Characterizing short-term evolution of DNA methylation in *A. thaliana* using next-generation sequencing

Dissertation

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

vorgelegt von
Jörg Hagmann
aus Karlsruhe

Tübingen
2015

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

Tag der mündlichen Qualifikation:	19.10.2015
Dekan:	Prof. Dr. Wolfgang Rosenstiel
1. Berichterstatter:	Prof. Dr. Detlef Weigel
2. Berichterstatter	Prof. Dr. Daniel Huson

Erklärung

Hiermit erkläre ich, dass ich diese Arbeit selbstständig und nur mit den angegebenen Hilfsmitteln angefertigt habe, und dass alle Stellen, die im Wortlaut oder dem Sinne nach anderen Werken entnommen sind, durch Angabe der Quellen kenntlich gemacht sind. Beiträge Dritter, auf die im Text eingegangen wird, sind in jedem Kapitel in den jeweiligen Abschnitten "Contributions" aufgeführt. Die Doktorarbeit basiert auf eigenen Publikationen, die ebenfalls in den Abschnitten "Contributions" aufgelistet sind.

Tübingen, 27.05.2015 Jörg Hagmann

Zusammenfassung

Unterschiede in der DNS Sequenz sind die treibende Kraft natürlicher Variation. In den letzten Jahren fanden jedoch zahlreiche Studien heraus, dass auch epigenetische Veränderungen unterschiedliche Merkmale von Individuen hervorrufen können. Epigenetische Markierungen regulieren die Aktivität der Gene, ohne dass dabei die DNS Sequenz verändert wird. Einer der häufigsten und am besten untersuchten Vertreter von epigenetischen Modifikationen ist die DNS Methylierung, bestehend aus einer zusätzlichen Methylgruppe an der DNS Base Cytosin. Dieses chemische Merkmal lässt sich über Zellteilungen, aber auch über Generationen hinweg stabil weitergeben, und somit lassen sich dauerhafte Unterschiede im Methylierungsmuster von verschiedenen Individuen finden, sogenannte Epimutationen. Solche natürlich vorkommenden Unterschiede haben drei hauptsächliche Ursachen: die meisten sind an genetische Mutationen geknüpft, sie können aber auch rein spontan auftreten oder als Antwort auf eine Veränderung in der Umgebung entstehen. Letzteres erlaubt eine kurzfristige Anpassung an veränderte externe Bedingungen, was über ungerichtete genetische Mutationen in so kurzer Zeit nicht möglich ist. Gegenwärtig findet eine Debatte unter Wissenschaftlern darüber statt, ob die Umwelt gezielt Epimutationen hervorrufen kann, die dauerhaft vererbt werden, was der klassischen Darwin'schen Evolutionslehre widersprechen würde.

Die meisten bisherigen Studien konnten aufgrund ihres experimentellen Aufbaus die drei genannten Faktoren für Methylierungsunterschiede nicht eindeutig trennen oder untersuchten potentiell mit Epimutationen verknüpfte DNS Mutationen nicht im gesamten Genom. Zusätzlich verfolgten sie Epimutationen meist nicht über zwei Generationen hinweg. So war bisher zum Einen unerforscht, wie häufig zufällig auftretende oder von der Umwelt hervorgerufene Epimutationen im gesamten Genom auftreten, zum Anderen blieb bislang unklar, ob und wie dauerhaft diese erworbenen Veränderungen vererbbar sind.

Diese Arbeit untersucht diese Fragen anhand von basengenauen und genomweiten Hochdurchsatz-Analysen des Methylierungsmusters zweier Populationen der Modellpflanze *Arabidopsis thaliana*, bei denen aufgrund des jeweiligen einzigartigen Versuchsaufbaus bestimmte Ursachen für Epimutationen vernachlässigt werden können. In genetisch identischen Individuen, die im Gewächshaus unter einheitlichen Bedingungen für 30 Generationen durch Selbstbefruchtung herangezogen, und bei denen somit genetische sowie umweltbedingte Einflüsse minimal gehalten wurden, zeigte sich, dass spontan auftretende Epimutationen häufig vorkommen, diese jedoch nur recht kurzlebig sind. Die Untersuchung von Pflanzen, die in unterschiedlichen natürlichen Umgebungen in Nordamerika gewachsen waren und die sich erst in den letzten zwei Jahrhunderten unabhängig voneinander entwickelt hatten, ergab ein größtenteils stabiles genomweites Methylierungsmuster, das zudem erstaunlicherweise stark dem der Pflanzen aus dem Gewächshaus ähnelte. Daraus lässt sich schließen, dass durch die Umwelt hervorgerufene Epimutationen wenig zu dauerhaften Methylierungsunterschieden beitragen, was im Widerspruch zu einigen jüngeren Veröffentlichungen steht, die eine weite Verbreitung und Vererbung von gezielten epigenetischen Veränderungen vermuten.

Zusätzlich zu diesen Erkenntnissen umfasst diese Arbeit verbesserte computergestützte Analysemethoden von Next-Generation Sequenzierungsdaten. Um möglichst umfassend den genomweiten genetischen Einfluss auf Epimutationen abzuschätzen, werden in dieser Arbeit Analyseschritte vorgestellt, die viele unterschiedliche Methoden zur Identifizierung genetischer Variation kombinieren und somit viele verschiedene Arten von DNA Sequenzunterschieden erkennen können. Zusätzlich beschreibt diese Arbeit eine umfassende Pipeline

zur Identifizierung von methylierten Cytosinen sowie Methylierungsunterschieden zwischen Individuen. Die meisten bisherigen epigenetischen Studien fanden variable Methylierung vornehmlich auf Cytosinen in einem bestimmten genomischen Kontext und benutzten grob vereinfachende statistische Tests. Diese Arbeit stellt ein neues, sensitives Verfahren vor, das im Vergleich zu den bisherigen Studien vermehrt Unterschiede in allen Regionen im Genom findet und einen robusten statistischen Test anwendet.

Zusammengefasst präsentiert diese Arbeit umfangreiche Analyseschritte zur Detektion von genetischer sowie epigenetischer Variation und untersucht die Rate und das Spektrum von natürlich vorkommenden Methylierungsunterschieden in Pflanzen. Sie leistet somit einen wichtigen Beitrag für unser Verständnis der Rolle von Epimutationen in der Evolution von Spezies.

Abstract

DNA sequence mutations are the principal source of natural variation. Over the last few decades, however, an increasing number of studies have suggested that also epigenetic components can be at the basis of differences in phenotypic traits. These epigenetic marks allow a flexible modulation of gene activity without changes in the DNA sequence. One of the most prominent epigenetic modifications is DNA methylation, which consists of cytosines that carry an additional methyl group. Such chemical marks can be inherited across cell divisions and generations, and there are many durable methylation differences between individuals, so-called epimutations. These can originate from mainly three different sources: most epimutations are coupled to genetic mutations, yet they can also arise spontaneously, or they can be induced by environmental stimuli. The latter case enables rapid adaptation to changing environments, which in the short term is usually not possible via genetic mutations. A current debate revolves around the question whether adaptive environmentally induced epimutations can be heritable, which would contradict the random mutagenesis assumption of Darwinian evolutionary theory.

However, the experimental setup of most studies that have examined epigenetic variation did not allow the clear separation of different sources of variable methylation. These studies typically did not inspect genome-wide genetic variation, or did not monitor environmentally induced changes for more than one or two generations. Thus it has remained largely unresolved how frequently methylation differences arise spontaneously on the whole-genome level, and how strongly and durably environmental conditions impact the methylation landscape.

This work addresses these questions in the model plant *Arabidopsis thaliana*. I present whole-genome DNA methylation analyses at base-pair resolution of two different populations, originating from unique experimental settings that largely eliminate specific sources of epimutations. Investigation of genetically quasi identical lines propagated for thirty generations in uniform greenhouse conditions – thus largely without genetic and environmental influences – revealed that spontaneously occurring epimutations emerged frequently, but seemed to be largely short-lived. Plants with minimal genetic divergence that had grown in diverse natural sites over a previously uncharted time period of over one hundred years exhibited a methylation pattern that was largely stable on the whole-genome level and that was in many aspects intriguingly similar to that of the greenhouse-grown lines. Thus, environmentally induced epimutations seem to be only minor contributors to heritable methylation differences, which challenges published claims of broad-scale inheritance of adaptive epigenetic variation.

This thesis also provides technical and methodological advances of next-generation sequencing (NGS) data analysis. To gauge the genome-wide genetic influence on epimutations, this work provides an iterative workflow that maximizes the detection of a wide range of DNA sequence variants using short NGS reads by integrating several different genetic variation detection approaches. Finally, while previous epigenetic studies in plants, due to rather simplistic statistical testing, largely revealed a biased picture of differential methylation in the genome, this work introduces a comprehensive DNA methylation pipeline for NGS data that includes a novel approach to obtain more sensitive and more unbiased calls of differentially methylated regions.

Together, this work presents advanced computational methods to profile genome-wide genetic and methylation variation, and inspects the rate and spectrum of naturally occurring methylation changes, thus contributing to elucidating the role of epimutations in evolution.

Contents

Prologue **1**

1 Introduction – Epigenetics **5**
1.1 The evolving definition(s) of epigenetics 5
1.2 Roles of epigenetics 7
1.3 Organization of DNA – the chromatin 9
1.4 Epigenetic modifications 10
1.4.1 DNA methylation 10
1.4.2 Chromatin modifications 14
1.4.3 What else is epigenetic? 15
1.5 Stability of epigenetic marks through mitosis and meiosis 16
1.6 DNA methylation in mammals 17
1.7 Sources of inter-individual epigenetic variation 18
1.7.1 Genetically induced epialleles 19
1.7.2 Spontaneously occurring epialleles 21
1.7.3 Environmentally induced epialleles 21
1.8 Contribution of this work to plant epigenetics 24

2 Introduction – Next-generation sequencing and analysis of genetic and epigenetic variation **27**
2.1 DNA sequencing 28
2.1.1 Next-generation sequencing 29
2.1.2 Next-generation sequencing platforms 30
2.1.3 Illumina's sequence-by-synthesis technology 32
2.1.4 Properties of Illumina read data 34
2.2 Analysis of short NGS data 35
2.2.1 Pre-processing reads 35
2.2.2 *De novo* genome assembly 35
2.2.3 Resequencing 36
2.2.4 Genotype calling at genomic positions 37
2.2.5 Structural variation calling 38
2.2.6 Objective of this work: An integrated method to detect genetic variation 40

2.3 DNA methylation sequencing . . . 41
2.3.1 Experimental methods to detect DNA methylation . . . 41
2.4 Analysis of bisulphite sequencing data . . . 45
2.4.1 Mapping of bisulphite treated reads . . . 45
2.4.2 Determining methylated positions . . . 46
2.4.3 Determining differential methylation at single sites . . . 48
2.4.4 Determining differential methylation in regions . . . 52
2.4.5 Objective of this work: WGBS-Seq pipeline and a novel approach to call DMRs . . . 54

3 Integrative detection of genetic variants by iterative re-alignment 57
3.1 General workflow . . . 57
3.1.1 Applicability and availability of the pipeline . . . 59
3.2 Tools for genetic variant detection . . . 60
3.2.1 SNP and small indel calling . . . 60
3.2.2 Structural variation calling . . . 61
3.2.3 Targeted *de novo* assembly . . . 61
3.3 Consolidating variants of different tools . . . 63
3.4 Building a branched reference sequence . . . 64
3.5 Population-aware calling of common and segregating variants . . . 65

4 A pipeline for the detection of differential methylation 67
4.1 General workflow . . . 68
4.2 Alignment of bisulphite-treated reads . . . 69
4.3 Determination of methylation rates . . . 70
4.4 Identification of methylated positions (MPs) . . . 71
4.5 Identification of differentially methylated positions (DMPs) . . . 72
4.5.1 Identification of DMP clusters (DMCs) . . . 74
4.6 Identification of differentially methylated regions (DMRs) . . . 74
4.6.1 Identification of methylated regions (MRs) . . . 75
4.6.2 Selecting regions to test for differential methylation . . . 76
4.6.3 A statistical test for differential methylation . . . 78
4.6.4 Identification of epiallele groups . . . 79
4.6.5 Identification of highly differentially methylated regions (hDMRs) 81

5 The rate and spectrum of natural DNA methylation variation 83
5.1 Data sets . . . 84
5.1.1 Mutation accumulation lines . . . 84
5.1.2 The haplogroup-1 population . . . 85
5.2 Spontaneous DNA sequence changes in the mutation accumulation lines 86
5.3 Genetic variation between the haplogroup-1 population and Col-0 . . . 88
5.4 Genetic variation among haplogroup-1 strains . . . 88
5.5 Spectrum of DNA methylation . . . 91

5.6 Spectrum of single-site epigenetic variation 92
5.7 The rate and recurrence of spontaneously occurring single-site epimutations 96
5.8 The rate and recurrence of single-site epimutations in nature 98
5.9 Determining methylated and differentially methylated regions 100
5.10 Spectrum of differentially methylated regions 102
5.11 The rate and recurrence of differentially methylated regions 104
5.12 Effect of differential methylation on gene expression 106
5.13 Recurrent epimutations under greenhouse and natural conditions 107
5.14 Linkage of epigenetic differences to genetic variation 110
5.15 Population structure of HPG1 strains based on methylation variation . 113
5.16 Methods 115
5.16.1 Sequencing and short read processing 115
5.16.2 Short read alignment 115
5.16.3 Determining high-quality positions 116
5.16.4 Determining DMP clusters 116
5.16.5 Comparison of epivariation between HPG1 strains and methylation-deficient mutants 116
5.16.6 Data accessibility 117

6 Discussion **119**
6.1 An integrated pipeline to call genetic variation 120
6.2 Bisulphite sequencing pipeline 123
6.3 Short-term evolution of DNA methylation 128
6.4 Outlook 135

A Supplemental Figures **139**

B Supplemental Tables **153**

C Command lines **159**

References **163**

Prologue

The story shall begin on the Roslagen archipelago, north of Stockholm, in 1742. The student Magnus Ziöberg, while roaming through his homeland, discovered an inconspicuous, peculiar little plant, which looked like toadflax, but had completely different flowers. Ziöberg was curious enough that soon after the plant found its way to the newly appointed professor of Uppsala University, Carl Linnaeus, later also named Carl von Linné after his ennoblement in 1761. Linnaeus, who was going to profoundly change the naming scheme of species to the consistent taxonomy used until today, had already gained a high reputation in botany at that time. Believing to know the complete Swedish flora, he first suspected someone had played a trick on him and had glued alien flowers to common toadflax. It turned out to be truly toadflax, or *Linaria vulgaris*, but with altered flower symmetry (Figure **0.1**). Linnaeus was fascinated and confused at the same time, because he based his classification of plants on flower anatomy. Consequently, this plant, albeit being identical to common toadflax in all its parts but the flowers, had to be placed in a taxonomic class other than Linaria. More mysteriously, progeny of flower-aberrant plants could produce conventional toadflax flowers. This profoundly shocked his and the common Christian belief of that time that all species had been placed on earth during Creation and new species could not arise. The direct offspring of a species had never before been reported to be of a different species than the parents. Maybe having this view in mind, Linnaeus gave the plant the common name Peloria, Ancient Greek for "monster". Consequently, this aberrant flower symmetry, which was later observed in other species than Linaria as well, is named peloric until today.

Although Linnaeus revised his view and eventually accepted that new species might arise, his initial explanations that Peloria must be a hybrid between an unknown plant and a common *Linaria vulgaris* plant did not hold up, since soon after, both flower shapes were seen on single plants. Bitterly despaired, he had to leave the mystery to numerous following generations to be fascinated by, including the poet Johann Wolfgang Goethe, who drew a sketch of the different plant shapes, and evolutionist Charles Darwin, experimenting on peloric phenomena in a different species, snapdragon.[1]

Since Darwin we have known that species can slowly and gradually change their traits, such as shape, size or the number of seeds. He proposed that random changes occur infrequently and are selected for by the natural environment. If changes have

[1]The section of text up to here is based on chapter 2 of ref. [Kegel, 2009].

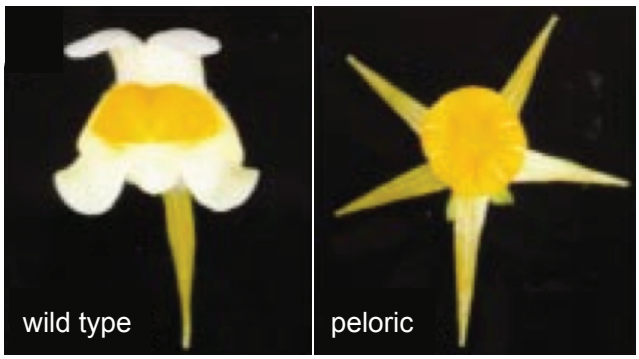

Figure 0.1: Wild type and peloric flowers of *Linaria vulgaris*. Modified from [Cubas et al., 1999].

for example advantages over competitors or better protection against enemies, in the long run such individuals will more likely pass down these traits rather than other potentially detrimental ones to their offspring. Molecular analyses improved throughout the twentieth century and led to the discovery of DNA as the carrier of heritable traits. It became conventional wisdom that the replacement of bases in the DNA sequence by different bases – known as DNA or genetic mutation – is the driving force of the variability and evolution of living organisms. Thus, small and maybe at first irrelevant DNA mutations could gradually accumulate and lead to visible differences in shape or function, or even to the emergence of new species. These theories are today summarized as the Modern Synthesis.

Hence, the peloric flower shape of toadflax, which puzzled so many generations, could in principle simply be explained by one or a few DNA mutations. Therefore, in their attempt to finally track down the molecular cause of Peloria, Pilar Cubas and colleagues focused on analyzing the DNA sequences of a candidate gene (*LCYC*) in both regular and peloric plants in 1999, no less than 257 years after Ziöberg‘s first sighting of the aberrant flowers. As in the past, this peculiar plant raised again some surprise as it turned out that the cause of the flower shape of the “first natural morphological mutant to be characterized” does not trace to DNA mutations: the sequences of the candidate gene in regular and peloric plants was identical [Cubas et al., 1999]. Instead, the researchers found many chemically modified cytosine bases that carried an additional methyl group. Such altered cytosines belong to so-called epigenetic modifications and are termed ‘DNA methylation’. These epigenetic marks were located throughout and around the *LCYC* gene and caused the deactivation of the gene. Typically, particular proteins called ‘transcription factors’ can bind to the region preceding the gene sequence, the ‘promoter’, in order to activate the gene. In case of methylation of specific sites, this binding can be blocked. Thus, the encoded protein of the *LCYC* gene in the peloric variant of toadflax cannot be generated that leads to the asymmetry in the development of the flower in the common, ‘wild type’ variant of Linaria. If this protein is not present at an early stage of development,

the flower develops symmetrically (Figure **0.1**). The methylated variant of the gene constitutes an alternative 'allele', and since the DNA sequence is not affected, it is termed 'epiallele'.

This story shall illustrate two aspects. First, there is more than just the DNA sequence involved in the generation of different shapes and functions, and even in the generation of such crucial traits like the reproductive organs in the case of Peloria. The existence of mutational processes beyond DNA changes was long suspected, and substantiated in the second half of the 20th century, but only in the last few decades a clearer mechanistic understanding of them has emerged. A natural question then is how large the impact of epigenetic modifications, or 'epimutations', is in comparison to DNA mutations on creating the substantial variation seen in nature across and within species. While throughout the last century the focus of research was on genetic mutations, it has remained less well studied how and at what frequency epigenetic modifications can affect traits, or how frequent random epigenetic modifications occur over time.

Second, the fact that the peloric flower shape is inherited from a plant to its offspring – albeit not always – indicates the potential for faithful propagation of epigenetic modifications across cell divisions and even generations. Numerous reports in the past few decades described that plants possess an 'epigenetic memory' established by altered epigenetic states upon exposure to a stress, which can enable facilitated or accelerated future responses to repeated stress periods. Some of these studies claim that this altered epigenetic state can be passed down to following generations as a long-term adaptation to changed environment. This would imply that the environment can influence the timing and genomic location of mutations, which profoundly contradicts the Modern Synthesis. Natural DNA mutations are widely acknowledged to occur at random time points and at random sites in the genome, and that natural selection will then potentially act on these changes. The recently suggested alternative concept is reminiscent of the theory of 'inheritance of acquired traits' by famous French botanist Jean-Baptiste Lamarck, who is today seen as the epitome of the common view of evolution historically before Darwin. However, the findings of the recent studies are ambiguous and constitute a matter of current scientific debate.

While generations of researchers had to rely solely on phenotypic characterizations or approximate observations of DNA methylation patterns, the rise of 'next-generation sequencing' at the beginning of this century revolutionized molecular analyses, including studies of the genome-wide DNA methylation pattern, termed the 'methylome' of an organism. Though still rather costly, this technology allows inspecting DNA methylation down to the single nucleotide level, yielding an almost complete picture of this epigenetic modification in an individual. This comes at the cost of a massive amount of data to analyze, which imposes challenges to resource management, statistical testing and handling of different sources of noise.

This work exhibits three major contributions. First, I introduce complete pipelines for next-generation sequencing data to identify a wide range of DNA mutations as well as DNA methylation differences between individuals, affecting both single nucleotides as well as larger regions of the genome.

Second, I applied these methods to identify methylation differences between lines of the model plant *Arabidopsis thaliana* grown for 30 generations in a controlled greenhouse environment. Because the individuals were genetically identical ('isogenic'), we could virtually rule out genetic effects on DNA methylation, which are commonly a major source of the establishment of epialleles. In addition, as these plants were grown under predominantly stable environmental conditions, differences in DNA methylation should constitute spontaneously occurring pure epialleles. While estimates of DNA mutation rates are widely known in many species, a spontaneous epimutation rate largely independent of genetic and environmental impacts has not yet been reported. We observed a surprising high rate of emerging epialleles with many of them showing an opposite behavior to genetic mutations.

Third, again using the newly developed pipelines, we compared genomes and methylomes of thirteen Arabidopsis strains that had grown in their natural settings at dispersed locations in North America. These lines likely derived from a common ancestor only around 200 years ago. Therefore, the strains were genetically nearly identical, but had been exposed to the natural fluctuating environment for over hundred generations. While other population epigenomic studies compared individuals separated by hundreds of thousands of years, or monitored differences in controlled settings for only a few generations, this study allowed us to estimate the impact of natural environments over a long time period, but largely independent of genetic variation. We addressed the questions if epialleles accumulate faster under environmental impact, and if they are more often adaptive, i.e. have an effect on the gene activity pattern, than epigenetic changes arising under uniform conditions.

The thesis is organized as follows: Chapter 1 introduces the wide field of epigenetics and its most important molecular players and presents the current knowledge about the sources of DNA methylation variation found between individuals. The incentives and contribution of this work to the field of plant epigenetics will be scientifically summarized at the end of the chapter. Chapter 2 will give an overview of the experimental techniques, the challenges and computational steps involved in analyzing whole-genome genetic and epigenetic variation between individuals and outlines the need for enhanced analytical workflows. Novel complete pipelines to accurately identify a wide range of DNA sequence differences and diverse features of differential methylation will be presented in detail in the chapters 3 and 4, respectively. Chapter 5 summarizes the results from the genome-wide characterization of the DNA sequence and methylation variation in the two introduced *A. thaliana* populations, using the newly developed computational analysis tools. This work ends with discussing the main findings and by setting this work in the context of current research in chapter 6.

Chapter 1

Epigenetics

Epigenetics has become a familiar expression in scientific and also popular literature. Studies investigating the important involvement of epigenetic processes mainly in cancer, but more and more in a plethora of other human diseases, have mushroomed in the last decades. An emerging notion that environment has a strong and lasting impact on our lives and on that of our descendants led to eye-catching and fairly exaggerated headlines like "You are what you eat" [Susiarjo and Bartolomei, 2014], "You are what your dad ate" [Ferguson-Smith and Patti, 2011], or even to a cover story in one of the biggest renowned German newspapers, headlining "The victory over the genes. Smarter, healthier, happier. How we can outwit our genome"[1]. Such excitement might reflect the constant struggle by proponents of epigenetics, first for the acceptance of their field, later for the awareness of its importance in the era of genetic determinism.

Yet, neither today nor in the past has there been agreement on the exact definition of epigenetics. A careful consensus about epigenetics as it is maybe most commonly seen today could be summarized as the study of gene activity changes that cannot be explained by primary DNA sequence changes. These differences can be heritable, meaning that they are transmitted across cell divisions, thereby remaining fairly stable during the lifetime of an organism, or that they are even passed from an individual to its progeny. Compared to the original meaning proposed by British scientist Conrad Hal Waddington in 1942 [Haig, 2004], the definition underwent several transitions, or co-existed with alternative descriptions, as will be briefly outlined in the following.

1.1 The evolving definition(s) of epigenetics

Waddington understood epigenetics as encompassing all mechanisms that lead from a single fertilized egg (the 'zygote') to a variety of cells of different shape and function, the process we call cell differentiation today. Two schools existed among the early embryologists. "Preformationists" believed that all adult characters of an organism are present and fully functional in the egg and solely need to be enlarged during

[1] http://www.spiegel.de/spiegel/print/d-73109479.html, last accessed May 2015

development. Proponents of the concept of "epigenesis" envisioned the development as a process of chemical reactions, potentially taking place in the cytoplasm, following a complex plan starting from an unknown primordial material [Felsenfeld, 2014]. Waddington already acknowledged that "the zygote contains certain preformed characters" [Waddington, 1956], but that "processes of epigenesis" must act on them during development. Such effectors, he envisioned, can be interactions of genes with each other or with the environment. Because the discipline of "preformed" characters at that time had already been called "genetics", the word epigenetics combined "epi" from epigenesis with genetics. The suitable ambiguity of this word comes from the fact that the Greek prefix 'epi-' translates to "on top of" or "in addition", implying that there is something in addition to genetics that enables cells to develop into various different cell types.

Indeed, studies over many decades in the 20$^{\text{th}}$ century [Muller, 1930, Briggs and King, 1952, Laskey and Gurdon, 1970] established the awareness that the DNA sequence might remain unchanged during an organism's development. That this view spread so slowly might be due to the decryption of the DNA structure and its replication mechanism by James Watson and Francis Crick in 1953, which steered the public attention on the DNA sequence and the genes and ascribed them to contain the complete plan of the body. Nonetheless, presumably geneticist David Nanney [Burggren and Crews, 2014] coined an at first parallel and later predominant usage of the term epigenetics in 1958 as more generally describing processes leading to differences in phenotype that are not encoded in the DNA sequence [Nanney, 1958].

In the following decades it became clear that the mechanisms taking place during an organism's development alter the activity of their genes, the 'gene expression'. Riggs, Holliday and Pugh were the first to assign a role of DNA methylation in modulating gene activity [Riggs, 1975, Holliday and Pugh, 1975]. In the 1970's, Adrian Bird could show that such distinguishing gene activity patterns were faithfully copied across clonal cell divisions ('mitosis'), as is the case for the DNA sequence [Bird, 1978]. Soon after, epigenetic marks were found to play a role in the long-known phenomenon of 'paramutation', when a locus represses gene expression of a remote locus with similar, derived sequence ('homologous locus in *trans*'). Such altered expression can be transmitted to following generations and even lead to expression changes in other homologous loci. Thus, it was shown that epigenetic states can be transmitted through meiosis as well. These findings together led to the addition of a heritability aspect in the definition of epigenetics, reflecting the current usage summarized *sensu* Riggs as "the study of mitotically and/or meiotically heritable changes in gene function that cannot be explained by changes in DNA sequence" [Riggs and Porter, 1996].

However, the inclusion of the heritability aspect remains questionable, since most contemporary views on epigenetics include all chromatin marks, transcriptional effects of RNA molecules, or higher order chromosomal organization, even if some of these marks are only short-lived and not ultimately proven to be directly transmissible through meiosis [Bird, 2007]. This current usage is maybe best illustrated by many emerging studies reporting 'epigenetic' changes in the brain, while these changes

clearly will not be passed down to daughter cells, since almost no neuron ever divides [Bird, 2007]. Indicatively, even large amply-funded projects like the NIH Roadmap Epigenomics Mapping Consortium explicitly include "stable, long-term alterations in the transcriptional potential of a cell that are not necessarily heritable"[2] in their focus of research.

Since the definition of epigenetics is based primarily on excluding what it is *not* certainly adds to the variety of existing definitions of it. Thus, one can find many more proposals to refine or restrict epigenetics in literature, e.g. the requirement that the initiation of a new epigenetic state should involve a transient mechanism (such as through transcription factors) separate from the one required to maintain it [Berger et al., 2009].

As the molecular mechanisms of epigenetic phenomena will be more and more elucidated, the evolution of the definition of epigenetics might continue in the foreseeable future. In this work, I follow the less stringent description of epigenetics by the NIH Consortium. It shall be noted here that throughout this work, expressions with the prefix 'epi' mostly refer to the epigenetic equivalent of genetic phenomena, e.g. epiallele refers to an epigenetic state and epimutation to an epigenetic modification.

1.2 Roles of epigenetics

When the new millennium started with the momentous decryption of complete genomes, especially the human DNA sequence, researchers and even former US president Bill Clinton announced big promises, including a revolution of diagnosis and therapy of human diseases[3]. Until today, a plethora of genome-wide association studies (GWAS) have been performed, which try to find correlations between DNA mutations and specific phenotypes or diseases [McCarthy et al., 2008, Visscher et al., 2012]. Yet, only few associations to increased disease risk were found and most dependencies – for almost all complex traits – remained obscurely hidden as "missing heritability" [Manolio et al., 2009]. Together with the disillusionment of cloning individuals and the widely observed phenotypic plasticity, researchers assigned more and more importance to non-genetic components and thus, by definition, to epigenetic effects. Paradoxically, soon after the peak of the era of genetic determinism at the beginning of the 21st century, the field of epigenetics experienced a reawakening, and most revelations of its molecular mechanisms have been gained in the last 20 years, many of them while studying these systems in plants.

Epigenetic processes include chemical modifications of the DNA base cytosine (DNA methylation) or the histone proteins around which the DNA is wound, as well as small RNA classes and higher order chromatin remodeling mechanisms (see Figure

[2] http://www.roadmapepigenomics.org/overview, last accessed April 2015

[3] http://web.ornl.gov/sci/techresources/Human_Genome/project/clinton1.shtml, last accessed April 2015

1.1 and section 1.4). Their concerted action modulates gene activity and creates the multiple different tissue types with specific gene expression patterns that constitute an organism. Once differentiated, these states can be maintained through myriads of cell divisions. Besides this historically original role, epigenetic marks are crucial for repressing transcription ('silencing') of large parts of the genome, mainly transposons and frequently occurring sequences (repeat elements). Transposons are mobile genetic elements that can propagate or relocate by integrating into random sites of the genome, thereby potentially disrupting the function of genes or gene regulatory sequences. Thus, epigenetic modifications serve to maintain genome integrity and therefore their general genome-wide pattern is kept stable across numerous generations [Vaughn et al., 2007, Zhang et al., 2008]. While the suppression of transposon or gene transcription by DNA binding proteins and micro- or small-interfering-RNAs (termed post-transcriptional silencing) relies on the steady *de novo* synthesis of these molecules, epigenetic marks can create and maintain a long-term repressed (transcriptionally silenced) or active environment by chemically modifying the DNA or histone proteins. Methylation of gene bodies, mostly in exons, is commonly associated with moderate gene expression [Zilberman et al., 2007] and is conserved over evolutionary time [Takuno and Gaut, 2013].

In contrast to this general inter-generational stability, epigenetic modifications can be highly dynamic intra-generationally, i.e. during the lifetime of an organism. Epigenetics acts as a mediator between the genome and the environment. Epigenetic modifications were found to be involved in adaptation mechanisms to adverse environmental influences and biotic and abiotic stresses, for example pathogen attacks [Gutzat and Mittelsten Scheid, 2012]. Such stress-induced epigenetic states were commonly thought to revert back to the initial state soon after the stress disappears. However, in some cases the epigenetic system seems to be capable of memorizing experienced stress and facilitate or accelerate future responses to the same unfavorable conditions [Pecinka and Mittelsten Scheid, 2012] (see section 1.7.3). Such an "epigenetic memory" was particularly often found in plants as a means to quickly adapt to changing environments that the plants cannot evade. Moreover, recent studies received both broad attention and critical skepticism by claiming that such acquired, targeted adaptations can also be stably transmitted to next generations, which would have a profound impact on our current understanding of undirected adaptation to changing environments and the evolution of species [Heard and Martienssen, 2014] (see section 1.7.3).

Multiple abnormalities are known to occur when enzymes of the DNA methylation machineries are disrupted in plant or mammalian germ lines, underpinning the essential role epigenetics plays in plants and animals. Numerous severe developmental phenotypes were reported in plants [Stroud et al., 2013b], alongside cognitive, neurological and behavioral abnormalities – besides many lethal cases – in humans [Brookes and Shi, 2014]. Furthermore, aberrations of epigenetic modifications in somatic cells were also found to be associated with human disease predisposition or onset – may it be in company with genetic mutations, as in the case of

cancer, autism, schizophrenia or congenital heart disease, or supposedly environmentally induced, as in age-related neurodegenerative diseases [Barrow and Michels, 2014, Brookes and Shi, 2014].

Some hypotheses assign an intermediate status towards evolutionary fixation to epigenetic modifications, since methylated cytosines mutate more frequently than unmethylated bases [Molaro et al., 2011]. Thus, advantageous traits might be assimilated by genetic mutations over time, and epigenetics solely serve as an experimental role in exploring the expression landscape without rare irreversible DNA sequence changes [Silveira et al., 2013, Molaro et al., 2011].

Together, epigenetic mechanisms are vital and involved in a plethora of cellular processes, as underlined – last but not least – by the fact that DNA methylation and even main components of the methylation machinery are conserved across almost all eukaryotes [Feng et al., 2010].

1.3 Organization of DNA – the chromatin

When concatenating all chromosomes of a single human cell, the resulting thread of DNA would be 1,80 meters in length (that of *Arabidopsis thaliana* still 6 centimeters), but an average human cell diameter is only 10μm, thus smaller by a factor of 180,000. Not surprisingly, instead of being randomly arranged, this fiber is highly structured in all organisms whose cells contain a nucleus ('eukaryotes'). The DNA is wrapped around proteins, so-called histones (Figure **1.1**). There are five major families of histones divided into four 'core histones' (H2A, H2B, H3, H4), and two kinds of histones summarized into one 'linker histone' family (H1, H5). Most commonly, two of each of the core histones form an octamer and bind a stretch of DNA spanning around 147 bases. The start and end of this DNA region can be covered by a linker histone, adding to the condensation of the double helix. This DNA-histone complex is the basic structural unit of the chromatin and is termed nucleosome (Figure **1.1**). The nucleosomes localize in specific distance to each other, interconnected by approximately 50 bp long 'linker DNA' devoid of histones. These units are further compacted into higher-level structures, ultimately reducing the length of the unwound DNA compared to compacted DNA by four orders of magnitude.

The occupancy of the genome with nucleosomes and the degree of condensation (e.g. presence of linker histones) determines the general structure of the genome, its 'chromatin'. Genome regions of tightly packed nucleosomes are in general transcriptionally inactive as a consequence of hampering the attachment of non-histone DNA-binding proteins such as transcription factors. Such regions are defined as heterochromatin (Figure **1.1**). In contrast, less compact DNA usually devoid of linker histones is present in a "beads on a string" shape, which is termed euchromatin and which is transcriptionally more active. Consistently, most genes localize to the euchromatin, while transposable elements, DNA repeats and highly repetitive, inactive ribosomal genes are predominantly contained in the heterochromatin [Arabidopsis Genome Initiative, 2000].

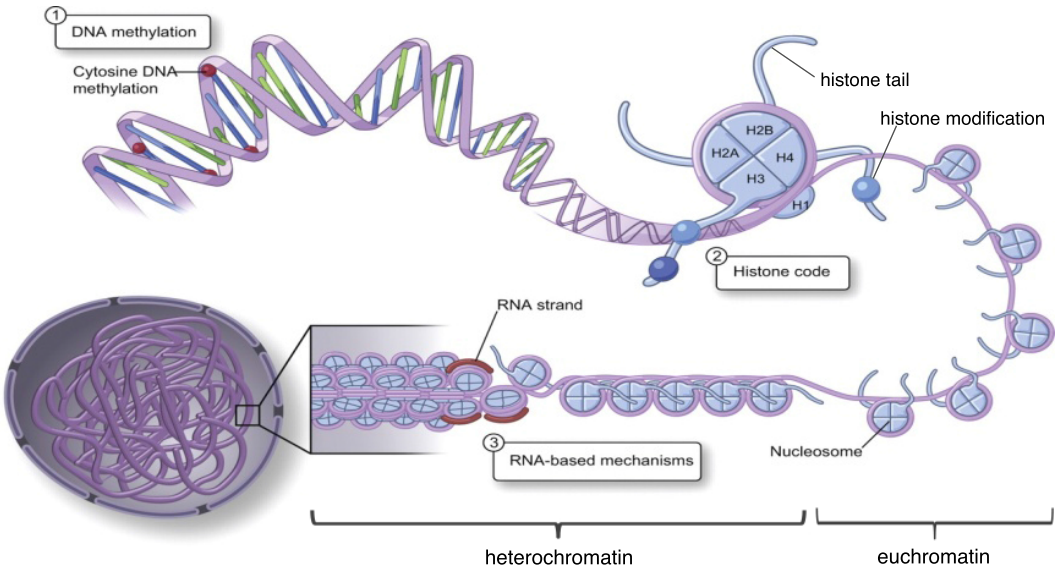

Figure 1.1: Organization of DNA in the nucleus and three main mechanisms of epigenetic gene regulation (enumerated boxes). Modified from ref. [Yan et al., 2010].

In *A. thaliana* as well as in humans, the chromosome arms are mostly euchromatic, and indeed nearly all protein-coding genes are found in these regions. The heterochromatin is situated in a consecutive stretch within the chromosome that contains the centromere, the region where both sister chromatids bind together during the strictly regulated cell division process before they are segregated into two daughter cells.

The spatial structure of the chromatin and thus the accessibility of the DNA to nuclear proteins is influenced by diverse chemical modifications to the DNA as well as to the histones, by the composition of histone variants, or by nucleosome positioning, all of which are considered to be epigenetic, since the DNA sequence is unaffected. Such modifications and their roles will be introduced in detail in the following sections.

1.4 Epigenetic modifications

1.4.1 DNA methylation

The best-studied epigenetic mark is DNA methylation, the addition of a methyl group (CH_3) to the fifth carbon in the aromatic ring of the DNA base cytosine (Figure **1.2**). The correct chemical notation is 5-methylcytosine, but since it is the most common form of methylation of DNA bases, the term DNA methylation generally refers to this modification. This mark is found in all vertebrates and most eukaryotes underpinning the important role it plays in cellular functioning [Suzuki and Bird, 2008, Feng et al., 2010]. Other, rather rare forms of cytosine methylation have been identified in humans (see section 1.6).

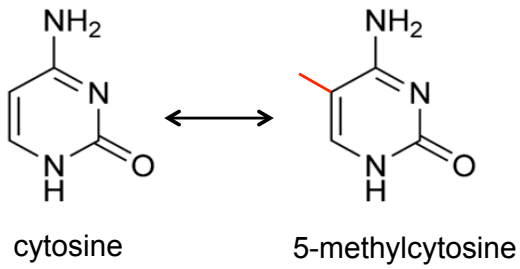

Figure 1.2: Chemical notations of cytosine and 5-methylcytosine. Their difference is the methyl group (CH_3), highlighted in red.

In plants, the complex DNA methylation machinery comprises at least four interwoven pathways: the maintenance system, *de novo* DNA methylation, DNA demethylation and a process termed RNA-directed DNA methylation. The different pathways act on different, yet partly overlapping patterns of sequence contexts and genomic annotations. Cytosines can occur within a CG dinucleotide on the DNA sequence, or in a CHG or CHH context, where H represents all bases except G (i.e. A, C, or T). The sequences CG and CHG are symmetrical since they read identically on both DNA strands (to be strict, except for the sequence CCG, which is not in CHG context on the opposite strand). This symmetry ensures the faithful reinforcement of methylation on double-stranded DNA, when only one of both strands carries methylation marks ('hemimethylation'), as is frequently found immediately after DNA replication. Enzymes termed DNA methyltransferases detect hemimethylated sites and place a methyl group to the opposite, unmethylated cytosine. This maintenance system comprises METHYLTRANSFERASE 1 (MET1) on CG and CHROMOMETHYLASE 3 (CMT3) on CHG sites [Law and Jacobsen, 2010]. MET1 strongly depends on auxiliary enzymes, VARIANT IN METHYLATION 1-3 (VIM1-3). Mutant plants with a disrupted MET1 gene (*met1* mutants) show severe developmental phenotypes [Kankel et al., 2003] and even have perturbed other methylation pathways, as demonstrated by a two-fold reduction of whole-genome CHH methylation levels [Lister et al., 2008]. CMT3 as well seems to rely on cofactors, such as the 'chromatin remodellers' SUPPRESSOR OF VARIEGATION 3-9 HOMOLOGUE 4 (SUVH4, also known as KRYPTONITE or KYP), SUVH5 and SUVH6 [Law and Jacobsen, 2010], but their interactions and dependencies are not entirely understood.

Throughout eukaryotes, DNA methylation occurs within gene bodies almost exclusively in the CG context. Studies based on different kinds of data and using different methodologies determined between 20-30% of analyzed genes to contain methylated cytosines [Zhang et al., 2006, Zilberman et al., 2007, Takuno and Gaut, 2012]. This marking is preferentially associated with moderate gene expression [Zilberman et al., 2007]. About 5% of *A. thaliana* genes have methylated sites in their promoter region [Zhang et al., 2006], which has a largely repressing effect on gene transcription. In contrast, most transposable or repetitive DNA elements show

high and dense levels of methylation in all three sequence contexts in seed plants like Arabidopsis [Zilberman et al., 2007, Lee et al., 2010], conferring transcriptional silencing.

Contrary to the CG and CHG contexts, the CHH sequence lacks a cytosine on the opposite strand, and thus potential methylation marks are lost after each DNA replication on one of the two newly synthesized double strands. So-called *de novo* methyltransferases can place methylation without a template methylcytosine on the opposite strand and they include CHROMOMETHYLASE 2 (CMT2), targeting almost exclusively CHH sites, as well as DOMAINS REARRANGED METHYLTRANSFERASE 1 (DRM1) and DRM2, targeting sites of all contexts, but mainly CHH. This process requires sequence specificity of the enzymes by still largely obscure mechanisms.

However, the specificity mainly of DRM2 has been shown to be dependent on small, non-coding RNAs. In a process called RNA-directed DNA methylation (RdDM), the plant-specific RNA polymerase Pol IV (a DNA-dependent RNA polymerase) copies small parts of the target DNA strand into RNA, which is then complemented to double-stranded RNA by RDR2 (RNA-DIRECTED RNASE 2; Figure **1.3**).

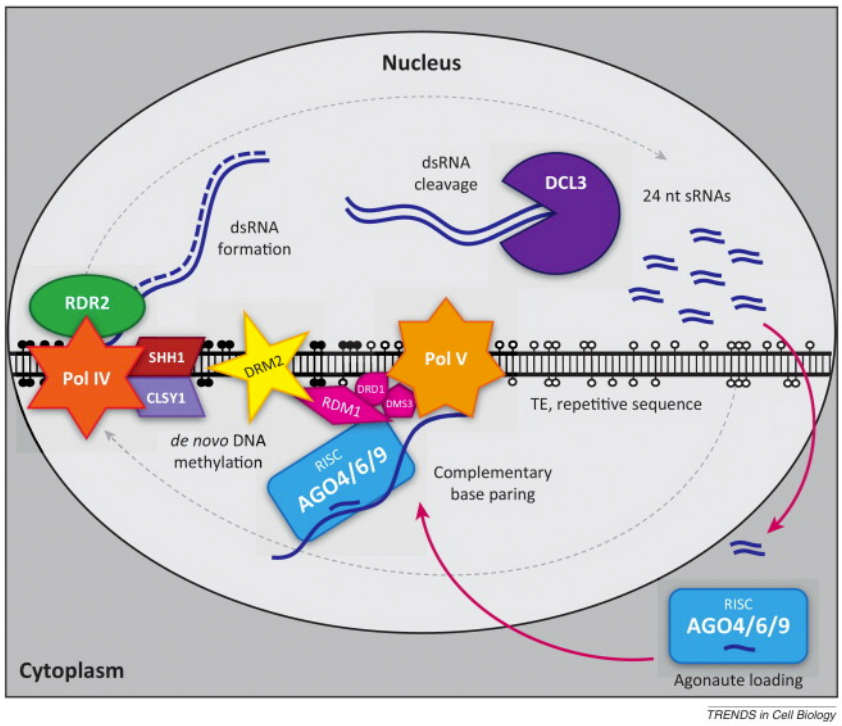

Figure 1.3: Schematic mechanism of RNA-directed DNA methylation (RdDM). See text for explanations. Filled and blank lollipops indicate methylated and unmethylated cytosines, respectively. dsRNA: double-stranded RNA, TE: transposable element. Modified from [Bond and Baulcombe, 2014].

Dicer-like proteins (DCL) cleave these molecules into small interfering RNAs (siRNAs), mostly 24 nucleotides in length. They are transported from the nucleus into the cytoplasm where they are loaded on Argonaute (AGO) proteins, thereby forming the RNA-induced silencing complex (RISC). This complex moves back into the nucleus and can bind to nascent RNA transcripts by complementary base pairing to the exact sequence of the siRNA in the RISC (Figure **1.3**). Such nascent RNAs are produced by yet another plant-specific RNA polymerase, Pol V. Ultimately, the *de novo* methyltransferase DRM2 is associated with or recruited by the protein-RNA complex and mediates methylation of the underlying DNA sequence [Bond and Baulcombe, 2014]. Plants with defective RdDM components show few obvious phenotypes [Matzke and Mosher, 2014, Stroud et al., 2013b], hinting at compensatory mechanisms. The maintenance methyltransferase CMT3 seems to act redundantly with DRM2, since both single mutants show no phenotypic effect, but the *cmt3drm2* double mutant has severe abnormalities [Mathieu et al., 2007].

In addition, a different version of the 'canonical' RdDM pathway has been suggested, which involves enzymes and reactions of the post-transcriptional gene silencing mechanisms [Matzke and Mosher, 2014, Bond and Baulcombe, 2015]. It might initially induce methylation that is then maintained by the canonical pathway [Nuthikattu et al., 2013], implicating a coupling of DNA methylation and transcription. Indeed, plants deficient in canonical RdDM components show reduced DNA methylation in transposable elements and repetitive regions nearby genes in largely actively transcribed chromatin [Zemach et al., 2013]. The plant-specific RNA polymerases IV and V evolved from the mRNA-producing Pol II, hinting at potentially derived targeting [Matzke and Mosher, 2014]. Thus, initial transcription might play a role in directing RNA-mediated DNA methylation.

The methylation-inducing siRNA molecules do not necessarily derive from the same locus that is methylated (*cis*-effect); they can also originate from remote locations in the genome (*trans*-effect) or even from other cells [Slotkin et al., 2009, Matzke and Mosher, 2014], expanding the scope of DNA methylation control.

Established methylation marks can also be removed, either passively or actively. Passive demethylation occurs in combination with DNA replication and the absence of methylation of newly synthesized strands. Active removal is independent of DNA replication. Especially the RdDM mechanism can be counteracted by active demethylation via so-called DNA glycosylases. They are capable of removing methylated cytosines through DNA repair mechanisms and replace them with unmethylated cytosines. Members of this enzyme family include ROS1, ROS3, DML2 and DML3, targeting preferentially RNA-directed methylated sites in vegetative tissue, and DEMETER (DME), which is involved in sex-specific demethylation ('imprinting') during gametogenesis [Gehring, 2013] (for review of DNA glycosylases, see [Law and Jacobsen, 2010]). Although their target specificity is largely unknown, studies with knockout mutants revealed that they seem to preferentially maintain a demethylated state on gene

promoters and 3' untranslated regions, on some transposons near genes and at boundaries between eu- and heterochromatin, which altogether suggests a role in protecting genes from silencing [Lister et al., 2008, Law and Jacobsen, 2010].

In addition to the processes discussed above, other marks might serve as a signal for DNA methylation, for example histone modifications or nucleosome positioning, as is described in the next section.

1.4.2 Chromatin modifications

The DNA is wound around histones in the nucleosomes. The cores of H2 and H3 histones as well as amino acid tails of the H3 and H4 histones protruding from the nucleosome are accessible to chemical modifications (Figure **1.1**). Those include, among others, the covalent addition of chemical groups in methylation, acetylation and phosphorylation, the binding of small proteins in ubiquitination, the addition of moieties like ADP-ribose, or the exchange of chemical groups in citrullination [Huang et al., 2014]. In some way, all of them may affect the charge of histones and thus the electrostatic affinity of the DNA bound to these regions, ultimately shaping the degree of DNA packing and the accessibility of the double helix to nuclear proteins.

Moreover, there are multiple copies of histone genes in eukaryotic genomes that differ in their genomic sequence, each giving rise to a histone variant. The combination of histone variants and histone modifications constitute the "histone code". Different genomic regions can have different histone marks that might confer specific functions on them. For example, a specific part of the centromeres of Arabidopsis is specifically bound by the histone variant CENH3, which replaces the H3 variant, and this region was shown to bind a protein complex called kinetochore that is responsible for chromosome segregation during cell division [Henikoff and Dalal, 2005]. Specific chromatin modifications are also crucial components or regulators of diverse DNA repair processes [House et al., 2014]. Integrated analyses across different species distinguish certain combinations of histone marks into broad functional genome annotations like promoter, enhancer, gene bodies or heterochromatin [Ho et al., 2014]. Based on twelve histone marks, another study in Arabidopsis grouped chromatin into regions of active genes, repressed genes, repressed repetitive elements and intergenic regions [Roudier et al., 2011], which was recently refined into six groups based on the analysis of 16 chromatin marks [Wang et al., 2015]. Thus, like DNA methylation, histone modifications can be associated with transcriptional activity. As examples, it was suggested that histone acetylation denotes mainly active, 'open' chromatin [Cedar and Bergman, 2009], and the tri-methylation of the fourth lysine (amino acid abbreviation: K) of histone H3 (H3K4me3) was found at two thirds of endogenous genes and promoters, but not at silent transposons [Zhang et al., 2009]. Furthermore, the histone variant H2A.Z is enriched at transcription start sites (TSS) but depleted from the bodies of expressed genes [Coleman-Derr and Zilberman, 2012, Zilberman et al., 2008]. The pattern of H2A.Z displays the opposite of the DNA methylation distribution,

and indeed such an inhibitory interplay between these two marks has been proposed [Coleman-Derr and Zilberman, 2012].

The latter case is only one of many known interconnections between the histone code and DNA methylation. Classified as a chromatin remodelling enzyme, DECREASE IN DNA METHYLATION 1 (DDM1) is required for the DNA methylation of a large portion of cytosines in all contexts [Zemach et al., 2013]. DDM1 is thought to synergize with the RdDM pathway in methylating short transposons near genes [Zemach et al., 2010, Ibarra et al., 2012], and with CMT2 independently of RdDM in methylating most transposable elements in the strongly heterochromatic genome regions at and around the centromeres [Zemach et al., 2013, Teixeira et al., 2009]. It was suggested that DDM1 removes the linker histone H1 from nucleosomes, allowing DNA methyltransferases better access to the double helix [Zemach et al., 2013]. Besides *met1*, *ddm1* plants show the most severe developmental abnormalities among single mutants of the methylation machinery, underlining the importance of chromatin remodelling for establishing the DNA methylation pattern [Kakutani et al., 1996].

In *met1* mutants, CG methylation is drastically reduced, as is the (mono-)methylation of H3K9, indicating a direct correlation between DNA methylation and histone marks [Lister et al., 2008, Tariq et al., 2003]. A similar pattern is observed for CMT3-mediated CHG methylation and H3K9me2, supported by known binding affinities of CMT3 to this histone mark, and of the chromatin remodeler KYP to CHG-methylated sites [Law and Jacobsen, 2010]. The latter two examples constitute positive feedback loops between symmetrical DNA methylation and histone modifications to reinforce a silent state. Asymmetrical CHH methylation might be regulated by a feedback loop as well, since it has been suggested that the sRNA-producing polymerases preferentially target methylated DNA [Matzke and Mosher, 2014].

A further epigenetic mark that could contribute to the site specificity of DNA methylation is the positioning of nucleosomes across the genome. Chodavarapu *et al.* reported correlations between DNA methylation and the location of nucleosomes, most notably that nucleosome-bound DNA is higher methylated in all contexts compared to flanking genomic regions, and that nucleosomes are enriched on exons, as are DNA methylation and the RNA polymerase II [Chodavarapu et al., 2010].

Together, DNA methylation marks and histone modifications, histone variant selection and histone and nucleosome placement tightly act together in a complex and still not fully understood interplay in shaping the chromatin and in modulating gene expression patterns. To date, in most cases it is not possible to pinpoint the origin of a chromatin state change; for example whether a DNA methylation change caused histone modifications, or vice versa, or whether transcription or chromatin changes are primary or secondary effects.

1.4.3 What else is epigenetic?

Reports infrequently include also chromatin-independent marks to be epigenetic. Self-sustaining feedback loops have been proposed between products of genes, i.e. mRNAs

or proteins such as transcription factors, which can reinforce their own transcription after cell divisions [Jablonka and Raz, 2009, Heard and Martienssen, 2014]. Furthermore, identical proteins can fold into different structures. It has been suggested that prions, infectious proteins that can induce a misfolding of other proteins to match their own structure, can be inherited across generations in fungi [Chien et al., 2004]. Other, more behavioral effects that parents can pass on to their offspring independent of their genotype are occasionally termed epigenetic and include ecological and cultural inheritance, such as niche construction (e.g. building webs or nests) or behavior learnt by others [Danchin et al., 2011]. However, the transmission of all of these marks through meiosis remains to be proven, at least in plants.

1.5 Stability of epigenetic marks through mitosis and meiosis

Differentiated cells stably maintain their chromatin states throughout the organism's whole lifetime. I elucidated mechanisms in the previous sections, which reinforce chromatin modifications via hemimethylated template strands or positive feedback loops during or after mitosis. Soluble molecules such as siRNAs might help initiating epigenetic marking in daughter cells as well, since they get nearly equally distributed during division.

In principle, these processes could act similarly during meiosis. However, plants and most other organisms with DNA methylation have developed mechanisms to erase epigenetic alterations to the chromatin in their germ cells or during embryo development (that is, just before or after meiosis), counteracting any heritability of acquired changes. This has obvious reasons, since germ cells must preserve and transmit the ability to differentiate into all possible tissue types (pluripotency) to their progeny cells, which can then again gradually undergo chromatin modifications towards their differentiated fate. Since germ cells in plants are generated from differentiated adult somatic cells late in development, epigenetic changes that might have been accumulated over time in these cells due to their environmental exposure would be transmitted to the next generation.

Indeed, in plants like Arabidopsis, pluripotent germ cells (the egg and the sperm cells) show a different genome-wide methylation profile than somatic cells. Although the detailed analysis of these single cells is technically challenging and thus not completely explored, an appealing view has emerged for the male gamete. The two sperm cells and their associated cell, the vegetative nucleus (VN), lose mostly CHH methylated sites from transposons, presumably due to the lack of DDM1 [Calarco et al., 2012]. In the VN, a large structural reorganization takes place, including high expression of the demethylation enzyme DME, loss of H3K9me2, and histone replacements (e.g. the centromere mark CENH3 is lost), resulting in a highly decondensed chromatin [Calarco et al., 2012]. This gives rise to ample transposon transcription and activation, which in turn produces siRNAs from these locations. Intriguingly, not the 24 nt

siRNAs as in the canonical RdDM pathway are generated, but 21 nt "epigenetically activated RNAs" (easiRNAs), which are coupled to transcription [Creasey et al., 2014]. Slotkin *et al.* detected these molecules in the sperm cell, where RdDM components are strongly repressed [Slotkin et al., 2009]. This suggests that these intercellularly mobile molecules could reinforce methylation and thus silencing of heterochromatic regions in the sperm via the RdDM pathway at later developmental stages, e.g. in the seed during or after fertilization [Calarco et al., 2012]. Alternatively or additionally, re-methylation in the seed or embryo could be directed by 24 nt siRNAs transmitted from maternal tissue like the seed coat or the endosperm [Mosher et al., 2009]. One reason why higher plants can risk the activation of potentially harmful transposons is that this happens only in the terminally differentiated vegetative nucleus, meaning that the effects are not transmitted to the progeny.

Similar processes have been reported in the female gametogenesis despite even more experimental limitations. The central cell, which is associated to the egg and eventually becomes the endosperm (i.e. the nutrition resource of the embryo), seems to be globally demethylated [Hsieh et al., 2009, Gehring et al., 2009], similar to the vegetative nucleus. Ibarra and colleagues found indications that siRNAs produced in the central cell can reinforce transposon silencing in the egg cell [Ibarra et al., 2012].

Even less is known about the transmission of histones through meiosis, but there are indications that the histone landscape is also profoundly reshaped. In the gametes, the H3 histone variants seem to be largely replaced [She and Baroux, 2014], and in the zygote the somatic H3 composition gets *de novo* synthesized [Ingouff et al., 2010].

Taken together, the restructuring processes that reset epigenetic marks ensure a seemingly robust propagation of the pluripotent epigenetic profile by a highly regulated developmental plan.

1.6 DNA methylation in mammals

To gain insight into the conservation of the methylation machinery and to understand the need for plant-specific methylome analyses, I briefly introduce the methylation machinery of mammals, including humans, and outline the most crucial differences compared to the plant system.

In contrast to seed plants, the vast majority of methylated sites are in a CG context in mammals, and non-CG methylation has only been found in low levels at transient developmental stages (germ cell progenitors, embryonic stem cells) or in neural cells [Lee et al., 2014]. Another major difference is that the vast majority (∼80%) of CG sites across all functional annotation classes are methylated, and only few regions in the genome remain largely unmethylated. These regions can be defined by a high density of CGs (>55%) compared to the general lack of CGs whole-genome wide (∼ 1%) and are referred to as CpG islands (CGIs) [Takai and Jones, 2002]. About half of them are located around transcription start sites [Long et al., 2013]; methylation around these sites presumably blocks transcription initiation [Jones, 2012]. The role

of inter- and intragenic CGIs is less well understood. CGIs are rarely methylated (3% [Maunakea et al., 2010]) and highly conserved across vertebrates [Long et al., 2013], but the methylation status of some of them is thought to change according to development or tissue [Jones, 2012].

The mammalian methylation maintenance system consists of a MET1 homolog, DNMT1 (DNA methyltransferase 1), with a preference for hemimethylated sites [Jones, 2012]. No CMT3 homolog has yet been found. The DRM2-homologous *de novo* DNA methyltransferases DNMT3A and DNMT3B form a complex with their cofactor DNMT3L to methylate CG sites [Law and Jacobsen, 2010]. Until now, there has been no evidence for siRNAs or an RdDM pathway in humans, but a similar class of 25-30 nt small RNAs, so-called PIWI-interacting RNAs (piRNAs), has been found conserved in mice, flies and worms [Castel and Martienssen, 2013]. Yet, piRNAs have only been found in animal germ lines, mediating H3K9 histone methylation and presumably also DNA methylation, but not in somatic cells [Castel and Martienssen, 2013].

In mammals, specific active demethylation mechanisms have been reported to occur either via deamination followed by DNA repair processes or via different intermediate forms of cytosine methylation, including 5-hydroxymethylcytosine, which is catalyzed via hydroxylation of the 5-methylcytosine by TET proteins (TEN-ELEVEN TRANSLOCATION) [Law and Jacobsen, 2010].

The developmental resetting also shows profound differences between the plant and mammalian systems. As mammalian germ lines are defined early even before meiosis and are highly protected from environmental impacts in the gonad anlagen, chances of experiencing epigenetic changes are low. Nevertheless, the developmental reprogramming of the chromatin occurs twice in mammalian individuals as opposed to only once in plants, namely during early germline development in primordial germ cells (PGCs), and after fertilization in the zygote [Lee et al., 2014, Kawashima and Berger, 2014]. Seemingly to prevent the inheritance of acquired chromatin changes even further, the extent of demethylation is more drastic than in plants, with whole-genome methylation levels dropping to ∼10% in PGCs and 30% in the zygote [Lee et al., 2014]. On the level of higher order chromatin restructuring, histones in sperm cells are almost entirely replaced by non-histone proteins, so-called protamines [Kawashima and Berger, 2014].

1.7 Sources of inter-individual epigenetic variation

Throughout the remaining work I will focus on DNA methylation, since this mark will be investigated in detail in the experimental part. Until now, I have introduced the general pattern of epigenetic modifications and mentioned its variability in the course of development, disease or as a response to environmental stress within a single individual. As the heritable DNA sequence accumulates differences between individuals over time, so does the heritable DNA methylation pattern. Changes of this epigenetic mark can have multiple causes. DNA methylation differences between individuals can be tightly linked to genetic variants, such that genetics directly controls methylation.

These cases were categorized as "obligate epialleles" sensu Richards [Richards, 2006]. At other loci, genetic differences might facilitate epigenetic changes in a probabilistic manner – those sites are referred to as "facilitated epialleles" [Richards, 2006]. Finally, "pure epialleles" [Richards, 2006] are independent of DNA mutations and can occur due to environmental signals or by putatively sheer coincidence. Lastly, individuals can be in different developmentally defined chromatin states, which needs to be accounted for in the experimental setup dealing with population studies.

In contrast to genetic mutations, which revert only extremely rarely, epialleles show a diverse range of stability over generations. Obligate epialleles are bound to stable genetic variants and are expected to be most faithfully maintained. Facilitated and spontaneously occurring epialleles show a varying range of stability across generations, and environmentally induced epivariation was commonly assumed to be transient and not transmissible through meiosis, but this view is challenged by recent studies suggesting that particular epigenetic changes can persist over subsequent generations, even in the absence of the epiallele-inducing signal.

The current state of the art in epigenetics is far from unambiguously telling the sources of epialleles apart, but for many cases there are fair indications that assign the reported mutation into one of the mentioned classes. The next sections will give an overview of the known epialleles from each source.

1.7.1 Genetically induced epialleles

The genomes of different Arabidopsis strains throughout Europe and Asia are rife with genetic differences [Cao et al., 2011, Gan et al., 2011, Long et al., 2013]. Similarly, comparative methylome studies in several plant species showed numerous single-site and regional DNA methylation differences between strains that were diverged for many thousands of years [Vaughn et al., 2007, Zhang et al., 2008, Eichten et al., 2011, He et al., 2013, Schmitz et al., 2013a, Schmitz et al., 2013b]. The first evidence that genetic events can cause DNA methylation changes that are also transmitted to following generations has been obtained by experimental studies in the late 20$^{\text{th}}$ century that artificially introduced transgenes into mammalian genomes (for review see [Daxinger and Whitelaw, 2012]). The animals were genetically identical to exclude confounding natural genetic effects. The transgene – once integrated into the genome – was methylated and silenced in following generations, although with varying transgenerational stability. Interestingly, the copy number of the transgene positively correlated with the degree of methylation and silencing [Garrick et al., 1998], suggesting an siRNA-dependent silencing, in which case more RNA template sequence leads to a higher siRNA production and increased DNA methylation. Later, introduction of viral DNA or transgenes in plant genomes equally invoked methylation changes, even at loci homologous to the sequence of the transgene [Jones et al., 2001, Kinoshita et al., 2007].

