Genome-wide analysis of organ-specific DNA methylation patterns in *Arabidopsis thaliana*

Dissertation
der Mathematisch-Naturwissenschaftlichen Fakultät
der Eberhard Karls Universität Tübingen
zur Erlangung des Grades eines
Doktors der Naturwissenschaften
(Dr. rer. nat.)

vorgelegt von
Jorge Enrique Quintana Kageyama
aus Ciudad de Mexico, Mexico

Tübingen
2016

Gedruckt mit Genehmigung der Mathematisch-Naturwissenschaftlichen Fakultät der Eberhard-Karls-Universität Tübingen.

Tag der mündlichen Qualifikation:	26.07.2016
Dekan:	Prof. Dr. Wolfgang Rosenstiel
1. Berichterstatter:	Prof. Dr. Daniel Huson
2. Berichterstatter:	Prof. Dr. Detlef Weigel

Zusammenfassung

Obwohl genetische Variation für eine lange Zeit als die einzige Quelle für phänotypische Variation angesehen wurde, sind heute epigenetische Marker als zusätzliche Quellen für phänotypische Diversität weitestgehend anerkannt. DNA Methylierung ist ein vererbbarer epigenetischer Marker, der in vielen Eukaryoten unverzichtbar ist für eine Vielzahl biologischer Prozesse, einschließlich der transkriptionellen Stummschaltung von Genen und Transposons. Darüber hinaus können Veränderungen in der DNA Methylierung durch endogene und exogene Signale herbeigeführt werden, so dass DNA Methylierung als Mechanismus zur Regulierung von Genexpression dienen kann.

Trotz ihrer Vererbbarkeit handelt es sich bei DNA Methylierung um einen äußerst dynamischen epigenetischen Marker. Bei Pflanzen und Tieren sind frühe Entwicklungsstadien gekennzeichnet durch genomweite Neuprogrammierung der DNA Methylierung. Dies wird der Regulierung von Entwicklungsprogrammen zugeschrieben. Nebst entwicklungsgesteuerter Veränderungen der DNA Methylierung tragen unter anderem auch spontane Variationen und Umweltfaktoren zur Variabilität von DNA Methylierung bei.

Frühere Pflanzenstudien erforschten, wie sich Methylierung im Laufe der Entwicklung verändert, wobei sie sich im Allgemeinen auf frühe Entwicklungsstadien konzentrierten und somit die Variation von DNA Methylierung zwischen sich später entwickelnden Organen größtenteils außer Acht ließen. Diese Studie präsentiert eine detaillierte vergleichende Analyse von DNA Methylierungsprofilen mehrerer Organe von *A. thaliana* und zieht hierfür hochauflösende genomweite Karten heran, die bis auf einzelne Basen genau DNA-Methylierung anzeigen. Die Studie kann grob in drei Teile untergliedert werden.

Der erste Teil beinhaltet eine Pipeline für die Analyse von durch Next Generation Sequencing-Technologien gewonnenen DNA Methylierungsdaten. Diese Pipeline identifiziert zum einen methylierte Stellen im Genom innerhalb einzelner Proben und sucht zum anderen nach statistischen Unterschieden in Methylierungsmustern zwischen mehreren Proben.

Der zweite Teil dieser Studie untersucht die Variabilität von DNA Methylierung zwischen den Blättern einer einzelnen Pflanze und zielt darauf ab, zu bestimmen, ob diese Unterschiede in DNA Methylierung ein Produkt stochastischer Varianz sind oder durch andere Faktoren erklärt werden können. Durch den Vergleich der Methylierungsprofile von 18 Blättern, die alle von einer einzelnen Pflanze stammen, ist es mir gelungen zu zeigen, dass die zwischen den einzelnen Blättern beobachteten Veränderungen in der Methylierung mit dem Entstehungszeitpunkt des jeweiligen Blattes korrelieren. Im Vergleich mit sich in der Entwicklung befindenen Blättern zeigten zu einem früheren Zeitpunkt entstandene Blätter einen relativen Verlust an DNA Methylierung.

Der dritte Teil dieser Studie widmet sich der Verbindung von DNA Methylierung mit Organidentität, und im Speziellen, wie Veränderungen der DNA Methylierung zwischen Organen mit Veränderungen in der Genexpression korrelieren. Zu diesem Zweck habe ich Methylierungskarten sowie Transkriptionsprofile für sechs verschiedene Organsysteme aus jeweils insgesamt drei Einzelpflanzen angefertigt. Meine Ergebnisse zeigen, dass im Vergleich von reproduktiven mit vegetativen Organen genomweite Unterschiede in der DNA Methylierung vorliegen. Ferner korrelieren manche dieser Veränderungen mit Unterschieden in der Expression Proteinkodierender Gene. Schließlich ist es mir gelungen, zu zeigen, dass DNA Methylierungsmarker in regulatorischen Regionen nicht ausreichend sind, um Genexpression zu kontrollieren; stattdessen müssen diese epigenetischen Marker an Transposons gekoppelt sein.

Zusammenfassend liefert meine Dissertation Einsichten in die Variabilität von DNA Methylierung im Laufe der pflanzlichen Entwicklung. Ich habe verschiedene Faktoren identifiziert, darunter Organidentität und Alter, welche zur Variabilität von DNA Methylierung im Genom beitragen, was darauf hinweisen kann, dass DNA Methylierung als eine zusätzliche Ebene von Genregulierung in der Pflanzenentwicklung fungiert.

Summary

Even though for a long time genetic variation was considered the only source of phenotypic variation, it is now widely accepted that epigenetic marks can also contribute to phenotypic diversity. This thesis focuses on DNA methylation, a heritable epigenetic mark that exists across eukaryotic lineages and is crucial for numerous biological processes, including regulation of gene expression and the transcriptional silencing of genes and transposable elements. Despite being a heritable epigenetic mark, DNA methylation is not entirely static. Early stages of animal and plant development are marked by genome-wide reprogramming events of DNA methylation. Changes in DNA methylation during later stages of development can be induced by endogenous and exogenous cues.

Previous studies in plants have focused on studying how methylation changes during development, generally focusing on early stages of development, leaving the variation of DNA methylation between organs that emerge at later stages largely unexplored. This study presents a detailed comparative analysis of the DNA methylation profiles of multiple organs of *A. thaliana* through the use of genome-wide single-base-resolution DNA methylation maps. The study can be broadly divided into three parts.

The first part of this study describes a pipeline for the analysis of DNA methylation data from next-generation-sequencing technologies. This pipeline is designed to identify methylated sites across the genome within samples and to test for statistical differences in methylation patterns between samples.

The second part of this study examines the variability of DNA methylation between leaves within an individual plant and aims to determine whether these differences in DNA methylation are a product of stochastic variation or could be explained by other factors. By comparing the methylation profiles of 18 individual leaves derived from a single plant I was able to show that the observed methylation changes between the individual leaves correlate with the time in which they emerged from the plant, where first leaves to emerge showed a relative loss of DNA methylation compared to the newly emerged leaves.

The third part of this study investigates the association of DNA methylation with organ identity, and in particular how changes in DNA methylation between organs correlate to changes in gene expression. For this purpose I produced methylation maps as well as transcriptional profiles for six different organ systems derived from three individual plants. My results show that there are genome-wide DNA methylation differences in reproductive organs compared to vegetative organs. Furthermore, some of these changes in DNA methylation correlate with changes in expression of protein coding genes. Finally, I was able to show that DNA methylation marks in regulatory regions are not sufficient to explain changes in the expression of nearby genes; instead they need to be coupled to a transposable element.

In conclusion, my thesis provides insights on the variability of DNA methylation across development. I have identified different factors such as organ identity and age that contribute to the DNA methylation variability across the genome, suggesting that DNA methylation might work as an additional layer of gene regulation during plant development.

Acknowledgments

I would like to thank,

Detlef Weigel, for providing me with the opportunity to be in one of coolest places to do science, his continuous advice, contagious passion for science and his tremendous support has made this an amazing experience.

Karsten Borgwardt and Daniel Huson who kindly agreed to be my supervisors and provided valuable advice and feedback for this work.

Kay Nieselt and Sascha Laubinger for agreeing to be part of my committee.

Claude Becker, without his guidance and his immense patience this work would have not been possible.

Rebecca Schwab, our dear lab manager, from thesis correcting, to making sure things got done. I sometimes wonder what we did before her.

Rena Stromann and Patricia Lang for helping me deal with the German language required for this thesis.

Joerg Hagmann, Jonas Mueller, and Damian Roqueiro for their insights and fruitful discussions on bioinformatics.

Daniel Koenig, Danelle Seymour, Beth Rowan, Hernan Burbano, Claude Becker, Joerg Hagmann, George Wang, Ignacio Rubio, Bonnie Fraser, Pablo Manavella and Moises Exposito for being great scientists, for their great advice, but above all for being great friends.

Beth Rowan, for all the corrections and help during the thesis writing process.

Julian Regalado and Sebastian Petersen for being Mexican.

Elena Najar, for all her support and all the fun times we had.

To all the people and staff of department 6, which made Germany feel like home.

To all the people at Google, Stack Overflow, Wikipedia and SeqAnswers for always being there to answer my questions.

To my parents and my brother, for their support and encouragement, I would not be here without them.

And finally to Rena Stromann, for ensuring that I would not starve to death, for believing in me and for always being there for me.

1 Introduction- Epigenetics

1.1 From Genetics to Epigenetics

Since ancient times, there have been significant human efforts in the study and manipulation of traits of living organisms. The word phenotype comes from the Greek *phainein*, which means to show, and *typos*, which means type, and it refers to any observable trait in living organisms. It has always been widely accepted that some proportion of the phenotypes is heritable- as the old phrase goes “Like father, like son”, but there has been much less agreement on the mechanisms by which these traits are inherited and how new phenotypes can arise. Even before understanding the scientific basis of heritability (Visscher, Hill et al. 2008), humans have inadvertently performed phenotypic selection on multiple species leading to the domestication of many crop plants and animals. Furthermore, natural selection can only act on phenotypes, and in order to understand evolution it is crucial to understand how phenotypic diversity arises.

Gregor Johann Mendel (20 July 1822– 6 January 1884) laid out the foundations needed for the development of theories to explain phenotypic inheritance and now he is regarded as the father of modern genetics. Mendel performed many hybridization studies in peas and was able to show that the inheritance of such traits followed certain rules, now known as *Mendel’s rules of segregation* (Mendel 1866). He hypothesized the existences of heritable discrete units responsible for the inheritance of phenotypes, which would be known in the future as genes.

A major breakthrough was the identification of such discrete hereditary units and their localization in the cell. The work done by Walther Flemming, Theodor Boveri (Boveri 1904), and Walter Sutton (Sutton 1902) on the characterization of chromosome and their movement during cell division led to the birth of the chromosome theory which hypothesized that chromosomes were the physical carriers of hereditary information. It was not until Thomas Hunt Morgan and his experiments on fruit flies (Morgan 1910) that direct evidence supporting the chromosome theory was available.

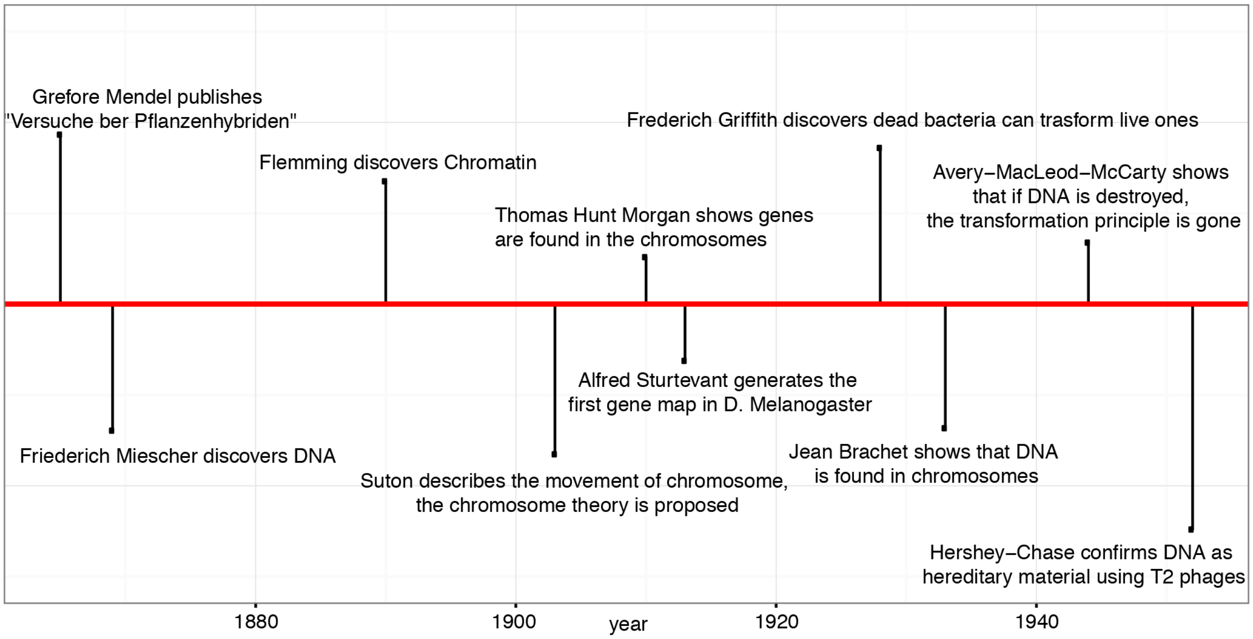


Figure 1 Timetable of relevant discoveries in the field of genetics

While the existence of informational molecules was a widely accepted fact, the scientific community was not sure what the chemical nature of these molecules was. Scientists knew that genes were carried on chromosomes, but the composition of chromosomes was known to be a heterogeneous mix of nucleic acids and proteins and as such both agents could be potentially responsible for the transformation principle. Furthermore, during this time it was known that proteins had complex structures and were chemically diverse compared to the much more stable and chemically homogenous DNA counterpart. As such, people were inclined to think that proteins were responsible for the complex behavior of genes. It took another 40 years, and the work of many notable scientists such as Frederic Griffith (Griffith 1928), Oswald Avery, Colin MacLeod and Maclyn McCarty (Avery, Macleod et al. 1944) and Alfred Hershey and Martha Chase (Hershey and Chase 1952). To show that DNA was the molecule responsible for the transmission of genetic information and end the long-standing debate on heritability, or so it was thought.

In the next 20 years there were many breakthroughs in molecular biology, such as the determination of the structure DNA (Franklin and Gosling 1953, Watson and

Crick 1953), the postulation of gene regulation (Jacob, Perrin et al. 1960) and the cracking of the genetic code (Crick, Barnett et al. 1961). During this time, scientists believed that all the information needed to understand an organism was coded in their genes, and the collection of genes within a genome determined the properties and traits of the host organism. While this was true for most of the biological traits, some traits appeared to have non-genetic components. Even before the identification of DNA as the hereditary molecule, H.J. Muller had observed that certain flies showed high phenotypical variation associated with specific DNA translocations (Muller 1930). These mutant flies had the same genetic material (arranged differently in the genome), but their phenotypes were different. This led him to hypothesize that there were forces acting on these regions, rather than acting at a single gene level. More evidence showed that the specific location of genes on the chromosome could be responsible for changes in phenotype, thus phenotype could not be explained by DNA sequence alone. A second phenomenon that this theory could not explain was the following: how can cells within an organism have different tissue identities (with unique phenotypic traits) even though they arose from the same set of embryonic cells and therefore share the same genetic code?

1.2 Epigenetics

C.H. Waddington was interested in such questions and coined the term epigenetics to address the problem in 1939 as an abstraction (Waddington 1939). He suggested of a principle of some kind affecting genes in order to reach and maintain specific cellular fate it. Similar to Sewall Wright’s Adaptive Landscape (Wright 1932), Waddington envisioned that cells would work like marbles rolling down a hill, an epigenetic landscape, (Figure 2), the marbles would be moving through grooves with branching points and the final position where the marbles landed would determine their final tissue type. Epigenetics, is derived from the prefix epi- with Greek roots επί which stands for “upon”, “over” or “above”. As the name suggests, epigenetics is focused on the study of heritable phenotypic variation that is not encoded in the DNA. Because epigenetics is defined on what it is not, rather than by what it is, the study of any heritable non-DNA unit falls under the study of epigenetics. This very broad definition of epigenetics can be problematic due to the fact that there are many different and chemically diverse epigenetic marks in nature, some of which are specific to only a limited set of organisms. Furthermore some of these can be very stable and be inherited throughout multiple generations, while others are only maintained throughout mitotic cell division but get reset in the progeny. The diversity of epigenetic marks has led to many variations of the definition of epigenetics.

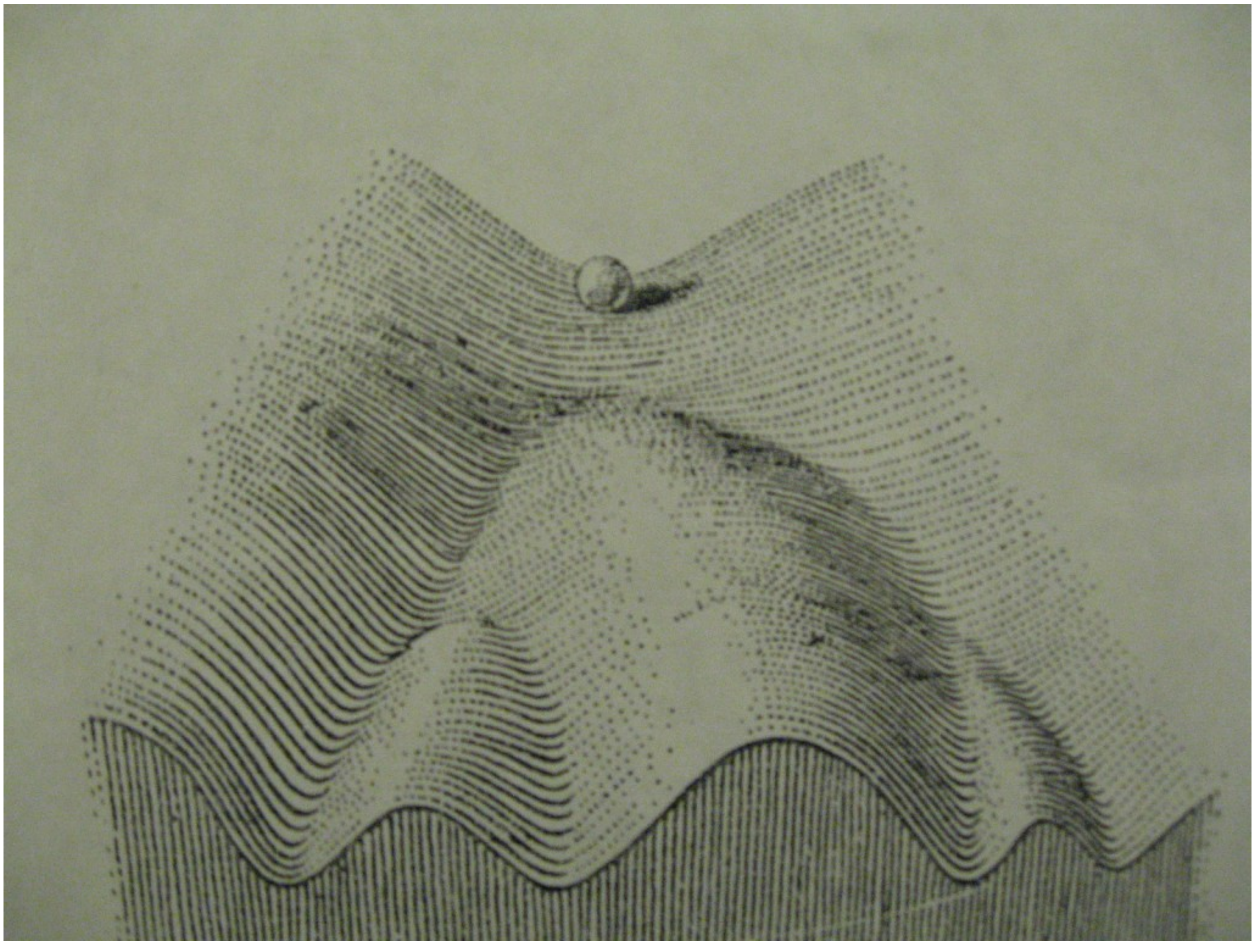


Figure 2 Waddington's epigenetic landscape. Waddington envisioned epigenetic states as a marble rolling down a hill. The marbles could traverse different grooves, which would represent different epigenetic states. Figure from (Waddington 1957).

Epigenetic marks can act at different organizational levels. The simplest epigenetic marks are direct chemical modification of nucleotide bases, which can be seen as the incorporation of non-canonical bases to the genetic code. These marks include DNA methylation among others, and have been shown to be able to affect gene expression. Another type of epigenetic marks is the modification of the packaging of DNA, such modifications are normally chemical modifications on proteins called histones that bind to DNA and are responsible for the regulation of many developmental genes (Tessarz and Kouzarides 2014). These modifications are tightly linked to DNA methylation, and their relationship will be discussed in section 1.3.3.

An important feature of epigenetic marks is that they can give rise to genetically identical but transcriptionally different cells, as in the case of neurons (Toyoda, Kawaguchi et al. 2014). While these changes are not heritable (due to the fact that some neurons won't be going through cell division), they are stable within the cell and are maintained through long periods of times.

Epigenetic changes can also be induced through endogenous or exogenous cues (Pecinka and Mittelsten Scheid 2012), which can in turn alter gene expression and give rise to new phenotypes. These environmentally induced epigenetic changes have been of particular interest to the scientific community, due to the fact that these modifications can be transmitted to future generations, providing a source of adaptation to offspring to a new environment without the need to be exposed to the new environment itself (Pecinka and Mittelsten Scheid 2012).

The field of epigenetics is a rapidly evolving field in biology, especially due to technological advances that have allowed for the characterization of epigenetic marks at a genome-wide level. While the definition used in the scientific community will most likely keep changing across time and between fields, such heritable marks are of crucial importance, especially when genetic determinism is so prominent. Even though the term epigenetics was coined to address questions regarding cell differentiation, the study of epigenetics has revealed the existence of many different epigenetic modifications involved in a wide span of biological processes. In the following section I will be focusing on one of such epigenetic modification and the main focus of this study: DNA methylation.

1.3 DNA methylation

A prominent form of epigenetic DNA modification is the addition of a methyl group to the 5th carbon of a cytosine in DNA. This chemical modification creates a non-canonical base called 5-methyl-cytosine (5mC) (Figure 3). When talking about DNA methylation, I will be referring exclusively to this particular methylation mark. Other less abundant methylated nucleotides such as N^6methyl-adenine have been described (Zhang, Huang et al. 2015), but their roles are less well understood and they were not investigated in this study. DNA methylation is an epigenetic mark involved in many diverse biological processes and can be found in both eukaryotes and prokaryotes, but is absent in some species like *Drosophila melanogaster* and *C. elegans*. Despite the abundance of organisms with DNA methylation, most of our understanding of DNA methylation, including the machinery and function, has been the product of studies in plants and animals.

The discovery of DNA methylation dates back almost as far as the identification of DNA as the hereditary material itself (Avery, Macleod et al. 1944, McCarty and Avery 1946). A couple of years after, with the use of paper chromatography, Hotchkiss (Hotchkiss 1948) was able to identify a chemically modified nucleotide in calf thymus and hypothesized that the modification was methylation in cytosine nucleotides. While scientists speculated that DNA methylation could be regulating gene expression, it was not until many years later that its function was shown (Holliday and Pugh 1975). Some of the early studies of methylation used 5-aza-cytidine, which acts as an inhibitor of DNA methylation, to show the functional importance of DNA methylation. For example, treatment of mouse embryo cells with 5-aza-cytidine has been shown to induce cell differentiation. Interestingly, after the chemical treatment was suspended, the differentiated state would persist, even after cell division (Taylor and Jones 1979). These were some of the first studies suggesting that DNA methylation played an important role in gene regulation and cell identity.

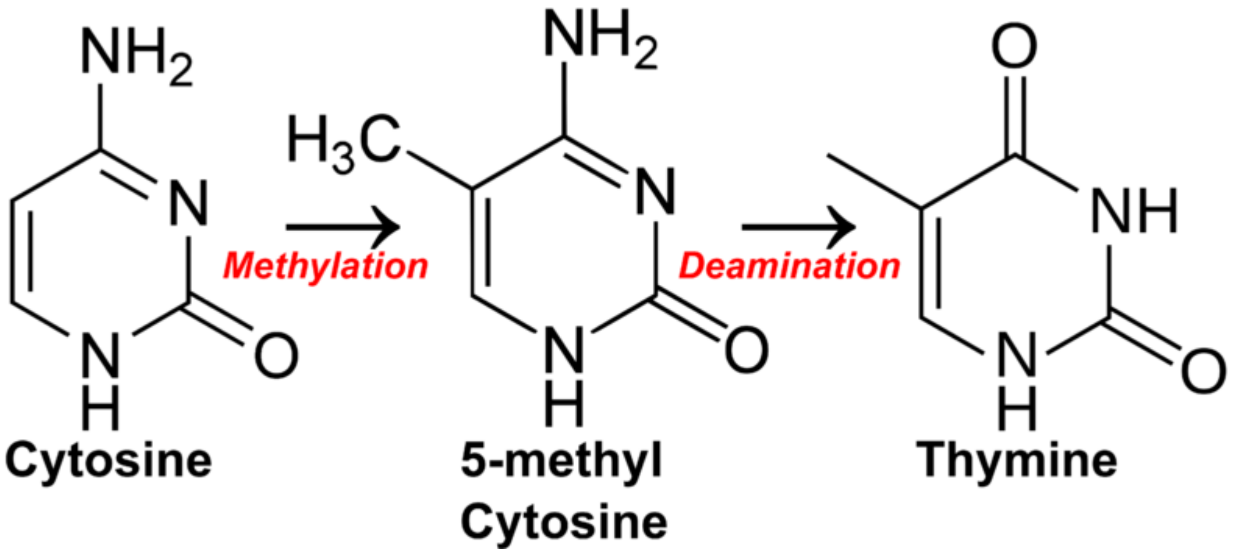


Figure 3 Chemical relationships between cytosine, 5-methyl-cytosine and thymine (figure from CGCF Wikimedia).

DNA methylation is established by the addition of methyl groups to cytosines by a family of enzymes called *methyltransferases.* These proteins use S-Adenosyl methionine as a methyl donor to establish new methylation marks on previously unmethylated DNA, or to maintain DNA methylation after DNA replication.

Even though methylation in plants and animals share many similarities, the proteins and pathways involved in the establishment and maintenance of DNA methylation are not all the same. Because they can differ in functions, I will describe first the functions of DNA methylation in both plants and animals separately, and then discuss the mechanisms for maintenance and establishment of DNA methylation.

1.3.1 DNA methylation in mammals

In mammals DNA methylation is not distributed equally across all cytosines in the genome, rather it is found almost uniquely in cytosines followed by a guanine. This type of methylation is called CG-methylation or CpG methylation (the term CpG is normally used in animal studies and it is used to denote “Cytosine phosphate Guanine” in order to avoid confusion with cytosine-guanine pairing). In general, most of the CG-sites in mammalian genomes (70-80%) are methylated (Ehrlich, Gama-Sosa et al. 1982). Unmethylated CG sites are normally localized in gene promoters, in regions called *CpG* *islands* (Suzuki and Bird 2008). These CpG islands are characterized by a higher occurrence of CG sites than would be expected by chance.

DNA methylation has been extensively studied in mice and humans, and has been shown to be important for development. Reprogramming events on DNA methylation including genome-wide gains and losses of methylation occur during early stages. For the formation of the germline, cells must go through an erasure of somatic epigenetic signatures and then go through an establishment of sex-specific epigenetic marks or *genomic imprinting* (Messerschmidt, Knowles et al. 2014). Imprinting is a process in which alleles are marked depending on their origin, either paternal or maternal, allowing for the expression of genes in a parent-of-origin-specific manner (Reik and Walter 2001).Even though these reprogramming events occur in both parental germlines, they occur at different time points, resulting in different levels of methylation in the sexual gametes (Figure 4). After fertilization, sperm cells show high methylation levels compared to the egg cell. This difference in methylation rapidly changes; the paternal gametes go through active demethylation and the maternal genome goes through a rapid wave of *de novo* DNA methylation (Figure 4). Most of the mutations causing disruption in the establishment of DNA methylation are embryonic lethal. Deficiency in DNA methylation pathways has also been associated to diseases in humans, such as ICF syndrome (Hansen, Wijmenga et al. 1999), Rett syndrome (Amir, Van den Veyver et al. 1999) and cancer predisposition (Barrow and Michels 2014). Both, losses and gains, of methylation can cause diseases, highlighting the importance of understanding the effects of DNA methylation.

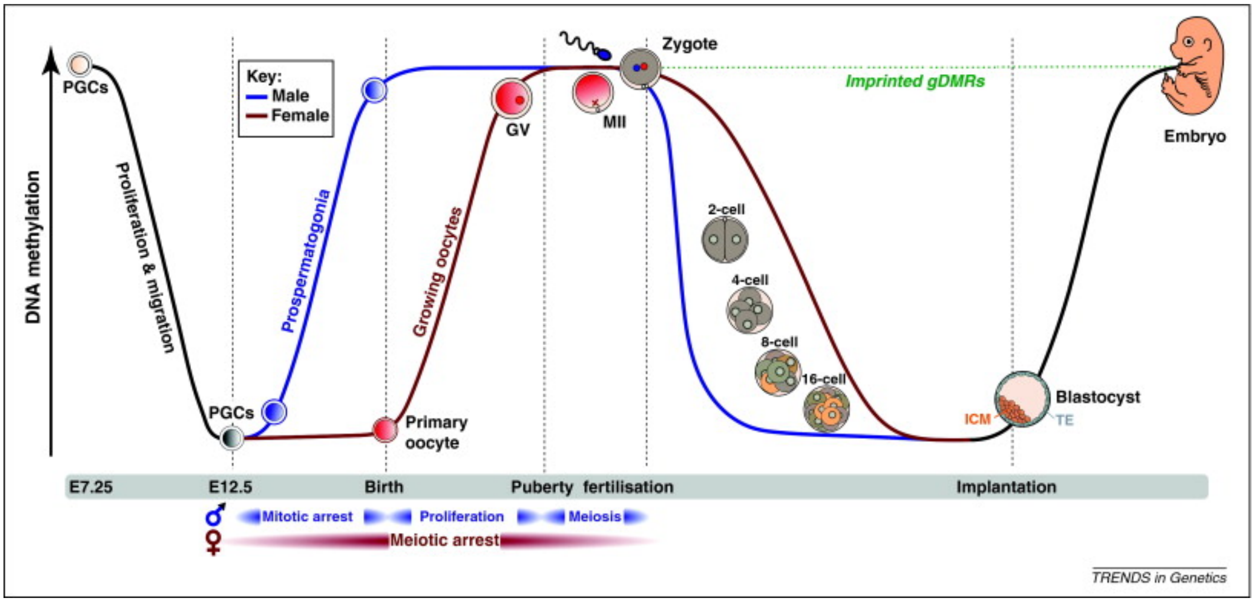


Figure 4 DNA Methylation changes through development in mammals. Primordial germ cell (PCGs) emerge at embryos at E7.5, DNA methylation is globally erased (black line). Following sex-determination, new DNA-methylation landscapes are established in germ-cell precursors in an asymmetrical fashion in male and female embryos. Following fertilization, a new wave of DNA demethylation takes place that is distinct on the parental genomes. In the zygote, DNA methylation of the paternal genome is rapidly erased by an active mechanism (blue line). Demethylation of the maternal genome is slower (red line) and is dependent on DNA replication (passive demethylation). Legend and figure from (Smallwood and Kelsey 2012).

