase Dot1 and has no H3K79me3 (Zhang *et al*, 2007), it is possible that H3K36me3 in plants serves a function equivalent to H3K79me3 in other eukaryotes. Furthermore, H3K36me2 could have a role similar to that attributed to H3K36me3 in other eukaryotes, as it peaks at the 3′-end of expressed genes in Arabidopsis (Oh *et al*, 2008).

Chromatin marks associated with transcription have been proposed to cross talk and serve as checkpoints in budding yeast and mammals (Suganuma and Workman, 2008; Weake and Workman, 2008; Lee *et al*, 2010a). A similar scenario could be envisioned in Arabidopsis based on the chromatin marks that predominate in CS1, whereby the RNA polymerase II-associated factor 1 complex would induce monoubiquitylation of H2B via the activity of the Rad6-Bre1 ubiquitin ligase homologues UBC1, 2 and 3 as well as HUB1 and 2, as shown at the *FLC* gene (Cao *et al*, 2008; Gu *et al*, 2009; Schmitz *et al*, 2009). H2Bub deposition would in turn help recruit COMPASS (COMplex Proteins ASsociated with Set1), thus mediating deposition of H3K4me3 and potentially H3K36me3 (in place of H3K79me3) as well as H3K36me2. Similarly to other eukaryotes, initiation of another round of transcription would require the activity of the Ubp8 ubiquitin protease homologue, UBP26, which catalyses H2B deubiquitylation (Sridhar *et al*, 2007). Consistent with this, H3K36me3 but not H3K36me2 nor H3K4me3 is almost lost at the 5′-end of the gene *FLC* in *ubp26* mutant plants and this loss is associated with a reduction of *FLC* expression (Schmitz *et al*, 2009). The steady-state distribution pattern of H2Bub observed over expressed genes presumably results from targeted deubiquitylation of H2B at the 5′-end and probably 3′-end of the transcribed region, rather than from an increased ubiquitylation of H2B towards the middle of the gene.

Our epigenomic profiling of the three forms of H3K27 indicates that methylation of this lysine residue is generally associated with repressive chromatin and that its indexing function depends on the degree of modification (mono-, di- and tri-methylation). Thus, in agreement with previous studies, H3K27me3, which is the hallmark of CS2, is almost exclusively present over transcriptionally repressed genes (Turck *et al*, 2007; Zhang *et al*, 2007), while H3K27me1 is prevalent over silent TEs in pericentromeric regions, where it is thought to prevent over-replication (Jacob *et al*, 2009, 2010). Our analysis reveals in addition that H3K27me2 is enriched over H3K27me3-marked genes, as well as of over TE sequences. Although immunolocalization of H3K27me2 at chromocenters (Fuchs *et al*, 2006) was proposed to result from cross-reactivity of antibodies with H3K27me1 in Arabidopsis (Jacob *et al*, 2009), we did not observe extensive cross-reactivity of the H3K27me2 antibodies used in our study with H3K27me1 (Supplementary Figure S5). Moreover, while all forms of methylated H3K27 can be found over genes and are associated with transcriptional repression, little overlap is observed between the small group of genes marked by H3K27me1 and the much larger set of genes marked by H3K27me2/3, suggesting that these modifications define two repressive pathways with distinct gene targets (Supplementary Tables VI and VII). Whereas H3K27me3 deposition is catalysed by the evolutionarily conserved Polycomb Repressive Complexes 2 (Kohler and Hennig, 2010; Bouyer *et al*, 2011), H3K27me1 deposition over TE sequences is partly dependent on the activity of the two SET-domain proteins ATXR5 and ATXR6 (Jacob *et al*, 2009). Whether H3K27me1 deposition over genes requires the same or different histone methyltransferases and whether it is associated with the control of DNA replication remain to be determined. Irrespective of the mechanisms involved, it is noteworthy that whereas H3K27me1-marked TE sequences are also co-marked with H3K9me2 and 5mC, this is not the case for H3K27me1-marked genes.

Acetylation of H3K56 is another chromatin mark that has been linked with the replication process. In Arabidopsis cell cultures, early replicating sequences form broad domains of H3K56ac (Lee *et al*, 2010b). Our epigenomic profiling of H3K56ac reveals mostly short domains located at the 5′-end of expressed genes, which correspond to the replication-independent incorporation of acetylated H3K56. However, a few large domains (~20 kb) are also detected, which span several genes, intergenic regions and TEs. As our epigenomic maps have been derived from whole seedlings that comprise only a small proportion of mitotic cells, these large H3K56ac domains might correspond to sequences frequently used as endoreplication origins.

Although most Arabidopsis genes are associated with chromatin states CS1 or CS2, ~10% are instead associated

with CS4, which is characterized by the absence of any prevalent chromatin mark among the 12 that were analysed in this work (Figure 3). Analysis of additional chromatin marks and proteins will be required to determine more precisely the nature of CS4 and notably the extent of its similarity to the repressive chromatin type BLACK of Drosophila (Filion *et al*, 2010).

To conclude, the first integrative view of the Arabidopsis epigenome provided here could be compared with a first sketch, which is progressively refined until a complete blueprint is produced. Importantly, key aspects of the Arabidopsis epigenome are already apparent in this first sketch, like the relative simplicity of designing principles, which appears to be shared with metazoans.

Materials and methods

Immunoprecipitation of chromatin and methylated DNA, labelling and microarray hybridization

All experiments were performed using wild-type *Arabidopsis thaliana* accession Columbia seedlings grown for 10 days either in liquid MS (whole seedlings) or on MS agar plates (roots and aerial parts) supplemented with 1% sucrose under long day conditions. ChIP and Me-DIP assays were carried out essentially as described (Lippman *et al*, 2005) using commercially available antibodies (Supplementary Table IX; Supplementary Figure S5). Specificity of the H3K27me2 and H3K9me3 antibodies was tested by peptide competition and western blotting analysis on nuclear extracts (Supplementary Figure S5) as described in Bouyer *et al* (2011) using H3K27me3, H3K27me2, H3K27me1, H3K9me3 and H3K36me3 peptides (Millipore, 12-565, 12-566, 12-567, 12-568 and Diagenode sp-058-050, respectively). Immunoprecipitated DNA (IP) and input DNA (INPUT) were amplified, differentially labelled and co-hybridized in dye-swap experiments as described (Lippman *et al*, 2004; Turck *et al*, 2007) for the chromosome 4 tiling array or according to the manufacturer's instructions for the Roche NimbleGen whole-genome tiling array. Two biological replicates were analySed (two dye-swaps). The chromosome 4 array contains 21 800 printed features, on average ~900 bp in size. The heterochromatic knob on the short arm and several megabases of pericentromeric heterochromatin are included and account for 16% of the 18.6 Mb covered by the array. Details of array design and production are described in Vaughn *et al* (2007). This platform has been deposited to GEO under accession number GPL10172. The whole-genome tiling array consists of 50–75 nt tiles, with 110 nt spacing on average, that are tiled across the entire genome sequence (TAIR7), without repeat masking. Tiles have a melting temperature of 74°C on average and 88% of them match a unique position in the genome. This custom design was either split into two arrays of 360 718 tiles each, with every other tile on each array (GEO accessions GPL10911 and GPL10918) or synthesized in triplicates of 711 320 tiles each on a single array (GEO accession GPL11005).

ChIP- and Me-DIP-chip data analysis

Hybridization data were normalized as described previously for the chromosome 4 array (Turck *et al*, 2007) or using an ANOVA model was applied to remove technical biases from data obtained using the whole-genome array. Data were averaged on the dye-swap to remove tile-specific dye bias. Normalized data were analysed using the ChIPmix method (Martin-Magniette *et al*, 2008), which was adapted to handle multiple biological replicates simultaneously. This method is based on a mixture model of regressions, the parameters of which are estimated using the EM algorithm. For each tile, a posterior probability is defined as the probability to be enriched given the log(Input) and log(IP) intensities, and is used to assign each tile into a normal or enriched class. A false-positive risk is determined by defining the probability of obtaining a posterior probability at least as extreme as the one that is actually observed when the tile is normal. False-positive risks are then adjusted by the Benjamini–Hochberg procedure and tiles for which the adjusted false-positive risk is <0.01 are declared enriched. Previously published data (Turck *et al*, 2007; Vaughn *et al*, 2007) were re-analysed using the same procedure. Neighbouring enriched tiles are joined into domains by requiring a minimal run of 1.6 kb or 400 bp and allowing a maximal gap of 800 or 200 bp for data obtained using the chromosome 4 or whole-genome arrays, respectively. Thus, 'singletons' are not considered for further analyses.

Computational analyses

General bioinformatics methods including positional, quantitative and class-based computations were conducted in Excel and using *ad hoc* scripts written in R, PERL or Python. Genes and transposable elements were annotated based on TAIR8 and other sequences are assumed to be intergenic. Gene Ontology analyses were done using the GOrilla (Eden *et al*, 2009) with an additional correction for multiple testing of the *P*-values. Pairwise association analysis, which is directional unlike correlation analysis, was calculated by scoring the frequency of co-occurrence of pairs of chromatin modifications among the 12 marks analysed on the chromosome 4 tiling array.

Whole seedlings transcriptome data were retrieved from Schmid *et al* (2005) and genes were binned into 20 expression percentiles according to their absolute expression values. Within each expression percentile, the number of genes marked by a given chromatin modification was calculated and represented as a percentage of all the genes marked by this modification. Shannon entropy for each set of marked genes was calculated as described (Zhang *et al*, 2006) using publicly available developmental expression series (Schmid *et al*, 2005), after filtering genes that showed no expression in any conditions.

Fuzzy *c*-means clustering using R MCLUST package was performed to classify tiles into principal chromatin states based on the 12 epigenomic maps. *c*-means clustering computes membership values for each tile towards all the clusters and all the membership values add up to 1. Each tile was assigned to one cluster only, based on a membership value equal or higher to 0.5. To identify the optimal number of clusters (k), cluster validity value, which is an estimate of homogeneity within the clusters and heterogeneity between them, was calculated for clusters from $k=2$–11.

Data availability

Raw and processed data have been deposited to NCBI's Gene Expression Omnibus (GEO; http://www.ncbi.nlm.nih.gov/geo/) under the super-series accession GSE24710 and to CATdb (http://urgv.evry.inra.fr/CATdb) (Samson *et al*, 2004; Gagnot *et al*, 2008). In addition, array data and genome annotation are displayed using a Generic Genome Browser, available for visualization at http://epigara.biologie.ens.fr/index.html.

Supplementary data

Supplementary data are available at *The EMBO Journal* Online (http://www.embojournal.org).

Acknowledgements

We thank members of the Colot group and Edith Heard for critical reading of the manuscript. This work was supported by grants from the Agence Nationale de la Recherche (ANR Genoplante TAG and REGENEOME, ANR blanc DDB1, ANR Sysbio) and by the European Union Network of Excellence 'The Epigenome'. IA, AS and SC were supported by PhD studentships from the ANR, the Centre National de la Recherche Scientifique (CNRS) and the Ministère de l'Enseignement Supérieur et de la Recherche (MESR), respectively.

Author contributions: FR and VC conceived and designed the experiments. FR, SC, DB, EC, LG, BD, SDr and FB performed the experiments. FR, IA, AS and VC analysed the data. CBe, TM-H, ED-B, LA-S, SDe, VB, SA, AS, CBo, M-LMM, SR and MC contributed reagents/materials/analysis tools. FR and VC wrote the paper.

Conflict of interest

The authors declare that they have no conflict of interest.

References

Berger SL (2007) The complex language of chromatin regulation during transcription. *Nature* **447:** 407–412

Bernatavichute YV, Zhang X, Cokus S, Pellegrini M, Jacobsen SE (2008) Genome-wide association of histone H3 lysine nine methylation with CHG DNA methylation in Arabidopsis thaliana. *PLoS One* **3:** e3156

Berr A, McCallum EJ, Menard R, Meyer D, Fuchs J, Dong A, Shen WH (2010) Arabidopsis SET DOMAIN GROUP2 is required for H3K4 trimethylation and is crucial for both sporophyte and gametophyte development. *Plant Cell* **22:** 3232–3248

Bouyer D, Roudier F, Heese M, Ellen D, Andersen ED, Gey D, Nowack MK, Goodrich J, Renou J-P, Grini PE, Colot V, Schnittger A (2011) Polycomb Repressive Complex 2 controls the embryo to seedling phase transition. *PLoS Genet* **7:** e1002014

Cao Y, Dai Y, Cui S, Ma L (2008) Histone H2B monoubiquitination in the chromatin of FLOWERING LOCUS C regulates flowering time in Arabidopsis. *Plant Cell* **20:** 2586–2602

Caro E, Castellano MM, Gutierrez C (2007) A chromatin link that couples cell division to root epidermis patterning in Arabidopsis. *Nature* **447:** 213–217

Charron JB, He H, Elling AA, Deng XW (2009) Dynamic landscapes of four histone modifications during deetiolation in Arabidopsis. *Plant Cell* **21:** 3732–3748

Chodavarapu RK, Feng S, Bernatavichute YV, Chen PY, Stroud H, Yu Y, Hetzel JA, Kuo F, Kim J, Cokus SJ, Casero D, Bernal M, Huijser P, Clark AT, Kramer U, Merchant SS, Zhang X, Jacobsen SE, Pellegrini M (2010) Relationship between nucleosome positioning and DNA methylation. *Nature* **466:** 388–392

Cokus SJ, Feng S, Zhang X, Chen Z, Merriman B, Haudenschild CD, Pradhan S, Nelson SF, Pellegrini M, Jacobsen SE (2008) Shotgun bisulphite sequencing of the Arabidopsis genome reveals DNA methylation patterning. *Nature* **452:** 215–219

Eden E, Navon R, Steinfeld I, Lipson D, Yakhini Z (2009) GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. *BMC Bioinformatics* **10:** 48

Ernst J, Kellis M (2010) Discovery and characterization of chromatin states for systematic annotation of the human genome. *Nat Biotechnol* **28:** 817–825

Feng S, Jacobsen SE (2011) Epigenetic modifications in plants: an evolutionary perspective. *Curr Opin Plant Biol* (advance online publication; doi:10.1016/j.pbi.2010.12.002)

Filion GJ, van Bemmel JG, Braunschweig U, Talhout W, Kind J, Ward LD, Brugman W, de Castro IJ, Kerkhoven RM, Bussemaker HJ, van Steensel B (2010) Systematic protein location mapping reveals five principal chromatin types in Drosophila cells. *Cell* **143:** 212–224

Fuchs J, Demidov D, Houben A, Schubert I (2006) Chromosomal histone modification patterns–from conservation to diversity. *Trends Plant Sci* **11:** 199–208

Gagnot S, Tamby JP, Martin-Magniette ML, Bitton F, Taconnat L, Balzergue S, Aubourg S, Renou JP, Lecharny A, Brunaud V (2008) CATdb: a public access to Arabidopsis transcriptome data from the URGV-CATMA platform. *Nucleic Acids Res* **36:** D986–D990

Gerstein MB, Lu ZJ, Van Nostrand EL, Cheng C, Arshinoff BI, Liu T, Yip KY, Robilotto R, Rechtsteiner A, Ikegami K, Alves P, Chateigner A, Perry M, Morris M, Auerbach RK, Feng X, Leng J, Vielle A, Niu W, Rhrissorrakrai K *et al* (2010) Integrative analysis of the Caenorhabditis elegans genome by the modENCODE project. *Science* **330:** 1775–1787

Gu X, Jiang D, Wang Y, Bachmair A, He Y (2009) Repression of the floral transition via histone H2B monoubiquitination. *Plant J* **57:** 522–533

Hon G, Wang W, Ren B (2009) Discovery and annotation of functional chromatin signatures in the human genome. *PLoS Comput Biol* **5:** e1000566

Jackson JP, Johnson L, Jasencakova Z, Zhang X, PerezBurgos L, Singh PB, Cheng X, Schubert I, Jenuwein T, Jacobsen SE (2004) Dimethylation of histone H3 lysine 9 is a critical mark for DNA methylation and gene silencing in Arabidopsis thaliana. *Chromosoma* **112:** 308–315

Jacob Y, Feng SH, LeBlanc CA, Bernatavichute YV, Stroud H, Cokus S, Johnson LM, Pellegrini M, Jacobsen SE, Michaels SD (2009) ATXR5 and ATXR6 are H3K27 monomethyltransferases required for chromatin structure and gene silencing. *Nat Struct Mol Biol* **16:** 763–796

Jacob Y, Stroud H, Leblanc C, Feng S, Zhuo L, Caro E, Hassel C, Gutierrez C, Michaels SD, Jacobsen SE (2010) Regulation of heterochromatic DNA replication by histone H3 lysine 27 methyltransferases. *Nature* **466:** 987–991

Jenuwein T, Allis CD (2001) Translating the histone code. *Science* **293:** 1074–1080

Jiang D, Wang Y, He Y (2008) Repression of FLOWERING LOCUS C and FLOWERING LOCUS T by the Arabidopsis Polycomb repressive complex 2 components. *PLoS One* **3:** e3404

Johnson L, Mollah S, Garcia BA, Muratore TL, Shabanowitz J, Hunt DF, Jacobsen SE (2004) Mass spectrometry analysis of Arabidopsis histone H3 reveals distinct combinations of post-translational modifications. *Nucleic Acids Res* **32:** 6511–6518

Kharchenko PV, Alekseyenko AA, Schwartz YB, Minoda A, Riddle NC, Ernst J, Sabo PJ, Larschan E, Gorchakov AA, Gu T, Linder-Basso D, Plachetka A, Shanower G, Tolstorukov MY, Luquette LJ, Xi R, Jung YL, Park RW, Bishop EP, Canfield TP *et al* (2011) Comprehensive analysis of the chromatin landscape in Drosophila melanogaster. *Nature* **471:** 480–485

Kohler C, Hennig L (2010) Regulation of cell identity by plant Polycomb and trithorax group proteins. *Curr Opin Genet Dev* **20:** 541–547

Kolasinska-Zwierz P, Down T, Latorre I, Liu T, Liu XS, Ahringer J (2009) Differential chromatin marking of introns and expressed exons by H3K36me3. *Nat Genet* **41:** 376–381

Kouzarides T (2007) Chromatin modifications and their function. *Cell* **128:** 693–705

Lee JS, Smith E, Shilatifard A (2010a) The language of histone crosstalk. *Cell* **142:** 682–685

Lee TJ, Pascuzzi PE, Settlage SB, Shultz RW, Tanurdzic M, Rabinowicz PD, Menges M, Zheng P, Main D, Murray JA, Sosinski B, Allen GC, Martienssen RA, Hanley-Bowdoin L, Vaughn MW, Thompson WF (2010b) Arabidopsis thaliana chromosome 4 replicates in two phases that correlate with chromatin state. *PLoS Genet* **6:** e1000982

Lippman Z, Gendrel AV, Black M, Vaughn MW, Dedhia N, McCombie WR, Lavine K, Mittal V, May B, Kasschau KD, Carrington JC, Doerge RW, Colot V, Martienssen R (2004) Role of transposable elements in heterochromatin and epigenetic control. *Nature* **430:** 471–476

Lippman Z, Gendrel AV, Colot V, Martienssen R (2005) Profiling DNA methylation patterns using genomic tiling microarrays. *Nat Methods* **2:** 219–224

Lister R, O'Malley RC, Tonti-Filippini J, Gregory BD, Berry CC, Millar AH, Ecker JR (2008) Highly integrated single-base resolution maps of the epigenome in Arabidopsis. *Cell* **133:** 523–536

Liu CL, Kaplan T, Kim M, Buratowski S, Schreiber SL, Friedman N, Rando OJ (2005) Single-nucleosome mapping of histone modifications in S. cerevisiae. *PLoS Biol* **3:** e328

Liu T, Rechtsteiner A, Egelhofer TA, Vielle A, Latorre I, Cheung MS, Ercan S, Ikegami K, Jensen M, Kolasinska-Zwierz P, Rosenbaum H, Shin H, Taing S, Takasaki T, Iniguez AL, Desai A, Dernburg AF, Lieb JD, Ahringer J, Strome S *et al* (2011) Broad chromosomal domains of histone modification patterns in C. elegans. *Genome Res* **21:** 227–236

Luco RF, Pan Q, Tominaga K, Blencowe BJ, Pereira-Smith OM, Misteli T (2010) Regulation of alternative splicing by histone modifications. *Science* **327:** 996–1000

Luo C, Lam E (2010) ANCORP: a high-resolution approach that generates distinct chromatin state models from multiple genome-wide datasets. *Plant J* **63:** 339–351

Martin-Magniette ML, Mary-Huard T, Berard C, Robin S (2008) ChIPmix: mixture model of regressions for two-color ChIP-chip analysis. *Bioinformatics* **24:** i181–i186

Minsky N, Shema E, Field Y, Schuster M, Segal E, Oren M (2008) Monoubiquitinated H2B is associated with the transcribed region of highly expressed genes in human cells. *Nat Cell Biol* **10:** 483–488

Oh S, Park S, van Nocker S (2008) Genic and global functions for Paf1C in chromatin modification and gene expression in Arabidopsis. *PLoS Genet* **4:** e1000077

Pekowska A, Benoukraf T, Ferrier P, Spicuglia S (2010) A unique H3K4me2 profile marks tissue-specific gene regulation. *Genome Res* **20:** 1493–1502

Rando OJ, Chang HY (2009) Genome-wide views of chromatin structure. *Annu Rev Biochem* **78:** 245–271

Riddle NC, Minoda A, Kharchenko PV, Alekseyenko AA, Schwartz YB, Tolstorukov MY, Gorchakov AA, Jaffe JD, Kennedy C, Linder-Basso D, Peach SE, Shanower G, Zheng H, Kuroda MI, Pirrotta V, Park PJ, Elgin SC, Karpen GH (2011) Plasticity in patterns of histone modifications and chromosomal proteins in Drosophila heterochromatin. *Genome Res* **21:** 147–163

Roudier F, Teixeira FK, Colot V (2009) Chromatin indexing in Arabidopsis: an epigenomic tale of tails and more. *Trends Genet* **25:** 511–517

Roy S, Ernst J, Kharchenko PV, Kheradpour P, Negre N, Eaton ML, Landolin JM, Bristow CA, Ma L, Lin MF, Washietl S, Arshinoff BI, Ay F, Meyer PE, Robine N, Washington NL, Di Stefano L, Berezikov E, Brown CD, Candeias R *et al* (2010) Identification of functional elements and regulatory circuits by Drosophila modENCODE. *Science* **330:** 1787–1797

Samson F, Brunaud V, Duchene S, De Oliveira Y, Caboche M, Lecharny A, Aubourg S (2004) FLAGdb++: a database for the functional analysis of the Arabidopsis genome. *Nucleic Acids Res* **32:** D347–D350

Schmid M, Davison TS, Henz SR, Pape UJ, Demar M, Vingron M, Scholkopf B, Weigel D, Lohmann JU (2005) A gene expression map of Arabidopsis thaliana development. *Nat Genet* **37:** 501–506

Schmitz RJ, Tamada Y, Doyle MR, Zhang X, Amasino RM (2009) Histone H2B deubiquitination is required for transcriptional activation of FLOWERING LOCUS C and for proper control of flowering in Arabidopsis. *Plant Physiol* **149:** 1196–1204

Schulze JM, Jackson J, Nakanishi S, Gardner JM, Hentrich T, Haug J, Johnston M, Jaspersen SL, Kobor MS, Shilatifard A (2009) Linking cell cycle to histone modifications: SBF and H2B monoubiquitination machinery and cell-cycle regulation of H3K79 dimethylation. *Mol Cell* **35:** 626–641

Sims III RJ, Reinberg D (2008) Is there a code embedded in proteins that is based on post-translational modifications? *Nat Rev Mol Cell Biol* **9:** 815–820

Squazzo SL, O'Geen H, Komashko VM, Krig SR, Jin VX, Jang SW, Margueron R, Reinberg D, Green R, Farnham PJ (2006) Suz12 binds to silenced regions of the genome in a cell-type-specific manner. *Genome Res* **16:** 890–900

Sridhar VV, Kapoor A, Zhang K, Zhu J, Zhou T, Hasegawa PM, Bressan RA, Zhu JK (2007) Control of DNA methylation and heterochromatic silencing by histone H2B deubiquitination. *Nature* **447:** 735–738

Strahl BD, Allis CD (2000) The language of covalent histone modifications. *Nature* **403:** 41–45

Suganuma T, Workman JL (2008) Crosstalk among histone modifications. *Cell* **135:** 604–607

Tanurdzic M, Vaughn MW, Jiang H, Lee TJ, Slotkin RK, Sosinski B, Thompson WF, Doerge RW, Martienssen RA (2008) Epigenomic consequences of immortalized plant cell suspension culture. *PLoS Biol* **6:** 2880–2895

Turck F, Roudier F, Farrona S, Martin-Magniette ML, Guillaume E, Buisine N, Gagnot S, Martienssen RA, Coupland G, Colot V (2007) Arabidopsis TFL2/LHP1 specifically associates with genes marked by trimethylation of histone H3 lysine 27. *PLoS Genet* **3:** e86

Vakoc CR, Mandat SA, Olenchock BA, Blobel GA (2005) Histone H3 lysine 9 methylation and HP1gamma are associated with transcription elongation through mammalian chromatin. *Mol Cell* **19:** 381–391

Vakoc CR, Sachdeva MM, Wang H, Blobel GA (2006) Profile of histone lysine methylation across transcribed mammalian chromatin. *Mol Cell Biol* **26:** 9185–9195

Vaughn MW, Tanurd Ic M, Lippman Z, Jiang H, Carrasquillo R, Rabinowicz PD, Dedhia N, McCombie WR, Agier N, Bulski A, Colot V, Doerge RW, Martienssen RA (2007) Epigenetic natural variation in Arabidopsis thaliana. *PLoS Biol* **5:** e174

Wang Z, Schones DE, Zhao K (2009) Characterization of human epigenomes. *Curr Opin Genet Dev* **19:** 127–134

Wang Z, Zang C, Rosenfeld JA, Schones DE, Barski A, Cuddapah S, Cui K, Roh TY, Peng W, Zhang MQ, Zhao K (2008) Combinatorial patterns of histone acetylations and methylations in the human genome. *Nat Genet* **40:** 897–903

Weake VM, Workman JL (2008) Histone ubiquitination: triggering gene activity. *Mol Cell* **29:** 653–663

Williams SK, Truong D, Tyler JK (2008) Acetylation in the globular core of histone H3 on lysine-56 promotes chromatin disassembly during transcriptional activation. *Proc Natl Acad Sci USA* **105:** 9000–9005

Zhang X, Bernatavichute Y, Cokus S, Pellegrini M, Jacobsen S (2009) Genome-wide analysis of mono-, di- and trimethylation of histone H3 lysine 4 in Arabidopsis thaliana. *Genome Biol* **10:** R62

Zhang X, Clarenz O, Cokus S, Bernatavichute YV, Pellegrini M, Goodrich J, Jacobsen SE (2007) Whole-genome analysis of histone H3 lysine 27 trimethylation in Arabidopsis. *PLoS Biol* **5:** e129

Zhang XY, Yazaki J, Sundaresan A, Cokus S, Chan SWL, Chen HM, Henderson IR, Shinn P, Pellegrini M, Jacobsen SE, Ecker JR (2006) Genome-wide high-resolution mapping and functional analysis of DNA methylation in Arabidopsis. *Cell* **126:** 1189–1201

Zhou VW, Goren A, Bernstein BE (2011) Charting histone modifications and the functional organization of mammalian genomes. *Nat Rev Genet* **12:** 7–18

Zilberman D, Coleman-Derr D, Ballinger T, Henikoff S (2008) Histone H2A.Z and DNA methylation are mutually antagonistic chromatin marks. *Nature* **456:** 125–129

Zilberman D, Gehring M, Tran RK, Ballinger T, Henikoff S (2006) Genome-wide analysis of Arabidopsis thaliana DNA methylation uncovers an interdependence between methylation and transcription. *Nat Genet* **39:** 61–69

Supplementary Figure legends

Supplementary Figure 1. Distribution of chromatin modifications along chromosome 4. Average enrichment level is shown for each chromatin modification in consecutive 50 kb windows. Green shading indicates heterochromatic regions. K: heterochromatic knob. The non-sequenced part of the centromere (C) is represented by the vertical black line.

Supplementary Figure 2. Epigenomic landscape at a euchromatin/heterochromatin transition. This partial view covers about 1% of the Arabidopsis genome. IP/INPUT ratios reporting significant enrichment are shown in dark colours. Distribution patterns of the four chromatin states CS1-4 are shown below the composite epigenomic map and using the same color code as in Figure 2.