Specific other genetic events, namely the deletion of particular genomic parts [Bender and Fink, 1995] or the disruption of components of the methylation machinery in genetic knockdown mutants (*met1* or *ddm1*) [Teixeira et al., 2009] also

caused heritable local or global methylation differences, respectively. Even when the mutant genetic alleles segregated away in progeny, many loci remained in the altered epigenetic state, while others infrequently reverted back, many only gradually over several generations [Teixeira et al., 2009]. Some of these inherited DNA methylation changes invoked obvious phenotypes in re-established wild type background [Teixeira et al., 2009, Cortijo et al., 2014].

Besides the artificially induced alleles, numerous other studies have provided evidence that also naturally occurring genetic variants cause heritable epigenetic changes. The phenomenon of "paramutation" has been known for many decades. (for review see [Chandler and Stam, 2004]). It refers to the process that a 'paramutagenic' allele causes a change in gene expression or epigenetic state of a second, mostly remote genomic locus, the paramutable allele. The altered epigenetic state can persist in future generations and become paramutagenic even when the initial paramutagenic allele is no longer present. Such heritable trans effects are most likely mediated by siRNAs produced at the paramutagenic site and acting at the paramutable locus, although the regulative impact of transposable elements in *cis* could also contribute to silencing [Hollick, 2012]. Epigenetic mechanisms acting in *trans* are also assumed to underlie non-additive DNA methylation at some sister alleles or remote homologous alleles in hybrids originating from two different accessions of the same species [Greaves et al., 2012, Eichten et al., 2011, Chodavarapu et al., 2012, Shivaprasad et al., 2012], and can even cause hybrid incompatibility [Durand et al., 2012]. Finally, causal relationships between naturally occurring phenotypic variants and altered DNA methylation have been uncovered in tomato and melon, affecting nothing less crucial than the ripening and sex determination, respectively [Manning et al., 2006, Martin et al., 2009]. Since transposable elements have been found in the vicinity of the phenotype-causing genes, it is an attractive speculation that the transposon in *cis* gained methylation, which spread into nearby genes.

Seemingly unifying features of gained methylation linked to genetic changes are repetitive elements in the genome, including homologous sequences or (parts of) transposons. This supports the suggestion that a major role of DNA methylation is the protection of the genome from abundant transcription of the putative noncoding elements and from potentially harmful spreading of mobile genetic elements. An impressive study revealed that when a mobile element exceeded a specific copy number, it triggered its own methylation, and that this methylation was stably inherited to further generations even if the copy number dropped below the threshold again [Marí-Ordóñez et al., 2013]. Consistently, Cruz and Houseley [Cruz and Houseley, 2014] showed higher selectivity of a synthetic RNAi system to high-copy genomic loci in yeast, providing a link to the RdDM pathway.

Together, transposable elements and repetitive sequences seem to be prone for methylation, implying that their translocation or variation is a major source of DNA methylation changes in *cis* or *trans*.

1.7.2 Spontaneously occurring epialleles

During the last decades, a few studies have emerged that report naturally occurring epigenetic changes without any (known) link to genetic variants. The peloric variant of common toadflax mentioned in the prologue may be one example [Cubas et al., 1999], and different skin color patterns have been associated with variable DNA methylation in apple [Telias et al., 2011], pear [Wang et al., 2013] and maize [Cocciolone and Cone, 1993, Cocciolone et al., 2001]. Moreover, a recently evolved *de novo* gene in Arabidopsis showed methylation in 29 out of 36 analyzed accessions, seemingly independent of DNA polymorphisms nearby [Silveira et al., 2013]. In yellow mustard, a gene controlling erucic acid content exists in two different natural expression states [Zeng and Cheng, 2014], dependent on the stochastic methylation of a retrotransposon in the 5' untranslated region, which qualifies this epiallele to be called facilitated [Richards, 2006]. A similar metastable epiallele causes dwarfism in rice due to highly fluctuating methylation of a tandem repeat sequence near the affected DWARF1 (D1) locus [Miura et al., 2009].

Like obligate epialleles, the naturally occurring stochastic switches in methylation status outlined in this section can affect important traits like flower or plant shape and can be inherited across generations, albeit with frequent reversions [Becker and Weigel, 2012]. This is supported by a study that experimentally induced an aberrant whole-genome methylation pattern resulting in phenotypic changes, and both methylation differences and their associated phenotypes were propagated across generations [Cortijo et al., 2014].

The spontaneous occurrence of epialleles might be explained by incorrect DNA replication [Alabert and Groth, 2012] or the error-proneness of the DNA methylation machinery [Genereux et al., 2005, Fu et al., 2010], or by stochastic inter-individual variance in small RNA compositions with a varying potential to direct methylation [Schmitz et al., 2011, Teixeira et al., 2009]. Additionally, repetitive sequences have been identified in proximity to the causal DNA methylation changes in some cases. Despite the detected invariant DNA methylation pattern of these repeat regions, it blurs the border between pure and facilitated, or even yet undetected obligate epialleles. Whole-genome profiling of genetic variation and advanced association methods might be necessary to rule out connections between genetic and epigenetic variation.

1.7.3 Environmentally induced epialleles

Epigenetic regulation of stress responses

Plants, as sessile organisms, have to regularly cope with diverse unfavorable environmental conditions, which can disrupt their physiology and may often lead to reduced growth or fertility. They have developed multiple strategies to withstand adverse environments [Hirayama and Shinozaki, 2010, Hauser et al., 2011]. The classic random genetic mutagenesis followed by natural selection is responsible for permanent adaptations to differing environmental conditions [Atwell et al., 2010,

Fournier-Level et al., 2011], which can contribute to the long-term evolution of protective mechanisms to delay, alleviate, or tolerate the negative impacts of stress (stress avoidance). In parallel, plants have also developed short-lived mechanisms to quickly counteract fluctuating, unfavorable conditions during their lifetime (stress tolerance). Genetic mutations seem to be too slow, undirected and irreversible for the latter case, although rapid local bursts of genetic hypermutability in prokaryotes and human [Boyko and Kovalchuk, 2011, Rando and Verstrepen, 2007] and a global increase of somatic homologous recombination rate in plants [Molinier et al., 2006, Boyko et al., 2007, Boyko and Kovalchuk, 2010, Kathiria et al., 2010] have been reported. In most cases, however, the plastic and transient nature of epialleles seems to be more suitable for more controlled, immediate and reversible acclimations. In some studies, the elevated recombination frequency upon biotic and abiotic stresses is accompanied by seemingly whole-genome DNA methylation level differences [Boyko et al., 2007, Boyko and Kovalchuk, 2010], and other analyses report effects on whole-genome methylation as well [Verhoeven et al., 2010, Kovalchuk et al., 2003].

There are also numerous cases of targeted stress-induced effects altering the regulation of only some responsive genes, may it be upon drought, heat, cold, elevated salt levels, wounding, or exposure to hormonal signals or heavy metals [Iwasaki and Paszkowski, 2014, Pecinka and Mittelsten Scheid, 2012, Hauser et al., 2011, Gutzat and Mittelsten Scheid, 2012]. Since Barbara McClintock we know that stress can activate the transcription and movement of transposons in maize [McClintock, 1984], and more recent studies in Arabidopsis and other plants suggest the possibility that stress-adaptive transposons or other repeat sequences can affect the activity of nearby genes, mostly defense related genes in response to pathogen attacks [Ito et al., 2011, Dowen et al., 2012, Yu et al., 2013, Le et al., 2014]. In most of these cases, there are fair indications that DNA methylation (or rather demethylation) or histone modifications and replacements [Talbert and Henikoff, 2014, Iwasaki and Paszkowski, 2014] might be involved, may it be directly or downstream of the stress signaling pathway.

Besides these immediate, short-lived adaptations that do not persist much longer than the stressful period, numerous cases have been reported where plants appear to be capable of memorizing exposures to environmental conditions, sometimes lasting their lifetime. The most prominent case in plants is "vernalization". Exposure to cold induces lifelong chromatin changes in annual plants, and these are essential for the plants to flower [Baulcombe and Dean, 2014]. An "epigenetic memory" of a stress seems to consist of a primed chromatin configuration poised for rapid and more pronounced adaptive responses on future exposures to the same stress conditions, including DNA methylation [Pecinka and Mittelsten Scheid, 2012, Iwasaki and Paszkowski, 2014]. Mild cold "hardens" plants so that their freezing tolerance is increased upon future exposures to cold [Iwasaki and Paszkowski, 2014, Baulcombe and Dean, 2014]. Similar mechanisms are assumed to be involved in systemic acquired resistance (SAR) towards various pathogen attacks [Iwasaki and Paszkowski, 2014, Baulcombe and Dean, 2014].

Transgenerational epigenetic effects

The mentioned stress adaptations are thought to be transient and to be reset in each new generation. However, several recent studies raise the possibility that some stress-induced phenotypes can also persist in unstressed progeny. A higher recombination rate was observed to be transmissible to subsequent generations in Arabidopsis plants upon UV light or flagellin exposure [Molinier et al., 2006, Boyko and Kovalchuk, 2010] as well as in tobacco plants upon infection with tobacco mosaic virus [Kathiria et al., 2010, Boyko et al., 2007]. After priming of Arabidopsis plants against the pathogen *Pseudomonas syringae*, they developed an increased resistance, which was also found in unexposed progeny [Luna et al., 2012]. Some studies found consistent DNA methylation alterations in stressed parental and unstressed progeny plants in response to several biotic and abiotic factors [Boyko et al., 2007, Boyko and Kovalchuk, 2010, Verhoeven et al., 2010]. There are also some studies claiming such transgenerational effects in humans [Daxinger and Whitelaw, 2012]. For example, undernutrition of pregnant mothers or overfeeding of grandfathers in their slow growth period was found associated with an increased risk of impaired glucose tolerance of daughters [Lumey et al., 2009], and with a higher risk of cardiovascular disease and diabetes of grandsons, respectively [Kaati et al., 2002].

These observations imply that acquired traits might be heritably propagated across generations; a scenario which is not explicable by Mendelian inheritance. If such specific adaptations were stably inherited over many generations, this would imply a coupling of mutagenesis and natural selection, contrary to the Modern Synthesis dogma of undirected mutagenesis that is "blind to environmental cues" [Heard and Martienssen, 2014]. Therefore, these supposedly revolutionary findings raised much attention and made headlines in diverse literature.

Possible mechanisms of transgenerational epigenetic inheritance

Since epimutations can arise in a non-random directed manner and can be faithfully propagated through mitosis, the underlying mechanisms of such inheritance of acquired traits are thought to be epigenetic. To be transmissible through meiosis, epimutations must persist the resetting mechanisms to a 'default' unstressed chromatin state in the germ lines (section 1.5). However, possibilities for escape mechanisms might exist especially in plants, since the reprogramming is largely limited to sites in CHH context in sperm cells [Calarco et al., 2012]. Thus, most CG or CHG sites are not reset in each generation. Another source of heritable DNA methylation changes could be novel small RNAs produced by transposons or repetitive sequences in associated cells of the gametes, which could direct DNA methylation in the gametes at or around homologous genome sequences. If the siRNA-induced methylation pattern can be propagated by the maintenance system, or if the siRNAs get *de novo* synthesized in each generation, these marks could be stably transmitted across generations. Moreover, plant germ cells derive late in development from somatic cells, and these somatic cells might have accumulated epimutations during much of the plant's lifetime, which could potentially

increase the chances of transgenerationally transmitted epialleles. Even though the resetting is more pronounced in mammals than in plants (section 1.6), indications of a few hundred methylated loci persisting in the reprogramming process were found in mice [Borgel et al., 2010, Hackett et al., 2013].

However, when inspecting the studies claiming transgenerational epigenetic inheritance more closely, numerous questions remain unaddressed. The heritable effect was often observed in only a subset of offspring, and frequent reversions occurred, sometimes after only a few generations. A direct causal relationship between epigenetic modification and heritable trait has not been found in any study so far, and no analysis inspected the whole-genome genetic variation to rule out DNA sequence changes as the cause for heritability. Furthermore, the experimental study design might be sensitive to fluctuating epigenetic variation, as illustrated by a report that disapproved the transgenerational inheritance of an increased recombination rate for the same Arabidopsis strains used in the study of Molinier *et al.* [Pecinka et al., 2009]. Furthermore, parental effects on the progeny have to be considered. For example, egg as well as sperm cells are rich in small RNA molecules, which could repress or degrade transcripts in the zygote and might be responsible for altered gene expression states in at least the next generation. They would not impose lasting changes to the epigenome, since the siRNA levels get diminished with each new cell division, unless they direct heritable DNA methylation changes. Consequently, bona fide transgenerational epigenetic inheritance that can affect adaptive fitness can only be approved if the trait is inherited across two unstressed generations (or three generations in the case of stressed pregnant mammals) since the stress can directly change the germ cells (or additionally the germ cells of the mammalian embryos). Up to now, several reviews do not see any published data set entirely addressing these questions, and thus, there is no unambiguous case of transgenerational inheritance of a stress-induced phenotype based on epigenetic causes [Pecinka and Mittelsten Scheid, 2012, Daxinger and Whitelaw, 2012, Heard and Martienssen, 2014]. Maybe the closest causal link between a molecule and transgenerational inheritance via gametes was found in mice: two micro RNAs (miRNAs) are the likely trigger of white tail tips and feet, which can be transmitted to offspring [Rassoulzadegan et al., 2006, Daxinger and Whitelaw, 2012].

Taken together, while there seems to be transgenerational effects due to environmental cues in plants, it remains to be elucidated if epigenetic mechanisms are causal or can establish a different epiallele that is stable over many generations.

1.8 Contribution of this work to plant epigenetics

Epigenetic variants between individuals can arise due to diverse factors, like developmental signals, genetic events or environmental cues, or they can occur spontaneously (section 1.7). Independent of the sources, epigenetic changes can be transmitted across generations and evoke obvious phenotypes. Thus, like genetic variation, epigenetic variation contributes to the extensive phenotypic vari-

ability found within and across species. Environmental signals are known to induce directed changes to the epigenome (section 1.7.3), and there is a current debate whether such changes can also be invoked in the germ line and therefore stably inherited over generations [Danchin et al., 2011, Bonduriansky, 2012, Boyko and Kovalchuk, 2011, Pecinka et al., 2010, Paszkowski and Grossniklaus, 2011, Hirsch et al., 2012, Heard and Martienssen, 2014]. These epimutations are frequently assumed to be more often adaptive than DNA mutations [Tricker et al., 2012, Danchin et al., 2011, Bonduriansky, 2012, Boyko and Kovalchuk, 2011], and thus the inheritance of such 'acquired' traits would contradict the random mutagenesis assumption of the Modern Synthesis.

Until today, only a few naturally occurring epialleles have been reported (section 1.7.2), raising the yet unaddressed questions how frequent random epimutations are genome-wide and how large their impact on phenotypic variation is in comparison to DNA mutations. This knowledge can help in assessing the contributions of the different sources of epigenetic variation. Most experimental setups of studies exploring epigenetic diversity were not able to distinguish possible sources. Population studies of natural accessions had to account for all genetic, environmental and stochastic influences accumulated over thousands of years. Experiments conducted in the greenhouse usually were restricted in their analyses to a specific stress and to one or two generations only, which did not rule out parental effects being responsible for a heritable transmission. Additionally, many epigenome studies inspected only a subset of all genome-wide epimutations as well as DNA mutations due to technological or resource constraints, which limits the identification of potential causal relationships.

To better gauge the stochastic, environmental and genetic contribution to epigenetic variation, we inspected the DNA methylation landscape of two unique Arabidopsis populations. The first set of individuals consisted of ten genetically identical lines, each grown for 30 generations of inbreeding under rather stable conditions in the greenhouse. This setting largely eliminated genetic and environmental influences and allowed us to assess the spectrum and rate of spontaneous epimutations. The second population consisted of thirteen natural, near-isogenic Arabidopsis strains grown at geographically dispersed locations in North America, which had diverged from a last common ancestor at least a century ago. These analyses allowed us to estimate the effect of a natural, fluctuating environment on the heritable fraction of epigenetic variation in the absence of large-scale DNA mutations over a previously uncharted timescale of a few hundred years. By comparing both populations, we were able to explore whether epimutations accumulated at higher rate or showed a different spectrum under century-long exposure to natural conditions compared to under uniform and benign conditions in the greenhouse.

This study contributes to the current discussion on whether the environment has a durable or even genome-wide effect on the DNA methylation pattern in *Arabidopsis thaliana*.

Chapter 2

Next-generation sequencing and analysis of genetic and epigenetic variation

The Modern Synthesis assigns genetic variation to be the main driving force of evolution that creates the extensive phenotypic diversity within and between species. Since the complete sequencing of eukaryotic genomes at the beginning of this century ushered in the era of genomics, huge efforts are underway to profile global genetic diversity, including ambitious collaborative projects that aim to sequence thousand [1000 Genomes Project Consortium et al., 2010, Weigel and Mott, 2009] up to hundred thousand genomes[1]. Their unifying goal is to ultimately detect associations between common single-nucleotide polymorphisms (SNPs) and complex traits in the population [Visscher et al., 2012, Atwell et al., 2010]. Mainly due to technical difficulties in their identification, genomic regions of inserted, deleted or relocated sequence, so-called structural variants (SVs), have been underrepresented in these analyses. However, they might have an equally important impact on phenotypes, as demonstrated by many known associations with human disease [Weischenfeldt et al., 2013] and with a plethora of traits in plants [Saxena et al., 2014]. In addition, epigenetic effects have become acknowledged to play an influential role in shaping natural variation, as described in the previous chapter 1. By now, similar-sized consortia than for DNA sequence variation attempt to chart the epigenomes of thousands of tissues or individuals[2] [Roadmap Epigenomics Consortium et al., 2015].

Thus, the extensive and accurate identification of genetic and epigenetic variants is crucial to find the sources of phenotypic variability and to understand and trace evolutionary processes. With the revolutionizing progress in DNA sequencing technology in the past decade, we can now assess the genome-wide scope of genetic and

[1] http://www.genomicsengland.co.uk/, last accessed April 2015

[2] http://www.roadmapepigenomics.org, http://www.blueprint-epigenome.eu/, http://cancergenome.nih.gov, last accessed April 2015

epigenetic diversity.

In this chapter, I will outline the current state-of-the-art approaches in discovering whole-genome DNA sequence and DNA methylation variation in terms of technology and computational analysis. I start by introducing next-generation sequencing with a special focus on Illumina technology and then systematically describe the analytical workflow of a typical sequencing experiment based on Illumina short reads. I then outline the main characteristics and advancements of a novel complete framework to identify a diverse spectrum of genetic variants that combines the benefits of several genetic variation detection approaches and tools, which is described in detail in chapter 3.

I further give an overview of different methodologies in detecting DNA methylation and chronologically review the computational steps necessary to identify DNA methylation on a whole genome scale in individuals and differences between individuals using the state-of-the-art approach, bisulphite sequencing. Finally, I will explain the contribution of a complete pipeline to detect epigenetic variation that implements a novel approach to identify more unbiased regions of differential methylation compared to previous methods. The detailed mode of operation is presented in chapter 4.

2.1 DNA sequencing

Two of the many methodological revolutions in molecular biology in the past 50 years relate to DNA sequencing. Sanger's chain termination method, published in 1977 [Sanger et al., 1977], allowed easy and accurate deciphering of DNA sequences of up to more than 1000 bp in length and is nowadays considered the "first-generation sequencing" technology. Its steady improvements and increased automation spurred the decryption of the complete human and *Arabidopsis thaliana* genomes at the beginning of the 21st century [Arabidopsis Genome Initiative, 2000, Lander et al., 2001, Venter et al., 2001], ushering in the era of genomics. Since these ventures were tremendously laborious (more than 10 years, more than 200 researchers) and costly (more than US$1 billion for the human genome), huge investments, including an over US$200-million US government funding program, have been made to support developers of new sequencing technologies [Hayden, 2014]. These efforts have been extremely fruitful since many second-generation, also called next-generation sequencing (NGS) technologies, hit the market from 2005 on, soon after the human genome sequence was released. I present a brief overview of different technologies, and give in-depth explanations of the Illumina procedures as this work's underlying data derives solely from this company's sequencing instruments.

2.1.1 Next-generation sequencing

The major improvements of NGS methods, which led to their promotion to a new generation, include that they are independent of bacterial cloning or gel electrophoresis (i.e. their cell-free systems), and that they do not require prior knowledge about the genome of interest (e.g. primer design), as well as their immense increase in parallelization and throughput. While the sequence output of the early Illumina instrument amounted to 1 gigabase (Gb), thus already outperforming Sanger sequencing, they quickly raised the output by two, soon reaching almost three orders of magnitude to presently 750 Gb (HiSeq systems) in a single sequencing run, retrieved usually within days. Similarly, the cost tremendously dropped within only a few years, from the billion range to a few thousand US dollars or less for a whole human genome. Recently, Illumina announced it has now broken a projected long-term barrier, the sequencing of a human genome for US$1000, although this is only theoretically achieved by factory-scale whole-year usage of ten sequencing machines (HiSeq X 10) [van Dijk et al., 2014], of which the acquisition alone amounts to US$10 million.

The drastic decrease in sequencing costs enabled a myriad of "-Seq" studies and led to the emergence of many entire research fields whose new one-word denotations end on "-omics". To name a few, genomics performs whole genome sequencing (WG-Seq) of individuals or populations, metagenomics analyzes the DNA inventory of organisms in entire ecosystems, epigenomics investigates epigenetic marks, and transcriptomics uses RNA-Seq and explores the transcription landscape of organisms. This brings the exploration of a plethora of biological questions into reach for current researchers, the most widely used being gene expression (mRNA-Seq), expression of regulatory small RNA (mi/sRNA-Seq), ribosome profiling and the localization of transcription factor binding, histone modifications, nucleosomes (Chip-Seq) or whole-genome DNA methylation marks (WGBS-Seq), and the three-dimensional organization of the genome (HiC-Seq), and many more modifications of these protocols.

NGS has led to impressive insights, such as the discovery of 1000 times more genes present in the human gut than in the human genome [Arumugam et al., 2011], the decryption of the genome of ancient species such as the Neanderthal [Prüfer et al., 2014], and the identification of devastating pathogens in history [Bos et al., 2011, Yoshida et al., 2013]. However, likely the deepest influence to the lives of most humans in the near future will be the progress made in clinical diagnostics, which eventually will become routine with further drops in sequencing costs. This could pave the way for personalized medicine, i.e. to design treatment depending on genetic setup [Kingsmore and Saunders, 2011]. Many common and rare genetic disorders causing various diseases have already been identified, and testing fetuses for genetic aberrations has already been successfully performed non-invasively in prenatal stages, even by sequencing only the parents [Kitzman et al., 2012].

2.1.2 Next-generation sequencing platforms

Current NGS methodologies generally start with randomly breaking the input DNA into smaller fragments. Commonly, these template molecules are attached or hybridized onto a solid surface. The spatial distribution of the templates allows for thousands to billions of separate sequencing reactions being observed in parallel. The current technologies can be distinguished by whether they require a template amplification step to intensify detection signals or whether they directly sequence single DNA molecules.

Table 2.1: Next-generation sequencing platforms and selected characteristics. Modified and updated from [Buermans and den Dunnen, 2014].

Platform (Instrument)	Sequence by	Detection	Run time	Read len (bp)	output per run
454 (GS FLX Titanium XL+)	Synthesis	Pyrophosphate release	23 h	700	700 Mb
Illumina (GAII)	Synthesis	Fluorescence	14 days	(2x) 150	85-95 Gb
Illumina (MiSeq)	Synthesis	Fluorescence	56 h	(2x) 300	15 Gb
Illumina (HiSeq 3000)	Synthesis	Fluorescence	3.5 days	(2x) 150	650-750 Gb
SOLiD (5500xl W, 2 FlowChips)	Ligation	Fluorescence	10 days	75/(2x) 50	300 Gb
Ion Torrent (Proton, IonP1 chip)	Synthesis	Proton release	4 h	125	8-10 Gb
Pacific Biosciences (RSII, 16 SMRT cells)	Single molecule synthesis	Fluorescence	4 h	50% > 20 kb	8-16 Gb

There were four widely used template-amplifying strategies, of which three are still on the market. The first commercialized sequencing platform, released in 2005 by 454 Life Sciences, was discontinued in 2013. This technology belongs to the sequence-by-synthesis (SBS) approaches, i.e. relying on strand extension by a DNA polymerase. The method captured DNA molecules on beads, amplified the DNA on them in a process termed emulsion PCR, and deposited single beads with hundreds of clonal copies of the same DNA fragment into separate wells [Metzker, 2010]. The sequencing reaction in each well relied on pyrosequencing, which measures the proportionate release of pyrophosphates during synthesis of the complementary template strand by DNA polymerase upon separate additions of each deoxynucleoside triphosphate (dNTP). The intensity of the bioluminescent signal translates to the number of sequential identical bases in the template. This complicates the determination of the exact length of homopolymer stretches. While this method yielded rather long sequences ($\sim$ 700 bp), the throughput was much more limited in comparison with other technologies ($<$ 1Gb; Table 2.1). Thus, the costs could not be decreased to competing ranges.

Soon after the 454 sequencer was released, Solexa/Illumina (Illumina acquired Solexa in 2006) launched their SBS-based sequencer. Illumina constitutes the market leader, having sold the most instruments and offering the lowest cost per base as a

result of its unbeaten output (up to 750 Gb and read lengths up to 300 bp; Table 2.1). Their technique will be explained in-depth in the next section.

The third technology on the market was Sequencing by Oligo Ligation Detection (SOLiD) by Applied Biosystems (now Life Technologies), relying on a sequencing-by-ligation principle, i.e. using DNA ligase instead of DNA polymerase. Fluorescently labeled oligonucleotide probes are sequentially ligated to and cleaved from emulsion-PCR-amplified DNA. Each binucleotide of the probe is attached to one of four different dyes so that four out of all 16 binucleotide combinations share the same dye. Thus, each nucleotide is translated into two consecutive colors, and the template DNA sequence can be determined by aligning the "color-space" reads to a "color-space" reference. While the number of reads per run is comparable to Illumina, the runtime is slightly longer and the read lengths are only 100 bp, resulting in a lower output compared to Illumina instruments (Table 2.1).

The most recently released SBS method is Ion Torrent, hitting the market in 2010 (now Life Technologies). Emulsion PCR-amplified DNA is loaded onto micro-arrayed wells with pH sensors. Unlike the 454 or Illumina techniques measuring fluorescence, the Ion Torrent technology measures the change of pH upon extension of each nucleotide. It obtains the order of bases according to the presence or absence of pH changes across the wells when adding each nucleotide and the number of identical consecutive bases according to the strength of the signal (since there are no terminator agents as in the Illumina method). Thus, similar to the 454 technique, Ion Torrent is strongly prone to erroneously called homopolymer stretches, but due to the saving of the imaging time it finishes sequencing runs within hours and not days. However, the throughput lags behind Illumina throughput by more than an order of magnitude [Buermans and den Dunnen, 2014] (Table 2.1).

Second-generation sequencing technologies rely on the amplification of DNA to intensify the signal to detect fluorescence or pH changes. This process is afflicted with biases against specific sequence compositions (e.g. AT- and GC-rich parts of genomes are underrepresented) and may introduce errors in the synthesis of copies. Furthermore, it can distort relative abundance measures in quantitative RNA or methylation level analyses. The first 'single molecule sequencer' that does not rely on any DNA amplification was distributed by Helicos Biosciences and performed similar sequencing steps to Illumina, but could only compete for three years (2009-2012) and is no longer available [van Dijk et al., 2014]. However, because of the combination of single molecule sequencing and detecting nucleotides in real time, the technology of Pacific Biosciences (PacBio) is now generally considered the first method assigned to 'third-generation sequencing'. Instead of fixing DNA adapters for capturing sequencing templates on a solid support, a single DNA polymerase is immobilized at the bottom of a well-like container, termed zero-mode wave-guide (ZMW). There are presently 150,000 ZMWs per sequencing run on a single molecule real-time (SMRT) cell. The synthesis of freely floating dye-labeled nucleotides into a single template strand in each ZMW is recorded by the fluorescence pulses in real time at 75 frames per second.

The main advantage of this technology is a long read length, of up to several tens of thousand base pairs, and currently achieving an average of 8-10 kb by improved chemistry (e.g. by better shielding polymerases from photo-damage). This method does not rely on separate steps to cleave dyes off the last added nucleotide; the fluorescence labels are removed during strand extension. Due to the lack of these reactions, sequencing of long reads is performed within a few hours (Table 2.1). However, the low throughput (lower than 1Gb) and accuracy with single-base error rates as high as 10-15% have room for improvement. Despite these shortcomings, the long read data and the lack of GC bias will allow PacBio to fill a unique niche for specific applications, foremost in facilitating the application and completion of *de novo* assemblies.

Other third-generation sequencing technologies are on the cutting edge of recent research. They measure changes of electric current or optical signals during the transit of a DNA molecule through protein or solid-state nanopores [Wang et al., 2014]. First experiences have been recently reported using hand-held sequencers [Mikheyev and Tin, 2014], still showing major weaknesses, but it is nonetheless possible that they could spearhead the next leap forward.

2.1.3 Illumina's sequence-by-synthesis technology

The procedure for Illumina sequencing can be summarized into three major steps: library preparation, cluster generation and sequencing by synthesis. To generate the library, the input genomic DNA has to be randomly fragmented into smaller double stranded DNA (dsDNA) with single stranded overhangs. Usually, only fragments of a few hundred base pairs in length are retained using gel electrophoresis. The length distribution is known as the "insert size" distribution of the library. After universal double stranded adapter sequences are ligated to the dsDNA, the library is spotted on a glass slide, called a "flow cell", with millions of both forward and reverse adapter sequences immobilized to it (Figure **2.1**A-E).

Cluster generation is initiated by the hybridization of denatured template DNA with its adapter sequences to the fixated complementary adapter oligonucleotides. In a process called bridge amplification, the single stranded protruding adapters of the DNA molecules hybridize to their nearby immobilized complements, shaping a bridge-like structure (Figure **2.1**F). Starting from a universal primer sequence in the adapters, the complementary strand of the DNA fragment is synthesized by DNA polymerases. Through repeated denaturating, bridging to nearby adapters and polymerization to dsDNA, one initial DNA molecule leads to a dense cluster of usually hundreds of clonal copies surrounding it (Figure **2.1**G). The latest instruments released in 2015 have billions of nanowells on the glass slide, each containing monoclonal clusters originating from a single template DNA fragment. This improves runtime and throughput by facilitating the identification of clusters and preventing overlapping clusters.

After annealing a common primer to the protruding adapters, the sequencing of each cluster of DNA fragments begins (Figure **2.1**H). In each sequencing 'cycle', the four dNTPs, each reversibly labeled by a base-specific fluorescent dye and bound to

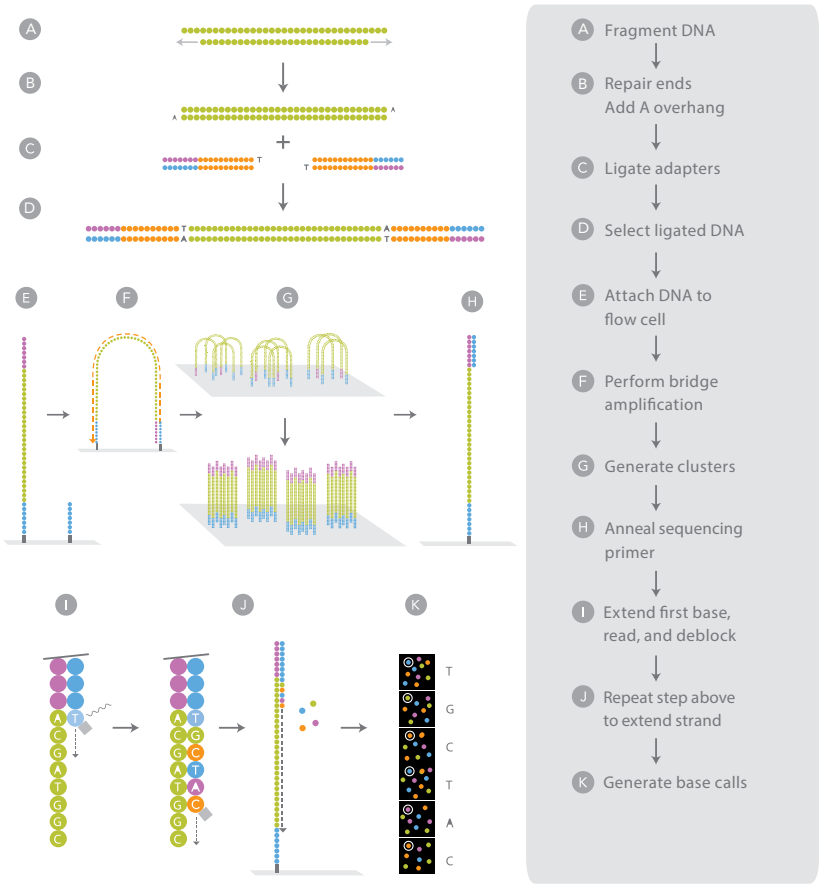

Figure 2.1: Workflow of the Illumina sequencing-by-synthesis approach. See more detailed explanations in text. Modified from Illumina Genome Analyzer brochure (http://support.illumina.com/content/dam/illumina-marketing/documents/products/brochures/brochure_genome_analyzer.pdf, last accessed March 2015).

a terminator molecule, are added to the flow cell. DNA polymerases add the respective nucleotides to complementary bases of the input DNA at the 3' end of the newly synthesized strands, while the attached terminator molecule ensures that only a single nucleotide is incorporated by preventing further strand elongation (Figure **2.1**I, J). Lasers excite the different newly added fluorophores, and CCD chips image the characteristic signals that identify the incorporated base (Figure **2.1**K). Due to the DNA amplification into clusters, the intensified signal of the whole cluster facilitates the correct dye identification. Before the addition of the next nucleotide, the terminating labels are enzymatically removed. Thus, each cycle consists of the addition of exactly one nucleotide to each DNA molecule.

Provided analysis software then performs cluster detection, noise reduction and base calling. Each step represents a source for errors, shaping the characteristics of the

output of the sequencing machine, the "reads", and I will briefly outline these properties in the next section, which determine the choice of downstream analysis methods.

2.1.4 Properties of Illumina read data

Typically, the sequencing reagents deteriorate with each cycle, limiting the length of the reads to around 100-150 bp in high throughput machines (HiSeqs). The range of read lengths has improved from 36 bp in the early Genome Analyzer machines to 300 bp in lower throughput MiSeq instruments.

Sequencing information from one DNA molecule can be increased by sequencing both ends of the input DNA fragments, achieved by two different sequencing primers matching to the different strands of the adapter sequence. This is referred to as paired-end (PE) sequencing. The main advantage of paired-end reads is that they provide linkage information over larger genomic distance, which can facilitate to span or anchor repetitive sequences in the genome.

The detection of the fluorescent dye can be ambiguous, first because many chemical steps are performed, making the method susceptible to incomplete reactions, improper dNTP incorporation or dNTP carry over from previous cycles. Therefore, different DNA molecules in a cluster might be out of phase, i.e. they do not contain the same base at a specific time, weakening the overall signal of the cluster. Second, the optical system is sensitive to inclusion of air or water droplets, or might not focus correctly at the borders of the flow cell. Third, clusters can overlap in older instruments (shipped before 2015), blurring the distinction of the respective DNA molecules. Thus, even though there are on average hundreds of molecules in each cluster, DNA sequences can contain ambiguous base calls, represented by the base 'N', when the base calling procedure cannot identify a unique base call. Additionally, even false base calls, so-called 'sequencing errors' can occur despite having marginal frequencies, below 0.5% per base in recent instruments. Single base pair substitutions outnumber inserted or deleted bases. To account for dubious calls, the Illumina base calling software calculates a quality value for each base as a measure of its reliability. This 'phred' value relates logarithmically to the base-calling error probability P:

$$Q = -10 \log_{10} P$$

Thus, a phred score of 10 (Q10) implies an error probability of 10%, Q20 translates to 1%, Q30 to 0.01% and so on. Due to deteriorating reagents over time, the base qualities and thus the reliability of the nucleotides decrease towards the read ends, and they are typically lower in second reads of paired-end sequencing runs compared to first reads.

The dependency on DNA amplification imposes other biases relevant for downstream *in silico* analyses. AT- and GC-rich genomic regions are underrepresented in the sequencing reads resulting in intrinsic genomic coverage fluctuations. This requires cautious interpretation, especially for quantitative analyses interested in representa-

tional abundance (e.g. RNA or methylation levels), since the bias can lead to a deviation of the observed to the true frequency or proportion of molecules. Furthermore, individual measurements might not be independent if the same input DNA fragment was sequenced multiple times (called "PCR duplicates").

Together, the short Illumina reads feature sequencing errors, low quality ends of reads, and their sampling underlies biases, which subsequent analyses have to account for.

2.2 Analysis of short NGS data

2.2.1 Pre-processing reads

The first step of any next-generation sequencing read analysis should be the filtering of low quality reads or bases to reduce the impact of such unreliable base information on any final analysis and to speed up further processes. Usually, filtering includes discarding reads containing many low quality or ambiguous bases and reads exhibiting low sequence complexity. Furthermore, potential barcode or adapter sequences and low quality ends of reads are removed.

Diverse available software packages perform these tasks, among them FASTX [Gordon, 2010], NGS QC toolkit [Patel and Jain, 2012], or SHORE [Ossowski et al., 2008]. Quality control software like FastQC [Andrews, 2012] can be used to spot and visualize possible problems or biases, and to help in finding suitable parameters for quality filtering.

2.2.2 *De novo* genome assembly

Possibly, the most intuitive use of the millions of next-generation sequencing reads is to place them one after another based on their matching overlapping parts. However, the short read lengths from most current high-throughput platforms prevent a unique ordering when sequences can be extended by different reads, resulting in the termination of the sequence, commonly referred to as an 'assembly gap'. To span repetitive regions, which constitute a high portion of most genomes of higher organisms and which can be several tens of kilobases in length, long-range linkage information is required. Paired-end reads (300-500 bp insert sizes), mate-pair reads (2-8 kbp), or fosmid ends (35kbp) can provide this information to some extent, but their generation is still costly and resource-consuming. Recently, long reads from different platforms, e.g. PacBio [Koren et al., 2012], or information about the genomic distance of DNA by proximity ligation based methods like Hi-C [Burton et al., 2013, Kaplan and Dekker, 2013], aid in connecting unlinked sequence pieces (scaffolding) and closing assembly gaps.

Although reference-quality genome assemblies currently seem to be within reach, finalizing assemblies relying mainly on short NGS reads remains a largely unsolved task and demands high sequencing depth and, thus, high costs. Whole-genome assembly [Gan et al., 2011] or local assembly of targeted genomic regions [Ossowski et al., 2008]

is often used in combination with the whole-genome 'resequencing' method described in the following.

2.2.3 Resequencing

To circumvent the costs and the computational challenge of *de novo* assemblies, a different strategy has been widely applied since the advent of NGS technology. For most model organisms, high-quality genome sequences have been made available and can serve as a reference to which newly sequenced individuals can be compared. This has the advantages of detecting genetic variants in a consistent coordinate system and allows adopting the reference genome's gene annotations.

In this process, termed 'resequencing', the first computational challenge is to find the location of a short query sequence in a large target genome where it most likely derived from ('read mapping'). Since the NGS reads commonly feature differences to the target genome due to genetic diversity, sequencing errors, or incorrect reference base calls, algorithms have to account for 'mismatching' bases between reads and the genome and seek to find the 'optimal alignment', i.e. with the least number of unmatched bases. This string-matching problem is an algorithmic problem of bioinformatics since the emergence of the field. Classic alignment algorithms such as Smith-Waterman or Needleman-Wunsch are impractical for large target sequences, and the widely used basic local alignment search tool (BLAST) [Altschul et al., 1990] is infeasible for millions of query sequences. Thus, short read alignment (or short read mapping) tools have been developed that efficiently place millions of short reads to their most likely location on large genome sequences, from which unmatched bases can be determined [Shang et al., 2014].

Their high performance is mainly achieved by building an index of all sequences of a given short length k of the genome (k-mers, or seeds), in which the location of seeds of the reads can be instantly looked up. Since this is most efficient if the k-mers perfectly match, k is chosen to be smaller than the read lengths, usually between 12-32 bp, to allow for mismatches in-between matching k-mers. Concordant seeds between read and reference can be merged with overlapping or adjacent matching seeds. This ultimately defines a set of possible locations ('hits') to which the whole read sequence can be aligned using computationally expensive alignment algorithms (e.g. using Needleman-Wunsch). By modulating the size of k, the number of k-mers considered per read, and the number of mismatches allowed within k-mers ("spaced seeds"), the trade-off between the number of genomic locations to be analyzed per read and the mapping accuracy can be regulated and adapted to particular applications.

There are numerous features and settings associated with the mapping process used in different tools. GenomeMapper allows for a large and user-defined number of mismatches and gaps, and can incorporate variation of several genomes into the reference sequence [Schneeberger et al., 2009]. The most widely used short read aligners utilize a compressed suffix array as the index, based on Burrows-Wheeler transformation [Li and Durbin, 2009, Langmead and Salzberg, 2012]. This reduces the memory foot-

print of genomes to their actual size in bytes, while non-compressed indices occupy multiples of the genome sizes in memory. Most current mapping tools penalize reads that map to several locations in the reference genome with the same amount of mismatches by decreasing a "mapping quality" score. Furthermore, mapping accuracy can be increased by accounting for base qualities, favorably placing mismatches on low quality bases, and by preferentially mapping both pairs of paired-end reads in a distance to each other according to the insert size of the sequencing library.

2.2.4 Genotype calling at genomic positions

Following the alignment of all reads to the reference genome, the DNA sequence can be reconstructed based on the frequency of bases contained in the reads overlapping each genomic position. There are different variant calling methods that rely on empirical thresholds or on probabilistic models and identify single nucleotide polymorphisms (SNPs) or inserted or deleted bases ('indels') compared to the reference sequence. Obviously, it is imperative to have sequenced many independent reads per position to distinguish sequencing errors from real genetic variants, or to reliably separate homozygous from heterozygous sites.

The alignment of short reads imposes several biases, which have to be accounted for in variant calling. The base calls of reads mapping to multiple genomic regions cannot be uniquely assigned to one instance of the repeated sequence. This is why reliable variant calling is commonly restricted to the uniquely mapping reads. Genomic segments that are only represented once in the reference sequence but multiple times in the read set, aggregate the reads from all repeat instances, potentially harboring pseudo heterozygote calls. Furthermore, cross-mappings can occur if sequences not covered in the reference are similar to other parts in the reference genome allowing an alignment within the mismatch limits. Although such regions can be theoretically detected by an increase in read coverage, in practice it is difficult to distinguish increased from average coverage, e.g. due to the confounding GC bias.

For performance reasons, short read mapping tools allow for a limited amount of differences between read and reference and therefore cannot detect longer indels or highly diverged regions. However, reads that overlap with diverged regions by only a few bases might align, but they typically exhibit many mismatching bases at their ends overlapping the indel (Figure **2.2**). These alignment artifacts likely lead to false positive SNP or small indel calls.

This is why genetic variant detection tools like SHORE [Ossowski et al., 2008] reduce the contribution of a specific number of bases at each read's end to variant calling and require a specific coverage of the 'core' region of reads at each variant. Other approaches try to identify and re-align affected reads to minimize the number of unmatched bases [DePristo et al., 2011, Li and Durbin, 2009].

Diverse tools for genetic variant calling have been developed to reliably call genotypes, e.g. by setting empirical filters for many criteria like base qualities, the number of mapping locations of reads or incorrect read alignments [Ossowski et al., 2008], by

Figure 2.2: Example of incorrect alignments of reads. Reads overlap with their ends into a diverged region harboring a 7-bp deletion. The alignments on top are the result of common short read mapping tools allowing for a limited number of mismatches or gaps, the correct alignments are shown at the bottom.

using statistical models [Li and Durbin, 2009, DePristo et al., 2011], or by calling haplotypes, thereby avoiding alignment artifacts [Garrison and Gabor, 2012].

2.2.5 Structural variation calling

The resequencing method can only provide genotype calls in largely conserved genomic regions that maximally show the level of divergence allowed by the read mapping. However, by exploiting alignment-related features like read coverage or insert size distribution, or spurred by methodological advancements mainly based on longer read lengths, a multitude of NGS-based tools have been developed that seek to find more diverged sequences that are inaccessible to the resequencing approach. They are used to detect structural variants (SVs) that encompass insertions, deletions, duplications, inversions or translocations. There are mainly four approaches to detect SVs based on next-generation sequencing [Medvedev et al., 2009, Alkan et al., 2011].

Methods based on depth of coverage model a genome-wide sequencing coverage, e.g. by a Poisson distribution, and segment the genome into regions of significantly reduced or elevated read depth, indicating deletions or duplications, respectively [Campbell et al., 2008, Abyzov et al., 2011]. Approaches based on paired-end mapping detect clusters of reads having a much shorter (insertion) or longer (deletion) insert size compared to the genome-wide average, modeled by a normal distribution [Cao et al., 2011, Rausch et al., 2012]. However, both read-depth and paired-end based approaches have difficulties in identifying short SVs (approximately shorter than 50 bp), and cannot pinpoint exact SV breakpoints at single nucleotide resolution. Facilitated by the technological progress in generating longer read lengths, newer approaches use local *de novo* assembly [Li et al., 2013] or map the right and left parts of reads separately, thereby bridging potential SVs ('split read' approaches) [Ye et al., 2009, Grimm et al., 2013]. These methods provide the exact location of SVs

to the single base pair level. Assembly-based tools are able to detect highly divergent regions consisting of multiple nearby variants, as well as insertions longer than the read length, which is impossible for split-read approaches, but they require high coverage.

Methods for consolidation of SV calls

Despite the progress made in SV detection methods, the application of different methods to identical data sets reveal little overlap, and false positive rates commonly exceed 10% [Mills et al., 2011], mainly due to limited sequencing depth or the repetitiveness of SV regions. The approaches show noticeable false negative rates as well, since each of the methods has limitations in terms of type and size of SVs they can detect, and no single tool can uncover the full range of structural variants.

Therefore, methods have been developed to exploit more than one approach in a single caller (e.g. [Rausch et al., 2012]), or to integrate calls from different SV detection tools (e.g. [Wong et al., 2010]). Comparing SVs across tools or samples is complicated by the fact that the same sequence can be aligned in various ways with different types or different amounts of variants (see Figure **3.4**). Simply defining SVs with the same genomic coordinates as being identical would discard many true positive calls, while simply collapsing overlapping SVs into shared SVs will increase false positive calls, especially in complex regions harboring several variants. Thus, while the limitation of single variation callers to specific types or lengths of SVs necessitates the application of several different tools to maximize genetic variant detection, the merging of the different calls is non-trivial [Lin et al., 2014].

There has not been any agreement on a standard way of integrating several SV callers, not even in the active field of human genomic research [Lin et al., 2014]. A few methods have been developed that either report the union set of SVs based on genomic overlap (SVmerge [Wong et al., 2010], iSVP [Mimori et al., 2013]) or only report intersecting SVs found by more than one tool (HugeSeq [Lam et al., 2012], intansv [Yao, 2014]). While SVmerge can integrate an arbitrary selection of tools, the other methods use a predetermined set of SV callers. However, each tool exhibits restrictions on the SV set. For example, SVmerge only reports SVs larger than 100bp and iSVP only reports deletions.

Advanced approaches for variation calling

Besides the consolidation of different genetic variant detection approaches and callers, another way to increase both specificity and sensitivity of polymorphism detection is to use the information of a population of samples. Calling variants in the consolidated data of the whole population increases read coverage and the ability to detect low frequency variants. Alternatively, during the short read mapping or the genotyping process, previously known variants of a population or of the species can be added and validated [Schneeberger et al., 2009], and there are databases of known polymorphisms providing extensive genetic variation (e.g., dbSNP [Sherry et al., 2001]). Furthermore, to access more complex regions harboring many variants, the resequencing approach

can be iteratively repeated after incorporating detected variants into a pseudo reference sequence [Gan et al., 2011].

2.2.6 Objective of this work: An integrated method to detect genetic variation

The main driving force of phenotypic diversity is variation in the DNA sequence, and a central goal of genetics research is to identify genetic causes of diseases or phenotypic traits. The advent of next-generation sequencing around a decade ago fueled the identification of genetic variants on a genome-wide level. Typically relying on short NGS reads and on the resequencing approach, these analyses restricted detected genetic variants mainly to SNPs and the location of variants to the rather conserved parts of the genome compared to a reference sequence. However, highly divergent regions and larger structural variants are assumed to be as frequent as SNPs and likely play a similarly important role for phenotypic variability. While reliable SNP calling is nowadays commonly considered a routine task, there is no agreement on how to best profile structural variants yet. Advancements in sequencing technology leading to longer read lengths improved *de novo* assemblies and spurred the development of a plethora of different SV identification tools in the last years. However, SV callers report only a limited subset of existing SVs while also exhibiting high false positive rates [Mills et al., 2011]. As a result, outputs of different tools show little overlap between each other [Mills et al., 2011, Lin et al., 2014], which necessitates an integrated use of diverse approaches to identify as much genetic variation as possible.

To this end, I propose a pipeline in chapter 3 that combines different approaches to call diverse classes of genetic variants by using resequencing, *de novo* assembly and several structural variation detection tools. At the same time, the variants are stringently evaluated based on re-alignment of reads to limit the false positive rate of variant detection.

The proposed strategy of consolidating predicted SVs from different callers improves existing SV merging approaches in several aspects. While most tools restrict SV detection on specific subsets of SVs, the method introduced here covers a large range of variant types and lengths. Additionally, it consolidates variants by comparing haplotype sequences rather than relying on identical or overlapping genomic coordinates of polymorphisms, which can vary from tool to tool, especially for SVs. Contrary to some other SV merging tools, my method does not rely on the intersection between SV callers and retains tool-specific SVs that are not contradicted by different sources, which increases sensitivity.

When handling multiple, closely related samples, the method exploits information of all samples for the detection of SVs in individuals, thereby reducing the false negative detection rate and compensating for the high false negative rate of SV callers. Finally, the strategy to detect genetic variation can be employed iteratively, thereby generating and increasingly refining a pseudo reference sequence, built from one or

multiple samples. Such a procedure facilitates the detection of complex variation and serves as an additional validation step of detected SNPs and SVs of each iteration.

Together, the strategy proposed in chapter 3 aims at maximizing the genetic variation detectable by short next-generation sequencing data that is present in individual samples or in a set of closely related samples, while at the same time rigorously validating the predicted variants to retain a high level of accuracy.

2.3 DNA methylation sequencing

One of the many different applications of next-generation sequencing is the detection of DNA methylation. The renewed interest in epigenetics in the last decade might have been closely linked to the technological advancement and cost reduction of next-generation sequencing. This led to the availability of information about millions of individual cytosines, which challenges the computational and statistical detection of differential DNA methylation.

In the following section, I first describe widely used experimental techniques to detect DNA methylation and then outline the common practices in computationally identifying this epigenetic mark using next-generation sequencing data, before reviewing diverse computational approaches to detect DNA methylation differences between samples. I focus on analytical methods applied in recent plant studies, but also discuss recently launched software solutions that were geared towards human samples and might become the state-of-the-art in detecting differential methylation based on whole-genome bisulphite sequencing (WGBS-Seq) data. This section ends by outlining a novel approach for plant WGBS-Seq data to identify regional differences in DNA methylation that is based on the strategy of the currently available tools for human data. The new method is described in detail in chapter 4.

2.3.1 Experimental methods to detect DNA methylation

Since standard molecular biology methods, such as polymerase chain reaction (PCR), erase DNA methylation marks, and since hybridization to microarrays and (next-generation) sequencing are insensitive to methylated bases, researchers developed several treatments for DNA that would allow detection of this epigenetic mark. These include three main approaches: methylation-specific enzyme digestion, affinity enrichment and chemical treatment with sodium bisulphite [Laird, 2010]. Subsequent analyses can then be performed by probe hybridization or (next-generation) sequencing methods to reveal the locations of methylated cytosines. The combination of these three approaches with different downstream analysis methods led to a multitude of techniques to detect DNA methylation [Plongthongkum et al., 2014]. Each methodology has specific advantages and limitations regarding cost, resolution, coverage, scalability and the amount of input DNA, and I will now review some of these techniques in more detail (Table 2.2).

Table 2.2: Established experimental approaches to detect DNA methylation and their characteristics. Input: amount of input DNA, Ref: need for reference sequence?

Approach	Cost	Resolution	Genome coverage	Scalability	Input	Ref
Enzyme digestion	low	1 bp	low	low	med	no
Affinity enrichment	med	$\gg$1 bp	high	med	high	yes
Bisulphite treatment	high	1 bp	near-complete	high	low	yes

Enzyme digestion methods

The earliest methods used restriction enzymes that cleaved DNA at characteristic motifs dependent on the methylation status of single cytosines therein [Laird, 2010]. For example, DNA methylation can be retrieved by comparing the different fragment sizes generated by two (or more) enzymes targeting the same motif, but showing different dependencies on the methylation status (such enzymes are termed "isoschizomers"). The first DNA methylation analyses, around 1980, separated digested fragments by gel electrophoresis followed by Southern blot hybridization [Laird, 2010]. Later, these fragments were amplified by PCR and today they are sequenced by array-based or next-generation sequencing methods, which have gradually opened the door for feasible analysis up to the whole genome level. These inexpensive methods do not require knowledge of the genome sequence of the analyzed species, making them suitable for non-model organisms or population-scale analyses. However, while they exhibit a resolution down to specific individual cytosines, they only uncover the methylation status of a single site within each recognition motif and are blind to the vast majority of cytosines in the genome, even if they lie within or close to a motif. Furthermore, the input DNA needs to be rather pure and abundant.

The comparison of isoschizomer-digested fragments is still used today. Several studies have used a derivative of this approach, termed methylation-specific amplification polymorphism (MSAP) (e.g., [Reyna-López et al., 1997, Lira-Medeiros et al., 2010, Medrano et al., 2014]). However, methylation information was typically obtained only for a non-representative set of a few hundred cytosines at most and the interpretation of MSAP data is not always consistent [Schulz et al., 2013] and associated with a series of potential caveats [Fulneček and Kovařík, 2014].

Affinity enrichment methods

Affinity enrichment methods represent a second strategy for detecting methylation. The underlying idea is to solely purify methylated DNA that can be analyzed using array-based or sequencing-based techniques. Purification is performed by immunoprecipitation of denatured genomic DNA with either methyl-binding proteins (methyl-CpG-binding domain protein sequencing; MBD) or antibodies having affinity to methylated DNA (methylation DNA immunoprecipitation sequencing; MeDIP).

In combination with next-generation sequencing, widely used protocols are MBD-Seq [Serre et al., 2010] and MeDIP-Seq [Down et al., 2008], respectively. While these methods allow for efficient detection of DNA methylation on large parts of the genome (60-90% CpG coverage in human/mouse [Plongthongkum et al., 2014]), they miss regions of sparsely distributed methylated sites or regions of low GC content. Furthermore, their resolution is not good enough for detection at the single cytosine level and they do not report a quantitative measure of methylation. Results obtained from these methods can only be interpreted as the relative frequency of methylated cytosines in genomic windows (or peaks) of variable lengths and this requires a bioinformatic adjustment for varying methylation density across regions or samples.

Chemical treatment with sodium bisulphite

The biggest leap forward in the detection of DNA methylation happened in the early 1990s by pretreating denatured DNA with sodium bisulphite [Frommer et al., 1992]. Bisulphite converts unmethylated cytosines to uracil, which is amplified to thymines in the PCR [Clark et al., 2006]. Thus, a difficult to detect epigenetic difference is transformed into an easily accessible genetic variant (Figure **2.3**).

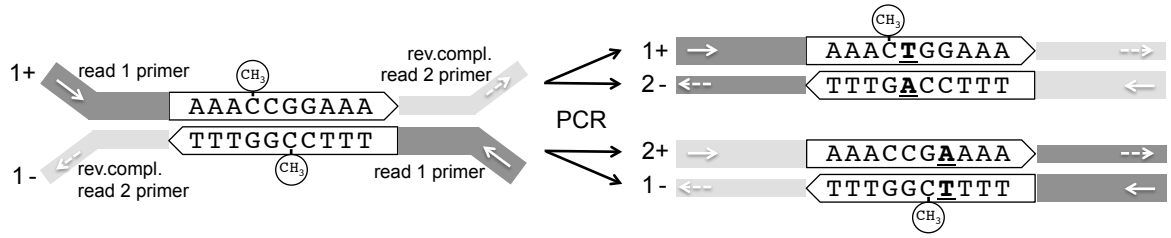

Figure 2.3: Illustration of bisulphite conversion. Shown are the possible combinations of derived strand and sequencing read of WGBS-Seq reads. Left: Double stranded DNA fragment with common adapters (grey Y shapes) harboring two different sequencing primers in paired-end mode. Right: After bisulphite treatment, unmethylated cytosines and their complementary guanines are replaced by thymines and adenines (bold and underlined bases), respectively, in the original bisulphite-treated fragments (1+ and 1–). Those reads are sequenced as reads 1 in paired-end mode. PCR amplification generates reverse complementary sequences (2+ and 2–) of the original bisulphite-treated fragments that are sequenced as reads 2 in paired-end mode.

First strategies using Sanger sequencing of bisulphite-converted DNA were expensive and hardly scalable. Nevertheless, huge efforts have been undertaken in charting the methylomes of three human chromosomes using this method [Eckhardt et al., 2006]. This in general low-throughput approach still constitutes a common and accurate method to validate single loci [Mensaert et al., 2014]. The use of bisulphite-treated DNA with arrays or NGS increased the throughput by several orders of magnitude. Since bisulphite reduces the sequence complexity to three bases, methylation detection using microarrays requires special array design or an

increased mismatch tolerance, which decreases hybridization efficiency. Thus, next-generation sequencing of bisulphite-treated DNA has emerged as the method of choice for high-throughput detection of methylation. Pioneering work in *Arabidopsis thaliana* and humans [Cokus et al., 2008, Lister et al., 2008, Lister et al., 2009] applied WGBS-Seq to generate precise and almost complete maps of DNA methylation. Combined with steadily decreasing sequencing costs, this prompted WGBS-Seq analyses in other species as well (e.g., [Li et al., 2012]).

Compared to previously discussed methods, WGBS-Seq offers improved resolution (single base pair), higher coverage (>90% of human genome), higher scalability (deep coverages and multiplexing numerous samples possible) and it can be used with small amounts of input DNA (down to 10 nanograms) [Plongthongkum et al., 2014] (Table 2.2). Additionally, WGBS-Seq provides quantification by digital counts, which allows the consolidation of data across sequencing libraries or even across different studies, assuming low batch effects. Yet, there are several drawbacks. The cost is still high, and bisulphite conversion can be incomplete, leading to false methylation calls. Lastly, bisulphite-treated samples are prone to considerable DNA damage or sample loss, requiring more amplification steps than typical DNA sequencing.

To reduce the cost of sequencing at the expense of coverage, numerous enrichment methods have been established (see [Laird, 2010, Plongthongkum et al., 2014]). They can be divided into non-targeted methods that randomly sequence fragments and targeted approaches, where genomic regions of interest can be pre-selected. These methods yield only a limited view on the methylome, justified by the fact that large parts of the genome of most organisms are unmethylated. For example, the non-targeted approach 'reduced representation bisulphite sequencing' (RRBS) restricts sequencing to only CG-rich and CG-dense genomic regions by digesting DNA with a methylation-insensitive restriction enzyme cutting the sequence CCGG, followed by fragment size selection [Meissner et al., 2008]. The maximal coverage of a human genome that has been achieved so far with this approach was nearly 20% in a study that used two different restriction enzymes [Wang et al., 2013]. In contrast, targeted approaches can be performed by PCR amplification of bisulphite-treated DNA, but it is difficult to scale and needs input DNA in the microgram range. Several methodologies capture target sequences by hybridizing to oligonucleotide libraries followed by NGS and provide high scalability and sensitivity (see [Plongthongkum et al., 2014]).

Current and future developments for the detection of DNA methylation

Despite the numerous advantages of bisulphite sequencing, there are still further improvements to be made. Recent efforts aim to reduce the amount of starting DNA down to picogram levels without the need for amplification (e.g. via post-bisulphite adaptor tagging, PBAT [Miura et al., 2012]), and even to levels that allow DNA sequencing of single cells [Smallwood et al., 2014]. This is especially useful for the analysis of cell types that are difficult to isolate, such as germ or embryo cells [Guo et al., 2014], and to investigate cell-to-cell variability in DNA methylation, which can occur in even rel-

atively homogeneous cell populations [Smallwood et al., 2014]. Yet, these most recent methods can only access less than half of the genomic cytosines (in humans) so far [Smallwood et al., 2014].

Advanced sequencing technologies, such as single molecule real-time sequencing by Pacific Bioscience or the emerging nanopore techniques can distinguish methylated from unmethylated cytosines on the fly without pretreatment and amplification of the DNA [Flusberg et al., 2010, Branton et al., 2008], although enzymatic conversion of 5-methylcytosines seems to noticeably enhance the detection sensitivity [Clark et al., 2013]. However, accuracy and throughput need to be improved before these approaches become the most efficient and affordable DNA methylation detection methods.

2.4 Analysis of bisulphite sequencing data

Due to the high throughput, the analysis of WGBS-Seq reads poses distinct computational challenges in terms of data processing, bias correction or statistical analyses. In the following, I focus on the methodologies that have been developed to call methylation and differential methylation based on large-scale next-generation sequencing.

2.4.1 Mapping of bisulphite treated reads

The first step in WGBS-Seq analyses is the standard quality control filtering as performed for conventional genomic reads (section 2.2.1). Due to the reduced complexity of bisulphite-treated reads (Figure **2.3**), the mapping process is the first step that requires specific adaptations. Alignment of these reads against a reference using established short read mapping tools would result in many mismatches of thymines in the read to cytosines in the reference. To retain mapping efficiency, two general strategies of tolerating these mismatches have been developed [Bock, 2012]. The first approach modifies the alignment process by either adapting the alignment scoring matrix such that C-to-T mismatches are not penalized and count as matches, or by replacing cytosines in the reference genome into the wild-card letter Y, which matches both cytosines and thymines in the reads (e.g., BSMAP [Xi and Li, 2009]). In contrast to such 'wild-card' aligners, 'three-letter' mapping tools internally convert all Cs in the read as well as in the reference sequence into Ts, thus reducing the sequence alphabet to three letters (e.g., Bismark [Krueger and Andrews, 2011]). This way, three-letter aligners avoid mismatches induced by bisulphite conversion and can therefore utilize standard short read alignment tools.

