Institut für Nutzpflanzenwissenschaften und Ressourcenschutz

Identification and computational analysis of differential H3K27me3 targets between *Arabidopsis thaliana* accessions

Inaugural-Dissertation
zur

Erlangung des Grades

Doktorin der Agrarwissenschaften

der

Landwirtschaftlichen Fakultät

der

Rheinischen Friedrich-Wilhelms-Universität Bonn

von

Xue Dong

aus

Pingyi, China

Referent: Prof. Dr. Heiko Schoof
Korreferent: Prof. Dr. Frank Hochholdinger
Tag der mündlichen Prüfung: 29.10.2012
Erscheinungsjahr: 2013

Abstract

Histone H3 lysine 27 trimethylation (H3K27me3) and lysine 9 dimethylation (H3K9me2) are two independent repressive chromatin modifications in *Arabidopsis thaliana*. H3K27me3 is established and maintained by Polycomb repressive complexes whereas H3K9me2 is catalyzed by histone methyltransferases SUVH(4-6). H3K27me3 mostly targets at protein coding genes in euchromatin which are reversible in repression. H3K9me2 mainly targets at transposons and repetitive sequences which should be constitutively silenced. Both marks can spread to flanking regions after initialization and they have been shown to be mutually exclusive in distribution in the *Arabidopsis* genome.

In this study, the extent of natural variation of H3K27me3 in the two accessions of *Arabidopsis thaliana*, Landsberg *erecta* (Ler) and Columbia (Col), and their hybrids was analyzed using chromatin immunoprecipitation followed by microarray or sequencing analysis (ChIP-chip and ChIP-seq). A computational workflow was implemented that includes remapping of probes to the Col and Ler genome assemblies in order to exclude differential signals due to genome polymorphisms.

The majority of genes that are H3K27me3 targets in Col are also targets in Ler and the F1 of reciprocal crosses. A small number of Ler-specific H3K27me3 targets were detected and well validated with independent ChIP-PCR whereas the Col-specific targets have not been confirmed so far. Ler-specific H3K27me3 targets showed an allele-specific H3K27me3 in both hybrids, consistent with a *cis*-regulatory mechanism for establishing H3K27me3.

Five Ler-specific H3K27me3 targets were marked by H3K4me3 in Col. Consistent with the activation role of H3K4me3 during transcription, the differential H3K27me3 of the these five genes accords to the expression variation between the two accessions. For the majority of Ler-specific H3K27me3 targets, no expression could be detected in Col, Ler or 17 other *Arabidopsis* accessions. Instead of H3K27me3, the antagonistic mark H3K9me2 and other heterochromatic features were observed at these loci in Col. More frequently than expected, transposable elements were found neighboring these loci in Col, and in many cases these transposable elements are missing in the Ler genome assembly. We propose a model where a transposon insertion specific to Col results in recruitment of H3K9me2, which spreads to neighboring genes already in a repressed state through H3K27me3, resulting in Ler-specific H3K27me3 as the ancestral state.

Zusammenfassung

Die Histon H3 Lysin 27 tri-Methylierung (H3K27me3) und Histon H3 Lysin 9 di-Methylierung (H3K9me2) sind zwei unabhängige Transkriptions-hemmende Chromatin Modifizierungen in *Arabidopsis thaliana*. H3K27me3 wird durch Polycomb Proteine etabliert und stabilisiert während H3K9me2 durch die Histon methyltransferases SUVH (4-6) katalysiert wird. H3K27me3 findet sich meist im Euchromatin Protein-codierende Gene, deren Repression reversibel ist. H3K9me2 hingegen findet sich zumeist in Transposons und repetitiven Sequenzelementen deren Transkription dauerhaft verhindert werden soll. Beide Modifizierungen können sich auf angrenzende Regionen ausdehnen und anhand ihrer Verteilung im Genom von *Arabidopsis* ist erkennbar, dass sie sich gegenseitig ausschließen.

In dieser Arbeit wurde die natürliche Variation von H3K27me3 in den zwei *Arabidopsis* Linien Landsberg erecta (Ler) und Columbia (Col) sowie ihren Hybriden mittels Chromatin Immunpräzipitation und anschließender Microarray- oder Sequenz-Analyse (ChIP-chip bzw. ChIP-seq) untersucht. Dazu wurde ein Computer-gestützter Arbeitsablauf entwickelt, der auch ein erneutes Mappen der DNA-Sonden gegen assemblierte Genom-Sequenzen von Col und Ler beinhaltet, um Signalunterschiede aufgrund von Polymorphismen auszuschließen.

Die Mehrzahl der Gene mit H3K27me3 in Col weisen diese Methylierung auch in Ler und den F1 Populationen aus wechselseitigen Kreuzungen auf. Es wurden nur wenige Gene mit Ler-spezifische H3K27me3 gefunden, die durch unabhängige ChIP-PCR gut validiert werden konnten. Col-spezifische Ziel-Gene konnten hingegen nicht bestätigt werden. Die Ler-spezifischen Ziel-Gene weisen in beiden Hybriden Allel-spezifische H3K27me3 auf, was die Annahme eines cis-regulatorischen Mechanismus für die Etablierung der Methylierung nahe legt.

Fünf der Ler-spezifischen H3K27me3 Ziel-Gene wiesen in Col eine H3K4me3 Modifizierung auf. In Übereinstimmung mit der aktivierenden Rolle der H3K4me3 während der Transkription, sind die Unterschiede in der H3K27me3 dieser fünf Gene zwischen den *Arabidopsis* Linien mit Unterschieden in der Gen-Expression korreliert. Für die Mehrzahl der Ler-spezifischen H3K27me3 Ziel-Gene kann weder in Col noch in Ler oder 17 anderen *Arabidopsis* Linien messbare Expression festgestellt werden. Anstelle

der H3K27m3 fanden sich an diesen Loci in Col die inhibitorische H3K9me3 und andere inhibitorische Modifikationen wie z.B. DNA-Methylierung oder H3K27me1. Häufiger als erwartet wurden in Col in der unmittelbaren Nähe dieser Loci Transposons gefunden die in vielen Fällen in der Ler Genom-Sequenz fehlen. Wir schlagen deshalb folgendes Modell vor: Die Insertion eines Transposons ausschließlich im Col Genom führt zur Rekrutierung der H3K9me2 und der anschließenden Ausbreitung dieser Modifizierung auf benachbarte Gene, deren Transkription zuvor bereits durch H3K27me3 unterdrückt war. Unserem Modell zufolge stellt die Ler-spezifische H3K27me3 also den ursprünglichen Zustand dar.

Table of Contents

1 Introduction

1.1 Introduction to chromatin features

In all eukaryotes, genomic DNA is tightly compacted into a complex structure that is known as chromatin (Ridgway 2001). The structure of chromatin undergoes dynamic changes at different phases of the cell cycle and in response to various intracellular and extracellular signals. In turn, the structure of chromatin can influence the accessibility of DNA regions and then further affect the recruitment of regulatory proteins to their target sites. Consequently, the regulation of nuclear processes such as transcription, replication and DNA repair (Misteli 2007) can be influenced. The basic unit of chromatin is the nucleosome that consists of a core nucleosome particle and a linker region that connects adjacent core histones. The core nucleosome particle is composed of about 147 base pairs of DNA wrapped around a histone octamer, which contains two subunits of each of the four core histone proteins H2A, H2B, H3, and H4. The N-terminal ends of histones are rich in basic amino acids that are flexible and subject to various post-translational modifications. Historically, based on microscopic observations, chromatin has been classified into two distinct functional forms: heterochromatin and euchromatin. Heterochromatin has been cytologically defined as deeply stained chromosomal regions that are highly condensed throughout cell cycle, whereas euchromatin is decondensed during interphase (Heitz, E. 1928; Harničarová Hor&aacut et al. 2010). Heterochromatin is rich in transposable elements and generally silenced during transcription but plays an important role in the organization and proper functioning of genomes (Pimpinelli 1995; Lippman et al. 2004). Euchromatin contains nearly all of the genes that are vital for cellular processes and often under active transcription. Heterochromatin and euchromatin occupy distinct genomic compartments and are associated with distinct chromatin signatures (Spada et al. 2005).

Heterochromatin can be further classified into two subtypes, constitutive and facultative heterochromatin. Constitutive heterochromatin includes the regions that contain a high density of repetitive DNA elements, such as centromeres, peri-centromeric regions and telomeres (Grewal & Moazed 2003). The genes in constitutive heterochromatin are generally transcriptional inactive and distributed in a low density. The active genes that

are inserted into the vicinity of constitutive heterochromatin by chromosomal rearrangement or insertion are often silenced (Brown et al. 1997; Bártová et al. 2002). Constitutive heterochromatin is an abundant component of eukaryotic genomes, forming about 5% of the genome in *Arabidopsis*, 30% in *Drosophila* and humans, and up to 70–90% in certain nematodes and plants (Moritz & Roth n.d.; Rossi et al. 2007; Anon 2000; Patrizio Dimitri, Nicoletta Corradini, et al. 2005; P Dimitri, N Corradini, et al. 2005; Patrizio Dimitri et al. 2009), The amount of heterochromatin was correlated to presence of TE and related sequences. Constitutive heterochromatin is marked by repressive chromatin modifications such as DNA methylation and trimethylated histone H3 Lys 9 (H3K9me3) in mammals, whereas in *Arabidopsis* it is mainly marked by H3K9me2 (Bernatavichute et al. 2008).

The facultative heterochromatin, on the other hand, is formed by inactivation of genes that was active during development, cellular differentiation and responses to external stimuli. One of the best-studied examples is the silencing of one of the two X chromosomes in female cells of mammals to equalize the dosage of X-linked gene expression with males (Brockdorff 2011). Facultative heterochromatin is associated with distinct histone modifications compared to constitutive heterochromatin. In mammals and plants, H3K27me3 is a common facultative heterochromatin mark that mediates reversible repression of euchromatic genes.

A variety of epigenetic events can occur on the chromatin level, for example DNA methylation, incorporation of histone variants, histone modifications (see **Figure 1**) and chromatin association of small interfering and long non-coding RNA. It is well accepted that different combinations of chromatin modifications exist and define distinct chromatin domains that correlate with certain transcriptional output or structural function (Margueron & Reinberg 2010). In the following sections, I will briefly introduce several well-studied chromatin marks with an emphasis on their differences between species or conditions. The characteristics of the two of histone modifications most relevant within this project, H3K27me3 and H3K9me2, will be further described in the sections 1.2 and 1.3.

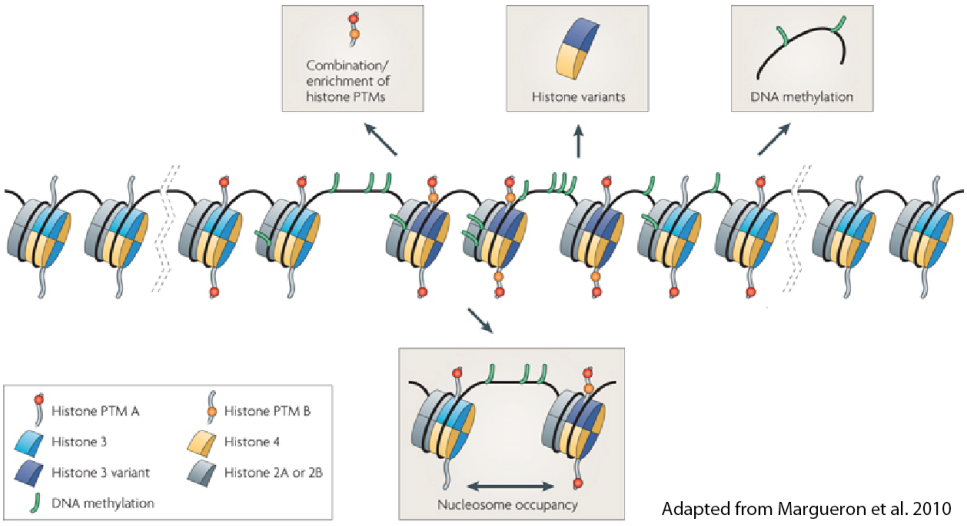


Figure 1. Scheme demonstrating modifications that define distinct chromatin domains. The types of modifications are shown in the grey boxes. The dashed lines indicate the break of two adjacent chromatin domains. PTM, post-translational modification. (Margueron & Reinberg 2010). Figure was modified after Margueron et al. (2010).

1.1.1 Introduction to DNA Methylation

DNA in nucleosomes can be covalently modified by the addition of methyl group at position 5 of cytosine. Cytosine DNA methylation is a conserved epigenetic silencing mechanism that participates in many diverse biological processes, such as the silencing of transposons and the regulation of gene expression (Birdg 2002; Xiaoyu Zhang et al. 2006). In mammals, DNA methylation occurs primarily in the CG context. About 70–80% of CG dinucleotides in the human genome are methylated (Jabbari & Bernardi 2004). But generally the active expression of genes are associated with unmethylated CpG sites in the CpG islands of promoters and enhancer (Rollins et al. 2006). Recently, it was shown that DNA methylation patterns are often conserved between humans and chimpanzees and the inter-species differences in promoter methylation could partially explain the differences in gene expression levels (Pai et al. 2011).

In plants, an increasing number of genome-wide DNA methylation studies have been conducted in *Arabidopsis* and rice (Xiaoyu Zhang et al. 2006; Zilberman et al. 2007; Cokus et al. 2008; Vaughn et al. 2007; Yan et al. 2010). DNA methylation can occur in all three cytosine contexts CG, CHG and CHH, where H can be A, C, or T. The occurrence of DNA methylation is predominantly on peri-centromeric heterochromatin, repetitive sequences, and regions producing small interfering RNAs (Xiaoyu Zhang et al.

2006; Cokus et al. 2008). Interestingly, genes that are methylated only in transcribed regions are highly expressed, whereas genes methylated in the promoter region show a negative correlation between methylation with gene expression and greater tissue specific differences in gene expression (Cokus et al. 2008). The comparison of DNA methylation pattern between the *Arabidopsis* accessions Columbia (Col) and Landsberg (L*er*) showed that genic methylation was highly polymorphic across ecotypes and is heritable, but is lost at a high frequency in segregating F_2 families (Vaughn et al. 2007). Using next-generation sequencing technique, recently it was shown that the F1 hybrids of Ler and C24 have increased DNA methylation across their entire genome relative to their patents and small RNAs might play a role in the differential methylation (H. Shen et al. 2012).

Plants utilize a pathway named RNA-directed DNA methylation (RdDM) to establish sequence-specific DNA methylation (Wassenegger et al. 1994; Chinnusamy & J.-K. Zhu 2009). Single-stranded RNA transcripts are first transcribed from transposons and repeat elements by plant-specific RNA polymerase IV, and then used as template to generate dsRNA by RNA-DEPENDENT RNA POLYMERASE 2 (RdRP2) (Law & Jacobsen 2010; Xie et al. 2004). DICER-LIKE 3 (DCL3) is thought to process the dsRNAs into 24-nucleotide (nt) small interfering RNAs (siRNAs) (Pontes et al. 2006). 24-nt siRNA can guide DNA methyltransferases to the siRNA-generating genomic loci and other loci that are homologous to the siRNAs for *de novo* DNA methylation.

Generally, DNA methylation is a repressive chromatin mark and widely distributed in pericentromeric heterochromatin, repetitive sequences and euchromatic genes. It acts together with other chromatin marks such as H3K9me2 to recruit factors that condense chromatin and then further confer the repression of their targets.

1.1.2 Introduction to histone variants

The *Arabidopsis* genome contains multiple genes encoding histone H1, H2A/B, H3 and H4(Y. Zhu et al. 2011). Most of the histone-encoding genes are for conventional histones (major subtypes of histones), which are highly conserved between species and can be produced and incorporated into nucleosomes during DNA replication (Y. Zhu et al. 2011; Kamakaka & Biggins 2005). Some other genes encode histone variants, which have different amino acid sequences mostly in the N-terminal region. The histone variants are subject to specific posttranslational modifications and then confer unique properties to the nucleosomes they are integrated in (Y. Zhu et al. 2011). The histone variants can replace

the conventional histones during the entire cell cycle without DNA replication (Kamakaka & Biggins 2005). The histone variants can be found at specific genomic regions and play an important role in various biological processes by regulating the stability and structure of nucleosomes in which they are embedded (Kamakaka & Biggins 2005; Y. Zhu et al. 2011). For example, H2A.X is a highly conserved variant of H2A and can be preferentially phosphorylated. The function it involved include DNA repair, recombination and transcription repression (Kamakaka & Biggins 2005). Another extensively studied H2A histone variance is H2A.Z. It has been implicated in multiple roles in divergent organisms. It was proposed that the incorporation of H2A.Z can promote histone turnover and chromatin accessibility (Mavrich et al. 2008). In eukaryotes, H2A.Z localizes to the strongly positioned +1 nucleosome downstream from the TSS and a few positioned nucleosomes further downstream, but is relatively depleted over gene bodies (Weber et al. 2010) (Zilberman et al. 2008). In mammals, H2A.Z is also enriched upstream of TSSs. It was also shown that H2A.Z had a coincident enrichment pattern with Polycomb group proteins Suz 12 in ES cells at the promoters of genes important for development and was necessary for ES cell differentiation (Creyghton et al. 2008). H2A.Z also can be a boundary element to protect euchromatin from spreading of heterochromatin (Meneghini et al. 2003). In *Arabidopsis*, the nucleosomes containing the histone H2A.Z variants are essential to perceive the ambient temperature correctly (Kumar & Wigge 2010). A recent study also suggested that H2A.Z is mutually antagonistic with DNA methylation and may protect genes from silencing by DNA methylation (Zilberman et al. 2008). Additionally, it was also shown in *Arabidopsis* that the complex required to H2A.Z generation is necessary for the high-level expression of FLC (Deal et al. 2007).

Histone variants, together with other chromatin marks, contribute to chromatin dynamics and the formation of distinct chromatin states.

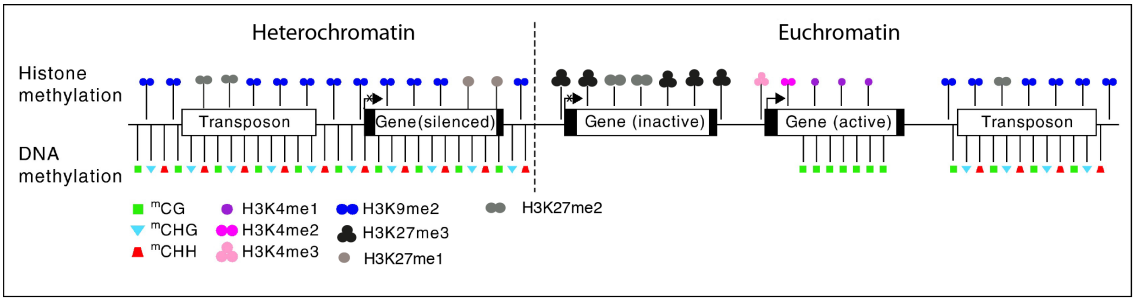
1.1.3 Introduction to histone modifications

The post-translational modification of histone proteins has been a topic of intense interest over the last decade as these marks play an important role in regulating chromatin states, thus influencing all chromatin dependent processes, such as maintenance and regulation of the dynamic chromatin structure, gene activation, DNA repair and many others. Histone modifications occur primarily on the N-terminal tails of H3 and H4 and include

acetylation, methylation, ubiquitination, phosphorylation, glycosylation, ADP ribosylation, and sumoylation (Kouzarides 2007). Particularly, histone lysine residues can be modified by addition of one, two, or three methyl groups. The methylation on the lysine 4, 9, and 27 positions of histone 3 is among the best-studied chromatin marks in both plants and animals. Lysine methyltransferases that contain the evolutionarily conserved SET domain is responsible for lysine methylation. In plant, the preferential distribution of these marks and the correlation of these marks with the expression of their target in various genomes have been explored (Turck et al. 2007; Xiaoyu Zhang et al. 2009; Bernatavichute et al. 2008; G. He et al. 2010) (see Figure 2). Distinct histone modifications are associated with euchromatin and heterochromatin. Chromatin acetylation and H3K4 methylation are associated with euchromatin, whereas DNA methylation, H3K27me1 and H3K9me1/2 are more likely in constitutively silenced heterochromatin in *Arabidopsis*. Strikingly, H3K27me3 and H3K9me3 are localized in euchromatin but not in heterochromatin as in *Drosophila* or mammal (Figure 2).

Histone modifications can recruitment other enzymes to change the chromatin structure to become more open or more closed, which is associated with activation or repression of target genes. For example, H3K4me3 is a mark of FACT-complex activity, which is required for transcription initiation (Iii et al. 2008). H3K36me3 is part of complexes that promote transcription elongation (Krogan et al. 2003; Lee & Shilatifard 2007). In general, H3K4 and H3K36 methylation are correlated with activation of gene transcription and called active marks. On the other hand, methylation at other lysine site such as H3K9, H3K27 and H3K20 methylation are correlated with gene repression and called repressive marks. Repressive mark H3K27me3 and H3K9me2 can spread to flanking regions from their primary targeting sites until it is blocked by chromatin boundary elements (Talbert & S. Henikoff 2006). Moreover, histone modifications are dynamic during differentiation or environmental changes and can be added or removed by respective enzymes (Zheng et al. 2010; van Dijk et al. 2010; Charron et al. 2009).

There is intensive crosstalk between the different histone modifications. The modification of one histone can influence (facilitate or inhibit) the modification on another histone protein (Luo & Lam 2010). For example, mono-ubiquitination of histone H2B at lysine 123 (K123) is required for the establishment of H3K4 and H3K79 methylation (Zheng et al. 2010; Sun & C David Allis 2002). But H3K4me3 and H3K36me2/3 marks can inhibite the establishment of H3K27me3 (Schmitges et al. 2011).



Modified from Current Opinion in Plant Biology

Figure 2. Schematic representation of the distribution of selected epigenetic marks in the *Arabidopsis* genome
H3K4 can be mono-, di-, and tri-methylated. H3K4 di- and tri-methylation peaks at the promoter and 5'genic regions, whereas H3K4me1 is localized at transcribed regions (Xiaoyu Zhang et al. 2009). The targets of H3K4 methylation are located in euchromatin. H3K4me3 target genes are highly expressed, but H3K4me2 and H3K4 me1 may not associate with gene activation directly (Xiaoyu Zhang et al. 2009; S. L. Berger 2007). H3K27 also can be mono-, di-, and trimethylated. H3K27me3 is localized at transcribed regions and H3K27me2 also spreads to flanking regions, while H3K27me peaks in the middle of genes (Roudier et al. 2011). H3K27 methylated genes are repressed or express in a tissues-specific manner (Xiaoyu Zhang et al. 2007; Roudier et al. 2011). H3K27me3 mark is prevalent in euchromatin, but H3K27me1 occurs mainly in heterochromatin and colocalizes with H3K9me2 and DNA methylation (Roudier et al. 2011). H3K27me2 is enriched along H3K27me3-marked genes and TE sequences (Roudier et al. 2011). The main targets of H3K27me/H3K9me2 are TEs and repetitive regions that are supposed to be silenced stably over the whole life of plants, while H3K27me2/3 target many tissue specific genes the expression of which needs to be dynamically regulated throughout the development of plants. These modifications at H3K27 occur at diverse target sites and therefore have different impact on regulating gene transcription.

Different histone modifications are thought to act sequentially or in a combinatorial way to regulate the expression of their targets genes indirectly or as part of transcription machinery for example H3K36me3 (Jenuwein & C D Allis 2001; S. L. Berger 2007; Roudier et al. 2011). Despite the association of histone modifications with gene activation, the transcriptional output of a gene also depends on the context and timing by which these modifications are introduced (Lee et al. 2010). Particularly, current genomic profiles of histone modifications in plants are based on a mixture of cell types, and it is not sufficient to determine the gene expression status by just looking at the pattern of chromatin modifications at a locus (Lee et al. 2010).

1.1.3 Classification of chromatin states

Recently, a large amount of genome-wide chromatin modification data has been generated and computationally analyzed. The chromatin states are the combinatorial patterns of these modifications. The genome can be partitioned into distinct chromatin

states according to genomic maps of these chromatin modifications. These chromatin states are preferentially associated with distinct gene annotations and thought to participate in selective regulation of transcriptional outcomes of the genomic sequences (Roudier et al. 2011). For example, genes actively transcribed are marked by H3K4me3 at their promoter and H3K36me3 along the transcribed region (Roudier et al. 2011). Despite the large number of theoretically possible combinations of the chromatin marks, only a relatively small number of these combinations has been shown to actually be present in different organisms (Mikkelsen et al. 2010; Roudier et al. 2011).

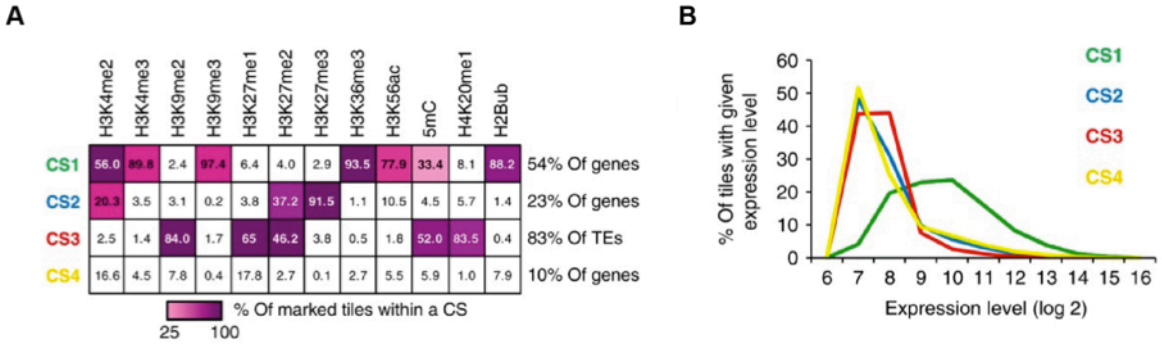


Figure 3. The association of chromatin states with various chromatin modifications and target genes expression.
(A) The table indicates the association (numbers inside cells) of four predominant chromatin states (CS) with each of the 12 of chromatin marks. The percentage of genes indexed by CS1, CS2 and CS4 and the percentage of TE annotations indexed by CS3 are also shown. CS1 and CS2 are both repressive states but associated with different categories of annotations. (B) Relationship between chromatin states and gene expression level. (See Roudier et al. 2011)

In particular, based on an integrative analysis of 12 chromatin marks, the *Arabidopsis* epigenome was shown to be organized around four predominant chromatin states (CS1, CS2, CS3 and CS4) (Roudier et al. 2011) (see **Figure 3**). CS1 and CS2 are antagonistically associated mainly with genes that are being either active or repressed/lowly expressed, respectively. CS3 is associated primarily with TE sequences while CS4 is depleted in any chromatin mark. The four CS correspond to distinct biochemical, transcriptional and sequence properties (Roudier et al. 2011). In mammalian cells, the correlation of chromatin states with gene expression and regulatory activity has been used to identify regulatory elements and cell-type-specific activators and repressors (Ernst et al. 2011).

1.1.4 Chromatin boundary elements

Chromatin boundary elements are DNA elements that protect active genes and regulatory regions from the spread of the adjacent heterochromatin formation or the effect mediated by repressive chromatin, for example H3K9me2 (Felsenfeld et al. 2004). The boundary elements can be recognized by boundary element binding proteins that can further recruit enzymes for chromatin modification. For example, CTCF (CCCTC-binding factor) is a well-known insulator-binding protein in vertebrates. It binds not only to many enhancer-blocking elements but also to the boundaries of repressive chromatin domains marked by H3K27me3 in mammalian cells (Cuddapah et al. 2007).

The formation of repressive chromatin marks must be blocked before it invades into important functional genes that need to be active. In *Drosophila*, Polycomb repressive complex 2 (PRC2) recognizes the Polycomb response elements and then catalyzes the formation of the repressive histone mark H3K27me3 in nearby nucleosomes. PRC2 complexes continually catalyze the formation of H3K27me3 (Simon & Kingston 2009). Once the reaction is started, the formation of heterochromatin will propagate until it meets some barrier. The barrier to the spread of heterochromatin could be some epigenomic landmark, such as an activating promoter, a strong enhancer or some specialized DNA element (Lin et al. 2011). Moreover, it was shown *in vitro* that the establishment of the repressive mark H3K27me3 by PRC2 is inhibited by H3K4me3 and H3K36me2/3 marks and the actively transcribed genes marked by H3K4me3 and H3K36me2/3 can serve as chromatin boundary that blocks the deposition of H3K27me3 (Schmitges et al. 2011). It was also proposed that Histone H2B ubiquitination is employed by chromatin boundaries to restrict the encroachment of repressive chromatin marks. In a study by Ma et al, after the depletion of ubiquitinated H2B, a striking collapse of the active histone modification signature including H3K4me3 at chromatin boundaries was observed (Ma et al. 2011). Furthermore, the repressive chromatin marks including H3K9me2 spread over the entire neighboring genes leading to the silencing of these genes (Ma et al. 2011). Since H2B ubiquitination is required for the establishment of H3K4me3 (Sun & C David Allis 2002), probably H3K4me3 is the signature that really stops the invasion of repressive marks.

1.2 Characteristics of H3K27me3

1.2.1 H3K27me3 mediated by Polycomb Group proteins

H3K27me3 is an abundant and repressive histone modification. It is catalyzed and maintained by Polycomb group proteins (PcG). PcG proteins were first discovered in *Drosophila* due to their crucial role in the regulation of *Drosophila* homeotic (HOX) genes. PcG proteins function in protein complexes, of which PRC1 and PRC2 are best characterized. PRC1 and PRC2 work sequentially and coordinately to tri-methylate the target genes and repress their expression (Pien & Grossniklaus 2007). After the establishment of H3K27me3 by PRC2, the H3K27me3 is recognized by PRC1, which binds to H3K27me3 to stably maintain the repression. Although the molecular mechanism leading to gene repression via H3K27me3 is not yet completely known, it was revealed in mammals that PRC1 can catalyze mono-ubiquitylation of lysine 119 at H2A, which leads to chromatin compaction and correlates with inhibition of transcription initiation (H. Wang & L. Wang 2004; Morey & Helin 2010).

PRC2 is a 600-kDa complex and composed of four core subunits that are conserved between Drosophila, mammals and plants. In Drosophila, The four core subunits are the SET domain protein Enhancer of Zeste (E(Z)), Suppressor of Zeste 12 (Su(Z)12), Extra Sex Combs (ESC) and Multicopy Suppressor of IRA (MSI). In *Arabidopsis*, each component of the PRC2 is encoded by small gene families, except for the single homolog of ESC, FERTILIZATION-INDEPENDENT ENDOSPERM (FIE) (Köhler & Villar 2008)(Morey & Helin 2010). E(Z) contains a SET domain, which is responsible for methylation activity of PRC2 complex. In *Arabidopsis*, E(Z) has three homologs including MEDEA (MEA), CURLY LEAF (CLF), and SWINGER (SWN); Su(z)12 also has three homologs FERTILIZATION- INDEPENDENT SEED2 (FIS2), EMBRYONIC FLOWER2 (EMF2), VERNALIZATION2 (VRN2); MSI has five homologs (MSI1-5) although only MSI1 and 4 are really connected to PcG so far (see Table1). Based on molecular and genetic evidence, it is thought that at least three PRC2 complexes exist in plants, FIS, VRN, and EMF complex, which differ in their subunit composition but regulate a largely overlapping set of genes and have their specific targets (Pien & Grossniklaus 2007).

