Exploring the Role of DNA Methylation in the Early Stages of Polyploidy with a Focus on Computational Reproducibility

---

Dissertation

zur

Erlangung der naturwissenschaftlichen Doktorwürde (Dr. sc. nat.)

vorgelegt der

Mathematisch-naturwissenschaftlichen Fakultät

der

Universität Zürich

von

Stefan Milosavljevic

von

Lugano TI und aus Serbien

Promotionskommission

Prof. Dr. Kentaro Shimizu (Vorsitz)

Dr. Rie Shimizu-Inatsugi (Leitung der Dissertation)

Prof. Dr. Mark Robinson (Leitung der Dissertation)

Prof. Dr. Andreas Wagner

Dr. Jun Sese

Zürich, 2022

To all of my friends, family and the love of my life

Summary

Polyploidization, also known as whole genome duplication (WGD), is an event that occurred in all domains of the tree of life and its prevalence is particularly pronounced in the evolutionary history of land plants. To understand how such events could become so prevalent, extensive studies investigated the effect of polyploidization in plants at different time scales. With this large body of research, two main findings emerged. First, a hypothesis stated that there might be no paradigm for polyploidy, meaning that different species respond differently to polyploidy. In parallel, to ultimately support or oppose this hypothesis, major gaps in current research need to be investigated to provide a more complete understanding of polyploidy.

Here, I focused on one of these gaps, namely the genomic effects right after the formation of a polyploid, specifically DNA methylation. Since large scale expression changes have been a common element in newly formed polyploids across different species, epigenetic changes were considered a good candidate as the underlying mechanism behind rapid expression control. Few studies on DNA methylation in early stages of polyploidy investigated whole genome patterns, but lack computational reproducibility and do not take into account important factors such as environmental stress.

In this thesis, I reassessed the role of DNA methylation after polyploid formation by combining reproducibility with an extensive experimental design including different environmental conditions and transcriptomic data. To set the grounds for reproducibility in polyploid studies, I developed ARPEGGIO, a polyploid-specific workflow to compare methylation patterns between two groups using whole genome bisulfite sequencing data. Next, I applied ARPEGGIO to real data from the allopolyploid *Arabidopsis kamchatica*, a species resulting from the crossing of its two progenitor species *A. halleri* and *A. lyrata*. Synthetic and natural *A. kamchatica* lines, together with their progenitor species, were grown in two conditions for several generations and their DNA methylation and expression pattern was analyzed.

Results revealed for the first time distinct trajectories of DNA methylation and expression changes in newly formed polyploids with respect to natural and progenitor species and the crucial role of environment in directing all of these changes. Novel polyploids showed diverging DNA methylation patterns from their progenitors and converging patterns towards their natural counterpart. The diverging changes were different for each condition, highlighting the significant influence of environment. Additionally, methylation changes were linked to functional expression changes related to polyploidy and environment.

Table of contents

General introduction

a. Polyploidy framework and evolutionary context

Polyploidy is defined as a state in which a cell or an organism possesses more than two sets of chromosomes. Polyploidization in species, also known as whole genome duplication (WGD), is an event leading to the formation of an organism with more than two sets of chromosomes. Polyploid species are divided into autopolyploids, resulting from the genome doubling of an individual species, and allopolyploids, resulting from the merging of two genomes from two different species by hybridization (1). This thesis will focus on allopolyploids and several important terms related to them need to be defined (Figure 1). The species of origin of an allopolyploid will be referred to as progenitors. The genome of an allopolyploid that originated from one of the two progenitors is defined as a subgenome (Figure 1a). Homoeolog genes are genes that had a common origin, diverged after speciation and were brought back together in the same genome by allopolyploidization (Figure 1b) (2).

From an evolutionary point of view, polyploidy was first considered a “dead-end” (3). More specifically, the ability of polyploid species to persist on the long term was limited because of their limited genetic potential. Assuming a single polyploidization event, polyploids were considered too genetically uniform and for a mutation in a gene to be fixated, the time required would be longer compared to diploids given the higher amount of alleles (in autopolyploids) or the presence of a homoeolog (in allopolyploids) (4). Given these assumptions, it followed that polyploidy did not possess the basis to be a driver of diversification. These views from the 1950s persisted in the scientific community until the early 2000s, where advances in genetic, genomic and computational tools led to a complete paradigm turnaround (4).

Polyploidy is now seen as a successful and major evolutionary force (5). New key findings led to this paradigm shift, namely the frequency of polyploidy across the tree of life, the discovery and association of ancient WGD events to diversification events and most importantly the extremely dynamic nature of polyploid genomes. Each one of these aspects will be developed in the next sections.

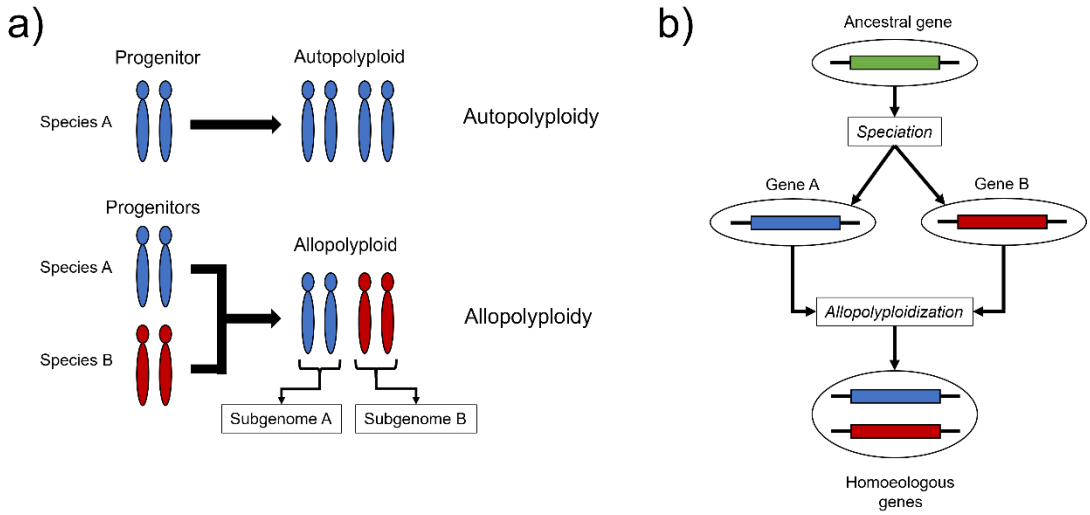


Figure 1: schematic view of two types of polyploidy and homoeologous genes with important terminology. Autopolyploidy is a process in which one progenitor species duplicates its own set of chromosomes to generate a new autopolyploid species (a, top). Allopolyploidy is a process in which two different progenitor species cross and provide their full set of chromosomes to generate a new allopolyploid species (a, bottom). Homoeologous genes are the two genes originating from the same ancestral gene, diverged by speciation and reunited under the same organism with allopolyploidization (b).

