

Measuring stress-induced DNA methylation in apomictic Dandelions

Thomas P. van Gorp

Thesis committee

Promotor

Prof. Dr W.H. van der Putten
Special Professor Functional Biodiversity
Wageningen University & Research

Co-promotors

Dr K.J.F. Verhoeven
Senior scientist
Netherlands Institute of Ecology, Wageningen

Dr A. Biere
Senior scientist
Netherlands Institute of Ecology, Wageningen

Other members

Prof. Dr F. Johannes, TUM School of Life Sciences, Germany
Prof. Dr M.E. Schranz, Wageningen University & Research
Prof. Dr D. de Ridder, Wageningen University & Research
Dr K. Kraaijeveld, VU Amsterdam

This research was conducted under the auspices of the Graduate School for
Experimental Plant Sciences (EPS)

Measuring stress-induced DNA methylation in apomictic Dandelions

Thomas P. van Gurp

Thesis

submitted in fulfillment of the requirements for the degree of doctor
at Wageningen University
by the authority of the Rector Magnificus,
Prof. Dr A.P.J. Mol,
in the presence of the
Thesis Committee appointed by the Academic Board
to be defended in public
on Monday 9th of October 2017
at 4 p.m. in the Aula.

Thomas P. van Gorp
Measuring stress-induced DNA methylation in apomictic Dandelions,
(to be included) pages.

PhD thesis, Wageningen University, Wageningen, the Netherlands (2017)
With references, with summaries in Dutch and English

ISBN: 978-94-6343-604-5

DOI: <https://doi.org/10.18174/417596>

Voor Tatiana, voor al je hulp en geduld

Content

Chapter 1	General introduction	7
Chapter 2	Consistent Errors in First Strand cDNA Due to Random Hexamer Mispriming	21
Chapter 3	epiGBS: reference-free reduced representation bisulfite sequencing	29
Chapter 4	Transgenerational inheritance of (stress-induced) DNA methylation in apomictic dandelions.	69
Chapter 5	Selection on early flowering in asexually reproducing Dandelions; using epigenetic variation as a substrate for artificial selection	103
Chapter 6	General discussion	137
	References	148
	Summary	162
	Samenvatting (Dutch)	165
	Acknowledgements / dankwoord	169
	About the author	172
	Education statement	174
	Propositions	177

General Introduction

1

1.1 Evolution, what mechanisms facilitate inheritance?

The mechanism of inheritance is a crucial component of the model of evolution. Inheritance and evolution have been studied for a long time. Charles Darwin made already many observations on how organisms passed their characters from one generation to the next, recognizing that evolution occurs through natural selection on heritable variation that shapes the phenotype, resulting in his “*On the origin of species by means of natural selection*” in 1859, laying the foundation for evolutionary biology. Jean-Baptiste Lamarck published his “*Philosophie Zoologique*” fifty years earlier in 1809, in which he detailed his theory of inheritance of acquired characteristics, called ‘soft inheritance’. Lamarck emphasized that the environment gives rise to changes in animals, explaining why moles have lost their sight and why teeth are present in mammals but absent in birds (Osborn 1896). Lamarck’s ideas were criticized during his lifetime by Lawrence and later by Weismann, who cut off the tails of 22 subsequent generations of mice, proving that the loss of tail is not inherited (I. Kovalchuk & O. Kovalchuk 2012). Darwin speculated on the mechanism that allowed for the inheritance of acquired characters. Lamarck did not propose an actual mechanism of inheritance, but suggested selective inheritance, stating that characters not used would disappear and thus not be transferred whereas important traits ‘practiced’ during the lifetime would be transferred and even enhanced in future generations. Darwin proposed the existence of ‘*gemmules*’, which were proposed to be minute particles shed by cells of the whole body that could be carried by the bloodstream to the reproductive organs where they accumulated in gametes.

1.2 Historic understanding of mechanism of inheritance

Better insight in the actual mechanism of inheritance was provided by the (rediscovery of the) work of Mendel. Gregor Mendel established the first explicit laws of heredity, and published his work on invisible “factors”, now called genes, in 1866, creating the modern science of genetics. It took until the turn of the 20th century before Mendel’s work was rediscovered independently by three botanists (Bowler 1989). The unification of Darwin’s theory of natural selection and Mendel’s laws of genetics by Ronald Fisher and others gave birth to the Modern evolutionary synthesis, the current paradigm in evolutionary biology (Mayr 2001), reflecting the consensus on how evolution works (Mayr & Provine 1998). Some of the major tenets of the this synthesis are that: **A)** populations have genetic variation that continuously arises by undirected processes such as mutation and recombination and **B)** populations evolve by changes in gene frequencies through drift, gene flow and natural selection (Futuyma 1986).

1.3 Gene-only model is too narrow and excludes epigenetic inheritance

The discovery of DNA and genes, and mathematical conclusions drawn on shared ancestry based on the similarity of the DNA sequence between related species led to the development of a narrow, gene-centered view on inheritance in the extended modern synthesis that excluded non-genetic inheritance (Jablonka et al. 2014). Genetic variation is thought to be the product of random mutations produced at a relatively constant rate that accumulate through positive selection or genetic drift (Takahata 1996). Although well-founded and mostly accurate, this gene centric paradigm explicitly excludes mechanisms of non-genetic inheritance such as the

inheritance of acquired characters: the only relevant information that is passed on between generations is the information that is encoded in the DNA. However, it has become clear that this is not always strictly true. Several forms of non-genetic inheritance are now known to affect heritable traits, also in absence of DNA sequence differences (Bonduriansky & Day 2009). Among the different pathways of non-genetic inheritance (Jablonka et al. 2014), epigenetic mechanisms such as DNA methylation and histone modifications have received increasing interest for their potential to affect heritable trait variation.

Such epigenetic mechanisms, which are chemical modifications of the DNA that can have stable but reversible effects on gene expression, have been demonstrated to generate heritable phenotypic variation that can be independent of the DNA sequence. In plants, several well-defined examples of epigenetic variation controlling a phenotype are described. One of the most striking examples of transgenerationally stable epigenetic variation determining the phenotype is found in the plant species *Linaria vulgaris*. Mutant flowers with radial instead of bilateral symmetry are caused by extensive methylation of the *Lcyc* gene controlling dorsoventral symmetry, rendering it transcriptionally silent (Cubas et al. 1999). Other such epialleles are found in *Zea mays* in which altered pigmentation patterns are brought about by altered DNA methylation states resulting from paramutation (Kalisz & Purugganan 2004), an interaction between two alleles at a single locus, whereby one allele induces a heritable change in the other allele (Chandler 2007).

In contrast to the tenets of the modern evolutionary synthesis, epigenetic variation can be rapidly generated in response to environmental stress (C. Jiang et al. 2014; Wibowo et al. 2016; Bruce et al. 2007), sometimes facilitating (transgenerational) adaptation (Mirouze & Paszkowski 2011). Environmental stress in parental plants can result in drastic phenotypic changes in stress-exposed offspring, which can be adaptive. Drought stressed *Polygonum persicaria* plants produce offspring with longer and more rapidly extending roots when grown in dry soil in comparison to offspring plants from non-stressed parents (Sultan et al. 2009), which is mediated by DNA methylation (Herman & Sultan 2016). Stress exposed plants face a balancing act. While stress induced plant responses are essential for survival, continued expression of stress responses under non-stress conditions might hinder recovery whereas repeated stress exposure would warrant a longer response (Crisp et al. 2016). These aspects make the study of environment-induced epigenetic variation very interesting as such heritable variation could affect our understanding of evolution. However, many important questions remain. The extent to which epigenetic variation is environmentally induced or simply determined by the genotype and the transgenerational stability of this epigenetic variation as well as its impact on adaptation are important open questions in ecological epigenetics (Verhoeven et al. 2016).

1.4 Epigenetic variation, relevance for transgenerational adaptation

The discovery and detailed characterization of epigenetic mechanisms regulating gene expression, or even determining the phenotype, by means of modifications on, and not in, the DNA sequence has fueled excitement on the potential impact of epigenetic variation on evolution (Jablonka & Lamb 1998; Jablonka et al. 2014). From an ecological and evolutionary perspective, epigenetic variation is highly

interesting as it could significantly improve our understanding of the mechanisms underlying natural phenotypic variation and the responses of organisms to environmental change (Bossdorf et al. 2007). Contrary to the tenets of the modern synthesis, clear evidence for the transmission of environmentally induced epigenetic variation is available (Verhoeven, Jansen, et al. 2009; Anway et al. 2005; Wibowo et al. 2016). Transgenerational phenotypic plasticity mediated by epigenetic variation is of ecological relevance, as it can increase offspring fitness but it does not necessarily influence adaptation or evolution. For epigenetic variation to shape the phenotype over longer periods of time and act as a source of variation upon which selection can act, it needs to be maintained for more than one generation. Selection acts on the phenotype, which is influenced by the genotype, epigenotype and environment (C. Richards & al 2010). Heritable epigenetic variation thus has a potential role in being subject to selection facilitating adaptation and ultimately evolution. An evolutionary model that includes the potential of intergenerational transfer of environment-induced changes is fundamentally different from a model based on only genetic changes that excludes such epigenetic effects, leading to different dynamics in simulation studies. Modelling studies that include epimutations suggest that the dynamics of genetic adaptation are significantly affected by such epimutations, even if the reversal rate of epimutations are high (Verhoeven et al. 2016). Heritable and transgenerationally stable epigenetic variation that is at least partially independent of genetic variation (E. J. Richards 2006) thus needs to be taken into account as a potential mechanism fueling (micro)evolution, for which theoretical models predict faster adaptation when epigenetic variation is included (Klironomos et al. 2013).

1.5 Epigenetic mechanisms

The most important and well-studied mechanisms that allow for changes in gene expression in the absence of DNA sequence changes are histone protein modifications, small RNAs and methylation of cytosines in DNA to form 5-methylcytosine (DNA methylation).

1.5.1 Histone modifications

Histone proteins form the core around which DNA is wrapped. Chemical modifications of histone proteins can lead to changes in the chromatin structure, inhibiting or enhancing gene expression (C. L. Peterson & Laniel 2004). The heritability of such histone modifications through mitotic and meiotic divisions is the subject of ongoing debate and research (Allis & Jenuwein 2016), mainly because mechanisms that would allow for such meiotic inheritance are poorly characterized. Recently, evidence was provided for transmission of H3K9 methylation through many mitotic but also meiotic divisions in yeast, showing that transgenerational inheritance of changed chromatin states is possible (Audergon et al. 2015). Similar results were obtained for repressive chromatin marks (H3K27) in the nematode *Caenorhabditis elegans* (Gaydos et al. 2014). Taken together, these results suggest that a detailed characterization of the molecular machinery involved in the incorporation of histones during or after DNA replication might provide more evidence for transgenerational transmission of histone modifications.

1.5.2 Small RNA

Small RNAs interact with, and presumably direct, chromatin modifying activities to genomic targets (Allis & Jenuwein 2016). RNA interference, involving sequence-specific gene regulation by small non-coding RNAs can inhibit the translation of genes, enabling post transcriptional gene silencing (Kamthan et al. 2015). In plants, transposable elements are silenced by small interfering RNA (siRNA) directing DNA methylation, in which small RNAs target the de novo DNA methylation at genomic loci homologous to the RNA sequence (Xie & Yu 2015). Transgenerational inheritance of such small RNAs has been demonstrated conclusively in *C. elegans*, whereas in plants maternally inherited 24nt small RNAs direct methylation of transposable elements derived from the pollen in developing plant seeds through an intriguing mechanism (Sarkies & Miska 2014). During pollen development, some DNA methylation is lost in the vegetative cells formed adjacent to developing generative cells within the pollen of angiosperms, triggering small RNA production that can move towards the germ cells, which is thought to reinforce silencing of TEs in gametes through RNA directed DNA methylation (Slotkin et al. 2009). Small RNAs generated during gametophyte development are thus capable of influencing the epigenetic profile of the developing seeds and as such can constitute a mechanism of transgenerational epigenetic inheritance in conjunction with other epigenetic marks such as DNA methylation.

1.5.3 DNA methylation

DNA methylation is the best-studied epigenetic mechanism that has shown high stability through both mitotic and meiotic divisions in plants. Although DNA methylation can occur in both adenine and cytosine it commonly refers to 5-methylcytosines, that can occur in different sequence contexts, CG, CHG and CHH (where H is any base besides G). Contrary to vertebrates, where DNA methylation occurs almost exclusively in CG context, plants have more diverse DNA methylation patterns in CG, CHG and CHH context, which are regulated by different enzymes and have a sequence-context dependent function (Law & Jacobsen 2010). Centromeric and pericentromeric regions in plant genomes, as well as other repetitive elements are heavily methylated (He et al. 2011). DNA methylation plays an important role in the silencing of transposable elements (TEs), with heavy methylation in all three sequence context preventing TEs from becoming active, preventing their often mutagenic activity (Ikeda & Nishimura 2015). High levels of CG methylation are often found in gene bodies of angiosperms, with moderately transcribed genes being most likely to be methylated, whereas low or highly expressed genes often lack gene body methylation (Zilberman et al. 2007). Complete loss of DNA methylation was observed for *Eutrema salsugineum*, suggesting it is non-essential or that redundant mechanisms for controlling gene activity exist (Bewick et al. 2016). In gene bodies, exons and especially splice sites have higher levels of DNA methylation than flanking introns and the splicing of about 22% of alternative exons is regulated by DNA methylation (Lev Maor et al. 2015).

1.6 Why DNA methylation and no other mechanisms in this thesis?

The field of epigenetics is characterized by an ever-increasing number of exciting new mechanisms that could all contribute to epigenetic inheritance (Plongthongkum et al. 2014; Allis & Jenuwein 2016). New discoveries point at a highly interconnected and increasingly complex network of epigenetic mechanisms, that all warrant

further study (Allis & Jenuwein 2016). Such endeavors however, are dependent on a high-quality reference genome and detailed knowledge of molecular pathways regulating such mechanisms. DNA methylation characterization for non-model systems on the other hand has been enabled by relatively simple techniques, such as methylation sensitive AFLP (MS-AFLP), that allow for limited characterization of DNA methylation variation. More detailed characterization of DNA methylation is enabled by sequencing bisulfite converted DNA. Bisulfite treatment results in conversion of unmethylated cytosines to uracil, resulting in a T base in sequencing. Methylated cytosines are however unchanged, resulting in both reads with Cs and Ts, where the ratio between these two nucleotides represents the average level of methylation for the cytosine under study from all cells in the tissue(s) from which it was sampled. Increasingly, bisulfite sequencing of reduced representation libraries allow for more detailed studies of DNA methylation in non-model systems (van Gurp et al. 2016; Trucchi et al. 2016; Verhoeven et al. 2016), to which I contribute in this thesis (see chapter 3 / van Gurp et al. 2016).

1.7 Why are plants an interesting system to study epigenetic variation?

Studying epigenetics in plants is highly interesting for several reasons. Plants are sessile organisms, which makes them especially vulnerable to environmental stress such as excessive or inadequate light, water, salt and temperature to which they need to adapt (Molinier et al. 2006). Generative cells can be formed continuously in many plants, facilitating transfer of acquired epigenetic changes to the germline, which unlike that of vertebrates, is not set apart from the soma early in development but develops from somatic tissue only during flower development. Gametogenesis and early embryonic development in vertebrates is characterized by two phases of extensive reprogramming of epigenetic marks (Heard & Martienssen 2014), resetting most but not all acquired DNA methylation changes. Plants in contrast have a much more stable inheritance of epigenetic marks, especially DNA methylation (Schmitz, He, et al. 2013; Becker et al. 2011). As mentioned previously, DNA methylation in plants is more complex than in vertebrates, with methylation occurring in three sequence contexts, CG, CHG and CHH (where H is A, C or T) controlled by different enzymes and carrying out different functions. Experimental modification of epigenetic patterns such as DNA methylation in plants is not subject to extensive legislation, as is the case with vertebrates (Lusser et al. 2012). Combined, these properties make plants attractive models to study epigenetic variation in the form of DNA methylation, which is easily characterized in non-model species.

1.8 Epigenetic studies in plants have mainly focused on *Arabidopsis thaliana*.

To understand the contribution of epigenetic variation to the phenotype, it is essential to understand the molecular mechanisms that facilitate this process. To this end, work in model species has taught us a great deal on epigenetic inheritance in the past decades. In plants, most work in epigenetics has focused on the model plant *Arabidopsis thaliana*. *Arabidopsis* rose to prominence as a model plant species for research in genetics (Rédei 1992), mainly due to its short generation time, and small (genome) size, facilitating rapid studies with big cohorts. Research on plant epigenetics has identified DNA methylation as an important regulator for the control of transposable elements (Ikeda & Nishimura 2015) and as a mechanism for gene expression regulation by promoter methylation, although the extent to which DNA

methylation contributes to gene expression variation appears to be limited compared to that of genetic variants in *A. thaliana* (Meng et al. 2016). Micro-array studies employing probes designed to capture and differentiate the bisulfite converted product of both methylated and non-methylated DNA allows for detecting global differences in DNA methylation. It was however not until the advent of whole genome bisulfite sequencing, allowing for single nucleotide resolution DNA methylation studies, that the full magnitude of differential DNA methylation and inheritance were fully understood and appreciated. Many exciting discoveries characterizing a variety of new epigenetic mechanisms and pathways were made in *A. thaliana* over the past decades (A. L. Jones & Sung 2014), but here I only focus on molecular pathways involved in the regulation of DNA methylation, as this thesis exclusively deals with this epigenetic mark.

1.9 Epigenetic landscape of *Arabidopsis thaliana*

Understanding the epigenetic landscape of *A. thaliana* has been enabled by complete sequencing of (mutant of) the methylome and the generation of artificial crosses involving parental plants devoid of DNA methylation. In early 2008, two groups almost simultaneously published the *A. thaliana* methylome, the first single nucleotide resolution plant methylome available (Cokus et al. 2008; Lister & Ecker 2009). Sequencing of both wild-type and mutants for DNA methyltransferases further elucidated the differential role of the different DNA methyltransferases DMT1, CMT3 and DRM1/DRM2 in maintaining DNA methylation in different sequence contexts available (Cokus et al. 2008; Lister & Ecker 2009). Solid evidence for the impact of epigenetic variation independent of the DNA sequence on plant phenotype was achieved by creating epigenetic recombinant inbred lines (epiRILs) in *A. thaliana* (Johannes et al. 2009). The *A. thaliana* epiRILs share the same DNA sequence but have a mosaic pattern of highly modified and wild-type DNA methylation resulting from a cross between a methylation deficient mutant with a ~70% reduction of overall DNA methylation and a wild type (normal) plant. These artificially induced changes in DNA methylation were (largely) inherited faithfully over at least 8 generations, and correlated strongly with a wide array of quantitative phenotypic traits such as flowering time and plant height (Johannes et al. 2009). Detailed genomic analysis involving both DNA methylation and genetic variation in these epiRILs indicated that 90% of the variation in flowering time and primary root length between epiRIL lines was caused by methylation variation (Cortijo et al. 2014). Many of these artificially induced methylation variants were also found in natural populations, suggesting that epigenetic variation might as well play a role in natural variation in flowering time (Cortijo et al. 2014).

1.10 Natural epigenetic variation in *Arabidopsis thaliana*

Natural epigenetic variation in *A. thaliana* has received considerable attention, with several studies documenting DNA methylation variation in wild. Overall, these studies indicate a tight relation between genetic and epigenetic variation, suggesting that much of the natural epigenetic variation observed in *A. thaliana* is (at least partly) determined by genetic variation in cis or trans. In one of the first population epigenomic studies in plants, Schmitz *et al.* found that DNA methylation variation in 155 *A. thaliana* plants with a worldwide distribution resulted in the same clustering as that obtained for genetic variation (Schmitz, Schultz, et al. 2013). Similarly, whole genome bisulfite sequencing studies characterizing progeny plants derived from a

geographically dispersed isogenic *A. thaliana* line in Northern America that diverged under natural conditions over at least a century ago, showed DNA methylation variation that by and large reflected DNA sequence variation. The DNA methylation differences that occurred overlapped with those that were previously detected in mutation accumulation lines of the same genotype grown in a greenhouse, suggesting that environmentally induced changes are only a minor contributor to heritable genome-wide epigenetic variation (Hagmann et al. 2015). In a different study, 150 Swedish *A. thaliana* accessions grown under different temperatures showed an increased level of methylation in CHH context of transposons at higher temperatures. Variation in the level of transposon CHH methylation was considerable, with genetic variation in genes associated with DNA methylation explaining a large extent of this variation. Differences in the level of gene body methylation were however found, with plants from more Nordic ancestors showing higher levels of gene body methylation and expression (Dubin et al. 2015). In the largest methylome sequencing effort till date, the 1001 epigenomes project sequenced 1,107 methylomes and 1,203 transcriptomes of *A. thaliana* derived from a wide variety of geographic origins. DNA methylation and gene expression variation was explained to a large extent by epialleles whose presence correlates with geographic origin. Methylation differences however were mainly found in transposable elements but not gene bodies, and were often accompanied by mutations near genes encoding enzymes involved in the methylation of transposable elements (Kawakatsu et al. 2016).

1.11 Environmental induction of DNA methylation variation

The role of stress and environmental perturbations in causing DNA methylation changes has long been recognized (Peng & Jing Zhang 2008). Studies in *A. thaliana* have indicated specific stress responses in genomic regions associated with differentially expressed genes. *A. thaliana* plants exposed to a bacterial pathogen, avirulent bacteria, or the plant hormone salicylic acid (SA) revealed numerous stress-induced differentially methylated regions, many of which were intimately associated with differentially expressed genes (Downen et al. 2012). Similarly, *A. thaliana* plants grown in a high-salinity soil exhibit an increased frequency of differentially methylated positions (DMPs) in genes (C. Jiang et al. 2014). A more detailed study employing repeated hyperosmotic stress in 5 subsequent generations of *A. thaliana* found improved salt resistance for germinating seedlings of parents exposed to hyperosmotic stress, which led to distinct DNA methylation changes in cytosines in CHG and CHH context and some changes in differentially methylated regions (DMRs) that occurred repeatedly (Wibowo et al. 2016), some of which were found previously (Becker et al. 2011; Hagmann et al. 2015). Interestingly, the unique or new salt-specific DMRs found in this study were often found near genes with functions related to metabolic responses and ion transport (Wibowo et al. 2016), suggesting a plant stress response that has both stress-specific and non-specific effects on DNA methylation.

1.12 From model plants to a wider taxonomic diversity

Studies in the model plant *A. thaliana* have indicated the presence of many exciting patterns but also indicate a high dependence of epigenetic on genetic variation. The enormous diversity in plants species, both in number, genome size and life history traits warrants a further exploration of epigenetic inheritance in a broader context

to determine the generality of and exceptions to the patterns detected in *A. thaliana*. *Arabidopsis thaliana* is an opportunistic annual weed, characterized by a rather simple life cycle. It has therefore long been ignored by ecologists and evolutionists, who deemed the species as ecologically uninteresting and an oddity from an evolutionary point of view (Pigliucci 2002). Recently, 34 plant methylomes covering a wide taxonomic variety have been sequenced (Niederhuth, Bewick, Ji, Alabady, Kim, Li, et al. 2016), uncovering widespread methylome variation, with CHG methylation levels being positively correlated to genome-size, and substantial variation in the amount of CHG and CHH methylation variation found in repeat elements (Gehring 2016). DNA methylation patterns in the Brassicaceae, which includes *A. thaliana*, are different from most angiosperms in having reduced CHG methylation levels and also reduced or lost CG gene body methylation (Niederhuth, Bewick, Ji, Alabady, Kim, Page, et al. 2016). At this point, the generality of the epigenetic contribution to phenotypic variation is not well understood, which warrants further research in a variety of species (Verhoeven et al. 2016). Such studies are facilitated by the increasing number of available reference genomes (Niederhuth, Bewick, Ji, Alabady, Kim, Page, et al. 2016) and the availability of reference-free reduced representation bisulfite sequencing (van Gurp et al. 2016). Epigenetic variation could be particularly important for asexual species. Although considered to be an evolutionary dead end (Lynch et al. 1993), clonal plant species persist and successfully expand in a wide range of different environments (Verhoeven & Preite 2014). The success of individual asexual lineages can be due to preadaptation of native genotypes to specific conditions in the new habitat or due to high phenotypic plasticity (general-purpose genotypes; (Baker 1965)). The persistence and ecological success of such lineages could be facilitated via epigenetic variation as a source of phenotypic plasticity (Verhoeven & Preite 2014). Asexual plant species such as apomictic dandelions therefore offer an interesting model to study epigenetic variation.

1.13 Ecological epigenetics: state of the field

Knowing to what extent epigenetic processes can explain heritable phenotypic variation that does not rely on the DNA sequence in natural populations is pivotal to understand the role of epigenetic variation in evolution. To this end, ecological epigenetics studies investigate the potential impact of epigenetic variation on ecology and evolution in an empirical fashion in ecological systems and natural populations. Ecological epigenetics was loosely defined by (Bossdorf et al. 2007) as the study of epigenetic processes in an ecological context. Epigenetic inheritance is particularly interesting for evolutionary biologists, as it offers a potential explanation for the heritable phenotypic variation observed in natural populations that cannot be explained by DNA sequence variation. It also could explain instances of (transgenerational) phenotypic plasticity, allowing organisms to respond to their environment (Bossdorf et al. 2007). The potential ecological importance of epigenetic variation is greater for plants reproducing asexually, such as apomictic dandelions which cannot rely on shuffling of genes allowed by meiotic recombination (Verhoeven & Preite 2014). Epigenetic variation in asexual plant species can be important both for enabling transgenerational phenotypic plasticity, as well as providing a source of heritable variation upon which (natural) selection can act. Most ecological epigenetics studies are conducted in plants, with the majority of these studies focusing on DNA methylation (Schrey et al. 2013; Kilvitis et

al. 2013). Many ecological epigenetics studies found evidence for DNA methylation-mediated responses to environmental stress (Verhoeven, Jansen, et al. 2009) or geographic differentiation of DNA methylation patterns (Busconi et al. 2015), which can be heritable across generations (Kilvitis et al. 2013).

1.14 Technological innovations needed to study DNA methylation variation.

Until recently, most ecological epigenetics studies employed methylation sensitive AFLPs (MS-AFLP) to determine the extent, structure and stability of DNA methylation variation in natural populations. MS-AFLP enable the distinction between presence and absence of DNA methylation in anonymous genomic loci using a pair of restriction enzymes with differential sensitivity to DNA methylation, resulting in presence-absence combination of gel bands that reflect the methylation status of the locus targeted. MS-AFLP studies offer a limited and incomplete estimate of DNA methylation variation for a number of reasons. First, MS-AFLP studies offer a very limited resolution of 100-200 markers compared to hundreds of thousands of polymorphic loci that can be uncovered using sequencing based approaches. Second, MS-AFLP offer a qualitative measure of what is essentially a quantitative state and do so in cytosine in CG context only. Third, interpretations of different combinations of polymorphic banding patterns to be preferentially associated to methylation in either CG or CHG context are incorrect (van Gorp et al. 2009; Fulneček & Kovarík 2013), leading to a persistent erroneous interpretation in the literature. Sequencing-based studies in *A. thaliana* and other species have uncovered detailed and very specific insights into differential DNA methylation by virtue of whole genome bisulfite sequencing and detailed knowledge of the genome architecture. Reduced representation bisulfite sequencing techniques offering nucleotide level and quantitative DNA methylation estimation while simultaneously allowing for functional characterization of the locus under study would offer a huge improvement over MS-AFLP based studies for non-model systems.

1.15 Research goals in this thesis:

The goal of the research described in this thesis is to empower and perform comprehensive research in ecologically interesting non-model species; focusing on the potential role of natural and stress induced epigenetic variation to be subjected to selection and contribute to adaptation and ultimately evolution. I performed my research in the triploid apomictic dandelion, *Taraxacum officinale*, which is characterized by its asexual reproduction and wide-spread occurrence in diverse habitats. To enable comprehensive characterization of DNA methylation in this non-model species, we developed epiGBS, a bisulfite sequencing extension of the flexible and highly successful genotyping by sequencing (GBS) method. Considerable effort has been spent on the development and validation of a comprehensive bioinformatics pipeline that enables the simultaneous interrogation of DNA methylation and genetic variation using bisulfite sequencing reads only while recreating the original reference sequence from which the bisulfite converted reads were derived. To investigate the heritability of DNA methylation variation in apomictic dandelions and assess the degree to which environmental stress can induce changes in DNA methylation I performed a transgenerational study of stress induced DNA methylation in Dandelion, enabling one of the first comprehensive bisulfite sequencing based study in a non-model organism that gives insight in the transgenerational stability of DNA methylation in apomictic dandelions. Finally, I

perform an artificial selection experiment on early flowering in isogenic apomictic dandelion lineages using both stressed and non-stressed lines to test if there is a response to sustained artificial selection on (induced) epigenetic variation.

1.16 Apomictic dandelions as model system for studying natural epigenetic variation

Apomictic dandelions are characterized by their worldwide distribution in various habitats, with several apomictic lineages showing widespread geographic distribution, seemingly in the absence of genetic recombination. Reproduction in apomictic plant species is characterized by an altered meiosis, which could impact the degree of resetting of epigenetic marks during gametogenesis (Sailer et al. 2016). Apomixis in plants exists in many forms, with each form being characterized by a distinct mode of developmental aberrations during meiosis leading to clonal offspring. Apomixis in dandelions is the result of a combination of diplospory, parthenogenesis and autonomous endosperm development of the seed. New apomictic dandelions are readily generated by crossing a triploid apomict (contributing diploid pollen) with a sexual diploid. Such crosses are known to give rise to natural epigenetic variation (Verhoeven, van Dijk, et al. 2009), upon which selection can act. As such, apomictic dandelions form an excellent model system to investigate the role of (natural) epigenetic variation in generating phenotypic variation as well as the way selection can act upon such variation.

1.17 Specific research questions addressed in this thesis per chapter

Chapter 2: Consistent errors in first strand cDNA due to random hexamer mispriming. Working with Next Generation Sequencing (NGS) data is still relatively new and therefore not always optimized for specific types of analysis. Here, I focus on one common error type, that I characterized in detail. While working on RNAseq data of a collaborator, a distinct error pattern was discovered in the data that indicated an artificial source of variation that could be traced back to the way the RNAseq libraries were constructed. One study stating that these errors were the results of RNA editing (M. Li et al. 2011), claimed a disruption in the information flow between a gene and its product, however, this proposal was heavily criticized (Lin et al. 2012; Pickrell et al. 2012; Kleinman & Majewski 2012) as the positions in the sequencing reads in which these supposed editing events took place were biased toward the first nucleotides of the reads. In our publication (van Gurp et al. 2013), we determined that this bias is caused by mispriming of random DNA hexamers to an RNA template during library construction (M. Li et al. 2011). Another misinterpretation, although not regularly leading to such high-profile publications, is made by many ecological epigeneticists who incorrectly assume the MS-AFLP variation they encounter to represent CHG methylation, due to an insufficient understanding of the technique and the DNA methylation profiles commonly found in plants (Fulneček & Kovarik 2013). Thorough knowledge of all the potential biases generated in the protocols that we use to generate (sequencing) data is essential for arriving at the right conclusions, being an ecologist can never be an excuse for not making the considerable effort required to (fully) understand the methods used to generate the data based on which we draw conclusions.

Chapter 3: epiGBS: reference-free reduced representation bisulfite sequencing

This chapter is the result of the quest for a quantitative bisulfite sequencing based method for measuring DNA methylation differences in species without a reference genome. This chapter addresses the challenges that were found and overwon for performing such research by adapting existing methods and developing new bioinformatics algorithms. In this publication, I describe epiGBS, a reduced representation bisulfite sequencing method for cost-effective exploration and comparative analysis of DNA methylation and genetic variation in hundreds of samples de novo. This method uses genotyping by sequencing of bisulfite converted DNA followed by reliable de novo reference construction, mapping, variant calling, and distinction of single-nucleotide polymorphisms (SNPs) versus methylation variation. The output of the pipeline can be loaded directly into a genome browser for visualization and into specialized methylation analysis software for analysis of differential methylation.

Chapter 4: Transgenerational inheritance of (stress-induced) DNA methylation in apomictic dandelions.

In this chapter I characterize the context-specific levels and transgenerational stability of DNA methylation variation in various genotypes of apomictic dandelions. I also determined the degree to which heritable DNA methylation variation is impacted by stress treatment with the chemical demethylation agent 5-azacytidine that will affect the level of DNA methylation and the plant hormone Jasmonic acid that is involved in plant defense signaling in response to attack by chewing insect herbivores and is therefore often used as a mimic of herbivory. EpiGBS allows for cost-effective and comprehensive sequencing-based characterization of DNA methylation over 6 generations of 4 different lines in 2 different genotypes, covering 178 plants. I only uncover weak patterns on CHG and CHH methylation indicating a transgenerationally transmitted stress response on DNA methylation, suggesting that either the stress response was not severe enough or transgenerational inheritance of such a response is limited. The experimental design allowed for a characterization of the stability of DNA methylation over generations, providing evidence for heritability of CHG methylation, which to our knowledge had not been characterized previously. This study also uncovered several improvement opportunities in the experimental design of epiGBS, such as switching to methylation insensitive enzymes, using a reduced number of PCR cycles and control nucleotides to detect PCR duplicates.

Chapter 5: Selection on early flowering in asexually reproducing Dandelions; using epigenetic variation as a substrate for artificial selection

Selection on a phenotypic trait which is underpinned by heritable DNA methylation variation should result in a shift towards trait values in the offspring in the direction that the trait was selected for. Here, I select on early flowering in isogenic lineages of stress pre-treated and control lineages of apomictic dandelions, which I expect to be under both genetic and epigenetic control. In this study, I aim to answer the following questions, (i) Did sufficient heritable variation arise during the three (treatment) generations to support a response on selection on flowering time? (ii) Is the response to selection stronger for plants that received a Jasmonic Acid or 5-azacytidine pre-treatment? And (iii) Is there a difference in the selection response between the three genotypes? Using parental stress treatments, I show a genotype and treatment specific stress response, which in some cases enhanced the selection response. I also uncover differences in the baseline levels of epigenetic variation, with a recent apomictic lineage showing a stronger response to selection regardless of treatment. Although some stress-induced epigenetic variation gave rise to an enhanced selection response this was not a general feature, suggesting that the stress responses were sufficiently severe enough to induce more heritable epigenetic variation on which selection can act. The fact that I did observe a significant shift in mean flowering time suggests that natural epigenetic variation underpinning phenotypic variation can be subjected to selection leading to a changed phenotype, giving support to the hypothesis that epigenetic variation can play a role in adaptation and evolution.

Consistent errors in first strand cDNA due to random hexamer mispriming.

2

Thomas P. van Gorp, Lauren M. McIntyre and Koen J.F. Verhoeven

*Published in a slightly different form in PLoS One (2013) 8(12): e85583.
doi:10.1371/journal.pone.0085583*

Abstract

Priming of random hexamers in cDNA synthesis is known to show sequence bias, but in addition it has been suggested recently that mismatches in random hexamer priming could be a cause of mismatches between the original RNA fragment and observed sequence reads. To explore random hexamer mispriming as a potential source of these errors, we analyzed two independently generated RNA-seq datasets of synthetic ERCC spikes for which the reference is known. First strand cDNA synthesized by random hexamer priming on RNA showed consistent position and nucleotide-specific mismatch errors in the first seven nucleotides. The mismatch errors found in both datasets are consistent in distribution and thermodynamically stable mismatches are more common. This strongly indicates that RNA-DNA mispriming of specific random hexamers causes these errors. Due to their consistency and specificity, mispriming errors can have profound implications for downstream applications if not dealt with properly.

Introduction

RNA-seq is a widely-used tool for transcriptome analysis and gene expression estimation. Most commonly, mRNA is fragmented followed by reverse transcription into first strand cDNA primed by random hexamers. Subsequently, second strand cDNA is synthesized from first strand cDNA by DNA polymerase, again initiated by random hexamer priming. Bias in hexamer distribution and CG content affects abundance estimates and several bias correction algorithms have been developed e.g. (Hansen et al. 2010; W. Zheng et al. 2011; Schwartz et al. 2011; L. Jiang et al. 2011; Benjamini & Speed 2012).

Apart from bias in hexamer priming sites, random hexamer mispriming has recently been implicated in sequence read to reference mismatches. Because mismatches mainly occur in the first seven nucleotides of first strand cDNA (L. Jiang et al. 2011; Pickrell et al. 2012) and are observed in transcriptomic but not genomic sequencing datasets (Hansen et al. 2010), RNA-DNA mispriming of random hexamers during first strand cDNA synthesis has been suggested as a likely explanation for the observed sequence mismatches (L. Jiang et al. 2011; Pickrell et al. 2012; Kleinman & Majewski 2012). It is important to recognize such technical artifacts because they might obviate biological interpretation of observed SNP patterns or RNA-editing (Lin et al. 2012; Pickrell et al. 2012; Kleinman & Majewski 2012; M. Li et al. 2011).

This paper deals with sequence read to reference mismatches commonly observed in RNA-sequencing data. Mismatches are defined as any position in sequencing reads that deviate from the reference to which these reads align. Mismatches can reflect 1) true biological variation, both in genomic DNA or caused by RNA-editing, or 2) errors in the library preparation process caused by hexamer mispriming or PCR errors, and finally 3) sequencing errors caused by the erroneous identification of bases in the sequencing process. We focus on errors that most likely arise in the library preparation phase. During RNA-seq library preparation, polyA⁺ RNA is fragmented and reverse transcribed into first strand cDNA initiated by random hexamer priming. We refer to mismatches between the reference fragment and the observed sequence that are caused by RNA-DNA hexamer mispriming during first strand cDNA synthesis as RD-mismatches. Similarly, we use the term DD-mismatches to describe errors caused by DNA-DNA hexamer mispriming during second strand cDNA synthesis.

Errors caused by random hexamer mispriming have thus far received limited attention. While read trimming to exclude error-rich first stretches of RNA-seq reads is done in some studies (Matvienko et al. 2013; Ashrafi et al. 2012), there is currently limited insight into the problem and no general consensus exists on how to efficiently deal with it in RNA-seq data. Here we perform a detailed analysis of both RD-mispriming and DD-mispriming in RNA-seq data in order to further our understanding of the causes and possible consequences of errors associated with random hexamer binding.

Materials and Methods

We conducted a RNA-seq experiment of *Taraxacum officinale* RNA mixed with ERCC RNA spikes. Twenty-three libraries were multiplexed using Illumina's multiplex sequencing assay and pooled with a 2% ERCC spike (Zook et al. 2012). Subsequently they were sequenced on 2 Hiseq lanes yielding a total of 2.4 million 100bp read pairs mapping to the ERCC spike set. Adapter trimming and quality filtering was done using Fastq-mcf v1.0.3-r152 with the following settings (-x 0 -k 0 -q 20). Mapping was done with BWA 0.6.1-r104 with default settings. Unpaired reads or reads from pairs with incorrect insert sizes containing adapter remnants at the 3' end were excluded from this analysis. A custom python script (available at <http://goo.gl/5c9DaZ>) was used to identify positional errors in forward and reverse mapping reads from the bam file, data are deposited in the SRA with reference number SRR954526. ERCC RNA-seq data described in (L. Jiang et al. 2011); GSM517062) were mapped and analyzed with the same settings.

Strand specific positional error rates in ERCC control RNAseq data

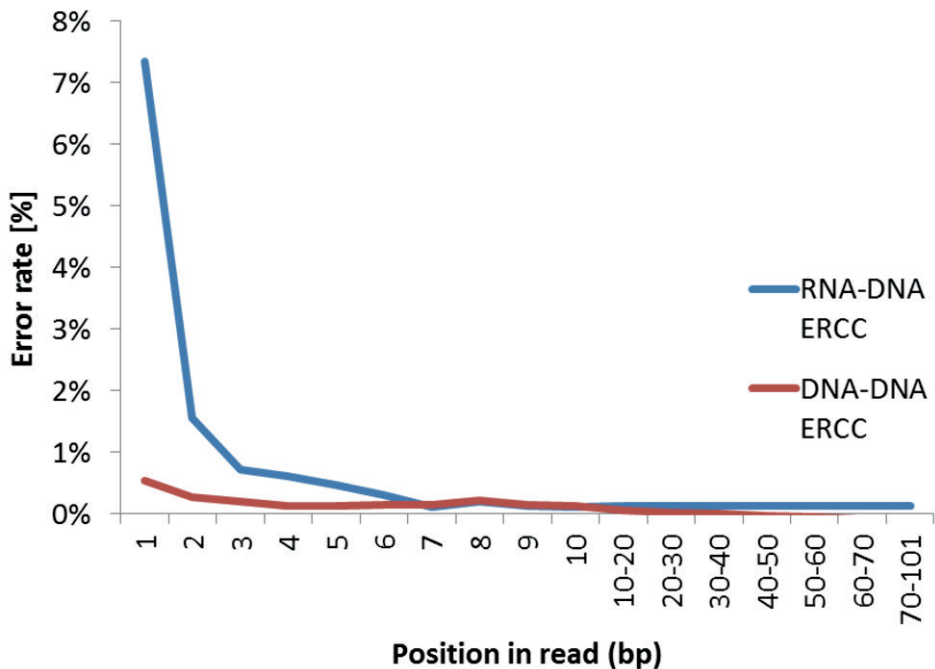


Figure 2.1 Read position effect on sequence mismatches.

Sequence mismatch rates in first (RNA-DNA) and second (DNA-DNA) strand cDNA of reads mapping to ERCC sequences.

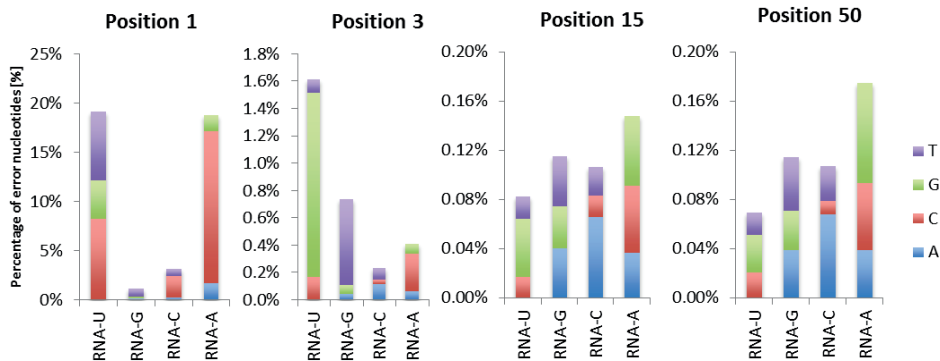


Figure 2.2 The mismatch rate and distribution for the first, third, 15th and 50th positions of first strand cDNA.

For all 4 nucleotides present in RNA the distribution of mismatching nucleotides is shown at selected positions. Mismatch rates are highest for first strand cDNA reads starting with T or A. For position three mismatches are mostly due to RNA-U vs DNA-G mispriming. Per nucleotide mismatch distributions are highly variable for the first seven positions, whereas they are consistent from position 7 onwards.

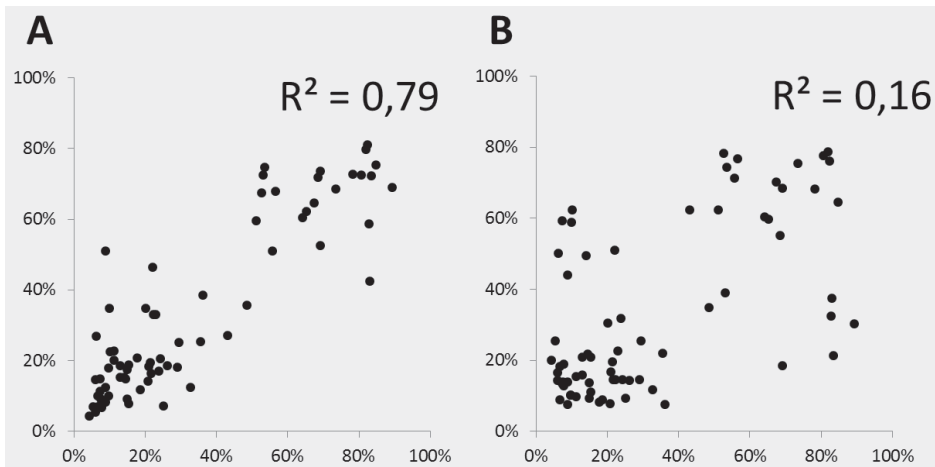


Figure 2.3 Mismatch error pattern correlation between two independent datasets. Mismatch error patterns observed in ERCC spike reads are correlated between independent data sets for read position 1-6 (panel A) but not for subsequent positions (position 7-20, panel B). For each position the distribution of errors over the 12 mispriming possibilities was determined (4 nucleotides x 3 mispriming options, summing to 100% per position) and plotted between the dandelion RNA-seq ERCC-dataset (see Figure 1) and ERCC RNA-seq data described in (L. Jiang et al. 2011; GSM517062). The correlation in error distributions between the two data sets shows consistency of specific mismatching errors only in the hexamer binding region (panel A).

Results and Discussion

We analyzed two independently generated RNA-seq datasets, focusing on reads mapping to ERCC spikes which are artificial RNA fragments of known sequence that are added during library preparation (L. Jiang et al. 2011). Reverse mapping reads represent first strand cDNA and forward mapping reads represent second strand cDNA. Because the reference strand is known for all ERCC's, deviations represent library preparation or sequencing errors, not true biological variation. For all positions in reverse and forward reads, substitution errors were calculated by parsing the sequence and MD tag of all reads in the SAM file. Here, the mismatching and expected nucleotide as well as position is denoted. Consistent with previous results (L. Jiang et al. 2011; Pickrell et al. 2012; Lin et al. 2012), we observed that per position nucleotide mismatch rates are higher for the first seven nucleotides compared to the rest of the sequence reads. As we can separate first and second strand synthesis we show that first strand synthesis mismatch rate is higher, consistent with RNA-DNA hexamer mispriming (Fig. 2.2.1). Of importance, these are bases called with high quality, and thus do not likely represent sequencing error. A slight increase in mismatch rate was also detected for initial nucleotides of second strand cDNA, suggesting that DNA-DNA hexamer binding is not insensitive to mispriming errors. Sequence to reference mismatches are not limited to the first nucleotide as commonly observed in Illumina sequencing data (Dohm et al. 2008), but similar to first strand cDNA, show higher rates in the first seven nucleotides compared to the rest of the read.

First strand cDNA mismatches in the first seven base pairs corresponding to the hexamer binding site and the base immediately downstream of this show position-dependent and nucleotide-dependent mismatch patterns (Fig. 2.2). These specific mismatch patterns differ markedly from the mismatch rates and distribution observed downstream of base seven caused by sequencing or PCR errors (Fig. 2.3). RNA-seq reads derived from first strand cDNA that start with A or T misprime in 20% of the cases, in which rA-dC and rU-dC mispriming are most common. In positions 2-6 ~65% of mispriming events consist of rU-dG and rG-dT, which are most stable among all 12 possible RNA-DNA misprimed pairs (Sugimoto et al. 2000). Overall, hexamer mispriming occurs most commonly at RNA binding sites with uracil, whereas cytosine in RNA prevents most hexamers from mispriming (Fig. 2.2).

Thus, we conclude that mispriming is non-random and can be heavily biased, as RNA binding sites with a U at positions three and four (relative to the hexamers 5' end) misprime with hexamers having a G at that position in ~88% of the mispriming cases. These specific mismatch patterns were observed consistently in two independent data sets (Fig. 2.3). The distribution, type and repeatability of mismatch patterns demonstrate that not all mispriming events have the same likelihood and RNA-DNA hexamer mispriming is the main source of error in the first seven nucleotides.

Consistent mismatch patterns observed in the first seven nucleotides of first strand cDNA will affect downstream applications such as de novo assembly, SNP calling and RNA-editing analysis. For instance, consistent and high (20%) mismatch rates can be problematic for k-mer assembly strategies, as these erroneous k-mers cannot be effectively combined with "true" k-mers. Because mispriming rates are not random

and for some positions heavily biased (Fig. 2.2) they can contribute to false positive variants that might not be detected based on filtering criteria such as read count or quality thresholds. Indeed, mismatches between human RNA-seq reads and the human reference genome have been interpreted as evidence for widespread RNA-editing (M. Li et al. 2011). Several reports highlight the overrepresentation of mismatches in the first six positions of first strand cDNA (Kleinman & Majewski 2012; Pickrell et al. 2012; Lin et al. 2012), suggesting hexamer mispriming on RNA as an explanation for the observed mismatches. Our study confirms the generality and specificity of hexamer mispriming, thus providing further support to this explanation. In fact, there are examples of filtering and preprocessing steps that counter the effect of random hexamer mispriming induced mismatches, however, their origin is not described. For example, empirically derived filtering parameters for putative somatic mutations used in Varscan2 exclude the first ten bases of reads (Koboldt et al. 2012), and 5' trimming has been applied to Illumina RNA-seq reads showing an increased N50, average and maximum contig length (Matvienko et al. 2013).

Identification of the hexamers that are most commonly involved in mispriming as well as the type and position of the mismatches this generates can aid in the design of strategies to counter the effects that these errors can have in downstream applications. Our results suggest that it could be useful to explore less aggressive approaches than trimming. Such approaches could include post-mapping correction of 5' mismatches, modifications to random hexamer design to exclude commonly mispriming hexamers and specific bias correction models that mask or remove the observed mismatches in the first seven bases of reads.

Conclusion

Our analyses show strong and consistent bias in sequence errors in 5' ends of RNA-seq reads, which (1) strongly supports the hypothesis that random hexamer mispriming during first strand cDNA synthesis causes the errors, and (2) highlights the risk of errors in downstream applications as well as suboptimal data use. We conclude that technical artifacts in sequencing data are insufficiently described. Further research on random hexamer mispriming will inform optimized strategies to mitigate their negative effect on downstream analysis.

Acknowledgments

We thank Joke van Vugt and Arjen Biere for their insightful comments and discussion on the manuscript, and Carla Oplaat and Alison Morse for generating the *Taraxacum* sequence libraries. This is publication 5497 of the Netherlands Institute of Ecology (NIOO-KNAW).

epiGBS: reference-free reduced representation bisulfite sequencing

Thomas P. van Gurp, Niels C.A.M. Wagemaker, Björn Wouters, Philippine Vergeer, Joop N.J. Ouborg and Koen J.F. Verhoeven

3

*Published in a slightly different form in Nature Methods (2016) 13, 322–324
doi:10.1038/nmeth.3763*

Abstract

We describe epiGBS, a reduced representation bisulfite sequencing method for cost-effective exploration and comparative analysis of DNA methylation and genetic variation in hundreds of samples *de novo*, without the need for a reference genome. This method uses genotyping by sequencing of bisulfite-converted DNA followed by reliable *de novo* reference construction, mapping, variant calling, and distinction of single-nucleotide polymorphisms (SNPs) versus methylation variation (software is available at <https://github.com/thomasvangurp/epiGBS>). The output can be loaded directly into a genome browser for visualization and into RnBeads for analysis of differential methylation.

Introduction

Epigenetic control by DNA methylation at cytosines (5_mC) is of paramount importance for cell regulation, differentiation and transposable element control (Suzuki & Bird 2008). Analysis of DNA methylation is enabled by bisulfite treatment, which converts unmethylated (but not methylated) cytosines to uracil. Subsequent sequencing of bisulfite-converted DNA allows for quantitative estimation of DNA methylation. Various methods enable bisulfite sequencing of specific subsets of, or all genomic DNA (Beck 2010). Reduced representation bisulfite sequencing (RRBS) focuses on a defined subset of genomic DNA combining restriction enzyme digestion with size selection. It was designed to target CpG islands (Gu et al. 2011), which are a common feature in mammalian, but not plant genomes, making it non-ideal for non-vertebrate systems. Currently published RRBS methods limit the researcher in terms of enzyme choice (J. Wang et al. 2013) and multiplexing level (Boyle et al. 2012) while requiring a reference genome for efficient mapping and variant calling (Gu et al. 2011). Here we present epiGBS, a method that allows for straightforward, cost-effective and reference-free RRBS of highly multiplexed libraries in an accurate and versatile way. The method calls both DNA methylation polymorphisms and SNPs from the same bisulfite-converted samples while reconstructing the consensus sequence of the targeted genomic loci.