Table 1. The components and complexes of PRC2 in *Arabidopsis*

PRC2 components	Domains	Homologues in Arabidopsis	EMF complex	FIS complex	VRN complex
E(Z)	SET	MEA, CLF, SWN	CLF/SWN	MEA/SWN	CLF/SMN
Su(Z)12	VEFS box, Zn-finger	FIS2, EMF2, VRN2	EMF2	FIS2	VRN2
ESC	WD40	FIE	FIE	FIE	FIE
MSI	WD40	MSI1-5	MSI1	MSI1	MSI1

In Drosophila, PRC1 complex consists of PC (polycomb), PH (polyhomeotic), PSC (posterior sex-comb) and RING proteins. For each subunit of PRC1 in Drosophila , there are several homologues in mouse and human, but all members contain a conserved domain (Morey & Helin 2010). However, the existence of PRC1 in *Arabidopsis* has been questioned since there is no clear genetic homologues of the core components of PRC1. But in *Arabidopsis*, a functional homologue PRC1-like complex LIKE HETEROCHROMATIN PROTEIN 1 (LHP1) was identified. LHP1 is a functional homologue of PC and it colocalizes with H3K27me3 along the genome of *Arabidopsis* (Xiaoyu Zhang et al. 2007; Turck et al. 2007). In animals, PRC1 bind to H3K27me3 and is required to stabilize the transcriptional repression. Whereas in *Arabidopsis*, mutant of LHP1 only have earlier flowing, curly leaf phenotype and do not disrupt the distribution of H3K27me3, indicating that LHP1 is not required for the maintenance of H3K27me3 (Turck et al. 2007). Recently, more likely PRC1 components have been identified. it was shown that two closest homologues to SCE in *Arabidopsis*, AtRING1A and AtRING1B can bind LHP1 *in vitro* (Sanchez-Pulido et al. 2008). And the homologues of PSC, AtBMI1A and AtBMI1B, also interact with LHP1 *in vitro*. More importantly,AtBMI1A/B proteins was proved to mediate H2A monoubiquitination in *Arabidopsis*, showing a similar function as PRC1 in animals (Bratzel et al. 2010). EMBRYONIC FLOWER 1 (EMF1) is also a candidate PRC1 component in *Arabidopsis*. It was shown to interact with MSI1, an *Arabidopsis* component of PRC2, as well as AtRING1A, AtRING1B, AtBMI1A and AtBMI1B *in vitro* (Calonje et al. 2008) (Bratzel et al. 2010). The PRC1-like component identified in *Arabidopsis* is growing but not complete yet.

1.2.2 The recruitment of PRC2

PRC2 regulate thousand of target genes by deposition of H3K27me3. It is essential to know how they are recruited to their targets. Currently, multiple mechanisms are likely

involved in this process in different organisms. In *Drosophila*, PRC2 is recruited to nucleosome-depleted regions of the genome called Polycomb response elements (PREs). PREs contain short DNA motifs that can be recognized by DNA binding proteins, which in turn recruit PRC2. However, the DNA sequences of PREs are not conserved in other species.

In mammals, it was reported that GC-rich elements depleted of activating transcription factor motifs can mediate PRC2 recruitment (Mendenhall et al. 2010). *HOTAIR*, a long intergenic noncoding RNA, transcribed from the *HOXC* locus, represses transcription *in trans* across 40 kilobases of the *HOXD* locus. It was shown that *HOTAIR* can interact with PRC2 and is required for histone H3 lysine-27 trimethylation of *HOXD* locus (Rinn et al. 2007). Later, *HOTAIR* was found to serve as a scaffold to tether two distinct histone modification complexes, methylase PRC2 and demethylase LSD1 complex, thereby specifying the pattern of histone modifications on target genes (Tsai et al. 2010).

In *Arabidopsis*, a specific *cis*-activating element for *LEAFY COTYLEDON2* (*LEC2*) was identified and shown to be required and sufficient to trigger H3K27me3 deposition (N. Berger et al. 2011). At the *FLC* locus, *COLDAIR* is transcribed from an intron of *FLC* and required for deposition of H3K27me3 through its interaction with PRC2 (Heo & Sung 2011).

Recently, more and more evidence supports the notion, that, PRC2 is recruited to chromatin by multiple mechanisms, besides the locus specific *cis*-regulatory elements, long ncRNAs are important participants in PRC2 function.

1.2.3 H3K27me3 targets in *Arabidopsis*

Whole-genome H3K27me3 profiling of the *Arabidopsis* Col genome has been performed by different laboratories. H3K27me3 modifications in *Arabidopsis* are preferentially enriched in transcribed regions of genes and seem to be confined to their target genes instead of spreading over hundreds of kilobases and cover many genes like in animals (Turck et al. 2007; Xiaoyu Zhang et al. 2007). Around 4400-8000 H3K27me3 targets were identified by different studies (Turck et al. 2007; Xiaoyu Zhang et al. 2007; Lafos et al. 2011). Though the amount of identified H3K27me3 target genes is different, the overlap of targets between studies is very high in spite of differences in analysis methods and plant materials. H3K27me3 target genes are enriched in transcription factors,

indicating that this histone modification plays a widespread role in the regulation of plant development. In particular, most Flowering-related-genes are H3K27me3 targets, for example, *FT* and *FLC* have been used as model genes to study the mechanism of H3K27me3 recruitment and regulation.

Lafos et al. (Lafos et al. 2011) compared H3K27 targets between undifferentiated cells of the shoot apical meristem and in differentiated leaf cells of *Arabidopsis*. Hundreds of genes gained or lost H3K27me3 upon differentiation, indicating the regulation of H3K27me3 in plants is dynamic during development (Lafos et al. 2011). A highly overlapping H3K27me3 targets were identified in seedlings and endosperm, in which transposable elements are specifically targeted by H3K27me3 (Weinhofer et al. 2010). In C24 and cvi, only a small amount of differential H3K27me3 targets were identified when compared with Col (Moghaddam et al. 2011).

Collectively, the profile of H3K27me3 in different tissues and *Arabidopsis* accessions shows that PRC2-mediated regulation is a stable repressive system and show limited dynamics between these studies. It controls the phase transitions during development, such as from vegetative phase to flowering and from embryonic phase to the seedling (Bouyer et al. 2011). In PcG mutant *fie*, The profile of H3K27me3 changed dramatically but no global changes in H3K4me3 levels were observed, indicating a repression of H3K27me3 targets by other mechanisms (Bouyer et al. 2011).

1.3 Characteristics of H3K9me2

Like methylation at H3K27, the methylation of H3K9 is also associated with gene repression. Both in animals and plants, lysine 9 of histone H3 can be mono-, di-, or tri-methylated. Like DNA methylation and H3K27me, H3K9me2 is an epigenetic mark that mainly targets at transposons, repetitive DNA and other DNA elements that need to be constitutively silenced in *Arabidopsis*. A very high concurrence rate between H3K9m2 and CHG methylation (where H is either A, T or C) was observed throughout the Col genome (Bernatavichute et al. 2008). About 3000-4000 H3K9me2 target genes (including protein coding gene and transposable element gene) have been identified in the *Arabidopsis* genome (Bernatavichute et al. 2008; Rehrauer et al. 2010). About 90% of them were located in the centromeric and pericentromeric heterochromatin, while about 10% were found in the euchromatic chromosome arms (Rehrauer et al. 2010). The

H3K9me2 targets were expressed only very weakly or even not at all (Bernatavichute et al. 2008). So H3K29me2 in *Arabidopsis* is a hallmark of silenced genes.

It was proposed that the H3K9me2 and H3K27me1 modifications represent two pathways controlling constitutive heterochromatin formation in parallel in arabidopisis (C. Liu, Lu, et al. 2010). The H3K9me2 pathway is DNA methylation–dependent. Histone H3K9 methyltransferases bind to methylated CHG target regions. The SET domains of these enzymes methylate neighboring histones to form a self-reinforcing loop between H3K9me2 and CHG DNA methylation. However, the H3K27me1 is set by histone methyltransferases ATXR5 and ATXR6 in a DNA methylation–independent manner (Jacob et al. 2009). In mammals, H3K9me2 is very stable and only shows distinct local changes before or after cellular differentiation (Lienert et al. 2011). Both in *Arabidopsis* and animals, H3K9me2 was shown to be mutually exclusive with H3K27me3 (Turck et al. 2007; Lienert et al. 2011).

1.4 Introduction to ChIP-chip and ChIP-Seq

Currently, two approaches are widely used for genome-wide identification and characterization of protein-DNA interactions *in vivo*. They are Chromatin immunoprecipitation (ChIP) followed by genomic tiling microarray hybridization (ChIP-chip) or massively parallel sequencing (ChIP-Seq). Both are used to identify the targeting sites of protein of interest in vivo. The fist step of the two techniques is ChIP, which is to immunoprecipitate native chromatin or chromatin covalently linked with proteins of interest. In the following step, the DNA separated from immunoprecipitated chromatin either is hybridized to microarray (ChIP-chip) or sequenced directly by various sequencing platforms (ChIP-Seq) (see Figure 4).

ChIP-chip appeared earlier than Chip-Seq. It is one of the earliest approaches to profile the protein-DNA interactions genome-wide and has been used widely used in the past 10 years and led to many important discoveries related to various fields of biology (Ho et al. 2011). Now ChIP-Seq is an attractive alternative of ChIP-chip and has the potential to replace ChIP-chip (Park 2009). ChIP-Seq has several advantages over ChIP-chip. For example, ChIP-Seq data can reach base pair resolution whereas the resolution of ChIP-chip depends on the probe length. Further, ChIP-Seq data does not have the noise caused by cross hybridization issues as ChIP-chip data (Park 2009). The process of hybridizing

DNA with probes is complex and influenced by many factors such as sequence composition of target and probes and experimental conditions. Moreover, ChIP-Seq can be used to analyze any species with a reference genome but not restricted to the ones with a specific microarray available as ChIP-chip. In a recent study, the performance of the two technologies was carefully compared (Ho et al. 2011). They show that both are highly reproducible within each platform and ChIP-Seq data has better signal-to-noise ratio and better balance of sensitivity and specificity, but the results of the two platforms can be significantly different due to the analysis methods used.

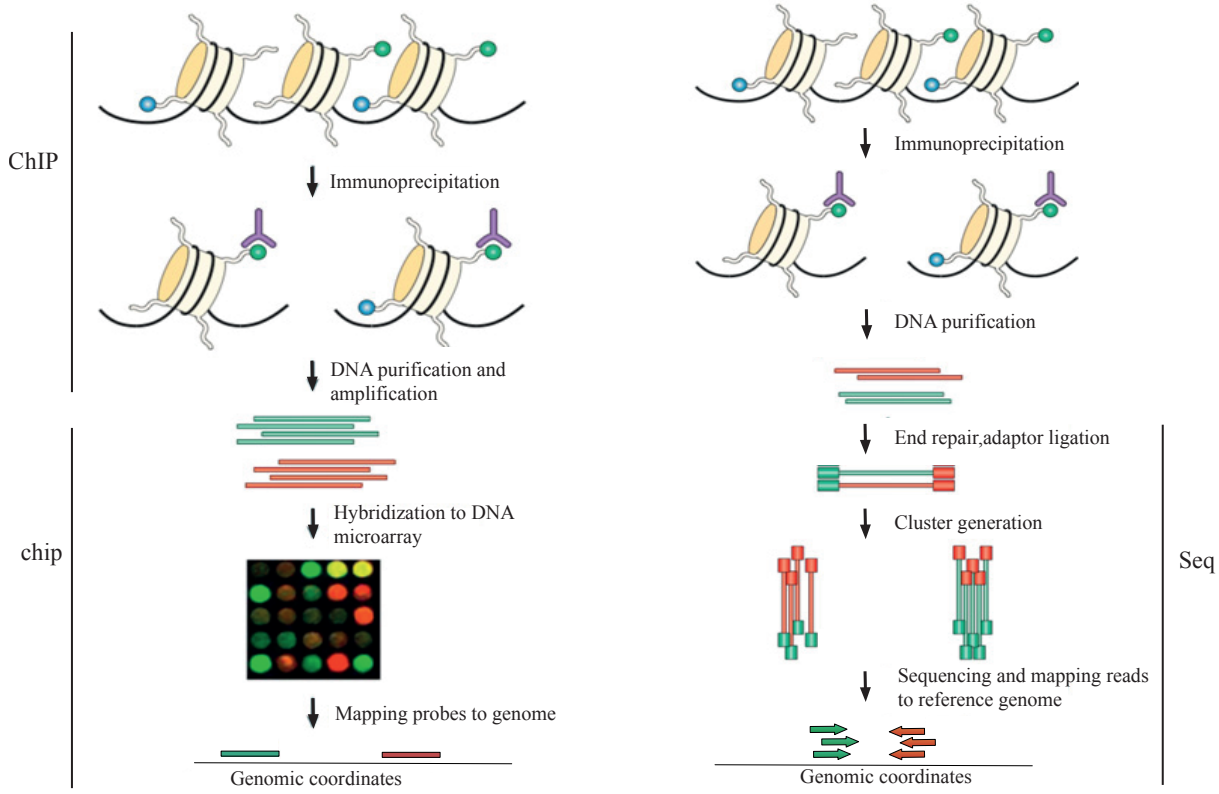


Figure 4**. Overview ofChIP-chip and ChIP-Seq**.
Using Chip-chip or ChIP-Seq, the specific targeting sites of histone modifications or chromatin-associated proteins can be identified in genome-wide scale. The process of ChIP is a common step in ChIP-chip and ChIP-Seq. It is to enrich the chromatin using an antibody specific to the histone modifications or protein associated with it. The DNA is firstly fragmented and then cross-linked with the protein that is associated with it. Using a specific antibody, only the chromatin with protein of interest is retained. The DNA in the antibody/histone/DNA complexes is purified, amplified and then hybridized on arrays or sequenced directly but different sequencing platform. Illumina use the technique of sequencing by synthesis to the cluster of clonal sequences that are generated by “bridged" amplification reaction. The reads after sequencing are mapped to reference genome to detect the targeting site of the protein to which the antibody bound. The colors for DNA sequence (DNA fragments, probes, reads) just indicate different DNA fragments. Figure is from Schones and Zhao (Schones & Zhao 2008).

It is worthy to note the process of CHIP-Seq require bulk processing of DNA fragments and massively parallel sequencing. A slightest bias in the ligation of linkers, in PCR amplification, or in hybridization might result in some platform-dependent biases (E. T. Liu, Pott, et al. 2010). So it is very necessary to generate a control library to adjust the bias generated in data preparation. No-antibody control can only produce little amount of DNA samples and is easy to be influenced by experimental conditions and subsequent global amplification. It therefore is not recommended to be used as background for normalization. The non-ChIP genomic DNA is normally sequenced and used as INPUT data for normalization. Ho et al also show that the quality of input DNA data, in terms of sequencing depth and GC content, is crucial for normalization in ChIP-Seq data analysis (Ho et al. 2011).

Considering the merits of ChIP-Seq and the cost of high-throughput sequencing technology continues to drop, ChIP-Seq will become more and more widely used.

1.5 Introduction to methodology in data analysis

The goal of ChIP-chip data analysis is to identify the genes or genomic regions that are associated with a protein of interested or a certain epigenetic mark. In a two-color ChIP-chip experiment, DNA fragments immunoprecipitated with a specific antibody (IP) and genomic DNA (INPUT) are differentially labeled and co-hybridized on the same array. For data analysis, the crucial step is to distinguish background signal from signal specific to the IP, thus reflecting the enrichment. The log ratio of the signal intensity of the two channels is often used for further data analysis.

Tiling arrays have been widely used for ChIP-chip. In tiling arrays, probes are positioned along the genome with a certain spacing. Not one probe but multiple probes are designed to cover each gene. Because of the high density of probes on a tiling array, if one probe is enriched, the adjacent probes probably also have high signal resulting in a ‘peak’ of intensity. Such spatial distribution of ChIP-chip data has been used by many peak detection methods to efficiently identify ChIP enriched regions. For example, peak detection methods by using sliding windows (Keleş 2007) or Hidden Markov Models (Ji & Wong 2005; W. Li et al. 2005).

Mixture model approaches are widely used to distinguish non-enriched probes from enriched probes **(**Sun et al. 2009**)**(Keleş 2007; Martin-Magniette et al. 2008; Toedling et

al. 2007). In a mixture model, the whole population of probes was dissected into a mixture of two distributions according to their signal: the distribution of IP-enriched genomic fragments, and the distribution of genomic DNA (INPUT). Depending on the abundance of probes from the ChIP-enriched regions, the two distribution can be bimodal or seem to be one distribution but with a heavy tail. Different statistical methods have been proposed to distinguish between the two populations by considering the distribution of the signals. A probe is then declared enriched when its signal exceeds a selected cutoff, which is fixed according to the data distribution.

Tiling array data analysis tool Tilemap has been used to identify H3K27me3 targets in *Arabidopsis* by Zhang et al (Ji & Wong 2005; Xiaoyu Zhang et al. 2007). A two-step approach was used by Tilemap to identify peaks. A test statistic is computed for each probe to measure probe-level binding signal. Then the average of the test statistic of probes in a sliding window or Hidden Markov Model (HMM) was used to estimate a window-level signal (Ji & Wong 2005). The resulting 4979 H3K27me3 targets in *Arabidopsis* highly overlap with those identified by other groups. But it is nevertheless a conservative prediction compared with other methods. ChIPmix uses a linear regression mixture model to identify actual binding targets of the protein under study and was also used to identify H3K27me3 targets in *Arabidopsis* (Martin-Magniette et al. 2008; Moghaddam et al. 2011). Instead of using the log ratio, ChIPmix directly works with the IP and INPUT signals of each probe by modeling the distribution of the IP signal conditional to the INPUT signal. They conclude that ChIPmix outperforms the standard approaches based on the log ratio.

Ringo is a package of the open-resource Bioconductor project (Toedling et al. 2007). It is implemented in the statistical programming language R (http://www.r-project.org). *Ringo* also uses a mixture model to get statistical significance for each probes. The null distribution used in *Ringo* is assumed to be symmetric. An upper bound y0 is estimated from the null distribution and probes with signal y > y0 are declared to be from the ChIP enrichment distribution rather than from the null distribution. The contiguous positive probes were further merged into ChIP enriched regions. The advantage of using *Ringo* is the facility to construct automated programmed workflows by using other Bionconductor packages, for example, *affy* for additional normalization, *limma* or *Rankprod* for differential data analysis.

RankProd is a also a Bioconductor project package and implemented in R (http://www.r-

project.org)(Hong et al. 2006). It has been widely used to detect differentially expressed genes in various studies (Gurvich et al. 2005; Suva et al. 2010; Dierssen et al. 2012). *RankProd* was developed from the rank product method (Breitling et al. 2004), which is a non-parametric statistical method making use of biologically intuitive criterion fold-change (FC). Items that are consistently highly ranked in a number of lists, for example genes that are consistently found among the most strongly unregulated (or down-regulated) in replicates are detected. The results of rank product have been shown to be more biologically relevant than those of other methods, especially in studies with noisy data and/or low numbers of replicates (Breitling et al. 2004). RankProd accepts pre-processed measurements in a matrix format and provides functions to perform meta-analysis as well as the analysis of a single experiment.

2 Aim of the Project

Previous studies in Drosophila have shown the importance of PRE in the recruitment of PRC2. It has been postulated that PRC2 will need cis-regulatory elements as targeting sites to establish H3K27me3 in *Arabidopsis*.

The main objective of this study is to check whether cis-regulatory variation caused differential H3K27me3 between accessions. To this aim, we first identified the genes that are differentially H3K27 tri-methylated between Col and Ler then explored the potential cause and the functional significance of the natural variation.

3 Materials and Methods

3.1 Tools and data used

The following list of tools was used in this project for data analysis.

1. Perl (http://www.perl.org), a programming language.
2. R (http://cran.r-project.org) (version 2.10.1), a programming language and environment for statistical computing.
3. Bionconductor (Gentleman et al. 2004) (version 2.6), a free, open source software project to provide tools for the analysis of genomic data. It is based primarily on the R language.
4. GBrowse (Stein et al. 2002), a local instance of GBrowse for visualization of genomic data.
5. BWA 0.5.9(H. Li & Durbin 2009), a short reads mapper for efficient mapping of probes or reads to a reference genome.
6. SAMtools (version 0.1.6_X86_64-linux) (H. Li et al. 2009), provides various utilities for manipulating alignments in the SAM format, including sorting, merging, indexing and generating alignments in a per position format.
7. *ChIPR* (Göbel et al. 2010), a R package with *Ringo* incorporated for ChIP-chip data analysis.
8. *RankProd* (Hong et al. 2006), an Bionconductor package for identification of differentially expressed genes from microarray data based on rank product statistics.
9. SICER (version v1.03)(Zang et al. 2009), a clustering approach for identification of enriched domains from ChIP-Seq data for histone modifications.
10. Picard (http://picard.sourceforge.net) (version 1.16), Java-based command-line utilities that manipulate SAM files.
11. MUMmer 3.22 (Kurtz et al. 2004) (http://mummer.sourceforge.net/), whole genome alignment tool.
12. CSAR(Muiño et al. 2011), a Bioconductor package or ChIP-Seq data analysis, was used to calculate coverage of reads per base.
13. Artemis Comparison Tool (ACT)(Carver et al. 2008), for displaying pairwise comparisons between Col and Ler DNA sequences.

14. The *Arabidopsis* Information Resource (TAIR) (Swarbreck et al. 2008), the database for downloading genome sequences and annotation of *Arabidopsis thaliana.*

15. TAIR genomic coordinates converter, for converting genomic coordinates from TAIR6-8 to TAIR9. Link: ftp://ftp.*Arabidopsis*.org/home/TAIR/Software/UpdateCoordBac/
16. Genesis (Sturn & Quackenbush 2002), Java suite for large-scale gene expression analysis.

The following list of data was used in this project. The other genomic data with accession numbers was summarized in Table 2.

1. Transcriptomic data from *Arabidopsis thaliana* Tiling Array Express (At-TAX) (Laubinger et al. 2008).
2. Transcriptomic data from 19 genome projects (Gan et al. 2011)
3. Ler genome assembly (Schneeberger et al. 2011), The Ler assembly released by 1001 genome project. (http://1001genomes.org/data/MPI/MPISchneeberger2011/releases/2010_09_30/Assemblies/High_Quality/)
4. SNP data between Col and Ler (Schneeberger et al. 2011). (http://1001genomes.org/data/MPI/MPISchneeberger2011/releases/2010_07_01/strains/Ler-1/).
5. TAIR9 genome sequence and annotations, TAIR9_GFF_genes.gff and TAIR9_GFF_genes_transposons.gff. ftp://ftp.*Arabidopsis*.org/home/tair/Genes/TAIR9_genome_release/TAIR9_gff3/
6. Small RNA, the same data as used in track 'Small RNAs (ASRP)' in GBrowse of TAIR, which was provided by *Arabidopsis* Small RNA Project (ASRP). http://asrp.cgrb.oregonstate.edu/db/download.html, downloaded on 26.10.2011.
7. Transposable element (TE) and Transposable element gene (TEG) as described at: ftp://ftp.*Arabidopsis*.org/home/tair/Genes/TAIR8_genome_release/Readme-transposons
8. Other genomic data from database and their accession codes (Table 2).

Table 2. Other genomic data used in the project.

Data	Accession code	Contact name	Database	Reference
H3K9me2	GSE12383	Steve E Jacobsen	GEO	(Bernatavichute et al. 2008)
H3K9me2	E-MEXP-2480	Lars Hennig	ArrayExpress	(Rehrauer et al. 2010)
H3K36me2	GSE7907	Van Nocker	GEO	(Oh et al. 2008)
H3K27me1	GSE22413	Hume Stroud,	GEO	(Jacob et al. 2009)
H3K4me3	GSE7907	Van Nocker	GEO	(Oh et al. 2008)
Nucleosome position	GSE21673	Matteo Pellegrini	GEO	(Chodavarapu et al. 2010)
DNA methylation	GSE5974	Jorja Henikoff	GEO	(Zilberman et al. 2007)

3.2 Biological experiments

Biological materials and experiments done by Julia Reimer were listed as following.

Biological materials and arrays used

Arabidopsis. thaliana of the accessions Col and Ler and their hybrids were grown in LD (16-hours light, 8-hours dark) for 10 days at 20°C on Murashige and Skoog medium supplemented with 1 % sucrose after stratification at 4°C for 2-4 days to synchronize germination. Light was provided by fluorescent tubes. For intraspecific crossing, Col and Ler plants were grown on soil in LD conditions. Five flower buds on the primary shoot and two side shoots were emasculated and manually cross pollinated, while all other flower buds were removed. The seeds of each genotype were pooled and the success rate of the crosses was determined by PCR using primers that detected and AFLP between Ler and Col (fw: 5'-ctggagatcatccaacaaagg-3', rv: 5'-ggcaatggaatgggctggtc-3'). Seed pools with less than 10 % maternal contamination were used in the F1 hybrid studies.

The arrays used were *Arabidopsis thaliana* ChIP-chip 385K Whole-Genome Tiling array set from Roche NimbleGen. The package of arrays contains three slides for the whole genome. Probe length is 50mer. Median probe spacing is 40bp. For ChIP-chip experiment, two biological replicates were performed for each accession respectively. During the ChIP-chip experiments, the samples immunoprecipitated (IP) against H3K27me3 and the Input were labeled with Cy5 (Red) and Cy3 (green), respectively and then hybridized with the probes on the arrays.

ChIP experiments

ChIP experiments were performed as described elsewhere (Göbel et al. 2010) except that chromatin was sonicated with a BioRuptor from Diagenode for 10 times 30s at high setting with 60s intermittant cooling in ice-water. The DNA fragment size of 300 to 1000 bp was controlled by running an aliquot of dee-crosslinked and purified DNA on a 1.5 % agarose gel. The following antibodies were used in immunoprecipations: anti-rat IgG (R9255, Sigma), anti-H3K27me3 (07-449, Millipore) and antiH3K9me2 (pAb-060-050, Diagenode). A very low signal was detected in anti-rat IgG-antibody precipitations and was subtracted as background. Quantitative real time PCR (qRT-PCR) data are shown as fold-enrichment of the input, the error bars represent SE of three technical replicates. At least two independent biological replicates were performed for each experiment and a representative one is shown.

ChIP-chip and ChIP-Seq experiments

ChIP-chip experiments were carried out as described elsewhere using 10 day old seedlings (J. J. Reimer & Turck 2010). DNA samples were amplified using a linker-mediated PCR and hybridized to two-color microarrays from Roche-NimbleGen, input samples were hybridized as reference. Two biological replicates were hybridized per accession. For ChIP-Seq, 80 % of a ChIP experiment precipitated with H3K27me3 antibodies was used to prepare libraries using the ChIP-Seq library preparation kit from Illumina (No. 11257047) according to the manufacturers' instructions. Each library was loaded on two lanes of the Illumina Genome Analyzers IIx to obtain single end 34mer reads. The sonicated input of a chromatin sample from Col was used as background reference.

Expression Analysis

Whole seedlings grown on GM medium were harvested for total RNA extraction at day 7, 10 and 12. The aerial part of soil-grown plants was collected on day 13 and 20 and tissue specific samples (rosette and cauline leaf, stem, open flower, apex enriched and silique) were obtained at day 27 and 34 to extract total RNA with the RNeasymini kit (Qiagen). Five micrograms of RNA were DNase treated using the DNA-free kit (Ambion) prior to cDNA synthesis with SuperScript II Reverse Transcriptase (18064-014,Invitrogen). Quantitative real-time RT-PCR was performed using a Roche Light Cycler and EVA Green dye detection. PP2A (At1g13320) was used as a housekeeping gene.

3.3 Computational methods

The following methods were used for analyze the genomic data related to this project.

3.3.1 Identification of ChIP enriched genes from ChIP-chip data

Before identifying ChIP (H3K27me3) enriched genes, the probes uniquely present in one genome or redundant in either genome should be excluded. Towards this, I remapped probe sequences to the Col genome (TAIR9) and Ler assembly. The following sub-sections are the steps of remapping and identification of ChIP enriched genes based on ChIP-chip data.

Remapping of probe sequences to the Col genome and Ler scaffolds

To map the probe sequences on arrays to Col genome (TAIR9), the short read mapping tool BWA was used. The probe sequences and their original placement on the microarrays were stored in three NimbleGen design (.ndf) files. The probe sequences were extracted from the ndf files and formatted into three FASTQ files (see Script 1), which is the desired input file format for BWA. The probe sequences were then mapped to the Col genome (TAIR9) using different maximal edit distances n. The edit distance includes mismatches and gaps. The following commands show the usage of BWA when mapping probe sequences in one FASTQ file to Col reference genome with n=2:

/path/to/BWA/bwa-0.4.9/bwa aln –n 2 */path/to/*TAIR9_chr12345.fas fastq1> fastq1.sa.sai

/path/to/BWA/bwa-0.4.9/bwa samse */path/to/*TAIR9_chr12345.fas fastq1 fastq1.sa.sai > fastq1.sam

*/path/to/samtools-0.1.6x86-64/samtools/ view –bq 1 /path/to/*TAIR9_chr12345.fas.fai –F 4 fastq1.sam > fastq1.bam

/path/to/samtools-0.1.6x86-64/samtools/sort fastq1.bam fastq1.bam.sort

/path/to/samtools-0.1.6x86-64/samtools/ view –h fastq1.bam.sort.bam > fastq1.bam.sort.bam.txt

The resulting text files listed the coordinates of probes that uniquely mapped to the TAIR9 reference allowing two of maximal number of mismatches since the appearance of gaps in a probe is rare. The problematic probes that were removed include probes having no match to the current genome or having multiple matches. The text output files were further converted into three pos (.pos) format files (Pos_Col) for subsequent data analysis. The original probe sequences were also mapped to the Ler assembly with the same parameters. The resulting output pos files (Pos_Ler), which contain the placement of probes in the Ler assembly, were generated. The Pos_Ler files were used for prediction of

differential H3K27me3 targets between the two accessions but not for identification of H3K27me3 targets in Ler, for which the same Pos_Col files was used as for Col.

Input files preparation based on remapped probes

To use Bioconductor package Ringo to analyze the ChIP-chip data, the pair format files containg the hybridization signal from both channels are needed as input files. The original pair files received from the experiments contain the intensities information for all original probes, including those that were discarded after remapping. To identify H3K27me3 targets in Col and Ler, based on Pos_Col, new pair files were regenerated. The ChIP-chip data for probes after such filtering was used for ChIP enriched genes prediction. The workflow for the data processing is shown in **Figure 4**.

The R package *ChIPR* (Göbel et al. 2010) was used to identify H3K27me3 enriched genes in Col and Ler. *ChIPR* has incorporated *Ringo* and was designed for convenient analysis of ChIP-chip data from NimbleGen arrays. For the two biological replicates for Col, based on Pos_Col, 12 pair files were generated to store the intensities for probes uniquely mapped to Col TAIR9 in the set of 3 arrays and 2 channels. Currently no array based on Ler assembly exists. In this project we thus used the arrays based on Col genome to predict H3K27me3 targets in Ler genome as Col and Ler share high sequence similarity.