Besides embryo development, DNA methylation plays a crucial role in establishing and maintaining tissue identity. Embryonic stem cells (ESC) with defects in DNA methylation pathways remain viable and are able to divide, but they are no longer able to go through correct cell differentiation (Jackson, Krassowska et al. 2004, Tsumura, Hayakawa et al. 2006). There have been identified multiple loci in ESC where factors associated to pluripotent states are hypo-methylated but after going through cell differentiation they become methylated (Farthing, Ficz et al. 2008). Furthermore, genome wide studies in DNA methylation have shown that tissues in humans show different methylation profiles, where changes in methylation between tissues correlate with changes in gene expression (Schultz, He et al. 2015), suggesting that DNA methylation is not only important for the maintenance of differentiated states but also for gene regulation.

DNA methylation is also essential for the control and silencing of transposable elements (TEs) in most of the eukaryotic species (Zemach, McDaniel et al. 2010). Mice lines with deficiencies in methylation pathways are unable to methylate TEs correctly, leading to TE reactivation (Bourc'his and Bestor 2004).

1.3.2 DNA methylation in plants

A major difference between DNA methylation in plants compared to animals is that it can occur in three different sequence contexts: CG, CHG and CHH, where H can be any non-G nucleotides. Contrary to mammals, most of the methylation in plant genomes is restricted to TEs as well as repetitive sequences and centromeric regions (Feng, Cokus et al. 2010, Zemach, McDaniel et al. 2010). Similar to its function in animals, DNA methylation in plants is crucial for TE silencing. Studies in mutants, which have defective pathways in charge of the establishment of DNA methylation, show reactivation of transposable elements (Miura, Yonebayashi et al. 2001, Singer, Yordan et al. 2001). Studies in pollen grains have shown that there are differences in methylation levels between the three cells that they are composed of (2 sperm cells and 1 vegetative) (Slotkin, Vaughn et al. 2009). The vegetative cell has low levels of methylation and an increased level of transposon activity (Zhao, Rank et al. 2009). In contrast, the sperm cells show high levels of DNA methylation causing silencing of transposable elements (Huh, Bauer et al. 2008). It has been hypothesized that the reactivation of transposon in the vegetative cell leads to an increase in 21nt RNA molecules that migrate to the sperm cells and reinforce methylation by a pathway dependent on small RNAs. This pathway, known as the RdDM pathway will be described in detail in section 1.4.2. Only the sperm cell will be contributing with DNA to the next generation; therefore their genome integrity is of higher importance.

DNA methylation in plants is also important for gene imprinting. For example *FLOWERING WAGENIGEN (FWA)* is a transcription factor that is expressed during seed development in the endosperm but is silenced in other tissues. *FWA* is regulated through the methylation of tandem repeats in its promoter, which need to be hypo-methylated in order to be transcribed. A set of mutants was identified that showed ectopic expression of *FWA* compared to the wild-type plants. Even though the *FWA* mutants and the corresponding wild type show identical genetic sequences at the *FWA* locus, but the repeat region in the promoter of the gene was lacking methylation in the mutant lines allowing the gene to escape silencing (Soppe, Jacobsen et al. 2000). This phenotype can also be induced by a *DEFICIENT IN DNA METHYLATION 1 (ddm1)* mutant, which is deficient in the establishment of DNA methylation (Soppe, Jacobsen et al. 2000). *FWA* is expressed in a parent-of-origin-specific way. This is achieved by DEMETER (DME), a protein (Kinoshita, Miura et al. 2004) that demethylates only the maternal genome (Kawashima and Berger 2014) allowing FWA to be transcribed.

Another identified function of DNA methylation, similar to mammals, is the regulation of developmental-associated genes. Studies in maize using mutants with defects in the establishment of DNA methylation show strong developmental defects and affect the expression of tissue-specific genes, which can control developmental phase transition and sex determination (Parkinson, Gross et al. 2007, Erhard, Stonaker et al. 2009).

Most of the examples outlined above are instances where DNA methylation in or near regulatory regions correlated is correlated with reduced gene expression. Not all methylation is associated to a reduction in gene expression. Methylation can also be found in the exons of genes (gene body methylation), this type of methylation is generally associated to constitutively expressed genes, as well as genes that show stable expression levels across tissues (Zhang, Yazaki et al. 2006, Zilberman, Gehring et al. 2007). While the mechanisms in which DNA methylation is regulating expression in such manner are not entirely clear, it has been hypothesized that DNA methylation is causing this pattern through the interaction with another type of epigenetic marks called histones which will be discussed in the next section (Coleman-Derr and Zilberman 2012).

1.3.3 DNA methylation and other epigenetic marks

In eukaryotic cells, DNA is wrapped around octameric complexes made of proteins called histones. These complexes, known as a nucleosomes, are composed of 8 histone core proteins (two of each the histone families H2A, H2B, H3 and H4), and a stretch of DNA of ~147 nucleotides surrounding the histone cores (Kornberg 1974). Nucleosomes can affect the spatial structure of DNA. Genomic regions with a tight packaging of nucleosomes (heterochromatin) are transcriptionally silent due to the poor accessibility of the transcription machinery to the packaged DNA. By contrast, less compacted regions (euchromatin) tend to be transcriptionally active (Venkatesh and Workman 2015).

Some of the changes in configuration are caused by chemical modification of the amino acids of histone tails such as phosphorylation, methylation and acetylation (Venkatesh and Workman 2015). The modification of histone tails and the presence of DNA methylation are intimately connected. In animals the DNA methyltransferase-like 3 (DNMT3L), a protein involved in the establishment of DNA methylation, interacts with the 4th lysine (unmethylated) of the H3 histone (H3K4)(Ooi, Qiu et al. 2007). In plants and animals, methylation of this residue is correlated negatively to CG-DNA methylation levels (Fournier, Goto et al. 2002, Zhang, Bernatavichute et al. 2009). In plants dimethylation or trimethylation of the 9th residue (lysine) of the H3 histone (H3K9me2 or H3K9me3) is highly correlated with CHG-DNA methylation in a reinforcing loop (Johnson, Bostick et al. 2007) (Bernatavichute, Zhang et al. 2008). In mutants lacking a functional KRYPTONITE protein, a histone methyltransferase involved in the establishment of the H3K9m2, reduction of global levels of CHG methylation can be observed (Jackson, Lindroth et al. 2002). More recent studies have been able to determine the structure of KRYPTONITE coupled to methylated DNA, show how DNA methylation helps recruit KRYPTONITE (Du, Johnson et al. 2014).

In addition to the wide variety of chemical modification of histones, in eukaryotes there are multiple copies of some histone cores, which differ in their amino acid sequence and function. The combination and ratio of these histone variants can confer different properties to the nucleosomes (Smolle and Workman 2013). Furthermore, there is evidence of a correlation between DNA methylation and histone variants (Cedar and Bergman 2009).

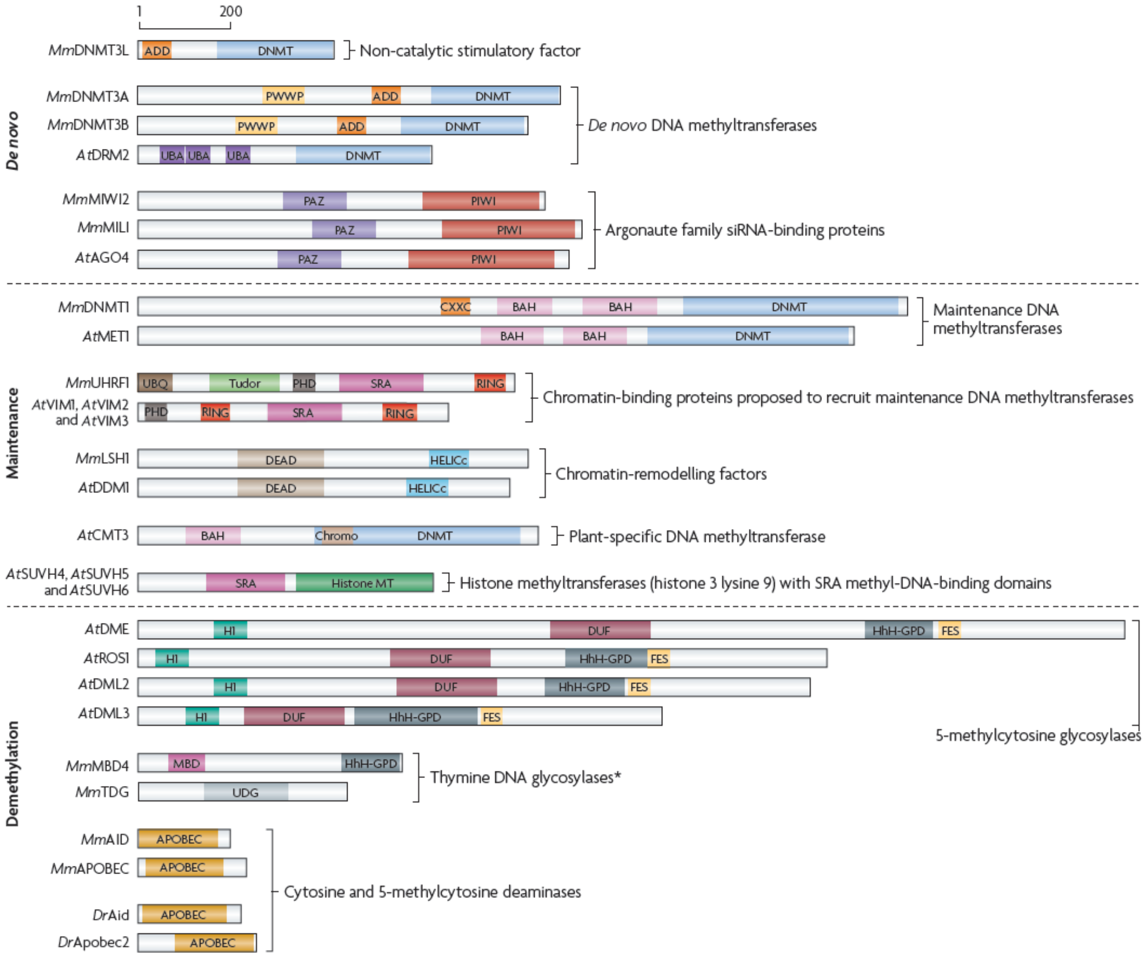
For example, one of the variants of the histone H2A is the histone H2A.Z. This histone has a 60% amino acid similarity to the histone H2A, and compromises 15% of the total amount of H2A histones in the genome of yeast (Zlatanova and Thakar 2008). This variant is commonly found near transcription start sites (TSS) and is associated to genes with either high or low expression levels. DNA methylation correlates negatively with the presence of histone variant H2A.Z (Guillemette and Gaudreau 2006). This association has led to the hypothesis that DNA methylation can regulate gene expression by modulating which histone variants are used near TSS. While the exact relationship of DNA methylation to histone methylation is not completely understood, many of the identified components of DNA methylation pathway interact with proteins involved in the establishment of histone modifications and will be discussed in the next section.

1.4 Establishment, maintenance and removal of DNA methylation

In both plants and animals multiple pathways have been identified that can modify the methylation in DNA. These molecular mechanisms are responsible for the propagation of DNA methylation after cellular division, as well as the establishment and removal of DNA methylation marks. In this section I will summarize the current knowledge about the different pathways involved in the regulation of DNA methylation.

1.4.1 Symmetrical DNA methylation

CHG and CG sequences are termed symmetrical in double stranded DNA (dsDNA), as the same configuration of bases is found in the complementary strand. In symmetrical sequence contexts, cytosine methylation is usually found on both strands of the DNA. During DNA replication, the newly synthesized strand of DNA is not methylated; this partially methylated dsDNA is termed hemi-methylated DNA. Some DNA methyltransferases can interact with hemi-methylated DNA and restore the methylation marks in the unmethylated newly synthesized strand, thereby ensuring that DNA methylation is faithfully transmitted after DNA replication.


1
200
MmDNMT3L
ADD
DNMT
Non-catalytic stimulatory factor
De novo
MmDNMT3A
PWWP
ADD
DNMT
MmDNMT3B
PWWP
ADD
DNMT
AtDRM2
UBA
UBA
UBA
DNMT
De novo DNA methyltransferases
MmMIWI2
PAZ
PIWI
MmMILI
PAZ
PIWI
AtAGO4
PAZ
PIWI
Argonaute family siRNA-binding proteins
Maintenance
MmDNMT1
CXXC
BAH
BAH
DNMT
AtMET1
BAH
BAH
DNMT
Maintenance DNA methyltransferases
MmUHRF1
UBQ
Tudor
PHD
SRA
RING
AtVIM1, AtVIM2 and AtVIM3
PHD
RING
SRA
RING
Chromatin-binding proteins proposed to recruit maintenance DNA methyltransferases
MmLSH1
DEAD
HELICc
AtDDM1
DEAD
HELICc
Chromatin-remodelling factors
AtCMT3
BAH
Chromo
DNMT
Plant-specific DNA methyltransferase
AtSUVH4, AtSUVH5 and AtSUVH6
SRA
Histone MT
Histone methyltransferases (histone 3 lysine 9) with SRA methyl-DNA-binding domains
Demethylation
AtDME
HI
DUF
HhH-GPD
FES
AtROS1
HI
DUF
HhH-GPD
FES
AtDML2
HI
DUF
HhH-GPD
FES
AtDML3
HI
DUF
HhH-GPD
FES
5-methylcytosine glycosylases
MmMBD4
MBD
HhH-GPD
MmTDG
UDG
Thymine DNA glycosylases*
MmAID
APOBEC
MmAPOBEC
APOBEC
DrAid
APOBEC
DrApobec2
APOBEC
Cytosine and 5-methylcytosine deaminases

Nature Reviews | Genetics

Figure 5 Cartoon representation of select mouse (Mm), Arabidopsis (At), and Zebrafish (Dr) proteins involved in maintenance methylation, de novo methylation, and demethylation. Original figure from (Law and Jacobsen 2010).

In mammals, the protein responsible for the maintenance of CG methylation is DNA methyltransferase 1 (DNMT1)(Yen, Vertino et al. 1992). The exact mechanism by which DNMT1 methylates cytosines is not completely understood, but studies have shown evidence of interaction between DNMT1 and proteins involved in DNA replication as such as proliferating nuclear antigen (PCNA)(Schermelleh, Haemmer et al. 2007), chromatin associated proteins like the ubiquitin-like plant homeodomain and RING finger domain (URF1) (Bostick, Kim et al. 2007) and chromatin remodeling factors such as the lymphoid-specific helicase1 (LSH1)(Dennis, Fan et al. 2001).

The homologous protein in plants (Figure 5), METHYL TRANSFERASE 1 (MET1) (Vongs, Kakutani et al. 1993), can also methylate CG sites, and uses accessory proteins homologous to the ones interacting with DNMT1 in mammals (Hirochika, Okamoto et al. 2000). This suggests deep conservation of the molecular pathway mediating maintenance of CG-methylation.

In plants, maintenance of CHG methylation requires the methyltransferase CHROMOMETHYLASE 3 (CMT3) (Lindroth, Cao et al. 2001). CMT3 interacts with SUPRESSOR OF VARIEGATION 3-9 HOMOLOGUE 4 (SUVH4, also known as KRYPTONITE), which is a histone methyltransferase in charge of the establishment of H3K9m2. DNA methylation guides SUVH4 for histone methylation (Du, Johnson et al. 2014), and CMT3 is associated to H3K9 methylation (Du, Zhong et al. 2012) creating a self-reinforcing loop.

1.4.2 De-novo DNA methylation

Contrary to DNA methylation in symmetrical sequence contexts, methylation at CHH sites is present only on one strand, and therefore one of the newly synthesized double strands after DNA replication will not have any methylation that could be used as template for DNA methylation. In this case, methylation needs to be (re-) established without a template of hemimethylated DNA to guide DNA methylation. This process is referred to *de novo* DNA methylation. The establishment of d*e novo* DNA methylation is not limited to maintenance of DNA methylation, as it also refers to the establishment of methylation in previously unmethylated genomic regions.

In both animals and plants, pathways have been identified that can establish *de novo* DNA methylation through the use of small RNAs. In plants, *RNA-directed DNA Methylation* (*RdDM*) (Wassenegger, Heimes et al. 1994) has been shown to trigger *de novo* DNA methylation (Matzke and Mosher 2014). This pathway includes two plant specific RNA polymerases (Pol IV and Pol V), members of the RNA interference (RNAi) pathway, DICER-like 3 (DCL3) and ARGONAUTE 4 (AGO4), as wells as small interfering RNAs (siRNA). Pol IV is involved in the synthesis of single-stranded RNA transcripts, which are converted to double-stranded RNA (dsRNA) by RNA-

DEPENDENT RNA POLYMERASE (RDR2). This is followed by the cleavage of dsRNA by DCL3 into 24nt siRNA products. The methyltransferase HUA ENHANCER 1 (HEN1) creates mature siRNAs by methylating the 3' ends of the siRNAs. The methylated siRNAs are loaded to proteins of the ARGONAUTE family, AGO4, AGO6, and AGO9. These protein-RNA complexes migrate to the nucleus where they interact with the RNA transcripts produced by the Pol IV. Such interaction is mediated by the base complementarity of the siRNA to the Pol IV transcripts. This complex is recruited by interacting with KOW DOMAIN-CONTAINING TRANSCRIPTION FACTOR 1 (KTF1). Finally, RNA-DIRECTED DNA METHYLATION 1 (RDM1) helps couple AGO4 with the protein DOMAINS REARRANGED METHYLTRANSFERASE 2 (DRM2), which then gives rise to *de novo* DNA methylation at the target site.

In mammals there is an equally complex pathway involving the use of small RNAs for the establishment of *de novo* DNA methylation. The piwi proteins are part of the ARGNOAUTE super family (see above) and interact with small RNAs of variable length (typically 24-32 nt) called Piwi-interacting RNAs (piRNAs). A high proportion of the identified piRNAs map to TEs and play an important role on the silencing of transposable elements (Kuramochi-Miyagawa, Kimura et al. 2004, Aravin, Hannon et al. 2007). Furthermore mutants on the piwi proteins *MILI* or *MIWI1* fail to establish *de novo* DNA methylation at some TEs, suggesting that the piRNA containing complexes serve to guide DNA methylation (Aravin, Sachidanandam et al. 2008, Kuramochi-Miyagawa, Watanabe et al. 2008), serving a similar role to the RdDM pathway in plants.

1.4.3 DNA demethylation

DNA methylation can be lost through passive or active mechanisms. If maintenance of DNA methylation does not take place after DNA replication, half of the methylation is lost after each replication cycle. Active DNA demethylation is achieved through proteins that remove methylation and thus is independent of DNA replication.

Active demethylation in plants is carried out by glycosylases such as REPRESSOR OF SILENCING 1 and 3 (ROS1 and ROS3), DME and DEMETER-LIKE 2 and 3 (DML2 and DML3) (Law and Jacobsen 2010). These proteins excise methylated bases from dsDNA. DNA repair mechanisms then re-replace the missing base with an un-methylated cytosine, effectively removing methylation from the DNA. Mutations in DNA glycosylases cause a general increase of DNA methylation across the genome (Gong, Morales-Ruiz et al. 2002). Active demethylation can have different biological roles. The DME protein is expressed during gametogenesis and is necessary for imprinting (Choi, Gehring et al. 2002). By contrast, ROS1, DML2 and DML3 (Gong, Morales-Ruiz et al. 2002, Ortega-Galisteo, Morales-Ruiz et al. 2008) are expressed in somatic tissue and target a similar set of sites, suggesting that they have overlapping functions. Their preferred sites are methylated TEs and gene ends (Penterman, Zilberman et al. 2007). While the particular role of ROS1 is not known, *ros1* mutants show increased methylation across the genome and reduced gene expression at such hyper-methylated loci (Zhu, Kapoor et al. 2007). These results suggest that ROS1 is important on removing the accumulation of DNA methylation. It has also been suggested that ROS1 could be triggered through stress, activating genes that under normal circumstances would be silent.

In mammals, the ten-eleven-translocation (TET) proteins have been identified as key proteins for DNA demethylation. This family of proteins is involved in the modification of 5mC residues into other non-canonical residues such as 5-hydroxymethylcytosine (5hmC), 5-formylmethylcytosine (5fC) and 5-carboxylmethylcytosine (5caC)(Tahiliani, Koh et al. 2009, He, Li et al. 2011, Ito, Shen et al. 2011). These modifications can cause a loss of methylation in both active and passive ways. Proteins in charge of the maintenance of DNA-methylation have much lower affinities to 5-hmC (Valinluck and Sowers 2007), causing an improper maintenance of DNA methylation, and leading to a passive loss of DNA methylation through DNA replication. Additionally thymine DNA glycosylase (TDG) coupled to base excision repair can actively replace 5-hmC residues with unmethylated cytosines.

In mammals, there are highly dynamic transitions of DNA-methylation patterns in the genomes of cells that give rise to sexual gametes (known as primordial germ cells (PGCs)), and early stage embryos. These cells have low levels of DNA methylation (Kafri, Ariel et al. 1992) and are caused by both passive and active demethylation mechanisms (Smith and Meissner 2013). In PGCs, DNA methylation is reset between the embryonic day 10.5 and 13.5 followed by a wave of *de novo* methylation in both male and female gametes as shown Figure 4. Studies in mouse have shown that PGCs can lose up 90% of the total DNA methylation (Popp, Dean et al. 2010). This resetting of DNA methylation is crucial for the removal of imprinting in genes (Reik, Dean et al. 2001). During embryo development, the paternal genome undergoes a transition within the first hours of fertilization, where remodeling of chromatin takes place followed by genome-wide erasure of DNA methylation by active demethylation (Mayer, Niveleau et al. 2000). DNMT1 (a protein responsible for the maintenance of DNA methylation) is excluded from the nucleus during cell division, resulting in passive DNA demethylation in the genome (Carlson, Page et al. 1992).

1.5 Sources of DNA methylation variability

Due to the fact that DNA methylation can modify gene expression and thus be a source phenotypic variability, there has been great interest in identifying the factors that can affect DNA methylation in the genome. In this section, I will present some of the sources that can give rise to epigenetic variability.

1.5.1 Genetically induced variation DNA methylation

William Bateson and Caroline Pellew discovered in 1915one of the first reported cases of a phenomenon that became to be called paramutation later*.* Some pea individuals displayed a “rogue” phenotype with curved pots and pointy leaflets. Such individuals when crossed with normal looking plants, produced offspring displaying the “rogue” phenotype. Furthermore, all the offspring deriving from the self-fertilization of the F_1 displayed the “rogue” phenotype (Bateson and Pellew 1915). Such behavior was puzzling as no normal phenotype was found in the F_2, thus violated Mendel’s laws of segregation.

Thirty years later, work on pigmentation in maize focusing on the *red1* locus (*r1*) described in more detail such phenomena. The cross of plants containing the normal pigment *R-r* allele with plants containing either the *R-marbled* (*R-mb*) or R-*stippled* (*R-*

st) allele, which display a mottled phenotype, resulted in offspring having a mottled phenotype as well (Brink 1956). Furthermore, this cross caused a modification of the normal *R-r* allele which now, despite being genetically identical to the original one, would also cause mottled phenotype (Brink 1956). This modification, which could be transmitted meiotically, was termed paramutation which is defined as an interaction between two alleles where one allele (paramutagenic) can induce a heritable epigenetic modification on the other one (paramutant). Such modification can persist even if the paramutagenic allele is no longer present. Furthermore paramutant can be propagated through meiosis and can be a source of phenotypic variation in genetically identical individuals.

While the molecular mechanisms that give rise to paramutation are still not entirely understood, there are some clues regarding the machinery involved. One protein that has been identified as necessary for paramutation in maize is MEDIATOR OF PARAMUTATION1 (MOP1), which is an RNA polymerase associated with the production of siRNAs (Sidorenko and Chandler 2008), suggesting that the RdDM pathway might be involved in such phenomena.

1.5.2 Environmentally induced changes of DNA methylation

Plants have very limited mobility and as such they need to be able to adapt quickly to an ever-changing environment. While the majority of this phenotypic plasticity needed to adapt to new environments is encoded in the genome, studies have shown that epigenetic marks such as DNA methylation or histone modifications can contribute to the responses to the environment (Dowen, Pelizzola et al. 2012, Secco, Wang et al. 2015). These epigenetic changes can also work as a heritable molecular memory, allowing the offspring of a stressed individual to be pre-adapted to a new environment, even though they have never experienced the stress. Many examples exist of epigenetic stress adaptation including salt stress, cold stress (Seymour, Koenig et al. 2014), biotic stress (Dowen, Pelizzola et al. 2012, Gutzat and Mittelsten Scheid 2012), and phosphate starvation (Secco, Wang et al. 2015, Yong-Villalobos, Gonzalez-Morales et al. 2015) among others.

The duration of stress memory in the progeny of stressed individuals varies. Some studies have shown that a single unstressed generation time is sufficient to reset the epigenetic state to pre-stress levels (Secco, Wang et al. 2015). Other studies could still measure the stress preadaptation after multiple generations (Iwasaki and Paszkowski 2014). It is worth noting that in the direct progeny of stressed plants, maternal effects could be the cause of apparent trans-generational adaptation, rather than trans-generational memory. In mammals, female gametes and progenitor cells of the male gametes are already present in the unborn offspring. Stress on a pregnant individual could therefore lead to epigenetic changes in the unborn offspring as well as in its germline. Consequently the third generation could also show an epigenetic adaptation, but such adaptation can be a direct result from the initial stress exposition rather than stress memory passing through meiosis.

1.5.3 Spontaneous variability

Due to the importance of DNA methylation in the regulation of biological processes, there has been a great interest in determining the rate at which DNA methylation varies across time. There are several examples where spontaneous changes of DNA methylation affect the expression of a gene. One of the best-characterized examples of such behavior in DNA methylation was ripening of an epimutant in tomato. The *Colorless non-ripening* locus (*Cnr*) encodes a regulator of fruit ripening in tomato, and its expression is regulated by DNA methylation (Manning, Tor et al. 2006). The epimutant showed stable methylation at the *Cnr* locus and produced fruits that never ripened (Figure 6). Adding 5-azacytidine, a known inhibitor of DNA methylation, to such plants generated a loss of methylation causing early ripening. Interestingly, individual cells in the epimutant could stochastically switch to an unmethylated state, causing partial ripening of the fruit.

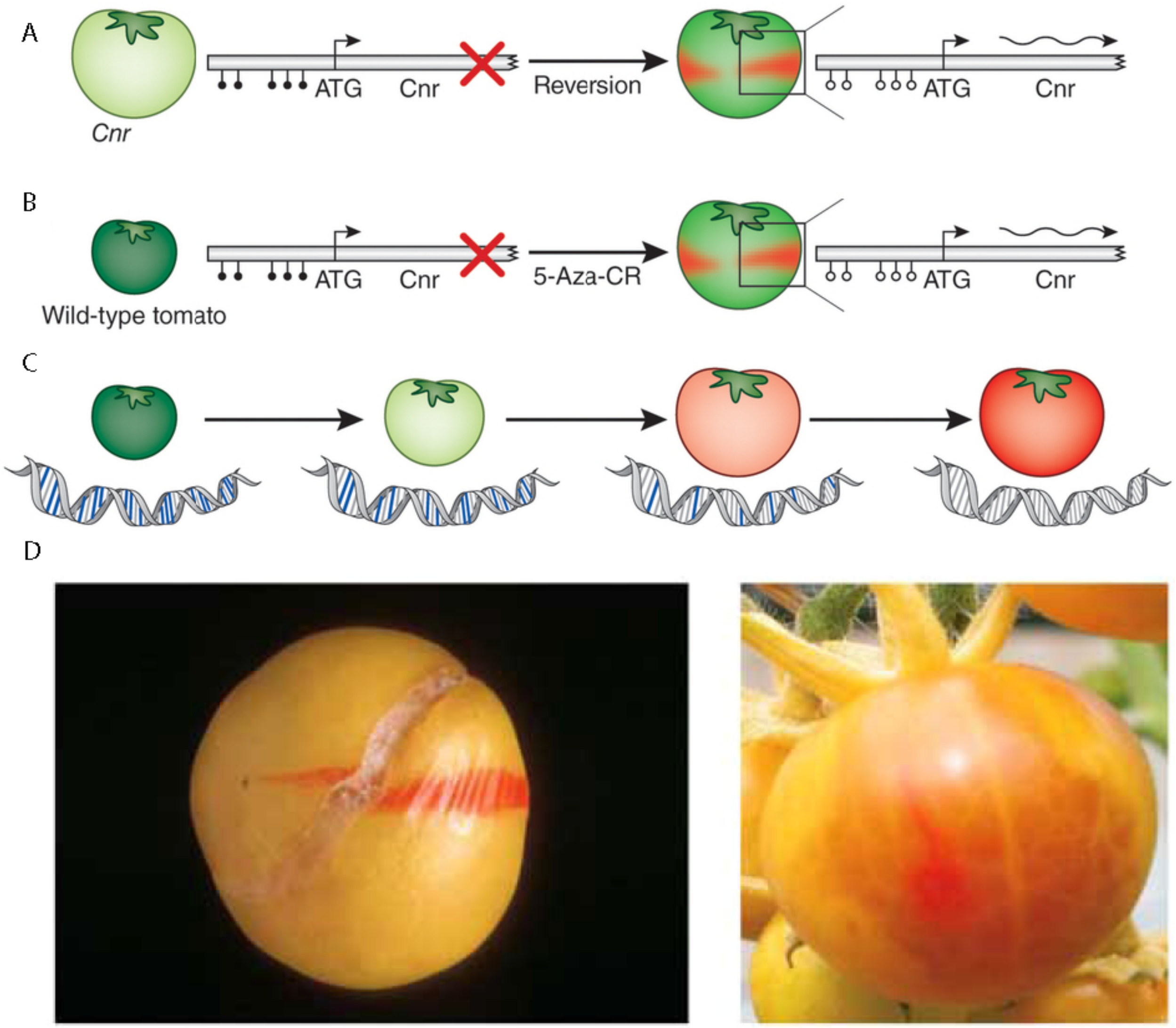


Figure 6 Random loss of methylation leads to expression across tomato. A) Spontaneous loss of DNA methylation leads to the expression of Cnr. B) the addition of 5-azacytidine leads to the demethylation of the Cnr locus, leading to expression. C) Losses of methylation leads to the ripening of tomato. D) Effect of the expression of CNR in tomato. Figure modified from (Ecker 2013) and (Manning, Tor et al. 2006).

Epimutant stability, such as the natural reversions found in tomato, have shed light on the dynamic nature of DNA methylation as an epigenetic mark. In order to understand the impact of DNA methylation on genome regulation as a whole, quantifying such spontaneous variability in the genome is necessary. With the availability of high throughput sequencing, it is now possible to study DNA methylation across the whole genome (Cokus, Feng et al. 2008, Lister, O'Malley et al. 2008). This has allowed the quantification of variability of DNA methylation across generations in some organisms such *Arabidopsis thaliana* (Becker, Hagmann et al. 2011, Schmitz, Schultz et al. 2011). Both groups used a set of individuals that originated from a founder plant 30 generation back, and have been propagated through single seed descent since (Shaw, Byers et al. 2000). By generating single base resolution methylation maps (Cokus, Feng et al. 2008, Lister, O'Malley et al. 2008) of several individuals from generation 3 and 30, the rate of change of methylation per generation (which I will be referring as the *epimutation rate* for the remaining of the text) could be measured. Much like DNA mutations, spontaneous changes at single positions where heritable and could persist in the following generations. This epimutation rate was determined to be over four orders of magnitude greater than the mutation rate of DNA (10^{-4} against 10^{-8}) when plants where grown in a stable and controlled environment. Furthermore, the epimutation rate didn’t seem to be the same for all positions in the genome. Some positions showed particularly high turnover rates between individuals. A second observation from these studies is that even though single sites show a high amount of variation, methylation across larger contiguous regions showed less variability. The variability of DNA methylation in those regions was estimated to be similar to DNA mutation rates, suggesting that methylation at a region level might play a stronger role in evolution than individual sites.