With any of these methods, the sequencing complexity is reduced, which can introduce slight biases [Krueger et al., 2012, Bock, 2012]. In WGBS-Seq analyses, reads mapping to multiple locations are typically discarded to avoid uncertain read counts. A reduction in sequence complexity leads to a larger number of reads aligning to more than one position in the reference sequence, which are then discarded. Since methylated

reads contain Cs, they are more likely to map uniquely than unmethylated reads and are therefore favored by wild-card aligners. Three-letter aligners typically achieve lower genomic coverage because cytosines are depleted from the reference and the read sequences and there is therefore an increased chance that reads covering cytosine sites will be ambiguous and thrown out. However, since these biases only affect rather repetitive regions and shorter reads, they are usually considered tolerable [Krueger et al., 2012].

The analysis is also affected by the fact that there are two types of sequencing libraries for WGBS-Seq reads [Krueger et al., 2012]. Sequencing adapters can be ligated to the DNA so that only the original bisulphite-treated DNA strands are sequenced ('directed library'; reads 1+ and 1– in Figure **2.3**) [Lister et al., 2008]. Alternatively, and most commonly used today are 'undirected libraries' [Cokus et al., 2008], where paired-end sequencing is performed and four possible DNA strands are sequenced after PCR amplification (Figure **2.3**). This way, an unmethylated cytosine on the reverse strand of the reference (a thymine) is amplified to an A on the opposite strand, which is aligned to a G on the forward reference sequence in the mapping process. Thus, when using undirected libraries, the aligner has to allow not only for C-to-T, but also for G-to-A mismatches. In this case, three-letters aligners generate an additional genome index where Gs have been replaced by As, and G-to-A-converted reads are mapped against it.

2.4.2 Determining methylated positions

Determining the methylation rate

After the read alignment, WGBS-Seq data consists of the counts of Cs and Ts contained in the reads covering each cytosine in the genomic DNA sequence. Most commonly, the methylation rate is simply calculated as the C/T fraction at a given site. When using undirected sequencing libraries, half of the reads are derived from the forward strand and the other half from the reverse strand of the original bisulphite-treated DNA molecule (Figure **2.3**). Thus, these two classes of reads have to be separated first. The combination of mapping direction and mismatch type (C-to-T or G-to-A changes), or mapping direction and sequencing read number in case of paired-end sequencing, determines the strand from which a read is derived (see section 4.3).

Methylation from WGBS-Seq is measured quantitatively and can take values between 0 and 1, or 0 and 100%. Typically, few cytosines have a methylation rate of 100%. This is because usually rather heterogeneous cell mixtures consisting of different cell types or even tissue types are sequenced, and they feature characteristic modifications of their methylomes [Ziller et al., 2013, Leung et al., 2015, Widman et al., 2014]. Furthermore, different developmental stages carry different DNA methylation fingerprints [Feng et al., 2010, Cantone and Fisher, 2013], but proper experimental design can minimize this variability. However, cell-to-cell variability can even occur in relatively homogeneous cell populations [Smallwood et al., 2014]. Some cytosines might be in dynamic states, e.g. due to steady competing methylation and demethylation reactions, or due

to frequent incomplete methylation maintenance during mitosis [Jones, 2012]. Lastly, another reason that methylation rates are not completely binary is the presence of allele-specific methylation, e.g. at imprinted loci, where the different parental chromosomes exhibit specific methylation patterns, leading to a mixture of methylated and unmethylated reads at such loci.

Biases in determining the methylation rate

The biological and experimental factors mentioned above determine the 'true' methylation rate. However, the WGBS-Seq analysis introduces four main biases that can cause the observed methylation rate to not accurately reflect the true methylation level:

- Incorrect base calls or mis-mappings of short reads can appear as a different methylation state. This can be countered by considering the mapping quality or assigning quality scores to positions, as is done for the genotyping in resequencing analyses (section 2.2.4).
- As for other quantitative sequencing approaches like RNA-Seq, the read counts can be biased by clonal reads derived from the amplification by PCR. Removing reads with identical start and end coordinates after mapping, and trimming of potentially overlapping sequences of read partners when using paired-end sequencing can eliminate this bias.
- Next-generation sequencing features substantial read depth fluctuations. The lower the sequencing coverage, the higher are the chances that the random sampling of reads during the sequencing process will not reflect the true methylation rate. Technical or biological replicates can alleviate this bias and also provide the advantage of compensating for low coverage sites. Moreover, biological replicates allow estimating the natural variance in methylation levels.
- Sodium bisulphite might leave unmethylated cytosines unchanged or, when overtreated, might convert methylated cytosines, although the latter rarely occurs. This can lead to false positive or false negative methylation calls, respectively, and adds a background variance to methylation levels. However, the sensitivity and specificity of the bisulphite conversion can be estimated by assessing the methylation level for control DNA with known methylation status. Most commonly, a 'false methylation rate' is monitored using unmethylated lambda DNA spiked in the sequencing library, or using typically unmethylated non-CG sites in humans or unmethylated chloroplast DNA in plants.

Calling methylated positions

The seminal first WGBS-Seq studies in *A. thaliana* and human [Lister et al., 2008, Cokus et al., 2008, Lister et al., 2009] established a common practice in calling methylated positions. Mis-alignments of reads were minimized by only considering uniquely

mapping reads, the PCR bias was avoided by discarding duplicate reads, and the read counts at each position were binomially tested against the incomplete bisulphite conversion rates. Thus, methylation levels that were unlikely to be explained by this 'false methylation rate' were considered as statistically significantly methylated positions.

2.4.3 Determining differential methylation at single sites

Most commonly, researchers are interested in comparing the methylation profiles of individuals within a population or between case and control samples. There are a multitude of approaches to determine differentially methylated positions (DMPs), which I roughly classify into three categories based on whether they perform statistical testing and whether they incorporate biological variance, estimated on biological replicates. The first category contains methods that neither incorporate statistical testing nor biological variance, the second group consists of approaches that implement statistical tests, but do not model biological variance, and methods of the third set fulfill both criteria. I will highlight some strategies from each of these groups.

The simplest method is to classify DMPs based on arbitrary cutoffs on the absolute methylation rate differences (applied for example in ref. [Laurent et al., 2010]). Since methylation rates underlie many sources of variance (as described above), this strategy is likely associated with high error rates, especially for lowly covered sites. A second approach that does not include statistical tests first identifies methylated positions by the best practice strategy described above, and then classifies sites as DMPs that are in different methylation states [Schmitz et al., 2011, Schmitz et al., 2013b, Schmitz et al., 2013a]. While such a strategy is efficient, it identifies solely presence/absence methylation and is not able to detect putatively relevant quantitative differences between methylated sites. Furthermore, this strategy includes DMPs that exhibit subtle differences in methylation rate near the switch between methylated/unmethylated calls, which strongly elevates the number of DMPs. This is especially true for CHH sites in plants, since the vast majority of methylated CHH positions show only 30% methylation or less. This bias might partly explain the finding of a recent study that has analyzed methylomes of 140 *A. thaliana* natural accessions and identified more than 90% of the genome-wide cytosines as being differentially methylated in at least one accession [Schmitz et al., 2013b]. The vast majority of DMPs was in the CHH context, which is different from what is seen in simple pairwise comparisons.

The second category of DMP detection approaches includes the currently most widely used method in plants. It tests sites in pairwise comparisons between samples using Fisher's exact test or a similar method developed by Altham [Altham, 1969]. This strategy compares read counts rather than methylation rates to account for coverage fluctuations [Lister et al., 2009, Lister et al., 2011, Qian et al., 2012, Calarco et al., 2012, Hodges et al., 2011]. A different strategy approximates methylation rates using a binomial model. For example, Chodavarapu *et al.* called positions as DMPs that have non-overlapping confidence intervals, retrieved from the binomial distributions [Chodavarapu et al., 2012]. That the authors additionally required a spe-

cific absolute methylation rate difference is likely because the binomial distribution allows for only a rather narrow variance [Chodavarapu et al., 2012]. While the statistical tests mentioned above account for technical variance caused by sequencing depth differences, they cannot model biological variance of methylation rate measurements from potential biological replicates. Instead, most studies discussed so far, in which replicate data was used, performed statistical testing on the accumulated read counts of the replicates.

Using biological replicates can compensate for low-coverage sites and also increases certainty of the methylation level, thus statistical power, which reduces the false positive DMP rate. The third category of DMP callers includes a linear mixed model to account for biological replicates, which is, however, restricted to exactly two replicates per sample [Dowen et al., 2012]. Recently, numerous specialized tools for the detection of differential methylation have been developed that incorporate biological variance. Among them are methylKit [Akalin et al., 2012] that utilizes logistic regression, and BSmooth [Hansen et al., 2012] that assumes local correlation between methylation rates by smoothing methylation levels over a large genomic window, thereby modeling read counts as binomially distributed. However, it is unclear how justified the assumption of local correlation is in plants. The most common approach to incorporate biological replicates in recent software implementations is to use a beta binomial model [Robinson et al., 2014], which will be introduced in the next section.

Beta binomial models for methylation

Numerous tools [Hebestreit et al., 2013, Park et al., 2014, Feng et al., 2014, Dolzhenko and Smith, 2014, Akman et al., 2014] and a research study in human [Ziller et al., 2013] have relied on the assumption that the number of methylation-supporting reads at a position is binomially distributed (given the methylation rate and coverage) and that the methylation rate follows a beta distribution (given the read counts). The binomial distribution models the within-sample variance of the methylation rates caused by stochastic read sampling variability, whereas the beta distribution accounts for between-sample variation of methylation ratios estimated from biological replicates. Combining the two distributions generates a model where the methylated and unmethylated read counts are assumed beta-binomially distributed. The beta binomial distribution can be described by two parameters: the mean and an over-dispersion parameter that represents the additional variance relative to the binomial variance. Thus, it allows more flexibility in modelling input data with a broad spectrum of variance (Figure **2.4**). Table 2.3 summarizes approaches that use beta binomial modeling and contrasts them to the early, most widely used studies relying on Fisher's exact test.

The current tools approximate the parameters of beta binomial distributions ([Ziller et al., 2013], MOABS [Sun et al., 2014], DSS [Feng et al., 2014], methylSig [Park et al., 2014]) or beta binomial regression models (BiSeq [Hebestreit et al., 2013], RADmeth [Dolzhenko and Smith, 2014]). The latter allows accounting for 'batch ef-

Table 2.3: Approaches to call differential methylation. 'Covariates' denotes whether the tools account for additional variables like batch effects. 'Grouping' denotes whether the tools provide a method to cluster samples into epiallele groups. *: additional references are [Lister et al., 2009, Lister et al., 2011, Schmitz et al., 2011, Qian et al., 2012, Calarco et al., 2012, Hodges et al., 2011], **: applies for [Becker et al., 2011] only. BH: Benjamini-Hochberg multiple correction method, DMP: differentially methylated position, DMR: differentially methylated region, FDR: false discovery rate, FNC: false negative (bisulphite) conversion rate, FPC: false positive conversion rate.

	early methods	BSmooth	Ziller *et al.*	BiSeq	methylSig	MOABS	DSS	RADmeth	BEAT	methpipeline
Reference	[Becker et al., 2011]*	[Hansen et al., 2012]	[Ziller et al., 2013]	[Hebestreit et al., 2013]	[Park et al., 2014]	[Sun et al., 2014]	[Feng et al., 2014]	[Dolzhenko et al., 2014]	[Akman et al., 2014]	[Hagmann et al., 2015]
Read count modeling	Hypergeometric	Binomial	Binomial	Binomial	Binomial	Binomial	Binomial	Binomial	Binomial mixture (accounts for FPC, FNC)	Hypergeometric
Methylation rate modeling	-	Smoothing (2kb or 70 CpGs)	Beta	Beta regression on smoothed m.rate (80bp)	Beta	Beta	Beta (with variance shrinkage)	Beta-Binomial regression	Beta	-
Single site testing	Fisher's exact test	Adapted *t*-test Significance cutoff on marginal empirical *t*-distribution	Beta difference distribution *P*-values against normal or simulation	Wald test *P*-values against normal	Likelihood-ratio test *P*-values against t^2 or chi^2	"Credible methylation difference" (CI of beta difference distribution)	Wald test *P*-values against normal	Likelihood-ratio test *P*-value retrieval unspecified	- (comparing probability thresholds of methylation rates)	Fisher's exact test
Replicate aware (DMPs)	no (DMPs betw. repl.removed**)	yes	yes	yes	yes	yes	yes	yes	no	no (DMPs betw. repl. removed)
Multi. testing correction	(in some cases)	no	no	FDR by Benjamini	FDR	no	no	?	no	FDR by Storey, context-dep.
Region selection	DMP clustering / Sliding window / Tiles	DMP clustering	DMP clustering	Predetermined regions (CpG-dense clusters)	Predetermined regions	DMP clustering using HMM (no details)	DMP clustering	DMP clustering	?	HMM-derived methylated regions
Region testing	Fisher's exact test / Kruskal-Wallis test / Chi^2 statistic / none	no	Random effects model	Wald test if >1 DMP in cluster *P*-values transformed to *z*-score, against normal	Likelihood-ratio test *P*-values against t^2 or chi^2	no	no	Merging of proximal *P*-values (*Z*-score transformation)	- (comparing probability thresholds of methylation rates)	Likelihood-ratio test (Beta-binomial) *P*-values against chi^2
Replicate aware (DMRs)	no	no	yes	yes	yes	*NA*	*NA*	no	*NA*	yes
Multi. testing correction	(yes, when regions tested)	*NA*	no	DMR size-weighted BH	FDR	*NA*	*NA*	no	no	FDR by Storey
Covariates	no	no	no	yes	no	no	no	yes	no	no
Grouping	no	no	no	no	no	no	no	no	no	yes

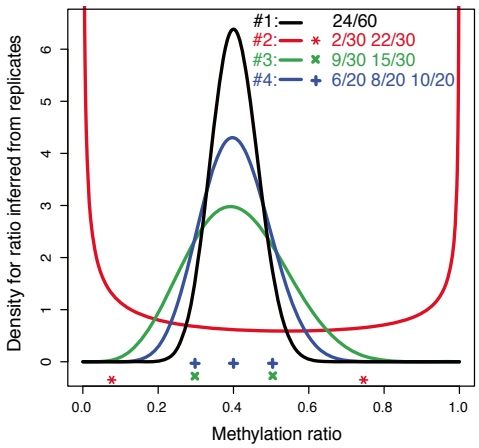

Figure 2.4: Exemplary beta distributions of methylation rates dependent on biological replicates. Symbols at the bottom denote the observed methylation rates of the replicates at the given site. The ratios on top indicate the number of methylated/number of total reads for each replicate. Note that the accumulated read counts for all four cases is identical. Reprinted from [Sun et al., 2014].

fects' like the gender or age of the samples, or more general properties like the day of sequencing library generation, which might be confounders in methylation analyses (for a general review of batch effects, see [Leek et al., 2010]). Another approach approximates read counts by a binomial mixture model incorporating false negative and false positive bisulphite conversion rates [Akman et al., 2014], but does not provide sound testing for differential methylation. While most programs utilize likelihood ratio tests or the similar (asymptotically equivalent) Wald test, MOABS [Sun et al., 2014] and Ziller *et al.* [Ziller et al., 2013] calculate beta difference distributions. MOABS defines a custom 'credible methylation difference' based on the confidence interval of the beta difference distribution that is assumed to combine statistical and biological significance. Ziller and colleagues calculate P values directly from the beta difference distribution by simulation, or by comparing against a normal distribution.

Since the described studies and tools were designed to operate on WGBS-Seq data (BiSeq on RRBS-Seq data), they perform millions of statistical tests. This large number necessitates using multiple hypothesis testing to minimize false positive significant tests that occur by chance. While most biological studies control the false discovery rate (FDR), only two of the presented tools offer the possibility to correct for multiple testing (methylSig and BiSeq).

Even though the beta binomial model seems most appropriate, it has been rarely applied beyond proof-of-principle studies carried out by the developers of the software programs that implement beta binomial models. Only a few biological studies in mammals utilized such modeling (e.g., [Ziller et al., 2013]), but to my knowledge there is no study in plants yet. One reason is that most developed software is designed for human

samples, i.e. assuming a single methylation rate for CG sites only, and adaptations for the more complex plant methylation patterns are still lacking.

2.4.4 Determining differential methylation in regions

The biological significance of DNA methylation differences at single sites is still not fully understood, and most studies that identified epialleles reported methylation differences in larger genomic regions (section 1.7). Thus, it is common to identify variable methylation that is clustered in regions. Methods to detect differentially methylated regions (DMRs) can be divided into approaches that determine regions *de novo*, or that test predetermined regions.

Most previous studies followed the approach of the first category. They first identified differentially methylated positions (DMPs) and consolidated them into DMRs by genomic distance, using criteria such as a minimum number of DMPs or a maximal allowed distance between DMPs (Figure **2.5**B,C) [Lister et al., 2009, Qian et al., 2012, Chodavarapu et al., 2012]. Other studies required significant differential methylation between the resulting DMP clusters using Fisher's exact test [Calarco et al., 2012, Ausin et al., 2012], Altham's method [Hodges et al., 2011], the Kruskal-Wallis test [Schmitz et al., 2011] or a random effects model [Ziller et al., 2013]. Although the tools mentioned in the previous section reasonably identify differential methylation at single sites, some of them (BSmooth, DSS, MOABS and RADmeth) define DMRs by simply clustering DMPs without any statistical test (Table 2.3). Instead, RADmeth combines neighboring P values of DMPs by transforming them into a Z-test [Dolzhenko and Smith, 2014].

All of these approaches require detectable methylation differences at single sites. However, there might be regions in which individual sites feature weak differential methylation and can only generate a significant signal when considered together. Thus, combining methylation data from multiple adjacent sites increases statistical power. In addition, correcting for multiple testing of a vast amount of single cytosines leaves only the strongest methylation differences as statistically significant. Thus, the decreased number of tests when analyzing regions compared to sites reduces the false negative rate for detecting differential methylation.

A few studies segmented the genome into 'tiling' regions (Figure **2.5**D) and statistically tested them using the Kruskal-Wallis test [Lister et al., 2011], Fisher's exact test [Stroud et al., 2013b, Stroud et al., 2013a, Yu et al., 2014] or the chi-square statistic [Regulski et al., 2013]. While these *de novo* DMR detection strategies reduce the number of statistical tests compared to DMP-based methods, they do not maximize statistical power, since they still test the whole genome, which can contain a considerable fraction of unmethylated 'tiles'. Additionally, these methods do not incorporate biological replicate data.

A different approach to identifying regions *de novo* is to test predetermined regions (Figure **2.5**E). Two tools from Table 2.3 operate on arbitrary, user-specified

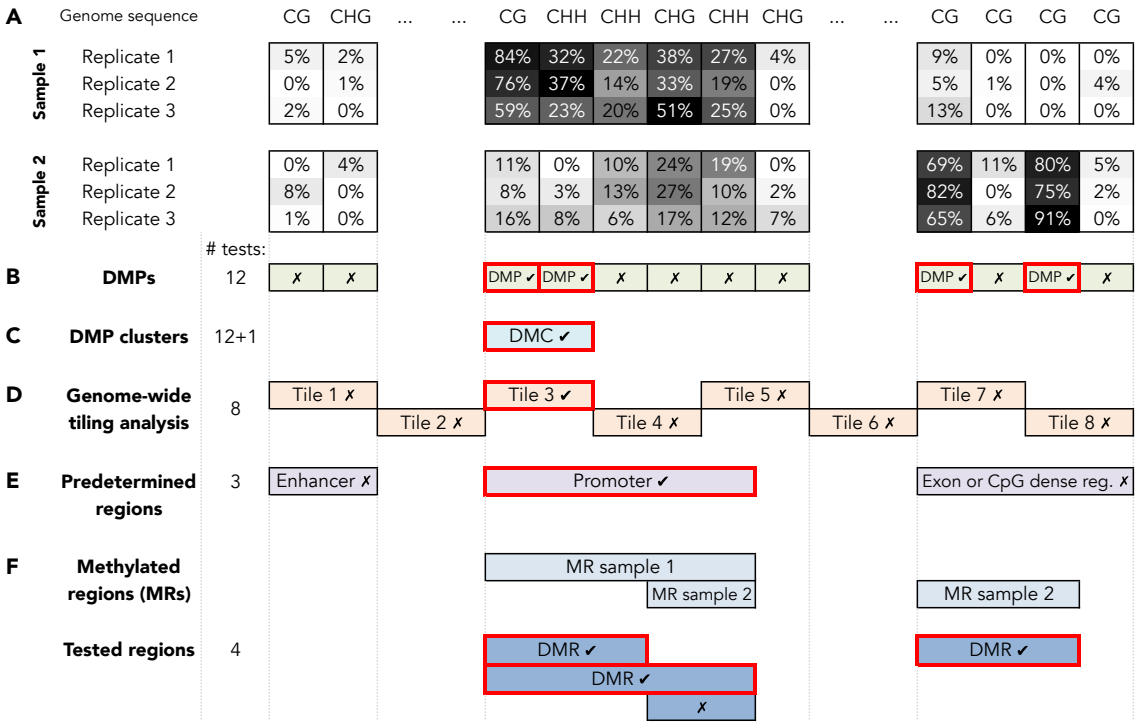

Figure 2.5: Different approaches to call differentially methylated regions (DMRs). (A) Methylation levels of three replicates from two samples each are shown and color-coded from white to black according to the range of minimal and maximal methylation rate of the respective sequence context. Statistical tests can be performed (B) at each single site to obtain differentially methylated positions (DMPs), (C) on each DMP cluster (DMCs) that spans a DMP-dense region, (D) on equally sized and spaced 'tiles' of the genome, (E) on pre-determined regions, defined based on annotated genomic features or selected by enrichment bisulphite sequencing experiments, or (F) on parts of methylated regions (MRs) that are determined by an additional method, e.g. by an HMM (chapter 4). Checks symbolize significant and crosses non-significant statistical tests.

target regions (methylSig [Park et al., 2014]) or regions rich in CG sites that are obtained by enrichment bisulphite sequencing methods, e.g. RRBS (BiSeq [Hebestreit et al., 2013]). These are the only tools to my knowledge that make use of the beta binomial model for testing regions rather than comparing individual cytosine sites and also account for biological replicate data. Both methods apply likelihood ratio tests and correct for multiple testing. BiSeq relies on smoothed methylation rates of CG sites over an 80-bp genomic window, and methylSig can also incorporate information from neighboring CG sites. While the authors of methylSig found that CG methylation rates were correlated within 200-300 bp in humans, it is still not known if this distance, known as 'linkage disequilibrium' (LD), is the same on the whole genome or across individuals. Therefore, the application of these tools to other species would require re-testing the LD of methylation levels to adapt the window size.

An additional caveat when this strategy should be adapted to plant data is that it is unknown how the LD of methylation ratios behave for the different sequence contexts that can be methylated in plants, since different methylation pathways act on partly overlapping, but largely different sequence contexts (section 1.4.1).

A common problem in biological studies is that statistical significance does not equal biological significance (cf. RNA-Seq). Since most discussed approaches call DMRs with rather subtle differences in methylation levels, a common way to potentially increase the biological significance of the detected DMRs is to require either a minimum absolute methylation rate difference (e.g., [Yu et al., 2014]) or a fold-change criterion (e.g., as stringent as an 8-fold difference requirement [Schmitz et al., 2011]). This might reflect the use of statistical tests that do not allow for a broad variance or are sensitive to large numbers. For example, Fisher's exact test tends to yield significant results more often when read counts are large, as can be the case for the accumulated read counts over regions.

2.4.5 Objective of this work: WGBS-Seq pipeline and a novel approach to call DMRs

The accurate identification of DNA methylation variation between individuals is crucial for understanding its role in shaping phenotypic diversity and for identifying potential short-term adaptations to environmental conditions. The current state-of-the-art in charting the whole-genome methylation landscape constitutes WGBS-Seq. Statistical analysis of these experiments poses many challenges due to fluctuating sequencing depth and methylation levels, small sample sizes and a large number of statistical tests performed. The challenge is to detect regional methylation differences, since most epialleles in natural populations consisted of methylation differences of several nearby positions. The vast majority of previous efforts in plant studies that called differentially methylated regions relied either on identifying variable sites first, or testing genome-wide sliding windows, which limits statistical power due to the high number of performed tests. These studies often visualized a few convincing DMRs in genome browser illustrations, but it remained unclear how representative these evident methylation differences are genome-wide.

Recently, numerous tools have been developed that accurately model methylation data using a beta binomial distribution, incorporating within-sample and between-sample variance of methylation rates, i.e. accounting for read depth fluctuations and biological replicate data, respectively (Table 2.3). However, those models have been applied to human samples only, where exclusively the CG context shows methylation, and adaptations for the more complex plant methylation pattern are still lacking.

In chapter 4, I introduce a workflow for next-generation sequencing bisulphite data that performs all steps from raw reads to the calling of DNA methylation in individual samples and differential DNA methylation between samples, both on a single base pair

and regional level (Table 2.3). The method to call DMRs constitutes a novel approach that introduces the use of beta binomial modeling for plant methylation data and uses a strategy that maintains a high level of statistical power. The latter is achieved by first segmenting the genome into methylated and unmethylated regions and then testing only the methylated space for differential methylation, which greatly reduces the number of statistical tests (Figure **2.5**F).

Methylated regions are identified unsupervised from all genome-wide cytosines by a Hidden Markov Model (HMM) that fits context-specific beta binomial distributions to the sequencing data of the regions. This accounts for the more complex methylation patterns that exist in plants compared with other species. This approach provides an unbiased, informed selection of regions to test for differential methylation that is independent from DMPs. Other DMR detection methods that are independent from DMPs target either biased regions from enrichment sequencing protocols (BiSeq), user-specified regions (methylSig), or test the whole genome independent of the methylation status (tiling region methods) (Table 2.3). Furthermore, my approach does not require the setting of arbitrary thresholds, like those used for clustering DMPs or smoothing bandwidths, which either use assumptions or require pre-knowledge about DMRs and their lengths.

Lastly, while current tools to detect differential methylation only report significant pairwise sample comparisons, my pipeline provides a unique method to classify multiple samples into groups based on the significance test between all pairs of samples and thus provides a way to determine epiallele groups.

Together, the accurate beta binomial modeling of methylation rates, combined with testing regions that have been determined in an informed manner make this strategy unique (Table 2.3). In addition, it might be the first method that implements beta binomial modeling for plant methylation data.

Chapter 3

Integrative detection of genetic variants by iterative re-alignment

In this chapter, I propose a strategy to identify diverse genetic variants by integrating multiple sources of variation detection using short next-generation sequencing data, when whole genome *de novo* assembly is infeasible. The pipeline integrates resequencing data, local *de novo* assembly and several structural variation detection tools and performs rigorous validation by re-mapping reads against the predicted polymorphisms.

I will begin this chapter with a brief overview of the proposed strategy and its suitability, before I introduce recommended variant detection tools that were used in the study presented in chapter 5. Subsequently, all further steps of the pipeline are elaborated on.

Contributions

This chapter describes the methods included in a publication: [Hagmann et al., 2015]. I designed and implemented all analytical steps and scripts presented in this chapter.

3.1 General workflow

The general strategy starts by collecting all genetic variants of a sample predicted by many genetic variation detection tools into a sample-specific consolidated variant set (CV set; Figure **3.1**a). I correct for putative false positives among the CVs by retaining only variants without contradicting resequencing reads (Figure **3.1**b). If multiple samples are used, the method determines common and potentially segregating variants by comparing haplotype sequences around CVs of the samples to each other. Subsequently, all polymorphisms not called in all strains of a population are tested for their presence in each sample. This compensates for potentially missed variants in the sample-specific calling (false negatives), but also constitutes a second filtering step to remove false positive calls in the sample-specific CV set. For this validation of variants in each sample, all non-shared variants of the population are incorporated into

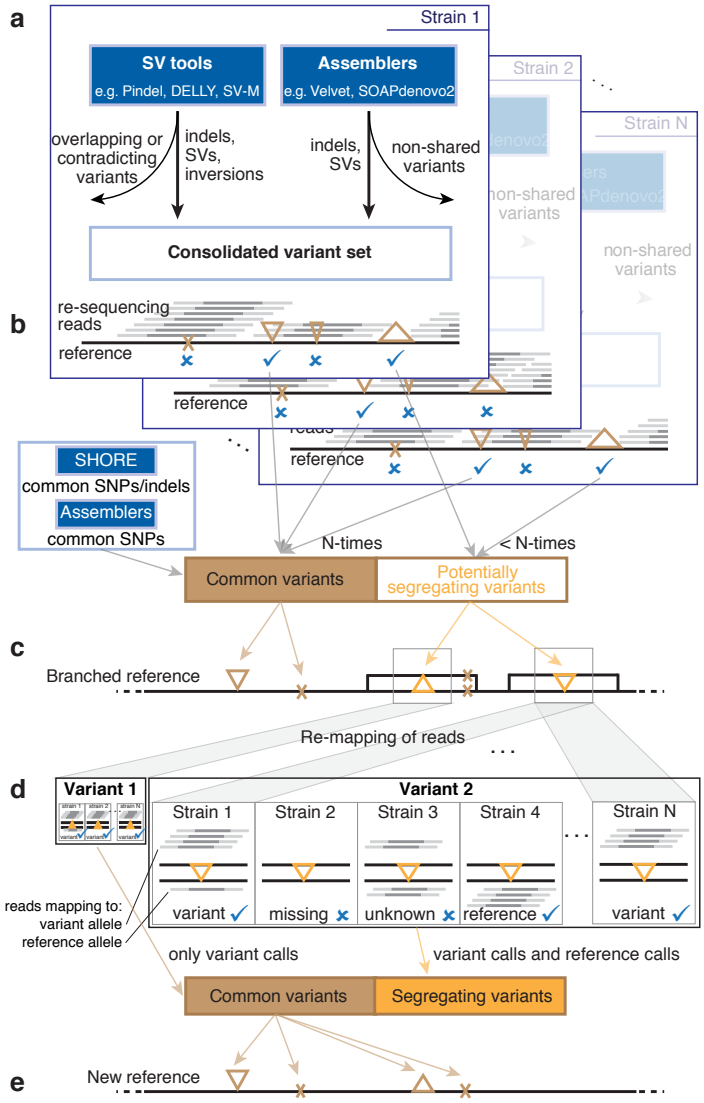

Figure 3.1: Workflow of the pipeline to generate a pseudo reference. (a) For each strain, variants called by diverse structural variation (SV) detection tools and assemblers are combined into a consolidated variant set. (b) Predicted variants that overlap with inner cores of mapped resequencing reads are filtered out and the remaining variants are classified into common and potentially segregating. Brown triangles: insertions/deletions, brown X: SNP. Checkmarks and crosses indicate that variants did or did not pass the filter, respectively. (c) A new reference genome incorporates all common variants and includes additional sequence containing the alternative haplotype sequence of segregating variants as 'branches'. (d) After mapping the reads against the branched reference, a binomial test assesses for each variant site which allele is present in each strain. Checkmarks and crosses indicate that there is or is not a valid variant call, respectively. Sites with the same variant allele call in all strains are classified as "common", those with significant tests for both a reference and a variant allele across all strains as "segregating" variants. (e) All common variants from the previous step can be incorporated into a new reference sequence, and a new iteration can start from (a).

a 'branched' reference sequence so that it harbors both the reference and the variant alleles (Figure **3.1**c). By re-aligning all reads against the branched reference, the presented method can compare the read support between the two alternative alleles. This validation of all detected variants in the population by whole-genome re-sequencing leads to a final distinction into common and segregating variants (Figure **3.1**d).

In addition, all common variants can be incorporated into a new unbranched reference genome, and this workflow can start from the beginning, thereby refining the reference sequence from iteration to iteration, following the rationale of Gan *et al.* [Gan et al., 2011] (Figure **3.1**e). A repeated application of this workflow constitutes an additional validation for all common (by re-aligning reads) and segregating variants (they have to be consistently re-called), including common SNPs and small indels found by the read alignment tool, which are not explicitly evaluated during a single iteration.

3.1.1 Applicability and availability of the pipeline

The strategy assumes homozygosity across the whole genome, which is appropriate for many plant species. In principle, it can be applied for single strains only. In this case, the validation by re-sequencing (Figure **3.1**c,d) can be performed for the single CV set of the sample only. However, since the workflow can use information from multiple samples for calling variants in each individual sample, it is particularly suited for the analysis of local populations of genetically similar strains, or more generally to strains that separated only recently from a common ancestor and share many variants (relative to a reference) among each other. To increase the detection of common genetic variation even more, especially to resolve highly diverged regions, the strategy can be repeated, thereby generating and iteratively refining a pseudo reference genome. A reference sequence that is more similar to the analyzed strains increases the number of detectable variants that are hidden to an analysis against a more distantly related reference genome, and can enhance analyses like population structure, particularly when the number of distinguishing polymorphisms between strains is low. Furthermore, when interested in the DNA methylation pattern, capturing more genetic diversity helps in identifying more differences in DNA methylation and can facilitate linking genetic to epigenetic divergence, since genetic variation is a major source of DNA methylation changes (section 1.7.1).

I tailored this pipeline to the analysis of a population of near-isogenic *A. thaliana* plants described in chapter 5. However, the pipeline can be applied to other data sets as well. It consists of several perl programs, which are wrapped in bash scripts, since it makes use of UNIX command line tools for basic file manipulations and performs several calls of third-party tools. The current implementation relies on SNP and small indel calls solely from SHORE [Ossowski et al., 2008]. However, it can utilize an arbitrary set of structural variation prediction tools with the requirement that SV calls are provided in the routinely used and standardized "variant call format" (vcf). I provide the set of

commands and the scripts with usage details online[1]. The parameters of the variant calling programs can be found in the Appendix C.

3.2 Tools for genetic variant detection

3.2.1 SNP and small indel calling

The genetic variation detection pipeline was tailored to use SHORE for SNP and small indel calling from read mapping [Ossowski et al., 2008]. Moreover, I slightly adapted SHORE's module 'consensus' to reduce false positive SNP and indel calls. I noticed that regions covered only by the ends of reads are enriched for many variant calls. Although SHORE reduces the impact of a few bases at either end of the reads on variant calling (see section 2.2.4), I observed that even a more trustable 'core region' setting of 10 bp on either side was insufficient to prevent potential false positive calls (Figure **3.2**). Therefore, I conceived a new criterion that requires a variant to be covered by at least one read with the inner 50% of its length (Figure **3.2**).

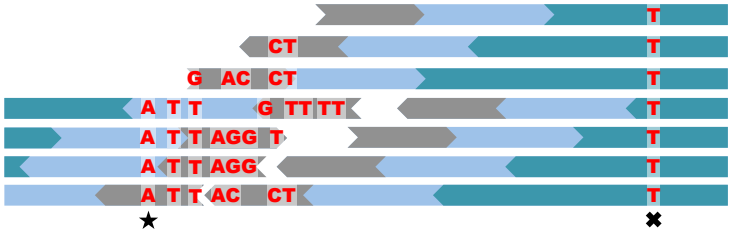

Figure 3.2: Real-world example of read alignments along a reference sequence. Letters refer to variant base calls. Grey bars represent the first and last 10 bases of reads, blue bars the 'core' region and petrol bars the 'inner core' region of reads (inner 50% of their length). Note that at the position marked by the star, all three 'core'-covered reads show the same variant call, which would potentially suffice to call a SNP. A putative reliable SNP is shown at the position marked by the cross, since there is sufficient 'inner core' read coverage.

I included this criterion into a scoring matrix, which is used by SHORE to determine variant quality scores. This matrix contains empirical thresholds for diverse sequence and alignment related features, against which each variant site is tested. Starting from an initial quality score of 40, there are three thresholds for each feature associated with specific relative reductions of the score (low penalty: 5%, intermediate: 20%, high: 40%) that are cumulatively applied. The list of features is presented in Table **S6** (containing the thresholds used for the analysis of chapter 5). A variant site is typically regarded empirically reliable if the quality score is equal to or greater than 25 (Q25), which disallows high penalties (40% of 40 is 16, i.e. a SNP with a high penalty can maximally have Q24).

[1]https://sourceforge.net/p/isvim, last accessed April 2015

3.2.2 Structural variation calling

Since there is typically only little overlap between predicted structural variants (SV) by different tools, it is recommended to apply multiple SV callers [Lin et al., 2014]. As this pipeline was designed to lead to a pseudo reference sequence, and because it is capable of comparing SVs found in different samples to each other, only tools that yield the exact breakpoints of SVs were used. These solely include split-read and assembly approaches (section 2.2.5). Although any tool predicting SVs at single base pair level can be chosen as long as the calls are in the vcf format, I briefly mention the programs used for the whole-genome analysis of *A. thaliana* populations described in chapter 5. Pindel [Ye et al., 2009] and SV-M [Grimm et al., 2013] perform split-read alignment of unmapped reads against a broad window around their mapped partner that spans most of the insert sizes found in the read set. DELLY [Rausch et al., 2012] maps split reads in a region identified as an SV before by analyzing paired-end information, thus combining two different signals for the presence of an SV in one calling process. While Pindel reports all SVs supported by at least two consistent read mappings, SV-M evaluates SV calls by applying a support vector machine that uses a multitude of alignment-related features, including nearby SNP calls found by resequencing, and relies on a training data set. I used the same set of Sanger-validated SVs of an *A. thaliana* natural accession as the SV-M publication [Grimm et al., 2013] to call SVs on our *A. thaliana* strains of chapter 5. In addition, I implemented a custom local *de novo* assembly method targeted towards uncovered genomic regions, which will be elaborated in the following section. This selection of tools makes use of three out of the common four currently available methodologies for SV detection using NGS (section 2.2.5).

3.2.3 Targeted *de novo* assembly

For several reasons, there are genomic regions without read coverage ("sequencing gaps") when performing a re-sequencing approach. First, the underlying region might be deleted in the newly sequenced sample. Second, the region might be too divergent to the reference sequence and beyond the mismatch limits imposed by the short read alignment. Third, the region might not be represented in the read set due to fluctuating read coverage along the genome. *De novo* assemblies can retrieve the diverged sequences of the first two cases, and I therefore developed a method that tries to bridge sequencing gaps by local assemblies.

Rather than using the whole read set, the method restricts the input for the assembly to only those sequences that map to the flanking regions of sequencing gaps, which reduces the complexity of the assembly and improves the quality of the contigs. Further, it targets not only regions with absent read coverage, but also regions without read core coverage, meaning regions that are overlapped by reads with few of their first and last bases only (section 2.2.4). This is because insertion breakpoints or small deletions are mostly covered by read ends overlapping the divergent regions with many mismatches

(see Figure **2.2**). Therefore, I will extend the term 'sequencing gaps' to also include such 'zero-core-covered' regions from hereon. The local assembly tool starts with combining all reads aligned to the 100 bp genomic regions surrounding sequencing gaps together with all unmapped reads and their partners (even if they mapped) into an assembly read set (Figure **3.3**a). Next, two *de novo* assembly tools are applied: SOAPdenovo2 v2.04 [Luo et al., 2012] and Velvet v1.2.0 [Zerbino and Birney, 2008] (see Appendix C for command lines).

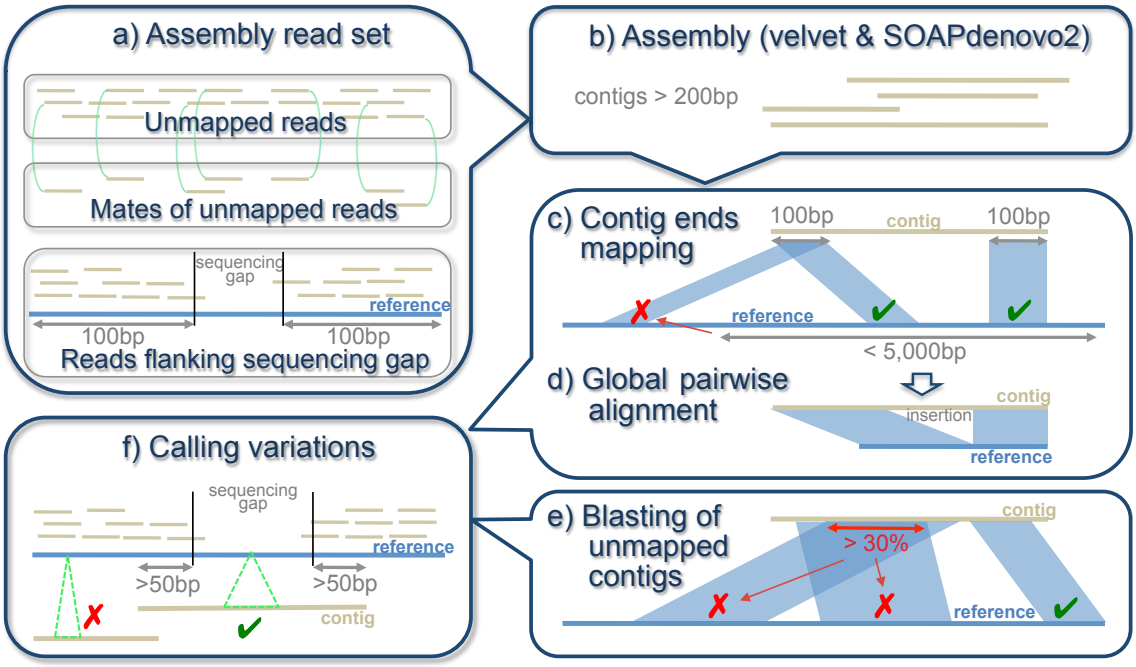

Figure 3.3: Targeted *de novo* assembly approach. See text for explanations.

The resulting contigs of maximal 200 bp in length are aligned to the iteration-specific reference genome. Since BLAST-based alignments do only provide local alignments of largely conserved sequence and thus do not directly report long diverged sequences, the first and last 100 bases of the contigs are aligned using GenomeMapper v0.4.5s [Schneeberger et al., 2009], allowing for 10 mismatches on each end (Figure **3.3**c). The mapped contig ends frame a genomic region, to which the entire contig is then aligned using the global Needleman-Wunsch alignment tool 'needle' from the EMBOSS package v6.3.1, if both contig ends do not map further than 5000 bp apart (Figure **3.3**d). Global alignments of contigs spanning an entire sequencing gap and a surrounding sequence of minimum 50 bp on either end are parsed for all differences between contig and reference sequence, including SNPs, inversions, long deletions and insertions (Figure **3.3**f). Additionally, to yield even more variants, non-mapping contigs are aligned using blastn (from BLAST package v2.2.23 [Altschul et al., 1990]), and variants from partly mapping contigs spanning sequencing gaps are parsed in the same way (Figure **3.3**e). The alignments of contig regions that map to multiple locations on

the reference genome are discarded if they reciprocally overlap by more than 30% of their lengths.

Thus, the method only reports variants within or surrounding sequencing gaps to limit false variant calls on contigs that align to regions of sufficient read support from resequencing. Finally, only identical sequence differences between both assembly tools are included in the final variant set of the targeted *de novo* assembly tool. The method calls shared variants by checking haplotype sequences surrounding the polymorphisms for identity.

3.3 Consolidating variants of different tools

After whole-genome read mapping, consensus calling with SHORE and application of the above described SV detection programs, the detected variant set consists of SNPs, indels and SVs. I refer to insertions and deletions up to a length of 7 bp that are detected by SHORE as indels, while SVs comprise all insertions and deletions from the assemblers, as well as insertions, deletions and inversions predicted by the specialized SV tools that are longer than 7 bp. The threshold of 7 bp is the consequence of short read mapping, which allows maximally 7 gapped positions in each read. Since SNPs and indels from SHORE are quality scored and covered by the inner core regions of reads, I assume they can be merged across samples by directly comparing their coordinates. SNPs from the assembly approach are summarized the same way. The following describes how the pipeline consolidates SVs found by multiple tools in each sample into a consolidated variant (CV) set (Figure **3.1**a).

Since different alignments of reads spanning the same SV can potentially result in overlapping and contradicting SV calls (Figure **3.4**), the first step of the consolidation step repositions the individual SVs to consistent coordinates so that they can later be compared across samples. I apply the tool Dindel v1.01 [Albers et al., 2011] to perform this task. It additionally filters out all but one out of multiple different calls for the same SV (Figure **3.4**A).

Subsequently, to tackle the problem of the high false positive rate of SV detection tools, the next step assesses the coverage of the resequencing reads within SV regions. While alignments of the ends of reads are prone to false mismatches or indels at diverged loci (see section 3.2.1), genomic regions covered by uniquely mapping reads with the inner 50% of their sequences ('inner core') are unlikely to contain long-range SVs when assuming homozygosity throughout the genome (section 3.2.1). Thus, I interpret coverage of reads with their inner core as evidence for the presence of the reference allele in the read set at the particular locus. Therefore, predicted SVs that overlap regions showing inner core coverage of uniquely mapping reads are discarded. The inner core coverage information along the genome is provided by my adapted version of 'SHORE consensus' (section 3.2.1).

Since different SV calls of different tools can lead to the same alternative DNA sequence compared to the reference sequence in a region, including different number and

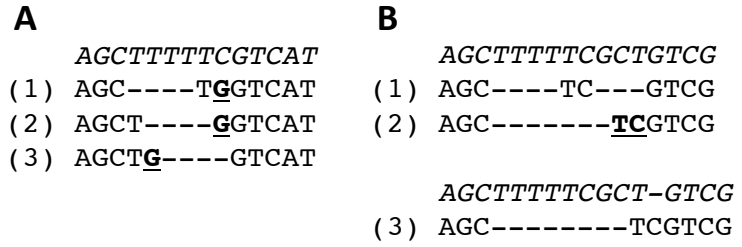

Figure 3.4: Alternative alignments of identical reads. Reads (enumerated) aligned to the same reference sequence region (italic sequence on top). Mismatches are shown as bold, underlined letters in the read sequences and gaps as dashes. (A) All reads show one mismatch and a 4-bp deletion, but the deleted sequence is different between the first two (TTTT) and the last read (TTTC). (B) Three different alignments exhibiting different number and types of variants.

types of variants (Figure **3.4**B), I furthermore provide a method to check for identical variation called by the different tools. To this end, I define haplotype sequences for each variant and tool containing the SV and, by default, 30 bp to the right and left of the reference sequence. SVs closer than 30 bases to each other are consolidated into one haplotype. In case of different haplotypes for a locus, the pipeline only retains SVs of the haplotype with the highest frequency across tools, or none if there is no majority.

The final set of consolidated variants (CVs) for each sample consists of SVs from non-contradicting haplotypes, SNPs and small indels called by SHORE, and SNPs detected by the assembly tools. Thus, the pipeline reports SVs that are specific to a single tool or, in case of contradicting SV calls, it retains only variants that are backed up by at least two programs.

3.4 Building a branched reference sequence

Since the consolidated variants of the samples have been consistently repositioned, chances are high that identical SVs across samples have the same coordinates. Thus, the pipeline compares all CVs for identity and incorporates variants found in all samples into the current reference sequence, replacing the previous reference allele. Similarly, common indels from SHORE and non-redundant common SNPs from SHORE and the assemblers are integrated into the reference sequence. By contrast, the non-common CVs are incorporated in addition to the respective reference allele as sequence "branches", so that read mapping in best-hit mode should reveal exclusive coverage of the allele that is present in the read set (Figure **3.1**c). This step compensates for an expected high false negative rate of SV detection tools, since the presence of all predicted non-common SVs in the population is checked in each individual sample. Thus, I refer to these SVs as potentially segregating, since many of them might be true

common SVs (Figure **3.1**b). Variants of a sample that are in a user-defined genomic distance to each other, or identical haplotypes across samples are merged into the same branch sequences. Contradicting segregating SV calls across samples lead to more than one variant branch at the corresponding locus. All non-reference branch sequences are concatenated, separated by a padding sequence ('N's) to prevent mapping across two independent branches. Their coordinates are stored and linked to the coordinates of the corresponding reference allele. Finally, the resulting sequence is included into the reference genome as an additional chromosome.

3.5 Population-aware calling of common and segregating variants

For each sample, all reads are aligned to the new reference sequence and read counts obtained using 'SHORE consensus'. Since the same variant can be included in several different haplotypes, reads supporting this variant would map at multiple locations in the reference. Therefore, the pipeline continues using the counts of all rather than only uniquely aligned reads. Next, it compares the inner core read coverages at the variant site of each branch (r_b) with the corresponding aligned site on the reference haplotype (r_{ref}). To increase certainty of variant calling and to rule out heterozygosity, the strategy is to test the read count of the most covered allele against a binomial distribution that assumes 95% allele frequency out of a total of $r_b + r_{\text{ref}}$ observations (95% and not 100% to account for slight read sampling bias). I require a total coverage over each pair of variant and reference allele of 4x, otherwise the site is marked as 'missing data' for this sample. After P value correction by Storey's method [Storey and Tibshirani, 2003] across the sufficiently covered loci, the null hypothesis of homozygosity is rejected at a maximal FDR of 5%.

Because low-coverage sites rarely yield statistical significance, I follow a "population-aware" approach to find more commonalities between samples. If there is at least one sample with a statistically significant homozygous call (q value below 0.05) at a variable site, the criterion for the other samples is relaxed so that they are more easily considered homozygous at the same site, namely if the read count of a haplotype exceeds the alternative haplotype read count by 2-fold. Finally, I classify variants as common in the population if all samples have a homozygous call for a variant 'branch' haplotype, and as segregating if both a homozygous reference and homozygous alternative allele are present in the population (Figure **3.1**d). The common variants can ultimately be incorporated into the current iteration's reference sequence to serve as a new reference for subsequent iterations (Figure **3.1**e). This procedure can terminate after a user-defined number of iterations, or when the reference genome converges, i.e. when the number of newly identified common variants is lower than a threshold of choice.

Chapter 4

A pipeline for the detection of differential methylation

In this chapter, I describe a complete workflow performing all steps from raw bisulphite next-generation sequencing reads to the statistically sound calling of DNA methylated positions in single samples and of differentially methylated loci between samples. The applied methodologies are specifically designed for the analysis of plants that exhibit DNA methylation in all three sequence contexts and are methylated at a small proportion of their genomes. While the identification of variably methylated single sites follows a similar strategy with previous plant studies, the pipeline introduces a novel approach to call differentially methylated regions (DMRs). It employs the current state-of-the-art representation of methylation data (using beta binomial modeling) to first define a set of methylated regions in an unbiased and informed manner (using an HMM method) and then compares these regions between samples with a sensitive beta binomial-based test. This strategy maintains a high level of statistical power and results in more unbiased DMR calls.

The chapter starts with outlining the general workflow of the pipeline and then chronologically follows and explains in detail all analytical steps from aligning reads to a reference genome to the calling of differential methylation between samples.

Contributions

This chapter describes the methods included in two publications: [Becker et al., 2011, Hagmann et al., 2015]. I adapted the short read mapping tool GenomeMapper to handle bisulphite data by modifying a version called Palmapper [Jean et al., 2010] as described in section 4.2. Based on the existing module 'consensus', I wrote the SHORE module 'methyl', which retrieves read count data for methylome studies and is described in section 4.3. I created and implemented all other programs and scripts of the pipeline with two exceptions: Dr. Oliver Stegle implemented the P value correction method used for the calling of differentially methylated positions and regions, and Jonas Müller wrote the software for statistically testing genomic regions for differential methylation

based on beta binomial models (section 4.6.3). The program to call methylated regions is based on the software MethPipe[1] v0.7.5 which I modified as described in section 4.6.1.

4.1 General workflow

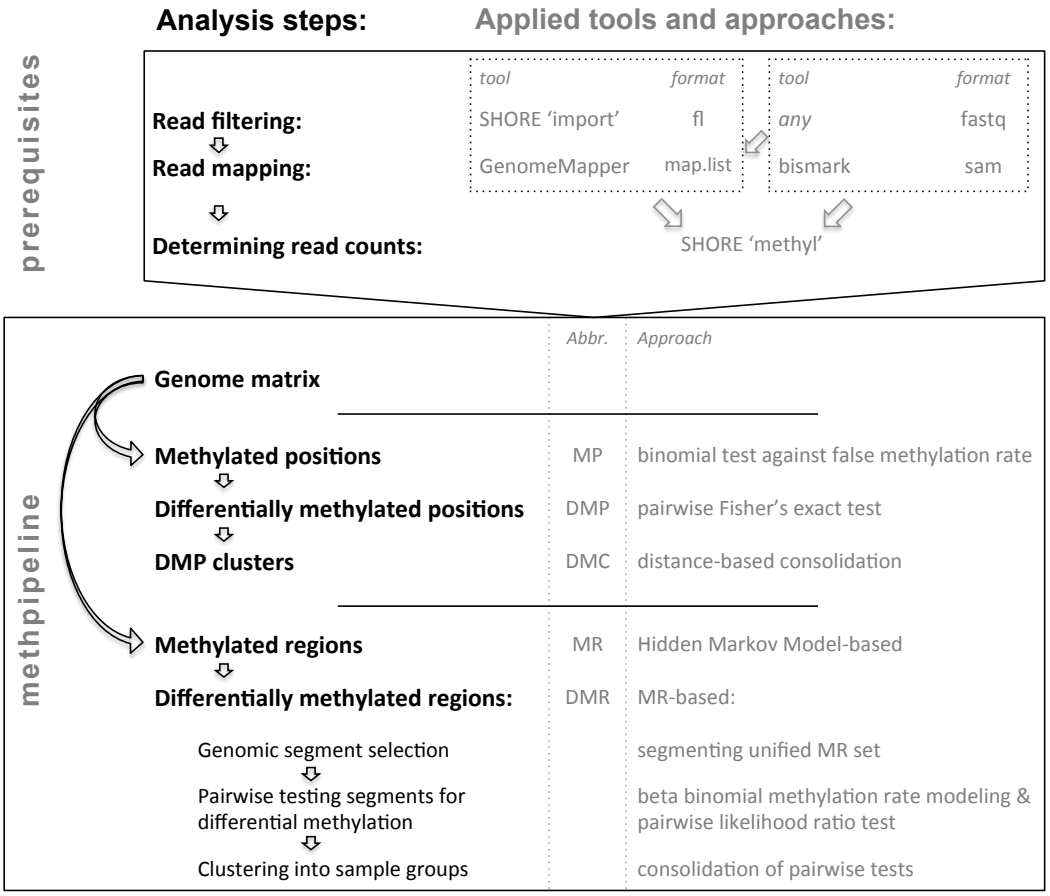

Figure 4.1: Workflow of the WGBS-Seq pipeline to call methylation in individuals and differential methylation between individuals.

The stand-alone methylation pipeline starts with filtered and aligned whole-genome bisulphite sequencing (WGBS-Seq) reads and requires the methylation data to be stored in a specific 'genome matrix' format (Figure **4.1**). I propose two routes to perform the pre-processing steps read filtering, read alignment and genome matrix generation.

Read filtering can be performed by any software solution that either produces read files in the SHORE-specific flat file format [Ossowski et al., 2008] or in the standard,

[1]http://smithlabresearch.org/software/methpipe/

widely used fastq format. The pipeline was tailored to two different bisulphite alignment tools: an adapted version of GenomeMapper [Schneeberger et al., 2009] and bismark [Krueger and Andrews, 2011]. Based on either of the two mapping file formats, my new SHORE module 'methyl' summarizes the alignment files into a genome matrix that contains the read counts of all genome-wide cytosine sites for all analyzed samples.

The methylation pipeline can then call methylated positions by binomially testing read counts against a false methylation rate (section 4.4). These positions can furthermore be tested between samples using Fisher's exact test to retrieve significantly differentially methylated positions (MPs; section 4.5). Finally, DMP-dense genomic regions can be reported as DMP clusters (DMCs; section 4.5.1).

An alternate, more sensitive and more unbiased method to detect regional methylation differences is to first identify methylated regions (MRs; section 4.6.1) and then compare them between samples (section 4.6.2 ff.). MRs can be calculated based on the genome matrix. Subsequently, DMRs can be obtained based on these MRs in further three main steps, elaborated in section 4.6.

Altogether, applying these analytical steps provides a comprehensive picture of the DNA methylation landscape of individual plants and of its variability within a population.

The source code, a user manual and usage recommendations are available online[2].

4.2 Alignment of bisulphite-treated reads

Bisulphite treatment converts unmethylated cytosines into thymines in the sequencing reads; thus short read mapping tools must account for this type of mismatch. I therefore adapted our previously developed alignment program GenomeMapper [Schneeberger et al., 2009] and implemented the changes into a version [Jean et al., 2010] that was created by merging GenomeMapper and QPALMA (PALMapper [De Bona et al., 2008]). GenomeMapper compares "seeds" (short sequences of up to 13 bases) contained in the reads to an index of all seeds found in the reference genome to retrieve possible genomic positions of the reads. It merges adjacent seeds on the reference into hits, which are ultimately aligned using classic alignment algorithms (an adapted Needleman-Wunsch algorithm). Thymines in a bisulphite-treated read can derive from true thymines or from unmethylated cytosines in the reference genome (Figure **2.3**). Similarly, adenines might match to guanines in the reference sequence. Hence, for each seed of a read, a new 'modified seed' is generated for each combination of all thymines in the seed being in one of two states: either unchanged or replaced by cytosine. Likewise, modified seeds are produced for all adenine configurations in the seed. This strategy imitates a reverse bisulphite conversion and allows the original genome index to be used without modification. The subsequent seed extension and alignment steps of GenomeMapper also remain unchanged. A sin-

[2]http://sourceforge.net/projects/methpipeline/, last accessed April 2015

gle final step was added that scans the read sequence for base conversions and reports them.

For each seed, this method obviously requires the additional analysis of $2^t + 2^a - 2$ modified seeds, where t and a are the number of thymines and adenines in the unmodified seed, respectively. To reduce this number by half, GenomeMapper exploits the fact that the two strands of the adapter sequence used in the standard Illumina protocol for paired-end bisulphite libraries contain different sequencing primers for read 1 and read 2. The amplification of bisulphite-treated DNA yields four different combinations of paired-end flag (1 or 2) and mapping strand (+ or –; Figure **2.3**). Reads originating from the original bisulphite-treated strands (1+ and 1–) can only be generated by the Illumina primer for read 1 and their mapping strand determines the type of conversion the read contains (either T-to-C for forward or A-to-G for reverse mapping reads; Figure **2.3**). Thus, when the paired-end flag and the mapping direction are known, only one of the two possible conversions has to be considered for each seed of the read, thereby reducing the number of modified seeds and the incidence of cross-mappings.

4.3 Determination of methylation rates

Bisulphite-treated reads contain only methylation information about one of the original treated strands. Reads with paired-end flag 1 mapping on the forward strand of the reference and second reads mapping on the reverse strand of the reference derive from the forward bisulphite-treated DNA strand (reads 1+ and 2– in Figure **2.3**), while the remaining reads contain information about the complementary bisulphite-treated strand (1– and 2+ in Figure **2.3**). To obtain strand-specific read counts, the base calls per genomic position have to be summarized separately for both of these groups.

I modified the module 'consensus' of the software SHORE [Ossowski et al., 2008] into a new module 'methyl' that records the number of reads supporting methylation (non-converted bases) and the number of reads indicating non-methylation (converted bases) for each cytosine on both reference strands. The estimated methylation rate of a site is calculated as the fraction of methylation supporting read bases from all read bases overlapping this site. To prevent the multiple counting of PCR duplicated reads, which can distort the quantitative measurement of the true methylation rate, the contribution of a set of duplicated reads at a position to the total read count is set to maximally 1. It can be less than 1 if duplicated reads show different base calls. In this case, the contribution of the set of duplicated reads is the ratio of methylation-supporting bases to all bases in this subset of reads; this accounts for the uncertain base call.

SHORE 'methyl' reports a methylation rate for different subsets of base calls. Most commonly, only uniquely mapping reads are used in methylation analyses. Additionally, SHORE 'methyl' can disregard a certain number of first and last bases of the reads and restrict analysis on the so-called 'core' region of reads, since read ends are

enriched in alignment errors (section 2.2.4). Lastly, to assess the mapping reliability of base calls, SHORE's genetic variant scoring scheme from the 'consensus' module was adapted in 'methyl' to provide a quality score (section 2.2.4).

4.4 Identification of methylated positions (MPs)

The estimated methylation rate obtained from the reads typically does not reflect the exact true methylation rate. It is influenced by the stochastic sampling of reads from heterogeneous input DNA and by bisulphite reaction conversion errors (section 2.4.2). While the rate of false positive bisulphite conversions is generally assumed negligible, the rate of incomplete, i.e. false negative, conversion events is appreciable. Incomplete bisulphite conversion leads to overestimating the methylation rates since unmethylated sites appear as methylated in the reads. Thus, a false negative bisulphite conversion equals a false positive methylation call. The scale of this bias can be easily assessed as the rate of (false) methylation on known unmethylated sequences such as spiked-in bacterial DNA, non-CG sites in most human samples, or plastid sequences such as the chloroplast in plants.

Following the best practices discussed in section 2.4.2, I implemented a method that determines statistically significantly methylated positions (MPs) by testing how likely the observed methylation rate at a genomic position can be explained by the false methylation rate ascertained at chloroplast sites. To allow for variation in read counts due to stochastic read sampling, the number of methylated reads is modeled using a binomial distribution. The maximum likelihood estimate (MLE) of the mean binomial rate of false positive methylation on *A. thaliana* chloroplast sites equals the sum of methylated reads across all chloroplast sites (S) divided by all reads:

$$\text{MLE} = \frac{\sum_{s \in S} m_s}{\sum_{s \in S} m_s + u_s}$$

where m_s and u_s are the number of methylated and unmethylated reads at site s, respectively.

I noticed, however, that most chloroplast positions are highly covered and have near-zero methylation levels (Figure **4.2**). These highly covered sites dominate the overall average false methylation rate, which leads to a markedly overestimated number of methylated sites in the nuclear genome for low-coverage sites. Therefore, I cluster chloroplast sites into coverage bins of multiples of fivefold and calculate the MLE of the false positive methylation rate (FMR) per bin (Figure **4.2**). This conservative, coverage-dependent testing represents a major difference to previous efforts.