To identify H3K27me3 targets in Ler, the same probe set retained after remapping to the Col genome was utilized for Ler. By combining the pos files, Pos_Col instead of Pos_Ler, and original pair files for Ler samples, 12 pair files were generated. So during the identification of H3K27me3 targets in Col and Ler, only the intensities associated with probes were different but the probe set used and the methods for sequential data processing were identical.

Quality control and normalization

The array design files (Pos_Col) and all probe intensity files (24 pair files) were imported into R environment for further data processing. The raw IP and INPUT intensities from double channels were plotted to check the hybridization efficiency for all arrays. The M values (M=log2(R/G), with R (red)=IP and G (green)=INPUT) were calculated for all probes and then used as input for the whole data analysis workflow. The M values were normalized first within arrays and subsequently between arrays. Intensity-dependent *dye bias* is a well-known artifact of two-color microarrays. To correct for this kind of bias, the

Loess (locally weighted scatterplot smoothing) normalization method was first applied within arrays. Afterwards, the scale normalization method within Bioconductor limma package was used between arrays to remove the inconsistency between arrays. The plots before and after normalization for samples of 2 biological replicates are shown in **Figure 7**.

ChIP enriched regions were identified using functions of the Bioconductor package *Ringo*. To reduce signal variance arising from systematic and stochastic noise, the data was separated into windows with a size of 600 bp for which median values were calculated to smooth the data before detection of enriched probes. The positive probes were defined later using the function upperBoundNull within *Ringo*. The definition of threshold in this function is independent of *priori* assumption on the fraction of positive probes among all probes on arrays. The ChIP Enriched Regions, so-called chers, were generated from adjacent positive probes. The regions were further merged if the distance between them were smaller than 300 bp.

Map chers to gene annotation

After identification of chers (H3K27me3 enriched regions in this project), the genes covered by such regions were identified and termed as H3K27me3 targets. The mapping of chers to *Arabidopsis* genome annotations (TAIR9) wad done within *ChIPR* (Göbel et al. 2010). A gene of which only a small proportion is covered by chers might not be a real H3K27me3 target. Genes were considered as H3K27me3 targets in this project, if either at least 30% of the *gene* and *300 bp* were covered by chers or at least 1000bp of it were covered by chers. With such a coverage filter, small genes shorter than 300 bp were excluded. Genes of intermediate length were required to be covered by chers of at least 30%. Genes longer than 3333 base pairs were defined as targets, if at least 1000bp were covered by chers. Furthermore, only genes consistently enriched in both replicates of Col or Ler were defined as H3K27me3 targets in the respective accession. The workflow for identification of H3K27me3 targets in Col and Ler is shown in **Figure 4**.

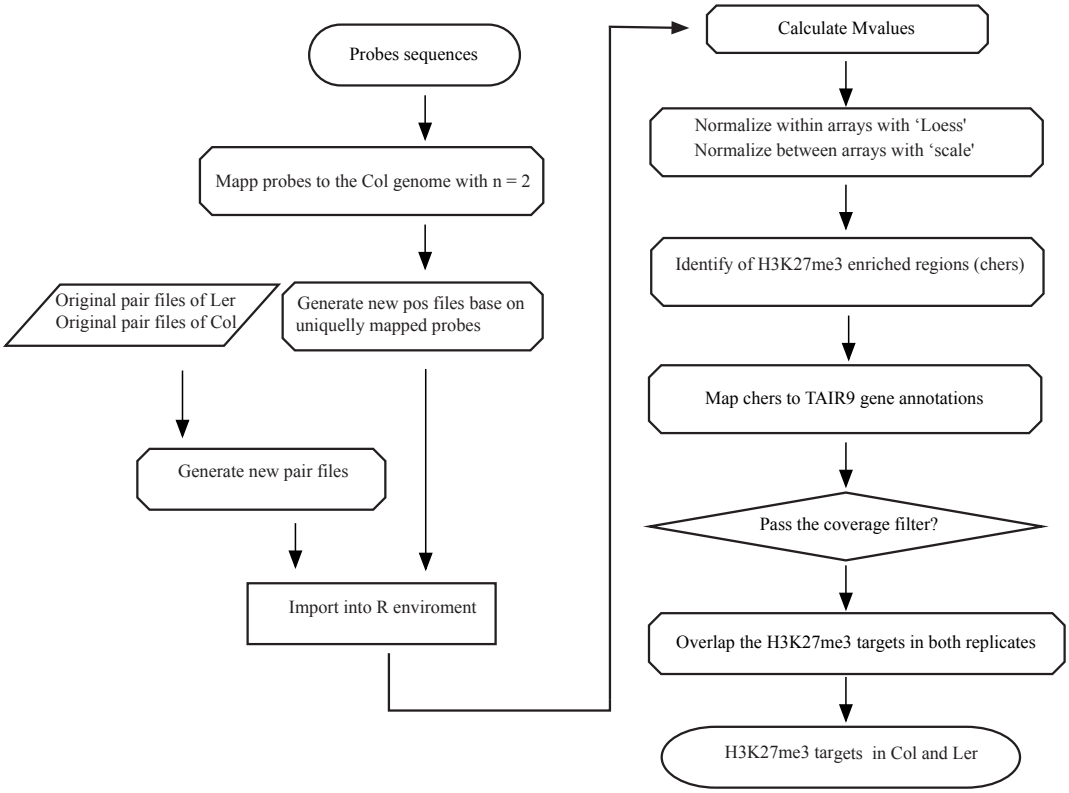


Figure 4. Workflow for identification of H3K27me3 targets in Col and Ler.

3.3.2 Identification of ChIP enriched genes from ChIP-Seq data

Mapping short reads to Col genome

Two lanes of single end short reads were produced to detect the H3K27me3 targets in F1 sample of Col x Ler and Ler x Col separately. The length of reads is 34mers. The raw short read data from Illumina was in FASTQ format. The Illumina FASTQ files were converted into Sanger FASTQ files with a script from Ulrike. The reads from two lanes but the same biological samples were merged then mapped to Col genome with BWA (H. Li & Durbin 2009). A maximal edit distance of n=3 including a maximal gap of one were allowed during mapping. The low quality bases at the end of short reads were trimmed with BWA while mapping by setting parameter q=15. The mapped reads were sorted according to their genome coordinates using SAMtools (H. Li et al. 2009). In order to completely exclude the effect of potential PCR artifacts, redundant reads mapped to the same position in the genome were cleaned with Picard by keeping just one copy. The short reads pileup format file showing which and how many reads pile up at genomic

coordinates were generated using SAMtools (H. Li et al. 2009). The pileup files were further used for allele specific H3K27me3 detection.

Identify H3K27me3-enriched regions and genes

The ChIP-Seq peak detection tool SICER (Zang et al. 2009) was used for identification of H3K27me3 marked regions in hybrids. The wig file containing the mapping information of non-redundant reads was used as input for SICER. The adjacent predicted peaks with gaps smaller than 200bp were merged into broader regions, which were further mapped to TAIR9 gene annotation with *ChIPR* (Göbel et al. 2010). Genes were considered as H3K27me3 targets in F1, if either at least 20% of the gene *and* 200 bp were covered by chers or at least 800 bp of it were covered by chers.

3.3.3 Identification of differentially ChIP enriched genes

Two approaches were used to identify H3K27me3 differentially enriched genes (DEGs) between Col and Ler. Before the release of Ler scaffolds, all original probes present on 3 arrays were used for predicting DEGs. After the release of the Ler scaffolds, only the probes uniquely present in both genomes were used to identify DEGs. The result from the later approach was used afterwards for further evaluation.

Identification of H3K27me3 DEGs with original probes on arrays

The Bioconductor package *RankProd* was used in this study to identify DEGs. For each gene, the median intensity of probes mapping to the gene body was calculated per replicate of Col and Ler. The matrix of medians for genes in all samples was used as input for the Bioconductor package *RankProd* (Hong et al. 2006). The genes detected at a confidence level of percentage of false prediction (pfp) less than 0.15 were defined as differentially enriched. The genes methylated at H3K27 in Ler but not Col are referred to as HLer, the genes methylated at H3K27 in Col but not Ler are referred to as HCol. To ensure that the H3K27me3 enrichment at DEGs was positive at least in one genome, HLer genes were required to be H3K27me3 targets in Ler and HCol genes were required to be targets in Col.

Identification of H3K27me3 DEGs with probes after remapping

To identify DEGs based on conserved, low-copy-number genes between Col and Ler, I remapped the probe sequences to Col genome and Ler scaffolds with BWA. To identify HLer genes, a certain number of mismatches were allowed during remapping. Different maximal numbers of mismatches n (n is 0, 2, 4 or 8) were tested. For each value of n, only the probes uniquely mapped in both genomes were used for subsequent analysis. Probes mapped to multiple positions in either genome were discarded. Best results were obtained using n=2 when compared with confirmation data from independent ChIP-PCR.

The probe set retained for further analysis was different at each value of n. To be consistent with the different probe sets, corresponding pos files and pair files were regenerated and imported into Bioconducor package *Ringo*. The median intensity of probes mapped to a gene body was calculated for each replicate of Col and Ler. The matrix of medians for genes per replicate per accession was then imported into Bioconductor package *RankProd* for DEGs identification. The used cutoff is pfp = 0.15.

In order to exclude false predictions, genes with less than four probes mapped or less than five probes per KB mapped were excluded. The genes that were neither H3K27me3 targets in Col nor in Ler based on co-existing probes in the two genomes were also excluded from further analysis. The workflow for identification of H3K27me3 DEGs with probes after remapping is shown in **Figure 5**.

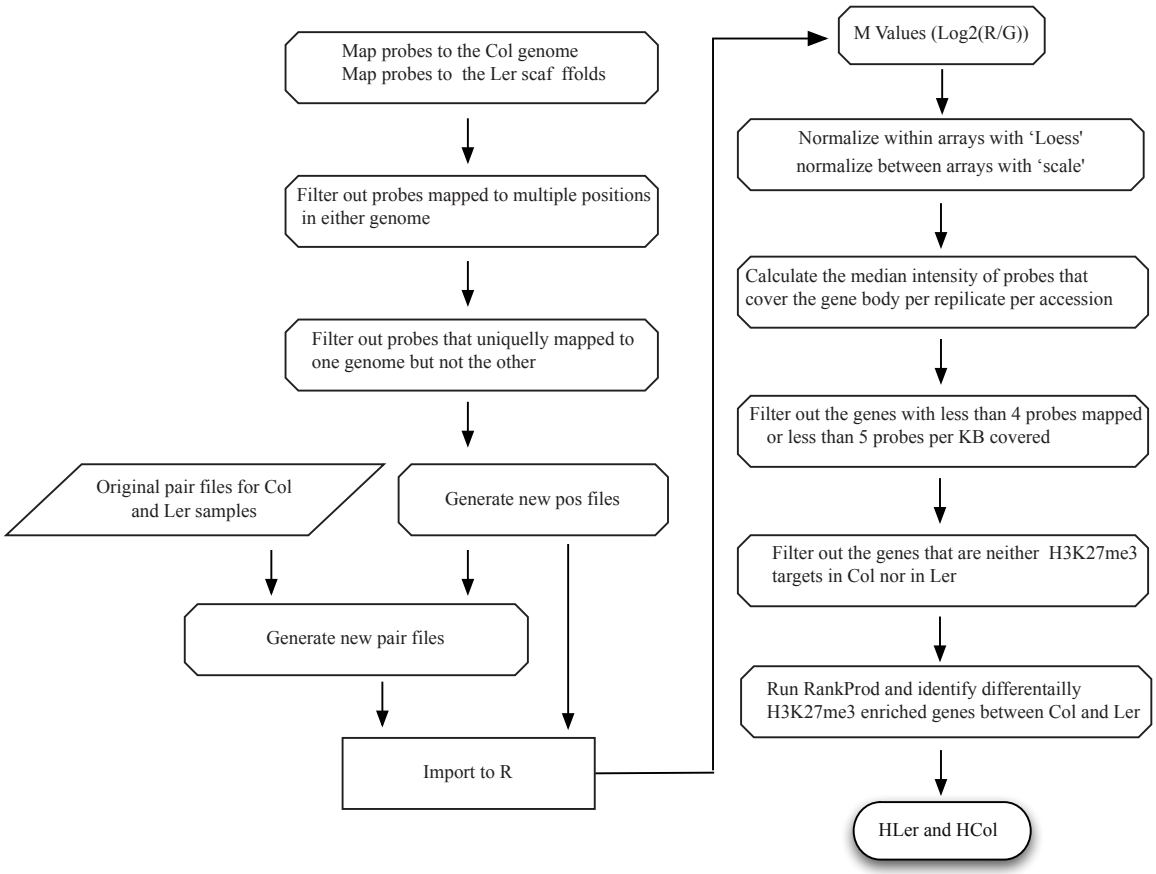


Figure 5. Workflow for identification of differential H3K27me3 targets using remapped probes.
The left panel indicates the process of remapping probes to Col and Ler assembly and the generation of input files for Ringo. The right panel presents the data processing after remapping for identification of differential H3K27me3 targets.

3.3.4 Characterization of differentially methylated genes.

Association between genes and transposable elements (TE)

Genes (except transposable element genes) and transposable elements (TE) in the TAIR9 annotation file were sorted according to their start position on the respective chromosomes. The coordinates of the genes of interest, HLer genes in this project, were extracted from the gene annotation of TAIR9 gene annotation release. It was then tested, whether the gene in front or behind a HLer gene was a TE or not. The occurrence of being adjacent to transposable element of HLer genes was calculated. To get a control distribution, the same amount of genes was randomly chosen from non-H3K27me3 targets in Col genome. The sampling process was repeated 200 times (see Script 2). For

each sampling, the percentage of genes being adjacent to a TE was calculated. The density plot of the percentage in control and HLer genes was plotted (see **Figure 15**).

Association between genes and other histone marks

For each gene in a set, the gene body and 5000bp up- and downstream of the gene were partitioned into 10 intervals, respectively. All together there were 30 intervals for each gene. Because the length of genes varies, the length of each interval in the gene body also varies between genes. For each interval in each gene, first its start and end position in the genome were determined, then it was compared whether it overlapped with histone marked regions or not. Finally, the fraction of all genes overlapping with certain histone marked regions in each bin was calculated and plotted. The Bioconductor package IRanges was used for the analysis. This process was applied for HLer genes, H3K27me3 target genes and non-H3K27me3 targets respectively (see Script 3).

Chromosomal distribution of HLer genes and heterochromatin regions

The five Col chromosomes were divided into bins with a fixed length of 200,000 bp. The amount of SNPs between Col and Ler in each bin was counted. The SNPs data was from 1001 projects (Schneeberger et al. 2011). The sum of SNPs in each bin was plotted along chromosomes. The location of HLer genes was plotted along chromosomes. H3K9me2 marked regions or TE annotations was marked in chromosomes. The percentage of bases coved by H3K9me2 marked region or TEs in each bin was calculated and plotted along chromosomes. (Script 4, see **Figure 16**).

Genome-wide sequence alignment and sequence comparison

The genome alignment tool MUMmer was used to align all scaffolds of Ler to the Col reference sequence. The alignment was performed following the instructions for "Mapping a draft sequence to a finished sequence" (*http://mummer.sourceforge.net/manual/#mappingdraft*). The parameter setting used was "nucmer --mum -b 1000 -l 35 -c 80 -f --prefix=outputFolder referenceSquence *LerassemblySequence*". With this setting, only anchors that were unique in both reference and query were allowed for alignment. In a second step, nucmer extends alignments across high diversity regions by maximally 1000bp. If the diverging regions or indels were larger than 1000bp, the alignment would break. Finally, we restricted the alignment to match only on the forward strand of the query.

To check whether the TEs flanking HLer in Col Genome exist or not in Ler genome, the HLer sequences at gene body and their up-/downstream regions were extracted from both genomes and annotated with a custom R script (Script 4). Gene sequences of each HLer gene body and its flanking regions in Col and Ler were aligned again with MUMmer. The alignment result was visualized with Artemis Comparison Tool (ACT) (see **Figure 17**).

Allele specific H3K27me3 detection

With SNP data between Col and Ler from the 1001 genomes project and the short read pileup file generated with SAMtools (H. Li et al. 2009)**,** the allele frequency of HLer, HCol and common H3K27me3 targets were calculated. SNPs to which with less than 3 non-redundant reads mapped, were not included in the calculation. Reads harboring nucleotides that were neither identical to the reference nor to Ler at SNP locations were discarded (see **Figure 19** B/C).

To test whether there is a high probability to observe the allele frequency of HCol and HLer in common H3K27me3 targets, the same amount of SNPs as in HCol or HLer was drawn from all SNPs 100000 times. The allele frequency was calculated for each randomization. The distribution of allele frequency from the 100000 samplings was compared with that from HCol and HLer. From this, I estimate the probability of observing an allele frequency as extreme as observed in HLer or HCol from common H3K27me3 targets to be less than 0.0001.

4 Results

4.1 Identification of H3K27me3 targets in Col and Ler

In *Arabidopsis*, H3K27me3 targets several thousands of genes in seedling, undifferentiated meristem and differentiated leaf (Turck et al. 2007; Xiaoyu Zhang et al. 2007; Lafos et al. 2011). It was suggested that H3K27me3 is required for stable gene repression throughout most of the plants life cycle. In order to uncover the natural variation of H3K27me3 distribution within *Arabidopsis*, the genome-wide distribution of H3K27me3 in two *Arabidopsis* accessions, Columbia (Col) and Landsberg *erecta* (Ler), was profiled using ChIP-chip technique. 10-day-old seedlings from the two accessions were used as plant material (see section 3.3). The DNA fragments from chromatin after immunopreciptation with antibodies against H3K27me3 (IP) and the control input samples were hybridized to two-color microarrays of *Arabidopsis*, which is *3*-slide NimbleGen *Arabidopsis* Tiling Array Set covering whole genome of Col. Since there is no array designed for the Ler genome and the genome sequences between Col and Ler are highly similar, the same arrays were used for Ler samples. Two biological replicates were hybridized per accession.

4.1.1 Remapping probes to *Arabidopsis* genome

The tiling array set used was designed based on the Col genome release of TAIR6. The coordinates for genomic features were not changed from TAIR6 to TAIR8. However, from TAIR6 to TAIR9, the genome sequences, coordinates and annotations have been updated significantly. A certain amount of probe sequences are possibly not unique or even not present in Col genome according to the TAIR9 annotation. This ambiguity of probes could cause false prediction in further data analysis.

In order to use only the probes that are present as a single copy in the Col genome to predict H3K27me3 targets, the probe sequences on the 3 slides of the array set were mapped to the TAIR9 genome sequences. Only probes mapped to unique positions in the genome were retained. To assess the extent of close matches (highly similar probes), a moderate number of mismatches (n) were allowed during remapping. Theoretically, when using increasing number of n, more probes that mapped to multiple positions and having

more cross hybridization issues can be excluded but more probes with few mismatches will be retained.

The effect of allowing two different maximal numbers of mismatches n (n=0 or n=2) has been evaluated (see **Table 3**). While using n=0 or n=2, a similar proportion of probes, 95.3% and 95.4% of original amount of probes, respectively, were retained for further analysis. The retained probes correspond to the same amount of 32496 genes, which cover a big proportion relative to the whole 33239 genes in the Col genome. The amount of probes filtered out is slightly different but corresponds to the same set of 530 genes. Additionally, 270 new genes in TAIR9 but not in TAIR8 were incorporated into analysis after remapping.

Since using n=0 or n=2 does not influence overall results, to be consistent with the number of maximal mismatches used during the identification of specific H3K27me3 targets unique in Col or Ler, the probes retained after remapping using n=2 have been used afterwards to identify H3K27me3 targets in Col and Ler.

Table 3. Statistics of probes and genes present on the arrays before and after remapping

	Genome release	**Probes number**					**Number of retained/all annotations***	**Specific genes**
	-	Slide1	Slide2	Slide3	All	Percentage after remapping	-	-
Before Remapping	TAIR8	385991	385991	385991	1157973		32756/33003	530
After Remapping	TAIR9 (n=0)**	363812	369190	370659	1103661	95.30%	32496/33239	270
After Remapping	TAIR9 (n=2)**	363936	369319	371020	1104275	95.40%	32496/33239	270

* the term annotation includes protein coding gene, transposable element gene and pseudogene, miRNA, snoRNA, tRNA, rRNA, snRNA, ncRNA, all AGI identifiers in the genome annotation file (TAIR9_GFF_genes.gff or TAIR8_GFF_genes.gff) released by TAIR.

** indicates the maximal number of mismatches per probe allowed during remapping to TAIR9 genome.

4.1.2 Quality control and normalization

New probe design files and intensity files for retained probes were generated (see section 4.2.1) for further analysis. To check the hybridization efficiency and reproducibility between biological replicates in Col and Ler, the log2 scaled raw intensities from red (CHIP) and green (INPUT) channels for retained probes were plotted individually per slide per accession (see **Figure 6**). The distributions of raw intensities from all slides are very similar and show normal distributions approximately, which means the hybridization

is reproducible and efficient. Only on slide 3 for the Col replicates and on slide2 for Ler replicates, the distributions do not overlap very well. Additionally, there are local small peaks (indicated by arrows in Figure 6) in the low intensity region of the plots for nearly all slides of Ler samples, indicating that certain amount of probes cannot hybridize with DNA samples coming from Ler. These probes are derived from Col specific genes that are absent in Ler.

To reduce the intensity-dependent dye bias and the different background among slides, the intensity signals of probes were normalized with the normalization method 'Loess' (locally weighted scatterplot smoothing) within slides and then 'scale' between slides (see section 3.3.1). Afterwards, the M-Values, log2 ratios of intensities from Red/Green channels (log2 (IP/INPUT)), for all probes of the according sample were plotted before normalization and after normalization (see Figure 7). After normalization, the M values are centralized approximately at zero for all slides and their distributions are nearly identical in the area of the major peak.

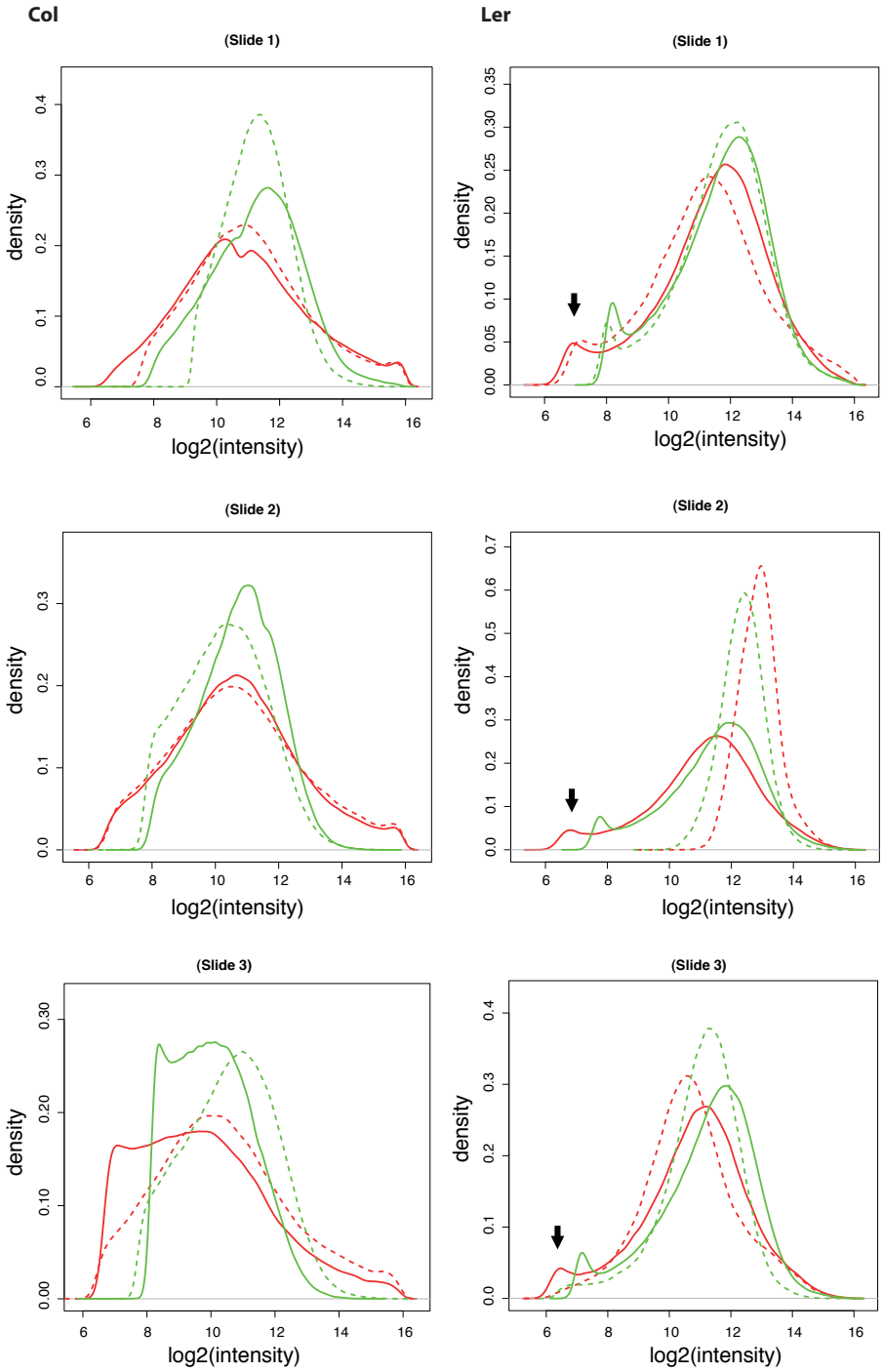


Figure 6. The density of raw intensities in red and green channels in 3 slides for all samples. The left panel is for Col and right panel for Ler. The x-axis shows the log2 scaled intensities of red (H3K27me3) and green (INPUT) channels. The y-axis shows the densities of the scaled intensities form both channels for all replicates. The red solid lines show the densities of intensities from red channels and green solid lines show that of green channels. The dashed red and green lines show the densities for red and green channels for the second biological replicate. The black arrows indicate the local peaks formed by non-hybridized probes. The set of three slides covering whole genome of Col used are indicated as slide1, slide2 and slide3.

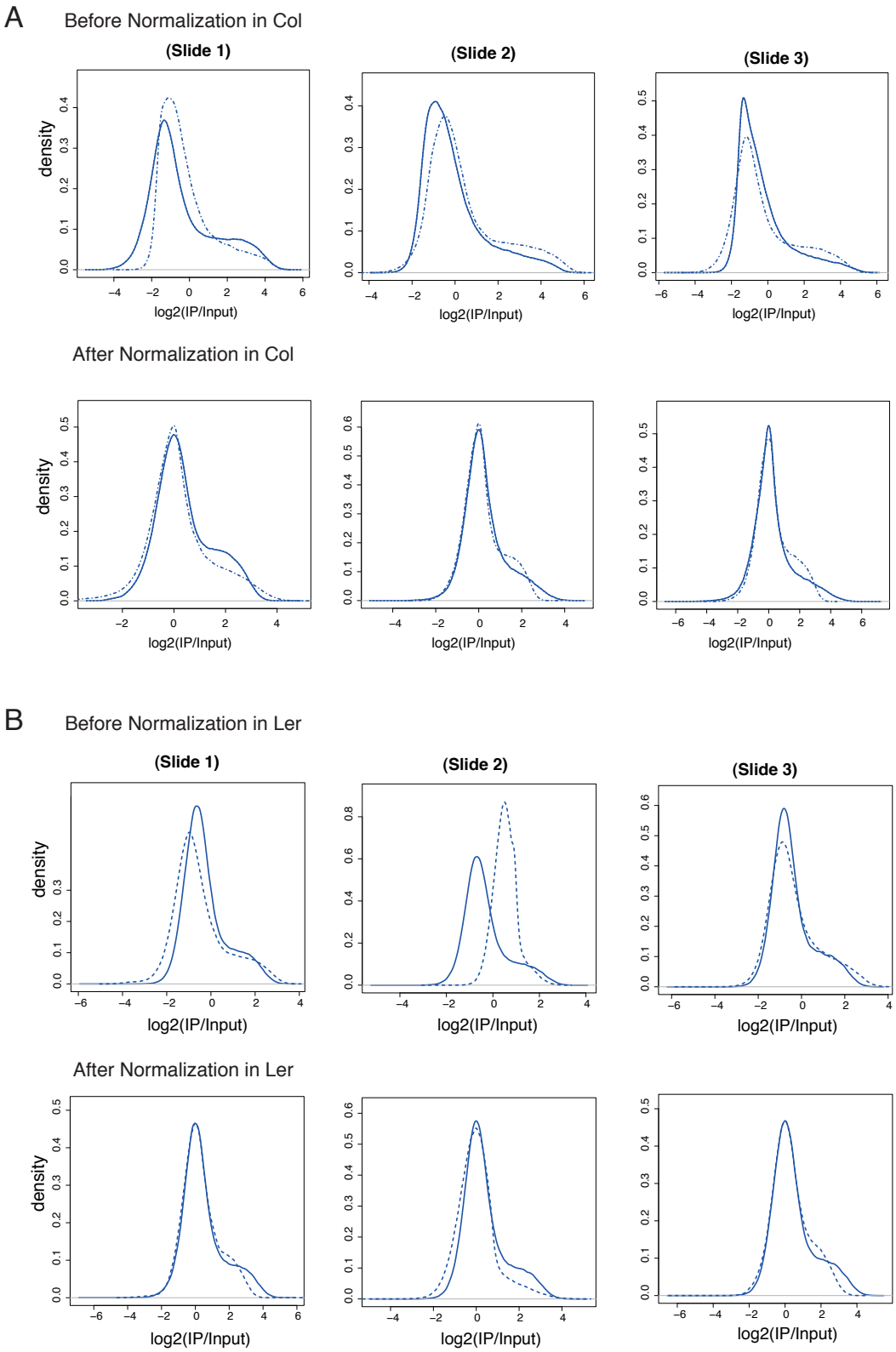


Figure 7. Distribution of log2 (IP/INPUT) before and after normalization in Col and Ler. The x-axis shows the log2 (IP/INPUT) for each replicate in each slide. The y-axis shows the density before and after normalization of the log2 (IP/INPUT). The upper panel (A) represents Col, the lower panel (B) represents Ler. The solid and dashed blue lines show the distributions for two replicates in each slide.

4.1.3 Profiles of H3K27me3 in Col and Ler

The normalized log2 (IP/INPUT) for Col and Ler samples were uploaded to a customized implementation of GBrowse and displayed as histograms aligned with the Col genome annotation and other features of interest. A visual analysis of the Col and Ler methylation levels revealed that the majority of loci exhibit very similar H3K27me3 enrichment along chromosomes (Figure 8).