b. Polyploidy and its influence on the tree of life

Evidence of WGD was found in *bacteria*, *archea* and *eukarya*, meaning that WGD touched all domains of life (6). Research in both *bacteria* and *archea* is still in its infancy, with scattered reports highlighting a variety of effects associated to WGD, mostly autopolyploidization, but limited evolutionary insights (7).

In eukaryotes, polyploidization has been reported in a multitude of lineages. These include fungi, oomycetes, chordates, nematodes, arthropods and land plants (6). This prevalence is so pronounced that rather than claiming that WGD is rare in a given lineage, it would be more cautious to state that such a lineage has not been studied enough (6).

The pervasiveness of WGD in eukaryotes is also accompanied by links to several key innovations in different species and taxonomic groups. One example comes from fungi, more specifically *Saccharomyces cerevisiae* (fermenting yeast), an important organism in the food industry for its ability to ferment sugars and producing ethanol, and one of the first eukaryotes with its entire genome sequenced (8). An ancient WGD event from ~100Mya was suggested to be the origin of duplicated genes in *S. cerevisiae* (9). This discovery had two implications. First, since some of the duplicated genes are related to the fermentation process, WGD may have been a major force contributing to this innovation that is unique to *Saccharomyces* compared to other yeasts (9). Additionally, *S. cerevisiae*, known to have haploid and diploid forms (10), was a “degenerate tetraploid”. Further analyses into this WGD event were able to classify it as an allopolyploid event (11). Similarly to yeast, studies in vertebrates discovered

two WGD events (1R and 2R) associated to their origin (12) and in fishes, two additional ancient WGD events (3R and 4R) were associated to innovations such as electric organs and a heart adapted to aquatic life (13).

Among all eukaryotic lineages, the most extensive studies on WGD come from land plants. There are several reasons behind this large body of research. First and foremost, most crop plants are polyploids (14). Some well-known examples among allopolyploid crops include hexaploid wheat (*Triticum aestivum*), canola (*Brassica napus*), tobacco (*Nicotiana tabacum*), peanut (*Arachis hypogaea*) and cotton (*Gossypium hirsutum*). In autopolyploid crops, we find watermelon (*Citrullus lanatus*), strawberries (*Fragaria x ananassa*), potato (*Solanumm tuberosum*) and alfalfa (*Medicago sativa*). Considering the economic and social importance of crops (15), research in polyploidy has the opportunity to improve crop production for some great future challenges such as climate change (16–18). The second aspect is related to the importance and frequency of polyploidization in the evolutionary history of land plants. It has been established that two ancient WGD events, also called paleopolyploidization events, happened right before the diversification of seed plants and angiosperms, suggesting an association between polyploidy and major innovations that lead to the success and spread of seed and flowering plants (19). In terms of frequency, data from over a thousand plant species and a robust phylogenomic approach revealed over 244 ancient inferred WGDs in the evolutionary history of green plants (20). These WGDs were found at least once in the ancestry of each analyzed plant lineage with very few exceptions (20).

c. Investigating genomic long- and short-term effects of allopolyploidy in land plants

Polyploid systems were studied on a variety of topics at different levels such as ecology, physiology, population genetics, molecular biology and genomics (21). Here we will focus on genomics, where a considerable number of studies investigated the effects of allopolyploidy on a variety of genera such as *Triticum* (22,23), *Gossypium* (24,25), *Arabidopsis* (26–28), *Brassica* (29,30) and others (21,31–33). With this extensive community effort, several important processes were discovered at different stages of polyploid evolution (21,33–35). These stages are commonly grouped into two time frames that are not well defined, one covering the period right after polyploid formation (short-term) and the other being much further in time (long-term). For the purpose of this study, we will define short-term as the period between 1 and 100 generations after polyploid formation, usually characterized by genomic instability, the medium-term as the period between 100 and 100’000 generations,

characterized by stabilization of the genome and establishment, and the long-term as a period over 100'000 generations characterized by genome downsizing (Figure 2).

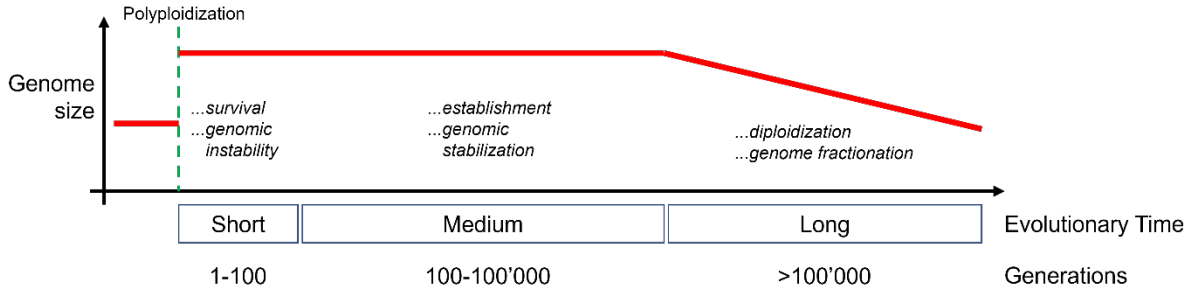


Figure 2: overview of genome size over evolutionary time grouped into three different time frames after polyploid formation. Polyploidization leads to doubling of genome size and short-term evolutionary time refers to the early generations right after polyploidization, here defined as the period between 1 and 100 generations. This time frame is characterized by the survival of the newly formed polyploid and genomic instability. Medium-term here is defined as the period between 100 and 100'000 generations characterized by the polyploid's establishment and genomic stabilization. Long term is the period after 100'000 generations characterized by the decrease in genome size towards diploidization driven by mechanisms such as genome fractionation.