Supplementary Figure 3. Combinatorial patterns of chromatin modifications over chromosome 4. (A) Tile-wide analysis of the co-occurrence of chromatin modifications. The y axis indicates the percentage of observed combinations between the 12 chromatin modifications analyzed. The x axis indicates the number of tiles associated with each combination. (B) Distribution of annotated features among the ten most frequent combinations.

Supplementary Figure 4. Length distribution of chromatin domains defined by the four chromatin states CS1-4. Grey color corresponds to tiles that cannot be unambiguously assigned to any of these four chromatin states. Arrows indicate median domain length.

Supplementary Figure 5. Analysis of antibody specificity using peptide competition assays on western blots of Arabidopsis nuclear extracts. Amounts of competing peptides are indicated.

Supplementary Table legends

Supplementary Table I. Chromosome-wide distribution of chromatin modifications over euchromatic and heterochromatic regions. Heterochromatin boundaries are as indicated in Figure 1.

Supplementary Table II. Domain organization of chromatin modifications and associated annotation on chromosome 4. Number of annotated genes and transposable elements (TAIR8) overlapping chromatin domains by at least 50 bp is indicated.

Supplementary Table III. Lists of marked genes and transposable elements detected in seedlings using the chromosome 4 tiling array.

Supplementary Table IV. Lists of marked genes and transposable elements detected in whole seedlings and roots using the whole-genome tiling array.

Supplementary Table V. Mean pairwise association values between the 12 chromatin modifications along chromosome 4 over all enriched tiles and over genic enriched tiles.

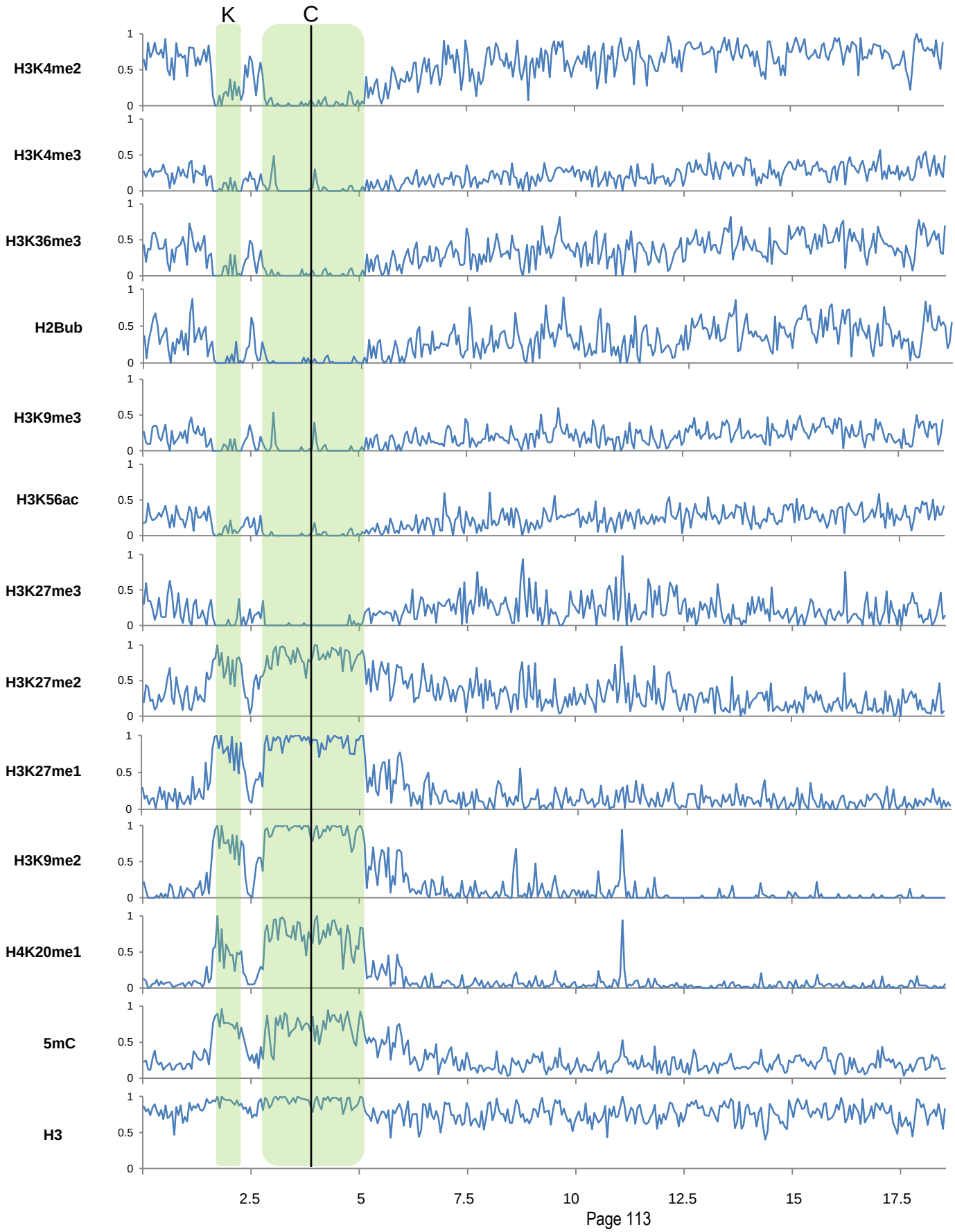
Supplementary Table VI. Gene Ontology analyses for each list of genes marked by a given chromatin modification on the chromosome 4 tiling array.

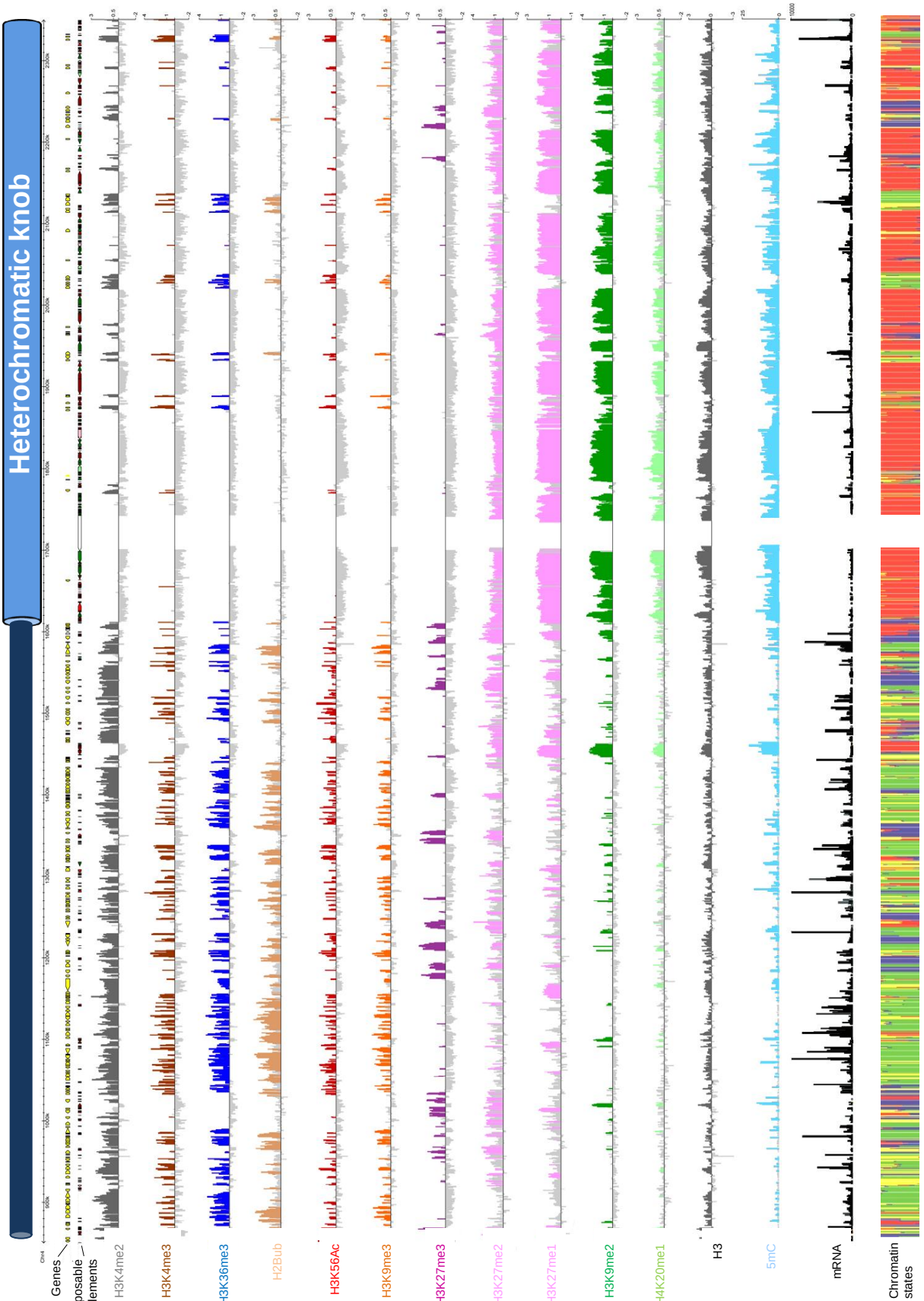
Supplementary Table VII. Gene Ontology analyses for each list of genes marked by a given chromatin modification on the whole genome tiling array.

Supplementary Table VIII. Lists of the different classes of genes co-marked by H3K4me3 and H3K27me3 in whole seedlings according to their marking in roots and aerial parts.

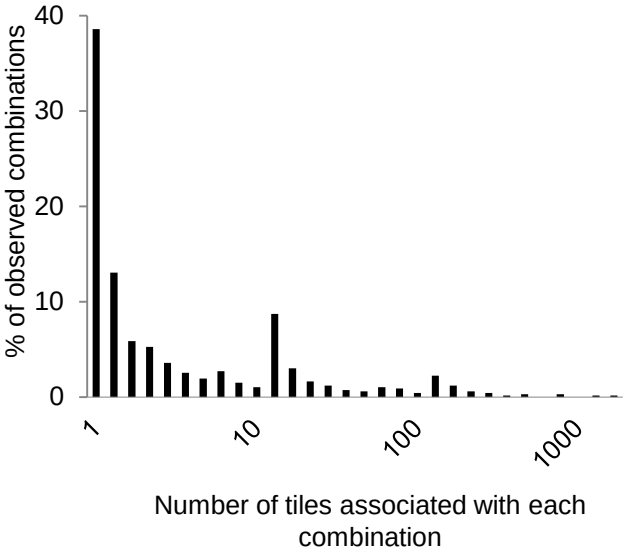
Supplementary Table IX. List of antibodies used in this study. Results of independent specificity assays are indicated and include peptide dot blot analysis (Egelhofer et al. 2010; http://compbio.med.harvard.edu/antibodies/) and peptide competition assays on western blots of Arabidopsis nuclear extracts (this study).

Supplementary Figure 1

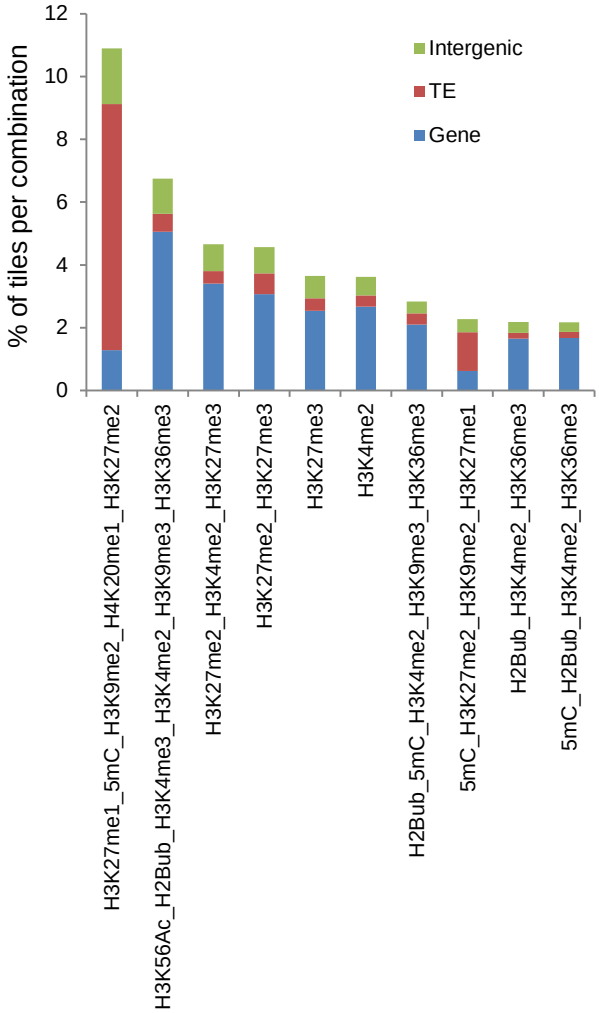


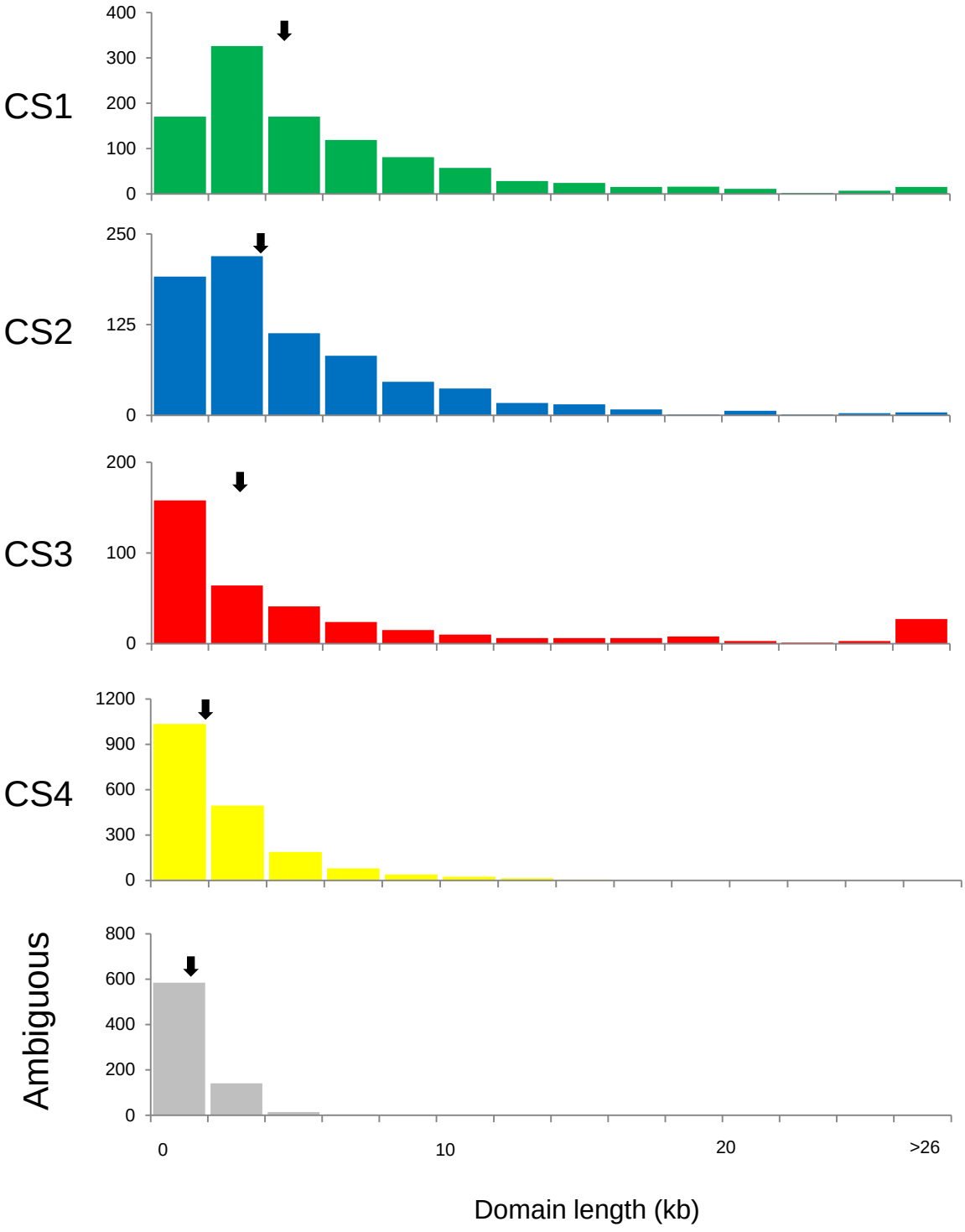

Heterochromatic knob
Genes
Transposable elements
H3K4me2
H3K4me3
H3K36me3
H2Bub
H3K56Ac
H3K9me3
H3K27me3
H3K27me2
H3K27me1
H3K9me2
H4K20me1
H3
5mC
mRNA
Chromatin states

A

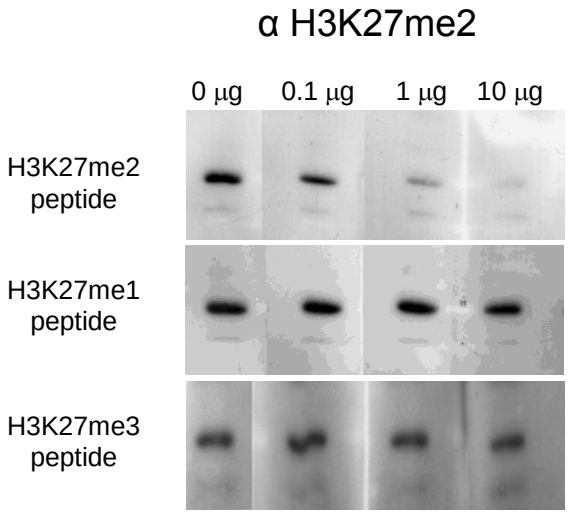

% of observed combinations
40
30
20
10
0
1
10
100
1000
Number of tiles associated with each combination

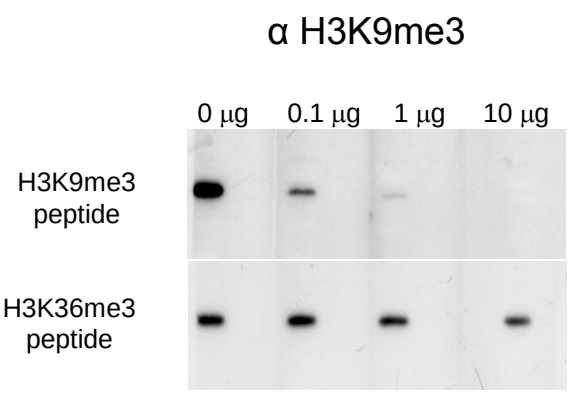
B


% of tiles per combination
12
10
8
6
4
2
0
Intergenic
TE
Gene
H3K27me1_5mC_H3K9me2_H4K20me1_H3K27me2
H3K56Ac_H2Bub_H3K4me3_H3K4me2_H3K9me3_H3K36me3
H3K27me2_H3K4me2_H3K27me3
H3K27me2_H3K27me3
H3K27me3
H3K4me2
H2Bub_5mC_H3K4me2_H3K9me3_H3K36me3
5mC_H3K27me2_H3K9me2_H3K27me1
H2Bub_H3K4me2_H3K36me3
5mC_H2Bub_H3K4me2_H3K36me3


CS1
400
300
200
100
0
CS2
250
125
0
CS3
200
100
0
CS4
1200
900
600
300
0
Ambiguous
800
600
400
200
0
0
10
20
>26
Domain length (kb)

Supplementary Figure 5





Supplementary Table I

		Chr4	H3	H3K4me2	H3K36me3	H2Bub	H3K56ac	H3K4me3	H3K9me3	H3K27me3	H3K27me2	H3K27me1	5mC	H3K9me2	H4K20me1
Genomic space covered (bp)	**Total**	1.85E+07	1.62E+07	1.09E+07	6.51E+06	5.72E+06	4.90E+06	4.26E+06	3.88E+06	3.84E+06	7.92E+06	5.33E+06	6.68E+06	4.32E+06	3.16E+06
	Euchromatin	1.55E+07	1.35E+07	1.08E+07	6.42E+06	5.66E+06	4.81E+06	4.08E+06	3.76E+06	3.79E+06	5.26E+06	2.66E+06	4.24E+06	1.58E+06	8.47E+05
	Heterochromatin	3.05E+06	2.63E+06	1.81E+05	9.93E+04	6.13E+04	9.06E+04	1.78E+05	1.27E+05	5.49E+04	2.66E+06	2.68E+06	2.44E+06	2.73E+06	2.32E+06
Relative distribution over Chr4 (%)	**Euchromatin**	83.6	73.1	58.1	34.6	30.5	25.9	22.0	20.3	20.4	28.4	14.3	22.9	8.5	4.6
	Heterochromatin	16.4	14.2	1.0	0.5	0.3	0.5	1.0	0.7	0.3	14.4	14.4	13.1	14.7	12.5
Relative coverage (%)	**Euchromatin**	100.0	87.5	69.5	41.4	36.5	31.0	26.4	24.3	24.5	33.9	17.2	27.4	10.2	5.5
	Heterochromatin	100.0	86.3	5.9	3.3	2.0	3.0	5.8	4.2	1.8	87.4	87.8	79.9	89.7	76.1
Relative distribution between compartments (%)	**Euchromatin**	83.6	83.7	98.4	98.5	98.9	98.2	95.8	96.7	98.6	66.4	49.8	63.5	36.7	26.7
	Heterochromatin	16.4	16.3	1.6	1.5	1.1	1.8	4.2	3.3	1.4	33.6	50.2	36.5	63.3	73.3

Supplementary Table II

	H3	H3K4me2	H3K4me3	H3K36me3	H3K56ac	H3K9me3	H2Bub	H3K27me3	H3K27me2	H3K27me1	H3K9me2	H4K20me1	5mC
Nb of marked domains	ND	898	1163	1042	1122	1000	838	638	908	610	306	337	ND
Median length of marked domains in euchromatin (bp)	ND	7272	2804	4527	3242	2995	4816	4283	4053	3109	3455	2461	1873
Nb of marked genes (total=4103)	3890 (95%)	3646 (87%)	2136 (52%)	2432 (59%)	2319 (57%)	1643 (40%)	2048 (50%)	1190 (29%)	1472 (36%)	704 (17%)	242 (6%)	182 (4%)	1574 (38%)
Nb of marked TEs (total= 3191)	2605 (82%)	293 (9%)	82 (3%)	72 (2%)	108 (3%)	62 (2%)	39 (1%)	523 (16%)	2473 (77%)	1859 (58%)	2143 (67%)	1387 (43%)	1866 (58%)

Supplementary Table V

All Tiles

H3K4me2	H3K4me3	H3K9me2	H3K9me3	H3K27me2	H3K27me3	H3K36me3	H3K56ac	5mC	H4K20me1	H2Bub	H3K27me1	
	0.93	0.09	0.96	0.35	0.54	0.97	0.96	0.43	0.18	0.94	0.31	*H3K4me2*
0.34		0.03	0.66	0.06	0.06	0.54	0.67	0.10	0.05	0.46	0.04	*H3K4me3*
0.04	0.03		0.03	0.47	0.08	0.02	0.03	0.51	0.85	0.02	0.66	*H3K9me2*
0.32	0.62	0.03		0.03	0.02	0.57	0.47	0.18	0.08	0.56	0.02	*H3K9me3*
0.24	0.11	0.82	0.06		0.69	0.06	0.19	0.55	0.87	0.05	0.69	*H3K27me2*
0.20	0.06	0.07	0.02	0.37		0.03	0.11	0.08	0.10	0.03	0.06	*H3K27me3*
0.56	0.86	0.03	0.97	0.05	0.05		0.75	0.31	0.09	0.85	0.09	*H3K36me3*
0.38	0.73	0.03	0.55	0.11	0.12	0.51		0.12	0.07	0.43	0.11	*H3K56ac*
0.24	0.15	0.72	0.30	0.45	0.11	0.30	0.17		0.76	0.33	0.61	*5mC*
0.05	0.04	0.63	0.07	0.37	0.08	0.05	0.05	0.40		0.04	0.47	*H4K20me1*
0.49	0.67	0.03	0.88	0.04	0.04	0.77	0.57	0.31	0.08		0.05	*H2Bub*
0.15	0.06	0.84	0.04	0.51	0.09	0.07	0.13	0.54	0.81	0.05		*H3K27me1*

Genic Tiles

H3K4me2	H3K4me3	H3K9me2	H3K9me3	H3K27me2	H3K27me3	H3K36me3	H3K56ac	5mC	H4K20me1	H2Bub	H3K27me1	
	0.95	0.25	0.98	0.53	0.56	0.97	0.96	0.69	0.50	0.97	0.64	*H3K4me2*
0.34		0.05	0.66	0.08	0.06	0.54	0.67	0.16	0.12	0.47	0.08	*H3K4me3*
0.03	0.02		0.02	0.22	0.07	0.02	0.03	0.23	0.58	0.02	0.36	*H3K9me2*
0.33	0.61	0.05		0.04	0.02	0.56	0.47	0.29	0.20	0.57	0.04	*H3K9me3*
0.23	0.10	0.75	0.05		0.68	0.06	0.19	0.32	0.70	0.05	0.55	*H3K27me2*
0.19	0.06	0.18	0.02	0.55		0.03	0.11	0.11	0.25	0.03	0.12	*H3K27me3*
0.57	0.87	0.08	0.98	0.08	0.06		0.76	0.51	0.27	0.87	0.18	*H3K36me3*
0.37	0.73	0.08	0.55	0.17	0.12	0.51		0.18	0.18	0.44	0.22	*H3K56ac*
0.23	0.15	0.60	0.29	0.25	0.10	0.30	0.16		0.59	0.33	0.43	*5mC*
0.05	0.03	0.45	0.06	0.16	0.07	0.05	0.05	0.17		0.04	0.22	*H4K20me1*
0.50	0.68	0.07	0.89	0.06	0.04	0.77	0.58	0.51	0.23		0.10	*H2Bub*
0.15	0.05	0.66	0.02	0.30	0.08	0.07	0.13	0.30	0.52	0.04		*H3K27me1*

THE EMBO JOURNAL

Manuscript EMBO-2010-76582

Integrative epigenomic mapping defines four major chromatin states in Arabidopsis

François Roudier, Ikhlak Ahmed, Caroline Bérard, Alexis Sarazin, Tristan Mary-Huard, Sandra Cortijo, Daniel Bouyer, Erwann Caillieux, Evelyne Duvernois-Berthet, Liza Al-Shikhley, Laurène Giraut, Barbara Després, Stéphanie Drevensek, Frédy Barneche, Sandra Dèrozier, Véronique Brunaud, Sébastien Aubourg, Arp Schnittger, Chris Bowler, Marie-Laure Martin-Magniette, Stéphane Robin, Michel Caboche and Vincent Colot

Corresponding author: Vincent Colot, IBENS

Review timeline:

Submission date:	23 November 2010
Editorial Decision:	21 December 2010
Revision received:	09 March 2011
Editorial Decision:	10 March 2011
Accepted:	10 March 2011

Transaction Report:

(Note: With the exception of the correction of typographical or spelling errors that could be a source of ambiguity, letters and reports are not edited. The original formatting of letters and referee reports may not be reflected in this compilation.)

1st Editorial Decision 21 December 2010

Thank you for submitting your manuscript for consideration by The EMBO Journal. It has been now been evaluated by three referees and I enclose their reports below. As you will see the referees find the analysis of the organization of the Arabidopsis genome to be interesting and important and recommend publication in The EMBO Journal once several issues are clarified, this includes a more direct comparison with previous studies including the recent study from the van Steensel lab. Given the interest from the referees should you be able to address these issues, we would be happy to consider a revised manuscript.

I should remind you that it is EMBO Journal policy to allow a single round of revision only and that, therefore, acceptance or rejection of the manuscript will depend on the completeness of your responses included in the next, final version of the manuscript. When you submit a revised version to the EMBO Journal, please make sure you upload a letter of response to the referees' comments. Please note that when preparing your letter of response to the referees' comments that this will form part of the Review Process File, and will therefore be available online to the community. For more details on our Transparent Editorial Process initiative, please visit our website: http://www.nature.com/emboj/about/process.html

Thank you for the opportunity to consider your work for publication. I look forward to your revision.

Yours sincerely,

Editor
The EMBO Journal

REFEREE COMMENTS

Referee #1 (Remarks to the Author):

This manuscript reports the systematic genome-wide mapping of a broad set of 11 histone marks plus DNA methylation in Arabidopsis seedlings. This unique dataset enabled the authors to conduct a detailed integrative analysis to understand the fine patterns and relationships among these marks. Among others, the authors find identify major combinations of marks (or actually three; the fourth lacks essentially all tested marks).