For all cytosine sites in the nuclear genome, P values are calculated by binomially testing the focal read counts against the false methylation rate of the respective coverage bin. After multiple testing correction of the genome-wide P values by calculating q values using the R package qvalue [Dabney and Storey, 2000], I define sites as statistically significantly methylated positions (MPs) if their methylation rate distribution

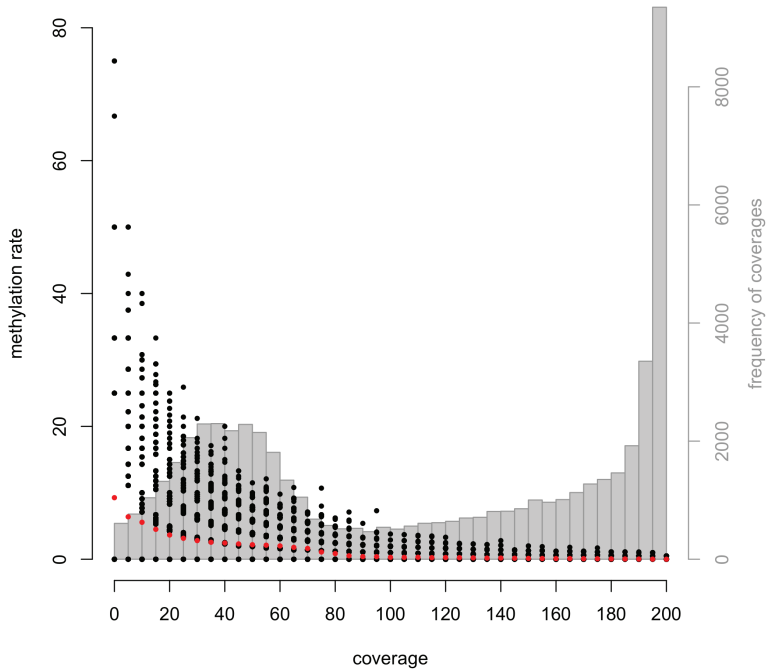

Figure 4.2: False methylation rate. False methylation rate by read coverage (black dots) and sequencing depth distribution (grey bars) at chloroplast sites of an *A. thaliana* sample (line 49), analyzed in chapter 5. Red dots represent the mean false methylation rate for each coverage bin (multiples of fivefold). Note that since reads had a maximum length of 100 bp and PCR duplicate reads were removed, the maximum coverage at each site was 200 (equaling maximally 200 different start positions of reads overlapping a single site).

is significantly different from the false methylation rate distribution, determined by a q value cutoff of 0.05 (adjustable). The q values are equivalent to a false discovery rate (FDR).

4.5 Identification of differentially methylated positions (DMPs)

High absolute methylation rate differences between samples are not necessarily indicative of strong differential methylation. An estimate of a methylation rate on low-coverage sites has a much higher variance caused by stochastic read sampling. Thus, a comparison of the methylation signal between samples should account for read coverage. I therefore apply a two-tailed Fisher's exact test, a significance test based on a 2x2 contingency table containing read counts that is suited for evaluating if an observed correlation might be random or real. The general strategy is to test cytosines in a pairwise sample comparison to obtain P values, followed by multiple testing correction by Storey's method, which calculates a false discovery rate for each

test [Storey and Tibshirani, 2003]. Sites at a maximal FDR (default: 5%) are defined as differentially methylated positions (DMPs).

The statistical power in determining DMPs mainly depends on two factors: the read coverage and the number of tests performed. Lower sequencing depth usually yields higher P values due to a higher read sampling variance. Correcting this bias cannot be performed on single sites, unless additional information are given or assumptions are made, e.g. about a correlation with the methylation rates of nearby sites. Without this information, subsampling the data can give a rough evaluation of a false negative rate of detecting DMPs, e.g. if or when a saturated state in the number of DMPs is assumed to be achieved.

The dependency of statistical power on the number of tests manifests in increased corrected P values (i.e. reduced power) with an increasing number of statistical tests performed. The challenge of retaining as much power as possible can be tackled in two ways: first by avoiding the correction of all P values at once and second, by limiting the number of statistical tests performed.

I employ an iterative procedure of two-level multiple testing correction that prevents correcting all performed tests at once, while nonetheless accounting for the fact that tests are performed across genome-wide cytosines and across pairwise comparisons. The method starts by adjusting the P values across all pairwise comparisons using Bonferroni correction at each tested genomic position. Then, the adjusted minimum P values at each genomic position are recorded in a vector. Consequently, this vector contains as many P values as there are analyzed cytosines in the genome. I calculate q values for all P values in the vector using Storey's method and report between-sample DMPs at a maximal FDR cutoff of 5%. The P values of the pairwise comparisons that are included in the vector are set to 1 and the procedure can be repeated to obtain a new vector and further differential pairwise comparisons on variable sites, up to a user-defined number of times (usually number of pairwise comparisons).

In addition to this multiple P value correction scheme, statistical power can be increased by limiting the number of statistical tests on those positions where there is a reasonable chance to retrieve low P values. This can be achieved by restricting tests on sufficiently covered sites (typically >3x) and sites that are methylated in at least one sample (both was done in our study, chapter 5). Additionally, a minimum absolute difference in methylation rate can be required, as was done in an analysis of larger genomes [Seymour et al., 2014].

Sequencing biological replicates helps in assessing the stochastic variance of methylation rates and substantiates methylation levels. Highly variable positions between replicates might lead to false positive calls of differential methylation between different samples. However, Fisher's exact test cannot incorporate replicate data and is usually performed on the accumulated read counts across replicates. One possibility to account for sites exhibiting the highest variance between replicates is to first identify DMPs between replicate samples by Fisher's exact test and then exclude these sites from the DMP analysis between different samples.

4.5.1 Identification of DMP clusters (DMCs)

Differential methylation at single cytosines might not be informative or functionally important. To explore whether and how many variable positions concentrate to specific regions in the genome, a simple distance-based consolidation of DMPs is implemented. Based on pairwise comparisons against a reference sample, I merge adjacent DMPs for a sample within a user-defined genomic distance to each other into clusters, perform a Fisher's exact test on the average read counts across sites in the region (accumulated over potential replicate data), and call sample-specific DMP clusters (DMCs) at a maximal FDR (5%). The set of DMCs can be further filtered by a minimum length and by the requirement to contain a minimum number of methylated and differentially methylated positions. Furthermore, DMCs from different samples are merged if they overlap by a specific fraction of their combined length and if the methylation change is in the same direction compared to the reference sample.

Please note that DMCs have been referred to as differentially methylated regions (DMRs) in most literature and in the introduction of this work (section 2.4.4). However, to better distinguish these regions from the DMRs obtained by my novel approach (explained in the following), I decided to rename the DMP-based regions.

4.6 Identification of differentially methylated regions (DMRs)

Most tools for detecting differentially methylated regions, including the method to identify DMCs in the previous section, are based on testing (commonly millions of) single cytosine sites and clustering them using arbitrary filter criteria (see section 2.4.4). We conceived an approach that first defines methylated regions and then compares them between samples, thereby limiting the number of tests to the methylated space and vastly reducing the number of multiple testing corrections. By using an unsupervised Hidden Markov Model (HMM) to call methylated regions, we bypass the need for setting any filter criteria to select regions to test. Moreover, since it combines information from neighboring sites, information of low-coverage regions contributes to the analysis and is not discarded, in contrast to most other approaches.

My DMR detection method starts by identifying methylated regions (MRs) for each sample separately by applying an HMM that approximates the methylated read counts with a beta binomial distribution. The method proceeds by selecting segments that are in different or highly methylated states between samples to statistically test them for differential methylation in all pairwise comparisons to retrieve significantly differentially methylated regions (DMRs). Finally, to summarize pairwise comparisons, I combine samples into epiallele groups for each DMR and test these groups again for significant differential methylation.

Hidden Markov Model (HMM)

The process of classifying a sequence of observations into discrete states can be achieved by a stochastic process termed a 'Markov chain'. States can change along the sequence of observations according to transition rules that solely depend on the current state of the system ('Markov process'). In Hidden Markov Models (HMM), the states have distinct probability distributions that lead to different outputs for the different observations ('emission probabilities'), and the states themselves cannot be observed (are 'hidden'). The parameters of an HMM consist of the number of states, the emission probabilities (determined by the parameters of each state's specific distribution), the transition probabilities and probabilities that determine in which state the HMM begins. Given initial start probabilities and prior distributions of each state's probability functions, an HMM can iteratively learn and refine the parameters of each state's distribution as well as the start and between-state transition probabilities on the actual data (the sequence of observations) without any training data set (it is 'unsupervised'). The transfer of the general principles to the calling of methylated regions is presented in the next section. In-depth explanations of Hidden Markov Models and the associated well-established algorithms are given in references [Durbin et al., 2007, Rabiner, 1989].

4.6.1 Identification of methylated regions (MRs)

To find hypomethylated regions in human samples, Molaro and colleagues implemented an HMM [Molaro et al., 2011], which I adapted to the distinct methylation pattern of *A. thaliana* (see below). The sequence of observations for this HMM consists of methylated and unmethylated read counts retrieved from bisulphite sequencing along the genome. The HMM utilizes a methylated and an unmethylated state, each containing specific beta binomial distributions that model the read counts (see section 2.4.3). It learns the parameters of these distributions and simultaneously estimates transition probabilities between the two states from genome-wide data at cytosine sites in an unsupervised manner, and irrespective of sequencing depth. The HMM is initialized with uniform start probabilities (0.5), specific transition probabilities (0.75 to remain in and 0.25 to switch between states) and beta binomial distributions shifted towards higher (methylated state) or lower methylation (unmethylated state). The training is performed iteratively (using the Baum-Welch algorithm) until probabilities converge (deviation $< 10^{-10}$) or until a user-defined number of steps is reached. In each iteration, the maximum likelihood parameters of the beta binomial distributions are estimated on the methylation ratios as if they were for a beta distribution. This allows the application of closed formulas [Molaro et al., 2011], which avoids computationally expensive numerical optimizations. On the final trained model, the "Posterior Decoding" method determines the most probable state at each position of the genome, resulting in a segmentation of the genome into regions of high and low methylation. The algorithms of the HMM are implemented in C++ exactly as specified in reference [Durbin et al., 2007].

The program of Molaro and colleagues performs post-processing steps to remove two biologically spurious cases [Molaro et al., 2011]. First, the HMM has no genomic distance information. It might not be desirable to span long uncovered genomic stretches between two consecutive positions in a methylated state. Thus, a user-defined maximal distance (termed 'desert size') splits the genome into consecutively covered parts, on which the HMM processes on independent 'sequences of observations'. Second, since the HMM can theoretically call very short hypomethylated regions down to single base pairs, Molaro and colleagues provide a method to test for the significance of such regions. To do so, hypomethylated regions are scored by the sum of the inverted methylation rates (1−methylation rate) within the region as a rough quantitative measure of the 'degree' of hypermethylation. These scores are compared against an empirical distribution of scores of hypomethylated regions obtained by random permutation of all methylation ratios throughout the genome and applying Posterior Decoding. From resulting P values for each region, false discovery rates are calculated using the Benjamini & Hochberg method and regions with a maximum FDR of 5% are retained as final hypomethylated regions (i.e. showing a higher 'degree' of hypomethylation than 95% of random hypomethylated regions).

The HMM implementation by Molaro *et al.* was tailored to human samples, where methylation is almost exclusively restricted to cytosines in the CG context and the vast majority of CG sites are methylated (section 1.6). To apply this method to plant methylation data, I modified their program in three ways. First, methylation in plants occurs in all three sequence contexts, whereby each context has a distinct methylation rate distribution across all sites of the genome (see Figure **5.5**C). Thus, the modified HMM models methylation rates for each sequence context separately, requiring three separate beta-binomial distributions. At each position in the genome, the two distributions of the respective sequence context of that site are used to calculate emission probabilities of the two states. The initializing beta binomial distributions are the same across sequence contexts. Second, as opposed to humans, only a small portion of cytosines are methylated in *A. thaliana*. Thus, the adapted HMM version inverts the methylation rates to find hyper- rather than hypomethylated regions, hereafter only referred to as methylated regions (MRs). Third, I included options for altering the borders of methylated regions. A user-specified distance between two methylated regions leads to their merging, and weakly methylated positions (user-defined threshold) can be trimmed at both ends of MRs.

4.6.2 Selecting regions to test for differential methylation

The identification of methylated regions yields a single segmentation per sample genome. Overlapping methylated regions between samples might not share the same start and end coordinates due to variability of coverage and methylation rate. An ultimate clustering of samples into epialleles requires a unified set of genomic regions to be tested in all pairwise sample comparisons.

To obtain such a set, I generate the unified MR space across all samples and define each resulting, potentially enlarged methylated region as a 'methylation island' (Figure **4.3**). For each methylation island, all combinations of start or end coordinates of the contained MRs define a set of regions (reg1-5 in Figure **4.3**). However, only regions that are completely covered by an MR in at least one sample are retained (this is why the entire methylation island is not tested in Figure **4.3**).

This can result in a large number of regions. To reasonably reduce this number, I defined greedy filter criteria to avoid the analysis of short regions and regions of low or unbalanced coverage, and to prevent the redundant analysis of nearly identical segments as follows (illustrated in a toy example in Figure **4.3**):

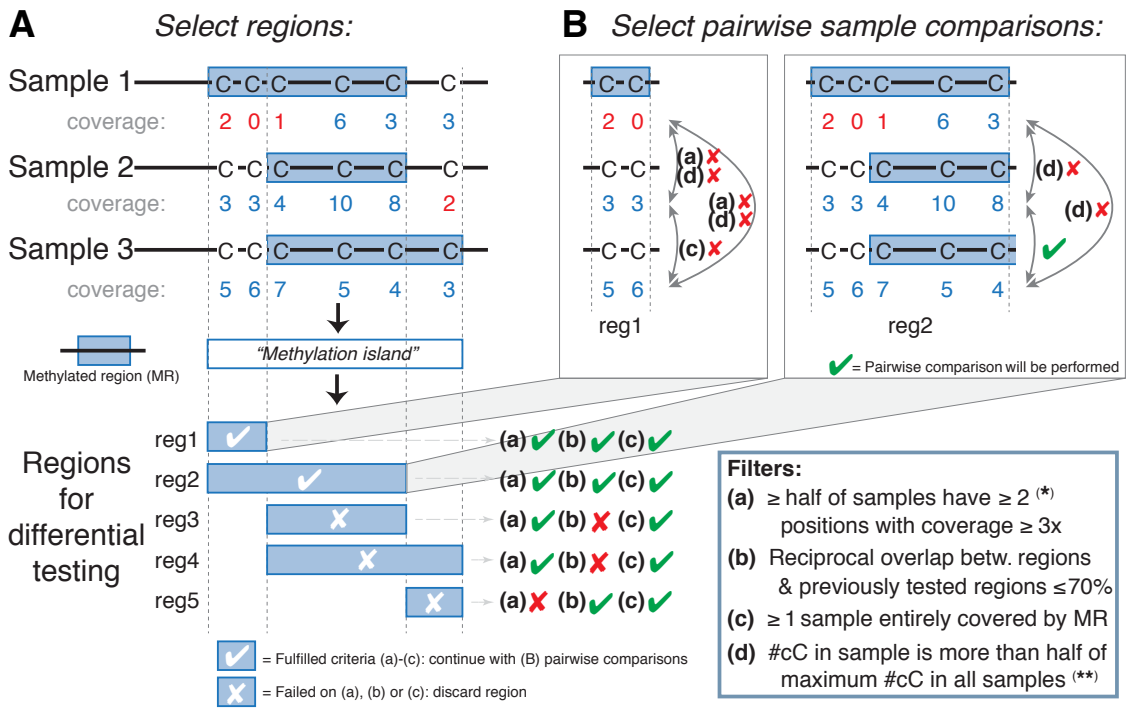

Figure 4.3: Selection and filtering of genomic regions in preparation for differential methylation testing. (A) Three methylated regions in three samples result in 5 segments for further analysis, from which only two fulfill the filters (a)-(c), listed in the 'Filters' box. Sites covered by less than three reads, marked by red-labeled coverage values, are considered uncovered, the others covered (blue). Parameters can be adjusted in the software implementation. (*): For simplicity, the illustration requires a minimum number of covered sites of two per region (10 sites for the real data set). (B) Filters (a), (c) and (d) are checked for each pairwise comparison, whereby filter (a) now requires the coverage criterion for each of the two samples. Criteria marked by a cross fail the test, criteria not listed are fulfilled for each pairwise comparison. (**): #cC denotes the number of covered cytosine sites in a region (i.e., having ≥3x coverage).

- Regions were discarded from any pairwise comparison if less than 2 samples contained at least 10 cytosines covered by at least 3 reads each (accumulated over strain replicates) in this region (Figure **4.3** (a)).

- Regions were discarded from any pairwise comparison if the reciprocal overlap of this region to at least one previously tested region was more than or equal to 70 %. This was done to prevent "similar" regions to be tested twice (Figure **4.3** (b)).

- Pairwise tests of a region were not performed if both strains were in low methylation state throughout the whole region (Figure **4.3** (c)).

- Strains were excluded from pairwise comparisons in a region if the number of positions covered by at least 3 reads each was less than half of the maximum number of such positions of all strains in the same region. This prevented comparing regions with unbalanced coverage to each other, e.g. a strain with 10 data points against another one with only 2 (Figure **4.3** (d)).

The parameters of the filter criteria can be adjusted in the software implementation.

4.6.3 A statistical test for differential methylation

Regions that pass the filter criteria described in the previous section are statistically tested for differential methylation in all pairwise sample comparisons. To also detect quantitative differences rather than solely presence/absence of methylation, I also compare entirely methylated regions in two samples to each other (e.g., reg2 in Figure **4.3**). Jonas Müller designed and implemented a log-likelihood ratio test for differential methylation of a given region between two samples. This method assumes that read counts along a genomic region follow a beta binomial distribution (section 2.4.3). It fits separate distributions for each of the three sequence contexts for each sample using a custom gradient-based numerical optimization method that determines the maximum likelihood estimates of the distribution's parameters. When replicate data is available, the distributions are approximated on the read count data of all replicate samples, increasing statistical power and yielding more accurate estimates of the biological variance. Moreover, for each sequence context, a joint distribution is fitted to the data of both samples together, treating them as replicates. Thus, there are nine beta binomial distributions per pairwise comparison: two times three sample-specific and three joint distributions (for each sequence context). The statistical test calculates the log odds ratio between the sample-specific and the joint likelihoods of the read count data of the region, using the corresponding beta binomial distributions. This ratio is calculated as follows:

$$\sum_{c=1}^{3} \log \frac{\max_{a,b}\left(\sum_{p=1}^{N_{Ac}} \binom{C_{Acp}}{x_{Acp}} B(a+x_{Acp}, b+C_{Acp}-x_{Acp})\right) \cdot \max_{a,b}\left(\sum_{p=1}^{N_{Bc}} \binom{C_{Bcp}}{x_{Bcp}} B(a+x_{Bcp}, b+C_{Bcp}-x_{Bcp})\right)}{\max_{a,b}\left[\left(\sum_{p=1}^{N_{Ac}} \binom{C_{Acp}}{x_{Acp}} B(a+x_{Acp}, b+C_{Acp}-x_{Acp})\right) \cdot \max_{a,b}\left(\sum_{p=1}^{N_{Bc}} \binom{C_{Bcp}}{x_{Bcp}} B(a+x_{Bcp}, b+C_{Bcp}-x_{Bcp})\right)\right]}$$

where N_{Sc} is the total number of cytosines of sample S in context c and C_{Scp} the number of reads at position p in context c, from which x_{Scp} are methylated.

This ratio is compared against a chi-squared distribution (with 6 degrees of freedom) to obtain P values, followed by FDR calculation with Storey's method [Storey and Tibshirani, 2003].

Calling DMRs

Since the statistical test is rather sensitive, regions are selected at a low FDR threshold of 1%. To improve the confidence of detected differences further, I additionally compare 'empirical confidence intervals' of each context-dependent distribution between both samples. This interval is defined as the mean plus and minus twice the standard deviation of the sample-specific beta binomial distributions. Differentially methylated regions (DMRs) are defined as those for which the confidence intervals of the samples do not overlap.

Since there is a beta-binomial distribution for each sequence context, the method allows calling differential methylation in a region for single contexts, termed CG-DMRs, CHG-DMRs and CHH-DMRs. C-DMRs are regions in which more than one context is statistically different.

4.6.4 Identification of epiallele groups

For comparisons of multiple samples, DMRs between the many pairwise comparisons could be difficult to summarize and interpret. Therefore, I developed a method utilizing graph theory to summarize pairwise comparisons and to obtain epiallele frequencies for each DMR.

For each region identified as a DMR in at least one pairwise comparison, the strategy is to construct a graph where vertices represent the samples and undirected edges connect samples if they feature significantly differential methylation (Figure **4.4**). The algorithmic challenge is to find the smallest number and composition of groups of vertices so that no two members of the same group are connected to each other (in graph theory, this number is the 'chromatic number' of the graph). In other words, DMRs cause separation into groups, and there is no DMR between samples within a group. This strategy assumes that samples within a group are then similarly methylated.

This problem is known as the "vertex coloring problem" and is NP-hard, but can be easily solved by 'brute force' for a few tens of vertices, i.e. samples. I implemented an exact recursive algorithm by iteratively increasing the number of different colors, i.e. groups, starting with two and stopping once all samples have been successfully assigned a color. In each iteration, all legal combinations of assigning the iteration-dependent number of colors to samples are tested recursively. The algorithm processes samples in descendent order of their degree (i.e. the number of edges a vertex is connected to), which speeds up the discovery of potential 'collisions'. In case there are different

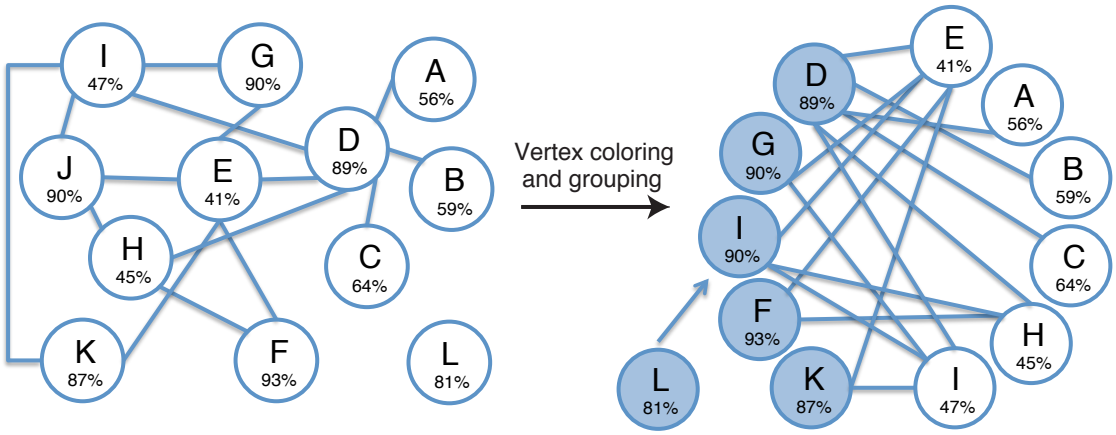

Figure 4.4: Illustrative example of the assignment of samples into different groups based on differential methylation. Left: An exemplary differentially methylated region (DMR) represented as a graph: a vertex reflects a sample and an edge a statistically significant test between two samples. Percentages denote the sample means, i.e. the accumulated means of the methylation rate distributions for each context. For simplicity, only a single sequence context is indicated here. Right: The solution of the "vertex coloring problem", i.e. finding the minimal number of sets of vertices (colors) so that no edge connects vertices of the same group. Groups are illustrated as blue and white vertices. The group mean of the blue samples would be (89+90+90+93+87)/5 = 89.8% and of the white group (41+56+59+64+45+47)/6 = 52%. Sample L shows no significantly differential methylation to any other sample and is assigned to the blue group, since its sample mean (81%) is closer to the blue group mean (90%) than to the white one (52%). The grouping diversity of the clustering shown here would be (from samples A to K): $(|56-52|+|59-52|+|64-52|+|89-89.8|+|41-52|+|93-89.8|+|90-89.8|+|45-52|+|47-52|+|90-89.8|+|45-52|+|87-89.8|)/11 = 2.84$.

possible groupings of the samples with the same number of colors, the clustering with the most similar methylation rates per group is selected. This measurement is roughly quantified by defining the following variables:

- The *sample mean* is the summation of the three mean values of the approximated beta-binomial distributions of each sequence context for a sample, i.e. the sum of the mean CG, CHG and CHH methylation rates in a region.
- The *group mean* is the average of the sample means of the group members.
- The *grouping diversity* is the accumulated deviation of all individual sample means from their respective group mean.

Thus, in case of conflicting groupings of samples, the clustering with the lowest grouping diversity is selected. Figure **4.4** illustrates this approach and the diversity of measurements in an example.

A sample that is not significantly differentially methylated from any other sample (i.e. its vertex has no edge) is assigned to the group where the group mean is closest to its sample mean (sample L in Figure **4.4**).

In some cases, I observed sporadic DMRs between only a few pairwise comparisons. To test whether a grouping reflects a clear separation of samples into putative epialleles and to distinguish this case from situations where the region exhibits many intermediate methylation levels in the population of samples, the same statistical test is used between the groups of samples as is also applied for the individual sample comparisons (see previous section 4.6.3). This selects for statistically significant groupings.

To this end, the method approximates group-specific beta binomial distributions from the read counts of all group members, thereby treating the samples in one group as replicates. DMRs of groups are called at a maximal FDR (5%) using Storey's method [Storey and Tibshirani, 2003].

Finally, since the selection of regions tested for differential methylation regularly produces overlapping segments (Figure **4.3**A), DMRs between groups of samples can overlap. I provide two ways in resolving superposed DMRs. From sets of overlapping DMRs, one can choose the non-overlapping DMRs with the lowest q values, or the non-overlapping DMRs with the highest number of samples that show significantly different methylation from any other sample(s).

4.6.5 Identification of highly differentially methylated regions (hDMRs)

Despite the many filtering steps for the final DMR set, some of the DMRs show only subtle methylation differences. This might occur if the approximation of the methylation rate distributions is based on many data points, for example in long or highly covered regions, leading to high confidence and low variance. Since small differences in methylation levels have dubious effect on transcription and therefore unclear biological significance, and since most reported naturally occurring epialleles show an obvious difference in methylation rate, I apply one last filter step. I define highly differentially methylated regions (hDMRs) as those DMRs that are longer than 50 bp, show more-than-three fold methylation rate difference in at least one sequence context when analyzing minimally five cytosines, and where the higher methylation rate of both samples exceeds 20%. This filters for more obvious cases of highly versus lowly methylated samples by avoiding short, potentially spurious variably methylated regions and regions of high fold-change, but low absolute difference (e.g. 2% versus 10% methylation).

Chapter 5

The rate and spectrum of natural DNA methylation variation in two *A. thaliana* populations

The study of DNA methylation variation in natural populations is important to understand the role of epigenetics in creating the extensive natural variation found within or between species. Naturally occurring DNA methylation differences can arise spontaneously, linked to DNA sequence changes, or due to environmental cues, and can be inherited over many generations (chapter 1). To disentangle the contributions of these factors in shaping the epigenetic landscape of *A. thaliana*, we analyzed the methylomes of two populations that represented unique settings, in which specific sources of epigenetic variation were kept minimal. This allowed us to mainly address two questions: how do spontaneously occurring epialleles behave, and how does a natural environment affect the pattern and emergence rate of epialleles.

The chapter starts by briefly introducing the data sets, then elucidates the genetic architecture of the populations to confirm their low genetic diversity and subsequently explores the DNA methylation and differential methylation landscapes.

Contributions

This chapter describes the studies that are included in two publications: [Becker et al., 2011, Hagmann et al., 2015]. I conceived the studies together with Prof. Detlef Weigel and Dr. Claude Becker. I performed the primary analyses of DNA sequence polymorphisms and DNA methylation calls according to the methods presented in chapters 3 and 4. Dr. Christa Lanz and Dr. Claude Becker performed Illumina sequencing. Dr. Claude Becker prepared the sequencing libraries and performed the transcriptome analysis. Dr. George Wang retrieved climate data. Dr. Oliver Stegle provided the program to calculate heritability values of differentially methylated regions. Dr. Rhonda Meyer and Prof. Thomas Altmann provided the phenotypic data. I

collaboratively performed and discussed all other analyses leading to most figures with Dr. Claude Becker and Prof. Detlef Weigel.

5.1 Data sets

5.1.1 Mutation accumulation lines

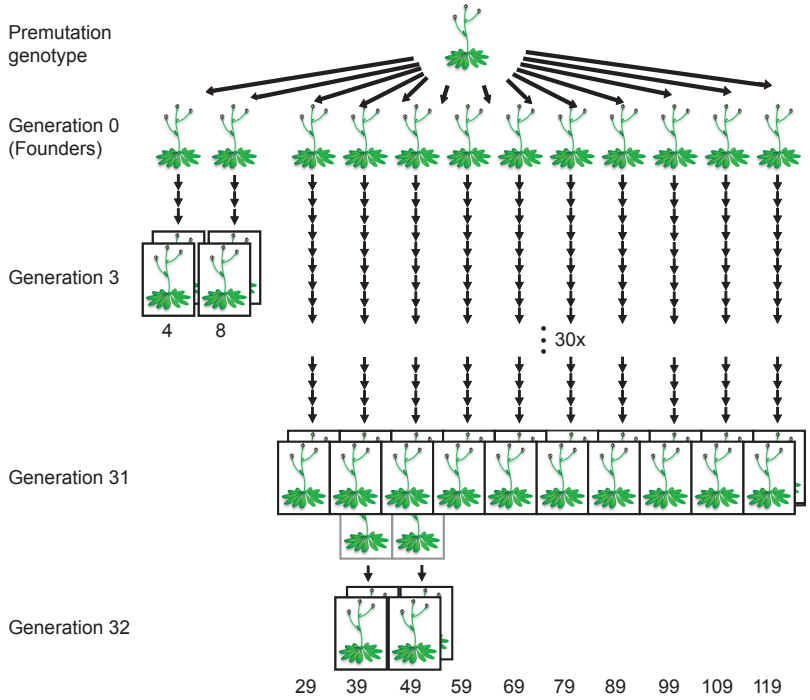

Figure 5.1: Experimental design of the mutation accumulation line study. Strains were derived from a single parent. The seeds were propagated by single-seed descent, with separate lineages for the 3rd and 31st generation individuals. Strains 39 and 49 were propagated for one more generation from siblings of the plants analyzed by sequencing (grey outlines).

We analyzed a set of ten *A. thaliana* mutation accumulation (MA) lines that had been propagated from a single individual by inbreeding for 30 generations in a uniform and benign greenhouse environment. The ancestor plant was genetically identical to the reference strain Col-0 that was used to generate the high-quality *A. thaliana* reference genome [Arabidopsis Genome Initiative, 2000]. Since seeds from the founder plants were not available, we assessed whole-genome cytosine methylation of two 3rd generation plants and compared it to the methylomes of ten independent 31st generation plants (Figure **5.1**). Moreover, we inspected the methylomes of direct descendants of two generation-31 plants (generation 32).

Hence, since the plants of the population were isogenic and experienced highly similar growth conditions, accumulated DNA methylation changes between generations should reflect spontaneous epimutations, largely independent of genetic and environmental influences.

In a previous study, Ossowski and colleagues had analyzed five of the ten 31st generation lines used in this study and reported that spontaneous genetic mutations arise at a rate of roughly one SNP per haploid genome and generation [Ossowski et al., 2010]. To verify the method to detect genetic variants presented in chapter 3 and to confirm the genetic mutation rate with higher coverage and longer read lengths, we also sequenced the genomes of our ten lines.

5.1.2 The haplogroup-1 population

The native range of the species *A. thaliana* is in Europe and Central Asia, and it has been naturalized to North America. In Eurasia, despite being mostly self-pollinating and thus homozygous at most loci, their genetic diversity is high and exceeds that of humans [Mitchell-Olds and Schmitt, 2006]. Near-isogenic strains are generally found only at the same geographic location. By contrast, about half of individuals sampled across North America were identical at 139 genome-wide markers [Platt et al., 2010]. Hence, we anticipated their potential to constitute genetically near-identical lines, grown at dispersed locations under varying natural environmental conditions for the past few centuries. I hereafter refer to this lineage as haplogroup-1 (HPG1), because it might dominate the North American population.

We selected 13 HPG1 individuals from seven locations in the Eastern Lake Michigan area, from one location in Western Illinois, and one location on Long Island. The maximum distance between sites was ∼400 km, the median ∼150 km. The set consisted of pairs of accessions from each of four sites, and single individuals from the other five sites (Figure **5.2**).

We were interested in long-term heritable patterns of DNA methylation, not in the potentially unstable, e.g. development-dependent, patterns the plants might have shown at their respective site of origin. Therefore, we grew plants under controlled conditions for two more generations after collection at the natural sites to erase potential parental effects on the DNA methylation pattern, e.g. derived from soil, climate or light. We performed whole genome, methylome and transcriptome sequencing of leaf tissue of these plants. For the methylome libraries, we pooled 8-10 individuals per replicate to reduce inter-individual methylation variation and fluctuations in methylation rate caused by stochastic coverage or read sampling bias.

To characterize the climatic conditions of the plants at their natural sites, we collected data from the nearest weather stations for the growing season preceding the collection of the plants and observed considerable variation in precipitation and temperature (Figures **S1** and **S2**). However, the analyzed plants did not show any visible phenotypic differences.

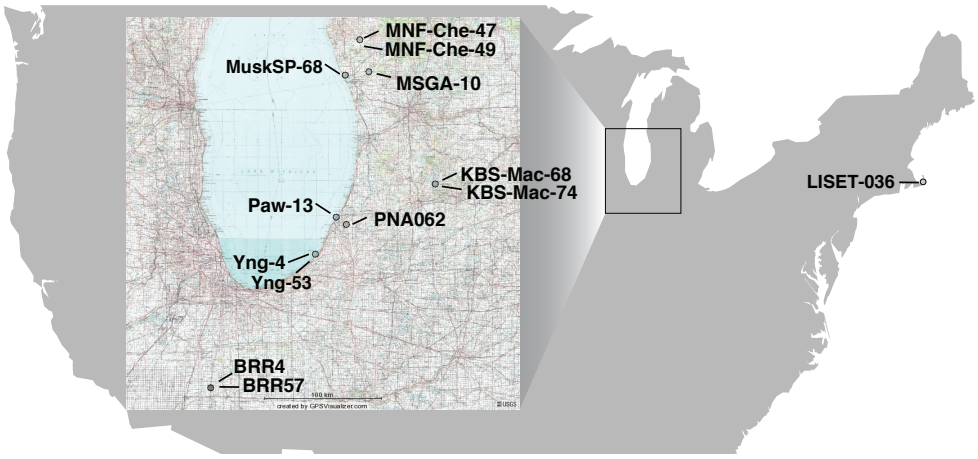

Figure 5.2: Sampling locations of the 13 haplogroup-1 (HPG1) strains in the United States of America.

Together, this population seems to constitute a collection of natural mutation accumulation lines, well suited to assess the long-term impact of natural environmental fluctuations on the heritable fraction of DNA methylation under minimal genetic influence.

5.2 Spontaneous DNA sequence changes in the mutation accumulation lines

We sequenced the genomes of the MA lines with 2x101 bp reads with a read depth of 40-fold on average (Table **S3**). I applied diverse structural variation (SV) detection tools and validated the SV calls by re-alignment of reads, following the workflow described in chapter 3 without repeated iterations. This procedure identified 2,203 polymorphisms that were shared between all strains, indicating errors in the reference sequence (12% of variants replaced N's in the TAIR9 genome) or genetic differences in the founder plant of the MA population compared to the Col-0 individual that had been used for the reference genome. By contrast, only 388 segregating variants were detected (76% SNPs, 14% deletions and 10% insertions). However, solely five deletions and one insertion longer than 10 bp were among those variants.

Since in genomes the size as that of *A. thaliana* (120 million bases) chances are negligible that a mutation independently occurs in more than one lineage within a few generations, it is expected that true variants are unique to one line. Out of the 388 segregating variants, 350 had an allele frequency of 1. Since this analysis was based on longer reads and higher sequencing coverage, I identified on average 33.6 variants per 31st generation line (Table **S1**), compared to 19.6 detected previously

[Ossowski et al., 2010]. This number is even closer to 1 SNP per generation and *A. thaliana* genome.

Validation of the genetic detection pipeline

I used this data set to roughly validate the genetic detection workflow of chapter 3. The fraction of variants being 'non-unique' to a single line can serve as an error estimate, although Ossowski and colleagues discussed the possibility that such variants might exist, potentially resulting from heterozygous positions in the founder plant [Ossowski et al., 2010]. Furthermore, in the previous study, variants found in all but one line were also reported as spontaneous mutations in regions of incorrect reference sequence. I identified eight variants with an allele frequency of 11 by using my pipeline (Table **S2**). This resulted in an approximated 'error rate' of 7.7% (30 out of 388), when assuming that the variants with allele frequencies 1 and 11 are true. Nonetheless, as this rate might reflect a combination of false positive rate (for low allele frequencies) and false negative rate (for high allele frequencies), I carefully deem the false positive rate of my genetic variant detection pipeline to be well below 10%.

To gauge the false negative detection rate, I calculated the overlap of the genetic variants that were unique to a single line with those reported previously [Ossowski et al., 2010]. Of 116 variants called and validated by Sanger sequencing in five mutation accumulation lines, 114 had read coverage in our analysis and 102 were also detected. This implies that my approach missed 10.5% of the previously identified variants. Ossowski and colleagues reported four long deletions that cannot be called by SHORE. However, the pipeline presented here reported only one 15-bp deletion as a 12-bp deletion one position distant to the true location. The remaining three were called at approximate lengths and positions by SV callers, but failed the filtering steps of the pipeline, since they overlapped with 'inner core' regions of resequencing reads (sections 3.2.1 and 3.3), i.e. the deleted region was not devoid of mapping reads. My pipeline did not call one insertion (out of five), since it failed the test for homozygosity against a binomial distribution with a mean of 95%, despite having 91% concordance of its alternative base. This was presumably due to high read coverage resulting in high confidence. From the remaining 7 missed variants, one deletion was detected at a position 3 bp near the true location, and the others were not found due to more stringent quality thresholds applied in SHORE for this study.

Hence, I conclude that the filtering steps and criteria of the genetic variation detection strategy of chapter 3 were overly conservative, owing partly to the different incentives of the detection algorithms. My pipeline set the focus on the detection of shared variation and was applied on the MA data set with identical parameters as for the haplogroup-1 population, for which it was designed. Ossowski and colleagues tailored the detection scheme specifically to the mutation accumulation lineages and used ∼500 Sanger-validated sequences to optimize their variant calling parameters.

5.3 Genetic variation between the haplogroup-1 population and Col-0

We sequenced the genomes of the 13 HPG1 lines to an average coverage of 39x (Table **S4**) and I processed reads as described in sections 5.16.1 and 5.16.2. As for the MA lines, I applied the pipeline of chapter 3 to find genetic variation. It is particularly suited to this population, since it takes into account the suggested low genetic divergence of the population [Platt et al., 2010] by testing all strains for the presence of all the variants predicted in any of the 13 lines. To increase the number of detected variants even more, I repeatedly detected 'common' variants found in all strains, incorporated them into a new HPG1 pseudo reference sequence and re-aligned the reads against the updated genome to call 'common' and 'segregating' variants. I analyzed the HPG1 data based on the second refinement of the reference genome after iteration 2, which included four re-alignment steps. For calling SNPs and small insertions and deletions (indels), I used the 'population-aware' SNP calling approach described in section 5.16.3. This iterative procedure reduced the number of unmappable reads by a third, increased the covered genome space by almost 1.4 Mb and the number of common and segregating variants by 12% and 18%, respectively (Figure **5.3**A).

To combine the common variants identified after the final iteration into potentially fewer evolutionary events, I aligned 200 bp around each variant of the last iteration's genome back to the TAIR9 Col-0 reference genome using a global alignment strategy. Ultimately, I identified 670,979 common SNPs and 170,998 insertions and deletions compared to the Col-0 reference sequence, among them 34,021 SVs with a length ≤ 8 bp (maximum: 12,346 bp, median: 15 bp), constituting more than 2 Mb of divergent sequence. The diverse tools that were used to call the genetic variants contributed to the variant calling differently depending on the variant type (Figure **5.3**B), and about half of their detected polymorphisms were falsified by the method described in chapter 3, underlining a rigorous filtering procedure (Figure **5.3**C). The number of insertions and deletions of any length and their distribution to annotation features was nearly identical (Figure **5.3**D, E). Compared to common SNPs, they were less often found in coding sequences and transposable elements, but slightly more often in introns. I found deletions in all 13 HPG1 strains spanning 35 whole genes, 134 entire transposable elements (TEs) and 11 entire non-coding RNAs, relying on the Col-0 TAIR10 annotation.

5.4 Genetic variation among haplogroup-1 strains

Compared to the many shared genetic variants, the iterative re-alignment strategy identified a much smaller number of segregating variants, namely 1,354 SNPs and 521 insertions and deletions. The segregating variants were used to construct a phylogenetic network, which reflected the approximate geographic origin of the strains (Figure **5.4**A): most accessions from the same sampling location clustered together, and only

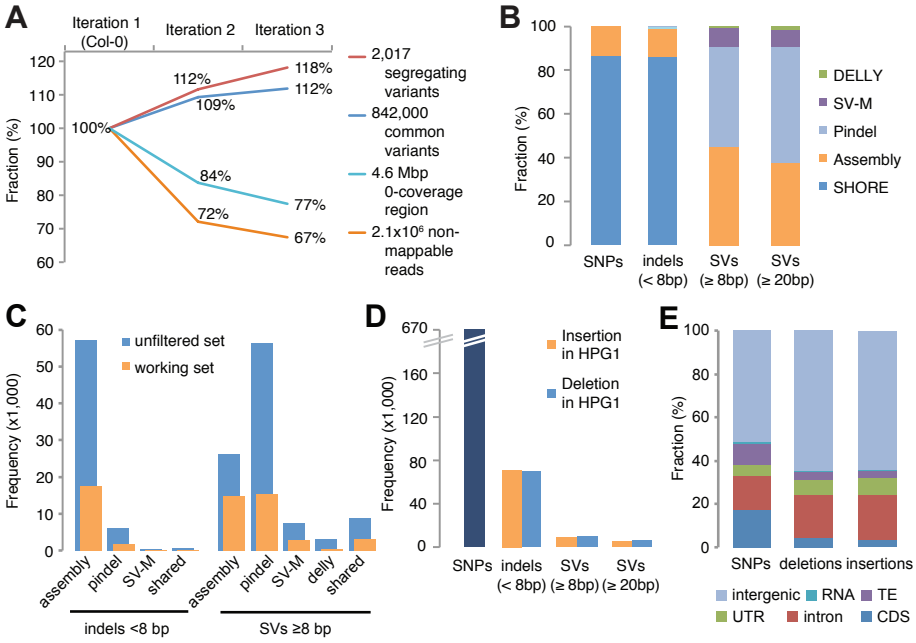

Figure 5.3: Determining and characterizing common genetic variation in the haplogroup-1 population. (A) Increase of detected variants and decrease of unsequenced genome space and unmappable reads by using the iterative mapping strategy of chapter 3. The legend on the right side denotes absolute values after iteration 2. The reference value (100%) derives from the mapping against the Col-0 genome (TAIR9), and for common variants it is the number of variants leading to the genome of iteration 1. Thus, ∼842,000 common variants led to the genome of iteration 2, ∼864,000 to the genome of iteration 3. (B) Composition of common variants by detection tool. (C) Number of predicted and verified insertions/deletions after iteration 1 (read mapping against Col-0) by detection tool. Shared category contains identical variants found by at least two different detection tools. (D) Number of common variants in the pseudo HPG1 sequence compared to Col-0. (E) Annotation of common variants in the pseudo HPG1 sequence.

the Yng strains showed admixture with two other lines (Paw-13 and 328PNA-062). However, the phylogenetic analysis did not reveal associations in larger geographical clusters, but revealed a star-like pattern of relatedness between the different sampling sites. Population structure analysis using 6 clusters mirrored this pattern of haplotype sharing (Figure **5.4**B). Intriguingly, the geographic outlier LISET-036 on Long Island clustered together with the Northern Michigan lines, and had not accumulated more variants than the other strains. On average, two HPG1 accessions differed by 294 SNPs, and strains from the same location by as few as 15 to 130. Those pairs of accessions were responsible for many alleles with a frequency of 2 in the sampled population (Figure **5.4**C). The allele frequency spectrum was very similar between SNPs and structural variants (SVs), indicating a reasonable calling of SVs.

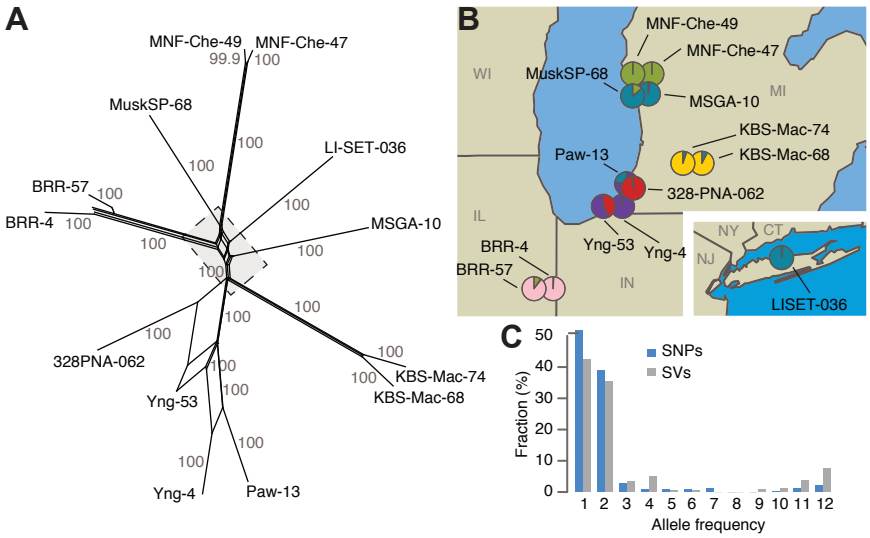

Figure 5.4: Genetic architecture of the haplogroup-1 (HPG1) population. (A) Phylogenetic network of HPG1 accessions based on segregating SNPs and structural variants (SVs) computed with SplitsTree v.4.12.3 [Huson and Bryant, 2006]. Numbers indicate bootstrap confidence values (10,000 iterations). Dashed line delimits close-up in Figure **S3**. (B) Sampling locations of the 13 haplogroup-1 strains. Pie charts indicate population structure inferred from segregating non-synonymous SNP data. We applied STRUCTURE v.2.3.4 [Pritchard et al., 2000], with $K = 6$ after determining the best K value using the δK method [Evanno et al., 2005] (we applied STRUCTURE with $K = 2$ to $K = 9$ with a burn-in of 50,000 and 200,000 chains for 10 repetitions). CT = Connecticut, IL = Illinois, IN = Indiana, MI = Michigan, NJ = New Jersey, NY=New York, WI=Wisconsin. (C) Allele frequencies of SNPs and SVs.

When relying on the spontaneous DNA mutation rate per generation that had been earlier derived from the greenhouse-grown MA lines [Ossowski et al., 2010], the HPG1 strains would be separated from each other between 15 (for replicates) to 384 generations (for distant strains). Thus, assuming a generation time of one year, their most common recent ancestor would have lived around 190 years ago, which is consistent with *A. thaliana* having been introduced to North America during colonization by European settlers. That *A. thaliana* truly populated North America around that time is indicated by the finding of several *A. thaliana* specimen from the mid-19th century in US herbarium collections (D. Weigel, C. Becker, H. Burbano, pers. communication).

The strains did not feature apparent, visible phenotypic differences. We also measured a phenotypic trait, the leaf area, on several days after sowing as described [Hagmann et al., 2015]. This analysis revealed only a marginal association of the trait to the segregating genetic variants (Figure **S4**). Thus, as the genetic variance is low and there seems to be no obvious phenotypic variation, we conclude that the HPG1 accessions constitute a near-isogenic population.

5.5 Spectrum of DNA methylation

We sequenced two replicates per strain at an average strand-specific depth of 20.0x for the MA and 17.7x for the HPG1 populations and filtered the sequencing reads as described in section 5.16.1. I aligned the high-quality reads of the MA lines against the *A. thaliana* TAIR9 reference sequence and the reads of the HPG1 strains against the HPG1 pseudo reference sequence (section 5.16.2). Using the pseudo reference genome instead of the Col-0 sequence for the HPG1 lines increased the number of sufficiently covered cytosines for statistical analysis by 5% on average, and the number of positions called as methylated by 7% (Table **S5**). Next, I retrieved positions that were covered by at least three reads and had a sufficient SHORE quality score for each strain (see section 5.16.3). Following the method described in section 4.4, these sites were binomially tested against false methylation rates, estimated on the chloroplast genome (Figure **S5**B). This resulted in 2.8 million statistically significantly methylated positions (MPs) out of 26.1 million sites on average per MA line, and 2.5 million MPs out of 24.3 million sites per HPG1 line (Tables **S3**, **S5**).

For subsequent analyses, I identified the set of positions that had at least threefold coverage in either all samples including replicates (allowing zero missing data values: NA0 data set) or in at least half of the samples (allowing at most 6 missing data values per sample: NA6 data set). This resulted in around 3 Mio positions, at which at least one replicate sample of a strain had a significantly methylated position in both populations using the NA0 criterion, and around 40% more MPs on average across populations in the NA6 data set (Table 5.1). Requiring a methylated state in both replicates of a strain (NA6r data set) limited the increase in methylated positions of the NA6 data set.

Table 5.1: Data filtering schemes. On the sets of 3-fold-covered cytosines (Cs) in all strains (NA0) or in at least half of the strains (NA6) for each population, the number of methylated (MPs) and differentially methylated (DMPs) positions were calculated from all pairwise strain comparisons (pw_cmp: all), from pairwise comparisons between 31st generation and 3rd generation MA lines, or by comparing HPG1 strains against the Long Island strain LISET-036. For the NA6r data, positions were called methylated only if both replicates of at least one strain showed significant methylation.

		MA lines			HPG1 lines		
data set	pw_cmp	Cs	MPs	DMPs	Cs	MPs	DMPs
NA0	all	13.9M	3.1M	254k	14.0M	2.8M	425k
NA0	31st vs. 3rd	13.9M	3.1M	186k	NA	NA	NA
NA6	all	25.3M	4.5M	376k	21.1M	3.8M	546k
NA6	vs. LISET-036	NA	NA	NA	21.1M	3.8M	513k
NA6r	all	25.3M	3.5M	407k	21.1M	3.0M	535k

Compared to a previous study, the frequency of methylated sites in our analysis was greater (on average 10.7% of all tested Cs per MA and 10.3% per HPG1 strain compared to 6.7% in Lister *et al.* [Lister et al., 2008]; Tables **S3**, **S5**). We assume that this is due to greater statistical power as a result of increased sequencing depth in our data sets. We investigated the three sequence contexts CG, CHG and CHH separately, since the plant methylation machinery operates differently on them (see section 1.4.1). Methylation was called at 31.7% of all tested CG, 16.8% of all tested CHG and 4.9% of all tested CHH sites in the genome on average per line of both populations (Tables **S3**, **S5**). The distribution of methylated cytosines along the genome was similar to previous studies [Lister et al., 2008, Cokus et al., 2008], with the vast majority of methylated sites being found in centromeric and pericentromeric regions (Figure **5.5**A). While methylated CHG and CHH sites were largely depleted from chromosome arms, CG methylation was found there at a low relative frequency (Figure **5.5**A). Notably, the haplogroup-1 lines showed a higher relative fraction than the MA lines. This is presumably due to lower read coverage of divergent repetitive or transposable elements in the HPG1 pseudo reference genome, indicated by the drops of MPs in the centromeres in Figure **5.5**A. These genomic features are preferentially located in and around the centromeres and contain most CHG and CHH methylated sites.

Compared to all and to the covered CG sites in the *A. thaliana* genome, the number of CG sites among methylated sites was about two-fold enriched (Figure **5.5**B). Besides reflecting actual biology, this can also partly be explained by the distinct methylation rate distributions of the different contexts, with most CG sites being highly and most CHH sites weakly methylated (Figure **5.5**C). High methylation is more likely to be significant in the binomial test against low false methylation rates.

5.6 Spectrum of single-site epigenetic variation

I analyzed the positions in each data set that were methylated in at least one sample (or both replicates) for significant differential methylation using the method described in section 4.5. First, to account for within-sample variance, I identified differentially methylated positions (DMPs) between replicates at a relaxed false discovery rate (FDR) of 10% and discarded them from further analysis. There were on average 6,300 DMPs between MA siblings (median 4,500), but only 46 between HPG1 replicates, because we sequenced pools of 8-10 individuals per haplogroup-1 sample, which compensated for this apparent inter-individual variation. Next, performing all pairwise strain comparisons at the remaining positions and requiring a minimum coverage of 3x in all strains (NA0 data) for each population resulted in 254,000 and 425,000 sites with significantly different methylation (false discovery rate <5%) in the MA and HPG1 population, respectively.

The capacity to detect DMPs mainly depends on two factors: the number of statistical tests and the read depth. Since the frequency of DMPs roughly linearly increased with the number of statistical tests (i.e. number of methylated positions) from NA0

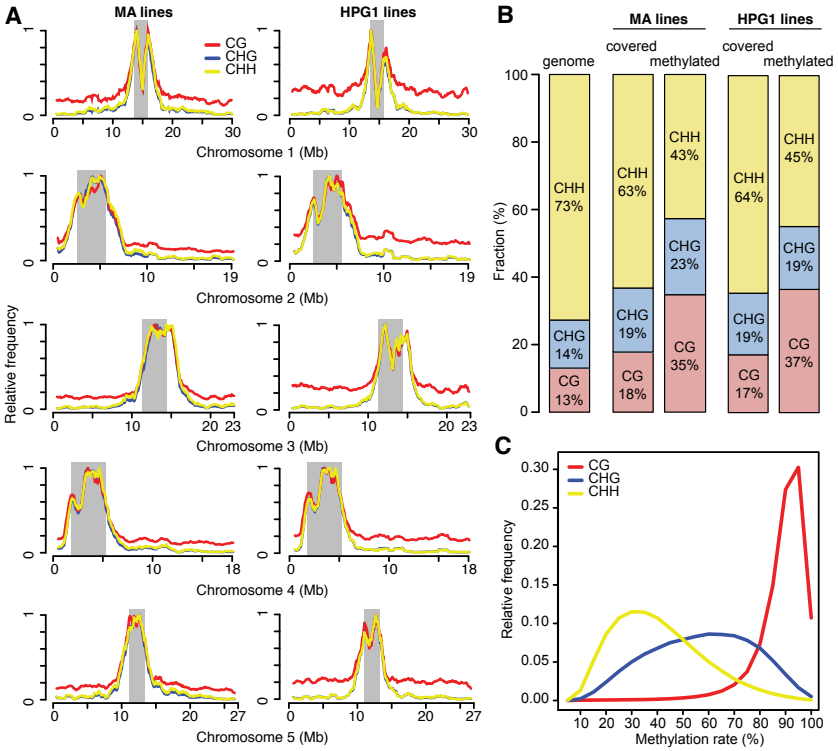

Figure 5.5: Spectrum of DNA methylation in two populations. (A) Distribution of methylated positions by sequence context for each population along each chromosome (centromeres in grey). Data were normalized to the highest value for each chromosome and class. (B) Contribution of CG, CHG and CHH sites to total cytosines in the genome, covered and methylated cytosines per population. (A) and (B) are based on the NA0 DMP set considering all pairwise strain comparisons (Table 5.1). (C) Relative frequency of methylation rates by sequence context. Data from 3rd generation MA lines for sites methylated in at least one sibling.

to NA6 data (Table 5.1), we conclude that in the range of our analyses, the single-site testing is largely insensitive to the number of tests performed. Read depth, however, might have a more pronounced effect. Fisher's exact test is known to yield more significant calls with increasing sequencing coverage. This is confirmed by simulations of decreased read depth on the mutation accumulation data set (Figure **5.6**). While I identified almost twice as many DMPs with 50% compared to 25% coverage, only 13% additional DMPs were called when increasing coverage from 75% to 100%. Thus, it is expected that further increasing the sequence coverage of our analyses would increase the number of detected DMPs by less than 13% based on the factor read depth.

Almost all DMPs were found in the CG context (96% and 97% for the MA and HPG1 populations), although CG sites constitute only around a third of the methylated sites in our data sets (Figure **5.5**B). Because CHG and especially CHH sites show much

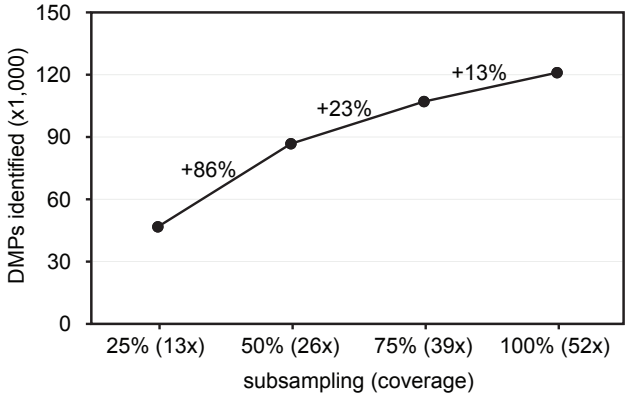

Figure 5.6: Detection of differentially methylated positions (DMPs) at different coverages. Number of identified DMPs between all pairwise comparisons of mutation accumulation lines when the data set was subsampled at different strand-specific coverages.

lower methylation levels compared to CG sites (Figure **5.5**C), differences in CHG and CHH methylation are mostly too weak to be called significant by the statistical test.

In contrast to the general distribution of methylated positions on the genome (Figure **5.5**A), CG-DMPs were enriched on chromosome arms (Figure **5.7**A). Chromosome arms contain most genes, and intriguingly, almost 90% of these variable sites located to genic regions (exons, introns and UTRs), which is an enrichment of more than three-fold for all sequence contexts compared to non-differentially methylated positions (N-DMPs) (Figure **5.7**B). In contrast, we found a depletion of CG-DMPs in transposable elements (TEs) and intergenic sequences. Despite the fact that genes mostly contain CG methylation only (section 1.4.1), CHG- and CHH-DMPs showed a bias towards genes and against TEs as well, albeit less pronounced than for CG-DMPs (Figure **5.7**B). However, we cannot exclude that the annotation of the Col-0 genome falsely classifies some TEs as genes.

It is known that CG methylation follows a specific distribution along exons [Cokus et al., 2008, Lister et al., 2008]. We investigated the profile of variable sites on exons for the MA data set and found that the distribution of DMPs in CG and CHG context largely followed the general pattern of methylation, showing a gradual increase towards the 3' end of the gene and a drop in the last exon (Figure **5.7**C). Similarly, the profiles of DMPs and N-DMPs were not much different at genes of varying lengths and at exon-intron boundaries or on transposable elements (Figure **S6**). Short genes up to the length of 1 kb had low methylation in general, consistent with previous reports [Cokus et al., 2008] (Figure **S6**A).

Methylation is preferentially found in TEs and at loci that can generate small interfering RNAs (siRNAs) [Henderson and Jacobsen, 2007, Lister et al., 2008], and these siRNAs can direct and maintain their own or remote methylation [Matzke and Mosher, 2014] (section 1.4.1). TEs and siRNA loci can be interspersed in euchromatin and might evoke methylation changes in nearby genes, either upon en-

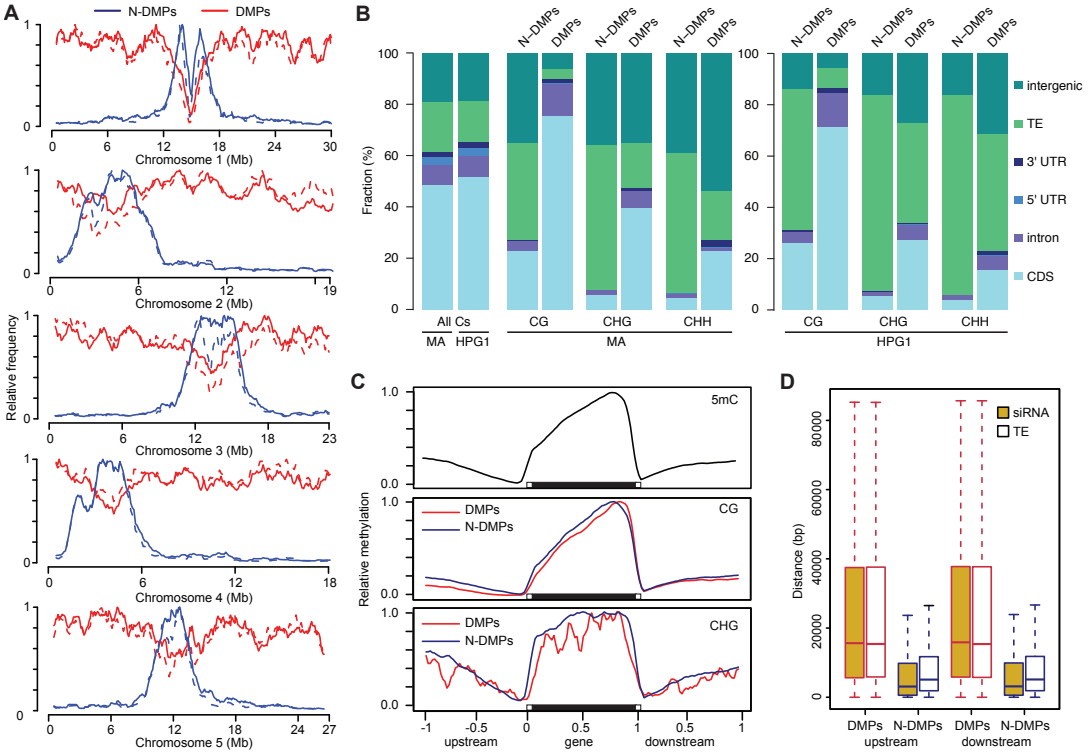

Figure 5.7: Single-site epivariation spectrum. (A) Distribution of differentially methylated positions (DMPs) and N-DMPs (non-differential) along each chromosome for the MA (solid) and HPG1 lines (dashed). Data were normalized to the highest value for each chromosome and class. Based on NA6 data on all pairwise strain comparisons. (B) Annotation of all cytosines (Cs), N-DMPs and DMPs for both populations. Sites were hierarchically assigned to CDS > intron > 5' UTR > 3' UTR > transposon > intergenic. (C) Averaged distribution of all methylated (^{5m}C) and (in)variably methylated CG and CHG sites in the MA lines along genes. Data were normalized to the highest value for each sequence context and class. The coding region is indicated by a black bar. (D) Distances of DMPs and N-DMPs of the MA lines to the closest upstream and downstream 24-nucleotide siRNA and transposable element (TE). Horizontal bar corresponds to median, whiskers indicate 75th percentile.

vironmental triggers [Dowen et al., 2012, Yu et al., 2013] or potentially spontaneously [Manning et al., 2006, Martin et al., 2009]. To check whether the variable methylation that we found on chromosome arms is mainly due to these loci, we compared the distances of DMPs and N-DMPs to their respective closest TE and siRNA-associated loci (as described in [Becker et al., 2011]). N-DMPs were consistently located closer to such elements (Figure **5.7**D) and overlapped seven times more often than DMPs with regions to which siRNAs mapped.

These findings agree with other reports [Vaughn et al., 2007, Zhang et al., 2008, Schmitz et al., 2013a] and confirm on a whole-genome level that single methylated CG sites in transposable elements are more stable than those in protein-coding genes.

5.7 The rate and recurrence of spontaneously occurring single-site epimutations

We performed hierarchical clustering of the MA lines to gauge systematic differences between strains. The siblings as well as the 3rd and 31st generation lines grouped together (Figure **5.8**A). By contrast, the clustering of N-DMPs did no longer resolve the differences between early and late generation plants, indicating that the DMP set seems to capture most differential methylation signals between the strains. This finding was confirmed by inspecting the pairwise distances between lines based on DMPs. Third generation strains were the most similar to each other, and late generation plants were more similar to 3rd generation strains, from which they had diverged for 34 generations, than to each other, being separated by 62 generations (Figure **5.8**B).