Short term processes are characterized by transcriptomic changes, epigenomic changes (DNA methylation, histone modifications and small RNA), and genomic changes (chromosome rearrangements and movement of transposable elements) (21,33–35). Newly formed polyploids undergo a sudden increase in gene copies and their expression pattern was studied extensively (6,21,36–41). Although there are differences between species, all recently formed polyploids showed some deviations in expression with respect to their progenitor species (42). These deviations can lead to expression patterns in the polyploid resembling one of the two progenitors in a process known as expression-level dominance, observed in *Spartina* and *Mimulus* (43,44), or in other cases the expression patterns showing higher or lower levels compared to the progenitors in a gene-by-gene case, as seen in *Arabidopsis* and *Triticum* (42). To better understand the underlying reasons behind these transcriptomic changes, epigenetic changes were investigated as a candidate mechanism to regulate expression. Few studies explored small RNA and histone modifications in newly formed polyploids (45–47). In the case of small RNA, novel expression patterns in synthetic *A. suecica* of small interfering RNA (siRNA) were linked to chromatin and genome stability while microRNA (miRNA) expression was associated to unequal gene expression between subgenomes (45). Histone modifications on the other hand were associated to flowering time variation between *A. suecica* and its progenitors by mediating expression of flowering genes (38). Both these results suggest a role for these epigenetic mechanisms in regulating expression to some extent, but their role in other polyploid systems is unexplored. DNA methylation on the other hand, has been extensively studied in terms of species and was found to be essential after polyploid formation (48). This study will focus on this epigenetic trait (see next section) and its link to expression changes in the short-term. Since such short-term

changes can affect polyploids in the long-term, the resulting picture would have a more comprehensive time frame to better understand the reasons behind the evolutionary success of polyploidy. An excellent example for this is a study on the allopolyploid *Mimulus peregrinus*, resulting from the crossing between its diploid progenitors *M. guttatus* and *M. luteus*. Researchers compared the DNA methylation pattern of a newly formed *M. peregrinus* to its two progenitors. DNA methylation changes were found and associated to a bias in homoeolog expression favoring the expression of genes from the *M. luteus* subgenome (44). Since this bias persisted over generations and was found to be strongest in natural *M. peregrinus*, there might be evidence of short-term DNA methylation changes driving expression changes that lead a polyploid's establishment and success (44). As for genomic changes, homoeologous exchanges (HEs) in polyploids are also a known mechanism where genetic material is exchanged between chromosomes of the diploid progenitors (49). This mechanism has been reported in several crop species such as *Brassica*, peanut, coffee and bread wheat (6). HEs are not only limited to the short-term and can occur for thousands of generations after polyploid formation (50). These changes can also be evolutionarily relevant. For example in *Brassica*, researchers were also able to associate HEs to phenotypic changes related to disease resistance and yield (51). Another relevant genomic change right after polyploidization is the movement of TEs, which has been studied in *Mimulus*, *Arabidopsis*, *Spartina*, *Nicotiana* and others (52). In most species, polyploidization didn't lead to an immediate 'burst' in TE movement, suggesting that this mechanism might not be a common phenomenon, but it can still contribute to gene expression divergence (53) and changes in DNA methylation that can repress expression of genes (54).

Even though the number of combinations of short-term effects is large, the long-term outcome for many polyploids is a consistent genome downsizing, also known as diploidization (33). The goal of studies on long-term effects of polyploidy is to understand the dynamics driving the process of genome downsizing. We divided these studies into two groups. In small-scale studies, given the sudden increase in gene copies after polyploidization, the general interest is often the fate of genes, while large-scale studies look at the genomics of polyploid populations to investigate genetic diversity, selection processes, population dynamics and structure, and try to define the most important evolutionary forces affecting the genome. At smaller scales, one of the best known processes is the uneven loss of genes after polyploidization, also known as biased genome fractionation (33,55). In paleohexaploid *Brassica rapa*, biased genome fractionation was observed on each of its three subgenomes and the difference between these fractionation rates was large enough to support a specific polyploid-formation model for *B. rapa* (56). Additional evidence of genome fractionation bias occurring, comes from maize, where loss of homeologs was found 2.3 times more frequently on one parental genome compared to the other (57). Biased genome fractionation can favor

one subgenome over the other and lead to subgenome dominance (6). This event was linked to known short term phenomena such as biased homeolog expression, expression level dominance and mobilization of transposable elements (TE) (33). At larger scales, studies require good quality genomic resources and large amounts of data (58), and as a consequence many of the studies come from crop species with breeding improvements as main driver (59). For example, with the publication of a chromosome-level assembly of wheat (60), researchers were able to explore its historical expansion and uncover several genetic mechanisms leading to its improvement and success (61). Similarly in cotton, after generating assemblies from five different species, researchers reported differences in their genomic diversification patterns that pointed to candidate genes for crop improvement (62). Nevertheless, population studies on non-crop species exist as well, such as *Capsella bursa-pastoris* (63), *Arabidopsis suecica* (64) and *Arabidopsis kamchatica* (26). Newly formed (natural) polyploid populations undergo a strong population bottleneck, meaning that the populations are small and with a limited gene pool, and high self-fertilization rates (65), leading to less genetic diversity and reducing the efficacy of natural selection. Researchers explored different aspects related to this common starting point. For example in *Capsella*, patterns of gene silencing and gene mutation after polyploid formation were strongly influenced by selection and expression patterns from the progenitor species, emphasizing a strong parental legacy (63). In *A. suecica* researchers focused on the rate and extent of change in the genome and epigenome after polyploidization. Since they couldn't find major changes, they suggested that polyploidization undergoes a gradual evolution process with selection being the major bottleneck (64). Finally in *A. kamchatica*, analyses of genome-wide diversity and selection patterns found high similarity between subgenomes but pairs of homoeologs showed low correlation between them, suggesting independent evolution (26). Further examination in other species is required to provide a more general picture of the evolutionary forces accompanying polyploidy in the long-term.