The strength of this manuscript lies in the thorough and systematic approach, which to my knowledge has not yet been done in any plant species. The data and all analyses are very robust. While the manuscript does not report many surprises, these kind of systematic surveys are very important for the "big picture". I therefore recommend (in principle) that it be published in EMBO Journal.

Major points:

- The data are restricted to histone marks (and 5mC), while none of hundreds of chromatin-associated proteins are considered here. I therefore feel that the claim to have identified "major chromatin types" (title, abstract) is premature. Twelve marks is also not enough to be sure that the entire 'space' of chromatin types has been sampled. It therefore would be more appropriate/accurate to use the term "major combinations of histone marks".

- The authors chose "well-characterized antibodies". Please substantiate this statement by adding a supplementary table that lists for each antibody how its specificity was confirmed in previous studies (peptide blots? peptide competition? etc), including relevant citations. Furthermore, it seems that the authors don't fully trust the specificity of the H3K27me2 and H3K9me3 antibodies. While I very much appreciate the authors' honesty about these doubts, I recommend that the most obvious candidates for crossreactivity (H3K27me1 and H3K27me3 for the H3K27me2 antibody, and H3K36me3 for the H3K9me3 antibody) are directly tested by peptide spotblot.

Minor points:

- It seems to me that CT4 may correspond to "black" chromatin as identified by Filion et al. Black chromatin also appears to lack known histone marks (so far).

- Fig1C: ordering along the horizontal axis seems arbitrary. Clustering is better visualized if both axes are ordered the same way.

- Fig4D: add grey scale bar for expression level

Referee #2 (Remarks to the Author):

The authors profiled eight histone modifications (H3K4me2 and 3, H3K27me1 and 2, H3K36me3, H3K56ac, H4K20me1, and H2Bub) of Arabidopsis seedlings, and took a systematic analysis both genome-wise and gene-wise, in combination with four other epigenetic maps (H3K9me2 and 3,

H3K27me3 and DNA methylation) obtained before. The authors identified the 12 epigenetic marks as 2 distinct groups by pair-wise association, distributing mainly in euchromatic regions or heterochromatic regions. They also described the biased association of some of the epigenetic marks with genes in terms of the transcription activity and gene length, as well as their spatial distribution relative to the gene annotations. They further identified 4 major chromatin types by c-means clustering based on the associations between the marks, which represent the heterchromatic regions, the actively transcribed chromatins, the repressed chromatins by PcG proteins, and ambiguous chromatins. Apart from the extended domains of heterchromatic regions, the euchromatic arms are largely interspersed by the 4 major types of chromatins. Finally, the authors discussed the chromatin composition in relation to expression specificity, focusing on the apparently double-indexed genes by both H3K4me3 and H3K27me3.

The data available to the community is valuable, and it's interesting to see the distinct types of chromatin compositions.

Major points:

- The value of the data sets as a ressource for the community is limited because most data are based on a tiling array that covers only one of the 5 chromosomes and has a relatively low resolution (~1kb). Having said that, it should be noted that the major conclusions are justified.

- Results part 1, 2 and 4 should be condensed. Given the over-interpretation often found with apparent bivalent chromatin in plants, the corresponding section of the manuscript is refreshing, however it does not fit well with the other parts. A much shortened version of this section would fit better into the discussion.

- The authors mentioned sub-types existing beneath the 4 major types. It would be very interesting to see the more detailed categorization.

- The section on gene length effects remained unclear to the reviewer and requires major re-writing. Examples of statements that confused the reviewer are "In particular,whereas H2Bub, H3K36me3 and H3K27me1 show similar distribution over the transcribed region of genes independently of their length, H2Bub and H3K36me3 are poorly detected over genes smaller than ~1.5 kb." and "Cytosine methylation is not frequently found over small genes either, although H3K27me3 deposition is somewhat biased towards smaller genes"

Minor points:

- In Figure 3A, how is Expression percentiles defined? Are all gene uniformly distributed in each percentile? If so then why would the "All genes" dashed line show an non-uniform distribution?

- On page 8, paragraph 2: "two of these four types change little over k values..." which two of the types change and which other don't? Explain clearly or do not mention it.

- On page 12, paragraph 2: "H3K79me3...a chromatin mark that is apparently missing in (Zhang et al, 2007)". Missing in Arabidopsis?

- "about 10% of the tiles do not show association with any of the chromatin modifications" - Failure to detect signals can always be a technical issue. The section describing and discussing missing signals should be shortened and phrased more carefully or even removed.

Referee #3 (Remarks to the Author):

The manuscript of Roudier et al. describes the genome wide view of 10 different chromatin marks, including selected histone marks and DNA methylation marks, and thus provides an organizational map of the Arabidopsis epigenome. Moreover, the genome wide maps of eight of these marks were generated for the first time in this study and these data will provide a very important resource for the epigenetic community, especially for plant epigeneticists. The authors discovered that the Arabidopsis epigenome can be confined into four major types of chromatin, which are not only

clearly associated with specific combinations of epigenetic marks but can also be well distinguished as different functional elements.

This is a very important and clearly presented study that provides readers with an essential and novel message.

There is only one minor point that should be addressed prior publication: the authors should provide a thorough discussion of their presented results and conclusions in the context of the similar, previously published studies performed for other organisms. This is especially important for the very recent discovery of five chromatin domains in Drosophila (Filion et al Cell 143 , 2010). Although the study of Filion at al,. and also others, are mentioned in the discussion the significance of the similarities and differences in the results presented in these papers with the results of the present work needs to be systematically discussed and interpreted. Such discussion including comparison of experimental approaches, number and sizes chromatin domains, their chromosomal distribution and functional relationships across various organisms would make this paper much more attractive for readers. Obviously this requires serious rewriting of the manuscript, with mentioning of the other studies already in the introduction, and a comparison of results throughout the manuscript: however, this extra effort would pay off in that it would allow the reader to gain a more global view of chromatin organization across different species.

1st Revision - authors' response 09 March 2011

Point by point response:

Referee #1:

Major points:

- The data are restricted to histone marks (and 5mC), while none of hundreds of chromatin-associated proteins are considered here. I therefore feel that the claim to have identified "major chromatin types" (title, abstract) is premature. Twelve marks is also not enough to be sure that the entire 'space' of chromatin types has been sampled. It therefore would be more appropriate/accurate to use the term "major combinations of histone marks".

We would like to emphasize that this study aimed at identifying prevalent chromatin states based on combinations of 12 chromatin marks. Although the 4 main chromatin states identified cover ~95% of the genome ‘space’, we agree with Referee #1 that our dataset cannot ascertain that the entire set of chromatin types has been sampled. To avoid confusion with the denomination ”chromatin types” used to describe the five principal combinations of chromatin-associated proteins identified via a much larger integrative analysis in Drosophila (Filion et al, 2010), we now use the denomination “chromatin states” (CS) and we clearly points in the Discussion to the possibility of additional such states in Arabidopsis. Nonetheless, we wish to stress that we have not only identified four major combinations of chromatin marks but have also determined that they have distinct functional properties.

- The authors chose "well-characterized antibodies". Please substantiate this statement by adding a supplementary table that lists for each antibody how its specificity was confirmed in previous studies (peptide blots? peptide competition? etc), including relevant citations. Furthermore, it seems that the authors don't fully trust the specificity of the H3K27me2 and H3K9me3 antibodies. While I very much appreciate the authors' honesty about these doubts, I recommend that the most obvious candidates for crossreactivity (H3K27me1 and H3K27me3 for the H3K27me2 antibody, and H3K36me3 for the H3K9me3 antibody) are directly tested by peptide spotblot.

We have removed this sentence form the text and we now provide peptide competition assays as recommended (Supplementary Figure 5). We also provide in a table (Supplementary Table IX) available information on commercial antibodies. Finally, we address more specifically the issue of antibody specificity in the Discussion.

Minor points:

- It seems to me that CT4 may correspond to "black" chromatin as identified by Filion et al. Black chromatin also appears to lack known histone marks (so far).

We now provide a more detailed comparison with the Filion et al article and mention the similarity between BLACK and what we now call CS4.

- Fig1C: ordering along the horizontal axis seems arbitrary. Clustering is better visualized if both axes are ordered the same way.

We feel that the directionality of association provides important information given the difference in abundancy between the chromatin marks analyzed. We therefore have included clustering information for the two axes and have amended figures legends accordingly.

- Fig4D: add grey scale bar for expression level

We have added a grey scale bar has been.

Referee #2:

The authors profiled eight histone modifications (H3K4me2 and 3, H3K27me1 and 2, H3K36me3, H3K56ac, H4K20me1, and H2Bub) of Arabidopsis seedlings, and took a systematic analysis both genome-wise and gene-wise, in combination with four other epigenetic maps (H3K9me2 and 3, H3K27me3 and DNA methylation) obtained before. The authors identified the 12 epigenetic marks as 2 distinct groups by pair-wise association, distributing mainly in euchromatic regions or heterochromatic regions. They also described the biased association of some of the epigenetic marks with genes in terms of the transcription activity and gene length, as well as their spatial distribution relative to the gene annotations. They further identified 4 major chromatin types by c-means clustering based on the associations between the marks, which represent the heterchromatic regions, the actively transcribed chromatins, the repressed chromatins by PcG proteins, and ambiguous chromatins. Apart from the extended domains of heterchromatic regions, the euchromatic arms are largely interspersed by the 4 major types of chromatins. Finally, the authors discussed the chromatin composition in relation to expression specificity, focusing on the apparently double-indexed genes by both H3K4me3 and H3K27me3.

The data available to the community is valuable, and it's interesting to see the distinct types of chromatin compositions.

Major points:

- The value of the data sets as a ressource for the community is limited because most data are based on a tiling array that covers only one of the 5 chromosomes and has a relatively low resolution (~1kb). Having said that, it should be noted that the major conclusions are justified.

We acknowledge the fact that referee #2 agrees that the major conclusions reached using the chromosome 4 tiling array are justified. In addition, we would like to stress that the genome-wide data that we present are for 7 of the 12 marks, including H3K36me3 and H2ub, which have not been analyzed at the genome scale before. Also, all of these these data will be publicly available at http://epigara.biologie.ens.fr/index.html , together with relevant epigenomic data published by other groups. This should therefore represent an important epigenomic resource for the community.

- Results part 1, 2 and 4 should be condensed. Given the over-interpretation often found with apparent bivalent chromatin in plants, the corresponding section of the manuscript is refreshing, however it does not fit well with the other parts. A much shortened version of this section would fit better into the discussion.

We have reorganized the Results part into 3 sections. Original parts 2 and 4 (which dealt with chromatin indexing of genes) have now been merged into a single section (section 3), which is much

more condensed. Original parts 1 and 3 (which dealt with the individual and combinatorial analyses of the marks across all genomic sequences) are now sections 1 and 2, and are essentially unchanged. This reorganization makes the results section much easier to read.

- The authors mentioned sub-types existing beneath the 4 major types. It would be very interesting to see the more detailed categorization.

Our intention was simply to mention that additional chromatin states likely exist. As splitting of CS1 by increasing the number of k in c-means clustering is neither robust nor biologically interpretable, we have decided to remove mention of possible additional subtypes/refined states from the Results section. The notion that the broad-level organization provided by our work might be further refined is now addressed more thoroughly in the Discussion.

- The section on gene length effects remained unclear to the reviewer and requires major re-writing. Examples of statements that confused the reviewer are "In particular,whereas H2Bub, H3K36me3 and H3K27me1 show similar distribution over the transcribed region of genes independently of their length, H2Bub and H3K36me3 are poorly detected over genes smaller than ~1.5 kb." and "Cytosine methylation is not frequently found over small genes either, although H3K27me3 deposition is somewhat biased towards smaller genes"

This part has been re-written and significantly shortened.

Minor points:

- In Figure 3A, how is Expression percentiles defined? Are all gene uniformly distributed in each percentile? If so then why would the "All genes" dashed line show an non-uniform distribution?

Whole seedling transcriptome data were retrieved from Schmid et al. (2005) and genes were binned into 20 expression percentiles according to their absolute expression values. For each bin, the number of genes marked by a given chromatin modification was scored and represented as a percentage of all the genes marked by this modification. The dashed line represents the distribution of all annotated genes on chromosome 4 across all expression percentiles.
Legend of Figure 4 has been amended and the procedure is now described in Materials and Methods.

- On page 8, paragraph 2: "two of these four types change little over k values..." which two of the types change and which other don't? Explain clearly or do not mention it.

See above.

- On page 12, paragraph 2: "H3K79me3...a chromatin mark that is apparently missing in (Zhang et al, 2007)". Missing in Arabidopsis?

Yes, this has been corrected.

- "about 10% of the tiles do not show association with any of the chromatin modifications" - Failure to detect signals can always be a technical issue. The section describing and discussing missing signals should be shortened and phrased more carefully or even removed.

This part has been removed.

Referee #3 (Remarks to the Author):

The manuscript of Roudier et al. describes the genome wide view of 10 different chromatin marks, including selected histone marks and DNA methylation marks, and thus provides an organizational map of the Arabidopsis epigenome. Moreover, the genome wide maps of eight of these marks were generated for the first time in this study and these data will provide a very important resource for the epigenetic community, especially for plant epigeneticists. The authors discovered that the Arabidopsis epigenome can be confined into four major types of chromatin, which are not only

clearly associated with specific combinations of epigenetic marks but can also be well distinguished as different functional elements.

This is a very important and clearly presented study that provides readers with an essential and novel message.

There is only one minor point that should be addressed prior publication: the authors should provide a thorough discussion of their presented results and conclusions in the context of the similar, previously published studies performed for other organisms. This is especially important for the very recent discovery of five chromatin domains in Drosophila (Filion et al Cell 143 , 2010). Although the study of Filion at al,. and also others, are mentioned in the discussion the significance of the similarities and differences in the results presented in these papers with the results of the present work needs to be systematically discussed and interpreted. Such discussion including comparison of experimental approaches, number and sizes chromatin domains, their chromosomal distribution and functional relationships across various organisms would make this paper much more attractive for readers. Obviously this requires serious rewriting of the manuscript, with mentioning of the other studies already in the introduction, and a comparison of results throughout the manuscript: however, this extra effort would pay off in that it would allow the reader to gain a more global view of chromatin organization across different species.

We agree with referee#3 that the comparison of the chromatin landscape in different organisms is an interesting topic. We now provide such comparisons in the Abstract, Introduction and Discussion, where we point in some detail to similarities and differences in chromatin indexing between Arabidopsis and metazoans, referring notably to the five recent papers of Kharchenko et al, 2010; Riddle et al, 2010; The modENCODE Consortium/Roy et al, 2010; Ernst & Kellis, 2010 and Filion et al. 2010. However, we have not extended these comparisons to the Results section, as we feel as this would have unnecessarily lengthen the manuscript.

2nd Editorial Decision 10 March 2011

I have read through your revised manuscript and I find that you have satisfactorily incorporated all the changes suggested by the referees, including an interesting comparison with the chromatin organisation in other organisms. I am happy to accept the manuscript for publication in The EMBO Journal. I believe it will make a great contibution to journal.

Yours sincerely,

Editor
The EMBO Journal

Additional Methods

Combinatorial Analysis

To identify groups of correlated marks and distinguish them among different annotated features like transposon sequences, genes and intergenic regions, a combinatorial analysis was performed on the entire set of 12 chromatin marks and all possible combinations were examined.

A combination is a way of selecting several parameters out of a larger group where order does not matter (unlike permutations). For example, given three balls of colours red, green and blue, there are three possible ways in which combinations of two can be drawn from this set i.e., red-green, red-blue and green-blue. Now, given a set of n elements, the total number of possible combinations for k objects where $k <= n$ is given by:

"n choose k" $$C_k^n = \frac{n!}{k!(n-k)!}$$

Each of our chromatin states defined by a given location on the chromosome can contain either none or one, two, three, …, up all the twelve chromatin marks taken in this study. Therefore, the total number of all possible combinations for 12 chromatin marks can be given by:

$$\sum_{k=0}^{n} C_k^n$$

Hence for each of the k values, number of combinations is:

k	0	1	2	3	4	5	6	7	8	9	10	11	12
C_k^n	1	12	66	220	495	792	924	792	495	220	66	12	1

Total=4096

Therefore,

$$\sum_{k=0}^{12} C_k^{12} = 4096$$

The table below contains the top 100 combinations of chromatin marks that represent ~90% of the *Arabidopsis* genome and the contribution of each combination towards four chromatin states as represented by CS1, CS2, CS3 and CS4. The combinations were computed using Perl (http://www.perl.org/) and the Math::Combinatorics package from CPAN (http://www.cpan.org/).

S.No	Marker_comb	%genome-occurrence	%Genes	%TEs	%CS1	%CS2	%CS3	%CS4
1	No-mark	9.63	3.81	10.77	0.00	0.00	0.00	39.16
2	H3K27me1_5mC_H3K9me2_H4K20me1_H3K27me2	9.85	2.09	31.34	0.00	0.00	51.74	0.00
3	H3K56ac_H2Bub_H3K4me3_H3K4me2_H3K9me3_H3K36me3	6.10	8.20	2.26	19.34	0.00	0.00	0.00
4	H3K27me2_H3K4me2_H3K27me3	4.21	5.52	1.59	0.00	21.21	0.00	0.00
5	H3K27me2_H3K27me3	4.13	4.98	2.63	0.00	20.78	0.00	0.00
6	H3K27me3	3.30	4.11	1.59	0.00	16.60	0.00	0.00
7	H3K4me2	3.27	4.33	1.44	0.00	0.00	0.00	13.31

S.No	Marker_comb	%genome-occurrence	%Genes	%TEs	%CS1	%CS2	%CS3	%CS4
8	H2Bub_5mC_H3K4me2_H3K9me3_H3K36me3	2.57	3.41	1.40	8.14	0.00	0.00	0.00
9	5mC_H3K27me2_H3K9me2_H3K27me1	2.06	1.01	4.93	0.00	0.00	10.81	0.00
10	H2Bub_H3K4me2_H3K36me3	1.98	2.68	0.71	6.27	0.00	0.00	0.00
11	5mC_H2Bub_H3K4me2_H3K36me3	1.96	2.71	0.79	6.22	0.00	0.00	0.00
12	H3K27me2	1.83	1.99	1.42	0.00	0.00	0.00	3.87
13	H2Bub_H3K4me2_H3K9me3_H3K36me3	1.80	2.45	0.58	5.71	0.00	0.00	0.00
14	H3K4me2_H3K27me3	1.77	2.44	0.56	0.00	8.91	0.00	0.00
15	H3K27me1_H3K4me2	1.43	1.84	0.50	0.00	0.00	0.00	5.83
16	H2Bub_H3K4me3_H3K4me2_H3K9me3_H3K36me3	1.40	1.95	0.63	4.44	0.00	0.00	0.00
17	H3K56ac_H3K4me2_H3K4me3_H3K36me3	1.35	1.89	0.54	4.28	0.00	0.00	0.00
18	H2Bub_H3K4me2	1.34	1.81	0.56	0.00	0.00	0.00	4.91
19	H3K27me2_H3K9me2_H3K27me1_H4K20me1	1.28	0.25	4.26	0.00	0.00	6.72	0.00
20	H2Bub_5mC_H3K56ac_H3K4me3_H3K4me2_H3K9me3_H3K36me3	1.12	1.37	0.73	3.56	0.00	0.00	0.00
21	H2Bub_H3K4me3_H3K56ac_H3K4me2_H3K36me3	1.07	1.45	0.42	3.40	0.00	0.00	0.00
22	H3K27me1_H3K9me2_5mC	1.03	0.36	2.92	0.00	0.00	5.43	0.00
23	H3K27me1_H3K9me2	0.99	0.37	2.76	0.00	0.00	0.00	4.01
24	H3K27me2_H3K27me1_H3K9me2	0.85	0.46	2.03	0.00	0.00	4.49	0.00
25	H3K27me2_H3K4me2	0.80	1.02	0.33	0.00	3.97	0.00	0.00
26	H3K4me3_H3K56ac_H3K4me2_H3K9me3_H3K36me3	0.80	0.92	0.44	2.53	0.00	0.00	0.00
27	5mC	0.77	0.89	0.46	0.00	0.00	0.00	3.13

S.No	Marker_comb	%genome-occurrence	%Genes	%TEs	%CS1	%CS2	%CS3	%CS4
28	H2Bub	0.72	0.05	0.00	0.00	0.00	0.00	2.93
29	H3K27me2_H3K4me2_H3K27me1	0.71	1.04	0.25	0.00	0.00	0.00	0.23
30	H3K4me2_H3K36me3	0.69	0.87	0.42	0.01	0.00	0.00	1.50
31	H3K27me2_H3K27me3_H3K56ac_H3K4me2	0.67	0.80	0.27	0.00	3.35	0.00	0.00
32	H2Bub_H3K4me3_5mC_H3K4me2_H3K9me3_H3K36me3	0.62	0.86	0.40	1.96	0.00	0.00	0.00
33	H3K27me1	0.61	0.02	0.02	0.00	0.00	0.00	2.47
34	H3K56ac_H3K4me2	0.59	0.73	0.25	0.00	0.00	0.00	2.28
35	H3K27me2_5mC	0.58	0.52	0.77	0.00	0.00	0.00	0.00
36	H3K9me2	0.56	0.42	0.81	0.00	0.00	0.00	2.28
37	H2Bub_H3K56ac_H3K4me2_H3K36me3	0.54	0.69	0.15	1.71	0.00	0.00	0.00
38	H2Bub_H3K4me2_5mC	0.50	0.63	0.21	1.57	0.00	0.00	0.00
39	H3K27me2_H3K27me3_5mC	0.47	0.54	0.44	0.00	2.35	0.00	0.00
40	H3K27me2_H3K27me3_H3K4me2_H3K27me1	0.43	0.56	0.19	0.00	2.18	0.00	0.00
41	H3K56ac_H3K4me2_H3K27me1	0.43	0.63	0.13	0.00	0.00	0.00	1.75
42	H3K4me3_H3K4me2	0.42	0.61	0.15	0.00	0.00	0.00	1.65
43	H2Bub_H3K4me2_H3K4me3_H3K36me3	0.40	0.56	0.06	1.26	0.00	0.00	0.00
44	H3K27me2_H3K27me1_5mC	0.39	0.30	0.75	0.00	0.00	2.06	0.00
45	H3K56ac_H3K4me2_H3K4me3	0.39	0.54	0.08	1.23	0.00	0.00	0.00
46	H3K56ac_H3K4me2_H3K36me3	0.38	0.57	0.08	1.21	0.00	0.00	0.00
47	H2Bub_H3K56ac_H3K4me2_H3K9me3_H3K36me3	0.38	0.48	0.21	1.21	0.00	0.00	0.00
48	H3K4me2_H3K4me3_H3K36me3	0.36	0.49	0.19	1.14	0.00	0.00	0.00

S.No	Marker_comb	%genome-occurrence	%Genes	%TEs	%CS1	%CS2	%CS3	%CS4
49	5mC_H3K9me2_H3K27me1_H4K20me1	0.35	0.09	1.04	0.00	0.00	1.86	0.00
50	H3K27me2_H3K9me2_5mC	0.35	0.29	0.58	0.00	0.00	1.83	0.00
51	H3K4me2_H3K27me1_H3K36me3	0.35	0.47	0.06	0.00	0.00	0.00	1.42
52	H3K4me2_5mC	0.34	0.46	0.21	0.00	0.00	0.00	1.40
53	H3K27me2_H3K56ac_H3K4me2_H3K27me1	0.34	0.48	0.08	0.00	0.36	0.00	0.00
54	5mC_H3K27me2_H3K27me3_H3K4me2	0.33	0.36	0.46	0.00	1.69	0.00	0.00
55	H2Bub_H3K56ac_5mC_H3K4me2_H3K36me3	0.33	0.47	0.06	1.05	0.00	0.00	0.00
56	H3K56ac_H3K27me2_H3K4me2	0.31	0.31	0.31	0.00	1.54	0.00	0.00
57	H3K4me2_H3K27me1_5mC	0.29	0.36	0.10	0.00	0.00	0.00	1.19
58	H3K27me2_H3K9me2	0.29	0.22	0.42	0.00	0.00	0.02	0.00
59	H3K27me1_5mC	0.27	0.22	0.42	0.00	0.00	0.00	1.11
60	H3K4me3	0.27	0.36	0.19	0.00	0.00	0.00	1.09
61	H2Bub_5mC_H3K4me2_H3K9me3_H3K36me3_H4K20me1	0.26	0.38	0.08	0.84	0.00	0.00	0.00
62	H3K56ac_H2Bub_H3K4me3_5mC_H3K4me2_H3K36me3	0.26	0.36	0.13	0.82	0.00	0.00	0.00
63	H3K27me2_H3K27me1	0.25	0.22	0.42	0.00	0.00	0.00	0.06
64	H2Bub_H3K27me2_H3K56ac_H3K4me3_H3K4me2_H3K9me3_H3K36me3	0.24	0.26	0.25	0.75	0.00	0.00	0.00
65	5mC_H3K27me2_H3K9me2_H4K20me1	0.21	0.13	0.46	0.00	0.00	1.12	0.00
66	H3K56ac_H2Bub_5mC_H3K4me2_H3K9me3_H3K36me3	0.21	0.25	0.08	0.67	0.00	0.00	0.00
67	H2Bub_H3K56ac_H3K4me2	0.21	0.30	0.02	0.66	0.00	0.00	0.00
68	H3K4me2_H3K36me3_5mC	0.21	0.22	0.17	0.66	0.00	0.00	0.00

S.No	Marker_comb	%genome-occurrence	%Genes	%TEs	%CS1	%CS2	%CS3	%CS4
69	H3K27me2_H3K27me3_H3K4me2_H4K20me1	0.20	0.19	0.17	0.00	1.02	0.00	0.00
70	5mC_H3K4me2_H3K27me1_H3K36me3	0.19	0.28	0.08	0.55	0.00	0.00	0.00
71	H3K4me2_H3K4me3_H3K9me3_H3K36me3	0.19	0.27	0.13	0.61	0.00	0.00	0.00
72	H2Bub_H3K4me2_H3K27me1	0.17	0.24	0.02	0.00	0.00	0.00	0.71
73	H3K27me2_H4K20me1_H3K9me2	0.17	0.06	0.31	0.00	0.00	0.89	0.00
74	H3K4me3_H3K56ac_H3K27me3_H3K4me2_H3K27me2	0.17	0.23	0.04	0.00	0.83	0.00	0.00
75	H3K56ac_H3K27me1_H3K27me3_H3K4me2_H3K27me2	0.17	0.22	0.06	0.00	0.83	0.00	0.00
76	H2Bub_H3K56ac_H3K4me3_H3K4me2_H3K9me3_H3K36me3_H4K20me1	0.17	0.22	0.06	0.52	0.00	0.00	0.00
77	H3K4me3_H3K56ac_H3K27me1_H3K4me2_H3K36me3	0.16	0.19	0.10	0.51	0.00	0.00	0.00
78	5mC_H3K27me2_H3K4me2_H3K27me1	0.15	0.19	0.10	0.00	0.00	0.79	0.00
79	H3K9me2_H3K27me2_H3K27me3_5mC_H3K4me2_H4K20me1	0.15	0.17	0.13	0.00	0.00	0.74	0.00
80	H3K9me2_H3K27me2_H3K4me3_H3K27me1_H3K9me3_H4K20me1	0.15	0.00	0.46	0.00	0.00	0.79	0.00
81	H2Bub_H3K4me3_5mC_H3K4me2_H3K36me3	0.15	0.19	0.04	0.46	0.00	0.00	0.00
82	H3K27me1_5mC_H3K4me2_H3K9me2_H3K27me2	0.15	0.15	0.13	0.00	0.00	0.77	0.00
83	H3K4me3_H3K56ac_H3K4me2_H3K36me3_H3K27me2	0.14	0.19	0.06	0.45	0.00	0.00	0.00
84	H3K27me2_H4K20me1_H3K27me3	0.14	0.16	0.08	0.00	0.69	0.00	0.00
85	H3K27me2_H3K9me2_H3K27	0.14	0.16	0.15	0.00	0.69	0.00	0.00

S.No	Marker_comb	%genome-occurrence	%Genes	%TEs	%CS1	%CS2	%CS3	%CS4
	me3							
86	H3K56ac_H3K4me2_H3K27me1_H3K36me3	0.14	0.18	0.10	0.42	0.00	0.00	0.00
87	H2Bub_5mC	0.13	0.14	0.10	0.00	0.00	0.00	0.27
88	5mC_H3K56ac_H3K4me2_H3K27me1	0.13	0.17	0.04	0.00	0.00	0.00	0.00
89	H3K9me2_H3K27me2_H3K27me3_5mC_H3K27me1_H4K20me1	0.13	0.14	0.15	0.00	0.00	0.69	0.00
90	H3K56ac	0.12	0.14	0.13	0.00	0.00	0.00	0.48
91	5mC_H3K56ac_H3K4me2_H3K36me3	0.12	0.14	0.06	0.37	0.00	0.00	0.00
92	H3K27me1_H4K20me1_H3K9me2	0.11	0.05	0.35	0.00	0.00	0.59	0.00
93	H2Bub_H3K27me1_5mC_H3K4me2_H3K36me3	0.11	0.17	0.00	0.36	0.00	0.00	0.00
94	5mC_H3K27me2_H3K9me2_H3K27me1_H3K27me3_H3K4me2_H4K20me1	0.11	0.12	0.13	0.00	0.00	0.59	0.00
95	H3K27me3_5mC	0.11	0.14	0.06	0.00	0.55	0.00	0.00
96	H2Bub_H3K4me3_H3K56ac_H3K9me3_H3K36me3	0.11	0.14	0.10	0.34	0.00	0.00	0.00
97	H3K56ac_H3K27me3_H3K4me2_H4K20me1_H3K27me2	0.11	0.16	0.06	0.00	0.55	0.00	0.00
98	H3K27me1_H3K27me3_5mC_H3K9me2_H3K27me2	0.11	0.08	0.17	0.00	0.00	0.57	0.00
99	H2Bub_H3K4me2_H3K4me3	0.10	0.15	0.00	0.33	0.00	0.00	0.00
100	H3K27me2_H3K27me3_H3K9me2_H3K4me2	0.10	0.11	0.08	0.00	0.52	0.00	0.00

Co-association

Co-association studies were conducted to identify the pairwise relationships between different chromatin modifications. Thus for any given mark, its co-association value with any other mark would be percentage of times that a mark co-occurs with it. In contrast to correlation analysis, co-association between two marks is directional, which means that a co-association value between mark1 versus mark2 can be different from mark2 versus mark1, e.g., for each probe identifying H3K4me3, 93% of the times H3K4me2 is also found, while for each probe identifying H3K4me2, H3K4me3 is found only 34% of the times. Thus in a pairwise co-association matrix the values in the upper triangular matrix are different than in lower triangular matrix.