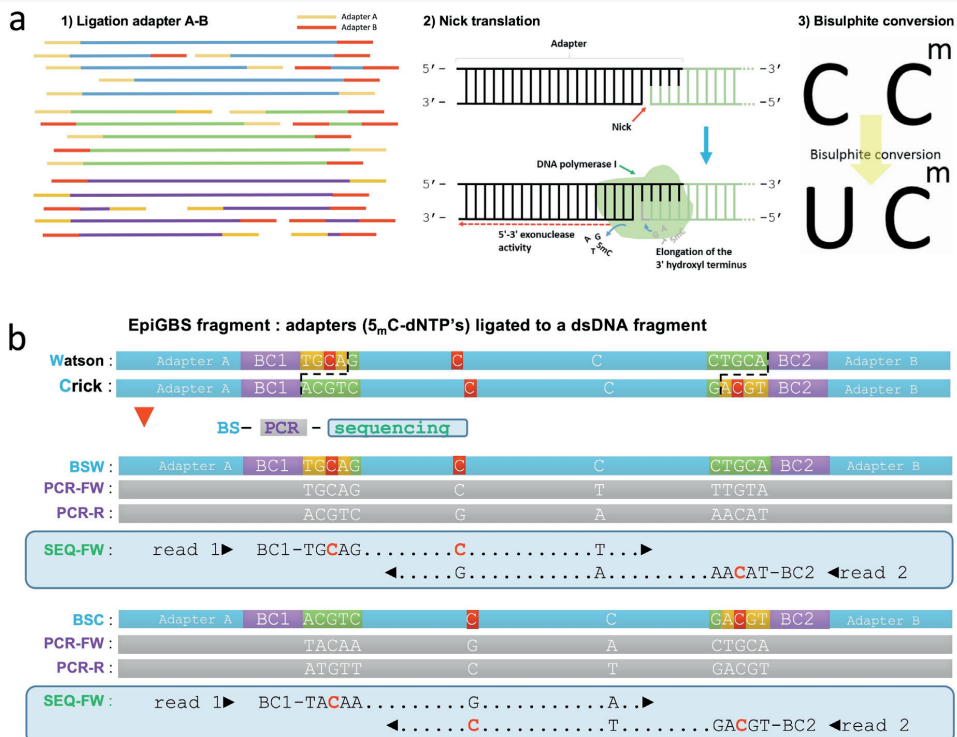


Figure 3.1 Method design and results. (a) Genomic DNA is digested with the selected restriction enzyme for sample 1-N. After ligation of barcoded adaptors A and B (1), fragments are pooled, PCR purified and subjected to 0.8× SPRI size selection. Nick translation (2) is used to repair the nicks between the adaptor and the restriction fragment (Supplementary Fig. 3.9). Fragments are bisulfite

converted (3) and PCR amplified to yield a sequencing library. **(b)** Unmethylated cytosines are bisulfite converted to uracil, whereas methylated cytosines remain intact. As only unmethylated PstI restriction sites result in digestion, the cytosines in enzyme-recognition sites (green) are converted. The cytosine in the adaptor is methylated (red) and thus remains unchanged. Eighteen cycles of PCR yield a sequencing library. Sequenced fragments contain a converted and an unconverted recognition site. The orientation of these sites is reversed for Watson and Crick reads; fragments with unconverted recognition sites on barcoded adaptor B are arbitrarily defined as Watson. BS, bisulfite sequencing; FW, forward; R, reverse.

Results and Discussion

Common protocols for RRBS call for MspI digestion of genomic DNA followed by end repair, A-tailing, adaptor ligation and several purification and gel-extraction steps (Gu et al. 2011). EpiGBS extends genotyping by sequencing (GBS) (Elshire et al. 2011) with bisulfite treatment, which allows for a much simpler protocol and substantial reductions in per-sample costs. Like GBS, our protocol involves enzymatic digestion of individual samples followed by barcoded adaptor ligation and pooling of samples. Subsequently, SPRI (solid-phase reversible immobilization)-based size selection, nick translation (Rigby et al. 1977), bisulfite treatment and PCR amplification are applied (Fig. 3.1a). We use nonphosphorylated adaptors to minimize adaptor dimerization. Nicks caused by the absence of 5' phosphate groups in adaptors are repaired using nick translation, which sequentially replaces adaptor nucleotides from 5'-3' using DNA polymerase I (Fig. 3.1a). By combining forward and reverse in-line barcoded adaptors compatible with Illumina sequencing primers, one can accomplish a 96-plex design with only 12 forward and 8 reverse adaptors containing 5-methylcytosines (Supplementary Fig. 3.1). Forward adaptors contain 4–6-base barcodes optimized for equal representation of per-cycle nucleotides during sequencing, thus minimizing phasing errors while maximizing signal intensity in Illumina sequencing (Elshire et al. 2011), allowing for high-quality epiGBS libraries.

We created PstI-epiGBS libraries for several species that were pooled in a 96-plex sequencing library (Supplementary Fig. 3.1e). Average per-species inserts ranged from 134 to 193 bases (Supplementary Fig. 3.1e). Included were four *Arabidopsis thaliana* (*Arabidopsis*) samples derived from lineages with known DNA methylation from a previous whole-genome bisulfite sequencing (WGBS) study (Becker et al. 2011). We carried out paired-end sequencing of these long fragments to facilitate efficient *de novo* clustering, reference reconstruction and BLAST-enabled functional classification.

Using the bisulfite-converted libraries only, we built a reference-free *de novo* bioinformatics pipeline by designing custom algorithms (<https://github.com/thomasvangurp/epiGBS>) aimed at clustering Watson and Crick reads derived from the same genomic location (same origin) (Fig. 3.1b). As bisulfite conversion targets only cytosines, unmodified guanines from the opposite strand provide the correct reference base after alignment (Fig. 3.1b). By aligning same-origin Watson and Crick reads, one can reconstruct the reference sequence *de novo* from bisulfite-converted reads (Supplementary Fig. 3.2). *De novo* reference

reconstruction for inserts larger than 240 nucleotides (nt) is limited to the ~120 nt at the start and end of the locus (Supplementary Fig. 3.3).

To validate our clustering approach, we mapped *de novo*-obtained reference sequences for four *Arabidopsis* epiGBS samples to the *Arabidopsis* reference genome (TAIR10). On the basis of an *in silico* digest of TAIR10 with PstI, we expected 2,094 fragments of 11–300 nt, from which 1,862 loci (89%) could be mapped by *de novo*-obtained clusters. To assess the technical performance of our *de novo* approach, we focused on a smaller subset of 1,691 genomic loci covered by at least ten Watson and ten Crick reads. An efficient *de novo* approach would produce the 1,691 clusters corresponding to the subset of genomic loci with sufficient mapped Watson and Crick reads. Clusters were mapped to 1,626 out of 1,691 loci (96%). Of these 1,626 loci, 5 were missed, 1,571 (97%) were covered once and only 50 (3%) were covered twice by the *de novo* generated clusters with a 0.1% rate of mismatch to the genomic reference sequence, confirming the sensitivity and accuracy of the *de novo* approach. The *de novo* clustering yielded 1,194 additional clusters larger than 300 nt that mapped back to the TAIR10 reference, as well as 16,498 clusters that mapped to microbial genomes or were unknown, which typically showed low coverage. In total, 96% of all mappable reads mapped to 3,056 *de novo*-obtained *Arabidopsis* clusters.

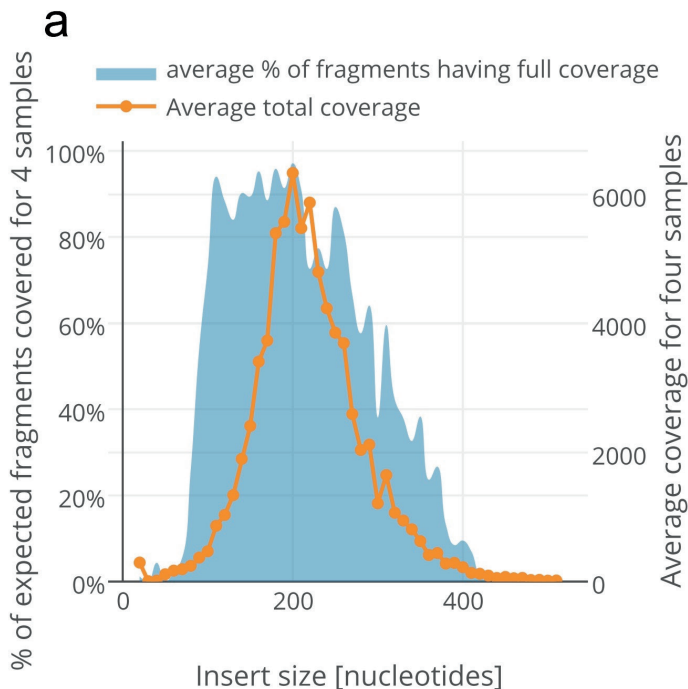
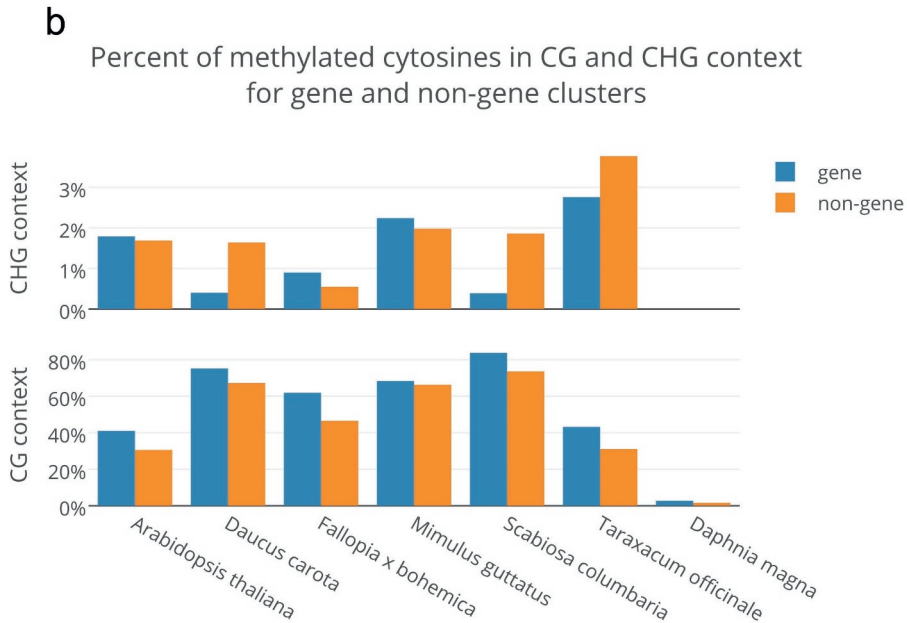


Figure 3.2 | Insert size and coverage in *Arabidopsis thaliana* and context-specific methylation distribution for seven species. (a) Insert size and coverage for *Arabidopsis*. The blue shading indicates the average percentage of sites covered per size bin of 10 nt for all four *Arabidopsis* samples sequenced, with coverage defined as present per individual per site if at least ten Watson and Crick reads are present. Approximately 90% of genomic loci with insert sizes between

170 and 220 were fully covered. The average coverage was maximal for fragments of 200 nt.



(b) The percentage of symmetrically methylated cytosines for seven species in CG and CHG contexts for gene-related and non-gene-related clusters. Per species, we chose four representative untreated individual specimens with the highest sequence coverage, except for *Mimulus guttatus*, *Daucus carota* and *Fallopia × bohemica*, for which we used only two specimens (Materials and Methods). Results for *Homo sapiens*, *Allium porrum* and phage λ are not shown because of insufficient coverage (Materials and Methods).

We use BWA-METH for mapping, as it provides accurate and reliable results and allows for the transfer of sample-identification information directly from FASTQ to BAM files (Pedersen et al. 2014). After mapping, the resulting BAM files are split into Watson and Crick strand-mapping reads, after which variant calling is done with FreeBayes (Garrison & Marth 2012). Combining variant calls from both Watson and Crick strand allows one to distinguish genetic from epigenetic variation (Liu et al. 2012). This is possible because methylation polymorphisms show as C/T (on the Watson strand) or A/G (on the Crick strand) but are nonvariable in the opposite strand, whereas genetic polymorphisms are variable on both strands. Genetic variations such as mutations in the enzyme-recognition site can lead to missing data for mutant samples. Likewise, mutant reads with structural variation can give rise to additional clusters to which only mutant reads will map, again leading to missing data. For a *de novo* analysis, samples should thus be genetically similar in order for epiGBS to find sufficient clusters with coverage for all samples. We estimated nonconversion rates on the basis of nonmethylated DNA of phage λ (Materials and Methods). Resulting methylation polymorphisms were converted into sample-

specific methylation BED files allowing for subsequent analysis with RnBeads (Assenov et al. 2014) or similar packages.

Comparisons between existing RRBS studies and WGBS (Lister et al. 2008; Cokus et al. 2008), the current gold standard, have shown that MspI RRBS provides results that benchmark very well against WGBS data, demonstrating the accuracy and sensitivity of RRBS (Beck 2010; Harris et al. 2010). To validate epiGBS, we compared epiGBS methylation calls for four *Arabidopsis* samples to WGBS-based methylation information on previous-generation plants from an earlier study (Becker et al. 2011). We compared methylation of cytosines with at least ten informative reads in WGBS data with positions covered by at least ten forward and ten reverse informative epiGBS reads (Supplementary Fig. 3.4). We obtained a Pearson R^2 of 0.95 for methylation levels at 12,389 cytosines in the CG context (Supplementary Fig. 3.4b). This is almost identical to the R^2 of 0.96 obtained for comparisons of intergenerational WGBS (Becker et al. 2011) values for CG methylation at the same genomic locations (Supplementary Fig. 3.4a). Correlations for cytosines in non-CG context were lower and showed a slight bias that is attributable to the CHG methylation sensitivity of PstI (Supplementary Fig. 3.4b). (This slight bias can be completely avoided if a methylation-insensitive restriction enzyme such as Csp6I is used.) Given sufficient coverage, this approach results in intergenerational cytosine methylation correlations for epiGBS data that are as strong as intergenerational WGBS-WGBS comparisons (Supplementary Fig. 3.4d). Also, the detection of differentially methylated positions (DMPs) on the basis of PstI-epiGBS data produced results nearly identical to those obtained via DMP calling using WGBS data (Supplementary Fig. 3.5). In epiGBS, the number and type of targeted loci vary with genome size and restriction-site distribution; the average coverage is typically highest for fragments of around 200 nt (Fig. 3.2a). Like GBS, epiGBS is flexible with respect to the use of different restriction enzymes, and this can be exploited to bias sequencing toward (or away from) specific genomic features. Using PstI in *Arabidopsis*, we targeted 2,260 loci with coverage of >100, representing ~0.37% of genomic DNA. As PstI is sensitive to CHG methylation, repetitive DNA in *Arabidopsis* is largely avoided (Supplementary Fig. 3.6), allowing the sequencing effort to focus on coding regions where most differentially methylated regions in *Arabidopsis* are located (Becker et al. 2011). Investigating a biased subset of the genome also means that PstI-based epiGBS methylation characterization is not necessarily representative of genome-wide methylation patterns, but such biases can be overcome easily with the use of a different (methylation-insensitive) enzyme, as we demonstrated using Csp6I (Supplementary Fig. 3.4c).

Our bioinformatics pipeline produces methylation and SNP variant call files as well as methylation-level tabular files suitable for visualization in a genome browser such as IGV (Supplementary Fig. 3.7) or the recently published RnBeads pipeline (Assenov et al. 2014), providing seamless integration and analysis of DMPs (see Material and Methods).

Our analysis confirms a prior observation in model species: cytosine methylation in the CG context is higher in genes than in non-gene clusters (Fig. 3.2b). We detected DMPs in all species. Strikingly, we detected more CHG DMPs in nonmodel plant species than in *Arabidopsis* (Supplementary Fig. 3.8). Our analysis of *Daphnia*

magna, an important ecotoxicological model species, provides compelling evidence for the occurrence of DNA methylation in the CG context in this aquatic species (Vandegheuchte et al. 2009)(Fig. 3.2b). Our method can thus be used to expand knowledge on the presence of DNA methylation in population or evolutionary studies of nonmodel species lacking a reference genome. Thus epiGBS provides a significant improvement over existing marker-based methylation screening tools such as MS-AFLP (Schrey et al. 2013), allowing for much more comprehensive studies on differences in DNA methylation in nonmodel species.

Materials and Methods

Description of Samples

DNA from individuals of several species was collected to showcase the versatility of the epiGBS method. For several species, samples treated with 5-azacytidine, genistein or other stressors were included. In the per species DNA methylation analysis as presented in Figure 2b and Supplementary Figure 8, only untreated control samples were included. Details for the individual treatments per species are presented here to facilitate further analysis on the data presented here (see Supplementary Data 1).

Scabiosa columbaria samples originated from an earlier experiment (Vergeer et al. 2012). For this experiment, seeds were collected in a large French calcareous grassland after which plants were grown in the greenhouse of Radboud University Nijmegen. Plants were crossed and F1 seeds were germinated on filter paper saturated with water or 50 μ M 5-azacytidine (Vergeer et al. 2012). Seedlings were transferred to either a soil mixture resembling calcareous soils (pH 7.5-8.0) or a soil mixture resembling acidic soil (pH 6.0-6.5). After 3 months, young leaves/plant were sampled for DNA extraction using the Macherey-Nagel nucleospin plant II kit for individual tubes. 400 ng gDNA was used for the epiGBS analysis. Plants that were included in the epiGBS analysis all originated from the same maternal line.

Mimulus guttatus seeds were collected in a large Dutch population in the vicinity of Arnhem, the Netherlands. The seeds were germinated and grown under controlled conditions in the greenhouse of Radboud University Nijmegen. Juvenile plants were divided over 4 experimental treatments: control, drought, water-logged or submerged. After 4 weeks, young leaves were sampled for DNA extraction using the Macherey-Nagel nucleospin plant II kit for individual tubes. Eight samples (2 control, 2 drought, 2 waterlogged and 2 submerged) were included in the epiGBS analysis (400 ng gDNA of each sample).

Allium porrum and *Daucus carota* seeds were provided by Nunhems Netherlands BV – Bayer CropScience. Of each crop, seeds of an inbred and hybrid line were germinated and grown under controlled conditions in the greenhouse of Radboud University Nijmegen. Seeds were germinated on filter paper saturated with water (control) or 40 μ M genistein solution (genistein). After 4 weeks, young leaf tissue was sampled for DNA extraction using the Macherey-Nagel nucleospin plant II kit for individual tubes. Per sample, 400 ng gDNA was used for the epiGBS analysis.

Daphnia magna samples were provided by Prof. De Meester, University of Leuven, Belgium. Inbred families were obtained through clonal selfing within clones that were isolated from fish pond LangeRodevijver near Leuven in Belgium. Outbred

families were derived from resting eggs, collected from the dormant egg bank in fish pond LangeRodevijver, that were hatched in the laboratory. For more detailed information, see (Swillen et al. 2015). Six samples (3 outbred and 3 inbred individuals from the second clutch) were used in the epiGBS analysis (400 ng gDNA of each sample).

Fallopia japonica and *Fallopia x bohemica* samples were provided by Prof. Christina Richards, University of South Florida, Tampa USA. Rhizomes were collected from field sites in Long Island, NY, USA which were subsequently propagated in the greenhouse to obtain fresh leaf tissue. DNA was extracted from the third uppermost fully expanded fresh leaf using Qiagen DNeasy Plant Mini kit. All samples were used to create a de novo reference but only two samples (301 and 314) of *F x bohemica* (both from roadside habitat) were included in the epiGBS analysis (400 ng gDNA of each sample).

Taraxacum officinale line A68 and A34 are clonal apomictic Dandelion lines (Verhoeven, van Dijk, et al. 2009), maintained under greenhouse conditions for several generations, of which leaf material of several individuals from an ongoing experiment were collected over the course of several years. Leaf material was stored at -80 °C for 1-2 years. DNA isolation was performed for all samples simultaneously as described for the other samples. Plants were grown until flowering was completed under controlled conditions in a greenhouse (day/night temperatures of 21°C/17°C, 16 h photoperiod).

Arabidopsis thaliana seeds were provided by C. Becker (Max Planck institute for developmental biology, Tübingen, Germany) via F. Johannes (Department of Plant Science, Technical University of Munich, Germany). Plants from the previous generation of the same lineage were subjected to Whole Genome Bisulfite Sequencing (WGBS) 8. For each lineage, up to 5 individual plants were grown for 4 weeks under controlled conditions in a greenhouse (day/night temperatures of 21°C/17°C, 16 h photoperiod). Per line, a single plant was used for DNA isolation.

Human DNA was bought from Promega (Human, Female, order nr. G1521)

Phage lambda DNA was bought from Promega (λ DNA, cl857 Sam7, order nr. D1521). Phage lambda cl857 Sam7 DNA was isolated from infected GM119, an *E. coli* strain lacking both the *dam* and *dcm* methylase activities (Promega).

Further sample details are available as a Supplementary Data 1 in which for all species per-sample genotypes, treatments, tissue, age and barcodes used are specified.

DNA Extraction

Plant material was homogenized by bead-beating frozen leaf tissue in a 2 mL eppendorf tube with 2-3 mm stainless steel beads. No more than 100 mg of fresh tissue was used per sample. Samples with beads were taken from the freezer and stored in liquid nitrogen. Per batch of 12 samples the tubes were put in a tissuelyzer 24 adapter set block, which was partially submerged in liquid nitrogen to avoid thawing during sample placement. After 30 seconds of shaking at 30 hz / 1800

oscillations per minute the tubes were submerged in liquid nitrogen after which a second round of 30 second shaking was performed.

For most plant species mentioned, we isolated DNA using the Macherey-Nagel nucleospin plant II kit (for details see Supplementary Data 1) for individual tubes. We followed the manufacturer's protocol with the following modifications. Cell lysis was performed using Cell lysis buffer PL1 for 30 instead of 10 minutes. After filtration through the pink nucleospin filter the flow-through is carefully pipetted to a fresh 1.5 mL tube, avoiding the pellet, which is often formed at this stage. An additional centrifugation step was used to avoid a small pellet; the clear supernatant was used in the following steps according to the manufacturer's protocol. As restriction enzymes are very sensitive to proteins and other contamination we only selected samples with high purity, specifically 260/280 and 260/230 ratios of at least 1.8 and 1.5 respectively). DNA concentration was determined using Qubit® 2.0 Fluorometric dsDNA HS Assay Kit (Q32851 Life technologies).

Restriction digestion

Per individual, 400 ng of genomic DNA (gDNA) was digested overnight (17hrs) at 37°C in a volume of 40 µL containing 1x NEBuffer 3.1, 125 µg BSA (NEB, B9000S) and 2 µL / 40 units of PstI (NEB, R0140S). Following digestion, barcoded adapters were ligated to the fragments.

Adapter ligation

To minimize the possibility of misidentifying samples as a result of sequencing or adapter synthesis errors, all pair-wise combinations of barcodes differed by a minimum of three mutational steps, barcode lengths were modulated from 4 to 6 bp to maximize the nucleotide balance of the bases at each position in the overall set of sequencing reads (Supplementary Fig. 3.1d). Samples were pooled and processed per species after ligation. For the ligation 1200 pg of both forward and reverse barcoded adapters (Supplementary Table 1b) were combined in a 60 µL reaction containing 40 µL gDNA digest, 1x T4 DNA ligase buffer and 4000 units T4 DNA ligase (NEB, M0202M/L) and ligated for 3hrs at 22°C followed by 4°C overnight, no inactivation afterward.

Cleanup and size selection

In order to assess the quality of libraries the pooling was performed per species. When pooled, the total library volume was reduced by Qiaquick PCR cleanup (Qiagen, 28104) to 60 µL. The libraries were size selected by a 0.8x Agencourt AMPure XP (Beckman coulter, A63880) purification favoring >200 bp DNA fragments and eluted in a total volume of 24 µL.

Nick translation

Due to the use of non-phosphorylated adapters, epiGBS libraries contain nicks between the 3' fragment overhang and the 5' non-phosphorylated adapter nucleotide. To prevent the loss of ssDNA adapter strands (at the nicked position) during bisulfite treatment the nick was repaired (see Supplementary Fig. 3.9) by a 1-hour nick translation reaction at 15°C in a reaction of 25 µL containing 18 µL of the purified library, 2.5 µL of 10 mM 5-methylcytosine dNTP Mix (Zymo research, D1030), 1x NEBuffer 2 and 7.5 units DNA polymerase I (NEB, M0209S).

Optional GBS PCR

At this stage, an optional GBS PCR was performed to check the library quality. GBS PCR was performed using the epiGBS PCR protocol (see below). The average size of the amplified GBS library was expected to be bigger than the amplified epiGBS library. The quality of these PCR-libraries was assessed by analyzing 1 μ L of the PCR product on a High Sensitivity DNA chip on a 2100 Bioanalyzer system (Agilent).

Bisulphite treatment and purification

For bisulfite treatment 20 μ L of the nick-translated library was used. Bisulfite treatment was performed using the EZ DNA Methylation-Lightning™ Kit (Zymo Research) with the following program: 8 minutes 98°C, 1 hour at 54 °C followed by up to 20h at 4°C, all according to the manufacturer's protocol.

EpiGBS PCR

Library amplification was performed per species in four individual 10 μ L reactions containing 1 μ L ssDNA template, 5 μ L KAPA HiFi HotStart Uracil+ ReadyMix (Kapa Biosystems), 3 pmol of each illumina PE PCR Primer (Supplementary Table 1b). Temperature cycling consisted of 95°C for 3 min followed by 18 cycles of 98°C for 10 s, 65°C for 15 s, 72°C for 15 s with a final extension step at 72°C for 5 min. Replicate PCR products were pooled and quantified using a Qubit® dsDNA HS Assay Kit (Life Technologies). The quality of the Libraries was assessed by analyzing 1 μ L on a High Sensitivity DNA chip on a 2100 Bioanalyzer system (Agilent). Libraries were considered suitable for sequencing if the majority of DNA fragments were between 150–400 bp and no adapter dimers were found. Typically, epiGBS PCR reactions of 18 cycles of a non-pooled plant sample yield 3-12 ng/ μ L of PCR-product.

When the 'per species' pooled libraries passed quality control they were further pooled according to concentration and number of samples in the species pool so that each individual sample was expected to yield an equal number of clusters on the Illumina flowcell. A 'nano run' was performed on the Illumina MiSeq to quantify per-sample read count. Based on the read counts obtained from this run, the individual nick-translated digestion-ligations were pooled in such a manner that an equal number of reads would be expected per individual. Finally, Rapid Run Mode Paired-End sequencing was performed on an Illumina HiSeq2500 sequencer using the HiSeq v4 reagents and the latest version of the HiSeq Control Software (v2.2.38), which optimizes the sequencing of low-diversity libraries (<http://res.illumina.com/documents/products/technotes/technote-hiseq-low-diversity.pdf>). As the first five cycles of a sequencing run are used to calculate the color matrix, our barcode design achieves almost perfect balance of the first 5 nucleotides when equal numbers of sequences are obtained per forward read or "A" barcode. The reverse read or "B" barcodes do not have this requirement, hence only barcodes of four nucleotides were used (see Supplementary Table 1a).

Methods Csp6I Laboratory work

The Csp6I epiGBS libraries were constructed in similar fashion as the PstI epiGBS libraries with the following modifications: The restriction digestion reaction contained 1x FD buffer and 4 μ L / 40 units of Csp6I (ThermoFisher Scientific, FD0214). The ligation reaction contained 2400 pg of both A and B adapters (both

adjusted for the Csp6I sticky end). While in the PstI protocol we used fully methylated adapters (both strand I and II methylated) for the Csp6I protocol we used hemi-methylated adapters. The adapter strands that were resynthesized (incorporating 5mC dNTP's) by nick translation were not methylated as all cytosines are replaced by methylated 5mC (see Supplementary Fig. 3.9). Final amplification for the Csp6I library yielded 4 - 8 ng / μ L product for an epiGBS PCR of 18 cycles of a library only containing Arabidopsis sample A29.

Bioinformatics Analysis

All custom python scripts mentioned are available on github (<https://github.com/thomasvangurp/epiGBS>). Raw reads are deposited at the NCBI short read archive (<http://www.ncbi.nlm.nih.gov/sra>) under bioproject PRJNA287755. Processed data is available at genomespace (<https://gsui.genomespace.org/jsui/gsui.html?pathOrUrl=/Home/thomasvangurp/epiGBS> Nature Methods). The bioinformatics pipeline design is visualized in Supplementary Fig. 3.10.

Demultiplexing

Paired-end reads for both the PstI-based titration miseq nano run as well as the Hiseq2500 run were demultiplexed using the custom python script demultiplex.py with an appropriate barcode to sample file. Up to two mismatches were allowed in the barcode and enzyme recognition site on both forward and reverse reads. During demultiplexing read group attributes specifying the sample name were added to the read name whereas the barcode was stripped from both forward and reverse reads. Additionally, a sample type tag (ST:Z:watson or ST:Z:crick) was added based on the orientation of the enzyme recognition sites in the paired reads (see Fig. 3.1b).

De novo reference construction

The paired end reads were split per species using unix grep. De novo reference construction was carried out on a per species level. Reads were merged using PEAR v0.9.5 (Jiajie Zhang et al. 2014) with the following settings (minimum p-value for accepting an overlap 0.001, minimum overlap 10, no trimming, minimum assembly length 0). This results in merged and non-merged reads. Non-merged reads were concatenated with 10 N nucleotides between the forward and reverse read (Supplementary Fig. 3.3). Both merged and non-merged reads were split in Watson and Crick reads using custom unix grep queries. Methylation polymorphisms in Watson and Crick reads were removed using unix sed, replacing all C's with T in Watson and all G's with A in Crick reads. Computationally derived "demethylated" sequences were dereplicated using usearch (<http://www.drive5.com/usearch/>) with the dereplication command. Only clusters consisting of 2 or more reads were retained. To enable pairing of same-origin Watson and Crick reads C's are converted to T's in Watson reads while G's are converted to A's in Crick reads; rendering a binary AT only sequence output, which is identical for same origin Watson and Crick reads (Supplementary Fig. 3.2). The reference sequence per pair is called for both Watson and Crick reads using combined bash piped queries using samtools mpileup and bcftools for variant calling followed by vcfutils.pl vcf2fq (<https://github.com/lh3/samtools/blob/master/bcftools/vcfutils.pl>) for creating the Watson and Crick consensus sequence. Watson and Crick consensus sequences are used to recreate the original sequence from which the bisulfite converted reads

were derived. Bases from both Watson and Crick consensus sequence are processed simultaneously per position with `create_consensus.py`. In case of Watson:T / Crick:C a C is added to the reference sequence output, whereas in case of Watson:G / Crick:A a G is added. If paired nucleotides match, that nucleotide is added. In all other cases where paired nucleotides do not match an ambiguous nucleotide (N) is added (Supplementary Fig. 3.11).

Trimming

Part of the read sequence in both Watson and Crick reads originates from fully methylated adapters (Supplementary Fig. 3.1c). To exclude these adapter nucleotides the first four nucleotides corresponding to adapter sequence ligated to the enzyme recognition site are trimmed. To achieve this, Watson reads were trimmed on forward (/1) reads whereas Crick reads were trimmed on reverse (/2) reads. In merged Watson reads the first four nucleotides were removed whereas in merged Crick reads the last four nucleotides were removed.

Mapping

Per species, merged and unmerged reads were mapped using a modified version of `bwameth.py` (see <https://github.com/thomasvangurp/epiGBS>) with default settings. The resulting bam files of both merged and unmerged reads were sorted and merged. Read group identifiers corresponding to the sample name identifiers present in the mapped reads were added using the `samtools` command `reheader`. Both Watson and Crick bam files were indexed using `samtools`.

Strand specific variant calling

Variant calling was done with `Freebayes` (Garrison & Marth 2012) per species on both Watson and Crick bam file separately using a custom Python script (`map_reads.py` module `runFreebayes`). This module runs `Freebayes` in parallel with settings that forces variants to be called on all positions (`freebayes -f consensus_cluster.renamed.fa -F 0 -E 1 -C 0 -G 0 --haplotype-length 1 --report-all-haplotype-alleles --report-monomorphic --report-genotype-likelihood-max --haplotype-length 1 -KkXuiwaq 21`). The resulting vcf files from the parallel runs done for both the Watson and Crick bam file were merged, compressed with `bgzip`, and indexed using `samtools` module “`tabix`” (<http://www.htslib.org/doc/tabix.html>) resulting in a VCF file for both the Watson and Crick strand.

Methylation calling

A custom python script `methylation_calling.py` uses the “`walktogether`” method of the `PyVCF` package (see <https://github.com/jamescasbon/PyVCF>) to simultaneously iterate over both Watson and Crick VCF files. By doing so, SNPs and methylation polymorphisms were distinguished and split. C/T polymorphisms in Watson combined with C on the Crick strand indicate a methylation polymorphism on the Watson strand whereas a G/A polymorphism on Crick combined with a G on the Watson strand indicates a methylation polymorphism on the Crick strand. Where combined SNP and methylation lead to a C/T or G/A polymorphism on both Watson and Crick strand only the SNP is called, as the methylation ratio cannot be reliably determined. Per site per individual methylation levels are available in various output formats, such as `.VCF`, `.IGV` and `.BED` files to provide for analysis and visualization of the methylation variation.

Methylation summary statistics calculation supplementary Table 3.1e

Per species, similarity searches for reference sequences of the clusters with an appropriate proteome were done with usearch ublast 23. A gene hit for a plant species was defined as a cluster having an e-value < 1e-5 after mapping to the reference proteome (refseq) of all eudicot plant species. For *Daphnia magna*, all *Daphnia* related proteins (NCBI) were used as a reference. Per species, each cluster was labeled “gene” or “non-gene” based on the results of usearch ublast. Methylation is summarized for cytosines in symmetric CG and CHG context for both gene and non-gene related clusters.

The following criteria were used for positions to be considered (methylated):

1. Per position minimum read coverage for forward and reverse informative reads > 10
2. For cytosines on the top strand only reads from the Watson pool are informative, whereas for cytosines on the bottom strand only Crick reads are informative. Positions with SNPs or ambiguous contexts due to neighboring SNPs are excluded from this analysis.
3. Minimum methylation ratio for both forward and reverse informative reads > 0.05, otherwise the position is considered non-methylated
4. Only symmetric positions in CG and CHG context with sufficient coverage according to criterion 1 are taken into account. For these symmetric positions, for both Watson and Crick strand nucleotides criterion 3 needs to apply, otherwise the symmetric pair is considered non-methylated.
5. Methylation is summarized for symmetric pairs for all positions meeting above criteria in up to four non-treated / control samples having the highest coverage per species. For *Mimulus guttatus*, *Daucus carota* and *Fallopia x bohemia* only two individuals could be included.
6. Methylation summary statistics calculation were performed for all species included with sufficient coverage. *Allium porrum* and *Homo sapiens* were excluded from the analysis due to insufficient sequence coverage. Statistics are computed in the submodule “summary_methylation” of the same python script used for DMP detection in all species “all_DMP_detection.py”.

Methods Csp6I Bioinformatics

For the Csp6I library we mapped all merged paired-end reads to the Arabidopsis TAIR10 genome. As the origin (Watson or Crick) of the merged reads could not be established due to the absence of a methylation sensitive restriction enzyme recognitions site as present for PstI, we have adjusted the procedure for mapping the reads. All merged reads were computationally demethylated using unix sed in two different ways, creating a C to T copy and a G to A copy for each read. These two different computationally derived copies were subsequently mapped with bwa-mem to a C to T converted and G to A converted genome (also created using unix sed). Only one of these four mapped copies has a meaningful and high-scoring match. To find that hit the four resultant bam files were sorted on name, were the highest scoring match was retained if it was matching over 80% of its length. This procedure mimics what bwa-meth normally does internally, but could not do given our Csp6I-based epiGBS data.

A Watson and Crick bam file were produced by adding the original non-converted sequence in its proper orientation to the appropriate bam file. Methylation calling was done using samtools mpileup through a custom python script (available on github repository <https://github.com/thomasvangurp/epiGBS> under *Arabidopsis_extra/Arabidopsis_Csp6i_analysis.py*). Comparisons between Csp6I Methylation estimates obtained for Arabidopsis generation 32 line 29 and corresponding generation 31 WGBS estimates from Becker et al 2011 were obtained in the same way as for the PstI-epiGBS data described above. Bam files with mapped reads for Csp6I are available on Genomespace ([https://gsui.genomespace.org/jsui/gsui.html?pathOrUrl=/Home/thomasvangurp/epiGBS Nature Methods/Csp6i](https://gsui.genomespace.org/jsui/gsui.html?pathOrUrl=/Home/thomasvangurp/epiGBS_Nature_Methods/Csp6i)).

DMP detection in all species

DMPs were detected in up to six pairwise comparisons per species for up to four individuals of non-treated samples (see Supplementary data 2 available online at <http://www.nature.com/nmeth/journal/v13/n4/extref/nmeth.3763-S3.xlsx>) having the highest sequence coverage (see materials and methods) in similar fashion as the detection carried out for Arabidopsis (Supplementary Fig. 3.5) with the exception being that for this analysis we did not require a minimum ratio difference of 70% or higher, as we also included the CHG context in which differences are typically smaller. DMP frequency was established for cytosines in CG and CHG context in the gene related and non-gene related (other) clusters separately (see materials and methods). We limited DMP detection to cytosines being either methylated or non-methylated symmetrically (on both Watson and Crick strand both below or both higher than 5%) in CG or CHG context. For positions to be considered in the DMP analysis we applied the same criteria as used for detection of methylation (see Materials and methods “Methylation summary statistics figure 2b”). All calculations were done with the custom python script “all_DMP_detection.py” available in the github repository (<https://github.com/thomasvangurp/epiGBS>).

Non-conversion and false methylation rate estimate

We computed non-conversion rates based on non-methylated phage lambda DNA. A total of 8,475 cytosines of the Phage Lambda genome were covered with 100 or more reads. On average, 0.39% of cytosines per position were non-converted. This non-conversion rate is not expected to have an impact on our methylation estimates as in our analysis we require coverage to be higher than 10 and methylation rates to be higher than 5% in read pools of both forward and reverse informative reads (see description Methylation summary statistics figure 2b). Using this criterion, not a single cytosine in phage lambda was found to be methylated at a rate above 5%.

Code availability

All custom python scripts mentioned are available on github (<https://github.com/thomasvangurp/epiGBS>).

Accession codes

Raw reads have been deposited at the NCBI short read archive under accession PRJNA287755.

Acknowledgements

We thank F. Johannes (Department of Plant Science, Technical University of Munich, Munich, Germany) and C. Becker (Max Planck Institute for Developmental Biology, Tübingen, Germany) for providing *Arabidopsis thaliana* samples, and we thank G. Maes, J. van Houdt and E. Schijlen for sequencing several epiGBS libraries as we developed the protocol. *Daphnia magna* samples were provided by L. De Meester (University of Leuven, Leuven, Belgium). *Fallopia japonica* and *Fallopia × bohemica* samples were provided by C. Richards (University of South Florida, Tampa, Florida, USA). *Allium porrum* and *Daucus carota* seeds were provided by Nunhems Netherlands BV–Bayer CropScience. Finally, we thank C. Bock, C. Richards, R. Elshire, W. van der Putten and A. Biere for critical comments and helpful suggestions on the manuscript. This study was funded by the Netherlands Organization for Scientific Research (NWO-ALW grants 864.10.008 K.J.F.V. and 820.01.025 K.J.F.V.).

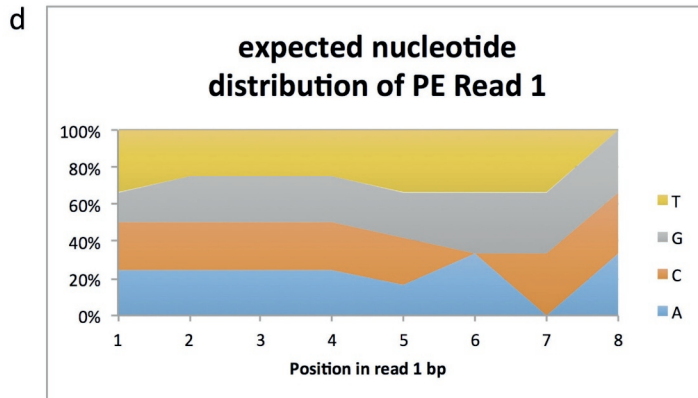
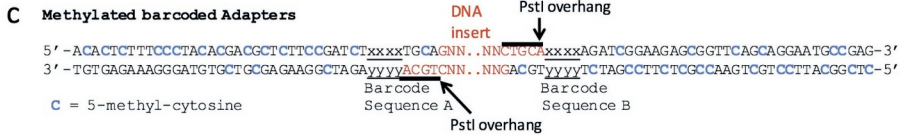
Author Contributions

T.P.v.G. conceived the method, performed lab work, developed software, analyzed data and wrote the paper. N.C.A.M.W. codeveloped the method, provided sample material, performed lab work and contributed to writing of the paper. B.W. codeveloped software and developed the RnBeads interface scripts. K.J.F.V. contributed to interpretation of the experiment and method and to writing of the paper. P.V. provided samples for the *Scabiosa* case study and contributed to interpretation of the experiment and method. J.N.J.O. contributed samples and to interpretation of the experiment and method.

Supporting information

Forward inline barcodes		Reverse inline barcodes	
Barcode A1	AACT	Barcode B1	AACT
Barcode A2	CCAG	Barcode B2	CCAG
Barcode A3	TTGA	Barcode B3	TTGA
Barcode A4	GGTA	Barcode B4	GGTA
Barcode A5	AACCA	Barcode B5	ATTG
Barcode A6	CCACG	Barcode B6	CGGT
Barcode A7	TTGTA	Barcode B7	TGCG
Barcode A8	GGTGG	Barcode B8	GTAT
Barcode A9	AATGCA		
Barcode A10	CCGTCA		
Barcode A11	TTACGA		
Barcode A12	TGCACA		

Barcoded adapter A top	5' - ACACTCTTTCCCTACACGACGCTCTTCCGATCTxxxTGCA -3'
Barcoded adapter A bottom	5' - yyyAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT -3'
Barcoded adapter B top	5' - xxxxAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAG -3'
Barcoded adapter bottom	5' - CTCGGCATTCTGCTGAACCGCTCTTCCGATCTyyyTGCA -3'
Illumina PE-PCR primer 1	5' - AATGATACGGGACCACCGAGATCTACACTCTTCCCTACACGACGCTCTTCCGATCT -3'
	5' -
Illumina PE-PCR primer 2	CAAGCAGAAGACGGCATACGAGATCGGTCTCGGCATTCTGCTGAACCGCTCTTCCGATCT -3'

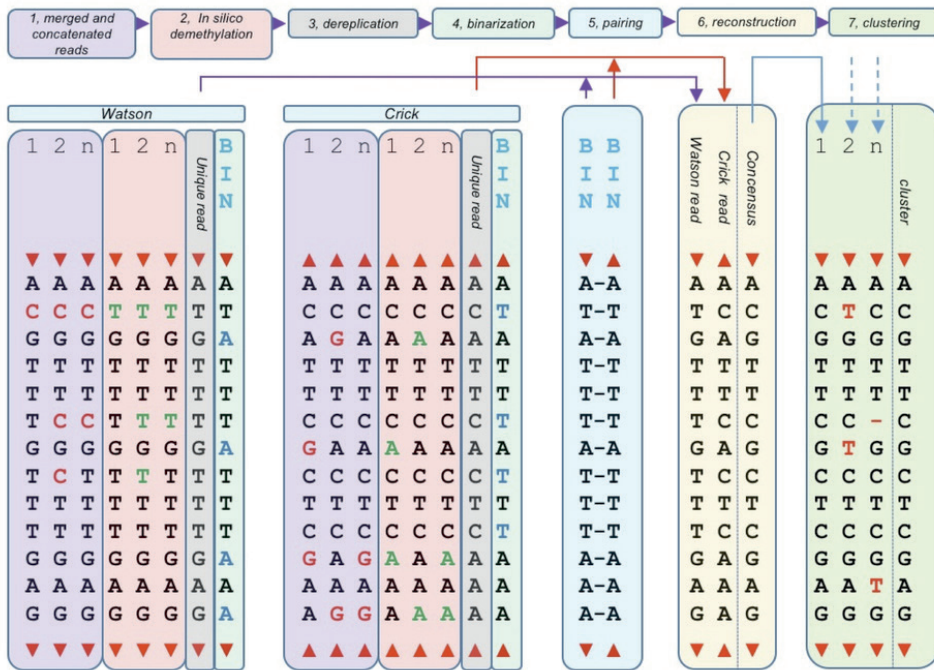


e

Species	Samples	Total reads	% merged reads	# clusters	CG content	Average size	Gene hits
<i>Allium porrum</i>	4	708,107	89%	7,291	40%	138	920
<i>Arabidopsis thaliana</i>	4	6,350,721	63%	19,318	57%	193	2,274
<i>Homo sapiens</i>	1	365,612	86%	2,641	49%	161	231
<i>Fallopia japonica & Fallopia x bohemica</i>	4	2,886,907	90%	13,745	45%	134	1,122
<i>Daucus carota</i>	4	2,471,270	92%	6,663	40%	144	1,487
<i>Daphnia magna</i>	6	890,587	82%	3,464	51%	155	583
<i>Scabiosa columbaria</i>	29	22,724,129	96%	33,604	59%	141	2,867
<i>Mimulus guttatus</i>	7	7,167,990	85%	63,184	58%	139	3,239
<i>Taraxacum officinale</i>	35	27,647,929	71%	33,517	55%	138	2,260
Phage lambda	1	1,289,004	87%	159	57%	237	-
	96	72,502,256					

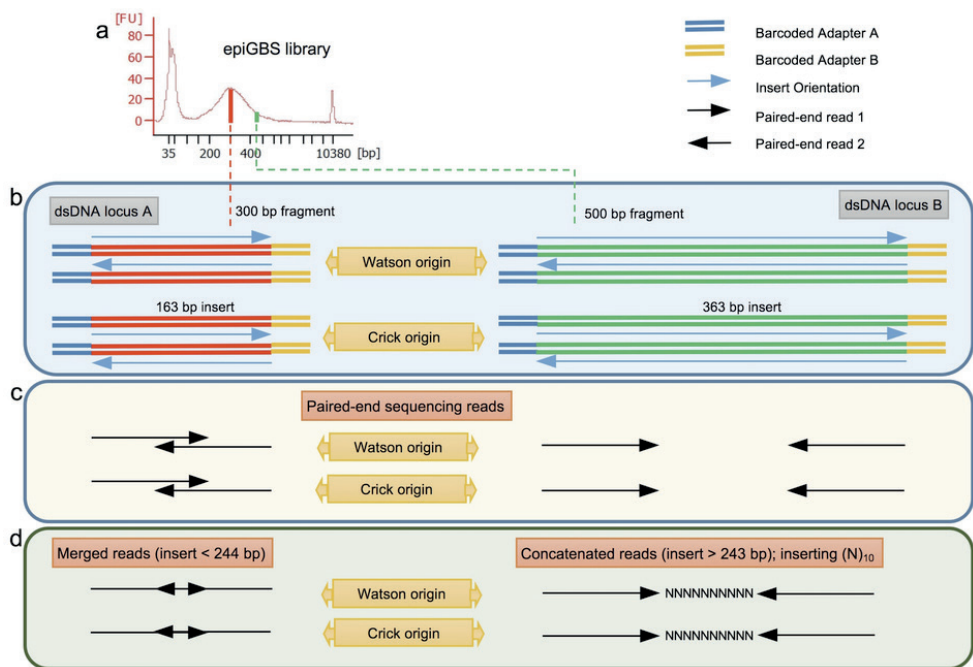
Supplementary Figure 3.1 Barcodes, barcoded adaptors and resulting expected nucleotide distributions and epiGBS clusters.

(a) The following forward and reverse inline barcodes were used to generate barcoded forward and reverse adapters (Supplementary Table 1b), (b) Barcoded adapter design is identical to the adapters used in Genotyping by Sequencing (Elshire et al. 2011) with the exception of the use of the B adapter barcode which is not present in GBS. Barcoded adapter sequences of both A and B adapters were generated using <http://www.deenabio.com/services/gbs-adapters> with 12 4-6 nucleotides barcodes for the A adapter and a subset of 8 4-nucleotide barcoded adapters were used for the B adapter. The sequence of Illumina PE-PCR primers 1 and 2 used to amplify the libraries are also listed. (c) Complimentary oligonucleotides with 5-methylcytosines instead of cytosines are annealed to form adapters. The four base 5'-3' overhang complements the restriction site overhang generated by PstI but can be modified depending on the specific enzyme used. Adapter A is identical to the barcoded adapter used in GBS. 4-6 nucleotide barcodes are designed to maximize diversity over the first cycles of the reads. Adapter B is identical to the common adapter used in GBS, with the exception that a barcoded sequence is placed before the enzyme overhang. For Csp6I we used 5'-CCTA-3' as forward and 5'-CTGG-3' as reverse barcode. The overhang of both Csp6I adapters was modified with respect to PstI, instead of 5'-TGCA-3' we designed adapters with 5'-AT-3' complementary to the overhang generated by Csp6I digestion. (d) Given equal representation of all 12 forward barcodes the expected per cycle nucleotide composition of the forward read is depicted. Up to position 5 the composition is mostly unbiased, aiding the calculation of proper run-specific parameters during Illumina sequencing and thus preventing phasing and pre-phasing detection errors that can lead to low quality sequencing libraries. (e) Per species the total number of paired-end sequencing reads, individuals per species, percentage of merged reads, number of de novo discovered clusters, CG content, average size and number of clusters having gene hits (see Materials and methods) is shown.



Supplementary figure 3.2 De novo reference creation.

De novo reference creation was performed at the species level. (1) paired-end reads are either merged using PEAR (Jiajie Zhang et al. 2014) or concatenated (See Supplementary Figure 3.3), after which Watson and Crick reads are split. For concatenated pairs only the ~120 nucleotides corresponding to the read length minus the barcode on both ends can be recovered. (2) Methylation variation is removed *in silico* by C to T conversion in Watson and G to A conversion in Crick reads. (3) Dereplication merges identical reads. (4) For each remaining Watson and Crick read a binary AT only representation is created by simultaneous C to T and G to A conversion (binarization). (5) Identical binary representative sequences are paired, yielding AT-only pairs containing both Watson and Crick reads. (6) The original Watson and Crick reads are paired using the AT-only pairs as template. Based on the aligned Watson Crick read-pair the original non-converted sequence is reconstructed. A and T are identical in the alignment, Watson G / Crick A yields a G whereas Watson T / Crick C yields a C. (7) Finally, consensus sequences are clustered with 95% identity to yield reference sequences, allowing for insertions, substitutions and partial matches.



Supplementary Figure 3.3 Creation of merged and concatenated reads.

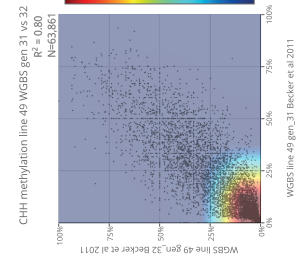
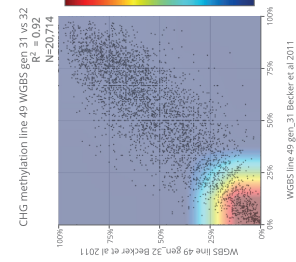
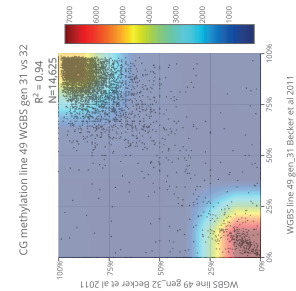
(a) A typical epiGBS library shows a peak on a Bioanalyzer image corresponding to fragments of ~300 basepairs (bp), although larger fragments are also obtained. A 300 bp fragment corresponds to an epiGBS fragment (insert) of 163 bp, as adapters are not sequenced. (b) Barcoded adapters A and B can be ligated on both ends of each fragment, so for any given genomic locus both forward and reverse oriented inserts exist. Fragments derive from either Watson or Crick strand. (c, d) Overlapping paired-end reads are merged whereas non-overlapping paired-end reads are concatenated to facilitate de novo reference reconstruction. Concatenation is performed by adding 10 N's between the forward and reverse read. Concatenated reads can be used to determine the sequence on both ends of inserts larger than ~1,9x the read length.