A similar study was performed using *Arabidopsis thaliana* wild populations across the East coast of the United States of America. The colonization event of *A. thaliana* in this area is believed to have happened around 300 years ago (Exposito-Alonso, Becker et al. 2016); therefore the variation of DNA methylation between these populations would be the product of the accumulation of naturally induced variants during this time (Hagmann, Becker et al. 2015). Strikingly, the epimutation rate of these wild populations was estimated to be very similar to the epimutation rate found in the laboratory-grown MA lines. These results suggest that while DNA methylation might provide a strategy for short-term adaption, DNA methylation is not likely to be a source of variation under strong selective pressure.

While the mechanisms that give rise to this type of variation remain elusive, there are some patterns found in such variation. First of all variable sites tend to be away from TEs and loci targeted by siRNAs, where methylation seems to be stable. This suggests that the density of TE's in a genome can affect the variability of DNA methylation. Additionally sites with high DNA methylation variability in natural populations had also high variability in green house conditions, suggesting that some sites in the genome are particularly prone to epimutations (Becker, Hagmann et al. 2011, Hagmann, Becker et al. 2015). Some of the epimutations identified in the transgenerational studies overlap with epimutations found in mutants on the DNA methylation maintenances machinery, suggesting that part of this spontaneous variation could be due to a failure in the maintenance of DNA methylation (Lister, O'Malley et al. 2008, Schmitz, Schultz et al. 2011).

1.6 Conclusion

DNA methylation is an epigenetic mark with a broad range of functions and a close connection to other epigenetic marks such as histones. DNA methylation is essential for the silencing of some genes, including transposable elements and is involved in the regulation of multiple developmental associated genes. There are many identified mechanisms in both plants and animals that can dynamically regulate DNA methylation. These include mechanisms in plants and animals to establish, maintain and remove methylation marks. Some of these mechanisms are shared between the two kingdoms, for example both have homologous methyltransferases that establish DNA methylation in hemi-methylated DNA, and they both use small RNAs to establish *de novo* DNA methylation.

While many of the pathways and their components have been elucidated, the dynamics of DNA methylation in the genome are still not well understood. Even though DNA methylation is a stable heritable mark, there are many sources that can alter methylation patterns. Some of the identified sources of variability are stochasticity, response to external stimuli, and genetic variability. All these attributes of DNA methylation serve to highlight why the study of DNA is an important and challenging topic.

2 Introduction-Next generation sequencing

As described in section 1.3, DNA methylation can affect gene expression, thereby constituting a source of phenotypic variability. Furthermore methylation can vary between populations, between individuals, or even within an individual's tissue types (Becker, Hagmann et al. 2011, Schultz, He et al. 2015, Exposito-Alonso, Becker et al. 2016). In order to understand the role of DNA methylation it is important to be able to measure these differences at a genome-wide scale. This section outlines the technological developments, focusing in particular on DNA sequencing technologies, which now allow for the study of DNA methylation at a genome wide level as well as its association to gene expression.

One of the major questions in the study of biology has been centered on deciphering the mechanisms that give rise to phenotypic variability in nature. With the help of DNA sequencing, it is now clear that genes and DNA mutations are the responsible for the majority of the phenotypic diversity. The "first generation" of sequencing technologies used Sanger's chain termination method for DNA sequencing, (Sanger, Nicklen et al. 1977), such method could only sequence a couple of kilobases of DNA but was suitable for the sequencing of entire genes. Further advancements in technology coupled with the automation of DNA sequencing (Smith, Sanders et al. 1986) lead to an exponential growth of output and decrease of cost of DNA sequencing, which could now sequence entire genomes. These sequencing technologies are now referred to next generation sequencing (NGS). The availability of high-throughput sequencing methodologies paved the way for a new era in Biology. By coupling NGS to other methodologies it was now possible to measure the presence of epigenetic marks at a genome wide levels or measure global levels of gene expression. In the following sections I will present how can NGS be used to study gene expression and DNA methylation.

2.1 Using NGS to study DNA methylation

While DNA methylation has been shown to be an important epigenetic mark involved in many diverse biological processes, it was not until the availability of high-throughput sequencing technologies, and in particular the development of bisulfite shotgun sequencing that it has been possible to study this mark at a genome-wide level. While there are many methods to measure DNA methylation, I will be describing three of the most used ones: restriction analysis, affinity enrichment and bisulfite sequencing. I will give a brief overview of the techniques, and will discuss as their advantages and limitations for genome-wide studies.

2.1.1 Restriction analysis of DNA methylation

One of the most widely used technologies to determine DNA methylation was restriction analysis of DNA methylation. This method uses pairs of restriction enzymes, called isoschizomers, that recognize the same DNA sequence, but one of them is unable to cleave the sequence when DNA methylation is present. An example of such set of enzymes is the HpaII-MspI pair. They both recognize a CCGG sequence, however HpaII is sensitive to cytosine methylation and won't cut the target sequence if it is methylated. By contrast MspI is insensitive to cytosine methylation and will perform its endonuclease activity at all sites. By comparing the restriction bands generated from the digestion of genomic DNA by the two enzymes, it is possible to infer the methylation status of the sites of interest. This method has some advantages: restrictions enzymes are relatively cheap, it has single base resolution and it does not require prior knowledge of the complete genome sequence of the organism of interest. While this method is still being used in *A. thaliana* (Cervera, Ruiz-Garcia et al. 2002), maize (Lu, Rong et al. 2008) and *O. sativa* (Wang, Pan et al. 2011*)*, there are some shortcomings to take into consideration: measurements are limited to cytosines contained in cleavage sites. This drastically limits the number of cytosines that can be interrogated across the genome. Most of the cytosines found in a genome are not part of such restriction sites. A second limitation is that the measurements are not quantitative; DNA methylation is measured as a binary feature with either presence or absence being the only possible outcomes.

2.1.2 Affinity enrichment

Methylated DNA immunoprecipitation (MeDIP) is based on the isolation of methylated DNA from fragmented genomic DNA by immunoprecipitation through antibodies that specifically bind to methylated DNA. Microarrays (MeDIP-chip) or next generation sequencing (MeDIP-seq) can be used to identify or sequence the immunoprecipitated fragments. While this method can be applied at a genome-wide level, it has the drawback that immunoprecipitation depends on the quality and specificity of the antibody and the distribution of methylated sites within a region. Furthermore, it lacks single base-resolution and instead reports methylation enrichment in the sequenced region. In plant genomes, this poses a problem due to the fact that most cytosines in the genome are not methylated. Furthermore there are lower levels of methylation at CHG and CHH sites than at CG sites, causing a reduction in this method's efficiency and generating biases towards CG-sites.

2.1.3 Whole-genome bisulfite sequencing

This technique consists of treating DNA with Bisulfite and the performing whole genome sequencing. It provides single base-pair resolution as well as quantitative measurements of DNA methylation, and is not affected by sequence context of the cytosines nor the abundance of methylation in a given region. Because bisulfite sequencing was the methodology used for the measurement of DNA methylation in this study, I will provide detailed overview:

In 1970, two groups described a chemical reaction in which sodium bisulfite ($NaHSO_3$) lead to the deamination of cytosines in DNA (Hayatsu, Wataya et al. 1970, Shapiro, Servis et al. 1970). This reaction occurs in three steps: the sulfite group binds the 6th carbon of the cytosine, followed by the loss of the amine group of the cytosine (cytosine converted to uracil sulfite), and finally the sulfite group is released, effectively converting a cytosines to uracils (Figure 7). This reaction acts 100 times faster on unmethylated cytosines compared to methylated cytosines.

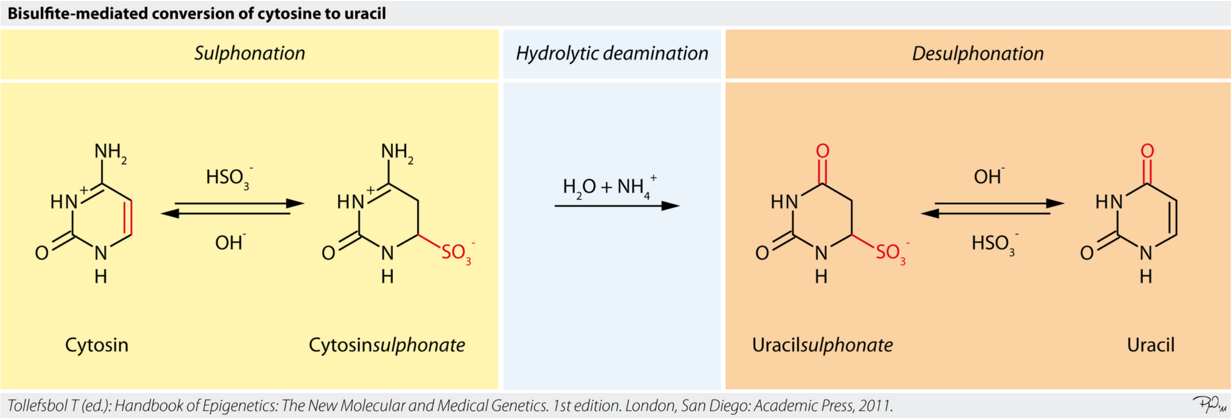

Bisulfite-mediated conversion of cytosine to uracil
Sulphonation
Hydrolytic deamination
Desulphonation
NH2
HN
O
N
H
HSO3-
OH-
SO3-
H2O + NH4+
O
HN
OH-
HSO3-
Cytosin
Cytosinsulphonate
Uracilsulphonate
Uracil
Tollefsbol T (ed.): Handbook of Epigenetics: The New Molecular and Medical Genetics. 1st edition. London, San Diego: Academic Press, 2011.

Figure 7 Conversion of cytosine to uracil through bisulfite treatment. Figure from (Tollefsbol 2011).

Bisulfite sequencing exploits this property of selective deamination of unmethylated cytosines (Frommer, McDonald et al. 1992), and is now one of the most widely used methods for (indirect) DNA methylation identification. This method consists of treating DNA with sodium bisulfite and then amplifying the DNA of interest with a PCR reaction. Unmethylated cytosines (from the untreated DNA) are converted to thymines, while 5-methyl-cytosines remain unaffected. To identify DNA methylation, one needs to distinguish which cytosines were converted to thymines. Initially, low throughput methods such as Sanger sequencing were used to determine the sequences of bisulfite-converted PCR products. However recent technologies in the field of DNA variant detection such as microarrays and next generation sequencing platforms have made possible the use of bisulfite sequencing for genome-wide DNA methylation analysis. Bisulfite sequencing can accurately survey most of the genome, it has a single-base resolution, gives strand specific information, and it is highly scalable. Furthermore, it provides a non-biased and highly sensitive quantitative measurement of DNA methylation, which allows cross experiment comparisons.

Technical confounders of bisulfite sequencing include 5-hydroxymethylcytosine, present at low levels in mammalian genomes, is also immune to bisulfite treatment, and therefore measured as 5-methylcytosine (Huang, Pastor et al. 2010). Also, PCR efficiency drops when amplifying DNA containing uracils; creating the need for more PCR cycles during sequencing library preparation, which can lead to PCR product biases and requiring higher amounts of starting DNA material. An additional complication is that bisulfite conversion is not 100% efficient. Some cytosines will not be converted, and are therefore incorrectly identified as methylated. This false positive methylation can be directly measured and used as a correction factor during downstream analysis. This is achieved by including in the sample to be sequenced DNA material that is known to lack methylation (such as phage lambda DNA or chloroplast DNA). Additionally, the analysis of bisulfite sequencing requires a reference genome. Methylation is estimated from the aligned reads to the reference genome; hence the quality of the reference genome will affect the correct mapping of sequenced reads, and therefore the methylation estimates. Furthermore areas where sequence alignment is either difficult or not possible (such as repetitive elements) cannot be interrogated. This can be problematic due to the fact that DNA methylation is a prominent mark of such regions. Finally, due to the reduction of sequence complexity of the bisulfite treated DNA, alignments to the reference sequence are less accurate, requiring more stringent quality standards or creating additional errors.

2.2 Generating transcription profiles

One of the central questions regarding DNA methylation is its relationship to gene expression. So far I have presented the tools needed to study DNA methylation at a genome wide level. In this section I will describe how NGS can be used to measure global gene expression in an organism.

RNA-seq is a method that involves the quantification of RNA transcripts in the cells through NGS. While the expression of a gene is hard to quantify, the quantity of messenger RNA for a given gene tends to correlate well with the amount of protein found in the cell, as such it has been used as a proxy for gene expression. To do this, the RNA of interest (such as mRNA or sRNA) is isolated from cells, and with the use of reverse transcriptase, complementary DNA (cDNA) fragments to the RNA molecules are generated. Such DNA fragments can then be sequenced, and from abundance of DNA molecules it is possible to infer the abundance of RNA transcripts. RNA-seq allows for the detection of exon/intron boundaries, as well as the identification of alternative transcripts. The currently most common platform for sequencing (Illumina HiSeq) generates short sequencing reads (<500 bases). These are easily processed when a reference transcriptome is available. While *de novo* assembly of transcripts is possible when such a reference is not available, it requires sophisticated software packages and the congruency of the results is low (Zhang, Jhaveri et al. 2014). Another challenge is that the abundance of different transcripts ranges across many orders of magnitude. To detect lowly abundant molecules, either enrichment or a very high sequencing depth is needed.

Despite these challenges, RNA-seq is the most common tool used in high throughput experiments. Advancements in technology and the development of better protocols, such as longer reads, enrichment techniques, and better kits, will help cope with the drawbacks mentioned above. Finally, more accurate and reliable pipelines for the downstream analysis are being developed, which will help on creating more robust analyses.

2.3 Conclusions

There are various ways of measuring DNA methylation, and new technologies are still being developed aiming at increased sensitivity and higher data output. These include methods for enrichment of desired DNA, thereby reducing the amount of required starting material, single molecule sequencing with long reads, e.g. using the Pacific Biosciences technology (Rhoads and Au 2015), and techniques that can measure DNA methylation directly without the need of nucleotide conversion are being developed, such as nanopore technology (Simpson, Workman et al. 2016). While there is great effort in developing these technologies they are still too expensive for large-scale experiments. Regardless of the technique, repetitive elements are difficult to analyze and to measure, this is particularly relevant in the field of epigenetics due to the fact that there are many epigenetic marks that are closely associated to these repetitive elements, including DNA methylation. Longer reads from the sequencing platforms and better bioinformatics tools will help to analyze these elements.

An inherent challenge that I will also address in the Results section is that measure average methylation (and also gene expression) across multiple cells (unless performing single-cell experiments). While the averages have been useful to describe the correlations between DNA methylation and gene expression, single cell sequencing has shown variability at both transcript and epigenetic level. Since my work also involved larger tissue samples, heterogeneity of cell types does limit the analyses described.

2.4 Analysis of bisulfite sequencing data

In section 2.1, I have described briefly the technologies available to sequence DNA, and how they can be applied to measure gene expression and to detect DNA methylation. In the following section, I will describe the computational strategies that have been used in previous studies to analyze bisulfite-sequencing data.

2.4.1 Processing of Illumina reads

The raw data from the Illumina instruments are images from the optical sensors of the machines. These images are then processed with “Base calling” software which generate a file with the DNA sequences of the reads in a fastq format (Cock, Fields et al. 2010). The length of the reads ranges from 100-300 bp depending on the instrument used. Each base of each read has “phred quality score” which is a measurement of the probability of an incorrect base call. Due to technology used for Illumina sequencing (sequencing by synthesis), the 3’ ends of the reads tend to have lower quality scores. Quality control is performed to filter out reads that could introduce errors in the downstream analysis, as well as to trim the ends of the reads, which have low quality scores. Some of the properties of Illumina reads that are used for filtering include number of bases with low quality, number of ambiguous bases and sequence complexity. Finally, the adapter sequences used for the indexing of sequencing libraries are removed from the reads. There are many different software packages designed to perform such tasks such as fastQC (Andrews) , STACKS (Catchen, Hohenlohe et al. 2013), GBSX (Herten, Hestand et al. 2015) and Shore (Schneeberger, Ossowski et al. 2009). In this study the SHORE toolkit was used for demultiplexing and quality control.

2.4.2 Alignment of Illumina reads

One of the biggest advantages of performing high throughput experiments in a model organism such as *A. thaliana* is the availability of a high quality reference genome. It allows for the identification of polymorphisms, such as the ones introduced during bisulfite conversion, through the comparison of the sequenced samples to the reference genome. To do this, it is necessary to identify the genomic regions from which the reads originated from which is equivalent to the problem of identifying sequences of high similarity. This problem has been widely studied in the field of bioinformatics and is known as sequence alignment problem, or in the case of matching individual reads to genomes, read mapping. Originally, the problem was limited to the alignment of one sequence to another and was solved by algorithms like Smith-Waterman (Smith and Waterman 1981) or Needleman-Wunsch (Needleman and Wunsch 1970). These algorithms were designed to find the optimal alignment between the two sequences. While these algorithms are still used for the alignment of sequences, they are computationally demanding and therefore not suitable for NGS data.

Current algorithms need to be able to handle millions of reads in a sensitive time while also producing accurate alignments. This type of software needs to be flexible enough to allow for mismatches, due to biological variability and sequencing errors found in the sequenced reads (Illumina machines have a 1% error rate), but sensitive enough to avoid reporting incorrect alignments. The repertoire of software implemented for the mapping of reads is vast, using many different algorithms and strategies. Some of these have focused on speed and a small memory footprint like Bowtie (Langmead and Salzberg 2012) and SOAP (Luo, Liu et al. 2012) and are suitable for desktop computers. Others have focused on the ability to use multiple reference sequences in order to create better alignments, such as Genomemapper (Schneeberger, Hagmann et al. 2009).

Read mapping of bisulfite sequencing data comes with an added layer of complexity. Due to the conversion of unmethylated cytosines to thymines after sequencing (caused by bisulfite treatment) the alignment of bisulfite reads to a reference genome would have an unusual number of mismatches, producing low score alignments (Bock 2012). There are two main strategies that have been used to address this problem. The first approach is to map to a modified reference genome, were all cytosines have been replaced with thymines (Krueger and Andrews 2011). This allows all converted and unconverted cytosines from the bisulfite sequencing to map correctly to the genome. A second widely sued approach is to modify the scoring matrix for the alignment to not penalize mismatches between cytosines and thymines (Xi and Li

2009). The problem with such strategies is that the lack of penalization of mismatches, or the reduction of sequence complexity lowers the accuracy of the alignments. Increasing the minimum quality scores of the alignments, as well as to remove any reads which map non-uniquely can help address this problem.

Once alignments are available, a second quality filter is performed in order to control for the PCR duplicates. During sequencing library preparation, there is a PCR step that might introduce a bias when not all fragments are amplified equally. This bias is stronger in bisulfite libraries due to the higher number of amplification cycles compared to normal genomic libraries (see section 2.1.3). To avoid redundant reads representing PCR duplicates, reads that map to the exact same starting position are removed. Because all reads are required to map at different locations, the maximum coverage at any given position is equal to the read length of the sequencing machine (assume a position *i* had a higher coverage than the read length *m,* that would mean there are at least *m+1* reads that map to different starts and cover *i*, but there are only *m* different positions at a distance <=m from *i*, by the pigeon hole principle, at least 2 reads would map to the exact position reaching a contraction).

Finally, a count table is generated with an entry for each cytosine in the genome containing the number of unconverted cytosines (methylated cytosines in the genomic DNA used to prepare sequencing libraries) and the total number of reads covering that site. This count table, which is referred to in subsequent sections as a genome matrix, is the only input needed for all subsequent analysis.

For each position, one can calculate the proportion of methylated reads from the total coverage. For the remaining of the text I will be referring to such quantity as the methylation rate. It is worth noting that even though DNA methylation at each position of a single DNA strand is a binary feature (it is either methylated or unmethylated), the methylation rate is a value that ranges from 0 to 1. This is because the material used for DNA library preparation usually comes from a large number of cells, and not all cells have the same DNA methylation profile (Smallwood, Lee et al. 2014). These differences can be due to stochastic differences or due to biological differences between cells (Feng, Jacobsen et al. 2010). Additionally, recently replicated DNA might still lack methylation on the newly synthesized strand, also contributing to these differences. This rate represents the proportion of DNA molecules that carry methylation, rather than the quantity of methylation within a cell.

2.5 Identifying methylated positions across the genome

In an ideal experiment, the methylation rate would be a faithful representation of the relative methylation found at each position. Unfortunately, as outlined in 2.1.3, errors from incomplete conversion during bisulfite treatment, sequencing errors, 'real' DNA mutations, and errors introduced by inaccurate alignments have to be taken into consideration. Some can be estimated easily: sequencing errors vary depending on the instrument used for sequencing and are normally very low. The Illumina HiSeq2000 instrument has an estimated sequencing error of ~1%. From these errors, only miss-calls of cytosine to thymines affect the methylation detection. By using only high quality reads one can further lower down this error rate. Due to these reasons, this error rate is normally ignored. Similar to sequencing errors, genetic variability in the analyzed DNA relative to the reference genomes can affect methylation estimates in the same way. While the genetic variability changes in a per study basis, its effect on DNA methylation is normally ignored.

Incomplete conversion during bisulfite treatment is another factor that can affect DNA methylation estimates. Incomplete conversion of unmethylated cytosines to thymines leads to an overestimation of DNA methylation; unmethylated reads will appear to be methylated. The magnitude of these errors can be many times higher than sequencing errors, greatly inflating the methylation rate. This incomplete conversion rate (referred to the false methylation rate) can vary between experiments, and even between replicates. It is possible to measure the false methylation rate (FMR) by treating control DNA composed entirely of unmethylated cytosines with bisulfite and measure directly the incomplete conversion rate. Some examples of commonly used control DNA are phage lambda DNA, chloroplast DNA in plants and, in the case of mammals, non-CG sites.

When testing a position for methylation, the Null hypothesis is that the position is not methylated and that the reads supporting methylation are a product of incomplete conversion from the bisulfite treatment. A binomial distribution is used to calculate the probability of seeing a number equal or greater of non- converted sites out of the total coverage, with the probability of a methylation call being equal to the false methylation rate. This procedure is performed for each cytosine position in each sample. Due to the great number of statistical inferences being made, multiple testing correction must be performed, which will be discussed in section 2.5.2.

2.5.1 Identifying differentially methylated positions

While identifying methylation across the region is an important task, sometimes it is more biologically relevant to identify positions where there are changes of methylation between samples. Early studies of DNA methylation have used relatively simple models to identify differentially methylated positions (DMPs). These include labeling positions as DMPs between samples if one sample is classified as methylated and another one is not (Schmitz, Schultz et al. 2011) or classifying a positions as DMPs if the difference in methylation rates between samples exceeds an arbitrary value (Laurent, Wong et al. 2010). While these methodologies are easy to implement, and they are not computationally demanding, they have important biases. The first approach fails to classify as DMPs positions with high differences in DNA methylation rates between two samples as long as both of them are methylated, while the second approach is biased to CG-sites (because the range of DNA methylation rates in CG-sites is higher) as well as sites with low coverage (standard deviation decreases proportionately with sample size). Finally, these approaches do not provide a confidence value for the classification of DMPs.

Instead, one can test directly if the two samples show differential methylation at a cytosine of interest by transforming the number of methylated reads/unmethylated reads of each sample into a contingency matrix and performing a two tail Fisher´s exact test. This allows for the identification of small but statistically significant differences between samples, making this method suitable to detect differences of DNA methylation in all cytosine contexts. A second advantage of this method is that is not affected by poor coverage; positions with low coverage but high methylation differences will have low confidence estimates. Finally, it provides a p-value, which serves as a measure of reliability of the statistical test. One of the drawbacks of this method though, is that small differences of DNA methylation between positions with high coverage are generally identified as differentially methylated. While these positions may show statistically significant differences in DNA methylation, these differences might not be biologically relevant. Furthermore, a Fisher's exact test must be performed for each position of interest for each pair of samples. If all pairwise comparisons are made, there is a combinatorial increase of the number of statistical tests needed, which can become a limiting factor in the analysis. Furthermore, due to the high number of tests performed, multiple testing correction is necessary, which lowers the power of detection.

2.5.2 Multiple testing correction

Prior to performing a statistical test, the experimenter must pick a significance value, which will serve as a cut off value to either reject or accept the Null hypothesis after performing the test. Once a significance value has been selected, it is compared to a p-value, which is defined as the probability of observing a case equal or more extreme given that the Null hypothesis is true. If the p-value is lower than the significance cutoff the test is called statistically significant. The problem arises when performing a large number of statistical tests, due to the fact that the probability of observing extreme or rare events increases proportionally to the number of tests being performed. In such cases, some instances of tests will report significant p-values, which will lead to the rejection of the Null hypothesis even though the Null hypothesis is true, and thus increasing the number of false positives (type I errors).

This problem is known as the multiple testing problem, and there have been many methods to control for this increase of type I errors. The simplest approach for multiple testing corrections is known as the Bonferroni correction (Dunn 1959). In this approach, one is interested in controlling the family-wise error rate (FWER), which is defined as the probability of making at least one or more type I errors. The Bonferroni procedure goes as follows, call p_i, the p-value of the H_i test. One will reject H_i if:

p_i< alpha/m

Where m is the total amount of statistical tests, and alpha is the desired FWER. It can easily be shown by Boole's inequality that this significance value guarantees an FWER equal or lower than alpha. The problem with this approach is that is conservative and significance values need to be extremely low when dealing with large samples in order to maintain a moderate FWER.

There are many other alternatives to calculate this correction value such as Westfall–Young permutation testing procedure, which calculates this correction factor empirically and is less conservative.

An alternative is to control for the proportion of type I errors instead known as the False Discovery Rate (FDR). Methods to control for the FDR have the advantage that are less stringent than methods controlling for the FWER, increasing the power of detection at the cost of a higher amount of type I errors. Some of the most used procedures to control for the FDR include Benjamini-Hochberg (Benjamini and Hochberg 1995) or Storey's method (Storey 2002).

Regardless of the method of choice, controlling the FWER or the FDR involves the establishing a new significance cutoff for the statistical tests that depends on the amount of tests being performed. As a consequence, this decreases with a larger sample size (and thus lower values are required to call a test significant).

2.6 Identifying methylated regions across the genome

Most studies where changes in DNA methylation are associated to differences in gene expression have been cases where methylation changes at a region level rather than at a single site level (Law and Jacobsen 2010, He, Chen et al. 2011). For this reason it has been of great interest to study DNA methylation at a regions level. The strategies used for the identification of methylated regions (MRs) can be roughly divided into two categories. The first strategy involves the testing of predetermined regions while the second involves the identification of methylated sites without any prior knowledge. I will introduce two approaches belonging to the first category, and summarize a more complex and novel method developed in the Weigel lab, which I have used for the analysis of my own samples.

2.6.1 Using individual sites to identify methylated regions

The most common approach for identifying methylated regions uses a sliding window. Methylated regions are defined as segments in the genome where there are multiple methylated positions within a specified genomic distance. This approach has been used in many studies due to its simplicity and ease of implementation (Lister, Pelizzola et al. 2009). It does however, not take into account the density of cytosines within a window. In areas of high GC-content, there is higher probability of finding methylated cytosines within the window size than in low GC-content areas. Furthermore, as described in 1.3.2 the levels of methylation vary depending on sequence context and in particular CG-sites are either completely methylated or completely unmethylated. The consequence is that the methylated regions identified with this approach are dominated by CG-methylation. Finally, the choice of window size is arbitrary, introducing further bias.

An alternative approach is to use predefined genomic coordinates to test for DNA methylation enrichment. It uses the number of methylated calls and the number of unmethylated reads in a given window, followed by a Pearson’s chi-squared test (Regulski, Lu et al. 2013) or a fisher exact tests (Stroud, Greenberg et al. 2013) to compare the numbers to the expected frequency assuming that the methylated reads are only a product of the false methylation rate . While this approach helps to classify known regions of interest as methylated and not methylated, it requires predefined regions for testing, which can introduce a bias. Furthermore, CHG, CHH and CG methylation marks have unique distribution patterns across the genome. Therefore, using a single distribution model for all methylated sites is far from ideal. Also, as discussed for individual cytosines, sites with high coverage are disproportionally identified as methylated, and thus contribute disproportionately to the classification of regions as methylated. Finally, this approach requires defining regions to be tested, which can introduce undesired biases as well.

2.6.2 Identifying methylated regions using beta-binomial models

As mentioned in the previous sections there are many factors that can affect the analysis of DNA methylation including sequencing coverage and sequence context. Furthermore the large amount of statistical inferences being made requires multiple testing correction lowering the power of detection. One promising alternative is the use of beta-binomial models (Molaro, Hodges et al. 2011). Such models use a beta-binomial distribution to model the distribution of DNA methylation counts from bisulfite sequencing data. These methods have the advantage that can easily incorporate the variation of methylation rates in the genome (which are sequence dependent), are not affected by sequencing depth and incorporate in their detection the variance found within and across replicate groups. These methods have been used in human studies (Molaro, Hodges et al. 2011), where methylation only occurs at CG sites, and therefore not suitable for plant studies. Recently, an implementation of such method has been developed for plant studies (Hagmann, Becker et al. 2015). Such method can model methylation for each sequence context, CG, CHG, CHHH, independently.

The implementation mentioned above (Hagmann, Becker et al. 2015) uses a Hidden Markov Model to classify genomic regions into methylated or non-methylated from single site information by fitting the data to a sequence-specific beta-binomial distribution. This strategy has the advantages that does not require prior knowledge of the regions and can model the unique distribution of each methylation context independently, allowing for a more accurate detection of methylated blocks. After MRs have being assigned, it is possible to tests these regions across samples for differential methylation using a log-likelihood test. This method calculates the log odds ratio between the likelihood with 2 different sets of beta-binomial distributions and the likelihood of the joint distribution.

3 Results

Many studies investigating DNA methylation have focused on two specific aspects: development-associated changes as well as heritability (and evolutionary conservation) of the methylation mark.

In this study, I assessed the variability of DNA methylation in a set of different *A. thaliana* organs

- To compare the inter-individual variability against intra-individual variability
- To address whether different developmental phases in the plant are associated with different DNA methylation patterns
- To study the relationship between organ-specific methylation pattern, organ identity, and gene expression profiles.

In the following sections, I present the obtained results and discuss the biological relevance of those results in combination with an outlook of this study.

3.1 A new pipeline for the analysis of DNA methylation

The initial version of the pipeline for detection of methylated positions (binomial tests) and differentially methylated positions (Fisher's exact tests) was first implemented in 2011 for the analysis of DNA methylation of the MA lines (Becker, Hagmann et al. 2011). This pipeline consisted of multiple python, Perl, R and Bash scripts that had to be run in a particular order, generating a great number of temporary files, each having many overlapping parameters.

For scientists without much experience using a terminal environment, running multiple programs in succession as well as keeping track of temporary files can prove to be difficult, furthermore having to take care of dependencies of each language being used increases the complexity of setting up the resources needed for the pipeline to work. It also makes comparing individual runs from different users more complicated due to the fact that each temporary file will be named differently, and parameters can vary. Due to these reasons, the first objective of this work was the implementation of self-contained and user friendly pipeline for the analysis of DNA methylation.

3.1.1 Outline

The only input needed to run the implementation of a pipeline is a methylation count file in a “genome matrix” format descried in (Becker, Hagmann et al. 2011). This count data file can be easily generated from an alignment file of bisulfite-sequencing reads to a reference genome using the Shore module “methyl”.