Co-association values were separately calculated for all tiles and tiles overlapping to genes only. Of all the tiles on chromosome IV, 11,441 tiles overlapped to genes and were used to compute the co-association between chromatin marks for genic tiles only. Co-association was also computed for genes only by declaring each gene as enriched or non-enriched in a given chromatin modification based on the status of one or more probes that reported the signal from within the gene. Of the total of ~5200 genes on chromosome IV, a set of only 4182 genes could be unambiguously declared as enriched or non-enriched for all the 12 chromatin marks. This set of 4182 genes was used to compute pairwise co-association between chromatin marks for genes only.

Heatmaps representing pairwise co-association values for 12 chromatin marks were plotted using R and the MADE4 package from Bioconductor.

Cluster Analysis

Cluster analysis or clustering is a standard technique of statistical data analysis based on unsupervised learning where one tries to identify the hidden structure in unlabelled data. The basic aim of any clustering method is to organise a set of objects (observations) into groups called clusters, wherein objects within a cluster are similar in some sense, and dissimilar to objects belonging to other clusters. The grouping of objects into clusters can be either conceptual wherein two or more objects belong to a cluster if they can be defined by a common concept, or distance-based where similarity criterion to compare the objects is dependent on some kind of a distance parameter. A popular measure of distance in the distance-based clustering methods is the Minkowski metric. The Minkowski distance between any two data records *a = (x1 , y1)* and *b = (x2 , y2)* is defined as:

$$d^p(a,b) = [|x_1 - x_2|^p + |y_1 - y_2|^p]^{1/p}$$

For higher dimensional data with dimensions *d*:

$$d^p(a,b) = \left(\sum_{k=1}^{d} |x_{a.k} - x_{b.k}|^p\right)^{1/p}$$

The Euclidean and Manhattan distances are the special cases of Minkowski metric where, the order of the Minkowski metric is $p=2$ and $p=1$ repectively.

Clustering algorithms may be classified as:

- Exclusive Clustering
- Overlapping Clustering
- Hierarchical Clustering
- Probabilistic Clustering

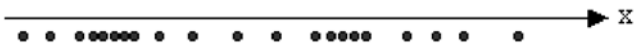
In the case of Exclusive clustering data are grouped in an exclusive way, so that each of the data objects belongs to one and only one cluster. On the contrary, in overlapping clustering, fuzzy logic is used to distribute the objects into clusters and each data object may belong to two or more clusters with different degrees of membership. A hierarchical clustering algorithm is instead based on the union between the two nearest clusters. The beginning condition is realized by setting every datum as a cluster. After a few iterations it reaches the final clusters desired. Finally, the last kind of clustering uses a completely probabilistic approach.

An exclusive clustering algorithm called $k\text{-}means$ clustering aims to partition n observations into k clusters fixed a priori, in which each observation belongs to the cluster with the nearest mean. The $k\text{-}means$ algorithm assigns each data object to the cluster whose centre (also called centroid) is nearest. The centre is the average of all the points in the cluster. The algorithm starts by randomly generating k clusters, and determines their cluster centers, or directly generates k random points as cluster centers. Each data point is then assigned to the nearest cluster center, where "nearest" is defined with respect to some distance measure, and

recomputes the new cluster centers. Finally these two steps are then repeated until some convergence criterion is met, i.e., the centroids do not move any more.

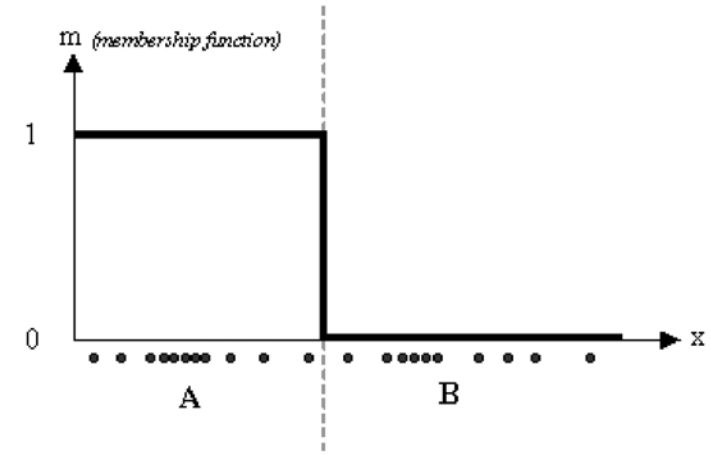
The Fuzzy *c-means* clustering, which is an overlapping clustering algorithm, assigns each data point to two or more than two clusters rather than belonging completely to just one cluster. Data are bound to each cluster by means of a Membership Function, which represents the fuzzy behaviour of this algorithm. Thus, data points on the edge of a cluster may be in the cluster to a lesser extent than data points in the centre of cluster. For each data point x, a coefficient gives the degree of being in the k^{th} cluster and usually, the sum of those coefficients for any given x is defined to be 1.

A very simple illustration of the *k-means* and fuzzy *c-means* clustering methods is given here. Let's say a certain dataset is distributed along the x-axis as shown below:

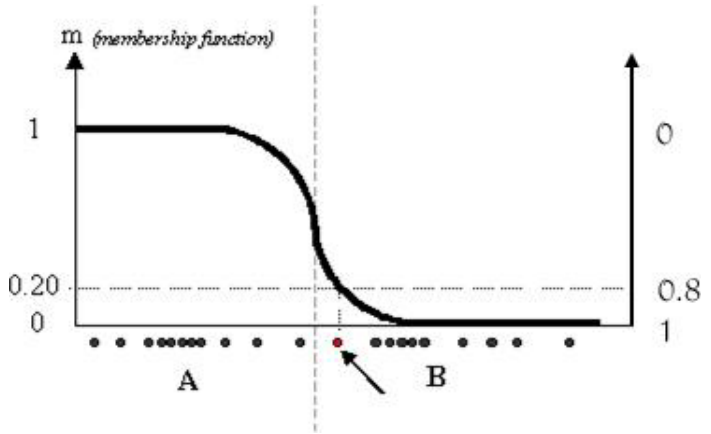


The figure shows two concentrations of the data points, which could roughly demarcate the centres of the two clusters and will be referred to as cluster A and cluster B.

In the *k-means* clustering algorithm each data point specifically belongs to one centroid only. Therefore the membership function (m) for a *k-means* algorithm returns either *0* or *1* for each data object towards a given cluster.



In the fuzzy *c-means* clustering method, the same data point does not belong exclusively to a well-defined cluster, but it can be placed in a middle way. In this case, the membership function follows a smoother line to indicate that every datum may belong to several clusters with different values of the membership coefficient.



The data point shown as a red marked spot (indicated by the arrow) belongs more to the cluster B rather than the cluster A and has a membership value of 0.8 for cluster B and a membership value of only 0.2 for cluster A.

Cluster Validity and clustering tendency

The most important parameter for all clustering algorithms is the correct identification of the number of clusters (k) because an inappropriate choice of k may yield poor results. This makes it important to run diagnostic checks for determining the number of clusters hidden in the data set. In most cases there does not seem to exist an ideal number of clusters, but it rather depends on the question being answered by the clustering results. However, an optimal number of clusters with respect to some quality criterion can be computed. As the primary focus of cluster analysis is to identify groups of objects where objects within the clusters are as similar as possible and objects between different clusters are as dissimilar as possible. These conditions can be evaluated by defining a measure of HOMOGENEITY within the clusters and HETEROGENEITY between them. Homogeneity of a cluster can be measured in terms of the maximum, minimum or average of the distances between all objects of a cluster, or an average distance of all the objects within a cluster from the cluster centre. Thus, one of the ways to measure the homogeneity S_j within a cluster j is

$$S_j = \sum_{i=1}^{n_j} |x_i^j - c_j|^2$$

Where n_j is the number of objects and c_j is the cluster centre of cluster j

Similarly, a measure of heterogeneity between two clusters can be computed based on the maximum, minimum, or average of all the pairwise distances between the objects of two clusters, or on the pairwise distances between the cluster centres. Thus one of the ways to measure heterogeneity D_{jl} between two clusters j and l is

$$D_{jl} = |c_j - c_l|^2$$

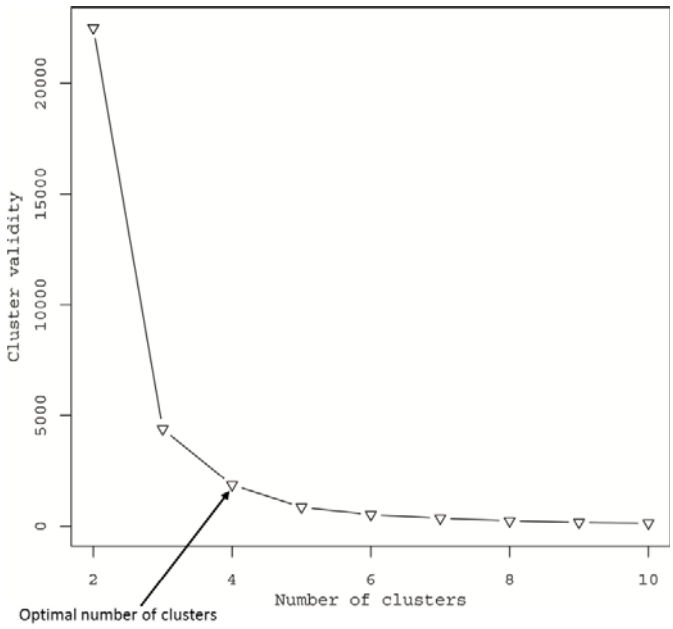
Where c_j and c_l are the cluster centres of j and l

Finally, a combination of the two parameters homogeneity & heterogeneity for all the k clusters would give a measure of cluster validity (V_k) for k clusters and is defined as:

$$V_k = \frac{\sum_{j=1}^{k} S_j}{\sum_{j<l=1}^{k} D_{jl}}$$

Since the average distances of objects in a homogenous cluster should be small from the cluster centre while as the distance between cluster centres should be large. Therefore, the homogeneity measure should be small and heterogeneity large. Thus, cluster validity (V_k), a measure of the optimality of the number of clusters (k) should be small. A graph showing the number of clusters versus the validity measure will often show a knee which indicates the optimal number of clusters as on the left of the knee the validity measure increase fast, whereas on the right on the knee only a marginal improvement on the validity measure can be made.

A cluster validity plot for the 21,406 probes on the Chromosome 4 tiling array, representing enrichment for all the 12 chromatin modifications analysed in this study, indicates an optimal number of clusters of 4 in our dataset.



The above plot was generated using an R script that requires e1071 package from Bioconductor and reads a tab delimited text file containing enrichment for the chromatin marks in the following format.

probe_id	H3K4 me2	H3K4 me3	H3K9 me2	H3K9 me3	H3K27 me2	H3K27 me3	H3K36 me3	H3K56 ac	5mC	H4K20 me1	H2B ub	H3K27 me1
ta01a01	0	0	1	1	1	0	0	0	0	1	0	0
ta01a02	0	0	1	1	1	0	0	0	0	1	0	0
ta01a03	0	0	1	1	1	0	0	0	0	1	0	0
ta01a04	0	1	1	1	1	0	0	0	0	1	0	0
ta01a05	0	1	1	1	1	0	0	0	0	1	0	0
ta01a06	0	1	1	1	1	0	0	0	0	1	0	0
ta01a07	1	1	1	1	1	0	0	0	0	1	0	0

ta01a08	1	1	1	1	1	0	0	0	0	1	0	0
ta01a09	0	0	1	1	1	0	0	0	0	1	0	0
ta01a10	0	0	1	1	1	0	0	0	0	1	0	0

R script to find the optimum fuzzy c-means clusters between k=2 to k=10. This script generates a cluster validity plot for the cluster range (k=2 to 10) like the one as shown above.

```
#The below code checks validity of fuzzy c-means clusters. The
#input_tab_delim_file.txt that the programme reads contains multidimensional
#data objects to cluster. Each row in the file is a data object and each
#column is a dimension


#the function cluster validity measure (clvm) requires a data frame (x) that
#contains the data to be clustered and clnumb is an array that gives range of
#clusters: default is 2:10

clvm <- function (x, clnumb = c(2:10))

{

    require(e1071)

    vali=c()

    for (j in 1:length(clnumb))

    {

        k = clnumb[j]

        set.seed(100)

        res = cmeans(x, k)

        wss = rep(NA, k)

        for (i in 1:k)

            {

               wss[i] = sum(apply(t(t(x[res$clu == i, ]) - res$cent[i,])^2,
               1, sum))

            }

        bss = dist(res$cent)

        vali[j] = sum(wss)/sum(bss)
```

```
    }
    return(vali)
 }


#read input file into a data frame
x <- read.table("input_tab_delim_file.txt",header=T,row.names=1)


#define cluster range to be checked
clr=2:10


#call clvm function with arguments as data frame and "cluster range". The
#function returns a vector of length "cluster range" in which each element
#contains the cluster validity index for the corresponding cluster number k
a <- clvm (x,clr)


#plot cluster number versus cluster validity index
plot (clr,a,xlab = "Number of clusters", ylab = "Cluster validity index",
      type = "b",ylim = c(0, max(a)), col = 1, lty = 1, pch = 6)
```

CHAPTER IV

SPATIAL & TEMPORAL DYNAMICS IN HISTONE H2BUB CHROMATIN MARK DURING LIGHT DRIVEN DEVELOPMENTAL ADAPTATION & THE ROLE THEREIN FOR FINE-TUNING OF GENE EXPRESSION

Introduction

Monoubiquitination of histones H2A and H2B plays an important role in the regulation of gene expression throughout the eukaryotic kingdom. While monoubiquited histone H2A is associated to Polycomb-mediated gene repression, the monoubiquitinated form of H2B (H2Bub) occurs genome-wide from unicellular yeast cells to multicellular eukaryotes such as mammals and plants and is associated with transcriptional activation. In mammals ubiquitination of H2B occurs at lysine 120 (H2BK120), and in the yeasts *Saccharomyces cerevisiae* and *Schizosaccharomyces pombe* at lysine 123 and 119, respectively. However, in plants such as Arabidopsis, H2B is monoubiquitinated at lysine 143 (H2BK143) in the C terminus, the equivalent of yeast H2BK123 (Sridhar, Kapoor et al. 2007). The site of ubiquitination in H2B and the primary amino acid sequence surrounding the ubiquitination site (AVT**K**FTSS) are highly conserved between yeast, human and Arabidopsis (Sridhar, Kapoor et al. 2007).

Ubiquitin (Ub) is a 76-amino acid protein that is universally distributed and highly conserved throughout eukaryotic organisms. Ubiquitination is a post-translational modification, and occurs when a ubiquitin moiety is attached to the side chain of a lysine residue in the acceptor protein. This involves a complex process that includes sequential actions of a three-step enzyme catalysed reaction including an E1 ubiquitin-activating enzyme, an E2 ubiquitin-conjugating enzyme, and E3 ubiquitin ligases. Proteins can be ubiquitinated in the form of either a single molecule attachment (monoubiquitination) or may occur as Ub-chains (polyubiquitination) with each subsequent Ub attached to a lysine of the prior ubiquitin moiety. Several lysine residues can be used for inter Ub-chain formation, including Lysine 6 (K6), Lysine 11 (K11), Lysine 29 (K29), Lysine 48 (K48), and Lysine 63 (K63). K48-linked polyubiquitin chains

primarily target proteins for proteasome-mediated destruction. K63 linkages perform more diverse functions by altering target protein structure, localization or activity. In all, ubiquitination of proteins has been linked to a variety of cellular processes including protein degradation, stress responses, cell-cycle regulation, protein trafficking, endocytosis signalling, differentiation & development, and transcriptional regulation (Pickart 2001).

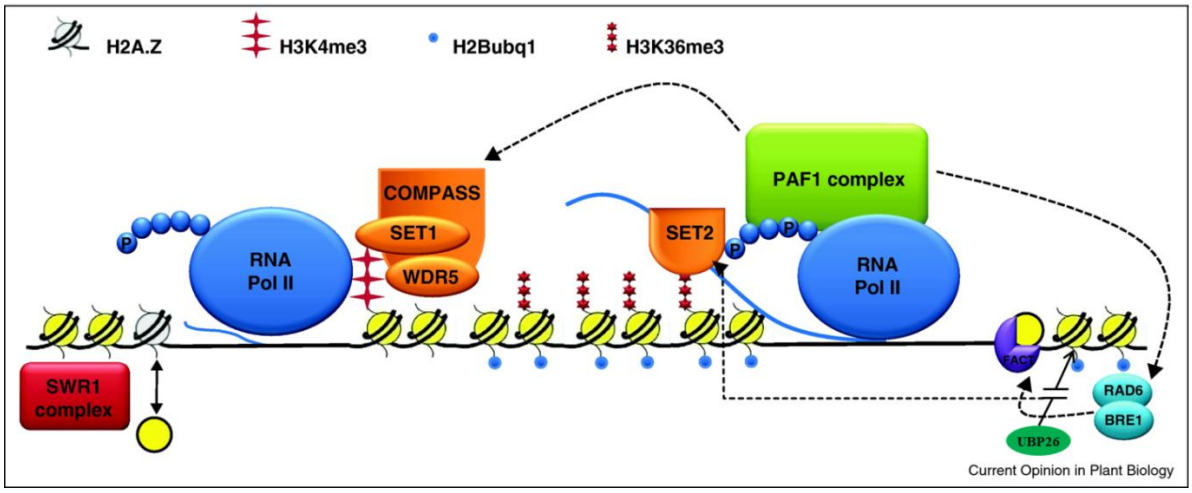
Rad6 and Bre1 were the first E2 and E3 enzymes described for the generation of H2Bub in yeast (Robzyk, Recht et al. 2000; Hwang, Venkatasubrahmanyam et al. 2003; Kao, Hillyer et al. 2004), but since then homologs of these genes have been identified in both animals (Bray, Musisi et al. 2005; Zhu, Zheng et al. 2005; Kim, Guermah et al. 2009) and plants (Fleury, Himanen et al. 2007; Liu, Koornneef et al. 2007; Gu, Jiang et al. 2009; Xu, Ménard et al. 2009). In yeast and humans, the Bre1 E3 ubiquitin ligase monoubiquitinates histone H2B and is required for addition of secondary modifications on histone H3, indicative of a regulatory unidirectional ‘‘crosstalk” between histone modifications. The RNA pol II-associated transcription elongation complex, known as Polymerase-associated factor 1 complex (Paf1C), is necessary for Rad6-Bre1 dependent H2B ubiquitination and, in turn, H2Bub is necessary for hypermethylation of H3 on lysine 4 (H3K4me3), catalysed by Set1, and on lysine 79 (H3K79me3) catalysed by Dot1. Paf1C, as well as the histone methyltransferases Set1 and Set2 that catalyse H3K4me3 and H3K36me3 reactions, respectively, have been found to be associated with the elongating form of Polymerase II (Xiao, Kao et al. 2005) to regulate the expression of a subset of the yeast genome (Betz, Chang et al. 2002). Still in yeast, the Rad6-Bre1 complex, which interacts with Paf1C and monoubiquitinates H2B at lysine 123 (H2BK123) within target genes (Wood, Krogan et al. 2003), helps recruiting COMPASS (COMplex Proteins ASsociated with SET1) to subsequently trigger

H3K4 hypermethylation and facilitate transcription (Ng, Robert et al. 2003). The histone methyltransferase Set2 that catalyses H3K36me2/3 is also essential for the proper transcriptional elongation of a subset of genes (Carrozza, Li et al. 2005). Unlike Set1, Set2 recruitment requires deubiquitination of histone H2B, which is catalyzed by SAGA and SAGA-like complexes via the activity of Ubiquitin protease 8 (Henry, Wyce et al. 2003). Therefore, the emerging picture is that transcriptional activation in yeast requires cycling of both ubiquitination and deubiquitination of histone H2B for sequential recruitment of Set1 and Set2 histone methyltransferase activities.

During the cyclic process, ubiquitin modification is proposed to enhance the movement of RNA pol II through nucleosomes by facilitating the function of the histone chaperone FACT (FAcilitates Chromatin Transcription). FACT allows elongation on chromatin templates by binding and displacing the H2A/H2B dimer from the core nucleosomes without the need for ATP hydrolysis (Belotserkovskaya, Oh et al. 2003; Pavri, Zhu et al. 2006; Reinberg and Sims 2006; Formosa 2008). Human FACT has been shown to assist RNA pol II activity by destabilizing nucleosomes in the path of the progressing RNA polymerase, facilitating the removal of a histone H2A/H2B dimer, while yeast FACT has been shown to reorganize nucleosomes into a form in which the nucleosomal DNA is more accessible, without the requirement for H2A/H2B displacement (Xin, Takahata et al. 2009). Importantly, FACT is also involved in restoring nucleosome structure behind the elongating RNA polymerase (Mason and Struhl 2003; Jamai, Puglisi et al. 2009; Lolas, Himanen et al. 2010).

In plants, the monoubiquitination of histone H2B has been recently investigated in the model species Arabidopsis *thaliana*, which has two Bre1 homologs, *HISTONE UBIQUITINATION 1* and *2* also denoted by *HUB1* and *HUB2*, and three Rad6 homologs, *UBIQUITIN-CONJUGATING ENZYME 1* to *3*, denoted by *UBC1* to *UBC3*. HUB1 and HUB2 proteins function non-redundantly and have been proposed to form a tetramer composed of two copies of each of them (Liu, Koornneef et al. 2007; Gu, Jiang et al. 2009). In contrast, the E2 conjugating enzymes UBC1 and UBC2 function redundantly to monoubiquitinate H2B (Gu, Jiang et al. 2009; Xu, Ménard et al. 2009). Loss of H2Bub in Arabidopsis, as in E2 or E3 mutant plants, does not lead to any detectable effects on global levels of H3K4me3 and H3K36me2 (Cao, Dai et al. 2008; Xu, Ménard et al. 2009). However, a defective H2Bub ubiquitinating machinery lowers H3K4me3 and H3K36me2 levels in a locus specific manner, as seen in the chromatin around *FLC* and *FLC* clade genes, and represses the expression of these genes leading to early flowering phenotypes (Cao, Dai et al. 2008; Gu, Jiang et al. 2009; Xu, Ménard et al. 2009). These findings therefore suggest that H2Bub is required for target gene-specific H3K4 trimethylation and regulation of gene expression in key developmental processes but is largely dispensable for the establishment of global H3K4me3 levels in Arabidopsis. An ad hoc histone crosstalk has not been demonstrated yet in plants, although recent advances suggest the existence of a COMPASS-like complex with HMT activity (Jiang, Gu et al. 2009; Jiang, Kong et al. 2011). Further, the Arabidopsis mutant *sup32*/*ubp26,* homolog of the yeast Ubp8 deubiquitinase, also shows reduced expression of the *FLC* gene that associates with weaker levels of H3K36me3 and increased levels of H3K27me3 but H3K4me3 is not affected (Schmitz, Tamada et al. 2009). This indicates a role for UBP26 in the transcriptional activation of gene expression through H2B

deubiquitination, which promotes H3K36me3 and prevents H3K27me3 at *FLC* chromatin. The current understanding of the working of transcriptional machinery at the *FLC* locus is summarised in Figure 4.1.



Modified from (Crevillén and Dean 2011)

Figure 4.1: Transcriptional regulation at the *FLC* gene locus.

The formation of preinitiation complex at gene promoter sequences is accompanied by phosphorylation at serine 5 of the carboxy terminal domain (CTD) of RNA polymerase II. This recruits COMPASS containing H3K4 methyltransferse SET1 (Arabidopsis homologs of SET1 are ATX1, ATX2 or ATR7) and a WD40 repeat domain-containing protein called WDR5a that binds H3K4 (Jiang, Gu et al. 2009). The elongating RNA pol II along with the PAF1 complex enhances the activity of SET1 to promote H3K4me3 around the transcriptional start site (TSS). Dynamic exchange of H2A.Z with H2A around the TSS is also required for full transcription of the gene. Transition into the elongation phase is associated with phosphorylation of the pol II CTD at serine 2 which recruits SET2 (Arabidopsis homolog EFS) to promote H3K36me3 in the body of the gene. H3K36me3 prevents initiation of cryptic transcripts from within the gene body and enhances transcriptional elongation. H2B ubiquitination by RAD6-BRE requires activity of PAF1 complex and active transcription which facilitate assembly and disassembly of nucleosomes in front of the elongating polymerase via the FACT-mediated removal of an H2A/H2B dimer. H2B deubiquitination by sup32/ubp26 is likely to further promote H3K36me3 and prevent H3K27me3. This kind of model presumes a scenario in which H3K4me3 occurs prior to H2B deubiquitination, whereas H3K36me3 occurs afterwards, with each cycle of ubiquitination pulling the polymerase ahead and deubiquitination setting up a flag in the form of H3K36me3 behind the progressing polymerase. H3K36me3 distribution on genes in Arabidopsis resembles that of H3K79me3 in mammals, and because Arabidopsis lacks a clear homolog of Dot1 (H3K79me3 methyltransferase), it is possible that H3K36me3 in plants performs a function akin to H3K79me3 in mammals (Roudier, Ahmed et al. 2011).