Considering the variety of genomic effects of polyploidy on different species leading to similar developments, Soltis and colleagues suggested that no paradigm might exist for polyploids (21). Instead, the differences observed should be attributed to species having different solutions to the same problem. To test this claim, the available knowledge about polyploid evolution is too scattered and diversified to be compared effectively. Because of this, Soltis and colleagues advocated for some major initiatives to focus on, to set a common goal and better unify research findings in the long term (21). Here, we will focus on one of the major community goals set by Soltis and colleagues, namely early stages of allopolyploid evolution and the study of recent and synthetic allopolyploid systems. As mentioned above, we will address this from an epigenetic point of view, specifically DNA methylation.

d. DNA methylation in synthetic allopolyploid systems

The role of DNA methylation was found to be essential for the survival of polyploids in their early generations. A milestone study on the synthetic allopolyploid *Arabidopsis suecica*, derived from a crossing between *A. thaliana* and *A. arenosa*, was able to highlight how essential DNA methylation was for the survival and reproduction of newly formed polyploids (48). In this study, seeds from both progenitor species, synthetic and natural *A. suecica* were treated with azadC to remove all DNA methylation. While seeds from progenitor species and natural *A. suecica* were able to grow and reproduce, plants from treated synthetic seeds consistently produced abnormal phenotypes and were not able to produce offspring (48). Additionally, increased transcriptional changes were reported in these demethylated synthetic plants, suggesting a link between DNA methylation and regulation of expression.

Further investigations on DNA methylation in newly formed polyploids were based on Methylation Sensitive Amplification Polymorphism (MSAP), providing an estimate of DNA methylation changes in newly formed polyploids, emphasizing the variation across species, but with very limited functional implications (Table 1). In short, MSAP takes advantage of two methylation-sensitive enzymes *HspII* and *MspI*, both cutting at CCGG sites along the genome. To obtain fragments with reasonable lengths, these enzymes are combined with *EcoRI* and their product is analysed through electrophoresis (66). The resulting bands would be counted and compared to a control sample to assess differences in DNA methylation. In a synthetic allopolyploid of *Aegilops sharonensis* x *T. monococcum ssp aegilopoides*, 11.3% of methylation changes were observed, compared to its diploid progenitors. The majority of these changes were the result of hybridization (6.9%), with some being associated to polyploidization as well (4.3%). Bands showing differential methylation were sequenced and analyzed, leading to 3 out of 12 bands showing similarity to repetitive DNA sequences and retrotransposons (67). Additionally, methylation changes occurred more frequently on one progenitor side compared to the other, but the sample size was low (67). In the case of synthetic *Brassica napus*, obtained by crossing *B. rapa* and *B. oleraceae*, an average of 6.84% of MSAP bands showed differential methylation over three different tissues: flowers and leaves in juvenile and mature plants (68). Similarly to *Aegilops/Triticum*, methylation changes occurred more often on one progenitor side (68). The highest amount of DNA methylation change was reported in synthetic *Spartina anglica*, a cross between *S. maritima* and *S. alterniflora*, with 28.6% via MSAP. Compared to other systems, methylation alteration was similar between subgenomes and sequencing didn't reveal any known gene (32). In synthetic *Senecio cambrensis*, originating from *S. squalidus* and *S. vulgaris*, 13.4% of methylation change were reported (69). In synthetic *A. arenosa* 8.3% of bands showed differential

methylation and sequencing was able to find only one gene structurally similar to retroelements (48).

MSAP represents a reproducible yet limited method to look at DNA methylation, particularly for plants. This method was originally developed to analyse CpG methylation in mammals, since CG is the most prevalent site for methylation in this lineage (70). For plants, there are two additional contexts that are methylated, CHG and CHH context (where H = A, T or C). This separation of contexts comes from studies in *A. thaliana* and *Zea mays* on the CHROMOMETHYLASES (CMTs) enzyme family, which found various enzymes of this class (and others) independently maintaining different contexts (71). Proportionally, CHH is the context with the highest count, followed by CHG and last CG. The average methylation level follows an opposite trend, with CG showing the highest average methylation, followed by CHG and last CHH (72). Taking into account the limitations associated with MSAP, it becomes clear that conclusions on DNA methylation changes through this method cannot be extended to the whole genome level. As an example, the *A. thaliana* genome includes ~21mio cytosines, out of which ~2.7mio (~12%) in CG context, ~3mio (~14%) in CHG context and ~15.3mio (74%) in CHH context. In (48) the total number of MSAP products analyzed was 623, meaning 1246 cytosines, 0.00006% of all the cytosines in the genome or roughly 0.0005% of all cytosines in CG context.

Few studies explored DNA methylation changes in synthetic polyploids at the whole genome level, even though the technology is available (see next section) and interesting functional implications were reported (Table 1). In synthetic *Arabidopsis suecica*, compared to its progenitors, most of the methylation changes occurred in CG context together with gene expression changes for genes associated to reproduction (39). These genes were hypothesized to be responsible for reproductive stability both in the short- and long-term evolution of *A. suecica*. In synthetic *Mimulus peregrinus*, a significant genome-wide change in CHH methylation was observed and was associated with the control of transposable elements (TEs) (44). Together with these methylation changes, subgenome expression level dominance was found and since TEs showed different methylation dynamics between subgenomes, they were presumed to contribute to expression dominance, possibly affecting long-term patterns of gene retention and loss.

Table 1: overview of DNA methylation studies in synthetic allopolyploid systems. For each studied genus, the method used to look at DNA methylation is specified, the amount of change reported, whether uneven DNA methylation changes happened between subgenomes, which tissue was used, the genetic targets of the methylation changes and reference articles.