CHH

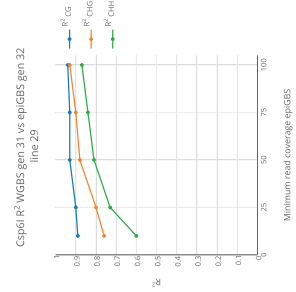
CHG

CG

a

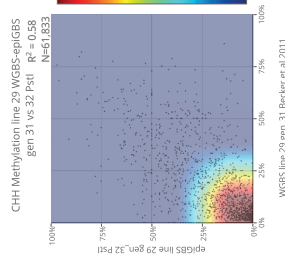
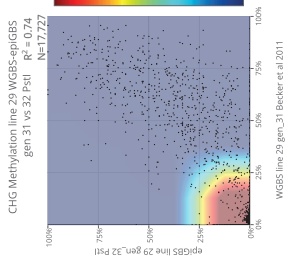
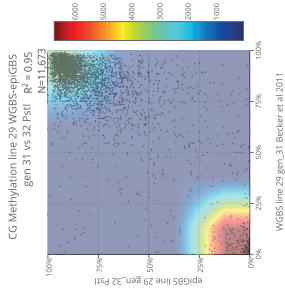


d

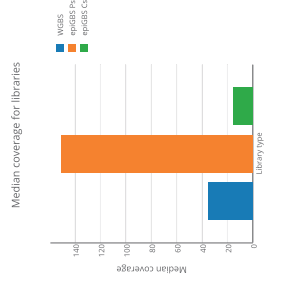


WGBS

b

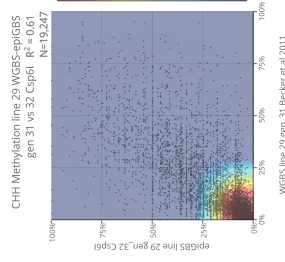
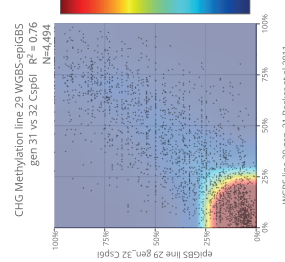
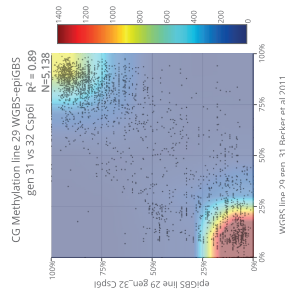


e

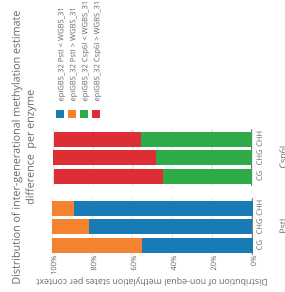


epiGBS
PstI

c



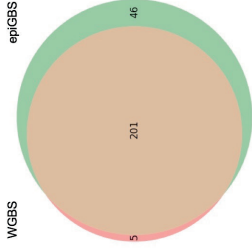
f



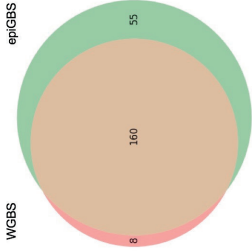
Supplementary Figure 3.4: Intergenerational methylation rate comparisons for different sequence contexts in epiGBS and WGBS.

Methylation rates are compared for CG, CHG and CHH context between generation 31 and 32 for WGBS vs WGBS data and WGBS vs epiGBS data. Only positions with 10 or more informative reads in both datasets were taken into account. **(a)** In the (Becker et al. 2011) WGBS data for line 49 we selected the subset of loci covered by PstI. Methylation rates between cytosines in generation 31 and 32 plants are well correlated. CG and CHG contexts have Pearson R^2 values of respectively 0.94 and 0.92 whereas rates in CHH context show a lower R^2 of 0.80. Heatmap legends on the right-hand side of all figures in panel a-c indicate the number of data points per 0.25*0.25 block in each scatter plot. **(b)** Methylation rates are compared in all sequence contexts between PstI-epiGBS data for line 29 generation 32 with the Becker et al WGBS data for generation 31. Compared to the WGBS vs WGBS correlations shown in panel a, the CG methylation correlates are equally good, whereas CHG and CHH correlations are slightly lower, with R^2 values of 0.74 and 0.59 respectively. The methylation sensitivity of PstI in CHG context could be causing this lower correlation, as CHG methylation in Arabidopsis is spatially auto-correlated (Cokus et al. 2008). Given partial methylation of CHG sites non-digestion of methylated PstI-sites could give rise to underestimating true methylation rates of cytosines inside these fragments as internal CHG and CHH methylation rates could be spatially correlated to methylation rates of the two CHG positions present in the PstI recognition site ('5- C_TGCA^G-3'). **(c)** To assess whether the methylation sensitivity of PstI contributed to the lower methylation estimates and lower R^2 for CHG and CHH context we have generated one additional epiGBS library with the methylation insensitive enzyme Csp6I for Arabidopsis line 29. Given the same coverage threshold (more than 10 informative reads) as used for the WGBS-WGBS and WGBS-PstI-epiGBS comparison shown in panel (a) and (b) we have obtained fewer positions with lower median coverage. This in turn leads to R^2 values for CG, CHG and CHH context of 0.89, 0.76 and 0.61. **(d)** To test whether the lower coverage of the Csp6I epiGBS library contributed to the lower R^2 values obtained compared to the WGBS data we calculated R^2 values between WGBS generation 31 and subsets of the generation 32 Csp6I epiGBS data with increasingly stringent minimum read coverage criteria. R^2 values increase asymptotically to the same levels as obtained for the WGBS vs WGBS comparisons made in panel a. Thus, given sufficient coverage, epiGBS estimates of methylation ratios with methylation insensitive enzymes are as accurate as WGBS based estimates. **(e)** Median coverage for positions in CHG context for WGBS data in line 49 gen 31 and for epiGBS data in line 29 gen 32 for PstI and Csp6I. Median coverage for other contexts was comparable to that of CHG. Median coverage was 151 in the epiGBS PstI library, whereas median coverage was 17 for the Csp6I data. **(f)** For positions having unequal inter-generational methylation rates we divided the positions in two groups per sequence context, one having a higher methylation estimate in epiGBS compared to WGBS whereas the other group shows the opposite pattern. For PstI, the group with positions showing a lower epiGBS rate is overrepresented for the CHG and CHH context. For Csp6I the groups are more or less of equal size, suggesting that the methylation sensitivity of PstI contributed to the overall lower methylation estimates obtained for CHG and CHH context.

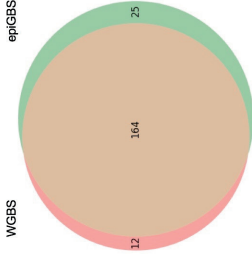
a) A89



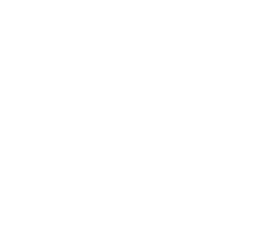
A109



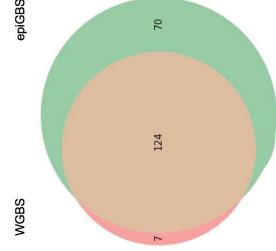
A119



A29

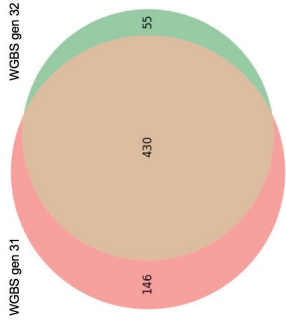


A89



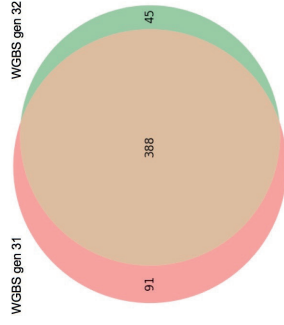
b)

A39



Average coverage gen 31 : 63
average coverage gen 32: 53

A49

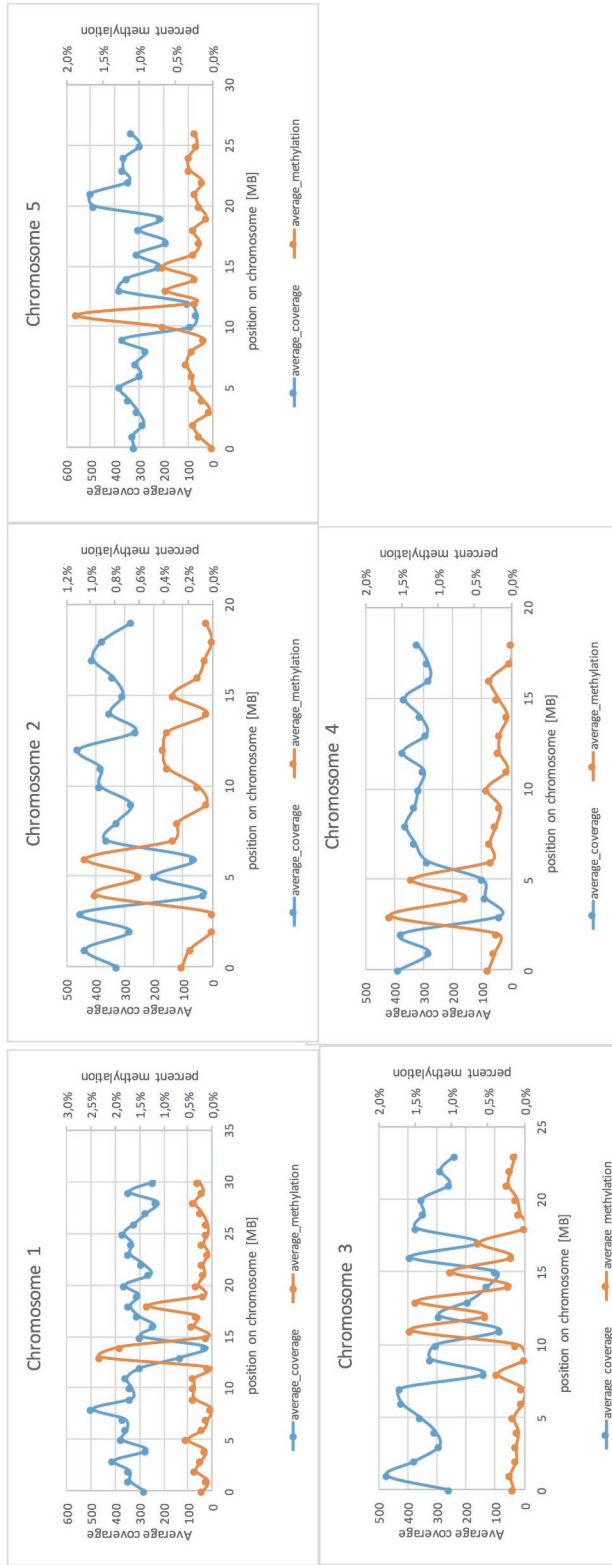


Average coverage gen 31 : 69
Average coverage gen 32: 45

A109

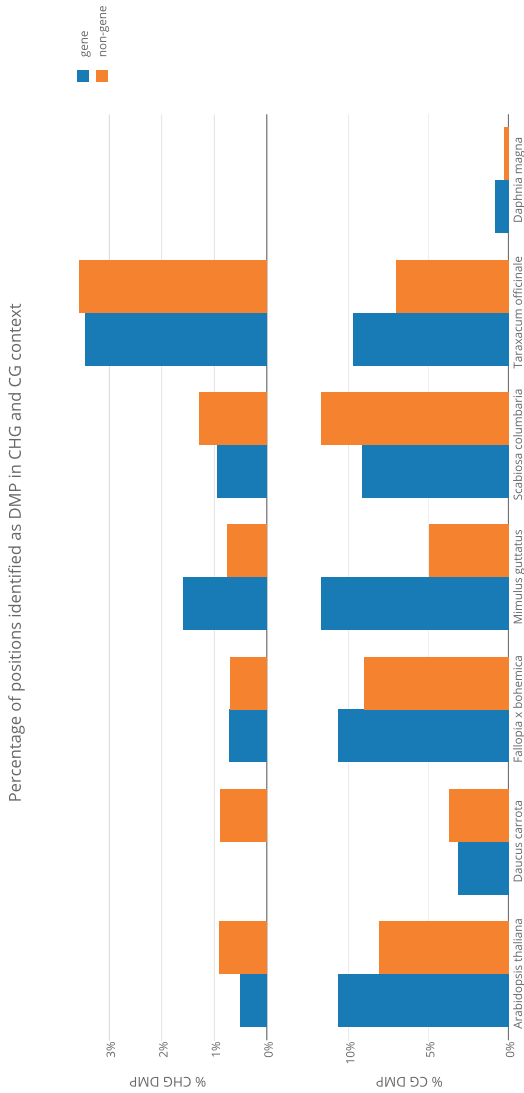
Supplementary Figure 3.5: Venn diagrams of DMPs detected in pairwise between-line comparisons of PstI-epiGBS versus WGBS data and of WGBS versus WGBS data.

Venn diagrams of DMPs detected in pairwise between-line comparisons **(a)** Pair-wise comparisons show the overlap and differences between DMPs detected between four genotypes using both WGBS generation 31 and epiGBS generation 32 data. Per diagram, only positions having sufficient informative read coverage in both WGBS and epiGBS data were taken into account. **(b)** Pair-wise comparisons show the overlap and differences between DMPs detected between the Arabidopsis line 39 and 49 using both WGBS generation 31 and WGBS generation 32 data. Per diagram, only positions having sufficient informative read coverage in both WGBS generation 31 and WGBS generation 32 data were taken into account.

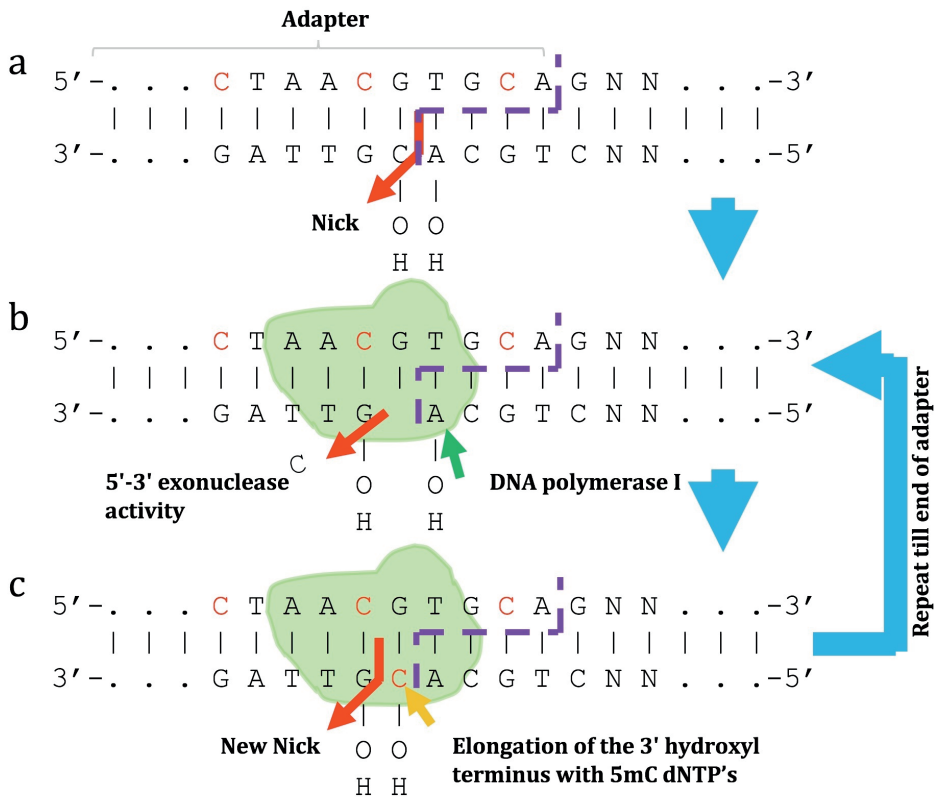


Supplementary Figure 6: Average coverage and methylation rate of the Arabidopsis thaliana genome as measured with PstI-epiGBS data.

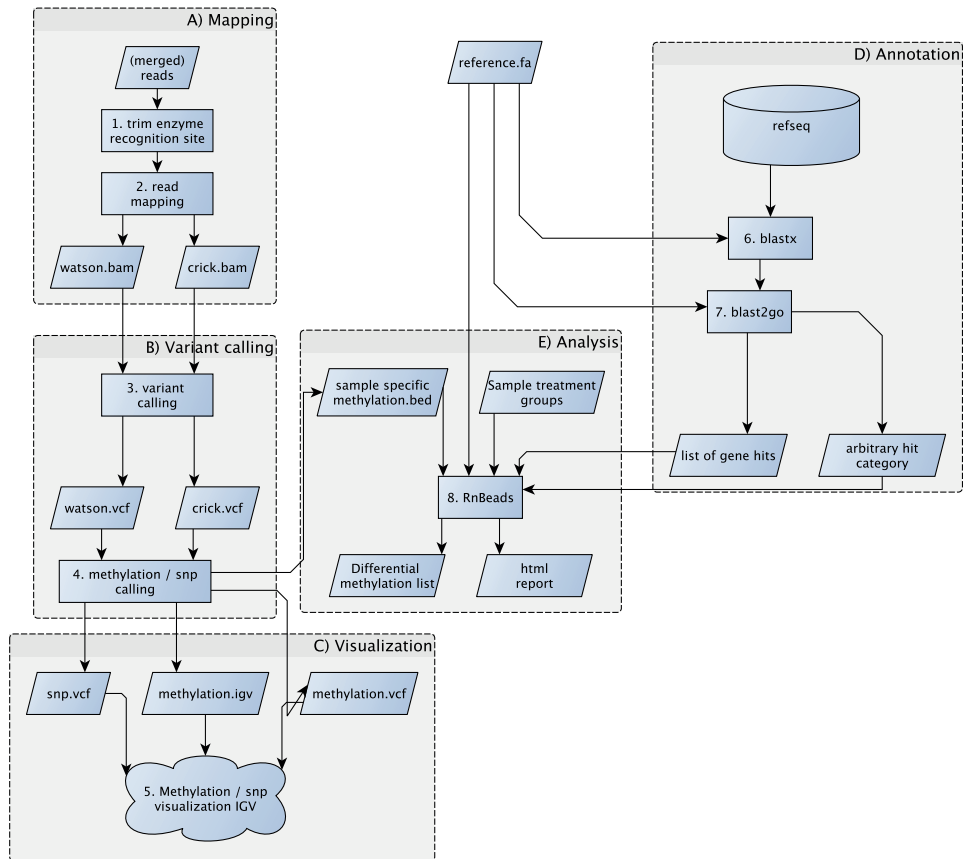
Per cytosine, average coverage and methylation rate were calculated per 1MB windows for all four Arabidopsis lines sequenced in PstI epiGBS for the Arabidopsis reference genome TAIR10 on chromosome 1-5. Only positions covered by at least 10 reads were taken into account. Average methylation is higher in centromeric regions but per cytosine coverage is lower in these regions, reflecting sensitivity of PstI to 5mC methylation.



Supplementary Figure 3.8: Percentage of positions identified as DMPs in CHG and CG contexts. The upper panel shows the percentage of cytosines in CHG context detected as DMP, whereas the lower panel shows the same for cytosines in CG context. Results are split for gene related and non-gene related clusters. Data for all positions per species is added as an excel file (Supplementary data 2 available online at <http://www.nature.com/nmeth/journal/v13/n4/extref/nmeth.3763-S3.xlsx>). DMP detection was carried out as described in materials and methods section "DMP detection in all species". For Arabidopsis we found that 11% of cytosines in CG context were differentially methylated in genes and 8% in non-genes. These values are lower than the 17.8% of cytosines in CG context reported as DMPs in (Becker et al. 2011). We suspect that our values are lower due to the stringent selection criteria requiring methylation to be present in both forward and reverse reads on both Watson and Crick strand. This might lead to under-estimation of actual DMP levels at lower sequencing coverage. Indeed, by selecting a higher minimum read coverage threshold of 50 forward and 50 reverse informative reads the percentage of CG dinucleotides found in genes was higher with 14.5 % and in non-genes 11.3 %. For most species, CG-DMPs are more prevalent in genes, with the exception of *Scabiosa columbaria* and *Daucus carota*, for which more DMPs were found in clusters not having significant blast hits to genes. CHG DMPs are absent in *Daphnia magna* and are most prevalent in *Taraxacum officinale*, consistent with the observations made for the absence of CHG methylation in *Daphnia magna* and the higher presence of CHG methylation in *Taraxacum officinale* (see Fig. 3.2b).



Supplementary Figure 3.9: Nick translation. (a) Due to the use of non-phosphorylated adapters, epiGBS libraries contain nicks between the 3' fragment overhang and the 5' non-phosphorylated adapter nucleotide (on the bottom strand in this Figure). To prevent the partial loss of adapter sequence during bisulfite treatment the nick was repaired by DNA polymerase I. DNA polymerase I is able to repair single stranded nicks in dsDNA in two steps. (b) first it removes the nucleotide (in this case C) at the 5' side of the nick by 5'-3' exonuclease activity after which (c) the 3' hydroxyl terminus is elongated with a single dNTP (in this case 5mC) using the nucleotide from the opposite strand (here G) as template. DNA polymerase I moves one nucleotide downstream towards the 3' end of the bottom adapter strand. Steps (b) and (c) are repeated until the 3' end of the adapter is reached, leaving no nicks in the adapter sequence.



A) Mapping

(1) Because part of the sequenced fragment originates from the methylated adapter sequence, this part needs to be excluded from the analysis. Therefore, the first 4 bases of merged and forward Watson reads as well as the last 4 of merged and first four of reverse Crick reads are removed.

(2) Reads are mapped against either de novo obtained or existing reference using BWA-METH (Pedersen et al. 2014) as it is more sensitive / accurate compared to similar bisulfite sequence-aligners and allows for easy transfer of read group tags from the sequence name to the bam output file (Pedersen et al. 2014).

B) Variant calling

(3) Variant calling is done using FreeBayes (Garrison & Marth 2012) separately for both Watson and Crick reads for all samples simultaneously. The settings force every position to be called for all samples. FreeBayes is used because it allows for indel realignment and sensitive variant calling.

(4) By simultaneously iterating over both Watson and Crick variant call files (VCF) SNPs and methylation polymorphisms can be distinguished. C/T polymorphisms in Watson combined with C on the Crick strand indicate a methylation polymorphism on the Watson strand whereas a G/A polymorphism on Crick combined with a G on the Watson

strand indicates a methylation polymorphism on the Crick strand. Where combined SNP and methylation lead to a C/T or G/A polymorphism on both Watson and Crick strand only the SNP is called as the methylation ratio cannot be determined.

C) Visualization

(5) SNPs and methylation polymorphisms are exported in VCF format. Furthermore, methylation ratios are calculated and exported in a tab separated IGV specific format (.IGV) including the context (CG, CHG or CHH). Datasets for all species studied are available on [genomespace](https://gsui.genomespace.org) (see <https://gsui.genomespace.org/jsui/gsui.html?pathOrUrl=/Home/thomasvangurp/epiGBS%20Nature%20Methods/>).

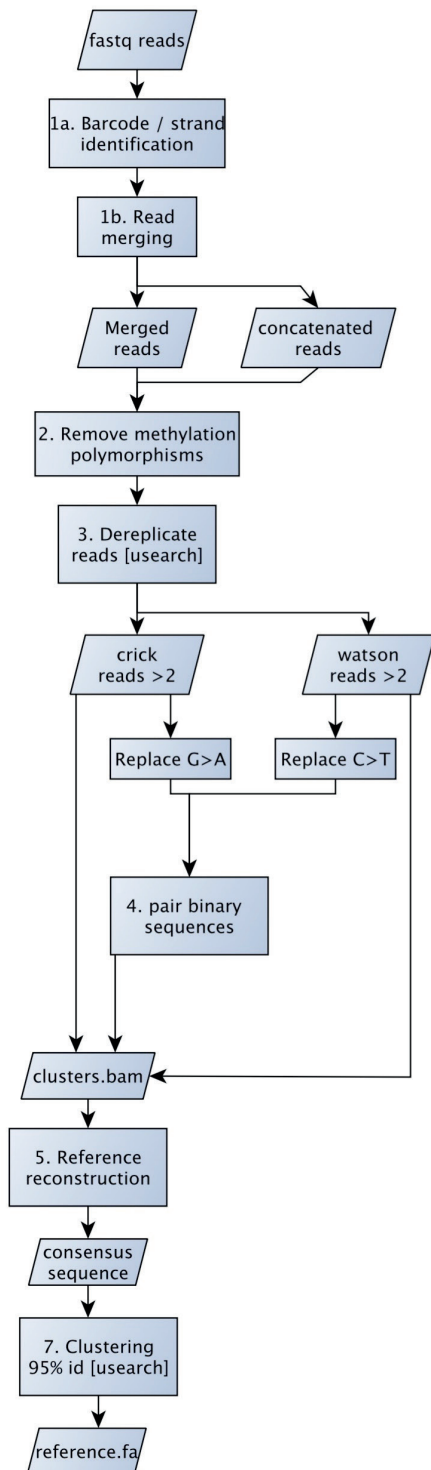
D) Annotation

(6) Usearch blastx (Edgar 2010) is performed against reference protein sequences related to the species sequenced.

(7) Resulting blastx reads are imported into blast2go for mapping to gene ontology terms and enzyme codes. A list of all annotated genes is exported. Optionally, a list of contigs mapping to specific genes or with specific GO terms can be exported allowing for a focused analysis in RnBeads (Assenov et al. 2014).

E) Analysis

(8) A pipeline for processing the sample specific methylation bed files and experimental details like treatment and/or sample groups allows for comparing differential methylation between combinations of treatments, including all possible 2-way interactions between treatment groups using RnBeads (Assenov et al. 2014) or other methylation analysis tools.



Supplementary Figure 11: De novo clustering algorithm.

1)
 a) Paired-end reads contain in-line barcodes in both forward and reverse reads. Barcodes are removed while sample specific tags are added to the read name. The original strand, Watson or Crick, is determined based on the (non)conversion of the enzyme recognition site at both reads (Fig. 3.1b).

b) Overlapping paired-end reads are merged using Pear (Jiajie Zhang et al. 2014) to (i) avoid adapter remnants for reads derived from fragment with an insert size shorter than the number of read cycles as they can lead to false polymorphisms or mapping errors, (ii) avoid double calling of bases from the same molecule in the overlapping region of the pair, and (iii) lowering the error rate by selecting the highest quality base from the two reads in the overlapping region. Non-overlapping reads are concatenated to allow for matching Watson and Crick pairs (Supplementary Fig. 3.3).

2) Methylation polymorphisms are removed by converting all C's in Watson to T and all G's in Crick to A to allow for clustering of same-origin reads with methylation polymorphisms.

3) Identical Watson and Crick reads are dereplicated using usearch (Edgar 2010). Only Watson and Crick reads that occur twice or more are retained to avoid sequencing errors.

4) Binarization: To aid pairing of same-origin Watson and Crick reads all C's are converted to T's in Watson reads while G's are converted to A's in Crick reads rendering a binary AT only sequence

output, that is identical for same origin Watson and Crick reads (see Fig. 3.1b).

5) Using usearch dereplication binary AT only watson and crick reads are paired.

6) Based on the alignment and orientation of the binary Watson and Crick reads the original unconverted Watson and Crick reads are aligned in a bam file. The reference sequence of each pair is called using combined bash piped queries using the following components of samtools (H. Li et al. 2009): mpileup and bcftools (H. Li et al. 2009) for variant calling followed by vcfutils.pl vcf2fq for creating the Watson and Crick consensus sequence. These consensus sequences are used to recreate the original sequence. In case of Watson:T / Crick:C a C is added, whereas in case of Watson:G / Crick:A a G is added. In all other cases where nucleotides do not match an ambiguous nucleotide (N) is added.

7) Consensus sequences with 95 % or higher identity are clustered, leading to the final reference sequence.

Sample Description										
Species	Common name	genotype	treatment	tissue	age	ng DNA used	Chr/ploidy	1C (Mbp)	# samples	DNA Origin
<i>Arabidopsis thaliana</i>	thale cress	columbia line 29,119,109 and 89	none	young leaves	4 weeks after sowing	400 ng	10	156	4	NucleoSpin® 8 Plant II kit from Macherery-Nagel
<i>Allium porrum</i> , hybrid	Leak	line 1 and 2	control or genistein (40 microM)	young leaves	4 weeks after sowing	400 ng	32	29340	4	NucleoSpin® 8 Plant II kit from Macherery-Nagel
<i>Taraxacum officinalis</i>	common dandelion	Homebred lines	none	young leaves		400 ng	24	1247	35	NucleoSpin® 8 Plant II kit from Macherery-Nagel
<i>Daphnia magna</i>	Water flea	variable	none	whole animal	adult	400 ng	20	469	6	NucleoSpin® 8 Plant II kit from Macherery-Nagel
<i>Fallopia japonica</i>	Japanese knotweed sensu lato	417_Fallopia- CMM 15 haplotype F	none	young leaves	young	400 ng	66	3389	2 pools	NucleoSpin® 8 Plant II kit from Macherery-Nagel
<i>Fallopia x bohemica</i>	Japanese knotweed sensu lato	297_Fallopia- RHC 1 haplotype G	none	young leaves	young	400 ng	88	4714	2 pools	NucleoSpin® 8 Plant II kit from Macherery-Nagel
<i>Homo sapiens</i>	Human	Promega (G1521) female Human	none	unknown	unknown	400 ng	46	3423	1	Human Genomic DNA (G1521, Promega)
Phage lambda	Phage lambda	λ DNA c1857 Sam7	none	whole phage	unknown	400 ng	-	0.048	1	λ DNA c1857 Sam7 (D1501, Promega)
<i>Daucus carota</i>	carrot	line 1 and 2	control or genistein (40 microM)	young leaves	4 weeks after sowing	400 ng	18	978	4	NucleoSpin® 8 Plant II kit from Macherery-Nagel
<i>Mimulus guttatus</i>	Common monkey-flower	population 'Rozendaa1'	variable	young leaves	4 weeks after sowing	400 ng	14	362	7	NucleoSpin® 8 Plant II kit from Macherery-Nagel
Scabiosa	small scabious	FR-3-4	variable	young leaves	3 months	400 ng	16	1048	29	NucleoSpin® 8 Plant II kit from Macherery-Nagel

Supplementary Data 1 | Species level descriptions

sample nr	Sample Description							BARCODES	
	Species	genotype'	treatment	tissue	age	ng DNA used	CO	BA	
44	Arabidopsis thaliana	a29	none	leaf	4 weeks after sowing	400 ng	GGTA	CCACG	
45	Arabidopsis thaliana	a109	none	leaf	4 weeks after sowing	400 ng	ATTG	CCACG	
46	Arabidopsis thaliana	a89	none	leaf	4 weeks after sowing	400 ng	CGGT	CCACG	
47	Arabidopsis thaliana	a119	none	leaf	4 weeks after sowing	400 ng	TGCG	CCACG	
71	Allium porrum, hybrid	line 1	control	leaf	4 weeks after sowing	400 ng	TGCG	AATGCA	
73	Allium porrum, hybrid	line 1	genistein (40 µM)	leaf	4 weeks after sowing	400 ng	AACT	CCGTCA	
74	Allium porrum, hybrid	line 2	genistein (40 µM)	leaf	4 weeks after sowing	400 ng	CCAG	CCGTCA	
75	Allium porrum, hybrid	line 2	control	leaf	4 weeks after sowing	400 ng	TTGA	CCGTCA	
1	Taraxacum officinalis	A34 JS 1.5.3.4	none	leaf	Mature plants after seeding	400 ng	AACT	AACT	
2	Taraxacum officinalis	A68 JS 5.2.1.1	none	leaf	Mature plants after seeding	400 ng	CCAG	AACT	
3	Taraxacum officinalis	A68 JS 5.2.1.2	none	leaf	Mature plants after seeding	400 ng	TTGA	AACT	
4	Taraxacum officinalis	A68 AS 3.2.4.2	none	leaf	Mature plants after seeding	400 ng	GGTA	AACT	
5	Taraxacum officinalis	A68 CS 3.5.3.4	none	leaf	Mature plants after seeding	400 ng	ATTG	AACT	
6	Taraxacum officinalis	A68 CS 3.7.1.4	none	leaf	Mature plants after seeding	400 ng	CGGT	AACT	
7	Taraxacum	A68 CS 6.6.2.4	none	leaf	Mature plants after seeding	400 ng	TGCG	AACT	

	officinalis		seeding					
8	Taraxacum officinalis	A68 AR 8.5.4.4	none	leaf	Mature plants after seeding	400 ng	GTAT	AACT
17	Taraxacum officinalis	A68 JR 8.6.1.1	none	leaf	Mature plants after seeding	400 ng	AACT	TTGA
18	Taraxacum officinalis	A68 JR 5.8.4.4	none	leaf	Mature plants after seeding	400 ng	CCAG	TTGA
19	Taraxacum officinalis	A68 JR 5.5.2.3	none	leaf	Mature plants after seeding	400 ng	TTGA	TTGA
20	Taraxacum officinalis	A68 CR 3.6.4.3	none	leaf	Mature plants after seeding	400 ng	GGTA	TTGA
21	Taraxacum officinalis	A34 AS 5.5.2.2	none	leaf	Mature plants after seeding	400 ng	ATTG	TTGA
22	Taraxacum officinalis	A34 CR 8.4.2.4	none	leaf	Mature plants after seeding	400 ng	CGGT	TTGA
23	Taraxacum officinalis	A34 JR 8.3.1.3	none	leaf	Mature plants after seeding	400 ng	TGCG	TTGA
24	Taraxacum officinalis	A34 AR 2.3.2.3	none	leaf	Mature plants after seeding	400 ng	GTAT	TTGA
25	Taraxacum officinalis	A34 CS 2.7.2.3	none	leaf	Mature plants after seeding	400 ng	AACT	GGTA
26	Taraxacum officinalis	A34 JR 7.6.4.4	none	leaf	Mature plants after seeding	400 ng	CCAG	GGTA
27	Taraxacum officinalis	A34 CR 2.3.1.2	none	leaf	Mature plants after seeding	400 ng	TTGA	GGTA
28	Taraxacum officinalis	A34 AC 2.3.2.1	none	leaf	Mature plants after seeding	400 ng	GGTA	GGTA
29	Taraxacum officinalis	A68 JS 7.2.4.2	none	leaf	Mature plants after seeding	400 ng	ATTG	GGTA
30	Taraxacum officinalis	A34 CR 2.3.1.3	none	leaf	Mature plants after seeding	400 ng	CGGT	GGTA
31	Taraxacum officinalis	A34 AR 8.7.2.3	none	leaf	Mature plants after seeding	400 ng	TGCG	GGTA

32	Taraxacum officinalis	A68 AS 3.2.4.3	none	leaf	Mature plants after seeding	400 ng	GTAT	GGTA
33	Taraxacum officinalis	A68 CR 2.6.3.3	none	leaf	Mature plants after seeding	400 ng	AACT	AACCA
34	Taraxacum officinalis	A68 CS 6.6.4.1	none	leaf	Mature plants after seeding	400 ng	CCAG	AACCA
35	Taraxacum officinalis	A68 CR 5.7.1.4	none	leaf	Mature plants after seeding	400 ng	TTGA	AACCA
36	Taraxacum officinalis	A34 CS 2.7.3.3	none	leaf	Mature plants after seeding	400 ng	GGTA	AACCA
37	Taraxacum officinalis	A34 AS 5.5.2.4	none	leaf	Mature plants after seeding	400 ng	ATTG	AACCA
38	Taraxacum officinalis	A34 JS 6.7.1.3	none	leaf	Mature plants after seeding	400 ng	CGGT	AACCA
39	Taraxacum officinalis	A34 JR 3.5.4.4	none	leaf	Mature plants after seeding	400 ng	TGCG	AACCA
40	Taraxacum officinalis	A68 AR 8.5.4.1	none	leaf	Mature plants after seeding	400 ng	GTAT	AACCA
41	Taraxacum officinalis	A34 CS 2.7.3.1	none	leaf	Mature plants after seeding	400 ng	AACT	CCACG
42	Taraxacum officinalis	A68 AS 3.2.4.1	none	leaf	Mature plants after seeding	400 ng	CCAG	CCACG
43	Taraxacum officinalis	A34 JS 2.4.2.2	none	leaf	Mature plants after seeding	400 ng	TTGA	CCACG
87	Daphnia magna	3_6_out_clone_8_F1	none	whole animal	adult	400 ng	TGCG	TTACGA
88	Daphnia magna	3_8_inb_famNF_2_F1	none	whole animal	adult	400 ng	GTAT	TTACGA
91	Daphnia magna	2_7_inb_famC.1	none	whole animal	adult	400 ng	TTGA	TGCACA
92	Daphnia magna	2_8_inb_famC.2	none	whole animal	adult	400 ng	GGTA	TGCACA
93	Daphnia magna	2_10_out_famD.1	none	whole	adult	400 ng	ATTG	TGCACA

				animal			
94	<i>Daphnia magna</i>	2_12_out_famD.3	none	whole animal	adult	400 ng	CGGT TGCACA
10	<i>Fallopia x bohemica</i>	301_Fallopia- RHC 10 haplotype G	none	leaf	young	400 ng	CCAG CCAG
83	<i>Fallopia x bohemica</i>	314_Fallopia- RHC 3 haplotype G	none	leaf	young	400 ng	TTGA TTACGA
89	<i>Fallopia japonica</i>	417_Fallopia- CMM 15 haplotype F	none	leaf	young	400 ng	AACT TGCACA
90	<i>Fallopia x bohemica</i>	297_Fallopia- RHC 1 haplotype G	none	leaf	young	400 ng	CCAG TGCACA
95	<i>Homo sapiens</i>	Promega (G1521) female Human	none	unknown	unknown	400 ng	TGCG TGCACA
48	<i>Parus major</i>	unknown	none	unknown	unknown	400 ng	GTAT CCAG
96	<i>Phage lambda</i>	λ DNA cl857 Sam7	none	whole phage	unknown	400 ng	GTAT TGCACA
72	<i>Daucus carota</i>	line 2	genistein (40 μM)	leaf	4 weeks after sowing	400 ng	GTAT AATGCA
76	<i>Daucus carota</i>	line 1	genistein (40 μM)	leaf	4 weeks after sowing	400 ng	GGTA CCGTCA
77	<i>Daucus carota</i>	line 2	Control	leaf	4 weeks after sowing	400 ng	ATTG CCGTCA
78	<i>Daucus carota</i>	line 1	Control	leaf	4 weeks after sowing	400 ng	CGGT CCGTCA
79	<i>Mimulus guttatus</i>	population 'Rozenaal'	Drought	leaf	4 weeks after sowing	400 ng	TGCG CCGTCA
80	<i>Mimulus guttatus</i>	population 'Rozenaal'	Control	leaf	4 weeks after sowing	400 ng	GTAT CCGTCA
81	<i>Mimulus guttatus</i>	population 'Rozenaal'	Control	leaf	4 weeks after sowing	400 ng	AACT TTACGA
82	<i>Mimulus guttatus</i>	population 'Rozenaal'	Drought	leaf	4 weeks after sowing	400 ng	CCAG TTACGA

84	Mimulus guttatus	population 'Rozendaal'	Waterlogged	leaf	4 weeks after sowing	400 ng	GGTA	TTACGA
85	Mimulus guttatus	population 'Rozendaal'	Submerged	leaf	4 weeks after sowing	400 ng	ATTG	TTACGA
86	Mimulus guttatus	population 'Rozendaal'	Submerged	leaf	4 weeks after sowing	400 ng	CGGT	TTACGA
9	Scabiosa	FR-3-4-1-S	calcareous+5azacytidine (50µM)	leaf	3 months	400 ng	AACT	CCAG
11	Scabiosa	FR-3-4-6-S	acidic	leaf	3 months	400 ng	TTGA	CCAG
12	Scabiosa	FR-3-4-1-S	acidic	leaf	3 months	400 ng	GGTA	CCAG
13	Scabiosa	FR-3-4-6-S	calcareous	leaf	3 months	400 ng	ATTG	CCAG
14	Scabiosa	FR-3-4-29-C	acidic+5azacytidine (50µM)	leaf	3 months	400 ng	CGGT	CCAG
15	Scabiosa	FR-3-4-4-S	calcareous	leaf	3 months	400 ng	TGCG	CCAG
16	Scabiosa	FR-3-4-4-S	acidic	leaf	3 months	400 ng	GTAT	CCAG
49	Scabiosa	FR-3-4-1-S	calcareous	leaf	3 months	400 ng	AACT	TTGTA
50	Scabiosa	FR-3-4-8-C	calcareous	leaf	3 months	400 ng	CCAG	TTGTA
51	Scabiosa	FR-3-4-35-C	calcareous+5azacytidine (50µM)	leaf	3 months	400 ng	TTGA	TTGTA
52	Scabiosa	FR-3-4-3-S	calcareous	leaf	3 months	400 ng	GGTA	TTGTA
53	Scabiosa	FR-3-4-3-C	calcareous	leaf	3 months	400 ng	ATTG	TTGTA
54	Scabiosa	FR-3-4-26-C	acidic+5azacytidine (50µM)	leaf	3 months	400 ng	CGGT	TTGTA
55	Scabiosa	FR-3-4-5-C	acidic	leaf	3 months	400 ng	TGCG	TTGTA
56	Scabiosa	FR-3-4-2-S	acidic	leaf	3 months	400 ng	GTAT	TTGTA
57	Scabiosa	FR-3-4-2-C	calcareous	leaf	3 months	400 ng	AACT	GGTGG
58	Scabiosa	FR-3-4-6-C	calcareous	leaf	3 months	400 ng	CCAG	GGTGG
59	Scabiosa	FR-3-4-4-C	acidic	leaf	3 months	400 ng	TTGA	GGTGG
60	Scabiosa	FR-3-4-18-C	acidic+5azacytidine (50µM)	leaf	3 months	400 ng	GGTA	GGTGG

61	Scabiosa	FR-3-4-1-S	acidic+5azacytidine (50µM)	leaf	3 months	400 ng	ATTG	GGTGG
62	Scabiosa	FR-3-4-2-S	calcarious+5azacytidine (50µM)	leaf	3 months	400 ng	CGGT	GGTGG
63	Scabiosa	FR-3-4-7-C	acidic	leaf	3 months	400 ng	TGCG	GGTGG
64	Scabiosa	FR-3-4-36-C	calcarious+5azacytidine (50µM)	leaf	3 months	400 ng	GTAT	GGTGG
65	Scabiosa	FR-3-4-5-S	calcarious	leaf	3 months	400 ng	AACT	AATGCA
66	Scabiosa	FR-3-4-6-C	acidic	leaf	3 months	400 ng	CCAG	AATGCA
67	Scabiosa	FR-3-4-5-C	calcarious	leaf	3 months	400 ng	TTGA	AATGCA
68	Scabiosa	FR-3-4-17-C	calcarious+5azacytidine (50µM)	leaf	3 months	400 ng	GGTA	AATGCA
69	Scabiosa	FR-3-4-12-C	calcarious+5azacytidine (50µM)	leaf	3 months	400 ng	ATTG	AATGCA
70	Scabiosa	FR-3-4-21-C	acidic+5azacytidine (50µM)	leaf	3 months	400 ng	CGGT	AATGCA

Supplementary Data 1 | Sample level descriptions

Supplementary data 2 | DMPs of all species is available online:

<http://www.nature.com/nmeth/journal/v13/n4/extref/nmeth.3763-S3.xlsx>

Supplementary data 3 Data for Supplementary Figures is available online at

<http://www.nature.com/nmeth/journal/v13/n4/extref/nmeth.3763-S4>.

**Transgenerational inheritance of (stress-induced) DNA methylation
in apomictic dandelions.**

T.P. van Gurp, K.J.F. Verhoeven

To be submitted

4

Abstract

Environmental perturbation or biotic stress by insect herbivory can result in DNA methylation changes in plants. Such stress-induced DNA methylation changes can induce a plant defense response against the herbivore and as such provide a mechanism for a systemic response by which gene activity is altered. DNA methylation patterns show random changes between generations, but both specific (targeted) and non-specific (random) changes that are environmentally induced can add to this baseline level. The stability of induced DNA methylation changes has hitherto not been studied in detail outside of the model plant species *Arabidopsis thaliana*. Apomictic dandelions are an interesting system to investigate the extent and stability of (induced) DNA methylation changes that are transmitted transgenerationally. As part of the DNA methylation changes can result from genetic changes that occur in sexual plant after meiosis, the apomictic reproduction mode in the dandelions used avoids this complication. Here, I aim to establish the extent to which heritable DNA methylation changes are the result of random processes or of 5-Azacytidine and Jasmonic acid treatment. I do this by tracking DNA methylation on a subset of positions covered by bisulfite sequencing reads on restriction-site associated set of DNA fragments using epiGBS. I find limited but clear evidence for environmental induction of heritable DNA methylation changes after JA application. Also, I characterize the stability of overall levels of DNA methylation of cytosines in CHG (where H is A, C or T) and CG sequence context over generational time. I find a significant negative relation between the similarity of DNA methylation levels and intergenerational distance. I conclude that DNA methylation in both CG and CHG context are heritable and that environmental perturbation can result in heritable DNA methylation changes which are not widespread.

4.1 Introduction

DNA methylation in the form of 5-methyl cytosine (5mC) is a well-studied epigenetic modification which is implicated in cell differentiation and gene regulation in eukaryotes. Mitotically heritable changes in DNA methylation have been observed in plants after exposure to stress. In some cases, these changes can also be transmitted meiotically, which hereafter I will refer to as environment-induced epigenetic inheritance. Epigenetic inheritance has received much interest for its potential role in rapid and reversible adaptation to changing environments experienced by species. As such, studying epigenetic variation is highly relevant in the study of ecology and evolution (Bossdorf et al. 2007). Currently, important open questions about the role of DNA methylation in ecology and evolution concern the extent of variation observed in natural plant methylomes, how this variation effects phenotypic variation, the degree of dependence of methylation variation on the genotype and if it contributes to evolutionary processes (Niederhuth & Schmitz 2014). So far, the number of studies focusing on the transgenerational inheritance of DNA methylation in plants at single nucleotide resolution are almost exclusively limited to *A. thaliana* and *Zea mays* (Regulski et al. 2013), often in (artificial epigenetic) recombinant Inbred lines (Schmitz, He, et al. 2013).

To understand the scope of epigenetics in ecology and evolution, further study of transgenerational inheritance of DNA methylation in a broader range of species is relevant for several reasons. First, DNA methylation levels are highly variable in different plant species, which to some extent can be explained by differences in genomic repeat content, which positively impact the methylation rate (Niederhuth, Bewick, Ji, Alabady, Kim, Li, et al. 2016). Many examples of phenotypic variation caused by heritable DNA methylation variants are found in non-model species, but they are poorly characterized (Verhoeven et al. 2016) which initially requires establishing their transgenerational stability. Second, the potential relevance of epigenetic variation may be different for species with different life history characteristics. For instance, relevance of epigenetic variation may be higher for asexual compared to sexual plant species, as asexual plants cannot generate genetic variation through recombination. Asexual reproduction also allows for bypassing epigenetic reprogramming mechanisms during meiosis, which could affect the stability of (induced) DNA methylation variation (Verhoeven & Preite 2014). Asexually reproducing plants like the apomictic dandelion *Taraxacum officinale* therefore offer an interesting system, allowing the study of epigenetic variation in a natural system devoid of interference by genetic variation.

DNA methylation in plants occurs in three different sequence contexts, CG, CHG and CHH (where H is A, C or T) which are controlled by different enzymes. Symmetric CG methylation is maintained faithfully during mitosis and often meiosis (Dalakouras et al. 2012). The relationship between epigenetic variation and underlying DNA sequence variation varies; often epigenetic variants are under direct genetic control (obligatory) or can only arise given a certain genotype (facilitated). Only pure epigenetic variation is truly independent on genetic variation, which may arise for instance due to imperfect copying of methylation patterns during cell divisions (E. J. Richards 2006). Assessing the degree of independence of epigenetic from genetic variation is important, as obligatory epigenetic variation only reflects genotypic variation. To understand the potential additional role in

ecology and evolution of epigenetic variation compared to genetic variation, researchers often try to focus on (partially) independent facilitated or pure epigenetic variation (Hauben et al. 2009; Cortijo et al. 2014; Johannes et al. 2009). For this purpose researchers often use inbred lines, or even inbred lines with artificially altered DNA methylation (Cortijo et al. 2014; Johannes et al. 2009). For example, epigenetic recombinant inbred lines (epiRILs) were created in *A. thaliana*, in which only the DNA methylation is different between genetically uniform individuals of the epiRIL population (Johannes et al. 2009; Cortijo et al. 2014). Absence of genetic differences permits observed epigenetic differences to be interpreted as 'pure'. *Arabidopsis thaliana* epiRILs show variation and high heritability for a number of quantitative traits as well as stable inheritance of a considerable part of the parental DNA methylation variants (epialleles) over at least eight generations (Johannes et al. 2009). These studies provide some of the best evidence that heritable, pure epigenetic variation can cause heritable phenotypic differences (Cortijo et al. 2014).

If epigenetic variation is heritable and impacts the phenotype of both current and subsequent generations, then there is potential for heritable epigenetic variation to be subject to natural selection and to drive differentiation. The reversible nature and lower stability of epigenetic variation limits the potential of epigenetic variation to contribute to a sustained selection response. However, modeling studies show that even under relatively high reversion rates there is a considerable impact of epigenetic variation on the dynamics of adaptation (Kronholm & Collins 2016). Multi-generation stability of naturally arising DNA methylation variation has been demonstrated for *A. thaliana* using isogenic (epi)mutation accumulation lines, with estimates for epimutation rates for methylation in CG context of $2.56 \cdot 10^{-4}$ for forward and $6.30 \cdot 10^{-4}$ for backward mutation rates (van der Graaf et al. 2015). Simple modelling studies indicate that these rates are high enough to rapidly uncouple genetic from epigenetic variation, but low enough for new epialleles to sustain long-term selection responses (van der Graaf et al. 2015). This suggests that heritable DNA methylation variation of cytosines in CG context can underpin selective phenotypic divergence.

DNA methylation differences do not only define different cellular types and functions (Widman et al. 2014), but it can also change on an organism-wide scale as the result of environmental perturbations or stress (Secco et al. 2015). These changes can impact the phenotype or stress response later in life (same-generation) or even be transgenerationally transmitted to subsequent generations. For instance, cold stress in maize can induce specific demethylation of the DNA associated with the nucleosome (part of the chromatin structure) of *ZmMI1*, inducing altered expression of this retrotransposon related gene by changing the chromatin structure. This hypomethylated state remained for up to seven days under normal temperatures (Steward et al. 2002), suggesting that the effect maintained for some time in the absence of cold-stress. In tobacco it was shown that several types of environmental stressors such as aluminium, salt, cold and oxidative stresses can induce DNA demethylation of the promotor region of a glycerophosphodiesterase gene (*NtGPD*), correlating with increased *NtGPD* expression (Choi & Sano 2007). Environmentally induced epigenetic changes can sometimes also be passed transgenerationally and as such constitute a potential mechanism for transgenerational phenotypic plasticity (Holeski et al. 2016). In *A. thaliana*, hyperosmotic stress was shown to target (de)methylation of distinct

Chapter 4 Transgenerational inheritance of (stress-induced) DNA methylation in apomictic dandelions.

regions that are conditionally inherited through the female line. These changes result in an adaptive phenotype for both the stressed parent and offspring plants. Reversal of the induced DNA methylation changes over the course of subsequent unstressed generations was observed, indicative of a transient short-term stress-dependent memory of the acquired changes (Wibowo et al. 2016). Although several examples of stress-induced heritable epigenetic changes are reported in plants (Steward et al. 2002; Verhoeven, Jansen, et al. 2009; Latzel et al. 2012), it is unclear how common such changes are (Heard & Martienssen 2014) and whether they persist for more than one unexposed offspring generation.

In non-model species lacking a reference genome studies aimed at understanding transgenerational inheritance of DNA methylation have largely relied on methylation sensitive AFLPs (MS-AFLP) (Verhoeven, Jansen, et al. 2009). MS-AFLP studies are limited to qualitative methylation estimates on few anonymous CCGG sites (Schrey et al. 2013). MS-AFLP results are difficult to interpret and often ambiguous (Fulneček & Kovarik 2013) and most versions of the protocol do not allow for distinction between hemimethylation on the external cytosines from the internal cytosine. Bisulfite sequencing based studies offer advantages in being able to detect DNA methylation variation in all sequence contexts in a quantitative fashion. Studying transgenerational methylome dynamics on a whole genome scale using sequencing requires a reference genome combined with whole genome bisulfite sequencing (WGBS), which to date is available for approximately 33 plant species (<http://schmitzlab.genetics.uga.edu/plantmethylomes>) that are mostly species of agronomic interest. Although the number of plants with a sequenced reference genome is increasing rapidly (Niederhuth, Bewick, Ji, Alabady, Kim, Page, et al. 2016), many ecologically interesting species do not yet have a sequenced reference genome. Even if they do, the costs associated with WGBS are substantial, especially for species with large genomes, limiting applicability to small experimental designs.