The pipeline is divided in five steps:

1) False methylation rate estimation
2) Binomial testing for differential methylation
3) Merging of the binomial test results
4) Differential Methylation Calculation
5) Clean up of intermediary files

The first step of the pipeline involves the estimation of the conversion rate of the bisulfite treatment (see section 2.5) or the FMR. These rates will be used to identify statistically methylated positions in step 2. This estimation uses the reads of DNA that is known to be completely unmethylated such as phage lambda DNA or plastid DNA such as mitochondria or chloroplast. The estimation of the FMR is done in a per sample basis. During this step the program also performs a round of filtering where sites that not meet a minimum coverage (default 5) and a minimum quality (default 16) are filtered out.

Most of the chloroplast reads are highly covered, and with very low false methylation rates (10^{-6} -10^{-8}), as such if one uses a single value for the false methylation rate (FMR) it will be dominated by highly covered single sites. For this reason I cluster sites by coverage bins (default value of bins of size 5, e.g. first bin consists of sites with 0-5 coverage, second bin 6-10, etc.) and calculate a false methylation for each bin. I use a conservative approach where bins with lower coverage must have a higher FMR.

To do this, if a low coverage bin has a lower FMR, then the FMR for that bin is set to the highest FMR among all the other bins with higher coverage (for that sample).

The second part of the pipeline is the calculation of p-values using a binomial distribution assuming that the presence of methylated reads is only due to incomplete conversion as described in section 2.5.Once all the p-values are calculated multiple testing correction is performed using Storey's method (Storey 2002). If one is only interested in identifying methylated positions one can stop here.

The next steps are for the identification of differentially methylated positions between samples. In the third step, the results from step 2 are summarized into a single file with a flag determining which positions passed the quality filters and the results of the binomial test. The default behavior of the pipeline is to test all samples against each other. In order to reduce the number of tests being performed, I only compare positions where at least one position was identified as methylated among all samples.

Finally for each pairwise comparison, all cytosines are tested for differential methylation using a Fisher's exact test using the number of methylated reads and unmethylated reads of each pair of samples. After all p-values are calculated, multiple testing correction is performed using Storey's method (Storey 2002).

One of the first improvements made was the consolidation of all individual scripts into a two R scripts, a C++ library and a C++ wrapper, which is the only platform in which the user interacts. This change improved the sharing of the pipeline as well as the simplicity of running it; also reducing the number of dependencies needed to run the pipeline. An additional advantage of this scheme is that all the intermediate files are named and handled automatically by the main function, ensuring a consistent experience across users. Each of these steps is implemented into a function in a library file with the Binomial testing being performed by C++ and the Differential testing being performed in an R helper file. In the simplest case, the user will have a genome matrix (refer to section 2.4.2) and will want to retrieve a list of methylated positions in all samples and a list of DMPs between all pairwise comparisons. In this particular case, the user only needs to provide a writable directory and the path of the genome matrix to get such output, illustrating how easy a new user could run this pipeline with default parameters. All necessary parameters have a default value, which are the ones used in previously published material, and these can be easily changed.

3.1.2 Handling of replicates

A second modification regarded the handling of replicates among samples. The original version of the pipeline could be provided with a list of replicates and for each group of replicates, only those positions that had concordant methylation status (methylated or unmethylated) across all replicates were tested. In my improved implementation, replicates can be handled in different ways. The first one is identical to the initial version, where only sites were all replicates were classified as either methylated or unmethylated are tested. The second one is testing directly for differential methylation between replicates using the same procedure used for testing differential methylation across samples and only using positions that were not identified as differentially methylated between replicates, regardless of classification as methylated or unmethylated. For example, a site can be classified as methylated in all replicates, but there can in addition be statistically significant differences in methylation level between the replicates; such a site would not be considered

In the original version, the reads from replicates were merged into a single sample by adding their respective methylated/unmethylated reads. This increase in coverage can cause problems due to the previously discussed coverage effect when applying a Fisher’s exact test. In my pipeline the user has two additional possibilities for handling replicate read counts. The first one is just using the sample with the highest coverage. The second method calculates the weighted arithmetic mean (weighted to the coverage) of the methylation rates of the replicates, and assigns methylated/unmethylated counts to that positions such that the coverage of that position is equal to the highest coverage across all replicates, and the methylation rate is as close as possible to the weighted average mean.

The implementation of this pipeline in C++ had also the expected advantage that numerical calculations where much faster compared to the counterpart in Perl. On a benchmark consisting of 500 sets with one millions simulated positions, the Perl implementation had an average runtime of 89 seconds (s.d.=1.4) compared to an average runtime of 9 seconds on C++ (s.d.=.2).

Furthermore, in contrast to the original pipeline where most of the operation had been designed to work within a single processor, my pipeline was implemented in order to be parallelizable and thus exploit the multicore capabilities that are present in most modern computers. Additionally my pipeline was designed to be able to run on stand-alone desktop or laptop computers by avoiding operations using large amounts of RAM. An alternative version of my pipeline was implemented such that it could exploit the resources of our computing cluster (through Sun Grid Engine scheduling system).

3.1.3 Using the minimum attainable p-value in multiple testing correction

Finally, I implemented a new method to reduce the loss of statistical power that is caused by multiple testing correction. As mentioned in 2.5.2, when performing multiple testing correction, one determines a new significance cutoff value that depends on the number of statistical inferences. This can become problematic when using a discrete test statistic such as the binomial test to detect MPs, or the Fisher’s exact test to detect DMPs. This is due to the discrete nature of the test: there is a finite number of possible significance values a test can output, and therefore there is a minimum attainable significance value for each test (known as the minimum attainable p-value). If the minimum attainable p-value of a test is greater than the significance cutoff set by the multiple testing correction procedure, the result of that test could never be considered significant. Such tests are not only non-informative, but are taken into consideration when calculating the new significance cutoff; as such it would be ideal to remove them completely from the analysis.

Because the minimum attainable value can be calculated without doing any statistical inferences, one can take advantage of this property to remove cases where the minimum attainable p-value is higher than the significance cutoff. This method has the advantage that less or an equal amount of statistical tests will be performed, while guaranteeing to be at least as powerful as just performing a single round of multiple testing correction.

The method works as follows:

1. Calculate the significance cutoff value *k*
2. Calculate the minimum attainable p-values for all tests
3. Sort them from lowest to highest

```
for $i$=1 to $m$ do
    If $P_{min(i)} > k$ then
      remove $P_{min(i)}$ from the set
      update $k$
    else
      return($k$)
```

This method guarantees that that all the remaining tests that are performed could be considered significant even after multiple testing correction. One drawback from this procedure is that one needs to calculate the minimum attainable p-values for each test, which can be computationally intensive. When testing for differential methylation it is possible to mitigate this problem by pre-calculating all the minimum attainable values for all possible cases. This is possible due to the way the methylation count table is generated where the maximum coverage of a single position is 200 (see section 2.4.2), consequently there are only 200x200 different possible Fisher's exact tests that can be performed, allowing for a fast implementation for such algorithm.

3.1.4 Conclusion

While previous pipelines were available for analysis of bisulfite data, my new implementation in C++ provides a more efficient and simpler to use alternative. It provides a tool that, can be easily used by non-expert users, while also giving more options for handling the analysis to advanced users. It also provides a novel method to preprocess data, reducing the number of statistical tests performed and thus reducing the impact of multiple testing correction as well as a moderate speedup. Finally, the pipeline was designed with flexibility and scalability in mind, with a version suited for stand-alone computers and another for a high-performance computing environment.

3.2 Methylation variation in leaf tissues

Many previous studies of genome-wide DNA methylation have used leaves as their study object as they are easily collected (Cokus, Feng et al. 2008, Lister, O'Malley et al. 2008). Usually, several leaves from a single individual or even leaves from multiple individuals are pooled, which helps to average out stochastic and inter-individual variation. Furthermore this allows for an easier detection of conserved methylation patterns. While this strategy is useful to investigate some aspects of DNA methylation, both the variation between individuals as well as between organs is lost in the pooling, which might contain relevant biological information.

Previous studies have shown that different epigenetic marks in plants can change with plant age and between different developmental stages (Feng, Jacobsen et al. 2010, Cantone and Fisher 2013), further pointing to a potentially biologically relevant role of such variation.

To study the variation of DNA methylation in leaf tissue, I produced single-base-resolution methylation maps using bisulfite-sequencing (Cokus, Feng et al. 2008, Lister, O'Malley et al. 2008) from 18 individual leaves from a single 5-week-old individual with Col-0 background, where the order in which each leaf arose in the plant was recorded. This allowed for studying the progressive effects of leaf age, or time point of leaf formation, on DNA methylation. In order to assess how pooling compares to individual leaf sequencing, I additionally sequenced bisulfite treated DNA of pools of six leaves from fourteen 5-week-old siblings originating from a single founder with a Col-0 background.

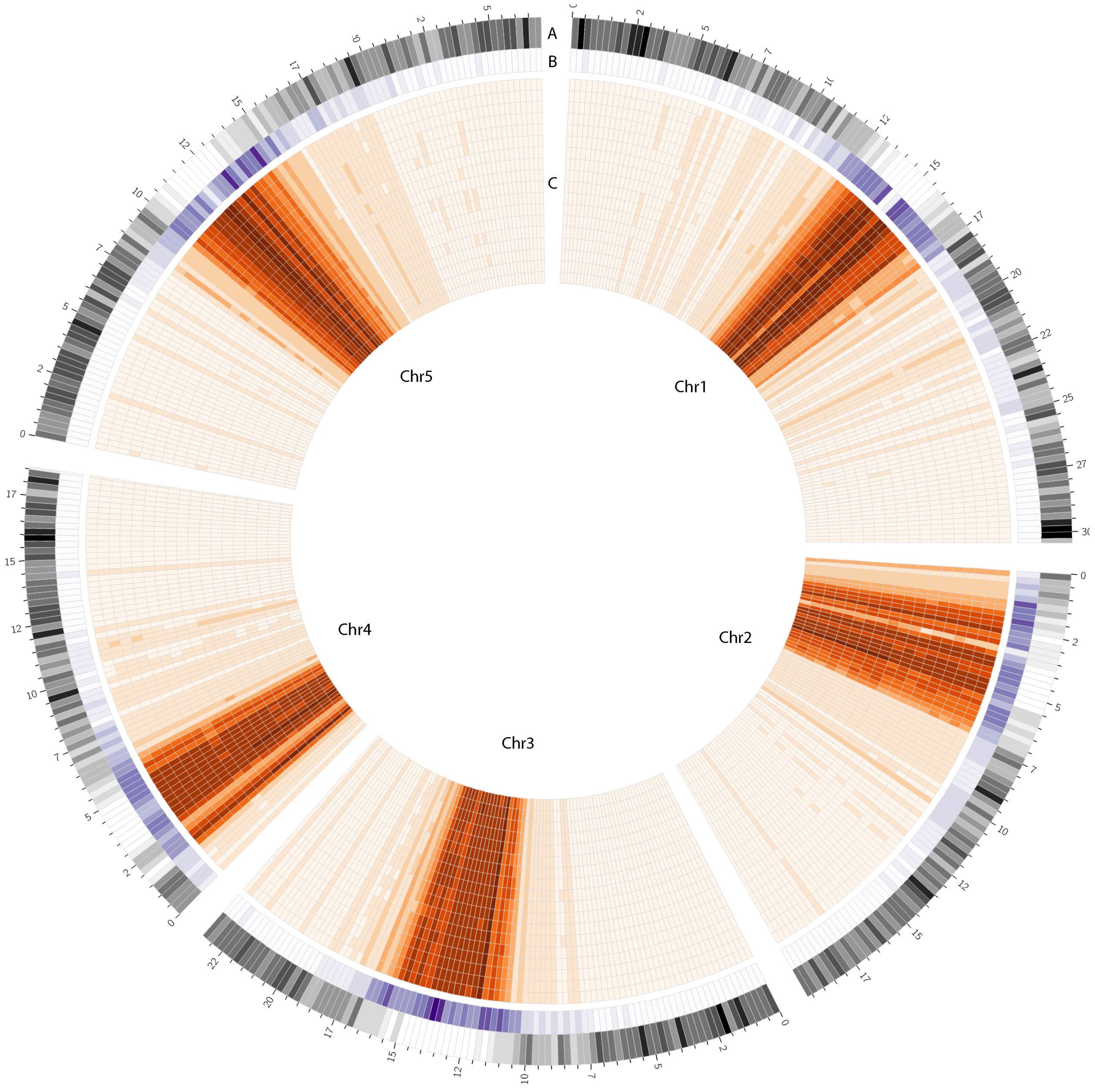


Figure 8 Distribution of DNA methylation of 18 individual leaves across the genome. Circos plot showing the normalized densities of a) genes density b) transposable element density c) DNA methylation (leaves are arranged according to their emergence time). Data was plotted using 250kb blocks and each track was normalized to the highest number in any block.

My samples were sequenced to an average genome coverage of 10x. I required each position analyzed to have a minimum coverage of 5x. Out of 43 million cytosines in the *A. thaliana* genome, an average of 27 million cytosines (s.d. 3 million) per sample passed the quality and coverage filters. I identified both MPs and DMPs using the pipeline described in section 3.1. Six million individual cytosines across the genomes were methylated in at least one sample, with an average of 4 million cytosines being methylated in each sample. Similar to previous studies (Cokus, Feng et al. 2008, Lister, O'Malley et al. 2008), most of the methylated positions were found near the centromeres, and there were fewer on the chromosome arms (Figure 8).

To reduce the impact of multiple testing correction when calling DMPs, I only performed biologically relevant comparisons, by separately assessing inter-individual variation with pairwise comparisons between the 14 sibling leaf samples, and intra individual variation with pairwise comparisons between the 18 leaves from the same individual.

3.2.1 Inter-individual variation and intra-individual variation

To assess the degree of conservation of methylation between individual leaves I calculated the frequency spectrum, which represents the proportion of sites that that are methylated by a given number of samples. My results showed that methylation in leaves was highly conserved, where most of the sites are either methylated or unmethylated in all leaf samples (Figure 9A). In order to compare the degree of conservation between individual leaves and siblings I calculated the overlap of methylated sites between them, in this specific context methylated referring to sites where at least 50% of the samples were methylated at that site. The majority of cytosines that were methylated in individual leaves were also methylated in the pools of leaves originating from 14 siblings (Figure 9B), suggesting that methylation in leaves is largely conserved.

I wanted to assess how does the variation between leaves compares to the variation between closely related individuals (siblings). I performed principal component analysis (PCA) using the methylation rates of all individual cytosines called as methylated in at least one sample (Figure 9C). To avoid imputing or handling missing information, only sites with complete information were taken into consideration (18.2 million sites). The PCA showed higher variation between individual leaves originating from a single individual compared to the variation found between pools of leaves from different individuals.

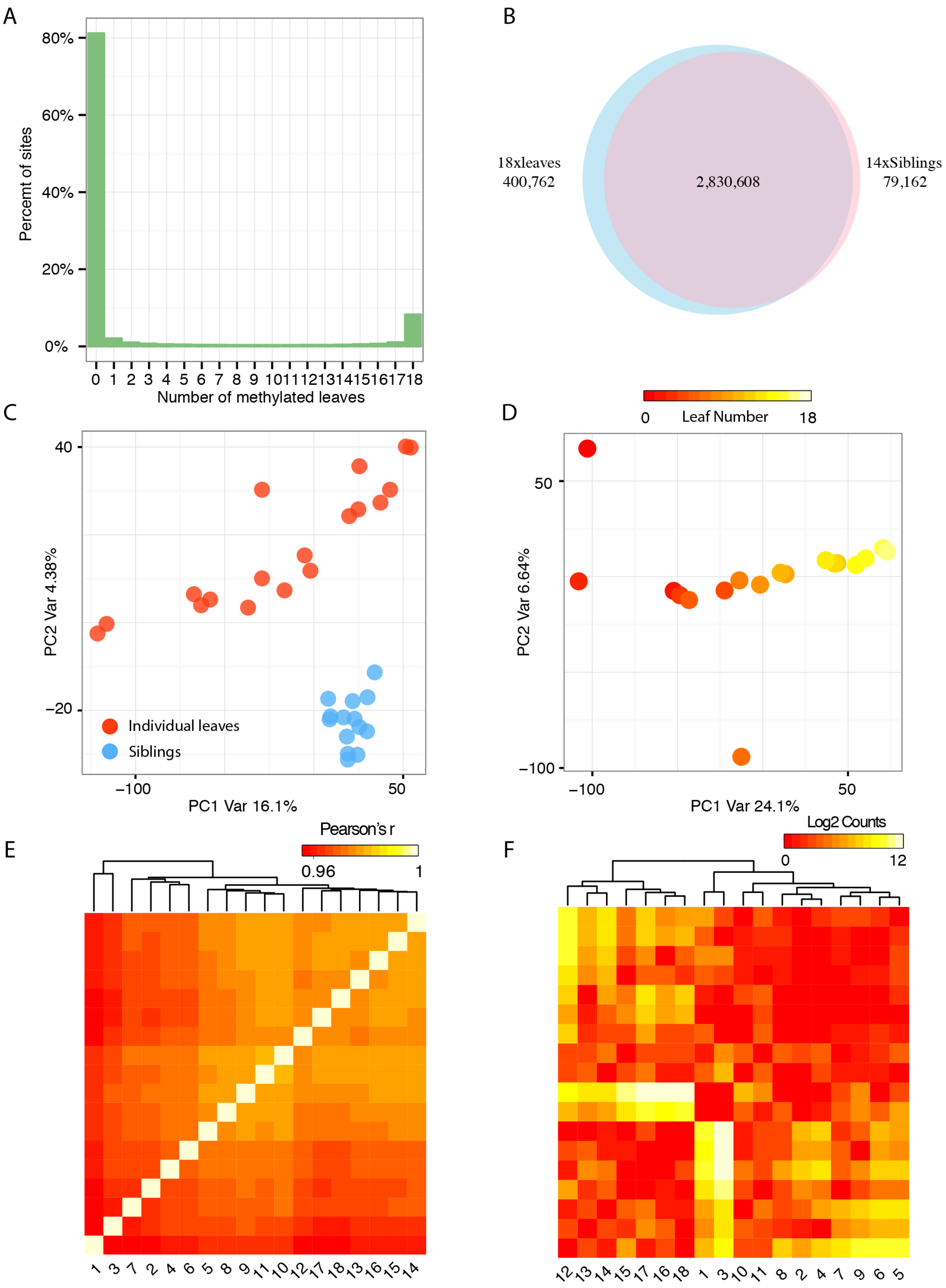

A
Percemt of sites
80%
60%
40%
20%
0%
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18
Number of methylated leaves
B
18xleaves
400,762
2,830,608
14xSiblings
79,162
C
40
−20
PC2 Var 4.38%
−100
50
PC1 Var 16.1%
Individual leaves
Siblings
D
0
Leaf Number
18
50
−100
PC2 Var 6.64%
−100
50
PC1 Var 24.1%
E
Pearson's r
0.96
1
1 3 7 2 4 6 5 8 9 11 10 12 17 18 13 16 15 14
F
Log2 Counts
0
12
12 13 14 15 17 16 18 1 3 10 11 8 2 4 7 9 6 5

Figure 9 Variation of DNA methylation in leaf tissue. A)Frequency spectrum of the number of methylated samples in the 18-leaf data set, "0" represents sites in which methylation was not detected in any leaf, "18" represents sites in which methylation was detected in all 18 leaves B) Overlap of methylated positions between the 14xSiblings set and the 18-leaf set. A position was considered methylated if at least half of the samples where methylated at that position C)Principal component analysis of methylation rates of methylated positions between siblings (blue) and individual leaves (red). D) Principal component analysis only using the individual leaf data set, the leaf number is represented by color where red is the oldest and yellow is the youngest. E) Hierarchical Clustering of individual leaves using Pearson's correlation coefficients as distance matrix. F) Hierarchical clustering of leaves using number of DMPs as distance matrix, colors represent the log2 of the counts between each sample.

3.2.2 Age-dependent changes of DNA methylation in leaves

I wanted to ask if age or time point of formation of the individual leaves was correlated to the observed variability in DNA methylation. In this case, and for the remaining of the text I define age as the order in which leaves emerged from the shoot apical meristem. I repeated the PCA analysis described in using only the data from individual leaves (Figure 9D). Leaf samples were numbered according to the order they emerged, where leaf 18 was the last one to emerge, and therefore the youngest. PC1 separated individual samples by order on the plant, and explained an increased amount of variance compared to the previous PCA (24% vs 21%, Figure 9), which included 14 individual siblings. To identify how related where the leaves between each other, I calculated the correlation coefficients between all leaf samples and used them as a distance matrix to perform hierarchical clustering (Figure 9E). Samples clustered in two groups where the 'oldest' (early emerging) and 'youngest' (late emerging) leaf samples grouped together. Clustering analysis using the number of DMPs between samples as a distance matrix (Figure 9F) showed a similar separation between 'young' and 'old' leaves.

To investigate whether a specific set of cytosines would be responsible for the difference between 'young' and 'old' leaves, I calculated the average methylation rate across all chromosomes in 250kb windows (Figure 10A). My previous results hinted at an age dependent effect of DNA methylation, therefore I normalized these methylation rates by the methylation rates found in the 'youngest' leaf sample (leaf #18). The normalization revealed a genome-wide correlation of DNA methylation with age. This correlation was more pronounced near the centromeres (Figure 10B). The oldest samples showed a generalized decrease of methylation, which was more pronounced near the centromere. Leaf development is a complex process and is composed of different stages with different transcriptional profiles. A possible explanation for the observed differences in DNA methylation could be due to changes in the expression of methylation machinery.

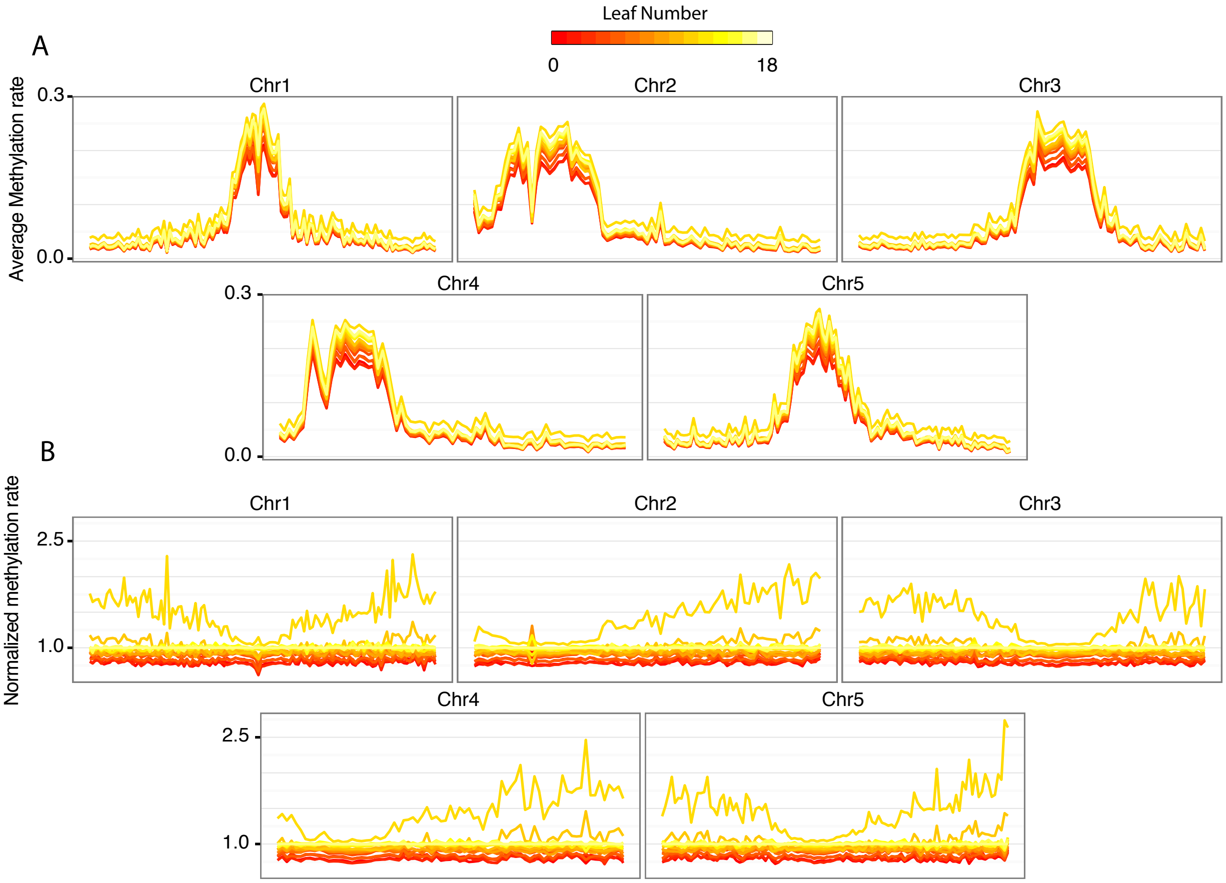


Figure 10 Genome wide age associated changes of DNA methylation patterns. A) Distribution of DNA methylation rates across the genome, each point is the average methylation rate of a 250kb block. B) Each block was divided by the average methylation rate found in the youngest leaf (Leaf #18).

To test this hypothesis, I used published transcriptome data (Schmid, Davison et al. 2005) to explore whether age-dependent changes in expression of methyltransferases genes might explain the relative loss of methylation in older leaves. Expression of both *METI* and *CMT3* (Figure 11A) (encoding the main methyltransferases responsible for CG and CHG) declined in older leaves and senescent leaves compared to young leaves. Reduced activity of DNA methyltransferases could thus be responsible for failure to faithfully maintain DNA methylation marks as leaves age.

I also wanted to whether variation of DNA methylation between leaves was similarly distributed across along the chromosomes as the variation found between individuals. In therefore compared the distribution of DMPs between individual leaves against the distribution of DMPs between siblings (Figure 11B). My data showed that DMPs between siblings were located mainly in chromosome arms and were rare in the centromere. Such distribution of DMPs has been previously reported in transgenerational studies (Becker, Hagmann et al. 2011). By contrast, DMPs between individual leaves were enriched in centromeres (Figure 11B) and were rare in chromosome arms. This suggested that the variation of DNA methylation individual leaves is different from the variation of DNA methylation between individuals.

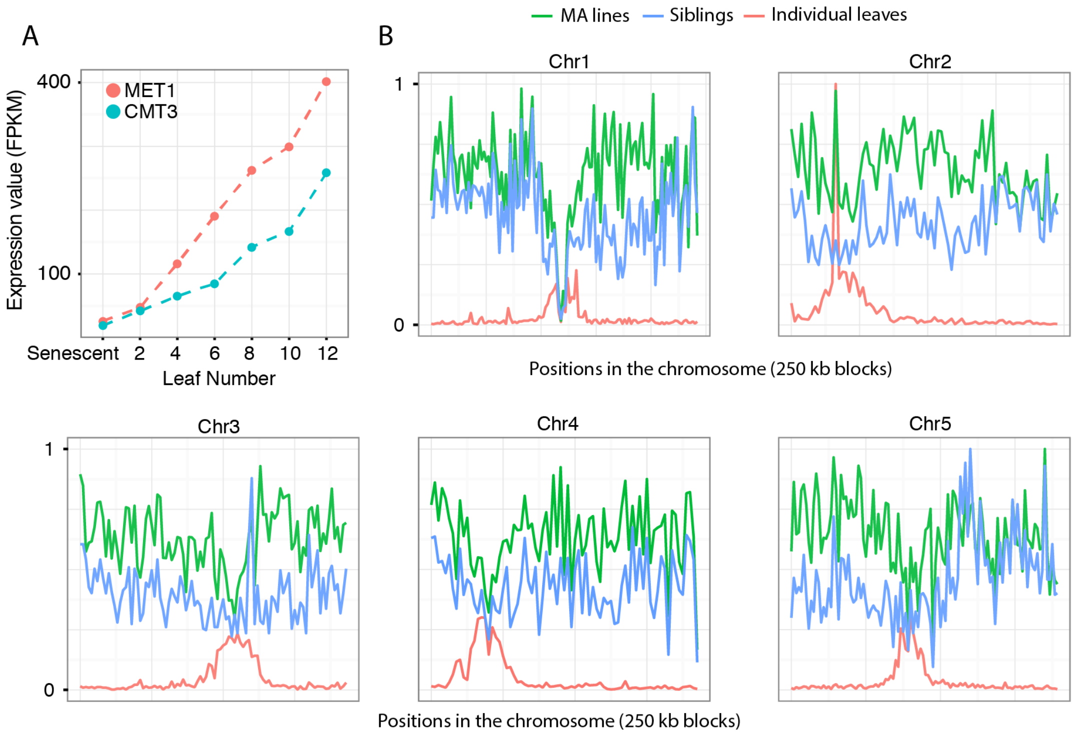


Figure 11 Changes in expression and distribution of DMPs between individual leaves. A) MET1 and CMT3 expression levels of different leaves at different ages. B) Relative density distribution of DMPs across the genome. Each block (size=250kb) was normalized by the highest DMP count for a block from their respective data set.

Next, I investigated whether the age-associated changes in DNA methylation were overrepresented in any specific genetic feature. For this purpose, I calculated the average methylation rate for pseudogenes, genes, 5'-UTRs and 3'-UTRs UTRs transposable elements (TEs) in each sample and compared them across samples Figure 12. Similar to previous results (Zilberman, Gehring et al. 2007), transposable elements had the highest average methylation rate among all the genetic features analyzed, followed by gene bodies and UTRs. All genomic features showed a similar relative loss of DNA methylation in older samples compared to the younger ones except for leaf #13. Leaf #13 had the highest false methylation rates among all samples, suggesting that the increase of DNA methylation is likely to be an artifact due to incomplete conversion after bisulfite treatment. This analysis suggests that the observed loss of DNA methylation was not specific to any particular genetic feature. This was surprising given that the mechanisms in which DNA methylation is established and maintained are different for each genetic feature.

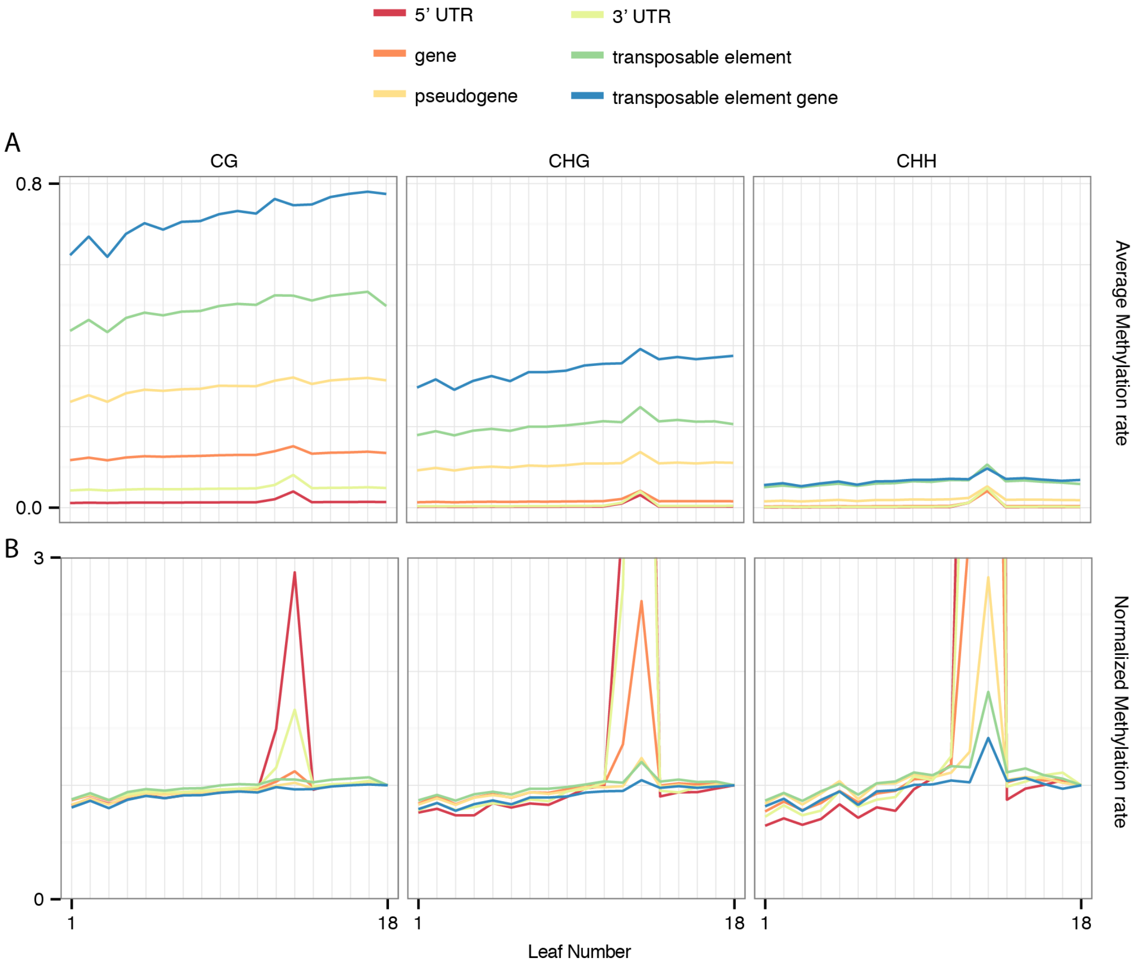


Figure 12 Distribution of DNA methylation across genetic features. A) Average DNA methylation of TEs, genes, 3' and 5' UTRs and pseudogenes B) Normalized average methylation of TEs, genes and 3' and 5' UTRs; samples were normalized to the average methylation levels of the youngest leaf (leaf #18).