Genus	Method	Allopolyploid DNA methylation change	Uneven sub-genome methylation changes	Tissue	Genetic targets	Reference
Arabidopsis	MSAP	8.3%	-	-	Various genes, one structurally similar to retroelements	(48)
	WGBS	Not reported	Yes	Young leaves	Reproduction-associated genes	(39)
Brassica	MSAP	6.84 - 9%	Yes	Young and mature leaves, flowers	Various genes, many with similarity to retroelements	(68,73)
Aegilops/Triticum	MSAP	11.3%	Yes (low sample size)	Young leaves	Repetitive DNA (retrotransposons) and low-copy DNA	(67)
Spartina	MSAP	28.6%	No	Leaves	Not recognizable	(32,74)
Senecio	MSAP	13.4%	-	Mature flower buds	-	(69)
Gossypium	MSAP	None	None	Young leaves	None	(75)
Mimulus	WGBS	-	Yes	-	Transposable elements	(44)

e. Technological and methodological advances in allopolyploid studies

In two decades, the cost of sequencing a genome has decreased dramatically. When looking at the human genome, the cost of sequencing went from $100'000'000 in 2001 to around $1000 in 2021 (76). This price drop has made read sequencing more accessible and affordable to scientists. Such progress was made possible with the help of increasingly powerful sequencing technologies, such as short and long read sequencing instruments. The former produce reads with a length between 75 and 250bp, with the advantage of high output (up to 6'000Gb), high accuracy and relatively low cost per base, while the latter produce reads between 10kb and >1000kb, with the advantage of resolving repeat sequences, detecting large scale genome rearrangements and easing the task of overlapping and merging sequences (77). For polyploid genomics, both of these technologies could offer a considerable amount of data, but obtaining the data represents just a first tep.

All the technological advances in sequencing caused a shift in terms of effort from data production to data analysis, representing the new bottleneck. When looking at genome assemblies, specifically in plants, the amount of genomes published every year has been increasing steadily (78), with the wheat genome assembly being one of the most recent and notable achievements in terms of complexity (60). Although other polyploid genome assemblies were generated together with wheat in the past years, their proportion remains low: out of 1031 plant genomes, only 62 were polyploids (78). This difference underlines the methodological challenges related to polyploidy, such as the increase in complexity caused by the increase in repeat content, TEs, contraction and expansion of gene families and other genomic rearrangements (79). As long-read sequencing technologies and algorithmic advances could help make polyploid genome assemblies easier, an alternative common solution is to generate assemblies from the (diploid) progenitor species instead (79). With this solution, the challenge of assigning polyploid reads to the correct sub-genome still remains. Several tools were developed to tackle this problem such as HyLiTE (80), HANDS (81) and its successor HANDS2 (82), PolyCat (83), PolyDog (84), SNiPloid (85), Homeoroq (86) and its successor EAGLE-RC (87). These tools can be grouped in two approaches: SNP-based and read classification-based. SNP-based approaches, as their name suggests, assign polyploid reads to a sub-genome based on the presence of single-nucleotide polymorphisms (SNPs). Such information can be already available for model-species with good genomic resources, but if that's not the case, some methods infer SNPs based on progenitor data aligned to their assembly. SNP-based tools include HyLiTE, HANDS, HANDS2, SNiPloid and PolyCat. Without SNP information or progenitor data, SNP-based methods cannot be used.

Read classification methods overcome these limitations by mapping polyploid reads to progenitors' assemblies and classifying reads to either assemblies (or none) based on different criteria applied to read mapping information. Homeoroq classifies reads based on the number of mismatches for a mapped read, while its successor EAGLE-RC computes a likelihood of a read given a reference genotype hypothesis that is used to assign reads to a subgenome. PolyDog applies four criteria in succession to classify reads: whether a read was mapped, the read mapping quality score, alignment length and the number of perfect matches in the alignment. An evaluation of these tools on transcriptomic data from cotton, focusing on homoeolog expression and co-expression analyses, revealed Poly-Cat and EAGLE-RC as tools with best overall performance (24).

To explore DNA methylation via sequencing in plants, two groups of methods exist, bisulfite-based and non-bisulfite-based methods (88). The latter are less popular and include methylated DNA immunoprecipitation (MeDIP)-seq, an approach where DNA is sheared, sonicated and precipitated with an antibody specific for 5-methylcytidine (89), Methyl-CpG binding domain protein sequencing (MBD-seq), with an approach similar to MeDIP-seq where instead of an antibody, a capture protein is used to enrich and precipitate DNA (90) and TET-assisted pyridine borane sequencing (TAPS), where DNA is treated with a borane reaction that converts methylated cytosines into thymines before being sequenced (91). The former, also known as bisulfite sequencing methods, are all based on bisulfite conversion. This chemical reaction is divided in three steps: DNA denaturation, bisulfite treatment and polymerase chain reaction (PCR). In the DNA denaturation step, the two strands are separated, then bisulfite treatment converts unmethylated cytosines into thymines, leaving methylated cytosines the same, and finally PCR amplifies the converted DNA to obtain material for library preparation and sequencing (92). There are two categories of bisulfite sequencing (BS): reduced representation (RRBS), which targets specific regions of the genome, and whole genome (WGBS), which represents the gold standard to explore single-nucleotide level methylation over the whole genome (92).

A lot of progress has been made for polyploid-specific data analysis tools, but all of them focus on RNA-seq data, while no polyploid-specific tools for whole genome DNA methylation data exist, even though a lot of sequencing technologies are available. In the case of WGBS, which will be used in this study, the data analysis process already requires a set of specific software and many steps to be analyzed without taking polyploidy into consideration (93).

f. The *Arabidopsis kamchatica* study system to explore the environmental impact on DNA methylation and expression in early stages of allopolyploidy

To investigate DNA methylation changes right after polyploidization, we used *Arabidopsis kamchatica* as a study system. The allotetraploid *Arabidopsis kamchatica* (2*n* = 4*x*) is a species resulting from the crossing between *A. halleri* (2*n* = 2*x*) and *A. lyrata* (2*n* = 2*x*) (94). *A. kamchatica* can be found in nature and its distribution spans East Asia (Far East Russia, China, Korea, Taiwan and Japan) and North America (Alaska, Canada and Northwestern United States) (95). Two subspecies were identified based on habitat, life history and morphology: *A. kamchatica* subsp. *kamchatica* (perennial) and *A. kamchatica* subsp. *kawasakiana* (annual).