In the absence of a reference genome, epigenetic Genotyping By Sequencing (epiGBS) and bsRadseq, both modified Reduced Representation Bisulfite sequencing (RRBS) techniques, offers sequencing-based alternatives for studying DNA methylation variation in non-model species (van Gurp et al. 2016; Trucchi et al. 2016). EpiGBS allows for reference free reduced representation sequencing on a restriction site-associated and replicable subset of genomic DNA of hundreds of individuals simultaneously for a low cost. This enables methylome characterization in transgenerational experiments involving non-model species. Here, I use epiGBS to study the transgenerational dynamics of DNA methylation variation, and its genotype-specificity, in a six-generation experiment with apomictic dandelions (*Taraxacum officinale*). I also evaluate how heritable methylation variation is affected by an environmental perturbation, specifically jasmonic acid treatment. Jasmonic acid (JA) is a plant hormone involved in regulating plant responses to abiotic and biotic stresses as well as plant growth and development (Delker et al. 2006). Jasmonic acid has been used widely as an artificial agent applied to induce systemic plant defense pathways typically associated with insect herbivory (Tytgat et al. 2013). Previous work in dandelion has shown that in some genotypes offspring of JA-exposed plants is avoided in a choice feeding experiment featuring the generalist herbivore *Spodoptera exigua* (Verhoeven & van Gurp 2012), suggesting that parental JA-exposure can alter the physiology of offspring plants. This

altered physiological state could be controlled by epigenetic mechanisms, such as altered DNA methylation patterns. Consistent with this idea, JA treatment was previously reported to show inherited DNA methylation changes in a MS-AFLP study (Verhoeven, Jansen, et al. 2009). My specific objectives are (1) to characterize context-specific DNA methylation levels in Apomictic dandelions, as well as (2) to examine the transgenerational stability of DNA methylation in CG and CHG context. Furthermore, I (3) investigate if JA exposure induces heritable changes in DNA methylation up to several offspring generations after exposure. I hypothesize that (1) DNA methylation patterns of Dandelion, like many other plant species (Niederhuth & Schmitz 2016), show higher values for DNA methylation in CG and CHG context compared to *Arabidopsis thaliana* and (2) high stability of DNA methylation only in CG context and (3) that JA exposure results in an increased number of DNA methylation changes that are transmitted transgenerationally. I have tested the hypothesis by measuring DNA methylation using the epiGBS method (van Gurp et al. 2016) in the offspring of apomictic dandelions derived from a multi-generational experiment with JA-treated and control lineages. Inter-generational comparisons of the same subset of (methylated) cytosines allow for assessing the transgenerational stability of DNA methylation patterns, whereas the treatment effect of JA can be determined by comparing DNA methylation patterns between the pre-treated and non-treated lineages.

4.2 Material and Methods

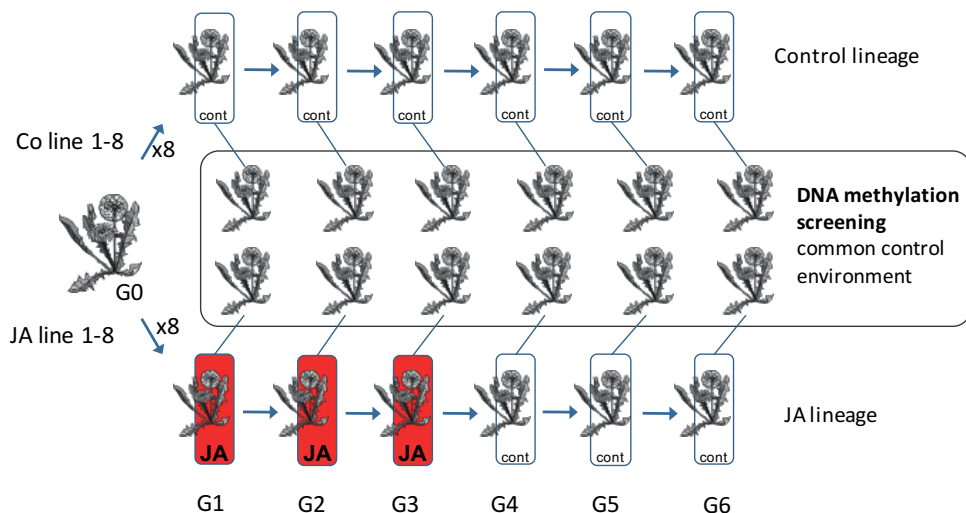


Figure 4.1. Experimental design. For both genotypes A34 and H72, two single-seed descent lineages with eight lines each (lines 1-8) were initiated. One lineage received a Jasmonic acid (JA) treatment lineage whereas the control (CO) lineage did not. Both lineages combined are referred to as “treatment” lineages. Descendants from all individuals of these treatment lineages were grown simultaneously in a common control environment for DNA methylation screening. I will refer to these plant as the DNA isolation lineage. As DNA methylation screening is performed in the offspring of the treatment lineages, the intergenerational distance between DNA methylation-screened plants from subsequent generations in the multigenerational experiment is augmented by two. For instance, the intergenerational distance between G1 and G2 plants of the same line in the treatment lineage is one generation, but as the DNA methylation is measured in the offspring of both G1 and G2 individuals a “distance” of two generations is added, resulting in a minimum intergenerational distance of three generations. Likewise, the intergenerational distance between the offspring of G1 and G6 of the same line, which are five generations apart, is seven generations.

4.2.1 Multi-generation experiment

Single seed-propagated lines of apomictic dandelions were established for the genotypes A34 and H72. A34 is a recent apomict that was produced in an experimental cross between a sexual diploid mother and diploid pollen from a triploid father (Tas & Van Dijk 1999; Verhoeven, van Dijk, et al. 2009). H72 is an established apomict collected from the field and lab-propagated for several generations prior to the experiment (Verhoeven & van Gurp 2012). Per genotype, two lineages with 8 lines each were established and propagated for 6 generations via single-seed descent (Fig. 4.1). During the first 3 generations of the experiment (G1-G3) the 8 lines of the JA lineage received a Jasmonic Acid (JA) treatment (see Fig. 4.1). In the rest of this chapter I refer to the lines that received JA in G1-G3 as ‘JA

lines' and the other lines are called 'control lines'. Note that plants from the JA lines did not receive JA treatment during generations G4-G6. In each JA line in generation G1-G3, JA was applied twice, when plants were 5 and 7 weeks old; 0.25 ml (week 5) and 0.75 ml (week 7) of a 10 mM JA solution (Sigma J-2500, dissolved in ethanol and diluted to the desired concentration with a 0.1% Triton X-100 surfactant solution) was applied and manually distributed over the surface of two leaves (week 5) or four leaves (week 7). Per genotype, all lines from both lineages were founded by a single ancestor individual (G0). In each generation of the experiment, all plants were grown in a common greenhouse environment. For all generations, seeds were germinated for 10-11 days (10 h dark: 14 h light; 14°C: 18°C) in a growth chamber. Seedlings were individually transplanted into 9cm pots containing a mixture of 80% potting soil and 20% pumice, and per genotype pots were fully randomized over a table in a greenhouse with (10 h dark : 14 h light; 16°C : 21°C). Plants were watered every ~2 days as required, and twice per week the table was flooded with half-strength Hoagland nutrient solution. Cohorts were grown until >95% of plants produced flowers and seed heads.

4.2.2 DNA methylation screening experiment

Apomictic seeds were collected from all individual plants from the multi-generation experiment described above and stored at 15°C in a seed storage cabinet with 30% humidity in the absence of light. After completion of the multi-generation experiment, a single offspring individual from each of the experimental plants was grown together in a common environment for DNA extraction and DNA methylation analysis. This DNA methylation screening experiment was designed for 192 plants (2 genotypes x 16 lines x 6 generations) but because some seeds were lost, the experiment was realized with 2 genotypes x 15 lines x 6 generations, for a total of 178 plants (two lines had germinating seeds for 5 instead of 6 generations). Seeds were sterilized using 0,05% bleach and germinated on 2% agar in petri dishes sealed with parafilm in a growth cabinet for 7 days (10 h dark: 14 h light; 14°C : 18°C). Seedlings were individually transplanted into 9cm pots containing a mixture of 80% potting soil and 20% pumice and were grown in a fully randomized design in a climate chamber (10 h dark: 14 h light; 15°C : 20°C). Plants were watered every two days, and twice per week the carts were flooded with half-strength Hoagland nutrient solution. Leaf samples were harvested after 30-36 days for DNA isolation. Per plant, 6 leaf punches were taken from the sixth leaf, avoiding the latex-rich mid-vein. Leaf punches were flash frozen in liquid nitrogen and stored at -80°C after sampling. Note that because the plants grown for DNA methylation screening did not receive any treatment, I can only detect Jasmonic acid effects on DNA methylation that are transmitted transgenerationally and not direct JA effects on the treated plants themselves.

4.2.3 DNA isolation and epiGBS library preparation

For DNA extraction leaf discs were homogenized by bead-beating the frozen leaf tissue in a 2-mL Eppendorf tube with 2-3mm stainless steel beads. No more than 100 mg of fresh tissue was used per sample. Samples with beads were taken from the freezer and stored in liquid nitrogen. For each batch of 12 samples, the tubes were put in a TissueLyser 24 adaptor set block that was partially submerged in liquid nitrogen to prevent thawing during sample placement. After 30 seconds of shaking at 30 Hz at 1,800 oscillations per minute, the tubes were submerged in liquid nitrogen, after which a second 30 second round of shaking

was performed at the same settings. DNA was isolated using the Macherey-Nagel Nucleospin Plant II kit. I followed the manufacturer's protocol with the following modifications. For cell lysis, I used cell lysis buffer PL1 for 30 min instead of 10 min. After filtration through the pink nucleospin filter, I carefully pipetted the flow-through into a fresh 1.5-mL tube, avoiding the pellet that is often formed at this stage. I used an additional centrifugation step to avoid a small pellet; the clear supernatant was used in the following steps according to the manufacturer's protocol. As restriction enzymes are very sensitive to proteins and other contamination, I selected only samples with high purity according to Nanodrop 2000 measurements (specifically, 260/280 and 260/230 ratios of at least 1.8 and 1.5, respectively). Based on Nanodrop measurements, per individual 400 nanogram of DNA was used. EpiGBS libraries were prepared as described previously (van Gurp et al. 2016) using PstI as enzyme. PstI (recognition site C_TGCAG) is a frequently used restriction enzyme in Genotyping By Sequencing (GBS) (Elshire et al. 2011), however it is methylation sensitive for CHG methylation, for which two sites are present in the enzyme recognition site. As such, a bias against CHG methylation was detected previously (van Gurp et al. 2016), making this enzyme better at unbiased quantification of CG methylation than of CHG methylation. In addition to the PstI epiGBS libraries, standard GBS libraries were prepared from the unconverted digestion ligation mixture. For both epiGBS and GBS libraries, 5 μ L out of a total of 60 μ L was used per sample. Pools of all samples for both the epiGBS and GBS libraries were cleaned up using Qiaquick PCR cleanup (Qiagen; 28104), elution was performed in the Qiagen elution buffer (EB) in a volume of 40 μ L. 0.8x SPRI size selection was applied, elution was done using Qiagen EB in a volume of 24 μ L. For the standard GBS libraries, nick repair and bisulfite conversion were skipped. 2 μ L of the size selected digestion ligation mixture was used as a template for the final PCR according to (van Gurp et al. 2016).

4.2.4 Sequencing and bioinformatics

4.2.4.1 Library quantification

For accurate quantification of the sequencing libraries, a miseq Nano run was performed on 178 pooled libraries from 178 samples of which 82 were GBS and 96 were epiGBS libraries, pooled in equal ratios. By combining GBS and epiGBS libraries in a single pool I only needed one sequencing run for quantification. As I have 12 barcoded adapters for the forward read and 8 for the reverse reads, a maximum of $8 \times 12 = 96$ samples could be multiplexed per lane. However, epiGBS and GBS libraries can be distinguished due to differential conversion of the cytosines in the unmethylated enzyme recognition site (see (van Gurp et al. 2016)), which is what allowed us to pool epiGBS and GBS beyond the 96-sample limitation. Sequencing reads were demultiplexed using demultiplex.py (see [https://github.com/thomasvangurp/epiGBS/tree/epiGBS-Nature methods](https://github.com/thomasvangurp/epiGBS/tree/epiGBS-Nature%20methods)) which was edited to allow for distinction between epiGBS and GBS reads. Read counts for GBS and epiGBS reads were made based on the forward and reverse barcoded tag present in the read names using the following bash commands for epiGBS (`pigz -cd R1samplecode123_000000000-D0BJK_s_1_fastq.txt.gz|grep "watson\|crick" | tr ' ' '\t'|cut -f2,3|tr '\t' '_' | awk '{count[$1]++} END {for (word in count) print word, count[word]}'|sed 's/BC:Z://g'|pbcopy)` and GBS (`pigz -cd R1samplecode123_000000000-`

DOBJK_s_1_fastq.txt.gz|grep "gbs" | tr ' ' '\t'|cut -f2,3|tr '\t' '_'| awk '{count[\$1]++} END {for (word in count) print word, count[word]]}'|sed 's/BC:Z://g'|pbcopy). Observed counts for barcodes were matched to the original samples. Per sample counts were used to adjust the sample pooling ratio for subsequent epiGBS library sequencing (see below) to correct for over- and underrepresented samples by adding less or more DNA from the digestion-ligation mixture to the final PCR.

4.2.4.2 epiGBS library sequencing

Two lanes of sequencing were performed on an Illumina HiSeq2500 (125-cycles paired-end reads) using a randomly distributed 96-plex (R1-R96) and 82-plex (R97-R178) epiGBS library for a total of 178 samples. The 96-plex library yielded 130,529,265 reads whereas the 82-plex library yielded 124,968,262 reads. Demultiplexing discards reads that cannot be assigned reliably due a high rate of mismatches in the barcode, yielding 103,400,923 usable reads for the 96-plex and 82,132,633 reads for the 82-plex library. Such reads are not filtered out here as such low-quality reads are discarded during the mapping process. Details of demultiplexing and mapping settings are described previously (van Gurp et al. 2016).

4.2.4.3 Methylation and SNP calling strategy

Reads were mapped against the PstI-based dandelion epiGBS pseudo-reference sequence published previously in (van Gurp et al. 2016). Variant calling was based on samtools mpileup version 1.3 combined with a custom python script for methylation and SNP calling (https://github.com/thomasvangurp/epiGBS/blob/master/mapping_varcall/methylation_calling_samtools.py). This script follows the same logic as the methylation calling script based on FreeBayes (van Gurp et al. 2016). However, compared to that script, the new version is more sensitive, runs faster and contains an improved methylation and SNP detection algorithm. Variant calls in gzipped VCF files were generated for the Watson and Crick strand separately using samtools mpileup v1.3 with the following command (samtools mpileup --reference ../output_denovo/consensus_cluster.renamed.fa -gt DP, AD, INFO/AD, ADF, INFO/ADF, ADR, INFO/ADR -d 10000000 -q 0 -Q 0 -v (watson.bam or crick.bam) > (watson.vcf.gz or crick.vcf.gz). For every non-reference allele called on all 178 samples, only alleles with a frequency of at least 5 percent of the reads were preserved to avoid potential sequencing errors. As presented in (van Gurp et al. 2016), Crick strand reads are converted to the reverse complement, to allow for representing (methylation) variants of both the Watson and Crick strand based on the same (Watson) reference sequence. On both strands, unmethylated cytosines are transformed into thymine, but given the reverse complement representation of the crick strand, methylation polymorphisms on the crick strand are denoted as G / A whereas on the Watson strand they are C/T. Methylation calling based on Watson and Crick allele observations follows the procedure described in (van Gurp et al. 2016). Briefly, the Python script methylation_calling.py uses the 'walktogether' method of the PyVCF package (<https://github.com/jamescasbon/PyVCF>) to simultaneously iterate over both Watson and Crick VCFs. In this way, SNPs and methylation polymorphisms were distinguished and split. C/T polymorphisms in the Watson strand combined with C on the Crick strand indicate a methylation polymorphism on the Watson strand, whereas a G/A polymorphism on the Crick strand combined with a G on the Watson strand indicates a methylation polymorphism on the Crick strand. Where

combined SNP and methylation lead to the same polymorphism (C/T or G/A) on both Watson and Crick strands, only the SNP is called, as the methylation ratio cannot be reliably determined. SNP variant calling was refined by writing a new algorithm according to a set of rules aimed at minimizing false positive SNPs. As a basic rule, SNP variants that are not obscured by bisulfite conversion, such as for example A/T SNPs, are required to be present on both strand in order for them to be called. Furthermore, SNPs containing a C or G are subject to rule 2 and 3:

1. Per sample SNP variant observations in Watson and Crick strand should match, e.g. if the reference is T and an A variant is observed on the Watson strand, that A should also be present as a variant on the Crick strand. Variants not meeting this criterion are not called
2. C/T and G/A SNPs can only be called based on respectively Crick or Watson strand as these genotype calls can only be interpreted as genetic variation with reads from one strand, as methylation variation can cause the same genotype call on the Watson strand (C/T) or Crick strand (G/A) after bisulfite conversion.
3. SNPs containing a C or T allele are called on both strands whenever possible. For example, a C/A SNP can be T/A on the Watson strand only if a T allele is absent from the Crick strand. In this case, the Watson T allele is interpreted as evidence for a C (which was bisulfite converted) and is used in conjunction with the C allele count from the Crick strand.

4.2.4.4 SNP filtering and sample exclusion

To assess whether all samples were genotyped correctly and no samples were accidentally swapped between genotypes I selected a number of high quality SNPs according to the following criteria:

1. Number of called samples is higher than 150
2. The number of samples (out of 178) called as homozygous reference, homozygous alternate or heterozygous are all below 100 in order to exclude spurious SNPs that are only found in a few individuals.

This filtering procedure resulted in 125 SNPs that together differentiate the A34 from the H72 genotype. Subsequently, the resulting filtered variant call file was used as an input to a custom R script (https://github.com/thomasvangurp/epiGBS/blob/master/mapping_varcall/SNP_tree.R) to make a dendrogram using the `snpGdsHCluster`, `snpGdsIBS` and `snpGdsCutTree` function from the `SNPRelate` library (X. Zheng et al. 2012). Based on evaluation of this dendrogram all samples from JA line 7 of A34 (A34_JA_7) were excluded, as three samples from this line (G4-G6) were clustered in the H72 group, which is most likely the result of an incorrectly labeled sample in G4 during the execution of the multi-generational greenhouse experiment. All individual samples clustering separately with a node height (Euclidean distance between samples) above 0.1 were excluded from the analysis (see Fig. 4.2), as the genetic dissimilarity of these samples compared to the other observed samples could indicate a problem with the genotyping quality. Even though I applied strict filtering on SNPs, technical artefacts due to for example excessive PCR-duplicates can occur. Samples A34_CO_5_g3 and H72_CO_7_g1, were discarded based on a low read count of respectively

4.2.4.5 Filtering of methylation calls

A custom Python script was written to analyze the methylation.bed file output of the methylation calling pipeline. I focused the analysis on CHG and CG sites, as CHH methylation is very low to absent and thus called with the least accuracy; CHH methylation was only assessed in describing the DNA methylation landscape. Only CG and CHG positions called in at least 134 out of 167 samples (>80%) were considered for comparative analysis. Per cytosine methylation ratios are calculated by dividing the unconverted read count by the total read count, resulting in a ratio of 0 (no evidence of methylation) to 1 (full methylation). For some downstream analyses a further minimum read coverage criterion was specified (see results) which restricts those analysis to a small subset of high-confidence methylation calls.

4.2.4.6 Calculating corrected correlation coefficients of inter-individual DNA methylation

One way of looking at transgenerational dynamics in DNA methylation is to determine the correlation of methylation levels (across all cytosines called) between samples of different generations. This was done separately for cytosines in CG and CHG context. As many loci are not called in all 167 samples, pairwise comparisons between individuals are based on different subsets of loci. In preliminary analysis, I observed that the number of loci called and the sequence coverage of these loci are important confounding factors, leading to a bias in the inter-individual Pearson R^2 for the methylation ratios. The number of loci called and the inter-individual Pearson R^2 statistic are positively correlated (see Fig. 4.3); comparisons between two samples with high sequencing coverage have a higher number of loci called with good coverage, making the individual estimates of the methylation level of these loci more accurate. Conversely, samples with low sequencing coverage have a lower number of loci called, each with lower coverage with methylation estimates that are less accurate. Initially, I tried to correct for this by including only loci with minimum 15X coverage in >80% of the samples. This filtering did not sufficiently remove the observed bias, which prompted us to design a more elaborate correction strategy. This strategy consists of (1) excluding correlations between individuals having less than 2,800 cytosines in CG context covered; and (2) statistically correcting the inter-individual Pearson correlation coefficients for the number of loci included in the pairwise comparison. The threshold of 2,800 cytosines in CG context was determined based on visual inspection of the distribution of the scatter plot (Fig. 4.3). The first filter discards inter-individual comparisons of individuals that have less than 2,800 cytosines in CG context in common, which also raises the minimum number of cytosines in CHG context to around 5,300. Based on this filter, 253 out of 1,526 comparisons for A34 (17%) and 388 out of 1,834 comparisons for H72 (21%) were discarded. The second filter corrects statistically for a coverage-dependent bias that I detected. Using Microsoft excel for Mac, trend lines were fitted using both linear and exponential equations. In case of exponential equations, excel reports the Pearson R^2 of a linear approximation of the $(x, \ln(y))$ values. The relation between the number of available positions and uncorrected Pearson R^2 is best described with exponential equations (see Fig. 4.3), which gave higher R^2 scores (based on the aforementioned linear approximation) compared to a simple linear regression (data not shown). The difference between the expected R^2 score (based on the regression line) and observed (uncorrected) R^2 score is added to an arbitrarily defined baseline value to yield a corrected Pearson R^2 score that is

independent of the number of loci available. The baseline value for the derived (corrected) R^2 statistic was determined by calculating the outcome of the derived equations (Table 1) for a number of cytosines close to the maximum of the range; 3,000 for CG and 5,800 for CHG. Although the baseline choice is arbitrary, I believe that the observed inter-individual correlations are most accurate in comparisons with these maximum number of loci included, making the R^2 values at this end of the scale a natural baseline choice. This leads to the baseline values listed in Table 1. Corrected Pearson R^2 coefficients were subsequently determined per genotype-treatment combination (A34 CO, A34 JA, H72 CO and H72 JA) for all pair-wise comparisons within and between the 7 to 8 lines of each genotype-treatment group. I suspect that especially for cytosines in CHG context these corrected Pearson R^2 coefficients likely underestimate the 'true' correlation coefficient and absolute R^2 values should thus be interpreted with caution. This is because the methylation sensitivity of the enzyme PstI I used, as well as PCR duplicates, have a negative impact on the accuracy of the correlation estimates (van Gurp et al. 2016). The corrected Pearson R^2 scores in this study are used to discern patterns in DNA methylation changes occurring over generations. For this purpose, it does not matter that these scores likely underestimate the true correlations as the technical issues that lead to this bias are constant. For between-line comparison the maximum inter-generational distance is 14 between offspring of the G4 individuals (see Fig. 4.1). The minimum inter-generational distance is 4, for the offspring of G1 individuals from different lines. Intra-line comparisons range in inter-generational distance from 3 (between offspring of subsequent generations) to 7 (from offspring of G1 to offspring of G6). Comparisons were made on 3,137 CG and 5,898 CHG positions distributed over 546 contigs.

Chapter 4 Transgenerational inheritance of (stress-induced) DNA methylation in apomictic dandelions.

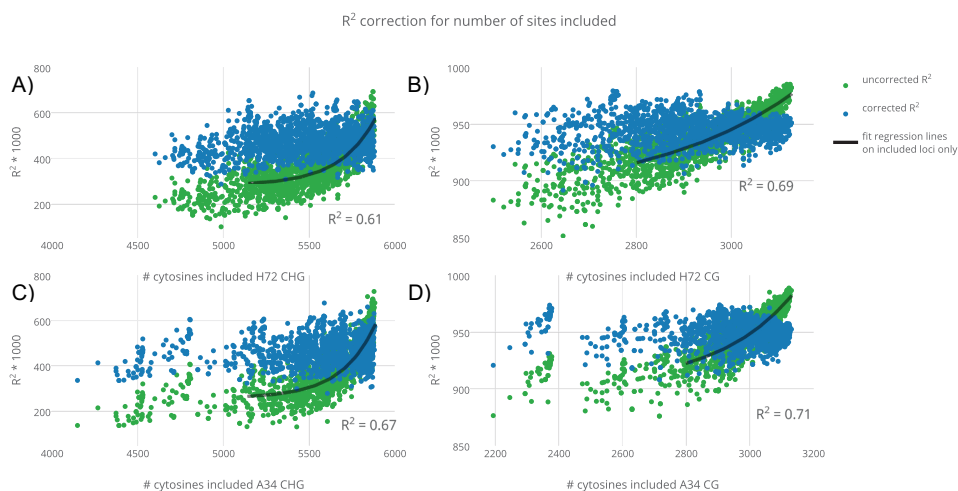


Figure 4.3. Number of cytosines available in inter-individual comparison vs corrected and uncorrected Pearson R² in both the JA and control lineage for H72 (A, B) as well as A34 (C, D). Exponential equations (see Table 4.1) explain 61%-71% of the observed variation between number of cytosines available for comparison and uncorrected Pearson R² between individuals sharing more than 2,800 cytosines in CG context. The uncorrected Pearson R² values are shown in green whereas the corrected Pearson R² are shown in blue. Regression lines are based on comparisons between individuals with at least 2,800 cytosines in CG context. A) Uncorrected and corrected Pearson R² for cytosines in CHG context in H72. B) Uncorrected and corrected Pearson R² for cytosines in CG context in H72. C) Uncorrected and corrected Pearson R² for cytosines in CHG context in A34. D) Uncorrected and corrected Pearson R² for cytosines in CG context in A34. The R² scores shown in the figure for the exponential fit are derived using MS Excel and represent the Pearson R² coefficient for a linear approximation of the (X, ln(Y)) values where X is the number of cytosines and Y is the corresponding R² score.

Table 4.1. Exponential equations derived for the corrected Pearson R² score.

genotype	Sequence context	Equation	Baseline value	Explained variance R ²
A34	CG	$Y(x) = 903 + 0.000133 \cdot e^{0.00425 \cdot x}$	0.948	0.71
A34	CHG	$Y(x) = 262 + 3.79e-11 \cdot e^{0.00505 \cdot x}$	0.461	0.67
H72	CG	$Y(x) = 887 + 0.00221 \cdot e^{0.00339 \cdot x}$	0.945	0.69
H72	CHG	$Y(x) = 288 + 7.73e-11 \cdot e^{0.00425 \cdot x}$	0.479	0.61

4.2.4.7 Annotation of clusters

Contigs of the pseudo-reference were evaluated for similarity to Gene and TE annotations using USEARCH UBLAST (Edgar 2010). A gene hit was defined for contigs with an e-value of $< 1e-5$ after mapping to the reference proteome (RefSeq) of all eudicot plant species. TE-related clusters were discovered using usearch blastx against a composite database, with

sequences derived from the plant repeat databases for all taxa for all repeat classes from Michigan state university downloaded in November 2015 (<http://plantrepeats.plantbiology.msu.edu/downloads.html>).

4.2.4.8 Detecting effects of JA treatment in previous generations on DNA methylation

In order to test whether parental Jasmonic acid application resulted in heritable DNA methylation changes, I performed both univariate and multivariate analysis, contrasting JA and control lines of both genotypes separately. In the univariate analysis, I aim to find differentially methylated cytosines between JA and control lines on a per generation per genotype basis. In the multivariate analysis, I visualized the DNA methylation distribution using principal component analysis (PCA) for an exploratory data analysis. I subsequently perform a redundancy analysis (RDA) to test for the statistical significance of the parental pre-treatment history (control or JA) on the methylation patterns for both cytosines in CG and CHG context per genotype.

4.2.4.8.1 univariate analysis to detect differentially methylated cytosines

For the univariate analysis, only positions with coverage of at least fifteen reads that are present in >80% of the 167 samples are considered. Furthermore, I subject only those cytosines to a statistical test that have at least 3 samples in each of the 2 experimental groups (JA or control) after the above filtering step. The total dataset subjected to univariate (and multivariate) analysis has 167 samples x 28,809 cytosines in the reference, which equals 4,811,103 data points. I used a logistic regression approach, testing the 'experimental group' effect on DNA methylation level using Wald Chi Square tests. If model fit was poor, as judged by significant Pearson or Deviance goodness of fit test, I repeated the logistic regression with the Williams scaling option to account for overdispersion. Overdispersion, or too much variability around the mean, which is often found in count and ratio data as the mean increases, was considered the most likely cause of poor model fit. I retained the p values from these models only if Williams scaling resulted in good model fit (non-significant Pearson and Deviance goodness of fit test). For those cytosines for which also the Williams scaling option did not result in models with good model fit I do not present p values. All retained p values were subjected to a false discovery rate correction, which was applied per genotype x context on all tests performed. Analyses were done using SAS 9.2 (SAS Institute, Cary, NC).

4.2.4.8.2 multivariate analysis to detect JA pretreatment effects on DNA methylation

For the multivariate analysis, I added two other filtering criteria to the ones defined for the univariate analysis, excluding non-variable sites and sites with missing data. The first criterion follows (Gugger et al. 2016) and aims to exclude loci that contribute noise but not signal for distinguishing between groups. Following (Gugger et al. 2016), cytosines were included in the analysis if the range of methylation levels across all samples (of that genotype and context) is higher than 0.1 (10%) (so for example: if all samples have methylation levels between 0 and 0.1 for a given cytosine, then that cytosine is excluded from the analysis, but if methylation levels range between 0 and 0.11 then the cytosine is included in the analysis). Secondly, I excluded positions that were not called in all samples for a given genotype (no missing data is allowed). For exploratory analysis and visualization, I used principal component analysis (PCA), coloring sample by lineage (JA or CO). To test for statistical significance of lineage (JA-treated or control) on the methylation

patterns, I used a redundancy analysis (RDA). Methylation levels were arcsin-transformed and centered (by cytosine) to make the data better suitable for RDA. For each combination of genotype and sequence context I fitted two models to test if treatment effects (lineage) methylation. First, I fitted a full model with treatment (lineage), generation, and the treatment (lineage) x generation interaction. If this model was significant, I proceeded with testing the main effect of treatment, and subsequently I fitted this model to subsets of the data that represent a single generation each. This is an attempt to tease out if treatment has an effect, and if so, in which generations this is expressed. Analysis were done using CANOCO v5.03.

4.2.5 Statistical analysis of decreasing correlation over generational distance

To determine if the negative trend for the corrected Pearson R^2 distance as a function of intergenerational distance as shown in Figure 4.5-4.7 is statistically significant, I performed one-way ANOVAs for cytosines in CG and CHG context for both lineages of the genotypes A34 and H72. Per lineage, I performed both analysis within individual lines and analysis between lines (excluding intra-line comparisons). Although the inter-line comparisons have a greater inter-generational distance, comparisons between individuals from different lines are not linear through time and as such can induce a bias, whereas intra-line comparisons are devoid of this bias (e.g. G6 was always grown at a different point in time than G4). Per lineage, all corrected Pearson R^2 distances are taken into account as individual data points. Analysis were performed using SPSS v24.

4.3 Results

4.3.1 Dandelion DNA methylation landscape

Depending on the genomic classification of the epiGBS cluster between 18% (in TE-related clusters) and 34% (in gene related clusters) of cytosines in CG context are methylated (> 10% methylation) (see Fig. 4.4). Only 4% of cytosines in CHG context are methylated (> 10% methylation). CHH methylation levels in dandelion are also low, with only 3% of cytosines in CHH context showing methylation above 10%. Gene-annotated clusters have higher level of CG methylation and lower levels of CHG and CHH methylation compared to non-annotated clusters, indicative of gene-body methylation (Takuno & Gaut 2013; H. Wang et al. 2015), whereas TE-related clusters have higher levels of CHG and CHH methylation, consistent with previous reports (Cokus et al. 2008; Song et al. 2013; X. Li et al. 2012; H. Wang et al. 2015) (Fig. 4.4).

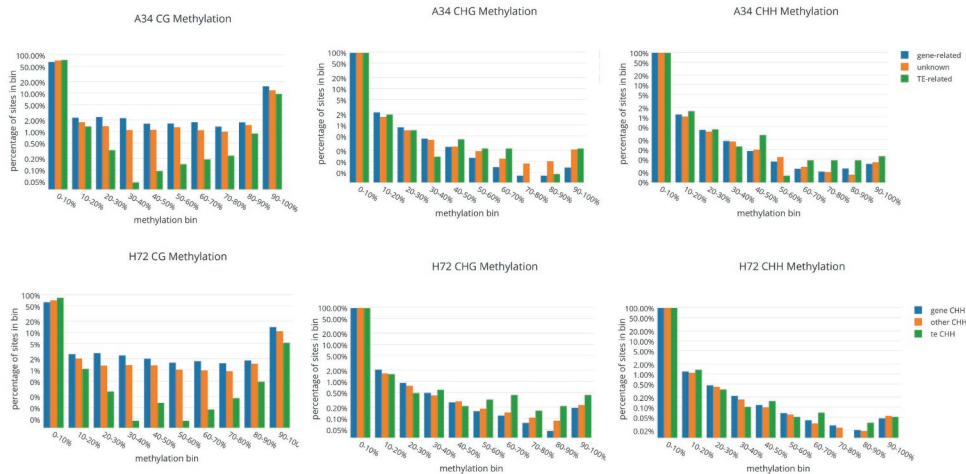


Figure 4.4. DNA methylation level distribution for cytosines in CG, CHG and CHH context in gene-related, TE-related or other clusters. Methylation levels are recorded for cytosines in contigs with >80% of the samples being called with a minimum coverage of 20 reads.

4.3.2 Transgenerational stability of DNA methylation variants

If DNA methylation is largely under genetic control and no mutations occur I would expect high correlations of DNA methylation ratios between generations. Similarly, DNA methylation variants that are not under genetic control but that are epigenetically inherited in a very stable way would lead to the same result. Unstable and non-heritable DNA methylation changes would lead to a lower intergenerational correlation, as would technical noise in the methylation scores (e.g. due to PCR duplicates, low coverage, or enzyme methylation sensitivity (van Gurp et al. 2016) (see discussion). Only novel but heritable epimutations should lead to a decrease of the correlation (Pearson R^2) over generational distance. In that case, the correlation of DNA methylation between closely related individuals should be higher than that of more distant relatives due to epigenetic divergence over generations. The slope of this declining correlation coefficient over generational distance provides an indication of the rate and the stability of heritable DNA methylation changes.

Overall, I observed a negative trend in the intergenerational correlation of DNA methylation between individuals over generational distance, for cytosines in CG and CHG context (see Fig. 4.5 and 6). I found considerable variation in the intergenerational methylation value correlations (Pearson R^2) for individual lines, showing weak or even absent patterns when considering some lines in isolation (see Supplementary Fig. 4.1), possibly related to technical noise in the PstI-based epiGBS methylation calls. Combined however, the average corrected Pearson R^2 values for all panels in Fig. 4.5 and panel A,C and D in Fig. 4.6 show a clear trend, with a negative regression coefficient that was significant for most inter-line comparisons in CG context and for most intra and inter line comparisons in CHG context

Chapter 4 Transgenerational inheritance of (stress-induced) DNA methylation in apomictic dandelions.

(see Supplementary Table 4.5). The correlation coefficients for CHG are markedly lower for cytosines in CHG context compared to cytosines in CG context, which could be explained by technical errors or a higher sensitivity to environmental changes (see discussion). The slope for the regression lines based on CHG methylation in control lines (Fig. 4.6) is 10-fold higher compared to the regression based on CG methylation in control lines (Fig. 4.5), indicating a higher rate of heritable methylation changes in CHG context than in CG context.

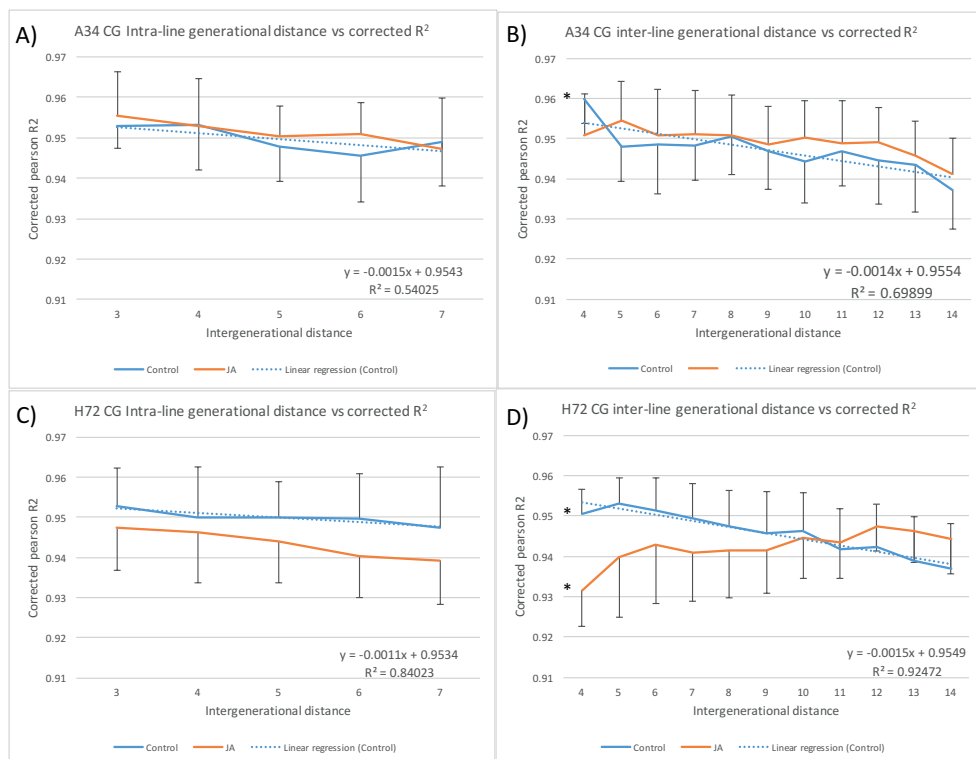


Figure 4.5. Generational distance versus average corrected Pearson R² for the JA and Control lineage for cytosines in CG context, both within (A, C) and between lines within a lineage (B, D). Error bars show standard deviations for average corrected Pearson R² scores. Linear regression lines are plotted for the control lineage. **A)** A34 CG context within line. **B)** A34 CG context between lines only. **C)** H72 CG context within line. **D)** H72 CG context between lines only. Asterisks next to the lines in all panels indicate a significant relation between the corrected Pearson R² score and intergenerational distance (see Supplementary Table 4.5)

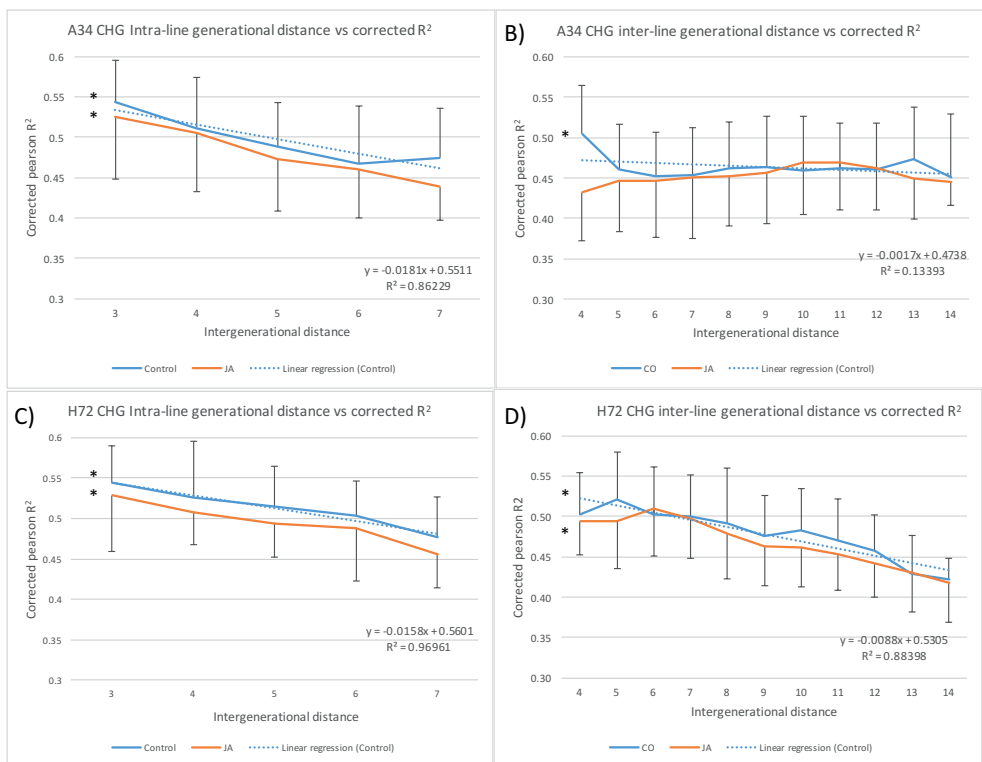


Figure 4.6. Generational distance versus average corrected Pearson R² for the JA and Control lineage for cytosines in CHG context, both within (A, C) and between lines (B, D). Error bars show standard deviations for average corrected Pearson R². Linear regression is plotted for control lineages. A) A34 CHG context within line. B) A34 CHG context between lines only. C) H72 CHG context within line. D) H72 CHG context between lines only. Asterisks next to the lines in all panels indicate a significant relation between the corrected Pearson R² score and intergenerational distance (see Supplementary Table 4.5)

4.3.3 Is there an effect of Jasmonic acid (JA) pre-treatment on DNA methylation in dandelions.

We assessed whether JA pre-treatment had an impact on intergenerational DNA methylation stability by comparing corrected Pearson R^2 scores of JA-treated (offspring) plants with control plants. I also investigated if JA treatment resulted in specific DNA methylation changes using both univariate and multivariate analysis of DNA methylation variation in both CG and CHG context.

4.3.3.1 (parental) JA treatment effect on intergenerational DNA methylation correlation

In principle, JA exposure could lead to heritable changes in DNA methylation in two different ways. First, JA exposure could trigger targeted modifications (shared between replicate plants). Second, JA exposure could enhance the rate of non-targeted methylation changes. From the literature, there is evidence for the occurrence of both random and targeted DNA methylation polymorphisms after stress exposure (C. Jiang et al. 2014; Wibowo et al. 2016), which gradually disappear after the stress is relieved. Targeted changes are expected to reduce the Pearson correlation coefficient after a shift in the environment. Nontargeted changes also reduce the correlation after each JA exposure. Thus, I expect an overall reduction in the intergenerational correlation in the JA lines compared to the control lines. Corrected Pearson R^2 scores for intergenerational comparisons in the JA lineage are indeed slightly lower across the board for H72 in both contexts and for A34 in CHG context. However, (almost) no significant differences in intergenerational correlations could be detected between JA and control lines based on a t-test (Supplementary Table 4.3), indicating that the data do not provide evidence for an overall increase in transgenerationally stable JA-induced epimutations.

Following the same argumentation, I expect that the average corrected Pearson R^2 scores for comparisons made between offspring plants from G1-G3 of the JA lineage to show a lower correlation compared to those from generation G4-G6 or the control lineage, assuming that the JA treatment effects on the (random) epimutation rate are transient and fade in the post-treatment generations. I therefore compared intra-line correlations of cytosine methylation ratios between subsequent generations (with an intergenerational distance of 3, see Fig. 4.7) between offspring of treated individuals (g1-g3) in the JA-lineage to the corresponding generations of the control lineage. I found a significant JA effect for the intra-line inter-generational average corrected Pearson R^2 score between g2 and g3 for H72 CG methylation, in which the six available JA lines showed an average corrected Pearson R^2 score of 0.952 compared to 0.940 of the six available control lines ($t=2.95$ $p=0.014$, see Supplementary Table 4.3). One marginally significant test results in H72 for cytosines in CG context was found for the average Pearson R^2 score between g1 and g2 ($t=0.75$ $p=0.07$), whereas no significant results were detected for cytosines CG context in A34 and cytosines in CHG context in either A34 or H72. The emerging patterns hint at a JA induced effect on CG methylation in H72, although this did not result in significant values for all comparisons that include offspring of JA treated plants (Supplementary Table 4.3).

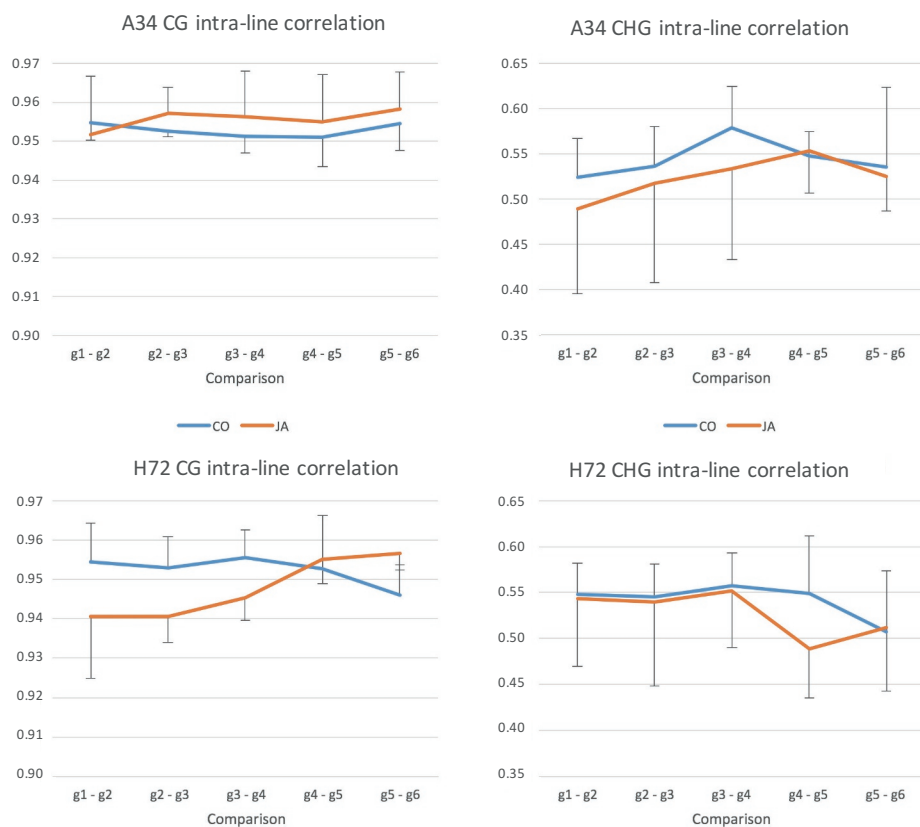


Figure 4.7. Corrected Pearson R^2 for offspring from adjacent generations (intergenerational distance of 3) for the JA and Control lineage for cytosines in CG and CHG context. Error bars show standard deviations for average corrected Pearson R^2 scores.

4.3.3.2 JA-treatment induced DNA methylation changes

To test if specific (targeted) changes occurred due to JA treatment I used logistic regressions to test whether there are individual cytosines that show a consistent difference in DNA methylation in the offspring of JA-treated plants compared to control plants of the same generation. I found only few positions with statistically significant differences in methylation rate between the JA-treated and control plants (see Table 4.2).

Table 4.2. Cytosines with significant JA-CO differences per genotype, differentiated per generation and sequence context.

genotype	context	generation	Tested cytosines	Significant JA-CO difference
A34	CG	G1	2,532	0
		G2	2,327	0
		G3	2,695	2
		G4	2,458	0
		G5	2,337	1
		G6	2,536	0
A34	CHG	G1	4,547	2
		G2	3,944	0
		G3	4,943	8
		G4	4,262	0
		G5	4,094	0
		G6	4,580	0
H72	CG	G1	2,228	0
		G2	2,533	1
		G3	2,337	0
		G4	2,495	0
		G5	2,291	0
		G6	2,223	0
H72	CHG	G1	3,812	0
		G2	4,542	0
		G3	4,018	0
		G4	4,355	0
		G5	3,862	0
		G6	3,672	1

Per genotype, context and generation the number of cytosines in the filtered data set and the number of cytosines with a significant JA-CO difference after controlling for multiple testing are shown.

Out of the total of 15 contigs that include a significantly differentially methylated position (data not shown), 3 have significant blast hits to genes (see Supplementary Table 4.4) whereas no TE-related contigs were found. Out of the total of 546 contigs included in the analysis that contained positions with sufficient sequencing coverage, 96 had significant blast hits to genes and 13 to transposable elements. Gene related contigs are thus significantly over-represented in the set of contigs in which I found differential methylation (Chi square 6.85 $p=0.01$), whereas the absence of TE-related contigs in this set was not significant (Chi square 0.713 $p=0.40$). The very few detected JA-induced DNA methylation differences were thus biased to occur in genes, not in transposable elements. However, this may be biased because genes are usually more easily annotated than transposable elements.

To check if the JA and control lines differ in DNA overall methylation profiles (evaluated across all positions simultaneously) I performed PCA on cytosines in CG and CHG contexts for both genotypes (see Fig. 4.8). In all cases, the amount of variation explained by the first 2 PCA axes is small. But, especially for CHG methylation in A34, the CO and JA sample distribution seem not random, a significant separation due to parental JA-treatment is present (see Fig. 4.8).

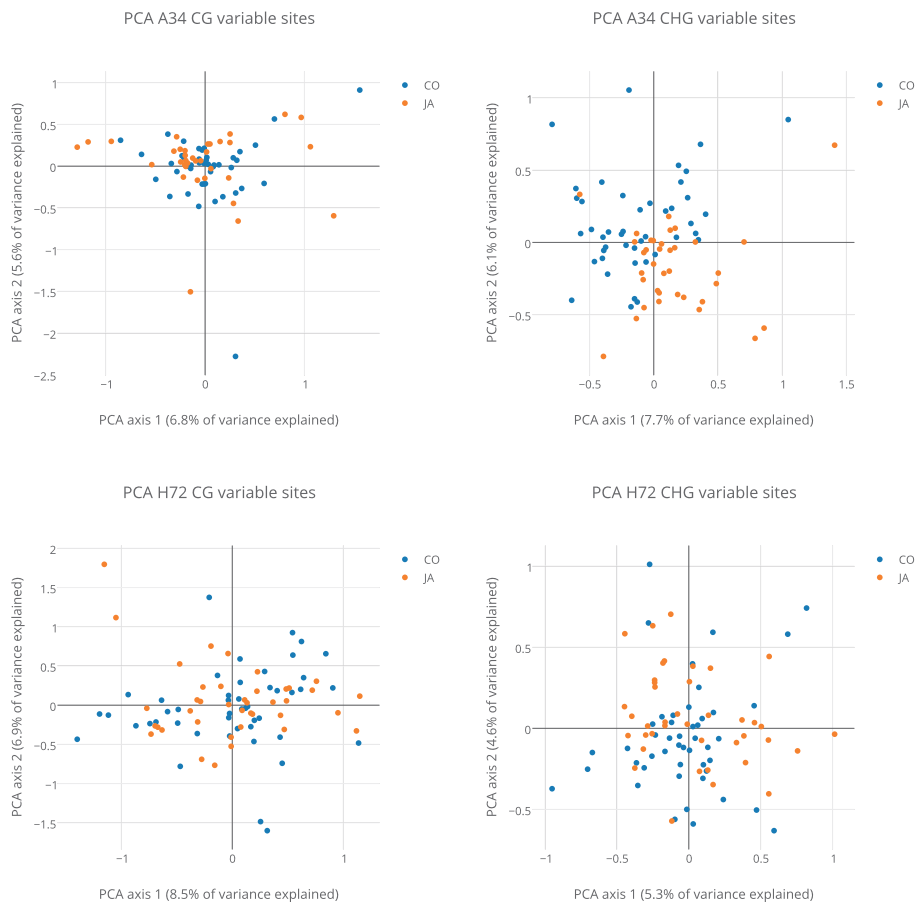


Figure 4.8. Principal component analysis (PCA) based on sites with variable DNA methylation, performed for CG and CHG context separately for both genotypes, including offspring from all six generation. **A) A34 CG B) A34 CHG C) H72 CG D) H72 CHG**

Table 4.3. Results of redundancy analysis tests of the effects of Treatment (lineage: JA or control) and Generation on DNA methylation profiles

genotype	Sequence context	Factors in model (G=generation, T=treatment)	explained variation ¹	Efficiency of constrained axis ²	Test	P-value
A34	CG	G, T, G x T	14.3%		Full model	0.178
		G, T, G x T	14.8%			0.015
	CHG	G	6.5%		Generation	0.294
		T	2.5%	31.9%		0.001
		T, G1 only	7.9%	41.7%	Treatment	0.322
		T, G2 only	9.3%	51.1%		0.008
		T, G3 only	9.1%	38.3%		0.147
		T, G4 only	11.1%	53.8%		0.014
		T, G5 only	8.1%	33.4%		0.656
		T, G6 only	8.4%	40.6%		0.15
H72	CG	G, T, G x T	12.5%		Full model	0.761
	CHG	G, T, G x T	12.9%			0.309

¹ % variation in the methylation data that is explained by the factors in the model

² efficiency of the first RDA axis is calculated by comparing two values of %explained variation in the methylation data: (1) %explained variation by the RDA axis (which is a constrained axis, meaning that it maximizes the variation that is associated with the factors in the model), compared to (2) %explained variation by the axis from the unconstrained method PCA (which maximizes the variation between samples irrespective of any association with factors in the model). This helps interpreting the strength of the effect of the model factors on methylation: the overall %explained variation may be small (often <10%) but a multivariate summary of the data along the levels of the model factors explains ~30-50% of a multivariate summary of the data that is blind to the model factors. So the model factors (treatment, specifically) is a nontrivial source of the variation in the methylation data.