Leaves undergo a series of developmental transitions during their life cycle. Even though the leaves I used for methylation profiling (leaves 1-18) were all harvested at the same time, they all emerged at different times and therefore some were in different developmental phases. The ‘oldest’ leaves were initiated during the juvenile rosette stage, while later arising leaves were initiated during the adult rosette stage that precedes the transition to flowering.

I wanted to determine whether the observed changes in DNA methylation between leaves of the same individual were dependent only on their age, i.e., the time that had passed since their initiation, or could be explained by their initiation at distinct developmental stages of the whole plant. In order to answer this question, I performed bisulfite sequencing of four individual leaves (leaves 1,3,5 and 7) of a 3-week-old plant and a 5-week-old plant. Methylation rates in each sample were calculated in 250kb blocks along the genome instead of at individual sites as the large sample number would have forced me to filter out a high proportion of sites when requiring complete information.

PCA separated samples by time of collection, but no age effects between individual leaves obtained from same individual could be observed (Figure 13A). Since these results were different from what I had observed in the 18-leaves-series, I wanted to exclude a technical error due to analyzing average methylation rate across blocks rather than methylation rates of individual positions. Therefore, I performed the same block analysis on the original 18 leaf samples (Figure 13B). This showed that the separation of samples by leaf age is still observable when using average methylation in blocks. The use of a different analyses pipeline can thus not explain the discrepancy between my results.

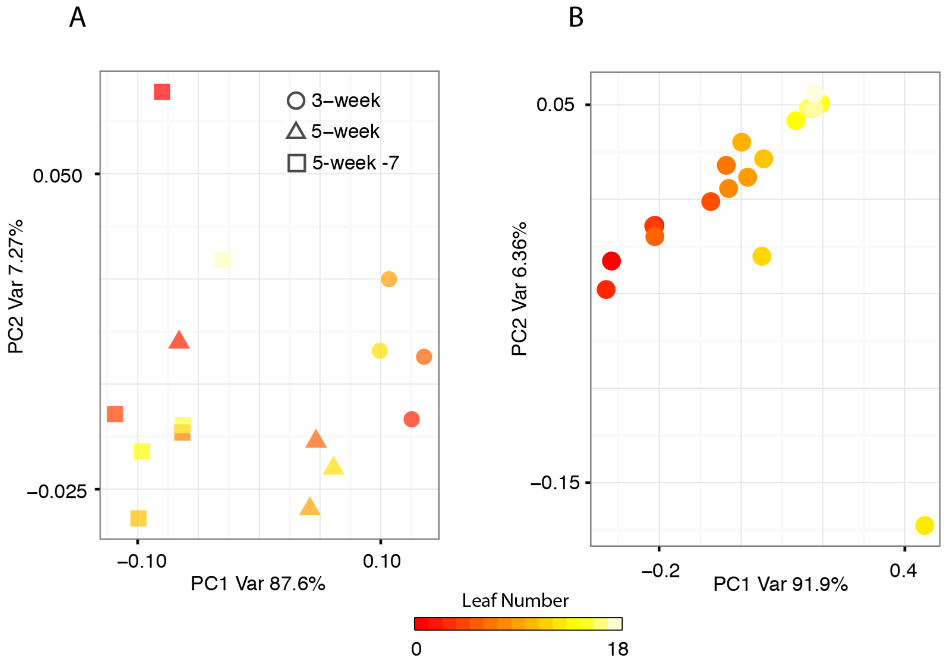


Figure 13 Effect of plant age on DNA methylation patterns A) PCA using average methylation from leaves of a 3 week-old and a 5-week old plant. Average methylation of 250kb blocks along the genome was used. B) PCA of the 18-leaves series using the methylation rate of 250kb blocks.

One major difference between the two datasets, the 18-leaves-series and the 3- and 5-week-old plant series, was the origin of the biological material. While all plants had a Col-0 reference background, the plant used for the 18-leaves-series was derived from an individual from the 30th generation of the MA collection line (Shaw, Byers et al. 2000), while the other plants were derived from an independent (not traceable) batch of seeds. To determine whether the discrepancy between experiments could be explained by seed origin, I performed another experiment using seven leaves (leaf number 1,3,5,7,9,11 and 13) of an independent 5-week old Col-0 individual from a third independent batch of seeds. Additionally I included in the analysis the methylation data from the MA lines, which are derived from whole rosettes of individual plants (Becker, Hagmann et al. 2011). The complete MA line data set compromises 20 individuals split by 30 generations from a common founder and two pairs of siblings that originated from the same founder plant but only 3 generations ago. This set includes the parental line from which the 18-leaves-series as well as the 14 siblings were derived from (one individual from generation 3). Furthermore, I also included in my analysis data from plants grown by a collaborating group, which sequenced 18 pools of the complete rosettes of 10x individuals each (Col-0 background) from 3 generations (Wibowo, Becker et al. 2016)

.

A PCA of the average methylation rate in 250kb blocks for all 89 samples showed that samples from the MA lines and samples generated from independent sets of seeds cluster separately (Figure 14). Figure 14 Differences in DNA methylation between the MA lines and other Col-0 accessions. PCA using methylation rates in 250 kb blocks of multiple whole genome bisulfite sequencing studies.

It remains an open question why the Col-0 derived MA lines and other Col-0 descendants have different methylation patterns. Only 15 mutations are shared across all MA individuals compared to the reference genome (13 substitution and 2 deletions) out of which none are in coding regions (Ossowski, Schneeberger et al. 2010), suggesting that genetic differences are unlikely to be the underlying cause. A second possibility could be that the MA lines and their relatives have an epigenetic memory different from individuals from independent lines. Further replication of the experiment using individuals derived from the MA lines is necessary to bring this study to a conclusion.

The aim of this study was to quantify the variation of DNA methylation within organs of a single individual. My results showed that there is variability present in the DNA methylation profiles between individual leaves originating from a single individual. Such variability was higher than the one found between pools of individual leaves from different individuals. Furthermore, in all cases individual leaves grouped together by their time of collection (3-week stage and 5-week stage) (Figure 13). Leaf development is a complex process composed of multiple developmental transitions such as leaf initiation and leaf elongation. Each of these developmental phases is accompanied by transcriptional changes. From tilling arrays I observed a correlation between the expression of different methyltransferases and leaf ageFigure 11). While it is tempting to speculate that DNA methylation might play a functional role during leaf development, My study showed that the changes of DNA methylation were not associated to any specific sequence context, nor enriched in particular genetic features such as transposable elements or gene bodies. Instead I hypothesize that these age-associated changes of DNA methylation are a product of improper maintenance of DNA methylation. This hypothesis is supported by the age–associated changes of expression of methyltransferases in leaves, arguing against the idea of leaf development being regulated by DNA methylation.

While the presented experiments revealed differences in DNA methylation between organs from the same individual, and an association of methylation with developmental stage, rosette leaves represent only a single organ type. Furthermore, different organ types can be both phenotypically and transcriptionally very different from each other (Schmid, Davison et al. 2005). Therefore I proceeded to focus on DNA methylation of different organs, its association to organ identity and gene expression.

3.3 Methylation variation across organs

Many studies in mammals and plants have shown that changes in DNA methylation are important for processes such as tissue differentiation and embryogenesis (Chan, Henderson et al. 2005, Smith and Meissner 2013). Developmental programs are tightly regulated and are accompanied by changes in transcription profiles as well as epigenetic changes, which are needed to accommodate to a new cellular fate. DNA methylation has been shown to be able to influence gene expression and to change through development (Law and Jacobsen 2010, He, Chen et al. 2011). In plants, most studies addressing methylation in a developmental context have focused on early stages, but the effects of DNA methylation on organ identity or its importance regarding dynamic gene regulation are still unclear.

To disentangle the contribution of DNA methylation to organ identity and gene expression, I performed bisulfite sequencing and RNA-seq of six different types of aerial organs of three *Arabidopsis thaliana* individuals. With this dataset I investigated tissue specific differences in methylation profiles, as well as their relationship to gene expression.

In this section, I will focus on two main questions:

- How do genome-wide methylation profiles differ between distinct plant organs?
- How does the intra-individual variation compare to inter-individual variation?
- Does methylation at different genetic elements, such as transposable elements and gene bodies, correlated with tissue identity and/or gene expression?

3.3.1 Organ-specific methylation profiles

To study differences of DNA methylation between organs, I generated single-base-resolution methylation maps from six types of aerial organs in three biological replicates: rosette leaves, cauline leaves, stems, siliques, open flowers, and closed flowers. Each plant was derived from a single individual of the third generation of the MA collection line (Col-0 background, refer to section 3.2 or (Becker, Hagmann et al. 2011)). These plants were grown in growth room conditions and have nearly identical genome sequences. Therefore differences in DNA methylation between identical organs from different siblings could be due to natural variation, while DNA methylation differences across organs could reflect changes associated to organ identity.

Samples were sequenced to an average coverage depth of 10x. I required all positions analyzed to have coverage equal or higher than 5. After filtering, 27.7 million cytosines remained out of the 43.1 million found in the repeat-masked genome. After using the pipeline described in section 3.1, I identified 6.5 million cytosines that were methylated in at least one sample with an average of 3.1 million methylated sites per sample (s.d. 0.5) and a range of 2.4 and 4.1 million samples depending on the organ type. Similar to previous studies, all organ types showed low levels of DNA methylation on chromosome arms and increased in centromeric regions (Figure 15).

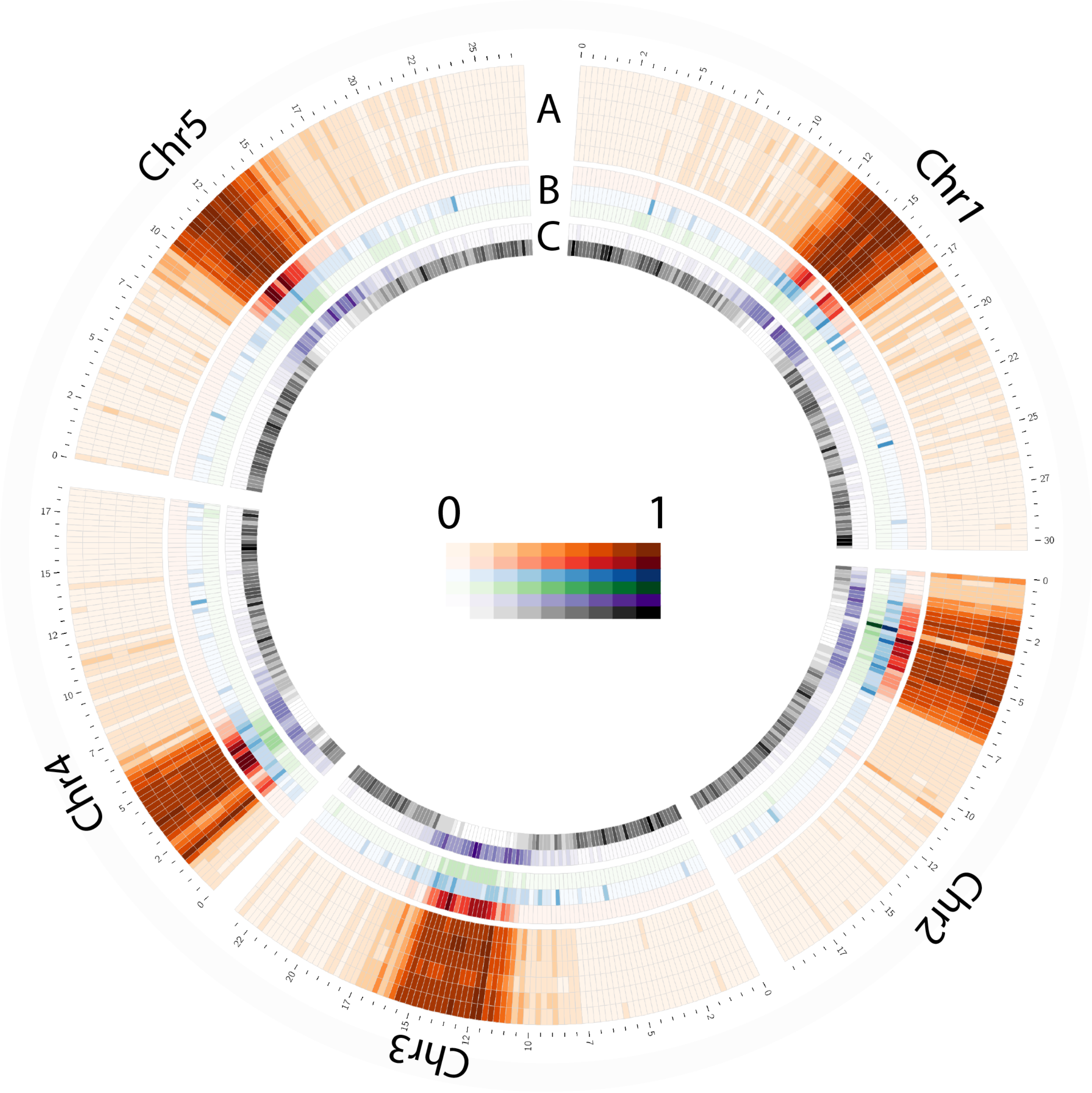


Figure 15 Distribution of DNA methylation across the genome. Circos plot showing the distribution of DNA methylation by A) organs (siliques, shoot, rosette leaves, open flowers, closed flowers, cauline leaves) B) Methylation density by sequence context CG (red), CHG (green) and CHH (blue) C) Density of TEs (gray) and genes (purple). Each track was normalized by the highest value in that track. Replicates were averaged together.

I wanted to determine whether DNA methylation was more strongly correlated with organ identity or with the individual from which the organs originated from. I performed Principal Component Analysis (PCA) using methylation rates for each cytosine in the genome. As shown in Figure 16A, samples clustered primarily according to organ systems (vegetative or reproductive). To assess the degree of similarity between the genome-wide methylation patterns, I calculated Pearson's correlation coefficient between all samples and performed hierarchical clustering using these correlation coefficients as a distance matrix (Figure 16B). In agreement with the PCA, the hierarchical clustering showed that samples belonging to either vegetative (leaves) or reproductive (flowers, siliques) organ systems grouped separately. Interestingly individual organ types did not cluster together. The first component of the PCA separated reproductive from vegetative organs when considering all cytosines (Figure 16A), as well as when separating cytosines by sequence context (Figure 16C). Interestingly the second component of the PCA using CG-methylation separated samples by the individual the organs originated from, such separation was not seen when looking at other sequence contexts.

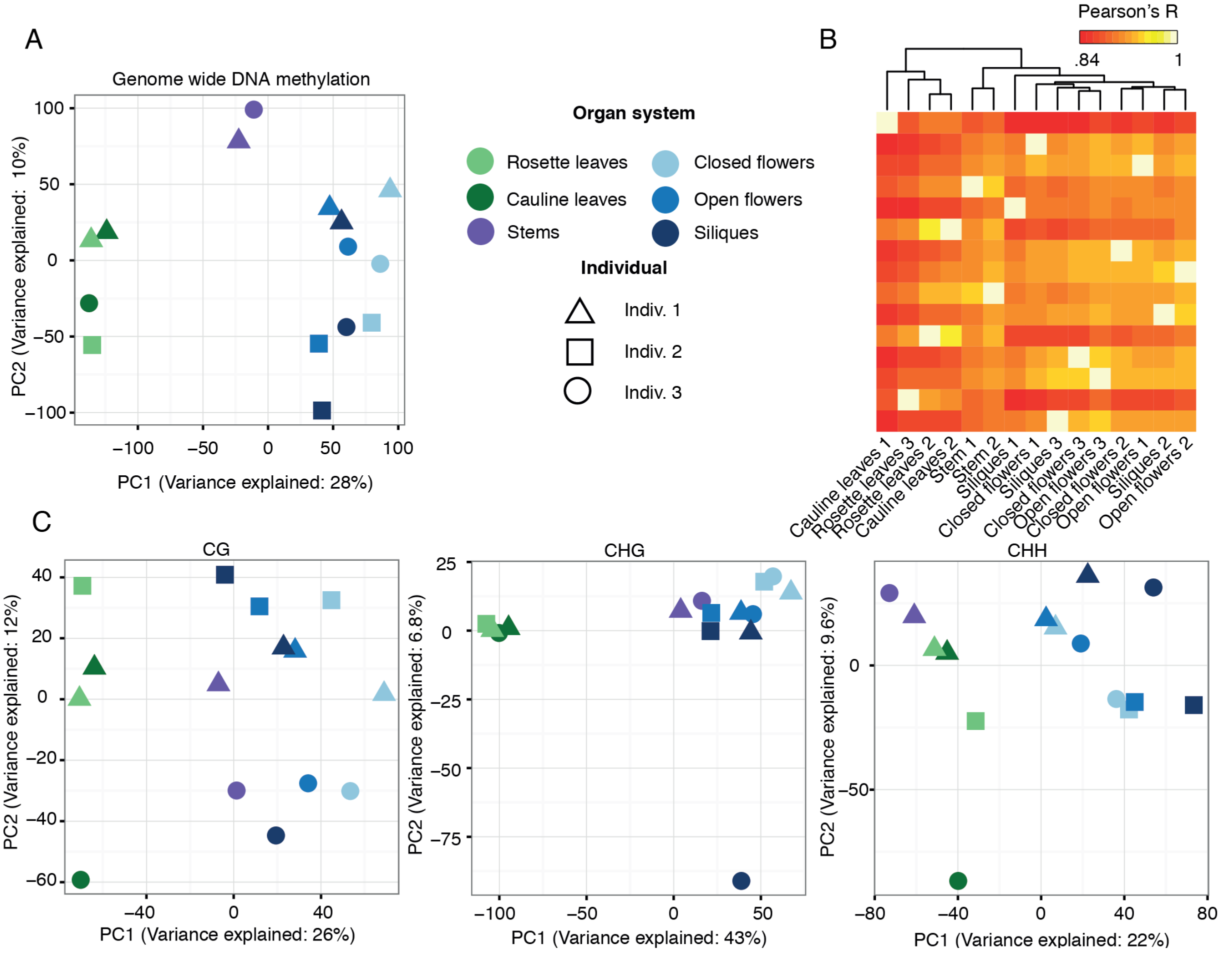


Figure 16 Genome-wide differences of DNA methylation between organ systems. A) PCA of the DNA methylation rates of individual cytosines from multiple organs of *A. thaliana*. B) Hierarchical clustering

analysis of organ types using the Pearson's correlation coefficient between samples. C) PCA of methylation rates from individual cytosines separated by sequence context.

The PCA showed that vegetative or reproductive organ systems have distinct methylation profiles. This could suggest that changes of DNA methylation occur during organ development, and that these changes are conserved across individuals. To measure the similarity of the methylation profiles, I calculated the correlation coefficients between all pairwise sample comparisons. The correlation coefficients ranged from 0.95 to 0.98 in all comparisons. This suggests that even though the genome-wide differences in methylation were sufficient to separate vegetative and reproductive organ systems, all organs showed an overall similar methylation profiles.

To identify DNA methylation differences between organs, I tested individual cytosines for differential methylation across all pairwise samples comparisons using previously described methods (see section 2.4). Differences between identical organs from different individuals can provide an estimate of spontaneous variation, while differences between organs could reflect organ-specific methylation patterns. I identified 813,461 differentially methylated positions (DMPs) with diverging methylation in at least one pair of samples, with an average of 26,888 differentially methylated positions per pairwise comparison.

I wanted to determine whether the identified DMPs were also correlated with organ identity. I performed a PCA using the methylation ratio of the identified DMPs (Figure 17). My results show the, similar to the previous analysis using whole-genome methylation rates, DMPs also separated vegetative and reproductive organ types apart. If the DMPs were product of stochastic variations between organ samples, the amount of variance explained by the PCA should remain the same. Instead, the amount of variance explained by PC1 was higher for DMPs than for whole-genome methylation rates, suggesting that the variability of these sites is not random and some sites responsible for most of the variation found across tissues. Furthermore the number of DMPs between tissues from similar types of organs (vegetative vs. reproductive) was lower than the number of DMPs found between tissues of organ systems, suggesting that the variability found within these positions reflect an association with organ identity.

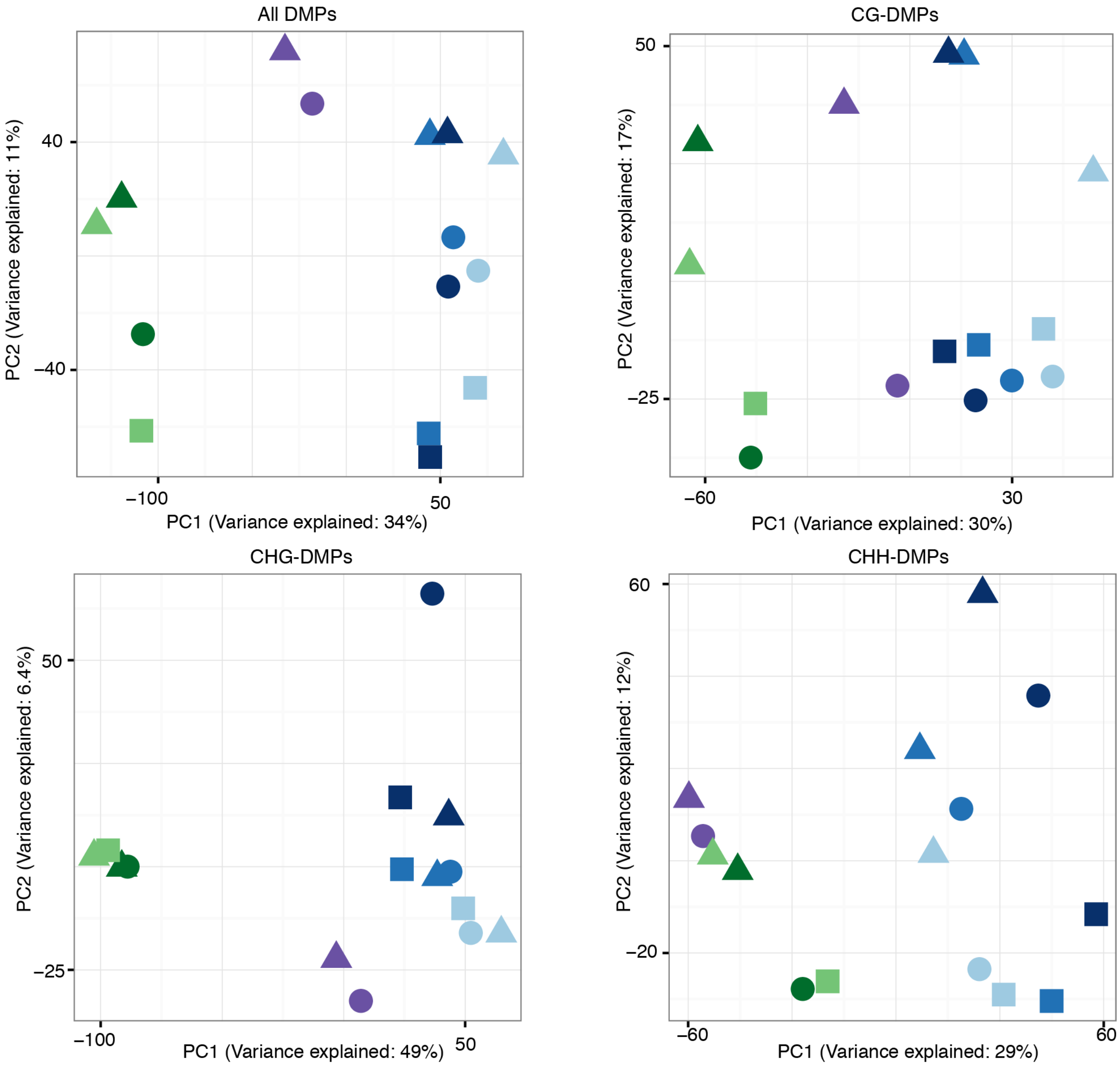


Figure 17 Differentially methylated positions separate samples by organ type. Principal component analysis methylation rates from DMPs separated samples by sequence context.

The functions of DNA methylation can change depending on sequence context (CG, CHG, and CHH). Furthermore, the DNA methylation maintenance mechanisms are different for each context. In order to investigate whether there was an association between sequence context and the variability found across tissues I repeated the clustering and PCA analysis separating individual cytosines by sequence context.

Hierarchical clustering separated vegetative and reproductive organ systems in all sequence contexts. My results also showed that identical organ types had similar

CHG-methylation profiles and therefore grouped together. In contrast, CG-methylation showed the least amount of organ specific differences, were samples were only separated between reproductive and vegetative tissues. Similar to the results presented above, the second component of the PCA using CG-sites separated samples according on the individual they originated from. These results suggested that methylation changes associated to development vary depending on sequence context.

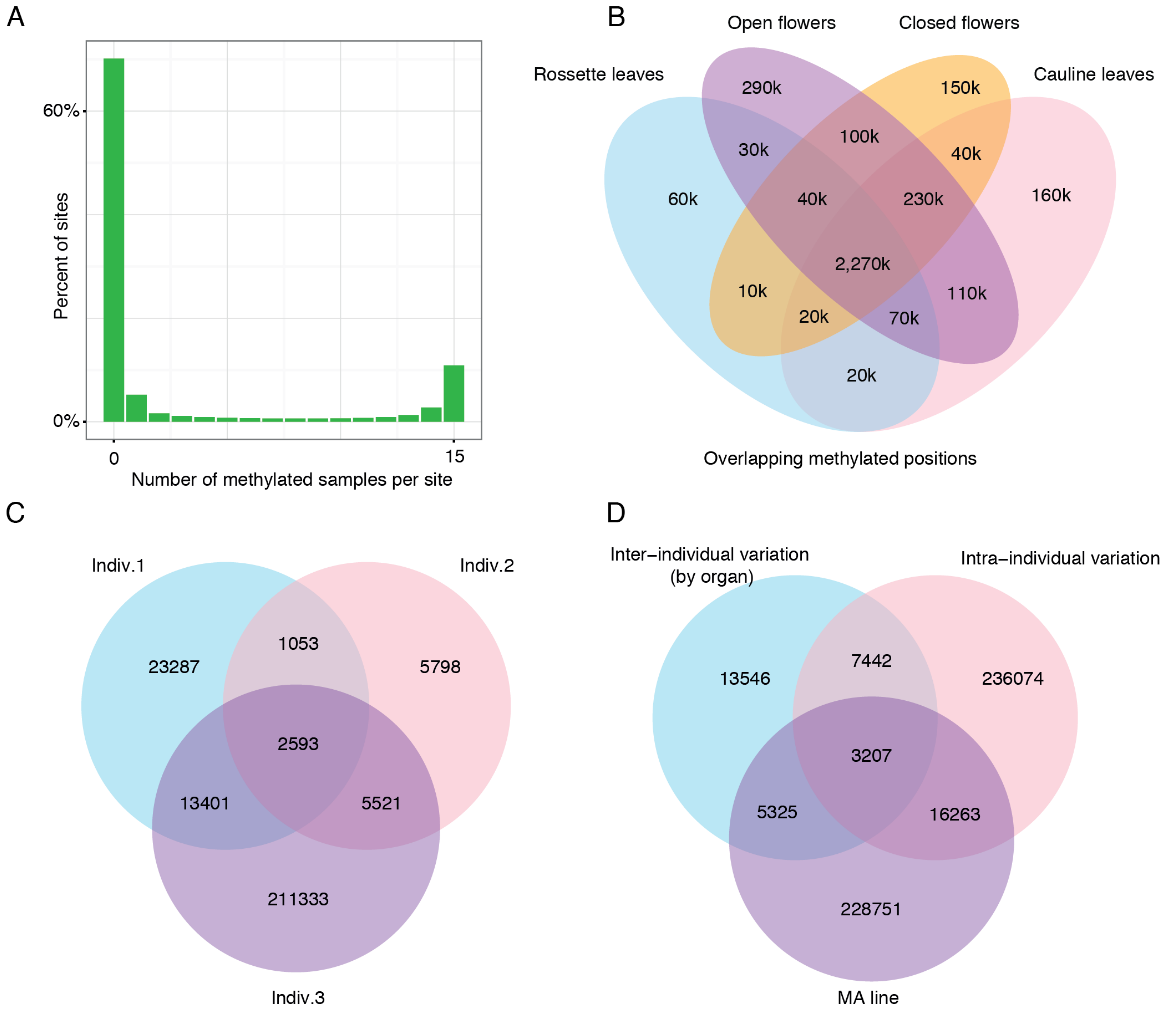


Figure 18 Inter- and intra-individual variation between organ systems. A) Proportion of cytosines according to the number of samples that are methylated, 0 represents sites where no sample was found methylated, 15 represents sites were all samples are methylated. B) Overlap of methylated positions between organ systems (k=1000). C) Overlap of the DMPs identified within each individual. D) Overlap of the DMPs found between identical organ types, with DMPs found between organs derived from a single individual and DMPs found in the MA collection line.

My previous results have shown that differences between methylation are sufficient to separate samples by organ systems, but it is still not clear which proportion of the genome is stable and which is variable. To measure the degree of stable DNA methylation between all samples, I calculated the frequency spectrum, resembling the site frequency spectrum but using DNA methylation as a polymorphism instead (Figure 18A). The majority of analyzed cytosines in the genome showed the same methylation status across all samples; 87.3 % of the cytosines that were methylated in one or more samples were methylated in all samples. This is in agreement with studies investigating transgenerational variability of DNA methylation as well as its changes in response to environmental stimuli, where most of the genome is stably methylated (Becker, Hagmann et al. 2011, Schmitz, Schultz et al. 2011, Dowen, Pelizzola et al. 2012, Seymour, Koenig et al. 2014).

Even though most of the cytosines in the genome were either methylated or unmethylated in most of the samples, the PCA still separated vegetative and reproductive organ systems. To identify the source of this variation, I calculated the overlap of methylated positions between all organs (Figure 18B). In agreement with my previous results, the majority of the positions that were methylated in one sample were methylated in all samples. The overlapping sites corresponded mainly to TEs and centromeric regions. As shown in Figure 18B, most of the positions that were methylated in leaves were also methylated in other organs, while reproductive organs had a higher number of methylated positions, mostly unique to those organs.