Chromatin structure and chromatin modifications play critical roles in response to environmental cues in plants. As a case in point, plants show major morphological and developmental changes following the perception of light. In particular, photomorphogenesis is a light-dependent developmental transition that first occurs in natural conditions when etiolated seedlings germinated in darkness reach the soil surface. Upon the first light exposure, a battery of photoreceptors triggers complex signal transduction pathways modulating gene expression to generate the photosynthetic machinery and ultimately allow both vegetative and reproductive development (Deng and Quail 1999; Neff, Fankhauser et al. 2000). Upon perception of light signals, this developmental transition is rapidly accompanied by the regulation of expression of hundreds of genes (Ma, Li et al. 2001; Jiao, Ma et al. 2005; Charron, He et al. 2009) and likely involves a massive genome-wide reprogramming of chromatin states. Such "reprogramming" of gene expression therefore offers an ideal system for dissecting the kinetics of transcriptional responses to an environmental stimulus that is ecologically and physiologically relevant. Evidence that chromatin-based mechanisms contribute to photomorphogenesis have recently emerged, with the role of histone acetylation on some light-responsive genes mediated by the evolutionarily conserved HAT GCN5, a component of SAGA complexes (Benhamed, Bertrand et al. 2006). Also, recent genome-wide approaches confirmed the highly dynamic nature of some candidate histone modifications as a response to light signals (Charron, He et al. 2009). In this last study, four histone modifications were investigated on a genome-wide scale, revealing that levels of H3K27me3, a mark characteristic of Polycomb-mediated activity, were dynamically regulated onto some light-responsive genes. Finally, our group has previously shown that DET1, a major photomorphogenic regulator, binds

chromatin through a direct interaction with histone H2B (Benvenuto, Formiggini et al. 2002). Considering this capacity of DET1 to interact with histone H2B, we analysed genome-wide profiles of histone H2B ubiquitination in the Arabidopsis *det1-1* mutant and compared it with the H2Bub enrichment profile in wild-type (data not shown). This revealed a dramatic reduction (50-75%) of global H2Bub levels in *det1-1* plants. The *det1-1* mutant shows a dramatic loss of the H2Bub mark that is almost exclusively associated with genic sequences. Residual levels of the H2Bub mark observed in *det1-1* are surprisingly due to a re-deposition onto specific families of transposable elements (TEs). Examination of *det1-1* transcripts by RT-qPCR indeed revealed a significant de-repression of some of these TEs exhibiting a high level of H2Bub in this mutant, which is in agreement with the recent demonstration that removal of H2Bub contributes to transcriptional gene silencing (Sridhar, Kapoor et al. 2007). These results indicate that DET1 may, directly or indirectly, have a general role on histone H2B ubiquitination on genes and TEs. *det1-1* is the only mutant identified with such a drastic effect on genome-wide levels of H2Bub, apart from the recently identified *hub* mutants for the H2B ubiquitin-ligase itself (HUB is the Arabidopsis homolog of yeast Bre1) that shows a complete loss of the H2Bub mark (Fleury, Himanen et al. 2007).

Although several genome-wide maps of chromatin modifications have recently been produced for the model plant species *Arabidopsis thaliana* (Zhang, Clarenz et al. 2007; Zhang, Bernatavichute et al. 2009; Roudier, Ahmed et al. 2011*)*, these reference epigenomes are usually static and rarely provide a dynamic view of chromatin-level responses to developmental or environmental cues. In this chapter, I used the transcriptional responses associated to the first light exposure as a paradigm for assessing the extent of genome-wide chromatin changes.

Considering (i) the dynamic nature of the H2Bub mark on gene regulation in yeast and mammals, (ii) the unraveled link between the light-signaling factor DET1 and H2B ubiquitination, and (iii) the fragmentary knowledge on the transcriptional coactivators acting through H2B ubiquitination, I choose to specifically focus on H2Bub chromatin mark. Importantly, early work in the laboratory has revealed that *Arabidopsis hub1-3* mutant seedlings display defects in the de-etiolation process, suggesting a role for H2Bub in this transition. In the current chapter, the *hub1-3* mutant was used as a tool to combine transcriptomic and epigenomic analyses that I integrated to assess the impact of H2Bub dynamics on light-driven transcriptional responses.

Methods

We generated genome-wide maps of histone H2B ubiquitination and analysed its role in the dynamics of gene expression as a response to the developmental transition triggered by light. Plants were cultivated for 5-day old Arabidopsis seedlings that were grown in darkness and either sampled before light exposure in darkness under a green safe light or exposed to white light for 1 or 6 hours (Figure 4.3). Samples were either directly used for RNA extraction and subsequent transcriptome analyses or crosslinked with 1% formaldehyde for chromatin extraction.

ChIP-chip analysis

Chromatin immunoprecipitation (ChIP) was carried out on crosslinked chromatin using H2Bub specific antibodies, followed by hybridisation to high-density Nimblegen tiling arrays like in Roudier et al. (2011). DNA recovered upon immunoprecipitation (IP fraction) or directly from

input chromatin (INPUT) was differentially labelled in a dye-swap and hybridised to whole-genome Nimblegen tiling arrays for two biological replicates. Data was normalised using ChIPmix method (Martin-Magniette, Mary-Huard et al. 2008) that has been adapted here to handle multiple biological replicates simultaneously. A two state hidden Markov model (HMM) implemented in CisGenome (Ji and Wong 2005; Ji, Jiang et al. 2008) package as TileMap application was used on normalised data to identify genomic regions that were significantly enriched in the H2Bub mark. Such regions, referred to as domains in this study, comprised at least three probes and represent enriched segments in the genome. TileMap computes the enriched regions from probe level test-statistic by a two-state HMM model, wherein neighbouring probes are joined together into a region based on a transition probability that depends whether an IP is significantly higher than the input. Further, TileMap was also used to identify regions that show statistically significant (posterior probability > 0.9) gain or loss of the H2Bub mark during the dark to light transition. The ChIP-chip chromatin data was visualised using a local implementation of the Generic Genome Browser and a Bio::DB::SeqFeature::Strore schema (see additional methods of Chapter II).

Transcriptome analysis

Transcriptome profiling of the same material as used for ChIP was performed for wild type (wt) and *hub1-3* plants using CATMA arrays (Allemeersch, Durinck et al. 2005). Because H2Bub ubiquitination is severely impaired in *hub1-3*, with no detectable levels of H2Bub in extracted chromatin (Figure 4.2), we used chromatin from this mutant as control material. RNA from two independent biological replicates in a dye swap experiment were hybridised on microarrays for wt and *hub1-3* (Figure 4.3). I mapped the Gene-specific Sequence Tags (GSTs) of the CATMA

array to the Arabidopsis genome TAIR8 version using SeqMap (Jiang and Wong 2008) while allowing a maximum of two mismatches only. The GSTs were assigned to a locus as defined in TAIR8 if the corresponding transcript aligned with at least 80% of the GST sequence length. The mean expression value for each GST was calculated by averaging the signals from two biological and two technical replicates (dye swap) for each of the five arrays. Linear models for microarray data analysis (Smyth 2005) was used to identify differential expression and GSTs having p-values < 0.05 were declared as differentially expressed and retained for the final analysis. For loci mapping to multiple GSTs, a mean value was computed if all of them had a p-value less than 0.05.

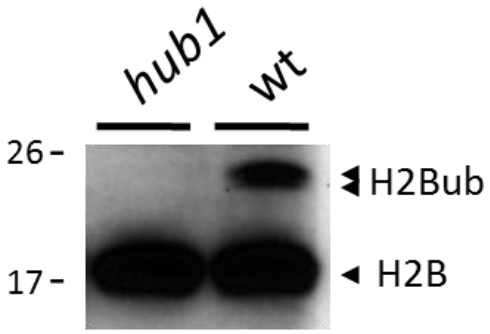


Figure 4.2: No H2Bub specific band is observed in the *hub1* mutant.

Equal amounts of chromatin extracts from wild type Col0 and *hub1-3* mutant were extracted and analyzed by immunoblot using an anti-H2B antibody. The H2Bub band is visible as a ~8KDa slower migrating doublet.

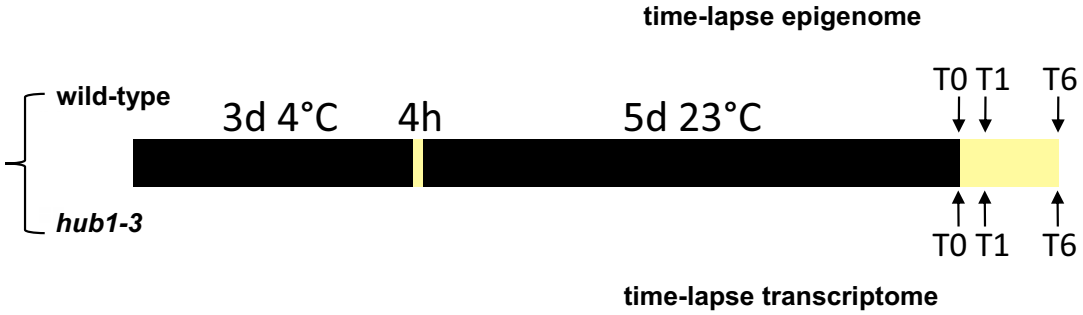


Figure 4.3: Experimental design

Upon stratification and pre-illumination, wild type and *hub1-3* plants were grown in complete darkness for five days. During the last hours, they were either maintained in darkness (T0) or exposed to white light for 1 (T1) or 6h (T6) to induce a photomorphogenic response. Plants were sampled simultaneously at midtime and directly used for RNA extraction or were crosslinked for chromatin extraction.

Results

The *hub1-3* mutant shows defects in fine-tuning of gene expression

An experimental scheme combining transcriptome and epigenome approaches was designed to study the effects of light on the dynamics of gene expression and the role of H2Bub in regulating these changes (Figure 4.3). Transcriptome analysis was performed on CATMA microarrays (http://www.catma.org) using RNAs extracted from 5-day-old seedlings with wild-type (wt) or *hub1-3* genotypes. These arrays are composed of ~31 000 Gene-specific Sequence Tags (GSTs) covering most *Arabidopsis thaliana* Col-0 genes. Aiming to provide a direct comparison of wild type and mutant plants, a first analysis was conducted on etiolated seedlings at T0 before light exposure. In this analysis the RNAs of wild type and *hub1-3* seedlings were processed and hybridized on the same arrays and therefore directly compared experimentally (wt vs *hub1-3* in dark; D). Then, for identifying early light-regulated genes, seedlings exposed to 1 hour of light were compared to similar plants of the same genotype sampled before light exposure (D vs 1h). The experiment was performed separately for wt and for *hub1-3* genotypes. Finally, a similar experiment was performed with plants exposed to 6 hour of white light (D vs 6h) for wt and for *hub1-3*, respectively. For each of these 5 different pairwise combinations, data were generated using two independent biological replicates. The resulting data were further validated by analyzing the RNA levels of ~20 candidate genes by quantitative RT-PCR analysis on independent biological replicates and proved to follow the same trend as the transcriptome analysis (data not shown). A summary of differentially regulated gene numbers is shown in Table 4.1.

	1 hour/Dark (*wt*)	6 hour/Dark (*wt*)	1 hour/Dark (*hub1-3*)	6 hour/Dark (*hub1-3*)	*hub1-3*/*wt* (dark)
Up Genes	454	872	511	905	156
Down Genes	256	707	172	631	582
2-fold Up	224	433	244	451	56
2-fold Down	80	262	49	214	267

1 hour = 5 days of dark plus 1 hour of illumination

6 hour = 5 days of dark plus 6 hours of illumination

Table 4.1: Differentially expressed genes.

Of the ~20,000 nucleus-encoded genes represented by at least one GST on the array, around 700 and 1500 genes are differentially expressed upon 1h or 6h of light exposure in the wild type, respectively, indicating that light had rapid and major effects on gene expression in our experimental conditions (Table 4.1). More globally, a total of 2711 transcripts were significantly differentially expressed in at least one of the experimental conditions.

In darkness, a set of 738 genes are misregulated in *hub1-3* compared to wt, and of these ~80% are downregulated by the *hub1* mutation (Table 4.1). Although many secondary effects might cumulate in this analysis, this proportion is in agreement with the predicted role of H2Bub in facilitating transcription. Furthermore, a significant fraction of the genes that are induced by light (differentially upregulated by light at 1h and/or 6h in the dark–to-light transition) or repressed by light (differentially downregulated by light at 1h and/or 6h in the dark–to-light transition) in wt are regulated in the same way in *hub1* prior to light exposure. Figure 4.4 A

(dark) shows that more than one-third of *hub1*-misregulated genes are also regulated in the same direction by light exposure. In this way, the *hub1-3* mutation partially mimics the effect of light on gene expression. This is in agreement with the partial de-etiolated phenotype of *hub1-3* in darkness (data not shown) and is reminiscent of the *det1* photomorphogenic mutant phenotype (Schroeder, Gahrtz et al. 2002). This observation may also suggest that HUB1 contributes to regulate many light-responding genes before illumination.

I then questioned whether *hub1-3* mutation would also affect photomorphogenic gene expression changes. Upon exposure to light for 1 or 6h, several hundreds of genes were differentially expressed in both wt and in *hub1-3* mutant. Figure 4.4 B shows how these genes are distributed in wt and *hub1-3.* In general terms, a majority of the genes induced or repressed by light in wt are also induced or repressed by light in *hub1-3*, which indicates either an H2Bub-independent transcription for these genes or a role for H2Bub in regulating the degree of induction/repression of gene expression. Nonetheless, a significant number of genes were not detected as being light-regulated in *hub1-3* mutant. Part of this could be explained by a differential expression level at T0, and this aspect will be further analysed by comparing gene expression levels.

(A) **(B)**

Figure 4.4: Venn diagram for induced and repressed genes in response to light in wild type and in *hub1-3* mutant seedlings.

(A) A large number of genes repressed by the *hub1-3* mutation in darkness are also repressed by light in wt and a significant number of genes (p-value 1.272E-27) induced in *hub1-3* mutant seedlings in darkness are also induced by light in wt. **(B)** Most light-induced genes in wt are also induced by light in *hub1* and a large fraction of light-repressed genes in wt are also repressed by light in *hub1*.

To further assess the role of H2B ubiquitination in regulating gene expression changes, transcriptome data for differentially expressed genes that undergo a high level of light-driven upregulation or downregulation (at least 2-fold) were taken for exploratory statistical analyses. Boxplots and analysis of frequency distribution were produced, aiming to compare the expression profiles of these transcripts in wild-type and in the *hub1* mutant (Figure 4.5). This analysis indicated that the dynamics of expression of the light-regulated genes is weaker in the absence of the H2Bub mark. Figure 4.5 A shows the analysis of genes upregulated more than 2-fold after 6h of light treatment, which mainly correspond to "late"-induced genes. This set of genes displays weaker amplitude of upregulation in *hub1-3* as compared to wild type. The differences become even clearer at the 6h point with about half of the genes being less upregulated in the mutant than in the wild-type. We concluded from these data that H2Bub might therefore contribute to the dynamic changes of gene expression, at least for the gain of transcriptional activity.

The analysis was then repeated for the whole set of light-downregulated genes (Figure 4.5 B), and showed a reciprocal effect with a milder downregulation of these genes in the *hub1-3* mutant. This effect is already detected at 1h and is more significant after 6h of light treatment. The observation of such a decreased dynamic change of mRNA levels in absence of H2Bub mark can lead to the hypothesis that histone H2B monoubiquitination is required for the attainment of an appropriate increased or decreased level of expression in response to a stimulus. The rapid slope in the decrease of mRNA levels further suggests that mRNA turnover might be involved, and therefore RNA stability and/or degradation might also be affected in the *hub1-3* mutant. This is further discussed below.

(A)

(B)

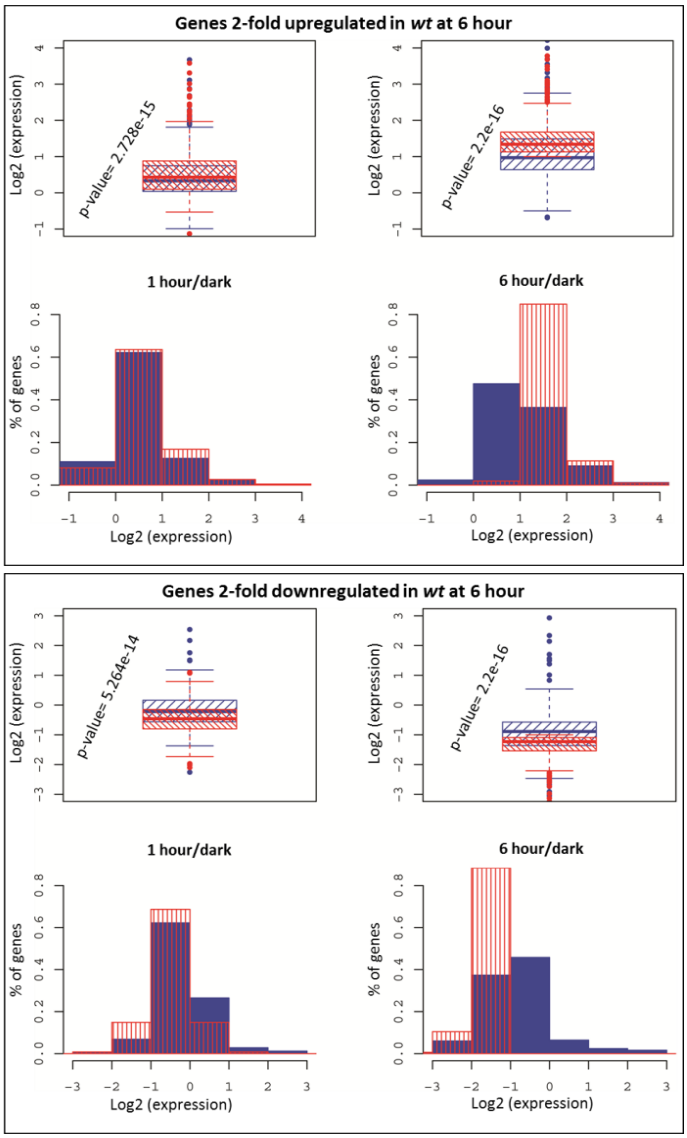


Figure 4.5: Kinetics of expression of light regulated genes in wt and *hub1-3*. (A) Genes 2-fold upregulated in wild type at 6h (n=423). **(B)** Genes 2-fold downregulated in wild type at 6h (n=249). Boxplots represent the distribution of the log2 expression ratios between dark and light for wt (red) and *hub1-3* (blue) at 1 h and 6 h. Boxes show centre quartiles (middle 50% of the data), and whiskers extend to the most extreme data points which is no more than 1.5 times the interquartile range. The outliers are shown as filled circles. Histograms show the frequency distribution of log2 expression ratios for wt (red) and *hub1-3* (blue) at 1 hour and 6 hours. The x-axis represents log2 expression ratios and the y-axis shows percentage of genes corresponding to a given expression scale on the x-axis. The p-value (Mann–Whitney U test) in each boxplot gives the significance of any differences in induction/repression of expression data between wt and *hub1-3*.

The same analysis was also conducted on the set of genes that are upregulated earlier, at the 1h point. The left panel in Figure 4.6 A shows that same effect is observed after 1h of light induction: again, upregulated genes are less efficiently induced in *hub1-3* than in the wild-type. In contrast, looking further at the 6h point it appeared that many of these genes are then subsequently downregulated in the wild-type, while this is not the case in *hub1-1* (right panel). The global picture observed is therefore in apparent contradiction with results in Figure 4.5 A. This population of "early-induced" genes was therefore splitted in two sets, the first one corresponding to genes that maintain a dynamic induction at the 6h point, and the second set corresponding to genes that are downregulated after the 1h peak. This secondary analysis allows clarifying the result. It notably shows that genes that are upregulated at 1h and pursue this induction at 6h are affected at each point, with a weaker induction dynamics (Figure 4.6 B). In contrast, the set of genes that undergo a downregulation at 6h after the early peak both exhibit a slight weaker induction at 1h and then a reduced donwregulation at 6h (Figure 4.6 C). This is therefore in agreement with the previous observation that H2Bub dynamics contributes to both increased and decreased gene expression.

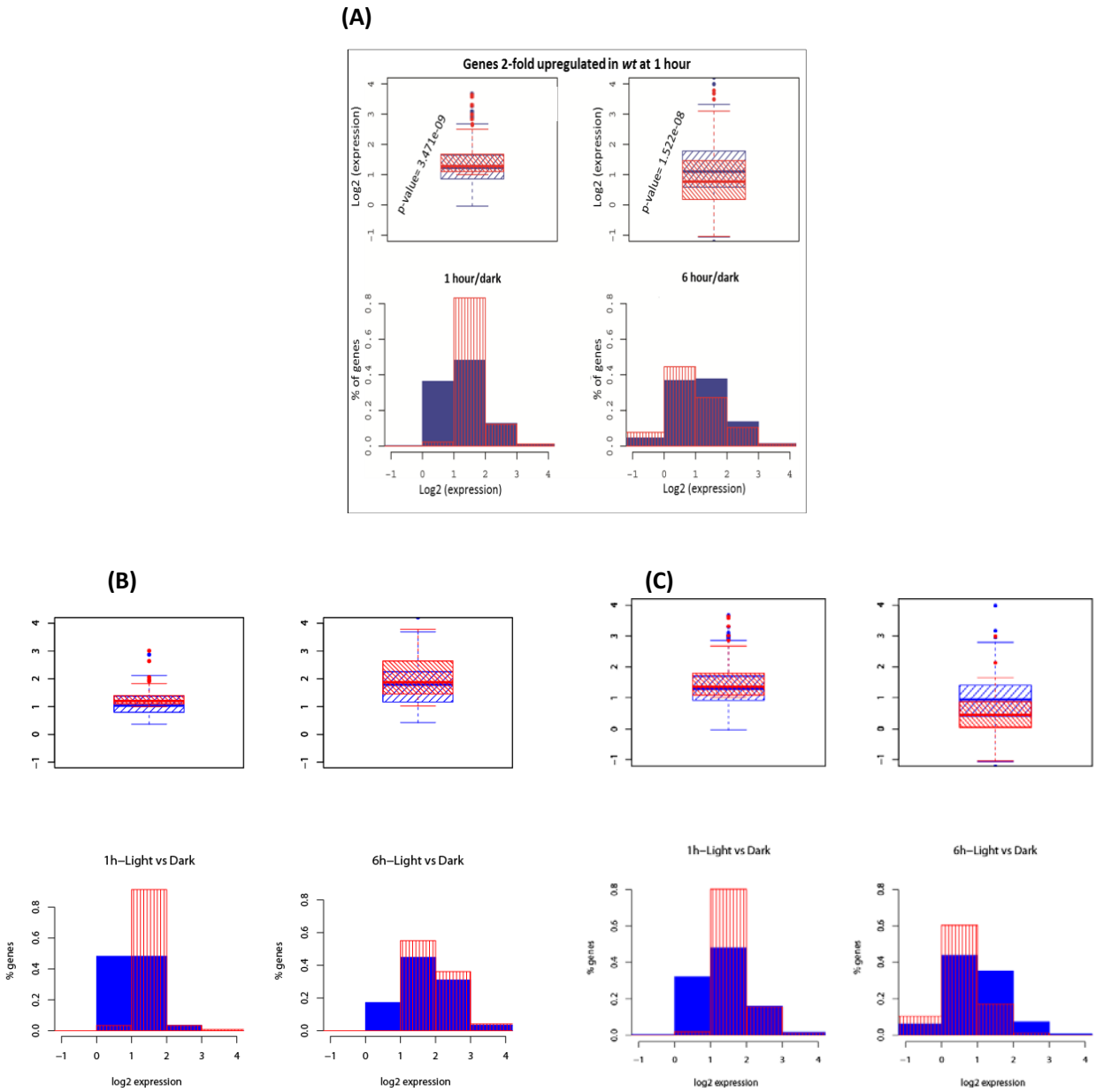


Figure 4.6: Kinetics of expression of early-induced genes. (A, B, C) Same analyses as in Figure 4.5 on the set of genes that are upregulated in wt after 1h of light exposure. **(A)** Genes 2-fold upregulated in wt at 1h (n=220). **(B)** Genes 2-fold upregulated in wt at 1h that maintain upregulation at 6h (n= 58) or **(C)** that initiates downregulation at 6h (n=162). upper panels contain boxplots representing the distribution of the log2 expression ratios (dark to light transition) for wt (red) and *hub1-3* (blue) at 1 h and 6 h, and lower panels give histogram representations of frequency distribution.

Because the H2Bub mark is most likely playing a role in transcriptional elongation, it was relevant to test the effect of loss of this mark in relation to gene size. To this end I took genes that were differentially upregulated by light at 1h or 6h and divided them into two groups based on their length. A boxplot analysis revealed there is no significant difference in fold change of expression (p-value = 0.2025) between wt and *hub1-3* for genes that are less than 2kb in length (Figure 4.7). In contrast, the genes upregulated by light and greater than 2kb in length showed a significantly lower induction in the *hub1* mutant compared to wt (p-value = 9.089e-06). Results of these analyses are compatible with the observation that long genes are more frequently marked by H2Bub (Roudier, Ahmed et al. 2011), and indicate that long genes may be more susceptible to H2Bub-dependent mechanisms to facilitate their transcription elongation.

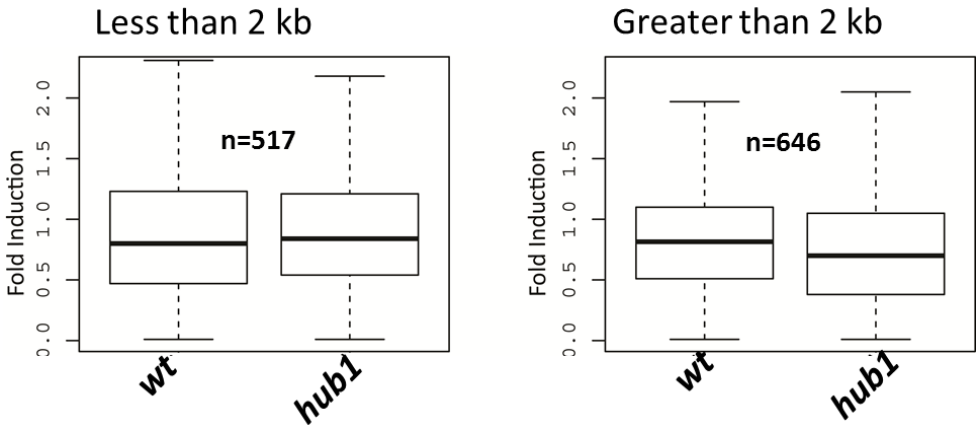


Figure 4.7: The *hub1-3* mutation mostly affects upregulation of long genes.

Genes that were differentially upregulated by light at 1h or 6h were divided into two groups: (i) genes less or equal to 2 kb in size, and (ii) genes longer than 2 kb in size. There is no detectable difference in induction levels between wt and *hub1-3* for genes less than 2 kb in size while genes greater than 2kb in size show a significantly higher induction in wt compared to *hub1-3*.

Cluster analysis of expression data for differentially expressed genes

Data from all 2711 transcripts showing differential expression in at least one of the experimental conditions were then taken for SOM (Self-Organising Maps; (Tamayo, Slonim et al. 1999) analysis to identify groups of co-expressed transcripts and to cluster them into functionally relevant gene classes based on their related expression profiles. SOM allows easy visualization of complex data sets using an unsupervised neural network algorithm that non-linearly maps data into a two-dimensional grid. This analysis revealed gene groups that have a tendency of progressive upregulation, downregulation, or maintaining constant transcript levels between 1h and 6h periods of illumination (Figure 4.8 A). The first group of 790 genes in the SOM plot comprises transcripts that show strong and steady upregulation by light in wt. These genes tend to be less upregulated in *hub1-3* mutant seedlings. Group 2 (271 genes) is similar to Group 1 but displays less pronounced upregulation by light compared to Group 1 genes, however, the rate of upregulation for these genes does not seem to be affected in *hub1-3*. Group 3 (pink) on the other hand, comprises 507 genes that do not appear to change expression in wt, between 1 and 6 hours of light, although the same genes are more upregulated at 6h of light compared to 1h in *hub1-3*. Further analysis shows that these genes are mostly downregulated in *hub1-3* in the dark (see the pink group of genes in Figure 4.8 C). This may indicate that a light-driven mechanism allows this set of genes to recover from a low level of expression in darkness and to tend towards wt levels of expression upon illumination. Finally, the last two groups, containing 295 and 848 genes, respectively, show downregulation of transcript levels between 1 and 6 hours of illumination in both wt and *hub1-3*. Nonetheless, the downregulation is stronger and more intense in Group 5 and less so in Group 4.