To determine statistical significance of the JA pre-treatment effect on overall DNA methylation profiles I performed a redundancy analysis (RDA) on the methylation ratios of cytosines in CG and CHG context for both genotypes. I found no evidence for an effect of treatment or generation for cytosines in CG context in either A34 or H72. A significant full model test was observed for genotype A34 for cytosines in CHG context (p=0.015; see Table 4.3), indicating that for this genotype the CHG methylation profile differed between generations and/or treatment groups. For A34 CHG, I thus proceeded with testing the effect of treatment within each of the generations separately. I found a significant effect for treatment across all generations simultaneously, and individually in generation 2 (p=0.008) and generation 4 (p=0.014). JA-treatment effects are thus significant, but they are not consistently detected in all generations. The difference also is not preferentially found in the offspring of treatment generations G1-G3, as no significant effects were detected for G1 and G3 (see Table 4.3).

4.4 Discussion

This study is the first to use epiGBS in an experiment designed to track transgenerational stability of DNA methylation. The aim of this study was to i) characterize DNA methylation in apomictic dandelions in detail using epiGBS, ii) determine the stability of DNA methylation over generations and iii) determine the impact of Jasmonic acid treatment on heritable DNA methylation changes. I performed this study using a recently established apomict A34 as well as the naturally occurring (and evolutionarily older) apomict H72, to evaluate whether the age of an apomictic lineage has an influence on the DNA methylation transmission and stability. Compared to other plant species (Niederhuth, Bewick, Ji, Alabady, Kim, Li, et al. 2016) , the DNA methylation patterns reported on here fall in line with previously described species, except for the somewhat lower levels of CHG methylation, which could well be attributed to technical artifact related to the methylation sensitive enzyme PstI used (as described in the methods section). I have found a significant JA treatment effect in a multivariate analysis on cytosines in CHG context for the recent apomictic genotype A34. I also found evidence of a few significantly differentially methylated loci using univariate analysis. The latter were over-represented in gene-related contigs. It thus appears that Jasmonic acid application results in some heritable DNA methylation changes, but this effect is not consistently observed in all genotypes or in all sequence contexts.

4.4.1 dandelion DNA methylation landscape

DNA methylation patterns are often distinct between plant families, with CHG methylation for instance being reduced in the Brassicaceae and reduced heterochromatic CHH methylation as characteristic of the Poaceae (Niederhuth, Bewick, Ji, Alabady, Kim, Li, et al. 2016). Although the number of fully sequenced plant methylomes is rapidly increasing, in the Asteraceae family to which *T. officinale* belongs no methylomes are yet available. I found most DNA methylation in CG context, consistent with observations in all other plant species (Niederhuth, Bewick, Ji, Alabady, Kim, Li, et al. 2016). The levels of CHG methylation I found were low compared to most other plant species, but this could be the consequence of a technical artefact (see materials and methods).

4.4.2 Transgenerational stability of DNA methylation variants

Stable DNA methylation, which is either genetically determined or pure (independent from genetic variation), is expected to result in a constant and high correlation coefficient over generational time when comparing the methylation profiles of ancestors with their offspring. Very unstable but non-heritable DNA methylation, which changes every generation, should contribute to a constant but lower correlation coefficient over generational time. In this study, I observed a declining correlation coefficient over generational time, regardless of the constant positive impact of stable variants and negative impact of unstable non-heritable methylation variants. This steady DNA methylation divergence over generations suggests the presence of novel methylation variants in each generation that are stably inherited. Besides these novel variants, existing variants can also revert to the original (non-methylated) state, such as observed in several studies (Becker et al. 2011; van der Graaf et al. 2015). I observed this negative trend in the corrected Pearson R^2 scores comparing DNA methylation in both CG and CHG contexts.

Correlations for cytosines in CHG context were lower than those in CG context while the decline in correlation over generational distance was stronger for cytosines in CHG context compared to cytosines in CG context. On average, closely related individuals with an intergenerational distance of three had a corrected Pearson R^2 for cytosines in CG context of 0.95 whereas that of cytosines in CHG context was 0.54. There are several potential contributing factors to this difference. First, as mentioned before, is the use of the enzyme PstI, which is blocked by methylation in CHG context. As methylation in CHG context is spatially auto-correlated (Cokus et al. 2008), sequences with a methylated restriction site are more likely to contain CHG methylated sites. As the enzyme is blocked, such observations would be missed, leading to an underestimate of the DNA methylation of cytosines in CHG context. Second, the distribution of DNA methylation in CG context is bimodal, whereas CHG and CHH methylation in plants have a unimodal continuous distribution (Cokus et al. 2008; Lister et al. 2008; van der Graaf et al. 2015). Changes in CG methylation are from very low to very high, or vice versa, making such events easily detectable, even with limited read coverage or severe PCR bias. CHG and CHH methylation ratios in plants on the other hand are not bimodally distributed, making detection of subtle changes in these ratios more challenging. These limitations on the accuracy of our estimates for methylation of cytosines in CHG context may contribute to a lower inter-individual correlation for methylation in CHG context. But they are not expected to be a cause of the negative trend in correlation over generational time that I observed, as they represent a technical artefact that should have a similar effect in each generation. The slope of the negative trend observed for intergenerational DNA methylation correlations was 10-fold higher for cytosines in CHG context compared to cytosines in CG context, suggesting more changes in DNA methylation. From our data, it is however not clear if this is due to a higher proportion of variable sites in CHG compared to CG context or a higher reversal rate (lower stability) of DNA methylation variants in CHG context than CG context.

Previous studies (van der Graaf et al. 2015; Becker et al. 2011) have not characterized heritability of DNA methylation in CHG context, which could be explained by the way heritability of DNA methylation variants was determined. The transgenerational stability of DNA methylation variants in these studies is typically assessed by calculating the number of differentially methylated positions that arise per generation. The statistical power for the detection of differentially methylated positions (DMPs) is higher for cytosines in CG context, as bimodal CG methylation typically switches from ~0% to ~100% or vice versa, which is thought to be a major contributor to their over-representation compared to differentially methylated positions in other contexts (Becker et al. 2011). Most transgenerational studies ignore non-CG DMPs, as these only represent a very small fraction of the total DMPs that can be detected (Schmitz et al. 2011; C. Jiang et al. 2014). A detailed study on the transgenerational stability of DNA methylation in *A. thaliana* did not find changes in CHG methylation between generations (van der Graaf et al. 2015). However, the presence-absence statistic used for DNA methylation is suitable for cytosines in CG but not CHG context. Where methylation of cytosines in CG context is bimodally distributed, methylation in CHG context in *A. thaliana* is nearly normally distributed, with average methylation around 55% (Becker et al. 2011). I suspect that by using the presence-absence metric subtle changes in the quantitative ratio of DNA methylation of cytosines in CHG context are missed, and therefore heritability of DNA methylation in CHG context has not been detected.

4.4.3 Stress induced DNA methylation variation (JA-effect)

We found subtle effects of parental jasmonic acid treatment on DNA methylation patterns that was distinct between the two genotypes A34 and H72. In the young apomict A34 parental JA treatment resulted in altered DNA methylation cytosines in CHG but not CG context (Fig. 4.8). In the established apomict H72 I did not find clustering reflecting parental (JA) treatment or significant model results from RDA, but I found a significantly lower intergenerational correlation for cytosines in CG context between offspring of JA-treated plants compared to offspring of non-treated plants. Using univariate analysis, I detected significant differentially methylated polymorphisms between the offspring of JA-treated plants and plants from the control line for corresponding generations. These changes were mainly found in CHG context and were overrepresented in contigs mapping to genes, indicating that cytosines in gene-related sequences are more likely to be impacted by stress induced DNA methylation changes.

Other studies focusing on the transgenerational stability of stress induced DNA methylation changes in *A. thaliana* found both similar and different results. (C. Jiang et al. 2014) exposed *A. thaliana* to high-salinity soil, which resulted in increased levels of differentially methylated positions (DMPs), mainly in CG context. Up to ~75% of these stress-induced DNA methylation changes were faithfully transmitted for up to two generations in plants grown in control environment. Clearly, these result contrast with our findings. I found almost no DMPs in CG context and the CG-DMPs that I found were not stably inherited in the absence of stress, as I detected hardly any CG-DMPs between the offspring of JA G4-G6 and the control line. A more detailed study on the transgenerational stability of repeated salt-stress-induced DNA methylation changes in *A. thaliana* uncovered that distinct epigenetically labile regions in the *A. thaliana* genome show extensive changes in DNA methylation (Wibowo et al. 2016). (Wibowo et al. 2016) found on average 6,866 DMPs (40% CG, 15% CHG and 45% CHH), detected per generation using pairwise comparisons between the offspring of salt-stressed and control plants, which represent 0.17% of the average number of methylated cytosines that could be detected. In Wibowo's results, in contrast with the findings described by (C. Jiang et al. 2014), second generation offspring of stressed plants did not show that these stress-induced DMPs were maintained, suggesting that these are transient. I do not find DMPs that are maintained for multiple generations, but I do find a few DMPs between the JA and control lineage in generation G4-G6 in both A34 and H72, suggesting that a minor proportion of stress-related differential methylation could have longer lasting effect.

A previous MS-AFLP based study on the impact of environmental stressors on changes in DNA methylation and transmission in apomictic dandelions found that i) more polymorphic MS-AFLP loci were found in plants subjected to stress treatment compared to control plants and ii) most of the changed loci were transmitted faithfully to the next generation (Verhoeven, Jansen, et al. 2009). Using our current dataset, I cannot determine if more changes in DNA methylation occurred in plant subjected to direct JA-treatment, as I did not collect DNA from these plants. Here, I only measure DNA methylation in the offspring of JA-treated plants. Based on the high heritability found of the stress-induced DNA methylation changes in the previous MS-AFLP based study, I can however say that the stress induced DNA methylation variants I find here appear much weaker than those found in Verhoeven

2010. Verhoeven subjected a single generation of A34, the same apomictic dandelion genotype used here, to the same Jasmonic acid treatment I used here. JA treatment resulted in a 10% increase (from 7.5 to 17.8%) in the number of polymorphic MS-AFLP loci compared to spontaneously occurring DNA methylation changes in control lines. These analyses were based on a small subset of labile MS-AFLP loci which were often polymorphic among replicates in multiple environments. Changes in DNA methylation detected using MS-AFLP are almost exclusively found in the CG context (Fulneček & Kovarík 2013). In most of the analyses I focused on all scorable loci, except for PCA and RDA analysis which required at least one difference in DNA methylation ratio higher than 0.1. This makes it difficult to compare the results between the two studies. Compared to MS-AFLP, epiGBS offers a higher resolution (more loci) a quantitative instead of a qualitative signal and information on all sequence contexts which should lead to more reliable results.

Stress-induced DNA methylation changes were distinct between A34 and H72. Specific stress induced changes in CHG DNA methylation were observed in the young apomict A34 and not in the established apomict H72 (based on the RDA analysis). Genotypic differences in the sensitivity to environmental stress have been found before. For instance, drought stress in *Persicaria maculosa* can lead to an adaptive transgenerational response in offspring growing under dry conditions, which develop a bigger root system compared to offspring of parents grown under control conditions (Herman & Sultan 2016). This adaptive transgenerational effect is underpinned by DNA methylation, as zebularine treatment which removes DNA methylation also removed the adaptive transgenerational effects. In comparing different genotypes of *P. maculosa*, the authors found substantial variation in the degree of both the adaptive transgenerational effect as well as the effect of zebularine treatment, suggesting that there is substantial diversity in the degree of sensitivity of environment induced DNA methylation between genotypes. This is consistent with the different response that I observed between A34 and H72, suggesting that the sensitivity of both genotypes to environmental stress might be different.

We uncovered specific stress induced changes in DNA methylation in the young apomict A34 using multivariate analysis in CHG and not CG context. The observed changes were variable between offspring plants from JA generations G1-G6, with a significant effect only in G2 and G4 and a low but non-significant effect in G3 ($p=0.147$; see Table 4.3). The impact of stress treatment on transgenerational DNA methylation changes is often stronger after multiple consecutive treatments. For instance, Wibowo *et al.* found no transgenerational effects on DNA methylation after the first stress generations, but did find such effects in the offspring of from a lineage with 3 and 5 subsequent generations of stress-exposure (Wibowo *et al.* 2016). This might explain why I did not find significant changes in DNA methylation in the offspring of the first JA treatment generation (G1). The indication of an effect in generations G2-G4, but not in later generations, suggests that induced effects are heritable but revert rapidly within 1-2 offspring generations, consistent with findings in *A. thaliana* (Wibowo *et al.* 2016). For genotype H72 I did not obtain significant results from the RDA analysis. Interestingly, the only sub-significant value (0.064) was obtained for cytosines in CHG context, suggesting that also in H72 cytosines in CHG context are more likely to be affected by JA treatment. The fact that I found specific effects in CHG and not CG context is corroborated by (Wibowo *et al.* 2016). Reanalysis by Wibowo *et al.* of the data of

Jiang *et al.* focusing on differentially methylated regions showed that CHH and CHG methylation in salt stressed plants correlated well with stress treatment (Wibowo *et al.* 2016), whereas changes in CG methylation did not, which is consistent with the JA-induced specific changes in CHG methylation I observed in A34.

4.4.4 Measuring DNA methylation in non-model organisms and epiGBS caveats

Measuring DNA methylation in species without a reference genome has hitherto mostly relied on global quantitative methods such as high-performance liquid chromatography (HPLC) or gel-based MS-AFLPs; but see (Platt *et al.* 2015) for semi-quantitative and quantitative (Gugger *et al.* 2016) bisulfite sequencing based approaches. The local sequence context (gene, TE, etc.) as well as the quantitative level of DNA methylation in a region can differ, which cannot be picked up by MS-AFLP. For species with a reference genome whole genome bisulfite sequencing (WGBS) has been used successfully to determine single base resolution quantitative DNA methylation profiles. In absence of a *T. officinale* reference genome, WGBS cannot be employed. Even with a reference genome available, applying WGBS on a large number of individuals would be prohibitively expensive. Reduced representation bisulfite sequencing (RRBS) allows focusing the sequencing effort on a random but defined subset of genomic DNA, allowing to study more individuals simultaneously with a limited sequencing effort. As the present study shows, the epi-genotyping by sequencing (epiGBS) method that we developed (van Gurp *et al.* 2016) is an efficient technique that enables researchers to study quantitative DNA methylation in species without a reference genome, simultaneously allowing the detection of DNA methylation and SNP polymorphisms while on a computationally derived pseudo-reference sequence from bisulfite converted reads. Our study showcases the potential of epiGBS to perform sequencing based studies in non-model organisms on a scale and level of details that were hitherto impossible.

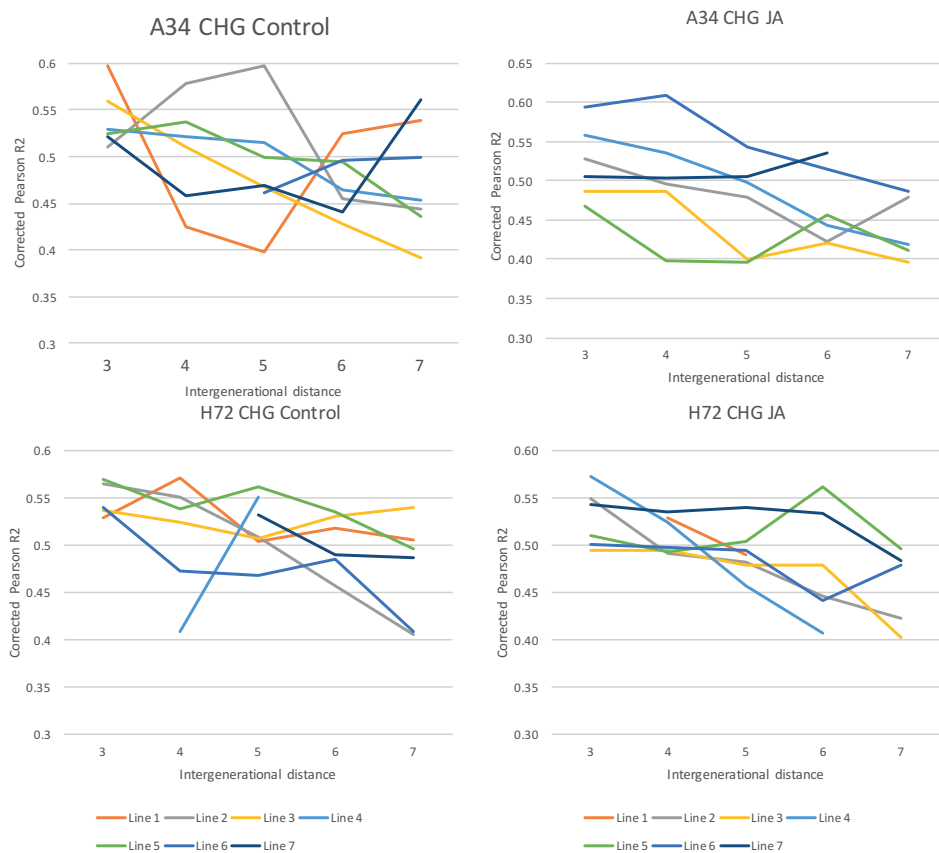
4.4.5 Conclusion and future perspectives

We provide the first bisulfite sequencing-based study on the transgenerational stability of DNA methylation in an asexual plant without a reference genome, providing evidence for heritable divergence over generations in DNA methylation in both CG and CHG contexts. I found few and subtle effects of stress-induced DNA methylation changes that are consistent with recent studies using whole genome bisulfite sequencing in *A. thaliana*. Studies on stress-induced DNA methylation in the model species *A. thaliana* paint an increasingly complex picture, which makes it difficult to assess the functional relevance of DNA methylation variation uncovered in most ecological epigenetics studies. Although stress can induce changes in the DNA methylation status of individual cytosines (DMPs), the functional significance of such changes are unclear (Hagmann *et al.* 2015). This is apparent when considering cytosines in CG context, which are strongly associated with constitutively expressed genes. (C. Jiang *et al.* 2014) found many CG DMPs, but their differentiation was not specifically associated with stress, in contrast to those found in CHH and CHG context (Wibowo *et al.* 2016). The relevance of differential DNA methylation patterns detection using MS-AFLPs, which focus on DNA methylation in CG and not CHG context (Fulneček & Kovarik 2013) is therefore questionable. Future work assessing the suitability of the algorithms currently used to detect DMRs in WGBS data for epiGBS data are however required. Benchmarking studies like we have done developing epiGBS comparing DMR

Chapter 4 Transgenerational inheritance of (stress-induced) DNA methylation in apomictic dandelions.

detection in WGBS data and epiGBS data from the same source can help to guide developments in custom algorithm development. Likewise, targeted GBS approaches employing oligonucleotide baits for specific genomic regions have been developed for genotyping by sequencing (Uitdewilligen et al. 2013). Similar targeted approaches can be easily envisioned to be applicable to epiGBS, allowing for targeted bisulfite sequencing. This approach could be useful even in model organisms, as it could be used to target promoters or other sequences of interest.

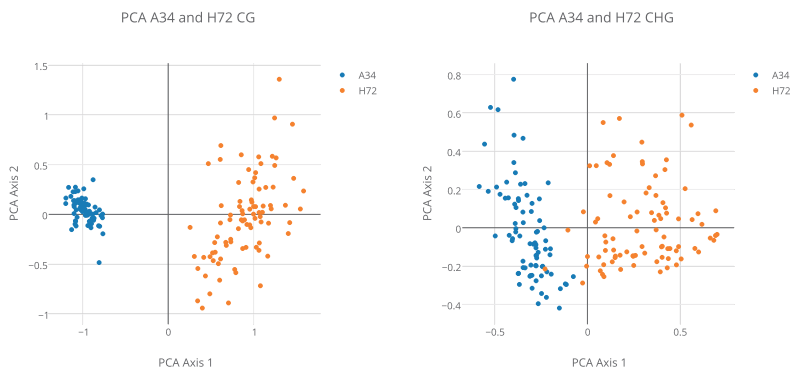
Supporting information



Supplementary Figure 4.1. Intergenerational correlation for CHG methylation for both A34 and H72 for all lines in both the control as well as the Jasmonic Acid treatment lineage. In general, a negative trend is clearly visible, although exceptions occur, most prominently in the control lineage of A34 for cytosines in CHG context, in which for Line 1 and 7 the negative trend is not at all clear.

Supplementary Table 4.2. Cytosines Included in multivariate analysis; without missing observations (called with at least 15 reads in all samples).

Genotype	Sequence context	Total positions	Called in all samples With coverage > 15	Called in all samples and variable across samples
A34	CG	3,137	787	554
A34	CHG	5,898	1,546	848
H72	CG	3,137	838	606
H72	CHG	5,898	1,682	978



Supplementary Figure 4.2. PCA using all samples, results indicate that for cytosines in CG and CHG context, methylation is different between the genotypes; showing clear separation of individuals from both genotypes on the First PCA Axis. Genotype H72 shows more variation in for both CG and CHG context compared to A34.

Chapter 4 Transgenerational inheritance of (stress-induced) DNA methylation in apomictic dandelions.

Supplementary Table 4.3. Independent sample t tests for corrected Pearson R² scores of intergenerational DNA methylation ratios obtained within lines, comparing the scores obtained within JA and control lines.

Genotype	context	generations included	intergenerational distance	t-statistic	p-value	Levene's test
A34	CG	All	3-7	-0.94	0.35	0.59
A34	CHG	All	3-7	0.97	0.33	0.53
A34	CG	All	3	-0.45	0.66	0.02
A34	CHG	All	3	0.62	0.54	0.41
A34	CG	All	4	-0.068	0.946	0.496
A34	CHG	All	4	0.222	0.825	0.798
A34	CG	All	5	-0.319	0.751	0.445
A34	CHG	All	5	0.558	0.581	0.809
A34	CG	All	6	-1.099	0.283	0.299
A34	CHG	All	6	0.488	0.63	0.493
A34	CG	All	7	-0.242	0.814	0.738
A34	CHG	All	7	1.14	0.279	0.187
A34	CG	g1 vs g2	3	0.30	0.77	0.59
A34	CHG	g1 vs g2	3	0.60	0.56	0.33
A34	CG	g2 vs g3	3	0.31	0.77	0.41
A34	CHG	g2 vs g3	3	0.34	0.74	0.62
A34	CG	g3 vs g4	3	0.56	0.59	0.10
A34	CHG	g3 vs g4	3	1.03	0.33	0.04
A34	CG	g4 vs g5	3	-1.70	0.13	0.34
A34	CHG	g4 vs g5	3	-0.93	0.38	0.71
A34	CG	g5 vs g6	3	-1.53	0.16	0.10
A34	CHG	g5 vs g6	3	0.30	0.77	0.09
H72	CG	All	3-7	1.93	0.06	0.33
H72	CHG	All	3-7	1.14	0.26	0.55
H72	CG	All	3	1.28	0.21	0.69
H72	CHG	All	3	0.73	0.47	0.56
H72	CG	All	4	0.17	0.87	0.44
H72	CHG	All	4	0.22	0.83	0.02
H72	CG	All	5	0.50	0.62	0.46
H72	CHG	All	5	-0.03	0.98	0.06
H72	CG	All	6	1.56	0.13	0.45
H72	CHG	All	6	0.64	0.53	0.54
H72	CG	All	7	1.01	0.33	0.82
H72	CHG	All	7	1.47	0.17	0.78
H72	CG	g1 vs g2	3	2.00	0.07	0.41
H72	CHG	g1 vs g2	3	1.11	0.29	0.22
H72	CG	g2 vs g3	3	2.95	0.01	0.98
H72	CHG	g2 vs g3	3	-0.29	0.78	0.31
H72	CG	g3 vs g4	3	0.75	0.47	0.07
H72	CHG	g3 vs g4	3	-0.66	0.53	0.98
H72	CG	g4 vs g5	3	-1.10	0.29	0.07
H72	CHG	g4 vs g5	3	1.37	0.20	0.17
H72	CG	g5 vs g6	3	-0.52	0.61	0.73
H72	CHG	g5 vs g6	3	0.17	0.87	0.93

Supplementary Table 4.4. Blast2go results for 3 gene-related contigs with differentially methylated positions (DMPs) in CHG and CG context.

Contig id	Description	Contig length	# blastHits	Min e-Value	sim mean
129	tetraspanin-3-like	136	20	3.37E-23	81.2%
419	E3 ubiquitin- ligase PRT6	242	20	3.35E-17	96%
466	nucleoporin (DUF3414)	251	20	6.69E-9	68.1%

Description shows the name of the top blast hit, length shows the sequence length of the contigs, min e-value is the e-value of the top blast hit. "Sim mean" is the average similarity of the translated contig sequence compared to that of the top blast hit protein sequence.

Supplementary Table 4.5. Regression results of corrected Pearson R² scores (dependent variable) on intergenerational distance (independent variable) for cytosines in CG and CHG sequence context for comparisons made within lines (intra-line) and between lines of the same lineage (inter-line).

context	line	genotype	lineage	F-statistic	p-value
CG	Intra-line	A34	CO	4.494	0.038
			JA	4.156	0.045
	Inter-line	H72	CO	1.502	0.224
			JA	4.663	0.034
		A34	CO	87.028	<0.000
			JA	17.331	<0.000
CHG	Intra-line	A34	CO	15.016	<0.000
			JA	21.028	<0.000
	Inter-line	H72	CO	11.102	0.001
			JA	10.051	0.002
		A34	CO	56.052	<0.000
			JA	5.047	0.025
H72	CO	77.299	<0.000		
	JA	145.476	<0.000		

**Selection on early flowering in asexually reproducing
Dandelions; using epigenetic variation as a substrate for
artificial selection**

T.P. van Gurp, A. Biere, K.J.F. Verhoeven

To be submitted

5

Abstract

Phenotypic variation in plants is essential for adaptation to adverse or changing environments. Phenotypic variation is determined by the complex interplay of genetic, epigenetic and environmental variation; however, the role of epigenetic variation remains poorly understood. I used apomictic dandelions that occur in a wide range of temperate environments and feature a world-wide distribution to determine the potential role of epigenetic variation in shaping the phenotype. Many dandelion lineages have an asexual reproductive mode and lack recombination which is typical for sexually reproducing organisms. Their asexual reproduction might imply that epigenetic variation is more important in such systems than in sexuals as a source of phenotypic variation, as new genetic variation cannot arise through recombination. Epigenetic variation can be caused by environmental stress and can cause phenotypic variation in traits that can be inherited by subsequent generations. To test if selection on epigenetic variants can drive heritable divergence, I performed artificial selection on an arbitrary yet important and easily identifiable phenotypic trait, flowering time, within genetically uniform apomictic dandelion lineages. I subjected three genotypes of apomictic dandelions to multi-generation stress-pretreatment with the phytohormone Jasmonic acid or single generation treatment with 5-azacytidine which disrupts DNA methylation. I hypothesized that these environmental treatments create additional epigenetic variation that is heritable, so that a selection response is expected to be stronger after these treatments. After exposure, I created selection lines for fast flowering as well as random lines and compared their flowering time for two generations. I found evidence for induction of (epi)genetic variation influencing flowering time in one of the genotypes, as well as clear indications for the presence of standing (epi)genetic variation upon which selection can act.

5.1 Introduction

Phenotypic variation in plants arises as the result of a three-way interaction between the genotype, the epi-genotype and the environment (Bossdorf & Yuanze Zhang 2011). The epi-genotype is defined by epigenetic marks on or associated with the DNA sequence, such as histone modifications or DNA methylation, which can influence DNA structure and gene expression. Part of this epigenetic variation is determined by genetic variation, which has been coined dependent epigenetic variation, whereas facilitated and pure epigenetic variation are respectively partly and completely independent on genetic variation (E. J. Richards 2006). Epigenetic marks serve an important role in, for instance, silencing transposable elements (TE), thus contributing to preserving genome integrity (Slotkin & Martienssen 2007). DNA methylation variation is of particular interest to ecologists and evolutionary biologists as, in contrast to histone modifications, DNA methylation changes can be transgenerationally heritable (Verhoeven et al. 2016). In plants, DNA methylation occurs at cytosines in CG, CHG and CHH context, where H is A, C or T. Differentially Methylated Positions (DMPs) are often heritable. The functional significance of DMPs is often low, whereas regional differences in DNA methylation (differentially methylation regions, DMRs) are often implicated in differential gene expression (Weigel & Colot 2012).

Contrary to genetic changes, epigenetic changes can occur rapidly and in response to environmental perturbations, which might allow plants to adapt to diverse environments without any genetic change which takes a very long time to accumulate in asexual plants. Genetically uniform *Arabidopsis thaliana* epigenetic recombinant inbred lines (epiRILs) with recombined genomes that are mosaics of artificially unmethylated and normally methylated chromosomes show a wide distribution in complex traits such as flowering time, plant height and primary root length. Variation in these traits was linked to stably transmitted DNA methylation variants which were detected as epigenetic quantitative trait loci (Johannes et al. 2009; Cortijo et al. 2014) accounting for 60% to 90% of the heritability of flowering time and primary root length (Cortijo et al. 2014). (Dubin et al. 2015) surveyed epigenetic and genetic variation in a natural *A. thaliana* population and found a strong association between epigenetic and genetic variation, suggesting that the genetic variation determines the epigenetic variation. However, also pure (uncoupled) epigenetic variants occur. Sophisticated analysis of genetic variation and differentially methylated regions (DMRs) associated with gene expression changes in naturally occurring *A. thaliana* accessions indicate the presence of DMRs that are not linked to genetic variation, suggesting that epigenetic variation independent of the DNA sequence can play a role in adaptation of plant populations (Schmitz, Schultz, et al. 2013).

Epigenetic variation that is heritable but uncoupled from genetic variation has the potential to contribute to selection responses in the absence of any genetic variation. Even when no permanent epigenetic adaptation may be produced because heritable epigenetic mutations are metastable and revert, selection acting on metastable epimutations has significant effects on the dynamics of adaptation even when reversal rates are high (Klironomos et al. 2013). Selection for improved energy use efficiency and drought tolerance in isogenic doubled haploid canola (*Brassica napus*) over three subsequent generations resulted in epilines with enhanced energy use efficiency (Hauben et al. 2009), drought tolerance and nitrogen use efficiency (Verkest et al. 2015). The selected lines showed differentially

expressed stress-tolerance regulating genes, which was accompanied by increased trimethylation of the permissive histone mark (H3K4me3) in these genes (Verkest et al. 2015), indicating that epigenetic variation was associated with, and presumably responsible for, the enhanced stress tolerance. Because of the limited scope for de novo genetic mutations as a basis for trait variation in these experiments (Verkest et al. 2015), strong evidence for the potential of epigenetic variation to effect phenotypic changes in plants is thus provided, as well as the ability to select on it. Because these studies in Canola (*Brassica napus*) used doubled haploid lines (Verkest et al. 2015; Hauben et al. 2009), and haploidization has been shown to impact DNA methylation (S. C. A. F. Chen 2014), it remains to be demonstrated to what extent natural heritable epigenetic variation can be a target of selection in genetically uniform wild-type plants.

Stressful environments can lead to both enhanced spontaneous genetic mutation rates as well as an enhanced epimutation rates in plants (Rapp & Wendel 2005; C. Jiang et al. 2014). A study tracking both spontaneous genetic mutations and epimutations in a mutation accumulation lineage of *A. thaliana* grown for ten successive generations under salt stress resulted in an increase of ~45% in intergenerational differentially methylated positions and a ~100% higher genetic mutation rate compared to rates observed for control plants grown in normal soil (C. Jiang et al. 2014). Even though this study resulted in a 2-fold higher increase in the mutation rate compared to the epimutation rate, the absolute increase in the number of epimutations per generation is much higher than the increase in number of mutations per generation as the mutation rate for *A. thaliana* is estimated at $7 \cdot 10^{-9}$ (Ossowski et al. 2010) whereas the epimutation rate in *A. thaliana* is estimated to be $4 \cdot 10^{-4}$ (van der Graaf et al. 2015), a difference of five orders of magnitude. Although the stress-induced genetic changes found in the salt stress mutation accumulation study showed a distinct pattern, with a specific increase in relative frequency of transversions and insertion / deletions, the stress-associated CG-DMPs were significantly overrepresented in genic regions. This suggests a specific response in DNA methylation changes that could potentially lead to heritable DNA methylation-induced gene expression changes (C. Jiang et al. 2014), although genic methylation in plants does not always have functional consequences (Niederhuth & Schmitz 2016).

Many more examples of induced DNA methylation changes are known (Peng & Jing Zhang 2008). Chemical alteration of DNA with 5-azacytidine, a cytidine analog that inhibits DNA methylation, has proven effective in changing genome-wide DNA methylation patterns (Christman 2002). 5-azacytidine can be incorporated in DNA during replication, thereby inhibiting methyltransferases and causing partial demethylation (P. A. Jones 1985; Burn et al. 1993; Fieldes & Amyot 2000; Bossdorf et al. 2010). Treatment with chemicals such as Jasmonic Acid (JA) or salicylic acid (SA), both signaling molecules involved in plant defense response pathways, can result in large and specific effects on DNA methylation patterns. Jasmonic acid is a natural plant hormone used as a signaling hormone to initiate a stress response (Pieterse & Dicke 2007) which can be used to experimentally mimic biotic attack and to induce defense pathways (Verhoeven & van Gorp 2012). SA application on cell cultures of *Vitis amurensis* induces specific changes in the DNA methylation in genes involved in secondary metabolite production, changing their expression (Kiselev et al. 2014). JA and SA-treated dandelions, *Taraxacum officinale* showed a significant increase in

polymorphic MS-AFLP bands compared to control plants. Such polymorphisms were by and large transmitted to the next generation, indicating that the treatment had caused heritable DNA methylation changes (Verhoeven, Jansen, et al. 2009), but see also chapter 4 of the present thesis. Stress can thus induce heritable DNA methylation variation, which can underpin (specific) stress related responses. However, the majority of stress-induced DNA methylation modifications may not be heritable (Grossniklaus et al. 2013; Hagmann et al. 2015) and little is known about the potential role of this induced DNA methylation variation to facilitate selection.

Few studies in plants have explored the potential of selection on phenotypic traits by means of stress-induced changes in DNA methylation (Xu 2016), whereas this could offer valuable insights in the role epigenetic variation can play in adaptation or plant breeding (Jaligot & Rival 2015; Peng & Jing Zhang 2008). Here, I select on early flowering in the genetically uniform offspring of apomictic dandelion populations, and I test if a selection response is facilitated after exposure to Jasmonic acid (JA) or 5-azacytidine. I hypothesize that flowering time in apomictic dandelion can be influenced by changes in DNA methylation. I expect that 5-azacytidine and JA treatment will trigger more heritable DNA methylation changes, which should result in a stronger selection response in these lineages. I performed the selection experiment on three different genotypes of apomictic dandelions, with two established and one recent triploid apomict to assess whether genotypic differences exist in selection response. In this study, I aim to answer the following questions: (1) Does sufficient heritable variation arise within a few generations to support a response on selection on flowering time? (2) Is the response to selection stronger for plants that received a Jasmonic Acid or 5-azacytidine pre-treatment? And (3) Is there a difference in the selection response among the three genotypes?

I hypothesize that (1) Artificial selection on early flowering can lead to a shift in flowering time in subsequent generations as (epi)genetic variants responsible for this early flowering phenotype are selected upon. Furthermore, (2) I expect that stress treatments with 5-azacytidine and Jasmonic Acid create more heritable epigenetic variation influencing flowering time that will result in a stronger selection response for pre-treated lines. Finally (3) I hypothesize that genetic and epigenetic differences between the three genotypes will lead to differences in the response to treatment, and the efficacy of selection on both stress treated and untreated lines. I expect a higher level of genomic instability for the recent apomictic line A34, which could facilitate a stronger selection response independent of pre-treatment based on natural (epi)genetic variation. Even though the trait that I select upon is chosen mainly due to the ease with which it can be scored, flowering time is an important phenotypic trait under strong selection in sexual plant populations (Ehrlén & Münzbergová 2009). For outcrossing plants, synchronous flowering is important as plants need to exchange pollen to reproduce. Early flowering plants have a demonstrated advantage over later germinating conspecifics (Collier & Rogstad 2004) and references therein, suggesting that selection for early flowering might occur in apomictic Dandelions. Flowering time is a tightly regulated process, controlled by complex independently developed multi-gene pathways incorporating sensitivity to both day length and temperature (vernalization) as parameters to adjust flowering time (Andrés & Coupland 2012). In dandelion, previous

work demonstrated that natural, heritable divergence in flowering time within an apomictic lineage is mediated by DNA methylation (Wilschut et al. 2016).

5.2 Materials and Method

5.2.1 General experimental set-up

For each of the three genotypes (A34, A68 and H72) used in this experiment, offspring from a single founding individual was exposed for three consecutive generations (without conscious selection) to either control conditions or Jasmonic acid treatment. Jasmonic acid was applied for three subsequent generations as repeated exposure to the same stress is thought to induce 'priming' or 'acclimation' via epigenetic changes ((Wibowo et al. 2016) and references therein). For the first two generations (T1 and T2, see Fig. 5.1), 8 replicate lines per group were maintained using single-seed descent. To maximize the potential for JA-induced epigenetic changes with an impact on flowering time, the cohort size was increased from 8 to 64 plants in T3 (8 offspring from each of 8 T2 individuals per group). From T3, single-seed derived offspring formed the base populations S0 from which I started the selection experiment. In addition to a 'control' and a 'JA' pre-treated base population, a third 'azacytidine' pre-treated base population was generated by exposing an additional group of 64 T3 plants from the control lineages to a 5-azacytidine treatment. I initially used two different azacytidine concentrations but I proceeded the selection generations with only one of these (described below). From each base population, I derived two selection lines: one selected for early flowering (selection line), and one selected randomly (random line). Per genotype, the different lines are denoted by the first letter of their parental treatment ("A" for 5-azacytidine, "J" for Jasmonic acid and "C" for Control) combined with "R" for random and "S" for the selection line respectively, resulting in six lines per genotype (AR, AS, JR, JS, CR and CS). In the S0 and S1 generations, I selected the 16 (25%) fastest flowering plants (selection line) and 16 random plants (random line) to each contribute four apomictic offspring individuals to the next generation for these two lines, respectively.

Experimental design selection lines

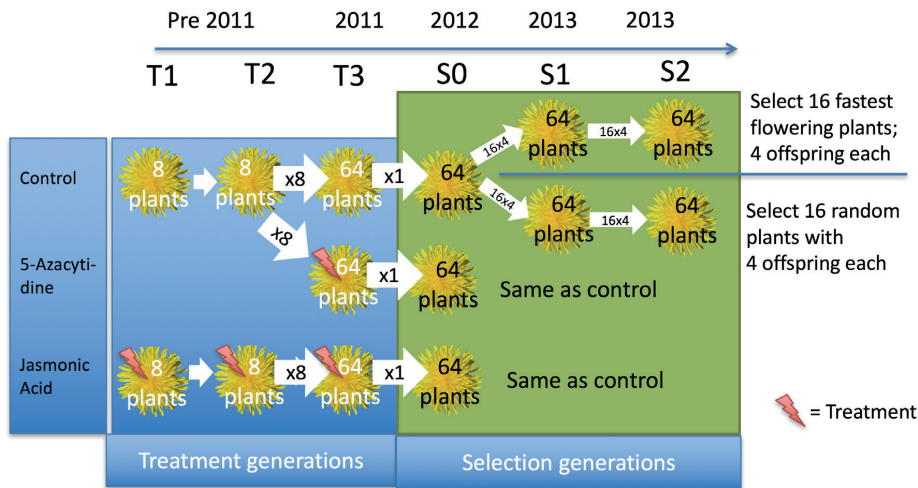


Figure 5.2 Experimental design for treatment generations and subsequent selection.

5.2.2 Plant material

I subjected three genotypes of the asexual apomictic common Dandelion, *Taraxacum officinale* Weber ex Wigg., A34, A68 and H72, to the selection experiment described above to evaluate whether selection results are genotype dependent. A34 is a recent synthetic apomict which was produced in an experimental cross between a sexual diploid mother and diploid pollen from a triploid apomictic father (Tas & Van Dijk 1999; Verhoeven, van Dijk, et al. 2009). A68 and H72 are established apomicts collected from the field (Verhoeven & van Gurp 2012). Asexual dandelions are polyploid (usually triploid, $3N=24$) obligate apomicts that produce clonal seeds in a process that involves unreduced egg cell formation (diplospory), parthenogenic embryo development and autonomous endosperm formation (Van Dijk et al. 1999).

5.2.3 Treatment generation T1 to T3

For all plants in the JA lines in the three treatment generations T1-T3, JA was applied twice, when plants were 5 and 7 weeks old. Applications were dosed at 0.25 ml (week 5) and 0.75 ml (week 7) of a 10 mM JA solution (Sigma J-2500, dissolved in ethanol and diluted to the desired concentration with a 0.1% Triton X-100 surfactant solution) which was applied and manually distributed over the surface of two leaves (week 5) or four leaves (week 7). T3 'azacytidine' seedlings were germinated on agar containing either 0.1 μ M or 1.0 μ M of 5-azacytidine (AZA). A 8,000 μ M stock solution was made by dissolving 100 mg of 5-azacytidine in 51.2 mL of sterile water. Appropriate dilutions were made by mixing this stock solution with 0.8% agar, which was poured in petri-dishes that were sealed with parafilm after seeds were put in. Seeds were germinated for 10-11 days (10 h dark: 14 h

light; 14°C: 18°C) after which seedlings were transplanted individually to pots. In T3, per experimental group 64 seedlings were individually transplanted into 9 cm pots containing a mixture of 80% potting soil and 20% pumice. T3 plants were placed together in blocks of 64 individuals per genotype-treatment combination on a single table. For each of the three genotypes, four different lines for the pre-treatments exist (control, JA, AZA 0.1 and 1.0 μM). These four lines with 64 plants each result in a total of 768 plants (64 plants x 4 lines x 3 genotypes) which were put on a single large table in a greenhouse with (10 h dark: 14 h light; 16°C: 21°C). Plants were watered every ~2 days, twice per week the table was flooded with Hoagland nutrient solution. Plants were grown until the large majority of plants produced flowers and seed heads. Flowering time was noted daily. Flowering time was defined as the number of days it took until the first flower head had opened and showed emerging yellow petals for more than half of the flower-head, with the seedling transplantation in the pot as day 1. Per plant, one seed-head was collected and stored in a seed storage chamber at 15 °C at 30% relative humidity (RH).

5.2.4 Base population S0

The base-population for selection was obtained using single-seed descent from T3. The expected number of plants for this generation is 768 (3 genotypes · 4 pre-treatments · 64 plants). Per line, between 79% - 96% of the parental plants produced seeds that germinated. 12 seeds per parent plant were germinated on 0.8% agar for 7 days. 82 parent plants did not produce germinating seeds or had failed to flower before termination of the T3 generation, leaving a total of 686 parent plants with seeds that could be included in S0. These 686 seedlings were transferred to 9cm pots containing a mixture of 80% potting soil and 20% pumice and put in a greenhouse under the same conditions as described for T3. In contrast to T3, plants from all lines were fully randomized per genotype on a single table. Pots were labeled with a barcode linked to a unique identifier to facilitate fast and accurate recording of flowering time per plant. A downey mildew infection occurred during growth in the greenhouse on several tables, for which treatment with Baycor Flow 1.0 ml/l (active ingredient Bitertanol) with Agral 1.0ml/l was given several times. Flowering time was recorded daily using a barcode-scanner. After seed-set, complete seed-heads were collected in a barcoded paper bag, after which the plant was discarded. Seeds were stored under seed storage conditions as described previously.

5.2.5 Correcting flowering time for spatial effects

A clear position effect was observed for flowering time: flowering varied with pot position on the greenhouse tables (see Fig. 5.2). I corrected for this environmental variation and used the corrected flowering time for subsequent analysis and selection (for details see Fig 5.2).

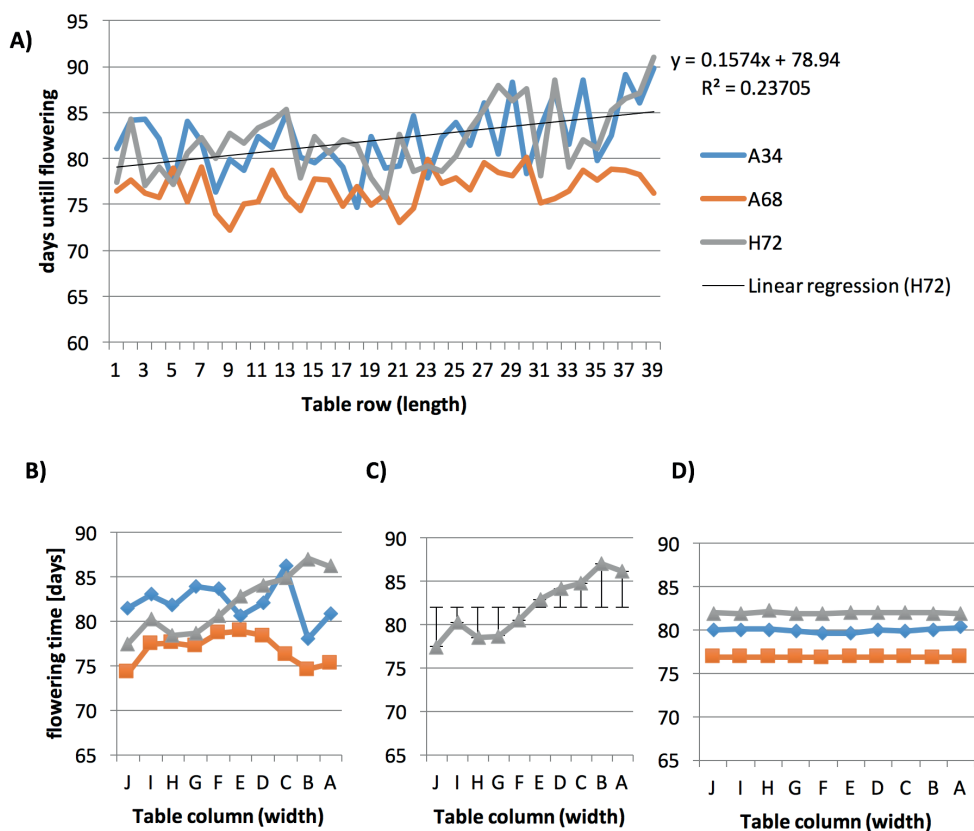


Figure 5.2 Correcting flowering time for table positions effects. **(a)** A positive trend is observed for flowering time dependent on the table row position for A34 and H72. Each table had 39 rows and 10 columns, providing space for 390 plants. A correction is applied per table (each with one genotype) by adding the difference between the average (per table) flowering time with the outcome of the equation derived from linear regression of flowering time on table row. Written as a formula this results in: Corrected flowering time = Observed flowering time + per table average flowering time (= 82 for H72) – position predicted flowering time. Predicted flowering time is calculated based on the equation derived from linear regression of table position (row) (independent variable) on flowering time (dependent variable). For H72 this relation is described as predicted flowering time = $0.1574 \cdot \text{row_position} + 78.94$; for A68 an equation was derived in identical fashion. **(b)** Position effects were also observed over the width of the tables, these effects were however not always linear, for A68 plants on both edges flowered earlier than central plants. A per-column correction was thus applied. Per genotype, the difference between the average per-column flowering time with the average flowering time of the whole table was added to the flowering time for every plant in that column (see **c**), leading to an average per column flowering time that is corrected for the observed table column effect (see **d**). Written as a

formula this results in: Corrected flowering time = Observed flowering time + per table average flowering time (= 82) – per column average flowering time.

5.2.6 Phenotypic selection

Based on the flowering time distribution analysis of T3 (data not shown), which showed that many AZA 1 μM plants for all genotypes flowered very late, I decided to discard plants with AZA 1 μM pre-treatment, leaving a control, JA and AZA 0.1 μM pre-treatment from which I started selection in S0 with 64 plants each, making for a total of six lines per genotype in subsequent generations S1-S2 (see Fig. 5.1b). As discarding the AZA 1 μM line leaves us with only the AZA 0.1 μM line I refer to this line as “AZA” from here on. As in earlier (treatment) generations fungal infections were sometimes present on germinating seeds, I proceeded to surface-sterilize seeds starting from S0 onwards. 20 seeds per plant were surface-sterilized for 5 minutes with 0.5% sodium hypochlorite including 0.05% Tween20 (Sigma-Aldrich, Zwijndrecht, The Netherlands) and afterwards washed with demineralized water. Sterilized seeds were germinated in individual petridishes sealed with parafilm on moist filter paper. Seeds were germinated for 10-11 days (10 h dark: 14 h light; 14°C : 18°C). Per petridish, four seedlings were individually transplanted into 9cm pots containing a mixture of 80% potting soil and 20% pumice. Plants were assigned a random number generated using random.org. Per genotype, 384 plants (3 pre-treatments x 2 selection lines x 64 plants) were put on a single table in ascending order according to the random number generated per plant. Plants were watered as described previously in a greenhouse maintained at the same conditions as described for previous generations. Selection was applied as described above for two consecutive generations in S0 and S1. A final evaluation of flowering time was done in S2, under common greenhouse conditions as described for the previous generations.

Data analysis

Position-corrected flowering time for all individuals from S0, S1 and S2 were imported in SPSS v24 for Mac OSX. To test whether pre-treatments resulted in an altered mean flowering time, I tested per genotype whether the mean position-corrected flowering time of the S0 plants that received a AZA 0.1 μM , AZA 1 μM or JA pre-treatment differed significantly from plants with a control pre-treatment using a one-way ANOVA (see Supplementary table 4). The distribution of the dependent variable position-corrected flowering time was not normal. I thus tested the difference on position-corrected flowering time on a LN-transformed position-corrected flowering time using a one-way ANOVA using pre-treatment (control, AZA 0.1 μM , AZA 1 μM or JA) as the independent and LN-transformed position-corrected flowering time as the dependent variable. I tested for equality of variance using Levene’s test on LN-transformed corrected flowering time, which did not result in any significant p-values. Standardized residuals were plotted to assess the normality of the distribution. Although slight deviations from normality were detected all distributions approximately followed a normal distribution, which improved after LN-transformation (data not shown).

I performed a two-way ANOVA with pre-treatment (control, AZA or JA) and selection regime (selection or random) as the independent variables and LN-transformed position-corrected flowering time as the dependent variable. Per genotype and generation, I begin by

describing the main selection effect detected using this two-way ANOVA. Subsequently, I pinpoint the pre-treatment in which a significant response to selection for early flowering was detected between the selection and random line using a one-way ANOVA; comparing the random and selection line within each pre-treatment group. For these one-way ANOVA's I tested for equality of variance between the random and control line using Levene's test. In some cases, significant deviations from variance equality were observed. For these cases, I performed tests with a robust ANOVA model that allows for such differences in variance, deriving p-values for both a Welch and Brown-Forsythe test. None of the statistically significant p-values changed after these corrections, likewise I did not observe statistically significant p-values from these tests that did not confirm p-values obtained from a standard ANOVA. I thus report Welch/Forsythe p-values (which were always equal) in case of significant Levene's test indicating unequal variance and report p-values on standard ANOVAs otherwise. Furthermore, I also explore whether there is a significant difference between the selection lines of the JA and AZA pre-treatment group (JS and AS) compared to the random line of the control pre-treatment group using one-way ANOVAs (Supplementary Table 5.2). This provides additional information on the coherence of the response observed within a pre-treatment group. In principle, when a significant selection effect is evident between the selection and random line of the JA or AZA pre-treatment group this should also give rise to a significant difference between the selection line of this pre-treatment group and the random line of the control group. Finally, I evaluate whether pre-treatment has a significant impact on the response to selection for early flowering by looking at the p-value of the interaction term pre-treatment x selection regime in the two-way ANOVA. In case the pre-treatment has an impact on the selection response, the interaction term pre-treatment x selection regime is expected to be significant. In contrast to the previous tests, the two-way ANOVAs were performed using SAS 9.2 as SPSS does not allow for defining a priori contrast tests of the selection response. I used a-priori defined contrast test to evaluate if the difference in mean position-corrected flowering time between the selection and random group (selection response) is different in the JA and AZA pre-treatment groups compared to the control group, which would indicate a dependence of the selection response on (stress) pre-treatment.

5.3 Results

5.3.1 Treatment effects on seed germination in T3

I observed a negative effect of 5-azacytidine treatment on germination. The percentage of 1.0 μM AZA seedlings failing to germinate was significantly higher than for control plants in both A68 (Chi-square 7.34 $p=0.007$) and H72 (Chi-square 9.3 $p=0.002$), whereas it did not affect germination in A34 (Chi-square 0.49 $p=0.49$). JA pre-treatment resulted in a 2- and 3-fold higher germination failure in A34 and H72, respectively (see Table 1) whereas it did not have any impact on the germination rate in A68. Overall, the germination failure rate of control plants without treatment was 3 times higher for the young apomict A34, potentially indicating genomic instability in this recent apomict.