Trans-generational DNA methylation studies in near-isogenic *A. thaliana* lines separated by 30 generations have identified spontaneously occurring DMPs between individuals (Becker, Hagmann et al. 2011, Schmitz, Schultz et al. 2011). Additional studies comparing the methylation profiles of *A. thaliana* individuals separated by hundreds of generations have found that DMPs between those populations have a higher than expected overlap with the DMPs identified in the 30 generation study (Hagmann, Becker et al. 2015), suggesting that methylation at such sites are inherently labile. I wanted to determine whether such labile sites were also present in my samples. For each of the three individuals, I extracted the set of identified DMPs between organs of the same individual (intra-individual variation), and then determined how many of these DMPs where present in more than one individual (Figure 18C). My results showed that the overlap of such positions between the three individuals was higher than expected by chance (binomial test, P-value<2x10^{-16}). While this overlap supports the idea of labile sites, these variable positions were identified by comparing different organ systems within an individual and therefore could be the result of conserved developmental changes rather than true lability. To determine whether these DMPs are genuinely labile or they are just the product of development-associated changes, I identified DMPs between identical organ systems (inter-individual variants), sites that were variable between organs of the same individual (intra-individual variants, which is the combined set from figure 16C), and sites that were found variable in the MA accumulation lines (trans-generational variants) (Becker, Hagmann et al. 2011)). The overlap between all data sets was statistically significant (binomial test, P-value<2x10^{-16}). One out of ten sites that were found variable within individuals and one out of three variable positions between individuals were also variable in the MA lines (Figure 18D), suggesting that these sites are inherently labile.

While the importance of changes of methylation of individual sites is still debated, it is generally accepted that changes of methylation in multiple contiguous nucleotides are able to affect gene expression. Therefore I decided to focus next on methylation differences across regions rather than individual sites.

3.3.2 Variation of methylation across genomic regions

I identified methylated regions (MRs) across all samples using the Hidden Markov Model described in section 2.6.2 (Hagmann, Becker et al. 2015) and then tested these regions for differential methylation between samples. Subsequently, for every sample I calculated the methylation rate of each sequence context separately for each DMR by averaging the methylation rate of all cytosines present in that region. I used these methylation rates as a distance matrix to perform hierarchical clustering.

While vegetative and reproductive organs showed clear differences in all sequence contexts, a more detailed grouping arose within the grouping. Unlike genome-wide methylation and DMPs, where organs types only showed similarities in CHG-methylation, samples grouped together by organ type across all contexts when considering DMRs. The fact that this pattern is only found when looking at regions but not when looking at individual sites suggests that the methylation of these regions are changing through organ development across all sequence contexts.

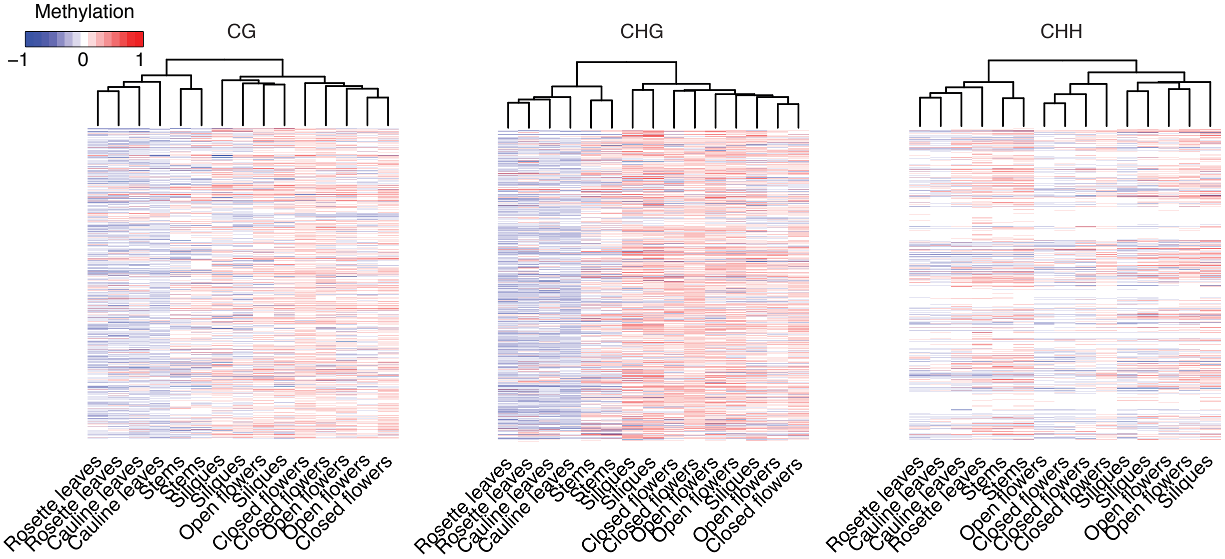


Figure 19 Differences in methylation at a region level. Hierarchical clustering of using the methylation rate of all identified DMRs. For each DMR, the average methylation rate of rosette leaves was subtracted from all samples. Red represents comparative gains of DNA methylation and blue losses of DNA methylation

A possible reason why identical organ systems clustered together when considering only DMRs is that the for each organ system, a single set of parameters for the beta-binomial distribution (which is used to model methylation) is calculated using the replicates for that organ system. This approach identified methylated regions within organ systems only when they were stable across replicates. While some bias could be introduced through this methodology, it is unlikely that this is enough to generate the observed organ-specific clustering. For this analysis, for each identified DMR, I calculated the average methylation rate of each sample independently and used these rates for the clustering. These include the methylation rates of samples that were not used to label a region as a DMR, and therefore are not biased by the beta-binomial model.

As seen in Figure 19, reproductive organs showed a relative increase of DNA methylation compared to vegetative organs in DMRs. Furthermore, this increase of methylation is found only in CHG and CG methylation, where CHG-methylation showed the relative gain. This difference in methylation is not equal in all organ systems, leaf tissues showed the lowest average gene body methylation, with shoot tissues showing intermediate levels, and reproductive tissues showing the highest levels.

My results have shown that even though vegetative and reproductive tissues show differences in DNA methylation in both single sites and regions, individual organ types within one class do not substantially differ from each other at the genome-wide methylation level. I thus decided to investigate methylation differences at specific genetic elements, such as at transposable elements and in gene bodies could have stronger association to organ identity.

3.3.3 Gene Body methylation

Gene body methylation, defined as the methylation present in exons of genes, has been shown to correlate with gene expression (Zhang, Yazaki et al. 2006, Zilberman, Gehring et al. 2007). This association depends on the sequence context of the methylation present. Gene expression levels correlate negatively with an increase of CHH and CHG methylation. In contrast CG-methylation is low in genes showing either high or low levels of expression, and is normally associated with constitutively expressed genes. Previous studies have shown that identical organ systems show similar expression profiles. I therefore wanted to determine whether tissue-dependent effects on gene expression could be correlated with organ-specific gene body methylation.

To address this question, for each gene in the TAIR10 annotation (Berardini, Reiser et al. 2015), I calculated the average gene body methylation by context and each sample Figure 20A. Similar to my previous results, reproductive and vegetative tissues could be separated by only using gene body methylation profiles. Surprisingly CHG-methylation showed strong organ-specific differences.

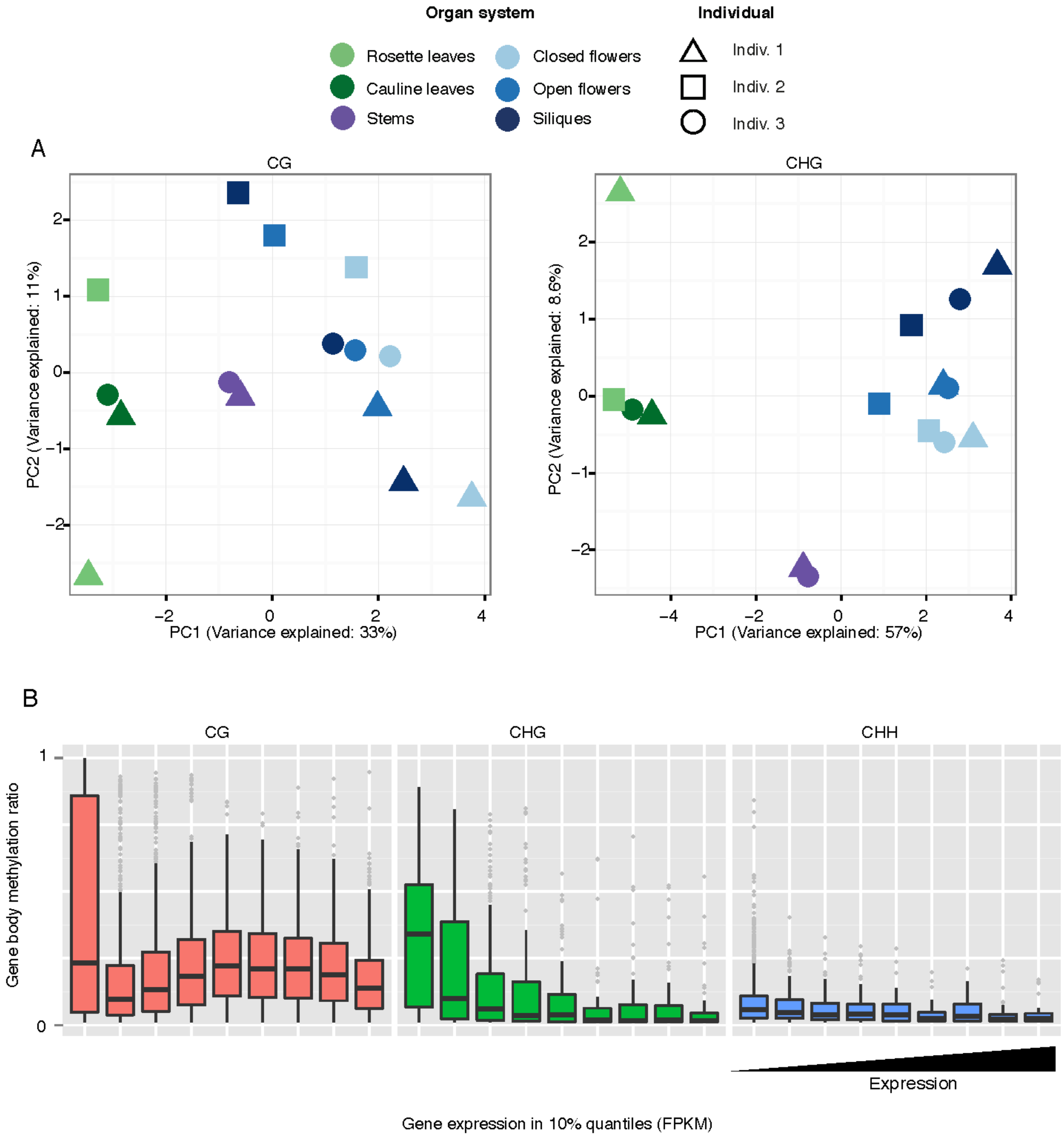


Figure 20 Gene body methylation reveals tissue specific patterns. A)Principal component analysis of DNA methylation rate in gene bodies separated by context. B)Distribution of DNA methylation in genes according to their expression levels. Genes were separated by expression levels in 10% quantiles.

I therefore hypothesized that gene body methylation could provide an added layer of tissue specific gene regulation. To explore the relationship between gene expression and DNA methylation, I linked the gene body methylation profiles to RNA-seq data produced from the same tissue type from the same individuals. Clustering analysis confirmed that samples from identical organ systems showed similar transcription profiles and therefore grouped together. Similar to published studies, all organ systems showed a strong correlation between CHG methylation within gene bodies and low levels of expression while genes with either high or low levels of CG methylation showed low levels of expression (Figure 20B).

Next, I wanted to determine whether changes in DNA methylation between organs were also associated with changes in gene expression, since I had already found that vegetative and reproductive organs have distinct methylation profiles. To reduce the dimensionality and complexity of the analysis of methylation and expression changes across all organ systems, I focused on analyzing two representatives organ systems from the reproductive and vegetative lineages: rosette leaves and closed flowers. In agreement with my previous results, closed flowers had a higher average methylation rate at gene bodies than rosette leaves Figure 21A. I selected the 500 genes with the highest amount of difference in methylation between both organs. From this set, all of them were relative gains of methylation in flowers compared to leaves. From each gene in this set I calculated the average methylation rate for all three sequence contexts, and compared these rates with their expression differences. Genes without read counts were excluded from the analysis. Linear regression revealed an association between changes in DNA methylation and changes in gene expression. CG and CHH showed the strongest association with R^2 values of 0.16 and 0.15 respectively (p. value $=2.81x10^{-5}$ and$4.97x10^{-5}$).

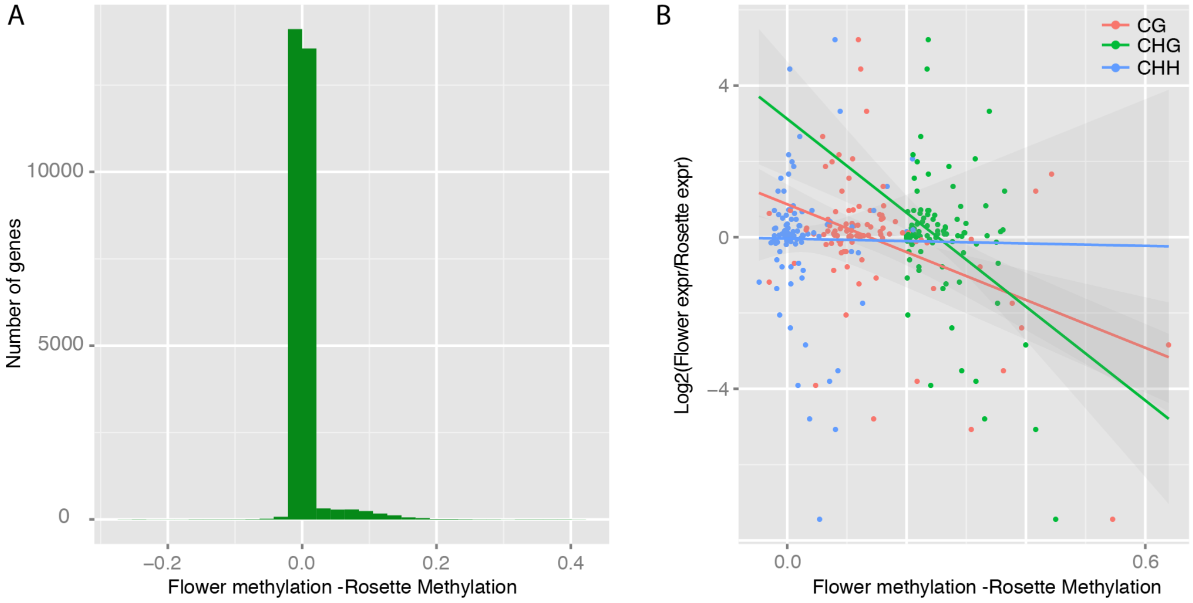


Figure 21 DNA methylation changes are associated with changes in gene expression. A) Histogram of genes according to the difference in methylation rates between closed flowers and rosettes. B) Changes in gene expression plotted against changes in methylation between closed flowers and rosette leaves. The expression changes are shown as log2 values of expression of flower divided by the expression in rosette leaves.

3.3.4 Transposable element, DNA methylation and gene expression

Large plant genomes are composed mainly of transposable elements (TE), and even in a species such as *A. thaliana* where TEs are not as highly abundant, TEs account for at least 10% of the genome (The Arabidopsis Genome Initiative 2000, Buisine, Quesneville et al. 2008). Because TEs can have harmful effects, plants have developed numerous mechanisms to inhibit TE mobility. One of these is to mark TEs with DNA methylation, which in turn recruits components of the TE silencing machinery (Slotkin and Martienssen 2007). TEs are normally heavily methylated which causes transcriptional silencing of these mobile elements preventing transposition (Zemach, McDaniel et al. 2010). In the event of loss of methylation, such as during gametogenesis, TEs get reactivated and start proliferating (Slotkin, Vaughn et al. 2009). Furthermore, the transposition of a TE can cause changes at a transcription level of genes near the site of insertion (Hollister, Smith et al. 2011). Some of these changes are caused by the spreading of TE-associated methylation to regulatory regions of neighboring genes (Arnaud, Goubely et al. 2000, Ahmed, Sarazin et al. 2011). This propagation of methylation from TEs represents a potential mechanism in which DNA methylation can affect gene regulation.

To investigate whether differences in DNA methylation of TEs near genes could explain some of the differences in expression levels across organs, I used the TAIR10 genome annotation to identify protein-coding genes with a TE within 5kb of either side of the transcribed region. I classified the genes into two categories, depending on whether or not they had a methylated region overlapping with the TE. Finally, I used the transcription profiles of each organ system and compared the expression levels of each gene to the distance of the TE to the gene.

As shown in Figure 22 the presence of unmethylated transposons near protein coding genes does not seem to affect the expression levels of genes. By contrast, protein-coding genes with a neighboring methylated transposon (methylated region overlapping the transposon) showed a decrease in gene expression levels. In fact, the expression level of genes was inversely proportional to the distance of the methylated TE to the gene. Expression was the lowest at a TE-gene distance of up to 1kb, the effect of expression was lost completely at a distance of 5kb or longer.

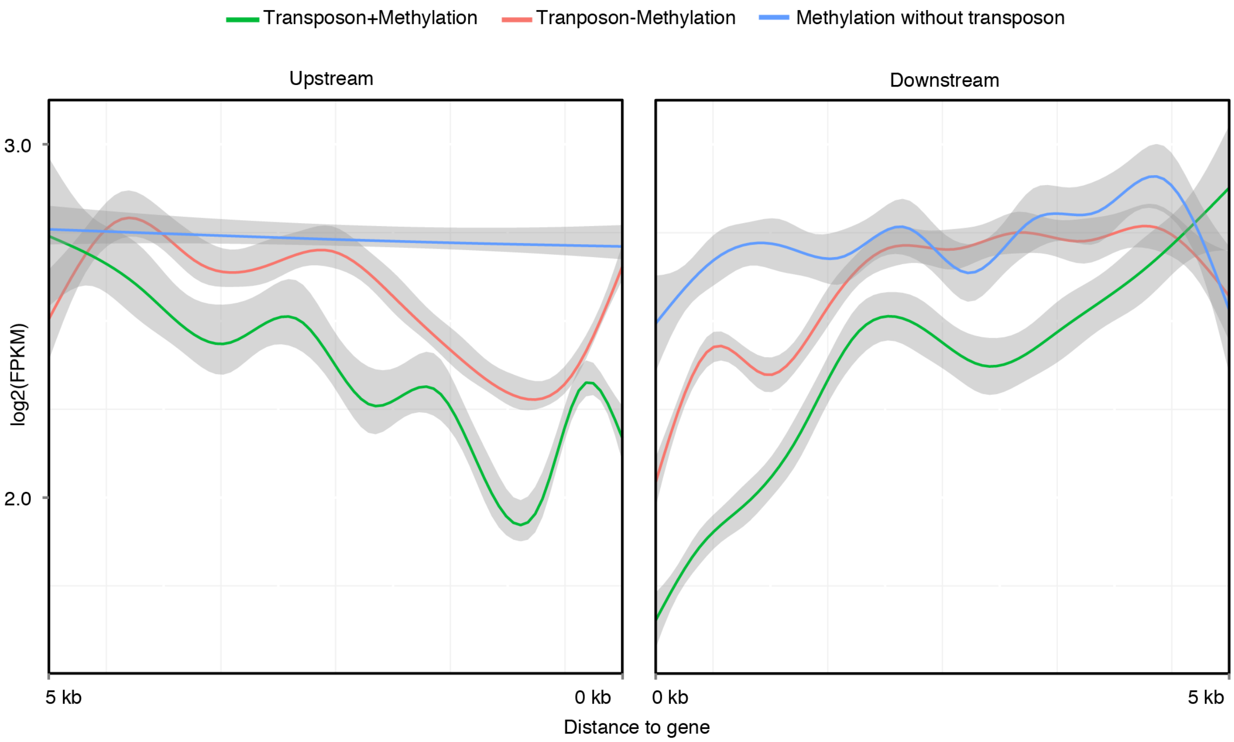


Figure 22 Effects of adjacent methylated regions to gene expression. Average gene expression of genes where a methylated TE (blue), unmethylated TE (red) or predicted methylated regions that did not overlap with a TE were present within 5kb upstream or downstream of the gene. Values are plotted against the distance of the TE or the methylated region to the gene. Methylation and expression were evaluated in a per individual basis. A generalized additive model was used for smoothening.

While my results suggest that methylation in TEs near regulatory regions affects gene expression, methylation alone could be causing the silencing of the genes, independent of the presence of a TE. To test this hypothesis, I used the TAIR10 genome annotation to look for genes without a TE within 10 kb upstream or downstream, but for which I had found an MR within 5kb. As shown in Figure 22, genes with a nearby MR but without TEs had a higher average expression than genes with either unmethylated or methylated TEs, suggesting that methylation alone is not sufficient (or at least less effective) to decrease gene expression when present near regulatory regions.

4 Discussion

The biological importance of epigenetic marks and in particular DNA methylation in plants and animal has been demonstrated repeatedly in the past couple of years (Law and Jacobsen 2010, He, Chen et al. 2011). Multiple reports have linked phenotypic differences to changes in DNA methylation, some of them including trait variation in commercially important species (Ong-Abdullah, Ordway et al. 2015). The sources of variation of DNA methylation have been studied extensively. Such variation can occur stochastically (Becker, Hagmann et al. 2011, Schmitz, Schultz et al. 2011), be induced by exposure to environmental conditions (Dowen, Pelizzola et al. 2012, Seymour, Koenig et al. 2014, Secco, Wang et al. 2015), or be associated to organ identity (Seymour, Koenig et al. 2014, Widman, Feng et al. 2014). Most of the studies in plants investigating the role of DNA methylation in development have focused on early stages where methylation is highly dynamic, but variation across organs at later stages remains largely unexplored. In this study I have performed a detailed genome-wide survey of the variability of DNA methylation across multiple organs and multiple individuals to describe inter- and intra-individual variation in *A. thaliana*.

To characterize such variation of DNA methylation across organ types I generated single-base-resolution DNA methylation maps using whole-genome bisulfite sequencing of three different collections of samples, each designed to capture different sources of epigenetic variation. The first set consisted of pools of leaves, where each pool was separately collected from 14 individuals, all of which were derived from a single parent. This dataset allowed me to investigate variability of DNA methylation between very closely related individuals that had experienced the same environmental conditions. The second set consisted of 18 separate leaves harvested from a single individual. This dataset was designed to answer questions regarding the effect of organ age on DNA methylation. The third set was composed of six different aerial organs, each collected from three very closely related individuals, allowing me to interrogate organ-specific patterns of DNA methylation. Furthermore, all samples were collected from individuals that shared a single parent allowing for cross comparisons.

My data showed that organ identity was the main factor contributing to variation of DNA methylation. Vegetative organs (rosette leaves and cauline leaves) showed a relative loss of DNA methylation compared to reproductive organs (flowers and siliques). A second source of variation in methylation was associated to age, were leaf samples of at different developmental stages showed genome wide differences in their methylation patterns.

By coupling DNA methylation data to RNA-seq data, I was also able to correlate methylation patterns and gene expression levels. This revealed that methylation at gene bodies correlated strongly with organ identity, and that differences in gene body methylation were associated with differences in gene expression. I could also show that cytosine methylation in regulatory regions alone is probably not sufficient to induce gene expression, rather it needs to be coupled to TE induced gene silencing.

Finally, based on a previous pipeline I implemented a new self-contained pipeline designed to identify DNA methylation and differential DNA methylation from bisulfite sequencing data. Compared to the original implementation pipeline, my improved pipeline requires fewer intermediate steps, and at the same time comes with significant speed-ups. The pipeline was designed with flexibility and scalability in mind, being able to be run in stand-alone desktop computers as well as being able to take advantage of high performance computing platforms. I also implemented a new method to mitigate the impact of multiple testing correction on the detection of methylation and differential methylation.

4.1 New methylation pipeline

I developed a new pipeline with tunable parameters to accommodate differences in available RAM as well as processing cores. A second version is available allowing the use of the Sun Grid Engine (SGE) directly in the pipeline, allowing the use of already efficient scheduling systems.

A main feature of the new pipeline is the use of a new method to mitigate the loss of statistical power caused by multiple testing correction. This method is based on the concept of minimum attainable p-value (Tarone 1990), and uses it to filter out comparisons where statistical tests are not informative. The reduction of the number of statistical inferences made, reduces the effect of multiple testing correction on the power of detection. This method requires a function to calculate a corrected significance value from only the number of statistical tests to be performed (refer to section 3.1). For a proof of concept, I used this method coupled with a Bonferroni correction procedure controlling for the FWER, but the obtained correction was still very conservative. While this method can in principle use any other correction function, the implementation with other multiple testing correction procedures is still lacking. Nevertheless it has been shown to be a viable in other studies (Llinares-Lopez, Grimm et al. 2015). When performing a large number of statistical tests, one might want to focus on the proportion of type I errors instead of controlling for the FWER. Methods to control for this value, known as the False Discovery Rate (FDR), have been widely implemented and are also incorporated in my pipeline. It is tempting to speculate that one can further improve such methods by incorporating the concept of a minimum attainable p-value into this type of procedures, but more research in this area would be needed for implementation.

Finally, the pipeline was designed for a simple user experience. After installing all necessary libraries (one C++ library and two R libraries), it only takes one single command to run the entire pipeline from start to end, including MP and DMP analyses. Such simplicity will make the analysis of such type of data accessible to any laboratory interested in performing bisulfite-sequencing experiments, even with very limited bioinformatics experience.

4.2 Short term variation of DNA methylation

I made use of my new bisulfite analysis pipeline to study intra- and inter-individual variation of cytosine methylation in *A. thaliana*. The analyses revealed different factors that contribute to the variation found on a genome-wide level.

4.2.1 Variation of DNA methylation in rosette leaves of *A. thaliana*

To study the variation of DNA methylation in rosette leaves, I generated methylation maps of multiple individual leaves from single plants. The first analyzed set composed of 18 individuals showed that older leaves (i.e. initiated earlier during the plant life cycle) have lower methylation levels compared to younger leaves. There are many biological processes that could lead to such an observation. The oldest leaves of the 5-week old plants I harvested had started the senescence program. During the final stages of leaf development, senescence-associated degradation and relocation of cellular components ensures recycling of nutrients in other organs (Tsukaya 2013).

Senescence is a highly regulated and coordinated process, characterized by many changes in metabolism, gene expression and cell structure. The maintenance of DNA methylation requires numerous different proteins; it would not be surprising that the correct maintenance of DNA methylation stopped being a priority in such late stages of development, which could lead to losses of DNA methylation. Furthermore, methyltransferases need metabolites to act as methyl donors (S-adenosyl methionine) to methylate cytosines (Wu and Santi 1987). Some studies have shown that metabolic changes can affect DNA methylation in the genome (Liu and Ward 2010). Another possible explanation is that the metabolic changes occurring during senescence are affecting the availability of methyl donors, causing the observed pattern.

A second factor to consider is that leaves can transition to endoreplication cell cycles, or endocycles for short (Lee, Davidson et al. 2009, Edgar, Zielke et al. 2014). During these cycles, DNA is replicated without cell division leading to an increase in cellular ploidy level. The ploidy levels have been shown to correlate with age and have been shown to be developmentally regulated (Galbraith, Harkins et al. 1991, Gendreau, Orbovic et al. 1999). While DNA methylation has yet to be studied in the context of leaf development and endoreduplication, studies using induced tetraploid *Arabidopsis* hybrids have shown that changes in ploidy levels lead to changes in DNA methylation (Ng, Lu et al. 2012, Tian, Li et al. 2014). Endocycles are accompanied by many transcriptional changes; such changes could also lead to changes in DNA methylation (Beemster, De Veylder et al. 2005). Expression of *MET1* and *CMT3* is not constant in different portions of *A. thaliana* leaves, which have different levels of cells going through endoreduplication (Galbraith, Harkins et al. 1991, Melaragno, Mehrotra et al. 1993). The regions with higher endoreduplication showed lower expression of methyltransferases, supporting this hypothesis. One problem of measuring DNA methylation on segments of leaves is that endoreduplication is not perfectly distributed across the leaf, furthermore more the ploidy level can vary significantly from an average of 8n to 64n, as such the exact effects of ploidy number cannot be directly assessed (Melaragno, Mehrotra et al. 1993).

I also performed whole-genome bisulfite sequencing of individual leaves from a 3-week-old and a 5-week-old individual. These replicates did not confirm the association of leaf age with methylation found in the 18-leaves-set. It is still unclear why each leaf set behaves differently, but it could be due to epigenetic/genetic differences between the lineages used in the two experiments. From the comparison of these datasets against published methylation data sets, plants derived from the MA lines cluster separately from the additional replicates supporting this hypothesis. Regardless, more replication is necessary to make any definitive conclusions.

While the effect of age in individual leaves remained inconclusive, one pattern was consistent in all data sets: the variability of DNA methylation between individual leaves originating from a single individual was higher than the variability between pools of leaves from distinct individuals. My analysis thus showed that DNA methylation varies even between organs of an individual.

4.3 Organ specific DNA methylation

A second aim of this study was to describe the association of DNA methylation with organ identity. The analysis of methylation maps of aerial organs showed that organ types had unique methylation profiles.

Previous studies have shown a considerable amount of spontaneous variation of DNA methylation after a few generations (Becker, Hagmann et al. 2011, Schmitz, Schultz et al. 2011). These changes can be a major source of inter-individual variability. One aim of this study was determine whether organ-specific differences are the product of the accumulation of somatic epimutations or are developmentally driven.

In order to test for this I used the epimutation rate calculated in transgenerational studies as a baseline of expected epimutations between organ types. My analysis showed that the number of epimutations between different organ types was higher than expected due to spontaneous variation, thus is unlikely that spontaneous variation alone is responsible for such differences in DNA methylation between organs. One caveat from using the epimutation rate from transgenerational studies as a Null expectation is that plants go through a partial resetting of DNA methylation during early stages of development. It could be possible that there is a high degree of somatic epimutations, but such epimutations are efficiently removed during this reset and therefore are not heritable, causing the transgenerational epimutation rate to be much lower than the “real” epimutation rate.

This could explain why the discrepancy in the number of epimutations between organs compared to what would be expected by the epimutation rate in trans-generational studies. Even if this was the case, one would expect that the epimutations (gains and losses of DNA methylation) should be equally distributed between the organ samples. My data shows that reproductive organs show a relative gain of DNA methylation compared to vegetative organs, supporting the hypothesis that these epimutations are not stochastic, rather they are systematic and directional among the organs of *A. thaliana*.

Genome-wide methylation data showed organ-specific differences (Figure 16). These differences became more evident when limiting the analyses to only DMPs (Figure 17). In both cases reproductive and vegetative organs showed distinct methylation profiles. This separation was also observed when separating the interrogated cytosines by sequence context and in all cases organ identity the largest source of variation. The CHG sequence context provided additional resolution as replicates of individual organs formed distinct clusters in the PCA (Figure 17). These findings suggest that if methylation is contributing to the maintenance of organ identity it is likely to be acting through CHG methylation, as this organ-specific separation is not observed in any other context. This hypothesis is further supported by my results in gene body methylation, where CHG methylation shows an even stronger separation between organ types, but is also not present in the other two sequence contexts. Furthermore, when comparing rosette leaves against closed flowers, the differences of DNA methylation in genes correlate with differences in gene expression. This correlation is found both in CG and CHG contexts independently. The overlap of CG and CHG methylation could be due to the interaction of methyltransferases with histone modifications. In particular H3K9 has been closely associated to both CG and CHG methylation (Jackson, Lindroth et al. 2002, Tariq, Saze et al. 2003).