Interestingly, line 69 showed a 40% increased number of DMPs to the early generation strains in comparison with the other late generation plants. The re-sequencing of its genome revealed a non-synonymous SNP in MATERNAL EFFECT EMBRYO ARREST 57 (MEE57), which is related on the protein level to the methyltransferase MET1, responsible for the maintenance of CG methylation (see section 1.4.1). This could potentially represent a case where a small change in the DNA has a large effect on the epigenome, as previously reported [Johannes et al., 2009, Stroud et al., 2013b, Shen et al., 2014, Willing et al., 2015]. However, it remains speculative, since although a study reported that MEE57 is essential for endosperm development [Pagnussat et al., 2005], several different *A. thaliana* strains lack a functional copy of MEE57 [Cao et al., 2011]. The siblings of this strain were as similar to each other as other sibling pairs (Figure **5.8**A), arguing against a generally increased epimutation rate.

Together, single-site epimutations show a gradual accumulation over time, similar to DNA mutations. Moreover, as few as 34 generations seem to be sufficient for detectable separation of strains based on epigenetic variation.

To better estimate and compare the frequency of DMPs across generations, I focused on differential methylation that separated early and late generation plants, i.e. that accumulated over 34 generations. To this end, I re-calculated false discovery rates (FDRs) for only the pairwise comparisons between 3rd and 31st generation lines after excluding ∼11,000 significantly different sites between both 3rd generation lines. I called DMPs as those sites that showed significant differential methylation (FDR < 0.05) between at least one 31st generation strain and both 3rd generation lines, considering only positions that were sufficiently covered in all strains (NA0 data). This resulted in a set of 186,000 DMPs, or about 30,000 on average per strain. Compared to the frequency of genetic mutations, which was determined on a subset of these lines as being less than 30 single-base mutations per 31st generation strain [Ossowski et al., 2010], the number of single-site DNA methylation changes in these lines was three orders of magnitudes higher. When assuming a theoretical equal share between ten lines, one would expect to find much fewer than 30,000 DMPs per line (18,600). That we identified more is because

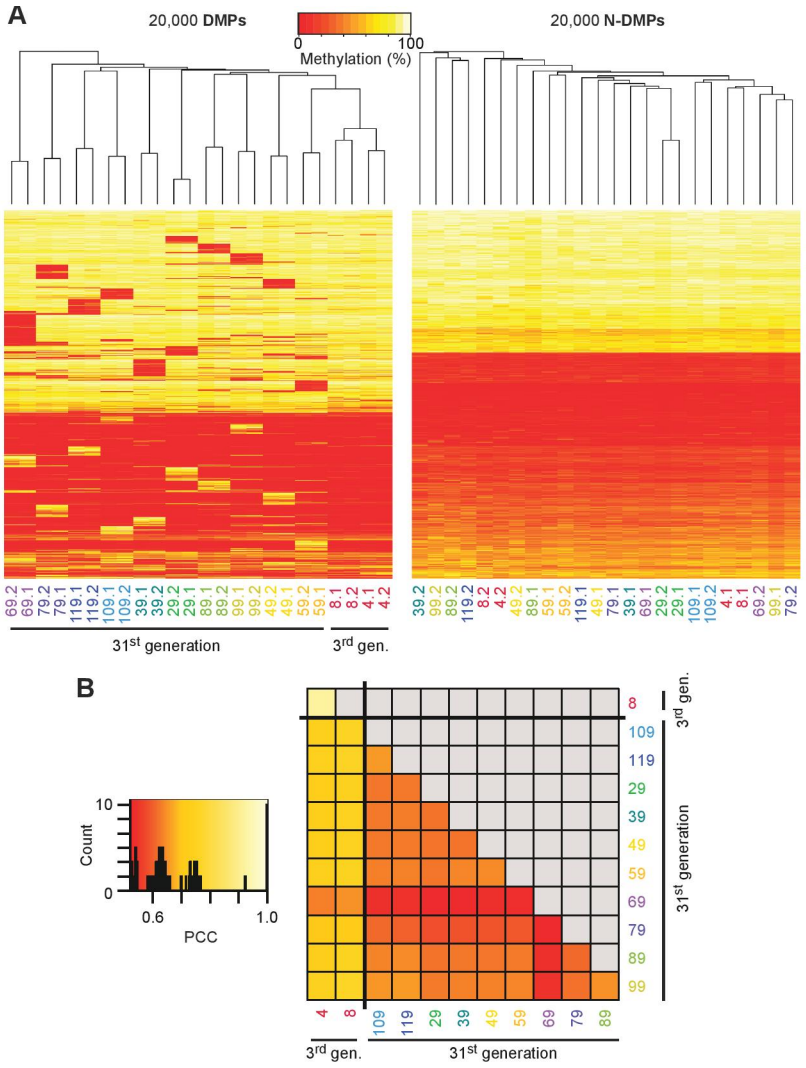

Figure 5.8: Single-site epigenetic diversity in the mutation accumulation lines. (A) Hierarchical clustering based on 20,000 sites each, drawn randomly from differentially methylated positions (DMPs) and invariant sites (N-DMPs). (B) Heat map representing pairwise Pearson's correlation coefficient (PCC) based on number of DMPs between individuals. PCCs between 3rd generation strains: 0.92; between 3rd and 31st generation: 0.63-0.77; between 31st generation lines: 0.52-0.66. The histogram on top of the color key indicates counts of PCC bins. Data for (A) and (B) based on all pairwise line comparisons and NA0 criterion (i.e. 253,000 DMPs in total).

32% of DMPs were shared in more than one independent lineage, and 13% in more than two, irrespective of the sequence context (Figure **5.9**A). The expected percentage of recurrently observing randomly distributed changes in ten independent strains is less than 1%, when assuming 30,000 DMPs per strain ($(30,000/3,000,000)^2 \cdot 45 = 0.45\%$).

This characteristic of recurrent DNA methylation changes is in stark contrast to DNA mutations that hardly affect the same sites in the short-term.

Hence, we can deduce that there are sites in the genome that are preferentially susceptible to methylation changes. In agreement with this assumption, I identified DMPs between two 32nd generation strains and their direct 31st generation ancestors and found that more than two-thirds of these DMPs were also found in other 31st generation lines. This comparison directly yielded the emergence rate of DMPs from one generation to the next, namely 3,300 on average. This number is in accordance with the number of DMPs between siblings that are separated by 2 generations and show on average 6,300 DMPs, but it is almost two times more compared to the 11,000 DMPs between both 3rd generation lines (6 generations apart). Moreover, when assuming a constantly linear accumulation of epimutations by 3,000 per generation, we would expect ∼100,000 DMPs per line after 34 generations, but we observed only 30,000.

Thus, the emergence rate of single-site epimutations does not seem to be linear, as opposed to DNA mutations. This leads to the conclusion that while few sites accumulate changes of their methylation states that are being maintained over generations, a fraction of them frequently revert back to an initial state.

5.8 The rate and recurrence of single-site epimutations in nature

To inspect whether accumulation rates of DMPs were different in controlled and natural environments, we compared the number of DMPs between any two strains within the HPG1 or MA set to the number of SNPs between them. The number of SNPs served as a molecular clock since the number of generations is unknown for the HPG1 strains. To increase the number of data points in the low range of genetic differences, we inferred the number of SNPs between MA siblings from the mutation rates determined by Ossowski *et al.* (2010) on a subset of these lines. Since natural environments are highly fluctuating and much more variable than greenhouse conditions, we expected the methylomes to be more highly diverged in the HPG1 than in the MA accessions.

In contrast to our null hypothesis, there was a similar trend of sub-linear accumulation of DMPs in both the HPG1 and MA populations (Figure **5.9**B). The broader distribution of MA line differences relative to HPG1 differences and its steeper initial ascent was most likely caused by the fact that methylome data in the MA lines were from individual plants, whereas we had sequenced pools of 8-10 individuals in the HPG1 experiment.

By pooling strains, low frequency epimutations are diluted and less likely to be detected. This assumption is corroborated by the fact that we see only around one hundred DMPs between replicates of HPG1 pools compared to several thousand in replicates of the MA lines. Moreover, we assumed the same genetic mutation rate in the two populations. A potentially faster genetic mutation rate in the wild, for example because of increased stress or exposure to UV, would result in a steeper slope of the

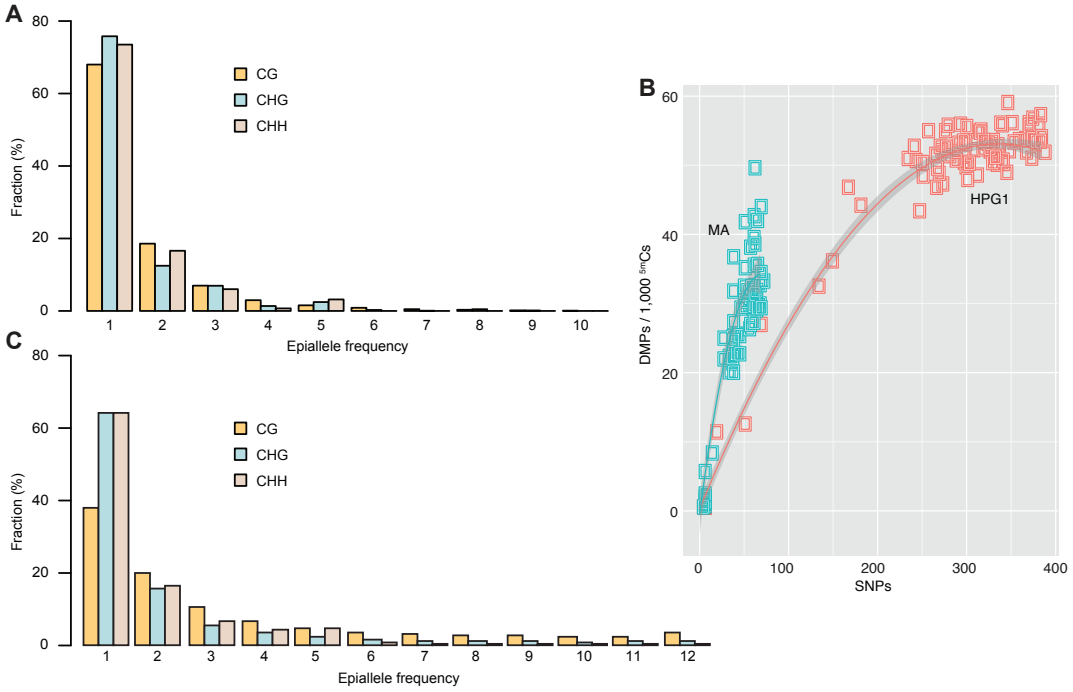

Figure 5.9: Recurrence and rate of single-site epimutations in two populations. (A) Epiallele frequency of differentially methylated positions (DMPs) between 31st and 3rd generation MA lines. (B) Number of DMPs (based on NA6r data) in relation to number of SNPs in pairwise comparisons. Data of mutation accumulation (MA) lines are based on single individuals, haplogroup1 (HPG1) data on pools of 8-10 individuals; each data point represents an independent comparison of two lines. DMPs in each pairwise contrast were scaled to the number of methylated sites compared. Comparisons with less than 40 DMPs per 1000 methylated positions were between strains from the same location, and between Paw-13 and Yng-4. (C) Epiallele frequency of DMPs between Lake Michigan HPG1 lines and LISET-036 from Long Island.

HPG1 curve, if plotted against the number of generations. Finally, the initial increase of the HPG1 epimutation rate is based on only few comparisons between strains from the same sampling site, which might not be sufficient for an accurate estimate. Hence, the reported DMP accumulation rate of the HPG1 accessions is likely an underestimation.

The mutation accumulation lines showed recurrent DMPs in independent lines. We next asked whether this was also the case in the haplogroup-1 lines. Since obtaining frequencies from all pairwise comparisons is difficult, we selected a reference strain, the geographic outlier LISET-036 from Long Island, and compared methylation in each of the 12 accessions from near Lake Michigan to it. From 513,283 DMPs in total (154,480 per strain), 61% of CG-DMPs were recurrent in at least two independent Lake Michigan accessions (Figure **5.9**C), which is almost double the number that was observed for the equi-distant MA lines. This can be partly explained by the fact that we sequenced four pairs of strains from the same location. Nearly half of all CG-DMPs

with a frequency of 2 were attributable to such pairs, while 6% would be expected if they were randomly distributed across strains. Moreover, since we compared all strains to LISET-036, DMPs unique to this strain are more likely to be shared among several accessions, skewing the distribution towards higher frequencies. By contrast, CHG- and CHH-DMPs showed similar epiallele frequencies as those found in the MA lines, which might again be explained by their more difficult detection.

Together, the profile of single-site methylation changes in the natural HPG1 population highly resembles the one from the greenhouse-grown MA lines, and despite the discussed caveats in the analysis, we deem it unlikely that DMPs accumulate at a much faster rate in the HPG1 than in the MA population.

5.9 Determining methylated and differentially methylated regions

Most naturally occurring epialleles consist of variable DNA methylation of a genomic region rather than of single sites only (section 1.7). To explore larger and potentially more influential epigenetic changes than single-site epimutations, I applied the novel method of calling differentially methylated regions (DMRs) described in chapter 4 on each data set. Thus, I first called methylated regions (MRs) for each line using a Hidden Markov Model-based approach (section 4.6.1), which operated on the accumulated read counts across replicates of all genome-wide cytosines, independent of the coverage. MRs were not extended beyond a region larger than 50 bp without a covered cytosine ('desert size' option) and positions of less than 10% methylation were trimmed from the beginning and end of a methylated region. An FDR threshold of 5% determined 22,446 (SD = 1,634) methylated regions per MA line and 32,529 (SD = 1,629) per HPG1 strain. However, the average length of the regions in the MA lines was twice that of the HPG1 accessions (Figure **5.10**), and the unified set across all MA lines covered 26.3 Mb of the Col-0 genome compared to 22.8 Mb of the HPG1 lines (22.6 Mb on the HPG1 genome). This difference is presumably due to the higher number of covered cytosines in the MA population and an increased number of sequencing gaps in the HPG1 lines, which leads to shorter MRs (Tables **S3** and **S5**).

For validation of our HMM-based methylated region detection method, we compared data from the mutation accumulation lines to data from methylated-DNA immunoprecipitation followed by sequencing (MeDIP-Seq; section 2.3) of a Col-0 sample (Vincent Colot and co-workers, pers. communication). Of the genome space enriched in MeDIP-seq, 91% was classified as a methylated region by my HMM-based approach in the union MR set of the mutation accumulation lines.

To determine DMRs based on methylated regions, my method divided the unified set of MRs across all strains into a set of segments (∼2.5 million for the HPG1 data set), which were filtered using empirical criteria (to ∼230,000 regions; section 4.6.2). On these regions, the statistical test for differential methylation, explained in section 4.6.3, was performed for all pairwise strain comparisons, thereby accounting for replicate

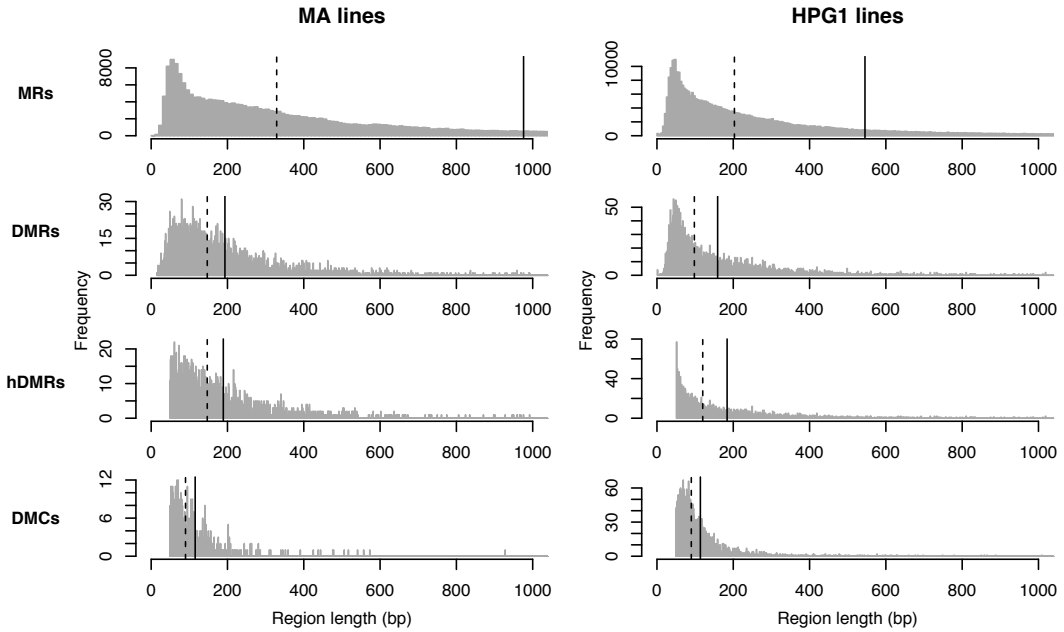

Figure 5.10: Regional length distributions. Length distributions of methylated regions (MRs), differentially methylated regions (DMRs), highly differentially methylated regions (hDMRs) and DMP clusters (DMCs) for the mutation accumulation (MA) and haplogroup-1 (HPG1) populations. Dashed line indicates median, solid line the mean.

data. This resulted in ∼221,000 significant pairwise comparisons on ∼11,000 regions for the HPG1 lines. Grouping the strains in each region based on the significant pairwise comparisons, testing the groups for differential methylation and resolving overlapping segments (section 4.6.4) resulted in a final set of 4,821 DMRs in the HPG1 population. For the MA lines, I detected 3,837 DMRs. Almost all DMRs (98%) reflected two alternative epialleles only (i.e. two groups of strains) across both data sets. Since the sensitive statistical test easily classifies regions with high sequencing information and low variance as significant, although their methylation rate difference is subtle, I empirically filtered the DMRs into 'highly' differentially methylated regions (hDMRs), requiring a more than three-fold difference in methylation levels between epialleles (section 4.6.5). The HPG1 and MA populations feature 3,199 and 2,352 hDMRs, respectively. DMRs and hDMRs showed similar length distributions compared to each other as well as across data sets (Figure **5.10**), spanning 767 kb (HPG1) and 742 kb (MA) of the genome.

Thus, we found that 97.2% and 96.7% of the methylated genome space remained largely stable on a regional scale in strains grown over 30 generations in the greenhouse and over a few hundred generations in nature, respectively.

For comparison with the most widely used approach to call differentially methylated regions, I clustered DMPs by genomic distance into regions as described in section 5.16.4 and according to ref. [Becker et al., 2011]. While I only called 600 DMCs

of the 31st generation strains against a single 3rd generation line, line 4, in the MA data, I detected many more of these regions (4,069) in the HPG1 strains in pairwise comparisons against LISET-036.

5.10 Spectrum of differentially methylated regions

Following the overall distribution of methylated cytosines, but in stark contrast to variably methylated sites, MRs and (h)DMRs were most frequent around the centromere (Figure **5.11**A) and mainly found in TEs and intergenic regions (Figure **5.11**B). This revealed a different profile of epigenetic variation when looking at regions rather than individual sites. However, although more DMRs were found in TEs than in genes, less than 2% of the global methylated TE space was covered by DMRs, compared to 5% and 9% of the methylated gene space in the MA and HPG1 population, respectively.

Only 1% of methylated CHH and 2% of methylated CHG positions were located outside methylated and differentially methylated regions (Figure **5.11**C), consistent with CHH and CHG methylation occurring almost exclusively in intergenic regions and silenced TEs [Zhang et al., 2006, Cokus et al., 2008, Lister et al., 2008] (Figure **5.5**A). By contrast, 45% of methylated CG sites (mCGs) were not included in MRs. Most of the mCGs outside MRs located to genes (94%), but compared to mCGs within MRs, they were less densely distributed (Figure **5.11**D) and were separated by many more unmethylated sites (Figure **5.11**E, F). Thus, the sparse distribution of methylated positions in coding regions explains the underrepresentation of genes in our (D)MR set, even though gene body methylation accounts for a considerable fraction of methylated CG sites. Relative to all methylated regions, however, gene-coding regions were two-fold overrepresented in the genome sequence covered by DMRs, and three-fold in the genome sequence covered by hDMRs (Figure **5.11**B). This might be due to the greater statistical power of detecting differential methylation at the typically higher methylated CG sites compared to CHG or CHH sites, similar to the DMP calling (section 5.6). Consistently, more than half of the (h)DMRs were differential in the CG context only (Figure **5.11**G), even though only one quarter of cytosines within DMRs were CG sites (Figure **5.11**C). On the other side, my DMR detection method classified regions as DMRs that show small differences in average DNA methylation rate, especially in non-CG contexts (Figure **5.12**), consistent with the general context-dependent methylation rate distributions (Figure **5.5**C).

Thus, whether the general bias of regional and particularly single-site DNA methylation towards increased variability in genes compared to TEs is solely attributable to higher statistical power in these mCG-rich regions or whether it reflects a real biological phenomenon (e.g. stable silencing of TEs) currently remains unclear.

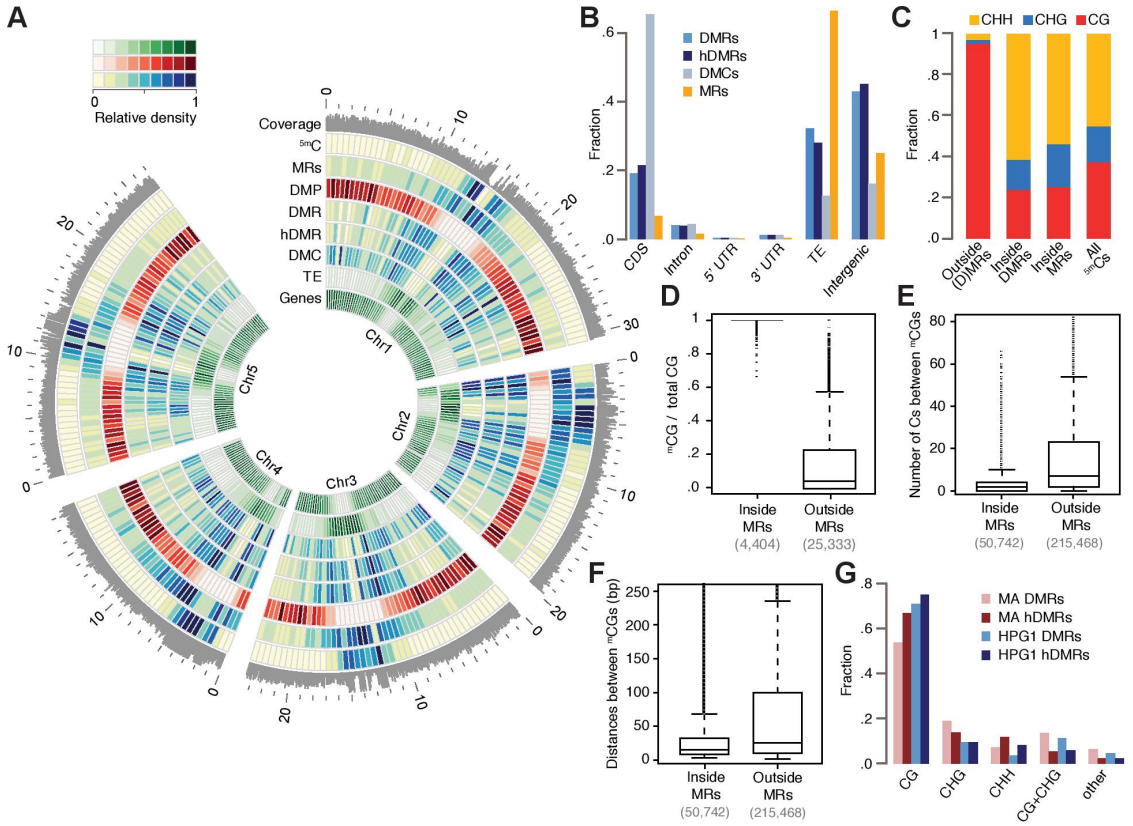

Figure 5.11: Spectrum of epigenetic variation. (A) Density of genome-wide features of the haplogroup-1 (HPG1) population: average coverage in 100 kb windows, all other tracks in 500 kb windows. Outside coordinates in Mb. (B) Annotation of cytosines in MRs and (h)DMRs for the HPG1 population. For comparison with the MA lines, see section 5.13. (hD)MR sequences were assigned to only one annotation in the following order: CDS > intron > UTR > transposon > intergenic. (C) Sequence context of methylated positions relative to MRs and DMRs; based on HPG1 data. (D)-(F) are based on data of one HPG1 strain (LISET-036). Number of data points in parentheses. (D) Fraction of methylated CG sites (mCG) among all CG sites for each gene and transposable element that contains at least 5 CGs. (E) Number of unmethylated cytosines (Cs) in-between methylated CG sites within genes in dependence of whether these sequences are inside or outside of MRs. (F) Distances in bp between methylated CG sites within genes in dependence of whether these sequences are inside or outside MRs (minimal distance 2 bp to exclude symmetrical sites). (G) DMRs and hDMRs of MA and HPG1 population by sequence contexts in which significant methylation differences were found. Abbreviations: ^{5m}C: methylated position, DMC: DMP cluster, DMP: differentially methylated position, DMR: differentially methylated region, hDMR: highly differentially methylated region, HPG1: haplogroup-1 lines, MA: mutation accumulation lines, MRs: methylated regions, SNP: single nucleotide polymorphism, TE: transposable element.

Comparison between DMCs and DMRs

For the haplogroup-1 data set, the 4,069 DMCs spanned 463 kb of the genome, which is around 60% of the genome space covered by DMRs. Moreover, only 169 kb (37%) of

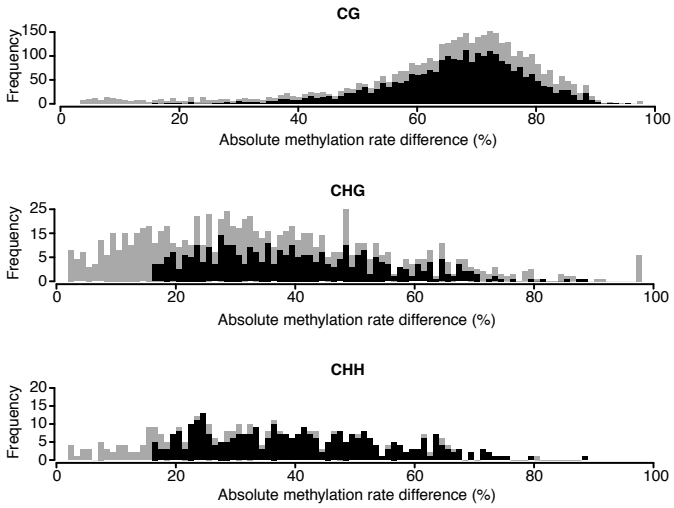

Figure 5.12: Methylation rate differences in regions of differential methylation. Histograms of the absolute mean methylation rate differences of differentially methylated regions (DMRs; grey) and highly differentially methylated regions (hDMRs; black, overlaid) of all different sequence contexts.

the DMC space was also covered by DMRs and 237 kb (51%) by methylated regions of the HPG1 population. The low overlap between (D)MRs and DMCs might again be explained by the sparse distribution of methylated CG sites in genic regions coupled with DMCs being composed primarily of CG sites. This is corroborated by the finding that only one quarter of the DMCs (988) were detected when the maximal allowed distance of DMPs between each other is reduced from 50 bp to 20 bp (section 5.16.4), and by the fact that methylated regions in genes mostly contain methylated positions spaced less than 50 bp apart (Figure **5.11**F).

The distribution along the genome and annotation of DMCs roughly followed that of DMPs (Figure **5.11**A, B), with around two thirds of DMCs covering genes. This underlines the rather distinct differential methylation patterns between DMP-based approaches and my novel DMR calling method.

5.11 The rate and recurrence of differentially methylated regions

To ascertain the spontaneous emergence rate of differentially methylation regions, I re-calculated DMRs in the MA population based on only pairwise comparisons between the 3rd and 31st generation lines. This resulted in 1,995 DMRs, and both early generation lines were classified in the same epiallele group in 1,234 of these DMRs. Based on the latter, I identified on average 203 DMRs per 31st generation line to at least one 3rd generation strain. Assuming a linear accumulation, this would yield an emergence

rate of ∼6 DMRs per line and generation, which is more than hundred times less than for single-site epimutations (∼880 DMPs per line/generation), but in the same range as DNA mutations (1 SNP per line/generation).

In addition, the comparison of two 32nd generation lines to the 31st generation lines, from which they derived, yielded 50 DMRs in total (27 in line 49 and 33 in line 39). For line 39, ten of the DMRs between the single generation were also identified as a DMR between generations 31 and 3. Intriguingly, the methylation profiles of these regions were similar between generations 32 and 3, but different from generation 31, i.e. the methylation status of the 31st generation line was reverted to the 'initial' state from 34 generations ago (one example is given in Figure **S7**). Such incidences were not found for the line 49. Hence, this observation demonstrates that large DNA methylation changes can occur even within a single generation.

Despite their different genomic distribution compared to DMPs, the frequency distribution of DMRs was similar to that of DMPs, with 30% and 36% shared within the MA and HPG1 population, respectively (Figure **5.13**). Many of the HPG1 DMRs with frequency 2 were also shared between strains of the same sampling location. Thus, recurrent differential methylation is also common in strains that diverged at least a century ago.

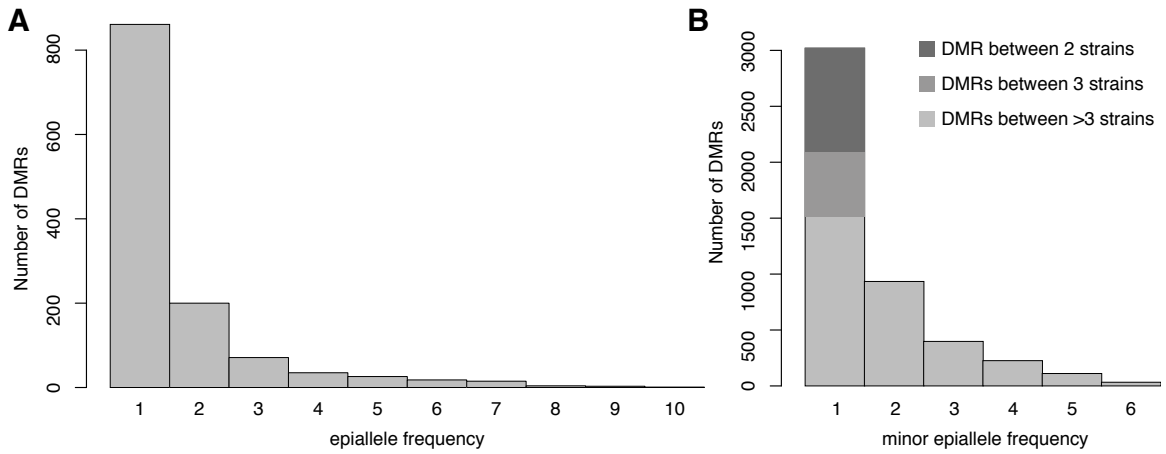

Figure 5.13: Allele frequency of regional variable methylation in two populations. (A) Epiallele frequency of differentially methylated regions (DMRs) of the mutation accumulation (MA) lines based on pairwise comparisons between 31st generation and 3rd generation lines. I considered 1,234 DMRs, for which both 3rd generation lines were assigned to the same epiallele group. (B) Minor epiallele frequency (MAF) of 4,722 DMRs that could be split into only two groups in the haplogroup-1 population. DMRs were based on all pairwise comparisons between strains. The frequency 1 class is subdivided dependent on the number of strains that were tested significant in pairwise comparisons (for 2 and 3 strains, MAF is always 1).

5.12 Effect of differential methylation on gene expression

We next asked whether the observed epigenetic variation has any phenotypic consequences. Since the 13 HPG1 strains did not show visible and only marginally measurable phenotypic divergence (see section 5.4 and Figure **S4**), we inspected gene expression levels. Therefore, we sequenced transcriptome libraries of all HPG1 strains and identified 269 differentially expressed (DE) genes across all possible pairwise comparisons as described [Hagmann et al., 2015]. The accessions did not cluster based on the overall expression levels, but when limiting the analysis to DE genes, accessions originating from the same geographical location were grouped together (Figure **S8**). The Yng strains were an exception, consistent with their admixed genetic architecture (see Figure **5.4**), and they were responsible for most of the DE genes identified in pairwise comparisons.

We identified only 28 DE genes that overlapped with an hDMR either in their coding or 1 kb upstream regions. Although they were not enriched in any GO term, a noticeable number of them were involved in stress response or disease resistance, as observed by manual inspection. However, we could not detect a systematic relationship between methylation and gene expression. By scanning hDMRs manually, we found not more than five instances of negative correlations between methylation and expression (and only one instance of positive correlation). Besides two genes of unknown function and a pseudogene, the Yng strains showed a lowly methylated and transcriptionally upregulated disease resistance gene (NB-ARC domain-containing protein), and the BRR strains and MNF-Che-47 were higher methylated in a known pathogen responsive gene (PCC1), whose expression was significantly downregulated compared to the other strains. This might indicate a mild infection of these individuals in the greenhouse. However, since these effects were shared between strains of the same location, we cannot rule out potential adaptations to the environment.

Since the regions of differential methylation only contain half of the single-site epimutations identified in the haplogroup-1 population, I next asked whether DMPs rather than DMRs have an effect on gene expression regulation. Since the impact of single polymorphic sites might be too weak, I focused on DMP clusters and identified 36 differentially expressed genes that overlapped with DMCs. Of these, 16 DE genes were uniquely covered by DMCs and not by any DMR. However, again no systematic associations were detected on these genes, and manual inspection could not identify any obvious negative correlation between DNA methylation and gene expression (there was one positive correlation in MYROSINASE-BINDING PROTEIN 2, which may be "involved in forming defence compounds to protect against herbivory"[1]).

In general, we observed a lack of correlation between gene expression and DNA methylation changes, which suggests that most transcriptional differences in the HPG1

[1] https://www.arabidopsis.org/servlets/TairObject?id=30525&type=locus, last accessed March 2015

population might be due to cryptic associations with DNA methylation, due to a secondary effect of a DMR changing the activity of a *trans*-regulatory locus, or independent of DNA methylation.

5.13 Recurrent epimutations under greenhouse and natural conditions

DNA methylation differences in the haplogroup-1 population can be due to genetic variation, environmental influences or they can occur randomly, whereas it is assumed that the impact of DNA mutations and the environment in the MA lines is minimal. Thus, by comparing epigenetic variation between the HPG1 accessions and the MA lines we assessed the fraction of DNA methylation changes in the HPG1 population that are likely to occur independent of genetic or environmental influences. Intriguingly, almost half of all positions that were classified as DMPs in the MA lines were also polymorphic in the HPG1 accessions (41%), whereby almost 30% had not even been tested in the HPG1 population because of insufficient coverage or lack of DNA methylation (Figure **5.14**A). The theoretical probability of a random methylated cytosine in the MA population being variably methylated also in the HPG1 population was only 7%. Similarly, about a third of HPG1-DMPs were also MA-DMPs and almost 40% were not considered for DMP testing (Figure **5.14**A). Conversely, DMPs from one population were more likely to be unmethylated in all strains of the other population than random methylated sites, indicating that many sites sporadically gain methylation.

Interestingly, on average 20% of DMPs between replicates of the individual MA lines are recurrently found in the HPG1 population. By contrast, of the DMPs that distinguish the 31st generation lines, on average 7% are also variable in the HPG1 population (Figure **5.14**B). This observation suggests that there are 'highly labile' sites that frequently change their methylation status already after a small number of generations, and that they are therefore found more often in independent populations.

The comparison of polymorphic regions between populations revealed similar degrees of overlap as for single sites: DMRs in the one population were 4-fold more likely to coincide with DMRs in the other population than with a random methylated region from the other population (Figure **5.14**C). These analyzed DMRs represented short-term variation over 34 and a few hundred generations. I next tested whether also long-term differential methylation affects same loci more than expected by chance by identifying DMRs between a randomly chosen MA and a randomly chosen HPG1 line, separated by hundred thousands of years. These DMRs were also enriched in each of the two sets of within-population DMRs (MA or HPG1) (Figure **5.14**D). Lastly, I compared the variable regions of the HPG1 strains to DMRs that had been identified with a different DMR detection method in a set of 140 natural, genetically diverse accessions from throughout the native world-wide habitat of *A. thaliana* [Schmitz et al., 2013b]

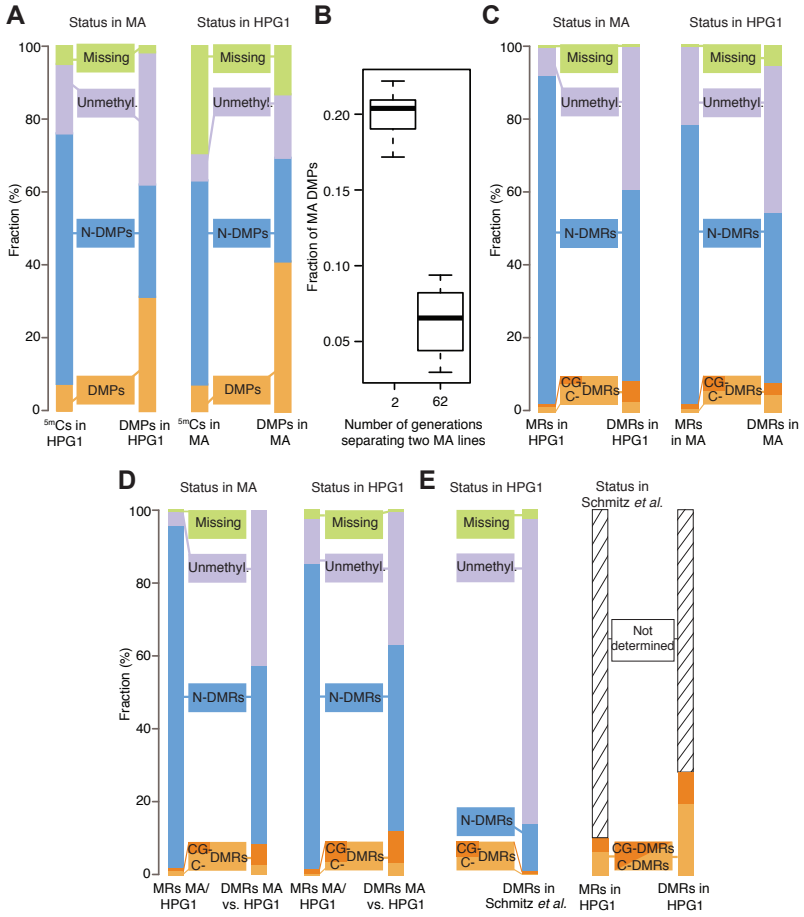

Figure 5.14: Shared epigenetic variation in independent populations. (A) Comparison of methylated positions (^{5m}Cs) and differentially methylated positions (DMPs) identified in pairwise comparisons of mutation accumulation (MA) and haplogroup-1 (HPG1) strains. Distinction in different sequence contexts has been omitted since almost all DMPs (>96%) are in CG context. Left: sites in HPG1 strains and their status in the MA data; right: sites in the MA strains and their status in the HPG1 data. (B) Overlap of MA and HPG1 DMPs according to MA generational distance. I computed DMPs between two randomly chosen MA strains separated by specific numbers of generations and plotted the fraction of those DMPs shared with a randomly chosen HPG1 strain. Each boxplot summarizes ten such random comparisons. (C) Comparison of methylated regions (MRs) and differentially methylated regions (DMRs) identified in pairwise comparisons of HPG1 and MA lines. Dark and light orange subsets of DMRs distinguish regions with differential methylation occurring exclusively in CG context (CG-DMRs) or in any additional or alternative context(s) (C-DMRs). Left: regions in HPG1 strains and their status in the MA data; right: regions in the MA strains and their status in the HPG1 data. (D) MRs and DMRs identified in comparison between one randomly chosen MA line (line 39) and one randomly chosen HPG1 line (MuskSP-68), denoted as "MRs MA/HPG1" and "DMRs MA vs. HPG1", and their overlap with within-population (D)MRs. (E) Comparison of HPG1 DMRs with CG-DMRs (dark orange) and C-DMRs (light orange) from ref. [Schmitz et al., 2013b] identified in 140 natural *A. thaliana* accessions. Because methylated regions were not reported in ref. [Schmitz et al., 2013b], the overlap of DMRs with the space not covered by DMRs could not be assessed. N-DMRs: non-differentially methylated regions.

(Figure **5.14**E). Although only 9,994 of 53,752 DMRs from the global accessions were completely covered by methylated regions in the MA or HPG1 strains, the overlap of DMRs was highly significant (F-score = 19.8; 100,000 permutations).

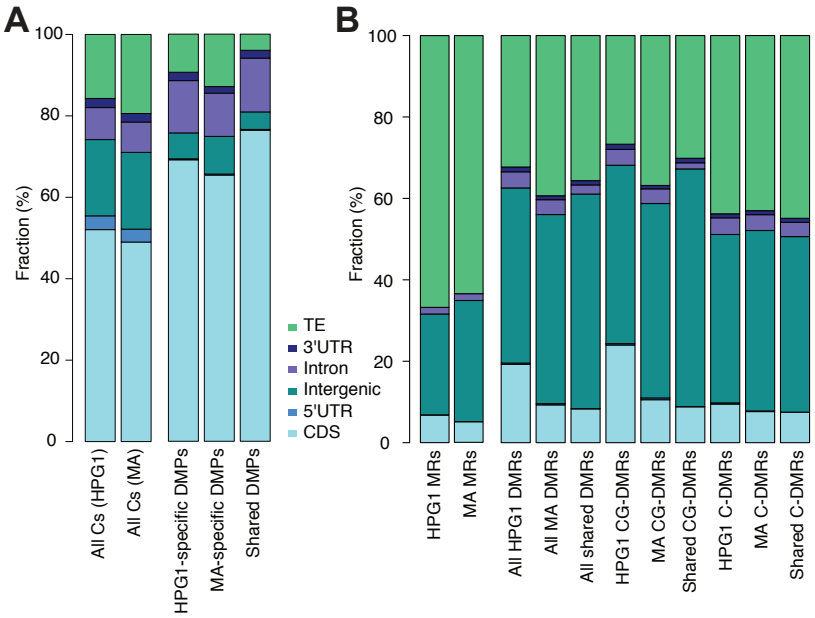

Figure 5.15: Annotation of epimutations in two populations. (A) Annotation of all cytosines (Cs) and differentially methylated positions (DMPs) dependent on the overlap between mutation accumulation (MA) and haplogroup-1 (HPG1) population. (B) Annotation of differentially methylated regions (DMRs) dependent on the overlap between MA and HPG1 populations for different sequence contexts. For (A) and (B), sites within the features were hierarchically assigned to CDS > intron > 5' UTR > 3' UTR > transposon > intergenic.

To ascertain whether there are differences between population-specific and shared epimutations, we investigated the annotation spectrum of these sets of epialleles. Shared DMPs were slightly more biased towards coding sequences than population-specific DMPs, but all cytosines accessible to our methylome analyses showed this bias as well (Figure **5.15**A). The reasons might be higher read coverage and a better TE annotation in the Col-0 reference compared to the HPG1 pseudo-reference genome. Indeed, except for chromosome 4, the average sequencing depth in the pericentromere was higher in the MA lines, and regions of increased coverage appear to have a higher epimutation frequency compared to the HPG1 population (Figure **S9**). For the same reason, variably methylated regions specific to the HPG1 population overlapped with genic sequences more often than MA DMRs (Figure **5.15**B). In contrast to shared DMPs, shared DMRs between both populations were not biased towards genic regions, but rather slightly towards TEs and intergenic regions, following the general DMR distribution (Figure **5.15**B). Together, HPG1-specific and MA-specific epimutations

show a similar annotation pattern, given the slightly different overall accessible set of cytosines in both populations.

To explore potential sources of labile sites that frequently change their methylation status independently of the genetic background, I compared the variable positions of the HPG1 lines with differential sites found in strains deficient in important components of the methylation machinery [Stroud et al., 2013b] (section 5.16.5). Intriguingly, almost all single-site as well as regional epimutations that covered genomic regions of exclusively CG methylation, were hypomethylated in mutants deficient in DNA methylation maintenance, notably in the *met1* single and *vim123* triple mutants (Figure **5.16**). This supports the hypothesis that the maintenance of symmetrical CG methylation during DNA replication is prone to errors [Genereux et al., 2005, Fu et al., 2010] (section 1.7.2) and seems to be imperfect at privileged loci in independent populations. These loci appear to be preferentially in regions that contain solely CG methylation, since positions in regions methylated in all sequence contexts are not enriched in the set of sites that lose methylation in *met1* and *vim123* mutants, supporting the observed higher variability of gene body methylation compared to TE methylation (Figure **5.16**C). Moreover, we observed that hypermethylated sites in the *rdd* triple mutant, which shows impaired demethylation, were also found slightly more often within variably methylated regions of all contexts (Figure **5.16**D), consistent with their role as antagonizing RNA-directed DNA methylation. IBM1, a histone demethylase [Saze et al., 2008], seems to be a strong and stable repressor of DNA methylation, as sites affected by a loss of IBM1 function are not more often variable than sites affected by other mutants (Figure **5.16**B, D).

In summary, a large fraction of the variability in DNA methylation is shared between independent populations of different genetic background and with different environmental history, and there are fair indications that the DNA methylation machinery is susceptible for errors on privileged loci.

5.14 Linkage of epigenetic differences to genetic variation

I described that a large portion of the epivariation in the HPG1 genome can be attributed to seemingly spontaneously occurring methylation changes. In addition, genetic events can have a profound impact on altered DNA methylation (section 1.7.1). The finding that the divergent sequence between the Col-0 and the HPG1 genome was much more likely to be methylated than conserved regions, especially in genic regions, provides an indication that variable DNA sequences themselves are susceptible for DNA methylation (Figure **5.17**A). To quantify how many methylation differences were linked to genome-wide genetic changes in *cis* or *trans*, I estimated the proportion of the epigenetic variance that is attributable to genetic variants (heritability value). I applied a linear mixed model-based method that is similar to variance component models used in genome-wide association studies [Kang et al., 2010, Lippert et al., 2011]. The model

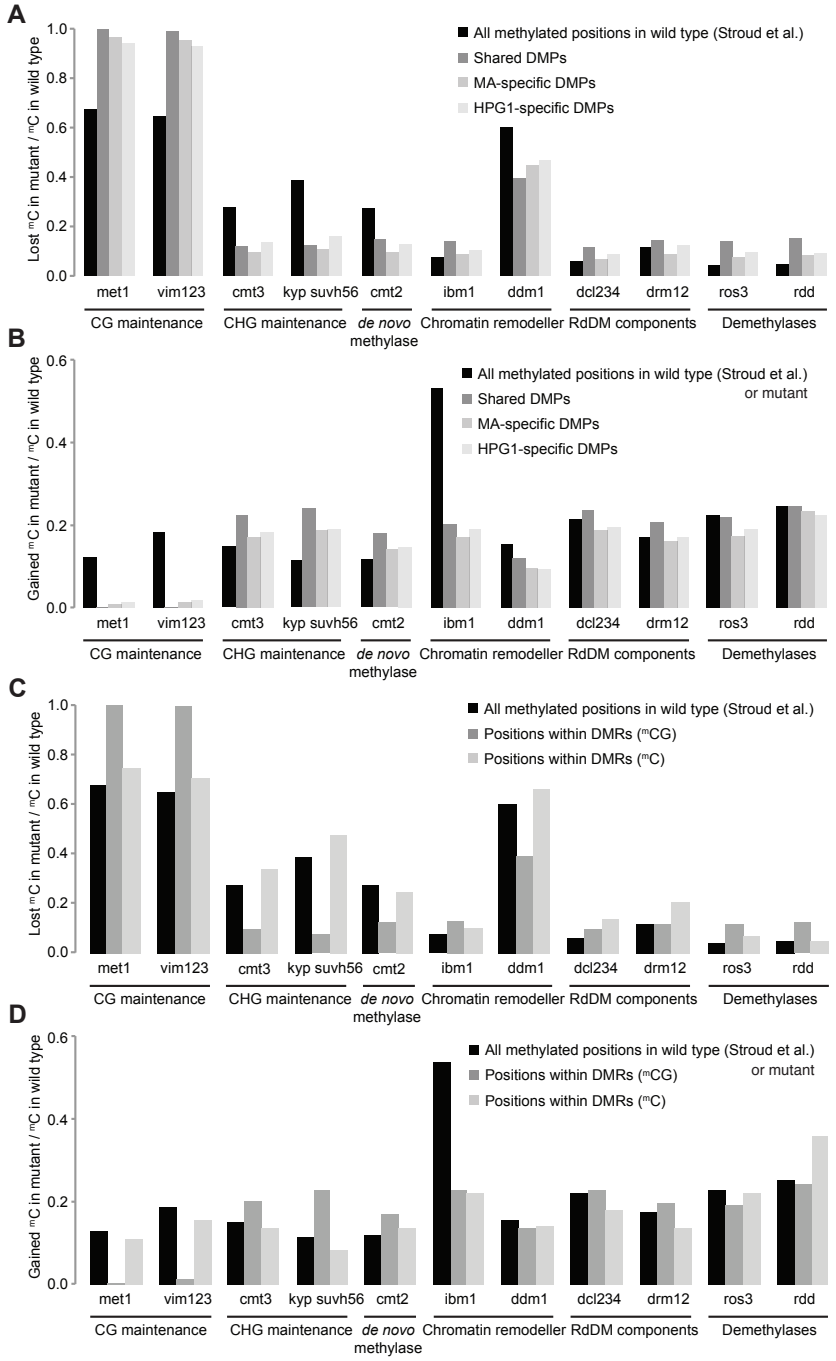

Figure 5.16: Overlapping epigenetic variation in HPG1 lines and methylation-deficient mutants. (A, B) Fraction of sites that lost (A) or gained (B) methylation in mutant samples compared to two wild type (WT) Col-0 samples [Stroud et al., 2013b], from all methylated positions in WT, for each subset of these sites according to the status in the haplogroup-1 (HPG1) and mutation accumulation (MA) population (DMP: differentially methylated position). (C, D) Fraction of sites that lost (C) or gained (D) methylation in mutant samples compared to WT samples, from all methylated positions in WT. Plotted are the fractions from all covered sites in all samples and from sites covered by DMRs within the haplogroup-1 lines that overlap regions of the genome with methylation occurring only in the CG context (mCG) or in any additional or alternative context(s) (mC). *rdd*: triple mutant of *ros1*, *drm1* and *drm2*.

considered the segregating sequence variants as the genotype information and the log average methylation rate of hDMRs as phenotype without allowing for any missing data.

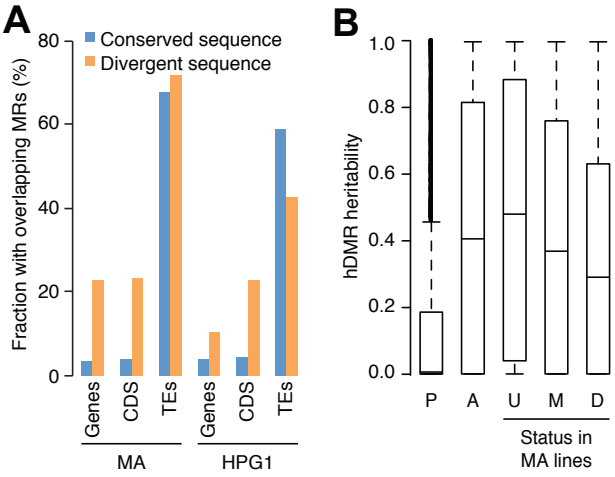

Figure 5.17: Genetic effects on epigenetic variation. (A) Correlation between structural variants (SVs) and probability of overlap with methylated regions (MRs) between the HPG1 pseudo reference and Col-0 sequence. Divergent sequences are insertions of at least 20 bp. This analysis is based on 3,256 SVs overlapping with genes, 641 with coding sequences (CDS) and 4,020 with transposable elements (TEs). (B) Heritability values based on genome-wide genetic differentiation for all haplogroup-1 highly differentially methylated regions (hDMRs), hDMRs with randomly permuted methylation rates and subsets of hDMRs depending on their overlap with methylated and differentially methylated regions of the mutation accumulation (MA) population, respectively. P: Permuted (2,945 hDMRs); A: All (2,945); U: Unmethylated in MA (1,310); M: Methylated in MA (1,243); D: DMR in MA (392).

The median heritability of all hDMRs was 0.41 (mean 0.44), which means that the genetic variance across the entire genome contributed with 41% to the methylation variance (Figure **5.17**B). Notably, HPG1-specific hDMRs that were not methylated in the greenhouse-grown MA lines had a higher median heritability, 0.48, than HPG1 hDMRs shared with the MA lines (0.29) (Figure **5.17**B). This trend was similarly found for all sequence contexts (Figure **S10**). Thus, HPG1-specific hDMRs, especially those in unmethylated regions of the MA lines, appear to be more linked to genotype than hDMRs shared between both populations.

I identified 19% of all hDMRs (21% CG-hDMRs, 14% CHG-hDMRs, 7% CHH-hDMRs) to be apparently strongly associated to genetic variants, since they show a heritability value greater than 0.9 (with a standard error of at most 0.1). For half of these 'heritable' hDMRs the genotype even explained more than 99% of their methylation differences. However, 6.7% of the sequence space of these heritable hDMRs

overlapped with MA DMRs (9.4% for the remaining, less heritable hDMRs), which again points to the presence of highly labile sites independent of genetic background.

To potentially pinpoint genetic variants that are directly linked to close-by methylation changes, I searched for DMRs that were within 1 kb of segregating SNPs or insertions/deletions. Of 159 segregating variants in the vicinity of 191 DMRs, 78% were SNPs, and there were only three indels longer than 10 bp. However, the strains containing a genetic variant compared to the HPG1 genome rarely shared the same methylation pattern, or other strains that did not have the genetic variant showed the same methylation status as well. I only found a single case of consistent linkage of a DMR to three SNPs nearby.

To summarize, while we see no evidence for direct linkage of DNA methylation and genetic changes, methylation variation specific to the HPG1 population is more likely associated to genetic variation than the epivariation shared with the greenhouse-grown lines. However, the extent of this association remains largely unclear.

5.15 Population structure of HPG1 strains based on methylation variation

Lastly, we asked whether differences in methylation in the HPG1 accessions reflected genetic relatedness, i.e. population structure. Hierarchical clustering by methylation rates of DMPs showed association of biological replicates, followed by weak association according to sampling location (Figure **5.18**A). As for the MA clustering, when we used information from methylated positions not classified as DMPs, no clear associations were detectable (Figure **5.18**B). Similar to DMPs, hierarchical clustering of average methylation rates of hDMRs reflected the geographical sampling location, with strains from the same site consistently being close to each other (Figure **5.18**C). However, differences in methylation rates of MRs not classified as DMRs still were sufficient to recover population structure, albeit with less confidence, as indicated by the shorter branch lengths separating the clusters (Figure **5.18**D). This suggests that my DMR calling algorithm is conservative.

As for the HPG1 data set, a clustering of hDMRs between the MA lines recapitulated their genetic distances by separating replicates as well as early and late generation strains (Figure **S11**), albeit not as pronounced as for single-site epimutations (Figure **5.8**).

The clustering of hDMRs in the HPG1 population revealed, however, that the geographic outlier LISET-036 was the most different strain, despite being genetically not more divergent to the most recent common ancestor of the HPG1 population than the other strains. However, this was only observed for DMRs in the CHG context (Figure **S12**). Even so, LISET-036 had the most private hDMRs among all accessions, but their spectrum in terms of context and overlap with genomic features did not deviate from that of the other strains (Figure **S13**). From the hDMRs specific to this strain, 44 overlapped with genes and 30 with the flanking 1 kb upstream or downstream

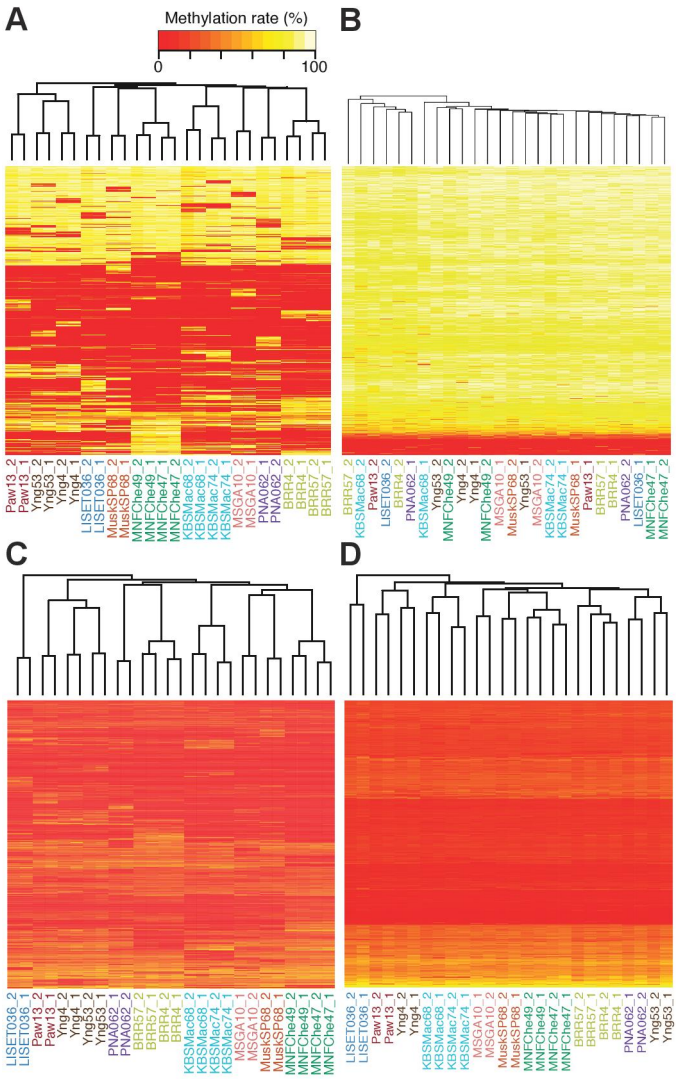

Figure 5.18: Hierarchical clustering by differentially and invariantly methylated positions and regions. (A), (B) Hierarchical clustering of haplogroup-1 (HPG1) strains based on methylation rates at 50,000 differentially methylated positions (A) and 50,000 invariantly methylated positions (B) in CG context. (C), (D) Hierarchical clustering of HPG1 strains based on average methylation rates of 2,829 hDMRs (C) and invariantly methylated regions with full information across all strains. Methylation rates per region were calculated as the average methylation rate of each methylated cytosine in that region.

region. The only GO term for which these 74 genes were enriched was "intrinsic to membrane" (P value 0.01). In addition, there were no overlapping differentially expressed genes. Thus, we found no evidence for a pronounced phenotypic effect of the epivariants unique to LISET-036.

Together, methylation data in general mirrored similarity between accessions at the genetic level, which can even be observed after only 34 generations, supporting the interpretation that methylation differences primarily reflect the number of generations since the last common ancestor.

5.16 Methods

Plant growth, nucleic acid extraction and library preparation was done as described for the mutation accumulation lines in ref. [Becker et al., 2011] and for the haplogroup-1 lines in ref. [Hagmann et al., 2015].

5.16.1 Sequencing and short read processing

For both data sets, we sequenced one genomic and two replicate bisulphite libraries per sample with 2x101-bp paired-end reads on an Illumina GAII instrument. Bisulphite libraries were constructed from individual plants for the MA lines and from pools of 8-10 individuals per HPG1 accession. For transcriptome sequencing, we used three biological replicates per HPG1 strain and produced 101 bp single end reads on an Illumina GAII instrument.

SHORE [Ossowski et al., 2008] was applied for all filtering steps of the raw reads. Following a similar approach as software provided by Illumina, it filters out reads that contain more than 2 or 5 bases in the first 12 or 25 positions, respectively, with a base quality of less than 3. In contrast to the Illumina software, it retains high-quality read partners of filtered out reads (termed 'single' reads). Furthermore, reads harboring 10% or more of ambiguous base calls ('N' bases in the reads) were discarded. Bases with quality less than 5 were removed from read ends up to the right-most occurrence of two adjacent bases with quality above 5. I discarded reads trimmed to less than 50 bp. A low complexity filter removed reads that contained less than three different 3-mers, i.e. if a read consisted entirely of a single repeated nucleotide or dinucleotide. I chose the filtering criteria relatively conservatively to retain as much sequencing information as possible, and since high-error reads still get filtered out by the subsequent mapping process.

5.16.2 Short read alignment

Next, I aligned the quality-filtered and trimmed reads using GenomeMapper (an adapted version thereof for bisulphite-treated reads; see section 4.2), allowing for up to 10% of single-base-pair substitutions relative to the read length for both genomic and bisulphite reads, and additionally for up to 7% of single-base-pair insertions or deletions for genomic reads. In best-hit mode, GenomeMapper only reports alignments with the least amount of mismatches for each read.

To discard putatively falsely mapping reads, a paired-end correction method compared the distance between reads and their partner to the average distance between

all read pairs and removed reads with abnormal distances (differing by more than two standard deviations) if there was at least one other alignment of this read in a concordant distance to one of its partner.

5.16.3 Determining high-quality positions

For each sample, base calls at all positions for genomic reads and read counts on all cytosine sites for bisulphite-treated reads were obtained with SHORE 'consensus' or SHORE 'methyl', respectively. SHORE 'methyl' focused the analysis only on uniquely mapping reads and accounted for PCR duplicated reads (section 2.4.2). SHORE 'consensus' required reads covered with their "core region" to call reliable genetic variants, i.e. ignoring the first and last 10 read bases, as described in section 3.2.1.

SHORE 'consensus' and 'methyl' assign a quality score to each position measuring the reliability of the base calls or read counts (see section 3.2.1, and Table **S6** for the scoring matrices that were used). To compensate for sporadic coverage or base quality fluctuations across samples, I applied a flexible quality score cutoff scheme, following the strategy of Cao *et al.* [Cao et al., 2011]. Assuming mostly conserved sequence and methylation pattern between samples, if a high quality position is observed in at least one sample, the quality requirement in all other samples can be lowered. Thus, I utilized an upper quality bound of Q32, allowing at maximum one intermediate penalty, and a lower of Q15, allowing no more than two intermediate penalties (see section 3.2.1). This strategy increases the number of 'trusted' positions across samples and reduces incidences of missing data.

5.16.4 Determining DMP clusters

I retrieved DMP clusters (DMCs) following the strategy described in section 4.5.1 and used the criteria of ref. [Becker et al., 2011]: DMPs within 50 bp were merged for each pairwise comparison into a sample-specific DMC and filtered by length ($>$ 50 bp), minimum number of methylated sites (10) and differentially methylated sites (5). Sample-specific DMCs were consolidated if they overlapped by 20% of their combined length and if the methylation change was in the same direction compared to a reference sample (3^{rd} generation strains in the MA, and LISET-036 in the HPG1 data set).