Several aspects make *A. kamchatica* an ideal system for this project. First, its natural distribution is broad, both overlapping with the niche of both progenitors and extending beyond them, suggesting a successful adaptation to a variety of environments. For other non-crops allopolyploid study species, the natural distribution can be limited. *Arabidopsis suecica* is found only in Northern Fennoscandia (64) while *Mimulus peregrinus* has been only reported in Scotland (96). Second, the progenitor species are known and both have available genome assemblies (26,97). Third, compared to other polyploids with large genomes such as wheat (17Gb), *A. kamchatica* has a relatively small genome size of about 475Mb (26). Another key aspect of this species is the ability to generate synthetic individuals (Figure 3). By crossing a female *A. halleri* to a male *A. lyrata*, F1 hybrid seeds can be obtained. After germinating and growing the seeds to the stage where cotyledons are fully developed, the shoot apical meristem can spontaneously undergo autopolyploidization or it be treated with a colchicine solution to induce autopolyploidization. With this process, F1 hybrids can produce seeds that can be germinated and grown to form first generation synthetic *A. kamchatica* individuals (86). With the ability to synthesize *A. kamchatica*, the transcriptomic and epigenomic background from the progenitors is known and can be used to explore changes in the early stages right after polyploidization.

With *A. kamchatica*'s broad environmental tolerance and the ability to generate synthetic individuals, the effect of environmental stress on both expression and methylation in a newly formed polyploid can be investigated for the first time. As seen in the previous sections, studies on DNA methylation in early stages of polyploidy suffer from a technological gap, but another major limitation is the exclusion of environmental stress. Studying synthetic polyploids in controlled conditions allows to attribute most of the changes observed to polyploidization *per se*, but this system is far from natural stressful conditions. This limitation might be critical

considering the strong link between polyploidy and stress (98). For example some ancient WGD events were found to occur in synchrony with major extinction events or in general periods with extreme global change (99,100). Another more recent example is the ability of polyploid species to adapt to dry and cold environments (101,102). Both these examples support the polyploidy and stress relationship indirectly. Few studies addressed a direct link in an experimental fashion and from a genomic point of view, but only at small scale (looking at few genes) or on established polyploids (103,104). It follows that a major gap exists in characterizing the effect of environmental stress on newly formed polyploids, particularly on both DNA methylation and expression at the whole-genome level.

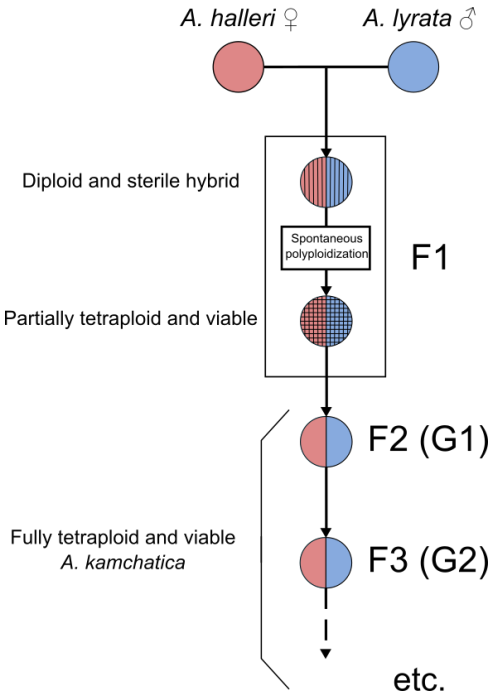


Figure 3: schematic overview of the steps required to generate a synthetic individual of A. kamchatica. First, pollen from A. lyrata is used to fertilize A. halleri, producing a diploid and sterile hybrid. This hybrid undergoes spontaneous polyploidization to produce a partially tetraploid individual that is viable. Viable seeds from this individual are fully tetraploid and are used to grow synthetic A. kamchatica individuals that can be propagated for several generations. Since the treated diploid hybrid represents the first filial generation (F1), the first viable A. kamchatica is considered F2, its offspring F3 and so on. Since F2 is the first fully polyploid generation from which our experiments start, to avoid confusion we named this generation G1, its offspring G2 and so on.

g. Goals of the thesis

This thesis has three main goals: 1) developing a reproducible polyploid-specific pipeline focusing on DNA methylation data, 2) applying this pipeline on real data to explore the relationship between DNA methylation patterns in early stages of polyploidy and environment, and 3) using expression data to assess the effect of DNA methylation in different conditions.

For the first goal, in Chapter 1 I developed, tested and published an Automated Reproducible Polyploid EpiGenetic GuIdance workflOw (ARPEGGIO) (105). The goal of ARPEGGIO was to offer a full set of analyses for researchers working with polyploid WGBS data, starting from raw data and obtaining a list of differentially methylated genes, requiring a minimal amount of well-documented steps. Additionally, ARPEGGIO's containerization and software management system were included to ensure reproducibility.

For the second goal, in Chapter 2 I investigated two novel aspects: the effect of environment on DNA methylation patterns in synthetic *A. kamchatica* and the trajectory of both patterns with respect to the progenitors and two natural *A. kamchatica* lines. This study showed for the first time how synthetic polyploids are able to respond rapidly and differently when exposed to different stresses for several generations. In addition we showed how DNA methylation patterns appear to converge between synthetic and natural species, highlighting the potential of this epigenetic mechanism to be a driver of adaptation.

For the third goal, in Chapter 3 I complemented the findings on DNA methylation with expression data from the same samples to assess the consistency between the two and their relationship with environmental stress. Results showed partial consistency between DNA methylation and expression patterns but the effect of environmental stress was reconfirmed. The amount of DNA methylation changes in genes could explain only a part of the large amount of expression changes observed. Finally, we suggested some candidate genes showing a consistent change both in expression and DNA methylation that might be important for both polyploidy- and stress-related response.