In the case of CG methylation, the second component of the PCA separated organs by the individual from which they originated, but only for CG-methylation. One possible explanation for this pattern is that due to the high levels of DNA methylation at CG-sites, small changes of DNA methylation are statistically significant, but biologically insignificant, in that they do not affect methylation at a functional level, and therefore are tolerated. In this scenario, spontaneous epimutations could accumulate in organs from an individual giving rise to the observed patterns. A second explanation could be that there are differences in the epimutation rates between different methylation contexts, and epimutations accumulate more rapidly in the CG contexts. This hypothesis is consistent with transgenerational studies that have showed a higher accumulation of epimutations in a CG context compared to other contexts. These differences in epimutation rates could be due to the fact that the establishment and maintenance of DNA methylation is mediated by different molecular machineries in each sequence context.

The strong association of CHG methylation with organ identity and the individual differences in CG methylation are reminiscent of the changes in DNA methylation triggered by a stress response (Dowen, Pelizzola et al. 2012, Wibowo, Becker et al. 2016). Previous studies have shown that stress affects methylation at each sequence context differently. In the case of salt stress, reproducible locus-specific gains and losses are observed in CHH and CHG contexts. By contrast salt stressed individuals show stochastic losses and gains of CG-methylation across the genome, similar to the patterns I observed at an organ level. The RdDM pathway (described in section 1.4.2) could provide a mechanism of tissue specific regulation, as this pathway could induce tissue-specific methylation through the regulation of the expression of siRNAs. There is evidence that siRNAs can spread between tissues (Molnar, Melnyk et al. 2010), which could facilitate gene regulation within tissues.

There have been studies showing that CG-methylation at gene bodies correlate with intermediate levels of gene expression through the exclusion of the histone variant H2A.Z. In contrast to CG methylation, CHH and CHG methylation is mostly associated with genes that are expressed at low levels (Zhang, Yazaki et al. 2006). This pattern was also observed in all organs investigated in this study. The fact that organ-specific differences in gene body methylation were not found in CG sequence contexts but only in CHG sequence contexts could be due to the preferential association of CG-methylation with constitutively expressed genes, which are expressed in all organ types. In contrast changes of CHG methylation might be involved in the activation or silencing of organ specific genes.

A central aim of this study was to assess the stability of methylation between different organs of *A. thaliana*. I found that the majority of cytosines in the genome are stably methylated across all organ types analyzed (Figure 18). Most of the methylated sites were found across TEs, repeat-rich regions and centromeric regions as expected (Cokus, Feng et al. 2008, Lister, O'Malley et al. 2008). When considering the amount of methylated sites that overlapped between different organs, rosette leaves had the highest amount proportion of sites that were methylated in at least another organ type (Figure 18). By contrast, a high proportion of methylated cytosines in later arising organs (reproductive organs) were not shared between organs that arise at earlier developmental stages. This correlation between gains of methylation and the temporal order in which organs were initiated was observable at both a single site and a region level.

In plants, organogenesis occurs throughout the life of the plant, and new organs arise at from the flanks of a pool of stem cells, this pool of cells is known as the shoot apical meristem (SAM) (Murray, Jones et al. 2012). At early stages, the SAM gives rise to the leaves and stems, and to flowers and siliques during the plants reproductive phase. One hypothesis is that the observed differences between early and late arising organ types is the product of the accumulation of methylation changes in SAM, this could explain why later arising organs have an increased proportion of non-shared methylated sites. The SAM is not going through active cellular division; epimutations caused by passive loss of DNA methylation due to DNA replication are rare, this could help explain why the observed epimutations between later arising tissues are directional gains of DNA methylation.

An alternative hypothesis is that there is a loss of DNA methylation in leaf organs instead. My results with of sets of individual leaves showed there is an age-dependent decrease of the expression of genes encoding proteins required for establishment of CG and CHG methylation. These transcriptional changes in leaves could lead to a generalized loss of methylation and explain the observed patterns.

Additional experiments surveying the methylome of the SAM over the course of the plant development would help determine the directionality of the changes in DNA methylation in leaves as well as to test if DNA methylation changes are accumulating in meristems.

Almost certainly, stochastic epimutations also contributed to the observed variability of DNA methylation between organs. A high proportion of the DMPs between identical organs from different individuals (inter-individual variation) overlapped with sites that have been identified as highly variable in transgenerational studies. Furthermore, a high proportion of the DMPs between organs of a single individual (intra-individual variation) also overlapped with these hyper variable sites. My study thus suggests that there is a high amount of stochastic variation of methylation in plants, and most of this variation is not conserved, nor transmitted to the next generation. This hypothesis raises the question: if methylation is so variable, how can it be subject to natural selection, and how can it have biological function. My hypothesis is that most of the DNA methylation in the genome is not subject to natural selection at a single site level; instead methylation is selected at a region level, where a combination of methylated sites have a functional roles. This point of view is consistent with my observations that DNA methylation show clearer organ specific pattern when focusing on regions as opposed to individual sites. I conclude that variability in methylation patterns results likely from a combination of systematic changes that occur through development and stochastic variation between and within individuals.

4.3.1 Transposable elements, DNA methylation and gene expression

One of the best-characterized functions of DNA methylation is the silencing of TEs (Slotkin and Martienssen 2007). Loss of TE methylation can cause transcriptional reactivation and induce transposon mobility (Miura, Yonebayashi et al. 2001). TE insertions can have deleterious effects, e.g. by disrupting genes or regulatory sequences. Additionally, TEs when integrated near or in gene regulatory regions can alter the expression pattern of genes through the recruitment of the transposon silencing machinery, which include components for the machinery for DNA methylation (Morgan, Sutherland et al. 1999). The recruitment of such machinery leads to the methylation of the newly inserted TE, this methylation can then spread to neighboring regions and cause silencing of genes in the vicinity (Jordan, Rogozin et al. 2003).

Several studies have addressed the effects of TEs on nearby gene expression at a genome-wide level. Some of them have inferred DNA methylation levels by using the presence of siRNAs target in TEs as a proxy (Hollister, Smith et al. 2011). Such indirect measurements can only provide information about the TEs that are target for DNA methylation, but they do not provide direct information regarding the effects of methylation per se. My study provides a more detailed picture of the relationship between DNA methylation, TEs and gene expression.

My findings are in agreement with previous studies (Hollister, Smith et al. 2011), which reported a lower average expression of genes containing a methylated TE in close proximity. The apparent repressive effect of methylated TEs on gene expression was on average less pronounced the further away the TE was located from the gene. I found that methylated TEs correlated negatively with gene expression both when found upstream and downstream of the genes (Figure 22). Only few instances have been described where TE methylation located downstream of genes affected gene expression. It has been recently shown in salt stress experiments that the a loss of DNA methylation in a TE downstream of the *CARBON/NITROGEN INSENSITIVE 1* (*CNI1*) gene in *A. thaliana* confers resistance to hyperosmotic stress (Wibowo, Becker et al. 2016). In this case the loss of DNA methylation leads to transcription of a long non-coding (lncRNA) RNA transcript that in turn reduces the expression of *CNI1*. Furthermore it was shown that DMRs in stressed plants overlap more frequently with downstream regions transcribing antisense lncRNAs than DMRs in transgenerational studies (Wibowo, Becker et al. 2016), suggesting that downstream methylation of TEs might play a more broadly role in stress acclimation.

My study suggests that downstream methylation could be contributing to changes in gene expression to a greater extent than previously envisioned. While it is clear that DNA methylation can modulate transcription dynamically, most TEs investigated in this study overlapped with methylated regions and were found methylated in all samples arguing against a role in tissue-specific gene regulation.

5 Outlook

I have presented a detailed comparison of methylation in organs of *A. thaliana*. To deepen our understanding of the role of DNA methylation in organ development of plants, numerous future avenues could be pursued, based on the data presented here.

One possible extension of the organ study would be to investigate not only the changes that occur between individual organs in plants at a defined stage, but to analyze also tissue of the same organ type at multiple developmental time points. Such an experiment would help to determine the temporal distribution of DNA methylation changes, and to elucidate how developmental transitions affect DNA methylation in organs and how age changes methylation in different organs of the plant.

In addition to this, the inclusion of a broader set of organs would provide a more detailed picture of the relationship between DNA methylation and organ identity. In particular, roots would be of high interest as they are derived from a different set of stem cells. It remains crucial to determine whether changes in DNA methylation are occurring in stem cells and at what frequencies (shoot apical meristem and root apical meristem), as these cells are the progenitors of all organs in plants. While it is technically difficult to isolate meristems, new technologies designed to extract nuclei only from specifically tagged cells, such as the INTACT method (Isolation of Nuclei TAgged in specific Cell Types) (Deal and Henikoff 2011) could help cope with this problem. An alternative is the use of fluorescent-activated cell sorting (FACS) to obtain pools of cells from a desired cell line. This approach has been used to study DNA methylation of different cell types in root of *Arabidopsis thaliana (Kawakatsu, Stuart et al. 2016)*, though it requires the availability of GFP reporter cell lines.

There are many sources that contribute to methylation variability, including but not limited to plant environment, stochastic variability, and temporal changes of DNA methylation and cell identity (Law and Jacobsen 2010, He, Chen et al. 2011, Becker and Weigel 2012). My data showed that there was more variability between individual leaves derived from a single individual than variability between pools of leaves from different individuals. This result showed that by pooling organs, some fraction of the biological variation gets masked in the analysis. An additional problem with using pools of cells to measure DNA methylation is that even relatively simple organs are composed of a heterogeneous mix of different tissue types. As an example, leaf architecture is complex, and one could argue that studying DNA methylation patterns of whole leaves is a gross over-simplification. Additionally, studies have shown that there is also variability in gene expression between individual cells. By sequencing pools of cells, it is not possible to determine whether differences in methylation and gene expression are caused by large-scale changes in a subset of cells, or whether these changes are coordinated and distributed homogenously across many cells. Single cell sequencing, the ultimate technology to address cell-specific questions, has already been applied successfully to study DNA methylation and gene expression profiling (Smallwood, Lee et al. 2014, Wu, Neff et al. 2014) and can circumvent many of the previously mentioned problems.

It has been shown that DNA methylation acts in concert with other epigenetic marks, more prominently with histone modifications and small RNA products (Cedar and Bergman 2009, Matzke and Mosher 2014). Despite this, the exact nature of the relationship between DNA methylation and such marks is not well understood. It remains a long-standing question whether DNA methylation is just a product of other epigenetic modifications or transcriptional changes. In order to gain full understanding of DNA methylation, it is crucial to start studying it in combination with other epigenetic players. The use of CHIP-seq and RNA-seq of sRNAs will provide valuable opportunities to study the interplay of DNA methylation with chromatin and transcriptional changes.

Arabidopsis thaliana is a model plant with a high quality reference genome, short generation time, and a small genome; these factors have facilitated the study of DNA methylation in plants. Despite these advantages, mutations that disrupt the establishment and maintenance of DNA methylation have small effects compared to other plant species like maize (Herr, Jensen et al. 2005, Pontier, Yahubyan et al. 2005, Parkinson, Gross et al. 2007, Erhard, Stonaker et al. 2009). This is likely due to the fact that these species have a much higher content of TE elements, thus DNA methylation might play a more prominent role on the regulation of genes in such species. A previous study comparing closely related plants from the *Brassicaceae* family have shown that genome architecture, and in particular TE element composition play a significant role shaping the evolution of DNA methylation (Seymour, Koenig et al. 2014), as such the expansion of DNA methylation profiling into plants with high TE content might provide valuable insights into the functional roles of DNA methylation.

With a true and full understanding of DNA methylation, through approaches such as the one I have performed and discussed here, it might become possible to exploit the properties of DNA methylation to create (epi)genetic regulatory circuits. One could potentially find regions of the genome that are targeted for methylation or demethylation, and use these DNA sequences as regulatory components of artificial regulatory circuits. With proper understanding of the variation of DNA methylation across development and environment, studies on DNA methylation could be used for synthetic biology.

6 Materials and methods

Plant growth and material

Sibling's data set, 18-leaves data set and organ data set

Seeds of a single individual from the MA collection line (3rd generation, Col-0 background, line ID 0-4-27) were stratified in soil for 3 days at 4^{o} in short day cycles (8 hours light, 16 hours dark). Plants were transferred to 23 C with long day conditions (16 hours light, 8 hours dark). After 5 weeks, 14 individuals were dissected and pools consisting between 4 to 6 leaves of each individual were collected and pooled separately. An additional individual was dissected and each of its 18 leaves were collected separately. A final set of 3 individuals was dissected to collect aerial tissues consisting of siliques, open flowers, closed flowers, stems, cauline leaves and rosette leaves. From rosette and cauline leaves, 5 leaves of each were pooled separately. For floral samples between 15 and 25 flowers were pooled, as well as 10-20 siliques were pooled per individual. All tissues were frozen immediately in liquid nitrogen after collection.

3-week and 5-week leaf data set (4 leaves)

Seeds of a single individual with a Col-0 background were stratified in soil for 3 days at 4^{o} in short day cycles (8 hours light, 16 hours dark). Plants were transferred to 23 C with long day conditions (16 hours light, 8 hours dark). After 3 weeks one individual was dissected, leaves 1,3,5 and 7 were collected. After a total of 5 weeks a second individual was dissected and leaves 1,3,5 and 7 were collected. All tissues were frozen immediately in liquid nitrogen after collection.

5-week replicate (7 leaves)

Seeds of a single individual with a Col-0 background were stratified in agar in a 4^{o} fridge for 3 days. Plants were transferred to 23 C with long day conditions (16 hours light, 8 hours dark). After 5 weeks, one individual was dissected and leaves 1,3,5,7,9,11 and 13 were collected separately. Samples were frozen in liquid nitrogen immediately after collection.

Bisulfite and RNA sequencing

Frozen plant material was ground with metal beads using a Qiagen Tissuelyzer II. Genomic DNA was extracted using the Qiagen Plant DNeasy kit (catalog # 69104) following the manufacturer's instructions. Bisulfite converted DNA libraries were prepared as described in (Becker, Hagmann et al. 2011) with a 300 bp insert size using the Qiagen Epitect Plus Kit (catalog # 59124). Total RNA was isolated using the RNAeasy Plant Mini Kit following manufacturer's instructions (catalog # 74904). Libraries for RNA-seq were prepared from 1 µg RNA using the TruSeq RNA sample prep kit (Illumina catalog # RS-122-2001) according to manufacturer's protocol. Libraries were sequenced on a HiSeq2000 instrument (Illumina) with 2x101 bp paired-end reads for bisulfite-treated DNA, and 101 bp single-end reads for RNA-seq. For image analysis and base calling, Illumina OLB software was used.

Read processing and alignment

For BS-sequencing, the SHORE pipeline v0.9.0 (Ossowski, Schneeberger et al. 2008) was used to trim and quality filter reads. Reads that had more than one base in the first 12 positions with a quality score below 4 were removed. Reads with 10% or more ambiguous bases were discarded. Reads were aligned using Shore against the TAIR9 (http://www.arabidopsis.org) version of the *A. thaliana* genome using a seed length of 40. Reads were required to have fewer than three mismatches within the seed. Paired-end reads were required to have a distance of 10-100 bp. Paired end read correction was performed after alignment using SHORE to discard reads that showed abnormal insert sizes (2 standard deviations from the average insert size). Finally, using SHORE's scoring matrix approach (Becker, Hagmann et al. 2011), unique read counts were retrieved. Command line arguments are provided in supplementary file commands.txt. RNA-seq reads were trimmed using the SHORE pipeline (Ossowski, Schneeberger et al. 2008).

Identification of methylated and differentially methylated positions

In all cases, MRs and DMRs were called using the pipeline described in sections 2.6.2 and 3.1 (Becker, Hagmann et al. 2011). A minimum coverage of 5 was required for each site. Quality scores were used as an additional filter, where sites with a quality score lower than 16 were removed from the analysis. Additionally all cytosine analyzed required at least one sample to have a quality score of 32. Storey's method was used in all cases for multiple testing correction.

Expression analysis

For transcriptome analysis, the cufflinks package was used. Bowtie2 (Langmead and Salzberg 2012) was used to map the RNA reads to the gene annotation of TAIR10 (Swarbreck, Wilks et al. 2008) allowing for up to 10% mismatches and 7% gaps. I used Cufflinks (Trapnell, Roberts et al. 2012) with default values to calculate expression values from pooled samples of the same organ type (Supplemental Data 1).

7 References

Ahmed, I., A. Sarazin, C. Bowler, V. Colot and H. Quesneville (2011). "Genome-wide evidence for local DNA methylation spreading from small RNA-targeted sequences in Arabidopsis." Nucleic Acids Res **39**(16): 6919-6931.

Amir, R. E., I. B. Van den Veyver, M. Wan, C. Q. Tran, U. Francke and H. Y. Zoghbi (1999). "Rett syndrome is caused by mutations in X-linked MECP2, encoding methyl-CpG-binding protein 2." Nat Genet **23**(2): 185-188.

Andrews, S. "FastQC: a quality control tool for high throughput sequence data." http://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

Aravin, A. A., G. J. Hannon and J. Brennecke (2007). "The Piwi-piRNA pathway provides an adaptive defense in the transposon arms race." Science **318**(5851): 761-764.

Aravin, A. A., R. Sachidanandam, D. Bourc'his, C. Schaefer, D. Pezic, K. F. Toth, T. Bestor and G. J. Hannon (2008). "A piRNA pathway primed by individual transposons is linked to de novo DNA methylation in mice." Mol Cell **31**(6): 785-799.

Arnaud, P., C. Goubely, T. Pelissier and J. M. Deragon (2000). "SINE retroposons can be used in vivo as nucleation centers for de novo methylation." Mol Cell Biol **20**(10): 3434-3441.

Avery, O. T., C. M. Macleod and M. McCarty (1944). "Studies on the Chemical Nature of the Substance Inducing Transformation of Pneumococcal Types : Induction of Transformation by a Desoxyribonucleic Acid Fraction Isolated from Pneumococcus Type Iii." J Exp Med **79**(2): 137-158.

Barrow, T. M. and K. B. Michels (2014). "Epigenetic epidemiology of cancer." Biochem Biophys Res Commun **455**(1-2): 70-83.

Bateson, W. and C. Pellew (1915). "On the genetics of “Rogues” among culinary peas (Pisum sativum)." Journal of Genetics **5**(1): 13-36.

Becker, C., J. Hagmann, J. Müller, D. Koenig, O. Stegle, K. Borgwardt and D. Weigel (2011). "Spontaneous epigenetic variation in the *Arabidopsis thaliana* methylome." Nature **480**(7376): 245-249.

Becker, C. and D. Weigel (2012). "Epigenetic variation: origin and transgenerational inheritance." Curr. Opin. Plant Biol. **15**(5): 562-567.

Beemster, G. T., L. De Veylder, S. Vercruysse, G. West, D. Rombaut, P. Van Hummelen, A. Galichet, W. Gruissem, D. Inze and M. Vuylsteke (2005). "Genome-wide analysis of gene expression profiles associated with cell cycle transitions in growing organs of Arabidopsis." Plant Physiol **138**(2): 734-743.

Benjamini, Y. and Y. Hochberg (1995). "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing." Journal of the Royal Statistical Society. Series B (Methodological) **57**(1): 289-300.

Berardini, T. Z., L. Reiser, D. Li, Y. Mezheritsky, R. Muller, E. Strait and E. Huala (2015). "The Arabidopsis information resource: Making and mining the "gold standard" annotated reference plant genome." Genesis **53**(8): 474-485.

Bernatavichute, Y. V., X. Zhang, S. Cokus, M. Pellegrini and S. E. Jacobsen (2008). "Genome-wide association of histone H3 lysine nine methylation with CHG DNA methylation in Arabidopsis thaliana." PLoS One **3**(9): e3156.

Bock, C. (2012). "Analysing and interpreting DNA methylation data." Nat Rev Genet **13**(10): 705-719.

Bostick, M., J. K. Kim, P. O. Esteve, A. Clark, S. Pradhan and S. E. Jacobsen (2007). "UHRF1 plays a role in maintaining DNA methylation in mammalian cells." Science **317**(5845): 1760-1764.

Bourc'his, D. and T. H. Bestor (2004). "Meiotic catastrophe and retrotransposon reactivation in male germ cells lacking Dnmt3L." Nature **431**(7004): 96-99.

Boveri, T. (1904). Ergebnisse über die Konstitution der chromatischen Substanz des Zellkerns. Von Theodor Boveri. Jena, G. Fischer.

Brink, R. A. (1956). "A Genetic Change Associated with the R Locus in Maize Which Is Directed and Potentially Reversible." Genetics **41**(6): 872-889.

Buisine, N., H. Quesneville and V. Colot (2008). "Improved detection and annotation of transposable elements in sequenced genomes using multiple reference sequence sets." Genomics **91**(5): 467-475.

Cantone, I. and A. G. Fisher (2013). "Epigenetic programming and reprogramming during development." Nat Struct Mol Biol **20**(3): 282-289.

Carlson, L. L., A. W. Page and T. H. Bestor (1992). "Properties and localization of DNA methyltransferase in preimplantation mouse embryos: implications for genomic imprinting." Genes Dev **6**(12B): 2536-2541.

Catchen, J., P. A. Hohenlohe, S. Bassham, A. Amores and W. A. Cresko (2013). "Stacks: an analysis tool set for population genomics." Mol Ecol **22**(11): 3124-3140.

Cedar, H. and Y. Bergman (2009). "Linking DNA methylation and histone modification: patterns and paradigms." Nat Rev Genet **10**(5): 295-304.

Cervera, M. T., L. Ruiz-Garcia and J. M. Martinez-Zapater (2002). "Analysis of DNA methylation in Arabidopsis thaliana based on methylation-sensitive AFLP markers." Mol Genet Genomics **268**(4): 543-552.

Chan, S. W., I. R. Henderson and S. E. Jacobsen (2005). "Gardening the genome: DNA methylation in Arabidopsis thaliana." Nat Rev Genet **6**(5): 351-360.

Choi, Y., M. Gehring, L. Johnson, M. Hannon, J. J. Harada, R. B. Goldberg, S. E. Jacobsen and R. L. Fischer (2002). "DEMETER, a DNA glycosylase domain protein, is required for endosperm gene imprinting and seed viability in arabidopsis." Cell **110**(1): 33-42.

Cock, P. J., C. J. Fields, N. Goto, M. L. Heuer and P. M. Rice (2010). "The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants." Nucleic Acids Res **38**(6): 1767-1771.

Cokus, S. J., S. Feng, X. Zhang, Z. Chen, B. Merriman, C. D. Haudenschild, S. Pradhan, S. F. Nelson, M. Pellegrini and S. E. Jacobsen (2008). "Shotgun bisulphite sequencing of the *Arabidopsis* genome reveals DNA methylation patterning." Nature **452**(7184): 215-219.

Coleman-Derr, D. and D. Zilberman (2012). "Deposition of histone variant H2A.Z within gene bodies regulates responsive genes." PLoS Genet **8**(10): e1002988.

Crick, F. H., L. Barnett, S. Brenner and R. J. Watts-Tobin (1961). "General nature of the genetic code for proteins." Nature **192**: 1227-1232.

Deal, R. B. and S. Henikoff (2011). "The INTACT method for cell type-specific gene expression and chromatin profiling in Arabidopsis thaliana." Nat Protoc **6**(1): 56-68.

Dennis, K., T. Fan, T. Geiman, Q. Yan and K. Muegge (2001). "Lsh, a member of the SNF2 family, is required for genome-wide methylation." Genes Dev **15**(22): 2940-2944.

Dowen, R. H., M. Pelizzola, R. J. Schmitz, R. Lister, J. M. Dowen, J. R. Nery, J. E. Dixon and J. R. Ecker (2012). "Widespread dynamic DNA methylation in response to biotic stress." Proc Natl Acad Sci U S A **109**(32): E2183-2191.

Du, J., L. M. Johnson, M. Groth, S. Feng, C. J. Hale, S. Li, A. A. Vashisht, J. Gallego-Bartolome, J. A. Wohlschlegel, D. J. Patel and S. E. Jacobsen (2014). "Mechanism of DNA methylation-directed histone methylation by KRYPTONITE." Mol Cell **55**(3): 495-504.

Du, J., X. Zhong, Y. V. Bernatavichute, H. Stroud, S. Feng, E. Caro, A. A. Vashisht, J. Terragni, H. G. Chin, A. Tu, J. Hetzel, J. A. Wohlschlegel, S. Pradhan, D. J. Patel and S. E. Jacobsen (2012). "Dual binding of chromomethylase domains to H3K9me2-containing nucleosomes directs DNA methylation in plants." Cell **151**(1): 167-180.

Dunn, O. J. (1959). "Estimation of the Medians for Dependent-Variables." Annals of Mathematical Statistics **30**(1): 192-197.

Ecker, J. R. (2013). "Epigenetic trigger for tomato ripening." <u>Nat Biotechnol</u> **31**(2): 119-120.

Edgar, B. A., N. Zielke and C. Gutierrez (2014). "Endocycles: a recurrent evolutionary innovation for post-mitotic cell growth." <u>Nat Rev Mol Cell Biol</u> **15**(3): 197-210.

Ehrlich, M., M. A. Gama-Sosa, L. H. Huang, R. M. Midgett, K. C. Kuo, R. A. McCune and C. Gehrke (1982). "Amount and distribution of 5-methylcytosine in human DNA from different types of tissues of cells." <u>Nucleic Acids Res</u> **10**(8): 2709-2721.

Erhard, K. F., Jr., J. L. Stonaker, S. E. Parkinson, J. P. Lim, C. J. Hale and J. B. Hollick (2009). "RNA polymerase IV functions in paramutation in Zea mays." <u>Science</u> **323**(5918): 1201-1205.

Exposito-Alonso, M., C. Becker, V. J. Schuenemann, E. Reitter, C. Setzer, R. Slovak, B. Brachi, J. Hagmann, D. G. Grimm, C. Jiahui, W. Busch, J. Bergelson, R. W. Ness, J. Krause, H. A. Burbano and D. Weigel (2016). "The rate and effect of de novo mutations in natural populations of Arabidopsis thaliana." <u>bioRxiv</u>.

Farthing, C. R., G. Ficz, R. K. Ng, C. F. Chan, S. Andrews, W. Dean, M. Hemberger and W. Reik (2008). "Global mapping of DNA methylation in mouse promoters reveals epigenetic reprogramming of pluripotency genes." <u>PLoS Genet</u> **4**(6): e1000116.

Feng, S., S. J. Cokus, X. Zhang, P. Y. Chen, M. Bostick, M. G. Goll, J. Hetzel, J. Jain, S. H. Strauss, M. E. Halpern, C. Ukomadu, K. C. Sadler, S. Pradhan, M. Pellegrini and S. E. Jacobsen (2010). "Conservation and divergence of methylation patterning in plants and animals." <u>Proc Natl Acad Sci U S A</u> **107**(19): 8689-8694.

Feng, S., S. E. Jacobsen and W. Reik (2010). "Epigenetic reprogramming in plant and animal development." <u>Science</u> **330**(6004): 622-627.

Fournier, C., Y. Goto, E. Ballestar, K. Delaval, A. M. Hever, M. Esteller and R. Feil (2002). "Allele-specific histone lysine methylation marks regulatory regions at imprinted mouse genes." <u>EMBO J</u> **21**(23): 6560-6570.

Franklin, R. E. and R. G. Gosling (1953). "Evidence for 2-chain helix in crystalline structure of sodium deoxyribonucleate." <u>Nature</u> **172**(4369): 156-157.

Frommer, M., L. E. McDonald, D. S. Millar, C. M. Collis, F. Watt, G. W. Grigg, P. L. Molloy and C. L. Paul (1992). "A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands." <u>Proc Natl Acad Sci U S A</u> **89**(5): 1827-1831.

Galbraith, D. W., K. R. Harkins and S. Knapp (1991). "Systemic Endopolyploidy in Arabidopsis thaliana." <u>Plant Physiol</u> **96**(3): 985-989.

Gendreau, E., V. Orbovic, H. H and J. Traas (1999). "Gibberellin and ethylene control endoreduplication levels in the *Arabidopsis thaliana* hypocotyl." Planta **209**(4): 513-516.

Gong, Z., T. Morales-Ruiz, R. R. Ariza, T. Roldan-Arjona, L. David and J. K. Zhu (2002). "ROS1, a repressor of transcriptional gene silencing in Arabidopsis, encodes a DNA glycosylase/lyase." Cell **111**(6): 803-814.

Griffith, F. (1928). "The Significance of Pneumococcal Types." J Hyg (Lond) **27**(2): 113-159.

Guillemette, B. and L. Gaudreau (2006). "Reuniting the contrasting functions of H2A.Z." Biochem Cell Biol **84**(4): 528-535.

Gutzat, R. and O. Mittelsten Scheid (2012). "Epigenetic responses to stress: triple defense?" Curr Opin Plant Biol **15**(5): 568-573.

Hagmann, J., C. Becker, J. Muller, O. Stegle, R. C. Meyer, G. Wang, K. Schneeberger, J. Fitz, T. Altmann, J. Bergelson, K. Borgwardt and D. Weigel (2015). "Century-scale methylome stability in a recently diverged Arabidopsis thaliana lineage." PLoS Genet **11**(1): e1004920.

Hansen, R. S., C. Wijmenga, P. Luo, A. M. Stanek, T. K. Canfield, C. M. Weemaes and S. M. Gartler (1999). "The DNMT3B DNA methyltransferase gene is mutated in the ICF immunodeficiency syndrome." Proc Natl Acad Sci U S A **96**(25): 14412-14417.

Hayatsu, H., Y. Wataya and K. Kazushige (1970). "The addition of sodium bisulfite to uracil and to cytosine." J Am Chem Soc **92**(3): 724-726.

He, X. J., T. Chen and J. K. Zhu (2011). "Regulation and function of DNA methylation in plants and animals." Cell Res **21**(3): 442-465.

He, Y. F., B. Z. Li, Z. Li, P. Liu, Y. Wang, Q. Tang, J. Ding, Y. Jia, Z. Chen, L. Li, Y. Sun, X. Li, Q. Dai, C. X. Song, K. Zhang, C. He and G. L. Xu (2011). "Tet-mediated formation of 5-carboxylcytosine and its excision by TDG in mammalian DNA." Science **333**(6047): 1303-1307.

Herr, A. J., M. B. Jensen, T. Dalmay and D. C. Baulcombe (2005). "RNA polymerase IV directs silencing of endogenous DNA." Science **308**(5718): 118-120.

Hershey, A. D. and M. Chase (1952). "Independent functions of viral protein and nucleic acid in growth of bacteriophage." J Gen Physiol **36**(1): 39-56.

Herten, K., M. S. Hestand, J. R. Vermeesch and J. K. Van Houdt (2015). "GBSX: a toolkit for experimental design and demultiplexing genotyping by sequencing experiments." BMC Bioinformatics **16**: 73.

Hirochika, H., H. Okamoto and T. Kakutani (2000). "Silencing of retrotransposons in arabidopsis and reactivation by the ddm1 mutation." Plant Cell **12**(3): 357-369.

Holliday, R. and J. E. Pugh (1975). "DNA modification mechanisms and gene activity during development." Science **187**(4173): 226-232.

Hollister, J. D., L. M. Smith, Y. L. Guo, F. Ott, D. Weigel and B. S. Gaut (2011). "Transposable elements and small RNAs contribute to gene expression divergence between Arabidopsis thaliana and Arabidopsis lyrata." Proc Natl Acad Sci U S A **108**(6): 2322-2327.

Hollister, J. D., L. M. Smith, Y. L. Guo, F. Ott, D. Weigel and B. S. Gaut (2011). "Transposable elements and small RNAs contribute to gene expression divergence between *Arabidopsis thaliana* and *Arabidopsis lyrata*." Proc. Natl. Acad. Sci. USA. **108**(6): 2322-2327.