5.16.5 Comparison of epivariation between HPG1 strains and methylation-deficient mutants

The data from Stroud and colleagues [Stroud et al., 2013b] contain position-wise methylation rates for each sample. I defined a single site as methylated in wild type (WT) if both Col-0 samples Col_WA034L3 and Col_WB023L8 had a methylation rate of 10% or higher, and if at least one of them is more than 20% methylated. I declared a site in a mutant sample as having 'lost' methylation where the wild type was methylated and the mutant showed a methylation rate of less than 10%. In contrast, a

'gained' methylation site had less than 10% methylation in at least one of the WT samples and more than 20% methylation in the mutant. To assess if epigenetic variation in the HPG1 lines is enriched at sites affected by impaired methylation machinery, for each mutant, I constructed a set of positions, which were methylated in WT, covered in the mutant sample (i.e. present with a rate in the mutant sample file), and which were covered in the HPG1 and MA populations. A site was considered covered in a population when more than half of the strains showed a high quality and a more than 3-fold covered base call (see section 5.16.3). For those positions and different subsets thereof, the fractions of sites with gained or lost methylation in the mutant compared to the wild type samples were plotted in Figure **5.16**.

5.16.6 Data accessibility

All sequencing data have been deposited at the European Nucleotide Archive under following accession numbers:

- DNA sequencing data of mutation accumulation lines: ERP000902
- DNA sequencing data of the haplogroup-1 lines: PRJEB5287
- RNA sequencing data of the haplogroup-1 lines: PRJEB5331

A GBrowse instance for DNA methylation profiles for the MA lines is available at http://gbrowse.weigelworld.org/fgb2/gbrowse/ath_methyl_ma.
Methylome and transcriptome data for the HPG1 strains is available in a GBrowse instance at http://gbrowse.weigelworld.org/fgb2/gbrowse/ath_methyl_haplotype1.

Chapter 6

Discussion

Today, DNA methylation changes are widely accepted as important contributors to phenotypic variation. These naturally occurring epimutations are often linked to genetic variation (section 1.7.1). In the last two decades, numerous spontaneously occurring and heritable epivariants have been identified (section 1.7.2). In addition, environmental signals are known to induce targeted, potentially adaptive epigenetic changes (section 1.7.3), but there is a current debate on whether such acquired epimutations can be stably transmitted to the progeny [Danchin et al., 2011, Boyko and Kovalchuk, 2011, Bonduriansky, 2012, Pecinka et al., 2010, Paszkowski and Grossniklaus, 2011, Hirsch et al., 2012, Heard and Martienssen, 2014] (section 1.7.3). Such transgenerational inheritance of acquired changes would implicate that the environment exhibits a direct impact on evolution, circumventing classical natural selection. Moreover, even genome-wide changes of the methylation landscape upon environmental stress have been recently reported [Kovalchuk et al., 2003, Boyko et al., 2007, Boyko and Kovalchuk, 2010, Verhoeven et al., 2010]. All these studies could not fully disentangle the different sources of epivariation and usually did not identify the whole range of genetic and epigenetic variation. Thus it has remained unclear how frequently stochastic DNA methylation changes emerge genome-wide, and how strongly and persistently natural environmental conditions impact the epigenetic landscape without broad genetic confounding effects.

To obtain a rate of spontaneous epigenetic variation and to gauge the impact of the environment on the epigenome over many decades, we have conducted detailed whole-genome DNA methylation analyses at single-base-pair resolution of two *A. thaliana* populations that each constituted a unique analytical framework to specifically assess different sources of epigenetic variation. Investigation of genetically identical 'mutation accumulation' (MA) lines that were raised in uniform and benign greenhouse conditions over thirty generations ensured minimal genetic and environmental influences. We found that spontaneously occurring pure epimutations emerge much more often than random genetic mutations and that there are privileged genomic loci that appear to frequently switch their methylation state. The second population consisted of

'haplogroup-1' (HPG1) accessions that had grown under natural conditions in North America and had diverged only over the past few hundred years. While previous studies typically exposed plants to single or few specific stresses, monitored DNA methylation changes over only one or two generations after stress treatment, or analyzed genetically distant accessions, we were able to gauge the impact of diverse and fluctuating environmental conditions on the epigenome over a previously uncharted period of a few centuries under minimal genetic influence. However, we found little evidence that the environment has a broad and durable effect on the spectrum and accumulation rate of epimutations in our *A. thaliana* strains.

Distinguishing the different sources of epigenetic variation requires knowledge of the genome-wide genetic variation. To this end, I developed a sensitive method to identify a wide range of classes of genome-wide genetic polymorphisms that facilitated the detection of shared variants in the HPG1 population to test their association with epigenetic differences. This concomitantly increased the accessible genome space for DNA methylation analyses.

Finally, I introduced a novel approach to call regional methylation differences (DMRs) between strains. Most previous methylation analyses in plants detected regional epivariation based on significant differences at single sites, which increases the multiple testing problem and restricts detected methylation differences to more highly methylated CG sites. By contrast, the method described in this work used sensitive statistical testing of whole regions allowing us to find more subtle methylation differences. This revealed a more unbiased distribution of epigenetic variability along the genome.

In this chapter, I will highlight selected aspects of the novel methods and discuss the main findings that I presented in this thesis, before drawing conclusions and providing an outlook on further research avenues to better elucidate the pattern of DNA methylation variability in plants.

6.1 An integrated pipeline to call genetic variation

Myriads of genetic variants have been identified in the last decades with the goal of explaining the extensive phenotypic diversity in nature [Visscher et al., 2012, Atwell et al., 2010]. Until today, most studies rely on rather short next-generation sequencing reads and profile genetic variation using the resequencing approach. For this data, reliable SNP calling is routinely performed, but the identification of larger structural variants (SVs) has not reached a consensus or high quality level. However, SVs might play an equally important role as SNPs in shaping phenotypic traits [Weischenfeldt et al., 2013, Saxena et al., 2014] and are a major source of epigenetic variation (section 1.7.1). Available SV detection methods differ substantially in the types and sizes of SVs they can predict and yield only a limited picture of the global set of genetic variation [Mills et al., 2011, Lin et al., 2014].

To enhance the detection of DNA sequence polymorphisms, I have presented a complete workflow that combines the benefits of diverse genetic variation calling approaches and that uses information about known or predicted variants from different samples. It relies on short next-generation sequencing data and is suitable when whole-genome assembly methods are infeasible due to read length, fragment size, coverage or cost limitations. The strategy uses the resequencing approach to call variants on genomic regions covered by uniquely mapping reads, and uses targeted *de novo* assembly and multiple SV calling methods to span regions with absent or very low resequencing coverage. The workflow consists of two major steps for each sample: the consolidation of SV calls from different sources and the validation of a potentially population-wide set of SVs by re-mapping reads.

Compared to most other methods that consolidate SV calls from different tools (section 2.2.5), my workflow can integrate a user-defined set of SV callers without restriction on specific types or lengths of variants. Furthermore, some SV merging tools report only the intersection of SV detection programs, which drastically reduces the set of bona fide variants and the sensitivity, given the high false negative rate of SV detection tools and the fact that identical SVs between different tools can have different genomic coordinates. Other merging tools rely on overlapping genomic coordinates, increasing the false positive rate, especially for more complex regions. To circumvent these problems, the presented pipeline combines variants by comparing resulting haplotypes for equivalence independent of genomic coordinates, which increases accuracy and considers possible nearby polymorphisms at the same time.

Accounting for the high false positive rate of single SV detection tools, predicted variants are validated by re-aligning reads to the reference genome and to the alternative haplotype sequences generated from shared but also tool-specific variants. This enables the direct comparison of the read support for different alleles and the testing for homozygosity, which reduces the detection of false positives. Moreover, it allows the incorporation and validation of additional variants, such as known polymorphisms or polymorphisms identified in other samples, which increases sensitivity by facilitating the detection of shared variants between closely related strains, like those of the haplogroup-1 population.

Since the pipeline was designed to use tools that predict variants at base-pair resolution, polymorphisms can be incorporated into the reference genome, thereby refining it from iteration to iteration. This increases the number of variants and the chances to detect complex variation.

A comparable approach was followed by Gan and colleagues [Gan et al., 2011], who established 18 reference genomes of natural *A. thaliana* accessions by iterative resequencing and also applied two different approaches to call genetic variation: resequencing and whole-genome *de novo* assembly. They mainly combined contradicting variant calls by prioritizing insertions and deletions over SNPs, which were more often reported by the assembly tool. By closely comparing the approaches, many unique features of my workflow are apparent. The method presented in this work makes use of known variant information in calling genetic variation (increasing sensitivity) and it generates

a reference sequence based on multiple samples, which makes it suitable for analyzing populations of strains. Instead of relying on a fixed number of a few approaches, my pipeline can integrate a plethora of SV detection tools, which potentially enables the detection of an enlarged set of variants of different types and lengths. Lastly, my workflow includes (statistical) testing of different, contradicting alleles between samples rather than only one alternative allele per genomic region, which also enhances sensitivity.

I applied the method on the near-identical mutation accumulation lines and compared the variants to previously called polymorphisms that had been validated by Sanger sequencing [Ossowski et al., 2010]. While the overlap of SNPs and short indels was high, my pipeline did not call the four validated larger SVs, mainly due to cross-mapping reads. This indicates that the pipeline is overly conservative in more repetitive and complex regions, which may be explained by the different incentives of this study compared to the previous MA line analysis. My pipeline accounted for the high false positive rate of SV detection algorithms and was optimized using rather stringent criteria to limit the number of false positives. By contrast, Ossowski and colleagues used ∼500 Sanger-validated sequences to optimize their variant calling parameters especially for the calling of singleton variants, i.e. they were optimizing for high sensitivity. The fact that 500 regions were sequenced might imply that there have been around 500 candidate variants, of which 116 turned out to be true variants. I roughly estimated that the false positive rate of the pipeline presented here is likely well below 10%.

Limitations and further improvements

The presented method has two main limitations. First, it assumes homozygosity throughout the whole genome because it was tailored to the analysis of *A. thaliana*, which is homozygous throughout the genome. Variants called by different tools are discarded if they overlap with the inner part of any resequencing read, and alleles are explicitly tested for homozygosity. However, these two main criteria can be adapted to be applicable for mating individuals. Omitting the resequencing coverage filter would, however, result in the validation of much more variants in the branched reference sequence.

The complexity and size of the branched reference sequence constitutes the second limiting factor and might render the pipeline applicable for only a moderately sized population of rather closely related strains. The number of strains, their genetic distance, the number of applied SV detection tools and their false positive rates all increase the number of different alleles and combinations of nearby alleles that are separately incorporated into the branched reference sequence. While the runtime is assumed to be feasible for small genome sizes such as that of *A. thaliana* and given the high performance of current short read aligners, it remains to be explored whether the overly enlarged genome space to which reads are mapped might become especially prone to cross-mappings. The total length of the branch sequences for the closely

related 13 HPG1 strains already amounted to ∼10% of the *A. thaliana* genome.

A possibility of improving the pipeline is a more elaborate merging of SVs. While the genetic variation pipeline prefers sequence evidence from the resequencing method to other sources, it treats SVs from different SV callers equally and discards contradicting haplotypes (unless there is a majority haplotype). A priority scheme for different methods based on their unique advantages might further improve the accuracy of variant calling [Lin et al., 2014].

Moreover, besides combining different SV detection tools based on NGS data, data from different sequencing platforms, such as from single molecule sequencing technologies (see section 2.1.2), can be used at low coverage to assist SV calling especially in complex repetitive regions, replacing (or improving) the *de novo* assembly method of short reads in the pipeline. Strategies integrating different sources of sequencing reads are now emerging (e.g., [Ritz et al., 2014, English et al., 2015]).

In summary, I presented a workflow for short second-generation sequencing data that aims at both maximizing the identification of genetic variation and maintaining a high level of accuracy for the analysis of groups of closely related strains such as natural local populations. While improved sequencing and assembly technologies are already on the market that easily allow the reconstruction of whole genomes and – once broadly affordable – will supersede complex SV detection approaches in the near future, the best option until that time might be the combination of a large number of different genetic variation calling methodologies [Lin et al., 2014].

6.2 Bisulphite sequencing pipeline

Exploring the DNA methylation landscape has become both increasingly popular and feasible in recent years. Epigenetic changes are involved in short-term phenotypic plasticity and might also contribute to phenotypic variation in the long run. Analyzing the whole-genome DNA methylation pattern using the state-of-the-art next-generation sequencing technology, however, imposes several computational challenges, from handling large data over corrections of several biases to sound statistical testing.

I have presented a comprehensive pipeline to statistically detect DNA methylation as well as differential DNA methylation signatures for NGS data from plant genomes. Identifying methylated cytosines mainly follows the best practice for WGBS-Seq data, while the unique advantage of my workflow is the application of a coverage-dependent incomplete bisulphite conversion rate, which decreases the false positive rate of methylation calls.

The pipeline employs Fisher's exact test to call differentially methylated positions, which is the most commonly used approach to call DMPs in plant studies. While this test might report considerable false positive DMPs, as it does not incorporate biological variance and only allows for a low sampling variance, tools that remedy this

disadvantage by relying on beta-binomial models have not been widely applied beyond proof-of-principle studies and were designed for human data, thus only accounting for CG methylation. Moreover, it remains to be shown in objective analyses that they accurately estimate per-site methylation level distributions based on (usually very) few biological replicates only [Robinson et al., 2014].

Novel DMR detection approach

The main novelty of the WGBS-Seq pipeline in this work is a unique approach to call differentially methylated regions that consists of two main steps and differs in many aspects to previously reported methods. Rather than relying on arbitrarily pre-defined regions (e.g. tiling regions, annotation features, CpG-rich regions) and independent of DMPs and user-specified DMP merging criteria, the workflow first identifies methylated regions (MRs) in an unbiased and informed manner by an unsupervised Hidden Markov Model (HMM) that is trained on genome-wide methylation data. The method was tailored to human methylation data and I adapted it to the more complex plant methylation pattern, mainly by handling the different sequence contexts separately. The validity of the resulting MRs was demonstrated by the high overlap to enriched methylated regions determined by an independent experimental approach (MeDIP-Seq; section 5.9). The HMM incorporates information from lowly covered sites and does not require a minimum per-site read depth. While an HMM was already applied to call hypomethylated regions in a study of the (largely methylated) maize genome [Regulski et al., 2013], no specific details of the implementation were given and no biological conclusions or further analyses like DMR calling were made.

The second analysis step is the statistical testing of MRs for differential methylation in pairwise sample comparisons. The statistical test employs the currently widely used beta binomial strategy that models the within-sample and between-sample variance of methylation rates, thus accounting for read depth fluctuations and biological replicate data to increase specificity. To my best knowledge, it has not yet been applied in any other plant study. In contrast to most available software packages, we apply the test on regions rather than on single positions only. This strategy, together with analyzing only the methylated space of the genome, heavily reduces the number of multiple testing corrections and ensures a high level of statistical power. Focusing on regions further enhances sensitivity for regions of weak methylation differences or of low sequencing coverage. Two available tools also directly test regions rather than single sites using beta binomial models [Hebestreit et al., 2013, Park et al., 2014], but they only account for CG positions in the genome and rely on pre-defined regions.

Lastly, another unique feature of my pipeline when studying a small number of samples is that it includes a method to classify samples into groups based on all pairwise significance tests and thus provides a way to determine epiallele groups without relying on an arbitrary reference strain.

The improvements of my DMR detection strategy over different other DMR detection approaches are summarized in Table 6.1.

Table 6.1: Comparison between my novel approach to detect differentially methylated regions (DMRs) to previous methods. DMP: differentially methylated positions.

DMR calling approach	**Improvements of my pipeline over approach on the left**
DMP-based *without testing regions:* [Lister et al., 2009, Qian et al., 2012, Chodavarapu et al., 2012] *with testing regions:* [Calarco et al., 2012, Ausin et al., 2012, Hodges et al., 2011, Schmitz et al., 2011, Ziller et al., 2013]	• reduced false positive rate compared to set of references (on the left) that do not perform statistical tests of regions • improved statistical power by drastically reduced number of statistical tests • no need for arbitrary DMP merging criteria or pre-knowledge about DMR lengths • less biased towards (highly methylated) CG sites • possible to detect weak regional methylation differences (e.g. without harboring DMPs) • information of low-coverage sites used by testing whole region • more sensitive test accounting for broader variances and biological replicates (beta binomial-based)
Whole-genome tiling regions and statistical test (Kruskal-Wallis, Fisher, chi-square) [Lister et al., 2011, Stroud et al., 2013b, Stroud et al., 2013a, Yu et al., 2014, Regulski et al., 2013]	• improved statistical power by reduced number of statistical tests • informed selection of regions (by HMM) • more sensitive test accounting for broader variances and biological replicates (beta binomial-based)
Beta binomial-based test on regions (pre-determined regions and test on CGs only) [Hebestreit et al., 2013, Park et al., 2014]	• *de novo* detection of regions for testing • context-specific testing (suitable for plant methylation data)
Exclusion of low-coverage sites and DMR reporting for each pairwise comparison only [most, if not all, DMR detection methods]	• improved specificity (lower false negative rate) • less biased towards more highly covered regions • reporting epialleles by grouping samples

As there is neither a gold standard nor a stand-alone DMR detection software for plant methylation data, I compared the DMRs of my novel method to those obtained by the most widely used approach in plant epigenetic studies that relies on clustering DMPs by genomic distance (DMCs). The DMRs of my novel strategy showed a different distribution along the genome that largely followed the overall distribution of

methylation. Furthermore, it detected more differences in lowly methylated CHG and CHH sites that are hardly identified with DMP-based methods. In contrast to DMCs, the identified DMRs contained fewer DMPs and less often overlapped with genic regions. Although DMCs therefore have higher chances to affect gene expression, we did not find examples of negative correlations between differential methylation and gene expression in regions covered by DMCs but not by DMRs, which indicates that DMRs might capture most biologically meaningful methylation changes.

A common problem in quantitative studies is to gauge biological significance from statistically significant results. Large read counts in a region due to high sequencing depth or a high number of cytosines can result in a highly confident estimate of the methylation rate. Thus, tiny methylation differences are more likely called significant. This is why quantitative studies typically use additional, arbitrary filter criteria to select the most obvious differences. In the presented pipeline, I require non-overlapping confidence intervals of the beta binomial distributions between two samples for DMRs (section 4.6.3) and a minimal three-fold change of the mean methylation rates for the identification of hDMRs (section 4.6.5).

However, it is largely unexplored whether and to which extent subtle differences are biologically relevant. A promising way to shed more light into this area is the sequencing of DNA from single cells, which will allow comparing binary methylation states (see section 6.4).

Runtime and scalability

Despite having applied the DMR pipeline solely on the two *A. thaliana* populations described in chapter 5, I expect that it is suitable for other data sets as well. The runtime of the HMM to detect methylated regions scales linearly with an increasing number of cytosines (Figure **S14**). For larger genomes such as those of crop species, the performance can be enhanced if the HMM is applied on each chromosome in parallel, when chromosomes are comparable in length to the whole *A. thaliana* genome to ensure a sufficiently large und unbiased training data set.

The step in the DMR pipeline that selects regions for testing operates independently on genomic regions separated by unmethylated space across samples ('methylation islands', section 4.6.2). The runtime of this method depends on the number of MRs with different coordinates within methylation islands, since the combinations of all start and end coordinates of MRs determine the set of regions to analyze. The number of combinations might become infeasible for a (large) population of samples exhibiting long methylation islands and many different, short MRs. However, such situations might be rare. Additionally, the filter steps prevent 'similar' regions from duplicate testing. Thus, I deem it likely that this step of the DMR pipeline is computationally feasible for many data sets.

By contrast, the statistical test of the regions requires extensive computation time due to the complex numerical optimization and is only feasible with high parallelization even for small sets of samples. Furthermore, the tests are performed for each

pairwise sample comparison. The number of pairwise comparisons grows quadratically with the number of samples and thus quickly becomes infeasible. To reduce pairwise comparisons, a reference strain can be selected to which all other strains are compared, resulting in a linear increase in the number of pairwise tests.

However, analyzing a large set of samples increases the multiple testing problem. Thus, for hundreds or thousands of samples, pre-filtering of the most promising candidate regions can be performed, based for example on absolute methylation rate differences, as similarly done for the detection of DMPs [Seymour et al., 2014]. Alternatively and maybe most efficiently, since the HMM statistically classifies methylated and unmethylated regions, it might be sufficient to define DMRs as regions that are in a different state between samples without explicitly testing for their difference. However, this brings along the problem of determining a unified set of regions across samples and identifies solely presence/absence methylation differences.

Limitations and further improvements

Besides the limited scalability of the presented DMR detection method in its current implementation, it also restricts the analyses on completely homozygous samples. Thus, it cannot detect imprinted loci [Gehring, 2013] or the whole spectrum of differential methylation in outcrossing species or in hybrids of different strains. Accounting for allele-specific methylation requires the distinction of the read set into two groups based on linked genetic variants [Chodavarapu et al., 2012], or based on two different methylation patterns, each of which nearly represents half of the complete data [Fang et al., 2012, Peng and Ecker, 2012].

The method to call methylated regions can be extended to explore the methylation pattern further. While the output is deemed to be highly accurate, as most strikingly documented by the high overlap of the HMM-based MRs with methylated domains retrieved by MeDIP-Seq, the use of only two states (methylated/unmethylated) might represent an oversimplification of the real methylation profile in plants. Since methylated genes contain exclusively CG methylation, interspersed unmethylated CHG and CHH sites might prevent the calling of methylated regions at these loci, as the HMM considers all sequence contexts in combination. Thus, a possible and feasible adaptation of the HMM to potentially better represent the plant methylation landscape would be to model a third state that represents gene body methylation. This might lead to the identification of more DMRs within genic regions and might result in a higher overlap between DMRs and DMCs. Along these lines, modeling even more states might detect additional, so far 'hidden' methylation patterns.

Although the statistical test accounts for technical and biological variance, it could also incorporate the incomplete bisulphite conversion rate as an additional source of variation. However, by comparing regions rather than single positions, this bias might be better compensated for, given similar per-sample false methylation rates. Moreover, a mathematically sound incorporation into beta binomial models remains to be shown. Although the software BEAT accounts for false positive and false

negative conversion rates in approximating methylation levels, it uses a simplified binomial model and does not propose a method to compare samples [Akman et al., 2014].

In summary, this work introduced a novel approach to call differentially methylated regions, which reduces the setting of largely arbitrary parameters as much as possible. While previous epigenetic studies employed rather simple (statistical) strategies, the presented method detects DMRs in an informed and more unbiased manner using latest sequencing data. Thus it might serve as a basis for more elaborate statistical methods in future methylation studies.

6.3 Short-term evolution of DNA methylation

By applying the previously discussed methods, I analyzed the whole-genome pattern of genetic and epigenetic variation in two *A. thaliana* populations to explore the spectrum, rate and possible sources of naturally occurring DNA methylation changes over short evolutionary time periods.

Low mutation and high epimutation frequencies

Genome sequencing of the MA and HPG1 strains revealed low genetic diversity within each population. I determined an emergence rate of approximately one DNA mutation per line and generation in the MA lines, which was slightly higher than that reported in a previous study on these lines [Ossowski et al., 2010]. This can be explained by the longer read lengths and higher read depths in our study. In the HPG1 population, I identified only around 2,000 segregating genetic variants, underlining its suitability as a near-isogenic "natural mutation accumulation line".

Intriguingly, methylome analyses of both populations revealed many orders of magnitude more naturally occurring DNA methylation changes at single sites (DMPs) than genetic mutations. However, only a small fraction of them was arranged in dense genomic regions, and the frequency of these DMP clusters (DMCs) was comparable to the rate of DNA mutations. The differentially methylated regions (DMRs) identified by the novel and more sensitive approach presented in this work (chapter 4) emerged at an intermediate rate between those of DMPs and DMCs.

Hence, changes of DNA methylation state occur frequently during even a few generations and are largely independent of genetic variation. This high scale of epigenetic variation in the MA population was in strong concordance with an independent study that independently performed methylome analyses on a subset of these MA lines [Schmitz et al., 2011].

Stable TE and less stable genic DNA methylation

The different approaches to call variable methylation revealed rather distinct distributions of DNA methylation changes in the genome. While DMPs and DMCs were mainly found in genic regions on chromosome arms, DMRs preferentially located in and around the centromere, in regions that are largely devoid of genes and rich in transposable elements (TEs). Although many DMRs were found in TEs, all types of epimutations were underrepresented in TEs and overrepresented in genes. Thus, variable DNA methylation occurs disproportionately outside of TEs, suggesting that the latter are stably maintained in a silent state to potentially prevent their transcription or spreading, whereas fluctuating methylation in genes is more common.

While this is consistent with previous reports [Vaughn et al., 2007, Zhang et al., 2008, Schmitz et al., 2011, Schmitz et al., 2013a], whether this reflects an actual biological phenomenon or is influenced by the bias of preferentially detecting differential methylation at CG sites remains unresolved. Compared to methylated CHG and CHH sites, which are almost exclusively contained in TEs and intergenic regions, methylated CG sites, prevalent in genes, usually have high methylation levels, which increase chances of calling significant differences. My novel DMR detection method identified much more CHG- and CHH-methylated regions that exhibited weaker methylation differences than those detected by DMP-based approaches. That the bias towards higher variability in genes still existed for the detected DMRs supports the hypothesis of higher TE methylation stability. However, the method was still conservative, since a clustering of invariantly methylated regions also showed systematic differences between the haplogroup-1 strains, albeit less pronounced as for the clustering of DMRs.

Regardless of this uncertainty, my novel DMR detection method revealed that CHG and CHH methylation is almost exclusively organized in densely methylated, consecutive regions in and around the centromere. By contrast, constitutively as well as variably methylated CG sites within genes are sparsely distributed, which is the reason that they are less often included in (differentially) methylated regions. Thus, variability of DNA methylation in plant genes often affects isolated cytosines.

The mechanisms leading to gene body methylation as well as to its variance are still unclear, e.g. whether gene body methylation is simply the consequence of transcription (section 1.4.1), or to what extent DMPs in genes are a readout of epigenetic changes other than DNA methylation (section 1.4.2). Similarly, the biological relevance of highly variable CG methylation in genes remains elusive: We observed only few correlations between differential methylation (DMRs and DMCs) and gene expression in our analysis of the HPG1 lines, and no such correlations in sparsely methylated genic regions that were overlapped by DMCs but not DMRs. On the other hand, methylation of homologous genes is in general evolutionarily conserved between species [Zemach et al., 2010, Takuno and Gaut, 2013, Seymour et al., 2014]. One interpretation of these findings would be that epivariation at single sites is tolerated as long as the gene is globally methylated to a specific extent.

Frequent reversions of epimutations

We found that epimutations that have accumulated over thirty generations were sufficient to separate the MA lines into early and late generation strains, as well as into groups of replicates. This implies that a fraction of seemingly random DNA methylation changes are transmitted across generations, which is also supported by many studies that identified natural epialleles (section 1.7.2).

However, by comparing the frequency of epimutations between MA lines that were separated by different numbers of generations from each other, we found that DNA methylation changes accumulated sub-linearly over time. This indicates that many of them were not stably inherited over the long term, and that loci reverted their differential methylation status to an initial state. A striking illustration that reversions even occur in larger regions from one generation to the next – thus without going through gradual, intermediate states – was our observation of a region that had become demethylated after 31 generations, but was re-methylated in the following generation. This constitutes a profound difference to DNA sequence mutations, where reverse mutations are exceedingly rare.

Similar epimutation rates in greenhouse and nature

Given the century-long exposure to natural environments and the absence of large-scale genetic variation, we deemed the HPG1 population suitable to test whether the environment has a broad and durable impact on the accumulation rate of epialleles. To assess only the heritable fraction of epigenetic changes, we grew the accessions for two extra generations in the greenhouse after collection at the natural sites. This also ruled out potential parental effects, which can induce epigenetic changes in immediate offspring. In light of the high responsiveness of the epigenome to environmental cues and of reports claiming a long-lasting epigenetic memory (section 1.7.3), this unique natural experiment allowed us to clarify whether long-lived epimutations accumulate at higher frequency in nature than in the greenhouse.

In contrast to this assumption, we found that the epimutation rate under natural growth conditions did not substantially differ from that in a benign greenhouse environment, although the rate in the HPG1 strains was slightly underestimated (section 5.8). Moreover, polymorphisms accumulated sub-linearly in both situations due to frequent reversions.

Given the commonly reported immediate but transient methylation changes upon environmental triggers (section 1.7.3), the extent of epigenetic variation would most likely have been higher if we had sequenced field-grown individuals directly. However, most of the DNA methylation changes do not seem to be heritable. In fact, the vast fraction of the methylated genome space in the HPG1 strains was predicted to be invariant (97%), i.e. devoid of DMRs. Notably, a very similar fraction of the genome was invariant in the MA lines. Although the DMR calling was slightly conservative,

these findings do not support a whole-genome effect of century-long exposure to a natural environment on the epigenome.

Little evidence for heritable local adaptations

Although we did not identify a global effect of the environment on the epigenome, there might be local adaptations. Environmentally induced epivariants have been proposed to be more often adaptive than random genetic variants, and to be potentially stably transmitted to following generations [Boyko and Kovalchuk, 2011, Danchin et al., 2011, Bonduriansky, 2012, Jablonka, 2013]. Thus, they should be found in the HPG1 population. To assess how many heritable epimutations might have phenotypic consequences in the HPG1 strains, we scanned for correlations between variably methylated regions and differentially expressed genes (DEGs), determined by transcriptome sequencing.

Strikingly, we found very few differentially expressed genes and thus little overlap between variable methylation and gene expression, independent of the approach to detect differentially methylated regions. We identified only two cases of evident negative correlation between DMRs and DEGs, and these genes were involved in disease resistance and pathogen response. Since these changes were consistent between strains from the same sampling location, this could indicate the presence of local adaptations related to the habitat. The fact that the relative fraction of HPG1-specific variably methylated regions overlapping with genes was higher compared to that of MA-specific DMRs could increase chances for the presence of adaptive epigenetic changes, but we explain this difference by an increased proportion of accessible sites in genes for the HPG1 population (section 5.13).

However, the low incidence of differentially expressed genes and the general lack of both phenotypic differences and correlation between methylation and gene expression changes suggest that epimutations are mostly neutral in nature, and thus comparable to genetic mutations. Hence, most transcriptional differences in the HPG1 population seem to be either due to cryptic associations with DNA methylation, due to epigenetic changes other than DNA methylation, or independent of epigenetic marks (e.g. due to DNA mutations). Thus, our analyses have revealed little evidence for large-scale and durable epigenetic differentiation that might have been induced by the variable and fluctuating environmental conditions experienced by the HPG1 accessions since they separated from each other.

The epigenetics community does not agree on whether local adaptations are more likely expected in short-lived plants like *A. thaliana* or in longer-lived species. On the one hand, one can argue that the need for epigenetic adaptations is higher in this "extreme inbreeding" plant because of lower genetic variation compared to outcrossing species [Tricker et al., 2012], or that the progeny of short-lived plants like *A. thaliana* is likely exposed to the same environments at similar locations, given the limited "seed

dispersal range" [Boyko and Kovalchuk, 2011]. On the other hand, environments frequently fluctuate, and it might not be beneficial to inherit adaptations to short-term growing conditions. A memory effect might even be less advantageous with increasing instability of the environmental conditions [Furrow and Feldman, 2014]. Investigating epigenetic inheritance in perennial plants would be an interesting avenue for further research, as they might record longer periods of climate. However, the long growth periods largely hamper controlled studies in these systems.

Obviously, a more direct method to identify correlations between stresses and induced epigenetic effects would be to perform experiments under controlled conditions. Indeed, the exact environmental conditions the haplogroup-1 accessions have been exposed to at their sampling sites are unknown, and the data provided by nearby weather monitoring stations only give a coarse overview of temperature and precipitation conditions. However, this data demonstrated the variability of the environment between the different sampling locations. Given the near-isogenic architecture of the HPG1 strains, we were able to test whether the entirety of all varied natural influences over more than 100 years leaves an epigenetic memory footprint, which represents a time period that has not been analyzed yet and could not have been achieved in a controlled greenhouse experiment.

Privileged loci of epigenetic variability

Irrespective of the methods to detect differential methylation, we found that variable single sites as well as regions were frequently shared between independent strains in both the MA and HPG1 data sets. In addition, the overlap of epivariation between both populations, as well as between accessions that diverged hundreds of thousands of years ago, was much higher than expected by chance, despite having experienced varied environmental conditions of different variability, and despite different genetic backgrounds. This suggests that there are loci in the *A. thaliana* genome that are privileged for DNA methylation changes.

We identified variable positions in the HPG1 population that overlapped more often with polymorphic sites between closely related MA lines than between MA lines separated by many more generations. Thus, there might be two classes of epimutations. One of these includes 'highly labile' sites that have a high mutation frequency and that are independent of the genetic background. These are therefore more likely found in different populations. By contrast, the other class contains more stable sites of lower epimutation rate that are less likely shared between populations.

The 'highly labile' sites likely derive from imperfections of the CG methylation maintenance machinery in *A. thaliana*, as almost all epimutations in CG methylated regions were hypomethylated in plants deficient for the main components of this pathway (MET1, VIM1-3). Error-proneness of the CG maintenance system might be common to a certain extent [Genereux et al., 2005, Fu et al., 2010], as also indicated by a study in *Arabis alpina* that discovered a putatively strongly decreased symmetrical CG

methylation maintenance [Willing et al., 2015]. However, the mechanisms that lead to the distinction into the proposed two classes of epimutations remain to be elucidated. Speculative hypotheses would be that metastable loci are more accessible to DNA methylases and glycosylases, or that loci that are more stably maintained are more strongly reinforced by other epigenetic marks such as histone modifications. Lastly, different DNA methylation pathways might have compensatory roles at more stable loci, e.g. in heterochromatic regions in which small RNAs and chromatin remodelers might contribute to maintaining methylation.

DNA methylation variation is overwhelmingly linked to genetic variation

DNA sequence changes are known to have a major effect on the epigenome (section 1.7.1), and we asked how much of the epigenetic variation in the haplogroup-1 population is either caused by, or stably co-segregated with genetic differences. We found that HPG1-specific highly differentially methylated regions (hDMRs) were more often linked to genotype variation than regions that were variably methylated in both the HPG1 and MA populations. Although we found only a single case of direct linkage of clustered DNA mutations with a methylation change in the HPG1 population, almost a fifth of the HPG1-hDMRs were highly associated with genome-wide genetic variation, as determined by a statistical approach used in genome-wide association studies.

Maybe the only other estimate of how many regional epimutations are associated with genetic variants based on a whole-genome level was performed by a study that analyzed 140 natural, distantly related *A. thaliana* strains from throughout the native worldwide habitat [Schmitz et al., 2013b]. Using a different methodology from that used in our study, Schmitz and colleagues estimated that ∼35% of variably methylated regions were largely associated with genetic variation in *trans* (they only found associations for regions exhibiting methylation in all three sequence contexts).

As these numbers likely represent underestimates due to missed genetic variation, they indicate the potential the genotype might have in shaping the DNA methylation profile. On the one hand, more associations could be identified by improved *de novo* assemblies or sequencing technologies that will reveal a more complete picture of genetic variants, especially in TE-rich (peri)centromeric regions. Furthermore, advanced association methods, that could potentially take biological replicates or even siRNA compositions into account, might prove beneficial in further studies to explore the genetic impact on the epigenome. On the other hand, identifying the complete set of correlations between genetic and epigenetic changes might be close to impossible because of the finding that altered DNA methylation states can be propagated even when the inducing genetic event is lost [Teixeira et al., 2009, Cortijo et al., 2014]. Moreover, DNA mutations in methylation machinery components can lead to a multitude of epigenetic changes [Stroud et al., 2013b, Shen et al., 2014, Willing et al., 2015], which is difficult to detect for association methods.

DNA methylation reflects population structure

Finally, we asked whether the general methylation pattern was different from the global genetic variation pattern. Hierarchical clustering based on variable methylation did not segregate the halogroup-1 lines into evident geographical clusters, similar to the clustering based on genetic variation. The few generations that separated the MA lines from each other were sufficient to reflect the pattern retrieved by hierarchical clustering based on DNA sequence changes. We conclude from this that the methylation pattern recapitulates genetic relatedness and that epimutations therefore accumulate like a molecular clock, and thus comparable to DNA mutations. This finding is in line with studies that show an increasing epigenetic divergence of duplicated genes with evolutionary age [Keller and Yi, 2014] and a gradual accumulation of DNA methylation changes from generation to generation, following a genetic event [Marí-Ordóñez et al., 2013, Silveira et al., 2013].

We observed that the geographic outlier LISET-036 from Long Island was the most different strain when clustering variably methylated regions (mainly due to differential CHG methylation). This strain also showed the highest number of DMRs unique to a single accession, despite being genetically no more diverged from the last common ancestor of HPG1 than the other strains. However, we found no clear evidence of genome-wide adaptive signatures: neither were LISET-036 specific regions of differential methylation in and near genes enriched for GO terms with an obvious connection to environmental adaptation, nor were there overlapping differentially expressed genes (section 5.15).

In light of the findings that epimutations arise frequently, but are mostly neutral and largely reflect the genetic distance, the conclusions of many studies using the MSAP method to detect methylation differences (section 2.3) might have to be put into perspective. Relying on data of a few hundred loci only, these studies find that epigenetic variation is larger than genetic variation and that samples can be better separated by habitat based on epigenetic rather than genetic markers, from which the authors infer the existence of frequent adaptations to different environmental conditions. Although these analyses are mainly performed in perennial plants, which might more likely accumulate local adaptations compared to annual plants like *A. thaliana*, it remains to be shown that the whole-genome epigenetic profile does not simply mirror the genome-wide genetic pattern.

Conclusions

Our analyses revealed that spontaneously occurring DNA methylation changes at single sites are much more frequent than random genetic mutations, while changes at extended clusters or regions of methylation were only moderately more frequent than DNA mutations. Although pure epimutations showed contrary characteristics compared to DNA mutations in that they are short-lived and seem to occur at privileged genomic loci independent of environment or genetic background, they appear to be

mostly neutral and reflect the genetic distance between strains, like genetic mutations. In addition, epigenetic variation did not accumulate at a faster rate in diverse natural environments than in uniform and benign greenhouse conditions.

These findings lead to the conclusions that heritable epigenetic variants accumulate rather as a function of time than as a consequence of rapid local adaptation caused by environmental signals, and that environment-induced DNA methylation changes are mainly short-lived and rarely – if at all – contribute to heritable epigenetic variation.

This study challenges recent reports claiming wide-spread heritable (and whole-genome) effects of environmental cues on the methylome. Our findings suggest that epimutations likely do not play a major role in evolution for *A. thaliana*, although we cannot rule out a limited number of subtle adaptive DNA methylation changes that are linked to specific growth conditions. It will be important for future studies that investigate the evolutionary role of the environment in durably shaping the epigenetic landscape to take into account the reversion rate of epimutations I have reported in this study. This high rate implies that most epialleles might not become subject to Darwinian selection. Furthermore, these studies need to profile the whole range of genetic mutations to rule out confounding genetic sources and use rigorous and sensitive statistical tests for differential methylation. My workflows to detect genetic and epigenetic variation might serve as an orientation for future studies.

6.4 Outlook

To better gauge the different contributions of stochastic, genetic and environmental factors to epigenetic variation in future research, I can envision several avenues of future research.

A valuable extension of our mutation accumulation line study would be to investigate the methylomes of even more lines, at best in consecutive generations, to retrieve a more accurate estimate of the emergence rate of epimutations as well as to more precisely estimate the set of positions that spontaneously fluctuate in their methylation states over time. Alternatively, the set of 'highly labile' sites could be refined by investigating methylation variation across many replicates of the same sample raised under uniform conditions. This might result in a clearer distinction of labile and more stable sites to spur exploratory analyses of possible sources of methylation instability.

Besides interrogating DNA methylation, further analyses need to go hand in hand with elucidating the interplay of all epigenetic players, including histone modifications, small RNA compositions or the positioning of nucleosomes to enhance our knowledge about epigenetic variation. Such integrated analyses are necessary to address the question whether, or which subset of DNA methylation marks are solely secondary effects following transcriptional, RNA-based or chromatin-level alterations.

Obviously, controlled stress experiments will reveal the direct influence of a specific environmental condition on the epigenome. However, to prove heritability of induced methylation changes, studies need to be designed to disprove the involvement of other

factors of epigenetic variation and sources of short-term epigenetic inheritance, such as parental effects. The latter can be achieved by monitoring epigenetic profiles over many following generations. Additionally, it might hamper our current understanding of the genetic source of epigenetic changes that we know least about the genetic variation of the most highly methylated regions in eukaryotic genomes: transposable elements and repetitive sequences. Thus, epigenetic studies need to be accompanied by the genome-wide detection of the entire genetic variation. I presented a pipeline to identify a wide range of DNA mutations, but it still suffers from the incomplete view on the genome that short reads can afford. Newer sequencing technologies such as the PacBio or nanopore single molecule sequencing technologies (section 2.1.2) are capable of producing reads in the kilobase-pair range and yield consecutive sequence information that can span long repetitive genomic regions [Huddleston et al., 2014, Krsticevic et al., 2015]. Already to date, these long reads assist in upgrading whole-genome *de novo* assemblies to a higher quality finished state, but further improvements in accuracy and read length promise to ultimately allow the identification of the complete set of genetic differences between strains. In combination with further advancements of tools that determine associations between genetic and epigenetic variants, this will enhance our understanding how many epimutations are linked to, or caused by DNA mutations.

The computational analysis of DNA methylation variation will benefit from more complete reference genomes as well, but will have to be advanced to a careful and sophisticated handling of repetitive sequences. Due to the kilobasepair-long reads, single molecule sequencing technologies can greatly facilitate methylation analyses in complex regions, and they are additionally able to distinguish methylated from unmethylated cytosines on the fly without pre-treatment of the DNA [Flusberg et al., 2010, Branton et al., 2008]. Omitting chemical conversion and PCR amplification steps that are associated with biases of current sequencing platforms such as the GC bias in Illumina sequencing has the potential to greatly enhance the technical quality of methylation calling, if the accuracy of identifying base modifications of these technologies will be reasonably high. Longer reads will furthermore enhance the elucidation of allele-specific methylation patterns.

A further leap forward in increasing our knowledge about epigenome variability constitutes single cell sequencing (section 2.1.2) [Kantlehner et al., 2011, Smallwood et al., 2014, Guo et al., 2014]. Most previous studies sequenced cell mixtures, but this setting does not allow distinguishing whether a small change in methylation level results from a small, homogeneous change across all cells, or from a large difference in a subset of cells only. Single cell sequencing will elucidate the scope of cell-to-cell variability and better measure the inherent noise in the epigenetic machinery. It will answer more detailed questions, e.g. why transposable elements rarely feature 100% methylation (in the CHG and CHH context in plants). This will determine whether single sites in TEs across a tissue are as fluctuating as sites within genes between individuals, or whether TEs are completely methylated in a small subset of tissue cells only. More importantly, single cell sequencing will enhance our knowledge about inheritance mechanisms of DNA methylation marks by sequencing individual germ

cells, and will help localizing possible subparts of a tissue that expresses responses to environmental cues.

Since methylation data derived from single cells should take values close to 0%, 50% and 100% for diploid organisms, it seems to make statistical testing obsolete for pairwise cell comparisons. However, the yet unknown technical variance as well as stochastic biological variance, or the chance sequencing of rare cells will introduce uncertainty and might necessitate the sequencing of multiple cells, thus leading to quantitative measurements again [Kantlehner et al., 2011]. Hence, analysis tools like that presented in this work might still be applicable for this kind of data.

In summary, progress in the molecular characterization of the epigenetic machinery systems, in the sequencing technologies and in the computational analysis of genomics and epigenomics data are underway and will enlarge our understanding of the mechanisms of epigenetic variation and epimutation stability in the near future. Until now, it seems the Modern Synthesis still holds without restrictions, but it will be exciting to follow whether advances in these fields will lead to the discovery of sporadic, unambiguous cases of transgenerational inheritance of acquired traits in nature.

Appendix A

Supplemental Figures

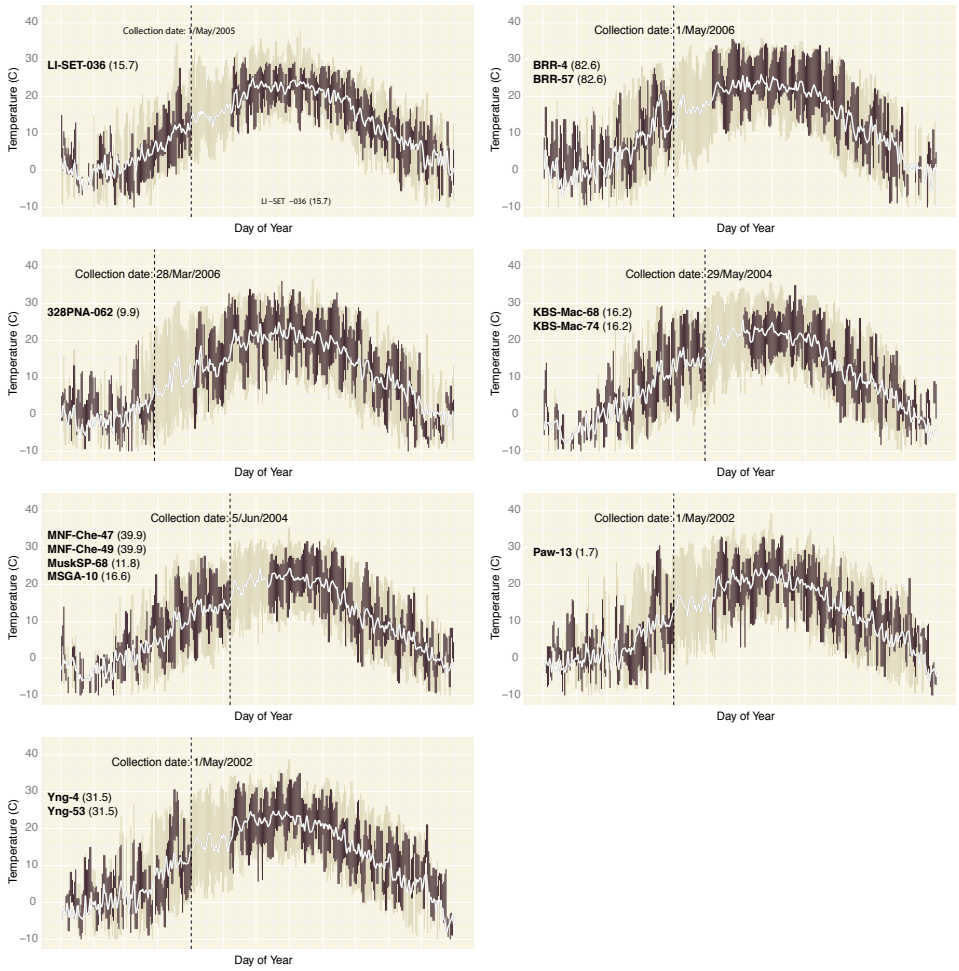

Figure S1: Recent temperature histories of samples. Local temperature data was calculated from National Climatic Data Center (NCDC) Global Summary of Day (GSOD) data. Collection locations were matched to the closest weather station (distance in km) with $< 5\%$ missing data for five years prior to the collection date. Daily temperature range (dark bars) and means (white points) of 330 days prior to the collection date, such that late year data reflect temperature ranges of the previous year. Five-year temperature ranges (light bars) and means (white line) indicate longer-term temperature variability. Different collection dates and locations both contribute to different means and variance of recent temperatures experienced by samples.

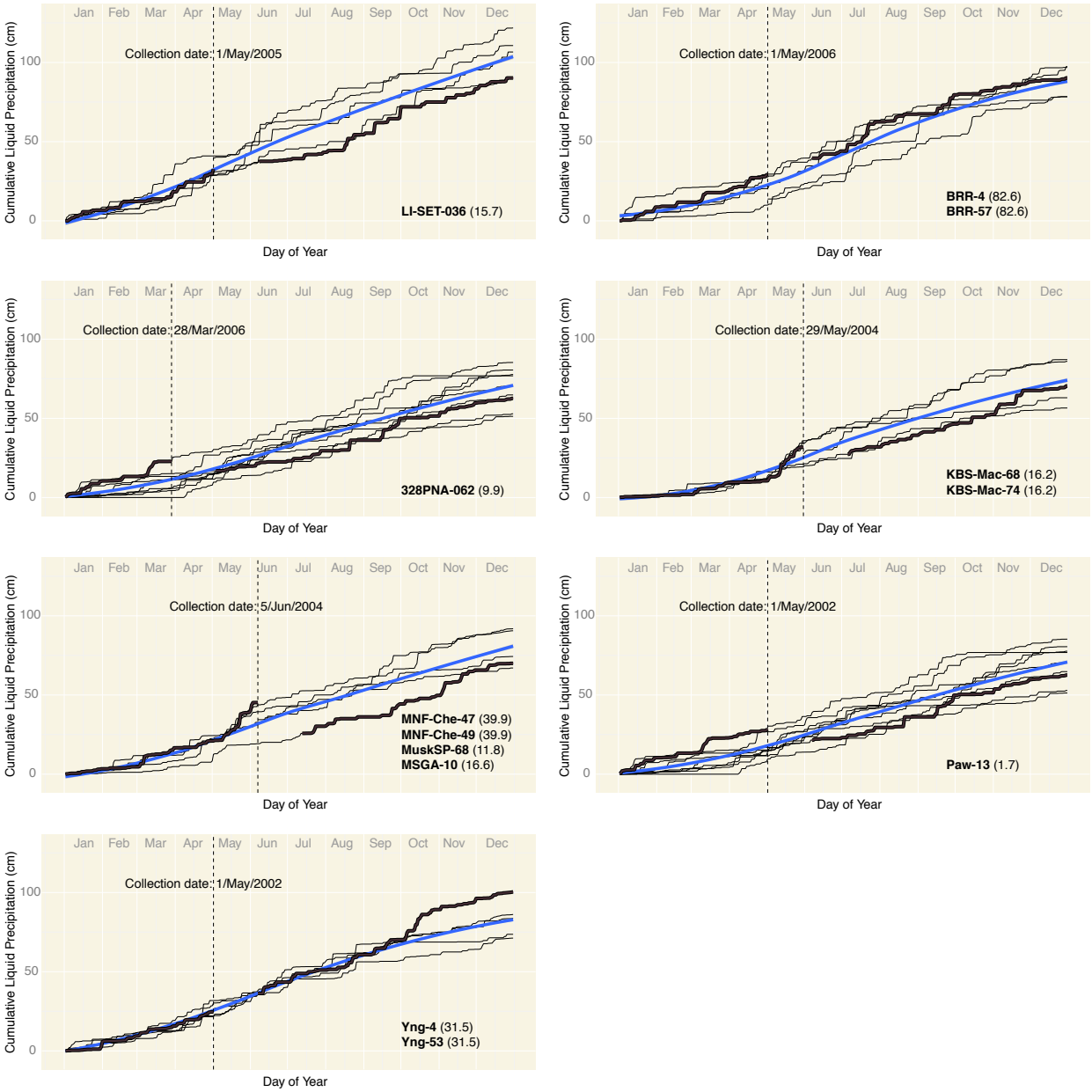

Figure S2: Recent precipitation histories of samples. Cumulative (from January 1) liquid precipitation at the weather stations closest to each collection site (distance in km), retrieved as for Figure **S1**. Black lines show yearly histories for five years prior to collection, the thick black line indicates the cumulative history of the previous 330 days. Blue line shows the LOESS estimate of mean precipitation accumulation over a year.

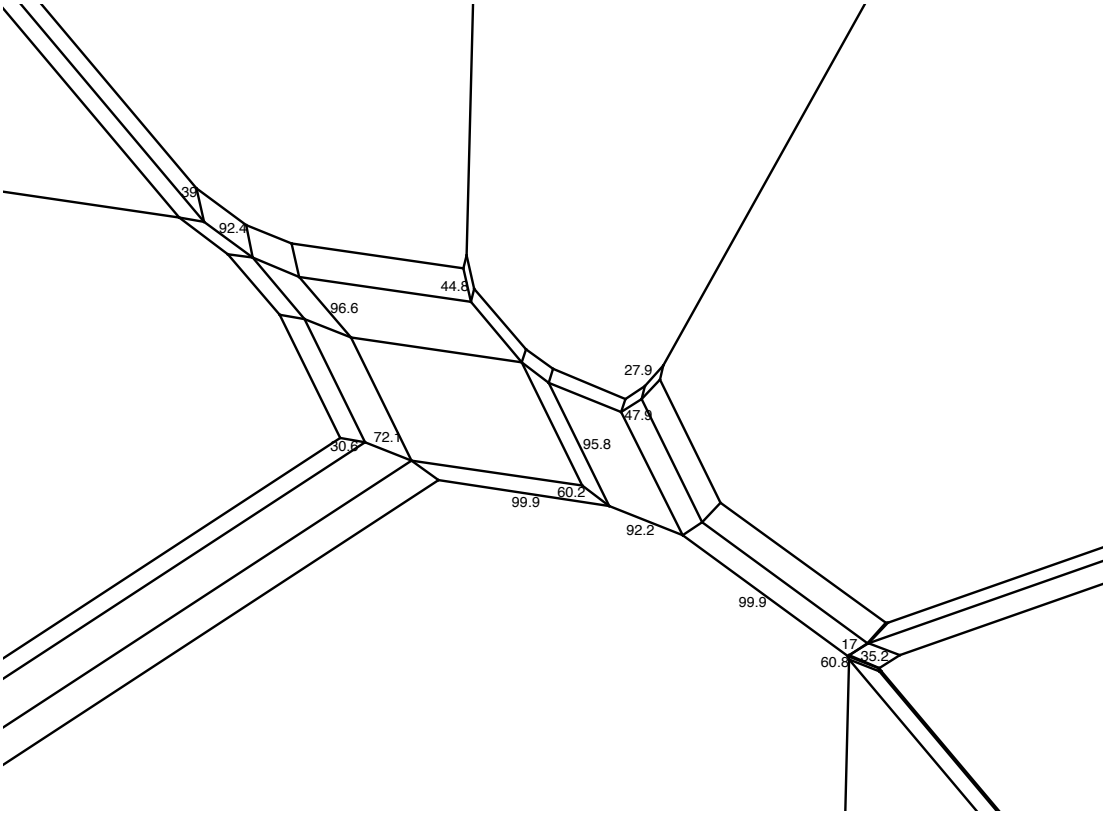

Figure S3: Magnification of the central area of the phylogenetic network in Figure **5.4**A. Numbers indicate bootstrap confidence values (10,000 iterations).

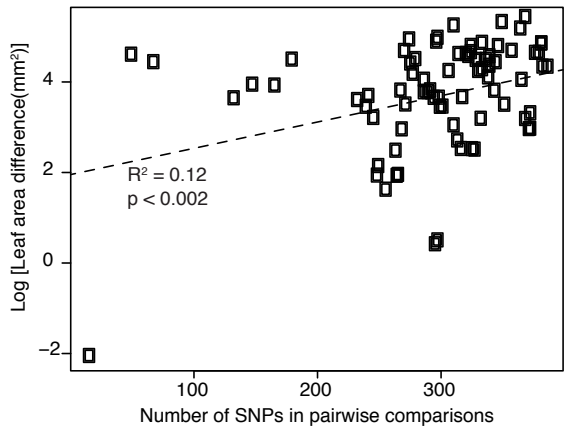

Figure S4: Phenotypic variation in the haplogroup-1 lines. Correlation of genetic distance, represented by the number of SNPs per pairwise comparison, and difference in leaf area at 21 days after germination.

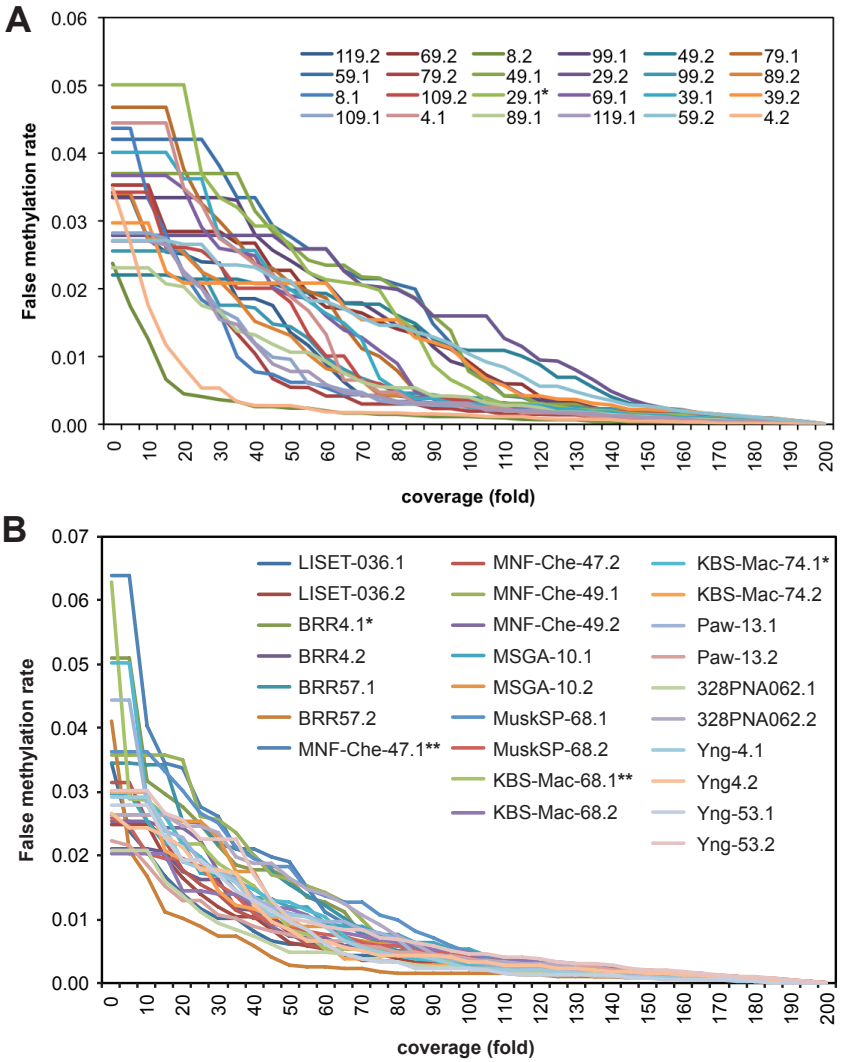

Figure S5: False methylation rates. False methylation rate (FMR) estimates and coverage bins from reads mapping to the chloroplasts for individual libraries of the (A) mutation accumulation lines and (B) haplogroup-1 accessions. *, **: strain for which FMR is above 5% and 6% for the lowest coverage bin (up to 5x), respectively.

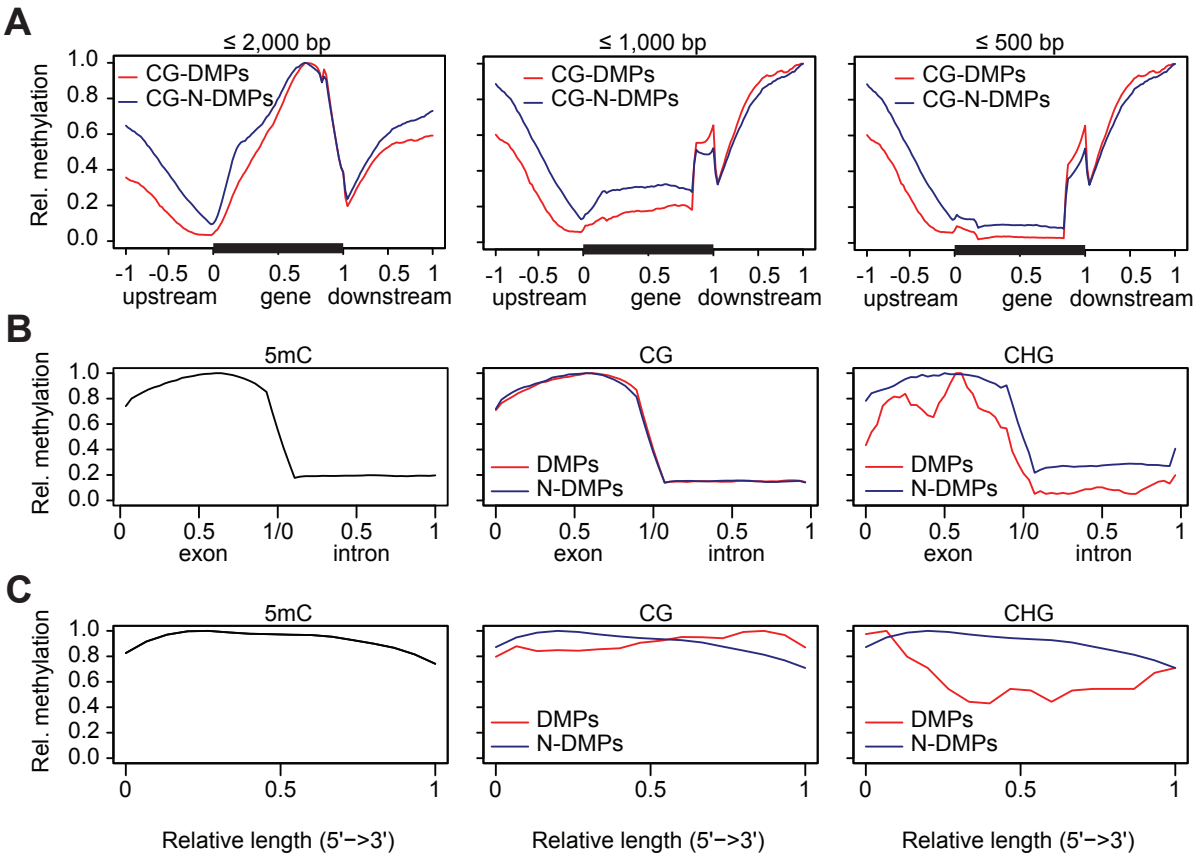

Figure S6: Frequency distribution of DMPs and N-DMPs. (A) Frequency along genes as shown in Figure **5.7**C, limited to genes of up to 2,000, 1,000 and 500 bp in length, respectively. (B) Frequency along exons and introns. (C) Frequency along transposable elements (TEs). Data were normalized to the highest value for each sequence context and class. CHH methylation is not shown due to the reduced statistical power in detecting differential methylation. 5mC: methylated position, DMP: differentially methylated position, N-DMP: invariantly methylated position.