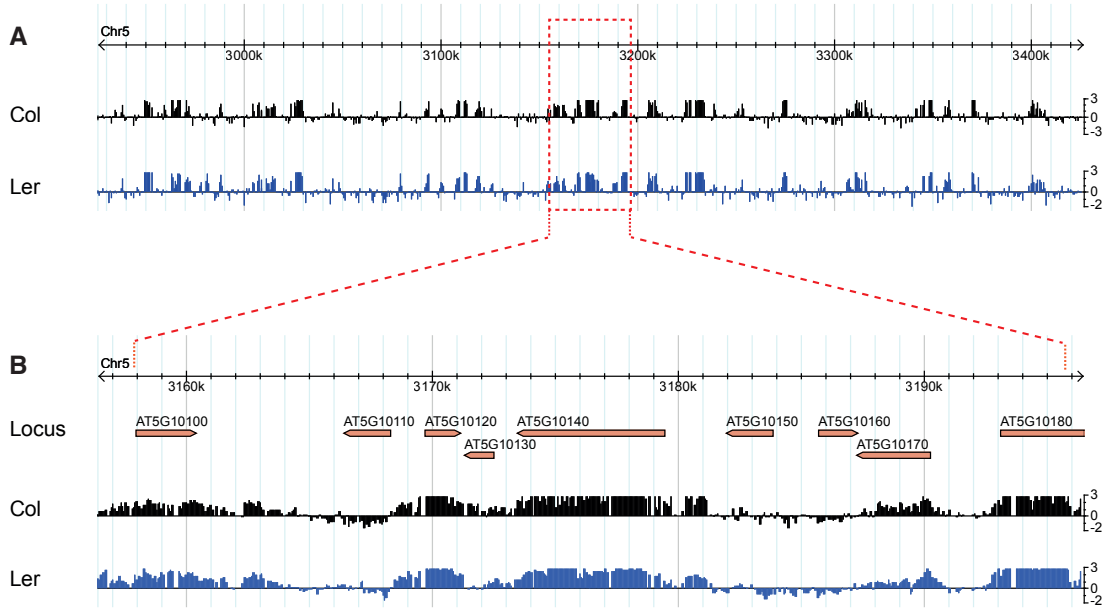


Figure 8. H3K27me3 profiles in Col and Ler in a representative region of Col.
H3K27me3 profiles in Col and Ler were shown in black and blue, respectively. (A) Mean log2(IP/INPUT) for each probe in two replicates for Col and Ler are displayed for about 400kb region of chromosome 5. (B) An enlarged region of about 30-kb of chromsosome 5 including the gene AT5G10140. The track 'Locus' shows the TAIR9 gene annotation in red bar.

The Bioconducter package *Ringo* was used for identification of H3K27me3 positive probes and H3K27me3 enriched regions (chers) in both genomes. The term 'chers' is defined in *Ringo* as ChIP enriched regions and means H3K27me3 positive regions in this project. The summary of chers in each sample is shown in Table 4. A A similar amount of chers was identified in replicates of Col and Ler.

Table 4. H3K27me3 positive regions were identified in two replicates of Col and Ler samples

Chers	Mismatches n	Replicate 1	Replicate 2
Col	0	9830	9086
	2	10482	9694
Ler	0	8912	9142
	2	9456	9792

The chers were mapped to TAIR9 gene annotations with R package ChIPR (Göbel et al. 2010). The coverage of chers on genes and the length of genes have been considered during the definition of H3K27me3 target in this project. In this project, *genes* of which at least 30% and *300 bp* were covered by chers were defined as H3K27me3 targets. But for very long genes, this threshold is too restrictive, so the genes of which at least 1000bp were covered by chers were also considered as H3K27me3 targets. Only genes consistently passing the threshold in both replicates of Col or Ler were defined as H3K27me3 targets in the respective accession. With re-annotated unique probes allowing maximal two mismatches during remapping, 6370 H3K27me3 targets were identified in Col and 6344 H3K27me3 targets were identified in Ler (see Table 5a, Figure 9). This number of H3K27me3 targets in Col is consistent with previous studies carried out on Col seedlings by different laboratories using various experimental platforms (Table 5b). The H3K27me3 targets identified in Col and Ler are highly overlapping (**Figure 9**). The two accessions share 5452 genes, which is also close to that of H3K27me3 targets identified in Col in other laboratories (Table 5b).

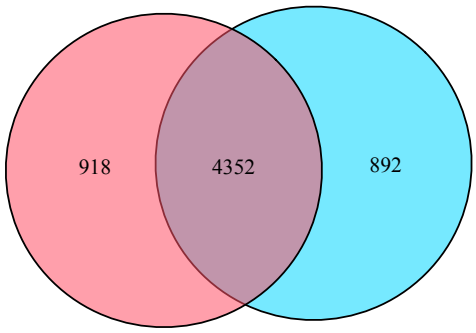


Figure 9. The H3K27me3 profile in in Col and Ler are highly similar.

The Venn diagram representing 4352 overlapping H3K27me3 targets in Col (red) and Ler (blue). 918 H3K27me3 targets specific in Col and 892 are specific in Ler according to the intersection analysis.

Table 5. The H3K27me3 targets identified in replicates, Col &Ler (a) and other groups (b)

a

Accessions	Replicates	Number	In R1 and R2	In Col and Ler
Col	R1	7568	6370	5452
	R2	6801		
Ler	R1	6976	6344	
	R2	6875		

b

Groups	Number	Overlap	Proportion*
Jacobsen	4979	4126	64.8%
Van Nocker	7856	5873	92.2%
Schubert	7463	5392	84.6%
Cao	5088	4293	67.4%

*Proportion means the overlapping H3K27me3 targets with other groups relative to that we identified in Col.

4.2 Prediction of differentially H3K27me3 enriched genes (DEGs)

4.2.1 Remapping probes to the *Arabidopsis* genome and Ler assembly

The ChIP-chip data for detecting the genome-wide distribution of H3K27me3 in Col and Ler was also used to identify differentially H3K27me3 enriched genes (DEGs) between the two accessions. As shown in **Figure 8** and **Figure 9**, the distributions of H3K27me3 in the two accessions are highly similar. There are about 14% genes that are H3K27me3 targets in one genome but not in the other according to the intersection analysis (Figure 4). This proportion is similar to that of non-overlapped H3K27me3 targets between two biological replicates using Chip-chip technique (Table 5). However, for identification of differentially enriched genes, it is not enough to just take the genes that are outside of the intersection of the H3K27me3 targets in two accessions. The non-overlapped targets could be due to structural variation such as copy number variation and presence-absence variation or when the enrichment in one accession is just above the threshold and the other just below.

To minimize the effects of probe copy number and genomic polymorphisms on the detected methylation levels, the probes on the arrays were remapped to the Col genome (TAIR9) and Ler scaffolds and then only probes uniquely mapped to both genomes were kept for detecting differentially H3K27me3 enriched genes. The problematic probes that mapped to multiple positions in either genome or only to one of the genomes were discarded. After this process, the genes having multiple closely related copies or being uniquely present in one genome but not the other are excluded from further analysis. Only single copy genes in Col that are also present in Ler are included for further analysis.
To identify H3K27me3 targets specifically in Ler (HLer), a maximal mismatch number n=2 was used during remapping to TAIR9 to keep the homologues genes in Ler that have moderate sequence diversity. In contrast, for the identification of H3K27me3 enriched genes specifically in Col (HCol), no mismatch (n=0) was allowed during remapping, which means only probes with perfect, unique matches in both accessions were included. This more stringent threshold helps to avoid false positive detection of HCol genes due to low efficiency of hybridization of Ler samples to probes in arrays, which was designed based on the genomic sequence of Col.

4.2.2 Identification of DEGs based on remapped probes

Based on set of the retained probes, probe design files and intensity files were generated accordingly. A matrix containing medians of M values of all probes mapped to corresponding genes per replicate and accession was generated. The statistical methods implemented in Bioconductor package RankProd were used to identify HLer and HCol genes (see section 3.3.3). At a maximal percentage of false prediction (pfp) of 0.15, a small number of genes were identified as differential H3K27me3 targets between Col and Ler. When using all the probes originally present on the whole set of chips without considering cross hybridization issues caused by redundant probes, 114 HLer and 76 HCol genes were identified. Such DEGs genes were poorly confirmed. For example, AT2G15327 was identified as HCol gene but later shown that it is only present in Col genome (Julia Reimer, personal communication). After excluding problematic probes and only keeping single copy genes present in both genomes, 32 HLer genes and 11 HCol genes were identified (Table 7, Table 6). The density of hybridization signals for probes mapping to HLer or HCol genes is higher in the accession where they are specifically H3K27me3 modified (see Figure S1). The higher H3K27me3 enrichment in Ler than in

Col for HLer genes were well validated (see Discussion). Gbrowse view of an example HLer gene AT5G35810 is shown in **Figure 10**. Several gene ontology (GO) analysis tools were used to find whether HLer or HCol genes are enriched in certain functional annotations, but no GO term is significantly overrepresented in HLer or HCol genes.

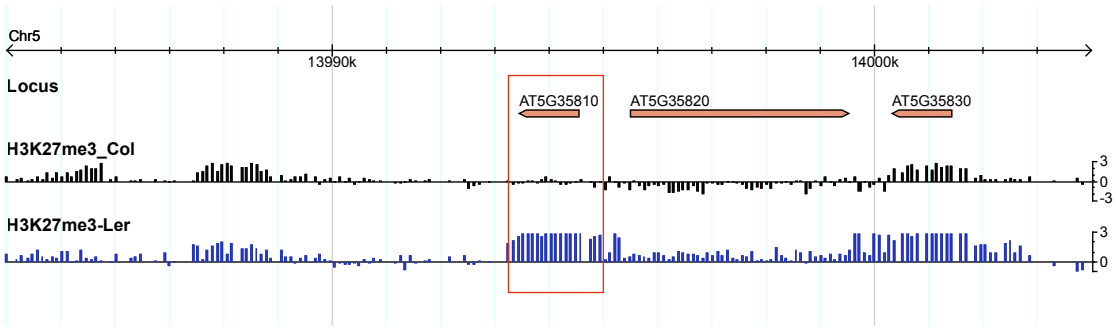


Figure 10. GBrowse view of an example of HLer gene AT5G35810.
GBrowse view is presented for 18kb of chromosome 5 that include a differentially methylated region (red box). Top track shows the gene annotation in TAIR9. The next two tracks show Col (black) and Ler (blue) relative methylation levels for all probes within these regions. Methylation levels are defined as the normalized log2 ratio of IP enriched sample relative to INPUT genomic DNA on a scale of -3 to 3. The methylation levels shown here are the mean of two biological replicates.

The gene name and annotations of HCol and HLer genes are listed in Table 6 and Table 7.

Table 6. The HCol genes identified with pfp 0.15 as threshold.

Gene name	FC(Ler/Col)	pfp	P-value	Annotations
AT5G60610	0.1617	0	0	F-box/RNI-like superfamily protein
AT5G28145	0.189	0.005	0	Transposable element gene; copia-like retrotransposon family
AT4G24420	0.1837	0.0067	0	RNA-binding (RRM/RBD/RNP motifs) family protein
AT1G11450	0.1998	0.0075	0	Nodulin MtN21 /EamA-like transporter family protein
AT4G29770	0.1911	0.008	0	Target of trans acting-siR480/255
AT1G31250	0.2107	0.015	0	Proline-rich family protein
AT5G36240	0.2063	0.0157	0	Zinc knuckle (CCHC-type) family protein
AT5G19875	0.2584	0.0412	1.00E-04	Unknown protein
AT4G22513	0.2595	0.0656	1.00E-04	Encodes a Protease inhibitor/seed storage/LTP family protein
AT4G32230	0.2525	0.064	1.00E-04	Unknown protein
AT5G11070	0.2696	0.0791	2.00E-04	Unknown protein

Table 7. The HLer genes identified with pfp 0.15 as threshold

Gene name	FC(Ler/Col)	pfp	P-value	Annotations
AT5G35810	8.0596	0	0	Ankyrin repeat family protein
AT5G56920	5.9724	0	0	Cystatin/monellin superfamily protein
AT5G35914	5.206	0.0067	0	Transposable element gene
AT1G30835	5.0965	0.0075	0	Member of Sadhu non-coding retrotransposon family
AT5G28463	4.8083	0.014	0	Unknown protein
AT3G60150	4.6442	0.0117	0	Protein of unknown function (DUF498/DUF598)
AT4G03566	4.8993	0.01	0	Unknown protein
AT5G42640	4.2458	0.0225	0.00E+00	C2H2 and C2HC zinc fingers superfamily protein
AT4G20480	4.1671	0.0233	0.00E+00	Putative endonuclease or glycosyl hydrolase
AT2G20910	3.9212	0.022	0.00E+00	Pseudogene
AT4G26350	3.9237	0.0264	0.00E+00	F-box/RNI-like/FBD-like domains-containing protein
AT5G28615	3.7386	0.0292	0.0001	RNA-directed DNA polymerase related family protein
AT1G35400	3.5033	0.0608	0.0001	CONTAINS InterPro DOMAIN/s: unknown protein
AT5G02700	3.4585	0.0664	0.0002	F-box/RNI-like superfamily protein
AT5G28610	3.5238	0.076	0.0002	BEST match is: glycine-rich protein (TAIR:AT5G28630.1)
AT1G65170	3.382	0.0838	0.0002	Ubiquitin carboxyl-terminal hydrolase family protein
AT2G16830	3.3951	0.0906	0.0002	Pseudogene, similar to plasma membrane intrinsic protein 3
AT3G60560	3.4693	0.0861	0.0003	Unknown protein;
AT1G66300	3.232	0.1042	0.0003	F-box/RNI-like/FBD-like domains-containing protein
AT4G10870	3.214	0.1125	4.00E-04	Unknown protein
AT5G12910	3.1846	0.1129	4.00E-04	Histone superfamily protein
AT4G09143	3.198	0.1086	4.00E-04	Pseudogene
AT2G01560	3.1715	0.1113	4.00E-04	Plant protein 1589 of unknown function
AT1G35186	3.0756	0.1317	0.0005	Similarity to non-LTR retroelement protein
AT3G46160	3.1138	0.1372	0.0006	Protein kinase superfamily protein
AT1G54230	3.032	0.1412	0.0006	Winged helix-turn-helix transcription repressor DNA-binding
AT1G21870	3.043	0.1367	0.0006	Encodes a Golgi-localized nucleotide-sugar transporter
AT2G36710	3.0791	0.1332	0.0006	Pectin lyase-like superfamily protein
AT2G34840	2.9816	0.1303	0.0006	Coatomer epsilon subunit
AT5G56910	3.0358	0.141	0.0007	Proteinase inhibitor I25, cystatin, conserved region
AT1G57565	3.0076	0.1384	0.0007	SWI-SNF-related chromatin binding protein
AT3G60965	2.9695	0.1384	7.00E-04	Transposable element gene; copia-like retrotransposon family

4.3 Computational analysis of HLer genes

This section includes the results from computational analysis of HLer genes. The computational analysis includes the expression analysis of HLer genes, association of HLer genes with other chromatin features and TE, sequence comparison of HLer genes between Col and Ler. Finally, a model is proposed to explain the specific loss of H3K27me3 modification in Col but presence in Ler.

4.3.1 Expression analysis of HLer genes

The expression analysis of HLer genes was done based on three different sources of data, including the expression data in different development stages in Col (published data), in Col and Ler for chosen HLer genes (see Discussion) and in seedlings of 19 *Arabidopsis* accessions (published data).

In the different developmental stages of the Col genome

H3K27me3 is associated with target gene repression generally. As defined before, HLer genes are H3K27me3 targets in the Ler genome but not in the Col genome. It is interesting to know whether the absence of this mark in Col caused higher gene expression in this accession or not. Whole genome gene expression measurement in Col has been done comprehensively. Therefore, I explored the expression level of HLer genes in the Col genome using published data.

The data used here for expression analysis of HLer in Col originated from At-TAX project, in which tiling arrays were used to measure the expression of genes in different developmental stages in Col (Laubinger et al. 2008). The matrix of expression values of HLer genes in different developmental stages was used as input for the clustering tool Genesis (Sturn & Quackenbush 2002). The expression values were adjusted by choosing "mean central experiments", afterwards the genes were clustered using the hierarchical clustering method. The clustering results show that HLer genes can be clustered into two groups (**Figure 11**). One group of genes almost is almost not expressed in all the stages studied, although they do not have the repressive mark of H3K27me3 in Col (Rep_Col). The other small group of genes, overall five genes: AT4G20480, AT2G34840, AT5G56910, AT3G60150, AT1G30835, show relatively high expression in almost all

studied stages (Exp_Col) (**Figure 11**). Further analysis reveal that the five Exp_Col genes are marked by the active mark H3K4me3 in Col (see **Figure 12**).

In seedlings of 19 accessions

To investigate how the expression pattern of HLer genes changes among other accessions of *Arabidopsis*, the expression of HLer genes in seedlings of 19 other accessions were compared. The transcriptome data for 19 *Arabidopsis* accessions were generated with RNA-Seq technique as part of 19 genomes project (Gan et al. 2011). In the downloaded files, each gene was assigned with a RPKM value in each replicate, where RPKM means reads per kilobase of exon model per million mapped reads. The mean value of three replicates for each gene was calculated, and then the mean expression values of HLer genes in 19 accessions were used as input for the clustering tool Genesis (Sturn & Quackenbush 2002). The expression value was not adjusted afterwards. To show the change of colors with change of expression value, the maximal value for viewing was set to 4. After hierarchical clustering analysis, the accessions were clustered into two big groups and Col and Ler are in two different groups (**Figure 11**). All HLer genes were also clustered into two groups according to the their expression pattern in the different accessions. One big group of HLer genes is not expressed in any accession but the other small group shows variable expression among accessions. Strikingly, nearly all Exp_Col (4/5) are in the group that shows variable expression among accessions with the exception of AT1G30835 which is not analyzable in the data used. Additionally, AT2G20910 also shows variable expression among accessions in seedlings. It is active in several accessions including Ler although it is H3K27me3 positive in Ler. All other genes (Rep_Col, defined based on expression in developmental data in Col, except AT2G20910) show constant repression in seedlings in all 19 accessions assayed. Two genes of AT1G30835 and AT5G35914 were not analyzable in this data set.

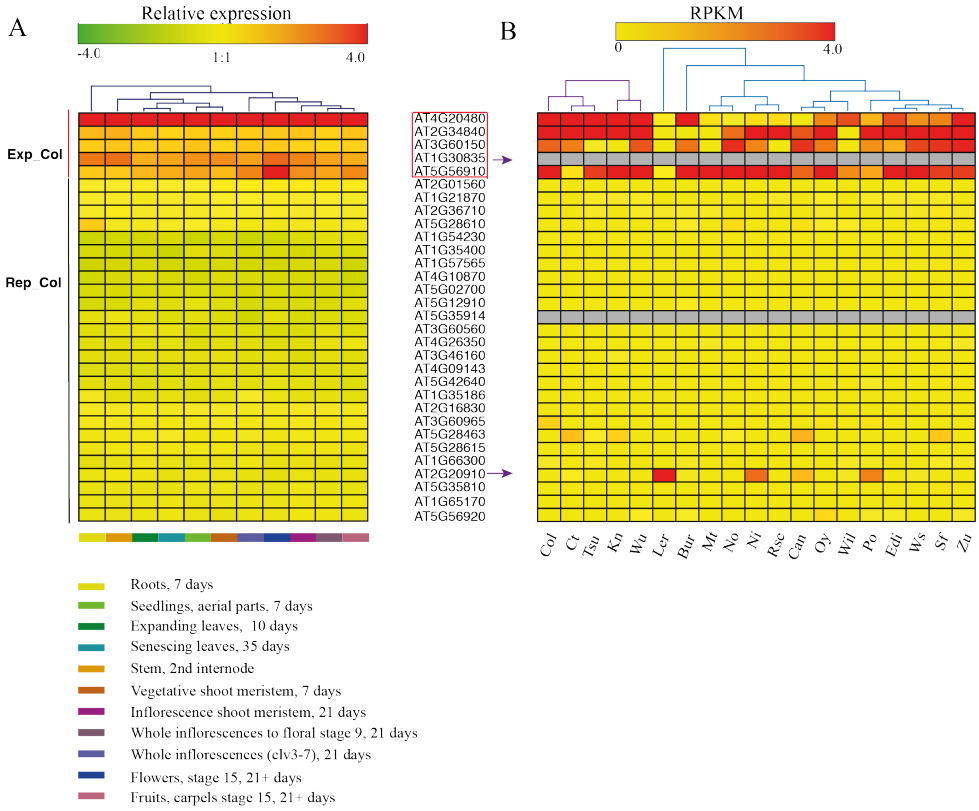


Figure 11. Two clusters of HLer genes according to their expression level
Two clusters of HLer genes according to their expression pattern in Col or 19 *Arabidopsis* accessions. Red or green indicate active or repressed state of a gene in that condition. (A) Expression pattern of HLer genes in Col. The expression value was normalized to the mean of experiments. The developmental stages and tissues are listed in lower panel and indicated in different color. The five HLer genes in red frame are a group of genes with relatively active (Exp_Col) and the rest are genes repressed (Rep_Col). (B) Expression pattern of HLer genes 19 *Arabidopsis* accessions. HLer genes again are grouped into two clusters based on their expression in seedlings of 19 accessions. Each row represents a gene and each column represents an accession. The accessions are indicated above and below the heatmap. These are Col, Ct, Tsu, Kn, Wu, Ler, Bur, Mt, No, Ni, Rsc, Can, Oy, Wil, Po, Edi, Ws, Sf, Zu. AT2G20910 (purple arrow) do not express in Col but in Ler and other 3 accessions. Gray indicate the genes that are not analyzable. AT1G30835 (purple arrow) has shown relatively high expression in seedlings in Col in At -TAX project and our lab, but it is not analyzable (gray) in the data of 19 genomes project (Gan et al. 2011).

In summary, we have evaluated the expression pattern of HLer genes from different sources, including the expression data at different time points in Col (Laubinger et al. 2008) and seedlings of other 17 accessions (Gan et al. 2011). Taken together, HLer genes can be classified into two groups based on the expression analysis. One group of HLer genes is the differentially H3K27me3 marked genes that are associated with gene expression variation between accessions. It includes Exp_Col plus AT2G20910. These genes show variable expression at time points in Col or in seedlings of all assayed

accessions. The rest of HLer genes, which are in the second group, are differentially H3K27me3 marked genes that are associated with constant repression in all the accession and at all time points assayed. The expression of random selection of genes in this two groups in Col and Ler show distinct, either active or repression in expression between groups but similar pattern between accessions although they are differential in H3K27me3 in seedlings (Julia Reimer, personal communications, see Discussion). The exception AT2G20910 shows a constant repression in Col (**Figure 11**A), but is active in seedlings in Ler (**Figure 11**B), which is in contrast to Exp_Col. Nevertheless, it belongs to genes associated with variable gene expression among accessions.

4.3.2 Association of HLer genes with various histone marks

Occupancy of H3K4me3, H3K27me3 and H2A.Z at HLer genes in Col genome

The pattern of gene expression is usually associated with various histone modifications. It has been shown in the last section that HLer genes can be classified into two groups based on their expression. To evaluate the association of different expression pattern of HLer genes in Col with some histone modifications, I investigated the signal profiles of several histone marks in the **Col** seedlings over the HLer genes and their up/downstream 5kb regions. Chromatin modifications explored in this section are always from Col genome not Ler. First of all, HLer genes are depleted of H3K27me3 modification whereas H3K27me3 targets are highly enriched with H3K27me3 especially in gene body (**Figure 12**A). Moreover, being consistent with their active expression in Col (**Figure 11**), the five HLer genes, Exp_inCol, are enriched for active mark H3K4me3 (D). However, active genes are only a small proportion among all HLer genes. On average, HLer genes are marked by the active mark H3K4me3 in low levels (**Figure 12**B), which is consistent with the low expression of most HLer in Col (**Figure 11**). H3K4me3 pattern of HLer in Col is similar to that of H3K27me3 targets although HLer genes are not H3K27me3 targets. In contrast, in non-targets, including numerous highly expressed genes, the H3K4me3 enrichment peaks at the 5' end of the transcribed region (**Figure 12**B).

It has been reported that H2A.Z is enriched at Polycomb Complex target genes by Suz12 in ES Cells and is necessary for lineage commitment (Creyghton et al. 2008). In *Arabidopsis*, H2A.Z and H3K27me3 are enriched along gene bodies and display a similar spatial patterning over H3K27me3 target genes (**Figure 12**A and **Figure 12**C). H2A.Z in *Arabidopsis* seems to be preferentially enriched in TSS but also extends across larger

regions into the gene bodies. In contrast to H3K27me3 targets, HLer genes are not associated with H2A.Z in Col.

In summary, HLer genes are not correlated with H3K4me3 except the five HLer genes of Exp_Col. HLer genes are not enriched in H2A.Z, which are usually co-localized with H3K27me3 targets in Col.

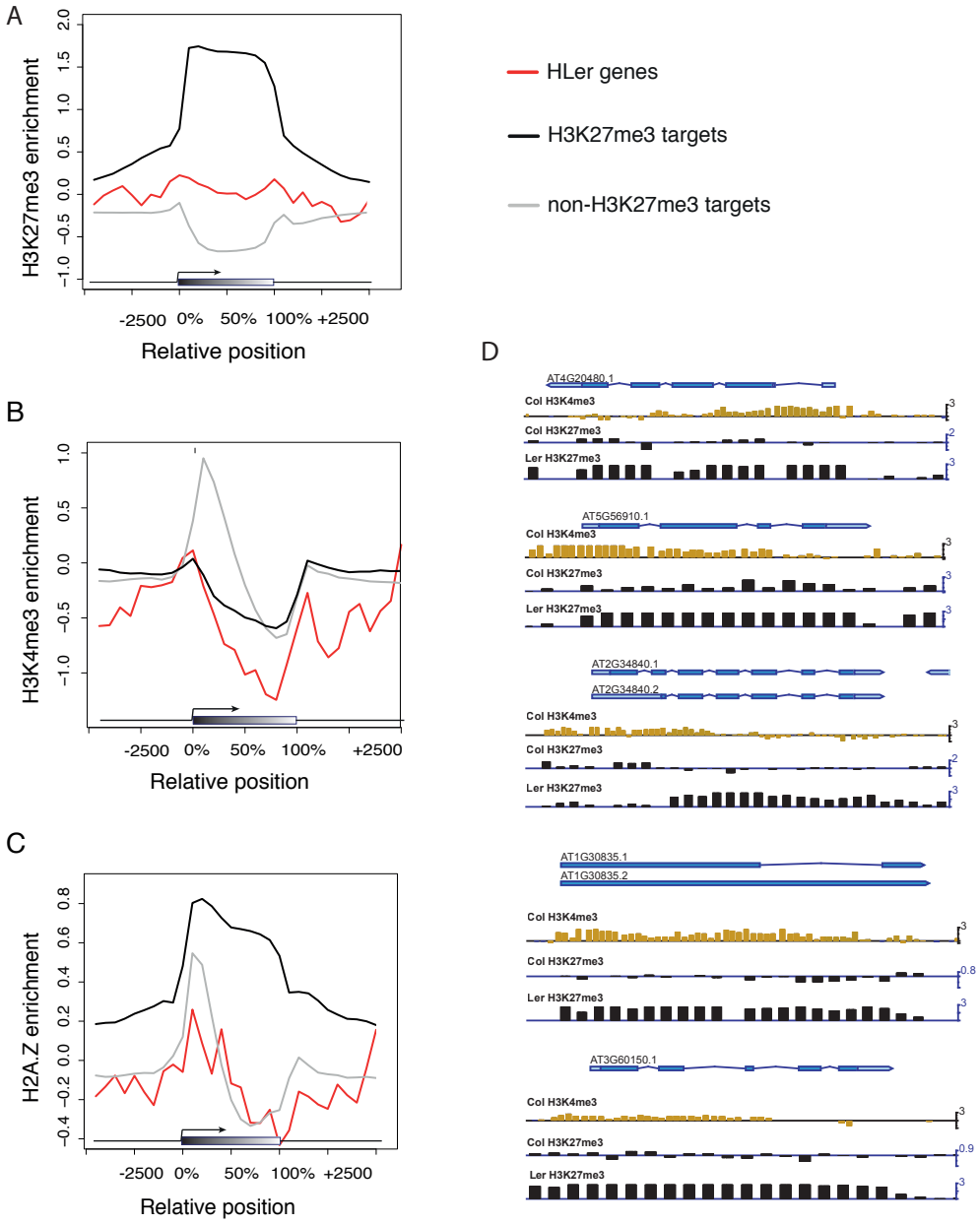


Figure 12. Enrichment of multiple chromatin features over genes in Col.
All the genes were aligned at 5' end. The gene body, 5-kb upstream and downstream of each gene were divided into10 bins, respectively. The enrichment of respective chromatin features was plotted for 10% length intervals along the gene body and for 500 bp sequence intervals for the 5-kb regions upstream and downstream of each gene. The red line traces HLer genes; the black line traces the H3K27me3 targets; the grey line traces genes not marked by H3K27me3. The x-axis shows the relative positions of the upstream, gene body and downstream of genes. The y-axis shows the corresponding mean signal of all genes in each bin. The mean signal of all probes mapped to that bin was taken for each gene in each bin. The grey bar represents the annotated gene body from transcription start (left) to transcription end (right). Arrows indicate the direction of transcription. (A) The enrichment of H3K27me3 over genes. (B) The enrichment of H3K4me3 over 3 lists of genes. (C) The enrichment of H2A.Z in 3 lists of genes. (D) The H3K4me3 signal in 5 HLer genes in the Col genome. The H3K4me3 intensity is shown in brown bars and H3K27me3 intensity is shown in black bars. The blue boxes with blue line connections in between represent the protein coding gene models for the 5 HLer genes (indicated above it). The track names above histone bars are the names of respective histone modifications.

Occupancy of repressive marks H3K9me2 and H3K27me at HLer genes in Col

Since most HLer genes show a low level of expression in Col although they do not have the repressive mark of H3K27me3, some other chromatin features might be the cause for the repression. Besides H3K27me3, there are several other repressive marks associated with gene repression, for example H3 dimethylation at Lys9 (H3K9me2), monomethylation at Lys27 (H3K27me1) and DNA methylation. In contrast to the euchromatic mark H3K27me3, these three marks are mainly located in constitutive heterochromatin. Therefore, the percentage of HLer being marked by these repressive modifications was calculated and compared with H3K27me3 targets and non-targets.

It has been shown that H3K9me2 is mutually exclusive with H3K27me3 in *Arabidopsis* (Turck et al. 2007). The H3K9me2 profiles over three gene lists (HLer, H3K27me3 targets and non-targets) were investigated. The lists of genes were aligned and partitioned as mentioned in **Figure 12**. For each bin, the proportion of genes being overlapped with H3K9me2 positive regions in their list was calculated and plotted in **Figure 13**A. The result shows that H3K9me2 frequently marks HLer genes, both in gene bodies and their surrounding regions. While H3K9me2 marks H3K27me3 targets and non-targets at a much lower level, especially within gene bodies. The same tendency was observed based on the H3K9me2 data from another group **Figure 13**E.

H3K27me1 is also a repressive histone modification. It has been proposed to be one pathway controlling constitutive heterochromatin formation in parallel with the H3K9me2 pathway (C. Liu, Lu, et al. 2010).The H3K27me1 enriched region was defined with SICER (Zang et al. 2009). The same method as described above was applied to H3K27me1 data respective to the three gene lists. The proportion of genes being marked by H3K27me1 in each bin is shown in **Figure 13**B. So HLer genes are more frequently marked by H3K27me1 when compared to H3K27me3 targets and non-targets.

DNA methylation is another mark associated with repressed genes. HLer genes are frequently marked by DNA methylation (**Figure 13**C). HLer genes are also frequently flanked by 24 nt small RNAs (**Figure 13**D), which can mediate DNA methylation in a RNA-directed manner.