Table 1) Pre-treatment dependent non-germination rates per genotype-treatment in T3.

Genotype	AZA 0.1 μM	AZA 1 μM	CONTROL	JA
A34	11%	13%	10%	21%
A68	14%	18%	3%	4%
H72	10%	20%	3%	9%

Percentages indicate the fraction of the 64 parental plants of T3 did not have any successfully germinating seeds.

5.3.2 Treatment induced transgenerational effects on flowering.

If 5-azacytidine or JA treatment would successfully alter heritable DNA methylation patterns influencing flowering time, this should result in a change in either the mean or variance of flowering time of the pre-treated plants (JA or AZA), for which the strongest response is expected in S0. I did not observe any effect of pre-treatment on either the mean or variance of position-corrected flowering time for A34 S0. For A68 S0 I found no difference in mean flowering time, but variance was significantly affected ($p=0.041$ Supplementary table 4). Subsequent pairwise one-way ANOVAs between control and pre-treatment groups AZA 0.1, AZA 1.0 and JA did however not result in a significant p -value for Levene's test. Comparing AZA 0.1 vs JA and AZA 0.1 vs AZA 1.0 in A68 S0 did however result in significant p -values (see Supplementary table 4), suggesting that contrasting parental effects due to line pre-treatments with JA and 0.1 μM 5-azacytidine could be picked up in the test including all pre-treatment lines, as the JA pre-treatment group showed a slightly increased variance whereas the AZA 0.1 pre-treatment group showed a decrease of the variance in flowering time distribution (Fig. 5.3 and Supplementary table 4). For H72 in the S0 generation, I found a significant delay of ~ 5 days in flowering for the offspring of 5-azacytidine treated plants for both the 1 μM ($p<0.001$) and 0.1 μM ($p<0.001$) 5-azacytidine concentrations (see Supplementary table 4 and Fig. 5.3). Levene's tests for one-way ANOVAs comparing variances of mean position-corrected flowering time between control and both 5-azacytidine concentrations used resulted in non-significant p -values (0.375 for AZA 1 μM and 0.403 for AZA 0.1 μM) indicating that the variance in flowering time was not affected. For the JA pre-treatment however, the ANOVA result was non-significant but the Levene's test for equality of variance was marginally significant ($p=0.06$), with a higher variance found for the JA pre-treatment group in S0 (see Supplementary table 4).

Table 5.2) Effects of pre-treatment and selection on flowering time in S1 and S2 generations.

Genotype	generation	ANOVA type	Factor / comparison	F-value	p-value	Levene's test
A34	S1	One-way	CS vs CR		0	0.318
			JS vs JR		0.139	0.868
			AS vs AR		0.004	0.07
		2-way	Pre-treatment	13.66	<.0001	
			Selection regime	9.51	0.0022	
			Pre. treatm. X sel. Reg.	8.00	0.0004	
	S2	One-way	Sel C vs AZA	0.97	0.3263	
			Sel C vs JA	14.60	0.0002	
			CS vs CR		0.012*	0.014
		2-way	JS vs JR		0.001*	0.01
			AS vs AR		0.078*	0.048
			Pre-treatment	2.43	0.0892	
		2-way	Selection regime	19.31	<.0001	
			Pre. treatm. X sel. Reg.	0.61	0.5450	
Sel C vs AZA	1.13		0.2879			
Sel C vs JA	0.08		0.7736			
A68	S1	One-way	CS vs CR		0.855	0.924
			JS vs JR		0.048	0.089
			AS vs AR		0.374	0.966
		2-way	Pre-treatment	2.44	0.0884	
			Selection regime	0.10	0.7556	
			Pre. treatm. X sel. Reg.	1.40	0.2488	
	S2	One-way	Sel C vs AZA	0.72	0.3974	
			Sel C vs JA	0.60	0.4392	
			CS vs CR		0.031	0.504
		2-way	JS vs JR		0.001	0.643
			AS vs AR		0.671	0.498
			Pre-treatment	3.36	0.0359	
		2-way	Selection regime	0.06	0.8074	
			Pre. treatm. X sel. Reg.	8.25	0.0003	
Sel C vs AZA	2.07		0.1510			
Sel C vs JA	16.07		<.0001			
H72	S1	One-way	CS vs CR		0.599	0.233
			JS vs JR		0.695	0.258
			AS vs AR		0.024*	0.002
		2-way	Pre-treatment	0.81	0.4440	
			Selection regime	1.51	0.2206	
			Pre. treatm. X sel. Reg.	1.99	0.1382	
	S2	One-way	Sel C vs AZA	3.85	0.0506	
			Sel C vs JA	0.44	0.5071	
			CS vs CR		0.644	0.256
		2-way	JS vs JR		0.475	0.244
			AS vs AR		0.250	0.146
			Pre-treatment	4.51	0.0117	
		2-way	Selection regime	1.82	0.1777	
			Pre. treatm. X sel. Reg.	0.14	0.8724	
Sel C vs AZA	0.27		0.6069			
Sel C vs JA	0.11		0.7401			

One-way and two-way ANOVA results on position corrected LN-transformed flowering time are presented per genotype and generation in Table 5.2. One-way ANOVA results are tests of the selection effect (comparing random versus selected lines) within each pre-treatment group. Significant p-values (<0.05) are shown in bold. In case of a significant result for Levene's test for equality of variance the Welch test p-value based for unequal variance is shown (*). Two-way ANOVAs are performed with independent variables pre-treatment group (control, AZA 0.1 or JA), selection regime (random or selection line) and the interaction term pre-treatment x selection regime. For two-way ANOVAs the results are also shown of a priori defined contrast tests that compare the selection response after AZA or JA pre-treatment to the selection response after control pre-treatment.

5.3.3 Response to selection in A34 S1

The main effect of selection regime in the two-way ANOVA was highly significant ($F=9.5$ $p=0.002$, Table 5.2 and Fig. 5.5), indicating a significant positive response to selection for early flowering. One-way ANOVAs performed between the random and selection line within each of the pre-treatment groups shows that selection resulted in significantly earlier flowering in plants from the control (6.5 days $p<0.0001$) and AZA pre-treatment group (4.2 days $p=0.005$), but not in the JA pre-treatment group (see Fig. 5.5). P-values for Levene's test for equality of variance were not significant for any of the one-way ANOVAs comparing selection and random line per pre-treatment group, indicating that the first selection round did not result in a significant change of the variation in flowering time. The differences in selection response among pre-treatment groups resulted in a significant pre-treatment x selection regime interaction for the 2-way ANOVA ($F=8.0$ $p<0.001$, Table 5.2). A priori contrast test comparing the selection response in the control group (CS-CR) with that of the AZA and JA group indicated a significantly different selection response in the control group compared to the JA group ($F=14.6$, $p<0.001$, Table 5.2) and no difference in selection response between the control and AZA group ($F=1.0$, $p=0.326$, Table 5.2), suggesting that pre-treatment with azacytidine did not lead to a stronger selection response. All A34 lines flowered earlier in S1 compared to S0 (Supplementary Table 5.3), except for CR which flowered 4.5 days later in S1 compared to S0, a significant delay (1-way ANOVA with Welch test for unequal variance $p=0.013$) suggesting that the significant interaction term pre-treatment · selection regime observed in the two-way ANOVA might reflect an unexpected delay in flowering in the CR group in S1 rather than a strong selection response in the AZA or JA pre-treatment groups. Both the random and selection line of the JA group (JS and JR) flowered significantly earlier than the random line of the control group (CR) (Supplementary Table 5.2). The AS line flowered significantly earlier than the CR line and so did all other lines, indicating a delayed flowering of the CR line compared to all other lines.

5.3.4 Response to selection in A34 S2

Consistent with the results obtained in S1, the main effect of selection regime in the two-way ANOVA was highly significant ($F=19.3$ $p<0.001$, Table 5.2) with selection lines flowering earlier than random lines in all but the AZA pre-treatment group (Fig. 5.5), indicating a significant positive response to selection for early flowering. In the control and the JA group, selection resulted in a significant 3.9 and 3.1 days earlier flowering, respectively. In the AZA group, selection only resulted in a marginally significant, 2 days earlier flowering ($p=0.076$). However, the interaction term pre-treatment x selection

regime of the two-way ANOVA was not significant ($F=0.61$ $p=0.55$, Table 5.2), indicating that the selection response was comparable in all three pre-treatment groups. The p-values for Levene's test for equality of variance of the one-way ANOVAs comparing the selection and random line per pre-treatment group were significant for all groups (see Table 5.2). Standard errors for the flowering time distribution of selection lines of all pre-treatment groups were significantly lower compared to both the random lines of S2 as well as the selection lines of the previous generation S1 (Supplementary Fig. 5.2), indicating that two subsequent rounds of selection on early flowering resulted in a significant decrease of variance in flowering time, irrespective of pre-treatment. The selection line of the AZA group (AS) flowered 4.6 days earlier than the random line of the control group (CR), a highly significant difference ($p<0.001$) (See Fig. 5.5 and Supplementary Table 5.2). JS also flowered significantly earlier than CR. In both generation S1 and S2 I find a pre-treatment group for which no significant difference in flowering time is detected between the selection and random line (JA for S1 and AZA for S2). In both cases these selection lines do flower earlier compared to the random group of the control line.

5.3.5 Response to selection in A68 S1

The main effect of selection regime in the two-way ANOVA was not significant ($F=0.10$ $p=0.76$), indicating that no clear overall response to selection for early flowering was achieved. Nevertheless, one-way ANOVA indicated a significant response to selection after JA pre-treatment. P-values for Levene's test for equality of variance of per pre-treatment group one-way ANOVAs were non-significant for all groups (see Table 5.2). The p-value for the interaction term pre-treatment * selection regime of the two-way ANOVA was not significant ($F=1.40$ $p=0.25$), indicating no significant difference in the selection response due to pre-treatment occurred. Likewise, no significant differences were detected between AS vs CR and JS vs CR (see Supplementary Table 5.2).

5.3.6 Response to selection in A68 S2

Consistent with the results obtained in S1, the main effect of selection regime in the two-way ANOVA was not significant ($F=0.06$ $p=0.81$, Table 5.2), indicating that no clear overall response to selection for early flowering was achieved. The absence of an overall effect was due to the large differences in selection response among pre-treatment groups. Per-group one-way ANOVAs revealed that in the control group, selection resulted in significant, 2.3 days, earlier flowering ($p=0.016$, Table 5.2). By contrast, in the JA group I observed a reverse trend; selection resulted in a significant ($p=0.001$, Table 5.2) three-day delay in flowering, whereas no response to selection was observed in the AZA group. The reverse trend in the JA group seemed to be caused by unexpected early flowering in the JR line, which flowered 2.3 days earlier than CR, a significant difference ($p=0.037$, Supplementary Table 5.2). This observation cannot be explained by a biased sampling of mother plants of S2 JR towards fast flowering plants, as the mean position corrected flowering time of the 16 randomly selected mother plants from JR S1 was 76 days whereas that for the 16 randomly selected mother plants from CR S1 was 74 days (data not shown). AS flowered 2.3 days earlier compared to CR, a significant difference ($p=0.019$, Supplementary Table 5.2). The differences in response resulted in a significant pre-treatment x selection regime interaction for the two-way ANOVA ($F=8.25$ $p<0.001$, Table 5.2). A priori contrast test comparing the selection response in the control group (CS-CR) with that of the AZA and JA

groups indicated a significant difference in the selection response between the control and JA group ($F=16.1$ $p<0.0001$, Table 5.2), which is to be expected given the reverse trend (negative selection response) observed in the JA group (see Fig. 5.5). Pre-treatment did not result in a significant increase in the selection response towards earlier flowering. P-values for Levene's test for equality of variance for one-way ANOVAs were non-significant for all groups (see Table 5.2), indicating that selection on early flowering did not result in a decreased variance as observed for A34 S2.

5.3.7 Response to selection in H72 S1

The main effect of selection regime in the two-way ANOVA was not significant ($F=1.51$ $p=0.22$, Table 5.2), indicating the absence of an overall significant positive response to selection for early flowering. Per-group one-way ANOVAs revealed that selection significantly affected flowering time for the AZA pre-treatment group only; plants within this group that had been subjected to selection (AS) flowered on average 5 days earlier than plants from the corresponding random line AR ($p=0.026$, Table 5.2 and Fig. 5.5). This difference was also significant when comparing AS to CR ($p=0.036$, supplementary Table 5.2). However, the interaction term pre-treatment x selection regime of the 2-way ANOVA was non-significant ($F=1.99$ $p=0.14$, Table 5.2), indicating that the observed differences in the selection response among pre-treatment groups were not statistically significant, although the contrast specifically testing difference in the strength of the selection response between control and AZA group was close to significance ($P=0.051$, Table 5.2), indicating that AZA pre-treatment tended to result in a stronger response to selection, consistent with pre-treatment induced epigenetic variation influencing flowering time. Levene's test for equality of variances showed a highly significant ($p=0.002$, Table 5.2) difference between plants from the AZA group that had been subject to selection (AS) compared to their randomly selected counterparts (AR), with plants from the selection line showing a lower standard error (see Supplementary Fig. 5.2). No differences in standard error were detected between the random and selection line of the control and JA group.

5.3.8 Response to selection in H72 S2

Consistent with the results obtained in S1, the main effect of selection regime in the two-way ANOVA was not significant ($F=1.82$ $p=0.18$, Table 5.2), indicating that no significant overall response to selection for early flowering was achieved. Also, one-way ANOVA's contrasting random and selection lines within each group did not show any significant difference, indicating that no selection effects were expressed in this generation. For the AZA group however, AS flowered ~2,3 days faster than CR, which was significant ($p=0.023$, Supplementary Table 5.2). The p-value for the interaction term pre-treatment x selection regime of the two-way ANOVA was non-significant ($F=0.14$ $p=0.87$, Table 5.2), indicating no significant difference in the selection response due to pre-treatment. The pretreatment term was however significant ($p=0.012$, Table 5.2). P-values for Levene's test for equality of variance for one-way ANOVAs were non-significant for all groups (see Table 5.2), indicating that selection on early flowering did not result in a decreased variance.

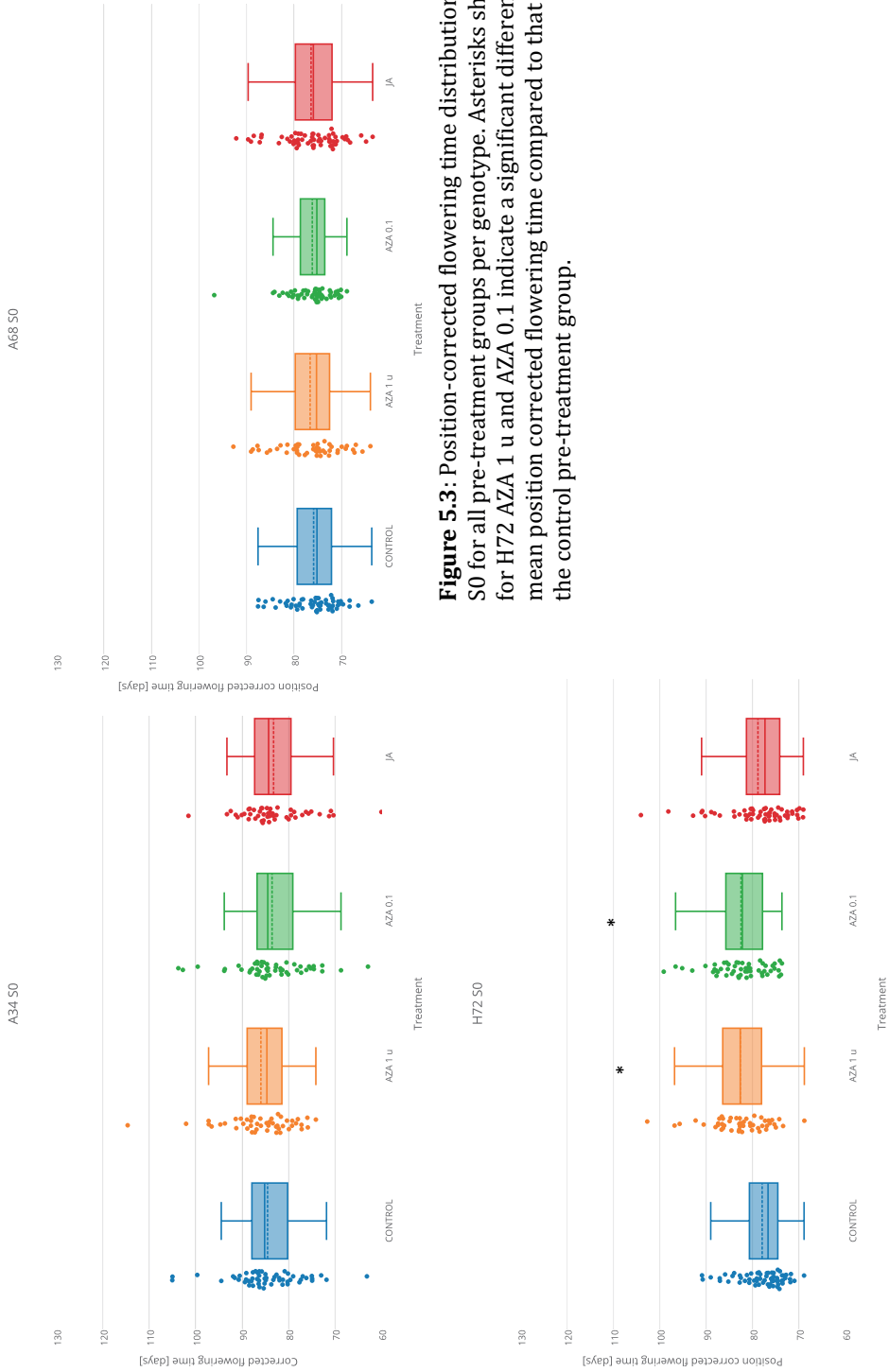


Figure 5.3: Position-corrected flowering time distribution in S0 for all pre-treatment groups per genotype. Asterisks shown for H72 AZA 1 u and AZA 0.1 indicate a significant difference in mean position corrected flowering time compared to that of the control pre-treatment group.

Chapter 5 Selection on early flowering in asexually reproducing Dandelions

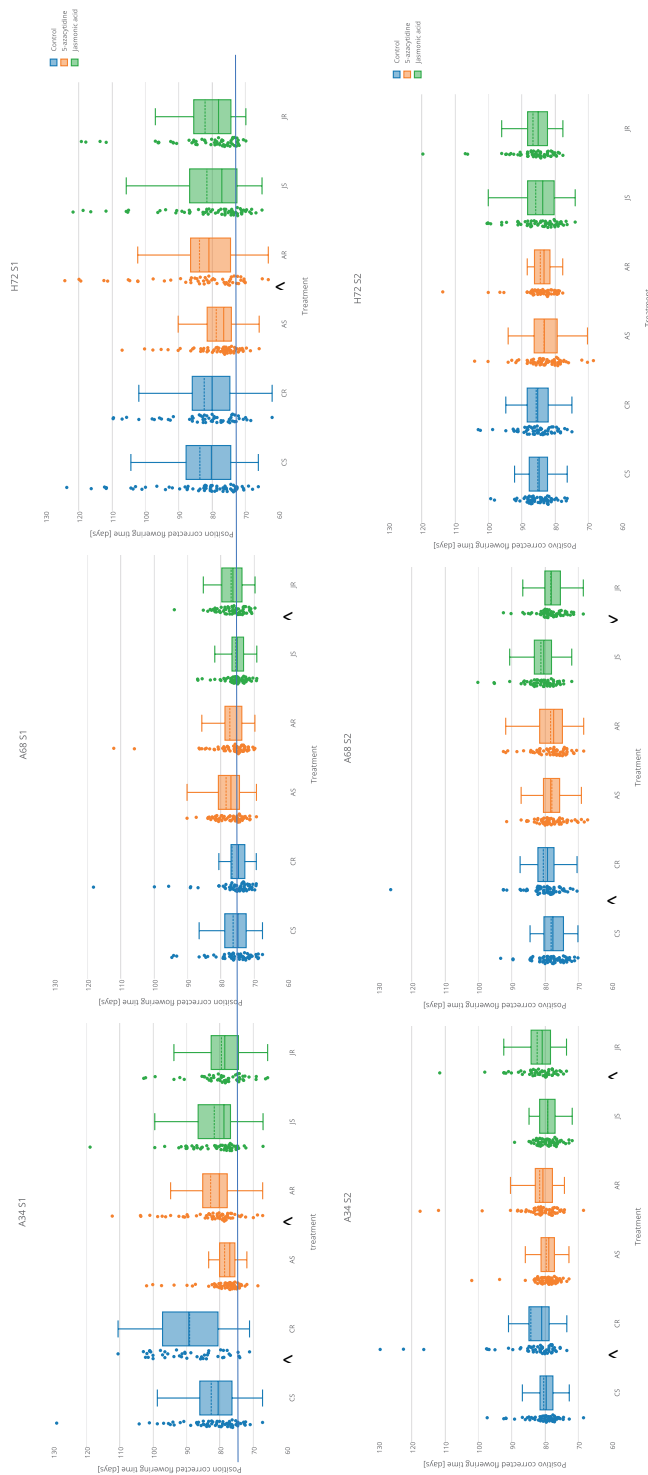


Figure 5.4: Position corrected flowering time distribution in S1 and S2 for all lines per genotype. Symbols ">" and "<" between lines per pre-treatment group indicate a statistically significant difference between the position-corrected flowering time of these lines.

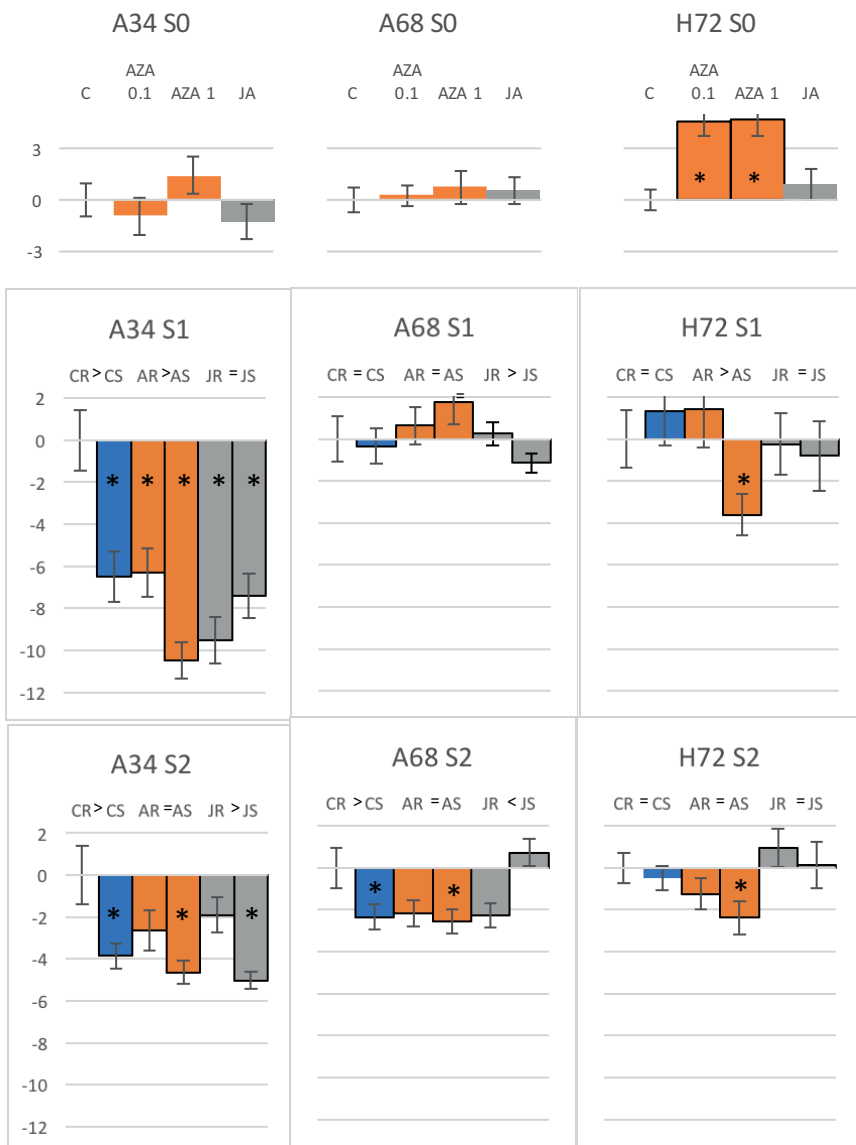


Figure 5.5: Mean difference of position-corrected flowering time compared to the control pre-treatment group (S0) or the random line of the control group (CR in S1 and S2) per genotype for S0, S1 and S2. Error bars indicate standard errors. Statistically significant differences in position-corrected LN-transformed mean flowering time between selection and control line within a group are indicated with the symbol ">" or "<", depending on the direction of the difference. In case no significant difference could be detected between lines

within a group this is indicated with the symbol “=”. Position corrected flowering time of both random and control lines in all groups are compared to that of the random line of the control group. Statistically significant differences compared to the (random line of) the control group are indicated with a star (*) on top of the bar representing the mean difference in position-corrected flowering time per group.

5.4 Discussion.

I have demonstrated that selection on early flowering is possible in isogenic apomictic dandelions, suggesting that epigenetic variation can fuel adaptation. The results suggest that pre-treatment with azacytidine can result in heritable (DNA methylation) variation, leading to significant selection results on early flowering. This effect is, however, genotype specific. Only in H72, pre-treatment with 5-azacytidine resulted in a significant response to selection on early flowering, consistent with broadening the opportunities for selection through stress-induced epigenetic variation. I did not find strong evidence for an enhanced early-flowering response to selection due to variation induced by parental Jasmonic acid treatment in any of the genotypes, suggesting little or no heritable epigenetic variation having an impact on flowering time was generated. Compared to A68 and H72, A34 showed the strongest selection response, suggesting that this young apomictic lineage contains more heritable (epigenetic) variation upon which (artificial) selection can act, without any evidence for an enhanced selection response due to 5-azacytidine or Jasmonic acid pre-treatment.

Pre-treatment with 5-azacytidine did impact the flowering time distribution in H72 S0 as well as the selection response in H72 S1, suggesting that this exposure lead to heritable (epigenetic) variation impacting flowering time. The same concentration of 5-azacytidine treatment did not result in phenotypic differences in the S0 generation of A34 or A68, suggesting that if it indeed effected the (epigenetic) variation in H72, the effective concentration for generating heritable epigenetic variation influencing flowering time differences might be genotype dependent. Alternatively, genetic differences between the genotypes, such as the presence of a transposable element in the vicinity of genes influencing flowering time, might explain these results, as the silencing by DNA methylation of transposable elements can also affect the expression of nearby genes.

Finally, the response to selection is variable between S1 and S2 within pre-treatment groups and there are significant changes in the variance (standard error) of flowering time distribution. These changes show pre-treatment and selection regime specific patterns, indicating both parental treatment effects and selection effects.

5.4.1 Selection on early flowering

Previous selection studies on energy use efficiency in isogenic canola (*Brassica napus*) used artificially generated doubled haploid lines (Hauben et al. 2009), whereas selection on early flowering in *Fragaria vesca* was only achieved after 5-azacytidine treatment (Xu 2016). My results in A34 provide the first indication that selection on natural (not stress-induced) epigenetic variation is possible. Contrary to my expectation, AZA and JA pre-treatments did not result in a significantly increased variance in the base population S0, suggesting that pre-treatments were not effective in generating enhanced levels of heritable epigenetic

variation influencing flowering time. I did find evidence of a negative impact of pre-treatment on flowering time for H72, in which both AZA groups flowered significantly later compared to the control group. Pre-treatment also had an impact on selection results in H72 S1, where the selection line in the AZA group flowered significantly faster compared to the random line of that group.

Stress treatments in plants have been shown to readily induce heritable epigenetic variation that can impact the phenotype (Molinier et al. 2006). Here, I used 5-azacytidine and Jasmonic acid to generate DNA methylation variation. For A34, I did not detect any difference in either the mean or the standard deviation of flowering time between the different pre-treatment groups, which suggests that the treatments were not effective in generating abundant epigenetic variation impacting the flowering time for this genotype. For A68 I found evidence for subtle pre-treatment dependent difference in the standard deviation of flowering time distribution, with JA and 0.1 μM AZA pre-treatment showing minor but opposing effects on the flowering time distribution. These differences were not detectable when compared to the control pre-treatment group. For H72, I observed a treatment induced delay in flowering, which suggests that the 5-azacytidine treatments influenced DNA methylation, resulting in a delayed flowering time, but also DNA methylation variants upon which selection can act. I did not find evidence for an effect of pre-treatment on the variance of the flowering time distribution in H72.

It is possible that my test of azacytidine-induced, selectable epigenetic variation was not successful in two of the three genotypes because of the relatively low concentrations of 5-azacytidine that I used compared to other studies. In a similar selection study on flowering time in *Fragaria vesca*, 5-azacytidine treatments with 3-4 orders of magnitude higher concentrations were used to generate phenotypic variation. Higher concentrations (50mM of 5-azacytidine) had a stronger effect on both DNA methylation as well as phenotypic variation (Xu 2016), although no statistical tests on flowering time distribution differences were employed, more outliers in flowering time distribution were found in plants treated with higher 5-azacytidine concentrations compared to those treated with mild concentrations. The 5-azacytidine concentrations applied in the *Fragaria vesca* study suggest that higher concentrations of 5-azacytidine might be required to generate sufficient DNA methylation variation to allow for fast and stable selection on phenotypic variation.

For A34 I obtained significant selection results regardless of pre-treatment. Given the genomic instability in young apomicts, I cannot exclude that besides epigenetic variation, genetic variation might have been subjected to my successful selection effort. For H72 this is less likely. In H72 I found a strong selection response in S1 and a weak non-significant response in S2, which is consistent with metastable induced epigenetic variation, as also in S0 there was evidence for a transgenerational effect of parental AZA exposure. Future (bisulfite-based) sequencing studies can aid in determining the type of heritable variation upon which selection acted.

To my knowledge, this study is among the first to select on phenotypic variation in asexual plants species, where a large the role of epigenetic variation is expected to play a large role in determining the phenotypic variation selected upon. Previous studies have demonstrated

that epigenetic variation in apomictic dandelions can be generated easily ((Verhoeven, Jansen, et al. 2009) and chapter 4 of this thesis). Although I did not measure DNA methylation in the present experiment, a large body of literature supports a causal link between changes in heritable DNA methylation and the phenotype (Cortijo et al. 2014; Johannes et al. 2009; Trerotola et al. 2015; Schmitz et al. 2011). Changes in DNA methylation can be generated using mutants for methyl-transferases involved in generating or maintaining DNA methylation or MSH1, a gene involved in mitochondrial genome rearrangements. It can also be generated by 5-azacytidine application or environmental stress. In some cases, environmental stress can result in a (transgenerational) adaptive response to the stress applied. Mutants for either DNA methyltransferases or a single gene (such as MSH1) show genome-wide changes in DNA methylation, often with large effects on phenotypic variation. In *A. thaliana*, epigenetic recombinant inbred lines (epiRILs) were created by crossing a methyltransferase mutant (*ddm1*) with a wild-type parent, resulting in offspring that differ in DNA methylation but not in their genomic sequence (Johannes et al. 2009). Substantial variation in quantitative traits such as flowering time and plant height were observed for epiRIL populations that were initiated from such mutant individuals, suggesting that epigenetic variation can control for variation in these important phenotypic characteristics (Johannes et al. 2009). In *Brassica rapa*, a hypomethylated isogenic (doubled haploid) plant population was created by 5-azacytidine treatment. Heritable variation for several quantitative traits including seed yield was generated as well as substantial DNA methylation variation (Amoah et al. 2012). Likewise, 5-azacytidine application to several *A. thaliana* genotypes resulted in strongly reduced growth and fitness of plants, delayed flowering and more phenotypic variation, but the degree of this response varied significantly among genotypes. These differences in response between genotypes were however only weakly related to their genetic relatedness suggesting (some) independence of epigenetic from genetic variation (Bossdorf et al. 2010). For the *Fragaria vesca* study mentioned before, selection on early and late flowering resulted in a significant deviation from the average flowering time of control plants (Xu 2016), suggesting that 5-azacytidine can cause heritable changes in DNA methylation that cause changes in the phenotype upon which selection can act. Adaptive transgenerational stress-responses linked to DNA methylation changes have also been found in non-model plant species, for which the implied relation to DNA methylation was established due to the fact that these responses disappeared after 5-azacytidine in the offspring; suggesting involvement of DNA methylation in transmitting the response between generations (see discussion in chapter 4).

5.4.2 Response to selection

If I disregard treatment-induced epigenetic variation in the JA and AZA pre-treatment groups and only compare the selection results between selection and random lines in the control group between the three genotypes, I observe a significantly faster position-corrected flowering time for the selection line of the control group (CS) compared to the random line of that group (CR) in the young apomict A34 in both generation S1 and S2 (see Fig. 5.5). In A68, I observe significantly faster flowering for the selection line CS compared to the random line CR in S2 but not S1, suggesting that multiple selection rounds are required for achieving a significant selection result in this genotype. In H72 I did not observe any significant selection effects in the control group of either S1 or S2. I can

interpret these different observations as evidence for genotype-specific differences in the baseline level of heritable epigenetic variation controlling flowering time. Alternatively, such differences could be caused by an interplay between genetic and epigenetic variation by, for example, a transposable element nearby a gene involved in the onset of flowering, for which the methylation status change could induce changes in flowering time.

I hypothesize that A34 has a higher baseline of epigenetic variation compared to both A68 and H72, whereas based on these limited observations I would expect A68 to have higher level of DNA methylation variation compared to H72. Higher levels of DNA methylation variation were previously shown for the recent apomict A34 compared to the established apomictic genotype A68 in an MS-AFLP study in which 4/122 markers were polymorphic for A34 instead of only 1/122 for A68 (Verhoeven, van Dijk, et al. 2009). New apomictic dandelions lineages arise through hybridization resulting from a cross between polyploid apomictic fathers and sexual diploid mothers. Hybridization has been shown to result in epigenetic changes (Greaves et al. 2012; Z. J. Chen 2013). I hypothesize that recent apomicts harbor a greater epigenetic diversity as well as a less stable epigenome. Increased levels of epigenetic variation in A34 could therefore contribute to the stronger selection response observed for this genotype. Alternatively, it could mean that the hybridization event leading to a new apomictic lineage causes changes in DNA methylation, which could trigger reactivation of previously silenced transposable elements (Verhoeven, van Dijk, et al. 2009; Salmon et al. 2005), resulting also in enhanced levels of within-lineage genetic variation due to transposition. Genotypic differences in the transgenerational stability of several phenotypic traits were found previously in the apomictic plant *Hieracium pilosella*. In a study comparing 12 newly generated apomictic *H. pilosella* lines, derived by crossing existing apomictic lineages with sexual lineages (Sailer et al. 2016) the stability of several phenotypic traits was assessed. The transgenerational stability of these traits was evaluated by comparing means across two subsequent generations. Most phenotypic traits (18/20) were stably inherited over two apomictic generations, grown simultaneously in a common control environment. However, one apomict showed different measures for most phenotypic traits in subsequent generations. Interestingly, among the transgenerationally labile phenotypic traits found for *H. pilosella* was age at flowering (flowering time), which did not show phenotypic stability across generations. This observation is in contrast with the transgenerationally stable mean flowering time I found for the three dandelion genotypes in the present study. Further studies using epiGBS or WGBS aimed at comparing DNA methylation variant in the selection lines of the recent apomict A34 and the more established apomict A68 and H72 should be done to assess whether differences in DNA methylation variation support the different outcome of the selection in these lines.

5.4.3 Variability in response by pre-treatment and selection regime

Our results indicate that the selection response can differ between S1 and S2, with some selection responses being genotype and / or pre-treatment specific. To systematically evaluate potential drivers of these differences I evaluate several flowering-time related statistics for the three genotypes separately on a per generation basis. Mean flowering time in random lines is not expected to change between generations because of selection. Nevertheless, minor and (often) non-significant changes in the flowering time of these random lines between generations do occur, for instance where random lines of treatment

groups flower earlier than the random line of the control group. Such events could indicate pretreatment, not selection, effects on flowering time.

A34: I found no evidence for any significant parental effects on mean or variance of flowering time due to pre-treatment in A34 for the S1 or S2 generation. The significant two-way ANOVA results for A34 S1 indicating a higher selection response in the control pre-treatment group could be explained by the significant delay in flowering of the CR group (see results). Selection was effective in both S1 and S2 for 2 out of 3 groups. I found evidence for a continued effect of selection on the flowering time distribution, as the variance (standard error) of all selection lines in S2 was significantly lower compared to that of the selection lines from S1 (Supplementary Fig. 5.2) as well as the random lines from S2 (Supplementary Table 5.2). Overall, the response to selection observed in S2 seemed lower than that of S1 in all pre-treatment groups except for JA. This pattern coincides with the slightly lower variation observed in S2 compared to S1 (Supplementary Fig. 5.2).

A68: I observed a reduction of variance in flowering time in the JA group from S0 to S1, irrespective of selection regime (both JS and JR have a significantly lower variance in S1 compared to S0) whereas for both random and selection line of the AZA and control group this variance was actually higher in S1 compared to S0 (see supplementary Fig. 5.2). The only positive result of selection in S1 was observed in the JA group. It is not clear what this implies. If JA treatment would have resulted in enhanced levels of epigenetic variation upon which successful selection took place in S1 I would have expected to observe treatment induced variation in S0 like I observed for H72. It is also difficult to explain the negative response to selection in the JA group for S2, although JS did not flower significantly later than the CR group it had the slowest flowering time on average (Fig. 5.5). My observations in A68 for the JA group suggest drift in mean flowering time in the random group, rather than a significant positive response to selection. Random changes in flowering time or negative responses to selection on flowering time have been reported previously (Burgess et al. 2007), suggesting that this might have been the case here.

H72: I suspect that H72 has lower baseline levels of epigenetic variation influencing flowering time compared to A34 and A68, as I did not observe positive selection results in the control pre-treatment group of H72 in either S1 or S2. The fact that the only significant selection result was obtained in the AZA group in S1 (H72 S1 AS vs H72 S1 AR) indicates that treatment induced heritable epigenetic variation can lead to successful selection. In contrast to the other genotypes, I found a significant difference in the standard error of most flowering time distribution between S0, S1 and S2, indicating that H72 is more sensitive to environmental conditions during growth compared to A34 and A68. The significant difference obtained for both two-way ANOVA and 1-way ANOVA in S2, showing that both lines of the AZA group flowered significantly earlier compared to both the control and JA group indicate that flowering time of the AZA group in S2 could still be influenced by the grand-parental pre-treatment, but this did not lead to a significant difference within the AZA group in the same generation (S2 AS vs S2 AR).

5.4.4 Summary for all genotypes

I observed no significant increase in the response to selection for A34 and A68 because of pre-treatment with 5-azacytidine or Jasmonic acid. In H72 I found a significant response to selection in S1, but I failed to maintain a clear difference in the subsequent selection generation S2, which I think might be attributed to reversal of azacytidine-induced DNA methylation changes two generations after treatment. I assume that the pre-treatments with 0.1 μM of 5-azacytidine and the JA-treatments did not alter heritable DNA methylation variation sufficiently to alter the selection response in A34 and A68. This is consistent with the absence of any significant treatment effect on the mean flowering time in S0 for both A34 and A68 (see Supplementary table 4). In H72 I have evidence for a 5-azacytidine pre-treatment effect on the selection response, but heterogeneous environmental conditions between the different generations and a relatively weak effect make it hard to evaluate whether stress-induced enhanced levels of epigenetic variation can lead to more effective selection in apomictic dandelions generally.

If I would have selected for both slow and fast flowering plants, I might have been able to observe a significant effect on flowering time distribution more easily by contrasting selection on early flowering with selection on late flowering. The contrasting responses to selection found between S1 and S2 of the JA group in A68 might be attributed to stochastic variation in the selection line, leading to a reverse response to selection, which was observed in a previous selection experiment on flowering time (Burgess et al. 2007). It is unclear if multiple successive rounds of selection on the offspring of individuals treated with a higher concentration of 5-azacytidine would increase the selection response observed.

Heritable epigenetic variation has a demonstrated capacity to determine phenotypic variation in many quantitative traits in a major way (Johannes et al. 2009). DNA methylation variation can be generated artificially, either through mutants or by applying chemicals. However, DNA methylation can also change naturally as the result of environmental perturbations or stress. It remains unclear to what extent stress-exposure in nature triggers more epigenetic variation in plants. Future epigenetic selection experiments should focus on selecting both adaptive phenotypic traits in response to an environmental stress as well as more neutral quantitative traits to assess whether stresses induce specific changes in DNA methylation allowing for selection on adaptive traits or if the induced changes are non-targeted, in which case the selection experiment on the neutral quantitative traits should work equally well. Of course, such selection experiments should be accompanied by DNA methylation profiling using epiGBS or WGBS and DNA demethylation treatments using 5-azacytidine in effective concentrations to ascertain whether DNA methylation variation or genetic variation can explain the observed phenotypic changes.

Besides (artificial) induction of epigenetic variation in plants by means of mutations or environmental stress, several studies have demonstrated successful artificial selection on epigenetic variation, resulting in changed phenotypes associated with DNA methylation variation (Xu 2016; Verkest et al. 2015). My study provides the first indication that natural heritable (epigenetic) variation (not generated due to haploidization or treatment)

provides for variation upon which (artificial) selection can act. So, although stress or chemical treatments of plants can increase the amount of epigenetic variation, such treatments are not a requirement, as baseline levels of epigenetic variation in plants can provide sufficient variation to allow for selection on deviant phenotypes as I have demonstrated in the present study.

Triploid apomictic dandelions can only accumulate genetic variation through somatic mutations or transposable element activity, as reproduction within a lineage is entirely asexual (Tas & Van Dijk 1999). This makes heritable epigenetic variation an intriguing potential resource for rapid adaptation to changing environmental conditions for apomictic lineages occurring in isolation of sexual lines, as meiotic recombination and segregation are not available (Verhoeven, van Dijk, et al. 2009). Although many studies have demonstrated transgenerational epigenetic inheritance in sexual plant species, this potential has received limited attention in asexual plant species (Verhoeven & Preite 2014). Successful colonization of asexual plant species is thought to occur due to high phenotypic plasticity, the so called “general purpose” genotype (Baker et al. 1965). This is also found in a north-south transect study with apomictic dandelions across Europe that expand northward (Preite 2016). Plants from the southern part of the range showed larger differences in biomass between test environments, which was mainly driven by a high performance in the optimal (southern) environment. Plants from central and northern European latitudes showed little variation in performance in different environments, consistent with a general purpose genotype (Oplaat & Verhoeven 2015). Besides phenotypic plasticity, local adaptation could also play a role in the successful invasion of novel environments by *T. officinale* (Molina-Montenegro et al. 2011). It is therefore important to consider and study both the potential role of transgenerationally stable epigenetic and genetic variation in apomictic dandelions in adaptation. Future studies focusing on tracking (selection on) epigenetic changes and evaluating fitness of apomictic dandelions from the same line in alien environments could help to further elucidate the mechanisms that cause local adaptation in asexual plant species.

Supporting information

Supplementary table 1) Descriptive statistics for position corrected flowering time S1 and S2 for all genotypes

genotype	generation	group	N	Mean corrected flowering time [days]	Std. Deviation	Std. Error		
A34	S1	AR	60	82.79	8.93	1.15		
		AS	63	78.60	6.89	0.87		
		CR	46	89.10	9.69	1.43		
		CS	64	82.61	9.73	1.22		
		JR	52	79.56	7.92	1.10		
		JS	59	81.69	8.12	1.06		
		AR	62	81.63	7.65	0.97		
	S2	AS	63	79.63	4.31	0.54		
		CR	57	84.27	10.64	1.41		
		CS	62	80.41	4.85	0.62		
		JR	56	82.37	6.39	0.85		
		JS	61	79.25	3.28	0.42		
		A68	S1	AR	63	77.33	7.03	0.89
				AS	59	78.45	8.22	1.07
CR	58			76.69	8.07	1.06		
CS	53			76.34	6.14	0.84		
JR	63			76.94	4.31	0.54		
JS	64			75.54	3.63	0.45		
AR	63			78.35	5.03	0.63		
S2	AS		63	77.96	4.55	0.57		
	CR		62	80.52	7.50	0.95		
	CS		62	78.17	4.62	0.59		
	JR		62	78.26	4.39	0.56		
	JS		60	81.25	5.10	0.66		
	H72		S1	AR	58	83.77	13.99	1.84
				AS	63	78.73	7.76	0.98
CR		59		82.35	10.56	1.37		
CS		64		83.66	12.91	1.61		
JR		61		82.10	11.38	1.46		
JS		63		81.55	13.17	1.66		
AR		60		84.44	5.64	0.73		
S2		AS	60	83.30	6.14	0.79		
		CR	62	85.68	5.64	0.72		
		CS	62	85.20	4.50	0.57		
		JR	61	86.65	7.13	0.91		
		JS	62	85.83	8.67	1.10		

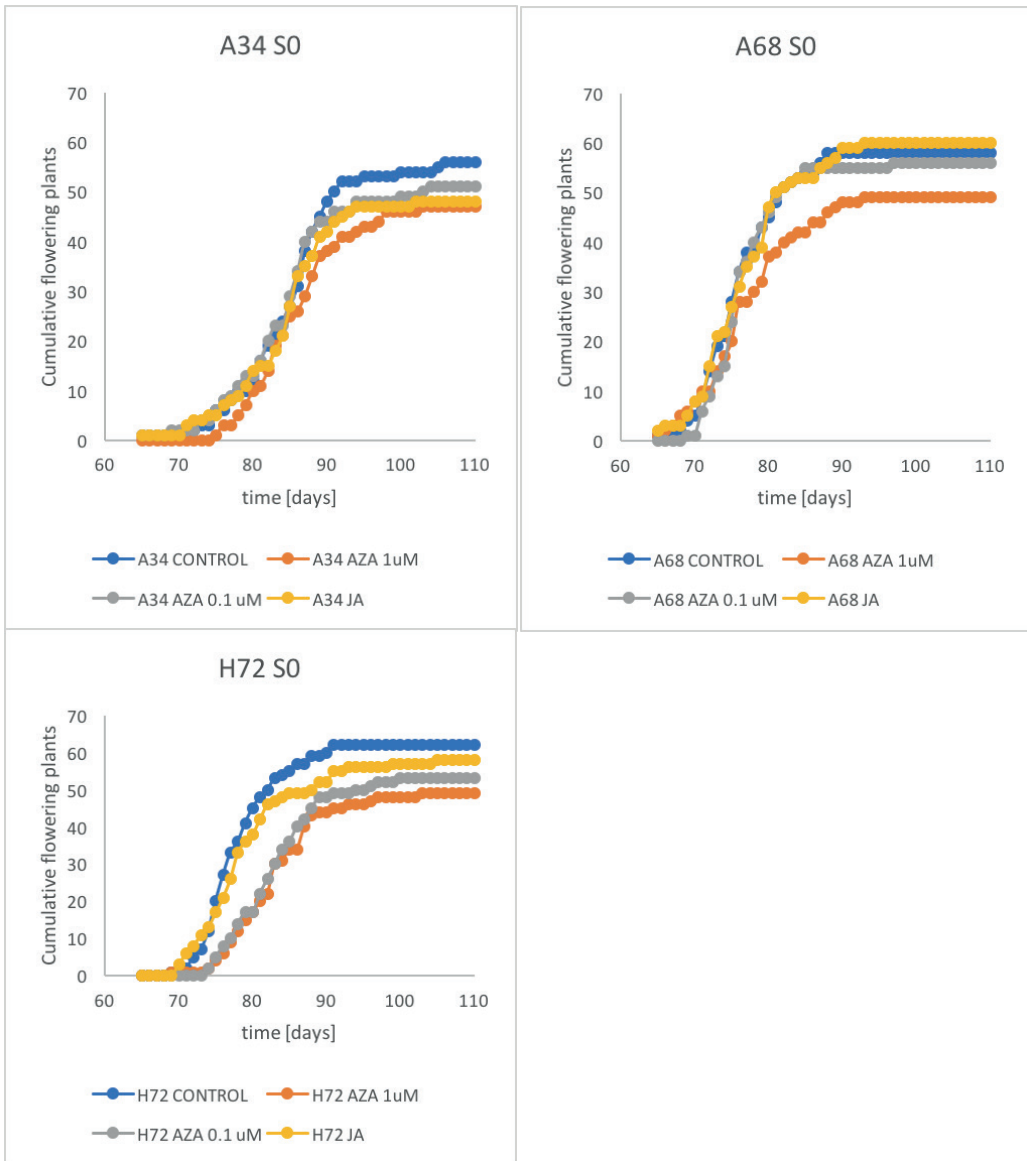
Supplementary Table 5.2) One-way ANOVA results for generation S1 and S2 comparing both the selection and random line of the AZA and JA pre-treatment group (AS, AR, JS and JR) with the random line of the control pre-treatment group (CR).

<i>Genotype</i>	<i>Generati on</i>	<i>compariso n</i>	<i>p-value</i>	<i>Levene's test</i>	<i>Welch p-value</i>	<i>Outcome</i>
A34	S1	JS vs CR	<0.001	0.055		JS<CR
A34	S1	AS vs CR		0.002	<0.001	AS<CR
A34	S2	JS vs CR		0.001	<0.001	JS<CR
A34	S2	AS vs CR		0.002	0.002	AS<CR
A34	S1	JR vs CR	<0.001	0.055		JR<CR
A34	S1	AR vs CR	0.001	0.186		AR<CR
A34	S2	JR vs CR	0.281	0.139		No diff
A34	S2	AR vs CR	0.107	0.181		No diff
A68	S1	JS vs CR	0.374	0.015	0.389	No diff
A68	S1	AS vs CR	0.169	0.169		No diff
A68	S2	JS vs CR	0.412	0.507		No diff
A68	S2	AS vs CR		0.516	0.019	AS<CR
A68	S1	JR vs CR	0.644	0.154		No diff
A68	S1	AR vs CR	0.552	0.739		No diff
A68	S2	JR vs CR	0.037	0.299		JR<CR
A68	S2	AR vs CR	0.050	0.877		No diff
H72	S1	JS vs CR	0.585	0.197	0.583	No diff
H72	S1	AS vs CR	0.036	0.058		AS<CR
H72	S2	JS vs CR	0.972	0.98		No diff
H72	S2	AS vs CR	0.023	0.603		AS<CR
H72	S1	JR vs CR	0.868	0.866		No diff
H72	S1	AR vs CR	0.634	0.150		No diff
H72	S2	JR vs CR	0.416	0.684		No diff
H72	S2	AR vs CR	0.209	0.290		No diff

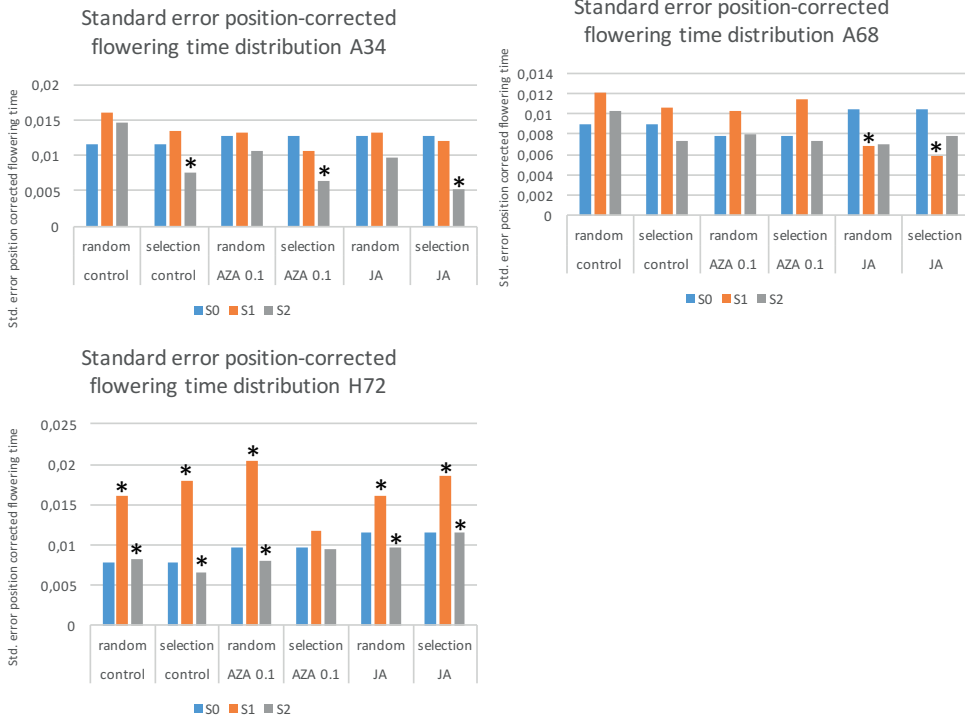
Supplementary table 5.3) ANOVA results for comparing per generation per group average position-corrected flowering time between generations S0 vs S1 and S1 vs S2.