Hotchkiss, R. D. (1948). "The quantitative separation of purines, pyrimidines, and nucleosides by paper chromatography." J Biol Chem **175**(1): 315-332.

Huang, Y., W. A. Pastor, Y. Shen, M. Tahiliani, D. R. Liu and A. Rao (2010). "The behaviour of 5-hydroxymethylcytosine in bisulfite sequencing." PLoS One **5**(1): e8888.

Huh, J. H., M. J. Bauer, T. F. Hsieh and R. L. Fischer (2008). "Cellular programming of plant gene imprinting." Cell **132**(5): 735-744.

Ito, S., L. Shen, Q. Dai, S. C. Wu, L. B. Collins, J. A. Swenberg, C. He and Y. Zhang (2011). "Tet proteins can convert 5-methylcytosine to 5-formylcytosine and 5-carboxylcytosine." Science **333**(6047): 1300-1303.

Iwasaki, M. and J. Paszkowski (2014). "Epigenetic memory in plants." EMBO J **33**(18): 1987-1998.

Jackson, J. P., A. M. Lindroth, X. Cao and S. E. Jacobsen (2002). "Control of CpNpG DNA methylation by the KRYPTONITE histone H3 methyltransferase." Nature **416**(6880): 556-560.

Jackson, M., A. Krassowska, N. Gilbert, T. Chevassut, L. Forrester, J. Ansell and B. Ramsahoye (2004). "Severe global DNA hypomethylation blocks differentiation and induces histone hyperacetylation in embryonic stem cells." Mol Cell Biol **24**(20): 8862-8871.

Jacob, F., D. Perrin, C. Sanchez and J. Monod (1960). "[Operon: a group of genes with the expression coordinated by an operator]." C R Hebd Seances Acad Sci **250**: 1727-1729.

Johnson, L. M., M. Bostick, X. Zhang, E. Kraft, I. Henderson, J. Callis and S. E. Jacobsen (2007). "The SRA methyl-cytosine-binding domain links DNA and histone methylation." Curr Biol **17**(4): 379-384.

Jordan, I. K., I. B. Rogozin, G. V. Glazko and E. V. Koonin (2003). "Origin of a substantial fraction of human regulatory sequences from transposable elements." Trends Genet **19**(2): 68-72.

Kafri, T., M. Ariel, M. Brandeis, R. Shemer, L. Urven, J. McCarrey, H. Cedar and A. Razin (1992). "Developmental pattern of gene-specific DNA methylation in the mouse embryo and germ line." Genes Dev **6**(5): 705-714.

Kawakatsu, T., T. Stuart, M. Valdes, N. Breakfield, R. J. Schmitz, J. R. Nery, M. A. Urich, X. Han, R. Lister, P. N. Benfey and J. R. Ecker (2016). "Unique cell-type-specific patterns of DNA methylation in the root meristem." Nat Plants **2**(5): 16058.

Kawashima, T. and F. Berger (2014). "Epigenetic reprogramming in plant sexual reproduction." Nat Rev Genet **15**(9): 613-624.

Kinoshita, T., A. Miura, Y. Choi, Y. Kinoshita, X. Cao, S. E. Jacobsen, R. L. Fischer and T. Kakutani (2004). "One-way control of FWA imprinting in Arabidopsis endosperm by DNA methylation." Science **303**(5657): 521-523.

Kornberg, R. D. (1974). "Chromatin structure: a repeating unit of histones and DNA." Science **184**(4139): 868-871.

Krueger, F. and S. R. Andrews (2011). "Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications." Bioinformatics **27**(11): 1571-1572.

Kuramochi-Miyagawa, S., T. Kimura, T. W. Ijiri, T. Isobe, N. Asada, Y. Fujita, M. Ikawa, N. Iwai, M. Okabe, W. Deng, H. Lin, Y. Matsuda and T. Nakano (2004). "Mili, a mammalian member of piwi family gene, is essential for spermatogenesis." Development **131**(4): 839-849.

Kuramochi-Miyagawa, S., T. Watanabe, K. Gotoh, Y. Totoki, A. Toyoda, M. Ikawa, N. Asada, K. Kojima, Y. Yamaguchi, T. W. Ijiri, K. Hata, E. Li, Y. Matsuda, T. Kimura, M. Okabe, Y. Sakaki, H. Sasaki and T. Nakano (2008). "DNA methylation of retrotransposon genes is regulated by Piwi family members MILI and MIWI2 in murine fetal testes." Genes Dev **22**(7): 908-917.

Langmead, B. and S. L. Salzberg (2012). "Fast gapped-read alignment with Bowtie 2." Nat Methods **9**(4): 357-359.

Laurent, L., E. Wong, G. Li, T. Huynh, A. Tsirigos, C. T. Ong, H. M. Low, K. W. Kin Sung, I. Rigoutsos, J. Loring and C. L. Wei (2010). "Dynamic changes in the human methylome during differentiation." Genome Res **20**(3): 320-331.

Law, J. A. and S. E. Jacobsen (2010). "Establishing, maintaining and modifying DNA methylation patterns in plants and animals." Nat Rev Genet **11**(3): 204-220.

Law, J. A. and S. E. Jacobsen (2010). "Establishing, maintaining and modifying DNA methylation patterns in plants and animals." Nat. Rev. Genet. **11**(3): 204-220.

Lee, H. O., J. M. Davidson and R. J. Duronio (2009). "Endoreplication: polyploidy with purpose." Genes Dev **23**(21): 2461-2477.

Lindroth, A. M., X. Cao, J. P. Jackson, D. Zilberman, C. M. McCallum, S. Henikoff and S. E. Jacobsen (2001). "Requirement of CHROMOMETHYLASE3 for maintenance of CpXpG methylation." Science **292**(5524): 2077-2080.

Lister, R., R. C. O'Malley, J. Tonti-Filippini, B. D. Gregory, C. C. Berry, A. H. Millar and J. R. Ecker (2008). "Highly integrated single-base resolution maps of the epigenome in Arabidopsis." Cell **133**(3): 523-536.

Lister, R., M. Pelizzola, R. H. Dowen, R. D. Hawkins, G. Hon, J. Tonti-Filippini, J. R. Nery, L. Lee, Z. Ye, Q. M. Ngo, L. Edsall, J. Antosiewicz-Bourget, R. Stewart, V. Ruotti, A. H. Millar, J. A. Thomson, B. Ren and J. R. Ecker (2009). "Human DNA methylomes at base resolution show widespread epigenomic differences." Nature **462**(7271): 315-322.

Liu, J. J. and R. L. Ward (2010). "Folate and one-carbon metabolism and its impact on aberrant DNA methylation in cancer." Adv Genet **71**: 79-121.

Llinares-Lopez, F., D. G. Grimm, D. A. Bodenham, U. Gieraths, M. Sugiyama, B. Rowan and K. Borgwardt (2015). "Genome-wide detection of intervals of genetic heterogeneity associated with complex traits." Bioinformatics **31**(12): i240-249.

Lu, Y., T. Rong and M. Cao (2008). "Analysis of DNA methylation in different maize tissues." J Genet Genomics **35**(1): 41-48.

Luo, R., B. Liu, Y. Xie, Z. Li, W. Huang, J. Yuan, G. He, Y. Chen, Q. Pan, Y. Liu, J. Tang, G. Wu, H. Zhang, Y. Shi, Y. Liu, C. Yu, B. Wang, Y. Lu, C. Han, D. W. Cheung, S.-M. Yiu, S. Peng, Z. Xiaoqian, G. Liu, X. Liao, Y. Li, H. Yang, J. Wang, T.-W. Lam and J. Wang (2012). "SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler." GigaScience **1**(1): 18.

Manning, K., M. Tor, M. Poole, Y. Hong, A. J. Thompson, G. J. King, J. J. Giovannoni and G. B. Seymour (2006). "A naturally occurring epigenetic mutation in a gene encoding an SBP-box transcription factor inhibits tomato fruit ripening." Nat Genet **38**(8): 948-952.

Matzke, M. A. and R. A. Mosher (2014). "RNA-directed DNA methylation: an epigenetic pathway of increasing complexity." Nat Rev Genet **15**(6): 394-408.

Mayer, W., A. Niveleau, J. Walter, R. Fundele and T. Haaf (2000). "Demethylation of the zygotic paternal genome." Nature **403**(6769): 501-502.

McCarty, M. and O. T. Avery (1946). "Studies on the Chemical Nature of the Substance Inducing Transformation of Pneumococcal Types : Ii. Effect of Desoxyribonuclease on the Biological Activity of the Transforming Substance." J Exp Med **83**(2): 89-96.

Melaragno, J. E., B. Mehrotra and A. W. Coleman (1993). "Relationship between Endopolyploidy and Cell Size in Epidermal Tissue of Arabidopsis." Plant Cell **5**(11): 1661-1668.

Mendel, G. (1866). Versuche über Pflanzen-Hybriden. Brünn :, Im Verlage des Vereines.

Messerschmidt, D. M., B. B. Knowles and D. Solter (2014). "DNA methylation dynamics during epigenetic reprogramming in the germline and preimplantation embryos." Genes Dev **28**(8): 812-828.

Miura, A., S. Yonebayashi, K. Watanabe, T. Toyama, H. Shimada and T. Kakutani (2001). "Mobilization of transposons by a mutation abolishing full DNA methylation in Arabidopsis." Nature **411**(6834): 212-214.

Molaro, A., E. Hodges, F. Fang, Q. Song, W. R. McCombie, G. J. Hannon and A. D. Smith (2011). "Sperm methylation profiles reveal features of epigenetic inheritance and evolution in primates." Cell **146**(6): 1029-1041.

Molnar, A., C. W. Melnyk, A. Bassett, T. J. Hardcastle, R. Dunn and D. C. Baulcombe (2010). "Small silencing RNAs in plants are mobile and direct epigenetic modification in recipient cells." Science **328**(5980): 872-875.

Morgan, H. D., H. G. Sutherland, D. I. Martin and E. Whitelaw (1999). "Epigenetic inheritance at the agouti locus in the mouse." Nat Genet **23**(3): 314-318.

Morgan, T. H. (1910). "Sex Limited Inheritance in Drosophila." Science **32**(812): 120-122.

Muller, H. J. (1930). "Types of visible variations induced by X-rays inDrosophila." Journal of Genetics **22**(3): 299-334.

Murray, J. A., A. Jones, C. Godin and J. Traas (2012). "Systems analysis of shoot apical meristem growth and development: integrating hormonal and mechanical signaling." Plant Cell **24**(10): 3907-3919.

Needleman, S. B. and C. D. Wunsch (1970). "A general method applicable to the search for similarities in the amino acid sequence of two proteins." J Mol Biol **48**(3): 443-453.

Ng, D. W., J. Lu and Z. J. Chen (2012). "Big roles for small RNAs in polyploidy, hybrid vigor, and hybrid incompatibility." Curr Opin Plant Biol **15**(2): 154-161.

Ong-Abdullah, M., J. M. Ordway, N. Jiang, S. E. Ooi, S. Y. Kok, N. Sarpan, N. Azimi, A. T. Hashim, Z. Ishak, S. K. Rosli, F. A. Malike, N. A. Bakar, M. Marjuni, N. Abdullah, Z. Yaakub, M. D. Amiruddin, R. Nookiah, R. Singh, E. T. Low, K. L. Chan, N. Azizi, S. W. Smith, B. Bacher, M. A. Budiman, A. Van Brunt, C. Wischmeyer, M. Beil, M. Hogan, N. Lakey, C. C. Lim, X. Arulandoo, C. K. Wong, C. N. Choo, W. C. Wong, Y. Y. Kwan, S. S. Alwee, R. Sambanthamurthi and R. A. Martienssen (2015). "Loss of Karma transposon methylation underlies the mantled somaclonal variant of oil palm." Nature **525**(7570): 533-537.

Ooi, S. K., C. Qiu, E. Bernstein, K. Li, D. Jia, Z. Yang, H. Erdjument-Bromage, P. Tempst, S. P. Lin, C. D. Allis, X. Cheng and T. H. Bestor (2007). "DNMT3L connects unmethylated lysine 4 of histone H3 to de novo methylation of DNA." Nature **448**(7154): 714-717.

Ortega-Galisteo, A. P., T. Morales-Ruiz, R. R. Ariza and T. Roldan-Arjona (2008). "Arabidopsis DEMETER-LIKE proteins DML2 and DML3 are required for appropriate distribution of DNA methylation marks." Plant Mol Biol **67**(6): 671-681.

Ossowski, S., K. Schneeberger, R. M. Clark, C. Lanz, N. Warthmann and D. Weigel (2008). "Sequencing of natural strains of *Arabidopsis thaliana* with short reads." Genome Res. **18**: 2024-2033.

Ossowski, S., K. Schneeberger, J. I. Lucas-Lledo, N. Warthmann, R. M. Clark, R. G. Shaw, D. Weigel and M. Lynch (2010). "The rate and molecular spectrum of spontaneous mutations in Arabidopsis thaliana." Science **327**(5961): 92-94.

Parkinson, S. E., S. M. Gross and J. B. Hollick (2007). "Maize sex determination and abaxial leaf fates are canalized by a factor that maintains repressed epigenetic states." Dev Biol **308**(2): 462-473.

Pecinka, A. and O. Mittelsten Scheid (2012). "Stress-induced chromatin changes: a critical view on their heritability." Plant Cell Physiol **53**(5): 801-808.

Penterman, J., D. Zilberman, J. H. Huh, T. Ballinger, S. Henikoff and R. L. Fischer (2007). "DNA demethylation in the Arabidopsis genome." Proc Natl Acad Sci U S A **104**(16): 6752-6757.

Pontier, D., G. Yahubyan, D. Vega, A. Bulski, J. Saez-Vasquez, M. A. Hakimi, S. Lerbs-Mache, V. Colot and T. Lagrange (2005). "Reinforcement of silencing at transposons and highly repeated sequences requires the concerted action of two distinct RNA polymerases IV in Arabidopsis." Genes Dev **19**(17): 2030-2040.

Popp, C., W. Dean, S. Feng, S. J. Cokus, S. Andrews, M. Pellegrini, S. E. Jacobsen and W. Reik (2010). "Genome-wide erasure of DNA methylation in mouse primordial germ cells is affected by AID deficiency." Nature **463**(7284): 1101-1105.

Regulski, M., Z. Lu, J. Kendall, M. T. Donoghue, J. Reinders, V. Llaca, S. Deschamps, A. Smith, D. Levy, W. R. McCombie, S. Tingey, A. Rafalski, J. Hicks, D. Ware and R. A. Martienssen (2013). "The maize methylome influences mRNA splice sites and reveals widespread paramutation-like switches guided by small RNA." Genome Res **23**(10): 1651-1662.

Reik, W., W. Dean and J. Walter (2001). "Epigenetic reprogramming in mammalian development." Science **293**(5532): 1089-1093.

Reik, W. and J. Walter (2001). "Genomic imprinting: parental influence on the genome." Nat Rev Genet **2**(1): 21-32.

Rhoads, A. and K. F. Au (2015). "PacBio Sequencing and Its Applications." Genomics Proteomics Bioinformatics **13**(5): 278-289.

Sanger, F., S. Nicklen and A. R. Coulson (1977). "DNA sequencing with chain-terminating inhibitors." Proc Natl Acad Sci U S A **74**(12): 5463-5467.

Schermelleh, L., A. Haemmer, F. Spada, N. Rosing, D. Meilinger, U. Rothbauer, M. C. Cardoso and H. Leonhardt (2007). "Dynamics of Dnmt1 interaction with the replication machinery and its role in postreplicative maintenance of DNA methylation." Nucleic Acids Res **35**(13): 4301-4312.

Schmid, M., T. S. Davison, S. R. Henz, U. J. Pape, M. Demar, M. Vingron, B. Scholkopf, D. Weigel and J. U. Lohmann (2005). "A gene expression map of Arabidopsis thaliana development." Nat Genet **37**(5): 501-506.

Schmid, M., T. S. Davison, S. R. Henz, U. J. Pape, M. Demar, M. Vingron, B. Schölkopf, D. Weigel and J. U. Lohmann (2005). "A gene expression map of *Arabidopsis thaliana* development." Nat. Genet. **37**(5): 501-506.

Schmitz, R. J., M. D. Schultz, M. G. Lewsey, R. C. O'Malley, M. A. Urich, O. Libiger, N. J. Schork and J. R. Ecker (2011). "Transgenerational epigenetic instability is a source of novel methylation variants." Science **334**(6054): 369-373.

Schneeberger, K., J. Hagmann, S. Ossowski, N. Warthmann, S. Gesing, O. Kohlbacher and D. Weigel (2009). "Simultaneous alignment of short reads against multiple genomes." Genome Biol. **10**(9): R98.

Schneeberger, K., S. Ossowski, C. Lanz, T. Juul, A. H. Petersen, K. L. Nielsen, J. E. Jørgensen, D. Weigel and S. U. Andersen (2009). "SHOREmap: simultaneous mapping and mutation identification by deep sequencing." Nat. Methods **6**(8): 550-551.

Schultz, M. D., Y. He, J. W. Whitaker, M. Hariharan, E. A. Mukamel, D. Leung, N. Rajagopal, J. R. Nery, M. A. Urich, H. Chen, S. Lin, Y. Lin, I. Jung, A. D. Schmitt, S. Selvaraj, B. Ren, T. J. Sejnowski, W. Wang and J. R. Ecker (2015). "Human body epigenome maps reveal noncanonical DNA methylation variation." Nature **523**(7559): 212-216.

Secco, D., C. Wang, H. Shou, M. D. Schultz, S. Chiarenza, L. Nussaume, J. R. Ecker, J. Whelan and R. Lister (2015). "Stress induced gene expression drives transient DNA methylation changes at adjacent repetitive elements." Elife **4**.

Seymour, D. K., D. Koenig, J. Hagmann, C. Becker and D. Weigel (2014). "Evolution of DNA methylation patterns in the Brassicaceae is driven by differences in genome organization." PLoS Genet **10**(11): e1004785.

Shapiro, R., R. E. Servis and M. Welcher (1970). "Reactions of Uracil and Cytosine Derivatives with Sodium Bisulfite." Journal of the American Chemical Society **92**(2): 422-424.

Shaw, R. G., D. L. Byers and E. Darmo (2000). "Spontaneous mutational effects on reproductive traits of *Arabidopsis thaliana*." Genetics **155**(1): 369-378.

Sidorenko, L. and V. Chandler (2008). "RNA-dependent RNA polymerase is required for enhancer-mediated transcriptional silencing associated with paramutation at the maize p1 gene." Genetics **180**(4): 1983-1993.

Simpson, J. T., R. Workman, P. C. Zuzarte, M. David, L. J. Dursi and W. Timp (2016). "Detecting DNA Methylation using the Oxford Nanopore Technologies MinION sequencer." bioRxiv.

Singer, T., C. Yordan and R. A. Martienssen (2001). "Robertson's Mutator transposons in A. thaliana are regulated by the chromatin-remodeling gene Decrease in DNA Methylation (DDM1)." Genes Dev **15**(5): 591-602.

Slotkin, R. K. and R. Martienssen (2007). "Transposable elements and the epigenetic regulation of the genome." Nat Rev Genet **8**(4): 272-285.

Slotkin, R. K., M. Vaughn, F. Borges, M. Tanurdzic, J. D. Becker, J. A. Feijo and R. A. Martienssen (2009). "Epigenetic reprogramming and small RNA silencing of transposable elements in pollen." Cell **136**(3): 461-472.

Smallwood, S. A. and G. Kelsey (2012). "De novo DNA methylation: a germ cell perspective." Trends Genet **28**(1): 33-42.

Smallwood, S. A., H. J. Lee, C. Angermueller, F. Krueger, H. Saadeh, J. Peat, S. R. Andrews, O. Stegle, W. Reik and G. Kelsey (2014). "Single-cell genome-wide bisulfite sequencing for assessing epigenetic heterogeneity." Nat Methods **11**(8): 817-820.

Smith, L. M., J. Z. Sanders, R. J. Kaiser, P. Hughes, C. Dodd, C. R. Connell, C. Heiner, S. B. Kent and L. E. Hood (1986). "Fluorescence detection in automated DNA sequence analysis." Nature **321**(6071): 674-679.

Smith, T. F. and M. S. Waterman (1981). "Identification of common molecular subsequences." J Mol Biol **147**(1): 195-197.

Smith, Z. D. and A. Meissner (2013). "DNA methylation: roles in mammalian development." Nat Rev Genet **14**(3): 204-220.

Smolle, M. and J. L. Workman (2013). "Transcription-associated histone modifications and cryptic transcription." Biochim Biophys Acta **1829**(1): 84-97.

Soppe, W. J., S. E. Jacobsen, C. Alonso-Blanco, J. P. Jackson, T. Kakutani, M. Koornneef and A. J. Peeters (2000). "The late flowering phenotype of fwa mutants is caused by gain-of-function epigenetic alleles of a homeodomain gene." Mol Cell **6**(4): 791-802.

Soppe, W. J., S. E. Jacobsen, C. Alonso-Blanco, J. P. Jackson, T. Kakutani, M. Koornneef and A. J. Peeters (2000). "The late flowering phenotype of *fwa* mutants is caused by gain-of-function epigenetic alleles of a homeodomain gene." Mol. Cell **6**(4): 791-802.

Storey, J. D. (2002). "A direct approach to false discovery rates." Journal of the Royal Statistical Society: Series B (Statistical Methodology) **64**(3): 479-498.

Stroud, H., M. V. Greenberg, S. Feng, Y. V. Bernatavichute and S. E. Jacobsen (2013). "Comprehensive analysis of silencing mutants reveals complex regulation of the Arabidopsis methylome." Cell **152**(1-2): 352-364.

Sutton, W. S. (1902). "ON THE MORPHOLOGY OF THE CHROMOSO GROUP IN BRACHYSTOLA MAGNA." The Biological Bulletin **4**(1): 24-39.

Suzuki, M. M. and A. Bird (2008). "DNA methylation landscapes: provocative insights from epigenomics." Nat Rev Genet **9**(6): 465-476.

Swarbreck, D., C. Wilks, P. Lamesch, T. Z. Berardini, M. Garcia-Hernandez, H. Foerster, D. Li, T. Meyer, R. Muller, L. Ploetz, A. Radenbaugh, S. Singh, V. Swing, C. Tissier, P. Zhang and E. Huala (2008). "The Arabidopsis Information Resource (TAIR): gene structure and function annotation." Nucleic Acids Res **36**(Database issue): D1009-1014.

Tahiliani, M., K. P. Koh, Y. Shen, W. A. Pastor, H. Bandukwala, Y. Brudno, S. Agarwal, L. M. Iyer, D. R. Liu, L. Aravind and A. Rao (2009). "Conversion of 5-methylcytosine to 5-hydroxymethylcytosine in mammalian DNA by MLL partner TET1." Science **324**(5929): 930-935.

Tariq, M., H. Saze, A. V. Probst, J. Lichota, Y. Habu and J. Paszkowski (2003). "Erasure of CpG methylation in Arabidopsis alters patterns of histone H3 methylation in heterochromatin." Proc Natl Acad Sci U S A **100**(15): 8823-8827.

Tarone, R. E. (1990). "A modified Bonferroni method for discrete data." Biometrics **46**(2): 515-522.

Taylor, S. M. and P. A. Jones (1979). "Multiple new phenotypes induced in 10T1/2 and 3T3 cells treated with 5-azacytidine." Cell **17**(4): 771-779.

Tessarz, P. and T. Kouzarides (2014). "Histone core modifications regulating nucleosome structure and dynamics." Nat Rev Mol Cell Biol **15**(11): 703-708.

The Arabidopsis Genome Initiative (2000). "Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*." Nature **408**(6814): 796-815.

Tian, L., X. Li, M. Ha, C. Zhang and Z. J. Chen (2014). "Genetic and epigenetic changes in a genomic region containing MIR172 in Arabidopsis allopolyploids and their progenitors." Heredity (Edinb) **112**(2): 207-214.

Tollefsbol, T. O. (2011). Chapter 1 - Epigenetics: The New Science of Genetics. Handbook of Epigenetics. San Diego, Academic Press**:** 1-6.

Toyoda, S., M. Kawaguchi, T. Kobayashi, E. Tarusawa, T. Toyama, M. Okano, M. Oda, H. Nakauchi, Y. Yoshimura, M. Sanbo, M. Hirabayashi, T. Hirayama, T. Hirabayashi and T. Yagi (2014). "Developmental epigenetic modification regulates stochastic expression of clustered protocadherin genes, generating single neuron diversity." Neuron **82**(1): 94-108.

Trapnell, C., A. Roberts, L. Goff, G. Pertea, D. Kim, D. R. Kelley, H. Pimentel, S. L. Salzberg, J. L. Rinn and L. Pachter (2012). "Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks." Nature Protoc. **7**(3): 562-578.

Tsukaya, H. (2013). "Leaf Development." The Arabidopsis Book / American Society of Plant Biologists **11**: e0163.

Tsumura, A., T. Hayakawa, Y. Kumaki, S. Takebayashi, M. Sakaue, C. Matsuoka, K. Shimotohno, F. Ishikawa, E. Li, H. R. Ueda, J. Nakayama and M. Okano (2006). "Maintenance of self-renewal ability of mouse embryonic stem cells in the absence of DNA methyltransferases Dnmt1, Dnmt3a and Dnmt3b." Genes Cells **11**(7): 805-814.

Valinluck, V. and L. C. Sowers (2007). "Endogenous cytosine damage products alter the site selectivity of human DNA maintenance methyltransferase DNMT1." Cancer Res **67**(3): 946-950.

Venkatesh, S. and J. L. Workman (2015). "Histone exchange, chromatin structure and the regulation of transcription." Nat Rev Mol Cell Biol **16**(3): 178-189.

Visscher, P. M., W. G. Hill and N. R. Wray (2008). "Heritability in the genomics era--concepts and misconceptions." Nat Rev Genet **9**(4): 255-266.

Vongs, A., T. Kakutani, R. A. Martienssen and E. J. Richards (1993). "Arabidopsis thaliana DNA methylation mutants." Science **260**(5116): 1926-1928.

Waddington, C. H. (1939). An introduction to modern genetics. New York,, The Macmillan company.

Waddington, C. H. (1957). The strategy of the genes; a discussion of some aspects of theoretical biology. London,, Allen & Unwin.

Wang, W. S., Y. J. Pan, X. Q. Zhao, D. Dwivedi, L. H. Zhu, J. Ali, B. Y. Fu and Z. K. Li (2011). "Drought-induced site-specific DNA methylation and its association with drought tolerance in rice (Oryza sativa L.)." J Exp Bot **62**(6): 1951-1960.

Wassenegger, M., S. Heimes, L. Riedel and H. L. Sanger (1994). "RNA-directed de novo methylation of genomic sequences in plants." Cell **76**(3): 567-576.

Watson, J. D. and F. H. Crick (1953). "Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid." Nature **171**(4356): 737-738.

Wibowo, A., C. Becker, G. Marconi, J. Durr, J. Price, J. Hagmann, R. Papareddy, H. Putra, J. Kageyama, J. Becker, D. Weigel and J. Gutierrez-Marcos (2016). "Hyperosmotic stress memory in Arabidopsis is mediated by distinct epigenetically labile sites in the genome and is restricted in the male germline by DNA glycosylase activity." eLife **5**: e13546.

Widman, N., S. Feng, S. E. Jacobsen and M. Pellegrini (2014). "Epigenetic differences between shoots and roots in Arabidopsis reveals tissue-specific regulation." Epigenetics **9**(2): 236-242.

Wright, S. (1932). The roles of mutation, inbreeding, crossbreeding, and selection in evolution.

Wu, A. R., N. F. Neff, T. Kalisky, P. Dalerba, B. Treutlein, M. E. Rothenberg, F. M. Mburu, G. L. Mantalas, S. Sim, M. F. Clarke and S. R. Quake (2014). "Quantitative assessment of single-cell RNA-sequencing methods." Nat Methods **11**(1): 41-46.

Wu, J. C. and D. V. Santi (1987). "Kinetic and catalytic mechanism of HhaI methyltransferase." J Biol Chem **262**(10): 4778-4786.

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

Yen, R. W., P. M. Vertino, B. D. Nelkin, J. J. Yu, W. el-Deiry, A. Cumaraswamy, G. G. Lennon, B. J. Trask, P. Celano and S. B. Baylin (1992). "Isolation and characterization of the cDNA encoding human DNA methyltransferase." Nucleic Acids Res **20**(9): 2287-2291.

Yong-Villalobos, L., S. I. Gonzalez-Morales, K. Wrobel, D. Gutierrez-Alanis, S. A. Cervantes-Perez, C. Hayano-Kanashiro, A. Oropeza-Aburto, A. Cruz-Ramirez, O. Martinez and L. Herrera-Estrella (2015). "Methylome analysis reveals an important role for epigenetic changes in the regulation of the Arabidopsis response to phosphate starvation." Proc Natl Acad Sci U S A **112**(52): E7293-7302.

Zemach, A., I. E. McDaniel, P. Silva and D. Zilberman (2010). "Genome-wide evolutionary analysis of eukaryotic DNA methylation." Science **328**(5980): 916-919.

Zhang, G., H. Huang, D. Liu, Y. Cheng, X. Liu, W. Zhang, R. Yin, D. Zhang, P. Zhang, J. Liu, C. Li, B. Liu, Y. Luo, Y. Zhu, N. Zhang, S. He, C. He, H. Wang and D. Chen (2015). "N6-methyladenine DNA modification in Drosophila." Cell **161**(4): 893-906.

Zhang, X., Y. V. Bernatavichute, S. Cokus, M. Pellegrini and S. E. Jacobsen (2009). "Genome-wide analysis of mono-, di- and trimethylation of histone H3 lysine 4 in Arabidopsis thaliana." Genome Biol **10**(6): R62.

Zhang, X., J. Yazaki, A. Sundaresan, S. Cokus, S. W. Chan, H. Chen, I. R. Henderson, P. Shinn, M. Pellegrini, S. E. Jacobsen and J. R. Ecker (2006). "Genome-wide high-resolution mapping and functional analysis of DNA methylation in *Arabidopsis*." Cell **126**(6): 1189-1201.

Zhang, Z. H., D. J. Jhaveri, V. M. Marshall, D. C. Bauer, J. Edson, R. K. Narayanan, G. J. Robinson, A. E. Lundberg, P. F. Bartlett, N. R. Wray and Q. Y. Zhao (2014). "A comparative study of techniques for differential expression analysis on RNA-Seq data." PLoS One **9**(8): e103207.

Zhao, Q., G. Rank, Y. T. Tan, H. Li, R. L. Moritz, R. J. Simpson, L. Cerruti, D. J. Curtis, D. J. Patel, C. D. Allis, J. M. Cunningham and S. M. Jane (2009). "PRMT5-mediated methylation of histone H4R3 recruits DNMT3A, coupling histone and DNA methylation in gene silencing." Nat Struct Mol Biol **16**(3): 304-311.

Zhu, J., A. Kapoor, V. V. Sridhar, F. Agius and J. K. Zhu (2007). "The DNA glycosylase/lyase ROS1 functions in pruning DNA methylation patterns in Arabidopsis." Curr Biol **17**(1): 54-59.

Zilberman, D., M. Gehring, R. K. Tran, T. Ballinger and S. Henikoff (2007). "Genome-wide analysis of *Arabidopsis thaliana* DNA methylation uncovers an interdependence between methylation and transcription." Nat. Genet. **39**(1): 61-69.

Zilberman, D., M. Gehring, R. K. Tran, T. Ballinger and S. Henikoff (2007). "Genome-wide analysis of Arabidopsis thaliana DNA methylation uncovers an interdependence between methylation and transcription." Nat Genet **39**(1): 61-69.

Zlatanova, J. and A. Thakar (2008). "H2A.Z: view from the top." Structure **16**(2): 166-179.