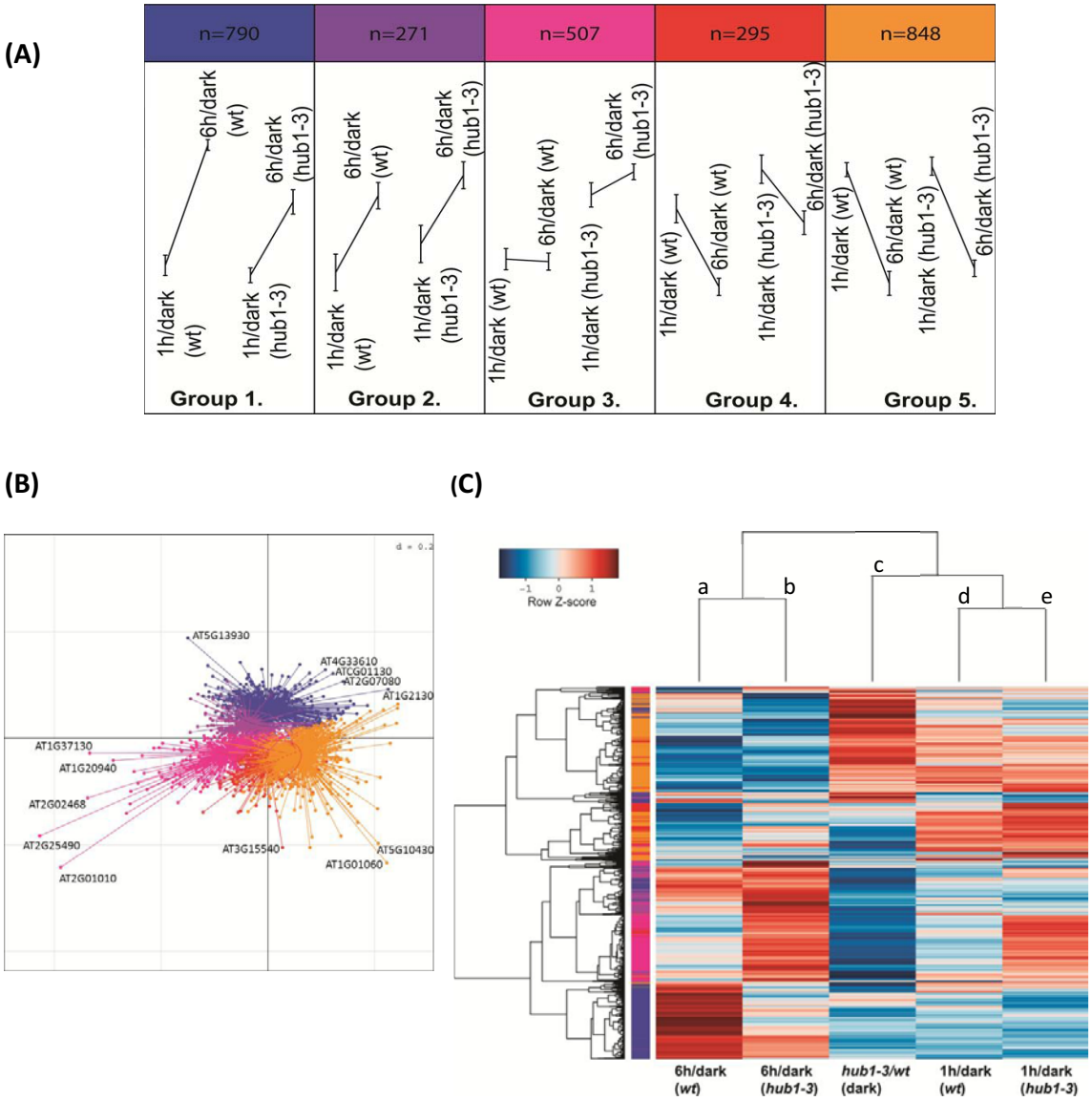


Figure 4.8: Clustering of gene expression data.

(A) Self organizing maps (SOM). In each partition the pattern reflects a general trend of expression gradient of the group of genes between 1 and 6 hours of light for wt and *hub1-3*. The four points in each group are the experimental conditions at which one of the four microarrays was performed (*hub1-3*/wt (dark) not included in this analysis) and the vertical bars at each point show variance in the group at that point. A gene is assigned to a single partition with similar groups placed in nearby partitions. **(B)** Principal component analysis. Each point in the plot is a gene whose coordinates correspond to principal component Axis1 (x-axis) and Axis2 (y-axis) values. Genes are coloured as per the groups in SOM and a concentration ellipse is drawn to enclose more than 80 percent of the data in each group. **(C)** Heatmap of the gene expression data for all five microarrays. Each horizontal line represents gene expression across the five experimental conditions with colours depicting normalized log2 expression ratios for the gene; red indicates upregulation while blue shows downregulation. The vertical colour bar next to the gene tree indicates genes belonging to each SOM group.

To corroborate the SOM analysis, I performed principal component analysis (PCA) of gene expression data as well as hierarchical clustering using the MADE4 package, displayed as a heat map of expression data in Figure 4.8 C (Culhane, Thioulouse et al. 2005). In close conformity to previous analysis, scatterplots of the first two components of PCA also clustered gene transcripts belonging to the same SOM group together. A heatmap of the hierarchically-clustered expression data revealed that many genes of the "blue" and "purple" groups are downregulated by the *hub1-3* mutation in darkness (Figure 4.8 C). The heatmap also shows that almost all transcripts that were early repressed by light (at 1h) were already strongly repressed in the *hub1-3* mutant in darkness (compare upregulated genes of column d of the heatmap to column b). Reciprocally, a large fraction of genes that are early induced by light are upregulated in *hub1-3* in the dark (compare upregulated genes of column d of the heatmap to column c). The heatmap analysis further indicates that a subset of genes from Groups 4 and 5 (red and orange) which are upregulated at 1h of light in wt are mostly early-induced genes (column d), since they further show a downregulation of their transcript levels after 6h light in both wt and *hub1-3* in columns a and b. Finally, the Group 1 genes (blue) that show a strong upregulation at 6h in wt (column a) and a relatively weaker upregulation in *hub1-3* (column b) are already downregulated at 1h as well as by the *hub1-3* mutation in darkness, indicating that these genes are mostly late induced genes.

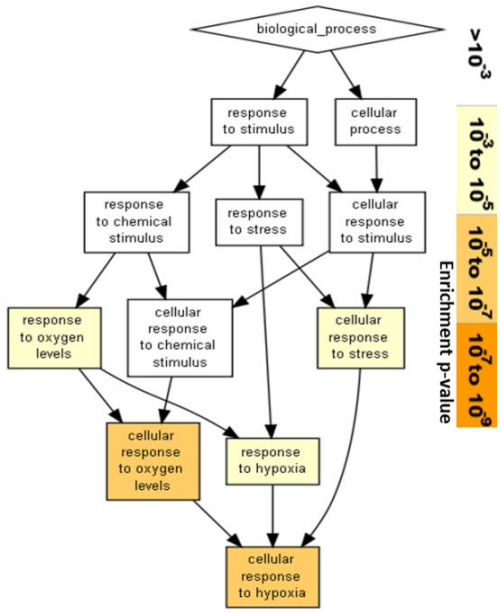
Gene Ontology annotation using Gorilla (Eden, Navon et al. 2009) was performed to identify enriched GO terms associated with all 2711 differentially expressed transcripts. Out of 2711 genes, only 1382 were found to be associated with a valid GO term and their enrichment for GO terms is depicted in Figure 4.9 B. This figure shows a highly significant enrichment for genes in

the differentially expressed gene set that are involved in 'response to stimulus' and 'stress conditions'. They also notably include most HIGH CHLOROPHYLL FLUORESCENCE genes (AT5G23120, HCF136; AT3G09650, HCF152; AT4G37200, HCF164; AT1G16720, HCF173; AT3G17040, HCF107), which mainly act as master regulators of post-transcriptional mechanisms of plastid gene expression during chloroplast biogenesis. Further a pie chart showing GO molecular function associated with the top genes in the PCA plot reveals a significant fraction of these genes having either a protein binding or nucleic acid binding function (Figure 4.9 C).

(A)

AT1G01060	AT1G44020	AT1G75780	AT2G46830	AT3G16400	AT4G02520	AT5G13930	ATCG00170
AT1G03870	AT1G51940	AT1G76930	AT3G02468	AT3G19170	AT4G11310	AT5G25250	ATCG00470
AT1G11610	AT1G59860	AT2G01010	AT3G02470	AT3G30390	AT4G14630	AT5G44120	ATCG01130
AT1G12080	AT1G61820	AT2G21660	AT3G07010	AT3G30775	AT4G15530	AT5G44440	
AT1G12110	AT1G65260	AT2G25490	AT3G09440	AT3G41768	AT4G16515	AT5G44585	
AT1G20940	AT1G66700	AT2G27050	AT3G11410	AT3G48360	AT4G33610	AT5G48485	
AT1G21310	AT1G67980	AT2G40080	AT3G12320	AT3G49620	AT4G37930	AT5G53250	
AT1G29910	AT1G72610	AT2G42870	AT3G13750	AT3G50770	AT5G02500	AT5G54770	
AT1G30290	AT1G73120	AT2G43050	AT3G15540	AT3G61160	AT5G10430	AT5G63190	
AT1G37130	AT1G74730	AT2G43610	AT3G16150	AT4G00720	AT5G12940	ATCG00050	

(B)



(C)

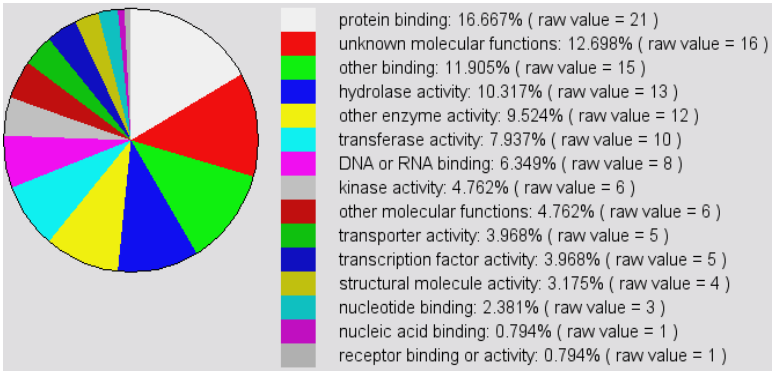


Figure 4.9: Clustering of gene expression data.

(A) Top genes in the PCA plot. **(B)** Plot showing enrichment of genes associated with a specific GO term for the differentially expressed genes. **(C)** Distribution of GO molecular function for top genes in PCA plot.

Genome-wide dynamics of H2Bub distribution

Chromatin immunoprecipitation (ChIP) followed by hybridisation to Nimblegen tiling arrays that covered the whole genome sequence of *Arabidopsis thaliana* at a resolution of 165 bp (Roudier, Ahmed et al. 2011) was used to obtain H2Bub epigenomic maps and identify changes in H2Bub levels during the dark to light transition. ChIP-chip experiments were performed using wild type dark-grown seedlings that were either sampled before light exposure (D), after 1 hour (1h) or 6 hour (6h) periods of illumination (Figure 4.3). All resulting data were incorporated in a Genome Browser. A representative chromosomal region demonstrating the IP signal over enriched probes and dynamics of the H2Bub mark during the dark to light transition is shown in Figure 4.10.

A detailed analysis of the genomic regions of H2Bub-enriched domains revealed that H2Bub enrichment is exclusively found within genes (Chapter III). Because H2Bub is known to peak in the transcribed region of a gene (Roudier, Ahmed et al. 2011) in this new analysis genes were declared H2Bub-enriched only if they overlapped in the middle 40% of their gene length with an enriched domain. Using this criterion, we could identify 4501, 4580 and 4513 genes as being enriched in dark, 1h and 6h conditions, respectively (Figure 4.11 A). The close similarity between the number of enriched domains (4105, 4014 and 3942) and the number of enriched genes for each of the three experimental conditions reflects the fact that most domains are short regions covering individual genes rather than broad domains covering clusters of genes.

Further, TileMap (see methods) was used to examine differential enrichment across the experimental conditions. This analysis to identify differentially enriched regions was performed

in two modes. A two-way compares two conditions (e.g. dark versus 6h) at a time, while a three-way compares all the three conditions simultaneously. Table 4.2 lists the number of differentially enriched regions for the two modes of comparison. The differentially enriched regions defined in Table 4.2 were further mapped to genes and any signal for differential enrichment was taken as a true signal if the gene also had a defined H2Bub enriched domain in the corresponding experimental condition, otherwise the detected differential enrichment of the gene was treated as noise. An illustration of this is shown in Figure 4.11 B&C for two-comparison conditions (6hr_gt_dark and dark_gt_6hr).

The repartition of H2Bub enrichment within genes was further analysed on marked genes sorted by length. Figure 4.11 D shows the distribution of enrichment levels for all and marked genes along with their 1 kb upstream and downstream sequences for each growth condition (D, 1h, 6h). As expected, the marking for enriched genes in all the three conditions is confined to the transcriptional units of the gene. H2Bub appears to predominantly mark longer genes.

Altogether this first analysis revealed that around 500 genes exhibit opposite marking by H2Bub as a result of plant light exposure. Comparing data in Figure 4.11 B and 4.11 C further indicates that hundreds of genes gain the mark upon illumination (~600 genes), while H2Bub seems to be pruned on only few of them during the period analysed (~25 genes). Because gene expression data comparing D, 1h and 6h do not show this almost ‘unidirectional’ property, these data suggest that acquiring the H2Bub mark can be a fast process while erasing it might require longer time-scales. These aspects will be further investigated below by integrating transcriptome and epigenome data.

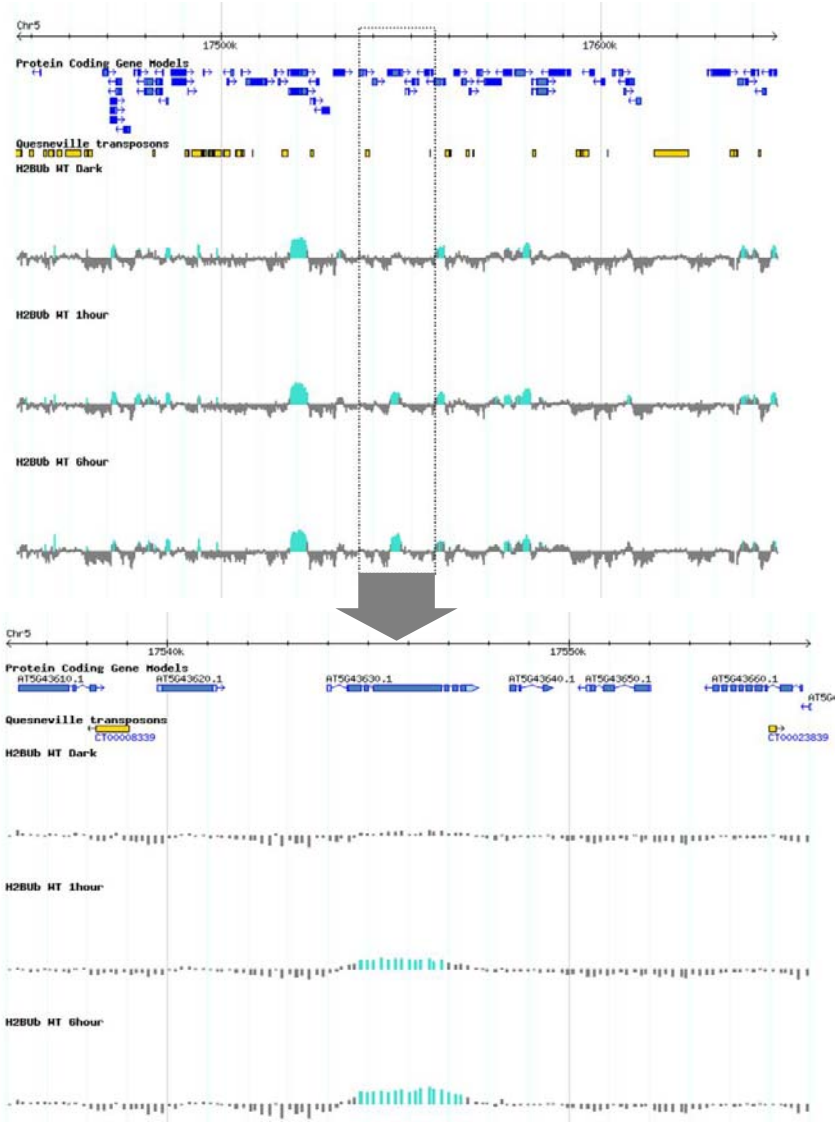


Figure 4.10: Genome Browser screenshot of a representative chromosomal region.

Shown in the figure is a 200 kb segment from Chr5 (17536000 to 17556000), and a zoomed in view of an 18 kb region (Chr5: 17538000 to 17556000) from this segment. The top 2 tracks are the loci defining annotated genes (TAIR8) and transposable elements, respectively. The three tracks immediately below show normalised H2Bub signal for dark and transition to 1h and 6h periods of illumination. Each vertical bar in the three tracks shows a probe reporting the H2Bub signal from chip (log2 of IP/INPUT). The blue colour represents a statistically enriched probe and the grey colour a non-enriched probe. A close up view from the region at bottom shows a gene (AT5G43630) that gains the H2Bub mark upon illumination. AT5G43630 is a TZP DNA-binding nuclear protein involved in regulation of transcription that is under circadian control to regulate morning-specific hypocotyl growth (Loudet, Michael et al. 2008).

(A)

2-way comparisons

Comparison	Name	No. of differentially enriched regions
dark > 6h	dark_gt_6r	1072
dark > 1h	dark_gt_1hr	228
6h > dark	6hr_gt_dark	1797
1h > dark	1hr_gt_dark	1159
6h > 1h	6hr_gt_1hr	2310

(B)

3-way comparisons

Comparison	Name	No. of differentially enriched regions
(dark > 1h) & (dark > 6h)	dark_gt_1hr&6hr	388
(6h > dark) & (6h > 1h)	6hr_gt_dark&1hr	1878
(1h > dark) & (1h > 6h)	1hr_gt_dark&6hr	596
(dark < 1h) & (dark < 6h)	dark_lt_1hr&6hr	1072
(6h < dark) & (6h < 1h)	6hr_lt_dark&1hr	123

Table 4.2: (A) 2-way comparisons for genomic regions differentially enriched with H2Bub. **(B)** 3-way comparisons for H2Bub differentially-enriched regions.

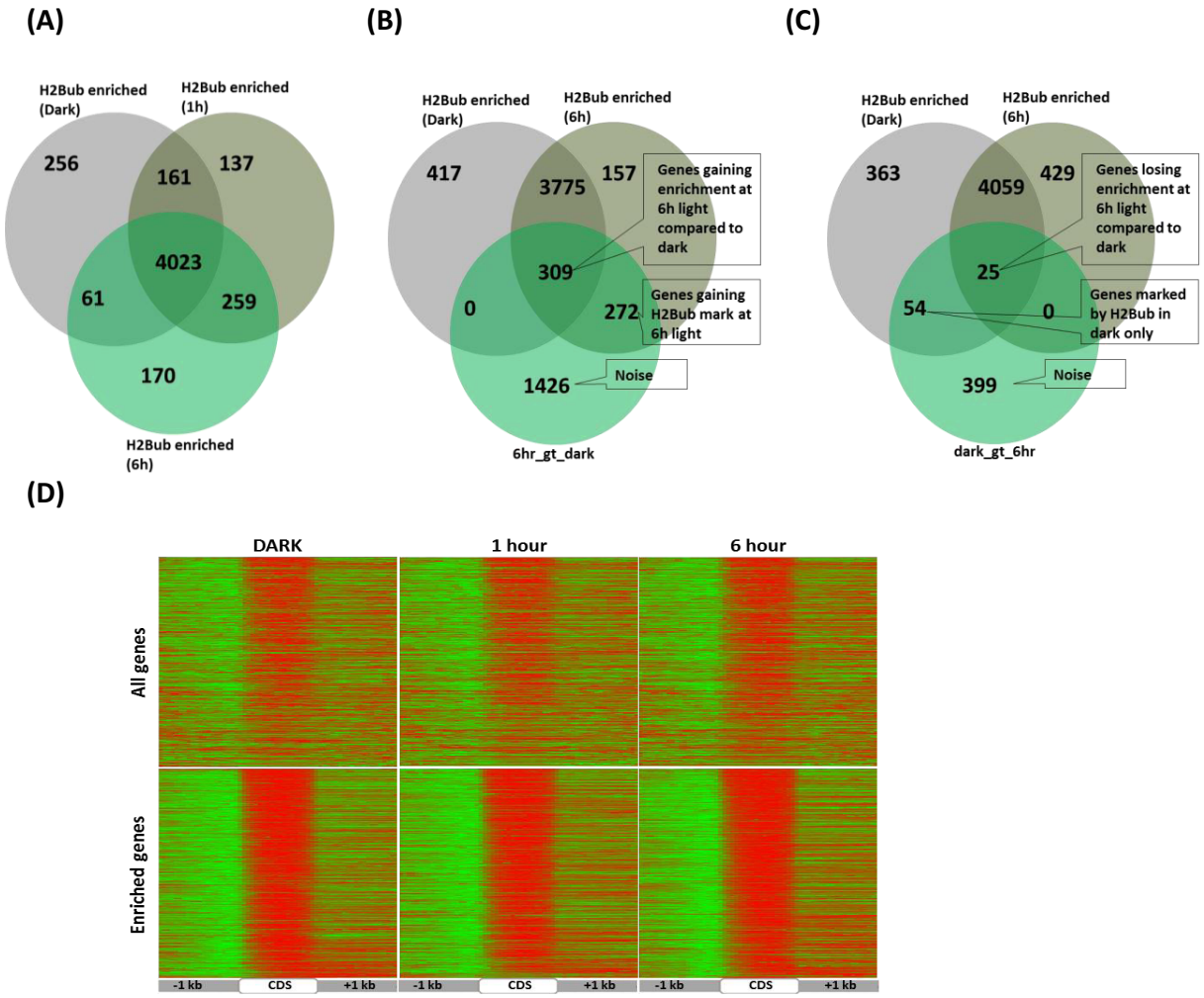


Figure 4.11: H2Bub enrichment over genes. (A) Venn diagram showing enriched genes in dark, 1 hour and 6 hours. **(B,C)** Genes that map to differentially enriched regions and show a loss or gain of the H2Bub mark in the dark to light transition. Genes were declared as being differentially enriched if either the gene is marked in both conditions but shows higher levels of H2Bub enrichment in one of the conditions or shows the mark in only one condition. Any differential enrichment for which an enriched domain could not be called in the corresponding experimental condition was treated as noise. **(D)** H2Bub enrichment on long genes. Each row is a gene along with its 1 kb upstream and downstream sequence and the genes are sorted by length with longer genes on top and shorter genes on the bottom of each of the heat maps. The red colour indicates maximum enrichment and green shows minimum enrichment. Transcriptional units are aligned in bin sizes of 1% of gene length and the upstream and downstream regions are binned for 10 bp. Top panel shows both enriched and non-enriched genes, while the bottom panel shows enriched genes only.

Light-driven induction is associated with dynamic changes of H2Bub

Aiming to determine dynamics of enrichment levels over genes, H2Bub mean levels were calculated within and around genes for each of the three conditions. Figure 4.12 A displays the distribution of H2Bub enrichment over marked genes and, as expected, shows a peak value in the middle of the gene. This steady-state distribution of the H2Bub mark over enriched genes resembles a Gaussian curve and probably reflects confounding effects of simultaneous ubiquitination and deubiquitination processes occurring at expressed genes rather than an increased ubiquitination in the middle. Comparison of mean enrichment along the marked genes in dark, 1h and 6h shows a slight increase, which is accompanied by a broadening of the H2Bub enrichment along gene length. This is further supported by comparison of domain width, shown as a box plot in the top right corner of Figure 4.12 A. Enriched domains for marked genes are significantly longer at 1h (p-value 4.524e-05; Wilcoxon rank sum test) and 6h (p-value 1.25e-06) compared to the dark condition. Visual inspection of H2Bub genic distribution may also indicate a slight shift of the peak towards the 3' end that can be observed in 6h light compared to dark. This shift might suggest a slight light-driven extension of H2Bub enrichment within expressed genes.

To better assess whether light driven extension of H2Bub domains within genes is associated to changes in expression levels, we plotted the dynamic H2Bub mean enrichment levels for the 189 genes that are marked by H2Bub and upregulated at 6h of light. As shown in Figure 4.12 B, these genes progressively gain H2Bub enrichment over time upon illumination. Broadening of H2Bub domains over time on these genes is also shown as a boxplot in the top right corner of Figure 4.12 B. We conclude from this analysis that genes gaining expression also globally gain

H2Bub enrichment. Nonetheless, this does not allow determining whether genes already marked by H2Bub in darkness gain an additional level of the mark or whether an increased number of cells acquire the mark on those genes.

Among the 189 genes marked by H2bub and upregulated at 6h, 118 genes were marked in all three conditions (dark, 1h and 6h). Plotting the distribution of mean enrichment levels over these 118 genes show no significant gain in H2Bub during gene induction (data not shown). In contrast, boxplot analysis of domain length reveals a slight but significantly increase in H2Bub domain lengths at 1h (p-value=1.93E-06; Mann–Whitney U test) and 6h (p-value=6.99E-09) compared to dark (Figure 4.12 B; inset). This suggests an association between H2Bub expansion and upregulation of expression for these genes, but determining if progressive gain in expression is associated with a similar gain of H2Bub requires further investigations (see below).

On the other hand 106 genes that were marked by H2Bub in the dark and downregulated at 6h compared to dark did not show any observable differences in H2Bub levels in any of the three experimental conditions (Figure 4.12 B). Of these 106 genes, 96 were H2Bub marked in all the three conditions, and box plot analysis does not reveal any significant differences in domain length between dark and 1h (p-value=0.844) or dark and 6h (p-value=0.8523). This may indicate that during downregulation of gene expression, H2Bub mark is more stable than the corresponding mRNAs. This result further suggests an H2Bub-independent downregulation of these genes.

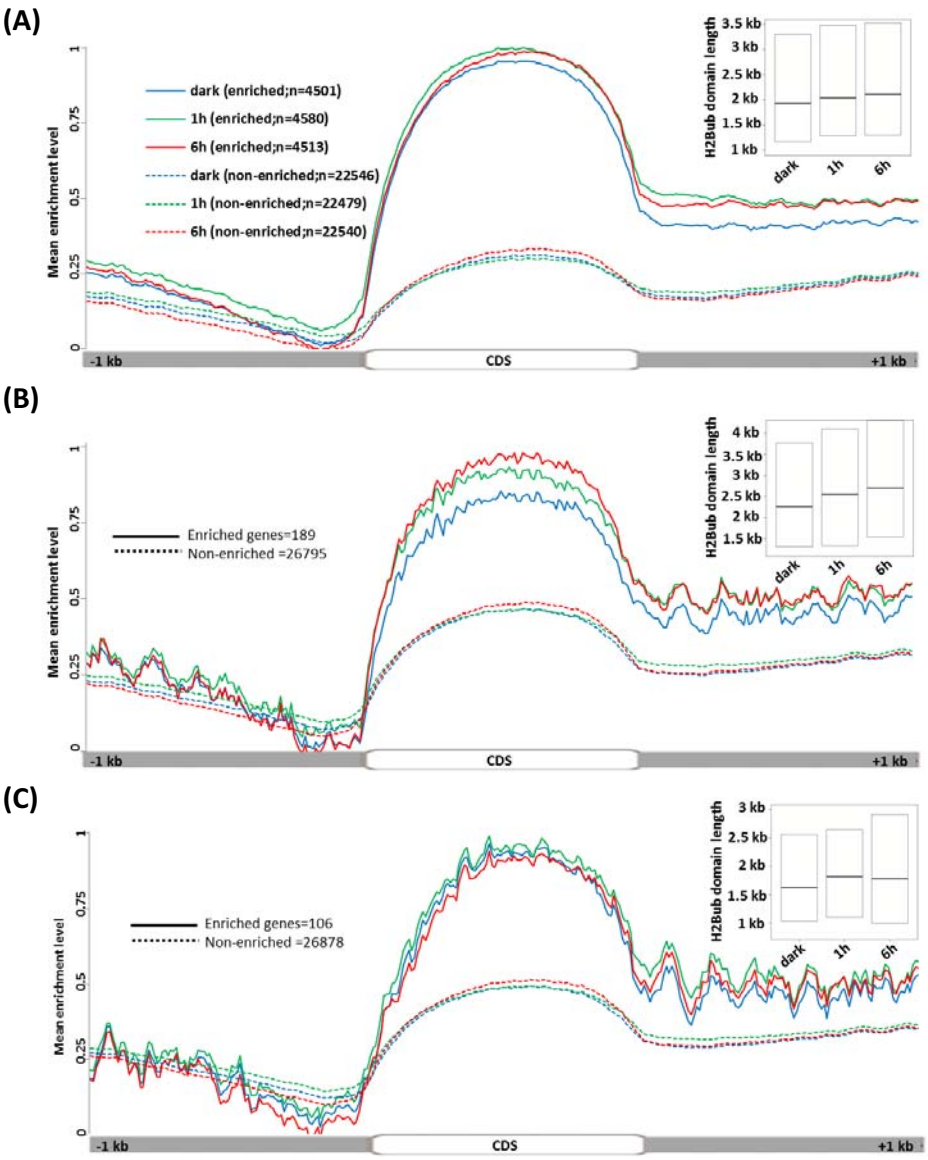


Figure 4.12: Distribution of H2Bub over genes.