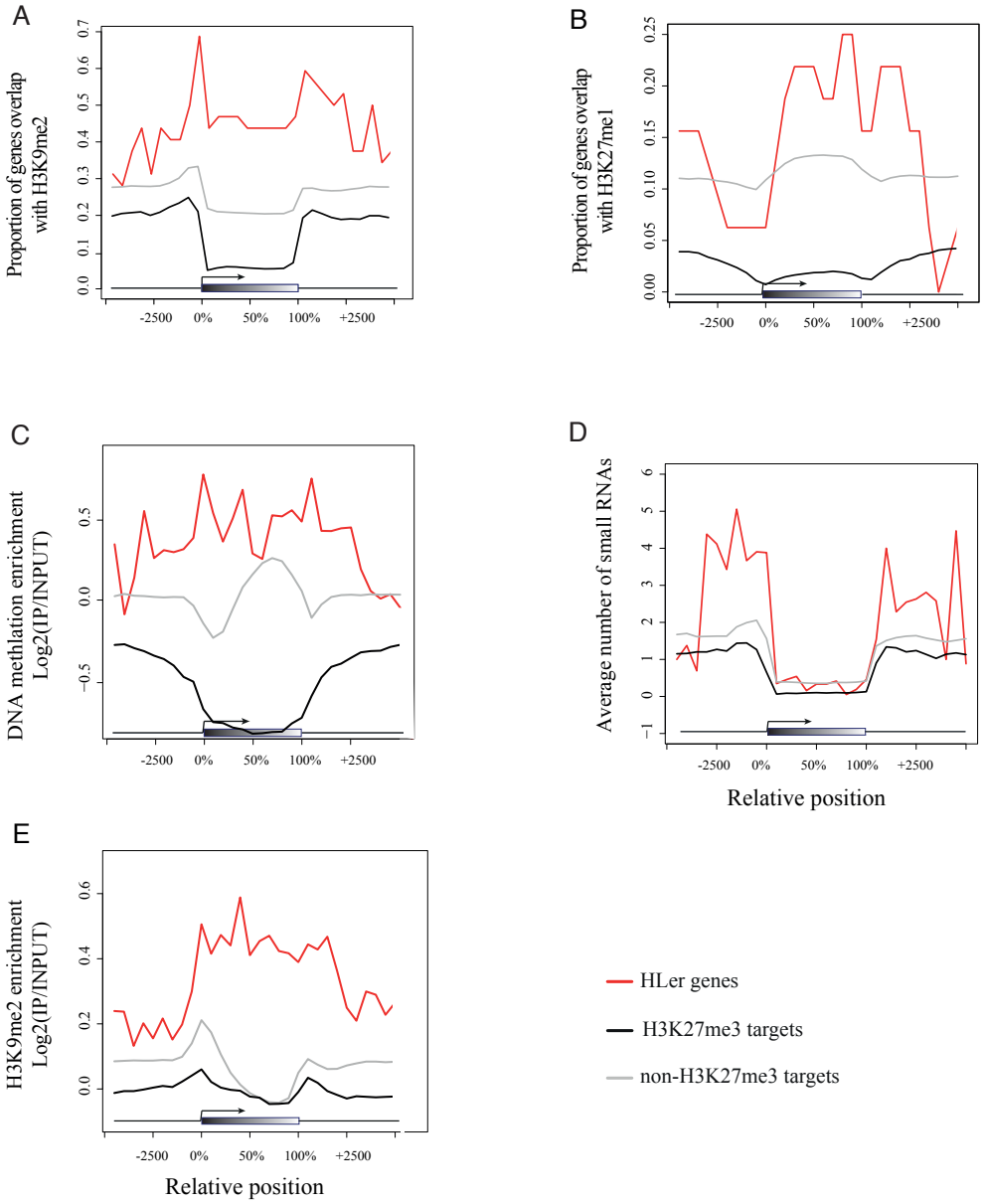


Figure 13. HLer genes are more frequently marked by repressive chromatin marks.
Genes were aligned at 5' end. The gene body, 5-kb upstream and downstream of each gene were divided into 30 bins. The percentage overlapping with histone-modification positive regions was plotted for 10% length intervals along the gene body and for 500 bp sequence intervals for the 5-kb regions upstream and downstream of each gene. The red line traces HLer genes; the black line marks the H3K27me3 targets; the grey line indicates genes not marked by H3K27me3. The x-axis shows the relative position in the upstream region, gene body and downstream region of genes. The y-axis represents the percentage of genes marked by given mark at each bin (A/B) or enrichment of respective chromatin modifications (C/E). The grey bar represents the annotated gene body from transcription start (left) to transcription end (right). (**A**) The proportion of genes overlapped with H3K9me2 marked regions. H3K9me2 data are from Jacobsen lab (Bernatavichute et al. 2008). (**B**) The proportion of genes overlapping with H3K27me1 marked regions (Jacob et al. 2009). (**C**) The enrichment of DNA methylation over 3 lists of genes (Zilberman et al. 2007). (**D**) Association of HLer genes with small RNAs. The y-axis shows the average number of 24nt small RNAs over genes. (**E**) The enrichment of H3K9me2 signal over 3 lists of genes. H3K9me2 data are from Hennig lab (Rehrauer et al. 2010).

Taken together, HLer gene are frequently marked by repressive H3K9me2, H3K27me1, DNA methylation and associated with flanking small interfering RNA. Small interfering RNAs (siRNAs) are known to cause RNA-directed DNA methylation, which can reinforce the formation of H3K9me2 in *Arabidopsis* (Chan et al. 2004; Xiaoyu Zhang et al. 2006). Whereas H3K9me2 and H3K27me3 are mutually exclusive(Turck et al. 2007). The association of HLer with repressive marks, especially H3K9me2, can largely explain the absence of H3K27me3 and their silence in expression in Col.

4.3.3 HLer genes are often neighbored by transposable elements

It is well known that H3K9me2 and H3K27me1 are mainly targeted at transposable element (TE) (Bernatavichute et al. 2008). Since HLer genes are also frequently marked by such repressive marks, it is possible that HLer genes are somehow related to TE. transposable element gene (TEG) is a gene encoded within a transposable element for example helicase, transposase etc. The definition of TE and TEG see file Readme-transposon at: (ftp://ftp.*Arabidopsis*.org/home/tair/Genes/ TAIR8_genome_release). To evaluate the association of HLer with TEG, first, the amount of TEGs in HLer genes was calculated. The result shows that there are four TEGs and three pseudogene in HLer. The rest of HLer are protein-coding genes and most of them have no precise gene annotation. So the composition of HLer cannot explain why H3K9me2 so frequently targets HLer genes.

It has been observed that the H3K9me2 recruited by TE can spread into nearby genes when the boundary sequences are absent (Ma et al. 2011). Next, I tested the percentage of HLer genes being flanked by TE and TEG respectively. The result shows that TEs and TEGs are preferred neighbors of HLer genes (**Figure 14**A and **Figure 15**A). The association of HLer genes with flanking TE and H3K9me2 is shown in **Figure 14**B. HLer genes, the H3K27me3 targets in Ler, frequently became the targets of heterochromatin mark H3K9me2 in Col. 19 of HLer genes are marked by H3K9me2 and 23 are flanked by an annotated TE. A big proportion of HLer genes (15 of 32) are associated with both. A summary of the flanking TE and other genomic features associated with HLer genes is shown in Table S1.

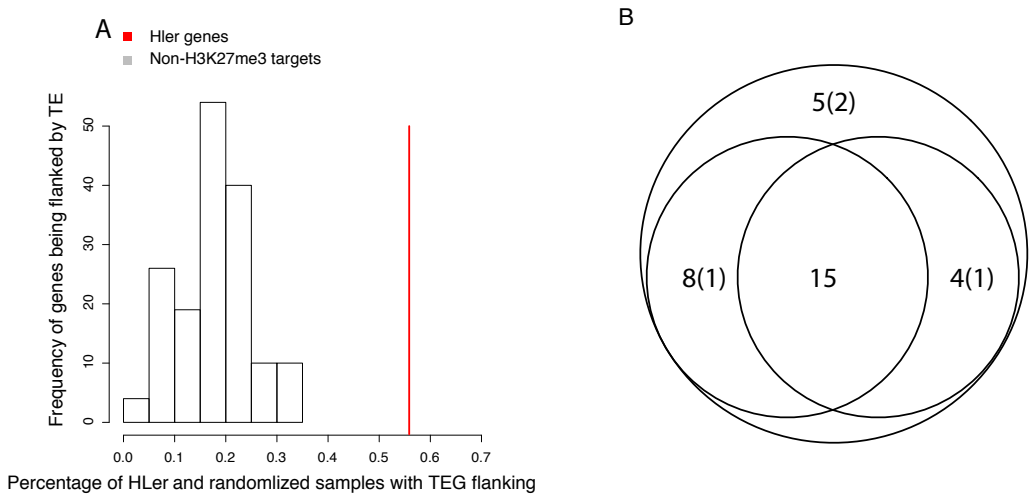


Figure 14. HLer genes are more frequenly flanked by TEGs than non-targtes and marked by H3K9me2 in the Col genome.
(A) HLer genes preferentially flanked by TEGs. The red line indicates the percentage of HLer genes being flanked by TEG. The histogram shows the distribution of the percentages from 100 times permutation in non-targets. (B) The Venn diagram shows the association of HLer genes with flanking TE and H3K9me2 mark. 23 of HLer genes are flanked by TE, 19 of HLer genes are marked by H3K9me2. 15 of HLer genes are associated with both. The number in brackets shows the number of transposable element genes in respective category.

The percentage of HLer genes being flanked by TE is 0.75. The value is significantly higher than the control (non-targets and H3K27me3 targets) (permutation test, $p<0.01$). Moreover, interestingly, H3K27me3 targets in general are more likely to be flanked by TE than non-targets genes (**Figure 15**). This observation was reproducible with sample size 28 and 1000 during permutation respectively (**Figure 15**). Taking together, HLer genes are preferentially marked by H3K9me2 and flanked by TEG or TE. But the association with TE is different with non-targets genes, but similar as H3K27me3 target genes.

Although we found that, TEs are generally more likely to neighbor H3K27me3 targets than non-targets, this observation does not confident enough to conclude that TEs play a role in the establishment of H3K27me3. The family of TE flanking H3K27me3 targets was explored, but no one family was overrepresented (Ulrike, personal communication). TE family of newly inserted TE in Col compared to Ler was also examined, no TE family was overrepresented either (data not shown). Inserted TEs in a genome are generally silenced via DNA methylation or other repressive histone modifications (Teixeira et al. 2009). It was reported previously that the methylated TEs have the deleterious effects on

the expression of neighboring genes and could be preferentially removed from gene-dense regions over time (Hollister & Gaut 2009). Now that non-targets of H3K27me3 include numerous highly active genes, the observed pattern of less TE neighboring them could be a result of purifying selection.

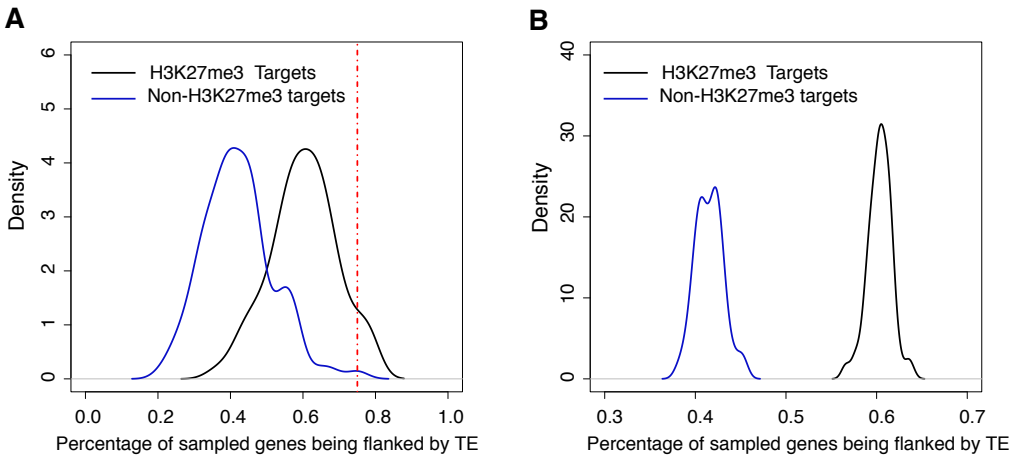


Figure 15. HLer genes are more frequently flanked by TE than non-targets in the Col genome.
(A) The density of percentage of genes flanked by TE in permutation with sample size 28. 28 non-TE genes were randomly sampled form H3K27me3 targets or non-targets for 100 times. The percentage of being flanked by TE in 28 non-TE genes for 100 samplings was calculated and shown in x-axis. The density of the percentage is shown in y-axis. The x-axis shows the percentage of genes flanked by TE for each time of sampling. The blue line indicates the distribution of the percentage derived from non-targets whereas the black line from H3K27me3 targets. The dashed red line shows the percentage being flanked by TE for the 28 non-TE genes in HLer. (B) The density of percentage of genes flanked by TE in permutation with sample size 1000. The bigger sample size reduce the percentage of getting a high percentage (above 0.5 for non-H3K27me3 targets; above 0.7 for H3K27me3 targets.) of genes flanked by TE.

4.3.4 HLer genes are not preferentially located in heterochromatic region

Since HLer genes are frequently marked by *repressive* chromatin *marks* typically in heterochromatic domain and frequently flanked by TE and TEG, it is possible that they are more physically located in heterochromatic regions such as centromeres and pericentromeres. To test this hypothesis, the chromosomal distribution of HLer together with H3K9me2 and TE density, which indicate the approximate location of centromeres and pericentromeres, was plotted (see **Figure 16**). According to the results shown in the

Figure 16, HLer genes do not preferentially locate in centromeres or pericentromeres but mostly in euchromatic regions. There is no clear association between HLer and the SNP density between Col and Ler.

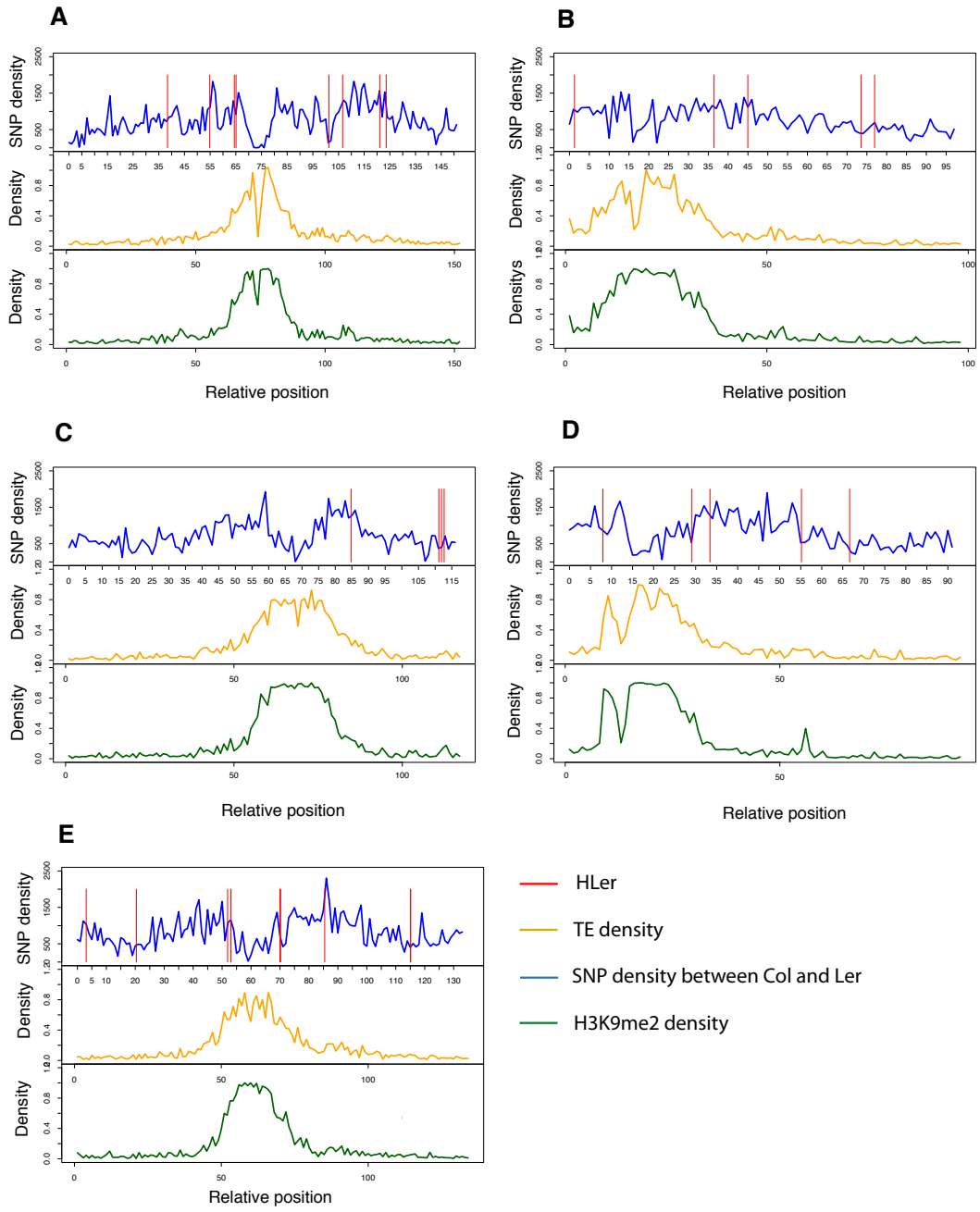


Figure 16. HLer genes are not preferentially located in heterochromatin regions.
(A) The distributions of HLer genes relative to heterochromatic regions in Chromosome 1 of the Col genome. (B), (C), (D), (E) are corresponding distribution in Chr2, Chr3, Chr4, Chr5. The red lines indicate the locations of HLer genes; the blue lines indicate the SNP density between Col and Ler; the brown lines indicate the density of TE; the green lines indicate the density of nucleotides marked by H3K9me2.

4.3.5 TE flanking HLer genes are often missing in Ler genome.

It is well known that TEs can influence the genes in their vicinity in many ways (Feschotte 2008; Hollister et al. 2011). The polymorphisms between species, such as insertions or deletions in one accession but not the other, could influence xthe chromatin modification states of nearby homologous genes. HLer genes are marked by different histone modifications in Col and Ler. At the same time, HLer genes are frequently flanked by TEs in Col. So it is interesting to check whether the TEs flanking HLer genes are present in Ler genome. The presence or absence will be helpful to investigate whether TE has been involved in the differential H3K27 trimethylation between Col and Ler.

To examine the gene structure changes between HLer and their surrounding regions, first, a whole genome alignment between the Col reference genome and the Ler draft genome was carried out with the tool MUMmer (Kurtz et al. 2004). Then, the sequence of HLer genes and their surrounding regions (5000bp up/down stream of gene body) were extracted and aligned with MUMmer again and then shown in the sequence comparison tool ACT (Carver et al. 2008). The gene annotation of the sequences extracted was also generated with custom scripts and shown in ACT(Carver et al. 2008) .

The sequence comparison shows that TEs flanking HLer are often missing in the Ler assembly (**Figure 17**). One example in **Figure 17**A shows that TEG AT5G35820 is missing in Ler while its flanking neighbors AT5G35810 (HLer gene) and AT5G35830 are highly similar between the two accessions according to the sequences extracted from Col genome and Ler scaffolds. For all the HLer genes with TE flanking, the gene sequences were compared between Col and Ler and the results were summarized in **Figure 17**B and Table S2. TE deletion in Ler was confirmed by PCR amplification with primers as shown in **Figure 17**A (Julia Reimer, personal communication). With this approach, 5 cases were confirmed. So we could show that the TE flanking HLer are often missing in Ler with both computational analysis and wet lab experiments.

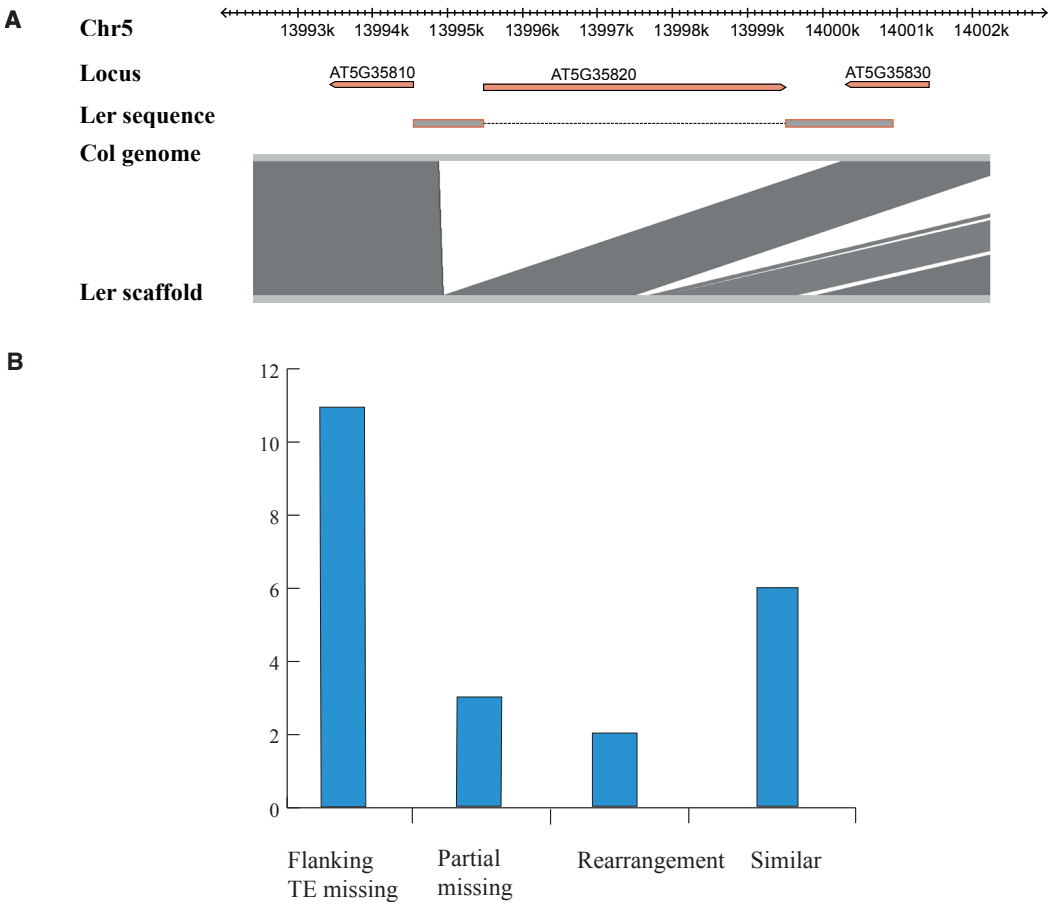


Figure 17. TE flanking HLer genes are often missing in Ler scaffolds.
(A) An example HLer gene with its flanking TEG missing in Ler. The scale shows the sequence coordinates in the Col genome. The red bars show the gene model of HLer AT5G38510, the TEG AT5G38520 flanking it and the following protein-coding gene AT5G35830. The short grey bars with red frame show the sequences in both accessions according to the sequencing of PCR products; the dashed grey line between short grey bars shows the sequence that is present in Col but missing in Ler according to the sequencing of PCR products (Julia Reimer, personal communication). The long grey bar flanking the grey and white regions shows the genome of Col and Ler. The grey regions between long grey bars show the sequences that can be aligned between Col genome and Ler scaffold; the white regions show the sequences that cannot be aligned.
(B) Number of HLer with flanking TE in Col but missing in Ler scaffolds. The HLer genes were classified into 4 groups based on the sequence comparison between Col and Ler. The names of groups are listed in the x-axis. The amount of genes in each group is shown in the y-axis. Group names indicate the sequence comparison results. 'Flanking TE missing' means the TE flanking HLer genes are missing in Ler assembly. 'Partial missing' means the HLer genes themselves are partially missing in Ler assembly. 'Rearrangement' means there is gene rearrangement neighboring HLer genes. 'Similar' means HLer genes in the two accessions are similar

4.3.6 Spreading of H3K9me2 from inserted TE to nearby genes in Col

As mentioned, H3K9me2 mark can spread from their targeting site into flanking regions (Talbert & S. Henikoff 2006; Locke & Martienssen 2006). Based on the observation that HLer are frequently flanked by TE and marked by H3K9me2, we proposed a model to explain the specific loss of H3K27me3 in Col (see Discussion). In this model, the inserted TE in Col could recruit the H3K9me2 mark, which in some cases can spread to nearby HLer genes. Due to the conflict of H3K9me2 and H3K27me3, HLer genes carry H3K9me2 but not H3K27me3 in Col and are repressed in expression. But in Ler genome, there is no TE insertion in corresponding regions or the insertion of TE did not spread H3K9me2 to nearby gene, and so HLer genes still carry the H3K27me3 mark. This model can explain the majority of the specific loss of H3K27me3 in Col. If the model is true, the protein coding gene that neighbor an inserted TE should more often carry H3K9me2 mark. To test this, I checked the occurrence of the protein-coding genes, which flank inserted TE in Col, being marked by H3K9me2. To do this, the polymorphic regions from 1001 project between Col and Ler were downloaded (Schneeberger et al. 2011). The inserted regions in Col compared to Ler were mapped to TE. The TE with at least 50bp overlap with the inserted regions was considered as inserted TE in Col but not Ler. For the 968 inserted TE, 965 flanking protein-coding genes without TE insertion directly within the gene were identified in Col. Among the 965 protein-coding genes that neighbor inserted TE, 89 are labeled with H3K9me2. For the 24495 protein-coding genes in TAIR9 without TE insertion within the gene body, 1080 of them are H3K9me2 targets. So the protein-coding genes flanking inserted TE in Col are more likely marked by H3K9me2 than whole genome level (Hypergeometric test, p=2.07848e-11). The TE family distribution of these inserted TE was shown in **Figure** 18. Compared with the distribution of all TEs in the Col genome, Copia family is overrepresented in these inserted TE when whose neighbor is H3K9me2 target. From this analysis we inferred that the high occurrence of H3K9me2 on these protein-coding genes could be the consequence of H3K9me2 spreading from nearby inserted TE in the Col genome. The TE family Copia might have stronger ability to spread their H3K9me2 to nearby region. The same H3K9me2 spreading event could have happened to HLer genes from nearby TE inserted in Col.

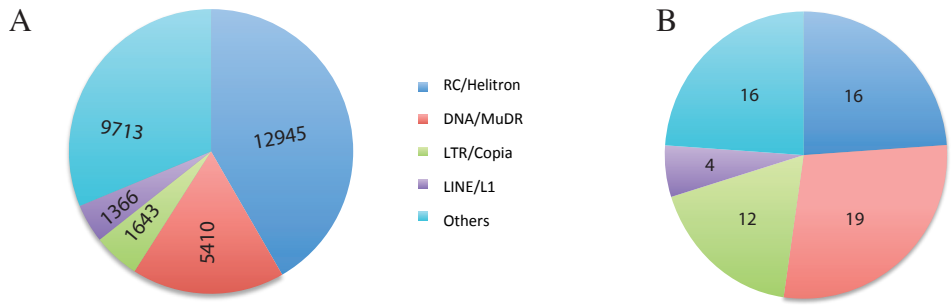


Figure 18. LTR/Copia family of TEs is overrepresented in the inserted TE in Col
(A) The TE family distribution of all TE in the Col genome. (B). The TE family of inserted TE when whose neighbor is H3K9me2 target in the Col genome. The LTR/Copia family is overrepresented compared with family distribution of all TEs in Col shown in (A).

4.4 Parental inheritance of H3K27me3

Genome-wide profiling of H3K27me3 targets in *Arabidopsis* has been carried out by several different groups. But how H3K27me3 is inherited is largely unknown. It has been implicated that H3K27me3 will be reset from one generation to the next (Ingouff et al. 2007; Feng et al. 2010). It is interesting to know how the establishment of H3K27me3 is regulated during the resetting. Now we have identified the genes uniquely methylated in Ler but not in Col. We can make use of F1 hybrids of Col and Ler to study the regulation of H3K27me3.

To study if the H3K27me3 mark is inherited in a *cis*-or *trans*-regulated manner, H3K27me3 profiling in F1 was carried out using ChIP-Seq technique. The ChIP experiments were done for whole 10-day-old seedlings from the F1 generation of reciprocal hybrids between Col and Ler (Col x Ler and Ler x Col) using the antibody against H3K27me3. The received reads after sequencing from two lanes for each biological sample were merged and then mapped to TAIR9 reference genome with the aligner BWA(H. Li & Durbin 2009) allowing maximally 3 mismatches, including maximally 1 gap. The data was further processed with SAMtools (H. Li et al. 2009) and Picard (http://picard.sourceforge.net/index.shtml) to pick out reads that mapped specifically and non-redundantly to the Col genome. 26,881,774 and 19,268,376 reads for F1:ColxLer and F1:LerxCol were mapped to Col TAIR9 reference genome to unique positions. After cleaning redundant reads frequently mapped to the same position in Col genome, 1,173,583 and 886,048 specific reads were kept for the two F1 hybrids for

further peaking calling, indicating an artifact caused by PCR amplification during sequencing library preparation. After excluding redundant reads, the sample from F1:Col x Ler has more high quality reads than the sample from F1:Ler x Col.

The H3K27me3 enriched regions in F1 were identified using SICER (Zang et al. 2009). The identified H3K27me3 enriched regions were mapped to TAIR9 gene annotation to identify the H3K27me3 targets in the F1 generation. 6648 and 6170 H3K27me3 targets were identified in F1 of Col x Ler and Ler x Col, respectively. 5586 of them were common between the two F1 hybrids. 5420 of 6370 H3K27me3 targets in Col were remains in the F1: Col x Ler. The saturation analysis showed that the coverage of reads from F1: Col x Ler was high enough for reliably identify H3K27me3 targets (Figure S2). So the amount of H3K27me3 targets identified were highly overlapping between the two F1 hybrids and between the parents and the F1 hybrids. This overlap is similar to that between targets in Col and Ler identified from ChIP-chip data (see **Figure 9**).

The overlap analysis of targets in F1 and both parents shows that almost all the common targets in parents and HCol genes are still H3K27me3 targets in F1 (**Figure 19**). In contrast, for HLer genes, about half were not detected as H3K27me3 targets according to the ChIP-Seq data, indicating that the mark could have been lost through a trans-regulated mechanism (**Figure 19**). To test whether these HLer genes really lost H3K27me3 in both alleles, we randomly chose four of them to verify their H3K27me3 state in F1 using ChIP-PCR. For the loci we tested, all of them are still H3K27me3 positive in F1 in both crosses between Col and Ler (see discussion, Julia Reimer, personal communication). So we see no evidence of difference in H3K27me3 signal between parents and both hybrids.

4.4.1 Allele-specific H3K27me3 in F1

For the genes still H3K27 trimethylated in F1 according to ChIP-Seq data, I calculated the allele frequency to infer which alleles of the genes are H3K27 trimehtylated in F1. Allele frequency means the reads originated from the Col allele relative to reads from both alleles. I used the available single nucleotide polymorphism (SNP) data in all enriched loci to evaluate which parental allele the reads originated from (Schneeberger et al. 2011). To ensure accuracy and reliability, only SNPs with more than three reads mapped were included in further analyses.