References

1. Chen ZJ. Molecular mechanisms of polyploidy and hybrid vigor. Trends Plant Sci [Internet]. 2010 Feb;15(2):57–71. Available from: https://linkinghub.elsevier.com/retrieve/pii/S136013850900301X
2. Glover NM, Redestig H, Dessimoz C. Homoeologs: What Are They and How Do We Infer Them? Trends Plant Sci [Internet]. 2016 Jul;21(7):609–21. Available from: https://linkinghub.elsevier.com/retrieve/pii/S1360138516000595
3. Stebbins GL. Chromosomal evolution in higher plants. Edward Arnold Ltd., London.; 1971. viii + 216 pp.
4. Soltis DE, Visger CJ, Soltis PS. The polyploidy revolution then...and now: Stebbins revisited. Am J Bot [Internet]. 2014 Jul 1;101(7):1057–78. Available from: http://doi.wiley.com/10.3732/ajb.1400178
5. Fox DT, Soltis DE, Soltis PS, Ashman T-L, Van de Peer Y. Polyploidy: A Biological Force From Cells to Ecosystems. Trends Cell Biol [Internet]. 2020 Sep;30(9):688–94. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0962892420301240
6. Blischak PD, Mabry ME, Conant GC, Pires JC. Integrating Networks, Phylogenomics, and Population Genomics for the Study of Polyploidy. Annu Rev Ecol Evol Syst [Internet]. 2018 Nov 2;49(1):253–78. Available from: https://www.annualreviews.org/doi/10.1146/annurev-ecolsys-121415-032302
7. Soppa J. Polyploidy in Archaea and Bacteria: About Desiccation Resistance, Giant Cell Size, Long-Term Survival, Enforcement by a Eukaryotic Host and Additional Aspects. J Mol Microbiol Biotechnol [Internet]. 2014;24(5–6):409–19. Available from: https://www.karger.com/Article/FullText/368855
8. Albertin W, Marullo P. Polyploidy in fungi: evolution after whole-genome duplication. Proc R Soc B Biol Sci [Internet]. 2012 Jul 7;279(1738):2497–509. Available from: https://royalsocietypublishing.org/doi/10.1098/rspb.2012.0434
9. Wolfe KH, Shields DC. Molecular evidence for an ancient duplication of the entire yeast genome. Nature [Internet]. 1997 Jun;387(6634):708–13. Available from: http://www.nature.com/articles/42711
10. Herskowitz I. Life cycle of the budding yeast Saccharomyces cerevisiae. Microbiol Rev. 1988;52(4):536–53.
11. Marcet-Houben M, Gabaldón T. Beyond the Whole-Genome Duplication: Phylogenetic Evidence for an Ancient Interspecies Hybridization in the Baker's Yeast Lineage. Hurst LD, editor. PLOS Biol [Internet]. 2015 Aug 7;13(8):e1002220. Available from: https://dx.plos.org/10.1371/journal.pbio.1002220
12. Van de Peer Y, Maere S, Meyer A. The evolutionary significance of ancient genome duplications. Nat Rev Genet [Internet]. 2009 Oct 4;10(10):725–32. Available from: http://www.nature.com/articles/nrg2600
13. Moriyama Y, Koshiba-Takeuchi K. Significance of whole-genome duplications on the emergence of evolutionary novelties. Brief Funct Genomics [Internet]. 2018 Sep 27;17(5):329–38. Available from: https://academic.oup.com/bfg/article/17/5/329/4951518
14. Udall JA, Wendel JF. Polyploidy and Crop Improvement. Crop Sci [Internet]. 2006 Nov;46(S1). Available from: https://onlinelibrary.wiley.com/doi/10.2135/cropsci2006.07.0489tpg
15. Crop Prospects and Food Situation #1, March 2021 [Internet]. FAO; 2021. Available from: http://www.fao.org/3/cb3672en/cb3672en.pdf
16. Levin DA. Plant speciation in the age of climate change. Ann Bot [Internet]. 2019 Nov 15;124(5):769–75. Available from: https://academic.oup.com/aob/article/124/5/769/5524583
17. Gao J. Dominant plant speciation types. A commentary on: 'Plant speciation in the age of climate change.' Ann Bot [Internet]. 2019 Nov 15;124(5):iv–vi. Available from: https://academic.oup.com/aob/article/124/5/iv/5626011
18. Moura RF, Queiroga D, Vilela E, Moraes AP. Polyploidy and high environmental tolerance increase the invasive success of plants. J Plant Res [Internet]. 2021 Jan 5;134(1):105–14. Available from: http://link.springer.com/10.1007/s10265-020-01236-6
19. Jiao Y, Wickett NJ, Ayyampalayam S, Chanderbali AS, Landherr L, Ralph PE, et al. Ancestral polyploidy in seed plants and angiosperms. Nature [Internet]. 2011 May 10;473(7345):97–100.