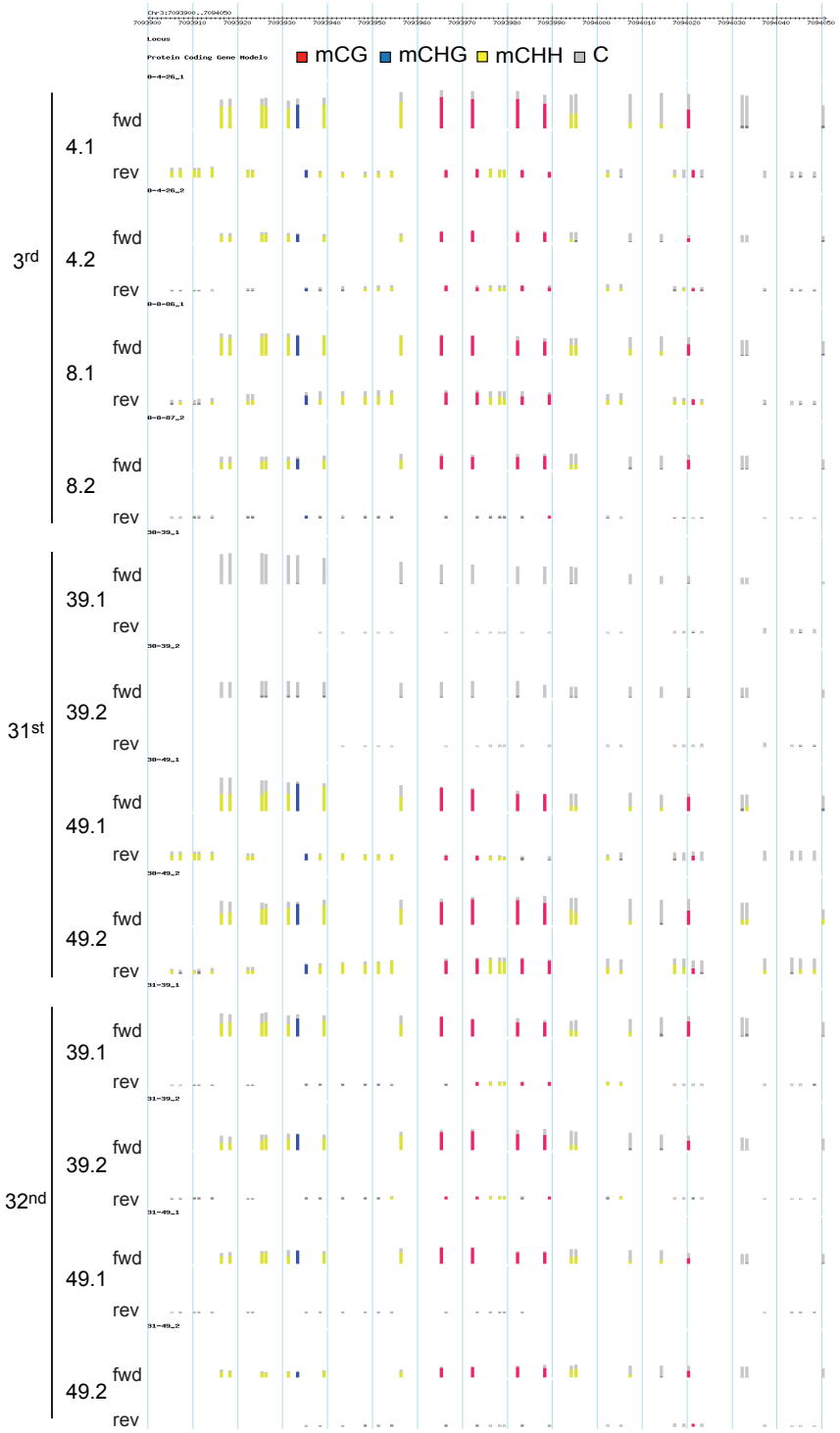

Figure S7: DMR identified between the 31st and 32nd generation of strain 39. Compared to the 3rd generation, this 150 bp region (Chr3: 7,093,900-7,094,050) showed a loss of methylation in the 31st generation, but a methylation pattern similar to the 3rd had been re-established in the 32nd generation. Line 49 showed no change. Methylation on both strands is indicated for each strain. Colors indicate methylated reads (red, CG; blue, CHG; yellow, CHH). Grey indicates reads supporting non-methylation.

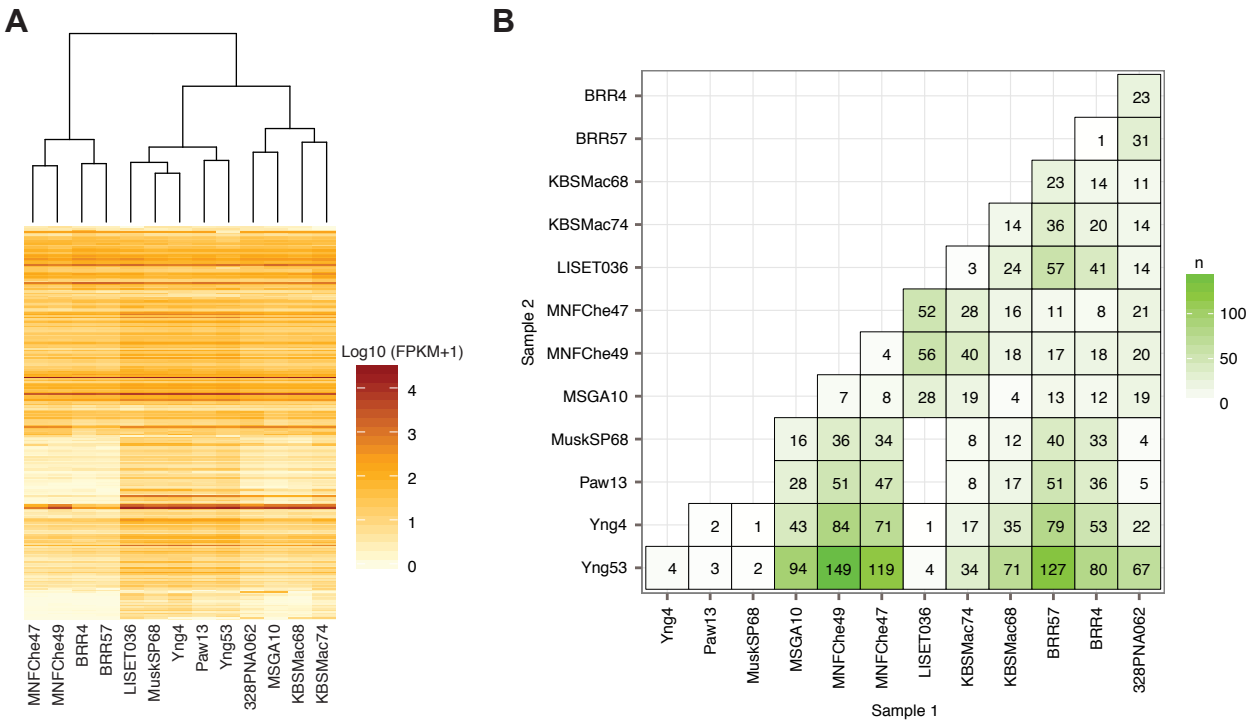

Figure S8: Differential gene expression pattern in the HPG1 accessions. (A) Hierarchical clustering of haplogroup-1 (HPG1) accessions by expression of differentially expressed genes. (B) Differentially expressed genes per pairwise comparison. FPKM: fragments per kilobase per million mapped reads.

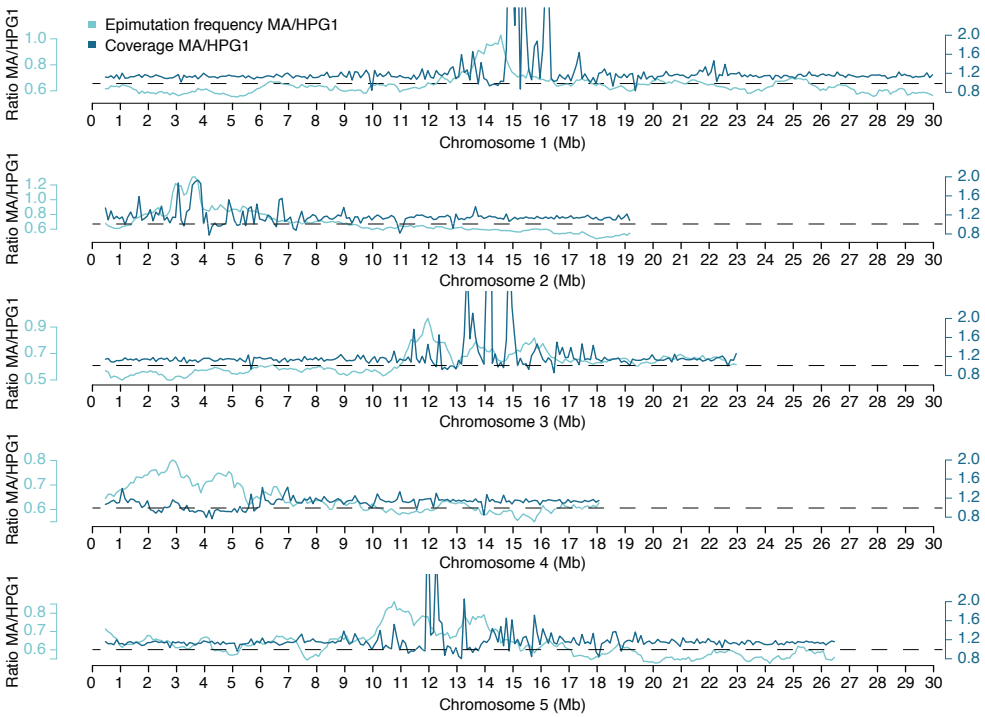

Figure S9: Local epimutation rate and read depths for the MA and HPG1 lineages. Ratios between epimutation frequencies and sequencing depth along the 5 chromosomes for mutation accumulation (MA) and haplogroup-1 (HPG1) lines. Epimutation frequencies were determined as the number of differentially methylated positions per cytosine with at least threefold coverage per window. Coverage is represented as average coverage per window across all accessions of each population. Dashed lines mark the balanced coverage ratio of 1. Sliding window; window size 100,000; step size 10,000 bp.

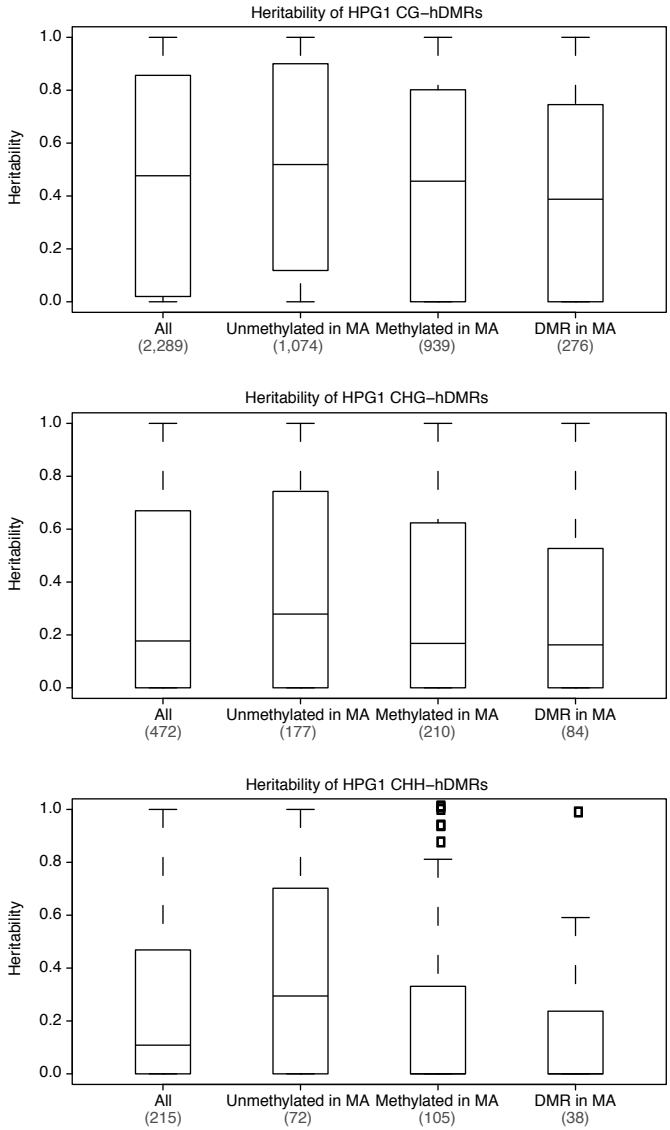

Figure S10: Heritability of HPG1-hDMRs by sequence context and status in MA lines. Distributions of heritability values of highly differentially methylated regions (hDMRs) of haplogroup-1 (HPG1) strains according to significant sequence context and methylation status of overlapping regions in the mutation accumulation (MA) lines.

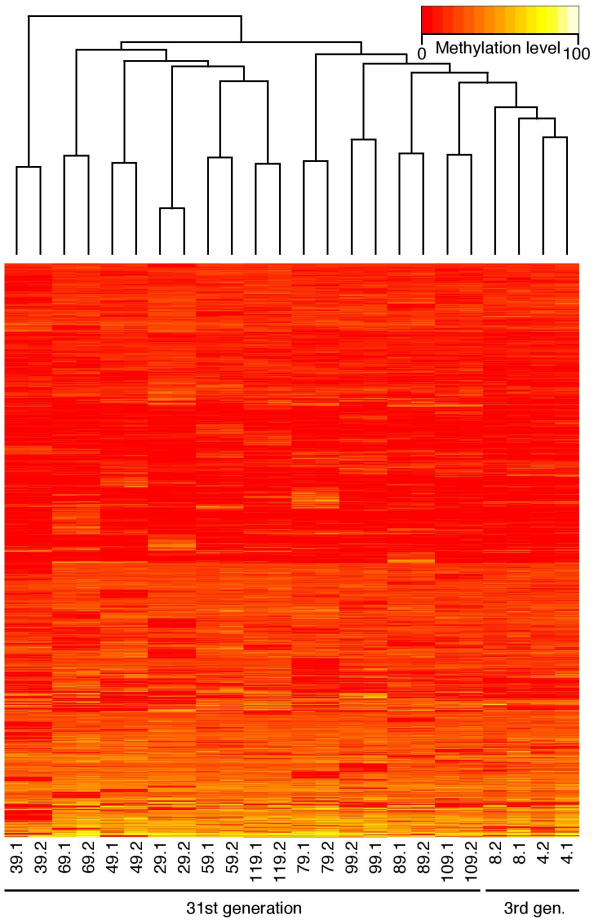

Figure S11: Hierarchical clustering of MA-hDMRs. Hierarchical clustering of mutation accumulation (MA) lines based on methylated sites in highly differentially methylated regions (hDMRs).

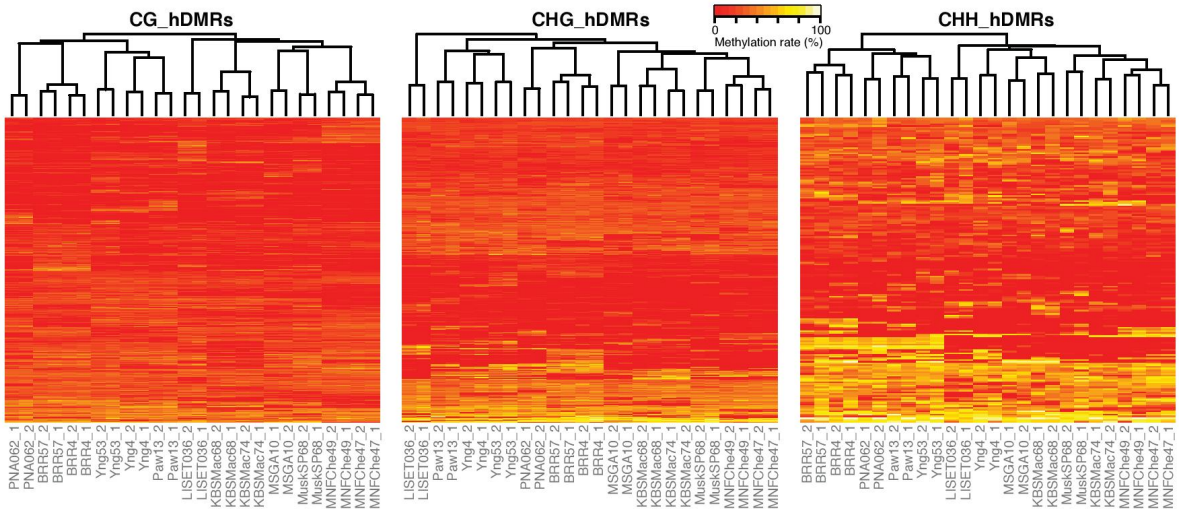

Figure S12: Hierarchical clustering of HPG1-hDMRs by sequence context. Hierarchical clustering of haplogroup-1 (HPG1) strains based on methylated sites in highly differentially methylated regions (hDMRs) by sequence context.

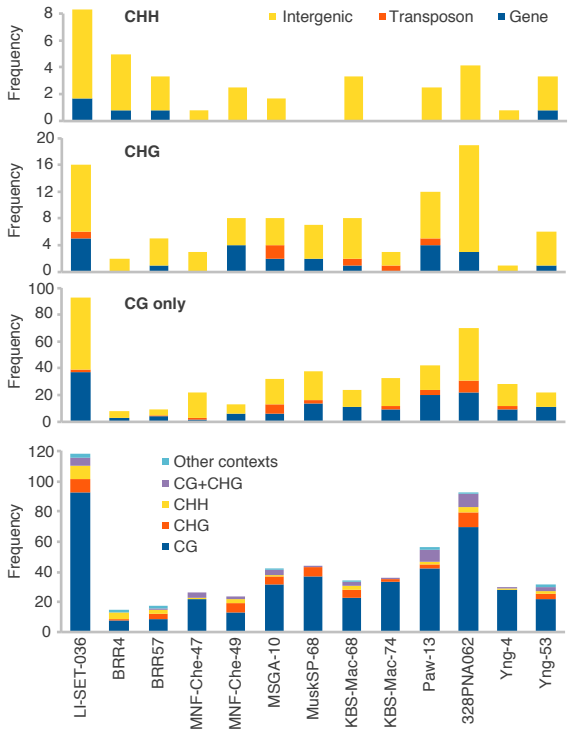

Figure S13: Analysis of hDMRs unique to LI-SET-036. Stacked bar plots showing the distributions of sequence contexts (bottom) and overlapping genomic features (top three plots) for hDMRs unique to each strain. 'CG only' exclusively considers CG-hDMRs whereas 'CHG' and 'CHH' might additionally include hDMRs of other contexts than CHG and CHH, respectively. The distribution across intergenic space, TEs and genes was similar for all strains.

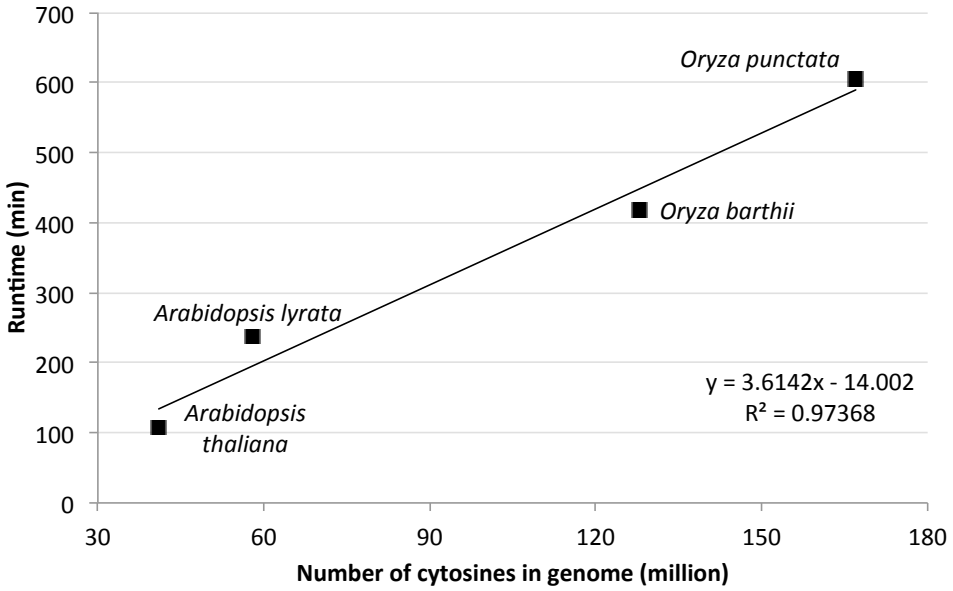

Figure S14: Runtime of HMM-based MR detection method for WGBS-Seq data of different species. Data for *A. thaliana* is represented by the mean runtime for all 13 haplogroup-1 samples, *A. lyrata* data is represented by the mean of four samples from ref. [Willing et al., 2015], and data for the two rice samples is based on one sample each (rice data retrieved from C. Becker, pers. communication). HMM: Hidden Markov Model, MR: methylated region, WGBS: whole-genome bisulphite sequencing.

Appendix B

Supplemental Tables

Table S1: Number of total variants and variants by type per mutation accumulation line, and the average and standard deviation (sd) of the 31st generation lines only.

Line	Total	SNP	Deletions	Insertions
4	7	6	1	0
8	7	5	0	2
29	31	23	7	1
39	33	26	2	5
49	39	26	10	3
59	28	18	7	3
69	30	23	3	4
79	46	31	8	7
89	26	21	3	2
99	31	25	3	3
109	35	27	5	3
119	37	35	1	1
Average (31st)	33.6	25.5	4.9	3.2
sd (31st)	5.9	4.9	3.0	1.8

Supplemental Tables

Table S2: Allele frequencies of detected variants in the mutation accumulation line population.

Allele frequency	Number of variants
1	350
2	9
3	1
4	1
5	3
6	3
7	3
8	5
9	0
10	5
11	8
Total	388

Table S3: Mutation accumulation lines and summary statistics on bisulphite sequencing.

		Strain	Sequencing depth (x)	Covered cytosine sites								Methylated cytosine sites			
				Coverage ≥ 1				Coverage ≥ 3				Coverage ≥ 3			
				all Cs	CG context	CHG context	CHH context	all Cs	CG context	CHG context	CHH context	all Cs	CG context	CHG context	CHH context
parental strains		4.1	47.3	39366222	5328029	5832538	28205655	33776987	4824357	5285998	23666632	3259484	1481845	751419	1026220
		4.2	30.6	30153541	4427411	4850924	20875206	19456542	3212894	3436288	12807360	2402085	1141166	599707	661212
		8.1	37.0	34817491	4909106	5376704	24531681	26337694	3969326	4371028	17997340	3041773	1321869	718866	1001038
		8.2	29.8	29017828	4298604	4709513	20009711	18403727	3059399	3269184	12075144	2316966	1097242	580151	639573
Descendants	Generation 30	109.1	38.4	36138189	5048852	5525008	25564329	27983214	4201628	4612498	19169088	2941472	1362284	697169	882019
		109.2	28.1	33545946	4824591	5257940	23463415	22165716	3611669	3862992	14691055	2410130	1207294	603006	599830
		119.1	36.1	34615649	4902016	5363998	24349635	25748084	3931379	4319140	17497565	2832361	1318003	693906	820452
		119.2	37.0	35132246	4976527	5431685	24724034	24657861	3935510	4223102	16499249	2849720	1319787	705476	824457
		29.1	48.0	38754671	5280828	5775597	27698246	32194819	4672128	5111847	22410844	2948321	1448774	741549	757998
		29.2	88.1 (*)	40359397	5406567	5916732	29036098	32611618	4743777	5120367	22747474	3458949	1509877	826756	1122316
		39.1	40.9	38032496	5223039	5713467	27095990	30847723	4535621	4964816	21347286	2902302	1410244	697775	794283
		39.2	24.4	36695554	5131341	5597090	25967123	25650981	4070908	4296902	17283171	2537779	1292286	599909	645584
		49.1	39.8	38419322	5255925	5743816	27419581	31452278	4606822	5028061	21817395	2817562	1434097	683328	700137
		49.2	33.3	37387546	5193665	5668723	26525158	27475296	4272300	4539317	18663679	2845504	1355229	661703	828572
		59.1	41.1	38829611	5289183	5781751	27758677	32282483	4688428	5120095	22473963	2748718	1441628	680904	626187
		59.2	35.5	35830330	5059592	5515767	25254971	25241024	4018558	4305193	16917273	2826409	1334113	702776	789520
		69.1	43.0	38443239	5252545	5748221	27442473	31547926	4603210	5041587	21903129	2900991	1412339	715824	772828
		69.2	46.5	36203457	5092089	5553148	25558220	26597480	4197471	4490328	17909681	3312380	1373651	764044	1174685
		79.1	40.9	37714492	5192863	5675528	26846101	30180157	4466420	4878916	20834821	2737172	1393844	668634	674694
		79.2	45.1	35790474	5026775	5491049	25272650	25958867	4093942	4413950	17450975	3400512	1370417	772137	1257958
		89.1	39.4	37121557	5141295	5628481	26351781	28951064	4286432	4726728	19937904	2882958	1349460	687791	845707
		89.2	28.4	32745585	4741870	5167715	22836000	21426202	3506739	3756027	14163436	2449304	1183989	607314	658001
		99.1	82.2 (*)	38758412	5282866	5779410	27696136	30272071	4525357	4897509	20849205	3346119	1463027	785578	1097514
		99.2	27.3	31739118	4636478	5048562	22054078	21354083	3423671	3719205	14211207	2521309	1179143	618334	723832
	Generation 31	39.1	34.6	34203426	4865951	5323242	24014233	23511514	3740485	4014488	15756541	2629829	1271740	640197	717892
		39.2	36.9	28994485	4286743	4695658	20012084	18551302	3058842	3288522	12203938	2570469	1124295	621160	825014
		49.1	33.6	28409980	4209592	4619082	19581306	17963795	2979502	3205013	11779280	2547351	1113837	624771	808743
		49.2	26.2	27121765	4068545	4453041	18600179	16655070	2822497	3001545	10831028	2277288	1050920	571320	655048
Mean			40.0	35155072	4941175	5401585	24812313	26044842	4002117	4332166	17710559	2811258	1312943	679339	818976

(*) These strains were sequenced as technical replicates on two independent flow cells, read data was then combined.

Table S4: Haplogroup-1 accessions and summary statistics on genome sequencing.

								Mapping to Col-0 reference			Mapping to HPG1 reference	
Accession ID	Accession name	Latitude	Longitude	State	Sampling date	Closest weather station	Distance to weather station (km)	Reads mapped to Col-0	Average coverage (x)	Positions Covered > 5x	Reads mapped to HPG-1 reference	Positions covered > 5x
8699	**328PNA-062**	42.0945	-86.3253	Michigan	Mar 28, 2006	726355-99999	10	71,064,284	48.0	107,019,787	72,348,378	108,702,343
470	**BRR-4**	40.8313	-87.7350	Illinois	May 1, 2006	724386-14835	83	69,155,280	45.2	106,981,548	70,401,862	108,680,008
504	**BRR-57**	40.8313	-87.7350	Illinois	May 1, 2006	724386-14835	83	41,119,975	28.9	106,831,177	41,814,926	108,627,670
1739	**KBS-Mac-68**	42.4050	-85.3980	Michigan	May 29, 2004	725396-14814	16	65,629,252	42.6	106,955,923	66,742,616	108,677,940
1741	**KBS-Mac-74**	42.4050	-85.3980	Michigan	May 29, 2004	725396-14814	16	61,925,749	39.0	106,926,171	62,923,233	108,665,771
742	**LI-SET-036**	40.9352	-73.1140	New York	April/May 2005	725035-04781	16	78,397,323	42.0	107,083,267	62,671,495	108,712,472
1942	**MNF-Che-47**	43.5251	-86.1843	Michigan	Jun 5, 2004	726360-14840	40	42,262,867	28.2	106,848,829	42,925,314	108,634,616
1943	**MNF-Che-49**	43.5251	-86.1843	Michigan	Jun 5, 2004	726360-14840	40	43,052,687	28.8	106,844,153	43,729,285	108,622,042
2081	**MuskSP-68**	43.2483	-86.3368	Michigan	Jun 4, 2004	726360-14840	12	63,558,211	42.7	106,989,784	42,298,899	108,688,853
2106	**MSGA-10**	43.2749	-86.0891	Michigan	Jun 4, 2004	726360-14840	17	41,684,061	26.0	106,809,482	64,718,515	108,608,920
2159	**Paw-13**	42.1480	-86.4310	Michigan	May 1, 2002	726355-99999	2	68,356,951	43.9	106,948,820	69,661,005	108,660,130
2370	**Yng-4**	41.8650	-86.6460	Michigan	May 1, 2002	725350-14848	32	69,221,983	41.7	106,965,670	70,384,702	108,679,280
2412	**Yng-53**	41.8650	-86.6460	Michigan	May 1, 2002	725350-14848	32	67,134,738	46.5	107,026,441	68,362,136	108,707,709

Table S5: Summary statistics on bisulphite sequencing of haplogroup-1 accessions.

Accession_ID	8699		470		504		1739		1741		742		1942		1943		2081	
Accession name	**328PNA-062**		**BRR-4**		**BRR-57**		**KBS-Mac-68**		**KBS-Mac-74**		**LI-SET-036**		**MNF-Che-47**		**MNF-Che-49**		**MuskSP-68**	
Replicate	1	2	1	2	1	2	1	2	1	2	1	2	1	2	1	2	1	2
Reads mapped to Col-0	61353974	58031068	55850977	65607151	54370962	43002155	36075984	55813465	54242907	63594061	53841607	56543883	60323516	61718557	52466876	65465682	50942064	65208383
Cytosines covered >= 3x	25860763	25380773	23921190	24974679	22557503	23172280	20333550	26660804	23093771	24329081	20118990	24700915	24093751	22276202	21413957	23205896	23102627	22626563
CG	3910937	3838859	3688663	3821587	3530880	3601637	3239682	4016321	3580308	3735945	3209230	3782972	3704936	3471235	3365057	3600995	3586751	3531143
CHG	4312700	4231781	4055419	4203308	3868646	3964333	3549399	4422816	3938250	4109775	3509147	4165707	4074707	3819524	3697425	3967203	3940594	3885016
CHH	17637126	17310133	16177108	16949784	15157977	15606310	13544469	18221667	15575213	16483361	13400613	16752236	16314108	14985443	14351475	15637698	15575282	15210404
Methylated cytosines	2387453	2335780	2197649	2568081	2149593	2282978	1919981	2309280	2151783	2401798	2654125	2372157	2177624	2358588	2006950	2321363	2073345	2246692
5mCG	1126346	1105811	1085198	1132913	1047072	1060006	951162	1138315	1048500	1105049	1035510	1114210	1085255	1072368	1005972	1063781	1051075	1050141
5mCHG	501930	492460	479168	533771	463997	482752	410238	484768	462056	509408	534057	501289	466067	506584	437184	493150	445481	484734
5mCHH	759177	737509	633283	901397	638524	740220	558581	686197	641227	787341	1084558	756658	626302	779636	563794	764432	576789	711817
Average coverage per strand (x)	19.7	18.2	15.7	20.4	15.8	13.2	11.0	17.0	16.4	20.2	18.6	16.9	17.8	19.2	15.6	21.8	14.7	21.8
Average FMR (%)*	0.23	0.34	0.30	0.26	0.38	0.22	0.39	0.25	0.33	0.24	0.30	0.20	0.41	0.24	0.48	0.41	0.40	0.45
Reads mapped to HPG1 reference	63017534	59436141	57073381	67247854	55531676	44145065	36906120	57386666	55489041	65098462	55012858	57945498	61692550	63067650	53595243	67086174	52089962	66801014
Cytosines covered >= 3x	27086015	26541252	25065761	26141729	23657531	24363548	21366459	27970328	24185039	25429150	21012690	25886645	25230221	23284813	22456874	24285937	24254852	23715088
CG	4081403	4005809	3852797	3986232	3691583	3773550	3394323	4197898	3738442	3892491	3342462	3951679	3867671	3618315	3518750	3758249	3753733	3689872
CHG	4480578	4395463	4217036	4365241	4026497	4134270	3702356	4600436	4093935	4263943	3641040	4331475	4234786	3964718	3849026	4121104	4104865	4041364
CHH	18524034	18139980	16995928	17790256	15939451	16455728	14269780	19171994	16352662	17272716	14029188	17603491	17127764	15701780	15089098	16406584	16396254	15983852
Methylated cytosines	2735919	2692113	2440066	2874140	2450913	2480393	2120310	2773910	2363257	2647181	2927447	2670797	2386925	2593870	2349141	2555996	2357787	2456184
5mCG	1221736	1195997	1155941	1215158	1134082	1132337	1018744	1250018	1115417	1177132	1102391	1197152	1155695	1141492	1102211	1140806	1137823	1122996
5mCHG	576485	569316	537622	600011	533042	532935	461767	579795	514408	566992	589951	568283	519938	561836	517005	549352	513455	536223
5mCHH	937698	926800	746503	1058971	783789	815121	639799	944097	733432	903057	1235105	905362	711292	890542	729925	865838	706509	796965

Accession_ID	2106		2159		2370		2412	
Accession name	**MSGA-10**		**Paw-13**		**Yng-4**		**Yng-53**	
Replicate	1	2	1	2	1	2	1	2
Reads mapped to Col-0	56083265	63634201	56369984	62647205	49230291	58518358	53933783	53618907
Cytosines covered >= 3x	20918212	23783022	22091966	23736569	20881243	21701501	24723011	23376345
CG	3292348	3685740	3431447	3655132	3294535	3391584	3779431	3620410
CHG	3620498	4052611	3782416	4049169	3618017	3758677	4157358	3995744
CHH	14005366	16044671	14878103	16032268	13968691	14551240	16786222	15760191
Methylated cytosines	2112174	2215814	2181790	2187930	2157088	2176477	2412495	2127348
5mCG	1006578	1083469	1057368	1042997	1007231	1001176	1122369	1054027
5mCHG	458066	472115	475943	469168	459806	471857	496530	452960
5mCHH	647530	660230	648479	675765	690051	703444	793596	620361
Average coverage per strand (x)	16.7	21.1	18.1	20.9	15.3	19.9	16.4	17.6
Average FMR (%)*	0.30	0.48	0.31	0.33	0.27	0.41	0.20	0.52
Reads mapped to HPG1 reference	57230842	65218510	57653604	64256352	50283004	59947599	55222846	55023087
Cytosines covered >= 3x	21886178	24910553	23116057	24856817	21855132	22736374	25877968	24535689
CG	3438423	3849680	3583888	3817105	3441946	3543924	3946964	3788510
CHG	3765318	4212585	3933759	4208645	3763742	3908748	4322296	4161489
CHH	14682437	16848288	15598410	16831067	14649444	15283702	17608708	16585690
Methylated cytosines	2372304	2452462	2353205	2474943	2459116	2388260	2747328	2351187
5mCG	1076961	1163187	1106714	1138250	1095887	1073249	1216900	1135201
5mCHG	518413	528506	519845	535121	528750	523408	569950	507403
5mCHH	776930	760769	726646	801572	834479	791603	960478	708583

* FMR = False Methylation Rate

Table S6: Scoring matrices for the assessment of the alignment quality at single sites. Matrices indicating the penalties applied during the assessment of the quality of the read and alignment data for genomic libraries (top) and bisulphite libraries (bottom) using the SHORE pipeline.

Features of short read alignments	Penalties		
	High	Intermediate	Low
Read support	<2	<3	<5
Read core support	<1	<2	<3
Read inner core support	<1	<1	<2
Concordance	<0.7	<0.8	<0.9
Concordance (core region)	<0.7	<0.8	<0.9
Max. quality of major allele base call	<10	<20	<30
Max. base quality of 3rd and 4th base	*NA*	>35	>30
Noise (frequency of 3rd and 4th base)	>0.3	>0.2	>0.1
Average number mismatches (relative to read length)	>10%	>7%	>5%
Support (left/right/forward/reverse)	*NA*	<1	<2
Coverage ratio (left vs. right half of read)	>1:99	>1:7	>1:5
Coverage ratio (forward vs. reverse strand)	>1:99	>1:7	>1:5
Average number of mappings per read (repetitiveness)	>1.8	>1.5	>1.2
Max. observed coverage (relative to average)	>10x	>7x	>5x
Max. expected to observed coverage ratio	>1:8	>1:5	>1:3
GC content	<5	<10	<15
Sequence complexity	*NA*	*NA*	<3
Zero-coverage region within x bp distance	<7	<11	<13

Features of short read alignments	Penalties		
	High	Intermediate	Low
Read support	<3	<4	<5
Max. non-converted base quality	<20	<30	<35
Difference between max. converted base quality and max. non-converted base quality	>20	>10	>5
Max. base quality of 3rd and 4th base	*NA*	>35	>30
Noise (frequency of 3rd and 4th base)	>0.3	>0.15	>0.05
Average number mismatches (relative to read length)	>10%	>7%	>5%
Coverage ratio (left vs. right half of read)	>1:30	>1:10	>1:5
Coverage ratio (forward vs. reverse strand)	>1:30	>1:10	>1:5
Average number of mappings per read (repetitiveness)	>2	>1.5	>1.1
Max. observed coverage (relative to average)	>10x	>7x	>5x
Max. expected to observed coverage ratio	>1:8	>1:5	>1:3
GC content	<5	<10	<15
Sequence complexity	*NA*	*NA*	<3
Zero-coverage region within x bp distance	<6	<9	<11

Appendix C

Command lines

Note:

- Most input file arguments and specific runtime arguments (e.g. number of threads) are omitted.
- File names and comments to be replaced are in squared brackets.

Genetic variation pipeline

I executed following commands for each strain and iteration:

- SHORE

```
shore mapflowcell -n 10% -g 7% -l 14 -s 1000
shore correct4pe -x 300
shore consensus -g 10 -h 10 -b .1 -v -a scoring_matrix_HPT.txt -i
    0.5
```

- bwa

```
bwa-0.6.2/bwa aln -n 10 -o 1 -e 7 -i 5 -l 25 -R 0 -f map.list_1.sai
     [for read2 respectively in map.list_2.sai]
bwa-0.6.2/bwa sampe -P -s -f map.list.bwa.sam [reference-file] map.
    list_1.sai map.list_2.sai [read-files]
samtools view -b -S -o map.list.bwa.bam map.list.bwa.sam
samtools sort map.list.bwa.bam map.list.bwa.sorted
samtools index map.list.bwa.sorted.bam
```

- DELLY

```
delly_v0.0.9/delly -o del_p_e30_s3_q20_m20.txt -p -e 0.3 -s 3 -q 20
     -m 20 -i $strain -b splitreads_p_e30_s3_q20_m20.txt map.list.
    bwa.sorted.bam
```

- Pindel

```
pindel024t/pindel --chromosome ALL --min_inversion_size 20
```

- SV-M
  Configuration file:

```
logging = 2
cpu = 1
genome_directory = [...]
chromosomes =
    1[13700000-15900000],2[2450000-5500000],3[11300000-14300000],
4[1800000-5150000],5[11000000-13350000]
mapped_reads = [...]
data_dir = [...]
unmapped_reads = left_over_all.fl
mu_read_output = [...]
mu_read_src = mu_reads
ref_sequence = [...]
output_dir = [...]
deletion_html = deletion
deletion_file = deletions
deletion_dublicates = deletion_dublicates.fl
insertion_html = insertion
insertion_file = insertions
insertion_dublicates = insertion_dublicates.fl
snp_data = [...]
svm_model_name = findel.model
svm_normalization_parameters = findel.norm
svm_model_name_insertions = findel_ins.model
svm_normalization_parameters_insertions = findel_ins.norm
alignment_display = 60
alignment_match = 5
alignment_mismatch = -4
alignment_gap_opening = -10
alignment_gap_extension = 0
alignment_min_length = 300
alignment_max_length = 5000
alignment_min_score = 34
alignment_prepost = 20
alignment_max_gap_opening = 2
alignment_allowed_mismatches = 10
traceback_threshold = 5
alignment_support = 2
```

- velvet

```
velveth [folder] 35 -shortPaired -fastq -separate assembly_reads_1.fq
    assembly_reads_2.fq -short2 -fastq assembly_reads_single.fq
velvetg [folder] -ins_length 350 -min_contig_lgth 200 -scaffolding no -
    read_trkg yes -unused_reads yes -max_coverage 100
```

- SOAPdenovo2

```
SOAPdenovo-63mer all -s SOAPdenovo_config.txt -K 35
```

Configuration file:

```
SOAPdenovo_config.txt:
#maximal read length
max_rd_len=101
[LIB]
#average insert size
avg_ins=300
#if sequence needs to be reversed
reverse_seq=0
#in which part(s) the reads are used
asm_flags=1
#in which order the reads are used while scaffolding
rank=1
#use only first 100 bps of each read
rd_len_cutoff=101
# cutoff of pair number for a reliable connection (at least 3 for short
     insert size)
pair_num_cutoff=3
#minimum aligned length to contigs for a reliable read location (at
    least 32 for short insert size)
map_len=32
#a pair of fastq file, read 1 file should always be followed by read 2
    file
q1=assembly_reads_1.fq
q2=assembly_reads_2.fq
#fastq file for single reads
q=assembly_reads_single.fq
```

Bisulphite analysis

SHORE:

```
shore mapflowcell -n 10% -g 0 -l 20 -s 200 -B
shore correct4pe -x 300
shore methyl -a scoring_matrix_meth.txt -m .05 -v -g 0
```

References

[1000 Genomes Project Consortium et al., 2010] 1000 Genomes Project Consortium, Abecasis, G. R., Altshuler, D., Auton, A., Brooks, L. D., Durbin, R. M., Gibbs, R. A., Hurles, M. E., and McVean, G. A. (2010). A map of human genome variation from population-scale sequencing. *Nature*, 467(7319):1061–1073.

[Abyzov et al., 2011] Abyzov, A., Urban, A. E., Snyder, M., and Gerstein, M. (2011). CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. *Genome Res*, 21(6):974–984.

[Akalin et al., 2012] Akalin, A., Kormaksson, M., Li, S., Garrett-Bakelman, F. E., Figueroa, M. E., Melnick, A., and Mason, C. E. (2012). methylKit: a comprehensive R package for the analysis of genome-wide DNA methylation profiles. *Genome Biol*, 13(10):R87.

[Akman et al., 2014] Akman, K., Haaf, T., Gravina, S., Vijg, J., and Tresch, A. (2014). Genome-wide quantitative analysis of DNA methylation from bisulfite sequencing data. *Bioinformatics*, 30(13):1933–1934.

[Alabert and Groth, 2012] Alabert, C. and Groth, A. (2012). Chromatin replication and epigenome maintenance. *Nat Rev Mol Cell Biol*, 13(3):153–167.

[Albers et al., 2011] Albers, C. A., Lunter, G., MacArthur, D. G., McVean, G., Ouwehand, W. H., and Durbin, R. (2011). Dindel: accurate indel calls from short-read data. *Genome Res*, 21(6):961–973.

[Alkan et al., 2011] Alkan, C., Coe, B. P., and Eichler, E. E. (2011). Genome structural variation discovery and genotyping. *Nat Rev Genet*, 12(5):363–376.

[Altham, 1969] Altham, P. M. E. (1969). Exact Bayesian Analysis of a 2 x 2 Contingency Table, and Fisher's "Exact" Significance Test. *Journal of the Royal Statistical Society. Series B (Methodological)*, 31, No. 2:261–269.

[Altschul et al., 1990] Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *J Mol Biol*, 215(3):403–410.

[Andrews, 2012] Andrews, S. (2012). FASTQC: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

[Arabidopsis Genome Initiative, 2000] Arabidopsis Genome Initiative, T. (2000). Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. *Nature*, 408(6814):796–815.

[Arumugam et al., 2011] Arumugam, M., Raes, J., Pelletier, E., Le Paslier, D., Yamada, T., Mende, D. R., Fernandes, G. R., Tap, J., Bruls, T., Batto, J.-M., Bertalan, M., Borruel, N., Casellas, F., Fernandez, L., Gautier, L., Hansen, T., Hattori, M., Hayashi, T., Kleerebezem, M., Kurokawa, K., Leclerc, M., Levenez, F., Manichanh, C., Nielsen, H. B., Nielsen, T., Pons, N., Poulain, J., Qin, J., Sicheritz-Ponten, T., Tims, S., Torrents, D., Ugarte, E., Zoetendal, E. G., Wang, J., Guarner, F., Pedersen, O., de Vos, W. M., Brunak, S., Doré, J., , M. I. T. C., Antolín, M., Artiguenave, F., Blottiere, H. M., Almeida, M., Brechot, C., Cara, C., Chervaux, C., Cultrone, A., Delorme, C., Denariaz, G., Dervyn, R., Foerstner, K. U., Friss, C., van de Guchte, M., Guedon, E., Haimet, F., Huber, W., van Hylckama-Vlieg, J., Jamet, A., Juste, C., Kaci, G., Knol, J., Lakhdari, O., Layec, S., Le Roux, K., Maguin, E., Mérieux, A., Melo Minardi, R., M'rini, C., Muller, J., Oozeer, R., Parkhill, J., Renault, P., Rescigno, M., Sanchez, N., Sunagawa, S., Torrejon, A., Turner, K., Vandemeulebrouck, G., Varela, E., Winogradsky, Y., Zeller, G., Weissenbach, J., Ehrlich, S. D.,

and Bork, P. (2011). Enterotypes of the human gut microbiome. *Nature*, 473(7346):174–180.

[Atwell et al., 2010] Atwell, S., Huang, Y. S., Vilhjálmsson, B. J., Willems, G., Horton, M., Li, Y., Meng, D., Platt, A., Tarone, A. M., Hu, T. T., Jiang, R., Muliyati, N. W., Zhang, X., Amer, M. A., Baxter, I., Brachi, B., Chory, J., Dean, C., Debieu, M., de Meaux, J., Ecker, J. R., Faure, N., Kniskern, J. M., Jones, J. D. G., Michael, T., Nemri, A., Roux, F., Salt, D. E., Tang, C., Todesco, M., Traw, M. B., Weigel, D., Marjoram, P., Borevitz, J. O., Bergelson, J., and Nordborg, M. (2010). Genome-wide association study of 107 phenotypes in Arabidopsis thaliana inbred lines. *Nature*, 465(7298):627–631.

[Ausin et al., 2012] Ausin, I., Greenberg, M. V. C., Simanshu, D. K., Hale, C. J., Vashisht, A. A., Simon, S. A., Lee, T.-f., Feng, S., Española, S. D., Meyers, B. C., Wohlschlegel, J. A., Patel, D. J., and Jacobsen, S. E. (2012). INVOLVED IN DE NOVO 2-containing complex involved in RNA-directed DNA methylation in Arabidopsis. *Proc Natl Acad Sci U S A*, 109(22):8374–8381.

[Barrow and Michels, 2014] Barrow, T. M. and Michels, K. B. (2014). Epigenetic epidemiology of cancer. *Biochem Biophys Res Commun*.

[Baulcombe and Dean, 2014] Baulcombe, D. C. and Dean, C. (2014). Epigenetic regulation in plant responses to the environment. *Cold Spring Harb Perspect Biol*, 6(9):a019471.

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

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

[Bender and Fink, 1995] Bender, J. and Fink, G. R. (1995). Epigenetic control of an endogenous gene family is revealed by a novel blue fluorescent mutant of Arabidopsis. *Cell*, 83(5):725–734.

[Berger et al., 2009] Berger, S. L., Kouzarides, T., Shiekhattar, R., and Shilatifard, A. (2009). An operational definition of epigenetics. *Genes Dev*, 23(7):781–783.

[Bird, 2007] Bird, A. (2007). Perceptions of epigenetics. *Nature*, 447(7143):396–398.

[Bird, 1978] Bird, A. P. (1978). Use of restriction enzymes to study eukaryotic DNA methylation: II. The symmetry of methylated sites supports semi-conservative copying of the methylation pattern. *J Mol Biol*, 118(1):49–60.

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

[Bond and Baulcombe, 2014] Bond, D. M. and Baulcombe, D. C. (2014). Small RNAs and heritable epigenetic variation in plants. *Trends Cell Biol*, 24(2):100–107.

[Bond and Baulcombe, 2015] Bond, D. M. and Baulcombe, D. C. (2015). Epigenetic transitions leading to heritable, RNA-mediated de novo silencing in Arabidopsis thaliana. *Proc Natl Acad Sci U S A*.

[Bonduriansky, 2012] Bonduriansky, R. (2012). Rethinking heredity, again. *Trends Ecol Evol*, 27(6):330–336.

[Borgel et al., 2010] Borgel, J., Guibert, S., Li, Y., Chiba, H., Schübeler, D., Sasaki, H., Forné, T., and Weber, M. (2010). Targets and dynamics of promoter DNA methylation during early mouse development. *Nat Genet*, 42(12):1093–1100.

[Bos et al., 2011] Bos, K. I., Schuenemann, V. J., Golding, G. B., Burbano, H. A., Waglechner, N., Coombes, B. K., McPhee, J. B., DeWitte, S. N., Meyer, M., Schmedes, S., Wood, J., Earn, D. J. D., Herring, D. A., Bauer, P., Poinar, H. N., and Krause, J. (2011). A draft genome of Yersinia pestis from victims of the Black Death. *Nature*, 478(7370):506–510.

[Boyko et al., 2007] Boyko, A., Kathiria, P., Zemp, F. J., Yao, Y., Pogribny, I., and Kovalchuk, I. (2007). Transgenerational changes in the genome stability and methylation in pathogen-infected plants: (virus-induced plant genome instability). *Nucleic Acids Res*, 35(5):1714–1725.

[Boyko and Kovalchuk, 2010] Boyko, A. and Kovalchuk, I. (2010). Transgenerational response to stress in Arabidopsis thaliana. *Plant Signal Behav*, 5(8):995–998.

[Boyko and Kovalchuk, 2011] Boyko, A. and Kovalchuk, I. (2011). Genome instability and epigenetic modification–heritable responses to environmental stress? *Curr Opin Plant Biol*, 14(3):260–266.

[Branton et al., 2008] Branton, D., Deamer, D. W., Marziali, A., Bayley, H., Benner, S. A., Butler, T., Di Ventra, M., Garaj, S., Hibbs, A., Huang, X., Jovanovich, S. B., Krstic, P. S., Lindsay, S., Ling, X. S., Mastrangelo, C. H., Meller, A., Oliver, J. S., Pershin, Y. V., Ramsey, J. M., Riehn, R., Soni, G. V., Tabard-Cossa, V., Wanunu, M., Wiggin, M., and Schloss, J. A. (2008). The potential and challenges of nanopore sequencing. *Nat Biotechnol*, 26(10):1146–1153.

[Briggs and King, 1952] Briggs, R. and King, T. J. (1952). Transplantation of Living Nuclei From Blastula Cells into Enucleated Frogs' Eggs. *Proc Natl Acad Sci U S A*, 38(5):455–463.

[Brookes and Shi, 2014] Brookes, E. and Shi, Y. (2014). Diverse Epigenetic Mechanisms of Human Disease. *Annu Rev Genet*.

[Buermans and den Dunnen, 2014] Buermans, H. P. J. and den Dunnen, J. T. (2014). Next generation sequencing technology: Advances and applications. *Biochim Biophys Acta*, 1842(10):1932–1941.

[Burggren and Crews, 2014] Burggren, W. W. and Crews, D. (2014). Epigenetics in comparative biology: why we should pay attention. *Integr Comp Biol*, 54(1):7–20.

[Burton et al., 2013] Burton, J. N., Adey, A., Patwardhan, R. P., Qiu, R., Kitzman, J. O., and Shendure, J. (2013). Chromosome-scale scaffolding of de novo genome assemblies based on chromatin interactions. *Nat Biotechnol*, 31(12):1119–1125.

[Calarco et al., 2012] Calarco, J. P., Borges, F., Donoghue, M. T. A., Van Ex, F., Jullien, P. E., Lopes, T., Gardner, R., Berger, F., Feijó, J. A., Becker, J. D., and Martienssen, R. A. (2012). Reprogramming of DNA methylation in pollen guides epigenetic inheritance via small RNA. *Cell*, 151(1):194–205.

[Campbell et al., 2008] Campbell, P. J., Stephens, P. J., Pleasance, E. D., O'Meara, S., Li, H., Santarius, T., Stebbings, L. A., Leroy, C., Edkins, S., Hardy, C., Teague, J. W., Menzies, A., Goodhead, I., Turner, D. J., Clee, C. M., Quail, M. A., Cox, A., Brown, C., Durbin, R., Hurles, M. E., Edwards, P. A. W., Bignell, G. R., Stratton, M. R., and Futreal, P. A. (2008). Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing. *Nat Genet*, 40(6):722–729.

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

[Cao et al., 2011] Cao, J., Schneeberger, K., Ossowski, S., Günther, T., Bender, S., Fitz, J., Koenig, D., Lanz, C., Stegle, O., Lippert, C., Wang, X., Ott, F., Müller, J., Alonso-Blanco, C., Borgwardt, K., Schmid, K. J., and Weigel, D. (2011). Whole-genome sequencing of multiple Arabidopsis thaliana populations. *Nat Genet*, 43(10):956–963.

[Castel and Martienssen, 2013] Castel, S. E. and Martienssen, R. A. (2013). RNA interference in the nucleus: roles for small RNAs in transcription, epigenetics and beyond. *Nat Rev Genet*, 14(2):100–112.

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

[Chandler and Stam, 2004] Chandler, V. L. and Stam, M. (2004). Chromatin conversations: mechanisms and implications of paramutation. *Nat Rev Genet*, 5(7):532–544.

[Chien et al., 2004] Chien, P., Weissman, J. S., and DePace, A. H. (2004). Emerging principles of conformation-based prion inheritance. *Annu Rev Biochem*, 73:617–656.

[Chodavarapu et al., 2010] Chodavarapu, R. K., Feng, S., Bernatavichute, Y. V., Chen, P.-Y., Stroud, H., Yu, Y., Hetzel, J. A., Kuo, F., Kim, J., Cokus, S. J., Casero, D., Bernal, M., Huijser, P., Clark, A. T., Krämer, U., Merchant, S. S., Zhang, X., Jacobsen, S. E., and Pellegrini, M. (2010). Relationship between nucleosome positioning and DNA methylation. *Nature*, 466(7304):388–392.

[Chodavarapu et al., 2012] Chodavarapu, R. K., Feng, S., Ding, B., Simon, S. A., Lopez, D., Jia, Y., Wang, G.-L., Meyers, B. C., Jacobsen, S. E., and Pellegrini, M. (2012). Transcriptome and methylome interactions in rice hybrids. *Proc Natl Acad Sci U S A*, 109(30):12040–12045.

[Clark et al., 2006] Clark, S. J., Statham, A., Stirzaker, C., Molloy, P. L., and Frommer, M. (2006). DNA methylation: bisulphite modification and analysis. *Nat Protoc*, 1(5):2353–2364.

[Clark et al., 2013] Clark, T. A., Lu, X., Luong, K., Dai, Q., Boitano, M., Turner, S. W., He, C., and Korlach, J. (2013). Enhanced 5-methylcytosine detection in single-molecule, real-time sequencing

via Tet1 oxidation. *BMC Biol*, 11:4.

[Cocciolone et al., 2001] Cocciolone, S. M., Chopra, S., Flint-Garcia, S. A., McMullen, M. D., and Peterson, T. (2001). Tissue-specific patterns of a maize Myb transcription factor are epigenetically regulated. *Plant J*, 27(5):467–478.

[Cocciolone and Cone, 1993] Cocciolone, S. M. and Cone, K. C. (1993). Pl-Bh, an anthocyanin regulatory gene of maize that leads to variegated pigmentation. *Genetics*, 135(2):575–588.

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

[Coleman-Derr and Zilberman, 2012] Coleman-Derr, D. and Zilberman, D. (2012). Deposition of histone variant H2a.Z within gene bodies regulates responsive genes. *PLoS Genet*, 8(10):e1002988.

[Cortijo et al., 2014] Cortijo, S., Wardenaar, R., Colomé-Tatché, M., Gilly, A., Etcheverry, M., Labadie, K., Caillieux, E., Hospital, F., Aury, J.-M., Wincker, P., Roudier, F., Jansen, R. C., Colot, V., and Johannes, F. (2014). Mapping the epigenetic basis of complex traits. *Science*, 343(6175):1145–1148.

[Creasey et al., 2014] Creasey, K. M., Zhai, J., Borges, F., Van Ex, F., Regulski, M., Meyers, B. C., and Martienssen, R. A. (2014). miRNAs trigger widespread epigenetically activated siRNAs from transposons in Arabidopsis. *Nature*, 508(7496):411–415.

[Cruz and Houseley, 2014] Cruz, C. and Houseley, J. (2014). Endogenous RNA interference is driven by copy number. *Elife*, 3:e01581.

[Cubas et al., 1999] Cubas, P., Vincent, C., and Coen, E. (1999). An epigenetic mutation responsible for natural variation in floral symmetry. *Nature*, 401(6749):157–161.

[Dabney and Storey, 2000] Dabney, A. and Storey, J. D. (2000). qvalue: Q-value estimation for false discovery rate control. *R package version 1.43.0*.

[Danchin et al., 2011] Danchin, É., Charmantier, A., Champagne, F. A., Mesoudi, A., Pujol, B., and Blanchet, S. (2011). Beyond DNA: integrating inclusive inheritance into an extended theory of evolution. *Nat Rev Genet*, 12(7):475–486.

[Daxinger and Whitelaw, 2012] Daxinger, L. and Whitelaw, E. (2012). Understanding transgenerational epigenetic inheritance via the gametes in mammals. *Nat Rev Genet*, 13(3):153–162.

[De Bona et al., 2008] De Bona, F., Ossowski, S., Schneeberger, K., and Rätsch, G. (2008). Optimal spliced alignments of short sequence reads. *Bioinformatics*, 24(16):i174–i180.

[DePristo et al., 2011] DePristo, M. A., Banks, E., Poplin, R., Garimella, K. V., Maguire, J. R., Hartl, C., Philippakis, A. A., del Angel, G., Rivas, M. A., Hanna, M., McKenna, A., Fennell, T. J., Kernytsky, A. M., Sivachenko, A. Y., Cibulskis, K., Gabriel, S. B., Altshuler, D., and Daly, M. J. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet*, 43(5):491–498.

[Dolzhenko and Smith, 2014] Dolzhenko, E. and Smith, A. D. (2014). Using beta-binomial regression for high-precision differential methylation analysis in multifactor whole-genome bisulfite sequencing experiments. *BMC Bioinformatics*, 15:215.

[Dowen et al., 2012] Dowen, R. H., Pelizzola, M., Schmitz, R. J., Lister, R., Dowen, J. M., Nery, J. R., Dixon, J. E., and Ecker, J. R. (2012). Widespread dynamic DNA methylation in response to biotic stress. *Proc Natl Acad Sci U S A*, 109(32):E2183–E2191.

[Down et al., 2008] Down, T. A., Rakyan, V. K., Turner, D. J., Flicek, P., Li, H., Kulesha, E., Gräf, S., Johnson, N., Herrero, J., Tomazou, E. M., Thorne, N. P., Bäckdahl, L., Herberth, M., Howe, K. L., Jackson, D. K., Miretti, M. M., Marioni, J. C., Birney, E., Hubbard, T. J. P., Durbin, R., Tavaré, S., and Beck, S. (2008). A Bayesian deconvolution strategy for immunoprecipitation-based DNA methylome analysis. *Nat Biotechnol*, 26(7):779–785.

[Durand et al., 2012] Durand, S., Bouché, N., Perez Strand, E., Loudet, O., and Camilleri, C. (2012). Rapid establishment of genetic incompatibility through natural epigenetic variation. *Curr Biol*, 22(4):326–331.

[Durbin et al., 2007] Durbin, R., Eddy, S., Krogh, A., and Mitchinson, G. (2007). *Biological sequence analysis: Probabilistic models of proteins and nucleic acids*. Cambridge University Press,

Cambridge, UK.

[Eckhardt et al., 2006] Eckhardt, F., Lewin, J., Cortese, R., Rakyan, V. K., Attwood, J., Burger, M., Burton, J., Cox, T. V., Davies, R., Down, T. A., Haefliger, C., Horton, R., Howe, K., Jackson, D. K., Kunde, J., Koenig, C., Liddle, J., Niblett, D., Otto, T., Pettett, R., Seemann, S., Thompson, C., West, T., Rogers, J., Olek, A., Berlin, K., and Beck, S. (2006). DNA methylation profiling of human chromosomes 6, 20 and 22. *Nat Genet*, 38(12):1378–1385.

[Eichten et al., 2011] Eichten, S. R., Swanson-Wagner, R. A., Schnable, J. C., Waters, A. J., Hermanson, P. J., Liu, S., Yeh, C.-T., Jia, Y., Gendler, K., Freeling, M., Schnable, P. S., Vaughn, M. W., and Springer, N. M. (2011). Heritable epigenetic variation among maize inbreds. *PLoS Genet*, 7(11):e1002372.

[English et al., 2015] English, A. C., Salerno, W. J., Hampton, O. A., Gonzaga-Jauregui, C., Ambreth, S., Ritter, D. I., Beck, C. R., Davis, C. F., Dahdouli, M., Ma, S., Carroll, A., Veeraraghavan, N., Bruestle, J., Drees, B., Hastie, A., Lam, E. T., White, S., Mishra, P., Wang, M., Han, Y., Zhang, F., Stankiewicz, P., Wheeler, D. A., Reid, J. G., Muzny, D. M., Rogers, J., Sabo, A., Worley, K. C., Lupski, J. R., Boerwinkle, E., and Gibbs, R. A. (2015). Assessing structural variation in a personal genome-towards a human reference diploid genome. *BMC Genomics*, 16(1):286.

[Evanno et al., 2005] Evanno, G., Regnaut, S., and Goudet, J. (2005). Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. *Mol Ecol*, 14(8):2611–2620.

[Fang et al., 2012] Fang, F., Hodges, E., Molaro, A., Dean, M., Hannon, G. J., and Smith, A. D. (2012). Genomic landscape of human allele-specific DNA methylation. *Proc Natl Acad Sci U S A*, 109(19):7332–7337.

[Felsenfeld, 2014] Felsenfeld, G. (2014). The evolution of epigenetics. *Perspect Biol Med*, 57(1):132–148.

[Feng et al., 2014] Feng, H., Conneely, K. N., and Wu, H. (2014). A Bayesian hierarchical model to detect differentially methylated loci from single nucleotide resolution sequencing data. *Nucleic Acids Res*, 42(8):e69.

[Feng et al., 2010] Feng, S., Jacobsen, S. E., and Reik, W. (2010). Epigenetic reprogramming in plant and animal development. *Science*, 330(6004):622–627.

[Ferguson-Smith and Patti, 2011] Ferguson-Smith, A. C. and Patti, M.-E. (2011). You are what your dad ate. *Cell Metab*, 13(2):115–117.

[Flusberg et al., 2010] Flusberg, B. A., Webster, D. R., Lee, J. H., Travers, K. J., Olivares, E. C., Clark, T. A., Korlach, J., and Turner, S. W. (2010). Direct detection of DNA methylation during single-molecule, real-time sequencing. *Nat Methods*, 7(6):461–465.

[Fournier-Level et al., 2011] Fournier-Level, A., Korte, A., Cooper, M. D., Nordborg, M., Schmitt, J., and Wilczek, A. M. (2011). A map of local adaptation in Arabidopsis thaliana. *Science*, 334(6052):86–89.

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

[Fu et al., 2010] Fu, A. Q., Genereux, D. P., Stöger, R., Laird, C. D., and Stephens, M. (2010). STATISTICAL INFERENCE OF TRANSMISSION FIDELITY OF DNA METHYLATION PATTERNS OVER SOMATIC CELL DIVISIONS IN MAMMALS. *Ann Appl Stat*, 4(2):871–892.