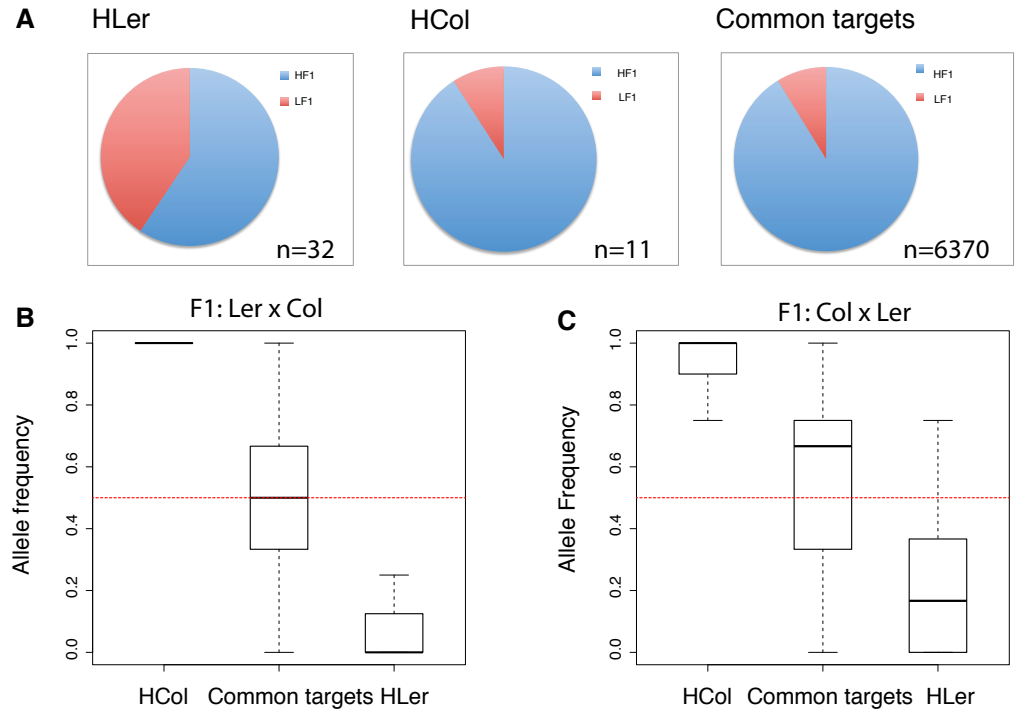


Figure 19. The parental inheritance of H3K27me3 targets.
(A) Pie chart showing the inheritance of H3K27me3 in hybrids of Col and Ler. The red sections represent the proportion of genes that lost H3K27me3 in hybrids, while the blue sections represent the proportion of genes keeping H3K27me3 in hybrids. The value for n indicates the number of genes in each gene list (HLer genes, HCol genes, common H3K27me3 targets between Col and Ler). (B) and (C), Allele-specific H3K27me3 in reciprocal hybrids of Col and Ler. The distribution of allele frequency in (B) was based on ChIP-Seq data from the F1 of Ler x Col. 22 SNPs for 12 HLer genes and 5 SNPs for 2 HCol genes were used for calculation of allele frequency. The allele frequency of (C) was based on ChIP-Seq data from the F1 of Col x Ler. 35 SNPs for 15 HLer genes and 11 SNPS for 2 HCol genes were used for calculation of allele frequency. The genome wide SNP list used for (B) and (C) is from 1001 projects (Schneeberger et al. 2011). The red dashed lines show the value of 0.5, which is the allele frequency value of heterozygous genes in case both alleles are equally H3K27 tri-methylated.

The distributions of the allele frequency of HCol, HLer and common H3K27me3 targets in reciprocal hybrids are shown as boxplot in **Figure 19**B and in **Figure 19**C. In both hybrids, the median allele frequency for HLer genes is 0, which means the reads covering SNPs are from the Ler allele; the median of allele frequency for HCol is 1, which means the reads covering SNPs are from the Col allele; The allele frequency for the common targets tends to be 0.5, which means the reads covering SNPs are from both alleles. The allele frequencies of random samples from common targets were significantly different from what was observed in HLer and HCol (permutation test, p value < 0.001). So the H3K27me3 mark showed a strict allele-specific H3K27me3 in reciprocal hybrids,

indicating a *cis*-regulated H3K27me3 deposition. Additionally, no significant parent-of-origin effect in the H3K27me3 modification of parental alleles in hybrids was detected.

4.5 A customized GBrowse instance for integrative data analysis

To visualize multiple genomic features along Col reference sequences, I maintained a local instance of GBrowse. Genomic data produced in our lab in this project and published data of interest, which was generated by other labs and available from public databases, were analyzed and uploaded into the database for the customized GBrowse instance. **Table 8** and **Table 9** show the summary of data chosen published data from other groups and generated in this project, respectively. A region containing two HLer genes and the genomic features associated with them are shown in **Figure 20**. The Gbrowse instance can be used as a public resource for convenient visualization of multiple genomic data and hypothesis generation. It will be of particular interest to the biologists working in the field of chromatin and epigenetics.

Table 8. Summary of selected genomic data integrated in local GBrowse from published data

Chromatin feature	Lab	Technique	Platform	Genotype	Reference
H3K27me3	Turck	ChIP-chip	Chromosome 4 tiling microarray	Col, Chr4	(Turck et al. 2007)
H3K9me2	Turck	ChIP-chip	Chromosome 4 tiling microarray	Col, Chr4	(Turck et al. 2007)
H3K9me3	Turck	ChIP-chip	Chromosome 4 tiling microarray	Col, Chr4	(Turck et al. 2007)
H3K27me3	Turck	ChIP-chip	Chromosome 4 tiling microarray	Col, *lhp1*, Chr4	(Turck et al. 2007)
H3K9me2	Jacobsen	ChIP-chip	Chromosome 4 tiling microarray	Col	(Bernatavichute et al. 2008)
H3K27me1	Jacobsen	ChIP-Seq	Illunima GAII sequencer	Col	(Jacob et al. 2010)
Nucleosome position	Jacobsen/Pellegrini	Chip-Seq	Illumina GAII sequencer	Col	(Chodavarapu et al. 2010)
H3K9me2	Hennig	ChIP-chip	Affymetrix AGRONOMICS1	Col	(Rehrauer et al. 2010)
H3K4me3	Van Nocker	ChIP-chip	Affymetrix Genechip Tiling 1.0R Array	Col	(Oh et al. 2008)
H3K27me2	Van Nocker	ChIP-chip	Affymetrix Genechip Tiling 1.0R Array	Col	(Oh et al. 2008)
H3K27me3	Van Nocker	ChIP-chip	Affymetrix Genechip Tiling 1.0R Array	Col	(Oh et al. 2008)
H3K36me2	Van Nocker	ChIP-chip	Affymetrix Genechip Tiling 1.0R Array	Col	(Oh et al. 2008)
H2AZ	Henikoff	Chip-chip	NimbleGen	Col	(Zilberman et al. 2008)
DNA methylation	Henikoff	Chip-chip	NimbleGen	Col	(Zilberman et al. 2007)
Small RNA	ASRP*	Multiple	Multiple	Col	(Gustafson et al. 2005)
sRNA in buds	Gregory/Ecker	Sequencing	Illumina GA	Col	(Backman et al. 2008)
sRNA in immature flower	Lister/Ecker	Sequencing	Illumina GA	Col	(Gregory et al. 2008)

*ASRP: Small RNA Project (http://asrp.cgrb.oregonstate.edu/db/download.html)

Table 9. Summary of genomic data generated in this project

Chromain feature	Technique	Platform	Genotype
H3K27me3	ChIP-chip	3-slide NimbleGen Arabidopsis Tiling Array Set	Col
H3K27me3	ChIP-chip	3-slide NimbleGen Arabidopsis Tiling Array Set	Ler
H3K27me3	ChIP-Seq	Illumina Solexa Sequencing	F1: Col x Ler
H3K27me3	ChIP- Seq	Illumina Solexa Sequencing	F1: Ler x Col
LHP1	ChIP-chip	3-slide NimbleGen Arabidopsis Tiling Array Set	Col
LHP1	ChIP-chip	3-slide NimbleGen Arabidopsis Tiling Array Set	Ler

A representative region showing two HLer genes and their associated genomic features are shown in **Figure 20** below.

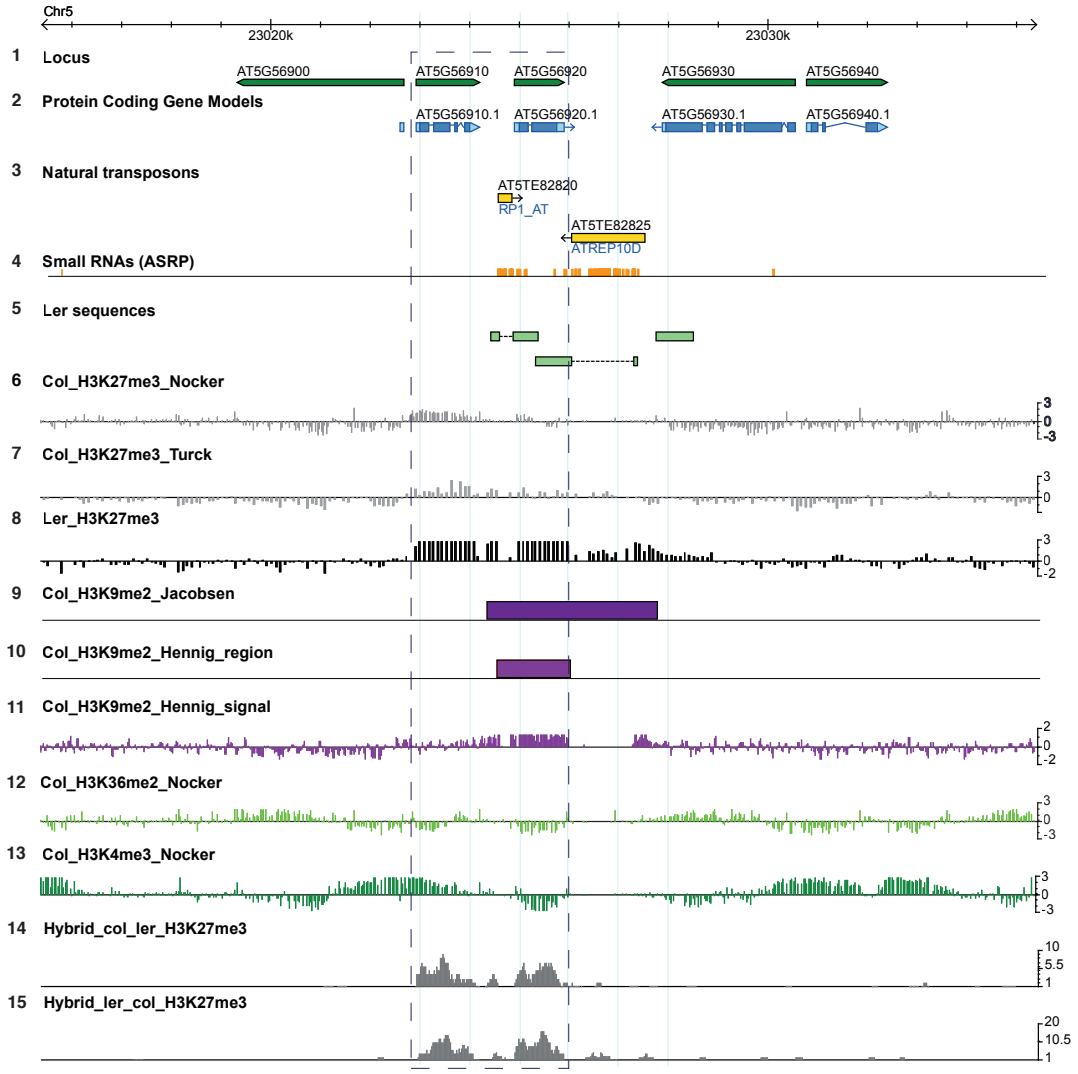


Figure 20. Visualization of multiple datasets in the same genomic region using locally maintained GBrowse.
The column of number on the left side is the track number; the text close to track name is the corresponding track name. The dashed gray box shows a region with differential H3K27me3 between Col and Ler. HLer gene AT5G56910 and AT5G56920 are shown in the dashed gray box. There are two TE in Col genome (Track 3); the two TE can generate small RNA in Col (Track 4); but the TE are not present in Ler genome (Track 5); according to H3K27me3 profiles generated in the lab of Nocker (Track6) and our lab (Track7), the two genes are not H3K27me3 targets in Col but in Ler (Track 8); AT5G56920 are heavily marked by H3K9me2 generated in two labs (Track 9,10,11); the Hler gene AT5G56920 is neither targets of active mark H3K36me2 (Track12) nor H3K4me3 (Track 13) in the Col genome, but AT5G56910 is a target of H3K4me3 (Track 13). The two genes are H3K27me3 targets in reciprocal hybrids of Col and Ler (Track 14,15).

5 Discussion

5.1 H3K27me3 targets in Col and Ler

To identify differential H3K27me3 targets between Col and Ler, we used chromatin immunoprecipitation (ChIP) coupled with high-density whole genome tiling arrays (ChIP-chip). We identified 6370 and 6344 H3K27me3 targets in Col and Ler, respectively. The targets identified in Col show a high degree of overlap (72-84%) with those identified by other groups, even though different conditions or tissues were used (Turck et al. 2007)(Xiaoyu Zhang et al. 2007; Moghaddam et al. 2011; Lafos et al. 2011) (see Table5). This could be expected, as majority of H3K27me3 targets are repressed in most tissues, for example, only a small amount of genes has a dynamic H3K27me3 regulation during differentiation and the embryo-to-seedling transition (Lafos et al. 2011; Bouyer et al. 2011), indicating that repressed state of H3K27me3 is likely the 'default state'. Nonetheless, the number of H3K27me3 targets identified by different groups ranges from about 5000 to 8000 (Xiaoyu Zhang et al. 2007; Lafos et al. 2011; Turck et al. 2007). The variance in the number of identified targets could be caused either by differences in the plant materials (including tissues or growth conditions), the assay platforms or the methods and thresholds used in the analysis. For example, Zhang et al (Xiaoyu Zhang et al. 2007) defines any gene as H3K27me3 target that overlaps with H3K27me3 positive regions, while Lafos et al (Lafos et al. 2011) took only those genes with at least 500bp overlap. We defined our own biologically motivated threshold by considering the length of gene (see 3.3.1).

The distributions of H3K27me3 in the two accessions are highly similar (see **Figure 9**). Although 918 and 892 genes are unique for the two accessions, respectively, according to the intersection analysis, this amount is similar to that of non-overlapped H3K27me3 targets between two biological replicates using Chip-chip technique (see **Table 5**a). It can be concluded that the majority of observed differences in H3K27me3 resulting from the intersection analysis was due to experimental conditions not to ecotype differences.

5.2 Differential H3K27me3 targets between Col and Ler

5.2.1 Workflow for DEGs identification

Since the same array was used for the two different *Arabidopsis* accessions, several issues should be considered during the identification of differential H3K27me3 targets because of the existence of sequence polymorphisms between accessions. The major sequence polymorphisms between Col and Ler include presence/absence polymorphisms, copy number differences and SNPs. Presence/absence polymorphisms could cause unique genes in one genome to be identified as differentially methylated ones; copy number differences cause cross hybridization problems; SNPs could cause low efficiency in hybridization and thus false identification of DEGs.

To overcome the limitations mentioned above, I designed a data analysis workflow including several filters and adjudgement to identify specific DEGs in this study. First, to solve the presence/absence problem, I remapped probes to Col genome and Ler scaffolds and only kept probes that mapped to both genomes. Next, to solve the copy number issue and only keep unique genes in the analysis, probes that mapped multiple positions were discarded. After excluding problematic probes, some genes had too few probes to judge their methylation state and had to be excluded. Thus, genes covered by less than four probes or less than 5 probes/Kb were also excluded. To exclude genes with low H3K27me3 signals in both accession but that showed quantitative differences, I filtered out the genes that have too little H3K27me3 signal to be called H3K27me3 targets in any accession. After applying the filtering procedure described above, the Bioconductor package *RankProd* was used to identify genes that are specifically H3K27me3 enriched in one genome but not the other.

Different numbers of mismatches during remapping were used for identification of HLer and HCol genes. Mismatches can help to exclude probes with high similarity and reduce cross hybridization. At the same time tolerance of a small number of mismatches can keep homologous genes with SNPs in Ler in the scope of data analysis. I tested several different values of mismatches and the resulting HLer genes lists were highly overlapping (data not shown). The list of HLer genes obtained with 2 mismatches was proved optimal in direct ChIP-PCR (see 4.2.3) and thus was used as final HLer gene list. Allowing mismatches can keep probes which do not have perfect matches with Ler DNA sequences in the analysis. During hybridization, these probes could have a low efficiency in

hybridization with Ler DNA sequences. Consequently, a lower signal in Ler sample could be generated for these probes even when their corresponding genes are H3K27 methylated as equally as in Col. so allowing mismatches can lead to false identification of HCol genes. Therefore, no mismatch was allowed during remapping for identification of HCol.

In other studies, to keep only conserved, single-copy genes in the genome-wide comparison data analysis, comparative genomic hybridization (CGH) has been often used as a first step to exclude genes with structural variations, such as presence/absence variation or copy number variation, from the analysis (Vaughn et al. 2007)(Eichten et al. 2011). This can help to detect conserved, single copy genes between different genomes but also could has some drawbacks depending on arrays used. For example, when using Col arrays for Ler samples, only decreases in copy number in the Ler genome can be detected, whereas an increase in copy number in Ler relative to Col or simple rearrangements can not be identified (Vaughn et al. 2007). In my workflow, by remapping probes to Col genome and Ler scaffolds and filter problematic probes, the same effect of excluding genes with structure variations was achieved as by CGH. But some limitations of the data analysis might exist that are mainly due to the incompleteness of the Ler assembly. First, remapping of probes to both genomes can exclude duplicated genes in Col but might not do so in Ler since the sequences for repetitive regions or duplicated genes are often absent in Ler assembly. Thus, some predicted DEGs could have multiple copies in the Ler genome. Second, remapping of probes to genomes could improve specificity but also reduce sensitivity. Some unique, high quality probes could have been filtered out by this procedure because they do not map to the current version of the Ler assembly. Additionally, since we are using arrays that were designed based on the Col genome to hybridize with DNA from Ler, some HLer genes might not be detected because of low hybridization efficiency caused by SNPs. Thus, the number of HLer genes we identified in this project could be underestimated. Nonetheless, the results of an independently performed ChIP-PCR for chosen HLer genes, confirmed the predicted HLer genes (see **Figure 21**).

Although the current Ler assembly is not complete, it allowed us to exclude genes with structure variations from differential H3K27me3 target prediction. The majority of consensus genes should be included, given that 96.3% of the sequences in Ler assembly can be mapped to Col genome and 77.8% of the Col genome sequence are covered by Ler

assembly according to the whole genome alignment result. The regions with largest gaps in the Ler assembly are highly repetitive sequences in the centromeric and peri-centromeric region (Schneeberger et al. 2011). The Ler assembly contains the majority of conserved, low copy number genes between Col and Ler and such genes are of principal interest in this project. By using the probes mapped to both the Col genome and the Ler assembly, we generated a list of HLer genes for further analysis. As mentioned above, HLer genes were well confirmed via an independently performed ChIP-PCR (Julia Reime, personal communication) (see **Figure 21**). We take this to indicate that our workflow reliably identified HLer genes. But it is not the case for HCol genes, see below.

5.2.2 DEGs identification and comparison

By analyzing the ChIP-chip data generated in our lab, only a very small amount of differentially H3K27me3 enriched genes were identified between Col and Ler (see **Table 7** and **Table 6**). The number of DEGs is much smaller than that identified between Col and Cvi or Col and C24 in a recent study (Moghaddam et al. 2011). In previous studies, the extent of divergence in genomic sequences, DNA methylation and histone modifications between accessions has been explored (Schmid et al. 2003; Kliebenstein et al. 2006; Schneeberger et al. 2011; Vaughn et al. 2007) (Cokus et al. 2008; Banaei Moghaddam et al. 2010). The more divergent the gene sequences, the more epigenetic polymorphisms and differences in gene expression have been detected among accessions. So the differences in the amount of DEGs compared to Col likely is due to the different accessions used. Indeed, the smaller number of differential H3K27me3 between Col and Ler is consistent with the closer genetic relationship of the two accessions (Clark et al. 2007) .

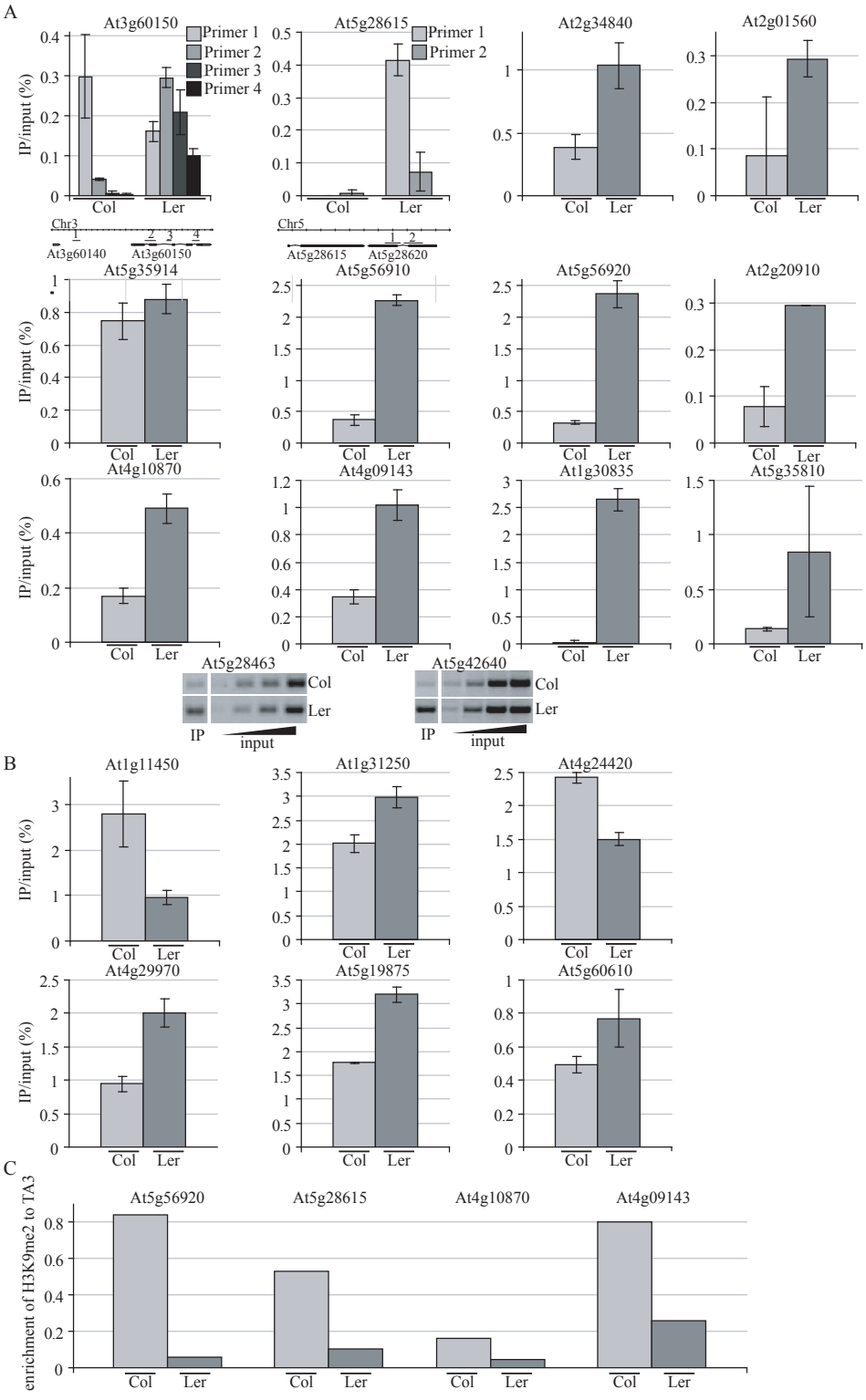
Besides the differences between the studied accessions, also differences in the methods used for detecting H3K27me3 polymorphisms among accessions could account for the number of DEGs identified. The intersection analysis method used by A. Moghaddam *et al.* simply defined the targets outside of an intersection as uniquely enriched genes in certain accession (Moghaddam et al. 2011). Using the same intersection analysis with our data, a similar number of DEGs could be identified as reported in C24 and Cvi (Moghaddam et al. 2011). However, the DEGs identified using this method could not be confirmed by independent experiments (Julia Reimer, personal communication).

Therefore we used a more stringent method to balance the specificity and sensitivity in DEGs identification.

5.2.3 Experimental validation of HLer and HCol genes

The HLer genes identified using my workflow can be well confirmed by ChIP-PCR (**Figure 21**, Julia Reimer, personal communication). Using ChIP-PCR or gel. 14 out of 15 randomly selected HLer genes were confirmed in 32 HLer genes. The gene AT5G35914 could not be validated although the signal in Ler is slightly higher than in Col. The occupancy of H3K9me2 at chosen HLer genes in Col but not Ler was also well validated (**Figure 21**C). In contrast, for the HCol genes the results of the independently performed ChIP-PCR for H3K27me3 were often inconsistent with the predictions based on the ChIP-chip data. According to the real-time ChIP-PCR results, two of six HCol genes show H3K27me3 modification high in Col but low in Ler, whereas the other four were H3K27me3 modified both in Ler and Col even showed slightly higher signal in Ler than in Col. It is noteworthy that two of the unconfirmed HCol genes, AT4G29770 and AT1G3125, show Col allele-specific H3K27me3 in F1 of Col x Ler and Ler x Col, respectively. This is a hint that they might be real HCol genes if there is only *cis*-effect (see 5.6.2). Thus these two genes should be considered as false negatives of real-PCR analysis if they are real HCol genes.

One cause for the potential false positives in HCol genes could be, that the difference in H3K27me3 is too small to be detected by PCR (fold change 3-8). However, the fold change in HLer genes is similar. Another cause could be the low hybridization efficiency of the Ler sample that contains a certain level of sequence polymorphisms. This can lead to an underestimation of FC in HLer but overestimation in HCol since few SNPs are in the region where the probe binds could cause a large reduction in the observed microarray signal. A similar effect would appear if multiple copies of a gene exist in the Ler genome,



t

Figure 21. The confirmation of DEGs by real-time ChIP-PCR and gel.
(A) Real-time ChIP-PCR and gel confirmation of HLer genes. Only AT5G35914 shows similar H3K27me3 in both accessions. The higher H3K27me3 in Ler versus Col was confirmed for 13 HLer genes out of 14. Another HLer genes AT4G20480 was also confirmed but are not shown here. (B) ChIP-PCR confirmation of HCol genes. Only 2 out of 6 (AT1G11450 and AT4g24420) show high H3K27me3 signal in Col versus Ler. (C) The enrichment of H3K9me2 at chosen HLer genes in Col and Ler. (Figure was provided by Julia Reimer).

these genes were not assembled correctly in the Ler scaffolds and only one copy was incorporated. During remapping, the probes for this gene seem to be unique in both Col and Ler and thus were retained. As we know, in the ChIP-chip procedure, immunoprecipated DNA (with H3K27me3 mark) and INPUT sample (genome DNA) are labeled with cy5 or cy3 respectively, mixed and then hybridized with probes on arrays in a specific buffer. If this gene now has multiple copies in Ler, but only one copy is H3K27me3 marked, all the DNA fragments of this gene family, regardless if they are methylated or not, will compete to hybridize with the limited number of probes. Consequently, only a relatively smaller proportion of ChIP DNA relative to the INPUT could get the chance to hybridize with probes, leading to a low signal intensity for these probes and a subsequent false prediction of HCol. I checked the input signal for HCol genes in Col and Ler samples. Although the overall signal in Ler is not higher than in Col samples, it could be that the probes on arrays were saturated by Ler samples already. This is consistent with the claim of Vaughn *et al* that increased copy number in Ler relative to Col could not be identified by comparative genome hybridization (Vaughn et al. 2007). Additionally, it is generally challenging to detect quantitative changes of H3K27me3 using ChIP-chip between Col and Ler due to the global amplification procedure of DNA samples. Nevertheless, we must conclude that because of the limitation in technique used, detection of HCol genes was not confidential, so we excluded HCol for further analysis.

5.3 Expression of HLer genes is coordinated with their histone modifications

5.3.1 Expression of HLer genes

Previous publications have revealed that several thousands of genes show variable gene expression in the same tissues between or within *Arabidopsis* species. The variable expression is associated with sequence polymorphisms and differential histone modifications, such as H3K27me3 (Kliebenstein et al. 2006; Xu Zhang & Borevitz 2009; F. He et al. 2012). Since HLer genes show differences in H3K27me3 between Col and Ler, we further investigated whether these differences correspond to variation in their expression.

The expression of all HLer genes was explored firstly using transcriptomic data in Col from the At-TAX project and then transcriptomic data from19 *Arabidopsis* genomes (Gan

et al. 2011). In Col, the HLer genes were clustered into two groups, five Exp_Col that are actively expressed and 26 Rep_Col that are repressed (see **Figure 11**). Interestingly, the HLer genes in these two groups show almost the same active or repressed state in the seedlings of 19 *Arabidopsis* accessions (Gan et al. 2011).

To explore the expression pattern of HLer genes in Ler, we used RT-PCR and measured the transcription of nine HLer genes in different tissues in Col and Ler (

Figure 22, provided by Julia Reimer, personal communication). The nine HLer genes chosen here include five Exp_Col genes that are actively expressed, and four Rep_Col genes that are repressed in Col based on the At-TAX data (see **Figure 11**) (Laubinger et al. 2008). The RT-PCR results show, that in Ler Exp_Col genes also have variable expression at the time points tested in Ler. The four Rep_Col genes, on the other hand, are constantly repressed in both accessions at all time points analyzed. So, the expression pattern of the analyzed HLer genes in Ler also supports the notion that HLer genes can be categorized into two groups. This is consistent with the classification based on the expression from the other two data resources, expression data in At-TAX (Laubinger et al. 2008) and 19 genome projects (Gan et al. 2011).

Since HLer genes were identified in Ler by the H3K27me3 state in 10 days seedlings and H3K27me3 is a well-known repressive mark, it is not surprising that all the 9 HLer genes assayed here are not expressed in Ler in 7-day-old seedlings. However, Exp_Col genes become active again at later time points. For example, AT2G34840 and AT1G30835 are expressed slightly at day 27 in the stem of Ler. The expression pattern of these HLer genes in Ler is slightly different to that in Col. The variable and temporal expression pattern of Exp_Col in Ler is in accordance with the characteristics of a typical H3K27me3 target, which means it is highly tissue specific and reversible. Nonetheless, the four HLer genes chosen from Rep_Col do not show any expression in Ler. The required conditions for expression might not fulfilled.