<i>Genotype</i>	<i>Generation</i>	<i>group</i>	<i>p-value</i>	<i>Levene's test</i>	<i>Welch p-value</i>	<i>Outcome</i>
A34	S0 vs S1	CR	0.011	0.012	0.013	S0<S1
A34	S0 vs S1	CS	0.164	0.174		No diff
A34	S0 vs S1	AR	0.587	0.456		No diff
A34	S0 vs S1	AS	<0.001	0.295		S0>S1
A34	S0 vs S1	JR	0.014	0.698		S0>S1
A34	S0 vs S1	JS	0.275	0.554		No diff
A34	S1 vs S2	CR	0.011	0.137		S1>S2
A34	S1 vs S2	CS	0.149	<0.001		No diff
A34	S1 vs S2	AR	0.461	0.055		No diff
A34	S1 vs S2	AS	0.223	0.031		No diff
A34	S1 vs S2	JR	0.026	0.133		S1<S2
A34	S1 vs S2	JS	0.042	<0.001	0.045	S1>S2
A68	S0 vs S1	CR	0.578	0.754		No diff
A68	S0 vs S1	CS	0.694	0.789		No diff
A68	S0 vs S1	AR	0.304	0.299		No diff
A68	S0 vs S1	AS	0.054	0.323		No diff
A68	S0 vs S1	JR	0.491	0.010	0.495	No diff
A68	S0 vs S1	JS	0.422	<0.001	0.430	No diff
A68	S1 vs S2	CR	0.002	0.581		S1<S2
A68	S1 vs S2	CS	0.049	0.152		S1<S2
A68	S1 vs S2	AR	0.264	0.638		No diff
A68	S1 vs S2	AS	0.782	0.345		No diff
A68	S1 vs S2	JR	0.086	0.733		No diff
A68	S1 vs S2	JS	0.000	0.094		S1<S2
H72	S0 vs S1	CR	0.006	<0.001	0.007	S0<S1
H72	S0 vs S1	CS	0.002	<0.001	0.002	S0<S1
H72	S0 vs S1	AR	0.818	<0.001	0.813	No diff
H72	S0 vs S1	AS	0.002	0.175		S0>S1
H72	S0 vs S1	JR	0.076	0.020	0.073	No diff
H72	S0 vs S1	JS	0.245	0.001	0.235	No diff
H72	S1 vs S2	CR	0.012	<0.001	0.014	S1<S2
H72	S1 vs S2	CS	0.158	<0.001	0.155	S1<S2
H72	S1 vs S2	AR	0.396	<0.001	0.403	No diff
H72	S1 vs S2	AS	0.000	0.110		S1<S2
H72	S1 vs S2	JR	0.002	0.001	0.002	S1<S2
H72	S1 vs S2	JS	0.009	0.001	0.009	S1<S2

Chapter 5 Selection on early flowering in asexually reproducing Dandelions



Supplementary Fig. 5.1. Cumulative flowering time S0. In all three genotypes the fraction of non-flowering plants is highest for the offspring of parent plants treated with 5-azacytidine.



Supplementary Figure 5.2). Standard errors for the flowering time distribution of both random and selection line per pre-treatment group from generation S0 – S1 are shown. For S0, values for the random and selection line are identical as no selection was yet made. Asterisks (*) indicate a significant value ($p < 0.05$ see Supplementary Table 5.3) for the Levene's test for equality of variance, derived from intergenerational ANOVA's on a per group per line basis (e.g. A34 S1 CS vs A34 S2 CS).

Supplementary table 4) 1-way ANOVA results for within between line comparisons of position-corrected flowering time in S0.

<i>Genotype</i>	<i>Generation</i>	<i>comparison</i>	<i>p-value</i>	<i>Levene's test</i>	<i>Outcome</i>
A34	S0	All	0.289	0.986	No diff.
A34	S0	AZA 0.1 vs C	0.491	0.764	No diff.
A34	S0	AZA 1.0 vs C	0.311	0.967	No diff.
A34	S0	JA vs C	0.373	0.975	No diff.
A68	S0	All	0.941	0.041	Diff. var.
A68	S0	AZA 0.1 vs JA	0.882	0.023	Diff. var.
A68	S0	AZA 0.1 vs AZA1	0.678	0.011	Diff. var.
A68	S0	AZA 0.1 vs C	0.723	0.170	No diff
A68	S0	AZA 1.0 vs C	0.567	0.157	No diff
A68	S0	JA vs C	0.652	0.303	No diff
H72	S0	All	<0.001	0.289	Diff mean
H72	S0	AZA 0.1 vs C	<0.001	0.403	Diff mean
H72	S0	AZA 1.0 vs C	<0.001	0.357	Diff mean
H72	S0	JA vs C	0.484	0.060	No diff.

Per genotype, an initial ANOVA is conducted on all lines simultaneously. Treatment lines are compared to the control line. For A68, significant p-values for Levene's test were also acquired for the comparison of AZA 0.1 vs JA and AZA 0.1 vs AZA 1.0, but not for any of the treatment vs control line comparison.

General Discussion

6

6.1 Aim and motivation for the research done in this thesis

The aim of my PhD research was to investigate the stability of (stress induced) DNA methylation variation in apomictic dandelions and the potential of phenotypic variation underpinned by DNA methylation variation to be subjected to selection. I therefore studied the transgenerational stability of both stress induced and natural DNA methylation variation in different genotypes of apomictic dandelions in a six-generation experiment, comparing DNA methylation patterns between generations and tracking changes in them. Apomictic dandelions reproduce asexually, which means that genetic changes between generations can only arise sporadically due to random mutations (including copy number variants and novel transposable element inserts). In contrast, heritable DNA methylation variation occur much more frequently (van der Graaf et al. 2015) and can also underpin phenotypic variation (Cortijo et al. 2014). A prerequisite for epigenetic variation to contribute to adaptation is that epigenetic variants that affect the phenotype are heritable. To test whether an epigenetics-based selection response is possible, at least over the time course of a few generations, I selected early flowering for two subsequent generations in three genotypes of apomictic dandelions. This selection effort included lines that received a stress pre-treatment with either Jasmonic Acid or 5-azacytidine, in order to determine if stress-induced DNA methylation variation would increase the capacity to respond to selection. The selection experiment on flowering time (chapter 5) resulted in a shift in flowering time for all lines in a young apomict, suggesting that natural and heritable epigenetic variation underpins quantitative traits such as flowering time. This pattern was not found for established apomicts A68 and H72, and a selection response was also not facilitated by Jasmonic Acid stress pretreatment. In one apomict, I obtained a positive selection result on early flowering for the first-generation offspring of the 5-azacytidine pre-treatment line, indicating that artificially induced DNA methylation variation can lead to changes in phenotypic traits upon which selection can act successfully.

Much has been learned on the dynamics and role of DNA methylation using the model plant species *Arabidopsis thaliana*, with whole genome bisulfite sequencing being key to uncover the complex and subtle patterns that govern (stress induced) DNA methylation variation and its transgenerational stability. Currently, it is not clear what the relevance of the DNA methylation variation uncovered using methylation sensitive AFLPs (MS-AFLP) in non-model species is, due to its quantitative and low resolution, the anonymous nature and the exclusion of non-CG methylation. Studying DNA methylation in a detailed and comprehensive manner in non-model organisms using bisulfite sequencing was not possible before this thesis. An important goal of my research was therefore to develop a bisulfite-sequencing based technique for non-model organisms. We have thus developed epiGBS, a reduced representation bisulfite sequencing technique based on the highly flexible and popular genotyping-by-sequencing technique (Elshire et al. 2011). It allows for single-nucleotide resolution bisulfite sequencing in non-model organisms, with an advanced bioinformatics pipeline that enables simultaneous interrogation of both DNA methylation and genetic variation.

Finally, a side-result of my research was the accidental discovery of highly irregular patterns in sequencing-mismatches in RNA sequencing data, which I confirmed to arise due to random hexamer mispriming. Such technical oddities are important to understand as

they have previously led to high profile papers claiming wide-spread RNA editing (M. Li et al. 2011), which is better explained by the random hexamer priming I discovered (van Gurp et al. 2013).

6.2 How do our research results contribute to answering open research questions in ecological epigenetics?

The field of ecological epigenetics, which primarily focuses on studying epigenetic inheritance in natural systems, is expanding rapidly. Although a wealth of knowledge has emerged from studying epigenetics in model species like *Arabidopsis thaliana*, understanding the importance of epigenetic variation for ecology and evolution requires extending this research to non-model organisms (Verhoeven et al. 2016). At this point, many basic questions remain open, in part due to technological limitations that limited the way epigenetic variation was measured in non-model systems. In this section, I review my results in the light of some of the open questions to which this research provides some answers, discussing them in the context of results of various studies in both model and non-model systems that addressed these issues. These questions are (i) How does epigenetic variation shape the phenotype? (ii) How much epigenetic variation is environmentally induced? And finally, (iii) Is environment-induced epigenetic variation transgenerationally stable?

6.2.1 How does epigenetic variation shape the phenotype?

Phenotypic variation is the product of genetic, epigenetic and environmental variation, with potential complex interaction terms between these components. When looking for an epigenetic contribution to adaptation that cannot already be explained by underlying genetic variation, only heritable epigenetic variation that is at least to some extent independent of genetic variation is relevant. In plants, several well defined examples of phenotypic variants that rely on epigenetic variation are known, such as the altered floral symmetry in *Linaria vulgaris* (Cubas et al. 1999), the methylation mutant colorless non-ripening in tomato (Manning et al. 2006) and the clark kent (clk) epiallele in *Arabidopsis*, resulting in increased numbers of stamens and carpels (Jacobsen & Meyerowitz 1997). Artificial induction of genome-wide changes in DNA methylation in *Arabidopsis* by crossing methylation deficient mutant lines with wild-type lines results in wild-type plants with aberrant DNA methylation patterns (Johannes et al. 2009; Cortijo et al. 2014). Importantly, these heritable DNA methylation changes determined variation in many important quantitative traits, such as flowering time and root length (Johannes et al. 2009; Cortijo et al. 2014). Environmental stress can induce changes in DNA methylation, which in some cases causes adaptive transgenerational responses in offspring of stressed parents (Sultan et al. 2009; Sultan 2011; Agrawal & al 1999). Such effects have been demonstrated also in non-model species, for which the dependence of such effects on DNA methylation is often established by erasing DNA methylation using the demethylating agents 5-azacytine or zebularine which removes these effects (Bossdorf et al. 2010; Xu 2016; Herman 2016). In our selection experiment (chapter 5), I obtained faster flowering plants, which most likely rely on epigenetic variation. I did not measure DNA methylation in the selection lines experiment, but multiple observations suggest that the phenotypic variation I selected on was caused by epigenetic and not genetic variation. First, I observed a significant positive selection response in the established apomict H72 in the first but not second generation

offspring of parental plants exposed to 5-azacytidine treatment. 5-azacytidine affects DNA methylation and not genetic changes (Christman 2002), proving that heritable epigenetic variation generated by 5-azacytidine can alter flowering time. Whereas stress-induced epigenetic changes are reversible, this is not the case for genetic changes. The patterns observed for the 5-azacytidine pre-treated H72 lines are thus consistent with epigenetic and not genetic changes. Second, genetic changes occur at a much lower frequency (1×10^{-9} in *Arabidopsis* (Becker et al. 2011)) compared to DNA methylation changes (4×10^{-4} (van der Graaf et al. 2015; Becker et al. 2011)), a difference of five orders of magnitude. If the changes in flowering time would be driven by rare genetic changes, these would most likely be limited to one or a few individuals, which are expected to reliably transmit the mutation to the four offspring individuals, which should also show a reduced flowering time. In the flowering time selection lines, I took seeds of the 25% fastest flowering plants which contribute four offspring individuals each. In none of the selection lines I found indications that showed a difference driven by four offspring individuals from one mother, which I would expect if genetic changes would underlie the change in flowering time, suggesting that epigenetic and not genetic changes were the driving force behind the shift in population mean observed for flowering time. Third, in a similar selection experiment on early flowering in highly inbred *Fragaria vesca* lines subjected to high concentrations of 5-azacytidine where DNA methylation patterns were measured using MS-AFLPs, researchers found distinct DNA methylation patterns for selection lines using principal component analysis (PCA) (Xu 2016), suggesting that heritable DNA methylation variation is an important epigenetic component determining flowering time.

Epigenetic variation thus has the capacity to influence the phenotype in a major way. Appreciating the importance of epigenetic variation compared to genetic and environmental variation is difficult, as controlling for each of the factors independently is needed. Apomictic dandelions provide an elegant model system that isolates epigenetic from genetic variation. (environment induced) epigenetic variation can result in changes in the phenotype, but the extent to which this epigenetic variation is transmitted transgenerationally determines the relevance of such variants for adaptation by selection. Epigenetic variants that exert an influence on the phenotype but are not transmitted transgenerationally contribute to phenotypic plasticity whereas variants that are transgenerationally transmitted contribute to heritable phenotypic variation upon which selection can act. Here, I demonstrate successful selection on early flowering in both stressed and non-stressed lineages of apomictic dandelions, which suggests that natural epigenetic variation can contribute to phenotypic changes in populations under selection. The relative importance of both heritable and non-heritable epigenetic variation in generating phenotypic changes that can be selected upon or contributing to phenotypic plasticity within a generation is not known for many species and should be investigated more thoroughly.

6.2.2 How much epigenetic variation is environmentally induced and how stable is it?

An important and distinguishing feature of epigenetic compared to genetic variation is the ease with which epigenetic but not genetic variation can be induced by the environment. Environmental induction of heritable epigenetic changes allows for rapid and reversible phenotypic changes in response to stress, which can support transgenerational phenotypic

plasticity. Many of these transgenerational effects which trigger changes in either gene expression, stress sensitivity or phenotypes are described, some of which are adaptive (Sultan 2011). Stress-induced changes can be divided in detection-based responses (Shea et al. 2011) that lead to targeted epigenetic changes in genes involved in a specific stress response, or random changes that increase the overall phenotypic variation in many traits, as a bet-hedging strategy (Simons 2011; O’Dea et al. 2016). Several studies on ecological epigenetics describe parental effects of stress exposure that are dependent on DNA methylation, but most of these studies fail to characterize the type and extent of the changes in epigenetic variation such as DNA methylation (Sultan 2011; Herman & Sultan 2016). Sophisticated studies in *Arabidopsis thaliana* have investigated the degree to which and where DNA methylation changes occur due to repeated salt-stress exposure (C. Jiang et al. 2014; Wibowo et al. 2016). These studies suggest that only 0.17% (6,866 out of 4.1 million) of all methylated positions are differentially methylated due to salt-stress (Wibowo et al. 2016). Stress-associated DNA methylation changes often occur in cytosines on contiguous stretches simultaneously, leading to differentially methylated regions (DMRs) which occur in different regions than and do not overlap with sparsely distributed differentially methylated positions (DMPs). Stress-induced DMRs, and not DMPs, were involved in the downregulation of genes implicated in stress tolerance (Wibowo et al. 2016). So, even though stress-induced DMPs can be detected, their functional significance is unknown (Hagmann et al. 2015). This makes it difficult to assess the relevance of ecological epigenetics studies employing MS-AFLP, focusing on few cytosines in CG context, that show heritable DNA methylation changes (in CG context) due to stress. My multivariate analyses results on JA-induced DNA methylation changes are in line with Wibowo *et al* 2016, in the sense that at least two rounds of consecutive stress application were needed to detect a significant effect, that subsequently disappeared over unexposed generations. I find some evidence for environment induced DNA methylation variation using epiGBS, but most patterns are subtle and limited to certain genotypes or to cytosines in CHG and not CG context. Detailed characterization of the type of epigenetic changes that result after multiple types of (natural) stress exposure as well as determining their transgenerational stability is needed to understand the role and importance of environment induced DNA methylation changes in nature.

The transgenerational stability of DNA methylation has hitherto only been studied in detail for a handful of plant species (using bisulfite sequencing) (Q. Li et al. 2014; Ou et al. 2012), with most studies focusing on the model plants *Arabidopsis thaliana* (Becker et al. 2011; van der Graaf et al. 2015). DNA methylation is very variable between plant species, with genome size and repeat content showing a positive correlation with DNA methylation (Niederhuth, Bewick, Ji, Alabady, Kim, Li, et al. 2016). The mode of reproduction of asexual plants potentially allows for less resetting of epigenetic variation between generations, depending on the type of asexual reproduction, as some steps are skipped in meiosis (Verhoeven & Preite 2014). The relevance of such variation might also be higher for asexual than sexual plants (Verhoeven & Preite 2014). My thesis contains the first high-resolution bisulfite sequencing based study that examines the transgenerational stability of DNA methylation in an asexual plant species, opening the frontier for ecological epigeneticist to start detailed exploration of such variation in a wide array of ecologically interesting non-model species. We employed epiGBS (chapter 3) to study stress induced DNA methylation

changes and the transgenerational stability of both natural and stress-induced DNA methylation variation (chapter 4). We found subtle and genotype-specific effects of parental stress treatment on DNA methylation for cytosines in CHG context. I demonstrated high stability of DNA methylation in CG sequence context in apomictic dandelions as evidenced by high intergenerational correlations of DNA methylation, which is consistent with previous reports (Becker et al. 2011; van der Graaf et al. 2015). Transgenerational correlations for cytosines in CHG context are lower, suggesting that these are less stably inherited compared to those in CG context. This pattern of lower transgenerational correlation, however, may (in part) also be caused by technical factors in the design of the epiGBS methods, which can be easily corrected in future applications (see section 3 below). I found evidence for the existence of *de novo* DNA methylation mutations that are inherited transgenerationally. The evidence for this is provided by the negative trend over intergenerational distance for correlations of DNA methylation between individuals (that is: gradually increasing divergence over generational time), which cannot be explained by technical artefacts. This trend is apparent for cytosines in both CG and CHG context; whereas previous studies have emphasized heritability of cytosines in CG and not CHG context (Becker et al. 2011; van der Graaf et al. 2015). This apparent discrepancy can be explained by the different approach (not based on DMPs) I took compared to previous studies. The stability of DNA methylation is typically assessed by counting the number of differentially methylated positions (DMPs) that arise per generation. Due to differences in the distribution of DNA methylation between cytosines with a different sequence context, detection is skewed in favor of those in CG context, which has led to a focus on cytosines in CG context in studies looking at intergenerational stability of DNA methylation (Becker et al. 2011). I took a different approach; by looking at the correlation of DNA methylation ratios of cytosines in CG and CHG context, I avoided this bias which helped us to uncover the negative trend for cytosines in CHG and CG context described earlier. Our study thus suggest that a more careful consideration of statistical methods is needed to appreciate the heritability of DNA methylation patterns.

6.3 Putting epiGBS into context compared to MS-AFLP and existing bisulfite sequencing methods

A persistent problem hampering many ecological epigenetics studies has been the difficulty in discerning epigenetic from genetic variation. MS-AFLP and AFLP based studies can provide insight into the level of genetic and epigenetic diversity; but (substantial) genetic variation goes undetected, given the low coverage of anonymous AFLP loci on plants genomes, which often have a large genome size. Similarly, MS-AFLP studies can miss many patterns of DNA methylation, especially if they do not occur in CG-context (Fulneček & Kovarik 2013; Trucchi et al. 2016; van Gurp et al. 2016). Another issue plaguing MS-AFLP studies is ambiguous if not downright incorrect interpretation of banding patterns. MspI and HpaII, two frequently used isoschizomers which differ in their sensitivity to DNA methylation in their recognition site are the most frequently used enzymes in MS-AFLPs. MspI and HpaII both cut unmethylated CCGG sites, but they show different activity for methylated CCGG sites. Whereas MspI can digest CCGG sites with internal (CG) methylation (C^{5m}CCGG), HpaII is able to digest hemi-methylated ^{5m}CCGG sites but not CCGG sites with internal CG methylation. Banding patterns showing a HpaII fragment but not an MspI fragment have thus been interpreted as evidence for CHG methylation, often analyzed

separately in MS-AFLP analysis. Whereas technically this could indeed be due to ^{5m}CCGG hemimethylation of the external cytosine(s), this can also occur due to variable methylation at an internal C^{5m}CCGG site within the MS-AFLP fragment, resulting in MspI to cut this fragment into two pieces, whereas HpaII does not (van Gurp et al. 2009). This was demonstrated in studies where the different bands were sequenced (Fulneček & Kovarík 2013) and references therein. Methylation of the external cytosines in plant methylomes in CCGG context almost always is accompanied by methylation of the internal cytosines (van Gurp et al. 2009; Fulneček & Kovarík 2013). As such, most studies interpreting single HpaII bands as evidence for CHG methylation, theorizing about its potential importance are most likely incorrect. Although improvements have been suggested for the MS-AFLP method, including a lane in which digestion with both MspI and HpaII simultaneously (Fulneček & Kovarík 2013) or using different enzymes sensitive to CHG methylation (Xu 2016; Bednarek et al. 2007); recent studies continue to incorrectly distinguish supposed CHG from CG methylation (Schulz et al. 2013), going as far as suggesting to analyze MspI/HpaII 0/1 and MspI/HpaII 1/0 patterns separately, as this would improve the understanding of the analysis outcome (Alonso et al. 2015). To resolve these issues, new sequencing-based techniques such as epiGBS are required to provide an improved resolution, functional annotation of the loci targeted, and unanimous quantitative DNA methylation estimates at a single nucleotide resolution. Here, I compare epiGBS to MS-AFLP, reduced representation and whole genome bisulfite sequencing (RRBS and WGBS respectively), exploring the strengths, weaknesses and improvement opportunities of epiGBS compared to these established methods.

6.3.1 *epiGBS vs MS-AFLP*

Compared to MS-AFLP, epiGBS offers various advantages such as a higher resolution (more markers), DNA methylation estimates in all sequence contexts (CG, CHG and CHH), quantitative instead of qualitative assessment of DNA methylation polymorphisms and functional annotation of sequenced regions, which are required in ecological epigenetics studies as described above. A potential disadvantage is the higher complexity of analysis methods that need to consider quantitative differences in DNA methylation ratios of individual cytosines, as compared to the simple presence-absence metric provided by MS-AFLP. Basic bioinformatics skills are required to use and implement the scripts required for analyzing epiGBS data whereas MS-AFLP data analysis is simpler.

6.3.2 *epiGBS vs RRBS*

Compared to the standard RRBS implementation (Meissner et al. 2005), epiGBS offers several advantages. Existing analysis methods for RRBS data require the presence of a reference genome. Using an advanced bioinformatics pipeline which can reconstruct the reference of the fragments targeted, epiGBS avoids this requirement. Second, existing RRBS protocols do not allow for cost-efficient multiplexing of multiple samples for sequencing. All existing RRBS techniques require 5-methylcytosines in the adapters, which are very expensive compared to normal cytosines. By using barcoded adapters combined with paired-end reads, a 96-plex design is enabled with only 12 forward and 8 reverse barcodes. DNA methylation variation is often determined by genetic variation. So, to understand the nature and contribution of the DNA methylation variation, genetic variation should also be measured. Simultaneous detection of both genetic and DNA methylation variation therefore

offers a more economical experimental design. The advanced epiGBS bioinformatics pipeline separates DNA methylation from genetic variation on a per sample basis, which also significantly improves the quality of DNA methylation calls, as SNPs that cause a sequence context to change from CG to CHH often cause a change in the DNA methylation status, which would be missed if these genetic variants would not be considered. Existing RRBS analysis pipelines lack this detection capacity. An advantage of RRBS over epiGBS is that RRBS data can also be produced by a single-end sequencing run, whereas epiGBS data require paired-end sequencing data for demultiplexing and de novo reference construction. This makes RRBS more flexible compared to epiGBS.

6.3.3 *epiGBS vs WGBS*

Compared to whole-genome bisulfite sequencing (WGBS), which is only available for species with a reference genome, epiGBS is only able to assess DNA methylation variation in a restriction site-associated and replicable subset of the genome, whereas whole genome bisulfite sequencing determines DNA methylation over the full genomic sequence. Although uneven coverage is a problem in most WGBS studies, this problem is much stronger in epiGBS, in which short fragments often get a 10 to 100 fold higher coverage compared to longer fragments (van Gurp et al. 2016). In the analysis, I proposed to eliminate fragments shorter than 100 bases, but even so, the PCR preference for shorter fragments cannot be completely avoided. Another issue that hampers all GBS-based sequencing technique is that it is not possible to eliminate PCR duplicates, which in WGBS are recognized by their identical beginning and end position on the reference genome. GBS-based methods lead to sequencing reads that always start and end in the same position regardless of their origin, making coordinate-based PCR duplicate removal impossible. The methylation sensitivity of the restriction enzyme used in epiGBS can lead to a bias in the detection of DNA methylation. I observed such a bias in DNA methylation estimates for cytosines in CHG context, for which I consequently detected a lower intergenerational correlation in both our dandelion epiGBS study (Chapter 4), as well as for epiGBS data in *Arabidopsis* (van Gurp et al. 2016), which I benchmarked against WGBS data from an intergenerational study on *Arabidopsis* (Becker et al. 2011). The CHG-methylation sensitivity of PstI blocks digestion of methylated restriction sites, leading to a bias in the restriction fragments obtained for bisulfite treatment and sequencing. As DNA methylation in CHG context is spatially auto-correlated, meaning that CHG methylation in the restriction site is often coupled with CHG methylation in the fragment, the CHG-methylation sensitivity of PstI can lead to an underestimation of the genome-wide levels of CHG methylation. Using the methylation insensitive enzyme Csp6I, the bias was largely avoided, leading to better intergenerational correlations for cytosines in CHG context at high sequencing depth (van Gurp et al. 2016). Finally, the efficiency with which epiGBS reads can be mapped is dependent on the heterozygosity of the fragments that result from digestion with the restriction enzyme employed. If a restriction enzyme targets repetitive regions, these regions will not be assembled or will be incorrectly grouped. Therefore, care should be taken in choosing an appropriate restriction enzyme for the species studied, as an incorrect choice can lead to a waste of sequencing data or worse, incorrect estimates of DNA methylation for repetitive regions based on improper merging of paralogous regions. Despite the drawbacks listed above, epiGBS allows for screening a defined subset of cytosines at an accuracy that is almost as high as WGBS for a fraction of the cost.

6.3.4 *epiGBS improvement opportunities*

Based on the comparisons against various existing techniques described in the previous paragraphs, several opportunities have become apparent for further improvement of the epiGBS technique. Controlling for PCR duplicates can be achieved by reducing the number of PCR cycles (Andrews et al. 2016; Boyle et al. 2012) or by using degenerate nucleotides in the adapters (Casbon et al. 2011; Schweyen et al. 2014; Tin et al. 2015), which help distinguish unique read (pairs) from PCR duplicates that map to the same locus. Another future improvement could be to enrich the epiGBS targets to specific loci of interest. For instance, in *Arabidopsis* epigenetic labile regions which are more prone to change are recognized (Wibowo et al. 2016), and functionally relevant DNA methylation changes often occur in differentially methylated regions (DMRs), which vary in size. In some studies, more economical epiGBS approaches could replace WGBS, if it could be modified to target those regions of interest. Implementing targeted sequencing in the epiGBS protocol seems feasible, as previous GBS and Radseq based techniques have achieved this with relative ease using oligonucleotide baits (Uitdewilligen et al. 2013), padlock probes (Deng et al. 2009) and other techniques as reviewed in (M. R. Jones & Good 2016) and (Hancock-Hanser et al. 2013). On the bioinformatics side, the detection of differentially methylated regions instead of only DMPs would improve the usability of the method. GBS (Elshire et al. 2011) and Radseq (Davey & Blaxter 2010) have seen high use and the development of several derived techniques (S. Wang et al. 2012; B. K. Peterson et al. 2012) as well as data analysis methods (<https://omictools.com/rad-seq-category>) greatly increase the use of both methods (Andrews et al. 2016). I expect that similar efforts for epiGBS will contribute to a more refined protocol, facilitating analysis with data similar in quality to WGBS.

6.4 *Forward looking statement*

Ecological epigenetics is an interesting and rapidly evolving field that so far has yielded intriguing observations that fuel speculation on the role and relevance of epigenetic variation in natural systems (Verhoeven et al. 2016). However, we still have more questions than answers. Compared to most non-model organisms, studies in the model plant *A. thaliana* offer more detailed insights in the complex patterns of DNA methylation dynamics and its relation to phenotypic variation. By removing technical limitations of MS-AFLPs and embracing bisulfite sequencing-based methods the field of ecological epigenetics can move forward in answering many of the outstanding questions. The studies conducted in this thesis have pioneered the use of epiGBS, allowing for high-resolution analysis of both stress induced and natural DNA methylation variation which hitherto where not possible in non-model organisms. Below, I propose how assaying more and diverse organisms and using advanced techniques such as epiGBS as well as established techniques can be used for conducting experiments that address some of the open questions in the field.

Comparative methylome studies in plants indicate a great variation in the level of DNA methylation in different sequence contexts, which to some extent are explained by repeat content and genome size (Niederhuth & Schmitz 2014). As such, appreciating the role that DNA methylation might play in shaping phenotypic diversity in different species will require studies in a variety of taxonomic groups. Besides these quantitative differences in DNA methylation ratios, the reproductive mode of the organism could affect the transgenerational transmission efficiency, as modified meiosis observed in asexual plant

species could lead to reduced resetting of epigenetic marks, increasing the transmission efficiency (Verhoeven & Preite 2014). The importance of epigenetic variation can also be influenced by life history traits such as the presence of negative biotic or abiotic interactions or generation time. Studies employing mutants for DNA methyltransferases in *Arabidopsis* have resulted in better insights in the dynamics and heritability of DNA methylation (Bartee et al. 2001; Law & Jacobsen 2010; Niederhuth & Schmitz 2014). Some enzymes are not present in all plant families, or are less effective. Plants in the Brassicaceae family have lower CHG methylation levels compared to other plants, whereas members of the Poaceae family show reduced CHH methylation levels (Niederhuth & Schmitz 2014). DNA methylation mutants exist in nature; *E. salsugineum* for example has no functional CMT3 enzyme, leading to almost complete absence of CHG methylation. Bisulfite sequencing based studies in such natural mutants or asexual plants, hitherto not yet done, can aid in determining the (diversity of) the role(s) that DNA methylation variation plays in shaping phenotypic variation, adaptation and ultimately (micro)evolution. Such studies are enabled by the epiGBS technique that we developed, even for species without a reference genome. Most published studies in ecological epigenetics have hitherto relied on MS-AFLPs. At this point, it is not clear what the relevance of the results of these studies is, given that they mostly target CG-DMPs which are not known to have a functional significance outside of transposons and promoters. Benchmarking such MS-AFLP based studies using epiGBS with the same enzymes used in MS-AFLP or even WGBS could shed light on what MS-AFLP does and does not detect. My epiGBS results do not provide evidence of heritable DNA methylation changes for cytosines in CG context, which is in contrast with the findings of a previous MS-AFLP based study, in which JA-stress application resulted in a significant increase in the number of polymorphic loci (Verhoeven, Jansen, et al. 2009). Unfortunately, although our epiGBS study used the same genotype, the genomic loci targeted are different and the results were obtained from a different experiment, making it difficult to compare both studies.

Bisulfite sequencing-based studies enabled by epiGBS offer many advantages over MS-AFLP, but the usefulness of the results can be increased significantly by knowing the local sequence context of the fragments obtained, for which an annotated reference genome is required. Even though the price of sequencing has exponentially declined over the past years (<https://www.genome.gov/sequencingcosts/>), assembling plant genomes has remained difficult, mainly due to their large genome size and high repeat content which make assembly difficult. Third generation sequencing techniques such as smart sequencing of pacific biosciences and the most recently developed nanopore sequencing offered by oxford nanopore can alleviate these problems by generating ultra-long (>100KB) sequencing reads (Lee et al. 2016). Additionally, whole genome bisulfite sequencing can be used to pinpoint DMRs and DMPs, based on which (targeted) epiGBS studies could be designed.

Understanding the role that natural epigenetic variation plays in shaping phenotypic variation is of great importance for understanding invasive species dynamics, response and adaptation to climate change but also for more applied fields such as plant breeding. Genetically identical *Fallopia japonica* plants, an invasive species in Europe and elsewhere, have different leaf shapes and grow to different heights depending on where they live,

which correlates with epigenetic variation that distinguishes these plants from different habitats (C. L. Richards et al. 2012). The role that epigenetic variation plays in adaptation to stressful conditions in natural populations is not yet clear (Bossdorf et al. 2007), but the reversible nature and rapid changes that can result from epigenetic changes coupled with the high number of documented cases of adaptive transgenerational stress responses suggest that epigenetic variation could play an important role in plant response to stress induced by climate change. Selection for phenotypic variation is the main goal of plant breeders, which aim for uniform traits and high yield. In *Brassica napus*, artificial selection on energy efficiency in isogenic population resulted in higher efficiency offspring which lead to a 5% yield increase (Hauben et al. 2009). Future studies addressing the role of epigenetic variation in shaping phenotypic diversity are thus important for understanding both natural events as well as artificial selection employed in breeding.

References

- Agrawal, A. & al, E., 1999. Transgenerational induction of defences in animals and plants. *Nature*.
- Allis, C.D. & Jenuwein, T., 2016. The molecular hallmarks of epigenetic control. *Nature reviews. Genetics*, 17(8), pp.487–500.
- Alonso, C. et al., 2015. MSAP markers and global cytosine methylation in plants: a literature survey and comparative analysis for a wild-growing species. *Molecular Ecology Resources*, 16(1), pp.80–90.
- Amoah, S. et al., 2012. A Hypomethylated population of *Brassica rapa* for forward and reverse Epi-genetics. *BMC plant biology*, 12.
- Andrews, K.R. et al., 2016. Harnessing the power of RADseq for ecological and evolutionary genomics. *Nature reviews. Genetics*, 17(2), pp.81–92.
- Andrés, F. & Coupland, G., 2012. The genetic basis of flowering responses to seasonal cues. *Nature reviews. Genetics*, 13(9), pp.627–639.
- Anway, M.D. et al., 2005. Epigenetic transgenerational actions of endocrine disruptors and male fertility. *Science*, 308(5727), pp.1466–1469.
- Ashrafi, H. et al., 2012. De novo assembly of the pepper transcriptome (*Capsicum annuum*): a benchmark for in silico discovery of SNPs, SSRs and candidate genes. *BMC genomics*, 13, p.571.
- Assenov, Y. et al., 2014. Comprehensive analysis of DNA methylation data with RnBeads. *Nature methods*, 11(11), pp.1138–1140.
- Audergon, P.N.C.B. et al., 2015. Epigenetics. Restricted epigenetic inheritance of H3K9 methylation. *Science*, 348(6230), pp.132–135.
- Baker, H.G., 1965. *Characteristics and Modes of Origin of Weeds*,
- Baker, H.G., Baker, H.G. & Stebbins, G.L., 1965. Characteristics and modes of origin of weeds. *The genetics of colonizing species*, pp.147–168.
- Bartee, L., Malagnac, F. & Bender, J.T.A.C.C.M.B.N.-C.M.A.S.O.A.E.G., 2001. Arabidopsis cmt3 chromomethylase mutations block non-CG methylation and silencing of an endogenous gene. *Genes & development*, 15(14), pp.1753–1758.
- Beck, S., 2010. Taking the measure of the methylome. *Nature Biotechnology*, 28(10), pp.1026–1028.
- Becker, C. et al., 2011. Spontaneous epigenetic variation in the *Arabidopsis thaliana* methylome. *Nature*, 480(7376), pp.245–249.

- Bednarek, P.T. et al., 2007. Quantification of the tissue-culture induced variation in barley (*Hordeum vulgare* L.). *BMC plant biology*, 7, p.10.
- Benjamini, Y. & Speed, T.P., 2012. Summarizing and correcting the GC content bias in high-throughput sequencing. *Nucleic acids research*, 40(10), p.e72.
- Bewick, A.J. et al., 2016. On the origin and evolutionary consequences of gene body DNA methylation. *Proceedings of the National Academy of Sciences*, 113(32), pp.9111–9116.
- Bonduriansky, R. & Day, T., 2009. Nongenetic inheritance and its evolutionary implications. *Annual Review of Ecology, Evolution, and Systematics*, 40, pp.103–125.
- Bossdorf, O. & Zhang, Yuanye, 2011. A truly ecological epigenetics study. *Molecular Ecology*, 20(8), pp.1572–1574.
- Bossdorf, O. et al., 2010. Experimental alteration of DNA methylation affects the phenotypic plasticity of ecologically relevant traits in *Arabidopsis thaliana*. *Evolutionary Ecology*, 24(3), pp.541–553. Available at: <http://www.springerlink.com/index/C847255UR67W2487.pdf>.
- Bossdorf, O., Richards, C.L. & Pigliucci, M., 2007. Epigenetics for ecologists. *Ecology Letters*, 0(0), pp.106–115.
- Bowler, P.J., 1989. *Evolution: the history of an idea*, Univ of California Press.
- Boyle, P. et al., 2012. Gel-free multiplexed reduced representation bisulfite sequencing for large-scale DNA methylation profiling. *Genome biology*, 13(10), p.R92.
- Bruce, T.J.A. et al., 2007. Stressful “memories” of plants: Evidence and possible mechanisms. *Plant Science*, 173(6), pp.603–608.
- Burgess, K.S., Etterson, J.R. & Galloway, L.F., 2007. Artificial selection shifts flowering phenology and other correlated traits in an autotetraploid herb. *Heredity*, 99(6), pp.641–648.
- Burn, J.E. et al., 1993. DNA methylation, vernalization, and the initiation of flowering. *Proceedings of the National Academy of Sciences of the United States of America*, 90(1), pp.287–291.
- Busconi, M. et al., 2015. AFLP and MS-AFLP Analysis of the Variation within Saffron Crocus (*Crocus sativus* L.) Germplasm M. Gijzen, ed. *PLoS One*, 10(4), p.e0123434.
- Casbon, J.A. et al., 2011. A method for counting PCR template molecules with application to next-generation sequencing. *Nucleic acids research*, 39(12), p.e81.
- Chandler, V.L., 2007. Paramutation: from maize to mice. *Cell*, 128(4), pp.641–645.

- Chen, S.C.A.F., 2014. Characterization of in vitro haploid and doubled haploid Chrysanthemum morifolium plants via unfertilized ovule culture for phenotypic traits and DNA methylation pattern. pp.1–10.
- Chen, Z.J., 2013. Genomic and epigenetic insights into the molecular bases of heterosis. *Nature reviews. Genetics*, 14(7), pp.471–482.
- Choi, C.-S. & Sano, H., 2007. Abiotic-stress induces demethylation and transcriptional activation of a gene encoding a glycerophosphodiesterase-like protein in tobacco plants. *Molecular genetics and genomics : MGG*, 277(5), pp.589–600.
- Christman, J.K., 2002. 5-Azacytidine and 5-aza-2'-deoxycytidine as inhibitors of DNA methylation: mechanistic studies and their implications for cancer therapy. *Oncogene*, 21(35), pp.5483–5495.
- Cokus, S.J. et al., 2008. Shotgun bisulphite sequencing of the Arabidopsis genome reveals DNA methylation patterning. *Nature*, 452(7184), pp.215–219.
- Collier, M.H. & Rogstad, S.H., 2004. Clonal variation in floral stage timing in the common dandelion *Taraxacum officinale* (Asteraceae). *American journal of botany*, 91(11), pp.1828–1833.
- Cortijo, S. et al., 2014. Mapping the Epigenetic Basis of Complex Traits. *Science*, 343(6175), pp.1145–1148.
- Crisp, P.A. et al., 2016. Reconsidering plant memory: Intersections between stress recovery, RNA turnover, and epigenetics. *Science advances*, 2(2), p.e1501340.
- Cubas, P., Vincent, C. & Coen, E., 1999. An epigenetic mutation responsible for natural variation in floral symmetry. *Nature*, 401(6749), pp.157–161.
- Dalakouras, A. et al., 2012. Transgenerational maintenance of transgene body CG but not CHG and CHH methylation. *Epigenetics : official journal of the DNA Methylation Society*, 7(9), pp.1071–1078.
- Davey, J.L. & Blaxter, M.W., 2010. RADSeq: next-generation population genetics. *Briefings In Functional Genomics*, 9(5-6), pp.416–423.
- Delker, C. et al., 2006. Jasmonate Biosynthesis in Arabidopsis thaliana - Enzymes, Products, Regulation. *Plant biology (Stuttgart, Germany)*, 8(03), pp.297–306–306.
- Deng, J. et al., 2009. Targeted bisulfite sequencing reveals changes in DNA methylation associated with nuclear reprogramming. *Nature Biotechnology*, 27(4), pp.353–360.
- Dohm, J.C. et al., 2008. Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic acids research*, 36(16), pp.e105–e105.

- Downen, R.H. et al., 2012. Widespread dynamic DNA methylation in response to biotic stress. *Proceedings of the National Academy of Sciences of the United States of America*, 109(32), pp.E2183–E2191.
- Dubin, M.J. et al., 2015. DNA methylation in *Arabidopsis* has a genetic basis and shows evidence of local adaptation. *eLife*, 4, p.e05255.
- Edgar, R.C., 2010. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*, 26(19), pp.2460–2461.
- Ehrlén, J. & Münzbergová, Z., 2009. Timing of flowering: opposed selection on different fitness components and trait covariation. *The American naturalist*, 173(6), pp.819–830.
- Elshire, R.J. et al., 2011. A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS One*, 6(5), p.e19379.
- Fieldes, M.A. & Amyot, L.M., 2000. Evaluating the potential of using 5-azacytidine as an epimutagen. *Canadian Journal of Botany*, 77(11), pp.1617–1622.
- Fulneček, J. & Kovarik, A., 2013. How to interpret methylation sensitive amplified polymorphism (MSAP) profiles? *BMC Genetics*, 15(1), pp.2–2.
- Futuyma, D.J., 1986. The evolution of interactions among species. *Evolutionary Biology*, Sinauer Associates, Inc., Sunderland, MA, pp.482–504.
- Garrison, E. & Marth, G., 2012. Haplotype-based variant detection from short-read sequencing. *arXiv:1207.3907v2*, q-bio.GN.
- Gaydos, L.J., Wang, W. & Strome, S., 2014. Gene repression. H3K27me and PRC2 transmit a memory of repression across generations and during development. *Science*, 345(6203), pp.1515–1518.
- Gehring, M., 2016. Prodigious plant methylomes. *Genome biology*, pp.1–3.
- Greaves, I.K. et al., 2012. Trans Chromosomal Methylation in *Arabidopsis* hybrids. *Proceedings of the National Academy of Sciences*, 109(9), pp.3570–3575.
- Grossniklaus, U. et al., 2013. Transgenerational epigenetic inheritance: how important is it? *Nature reviews. Genetics*, 14(3), pp.228–235.
- Gu, H. et al., 2011. Preparation of reduced representation bisulfite sequencing libraries for genome-scale DNA methylation profiling. *Nature Protocols*, 6(4), pp.468–481.
- Gugger, P.F. et al., 2016. Species-wide patterns of DNA methylation variation in *Quercus lobata* and their association with climate gradients. *Molecular Ecology*, 25(8), pp.1665–1680.

- Hagmann, J. et al., 2015. Century-scale Methylome Stability in a Recently Diverged *Arabidopsis thaliana* Lineage T. Kakutani, ed. *PLoS Genetics*, 11(1), p.e1004920.
- Hancock-Hanser, B.L. et al., 2013. Targeted multiplex next-generation sequencing: advances in techniques of mitochondrial and nuclear DNA sequencing for population genomics. *Molecular Ecology Resources*, 13(2), pp.254–268.
- Hansen, K.D., Brenner, S.E. & Dudoit, S., 2010. Biases in Illumina transcriptome sequencing caused by random hexamer priming. *Nucleic acids research*, 38(12), pp.e131–e131.
- Harris, R.A. et al., 2010. Comparison of sequencing-based methods to profile DNA methylation and identification of monoallelic epigenetic modifications. *Nature Biotechnology*, 28(10), pp.1097–1105.
- Hauben, M. et al., 2009. Energy use efficiency is characterized by an epigenetic component that can be directed through artificial selection to increase yield. *Proceedings of the National Academy of Sciences*, 106(47), pp.20109–20114.
- He, X.-J., Chen, T. & Zhu, J.-K., 2011. Regulation and function of DNA methylation in plants and animals. *Cell research*, 21(3), pp.442–465.
- Heard, E. & Martienssen, R.A., 2014. Transgenerational Epigenetic Inheritance: Myths and Mechanisms. *Cell*, 157(1), pp.95–109.
- Herman, J.J., 2016. *Epigenetics of Adaptive Plasticity: An Investigation of Plant Responses to Environmental Stress Within and Across Generations*, Available at: http://wescholar.wesleyan.edu/etd_diss/70.
- Herman, J.J. & Sultan, S.E., 2016. DNA methylation mediates genetic variation for adaptive transgenerational plasticity. *Proceedings. Biological sciences*, 283(1838).
- Holeski, L.M., Jander, G. & Agrawal, A.A., 2016. Transgenerational defense induction and epigenetic inheritance in plants. *Trends in Ecology & Evolution*, pp.1–9.
- Ikeda, Y. & Nishimura, T., 2015. The Role of DNA Methylation in Transposable Element Silencing and Genomic Imprinting. In *Nuclear Functions in Plant Transcription, Signaling and Development*. New York, NY: Springer, pp. 13–29. Available at: http://link.springer.com/10.1007/978-1-4939-2386-1_2.
- Jablonka, E. & Lamb, M.J., 1998. Epigenetic inheritance in evolution. *Journal of Evolutionary Biology*, 11(2), pp.159–183.
- Jablonka, E., Lamb, M.J. & Zeligowski, A., 2014. *Evolution in four dimensions, revised edition: Genetic, epigenetic, behavioral, and symbolic variation in the history of life*, MIT press.
- Jacobsen, S.E. & Meyerowitz, E.M., 1997. Hypermethylated SUPERMAN epigenetic alleles in *Arabidopsis*. *Science*, 277(5329), pp.1100–1103.

- Jaligot, E. & Rival, A., 2015. Applying Epigenetics in Plant Breeding: Balancing Genome Stability and Phenotypic Plasticity. In Cham: Springer International Publishing, pp. 159–192.
- Jiang, C. et al., 2014. Environmentally responsive genome-wide accumulation of de novo *Arabidopsis thaliana* mutations and epimutations. *Genome Research*.
- Jiang, L. et al., 2011. Synthetic spike-in standards for RNA-seq experiments. *Genome Research*, 21(9), pp.1543–1551.
- Johannes, F. et al., 2009. Assessing the Impact of Transgenerational Epigenetic Variation on Complex Traits. *PLoS Genetics*, 5(6).
- Jones, A.L. & Sung, S., 2014. Mechanisms Underlying Epigenetic Regulation in *Arabidopsis thaliana*. *Integrative and Comparative Biology*, 54(1), pp.61–67.
- Jones, M.R. & Good, J.M., 2016. Targeted capture in evolutionary and ecological genomics. *Molecular Ecology*, 25(1), pp.185–202.
- Jones, P.A., 1985. Altering gene expression with 5-azacytidine. *Cell*, 40(3), pp.485–486.
- Kalisz, S. & Purugganan, M.D., 2004. Epialleles via DNA methylation: consequences for plant evolution. *Trends in Ecology & Evolution*, 19(6), pp.309–314.
- Kamthan, A. et al., 2015. Small RNAs in plants: recent development and application for crop improvement. *Frontiers in Plant Science*, 6, p.208.
- Kawakatsu, T. et al., 2016. Epigenomic Diversity in a Global Collection of *Arabidopsis thaliana* Accessions. *Cell*, 166(2), pp.492–505.
- Kilvitis, H.J. et al., 2013. Ecological Epigenetics. In C. R. Landry & N. Aubin-Horth, eds. *Ecological Genomics: Ecology and the Evolution of Genes and Genomes*. Ecological Genomics: Ecology and the Evolution of Genes and Genomes. Dordrecht: Springer Netherlands, pp. 191–210.
- Kiselev, K.V., Tyunin, A.P. & Karetin, Y.A., 2014. Salicylic acid induces alterations in the methylation pattern of the VaSTS1, VaSTS2, and VaSTS10 genes in *Vitis amurensis* Rupr. cell cultures. *Plant Cell Reports*, 34(2), pp.311–320.
- Kleinman, C.L. & Majewski, J., 2012. Comment on “Widespread RNA and DNA Sequence Differences in the Human Transcriptome.” *Science*, 335(6074), pp.1302–1302.
- Klironomos, F.D., Berg, J. & Collins, S., 2013. How epigenetic mutations can affect genetic evolution: model and mechanism. *BioEssays : news and reviews in molecular, cellular and developmental biology*, 35(6), pp.571–578.
- Koboldt, D.C. et al., 2012. VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Research*, 22(3), pp.568–576.

- Kovalchuk, I. & Kovalchuk, O., 2012. *Epigenetics in health and disease*, FT Press.
- Kronholm, I. & Collins, S., 2016. Epigenetic mutations can both help and hinder adaptive evolution. *Molecular Ecology*, 25(8), pp.1856–1868.
- Latzel, V. et al., 2012. Epigenetic variation in plant responses to defence hormones. *Annals of botany*.
- Law, J.A. & Jacobsen, S.E., 2010. Establishing, maintaining and modifying DNA methylation patterns in plants and animals. *Nature reviews. Genetics*, 11(3), pp.204–220.
- Lee, H. et al., 2016. Third-generation sequencing and the future of genomics. *Biorxiv*, p.048603.
- Lev Maor, G., Yearim, A. & Ast, G., 2015. The alternative role of DNA methylation in splicing regulation. *Trends In Genetics*, 31(5), pp.274–280.
- Li, H. et al., 2009. The Sequence Alignment/Map format and SAMtools.
- Li, M. et al., 2011. Widespread RNA and DNA sequence differences in the human transcriptome. *Science*, 333(6038), pp.53–58.
- Li, Q. et al., 2014. Inheritance patterns and stability of DNA methylation variation in maize near-isogenic lines. *Genetics*, 196(3), pp.667–676.
- Li, X. et al., 2012. Single-base resolution maps of cultivated and wild rice methylomes and regulatory roles of DNA methylation in plant gene expression. *BMC genomics*, 13, p.300.
- Lin, W. et al., 2012. Comment on “Widespread RNA and DNA Sequence Differences in the Human Transcriptome.” *Science*, 335(6074), pp.1302–1302.
- Lister, R. & Ecker, J.R., 2009. Finding the fifth base: Genome-wide sequencing of cytosine methylation. *Genome Research*, 19(6), pp.959–966.
- Lister, R. et al., 2008. Highly integrated single-base resolution maps of the epigenome in Arabidopsis. *Cell*, 133(3), pp.523–536.
- Liu, Y. et al., 2012. Bis-SNP: Combined DNA methylation and SNP calling for Bisulfite-seq data. *Genome biology*, 13(7), p.R61.
- Lusser, M. et al., 2012. Deployment of new biotechnologies in plant breeding. *Nature Biotechnology*, 30(3), pp.231–239.
- Lynch, M. et al., 1993. The mutational meltdown in asexual populations. *Journal Of Heredity*, 84(5), pp.339–344.

- Manning, K. et al., 2006. A naturally occurring epigenetic mutation in a gene encoding an SBP-box transcription factor inhibits tomato fruit ripening. *Nature genetics*, 38(8), pp.948–952.
- Matvienko, M. et al., 2013. Consequences of Normalizing Transcriptomic and Genomic Libraries of Plant Genomes Using a Duplex-Specific Nuclease and Tetramethylammonium Chloride. *PLoS One*, 8(2), p.e55913.
- Mayr, E., 2001. *What evolution is*, Basic books.
- Mayr, E. & Provine, W.B., 1998. *The evolutionary synthesis: perspectives on the unification of biology*, Harvard University Press.
- Meissner, A. et al., 2005. Reduced representation bisulfite sequencing for comparative high-resolution DNA methylation analysis. *Nucleic acids research*, 33(18), pp.5868–5877.
- Meng, D. et al., 2016. Limited Contribution of DNA Methylation Variation to Expression Regulation in *Arabidopsis thaliana*. B. S. Gaut, ed. *PLoS Genetics*, 12(7), p.e1006141.
- Mirouze, M. & Paszkowski, J., 2011. Epigenetic contribution to stress adaptation in plants. *Current opinion in plant biology*, 14(3), pp.267–274.
- Molina-Montenegro, M.A. et al., 2011. Functional differences in response to drought in the invasive *Taraxacum officinale* from native and introduced alpine habitat ranges. *Plant Ecology & Diversity*, 4(1), pp.37–44.
- Molinier, J. et al., 2006. Transgeneration memory of stress in plants. *Nature*, 442(7106), pp.1046–1049.
- Niederhuth, C.E. & Schmitz, R.J., 2014. Covering your bases: inheritance of DNA methylation in plant genomes. *Molecular plant*, 7(3), pp.472–480.
- Niederhuth, C.E. & Schmitz, R.J., 2016. Putting DNA methylation in context: from genomes to gene expression in plants. *Biochimica et biophysica acta*.
- Niederhuth, C.E., Bewick, A.J., Ji, L., Alabady, M., Kim, K.D., Page, J.T., et al., 2016. *Widespread natural variation of DNA methylation within angiosperms*,
- Niederhuth, C.E., Bewick, A.J., Ji, L., Alabady, M.S., Kim, K.D., Li, Q., et al., 2016. Widespread natural variation of DNA methylation within angiosperms. *Genome biology*, 17(1), p.194.
- Oplaat, C. & Verhoeven, K.J.F., 2015. Range expansion in asexual dandelions: selection for general-purpose genotypes? N. Rafferty, ed. *Journal of Ecology*, 103(1), pp.261–268.
- Osborn, H.F., 1896. *From the Greeks to Darwin: an outline of the development of the evolution idea*, Macmillan.