[Fulneček and Kovařík, 2014] Fulneček, J. and Kovařík, A. (2014). How to interpret methylation sensitive amplified polymorphism (MSAP) profiles? *BMC Genet*, 15:2.

[Furrow and Feldman, 2014] Furrow, R. E. and Feldman, M. W. (2014). Genetic variation and the evolution of epigenetic regulation. *Evolution*, 68(3):673–683.

[Gan et al., 2011] Gan, X., Stegle, O., Behr, J., Steffen, J. G., Drewe, P., Hildebrand, K. L., Lyngsoe, R., Schultheiss, S. J., Osborne, E. J., Sreedharan, V. T., Kahles, A., Bohnert, R., Jean, G., Derwent, P., Kersey, P., Belfield, E. J., Harberd, N. P., Kemen, E., Toomajian, C., Kover, P. X., Clark, R. M., Rätsch, G., and Mott, R. (2011). Multiple reference genomes and transcriptomes for Arabidopsis

thaliana. *Nature*, 477(7365):419–423.

[Garrick et al., 1998] Garrick, D., Fiering, S., Martin, D. I., and Whitelaw, E. (1998). Repeat-induced gene silencing in mammals. *Nat Genet*, 18(1):56–59.

[Garrison and Gabor, 2012] Garrison, E. and Gabor (2012). Haplotype-based variant detection from short-read sequencing. *http://arxiv.org/abs/1207.3907*.

[Gehring, 2013] Gehring, M. (2013). Genomic imprinting: insights from plants. *Annu Rev Genet*, 47:187–208.

[Gehring et al., 2009] Gehring, M., Bubb, K. L., and Henikoff, S. (2009). Extensive demethylation of repetitive elements during seed development underlies gene imprinting. *Science*, 324(5933):1447–1451.

[Genereux et al., 2005] Genereux, D. P., Miner, B. E., Bergstrom, C. T., and Laird, C. D. (2005). A population-epigenetic model to infer site-specific methylation rates from double-stranded DNA methylation patterns. *Proc Natl Acad Sci U S A*, 102(16):5802–5807.

[Gordon, 2010] Gordon (2010). FastX toolkit: http://hannonlab.cshl.edu/fastx_toolkit/.

[Greaves et al., 2012] Greaves, I. K., Groszmann, M., Ying, H., Taylor, J. M., Peacock, W. J., and Dennis, E. S. (2012). Trans chromosomal methylation in Arabidopsis hybrids. *Proc Natl Acad Sci U S A*, 109(9):3570–3575.

[Grimm et al., 2013] Grimm, D., Hagmann, J., Koenig, D., Weigel, D., and Borgwardt, K. (2013). Accurate indel prediction using paired-end short reads. *BMC Genomics*, 14:132.

[Guo et al., 2014] Guo, H., Zhu, P., Yan, L., Li, R., Hu, B., Lian, Y., Yan, J., Ren, X., Lin, S., Li, J., Jin, X., Shi, X., Liu, P., Wang, X., Wang, W., Wei, Y., Li, X., Guo, F., Wu, X., Fan, X., Yong, J., Wen, L., Xie, S. X., Tang, F., and Qiao, J. (2014). The DNA methylation landscape of human early embryos. *Nature*, 511(7511):606–610.

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

[Hackett et al., 2013] Hackett, J. A., Sengupta, R., Zylicz, J. J., Murakami, K., Lee, C., Down, T. A., and Surani, M. A. (2013). Germline DNA demethylation dynamics and imprint erasure through 5-hydroxymethylcytosine. *Science*, 339(6118):448–452.

[Hagmann et al., 2015] Hagmann, J., Becker, C., Müller, J., Stegle, O., Meyer, R. C., Wang, G., Schneeberger, K., Fitz, J., Altmann, T., Bergelson, J., Borgwardt, K., and Weigel, D. (2015). Century-scale Methylome Stability in a Recently Diverged Arabidopsis thaliana Lineage. *PLoS Genet*, 11(1):e1004920.

[Haig, 2004] Haig, D. (2004). The (dual) origin of epigenetics. *Cold Spring Harb Symp Quant Biol*, 69:67–70.

[Hansen et al., 2012] Hansen, K. D., Langmead, B., and Irizarry, R. A. (2012). BSmooth: from whole genome bisulfite sequencing reads to differentially methylated regions. *Genome Biol*, 13(10):R83.

[Hauser et al., 2011] Hauser, M.-T., Aufsatz, W., Jonak, C., and Luschnig, C. (2011). Transgenerational epigenetic inheritance in plants. *Biochim Biophys Acta*, 1809(8):459–468.

[Hayden, 2014] Hayden, E. C. (2014). Technology: The $1,000 genome. *Nature*, 507(7492):294–295.

[He et al., 2013] He, G., Chen, B., Wang, X., Li, X., Li, J., He, H., Yang, M., Lu, L., Qi, Y., Wang, X., and Deng, X. W. (2013). Conservation and divergence of transcriptomic and epigenomic variation in maize hybrids. *Genome Biol*, 14(6):R57.

[Heard and Martienssen, 2014] Heard, E. and Martienssen, R. A. (2014). Transgenerational epigenetic inheritance: myths and mechanisms. *Cell*, 157(1):95–109.

[Hebestreit et al., 2013] Hebestreit, K., Dugas, M., and Klein, H.-U. (2013). Detection of significantly differentially methylated regions in targeted bisulfite sequencing data. *Bioinformatics*, 29(13):1647–1653.

[Henderson and Jacobsen, 2007] Henderson, I. R. and Jacobsen, S. E. (2007). Epigenetic inheritance in plants. *Nature*, 447(7143):418–424.

[Henikoff and Dalal, 2005] Henikoff, S. and Dalal, Y. (2005). Centromeric chromatin: what makes it unique? *Curr Opin Genet Dev*, 15(2):177–184.

[Hirayama and Shinozaki, 2010] Hirayama, T. and Shinozaki, K. (2010). Research on plant abiotic stress responses in the post-genome era: past, present and future. *Plant J*, 61(6):1041–1052.

[Hirsch et al., 2012] Hirsch, S., Baumberger, R., and Grossniklaus, U. (2012). Epigenetic variation, inheritance, and selection in plant populations. *Cold Spring Harb Symp Quant Biol*, 77:97–104.

[Ho et al., 2014] Ho, J. W. K., Jung, Y. L., Liu, T., Alver, B. H., Lee, S., Ikegami, K., Sohn, K.-A., Minoda, A., Tolstorukov, M. Y., Appert, A., Parker, S. C. J., Gu, T., Kundaje, A., Riddle, N. C., Bishop, E., Egelhofer, T. A., Hu, S. S., Alekseyenko, A. A., Rechtsteiner, A., Asker, D., Belsky, J. A., Bowman, S. K., Chen, Q. B., Chen, R. A.-J., Day, D. S., Dong, Y., Dose, A. C., Duan, X., Epstein, C. B., Ercan, S., Feingold, E. A., Ferrari, F., Garrigues, J. M., Gehlenborg, N., Good, P. J., Haseley, P., He, D., Herrmann, M., Hoffman, M. M., Jeffers, T. E., Kharchenko, P. V., Kolasinska-Zwierz, P., Kotwaliwale, C. V., Kumar, N., Langley, S. A., Larschan, E. N., Latorre, I., Libbrecht, M. W., Lin, X., Park, R., Pazin, M. J., Pham, H. N., Plachetka, A., Qin, B., Schwartz, Y. B., Shoresh, N., Stempor, P., Vielle, A., Wang, C., Whittle, C. M., Xue, H., Kingston, R. E., Kim, J. H., Bernstein, B. E., Dernburg, A. F., Pirrotta, V., Kuroda, M. I., Noble, W. S., Tullius, T. D., Kellis, M., MacAlpine, D. M., Strome, S., Elgin, S. C. R., Liu, X. S., Lieb, J. D., Ahringer, J., Karpen, G. H., and Park, P. J. (2014). Comparative analysis of metazoan chromatin organization. *Nature*, 512(7515):449–452.

[Hodges et al., 2011] Hodges, E., Molaro, A., Dos Santos, C. O., Thekkat, P., Song, Q., Uren, P. J., Park, J., Butler, J., Rafii, S., McCombie, W. R., Smith, A. D., and Hannon, G. J. (2011). Directional DNA methylation changes and complex intermediate states accompany lineage specificity in the adult hematopoietic compartment. *Mol Cell*, 44(1):17–28.

[Hollick, 2012] Hollick, J. B. (2012). Paramutation: a trans-homolog interaction affecting heritable gene regulation. *Curr Opin Plant Biol*, 15(5):536–543.

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

[House et al., 2014] House, N. C. M., Koch, M. R., and Freudenreich, C. H. (2014). Chromatin modifications and DNA repair: beyond double-strand breaks. *Front Genet*, 5:296.

[Hsieh et al., 2009] Hsieh, T.-F., Ibarra, C. A., Silva, P., Zemach, A., Eshed-Williams, L., Fischer, R. L., and Zilberman, D. (2009). Genome-wide demethylation of Arabidopsis endosperm. *Science*, 324(5933):1451–1454.

[Huang et al., 2014] Huang, H., Sabari, B. R., Garcia, B. A., Allis, C. D., and Zhao, Y. (2014). SnapShot: Histone Modifications. *Cell*, 159(2):458–458.e1.

[Huddleston et al., 2014] Huddleston, J., Ranade, S., Malig, M., Antonacci, F., Chaisson, M., Hon, L., Sudmant, P. H., Graves, T. A., Alkan, C., Dennis, M. Y., Wilson, R. K., Turner, S. W., Korlach, J., and Eichler, E. E. (2014). Reconstructing complex regions of genomes using long-read sequencing technology. *Genome Res*, 24(4):688–696.

[Huson and Bryant, 2006] Huson, D. H. and Bryant, D. (2006). Application of phylogenetic networks in evolutionary studies. *Mol Biol Evol*, 23(2):254–267.

[Ibarra et al., 2012] Ibarra, C. A., Feng, X., Schoft, V. K., Hsieh, T.-F., Uzawa, R., Rodrigues, J. A., Zemach, A., Chumak, N., Machlicova, A., Nishimura, T., Rojas, D., Fischer, R. L., Tamaru, H., and Zilberman, D. (2012). Active DNA demethylation in plant companion cells reinforces transposon methylation in gametes. *Science*, 337(6100):1360–1364.

[Ingouff et al., 2010] Ingouff, M., Rademacher, S., Holec, S., Soljić, L., Xin, N., Readshaw, A., Foo, S. H., Lahouze, B., Sprunck, S., and Berger, F. (2010). Zygotic resetting of the HISTONE 3 variant repertoire participates in epigenetic reprogramming in Arabidopsis. *Curr Biol*, 20(23):2137–2143.

[Ito et al., 2011] Ito, H., Gaubert, H., Bucher, E., Mirouze, M., Vaillant, I., and Paszkowski, J. (2011). An siRNA pathway prevents transgenerational retrotransposition in plants subjected to stress. *Nature*, 472(7341):115–119.

[Iwasaki and Paszkowski, 2014] Iwasaki, M. and Paszkowski, J. (2014). Epigenetic memory in plants. *EMBO J*.

[Jablonka, 2013] Jablonka, E. (2013). Epigenetic inheritance and plasticity: The responsive germline. *Prog Biophys Mol Biol*, 111(2-3):99–107.

[Jablonka and Raz, 2009] Jablonka, E. and Raz, G. (2009). Transgenerational epigenetic inheritance: prevalence, mechanisms, and implications for the study of heredity and evolution. *Q Rev Biol*, 84(2):131–176.

[Jean et al., 2010] Jean, G., Kahles, A., Sreedharan, V. T., De Bona, F., and Rätsch, G. (2010). RNA-Seq read alignments with PALMapper. *Curr Protoc Bioinformatics*, Chapter 11:Unit 11.6.

[Johannes et al., 2009] Johannes, F., Porcher, E., Teixeira, F. K., Saliba-Colombani, V., Simon, M., Agier, N., Bulski, A., Albuisson, J., Heredia, F., Audigier, P., Bouchez, D., Dillmann, C., Guerche, P., Hospital, F., and Colot, V. (2009). Assessing the impact of transgenerational epigenetic variation on complex traits. *PLoS Genet*, 5(6):e1000530.

[Jones et al., 2001] Jones, L., Ratcliff, F., and Baulcombe, D. C. (2001). RNA-directed transcriptional gene silencing in plants can be inherited independently of the RNA trigger and requires Met1 for maintenance. *Curr Biol*, 11(10):747–757.

[Jones, 2012] Jones, P. A. (2012). Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nat Rev Genet*, 13(7):484–492.

[Kaati et al., 2002] Kaati, G., Bygren, L. O., and Edvinsson, S. (2002). Cardiovascular and diabetes mortality determined by nutrition during parents' and grandparents' slow growth period. *Eur J Hum Genet*, 10(11):682–688.

[Kakutani et al., 1996] Kakutani, T., Jeddeloh, J. A., Flowers, S. K., Munakata, K., and Richards, E. J. (1996). Developmental abnormalities and epimutations associated with DNA hypomethylation mutations. *Proc Natl Acad Sci U S A*, 93(22):12406–12411.

[Kang et al., 2010] Kang, H. M., Sul, J. H., Service, S. K., Zaitlen, N. A., Kong, S.-Y., Freimer, N. B., Sabatti, C., and Eskin, E. (2010). Variance component model to account for sample structure in genome-wide association studies. *Nat Genet*, 42(4):348–354.

[Kankel et al., 2003] Kankel, M. W., Ramsey, D. E., Stokes, T. L., Flowers, S. K., Haag, J. R., Jeddeloh, J. A., Riddle, N. C., Verbsky, M. L., and Richards, E. J. (2003). Arabidopsis MET1 cytosine methyltransferase mutants. *Genetics*, 163(3):1109–1122.

[Kantlehner et al., 2011] Kantlehner, M., Kirchner, R., Hartmann, P., Ellwart, J. W., Alunni-Fabbroni, M., and Schumacher, A. (2011). A high-throughput DNA methylation analysis of a single cell. *Nucleic Acids Res*, 39(7):e44.

[Kaplan and Dekker, 2013] Kaplan, N. and Dekker, J. (2013). High-throughput genome scaffolding from in vivo DNA interaction frequency. *Nat Biotechnol*, 31(12):1143–1147.

[Kathiria et al., 2010] Kathiria, P., Sidler, C., Golubov, A., Kalischuk, M., Kawchuk, L. M., and Kovalchuk, I. (2010). Tobacco mosaic virus infection results in an increase in recombination frequency and resistance to viral, bacterial, and fungal pathogens in the progeny of infected tobacco plants. *Plant Physiol*, 153(4):1859–1870.

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

[Kegel, 2009] Kegel, B. (2009). *Epigenetik: wie Erfahrungen vererbt werden*. DuMont, Köln, 1. aufl. edition.

[Keller and Yi, 2014] Keller, T. E. and Yi, S. V. (2014). DNA methylation and evolution of duplicate genes. *Proc Natl Acad Sci U S A*, 111(16):5932–5937.

[Kingsmore and Saunders, 2011] Kingsmore, S. F. and Saunders, C. J. (2011). Deep sequencing of patient genomes for disease diagnosis: when will it become routine? *Sci Transl Med*, 3(87):87ps23.

[Kinoshita et al., 2007] Kinoshita, Y., Saze, H., Kinoshita, T., Miura, A., Soppe, W. J. J., Koornneef, M., and Kakutani, T. (2007). Control of FWA gene silencing in Arabidopsis thaliana by SINE-related direct repeats. *Plant J*, 49(1):38–45.

[Kitzman et al., 2012] Kitzman, J. O., Snyder, M. W., Ventura, M., Lewis, A. P., Qiu, R., Simmons, L. E., Gammill, H. S., Rubens, C. E., Santillan, D. A., Murray, J. C., Tabor, H. K., Bamshad, M. J., Eichler, E. E., and Shendure, J. (2012). Noninvasive whole-genome sequencing of a human fetus. *Sci Transl Med*, 4(137):137ra76.

[Koren et al., 2012] Koren, S., Schatz, M. C., Walenz, B. P., Martin, J., Howard, J. T., Ganapathy, G., Wang, Z., Rasko, D. A., McCombie, W. R., Jarvis, E. D., and Adam M Phillippy (2012).

Hybrid error correction and de novo assembly of single-molecule sequencing reads. *Nat Biotechnol*, 30(7):693–700.

[Kovalchuk et al., 2003] Kovalchuk, O., Burke, P., Arkhipov, A., Kuchma, N., James, S. J., Kovalchuk, I., and Pogribny, I. (2003). Genome hypermethylation in Pinus silvestris of Chernobyl–a mechanism for radiation adaptation? *Mutat Res*, 529(1-2):13–20.

[Krsticevic et al., 2015] Krsticevic, F. J., Schrago, C. G., and Carvalho, A. B. (2015). Long-Read Single Molecule Sequencing To Resolve Tandem Gene Copies: The Mst77y Region on the Drosophila melanogaster Y Chromosome. *G3 (Bethesda)*.

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

[Krueger et al., 2012] Krueger, F., Kreck, B., Franke, A., and Andrews, S. R. (2012). DNA methylome analysis using short bisulfite sequencing data. *Nat Methods*, 9(2):145–151.

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

[Lam et al., 2012] Lam, H. Y. K., Pan, C., Clark, M. J., Lacroute, P., Chen, R., Haraksingh, R., O'Huallachain, M., Gerstein, M. B., Kidd, J. M., Bustamante, C. D., and Snyder, M. (2012). Detecting and annotating genetic variations using the HugeSeq pipeline. *Nat Biotechnol*, 30(3):226–229.

[Lander et al., 2001] Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczky, J., LeVine, R., McEwan, P., McKernan, K., Meldrim, J., Mesirov, J. P., Miranda, C., Morris, W., Naylor, J., Raymond, C., Rosetti, M., Santos, R., Sheridan, A., Sougnez, C., Stange-Thomann, N., Stojanovic, N., Subramanian, A., Wyman, D., Rogers, J., Sulston, J., Ainscough, R., Beck, S., Bentley, D., Burton, J., Clee, C., Carter, N., Coulson, A., Deadman, R., Deloukas, P., Dunham, A., Dunham, I., Durbin, R., French, L., Grafham, D., Gregory, S., Hubbard, T., Humphray, S., Hunt, A., Jones, M., Lloyd, C., McMurray, A., Matthews, L., Mercer, S., Milne, S., Mullikin, J. C., Mungall, A., Plumb, R., Ross, M., Shownkeen, R., Sims, S., Waterston, R. H., Wilson, R. K., Hillier, L. W., McPherson, J. D., Marra, M. A., Mardis, E. R., Fulton, L. A., Chinwalla, A. T., Pepin, K. H., Gish, W. R., Chissoe, S. L., Wendl, M. C., Delehaunty, K. D., Miner, T. L., Delehaunty, A., Kramer, J. B., Cook, L. L., Fulton, R. S., Johnson, D. L., Minx, P. J., Clifton, S. W., Hawkins, T., Branscomb, E., Predki, P., Richardson, P., Wenning, S., Slezak, T., Doggett, N., Cheng, J. F., Olsen, A., Lucas, S., Elkin, C., Uberbacher, E., Frazier, M., Gibbs, R. A., Muzny, D. M., Scherer, S. E., Bouck, J. B., Sodergren, E. J., Worley, K. C., Rives, C. M., Gorrell, J. H., Metzker, M. L., Naylor, S. L., Kucherlapati, R. S., Nelson, D. L., Weinstock, G. M., Sakaki, Y., Fujiyama, A., Hattori, M., Yada, T., Toyoda, A., Itoh, T., Kawagoe, C., Watanabe, H., Totoki, Y., Taylor, T., Weissenbach, J., Heilig, R., Saurin, W., Artiguenave, F., Brottier, P., Bruls, T., Pelletier, E., Robert, C., Wincker, P., Smith, D. R., Doucette-Stamm, L., Rubenfield, M., Weinstock, K., Lee, H. M., Dubois, J., Rosenthal, A., Platzer, M., Nyakatura, G., Taudien, S., Rump, A., Yang, H., Yu, J., Wang, J., Huang, G., Gu, J., Hood, L., Rowen, L., Madan, A., Qin, S., Davis, R. W., Federspiel, N. A., Abola, A. P., Proctor, M. J., Myers, R. M., Schmutz, J., Dickson, M., Grimwood, J., Cox, D. R., Olson, M. V., Kaul, R., Raymond, C., Shimizu, N., Kawasaki, K., Minoshima, S., Evans, G. A., Athanasiou, M., Schultz, R., Roe, B. A., Chen, F., Pan, H., Ramser, J., Lehrach, H., Reinhardt, R., McCombie, W. R., de la Bastide, M., Dedhia, N., Blöcker, H., Hornischer, K., Nordsiek, G., Agarwala, R., Aravind, L., Bailey, J. A., Bateman, A., Batzoglou, S., Birney, E., Bork, P., Brown, D. G., Burge, C. B., Cerutti, L., Chen, H. C., Church, D., Clamp, M., Copley, R. R., Doerks, T., Eddy, S. R., Eichler, E. E., Furey, T. S., Galagan, J., Gilbert, J. G., Harmon, C., Hayashizaki, Y., Haussler, D., Hermjakob, H., Hokamp, K., Jang, W., Johnson, L. S., Jones, T. A., Kasif, S., Kaspryzk, A., Kennedy, S., Kent, W. J., Kitts, P., Koonin, E. V., Korf, I., Kulp, D., Lancet, D., Lowe, T. M., McLysaght, A., Mikkelsen, T., Moran, J. V., Mulder, N., Pollara, V. J., Ponting, C. P., Schuler, G., Schultz, J., Slater, G., Smit, A. F., Stupka, E., Szustakowski, J., Thierry-Mieg, D., Thierry-Mieg, J., Wagner, L., Wallis, J., Wheeler, R., Williams, A., Wolf, Y. I., Wolfe, K. H., Yang, S. P., Yeh, R. F., Collins, F., Guyer,

M. S., Peterson, J., Felsenfeld, A., Wetterstrand, K. A., Patrinos, A., Morgan, M. J., de Jong, P., Catanese, J. J., Osoegawa, K., Shizuya, H., Choi, S., Chen, Y. J., Szustakowki, J., and , I. H. G. S. C. (2001). Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921.

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

[Laskey and Gurdon, 1970] Laskey, R. A. and Gurdon, J. B. (1970). Genetic content of adult somatic cells tested by nuclear transplantation from cultured cells. *Nature*, 228(5278):1332–1334.

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

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

[Le et al., 2014] Le, T.-N., Schumann, U., Smith, N. A., Tiwari, S., Au, P., Zhu, Q.-H., Taylor, J., Kazan, K., Llewellyn, D. J., Zhang, R., Dennis, E. S., and Wang, M.-B. (2014). DNA demethylases target promoter transposable elements to positively regulate stress responsive genes in Arabidopsis. *Genome Biol*, 15(9):458.

[Lee et al., 2014] Lee, H. J., Hore, T. A., and Reik, W. (2014). Reprogramming the methylome: erasing memory and creating diversity. *Cell Stem Cell*, 14(6):710–719.

[Lee et al., 2010] Lee, T.-F., Zhai, J., and Meyers, B. C. (2010). Conservation and divergence in eukaryotic DNA methylation. *Proc Natl Acad Sci U S A*, 107(20):9027–9028.

[Leek et al., 2010] Leek, J. T., Scharpf, R. B., Bravo, H. C., Simcha, D., Langmead, B., Johnson, W. E., Geman, D., Baggerly, K., and Irizarry, R. A. (2010). Tackling the widespread and critical impact of batch effects in high-throughput data. *Nat Rev Genet*, 11(10):733–739.

[Leung et al., 2015] Leung, D., Jung, I., Rajagopal, N., Schmitt, A., Selvaraj, S., Lee, A. Y., Yen, C.-A., Lin, S., Lin, Y., Qiu, Y., Xie, W., Yue, F., Hariharan, M., Ray, P., Kuan, S., Edsall, L., Yang, H., Chi, N. C., Zhang, M. Q., Ecker, J. R., and Ren, B. (2015). Integrative analysis of haplotype-resolved epigenomes across human tissues. *Nature*, 518(7539):350–354.

[Li and Durbin, 2009] Li, H. and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, 25(14):1754–1760.

[Li et al., 2013] Li, L., Petsch, K., Shimizu, R., Liu, S., Xu, W. W., Ying, K., Yu, J., Scanlon, M. J., Schnable, P. S., Timmermans, M. C. P., Springer, N. M., and Muehlbauer, G. J. (2013). Mendelian and non-Mendelian regulation of gene expression in maize. *PLoS Genet*, 9(1):e1003202.

[Li et al., 2012] Li, X., Zhu, J., Hu, F., Ge, S., Ye, M., Xiang, H., Zhang, G., Zheng, X., Zhang, H., Zhang, S., Li, Q., Luo, R., Yu, C., Yu, J., Sun, J., Zou, X., Cao, X., Xie, X., Wang, J., and Wang, W. (2012). Single-base resolution maps of cultivated and wild rice methylomes and regulatory roles of DNA methylation in plant gene expression. *BMC Genomics*, 13:300.

[Lin et al., 2014] Lin, K., Smit, S., Bonnema, G., Sanchez-Perez, G., and de Ridder, D. (2014). Making the difference: integrating structural variation detection tools. *Brief Bioinform*.

[Lippert et al., 2011] Lippert, C., Listgarten, J., Liu, Y., Kadie, C. M., Davidson, R. I., and Heckerman, D. (2011). FaST linear mixed models for genome-wide association studies. *Nat Methods*, 8(10):833–835.

[Lira-Medeiros et al., 2010] Lira-Medeiros, C. F., Parisod, C., Fernandes, R. A., Mata, C. S., Cardoso, M. A., and Ferreira, P. C. G. (2010). Epigenetic variation in mangrove plants occurring in contrasting natural environment. *PLoS One*, 5(4):e10326.

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

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

[Lister et al., 2011] Lister, R., Pelizzola, M., Kida, Y. S., Hawkins, R. D., Nery, J. R., Hon, G.,

Antosiewicz-Bourget, J., O'Malley, R., Castanon, R., Klugman, S., Downes, M., Yu, R., Stewart, R., Ren, B., Thomson, J. A., Evans, R. M., and Ecker, J. R. (2011). Hotspots of aberrant epigenomic reprogramming in human induced pluripotent stem cells. *Nature*, 471(7336):68–73.

[Long et al., 2013] Long, Q., Rabanal, F. A., Meng, D., Huber, C. D., Farlow, A., Platzer, A., Zhang, Q., Vilhjálmsson, B. J., Korte, A., Nizhynska, V., Voronin, V., Korte, P., Sedman, L., Mandáková, T., Lysak, M. A., Seren, Ü., Hellmann, I., and Nordborg, M. (2013). Massive genomic variation and strong selection in Arabidopsis thaliana lines from Sweden. *Nat Genet*, 45(8):884–890.

[Lumey et al., 2009] Lumey, L. H., Stein, A. D., Kahn, H. S., and Romijn, J. A. (2009). Lipid profiles in middle-aged men and women after famine exposure during gestation: the Dutch Hunger Winter Families Study. *Am J Clin Nutr*, 89(6):1737–1743.

[Luna et al., 2012] Luna, E., Bruce, T. J. A., Roberts, M. R., Flors, V., and Ton, J. (2012). Next-generation systemic acquired resistance. *Plant Physiol*, 158(2):844–853.

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

[Manning et al., 2006] Manning, K., Tör, M., Poole, M., Hong, Y., Thompson, A. J., King, G. J., Giovannoni, J. J., and Seymour, G. B. (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.

[Manolio et al., 2009] Manolio, T. A., Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorff, L. A., Hunter, D. J., McCarthy, M. I., Ramos, E. M., Cardon, L. R., Chakravarti, A., Cho, J. H., Guttmacher, A. E., Kong, A., Kruglyak, L., Mardis, E., Rotimi, C. N., Slatkin, M., Valle, D., Whittemore, A. S., Boehnke, M., Clark, A. G., Eichler, E. E., Gibson, G., Haines, J. L., Mackay, T. F. C., McCarroll, S. A., and Visscher, P. M. (2009). Finding the missing heritability of complex diseases. *Nature*, 461(7265):747–753.

[Marí-Ordóñez et al., 2013] Marí-Ordóñez, A., Marchais, A., Etcheverry, M., Martin, A., Colot, V., and Voinnet, O. (2013). Reconstructing de novo silencing of an active plant retrotransposon. *Nat Genet*, 45(9):1029–1039.

[Martin et al., 2009] Martin, A., Troadec, C., Boualem, A., Rajab, M., Fernandez, R., Morin, H., Pitrat, M., Dogimont, C., and Bendahmane, A. (2009). A transposon-induced epigenetic change leads to sex determination in melon. *Nature*, 461(7267):1135–1138.

[Mathieu et al., 2007] Mathieu, O., Reinders, J., Caikovski, M., Smathajitt, C., and Paszkowski, J. (2007). Transgenerational stability of the Arabidopsis epigenome is coordinated by CG methylation. *Cell*, 130(5):851–862.

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

[Maunakea et al., 2010] Maunakea, A. K., Nagarajan, R. P., Bilenky, M., Ballinger, T. J., D'Souza, C., Fouse, S. D., Johnson, B. E., Hong, C., Nielsen, C., Zhao, Y., Turecki, G., Delaney, A., Varhol, R., Thiessen, N., Shchors, K., Heine, V. M., Rowitch, D. H., Xing, X., Fiore, C., Schillebeeckx, M., Jones, S. J. M., Haussler, D., Marra, M. A., Hirst, M., Wang, T., and Costello, J. F. (2010). Conserved role of intragenic DNA methylation in regulating alternative promoters. *Nature*, 466(7303):253–257.

[McCarthy et al., 2008] McCarthy, M. I., Abecasis, G. R., Cardon, L. R., Goldstein, D. B., Little, J., Ioannidis, J. P. A., and Hirschhorn, J. N. (2008). Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nat Rev Genet*, 9(5):356–369.

[McClintock, 1984] McClintock, B. (1984). The significance of responses of the genome to challenge. *Science*, 226(4676):792–801.

[Medrano et al., 2014] Medrano, M., Herrera, C. M., and Bazaga, P. (2014). Epigenetic variation predicts regional and local intraspecific functional diversity in a perennial herb. *Mol Ecol*, 23(20):4926–4938.

[Medvedev et al., 2009] Medvedev, P., Stanciu, M., and Brudno, M. (2009). Computational methods

for discovering structural variation with next-generation sequencing. *Nat Methods*, 6(11 Suppl):S13–S20.

[Meissner et al., 2008] Meissner, A., Mikkelsen, T. S., Gu, H., Wernig, M., Hanna, J., Sivachenko, A., Zhang, X., Bernstein, B. E., Nusbaum, C., Jaffe, D. B., Gnirke, A., Jaenisch, R., and Lander, E. S. (2008). Genome-scale DNA methylation maps of pluripotent and differentiated cells. *Nature*, 454(7205):766–770.

[Mensaert et al., 2014] Mensaert, K., Denil, S., Trooskens, G., Van Criekinge, W., Thas, O., and De Meyer, T. (2014). Next-generation technologies and data analytical approaches for epigenomics. *Environ Mol Mutagen*, 55(3):155–170.

[Metzker, 2010] Metzker, M. L. (2010). Sequencing technologies - the next generation. *Nat Rev Genet*, 11(1):31–46.

[Mikheyev and Tin, 2014] Mikheyev, A. S. and Tin, M. M. Y. (2014). A first look at the Oxford Nanopore MinION sequencer. *Mol Ecol Resour*, 14(6):1097–1102.

[Mills et al., 2011] Mills, R. E., Walter, K., Stewart, C., Handsaker, R. E., Chen, K., Alkan, C., Abyzov, A., Yoon, S. C., Ye, K., Cheetham, R. K., Chinwalla, A., Conrad, D. F., Fu, Y., Grubert, F., Hajirasouliha, I., Hormozdiari, F., Iakoucheva, L. M., Iqbal, Z., Kang, S., Kidd, J. M., Konkel, M. K., Korn, J., Khurana, E., Kural, D., Lam, H. Y. K., Leng, J., Li, R., Li, Y., Lin, C.-Y., Luo, R., Mu, X. J., Nemesh, J., Peckham, H. E., Rausch, T., Scally, A., Shi, X., Stromberg, M. P., Stütz, A. M., Urban, A. E., Walker, J. A., Wu, J., Zhang, Y., Zhang, Z. D., Batzer, M. A., Ding, L., Marth, G. T., McVean, G., Sebat, J., Snyder, M., Wang, J., Ye, K., Eichler, E. E., Gerstein, M. B., Hurles, M. E., Lee, C., McCarroll, S. A., Korbel, J. O., and , . G. P. (2011). Mapping copy number variation by population-scale genome sequencing. *Nature*, 470(7332):59–65.

[Mimori et al., 2013] Mimori, T., Nariai, N., Kojima, K., Takahashi, M., Ono, A., Sato, Y., Yamaguchi-Kabata, Y., and Nagasaki, M. (2013). iSVP: an integrated structural variant calling pipeline from high-throughput sequencing data. *BMC Syst Biol*, 7 Suppl 6:S8.

[Mitchell-Olds and Schmitt, 2006] Mitchell-Olds, T. and Schmitt, J. (2006). Genetic mechanisms and evolutionary significance of natural variation in Arabidopsis. *Nature*, 441(7096):947–952.

[Miura et al., 2012] Miura, F., Enomoto, Y., Dairiki, R., and Ito, T. (2012). Amplification-free whole-genome bisulfite sequencing by post-bisulfite adaptor tagging. *Nucleic Acids Res*, 40(17):e136.

[Miura et al., 2009] Miura, K., Agetsuma, M., Kitano, H., Yoshimura, A., Matsuoka, M., Jacobsen, S. E., and Ashikari, M. (2009). A metastable DWARF1 epigenetic mutant affecting plant stature in rice. *Proc Natl Acad Sci U S A*, 106(27):11218–11223.

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

[Molinier et al., 2006] Molinier, J., Ries, G., Zipfel, C., and Hohn, B. (2006). Transgeneration memory of stress in plants. *Nature*, 442(7106):1046–1049.

[Mosher et al., 2009] Mosher, R. A., Melnyk, C. W., Kelly, K. A., Dunn, R. M., Studholme, D. J., and Baulcombe, D. C. (2009). Uniparental expression of PolIV-dependent siRNAs in developing endosperm of Arabidopsis. *Nature*, 460(7252):283–286.

[Muller, 1930] Muller, H. (1930). Types of visible variations induced by X-rays in Drosophila. *J Genet*, 22:299–334.

[Nanney, 1958] Nanney, D. L. (1958). Epigenetic control systems. *Proc Natl Acad Sci U S A*, 44(7):712–717.

[Nuthikattu et al., 2013] Nuthikattu, S., McCue, A. D., Panda, K., Fultz, D., DeFraia, C., Thomas, E. N., and Slotkin, R. K. (2013). The initiation of epigenetic silencing of active transposable elements is triggered by RDR6 and 21-22 nucleotide small interfering RNAs. *Plant Physiol*, 162(1):116–131.

[Ossowski et al., 2008] Ossowski, S., Schneeberger, K., Clark, R. M., Lanz, C., Warthmann, N., and Weigel, D. (2008). Sequencing of natural strains of Arabidopsis thaliana with short reads. *Genome Res*, 18(12):2024–2033.

[Ossowski et al., 2010] Ossowski, S., Schneeberger, K., Lucas-Lledó, J. I., Warthmann, N., Clark,

R. M., Shaw, R. G., Weigel, D., and Lynch, M. (2010). The rate and molecular spectrum of spontaneous mutations in Arabidopsis thaliana. *Science*, 327(5961):92–94.

[Pagnussat et al., 2005] Pagnussat, G. C., Yu, H.-J., Ngo, Q. A., Rajani, S., Mayalagu, S., Johnson, C. S., Capron, A., Xie, L.-F., Ye, D., and Sundaresan, V. (2005). Genetic and molecular identification of genes required for female gametophyte development and function in Arabidopsis. *Development*, 132(3):603–614.

[Park et al., 2014] Park, Y., Figueroa, M. E., Rozek, L. S., and Sartor, M. A. (2014). MethylSig: a whole genome DNA methylation analysis pipeline. *Bioinformatics*, 30(17):2414–2422.

[Paszkowski and Grossniklaus, 2011] Paszkowski, J. and Grossniklaus, U. (2011). Selected aspects of transgenerational epigenetic inheritance and resetting in plants. *Curr Opin Plant Biol*, 14(2):195–203.

[Patel and Jain, 2012] Patel, R. K. and Jain, M. (2012). NGS QC Toolkit: a toolkit for quality control of next generation sequencing data. *PLoS One*, 7(2):e30619.

[Pecinka et al., 2010] Pecinka, A., Dinh, H. Q., Baubec, T., Rosa, M., Lettner, N., and Mittelsten Scheid, O. (2010). Epigenetic regulation of repetitive elements is attenuated by prolonged heat stress in Arabidopsis. *Plant Cell*, 22(9):3118–3129.

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

[Pecinka et al., 2009] Pecinka, A., Rosa, M., Schikora, A., Berlinger, M., Hirt, H., Luschnig, C., and Mittelsten Scheid, O. (2009). Transgenerational stress memory is not a general response in Arabidopsis. *PLoS One*, 4(4):e5202.

[Peng and Ecker, 2012] Peng, Q. and Ecker, J. R. (2012). Detection of allele-specific methylation through a generalized heterogeneous epigenome model. *Bioinformatics*, 28(12):i163–i171.

[Platt et al., 2010] Platt, A., Horton, M., Huang, Y. S., Li, Y., Anastasio, A. E., Mulyati, N. W., Agren, J., Bossdorf, O., Byers, D., Donohue, K., Dunning, M., Holub, E. B., Hudson, A., Le Corre, V., Loudet, O., Roux, F., Warthmann, N., Weigel, D., Rivero, L., Scholl, R., Nordborg, M., Bergelson, J., and Borevitz, J. O. (2010). The scale of population structure in Arabidopsis thaliana. *PLoS Genet*, 6(2):e1000843.

[Plongthongkum et al., 2014] Plongthongkum, N., Diep, D. H., and Zhang, K. (2014). Advances in the profiling of DNA modifications: cytosine methylation and beyond. *Nat Rev Genet*, 15(10):647–661.

[Pritchard et al., 2000] Pritchard, J. K., Stephens, M., and Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics*, 155(2):945–959.

[Prüfer et al., 2014] Prüfer, K., Racimo, F., Patterson, N., Jay, F., Sankararaman, S., Sawyer, S., Heinze, A., Renaud, G., Sudmant, P. H., de Filippo, C., Li, H., Mallick, S., Dannemann, M., Fu, Q., Kircher, M., Kuhlwilm, M., Lachmann, M., Meyer, M., Ongyerth, M., Siebauer, M., Theunert, C., Tandon, A., Moorjani, P., Pickrell, J., Mullikin, J. C., Vohr, S. H., Green, R. E., Hellmann, I., Johnson, P. L. F., Blanche, H., Cann, H., Kitzman, J. O., Shendure, J., Eichler, E. E., Lein, E. S., Bakken, T. E., Golovanova, L. V., Doronichev, V. B., Shunkov, M. V., Derevianko, A. P., Viola, B., Slatkin, M., Reich, D., Kelso, J., and Pääbo, S. (2014). The complete genome sequence of a Neanderthal from the Altai Mountains. *Nature*, 505(7481):43–49.

[Qian et al., 2012] Qian, W., Miki, D., Zhang, H., Liu, Y., Zhang, X., Tang, K., Kan, Y., La, H., Li, X., Li, S., Zhu, X., Shi, X., Zhang, K., Pontes, O., Chen, X., Liu, R., Gong, Z., and Zhu, J.-K. (2012). A histone acetyltransferase regulates active DNA demethylation in Arabidopsis. *Science*, 336(6087):1445–1448.

[Rabiner, 1989] Rabiner, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77, Issue: 2:257 – 286.

[Rando and Verstrepen, 2007] Rando, O. J. and Verstrepen, K. J. (2007). Timescales of genetic and epigenetic inheritance. *Cell*, 128(4):655–668.

[Rassoulzadegan et al., 2006] Rassoulzadegan, M., Grandjean, V., Gounon, P., Vincent, S., Gillot, I., and Cuzin, F. (2006). RNA-mediated non-mendelian inheritance of an epigenetic change in the mouse. *Nature*, 441(7092):469–474.

[Rausch et al., 2012] Rausch, T., Zichner, T., Schlattl, A., Stütz, A. M., Benes, V., and Korbel, J. O. (2012). DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics*, 28(18):i333–i339.

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

[Reyna-López et al., 1997] Reyna-López, G. E., Simpson, J., and Ruiz-Herrera, J. (1997). Differences in DNA methylation patterns are detectable during the dimorphic transition of fungi by amplification of restriction polymorphisms. *Mol Gen Genet*, 253(6):703–710.

[Richards, 2006] Richards, E. J. (2006). Inherited epigenetic variation–revisiting soft inheritance. *Nat Rev Genet*, 7(5):395–401.

[Riggs and Porter, 1996] Riggs, A. and Porter, T. (1996). Overview of epigenetic mechanisms. In *Epigenetic mechanisms of gene regulation* (ed. Russo VEA, Martienssen R, Riggs AD). Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

[Riggs, 1975] Riggs, A. D. (1975). X inactivation, differentiation, and DNA methylation. *Cytogenet Cell Genet*, 14(1):9–25.

[Ritz et al., 2014] Ritz, A., Bashir, A., Sindi, S., Hsu, D., Hajirasouliha, I., and Raphael, B. J. (2014). Characterization of structural variants with single molecule and hybrid sequencing approaches. *Bioinformatics*, 30(24):3458–3466.

[Roadmap Epigenomics Consortium et al., 2015] Roadmap Epigenomics Consortium, Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Heravi-Moussavi, A., Kheradpour, P., Zhang, Z., Wang, J., Ziller, M. J., Amin, V., Whitaker, J. W., Schultz, M. D., Ward, L. D., Sarkar, A., Quon, G., Sandstrom, R. S., Eaton, M. L., Wu, Y.-C., Pfenning, A. R., Wang, X., Claussnitzer, M., Liu, Y., Coarfa, C., Harris, R. A., Shoresh, N., Epstein, C. B., Gjoneska, E., Leung, D., Xie, W., Hawkins, R. D., Lister, R., Hong, C., Gascard, P., Mungall, A. J., Moore, R., Chuah, E., Tam, A., Canfield, T. K., Hansen, R. S., Kaul, R., Sabo, P. J., Bansal, M. S., Carles, A., Dixon, J. R., Farh, K.-H., Feizi, S., Karlic, R., Kim, A.-R., Kulkarni, A., Li, D., Lowdon, R., Elliott, G., Mercer, T. R., Neph, S. J., Onuchic, V., Polak, P., Rajagopal, N., Ray, P., Sallari, R. C., Siebenthall, K. T., Sinnott-Armstrong, N. A., Stevens, M., Thurman, R. E., Wu, J., Zhang, B., Zhou, X., Beaudet, A. E., Boyer, L. A., De Jager, P. L., Farnham, P. J., Fisher, S. J., Haussler, D., Jones, S. J. M., Li, W., Marra, M. A., McManus, M. T., Sunyaev, S., Thomson, J. A., Tlsty, T. D., Tsai, L.-H., Wang, W., Waterland, R. A., Zhang, M. Q., Chadwick, L. H., Bernstein, B. E., Costello, J. F., Ecker, J. R., Hirst, M., Meissner, A., Milosavljevic, A., Ren, B., Stamatoyannopoulos, J. A., Wang, T., and Kellis, M. (2015). Integrative analysis of 111 reference human epigenomes. *Nature*, 518(7539):317–330.

[Robinson et al., 2014] Robinson, M. D., Kahraman, A., Law, C. W., Lindsay, H., Nowicka, M., Weber, L. M., and Zhou, X. (2014). Statistical methods for detecting differentially methylated loci and regions. *Front Genet*, 5:324.

[Roudier et al., 2011] Roudier, F., Ahmed, I., Bérard, C., Sarazin, A., Mary-Huard, T., Cortijo, S., Bouyer, D., Caillieux, E., Duvernois-Berthet, E., Al-Shikhley, L., Giraut, L., Després, B., Drevensek, S., Barneche, F., Dèrozier, S., Brunaud, V., Aubourg, S., Schnittger, A., Bowler, C., Martin-Magniette, M.-L., Robin, S., Caboche, M., and Colot, V. (2011). Integrative epigenomic mapping defines four main chromatin states in Arabidopsis. *EMBO J*, 30(10):1928–1938.

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

[Saxena et al., 2014] Saxena, R. K., Edwards, D., and Varshney, R. K. (2014). Structural variations in plant genomes. *Brief Funct Genomics*, 13(4):296–307.

[Saze et al., 2008] Saze, H., Shiraishi, A., Miura, A., and Kakutani, T. (2008). Control of genic dna methylation by a jmjc domain-containing protein in arabidopsis thaliana. *Science*, 319(5862):462–465.

[Schmitz et al., 2013a] Schmitz, R. J., He, Y., Valdés-López, O., Khan, S. M., Joshi, T., Urich, M. A.,

Nery, J. R., Diers, B., Xu, D., Stacey, G., and Ecker, J. R. (2013a). Epigenome-wide inheritance of cytosine methylation variants in a recombinant inbred population. *Genome Res*, 23(10):1663–1674.

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

[Schmitz et al., 2013b] Schmitz, R. J., Schultz, M. D., Urich, M. A., Nery, J. R., Pelizzola, M., Libiger, O., Alix, A., McCosh, R. B., Chen, H., Schork, N. J., and Ecker, J. R. (2013b). Patterns of population epigenomic diversity. *Nature*, 495(7440):193–198.

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

[Schulz et al., 2013] Schulz, B., Eckstein, R. L., and Durka, W. (2013). Scoring and analysis of methylation-sensitive amplification polymorphisms for epigenetic population studies. *Mol Ecol Resour*, 13(4):642–653.

[Serre et al., 2010] Serre, D., Lee, B. H., and Ting, A. H. (2010). MBD-isolated Genome Sequencing provides a high-throughput and comprehensive survey of DNA methylation in the human genome. *Nucleic Acids Res*, 38(2):391–399.

[Seymour et al., 2014] Seymour, D. K., Koenig, D., Hagmann, J., Becker, C., and Weigel, D. (2014). Evolution of DNA Methylation Patterns in the Brassicaceae is Driven by Differences in Genome Organization. *PLoS Genet*, 10(11):e1004785.

[Shang et al., 2014] Shang, J., Zhu, F., Vongsangnak, W., Tang, Y., Zhang, W., and Shen, B. (2014). Evaluation and comparison of multiple aligners for next-generation sequencing data analysis. *Biomed Res Int*, 2014:309650.

[She and Baroux, 2014] She, W. and Baroux, C. (2014). Chromatin dynamics during plant sexual reproduction. *Front Plant Sci*, 5:354.

[Shen et al., 2014] Shen, X., De Jonge, J., Forsberg, S. K. G., Pettersson, M. E., Sheng, Z., Hennig, L., and Carlborg, O. (2014). Natural CMT2 Variation Is Associated With Genome-Wide Methylation Changes and Temperature Seasonality. *PLoS Genet*, 10(12):e1004842.

[Sherry et al., 2001] Sherry, S. T., Ward, M. H., Kholodov, M., Baker, J., Phan, L., Smigielski, E. M., and Sirotkin, K. (2001). dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res*, 29(1):308–311.

[Shivaprasad et al., 2012] Shivaprasad, P. V., Dunn, R. M., Santos, B. A., Bassett, A., and Baulcombe, D. C. (2012). Extraordinary transgressive phenotypes of hybrid tomato are influenced by epigenetics and small silencing RNAs. *EMBO J*, 31(2):257–266.

[Silveira et al., 2013] Silveira, A. B., Trontin, C., Cortijo, S., Barau, J., Del Bem, L. E. V., Loudet, O., Colot, V., and Vincentz, M. (2013). Extensive natural epigenetic variation at a de novo originated gene. *PLoS Genet*, 9(4):e1003437.

[Slotkin et al., 2009] Slotkin, R. K., Vaughn, M., Borges, F., Tanurdzić, M., Becker, J. D., Feijó, J. A., and Martienssen, R. A. (2009). Epigenetic reprogramming and small RNA silencing of transposable elements in pollen. *Cell*, 136(3):461–472.

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

[Storey and Tibshirani, 2003] Storey, J. D. and Tibshirani, R. (2003). Statistical significance for genomewide studies. *Proc Natl Acad Sci U S A*, 100(16):9440–9445.

[Stroud et al., 2013a] Stroud, H., Ding, B., Simon, S. A., Feng, S., Bellizzi, M., Pellegrini, M., Wang, G.-L., Meyers, B. C., and Jacobsen, S. E. (2013a). Plants regenerated from tissue culture contain stable epigenome changes in rice. *Elife*, 2:e00354.

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

[Sun et al., 2014] Sun, D., Xi, Y., Rodriguez, B., Park, H. J., Tong, P., Meong, M., Goodell, M. A.,

and Li, W. (2014). MOABS: model based analysis of bisulfite sequencing data. *Genome Biol*, 15(2):R38.

[Susiarjo and Bartolomei, 2014] Susiarjo, M. and Bartolomei, M. S. (2014). Epigenetics. You are what you eat, but what about your DNA? *Science*, 345(6198):733–734.

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

[Takai and Jones, 2002] Takai, D. and Jones, P. A. (2002). Comprehensive analysis of CpG islands in human chromosomes 21 and 22. *Proc Natl Acad Sci U S A*, 99(6):3740–3745.

[Takuno and Gaut, 2012] Takuno, S. and Gaut, B. S. (2012). Body-methylated genes in Arabidopsis thaliana are functionally important and evolve slowly. *Mol Biol Evol*, 29(1):219–227.

[Takuno and Gaut, 2013] Takuno, S. and Gaut, B. S. (2013). Gene body methylation is conserved between plant orthologs and is of evolutionary consequence. *Proc Natl Acad Sci U S A*, 110(5):1797–1802.

[Talbert and Henikoff, 2014] Talbert, P. B. and Henikoff, S. (2014). Environmental responses mediated by histone variants. *Trends Cell Biol*, 24(11):642–650.

[Tariq et al., 2003] Tariq, M., Saze, H., Probst, A. V., Lichota, J., Habu, Y., and Paszkowski, J. (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.

[Teixeira et al., 2009] Teixeira, F. K., Heredia, F., Sarazin, A., Roudier, F., Boccara, M., Ciaudo, C., Cruaud, C., Poulain, J., Berdasco, M., Fraga, M. F., Voinnet, O., Wincker, P., Esteller, M., and Colot, V. (2009). A role for RNAi in the selective correction of DNA methylation defects. *Science*, 323(5921):1600–1604.

[Telias et al., 2011] Telias, A., Lin-Wang, K., Stevenson, D. E., Cooney, J. M., Hellens, R. P., Allan, A. C., Hoover, E. E., and Bradeen, J. M. (2011). Apple skin patterning is associated with differential expression of MYB10. *BMC Plant Biol*, 11:93.

[Tricker et al., 2012] Tricker, P. J., Gibbings, J. G., Rodríguez López, C. M., Hadley, P., and Wilkinson, M. J. (2012). Low relative humidity triggers RNA-directed de novo DNA methylation and suppression of genes controlling stomatal development. *J Exp Bot*, 63(10):3799–3813.

[van Dijk et al., 2014] van Dijk, E. L., Auger, H., Jaszczyszyn, Y., and Thermes, C. (2014). Ten years of next-generation sequencing technology. *Trends Genet*, 30(9):418–426.

[Vaughn et al., 2007] Vaughn, M. W., Tanurdzić, M., Lippman, Z., Jiang, H., Carrasquillo, R., Rabinowicz, P. D., Dedhia, N., McCombie, W. R., Agier, N., Bulski, A., Colot, V., Doerge, R. W., and Martienssen, R. A. (2007). Epigenetic natural variation in Arabidopsis thaliana. *PLoS Biol*, 5(7):e174.

[Venter et al., 2001] Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., Smith, H. O., Yandell, M., Evans, C. A., Holt, R. A., Gocayne, J. D., Amanatides, P., Ballew, R. M., Huson, D. H., Wortman, J. R., Zhang, Q., Kodira, C. D., Zheng, X. H., Chen, L., Skupski, M., Subramanian, G., Thomas, P. D., Zhang, J., Gabor Miklos, G. L., Nelson, C., Broder, S., Clark, A. G., Nadeau, J., McKusick, V. A., Zinder, N., Levine, A. J., Roberts, R. J., Simon, M., Slayman, C., Hunkapiller, M., Bolanos, R., Delcher, A., Dew, I., Fasulo, D., Flanigan, M., Florea, L., Halpern, A., Hannenhalli, S., Kravitz, S., Levy, S., Mobarry, C., Reinert, K., Remington, K., Abu-Threideh, J., Beasley, E., Biddick, K., Bonazzi, V., Brandon, R., Cargill, M., Chandramouliswaran, I., Charlab, R., Chaturvedi, K., Deng, Z., Di Francesco, V., Dunn, P., Eilbeck, K., Evangelista, C., Gabrielian, A. E., Gan, W., Ge, W., Gong, F., Gu, Z., Guan, P., Heiman, T. J., Higgins, M. E., Ji, R. R., Ke, Z., Ketchum, K. A., Lai, Z., Lei, Y., Li, Z., Li, J., Liang, Y., Lin, X., Lu, F., Merkulov, G. V., Milshina, N., Moore, H. M., Naik, A. K., Narayan, V. A., Neelam, B., Nusskern, D., Rusch, D. B., Salzberg, S., Shao, W., Shue, B., Sun, J., Wang, Z., Wang, A., Wang, X., Wang, J., Wei, M., Wides, R., Xiao, C., Yan, C., Yao, A., Ye, J., Zhan, M., Zhang, W., Zhang, H., Zhao, Q., Zheng, L., Zhong, F., Zhong, W., Zhu, S., Zhao, S., Gilbert, D., Baumhueter, S., Spier, G., Carter, C., Cravchik, A., Woodage, T., Ali, F., An, H., Awe, A., Baldwin, D., Baden, H., Barnstead, M., Barrow, I., Beeson, K., Busam, D., Carver, A., Center, A., Cheng, M. L., Curry, L., Danaher, S., Davenport, L., Desilets, R., Dietz, S., Dodson, K., Doup,

L., Ferriera, S., Garg, N., Gluecksmann, A., Hart, B., Haynes, J., Haynes, C., Heiner, C., Hladun, S., Hostin, D., Houck, J., Howland, T., Ibegwam, C., Johnson, J., Kalush, F., Kline, L., Koduru, S., Love, A., Mann, F., May, D., McCawley, S., McIntosh, T., McMullen, I., Moy, M., Moy, L., Murphy, B., Nelson, K., Pfannkoch, C., Pratts, E., Puri, V., Qureshi, H., Reardon, M., Rodriguez, R., Rogers, Y. H., Romblad, D., Ruhfel, B., Scott, R., Sitter, C., Smallwood, M., Stewart, E., Strong, R., Suh, E., Thomas, R., Tint, N. N., Tse, S., Vech, C., Wang, G., Wetter, J., Williams, S., Williams, M., Windsor, S., Winn-Deen, E., Wolfe, K., Zaveri, J., Zaveri, K., Abril, J. F., Guigó, R., Campbell, M. J., Sjolander, K. V., Karlak, B., Kejariwal, A., Mi, H., Lazareva, B., Hatton, T., Narechania, A., Diemer, K., Muruganujan, A., Guo, N., Sato, S., Bafna, V., Istrail, S., Lippert, R., Schwartz, R., Walenz, B., Yooseph, S., Allen, D., Basu, A., Baxendale, J., Blick, L., Caminha, M., Carnes-Stine, J., Caulk, P., Chiang, Y. H., Coyne, M., Dahlke, C., Mays, A., Dombroski, M., Donnelly, M., Ely, D., Esparham, S., Fosler, C., Gire, H., Glanowski, S., Glasser, K., Glodek, A., Gorokhov, M., Graham, K., Gropman, B., Harris, M., Heil, J., Henderson, S., Hoover, J., Jennings, D., Jordan, C., Jordan, J., Kasha, J., Kagan, L., Kraft, C., Levitsky, A., Lewis, M., Liu, X., Lopez, J., Ma, D., Majoros, W., McDaniel, J., Murphy, S., Newman, M., Nguyen, T., Nguyen, N., Nodell, M., Pan, S., Peck, J., Peterson, M., Rowe, W., Sanders, R., Scott, J., Simpson, M., Smith, T., Sprague, A., Stockwell, T., Turner, R., Venter, E., Wang, M., Wen, M., Wu, D., Wu, M., Xia, A., Zandieh, A., and Zhu, X. (2001). The sequence of the human genome. *Science*, 291(5507):1304–1351.

[Verhoeven et al., 2010] Verhoeven, K. J. F., Jansen, J. J., van Dijk, P. J., and Biere, A. (2010). Stress-induced DNA methylation changes and their heritability in asexual dandelions. *New Phytol*, 185(4):1108–1118.

[Visscher et al., 2012] Visscher, P. M., Brown, M. A., McCarthy, M. I., and Yang, J. (2012). Five years of GWAS discovery. *Am J Hum Genet*, 90(1):7–24.

[Waddington, 1956] Waddington, C. (1956). Embryology, epigenetics and biogenetics. *Nature*, 177(1241).

[Wang et al., 2015] Wang, C., Liu, C., Roqueiro, D., Grimm, D., Schwab, R., Becker, C., Lanz, C., and Weigel, D. (2015). Genome-wide analysis of local chromatin packing in Arabidopsis thaliana. *Genome Res*, 25(2):246–256.

[Wang et al., 2013] Wang, X., Weigel, D., and Smith, L. M. (2013). Transposon variants and their effects on gene expression in Arabidopsis. *PLoS Genet*, 9(2):e1003255.

[Wang et al., 2014] Wang, Y., Yang, Q., and Wang, Z. (2014). The evolution of nanopore sequencing. *Front Genet*, 5:449.

[Weigel and Mott, 2009] Weigel, D. and Mott, R. (2009). The 1001 genomes project for Arabidopsis thaliana. *Genome Biol*, 10(5):107.

[Weischenfeldt et al., 2013] Weischenfeldt, J., Symmons, O., Spitz, F., and Korbel, J. O. (2013). Phenotypic impact of genomic structural variation: insights from and for human disease. *Nat Rev Genet*, 14(2):125–138.

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

[Willing et al., 2015] Willing, E.-M., Rawat, V., Mandáková, T., Maumus, F., James, G. V., Nordström, K. J., Becker, C., Warthmann, N., Chica, C., Szarzynska, B., Zytnicki, M., Albani, M. C., Kiefer, C., Bergonzi, S., Castaings, L., Mateos, J. L., Berns, M. C., Bujdoso, N., Piofczyk, T., de Lorenzo, L., Barrero-Sicilia, C., Mateos, I., Piednoël, M., Hagmann, J., Chen-Min-Tao, R., Iglesias-Fernández, R., Schuster, S. C., Alonso-Blanco, C., Roudier, F., Carbonero, P., Paz-Ares, J., Davis, S. J., Pecinka, A., Quesneville, H., Colot, V., Lysak, M. A., Weigel, D., Coupland, G., and Schneeberger, K. (2015). Genome expansion of Arabis alpina linked with retrotransposition and reduced symmetric DNA methylation. *Nature Plants*, 1:14023.

[Wong et al., 2010] Wong, K., Keane, T. M., Stalker, J., and Adams, D. J. (2010). Enhanced structural variant and breakpoint detection using SVMerge by integration of multiple detection methods and local assembly. *Genome Biol*, 11(12):R128.

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

[Yan et al., 2010] Yan, M. S.-C., Matouk, C. C., and Marsden, P. A. (2010). Epigenetics of the vascular endothelium. *J Appl Physiol (1985)*, 109(3):916–926.

[Yao, 2014] Yao, W. (2014). intansv: Integrative analysis of structural variations. *R package version 1.8.0*.

[Ye et al., 2009] Ye, K., Schulz, M. H., Long, Q., Apweiler, R., and Ning, Z. (2009). Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics*, 25(21):2865–2871.

[Yoshida et al., 2013] Yoshida, K., Schuenemann, V. J., Cano, L. M., Pais, M., Mishra, B., Sharma, R., Lanz, C., Martin, F. N., Kamoun, S., Krause, J., Thines, M., Weigel, D., and Burbano, H. A. (2013). The rise and fall of the Phytophthora infestans lineage that triggered the Irish potato famine. *Elife*, 2:e00731.

[Yu et al., 2013] Yu, A., Lepère, G., Jay, F., Wang, J., Bapaume, L., Wang, Y., Abraham, A.-L., Penterman, J., Fischer, R. L., Voinnet, O., and Navarro, L. (2013). Dynamics and biological relevance of DNA demethylation in Arabidopsis antibacterial defense. *Proc Natl Acad Sci U S A*, 110(6):2389–2394.

[Yu et al., 2014] Yu, W., McIntosh, C., Lister, R., Zhu, I., Han, Y., Ren, J., Landsman, D., Lee, E., Briones, V., Terashima, M., Leighty, R., Ecker, J. R., and Muegge, K. (2014). Genome-wide DNA methylation patterns in LSH mutant reveals de-repression of repeat elements and redundant epigenetic silencing pathways. *Genome Res*, 24(10):1613–1623.

[Zemach et al., 2013] Zemach, A., Kim, M. Y., Hsieh, P.-H., Coleman-Derr, D., Eshed-Williams, L., Thao, K., Harmer, S. L., and Zilberman, D. (2013). The Arabidopsis nucleosome remodeler DDM1 allows DNA methyltransferases to access H1-containing heterochromatin. *Cell*, 153(1):193–205.

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

[Zeng and Cheng, 2014] Zeng, F. and Cheng, B. (2014). Transposable Element Insertion and Epigenetic Modification Cause the Multiallelic Variation in the Expression of FAE1 in Sinapis alba. *Plant Cell*, 26(6):2648–2659.

[Zerbino and Birney, 2008] Zerbino, D. R. and Birney, E. (2008). Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res*, 18(5):821–829.

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

[Zhang et al., 2008] Zhang, X., Shiu, S.-H., Shiu, S., Cal, A., and Borevitz, J. O. (2008). Global analysis of genetic, epigenetic and transcriptional polymorphisms in Arabidopsis thaliana using whole genome tiling arrays. *PLoS Genet*, 4(3):e1000032.

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

[Zilberman et al., 2008] Zilberman, D., Coleman-Derr, D., Ballinger, T., and Henikoff, S. (2008). Histone H2a.Z and DNA methylation are mutually antagonistic chromatin marks. *Nature*, 456(7218):125–129.

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

[Ziller et al., 2013] Ziller, M. J., Gu, H., Müller, F., Donaghey, J., Tsai, L. T.-Y., Kohlbacher, O., De Jager, P. L., Rosen, E. D., Bennett, D. A., Bernstein, B. E., Gnirke, A., and Meissner, A. (2013). Charting a dynamic DNA methylation landscape of the human genome. *Nature*, 500(7463):477–481.