It was speculated previously that the H3K27me3 polymorphisms observed between Col and C24/Cvi could correspond to differences in gene expression patterns (Moghaddam et al. 2011). Now we show that the expression of HLer genes, which are marked by H3K27me3 in Ler but not Col, cannot always be predicted by the state of H3K27me3 alone. We observed very distinct expression patterns of HLer gene in Col, Ler and seedlings of 19 accessions. The investigated expression data is highly consistent between different studies (except AT2G20190). It is also worthy to note that the HLer genes that

exhibit variable expression among accessions (19 accessions) also display variable expression at time points or tissues within the same accession (Col or Ler). This observation is consistent with the flexible and dynamic nature of the H3K27me3 they carry. In contrast, HLer genes in the other group are repressed in the 19 Arabidosis accessions explored irrespective of the differential H3K27me3 at these loci between Col and Ler.

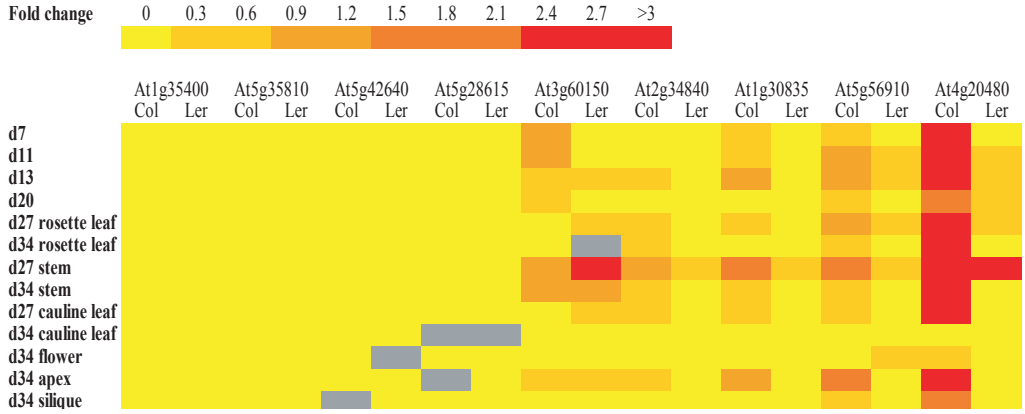


Figure 22. Expression of nine HLer genes in Col and Ler in different times or tissues.
Heat map depicting patterns of changing gene expression between different conditions. Red to yellow indicates active to repressed state of gene expression. Columns indicate expression changes of HLer genes at different time points in certain tissues. Four of the Rep_Col depicted here (AT1G35400, AT5G35810, AT5G42640, AT5G28615) do not express in either genome at any of the assayed time points. Exp_Col genes (AT3G60150, AT2G34840, AT1G30835, AT5G56910, AT4G20480) show variable expression between Col and Ler. The analyzed time points and tissues (left side of the heat map) are whole seedlings in day 7, day 11, day 13 and day 20; rosette leaf in day 27 and day34; stem in day 27 and day34; cauline leaf in day 27 and day34; flower in day 34; apex in day 34; silique in day 34. Grey indicates no detection. (Figure was provided by Julia Reimer, personal communication).

5.3.2 Association of HLer gene expression and their histone modifications

Consistent with their distinct expression patterns within or between accessions, coordinated chromatin modifications were observed at HLer genes. The majority of Rep_inCol HLer genes (except AT2G20910) are not expressed at any of the time points or in any of the tissues and accessions analyzed. Consistent with their low expression, subsequent analysis showed that these genes are associated with repressive chromatin marks, i.e. H3K9me3 and DNA methylation in Col and H3K27me3 in Ler. H3K9me2 and

H3K27me3 belong to two distinct repressive pathways maintained by different mechanisms. Both of them stably repress the activation of HLer genes in both accessions based on the time points we examined, although H3K27me3 is a reversible repressive mark in contrast to H3K9me2 (Lafos et al. 2011).

The five Exp_inCol HLer genes show relatively strong expression in Col, Ler and 17 other accessions. The expression pattern of these genes in Ler agrees with the tissue-specific-expression pattern of a typical H3K27me3 target as previously reported (Turck et al. 2007). In seedlings, the five HLer genes are modified differently between Col and Ler. In Col, they are expressed and marked by the active mark H3K4me3, while in Ler, they are repressed and marked by the repressive mark H3K27me3. So the distinct histone modifications on these five HLer genes are correlated with their expression pattern. Vice versa, the genes that show variable expression among the 19 accessions in seedlings (see **Figure 11**) could be associated with different histone modifications, either H3K27me3 or H3K4me3 but not H3K9me2. The latter modification could not occur in this case, as the genes marked with it seem not to be expressed at all, neither at the time points assayed for Col and Ler in this project nor in the studies previously performed (Bernatavichute et al. 2008) (Rehrauer et al. 2010). So for the HLer genes analyzed here, the expression pattern fits the expectation based on the chromatin state.

5.3.3 Expression of HLer genes and flanking TEs

The expression pattern of HLer genes could have been influenced by TEs that frequently are flanking them. TE sequences have been shown to be able to repress their adjacent genes' transcription through deposition of repressive chromatin modifications like H3K9me2. For instance, an insertion of a TE into an intron of *FLOWERING LOCUS C* (*FLC*) in Ler causes reduced expression of this locus, and consequently earlier flowering of Ler in comparison to Col (J. Liu et al. 2004). TEs also can modify the expression of neighboring genes in wheat, maize, and rice through disruption of native promoter regulation or introduction of new regulatory elements (Kashkush et al. 2003; Pooma et al. 2002; Huang et al. 2008). It has been shown that transposable elements and small RNAs can contribute to gene expression divergence between *Arabidopsis thaliana* and *A. lyrata* (Hollister et al. 2011; F. He et al. 2012).

The insertion of TE was reported to be able to not only repress but also activate nearby genes (Fernandez et al. 2010). Consistent with this, interestingly, not only the HLer genes that are not expressed in Col, but also the HLer genes that are expressed in Col can be adjacent to TE, for example, AT5G56910 and AT4G20480. This occurrence of both activation and inactivation suggest a possible dual role of TE on respective adjacent HLer genes.

5.4 Characteristics of HLer in Col

5.4.1 The chromatin modifications of HLer genes in Col

In Col, HLer genes are associated with distinct histone modifications compared to H3K27me3 targets or non-targets. Consistent with their expression, H3K4me3 is associated with the smaller group of genes (five Exp_Col) that is actively expressed **Figure 12**D, whereas H3K9me2 is associated with the other group of genes (Rep_Col) that is not expressed (see **Figure 13**A/E). The H3K9me2 data generated in two independent studies has been used to investigate the occupancy of H3K9me2 at gene bodies and surrounding regions of HLer genes in the Col genome. The same tendency was observed in the two data sets, that HLer genes were highly enriched in H3K9me2 (**Figure 13**), which is typically associated with heterochromatin in *Arabidopsis*. Besides, higher DNA methylation and more frequent H3K27me1 were also found over HLer regions and more small RNAs were mapped to the flanking region of HLer (see **Figure 13**). Small RNAs might directly contribute to the recruitment of DNA methylation and subsequent H3K9me2 establishment. The combined existence of these chromatin modifications indicates a rather condensed chromatin structure, which is inaccessible to the transcriptional machinery.

The combination of chromatin modifications observed at Rep_Col represents a typical chromatin state, CS3, a state that marks 83% of TEs in the Col genome. CS3 is a stable repression state for preventing the deleterious mobility of TEs or repeat elements (Roudier et al. 2011). While in the Ler genome, HLer genes are marked by H3K27me3, which is a typical repressive but reversible euchromatin mark for genes and characteristic for CS2, a state that marks 23% of genes. These genes are reversibly repressed (Roudier et al. 2011). CS2 and CS3 are antagonistic although both CS2 and CS3 are repressive chromatin states (Roudier et al. 2011). Additionally, five HLer genes (Exp_inCol) are

modified with H3K4me3, which is a typical mark for active genes in CS1. So HLer genes are associated with completely different chromatin states in Col and Ler. They are indexed with CS1 or CS3 in Col, but CS2 in Ler. The expression of HLer genes fits to the chromatin state they have in respective genomes.

However, the reversible expression of Rep_inCol was not detected by the experiments that were published or performed in our lab (personal communication, Julia Reimer). On the one hand, it could be that the expression is transient and thus not captured in our experiments or the condition for their expression is irregular and not fulfilled in current or previous studies (Laubinger et al. 2008; Gan et al. 2011). On the other hand, it could be that the HLer genes that are expressed in neither genome are surplus genes that are not essential during any stage of the life cycle. In Col these genes are completely silenced, while in Ler they are repressed but could be activated again. In this case, the repressive pathways of H3K27me3 and H3K9me2 are compensative to each other to maintain the silencing system in the plant between different accessions. This kind of compensative role is probably present within Col. For example, the majority of TEs are silenced by H3K9me2 and some by H3K27me3. In seedlings of Col, some genes (TEs and protein coding genes) were densely marked by DNA methylation (probably together with H3K9me2), but become specific H3K27me3 targets in endosperm (Weinhofer et al. 2010). These findings support the notion that the repression of the genes can be controlled by either DNA methylation/H3K9me2 or H3K27me3, the two independent, alternative repressive pathways in fine-tuning the transcript levels of specific target genes. The H3K27me3 targets that do not show active transcription in any phase of a typical plant life could be activated in special conditions. We have not observed reversible expression of Rep_Col HLer genes, so it is unknown if the difference in repressive histone modifications causes any ecological consequences.

5.4.2 TEs are more likely neighboring HLer genes in Col and missing in Ler genome

We found that HLer genes are more likely flanked by transposons in the Col genome, which might contribute to the generation of the small RNAs surrounding HLer genes. According to the description file for transposons released by TAIR8, a TEG is a TE that embeds a gene for example a helicase, transposase etc. The same tendency was observed for TE and TEG respectively (see **Figure 14** and **Figure 15**). I also randomly chose the

same number of non-H3K27me3 targets from the Col genome for 100 times and calculated their percentage of being flanked by TEs and HLer genes are much more often flanked by TEs than random selections (Permutation test, p = 0.01) (see **Figure 15**).

TEs are enriched in pericentromeric regions. The highly association of HLer genes with TEs can be explained if these HLer genes are preferentially located in pericentromeric regions. In fact, HLer genes are not preferentially enriched in pericentromeric regions according to the distribution of them along 5 chromosomes (see **Figure 16**). So the preferential association of TEs with HLer genes can not be explained simply by the location of HLer genes.

Table 10. The validation of missing TE in Ler by sequence comparison following Sanger sequencing

HLer gene	Missing TE in Ler	PCR-size in Col	PCR-size in Ler	Type of deleted TE	Duplicated or missing nucleotides
AT5G56910	AT5TE82820	2.071 kb	1.8 kb	RP1-AT	gca - in Ler missing
AT5G56920	AT5TE82825	3.286 kb	2 kb	AtREP10D	ATTAAGTAA - duplicated in Col
AT2G34840	AT2TE65230	2.839 kb	1.8 kb	AtMU1	atttg - duplicated in Col
AT5G42640	At5g42645	7.889 kb	3 kb	AT COPIA	ccgca - duplicated in Col
AT5G35810	At5g35820	6.568 kb	1 kb	AT COPIA	ATACCT - duplicated in Col

* Table was generated based on the confirmation results of Julia Reimer

However, TEs or TEGs are often missing in Ler scaffolds based on the whole genome alignment between Col genome and Ler scaffolds (see **Figure 17**). The missing of TE/TEG in Ler could be caused by the incompleteness of the Ler assembly. To test if the flanking TE of HLer genes are really often missing in Ler genome, PCR with specific primers was used to detect the presence of the flanking TE in Col and Ler, and five deletions of TEs in the Ler genome out of a random selection were confirmed by this test (see Table 10) (Julia Reimer, personal communication). So we have shown that the TEs flanking HLer genes are often missing in Ler with both genomic sequence analysis and specific PCR experiments.

5.5 Replacement of H3K27me3 by H3K9me2/H3K4me3 in Col at HLer genes

We observed three interesting patterns associated with HLer genes. First, most of the HLer genes are modified by H3K9me2 in Col but by H3K27me3 in Ler, and they are not expressed in Col. An exception are the five Exp_inCol genes which are expressed in Col but not in Ler and carry H3K4me2 in Col. Second, TEs are likely to neighbor HLer genes. Last, TEs flanking HLer genes in Col are often missing in the Ler genome. Based on these observations, we propose a model in which, during the evolution of Col, the insertion of TEs in Col recruits heterochromatin mark H3K9me2 in combination with DNA methylation, which further spreads to neighboring genes in certain situations, thereby replacing H3K27me3. This leads to the loss of H2K27me3 in some genes in Col compared to Ler and the subsequent identification as HLer genes in our study.

TEs are well known to be more polymorphic than genes in *Arabidopsis* or other species (Moghaddam et al. 2011; Springer et al. 2009). Via comparative genome hybridization (CGH) analysis of chromosome 4 Vaughn et al. showed that tiles corresponding to TEs and repeats were often missing in the Ler genome (Vaughn et al. 2007). The comparison between the Col genome and the Ler assembly based on re-sequencing data also showed that the deleted regions in the Ler assembly are significantly enriched for transposable elements (Schneeberger et al. 2011). Similarly, in the polymorphic regions between C24/Cvi and Col, TEs are also overrepresented (Moghaddam et al. 2011). TE insertions in a genome are often associated with reduced expression of nearby genes (Hollister & Gaut 2009) (Hollister et al. 2011). It was found that proximal TEs are associated with lower expression, especially when the TE is targeted by siRNA, which is crucial for the initiation and maintenance of DNA methylation (Hollister et al. 2011).

Considering the self-reinforcing loop between DNA methylation and H3K9me2, the insertion of TEs in Col could have recruited H3K9me2 as well, which occasionally can spread to nearby regions when a boundary element/signal/region for stopping the heterochromatic mark is absent. To test this possibility, I examined the H3K9me2 state of protein coding genes neighboring newly-inserted TEs in Col compared to Ler. Indeed, on the genome wide level, compared with genes that do not have newly-inserted TEs flanking in Col, the genes with flanking newly-inserted TEs are more likely marked by H3K9me2 (see 4.3.6), supporting the hypothesis that the spreading of H3K9me2 caused by TE insertion can lead to the absence of H3K27me3 at HLer loci in Col.

We propose this model while recognizing some limitations of the model. HLer genes are not always flanked by TEs and the corresponding TEs in Col are not always missing in Ler. So the presence/absence of TEs contributes to the differential H3K27me3 modification at HLer genes but does not seem to be the only cause. Thus, other unknown mechanisms may also be involved in the differential H3K27me3 between Col and Ler. Nevertheless, considering the antagonistic nature of H3K9me2 and H3K27me3, the introduction of H3K9me2 (Results, Figure 11B) can explain the disappearance of H3K27me3 in Col in most cases.

The five Exp_Col genes are an exception from this model. They have the active mark H3K4me3 in Col but repressive mark H3K27me3 in Ler. TE insertion in Col could also be involved in the activation of HLer genes in Col seedlings. Two of the Exp_Col, AT4G20480 and AT5G56910, are flanked by TE. The absence of the TEs in Ler was detected by sequence comparison. For AT5G56910, the deletion of the neighboring TE AT5TE82820 was even validated by PCR following Sanger sequencing. The occurrence of H3K4me3 at Exp_Col can be partially due to the insertion of TE in Col. It has been shown that H3K4me3 can inhibit the recruitment of H3K27 trimethylation by PRC2 in mammals in vitro (Schmitges et al. 2011). The existence of H3K4me3 probably functions as a barrier to prevent the invasion of H3K27me3 in Col.

5.6 Inheritance of H3K27me3 in reciprocal hybrids of Col and Ler

5.6.1 F1 hybrids inherited H3K27me3 from any parent

The H3K27me3 profiles are highly similar between parents (Col and Ler) and the reciprocal hybrids. The number of common H3K27me3 targets between the two F1 or Col and F1 is similar to that between replicates of Col or Ler based on ChIP-chip data (see Table 5). Most non-overlapping targets could be caused by differences in the technique for data generation and the threshold for the target definition. However, based on ChIP-Seq data in F1 of Col x Ler, about half of HLer genes were not enriched in H3K27me3 according to the number of reads mapped to them. To test whether these HLer genes really lost H3K27me3 in both alleles, four of these HLer genes were chosen to verify their H3K27me3 level using ChIP-PCR. Strikingly, all of the chosen HLer genes still show a similarly strong H3K27me3 signal in both hybrids as in Ler (**Figure 23**, by

Julia, personal communication). So, none of the HLer genes lost the H3K27me3 mark in any of their hybrids.

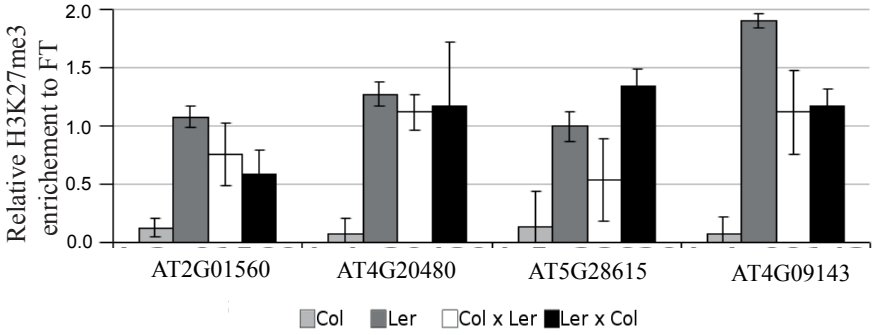


Figure 23. HLer genes show similar H3K27me3 signal in F1 as in parental Ler allele.
Four HLer genes which show low signal in ChIP-Seq data were chosen to test the level of H3K27me3 in reciprocal F1 hybrids using ChIP-PCR. The H3K27me3 signal from the four genes was normalized with the signal from *FLOWRING LOCUS T* (*FT*) (Figure was provided by Julia Reimer).

The H3K27me3 in the F1 hybrids was not detected from the ChIP-Seq data. These could be caused by two reasons, one reason is that, for HLer genes, probably only Ler allele are H3K27me3 modified, so less reads are produced in hybrids. The other reason is the difficulty to map Ler reads to the Col reference genome because of SNPs. The two reasons resulted in a low signal that passed below the threshold for H3K27me3 target detection in hybrids. Nevertheless, we see no evidence of difference in H3K27me3 signal between parents and both hybrids. The parental alleles tend to keep their state of histone modifications in the next generation.

5.6.2 *Cis*-effect of H3K27me3 inheritance in reciprocal hybrids of Col and Ler

In a *cis*-effect model, only the Ler allele of HLer genes can be H3K27 methylated in hybrids since only the Ler allele is H3K27me3 methylated in parents. In a *trans*-effect model, a *trans*-acting factor could either recruit H3K27me3 to the Col allele or repress H3K27me3 of the Ler allele. In our data set, we see clear allele-specific H3K27me3 modifications in the F1. I calculated the allele frequency of HLer, HCol genes and common targets between Col and Ler in both hybrids (see **Figure 19**B/C). For common targets, the allele frequency tends to be 0.5, which means the alleles from Col and Ler are H3K27 trimethylated equally; For HLer genes, the allele frequency tends to be 0, which

means only the Ler allele was H3K27 trimethylated, whereas the Col allele was not; For HCol genes, the allele frequency tends to be 1, suggesting that the Col allele was H3K27 trimethylated, whereas the Ler allele was not. The allele frequencies of HLer and HCol genes were significantly different from random samples from common targets (permutation test, p value < 0.001).

We found a slight bias towards the Col alleles in crosses where Col had been the mother. This bias was due to a contamination with non-hybrid seeds in the crosses. About 10% seedlings were detected by PCR to be pure mother plant in both F1 hybrid cohorts (Julia Reimer, personal communication). We did not find the bias towards the Ler alleles in crosses where Ler had been the female. This can be explained by the counteraction of two biases from two directions. One bias was caused by the inefficiency in mapping Ler reads to Col genome and the other bias was introduced by a contamination with non-hybrid seeds.

This allele-specific H3K27me3 in F1 hybrids supports a *cis*-effect in the regulation of H3K27me3 inheritance and is consistent with a previous study that compared histone methylation patterns in hybrids between rice cultivars (G. He et al. 2010). However, the number of chosen HLer genes for confirming the H3K27me3 modification in F1 is small because some HLer genes do not have SNPs or enough reads mapped to detect allele frequency. Additionally, HCol genes are not specific according to ChIP-PCR and could contain false predictions. We cannot completely exclude the possible existence of a *trans*-effect for some genes in the whole genome. However, based on this data, for the majority of genes H3K27me3 inheritance must be regulated through a *cis*-effect.

6 Conclusions and perspectives

We demonstrate local variation of the repressive chromatin mark H3K27me3 between *Arabidopsis thaliana* accessions Col and Ler. These variations are associated with differences in other chromatin modifications and transcriptional output of target genes. The distribution of H3K27me3 in Col, Ler and their reciprocal F1 hybrids is highly similar and only a small number of genes are H3K27me3 targets in Ler but not Col (HLer). These HLer genes were found to be either marked by the active chromatin mark H3K4me3 or the repressive mark H3K9me2 in the Col genome instead of the H3K27me3 found in Ler. In the reciprocal hybrids, allele specific H3K27me3 was observed, indicating a *cis*-regulatory mechanism. We propose a model where the insertion of TE into the Col genome influences the neighboring HLer genes by spreading of H3K9me2, which is antagonistic to H3K27me3. This model is consistent with the majority of HLer genes found. In five cases H3K4me3 was found at these genes in Col, which may also be caused by TE insertion, but with TE having an expression activating effect on neighboring genes. The number of HLer genes which could not be explained by occupancy of H3K9me2/H3K4me3 was small and did not allow the identification of common sequence patterns that may be responsible for differential H3K27me3 occurrence.

After comparing the genomic sequences of HLer genes identified in Col and Ler, presence or absence of TE became a topic of interest in this study. To understand the variation of H3K27me3 in Col, it is not enough to only compare the gene sequences in both genomes or just look at the histone modifications in Col, but also the maps of other histone modifications in Ler are needed considering that the chromatin state are characterized by combinations of different histone modifications. In the future, it will be helpful to get a full set of histone marks in both genomes, which can then be analyzed in a systematic way.

Col and Ler are two relative close accessions in *Arabidopsis thaliana*. There is only a small amount of genes differentially marked by H3K27me3 and most of them do not correlate with variation in gene expression. It would be interesting to determine the natural variation of histone modifications between more divergent species such as Col and *Arabidopsis lyrata* and evaluate the extent to which the variation at H3K27me3

influences the variation in expression. We have shown in another study that in many genes and TEs the Col alleles were less expressed than lyrate allele in the F1 hybrids of the two species and the allele-specific expression correlated with differential H3K27me3 (He et al. 2012; He et al. 2012). This founding indicates that a substantial amount of genes are differentially decorated by H3K27me3 between the two species and behave as the five Exp_Col that we identified in this project. However, H3K9me2 occupancy may also accounts for the differences of H3K27me3 between species.

7 References

Anon, 2000. Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. *Nature*, 408(6814), pp.796-815.

Banaei Moghaddam, A.M. et al., 2010. Intraspecific hybrids of Arabidopsis thaliana revealed no gross alterations in endopolyploidy, DNA methylation, histone modifications and transcript levels. *TAG. Theoretical and applied genetics. Theoretische und angewandte Genetik*, 120(2), pp.215-26.

Berger, N. et al., 2011. Transcriptional regulation of Arabidopsis LEAFY COTYLEDON2 involves RLE, a cis-element that regulates trimethylation of histone H3 at lysine-27. *The Plant cell*, 23(11), pp.4065-78.

Berger, S.L., 2007. The complex language of chromatin regulation during transcription. *Nature*, 447(7143), pp.407-12.

Bernatavichute, Y.V. et al., 2008. Genome-wide association of histone H3 lysine nine methylation with CHG DNA methylation in Arabidopsis thaliana. *PloS one*, 3(9), p.e3156.

Bird, A., 2002. DNA methylation patterns and epigenetic memory. *Genes & development*, 16(1), pp.6-21.

Bouyer, D. et al., 2011. Polycomb repressive complex 2 controls the embryo-to-seedling phase transition. *PLoS genetics*, 7(3), p.e1002014.

Bratzel, F. et al., 2010. Keeping cell identity in Arabidopsis requires PRC1 RING-finger homologs that catalyze H2A monoubiquitination. *Current biology : CB*, 20(20), pp.1853-9.

Breitling, R. et al., 2004. Rank products: a simple, yet powerful, new method to detect differentially regulated genes in replicated microarray experiments. *FEBS Letters*, 573(1-3), pp.83-92.

Brockdorff, N., 2011. Chromosome silencing mechanisms in X-chromosome inactivation: unknown unknowns. *Development*, 138(23), pp.5057-5065.

Brown, K.E. et al., 1997. Association of transcriptionally silent genes with Ikaros complexes at centromeric heterochromatin. *Cell*, 91(6), pp.845-54.

Bártová, E. et al., 2002. Nuclear structure and gene activity in human differentiated cells. *Journal of structural biology*, 139(2), pp.76-89.

Calonje, M. et al., 2008. EMBRYONIC FLOWER1 participates in polycomb group-mediated AG gene silencing in Arabidopsis. *The Plant cell*, 20(2), pp.277-91.

Carver, T. et al., 2008. Artemis and ACT: viewing, annotating and comparing sequences stored in a relational database. *Bioinformatics (Oxford, England)*, 24(23), pp.2672-6.

Chan, S.W. et al., 2004. RNA Silencing Genes Control de Novo DNA Methylation. *Science*, 303(February), p.2004.

Charron, J.-B.F. et al., 2009. Dynamic landscapes of four histone modifications during deetiolation in Arabidopsis. *The Plant cell*, 21(12), pp.3732-48.

Chinnusamy, V. & Zhu, J.-K., 2009. RNA-directed DNA methylation and demethylation in plants. *Science in China. Series C, Life sciences / Chinese Academy of Sciences*, 52(4), pp.331-43.

Clark, R.M. et al., 2007. Common sequence polymorphisms shaping genetic diversity in Arabidopsis thaliana. *Science (New York, N.Y.)*, 317(5836), pp.338-42.

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

Creyghton, M.P. et al., 2008. H2AZ is enriched at polycomb complex target genes in ES cells and is necessary for lineage commitment. *Cell*, 135(4), pp.649-61.

Cuddapah, S. et al., 2007. Global analysis of the insulator binding protein CTCF in chromatin barrier regions reveals demarcation of active and repressive domains. *Genome Research*, pp.24-32.

Deal, R.B. et al., 2007. Repression of flowering in Arabidopsis requires activation of FLOWERING LOCUS C expression by the histone variant H2A.Z. *The Plant cell*, 19(1), pp.74-83.

Dierssen, M. et al., 2012. Reduced Mid1 Expression and Delayed Neuromotor Development in daDREAM Transgenic Mice. *Frontiers in molecular neuroscience*, 5(May), p.58.

van Dijk, K. et al., 2010. Dynamic changes in genome-wide histone H3 lysine 4 methylation patterns in response to dehydration stress in Arabidopsis thaliana. *BMC plant biology*, 10(1), p.238.

Dimitri, P, Corradini, N, et al., 2005. Transposable elements as artisans of the heterochromatic genome in Drosophila melanogaster. *Cytogenetic and genome research*, 110(1-4), pp.165-72.

Dimitri, Patrizio et al., 2009. Constitutive heterochromatin: a surprising variety of expressed sequences. *Chromosoma*, 118(4), pp.419-35.

Dimitri, Patrizio, Corradini, Nicoletta, et al., 2005. The paradox of functional heterochromatin. *BioEssays : news and reviews in molecular, cellular and developmental biology*, 27(1), pp.29-41.

Eichten, S.R. et al., 2011. Heritable Epigenetic Variation among Maize Inbreds G. P. Copenhaver, ed. *PLoS Genetics*, 7(11), p.e1002372.

Ernst, J. et al., 2011. Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature*, 473(7345), pp.43-9.

Felsenfeld, G. et al., 2004. Chromatin boundaries and chromatin domains. *Cold Spring Harbor symposia on quantitative biology*, 69, pp.245-50.

Feng, S., Jacobsen, S.E. & Reik, W., 2010. Epigenetic reprogramming in plant and animal development. *Science (New York, N.Y.)*, 330(6004), pp.622-7.

Fernandez, L. et al., 2010. Transposon-induced gene activation as a mechanism generating cluster shape somatic variation in grapevine. *The Plant journal : for cell and molecular biology*, 61(4), pp.545-57.

Feschotte, C., 2008. Transposable elements and the evolution of regulatory networks. *Nature reviews. Genetics*, 9(5), pp.397-405.

Gan, X. et al., 2011. Multiple reference genomes and transcriptomes for Arabidopsis thaliana. *Nature*, pp.1-5.

Gentleman, R.C. et al., 2004. Bioconductor: open software development for computational biology and bioinformatics. *Genome biology*, 5(10), p.R80.

Grewal, S.I.S. & Moazed, D., 2003. Heterochromatin and epigenetic control of gene expression. *Science (New York, N.Y.)*, 301(5634), pp.798-802.

Gurvich, N. et al., 2005. Association of valproate-induced teratogenesis with histone deacetylase inhibition in vivo. *FASEB journal : official publication of the Federation of American Societies for Experimental Biology*, 19(9), pp.1166-8.

Guttman, M. et al., 2009. Chromtin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. , 458(7235), pp.223-227.

Göbel, U., Reimer, J. & Turck, F., 2010. Plant Epigenetics I. Kovalchuk & F. J. Zemp, eds. *plant genetics*, 631, pp.161-184.

Harničarová Hor&aacut, A., Bártová, E. & Kozubek, S., 2010. Chromatin Structure with Respect to Histone Signature Changes during Cell Differentiation. *Cell Structure and Function*, 35(1), pp.31-44.

He F, Zhang X, Hu J, Turck F, Dong X, Goebel U, Borevitz J, de M.J., 2012. Genome-wide analysis of cis-regulatory divergence between species in the Arabidopsis genus. *Mol Biol Evol*.

He, F. et al., 2012. Widespread interspecific divergence in cis-regulation of transposable elements in the Arabidopsis genus. *Molecular biology and evolution*, 29(3), pp.1081-91.

He, G. et al., 2010. Global epigenetic and transcriptional trends among two rice subspecies and their reciprocal hybrids. *The Plant cell*, 22(1), pp.17-33.

Heo, J.B. & Sung, S., 2011. Vernalization-mediated epigenetic silencing by a long intronic noncoding RNA. *Science (New York, N.Y.)*, 331(6013), pp.76-9.

Ho, J.W.K. et al., 2011. ChIP-chip versus ChIP-seq: lessons for experimental design and data analysis. *BMC genomics*, 12(1), p.134.

Hollister, J.D. et al., 2011. Transposable elements and small RNAs contribute to gene expression divergence between Arabidopsis thaliana and Arabidopsis lyrata. *Proceedings of the National Academy of Sciences of the United States of America*, 108(6), pp.2322-7.

Hollister, J.D. & Gaut, B.S., 2009. Epigenetic silencing of transposable elements: a trade-off between reduced transposition and deleterious effects on neighboring gene expression. *Genome research*, 19(8), pp.1419-28.

Hong, F. et al., 2006. BIOINFORMATICS APPLICATIONS NOTE RankProd : a bioconductor package for detecting differentially expressed genes in meta-analysis. *Bioinformatics*, 22(22), pp.2825-2827.