(A) Mean level of H2Bub enrichment for enriched and non-enriched genes. **(B)** Mean H2Bub enrichment for genes upregulated by light at 6h compared to dark and marked by H2Bub at 6h. **(C)** Mean H2Bub enrichment for genes downregulated by light at 6h compared to dark and marked by H2Bub in darkness. The values are arbitrarily scaled from 0 to 1 depicting minimum and maximum enrichments. The figure also shows a box plot comparison of H2Bub enriched domain lengths in dark, 1h and 6h conditions. All enriched domains that overlap middle 40% of genes were taken for the analysis and their domain lengths were compared to see if H2Bub enrichment extends in 1h and 6h compared to dark. Only the middle 50% of data is shown. Internal panels give mean H2Bub domain length for each class of gene.

Aiming to test any significance of correlation between loss/gain of expression with loss/gain of H2Bub mark, I then analysed the 3099 genes that had a reported signal on the CATMA expression array and were marked by H2Bub in both dark and 6h. The results, displayed as a scatterplot in Figure 4.13, show that most genes upregulated at 6h compared to dark are accompanied by a gain of the H2Bub mark in 6h. This is indicated by the gene cloud being primarily shifted to the second quadrant as against the fourth quadrant. In contrast, the association of H2Bub with downregulation is somewhat less pronounced, as almost equal numbers of genes downregulated by light show loss or gain of the H2Bub mark.

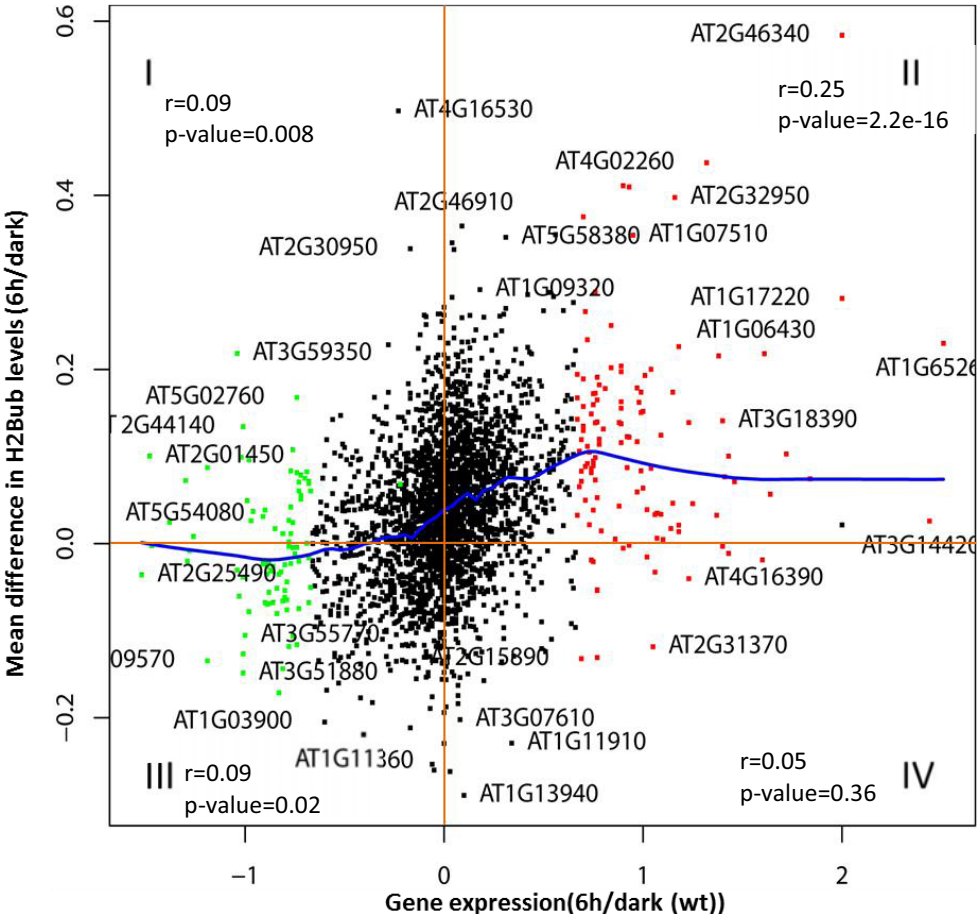


Figure 4.13: Scatterplot showing changes in expression versus changes in the H2Bub mark for genes marked in darkness and upon 6h of illumination.

The x-axis shows the differential expression of genes in response to 6h light and the y-axis represents the average difference in H2Bub levels for all probes contained within the gene. Red dots represent genes differentially upregulated and green dots show genes differentially downregulated in the light. The thick blue curve shows trend-line from LOWESS smoother function. Quadrant I represents light-downregulated genes that gain the H2Bub mark, quadrant II shows genes upregulated by light and which gain the H2Bub mark, quadrant III shows downregulated genes that lose the mark, while quadrant IV shows upregulated genes that lose the H2Bub mark in the light. The correlation coefficient (r) along with the significance of correlation is shown for each quadrant and there is a highly significant (p-value < 2.2e-16) positive correlation (r=0.6) between loss/gain of the H2Bub mark with loss/gain in gene expression.

Finally, to explore quantitatively the relationship between transcript level changes and the dynamics in H2Bub mark, H2Bub mean enrichment level was calculated for each light-regulated gene. These levels were analysed by individual boxplots for D, 1h and 6h samples and compared to all other genes. In this analysis, light-induced genes show considerably higher levels of the H2Bub mark in 1h and 6h compared to dark (Figure 4.14 A). In contrast, light-repressed genes do not display significant differences of the mark enrichment in this timescale.

Aiming to refine this result, we asked whether the most upregulated genes would show higher gain of the mark than moderately induced genes. To achieve this, the list of light-regulated genes was divided in expression quartiles for downregulation and upregulation based on mRNA levels at 6h light compared to dark, such that quartile 1 contains the least and quartile 4 contains the top-most upregulated or downregulated genes. Figure 4.14 B shows dynamics of the H2Bub mark for the four quartiles of downregulated and upregulated genes. For each of the quartiles, boxplots give the distribution of H2Bub enrichment levels in the dark, 1h and 6h conditions. Again, there does not appear to be a clear relationship between H2Bub content and the level of downregulation of the genes, as for most genes downregulation is not accompanied by a reduction of H2Bub enrichment level. Only for the fourth quartile, the topmost downregulated genes, the H2Bub mark is slightly lower in 6h compared to dark. This fourth quartile notably displays elevated H2Bub levels, a feature in agreement with expression being high in darkness for this class of genes. Based on these analyses, this result again suggests that, the H2Bub mark appears to be more stable than the corresponding mRNAs during downregulation. Remarkably, for the upregulated genes, a steady increase in the level of H2Bub enrichment is found as the degree of induction increases from quartile 1 to quartile 4. The

H2Bub levels in the dark are significantly lower than 1h and 6h in the last two expression quartiles of upregulated genes, indicative of a strong positive correlation between H2Bub enrichment levels and fold induction of transcription for light regulated genes. In addition, the fourth quartile is also characterized by elevated levels of H2Bub mark in the three conditions. Altogether these results therefore provide persuasive evidence implicating H2Bub in the selective regulation of gene expression. The analysis was also corroborated by taking an independent set of genes that corresponded to the pattern 6hr_gt_dark (6h > dark) in Table 4.2 A. These genes display a statistically significant gain of the H2Bub mark in the light (Figure 4.14 C), so I asked if there is a corresponding upregulation of gene expression. The results shown in Figure 4.14 D clearly indicate gain in H2Bub is markedly associated with a gain in gene expression from 1h to 6h in wt. Analysis of expression levels of these genes in the *hub1-3* mutant reveal that, as expected, they display a less pronounced gain of expression in the *hub1-3* mutant, again in agreement with a possible direct involvement of H2B ubiquitination in facilitating gene expression (Figure 4.14 D).

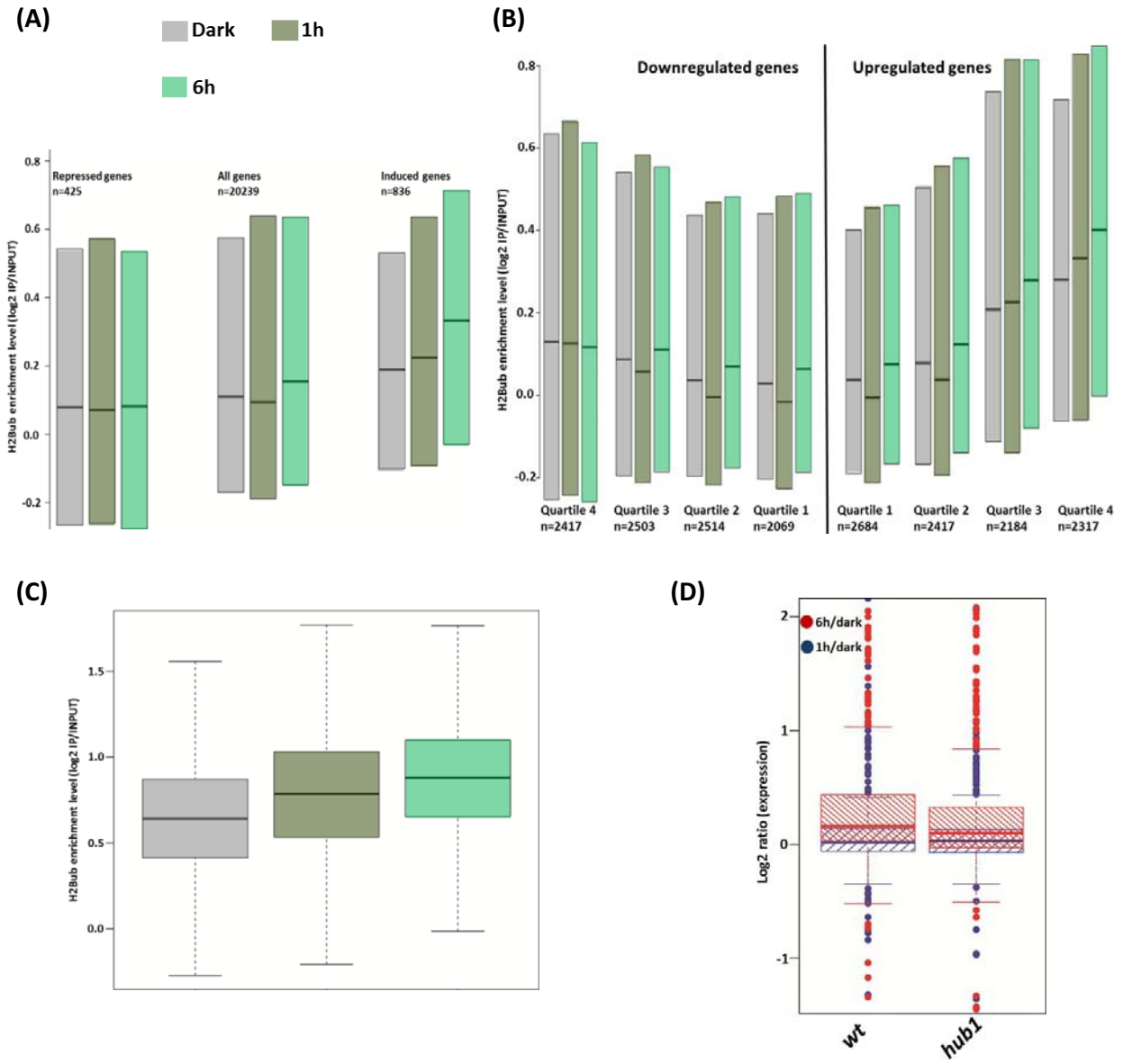


Figure 4.14: Gain of the H2Bub mark within genes associates to increased expression.

(A,B) Genes induced in the light show an increase in the amount of H2Bub marking in light compared to dark whereas repressed genes exhibit complex relationship with the H2Bub mark. **(C, D)** Conversely, genes that gain H2Bub mark in the light (n=408; as defined by the relationship 6hr_gt_dark) show a corresponding increase in transcript levels in the light from 1h to 6h. These genes display a less pronounced change in expression in the *hub1-3* mutant.

Discussion

It has recently emerged that mono-ubiquitination and deubiquitination of histones is a basal mechanism associated with transcription in a chromatin context and can contribute to regulate gene expression in major cellular processes (He 2009; Crevillén and Dean 2011; Frappier and Verrijzer 2011). In this study I have presented an integrated analysis of H2Bub dynamics onto light-regulated genes and tested the role of this chromatin mark in this process. To achieve this, I constructed genome-wide maps of ubiquitinated histone H2B during the dark to light transition in Arabidopsis, and analysed them together with expression data of wild type and *hub1* mutant plants to understand the role of H2Bub in the dynamics of gene expression changes.

First, before light exposure, I observed that in darkness a large majority (~80%) of the misregulated genes in *hub1* are downregulated, in agreement with a primary effect of HUB1 on gene activation. Although to a lower extent, this was also observed in early transcriptome analyses conducted on shoot apices using the *hub1-1* mutant in another *A. thaliana* accession and using an Affimetrix platform (Fleury, Himanen et al. 2007). This was the only direct comparison of wild type and *hub1* RNAs in the experimental design that we selected (Figure 4.3). For the purpose of this project, the genes misregulated in darkness were further compared to the set of light-regulated genes. This revealed that *hub1* mutation partially mimics the effect of light on gene expression, although the light effect is stronger compared to the *hub1* mutation. This property is reminiscent of *constitutive photomorphogenic* mutants such as *cop1* and *det1*, in which many light-induced genes are already de-repressed before illumination (Chory, Peto et al. 1989; Ma, Li et al. 2001; Schroeder, Gahrtz et al. 2002). This result is of

interest for studying the potential link between photomorphogenic regulators such as DET1 and the HUB1 pathway, as other work conducted in the laboratory recently revealed that DET1 has a general effect on H2Bub deposition over genes. Nonetheless, the question that I have been asking in this project was more particularly focused on the potential impact of H2Bub during gene expression changes, by using light signals as a paradigm to investigate the extent and the impact of H2Bub dynamics on gene regulation. This was primarily tested upon illumination, by sampling seedlings at 1 and 6h, two time points that had first been determined experimentally to be relevant for studying early and late-responding genes.

Combining epigenomic and transcriptomic data, the most notable results were on one hand the demonstration that H2Bub mark is highly dynamic on hundreds of genes and can be significantly enriched on many genes within as little as 1h, and on the other hand that a mutant with no H2Bub mark exhibit weakened expression changes. More precisely, many genes that are induced or repressed by light in wt seedlings were mostly induced or repressed by light also in the *hub1* mutant, albeit to a lesser extent. This is in agreement with a direct role for H2Bub in the regulation of gene expression changes by allowing appropriate levels of gene expression to be reached in response to a stimulus. Rather subtle effects could be detected, as for example we observed that a large fraction (73%) of early induced genes in wt show downregulation upon longer periods of light exposure and that, after their peak of induction, these genes show a weaker downregulation in the *hub1* mutant. Moreover, a fraction of the early-induced genes (27%) that still show upregulation in wt between 1h and 6h of illumination show weaker induction in *hub1*. More surprisingly, it was observed both for early and late-repressed genes that *hub1-1* mutation can affect downregulation (Figure 4.6 B and 4.6 C). Because of these

properties, we have been able to dissect an apparent increase level of early-induced transcripts in *hub1-3* at 6h: slower early accumulation of transcripts and reduced mRNA decrease after the peak of induction can both contribute to explain this (Figure 4.6 A).

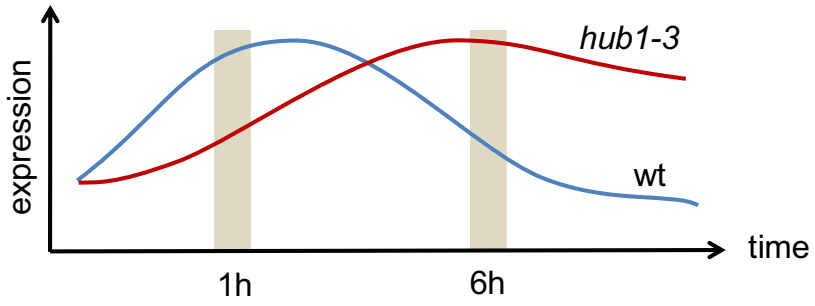


Figure 4.15: Model for the apparent increase of H2Bub levels on early light-induced genes at 6h.

Although few genes (25, Figure 4.11 C) were detected as loosing H2Bub mark within 6h, we observed that gene downregulation was usually not associated with a significant loss of the mark. This was clearly opposite to the concomitant increase in expression and in H2Bub enrichment observed for many light-induced genes, and might be explained by a higher stability of this chromatin mark onto genes than the corresponding transcripts. I therefore propose that the H2Bub mark could affect downregulation at two levels, either at the transcriptional level where a decrease of H2Bub would affect rates of transcription, or at a posttranscriptional level, where the ubiquitin system is involved either directly or indirectly in the degradation of mRNA transcripts (Cano, Miranda-Saavedra et al. 2010). This result therefore appeals novel experiments, such as longer kinetics to determine the average time for H2Bub pruning after

stopping the light stimulus, as a function of gene activity. It might also open the way to more pioneering experimental designs, in which wild type and *hub1* mutants would be compared for their capacity to respond to a second light stimulus, thereby assessing if the H2Bub mark could serve as a “memory" of previously elevated transcriptional activity and would predispose to novel gene inductions.

Finally, we observed that the H2Bub mark is most pronounced on longer genes. Indeed, light upregulated genes less than 2 kb in size do not show any significant differences in expression between wt and *hub1*, while genes longer than 2 kb show a significantly higher induction in wt compared to *hub1* (Figure 4.7). This result fits with the predicted role of H2Bub mark in facilitating transcription elongation, longer genes being more particularly sensitive to the kinetics of nucleosome eviction and/or opening of nucleosomal structures for RNA Polymerase processivity.

The observed defective mRNA decrease during light-driven downregulation in *hub1* mutant also raises some novel interrogations. One hypothesis has to consider possible effects of HUB1 on mRNA stability and/or degradation. Indeed, the time-course in our experiments is rapid, within 1 and 6h, and transcriptome analyses only inform on RNA steady-state levels. Pirngruber et al. (Pirngruber, Shchebet et al. 2009) reported that H2Bub is required for correct 3'-end mRNA processing in mammals, a role that might affect mRNA turnover and possibly explain a delayed mRNA degradation in the absence of HUB1. Another hypothesis, more directly linked to altered chromatin patterns in *hub1*, may reflect a scenario in which cryptic transcripts generated in the absence of the H2Bub mark and more stable than correct mRNAs, would be spuriously

detected. This is plausible because the impaired H2Bub machinery in *hub1* would be expected to lower the levels of H3K36me3, a mark that was shown to prevent initiation of cryptic transcripts from within the gene body and enhance transcriptional elongation (Crevillén and Dean 2011). This last scenario would nonetheless not be fully compatible with the observed weakened mRNA increase for light-induced genes in the *hub1* mutant.

Using several statistical analyses, I have shown that light-driven gene expression is globally associated with both increased levels and wider distribution of H2Bub over marked genes. A slight light-driven extension of H2Bub-enriched domains towards the 3' end of expressed genes could also be observed. Nonetheless, since chromatin extracts were prepared on whole organisms, this approach does not allow determining whether genes already marked by H2Bub in darkness gain an additional level of the mark or whether an increased number of cells acquire the mark on those genes. The progressive increase of H2Bub on light-induced genes observed in our experiments is compatible with both hypotheses.

References:

Allemeersch, J., S. Durinck, et al. (2005). "Benchmarking the CATMA Microarray. A Novel Tool forArabidopsis Transcriptome Analysis." Plant Physiology **137**(2): 588 -601.

Belotserkovskaya, R., S. Oh, et al. (2003). "FACT facilitates transcription-dependent nucleosome alteration." Science (New York, N.Y.) **301**(5636): 1090-1093.

Benhamed, M., C. Bertrand, et al. (2006). "Arabidopsis GCN5, HD1, and TAF1/HAF2 interact to regulate histone acetylation required for light-responsive gene expression." The Plant Cell **18**(11): 2893-2903.

Benvenuto, G., F. Formiggini, et al. (2002). "The photomorphogenesis regulator DET1 binds the amino-terminal tail of histone H2B in a nucleosome context." Current Biology: CB **12**(17): 1529-1534.

Betz, J. L., M. Chang, et al. (2002). "Phenotypic analysis of Paf1/RNA polymerase II complex mutations reveals connections to cell cycle regulation, protein synthesis, and lipid and nucleic acid metabolism." Molecular Genetics and Genomics: MGG **268**(2): 272-285.

Bray, S., H. Musisi, et al. (2005). "Bre1 is required for Notch signaling and histone modification." Developmental Cell **8**(2): 279-286.

Cano, F., D. Miranda-Saavedra, et al. (2010). "RNA-binding E3 ubiquitin ligases: novel players in nucleic acid regulation." Biochemical Society Transactions **38**(6): 1621-1626.

Cao, Y., Y. Dai, et al. (2008). "Histone H2B monoubiquitination in the chromatin of FLOWERING LOCUS C regulates flowering time in Arabidopsis." The Plant Cell **20**(10): 2586-2602.

Carrozza, M. J., B. Li, et al. (2005). "Histone H3 methylation by Set2 directs deacetylation of coding regions by Rpd3S to suppress spurious intragenic transcription." Cell **123**(4): 581-592.

Charron, J.-B. F., H. He, et al. (2009). "Dynamic landscapes of four histone modifications during deetiolation in Arabidopsis." The Plant Cell **21**(12): 3732-3748.

Chory, J., C. Peto, et al. (1989). "Arabidopsis thaliana mutant that develops as a light-grown plant in the absence of light." Cell **58**(5): 991-999.

Crevillén, P. and C. Dean (2011). "Regulation of the floral repressor gene FLC: the complexity of transcription in a chromatin context." Current Opinion in Plant Biology **14**(1): 38-44.

Culhane, A. C., J. Thioulouse, et al. (2005). "MADE4: an R package for multivariate analysis of gene expression data." Bioinformatics (Oxford, England) **21**(11): 2789-2790.

Deng, X. W. and P. H. Quail (1999). "Signalling in light-controlled development." Seminars in Cell & Developmental Biology **10**(2): 121-129.

Eden, E., R. Navon, et al. (2009). "GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists." BMC Bioinformatics **10**(1): 48.

Fleury, D., K. Himanen, et al. (2007). "The Arabidopsis thaliana homolog of yeast BRE1 has a function in cell cycle regulation during early leaf and root growth." The Plant Cell **19**(2): 417-432.

Formosa, T. (2008). "FACT and the reorganized nucleosome." Molecular bioSystems **4**(11): 1085-1093.

Frappier, L. and C. P. Verrijzer (2011). "Gene expression control by protein deubiquitinases." Current Opinion in Genetics & Development **21**(2): 207-213.

Gu, X., D. Jiang, et al. (2009). "Repression of the floral transition via histone H2B monoubiquitination." The Plant Journal: For Cell and Molecular Biology **57**(3): 522-533.

He, Y. (2009). "Control of the transition to flowering by chromatin modifications." Molecular Plant **2**(4): 554-564.

Henry, K. W., A. Wyce, et al. (2003). "Transcriptional activation via sequential histone H2B ubiquitylation and deubiquitylation, mediated by SAGA-associated Ubp8." Genes & Development **17**(21): 2648-2663.

Hwang, W. W., S. Venkatasubrahmanyam, et al. (2003). "A conserved RING finger protein required for histone H2B monoubiquitination and cell size control." Molecular Cell **11**(1): 261-266.

Jamai, A., A. Puglisi, et al. (2009). "Histone chaperone spt16 promotes redeposition of the original h3-h4 histones evicted by elongating RNA polymerase." Molecular Cell **35**(3): 377-383.

Ji, H., H. Jiang, et al. (2008). "An integrated software system for analyzing ChIP-chip and ChIP-seq data." Nature Biotechnology **26**(11): 1293-1300.

Ji, H. and W. H. Wong (2005). "TileMap: create chromosomal map of tiling array hybridizations." Bioinformatics (Oxford, England) **21**(18): 3629-3636.

Jiang, D., X. Gu, et al. (2009). "Establishment of the winter-annual growth habit via FRIGIDA-mediated histone methylation at FLOWERING LOCUS C in Arabidopsis." The Plant Cell **21**(6): 1733-1746.

Jiang, D., N. C. Kong, et al. (2011). "Arabidopsis COMPASS-like complexes mediate histone H3 lysine-4 trimethylation to control floral transition and plant development." PLoS Genetics **7**(3): e1001330.

Jiang, H. and W. H. Wong (2008). "SeqMap: mapping massive amount of oligonucleotides to the genome." Bioinformatics (Oxford, England) **24**(20): 2395-2396.

Jiao, Y., L. Ma, et al. (2005). "Conservation and divergence of light-regulated genome expression patterns during seedling development in rice and Arabidopsis." The Plant Cell **17**(12): 3239-3256.

Kao, C.-F., C. Hillyer, et al. (2004). "Rad6 plays a role in transcriptional activation through ubiquitylation of histone H2B." Genes & Development **18**(2): 184-195.

Kim, J., M. Guermah, et al. (2009). "RAD6-Mediated transcription-coupled H2B ubiquitylation directly stimulates H3K4 methylation in human cells." Cell 137(3): 459-71.

Liu, Y., M. Koornneef, et al. (2007). "The absence of histone H2B monoubiquitination in the Arabidopsis hub1 (rdo4) mutant reveals a role for chromatin remodeling in seed dormancy." The Plant Cell **19**(2): 433-444.

Lolas, I. B., K. Himanen, et al. (2010). "The transcript elongation factor FACT affects Arabidopsis vegetative and reproductive development and genetically interacts with HUB1/2." The Plant Journal: For Cell and Molecular Biology **61**(4): 686-697.

Loudet, O., T. P. Michael, et al. (2008). "A zinc knuckle protein that negatively controls morning-specific growth in Arabidopsis thaliana." Proceedings of the National Academy of Sciences of the United States of America **105**(44): 17193-17198.

Ma, L., J. Li, et al. (2001). "Light control of Arabidopsis development entails coordinated regulation of genome expression and cellular pathways." The Plant Cell **13**(12): 2589-2607.

Martin-Magniette, M.-L., T. Mary-Huard, et al. (2008). "ChIPmix: mixture model of regressions for two-color ChIP-chip analysis." Bioinformatics (Oxford, England) **24**(16): i181-186.

Mason, P. B. and K. Struhl (2003). "The FACT complex travels with elongating RNA polymerase II and is important for the fidelity of transcriptional initiation in vivo." Molecular and Cellular Biology **23**(22): 8323-8333.

Neff, M. M., C. Fankhauser, et al. (2000). "Light: an indicator of time and place." Genes & Development **14**(3): 257-271.

Ng, H. H., F. Robert, et al. (2003). "Targeted recruitment of Set1 histone methylase by elongating Pol II provides a localized mark and memory of recent transcriptional activity." Molecular Cell **11**(3): 709-719.

Pavri, R., B. Zhu, et al. (2006). "Histone H2B monoubiquitination functions cooperatively with FACT to regulate elongation by RNA polymerase II." Cell **125**(4): 703-717.

Pickart, C. M. (2001). "Mechanisms underlying ubiquitination." Annual Review of Biochemistry **70**: 503-533.

Pirngruber, J., A. Shchebet, et al. (2009). "CDK9 directs H2B monoubiquitination and controls replication-dependent histone mRNA 3'-end processing." EMBO Reports **10**(8): 894-900.

Reinberg, D. and R. J. Sims, 3rd (2006). "de FACTo nucleosome dynamics." The Journal of Biological Chemistry **281**(33): 23297-23301.

Robzyk, K., J. Recht, et al. (2000). "Rad6-dependent ubiquitination of histone H2B in yeast." Science (New York, N.Y.) **287**(5452): 501-504.

Roudier, F., I. Ahmed, et al. (2011). "Integrative epigenomic mapping defines four main chromatin states in Arabidopsis." The EMBO Journal **30**(10): 1928-1938.

Schmitz, R. J., Y. Tamada, et al. (2009). "Histone H2B deubiquitination is required for transcriptional activation of FLOWERING LOCUS C and for proper control of flowering in Arabidopsis." Plant Physiology **149**(2): 1196-1204.

Schroeder, D. F., M. Gahrtz, et al. (2002). "De-etiolated 1 and damaged DNA binding protein 1 interact to regulate Arabidopsis photomorphogenesis." Current Biology: CB **12**(17): 1462-1472.

Smyth, G. K. (2005). Limma: Linear Models for microarray data. New York, Springer.

Sridhar, V. V., A. Kapoor, et al. (2007). "Control of DNA methylation and heterochromatic silencing by histone H2B deubiquitination." Nature **447**(7145): 735-738.

Tamayo, P., D. Slonim, et al. (1999). "Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation." Proceedings of the National Academy of Sciences of the United States of America **96**(6): 2907-2912.