- Ossowski, S. et al., 2010. The rate and molecular spectrum of spontaneous mutations in *Arabidopsis thaliana*. *Science*, 327(5961), pp.92–94.
- Ou, X. et al., 2012. Transgenerational Inheritance of Modified DNA Methylation Patterns and Enhanced Tolerance Induced by Heavy Metal Stress in Rice (*Oryza sativa* L.) K. Wu, ed. *PLoS One*, 7(9), p.e41143.
- O’Dea, R.E. et al., 2016. The role of non-genetic inheritance in evolutionary rescue: epigenetic buffering, heritable bet hedging and epigenetic traps. *Environmental Epigenetics*, 2(1), p.dvv014.
- Pedersen, B.S. et al., 2014. Fast and accurate alignment of long bisulfite-seq reads. *arXiv:1401.1129v2*.
- Peng, H. & Zhang, Jing, 2008. Plant genomic DNA methylation in response to stresses: Potential applications and challenges in plant breeding. ..., 19(9), pp.1037–1045.
- Peterson, B.K. et al., 2012. Double digest RADseq: an inexpensive method for de novo SNP discovery and genotyping in model and non-model species. *PLoS One*, 7(5), p.e37135.
- Peterson, C.L. & Laniel, M.-A., 2004. Histones and histone modifications. *Current biology*, 14(14), pp.R546–51.
- Pickrell, J.K., Gilad, Y. & Pritchard, J.K., 2012. Comment on "Widespread RNA and DNA sequence differences in the human transcriptome". *Science*, 335(6074), pp.1302–author reply 1302.
- Pieterse, C.M. & Dicke, M., 2007. Plant interactions with microbes and insects: from molecular mechanisms to ecology. *Trends in Plant Science*, 12(12), pp.564–569.
- Pigliucci, M., 2002. Ecology and evolutionary biology of *Arabidopsis*. *The Arabidopsis book*, 1, p.e0003.
- Platt, A. et al., 2015. Genome-wide signature of local adaptation linked to variable CpG methylation in oak populations. *Molecular Ecology*, 24(15), pp.3823–3830.
- Plongthongkum, N., Diep, D.H. & Zhang, K., 2014. Advances in the profiling of DNA modifications: cytosine methylation and beyond. *Nature reviews. Genetics*, 15(10), pp.647–661.
- Preite, V., 2016. *Epigenetic inheritance in apomictic dandelions: stress-induced and heritable modifications in DNA methylation and small RNA*. Wageningen: Wageningen University.
- Rapp, R.A. & Wendel, J.F., 2005. Epigenetics and plant evolution. *The New phytologist*, 168(1), pp.81–91.

- Regulski, M. et al., 2013. The maize methylome influences mRNA splice sites and reveals widespread paramutation-like switches guided by small RNA. *Genome Research*, 23(10), pp.1651–1662.
- Rédei, G.P., 1992. A heuristic glance at the past of Arabidopsis genetics. *Methods in Arabidopsis research*, pp.1–15.
- Richards, C. & al, E., 2010. What Role Does Heritable Epigenetic Variation Play in Phenotypic Evolution? *BioScience*.
- Richards, C.L. et al., 2012. Invasion of diverse habitats by few Japanese knotweed genotypes is correlated with epigenetic differentiation. C. L. Richards, A. W. Schrey, & M. Pigliucci, eds. *Ecology Letters*, 15(9), pp.1016–1025.
- Richards, E.J., 2006. Inherited epigenetic variation--revisiting soft inheritance. *Nature reviews. Genetics*, 7(5), pp.395–401.
- Rigby, P.W.J. et al., 1977. Labeling deoxyribonucleic acid to high specific activity in vitro by nick translation with DNA polymerase I. *Journal of molecular biology*, 113(1), pp.237–251.
- Sailer, C., Schmid, B. & Grossniklaus, U., 2016. Apomixis Allows the Transgenerational Fixation of Phenotypes in Hybrid Plants. *Current biology*, 26(3), pp.331–337.
- Salmon, A., Ainouche, M.L. & Wendel, J.F., 2005. Genetic and epigenetic consequences of recent hybridization and polyploidy in *Spartina* (Poaceae). *Molecular Ecology*, 14(4), pp.1163–1175.
- Sarkies, P. & Miska, E.A., 2014. Small RNAs break out: the molecular cell biology of mobile small RNAs. *Nature reviews. Molecular cell biology*, 15(8), pp.525–535.
- Schmitz, R.J. et al., 2011. Transgenerational Epigenetic Instability Is a Source of Novel Methylation Variants. *Science*, 334(6054), pp.369–373.
- Schmitz, R.J., He, Y., et al., 2013. Epigenome-wide inheritance of cytosine methylation variants in a recombinant inbred population. *Genome Research*, 23(10), pp.1663–1674.
- Schmitz, R.J., Schultz, M.D., et al., 2013. Patterns of population epigenomic diversity. *Nature*, 495(7440), pp.193–198.
- Schrey, A.W. et al., 2013. Ecological Epigenetics: Beyond MS-AFLP. *Integrative and Comparative Biology*, 53, pp.340–350.
- Schulz, B., Eckstein, R.L. & Durka, W., 2013. Scoring and analysis of methylation-sensitive amplification polymorphisms for epigenetic population studies. *Molecular Ecology Resources*, 13(4), pp.642–653.

- Schwartz, S., Oren, R. & Ast, G., 2011. Detection and Removal of Biases in the Analysis of Next-Generation Sequencing Reads P. Lopez-Garcia, ed. *PLoS One*, 6(1), p.e16685.
- Schweyen, H., Rozenberg, A. & Leese, F., 2014. Detection and removal of PCR duplicates in population genomic ddRAD studies by addition of a degenerate base region (DBR) in sequencing adapters. *The Biological bulletin*, 227(2), pp.146–160.
- Secco, D. et al., 2015. Stress induced gene expression drives transient DNA methylation changes at adjacent repetitive elements. *eLife*, 4.
- Shea, N., Pen, I. & Uller, T., 2011. Three epigenetic information channels and their different roles in evolution. *Journal of Evolutionary Biology*, 24(6), pp.1178–1187.
- Simons, A.M., 2011. Modes of response to environmental change and the elusive empirical evidence for bet hedging. *Proceedings. Biological sciences*, 278(1712), pp.1601–1609.
- Slotkin, R.K. & Martienssen, R., 2007. Transposable elements and the epigenetic regulation of the genome. *Nature reviews. Genetics*, 8(4), pp.272–285.
- Slotkin, R.K. et al., 2009. Epigenetic reprogramming and small RNA silencing of transposable elements in pollen. *Cell*, 136(3), pp.461–472.
- Song, Q.-X. et al., 2013. Genome-wide analysis of DNA methylation in soybean. *Molecular plant*, 6(6), pp.1961–1974.
- Steward, N. et al., 2002. Periodic DNA methylation in maize nucleosomes and demethylation by environmental stress. *The Journal of biological chemistry*, 277(40), pp.37741–37746.
- Sugimoto, N., Nakano, M. & Nakano, S., 2000. Thermodynamics-structure relationship of single mismatches in RNA/DNA duplexes. *Biochemistry*, 39(37), pp.11270–11281.
- Sultan, S.E., 2011. Adaptive transgenerational plasticity in plants: case studies, mechanisms, and implications for natural populations. pp.1–10.
- Sultan, S.E., Barton, K. & Wilczek, A.M., 2009. Contrasting patterns of transgenerational plasticity in ecologically distinct congeners. *Ecology*, 90(7), pp.1831–1839.
- Suzuki, M.M. & Bird, A., 2008. DNA methylation landscapes: provocative insights from epigenomics. *Nature reviews. Genetics*, 9(6), pp.465–476.
- Swillen, I., Vanoverbeke, J. & De Meester, L., 2015. Inbreeding and adaptive plasticity: an experimental analysis on predator-induced responses in the water flea *Daphnia*. *Ecology and evolution*, 5(13), pp.2712–2721.
- Takahata, N., 1996. Neutral theory of molecular evolution. *Current Opinion in Genetics & Development*, 6(6), pp.767–772.

- Takuno, S. & Gaut, B.S., 2013. Gene body methylation is conserved between plant orthologs and is of evolutionary consequence. *Proceedings of the National Academy of Sciences*, 110(5), pp.1797–1802.
- Tas, I.C. & Van Dijk, P.J., 1999. Crosses between sexual and apomictic dandelions (*Taraxacum*). I. The inheritance of apomixis. *Heredity*, 83 (Pt 6), pp.707–714.
- Tin, M.M.Y. et al., 2015. Degenerate adaptor sequences for detecting PCR duplicates in reduced representation sequencing data improve genotype calling accuracy. *Molecular Ecology Resources*, 15(2), pp.329–336.
- Trerotola, M. et al., 2015. Epigenetic inheritance and the missing heritability. *Human genomics*, 9, p.17.
- Trucchi, E. et al., 2016. BsRADseq: screening DNA methylation in natural populations of non-model species. *Molecular Ecology*, 25(8), pp.1697–1713.
- Tytgat, T.O.G. et al., 2013. Plants Know Where It Hurts: Root and Shoot Jasmonic Acid Induction Elicit Differential Responses in Brassica oleracea G. Bonaventure, ed. *PLoS One*, 8(6), p.e65502.
- Uitdewilligen, J.G.A.M.L. et al., 2013. A Next-Generation Sequencing Method for Genotyping-by-Sequencing of Highly Heterozygous Autotetraploid Potato L. Lukens, ed. *PLoS One*, 8(5), p.e62355.
- van der Graaf, A. et al., 2015. Rate, spectrum, and evolutionary dynamics of spontaneous epimutations. *Proceedings of the National Academy of Sciences*, 112(21), pp.6676–6681.
- Van Dijk, P.J. et al., 1999. Crosses between sexual and apomictic dandelions (*Taraxacum*). II. The breakdown of apomixis. *Heredity*, 83 (Pt 6), pp.715–721.
- van Gurp, T.P. et al., 2016. epiGBS: reference-free reduced representation bisulfite sequencing. *Nature methods*.
- van Gurp, T.P., McIntyre, L.M. & Verhoeven, K.J.F., 2013. Consistent errors in first strand cDNA due to random hexamer mispriming. *PLoS One*, 8(12), p.e85583.
- van Gurp, T.P., Wagemaker, N. & Bakker, F., 2009. *Measuring the Arabidopsis methylome using secondary digest MSAP: an in vitro & in silico approach*. Wageningen University.
- Vandegheuchte, M.B. et al., 2009. Occurrence of DNA methylation in *Daphnia magna* and influence of multigeneration Cd exposure. *Environment International*, 35(4), pp.700–706.
- Vergeer, P., Wagemaker, N.C.A.M. & Ouborg, N.J., 2012. Evidence for an epigenetic role in inbreeding depression. *Biology letters*, 8(5), pp.798–801.

- Verhoeven, K.J.F. & Preite, V., 2014. Epigenetic variation in asexually reproducing organisms. *Evolution*, 68(3), pp.644–655.
- Verhoeven, K.J.F. & van Gorp, T.P., 2012. Transgenerational Effects of Stress Exposure on Offspring Phenotypes in Apomictic Dandelion. *PLoS One*, 7(6), p.e38605.
- Verhoeven, K.J.F., Jansen, J.J., et al., 2009. Stress-induced DNA methylation changes and their heritability in asexual dandelions. *New Phytologist*, 185(4), pp.1108–1118. Available at: <http://onlinelibrary.wiley.com/doi/10.1111/j.1469-8137.2009.03121.x/full>.
- Verhoeven, K.J.F., van Dijk, P.J. & Biere, A., 2009. Changes in genomic methylation patterns during the formation of triploid asexual dandelion lineages. *Molecular Ecology*, 19(2), pp.315–324.
- Verhoeven, K.J.F., vonHoldt, B.M. & Sork, V.L., 2016. Epigenetics in ecology and evolution: what we know and what we need to know. *Molecular Ecology*, 25(8), pp.1631–1638.
- Verkest, A. et al., 2015. Selection for Improved Energy Use Efficiency and Drought Tolerance in Canola Results in Distinct Transcriptome and Epigenome Changes. *Plant physiology*, 168(4), pp.1338–1350.
- Wang, H. et al., 2015. CG gene body DNA methylation changes and evolution of duplicated genes in cassava. *Proceedings of the National Academy of Sciences*, 112(44), pp.13729–13734.
- Wang, J. et al., 2013. Double restriction-enzyme digestion improves the coverage and accuracy of genome-wide CpG methylation profiling by reduced representation bisulfite sequencing. *BMC genomics*, 14(1), pp.1–1.
- Wang, S. et al., 2012. 2b-RAD: a simple and flexible method for genome-wide genotyping. *Nature Publishing Group*.
- Weigel, D. & Colot, V., 2012. Epialleles in plant evolution. *Genome biology*, 13(10), p.249.
- Wibowo, A. et al., 2016. Hyperosmotic stress memory in Arabidopsis is mediated by distinct epigenetically labile sites in the genome and is restricted in the male germline by DNA glycosylase activity. *eLife*, 5.
- Widman, N. et al., 2014. Epigenetic differences between shoots and roots in Arabidopsis reveals tissue-specific regulation. *Epigenetics : official journal of the DNA Methylation Society*, 9(2), pp.236–242.
- Wilschut, R.A. et al., 2016. Natural epigenetic variation contributes to heritable flowering divergence in a widespread asexual dandelion lineage. *Molecular Ecology*, 25(8), pp.1759–1768.
- Xie, M. & Yu, B., 2015. siRNA-directed DNA Methylation in Plants. *Current genomics*, 16(1), pp.23–31.

- Xu, J., 2016. DEVELOPING A NOVEL HYPOMETHYLATED POPULATION AND CHARACTERIZING A STABLE EARLY FLOWERING EPIMUTANT IN STRAWBERRY (FRAGARIA VESCA). pp.1–178.
- Zhang, Jiajie et al., 2014. PEAR: a fast and accurate Illumina Paired-End reAd mergeR. *Bioinformatics*, 30(5), pp.614–620.
- Zheng, W., Chung, L.M. & Zhao, H., 2011. Bias detection and correction in RNA-Sequencing data. *Bmc Bioinformatics*, 12(1), p.290.
- Zheng, X. et al., 2012. A high-performance computing toolset for relatedness and principal component analysis of SNP data. *Bioinformatics*, 28(24), pp.3326–3328.
- Zilberman, D. et al., 2007. Genome-wide analysis of Arabidopsis thaliana DNA methylation uncovers an interdependence between methylation and transcription. *Nature genetics*, 39(1), pp.61–69.
- Zook, J.M. et al., 2012. Synthetic spike-in standards improve run-specific systematic error analysis for DNA and RNA sequencing. J. M. Zook et al., eds. *PLoS One*, 7(7), p.e41356.

Summary

The success or continuous existence of species requires continuous adaptation to changes in the environment to survive and contribute offspring to the next generation. Selection acts on the phenotype, which is in turn determined by the complex interplay of genetic, epigenetic and environmental variation. (Natural) selection leads to 'survival of the fittest' or best-adapted individuals to their local environment, ultimately determining which individuals contribute offspring to the next generation. Understanding the mechanisms by which epigenetic and genetic variation can arise and get passed on through generations determines our understanding of inheritance and evolution. Hitherto, the mechanistic understanding of genetics has shaped the scientific view of inheritance and evolution, leading to the gene-centered paradigm of Neo-Darwinism. However, recent studies indicate that besides genetic (DNA sequence) variation, epigenetic variation can also be transmitted between generations. Further studies on the properties and transgenerational dynamics of epigenetic variation are needed to enhance our understanding of heritability and evolution.

Epigenetic variation has distinct properties and different transgenerational dynamics compared to genetic variation. Epigenetic variation helps to regulate gene expression and determines the different cell types and function in eukaryotes. The main function of DNA methylation, an important part of the epigenetic code, is to prevent the spread of selfish genetic elements in the genome and to establish the different cellular profiles observed in multicellular organisms. One differentiating feature of epigenetic variation compared to genetic variation is that (specific) epigenetic variation can arise under the influence of stress. This can enable a trans-generational stress-response of organisms which can have a positive influence on the phenotype and (natural) selection on either the (enhanced level of) transgenerational phenotypic plasticity or the epigenetic variation itself, potentially influencing natural selection and ultimately evolution. Where genetic variation can be characterized as hard-inheritance, the inheritance of epigenetic variation is often referred to as 'soft-inheritance' due to the lower transgenerational stability and resetting that occurs in the intergenerational transfer of epigenetic variation. Epigenetic variation is also often dependent on, or a downstream consequence, of genetic variation, suggesting that it is (in part) determined by genetic variation.

Mechanistic studies in model species have contributed greatly to the understanding of the molecular mechanisms that control the dynamics of different epigenetic marks present in multicellular organisms. In plants, studies in the model plant *Arabidopsis thaliana* have resulted in deciphering the most important molecular mechanisms and actors, giving an ever-increasing insight into the dynamics of epigenetic regulation of cells and organisms. A key feature of model systems is the ability to 'switch' off certain genes or molecular pathways, for instance via the experimental use of mutants, enabling the study of their role in the heritability of epigenetic marks. DNA methylation is a well-studied epigenetic mark, which has shown high stability even in transgenerational experiments.

From the perspective of studying epigenetic variation, plants are particularly interesting for several reasons, most importantly: 1) The separation between soma and germline, the Weismann barrier, is less strict in plants compared to other eukaryotes, as in higher plants

germline cells are formed during floral development from somatic cells (which can occur throughout the life of the plant), whereas in most eukaryotes germline cell development is restricted to a defined point (early) in the organismal development. 2) The sessile nature of plants makes an adaptive plastic response to changing environments an important feature, a plant cannot just walk away when the going gets tough. 3) The transgenerational stability of DNA methylation is higher in plants compared to other eukaryotes such as mammals, in which epigenetic information is erased during germline reprogramming. These factors combined suggest that the potential importance of epigenetic variation in plants might be high.

In this thesis, I focus on studying DNA methylation in apomictic Dandelions, applying Next Generation Sequencing (NGS) approaches to the study of this non-model plant species. Apomictic dandelions produce seeds that are genetically identical to the 'mother' plant, which makes it easier to study the influence of epigenetic variation without confounding effects of genetic variation. Working with Next Generation Sequencing data is still relatively new and therefore not always optimized for specific types of analysis. I discovered a distinct error pattern in RNAseq data that indicated an artificial source of variation that could be traced back to the way the RNAseq libraries were constructed. The first publication of this thesis contains a technical analysis of such artefacts present in RNAseq data, suggesting that these errors are related to random hexamer mispriming during library construction (Chapter 2).

The main goal of my work is to better understand the role of epigenetic variation in adaptation and plasticity of plants. This role remains poorly understood. This is in part due to the lack of high-resolution techniques that allow for the detailed study of epigenetic marks such as DNA methylation in non-model organisms. Existing techniques for measuring DNA methylation such as methylation sensitive AFLPs offer only information on DNA methylation variation in an anonymous and limited fashion. The plummeting costs of sequencing techniques have enabled large-scale genotyping efforts (focusing on genetic variation only) for a wide variety of non-model organisms. Here, I extended this popular genotyping by sequencing technique, to allow for sequencing-based epigenotyping or epiGBS (chapter 3), which allows for measuring DNA methylation and genetic variation in hundreds of samples simultaneously. I have extensively validated the approach, providing evidence that with the right design, the accuracy of the DNA methylation measurements with epiGBS are as high as those with the gold standard Whole Genome Bisulfite Sequencing.

An important aim of my PhD research was to investigate the stability of (stress induced) DNA methylation variation in apomictic dandelions and the potential of phenotypic variation underpinned by DNA methylation variation to be subjected to selection. I therefore studied the transgenerational stability of both stress induced and natural DNA methylation variation in different genotypes of apomictic dandelions in a six-generation experiment, comparing DNA methylation patterns between generations and tracking changes in them (chapter 4) using epiGBS. I found clear but limited evidence for environmental induction of heritable DNA methylation changes after application of Jasmonic Acid. Furthermore, I found a significant negative relation between the similarity of

DNA methylation patterns and intergenerational distance, indicating epigenetic divergence over generations. I conclude that DNA methylation in both CG and CHG (where H can be any nucleotide except for G) sequence context are heritable and that environmental perturbation can result in heritable DNA methylation changes which are however not widespread.

A prerequisite for epigenetic variation to contribute to adaptation is that epigenetic variants that affect the phenotype are heritable. To test whether an epigenetics-based selection response is possible, at least over the time course of a few generations, I selected early flowering for two subsequent generations in three genotypes of apomictic dandelions. This selection effort included lines that received a stress pre-treatment with either Jasmonic Acid or 5-azacytidine, to determine if stress-induced DNA methylation variation would increase the capacity to respond to selection. The selection experiment on flowering time (chapter 5) resulted in a shift in flowering time for all treatments in a young apomict, suggesting that natural and heritable epigenetic variation can underpin quantitative traits such as flowering time. I also found evidence for treatment induced (epi)genetic variation leading to a stronger selection response in one out of 3 genotypes. This suggests that stress-induced heritable epigenetic variation can lead to a selection response. Further study is however required to rule out genetic variants and to study the long-term stability of the variation selected upon.

Finally, in the General Discussion I summarize the findings, putting them in context with recently published studies. I reflect on the state of the field of ecological epigenetics and in what sense the epiGBS technique that I developed and other emerging techniques can contribute to a better understanding of the role of epigenetic variation in ecology and evolution. I reflect on the place of epiGBS compared to other techniques. I point out that with the growing evidence of the inadequacy and misinterpretation of MS-AFLP results a systematic review of such results by replicating the studies employing sequencing based techniques such as epiGBS instead of MS-AFLP is in order.

Samenvatting

Het succes of blijven voortbestaan van soorten vereist continue aanpassing aan veranderende milieumomstandigheden om te overleven en nageslacht te produceren. Selectie vindt plaats op fenotypische eigenschappen, welke worden bepaald door een complex samenspel van genetische en epigenetische variatie en omgevingsfactoren. (Natuurlijke) selectie leidt tot 'survival of the fittest' die het beste zijn aangepast aan hun lokale omgeving. Alleen deze goed aangepaste individuen kunnen nageslacht doorgeven aan de volgende generatie. Kennis over de mechanismen waarmee epigenetische en genetische variatie kunnen worden gegenereerd en doorgegeven bepalen ons inzicht in erfelijkheid en evolutie. Tot nu toe is de mechanistische kennis van de moleculaire genetica bepalend geweest voor de wetenschappelijk inzichten in erfelijkheid en evolutie, uitmondend in het gen-centristische paradigma van het Neo-darwinisme. Nieuwe studies laten echter zien dat naaste genetische variatie (in het DNA), epigenetische variatie ook erfelijk kan zijn. Verdere studies naar de eigenschappen en de transgenerationale dynamiek van epigenetische variatie is nodig om ons begrip van erfelijkheid en evolutie te verbreden en verdiepen.

Epigenetische variatie heeft, vergeleken met genetische variatie, andere eigenschappen en vertoont een andere transgenerationale dynamiek. Epigenetische variatie helpt onder andere bij de regulatie van genexpressie en het bepalen van de verschillende typen cellen in eukaryoten. De belangrijkste functie van DNA methylering, welke een belangrijk onderdeel vormt van de ons bekende epigenetische variatie, is het onderdrukken van mobiele genetische elementen zoals transposons, die het genoom kunnen ontregelen door zichzelf veelvuldig te kopiëren. Daarnaast bepaald epigenetische variatie de vorm en functie van verschillende celtypen in meercellige organismen (zoals bijvoorbeeld een haar of zenuwcel). Een onderscheidend kenmerk van epigenetische variatie ten opzichte van genetische variatie is dat onder invloed van stress (specifieke) epigenetische veranderingen kunnen optreden. Dit maakt een transgenerationale stress-respons mogelijk die mogelijk adaptief is. Daarnaast kan er (natuurlijke) selectie plaatsvinden op de mate waarin stress-geïnduceerde epigenetische variatie wordt doorgegeven of op de doorgegeven variatie zelf in het nageslacht, wat invloed kan hebben op natuurlijke selectie en uiteindelijk evolutie. Waar genetische variatie veelal onveranderd wordt doorgegeven tussen generaties is dit bij epigenetische variatie niet het geval; veel nieuw gegenereerde variatie valt weer terug. Epigenetische erfelijkheid wordt daarom ook wel 'softe' erfelijkheid genoemd. Epigenetische variatie is ook vaak afhankelijk van of direct bepaald door genetische variatie.

Mechanistische studies in modelsoorten hebben een belangrijke bijdrage geleverd aan ons begrip van de moleculaire mechanismen die de dynamiek en erfelijkheid van verschillende typen van epigenetische variatie bepalen in meercellige organismen. Studies in de modelplant *Arabidopsis thaliana* hebben inzicht gegeven in de belangrijkste moleculaire mechanismen, welke leiden tot een steeds ingewikkelder plaatje van de dynamische epigenetische regulatie van cel tot systemisch niveau. Een belangrijke eigenschap van modelsystemen is dat het mogelijk is om bepaalde genen of moleculaire pathways uit te schakelen met mutanten, wat de studie naar hun rol in de erfelijkheid van epigenetische variatie mogelijk maakt. DNA methylering

is een van de best-bestudeerde epigenetische merkers, die een hoge mate van stabiliteit heeft laten zien in transgenerationale experimenten.

Voor de studie van epigenetische variatie zijn planten bijzonder geschikt voor een aantal redenen, waaronder: 1) De scheiding tussen soma en kiembaan (cellen betrokken bij productie van zaad en eicellen), de Weismann barrière, is minder strikt in planten vergeleken met andere eukaryoten, omdat in hogere planten de geslachtscellen worden gevormd uit somatische cellen gedurende de bloemontwikkeling (die gedurende het hele leven van de plant kan plaatsvinden). In andere eukaryoten is de ontwikkeling van cellen in de kiembaan beperkt tot een gedefinieerde en vaak korte periode in de ontwikkeling. 2) De sessiele levenswijze van planten stelt hoge eisen aan hun adaptief vermogen; een plant kan niet weglopen als de omstandigheden verslechteren. 3) De transgenerationale stabiliteit van DNA methylatie is hoog in planten vergeleken met andere eukaryoten zoals zoogdieren, waarbij gedurende de ontwikkeling het merendeel van de epigenetische merkers wordt gereset. Gecombineerd, suggereren deze factoren dat het potentiële belang van erfelijke epigenetische variatie in planten hoog zou kunnen zijn.

In dit proefschrift focus ik op de studie van DNA methylatie in apomictie paardenbloemen; waarbij ik Next Generation Sequencing (NGS) methoden toepas en ontwikkel die geschikt zijn voor zowel model als niet-model planten. Apomictie paardenbloemen produceren zaad dat genetisch identiek is aan dat van de moederplant, wat het makkelijk maakt om de rol van epigenetische variatie in isolatie te bestuderen, omdat genetische variatie in principe geen rol speelt. Het werken met Next Generation Sequencing (NGS) data is nog relatief nieuw en daardoor nog niet altijd geoptimaliseerd voor specifieke analyses. Zo heb ik een specifiek foutpatroon in RNAseq data ontdekt dat beruiste op de manier waarop de sequencing-libraries worden gemaakt. De eerste publicatie in dit proefschrift (hoofdstuk 2) omvat een technische analyse van dergelijke artefacten in RNAseq data, waarbij ik suggereer dat deze fouten worden veroorzaakt door het misprimen van random hexameren gedurende de library constructie.

Het hoofddoel van mijn werk in dit proefschrift is om de rol van epigenetische variatie in de plasticiteit en adaptatie van planten beter te leren begrijpen. Van deze rol is nog niet veel bekend. Dit komt onder andere door een gebrek aan geschikte technieken waarmee epigenetische merkers zoals DNA methylatie in detail kunnen worden bestudeerd in niet-model organismen waarin een rol voor deze variatie ecologisch voor de hand ligt. Bestaande technieken zoals methylatie sensitieve AFLPs geven slechts gelimiteerde en 'anonieme' informatie waarbij de kennis over de genomische locatie van de aanwezige variatie ontbreekt. De snel dalende kosten van sequencing-data hebben geleid tot een snelle adoptie van genotyperingstechnieken die het mogelijk maken op grote schaal genetische variatie te meten in niet-model organismen. In dit proefschrift beschrijf ik de ontwikkeling van een aanpassing op de populaire genotyping by sequencing techniek die het mogelijk maakt om DNA methylatie en genetische variatie gelijktijdig te meten. Deze epigenotyping by sequencing techniek (epiGBS hoofdstuk 3) maakt het mogelijk om in honderden samples voor soorten zonder referentiegenoom bisulfiet-sequencing data te genereren die inzicht verschaft in zowel genetische als DNA-methylatie (epigenetische) variatie. Ik heb epiGBS uitgebreid gevalideerd en

getoetst aan bestaande methodes. Met de juiste experimentele opzet is de nauwkeurigheid van de bepaling van DNA-methylatie net zo groot als die van de huidige standaard, Whole Genome bisulfite sequencing (WGBS).

Een ander belangrijk doel van mijn promotieonderzoek was het bestuderen van de stabiliteit van (stress-geïnduceerde) DNA methylatievariatie in apomictie paardenbloemen en het potentieel van het uitvoeren van (artificiële) selectie op fenotypische variatie gebaseerd op DNA-methylatie variatie. Ik heb hiervoor de transgenerationale stabiliteit van zowel stress-geïnduceerde als natuurlijke DNA methylatievariatie bepaald in een zes-generatie experiment, waarbij ik de veranderingen in DNA methylatie in opvolgende generaties bepaalde met behulp van epiGBS. Ik heb duidelijke aanwijzingen voor subtiele veranderingen in DNA methylatie na toepassing van jasmonzuur. Hiernaast vond ik een significante negatieve relatie tussen de similariteit van DNA methylatiepatronen en intergenerationale afstand, wat duidt op epigenetische divergentie over generaties. Ik concludeer dat DNA methylatie in zowel CG als CHG (waar H elke nucleotide kan zijn behalve G) sequence context erfelijk is; en dat milieu-invloeden kunnen resulteren in (erfelijke) veranderingen in DNA methylatie, hoewel de schaal waarop deze variatie transgeneratieel wordt doorgegeven relatief beperkt is.

Een vereiste voor een mogelijke bijdrage van epigenetische variatie aan adaptatie is dat epigenetische variatie die een afwijkend fenotype bepaald erfelijk is. Om te bepalen of het mogelijk is te selecteren op basis van epigenetische variatie, voor tenminste enkele generaties, heb ik snel-bloeiende paardenbloemen voor twee opeenvolgende generaties geselecteerd in drie verschillende genotypes van apomictie paardenbloemen. Dit selectie-experiment bevatte lijnen die een stress voorbehandeling met jasmonzuur ofwel 5-azacytidine hadden ontvangen, om te bepalen of stress-geïnduceerde DNA methylatievariatie de selectierespons zou verhogen. Het selectie-experiment op bloeitijd (hoofdstuk 5) resulteerde in een verschuiving in bloeitijd voor zowel controle als stress-voorbehandelingslijnen, wat suggereert dat natuurlijke, erfelijke epigenetische variatie kwantitatieve eigenschappen zoals bloeitijd kan beïnvloeden. Daarnaast vond ik in een van de genotypes een sterkere selectierespons in een lijn met stress-voorbehandeling ten opzichte van de controle-lijn. Dit suggereert dat stress-geïnduceerde, erfelijke epigenetische variatie aan de basis kan liggen voor (artificiële) selectie op een veranderd fenotype. Diepgravende studies zijn echter nodig om genetische variatie uit te sluiten en om inzicht te verkrijgen in de stabiliteit van de variatie waarop geselecteerd is.

Tot slot vat ik mijn bevindingen samen in de general discussion, waarbij ik mijn resultaten aan de hand van de recente literatuur bediscussieer en in context plaats. Ik reflecteer op de status van het vakgebied van de ecologische epigenetica en in welke mate de epiGBS techniek die ik ontwikkeld heb en andere opkomende technieken kunnen bijdragen aan een beter begrip van de rol van epigenetische variatie in een ecologische en evolutionaire context. Ik reflecteer op de mogelijke inzet van epiGBS naast of in plaats van andere technieken. Verder beschrijf ik dat in het licht van de groeiende bewijslast die de ontoereikendheid en misinterpretatie van MS-AFLP resultaten aantoont een systematische review en revisie van deze resultaten middels het repliceren van (enkele van) deze studies met op sequencing gebaseerde technieken zoals

epiGBS nodig is om de relevantie van de tot nu toe opgedane kennis in de ecologische epigenetica te staven.

Acknowledgements / dankwoord

De kans om dit promotieonderzoek te mogen doen en met zo veel vrijheid en ambitie te kunnen werken aan het verbeteren van methodes en deze toepassen was een unieke kans.

Ik liet een bestaande baan achter om me in dit avontuur te storten. Koen, je hebt in je hele wetenschappelijke carrière keihard gewerkt om je ideeën te realiseren en je groep te starten. Ik ben trots en gelukkig daar onderdeel van te mogen zijn geweest.

Mijn begeleiders, Koen, Arjen en Wim wil ik heel erg bedanken voor het vertrouwen en de ondersteuning om mijn PhD op mijn manier invulling te geven. Na elke meeting waarin we de plannen of analysemogelijkheden bespraken kon ik weer vol goede moed aan de gang, de sfeer en het onderlinge vertrouwen waren groot. Het epiGBS project, wat technisch erg ambitieus was en ook nog eens relatief ver afstond van jullie expertise, heeft verreweg de meeste tijd en energie gekost gedurende mijn promotieonderzoek. De uitkomst, het publiceren van de techniek in Nature Methods, was allerm minst gegarandeerd en vormde een groot risico. Dit hebben jullie met mij aan willen gaan. Ik ben blij en trots dat dit gelukt is. Na 'de' publicatie lagen er echter nog drie hoofdstukken om te schrijven, wat ik grotendeels vanuit Bolivia gedaan heb. Jullie snelle feedback en uitgebreide tijdsinvestering in skype-gesprekken hebben ertoe bijgedragen dat ik deze hoofdstukken relatief snel heb kunnen afwerken.

Niels, het realiseren van deze thesis was nooit gelukt zonder jou. Vanaf 2011 ben je bijna fulltime betrokken geweest bij het ontwikkelen van de epiGBS techniek. Het was en is 'ons' project, waarin we beide al onze tijd en energie stopten in de ontwikkeling. Je realisatie dat door een type adapter te gebruiken aan beide kanten van de DNA-fragmenten we alleen maar hairpins kregen in ons product, was Vos et al al in 1995 hadden beschreven heeft dit project gered. Vanaf toen zijn we op de goede weg gekomen. Volgens mij ben ik meer in Nijmegen op het lab geweest dan in Wageningen, de eerste succesvolle epiGBS libraries hebben we daar gemaakt. Bedankt voor je geduld en begrip over mijn veeleisendheid bij het optimaliseren van het protocol (minder PCR-stappen! ☺). Je praktische kennis en ervaring zijn enorm belangrijk gebleken bij het optimaliseren van het protocol. Philippine en Joop, jullie betrokkenheid bij het project was groot, bedankt voor al jullie input en het beschikbaar houden/stellen van Niels op dit project. In the initial and final phases of the epiGBS project numerous colleagues have helped me. First and foremost, thanks Rob Elshire for giving such a great workshop on genotyping by sequencing in 2011, it gave me the thorough understanding needed for developing epiGBS afterwards. It was a lot of fun to work on the barcoded adapter generator and to continue discussing the bioinformatics and library preparation steps. In writing up the paper and preparing it for Nature methods, I got some good advice from Bob Schmitz, Christina Richards, Christoph Bock, Claude Becker. Thanks to Elio Schijlen, Gregory Maes, Jeroen van Houdt and Baseclear for helping me sequence the many test-libraries that we needed for developing the epiGBS protocol. Thanks to Frank Johannes, Claude Becker, Luc de Meester, Christina Richards, Philippine Vergeer and Nunhems for providing samples for the epiGBS paper.

The Dandelion team, Koen, Veronica, Carla and Julie, thanks for providing such a warm and supportive and fun environment! Veronica, we started our PhD almost at the same time, we had such a great time on the dandelion collection trip across Europe! You were always in for listening to my laborious explanations on the technical details of the epiGBS protocol. But I also remember the pizza and movie nights, some dinners at droevendaal and the camaraderie in the office that we shared for almost 5 years! Carla, bedankt voor de ondersteuning bij alle technische werk met DNA-isolaties en library-preps en de lekker appeltaart die je zo regelmatig meenam als er iets te vieren was. July, I enjoyed working wit and next to you. Thanks for all the dinners, get-togethers and inviting us to your wedding in France, I had a great time.

To all my colleagues at the NIOO, thanks for providing such a great atmosphere and inspiration. Louise, you are the leader that shows by example that every organization could hope for. I have learned a lot from your enthusiasm, drive and effort and believing that you can achieve ambitious goals and change the world together! Thanks to my PhD colleagues at terrestrial ecology Sabrina, Taia, Roeland, Stijn, Mingang, Moniek, Jasper, Minghui, Marta, Rutger, Viola and Sigrid for sharing the tough but rewarding journey. Also, a everyone else in the department Tanja, Freddy, Slavica, Carolin, Roel, Ciska V and Ciska R, Martine, Olga, Elly, Kadri, Stefan, Basten, Joke, Kelly, Gerard, Jeff, Pella, Gera, Stefan, Martijn, thanks for providing a great atmosphere. Petra, bedankt voor je input en bijdrage in het onderzoek of het zinvol zou zijn om de epiGBS-methode te patenteren. Jeff, thanks for the crazy discussions, wild ideas and unconditional support and believe in my abilities.

Als fanatiek klimmer is de klimhal in Arnhem jarenlang mijn tweede thuis geweest. Jaap, Evert, Ineke, Carlo, Nori, Noël, Sanne, Taimoor en Jasper: bedankt voor alle support en kameraadschap in de afgelopen jaren in en buiten de klimhal. Jaap, je hebt mij altijd op de juiste manier weten te inspireren en motiveren om doelen te bereiken, zowel op klimvlak als in mijn werk. Jaap, Ineke en Evert, bedankt voor de "Nature Matters" foto, dit betekende veel voor mij. Niels en Aukje, jullie kennen mij vanaf mijn jeugd, ik bedankt jullie voor alle momenten die we de afgelopen jaren gedeeld hebben en de steun en begrip voor het proces waarin ik in de PhD doorheen ging.

Mijn familie wil ik bedanken voor de steun en het geduld. Mam, als je me die DVD van 'the ghost in your genes' niet had gegeven in 2007 was deze thesis er niet gekomen. Het was een eer en uitdaging om bij jou in de klas de scholieren iets te vertellen over de epigenetica. Ook hebben we vaak gesproken over nieuwe ontwikkelingen in de Biologie. Ymke, Mirte en Tijmen: bedankt voor jullie geduld en luisterend oor als het aan de keukentafel weer eens over Biologie en wetenschap ging. Pieter, je was in Wageningen dichtbij, wat mooi om je door de week in de winkel te zien en wat leuk dat je zo vaak bij ons kwam eten. Je interesse en warme betrokkenheid betekenen veel.

A mi familia en Bolivia a la que conocí el 2009, un abrazo enorme y unas gracias por su apolo en todo. Lola y Jaime, queridos suegros, gracias para cuidar nos tan bien mientras estamos viviendo como vecinos en el departamento en Cochabamba. Siempre han tratado de traer nos todo, cocinar cosas ricas y compartir la vida en familia. Jaimito, Juancio, Carla y Quesia, gracias

Acknowledgements / dankwoord

para compartir su tiempo y ideas escuchando lo que estaba haciendo. Un abrazo y un saludo a todos los tíos, tías, primos y amigos que eran y son parte de nuestra vida en Bolivia y el mundo. Su apolo y amistad nos han llevado adelante siempre.

Als laatste wil ik graag mijn Vrouw bedanken. Tatiana, mi amor, ik had dit traject nooit zonder jouw kunnen afleggen. Je hebt me door dik en dun gesteund en met met heel veel zaken willen helpen. In de kas, met het vullen van potjes, het nakijken van figuren, het doorlezen van manuscripten, of zorgen dat ik aan niets anders dan mijn PhD hoefde te werken. Je geduld is danig op de proef gesteld. Bedankt hiervoor!

Nu weet ik zeker dat ik een aantal mensen ben vergeten te noemen, mijn excuses en alsnog hartelijk bedankt!

About the author

About the Author



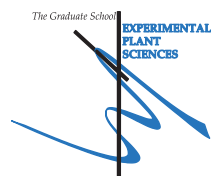
Thomas van Gulp was born on the 24th of February 1983 in Wageningen, the Netherlands. He grew up in Oosterbeek and went to high school in Arnhem. During his youth, he was part of the Dutch Youth organization for nature study (NJN), where he met many aspiring biologists. After starting the first year of forest and nature conservation at Larenstein in Velp he decided to switch to studying Biology in Wageningen where he accomplished his Masters's degree in 2009 specializing in ecological genetics. After finishing his internship studying retrotransposons in *Drosophila* in France, he took a gap year to become chairman of the student sport council in 2007/2008. For his

Master thesis, supervised by Dr. Freek Bakker, he investigated how methylation sensitive AFLP compares to bisulfite sequencing *in silico* using data of *Arabidopsis thaliana*. After graduating, he spent 3 months at the Netherlands Institute of Ecology (NIOO-KNAW) in Heteren to write an article together with his later thesis supervisor Koen Verhoeven. Afterwards, he worked as a junior researcher in the plant breeding group of Wageningen University with Dr. Olga Scholten and Dr. Ben Vosman working on SNP discovery in Leek and Onion for one year. In 2011, he started his PhD-research at the Netherlands Institute of Ecology (NIOO-KNAW) supervised by Dr. Koen Verhoeven, Dr. Arjen Biere and Prof. Wim van de Putten. He investigated the heritability of (stress induced) DNA methylation variation in apomictic dandelions and his findings are presented in this thesis. Besides his PhD, he works as a bioinformatics consultant for his own company Deena Bioinformatics since 2011 for 1-2 days per week. Since July 2017 Thomas works as a Bioinformatics Scientist at Naktuinbouw (the Netherlands Inspection Service for Horticulture).

Peer-reviewed publications

- Verhoeven, K. J. F., & **van Gurp, T. P.** (2012). Transgenerational Effects of Stress Exposure on Offspring Phenotypes in Apomictic Dandelion. *PLoS One*, 7(6), e38605. <http://doi.org/10.1371/journal.pone.0038605>
- Shahin, A., **van Gurp, T.**, Peters, S. A., Visser, R. G., van Tuyl, J. M., & Arens, P. (2012). SNP markers retrieval for a non-model species: a practical approach. *BMC Research Notes*, 5(1), 79.
- van Gurp, T. P.**, McIntyre, L. M., & Verhoeven, K. J. F. (2013). Consistent errors in first strand cDNA due to random hexamer mispriming. *PLoS One*, 8(12), e85583. <http://doi.org/10.1371/journal.pone.0085583>
- van Gurp, T. P.**, Wagemaker, N. C. A. M., Wouters, B., Vergeer, P., Ouborg, J. N. J., & Verhoeven, K. J. F. (2016). epiGBS: reference-free reduced representation bisulfite sequencing. *Nature Methods*. <http://doi.org/10.1038/nmeth.3763>
- de Carvalho, J. F., de Jager, V., **van Gurp, T. P.**, Wagemaker, N. C. A. M., & Verhoeven, K. J. F. (2016). Recent and dynamic transposable elements contribute to genomic divergence underasexuality. *BMC Genomics*, 1–12. <http://doi.org/10.1186/s12864-016-3234-9>
- Richards, C. L., Alonso, C., Becker, C., Bossdorf, O., Bucher, E., **van Gurp, T. P.**, Colomé-Tatché, M., et al. (2017). Ecological plant epigenetics: Evidence from model and non-model species, and the way forward. <http://doi.org/10.1101/130708>

Education Statement of the Graduate School
Experimental Plant Sciences



Issued to: Thomas van Gorp
Date: 9 October 2017
Group: Terrestrial Ecology, NIOO-KNAW and Nematology, WUR
University: Wageningen University & Research

1) Start-up phase	<u>date</u>
<ul style="list-style-type: none"> ▶ First presentation of your project <i>Title:</i> Natural epigenetic variation in apomictic dandelion lineages ▶ Writing or rewriting a project proposal ▶ Writing a review or book chapter ▶ MSc courses ▶ Laboratory use of isotopes 	08 Dec 2011
<i>Subtotal Start-up Phase</i>	<i>1.5 credits*</i>

2) Scientific Exposure	<u>date</u>
<ul style="list-style-type: none"> ▶ EPS PhD student days Annual EPS PhD Student Day, Utrecht University Annual EPS PhD Student Day 'Get2Gether 2017', Soest, NL ▶ EPS theme symposia EPS Theme 4 Symposium 'Genome Biology', Nijmegen, NL EPS Theme 2 symposium 'Interactions between Plants and Biotic Agents', Wageningen, NL ▶ National meetings (e.g. Lunteren) and other National Platforms National Ecogenomics day, Wageningen, NL Annual meeting 'Experimental Plant Sciences', Lunteren, NL ▶ Seminars (series), workshops and symposia Clive Jones - On the purpose, meaning and usage of the physocial ecosystem engineering concept QIAGEN's Epigenetics and Pyrosequencing Seminar Nicole M. Gerardo - Using Insect Systems to Study How Organisms Interact with Beneficial and Harmful Microbes Judith Becerra - Bursera plants and their herbivores: the ecological theatre and the evolutionary play Joris M. Koene - Sex and conflict in simultaneous hermaphrodites Seminar Prof. dr. Rosie Bradshaw 'The genome of Dothiostroma septosporum, a close relative of Cladosporium fulvum; what have we learnt so far?' William Ratcliff - Experimental Evolution of Multicellularity Javier Palatnik - 'Biogenesis and function of plant microRNAs' Marcel Visser - Seasonal timing in a warming world Angus Buckling - Bacteria-phage evolutionary ecology: lab, wild and applications Anne Yoder - Madagascar: a case study for integrating population genetic and GIS methods for guiding conservation policies Frans Vera - Shifting images by shifting baselines NBIC genome browser seminar WEES seminar Patrizia d'Ettorre "Recognition of social identity in ants: pheromones & signature mixtures" Prof. Nicole van Dam - Multiple-stress management: what can we learn from plants? Julie Ferreira - developing genomic resources for the hybrid, polyploid & invasive spartina Phillipa Griffin - Population genetics and phylogeography of Arabidopsis lyrata Dr Björn Podola - The algal leaf: immobilized cultivation of microalgae using a new photobioreactor concept WEES seminar Marten Scheffer - tipping points in complex systems NIOO-Science days 2013 6th ecogenomics day Utrecht university NIOO-Science days 2015 Oxford nanopore sequencing seminar ▶ Seminar plus ▶ International symposia and congresses ESF-EMBO Symposium: Epigenetics in Context: From Ecology to Evolution, Sant Feliu de Guixols, Spain Plant Epigenetics, Stress and Evolution - Suzhou Dushu Lake Conference Center China workshop sEpiDiv, Leipzig, Germany Evolution 2015 Brazil 	01 Jun 2010 09-10 Feb 2017 14 Dec 2012 23 Jan 2017 07 Jun 2012 22-23 Apr 2013 19 May 2011 30 Jun 2011 21 Apr 2011 26 May 2011 23 Jun 2011 04 Aug 2011 26 Aug 2011 25 Aug 2011 22 Sep 2011 20 Oct 2011 10 Nov 2011 12 Dec 2011 19 Dec 2011 20 Sep 2012 20 Dec 2012 16 Jan 2013 17 Jan 2013 24 Jan 2013 23 May 2013 11-12 Nov 2013 24 Jun 2014 11-12 Nov 2015 14 Apr 2016 18-23 Sep 2011 29 Oct-02 Nov 2012 02-06 Mar 2015 26-30 Jun 2015

Education certificate

<ul style="list-style-type: none"> ▶ Presentations <i>Talk:</i> Genotyping by sequencing presentation - Institute wide NIOO-KNAW <i>Poster:</i> ESF-EMBO Symposium Barcelona, Spain <i>Talk:</i> De wereld leert door, NL, national TV <i>Poster:</i> NIOO-days entire institute <i>Talk:</i> Workshop sEpiDiv Leipzig, Germany <i>Talk:</i> Evolution conference 2015 Brazil <i>Talk:</i> Frank Johannes group, Groningen <i>Talk:</i> epiGBS workshop for Christina Richards and Students <i>Talk:</i> CIPY GBS Workshop, Wageningen ▶ IAB interview ▶ Excursions 	<ul style="list-style-type: none"> 01 Jul 2011 18-23 Sep 2011 22 Jan 2013 11-12 Nov 2013 02 Mar 2015 30 Jun 2015 16 Oct 2015 Jul 2015 13 Dec 2015
--	--

Subtotal Scientific Exposure 21.3 credits*

<p>3) In-Depth Studies</p> <ul style="list-style-type: none"> ▶ EPS courses or other PhD courses Next Generation Sequencing data analysis Bioinformatics, a user's approach Cornell course Genotyping By Sequencing ▶ Journal club Member of literature discussion group at Ecological epigenetics group ▶ Individual research training 	<p style="text-align: center;"><u>date</u></p> <ul style="list-style-type: none"> 17-19 Mar 2010 03 Aug-03 Sep 2010 06-09 Jun 2011 2012-2014
--	--

Subtotal In-Depth Studies 5.9 credits*

<p>4) Personal development</p> <ul style="list-style-type: none"> ▶ Skill training courses Organising and supervising Msc theses Wisconsin Entrepreneurial Bootcamp Wageningen Entrepreneurial bootcamp ▶ Organisation of PhD students day, course or conference ▶ Membership of Board, Committee or PhD council 	<p style="text-align: center;"><u>date</u></p> <ul style="list-style-type: none"> 30 Jan & 01 Feb 2011 19-24 Jun 2011 13-18 Nov 2011
--	---

Subtotal Personal Development 4.2 credits*

TOTAL NUMBER OF CREDIT POINTS*	32.9
---------------------------------------	-------------

Herewith the Graduate School declares that the PhD candidate has complied with the educational requirements set by the Educational Committee of EPS which comprises of a minimum total of 30 ECTS credits

* A credit represents a normative study load of 28 hours of study.

Colophon

Colophon

The research presented in this thesis was conducted at the Department of Terrestrial ecology at the Netherlands Institute of Ecology (NIOO-KNAW) in Wageningen.

The research described in this thesis was financially supported by the NWO-ALW grants: 820.01.025 from the Netherlands Organization for Scientific Research.

This is NIOO thesis 146.

Figure Cover (background): Istockphotos ID 587200764, creator: traffic_analyzer

Figure back cover: iTZiT BV Gouda

All other text & figures: T.P. van Gulp (thomasvangulp@gmail.com)

Printed by: Digiforce | Proefschriftmaken.nl

Propositions

1. Understanding the relevance of epigenetic variation in ecology and evolution requires studying this in a variety of species in natural settings, and methods that are capable of measuring quantitative DNA methylation levels at high resolution in non-model species.
(this thesis)
2. Genetically independent DNA methylation variation in apomictic dandelions can determine phenotypic variation upon which selection may act.
(this thesis)
3. Aspects of our health and even behavior are determined by our microbiome which can change due to environmental factors and be transmitted transgenerationally, making this an extra-cellular inheritance system whose properties are similar to epigenetic variation.
4. The use of genetically modified mosquitos in fighting global pandemics such as ZIKA, leads to lower environmental impact and higher efficacy than chemical control of mosquitos.
5. As future generations of students will design their own curriculum based on the widely available and high-quality on-line courses from international top universities, the existence of local universities will be threatened.
6. As the skills required to be a successful scientific researcher are increasingly indistinguishable from the skills required to be successful in industry, there will be an increased brain drain of highly skilled scientists from academia to industry
7. Ideas, and not viruses, are the most contagious parasites that, once they take hold of the brain are almost impossible to eradicate.

Propositions belonging to the thesis, entitled:

Measuring stress-induced DNA methylation in apomictic Dandelions

Thomas Pieter van Gurp
Wageningen, 09 October 2017