Huang, X. et al., 2008. Genome-wide analysis of transposon insertion polymorphisms reveals intraspecific variation in cultivated rice. *Plant physiology*, 148(1), pp.25-40.

Iii, R.J.S. et al., 2008. Recognition of Trimethylated Histone H3 Lysine 4 Facilitates the Recruitment of Transcription Post-Initiation Factors and pre- mRNA Splicing. , 28(4), pp.665-676.

Ingouff, M. et al., 2007. Distinct dynamics of HISTONE3 variants between the two fertilization products in plants. *Current biology : CB*, 17(12), pp.1032-7.

Jabbari, K. & Bernardi, G., 2004. Cytosine methylation and CpG, TpG (CpA) and TpA frequencies. *Gene*, 333, pp.143-9.

Jacob, Y. et al., 2009. ATXR5 and ATXR6 are H3K27 monomethyltransferases required for chromatin structure and gene silencing. *Nature structural & molecular biology*, 16(7), pp.763-8.

Jenuwein, T. & Allis, C D, 2001. Translating the histone code. *Science (New York, N.Y.)*, 293(5532), pp.1074-80.

Ji, H. & Wong, W.H., 2005. TileMap: create chromosomal map of tiling array hybridizations. *Bioinformatics (Oxford, England)*, 21(18), pp.3629-36.

Kamakaka, R.T. & Biggins, S., 2005. Histone variants: deviants? *Genes & development*, 19(3), pp.295-310.

Kashkush, K., Feldman, M. & Levy, A. a, 2003. Transcriptional activation of retrotransposons alters the expression of adjacent genes in wheat. *Nature genetics*, 33(1), pp.102-6.

Keleş, S., 2007. Mixture modeling for genome-wide localization of transcription factors. *Biometrics*, 63(1), pp.10-21.

Kliebenstein, D.J. et al., 2006. Genomic survey of gene expression diversity in Arabidopsis thaliana. *Genetics*, 172(2), pp.1179-89.

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

Krogan, N.J. et al., 2003. Methylation of Histone H3 by Set2 in Saccharomyces cerevisiae Is Linked to Transcriptional Elongation by RNA Polymerase II. , 23(12), pp.4207-4218.

Kumar, S.V. & Wigge, P. a, 2010. H2A.Z-containing nucleosomes mediate the thermosensory response in Arabidopsis. *Cell*, 140(1), pp.136-47.

Kurtz, S. et al., 2004. Versatile and open software for comparing large genomes. *Genome biology*, 5(2), p.R12.

Köhler, C. & Villar, C.B.R., 2008. Programming of gene expression by Polycomb group proteins. *Trends in cell biology*, 18(5), pp.236-43.

Lafos, M. et al., 2011. Dynamic regulation of H3K27 trimethylation during Arabidopsis differentiation. *PLoS genetics*, 7(4), p.e1002040.

Laubinger, S. et al., 2008. At-TAX: a whole genome tiling array resource for developmental expression analysis and transcript identification in Arabidopsis thaliana. *Genome biology*, 9(7), p.R112.

Law, J. a & Jacobsen, S.E., 2010. Establishing, maintaining and modifying DNA methylation patterns in plants and animals. *Nature reviews. Genetics*, 11(3), pp.204-20.

Lee, J.-S. & Shilatifard, A., 2007. A site to remember: H3K36 methylation a mark for histone deacetylation. *Mutation research*, 618(1-2), pp.130-4.

Lee, J.-S., Smith, E. & Shilatifard, A., 2010. The language of histone crosstalk. *Cell*, 142(5), pp.682-5.

Li, H. et al., 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics (Oxford, England)*, 25(16), pp.2078-9.

Li, H. & Durbin, R., 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics (Oxford, England)*, 25(14), pp.1754-60.

Li, W., Meyer, C. a & Liu, X.S., 2005. A hidden Markov model for analyzing ChIP-chip experiments on genome tiling arrays and its application to p53 binding sequences. *Bioinformatics (Oxford, England)*, 21 Suppl 1(2004), pp.i274-82.

Lienert, F. et al., 2011. Genomic Prevalence of Heterochromatic H3K9me2 and Transcription Do Not Discriminate Pluripotent from Terminally Differentiated Cells. *PLoS genetics*, 7(6), p.e1002090.

Lin, N. et al., 2011. A barrier-only boundary element delimits the formation of facultative heterochromatin in Drosophila melanogaster and vertebrates. *Molecular and cellular biology*, 31(13), pp.2729-41.

Lippman, Z. et al., 2004. Role of transposable elements in heterochromatin and epigenetic control. *Nature*, 430(6998), pp.471-6.

Liu, C., Lu, F., et al., 2010. higher plant histone modifcation. *Annual review of plant biology*, 61, pp.395-420.

Liu, E.T., Pott, S. & Huss, M., 2010. Q&A: ChIP-seq technologies and the study of gene regulation. *BMC biology*, 8, p.56.

Liu, J. et al., 2004. siRNAs targeting an intronic transposon in the regulation of natural flowering behavior in Arabidopsis. *Genes & development*, 18(23), pp.2873-8.

Locke, S.M. & Martienssen, R. a, 2006. Slicing and spreading of heterochromatic silencing by RNA interference. *Cold Spring Harbor symposia on quantitative biology*, 71, pp.497-503.

Luo, C. & Lam, E., 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), pp.339-51.

Ma, M.K.-W. et al., 2011. Histone Crosstalk Directed by H2B Ubiquitination Is Required for Chromatin Boundary Integrity J. T. Lee, ed. *PLoS Genetics*, 7(7), p.e1002175.

Margueron, R. & Reinberg, D., 2010. Chromatin structure and the inheritance of epigenetic information. *Nature reviews. Genetics*, 11(4), pp.285-96.

Martin-Magniette, M.-L. et al., 2008. ChIPmix: mixture model of regressions for two-color ChIP-chip analysis. *Bioinformatics (Oxford, England)*, 24(16), pp.i181-6.

Mavrich, T.N. et al., 2008. Nucleosome organization in the Drosophila genome. *Nature*, 453(7193), pp.358-62.

Mendenhall, E.M. et al., 2010. GC-rich sequence elements recruit PRC2 in mammalian ES cells. H. D. Madhani, ed. *PLoS genetics*, 6(12), p.e1001244.

Meneghini, M.D. et al., 2003. Conserved Histone Variant H2A . Z Protects Euchromatin from the Ectopic Spread of Silent Heterochromatin. , 112, pp.725-736.

Mikkelsen, T.S. et al., 2010. Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. , 448(7153), pp.553-560.

Misteli, T., 2007. Beyond the sequence: cellular organization of genome function. *Cell*, 128(4), pp.787-800.

Moghaddam, A.M.B. et al., 2011. Additive inheritance of histone modifications in Arabidopsis thaliana intra-specific hybrids. *The Plant journal : for cell and molecular biology*, 67(4), pp.691-700.

Morey, L. & Helin, K., 2010. Polycomb group protein-mediated repression of transcription. *Trends in biochemical sciences*, 35(6), pp.323-32.

Moritz, K.B. & Roth, G.E., Complexity of germline and somatic DNA in Ascaris. *Nature*, 259(5538), pp.55-7.

Muiño, J.M. et al., 2011. ChIP-seq Analysis in R (CSAR): An R package for the statistical detection of protein-bound genomic regions. *Plant methods*, 7(1), p.11.

Oh, S., Park, S., & Van Nocker, S. (2008). Genic and global functions for Paf1C in chromatin modification and gene expression in Arabidopsis. *PLoS genetics*, *4*(8), e1000077. doi:10.1371/journal.pgen.1000077

Pai, A. a et al., 2011. A genome-wide study of DNA methylation patterns and gene expression levels in multiple human and chimpanzee tissues. *PLoS genetics*, 7(2), p.e1001316.

Park, P.J., 2009. ChIP-seq: advantages and challenges of a maturing technology. *Nature reviews. Genetics*, 10(10), pp.669-80.

Pien, S. & Grossniklaus, U., 2007. Polycomb group and trithorax group proteins in Arabidopsis. *Biochimica et biophysica acta*, 1769(5-6), pp.375-82.

Pimpinelli, S., 1995. Transposable Elements are Stable Structural Components of Drosophila melanogaster Heterochromatin. *Proceedings of the National Academy of Sciences*, 92(9), pp.3804-3808.

Pontes, O. et al., 2006. The Arabidopsis chromatin-modifying nuclear siRNA pathway involves a nucleolar RNA processing center. *Cell*, 126(1), pp.79-92.

Pooma, W., Gersos, C. & Grotewold, E., 2002. Transposon insertions in the promoter of the Zea mays a1 gene differentially affect transcription by the Myb factors P and C1. *Genetics*, 161(2), pp.793-801.

Rehrauer, H. et al., 2010. AGRONOMICS1: a new resource for Arabidopsis transcriptome profiling. *Plant physiology*, 152(2), pp.487-99.

Reimer, J.J. & Turck, F., 2010. Plant Epigenetics I. Kovalchuk & F. J. Zemp, eds. *Plant Epigenetics*, 631(1), pp.139-160.

Ridgway, P., 2001. Chromatin assembly and organization. *Journal of cell science*, 114(Pt 15), pp.2711-2712.

Rinn, J.L. et al., 2007. Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs. *Cell*, 129(7), pp.1311-23.

Rollins, R. a et al., 2006. Large-scale structure of genomic methylation patterns. *Genome research*, 16(2), pp.157-63.

Rossi, F. et al., 2007. Cytogenetic and molecular characterization of heterochromatin gene models in Drosophila melanogaster. *Genetics*, 175(2), pp.595-607.

Roudier, F. et al., 2011. Integrative epigenomic mapping defines four main chromatin states in Arabidopsis. *The EMBO journal*, 30(10), pp.1928-38.

Sanchez-Pulido, L. et al., 2008. RAWUL: A new ubiquitin-like domain in PRC1 Ring finger proteins that unveils putative plant and worm PRC1 orthologs. *BMC Genomics*, 9(1), p.308.

Schmid, K.J. et al., 2003. Large-scale identification and analysis of genome-wide single-nucleotide polymorphisms for mapping in Arabidopsis thaliana. *Genome research*, 13(6A), pp.1250-7.

Schmitges, F.W. et al., 2011. Histone methylation by PRC2 is inhibited by active chromatin marks. *Molecular cell*, 42(3), pp.330-41.

Schneeberger, K. et al., 2011. Reference-guided assembly of four diverse Arabidopsis thaliana genomes. *Proceedings of the National Academy of Sciences of the United States of America*, 108(25), pp.10249-54.

Schones, D.E. & Zhao, K., 2008. Genome-wide approaches to studying chromatin modifications. *Nature reviews. Genetics*, 9(3), pp.179-91.

Shen, H. et al., 2012. Genome-Wide Analysis of DNA Methylation and Gene Expression Changes in Two Arabidopsis Ecotypes and Their Reciprocal Hybrids. *Plant Cell*.

Simon, J.A. & Kingston, R.E., 2009. Mechanisms of polycomb gene silencing: knowns and unknowns. *Nature reviews. Molecular cell biology*, 10(10), pp.697-708.

Spada, F., Vincent, M. & Thompson, E.M., 2005. Plasticity of histone modifications across the invertebrate to vertebrate transition: histone H3 lysine 4 trimethylation in heterochromatin. *Chromosome research : an international journal on the molecular, supramolecular and evolutionary aspects of chromosome biology*, 13(1), pp.57-72.

Springer, N.M. et al., 2009. Maize inbreds exhibit high levels of copy number variation (CNV) and presence/absence variation (PAV) in genome content. J. R. Ecker, ed. *PLoS genetics*, 5(11), p.e1000734.

Stein, L.D. et al., 2002. The generic genome browser: a building block for a model organism system database. *Genome research*, 12(10), pp.1599-610.

Sturn, A. & Quackenbush, J., 2002. Genesis: cluster analysis of microarray data. *Bioinformatics*, 18(1), pp.207-208.

Sun, Z.-W. & Allis, C David, 2002. Ubiquitination of histone H2B regulates H3 methylation and gene silencing in yeast. *Nature*, 418(6893), pp.104-8.

Suva, M.-luca et al., 2010. EWS-FLI-1 modulates miRNA145 and SOX2 expression to initiate mesenchymal stem cell reprogramming toward Ewing sarcoma cancer stem cells. , pp.916-932.

Swarbreck, D. et al., 2008. The Arabidopsis Information Resource (TAIR): gene structure and function annotation. *Nucleic acids research*, 36(Database issue), pp.D1009-14.

Talbert, P.B. & Henikoff, S., 2006. Spreading of silent chromatin: inaction at a distance. *Nature reviews. Genetics*, 7(10), pp.793-803.

Teixeira, F.K. et al., 2009. A role for RNAi in the selective correction of DNA methylation defects. *Science (New York, N.Y.)*, 323(5921), pp.1600-4.

Toedling, J. et al., 2007. Ringo--an R/Bioconductor package for analyzing ChIP-chip readouts. *BMC bioinformatics*, 8, p.221.

Tsai, M.-C. et al., 2010. Long noncoding RNA as modular scaffold of histone modification complexes. *Science (New York, N.Y.)*, 329(5992), pp.689-93.

Turck, F. et al., 2007. Arabidopsis TFL2/LHP1 specifically associates with genes marked by trimethylation of histone H3 lysine 27. *PLoS genetics*, 3(6), p.e86.

Vaughn, M.W. et al., 2007. Epigenetic natural variation in Arabidopsis thaliana. *PLoS biology*, 5(7), p.e174.

Wang, H. & Wang, L., 2004. Role of histone H2A ubiquitination in Polycomb silencing. *Nature*, 431(7010), pp.873-877.

Wassenegger, M. et al., 1994. RNA-directed de novo methylation of genomic sequences in plants. *Cell*, 76(3), pp.567-76.

Weber, C.M., Henikoff, J.G. & Henikoff, S., 2010. H2A.Z nucleosomes enriched over active genes are homotypic. *Nature structural & molecular biology*, 17(12), pp.1500-7.

Weinhofer, I. et al., 2010. H3K27me3 profiling of the endosperm implies exclusion of polycomb group protein targeting by DNA methylation. *PLoS genetics*, 6(10), pp.1-14.

Xie, Z. et al., 2004. Genetic and functional diversification of small RNA pathways in plants. *PLoS biology*, 2(5), p.E104.

Yan, H. et al., 2010. Genome-wide mapping of cytosine methylation revealed dynamic DNA methylation patterns associated with genes and centromeres in rice. *The Plant journal : for cell and molecular biology*, pp.353-365.

Zang, C. et al., 2009. A clustering approach for identification of enriched domains from histone modification ChIP-Seq data. *Bioinformatics (Oxford, England)*, 25(15), pp.1952-8.

Zhang, Xiaoyu et al., 2009. Genome-wide analysis of mono- , di- and trimethylation of histone H3 lysine 4 in Arabidopsis thaliana. *Genome Biology*, pp.1-14.

Zhang, Xiaoyu et al., 2006. Genome-wide high-resolution mapping and functional analysis of DNA methylation in arabidopsis. *Cell*, 126(6), pp.1189-201.

Zhang, Xiaoyu et al., 2007. Whole-genome analysis of histone H3 lysine 27 trimethylation in Arabidopsis. *PLoS biology*, 5(5), p.e129.

Zhang, Xu & Borevitz, J.O., 2009. Global analysis of allele-specific expression in Arabidopsis thaliana. *Genetics*, 182(4), pp.943-54.

Zheng, S., Wyrick, J.J. & Reese, J.C., 2010. Novel trans-tail regulation of H2B ubiquitylation and H3K4 methylation by the N terminus of histone H2A. *Molecular and cellular biology*, 30(14), pp.3635-45.

Zhu, Y., Dong, A. & Shen, W.-H., 2011. Histone variants and chromatin assembly in plant abiotic stress responses. *Biochimica et biophysica acta*, 1819(3-4), pp.343-348.

Zilberman, D. et al., 2007. Genome-wide analysis of Arabidopsis thaliana DNA methylation uncovers an interdependence between methylation and transcription. *Nature genetics*, 39(1), pp.61-9.

Zilberman, D. et al., 2008. Histone H2A.Z and DNA methylation are mutually antagonistic chromatin marks. *Nature*, 456(7218), pp.125-9.

8 Appendix

8.1 Supplementary Tables

Table S1. The genomic features associated with HLer genes.

HLer gene type	TE_left	HLer name	TE_right	Confirmed HLer	Sequence polymorphism	Confirmed insertion	Expression in Col	H3K4me3 marked	H3K9me2 marked
PCG	AT5TE49910	AT5G35810	AT5TE49915	yes	Col_insertion	yes	no	no	yes
PCG	AT5TE82820	AT5G56920	AT5TE82825	yes	Col_insertion	yes	no	no	yes
TE	AT5TE50065	AT5G35914	no	not	no		no	no	no
TE	no	AT1G30835	no	yes	no		yes	yes	no
PCG	AT5TE37885	AT5G28463	AT5TE37900	yes	no		no	no	yes
PCG	no	AT3G60150	no	yes	no, small ler_insertion		yes	yes	yes
PCG	no	AT4G03566	no		no		no	no	yes
PCG	AT5TE61720	AT5G42640	AT5TE61725	yes	Col_insertion	yes	no	no	yes
PCG	AT4TE50620	AT4G20480	AT4TE50630	yes	Col_insertion		yes	yes	no
Pseudogene	AT2TE37940	AT2G20910	no	yes	Col_insert_itself		no	no	yes
PCG	no	AT4G26350	AT4TE62660		NA		no	no	yes
PCG	no	AT5G28615	no	yes	no		no	no	yes
PCG	AT1TE42435	AT1G35400	AT1TE42440		no		no	no	yes
PCG	no	AT5G02700	AT5TE02190		rearrangement		no	no	yes
PCG	AT5TE38680	AT5G28610	no		no		no	no	yes
PCG	no	AT1G65170	no		Col_insertion_itself		no	no	no
Pseudogene	AT2TE29725	AT2G16830	AT2TE29730		Col_insertion		no	no	yes
PCG	no	AT3G60560	AT3TE91170		Col_insertion		no	no	no
PCG	no	AT1G66300	AT1TE81190		Col_insertion		no	no	no
PCG	AT4TE28665	AT4G10870	AT4TE28670		Col_insertion		no	no	yes
PCG	AT5TE14765	AT5G12910	no		no		no	no	no
Pseudogene	no	AT4G09143	AT4TE24480	yes	Col_insertion		no	no	yes
PCG	AT2TE01000	AT2G01560	AT2TE01010	yes	Col_insertion		no	no	no
TE	no	AT1G35186	no		no		no	no	yes
PCG	no	AT3G46160	no		no		no	no	no
PCG	no	AT1G54230	no		no		no	no	no
PCG	no	AT1G21870	no		Col_insertion_noTE		no	no	no
PCG	AT2TE68590	AT2G36710	no		rearrangement		no	no	yes
PCG	AT2TE65225	AT2G34840	AT2TE65230	yes	Col_insertion_itself	yes	yes	no	yes
PCG	no	AT5G56910	AT5TE82820	yes	Col_insertion	yes	yes	yes	no
PCG	AT1TE70420	AT1G57565	no		no		no	no	yes
TE	AT3TE91830	AT3G60965	no		Col_insertion_itself		no	no	no

PCG: Protein-Coding Gene

Table S2. The sequence polymorphism of TE-flanking HLer genes between Col and Ler.

TE-flanking HLer genes	Sequence polymorphism	TE-flanking HLer genes	Sequence polymorphism
AT2G16830	Col_insertion	AT5G02700	Rearrangement
AT1G66300	Col_insertion	AT2G36710	Rearrangement
AT3G60560	Col_insertion	AT2G34840	Col_insertion_itself
AT4G09143	Col_insertion	AT2G20910	Col_insertion_itself
AT4G10870	Col_insertion	AT3G60965	Col_insertion_itself
AT4G20480	Col_insertion	AT5G35914	no
AT2G01560	Col_insertion	AT1G57565	no
AT5G56910	Col_insertion	AT1G35400	no
AT5G35810	Col_insertion	AT5G28463	no
AT5G56920	Col_insertion	AT5G28610	no
AT5G42640	Col_insertion	AT5G12910	no
AT4G26350	NA*		

8.2 Supplementary Figures

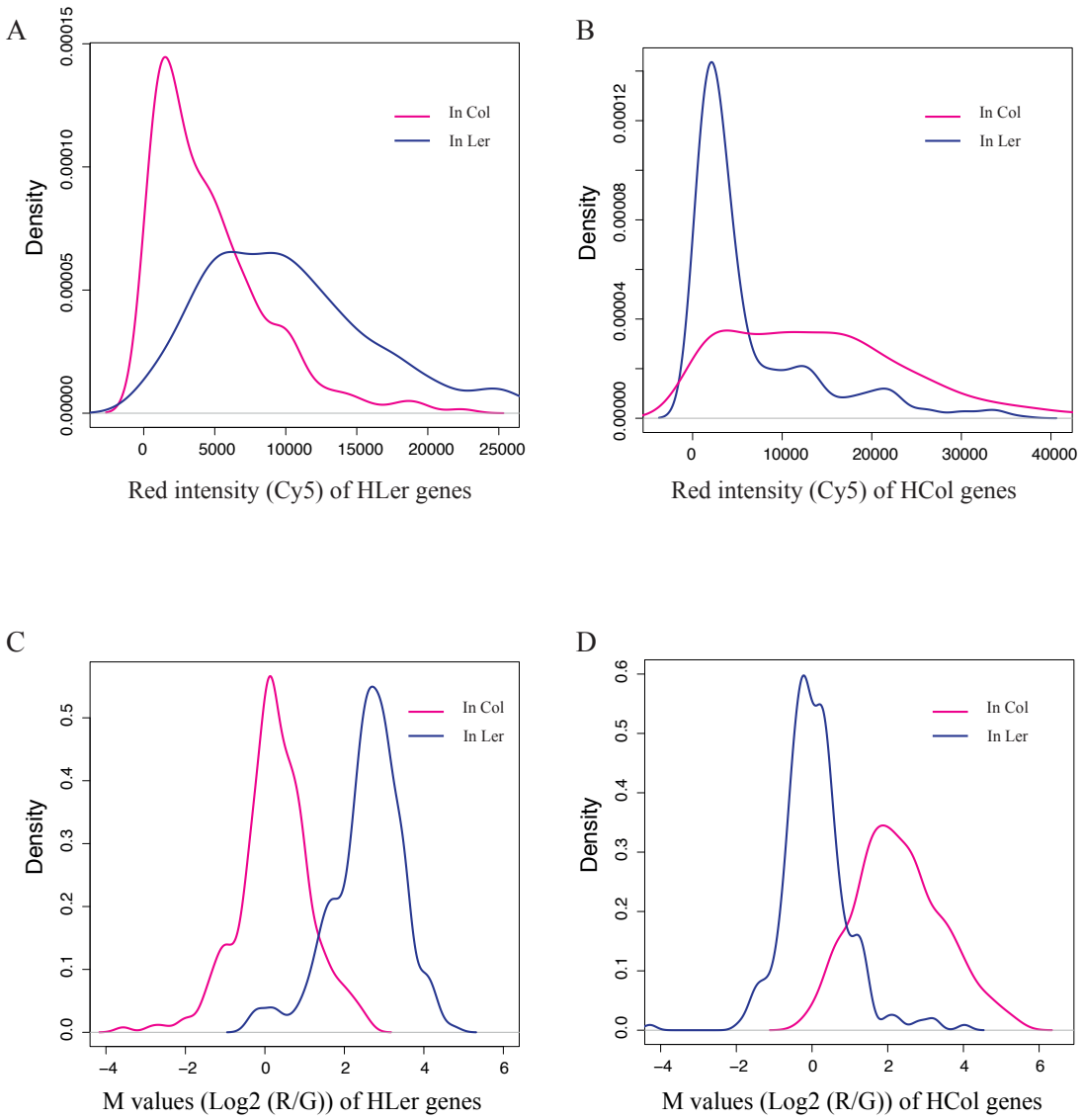


Figure S1. The distributions of red intensity (Cy5, ChIP) and M values of HLer/HCol genes in Col and Ler samples.
The red lines indicate that the values were from Col samples and blue lines indicate that the values were from Ler samples. The red intensity indicate the absolute H3K27me3 signal. The M values indicate the H3K27me3 signal relative to INPUT background.
(A) The distribution red intensity of HLer genes in Col and Ler. (B) The distribution red intensity of HCol genes in Col and Ler. (C) The distribution M values of HLer genes in Col and Ler. (D) The distribution M values of HCol genes in Col and Ler.

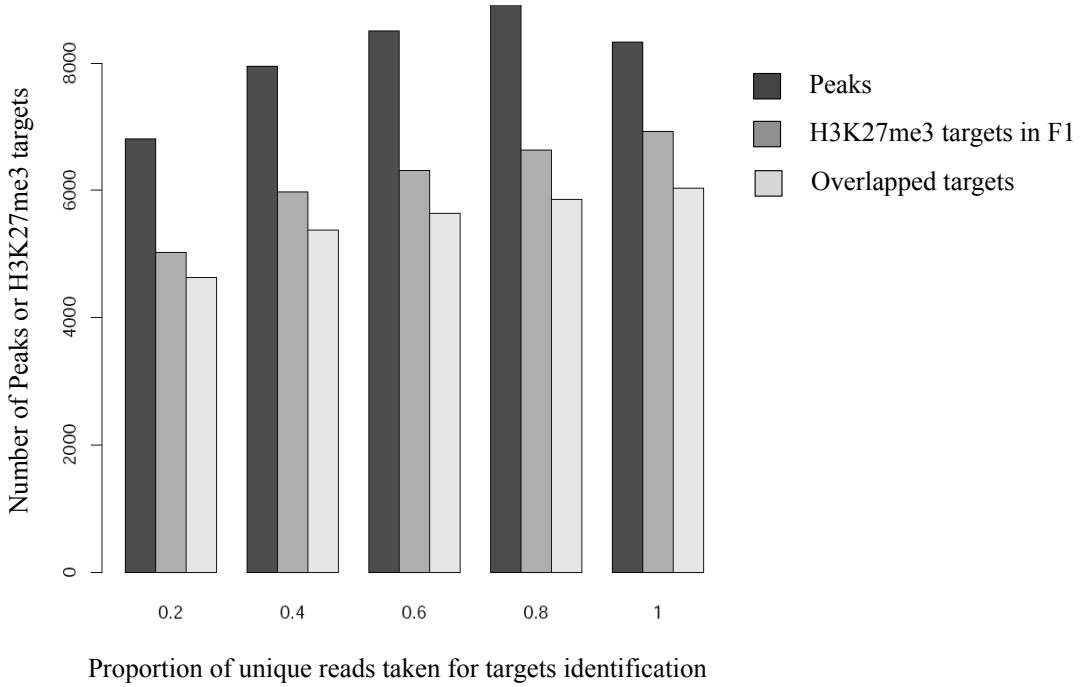


Figure S2. The saturation analysis of Chip-Seq data for F1: Col x Ler.
The dark gray bar shows the number of peaks identified from different proportion of reads; the gray bar shows the number of H3K27me3 targets identified based on the corresponding peaks; the light gray bar shows the overlapping H3K27me3 targets for ChIP-chip and corresponding peaks from ChIP-Seq data. The H3K27me3 targets was definition as at least 500bp or 30% of the gene was covered by H3K27me3 peaks.

8.3 Abbreviations

ACT	Artemis Comparison Tool
ASRP	*Arabidopsis* Small RNA Project
Arabidopsis	*Arabidopsis thaliana*
At-TAX	*Arabidopsis thaliana* Tiling Array Express
BWA	Burrows-Wheeler Aligner
CGH	Comparative Genomic Hybridization
CSAR	ChIP-Seq Analysis in R
ChIP-PCR	Chromatin immunoprecipitation (ChIP) following PCR
ChIP-chip	ChIP followed by hybridization with to whole genome tiling arrays
ChIP-Seq	ChIP followed by sequencing
Col	*Arabidopsis thaliana* Columbia
DEG	Differentially enriched gene
DNA	Deoxyribonucleic acid
Drosophila	Drosophila melanogaster
Exp_Col	The group of HLer genes expressing in Col
FIE	FERTILIZATION INDEPENDENT ENDOSPERM
FIS	FERTILIZATION INDEPENDENT SEED
FLC	FLOWERING LOCUS C
FT	FLOWERING LOCUS T
GEO	Gene Expression Omnibus
GO	Gene ontology
H2A.Z	Histone 2A.Z
H2B	Histone 2B
H3	Histone 3
H4	Histone 4
H3K27me3	H3 trimethylation at Lys27
H3K27me	H3 monomethylation at Lys27
H3K9me2	H3 dimethylation at Lys9
H3K4me3	tri-methylated lysine 4 at histone 3
HCol	Highly H3K27 trimethylated genes in Col

HLer	Highly H3K27 trimethylated genes in Ler
Ler	*Arabidopsis thaliana* Landsberg erecta
LHP1	LIKE HETEROCHROMATIN PROTEIN 1
PcG	Polycomb group
PCR	Polymerase chain reaction
PRC	Polycomb repressive complex
PRC1	Polycomb repressive complex 1
PRC2	Polycomb repressive complex 2
PcG	Polycomb group protein
RNA	Ribonucleic acid
RNA-Seq	RNA sequencing
RPKM	Reads per kilobase per million mapped reads
Rep_Col	The group of Hler genes repressed in expression in Col
SNP	single nucleotide polymorphism
TE	Transposable Element
TEG	Transposable Element Gene
TxG	Trithorax-group protein
bp	base
me	any methylation state of an arginine or lysine
me1	mono-methylation
me2	di-methylation
me3	tri-methylation
MPIPZ	Max Planck Institute for Plant Breeding Research
qRT-PCR	quantitative real time polymerase chain reaction
SU(Z)12	Suppressor of Zeste 12
TAIR	The *Arabidopsis* Information Resources

9 List of Figures and Tables

9.1 List of Figures

9.2 List of Tables

9 Acknowledgements

I would like to thank everyone who helped me during the time of doing my project. This work would not have been accomplished without the help of them. Special thanks go to:

Dr. Franziska Turck for her ideas, inspiration and supervision especially in biological science.

Prof. Dr. Heiko Schoof for giving me the chance to work on this cooperative project and his supervision in bioinformatics and biology, his warm encouragement and supports.

Prof. Frank Hochholdinger for being my second examiner and for all the questions and comments.

Prof. George Coupland for giving me the opportunity to join his department and attend all the department seminars.

Prof. Dr. Thomas Wiehe for supporting my study in university of Cologne.

Dr. Ulrike Goebel for her helpful recommendation and great advice during all the time especially at the beginning of my thesis.

Dr. Julia Reimer for her warm smiles and cooperative work in lab and critical reading of the manuscript.

Dr. Barbara Kracher, Dr. Nahal Ahmadinejad for critical reading of the manuscript.

Everyone in the groups of Schoof, Turck and BIT for good times and scientific discussions.

Prof. Heinz Saedler for the financial support in doing my internship together with Dr. Ulrike Goebel before start this thesis. Without the internship, I would not have had the opportunity to work on this project.

Max Planck Institute for Plant Breeding for offering me great scientific environment and facilities to do my project there. Thanks SUSAN for the technique supports.

All the Chinese colleagues in the MPIPZ for the BBQ and the great atmosphere; especially **Dr. Fei He** for great collaboration.

My family for being together with me, supporting me and loving me.