Wood, A., N. J. Krogan, et al. (2003). "Bre1, an E3 ubiquitin ligase required for recruitment and substrate selection of Rad6 at a promoter." Molecular Cell **11**(1): 267-274.

Xiao, T., C.-F. Kao, et al. (2005). "Histone H2B ubiquitylation is associated with elongating RNA polymerase II." Molecular and Cellular Biology **25**(2): 637-651.

Xin, H., S. Takahata, et al. (2009). "yFACT induces global accessibility of nucleosomal DNA without H2A-H2B displacement." Molecular Cell **35**(3): 365-376.

Xu, L., R. Ménard, et al. (2009). "The E2 ubiquitin-conjugating enzymes, AtUBC1 and AtUBC2, play redundant roles and are involved in activation of FLC expression and repression of flowering in Arabidopsis thaliana." The Plant Journal: For Cell and Molecular Biology **57**(2): 279-288.

Zhang, X., Y. V. Bernatavichute, et al. (2009). "Genome-wide analysis of mono-, di- and trimethylation of histone H3 lysine 4 in Arabidopsis thaliana." Genome Biology **10**(6): R62.

Zhang, X., O. Clarenz, et al. (2007). "Whole-genome analysis of histone H3 lysine 27 trimethylation in Arabidopsis." PLoS Biology **5**(5): e129.

Zhu, B., Y. Zheng, et al. (2005). "Monoubiquitination of human histone H2B: the factors involved and their roles in HOX gene regulation." Molecular Cell **20**(4): 601-611.

CHAPTER V

DISCUSSION

Methylation of DNA is a hallmark of epigenetic inactivation and heterochromatin in both plants and mammals, typically associated with a silenced chromatin state, and is largely confined to transposable element (TE) and other repeat sequences. Transposable elements, particularly retrotransposons, are known to be activated by biotic and abiotic stresses (Wessler 1996) through epigenetic changes like reduction in DNA methylation or global hypomethylation. However, the fixation of new TE insertion depends on events not only at the level of the nucleus, but also at the population level as it is difficult for a genetic variant to become fixed except in small inbred populations (Husband 2004; Belyayev, Kalendar et al. 2010). TE sequences, because of their inherent capability to move from one position to another in a genome, are potentially deleterious, and are the main targets of RNA-directed DNA methylation (RdDM)-mediated transcriptional gene silencing (TGS). RdDM is important for the establishment of DNA methylation in TEs and comparison of genome-wide methylation patterns and small RNAs in Arabidopsis indicated that at least one-third of methylated loci are associated with siRNA clusters (Lister, O'Malley et al. 2008). Notwithstanding, it is not very clear how the RdDM machinery can recognize and specifically act on transposons and not on host genes. It has been postulated that TE sequences may possess some features like specific secondary structures or long untranslated regions that make them more susceptible to the RdDM methylation machinery; however, several studies in Arabidopsis have indicated that nothing intrinsic to the TE sequences provides this recognition feature (Miura, Yonebayashi et al. 2001; Kato, Takashima et al. 2004; Lippman, Gendrel et al. 2004; Mirouze, Reinders et al. 2009). A possible characteristic feature of transposons can be considered to be their capability of producing double-stranded RNAs (dsRNAs) either through a convergent/divergent

transcription from the two strands or through the action of RNA-dependent RNA polymerase (RdRP). These dsRNAs act as precursors to produce the siRNAs that target the RdDM machinery to the transposon sequences in the genome.

Loss of DNA methylation, as in the mutants of DNA methyltransferases or chromatin remodeler protein DDM1 (decrease in DNA methylation 1), de-represses and mobilizes several TE sequences (Miura, Yonebayashi et al. 2001; Lippman, Gendrel et al. 2004; Mirouze, Reinders et al. 2009). TEs activated in these mutants remain active even after introduction into a wild-type background which indicates that the silencing of TE sequences depends on the silent state in the previous generations. An efficient remethylation mechanism that corrects for the transgenerational DNA methylation defects was discovered by Teixeira et al. (Teixeira, Heredia et al. 2009), who showed that a subset of TE sequences is remethylated after introduction into a wild-type background. This selective and progressive remethylation is facilitated by RNA interference (RNAi) and is dependent on the activity of RDR2 (RNA-dependent RNA polymerase), a known component of RdDM.

The RNA-directed DNA methylation thus is regarded as an important player that dictates the establishment of methylation and maintains its inheritance in a transgenerational manner. It is often considered to be associated with dense methylation of TE sequences at CG and non-CG sites. However, my analysis in Chapter II of this thesis shows that an unexpectedly high number of the densely methylated TE sequences are characterized by the absence or near absence of matching siRNAs. I have provided evidence that in such cases DNA methylation most likely results from local spreading from flanking, siRNA-targeted sequences. This spreading of DNA

methylation from siRNA-targeted sequences could be seen in both heterochromatin and euchromatin albeit to different extents (~500 bp and ~200 bp, respectively), which reflects either a facilitating role of heterochromatin in promoting DNA methylation spread, an inhibitory effect of euchromatin, or a higher DNA demethylation activity in euchromatin. The results presented in Chapter II provide evidence that the phenomenon of DNA methylation spread could result from a complex set of interactions between different DNA methyltransferases that promote or limit the extent of DNA methylation spread and does not involve any significant contribution from active DNA demethylation. However, active DNA demethylation cannot be ruled out for a small subset of methylated TE sequences located near to critical genes, as recently reported by Lister et al., where they identified a subset of TE insertions near to genes that are unmethylated in the wild-type background but are hypermethylated in the DNA demethylase triple mutant *rdd* (Lister, O'Malley et al. 2008).

I have also provided evidence that TE sequences in euchromatin have an insertional preference for open chromatin and are more abundant closer to genes. However, the distribution of unmethylated versus methylated TE sequences is somewhat skewed from the 5'-end of genes, and methylated TE sequences associated with siRNAs tend to be least abundant and uniformly distributed within the first 500 bp flank from the 5'-end of genes. These results therefore suggested that methylated TE sequences have more deleterious effects on transcription initiation than termination and that these effects are more severe when methylated TE sequences have matching siRNAs. Given that DNA methylation has an inhibitory effect on promoter activity, a DNA methylation spread could therefore explain in part the negative

impact of methylated TE sequences on neighbouring gene expression and hence partial purging of methylated compared to unmethylated TE sequences from the close vicinity of genes.

In light of these new results, it would be interesting to determine and capture the de novo insertions of TE sequences and examine how they impact the epigenomic context of their neighbouring genes and influence their expression. An experimental approach to determine transposon locations and detect any new TE insertions would be probably based on the identification of the contiguous flanking DNA sequences of the TEs. While individual members of a family of transposons are highly conserved, their flanking sequences are likely to be unique. Thus one of the approaches to identify new TE insertions would be to determine their contiguous DNA sequences either by deep sequencing or by use of microarray technology to capture TE borders (Gabriel, Dapprich et al. 2006; Wheelan, Scheifele et al. 2006). These data could be combined with genome-wide maps of epigenetic marks like DNA methylation to ascertain the time frame for establishing new epigenetic states and follow the stability of these epigenetic changes over time.

In Chapter III, the genome-wide mapping of a broad set of 11 histone marks was reported, along with DNA methylation. We first showed that although a large number of theoretical permutations of DNA methylation and histone marks are possible, only a limited number of such combinations are actually observed, reflecting the likelihood that locus-specific combinations of chromatin marks are involved in epigenetic control mechanisms and are not merely random chance events. Further analysis using fuzzy c-means clustering partitioned the Arabidopsis epigenome into four major chromatin types with distinct functional properties.

These chromatin states can be mainly distinguished into two groups as evidenced by pair-wise co-association analysis, and form either constitutive heterochromtin (denoted CS3) or three different flavours of euchromatic regions (denoted CS1, CS2 and CS4). The first chromatin state (CS1) corresponds to transcriptionally active genes and is typically enriched in H3K4me3, H3K36me3, H3K9me3, H2Bub, H3K4me2 and H3K56 chromatin marks. The second chromatin state (CS2) represents repressive chromatin and corresponds to H3K27me3-marked regions that are mainly associated with genes under PRC2-mediated repression (Turck, Roudier et al. 2007; Zhang, Clarenz et al. 2007). The other prevalent signature associated with this chromatin state is H3K27me2. The third chromatin state (CS3) that corresponds to classical heterochromatin is almost exclusively located over silent TEs (Lippman, Gendrel et al. 2004; Bernatavichute, Zhang et al. 2008) and comprises regions marked by H3K9me2 and H4K20me1. This state is often co-marked by mono- or dimethylated forms of H3K27 and DNA methylation (5mC). However, the two marks 5mC and H3K27me2 are also found on genes and mark ~33% and ~31% of the genes, respectively. A fourth chromatin state (CS4) is characterized by the absence of any prevalent mark and is associated with weakly expressed genes and intergenic regions.

Genome-wide studies of epigenomic landscapes have also been conducted in yeast (Liu, Kaplan et al. 2005), *C. elegans* (Gerstein, Lu et al. 2010; Liu, Rechtsteiner et al. 2011), Drosophila (Roy, Ernst et al. 2010; Kharchenko, Alekseyenko et al. 2011; Riddle, Minoda et al. 2011) and human cells (Wang, Zang et al. 2008; Hon, Wang et al. 2009; Ernst and Kellis 2010; Zhou, Goren et al. 2011), all of which have similarly indicated a relatively low combinatorial complexity of chromatin marks. The two silent chromatin states defined in Arabidopsis (CS2 & CS3) have

similar counterparts in metazoans, indicating the high level of conservation for these states in plants and animals. On the contrary several chromatin states have been defined for transcriptionally active genes in other organisms which correspond to a single predominant cluster (CS1) of transcriptionally active genes in Arabidopsis. This discrepancy might be due to the fact that the Arabidopsis genome is comparatively smaller with shorter gene size and intergenic length, suggesting that gene length might have an important role in the creation of chromatin states.

Using an integrative analysis of the distribution of these 12 chromatin marks, we demonstrated that the Arabidopsis epigenome shows simple principles of organisation and that chromatin states are mainly defined at the level of single transcription units. Our analysis also confirmed that while H2Bub, H3K36me3, 5mC, as well as H3K4me2 and H3K4me3, tend to be associated with longer genes, H3K27me3 on the other hand preferentially marks smaller genes as noted before (Luo and Lam 2010). Unlike these chromatin modifications, H3K27me1 does not exhibit any preferential association in relation to gene length. Further we also confirmed that H3K4me3 and H3K56ac tend to mark highly and broadly expressed genes (Oh, Park et al. 2008; Zhang, Bernatavichute et al. 2009), and that H3K27me3 preferentially associates with genes that are expressed at lower levels or in a tissue-specific manner (Turck, Roudier et al. 2007; Zhang, Clarenz et al. 2007; Oh, Park et al. 2008), while 5mC tends to mark moderately expressed genes (Zilberman, Gehring et al. 2007). In addition, we noted that H2Bub, H3K36me3 and H3K9me3 tend to mark highly expressed genes, like H3K4me3 and H3K56ac. H3K4me2 does not show any particular association with gene expression, which could be rather dependent on its co-occurrence with other marks like H3K4me3 or H3K27me3. On the other

hand, genes marked by H3K27me1 or H3K27me2 are expressed at lower levels or in a tissue-specific manner. However, H3K27me1 and H3K27me2/3 mark largely non-overlapping sets of genes with different ontologies, which suggest the existence of two distinct gene repression systems associated with methylation of H3K27. We also observed that enrichment levels for most chromatin marks are correlated with expression levels, but whether these correlations reflect expression of genes in a variable number of cells, or true differential enrichment in relation to expression level, remains to be determined. As each cell-type possesses a unique epigenome, which defines its gene regulatory programme, it is ultimately only the knowledge of cell-type specific epigenomes that will enable a full understanding of the functional impact of chromatin-level regulation on genome activity. Next generation sequencing technologies are constantly being improved to obtain high-quality sequence data from a genome isolated from a single cell. Epigenetic studies by single-molecule real-time sequencing (Pushkarev, Neff et al. 2009; Korlach, Bjornson et al. 2010) are likely to revolutionise epigenetic research in the future and to eliminate many problems of interpreting epigenomes, which currently are the products of a mixed population of cells/chromosomes.

Finally, in the last Results chapter of this thesis (Chapter IV), a specifically chosen histone modification (H2Bub) was used to provide a dynamic view of this mark and its effects on the transcriptional state of the underlying chromatin in response to light-driven developmental adaptation. The H2Bub mark associates with more than 10,000 genes in *A. thaliana* wild-type seedlings. As expected from its role in yeast and mammals, we observed that H2Bub is restricted to transcribed regions and mostly associates with active genes (Roudier, Ahmed et al. 2011). Moreover, the H2Bub histone mark in Arabidopsis correlates with several other histone

marks associated with active gene transcription, namely histone H3 trimethylation on lysines 4, 9 and 36. The mutant plants of H2B ubiquitin-ligase (HUB),which is a homolog of the yeast Bre1 protein (Fleury, Himanen et al. 2007), show complete loss of the H2Bub mark throughout the genome. Indeed, HUB proteins act as a heterodimer of the HUB1 and HUB2 paralogs, and consequently a null mutation in the HUB1 or in the HUB2 gene abolishes H2Bub deposition. The HUB mutant (*hub1-3*) exhibits no visible residual H2Bub mark and provides an ideal tool to probe global roles of this histone PTM.

Considering the dynamic nature of the H2Bub mark on gene regulation in yeast and mammals, and the fragmentary knowledge on the transcriptional coactivators acting through H2B ubiquitination, I chose to specifically focus on the H2Bub chromatin mark. More generally, the role of histone mono-ubiquitination has still not been properly investigated in plants, apart from the recent description of H2A ubiquitination associated to Polycomb-related machineries and the role of H2Bub deposition in the control of *FLC* gene expression in *A. thaliana* (Cao, Dai et al. 2008; Bratzel, López-Torrejón et al. 2010).

Furthermore, early work in the laboratory has revealed that Arabidopsis *hub1-3* mutant seedlings, in which H2Bub is completely abolished, display defects in the de-etiolation process, suggesting a role for H2Bub in the dark to light transition. Therefore, in Chapter IV of this thesis, the *hub1-3* mutant was used as a tool to combine transcriptomic and epigenomic analyses that I have integrated to assess the impact of H2Bub dynamics on light-driven transcriptional responses. My analysis has indicated that H2Bub may not be required for determining ON/OFF gene activation states, but may rather contribute to the dynamics of expression and selectively

regulate the fine-tuning of gene expression in response to light signals. H2Bub enrichment over marked genes shows a peak value in the middle of genes. This steady-state distribution of the H2Bub mark over enriched genes resembles a Gaussian curve and probably reflects confounding effects of simultaneous ubiquitination and deubiquitination processes occurring on expressed genes rather than an increased ubiquitination in the middle. I have demonstrated that the H2Bub mark is highly dynamic on hundreds of genes and significant changes in the enrichment levels occur on many genes within as little as 1 hour of light illumination. Although only a few genes were found to show a loss of H2Bub enrichment within the examined 6 hour period of light, we observed that gene downregulation was usually not associated with a significant loss of the mark. This is in clear contrast with the steady gain of this mark associated with a corresponding gain in expression observed for many light-induced genes. These observations might be explained by a higher stability of this chromatin mark onto genes than the corresponding transcripts which therefore prompts to propose that the H2Bub mark could affect downregulation at two levels, either at the transcriptional level where a decrease of H2Bub would affect rates of transcription, or at a posttranscriptional level, where the ubiquitin system is involved either directly or indirectly in the degradation of mRNA transcripts (Cano, Miranda-Saavedra et al. 2010). This result therefore suggests additional experiments, such as performing longer kinetics to determine the average time for H2Bub pruning after stopping the light stimulus, as a function of gene activity. Finally, we observed that the H2Bub mark is most pronounced on longer genes, because light upregulated genes less than 2 kb in size did not show any significant differences in expression between *wt* and *hub1*, while genes longer than 2 kb showed a significantly higher induction in *wt* compared to *hub1*. This result fits with the

predicted role of the H2Bub mark in facilitating transcription elongation, longer genes being more particularly sensitive to the kinetics of nucleosome eviction and/or opening of nucleosomal structures for RNA Polymerase processivity.

In this thesis I have provided a spatial and temporal view of the *A. thaliana* epigenome using significant contributions from bioninformatics. All experimental methods for epigenome mapping generate enormous quantities of data that require efficient methods for low-level data processing and quality control. While the development of computational tools and resources for epigenome-wide data analysis is accelerating fast, various bioinformatics challenges still exist that arise from the analysis of epigenomic data. These include, but are not limited to, considerations for hybridization-based methods like image and scanning artefacts, background correction, array normalization, and for sequence-based analyses involve alignment to a reference genome with accuracy, speed, memory usage and flexibility of the alignment programmes. Other considerations for sequencing-based methods would be to allow for mis-matches and multiple matches in the alignment, and the requirement for longer or paired-end reads. A great deal of algorithms have been written for finding the peaks or bound regions for ChIP-on-chip or ChIP-seq datasets, but the peak finding problem has still not been fully solved. For example, most of the currently available peak finding algorithms do not work particularly well with the histone modifications that cover extended domains like H3K27me3 and they seem to miss a substantial number of weak binding regions for such modifications. Finally, as epigenomic data is growing at an unprecedented rate, the development of powerful and user-friendly Genome Browsers is the need of the hour to ease our understanding and interpretation of the biologically meaningful information from the data.

References:

Belyayev, A., R. Kalendar, et al. (2010). "Transposable elements in a marginal plant population: temporal fluctuations provide new insights into genome evolution of wild diploid wheat." Mobile DNA **1**(1): 6.

Bernatavichute, Y. V., X. Zhang, et al. (2008). "Genome-wide association of histone H3 lysine nine methylation with CHG DNA methylation in Arabidopsis thaliana." PloS One **3**(9): e3156.

Bratzel, F., G. López-Torrejón, et al. (2010). "Keeping cell identity in Arabidopsis requires PRC1 RING-finger homologs that catalyze H2A monoubiquitination." Current Biology: CB **20**(20): 1853-1859.

Cano, F., D. Miranda-Saavedra, et al. (2010). "RNA-binding E3 ubiquitin ligases: novel players in nucleic acid regulation." Biochemical Society Transactions **38**(6): 1621-1626.

Cao, Y., Y. Dai, et al. (2008). "Histone H2B monoubiquitination in the chromatin of FLOWERING LOCUS C regulates flowering time in Arabidopsis." The Plant Cell **20**(10): 2586-2602.

Ernst, J. and M. Kellis (2010). "Discovery and characterization of chromatin states for systematic annotation of the human genome." Nature Biotechnology **28**(8): 817-825.

Fleury, D., K. Himanen, et al. (2007). "The Arabidopsis thaliana homolog of yeast BRE1 has a function in cell cycle regulation during early leaf and root growth." The Plant Cell **19**(2): 417-432.

Gabriel, A., J. Dapprich, et al. (2006). "Global Mapping of Transposon Location." PLoS Genet **2**(12): e212.

Gerstein, M. B., Z. J. Lu, et al. (2010). "Integrative analysis of the Caenorhabditis elegans genome by the modENCODE project." Science (New York, N.Y.) **330**(6012): 1775-1787.

Hon, G., W. Wang, et al. (2009). "Discovery and annotation of functional chromatin signatures in the human genome." PLoS Computational Biology **5**(11): e1000566.

Husband, B. C. (2004). "Chromosomal variation in plant evolution." American Journal of Botany **91**(4): 621 -625.

Kato, M., K. Takashima, et al. (2004). "Epigenetic control of CACTA transposon mobility in Arabidopsis thaliana." Genetics **168**(2): 961-969.

Kharchenko, P. V., A. A. Alekseyenko, et al. (2011). "Comprehensive analysis of the chromatin landscape in Drosophila melanogaster." Nature **471**(7339): 480-5.

Korlach, J., K. P. Bjornson, et al. (2010). "Real-time DNA sequencing from single polymerase molecules." Methods in Enzymology **472**: 431-455.

Lippman, Z., A. V. Gendrel, et al. (2004). "Role of transposable elements in heterochromatin and epigenetic control." Nature **430**(6998): 471-6.

Lister, R., R. C. O'Malley, et al. (2008). "Highly integrated single-base resolution maps of the epigenome in Arabidopsis." Cell **133**(3): 523-36.

Liu, C. L., T. Kaplan, et al. (2005). "Single-nucleosome mapping of histone modifications in S. cerevisiae." PLoS Biology **3**(10): e328.

Liu, T., A. Rechtsteiner, et al. (2011). "Broad chromosomal domains of histone modification patterns in C. elegans." Genome Res **21**(2): 227-36.

Luo, C. and E. Lam (2010). "ANCORP: a high-resolution approach that generates distinct chromatin state models from multiple genome-wide datasets." The Plant Journal: For Cell and Molecular Biology **63**(2): 339-351.

Mirouze, M., J. Reinders, et al. (2009). "Selective epigenetic control of retrotransposition in Arabidopsis." Nature **461**(7262): 427-430.

Miura, A., S. Yonebayashi, et al. (2001). "Mobilization of transposons by a mutation abolishing full DNA methylation in Arabidopsis." Nature **411**(6834): 212-214.

Oh, S., S. Park, et al. (2008). "Genic and global functions for Paf1C in chromatin modification and gene expression in Arabidopsis." PLoS Genetics **4**(8): e1000077.

Pushkarev, D., N. F. Neff, et al. (2009). "Single-molecule sequencing of an individual human genome." Nature Biotechnology **27**(9): 847-850.

Riddle, N. C., A. Minoda, et al. (2011). "Plasticity in patterns of histone modifications and chromosomal proteins in Drosophila heterochromatin." Genome Research **21**(2): 147-163.

Roudier, F., I. Ahmed, et al. (2011). "Integrative epigenomic mapping defines four main chromatin states in Arabidopsis." The EMBO Journal **30**(10): 1928-1938.

Roy, S., J. Ernst, et al. (2010). "Identification of functional elements and regulatory circuits by Drosophila modENCODE." Science (New York, N.Y.) **330**(6012): 1787-1797.

Teixeira, F. K., F. Heredia, et al. (2009). "A role for RNAi in the selective correction of DNA methylation defects." Science **323**(5921): 1600-4.

Turck, F., F. o. Roudier, et al. (2007). "Arabidopsis TFL2/LHP1 specifically associates with genes marked by trimethylation of histone H3 lysine 27." PLoS Genetics **3**(6): e86.

Wang, Z., C. Zang, et al. (2008). "Combinatorial patterns of histone acetylations and methylations in the human genome." Nat Genet **40**(7): 897-903.

Wessler, S. R. (1996). "Turned on by stress. Plant retrotransposons." Current Biology: CB **6**(8): 959-961.

Wheelan, S. J., L. Z. Scheifele, et al. (2006). "Transposon insertion site profiling chip (TIP-chip)." Proceedings of the National Academy of Sciences **103**(47): 17632 -17637.

Zhang, X., Y. V. Bernatavichute, et al. (2009). "Genome-wide analysis of mono-, di- and trimethylation of histone H3 lysine 4 in Arabidopsis thaliana." Genome Biology **10**(6): R62.

Zhang, X., O. Clarenz, et al. (2007). "Whole-genome analysis of histone H3 lysine 27 trimethylation in Arabidopsis." PLoS Biol **5**(5): e129.

Zhou, V. W., A. Goren, et al. (2011). "Charting histone modifications and the functional organization of mammalian genomes." Nat Rev Genet **12**(1): 7-18.

Zilberman, D., M. Gehring, et al. (2007). "Genome-wide analysis of Arabidopsis thaliana DNA methylation uncovers an interdependence between methylation and transcription." Nature Genetics **39**(1): 61-69.

Résumé

Les génomes nucléaires eucaryotes sont empaquetés au sein d'une structure nucléoprotéique appelée chromatine et dont l'unité fondamentale est le nucléosome. Celui-ci est composé d'un octamère d'histones, contenant deux molécules de chacune des histones H2A, H2B, H3 et H4, autour duquel 147 pb d'ADN sont enroulées. Les modifications post-traductionnelles (PTMs) des histones et de l'ADN (méthylation des cytosines) constituent des marqueurs épigénomiques primaires qui participent à la régulation et au contrôle de l'accessibilité des différentes régions du génome. Ainsi, la chromatine forme une structure dynamique influencée par les changements environnementaux et développementaux et contribue à orchestrer diverses fonctions du génome. L'objectif principal de ma thèse était de caractériser l'organisation spatiale et la dynamique temporelle des états chromatiniens chez Arabidopsis par des approches à l'échelle du génome permettant l'étude des profils de méthylation de l'ADN et d'un ensemble de modifications post-traductionnelles des histones.

La méthylation de l'ADN, une marque caractéristique de l'inactivation épigénétique et de l'hétérochromatine chez les plantes et les mammifères, est largement confinée aux séquences répétées, dont les éléments transposables (TEs). Par mon travail de thèse, j'ai montré que chez Arabidopsis les séquences de TEs faiblement méthylées ou non associées à des petits ARN interférents (siRNAs), donc potentiellement non régulées par la machinerie de RNA-directed DNA methylation (RdDM), peuvent acquérir une méthylation de l'ADN par diffusion à partir de séquences adjacentes ciblées par les siRNAs. Cette diffusion de la méthylation de l'ADN sur des régions promotrices pourrait expliquer, au moins en partie, l'impact négatif des TEs associés à des siRNAs sur l'expression des gènes à proximité immédiate. Dans une seconde partie de ma thèse, j'ai contribué à l'analyse intégrée de la méthylation de l'ADN et de onze modifications des histones. L'utilisation d'analyses combinatoires et en cluster m'a permis de montrer que l'épigénome d'Arabidopsis présente des principes simples d'organisation. En effet, ces analyses nous ont conduit à distinguer quatre états fondamentaux de la chromatine chez Arabidopsis, préférentiellement associés aux gènes actifs, aux gènes inactifs, aux TEs et aux régions intergéniques. Dans une troisième partie, j'ai intégré des données épigénomiques et transcriptomiques obtenues à différents temps afin d'étudier les dynamiques temporelles des états chromatiniens en réponse à un stimulus externe, la première exposition à la lumière des plantules suite à la germination. Ces travaux nous ont permis de montrer que la monoubiquitination de l'histone H2B participe à la modulation fine et sélective des changements rapides de l'expression de gènes.

L'ensemble du travail présenté contribue à une meilleure compréhension de l'organisation de la chromatine le long du génome des plantes et de la dynamique des états chromatiniens en réponse aux changements de l'environnement.

Abstract

Eukaryotic genomes are packed into the confines of the nucleus through a nucleoproteic structure called chromatin. Chromatin is a dynamic structure that can respond to developmental or environmental cues to regulate and orchestrate the functions of the genome. The fundamental unit of chromatin, the nucleosome, consists of a protein octamer, which contains two molecules of each of the core histone proteins (H2A, H2B, H3, H4), around which 147 bp of DNA is wrapped. The post-translational modifications (PTMs) of histones and methylation of the cytosine residues in DNA (DNA methylation) constitute primary epigenomic markers that dynamically alter the interaction of DNA with nucleosomes and participate in the regulation and control access to the underlying DNA. The main objective of my thesis was to understand the spatial and temporal dynamics of chromatin states in Arabidopsis by investigating on a genome-wide scale, patterns of DNA methylation and a set of well-characterized histone post-translational modifications.

DNA methylation, a hallmark of epigenetic inactivation and heterochromatin in both plants and mammals, is largely confined to transposable elements and other repeat sequences. I show in this thesis that in Arabidopsis, methylated TE sequences having no or few matching siRNAs, and therefore unlikely to be targeted by the RNA-directed DNA methylation (RdDM) machinery, acquire DNA methylation through spreading from adjacent siRNA-targeted regions. Further, I propose that this spreading of DNA methylation through promoter regions can explain, at least in part, the negative impact of siRNA-targeted TE sequences on neighbouring gene expression. In a second part, I have contributed to integrative analysis of DNA methylation and eleven histone PTMs. I have shown through combinatorial and cluster analysis that the Arabidopsis epigenome shows simple principles of organisation and can be distinguished into four primary types of chromatin that preferentially index active genes, repressed genes, TEs, and intergenic regions. Finally, in a third part, I integrated epigenomics with transcriptome data at three different time points in a developmental window to investigate the temporal dynamics of chromatin states in response to an external stimulus. This used the light-induced transcriptional response as a paradigm to assess the impact of histone H2B monoubiquitination (H2Bub), and showed that this PTM is associated with active transcription and implicated in the selective fine-tuning of gene expression.

Taken together, the work presented here contributes significantly to our understanding of the spatial organisation of chromatin states in plants, its dynamic nature and how it can contribute to allow plants to respond to a signal from the environment.