Available from: http://www.nature.com/articles/nature09916
20. One thousand plant transcriptomes and the phylogenomics of green plants. Nature [Internet]. 2019 Oct 23;574(7780):679–85. Available from: http://www.nature.com/articles/s41586-019-1693-2
21. Soltis DE, Visger CJ, Marchant DB, Soltis PS. Polyploidy: Pitfalls and paths to a paradigm. Am J Bot [Internet]. 2016 Jul;103(7):1146–66. Available from: http://doi.wiley.com/10.3732/ajb.1500501
22. Chagué V, Just J, Mestiri I, Balzergue S, Tanguy A-M, Huneau C, et al. Genome-wide gene expression changes in genetically stable synthetic and natural wheat allohexaploids. New Phytol [Internet]. 2010 Sep;187(4):1181–94. Available from: http://doi.wiley.com/10.1111/j.1469-8137.2010.03339.x
23. Ramírez-González RH, Borrill P, Lang D, Harrington SA, Brinton J, Venturini L, et al. The transcriptional landscape of polyploid wheat. Science (80-) [Internet]. 2018 Aug 17;361(6403):eaar6089. Available from: https://www.sciencemag.org/lookup/doi/10.1126/science.aar6089
24. Hu G, Grover CE, Arick MA, Liu M, Peterson DG, Wendel JF. Homoeologous gene expression and co-expression network analyses and evolutionary inference in allopolyploids. Brief Bioinform [Internet]. 2020 Mar 27; Available from: https://academic.oup.com/bib/advance-article/doi/10.1093/bib/bbaa035/5811916
25. Yoo M-J, Szadkowski E, Wendel JF. Homoeolog expression bias and expression level dominance in allopolyploid cotton. Heredity (Edinb) [Internet]. 2013 Feb 21;110(2):171–80. Available from: http://www.nature.com/articles/hdy201294
26. Paape T, Briskine R V., Halstead-Nussloch G, Lischer HEL, Shimizu-Inatsugi R, Hatakeyama M, et al. Patterns of polymorphism and selection in the subgenomes of the allopolyploid Arabidopsis kamchatica. Nat Commun [Internet]. 2018 Dec 25;9(1):3909. Available from: http://www.nature.com/articles/s41467-018-06108-1
27. Wang J, Tian L, Madlung A, Lee H-S, Chen M, Lee JJ, et al. Stochastic and Epigenetic Changes of Gene Expression in Arabidopsis Polyploids. Genetics [Internet]. 2004 Aug;167(4):1961–73. Available from: http://www.genetics.org/lookup/doi/10.1534/genetics.104.027896
28. Bomblies K, Madlung A. Polyploidy in the Arabidopsis genus. Chromosom Res [Internet]. 2014 Jun 1;22(2):117–34. Available from: http://link.springer.com/10.1007/s10577-014-9416-x
29. Baker RL, Yarkhunova Y, Vidal K, Ewers BE, Weinig C. Polyploidy and the relationship between leaf structure and function: implications for correlated evolution of anatomy, morphology, and physiology in Brassica. BMC Plant Biol [Internet]. 2017 Dec 5;17(1):3. Available from: https://bmcplantbiol.biomedcentral.com/articles/10.1186/s12870-016-0957-3
30. Liu S, Liu Y, Yang X, Tong C, Edwards D, Parkin IAP, et al. The Brassica oleracea genome reveals the asymmetrical evolution of polyploid genomes. Nat Commun [Internet]. 2014 Sep 23;5(1):3930. Available from: http://www.nature.com/articles/ncomms4930
31. Soltis DE, Buggs RJA, Barbazuk WB, Chamala S, Chester M, Gallagher JP, et al. The Early Stages of Polyploidy: Rapid and Repeated Evolution in Tragopogon. In: Polyploidy and Genome Evolution [Internet]. Berlin, Heidelberg: Springer Berlin Heidelberg; 2012. p. 271–92. Available from: http://link.springer.com/10.1007/978-3-642-31442-1_14
32. SALMON A, AINOUCHE ML, WENDEL JF. Genetic and epigenetic consequences of recent hybridization and polyploidy in Spartina (Poaceae). Mol Ecol [Internet]. 2005 Mar 16;14(4):1163–75. Available from: http://doi.wiley.com/10.1111/j.1365-294X.2005.02488.x
33. Wendel JF, Lisch D, Hu G, Mason AS. The long and short of doubling down: polyploidy, epigenetics, and the temporal dynamics of genome fractionation. Curr Opin Genet Dev [Internet]. 2018 Apr;49:1–7. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0959437X17301557
34. Cheng F, Wu J, Cai X, Liang J, Freeling M, Wang X. Gene retention, fractionation and subgenome differences in polyploid plants. Nat Plants [Internet]. 2018 May 30;4(5):258–68. Available from: http://www.nature.com/articles/s41477-018-0136-7
35. Nieto Feliner G, Casacuberta J, Wendel JF. Genomics of Evolutionary Novelty in Hybrids and Polyploids. Front Genet [Internet]. 2020 Jul 28;11. Available from: https://www.frontiersin.org/article/10.3389/fgene.2020.00792/full
36. Doyle JJ, Flagel LE, Paterson AH, Rapp RA, Soltis DE, Soltis PS, et al. Evolutionary Genetics of Genome Merger and Doubling in Plants. Annu Rev Genet [Internet]. 2008 Dec 1;42(1):443–61. Available from: https://www.annualreviews.org/doi/10.1146/annurev.genet.42.110807.091524

37. Adams KL, Percifield R, Wendel JF. Organ-Specific Silencing of Duplicated Genes in a Newly Synthesized Cotton Allotetraploid. Genetics [Internet]. 2004 Dec 1;168(4):2217–26. Available from: https://academic.oup.com/genetics/article/168/4/2217/6059384
38. Wang J, Tian L, Lee H-S, Wei NE, Jiang H, Watson B, et al. Genomewide Nonadditive Gene Regulation in Arabidopsis Allotetraploids. Genetics [Internet]. 2006 Jan 1;172(1):507–17. Available from: https://academic.oup.com/genetics/article/172/1/507/6065206
39. Jiang X, Song Q, Ye W, Chen ZJ. Concerted genomic and epigenomic changes accompany stabilization of Arabidopsis allopolyploids. Nat Ecol Evol [Internet]. 2021 Oct 19;5(10):1382–93. Available from: https://www.nature.com/articles/s41559-021-01523-y
40. Chen ZJ. Genetic and Epigenetic Mechanisms for Gene Expression and Phenotypic Variation in Plant Polyploids. Annu Rev Plant Biol [Internet]. 2007 Jun;58(1):377–406. Available from: http://www.annualreviews.org/doi/10.1146/annurev.arplant.58.032806.103835
41. Chagué V, Just J, Mestiri I, Balzergue S, Tanguy A, Huneau C, et al. Genome-wide gene expression changes in genetically stable synthetic and natural wheat allohexaploids. New Phytol [Internet]. 2010 Sep 25;187(4):1181–94. Available from: https://onlinelibrary.wiley.com/doi/10.1111/j.1469-8137.2010.03339.x
42. Yoo M-J, Liu X, Pires JC, Soltis PS, Soltis DE. Nonadditive Gene Expression in Polyploids. Annu Rev Genet [Internet]. 2014 Nov 23;48(1):485–517. Available from: https://www.annualreviews.org/doi/10.1146/annurev-genet-120213-092159
43. Chelaifa H, Monnier A, Ainouche M. Transcriptomic changes following recent natural hybridization and allopolyploidy in the salt marsh species Spartina × townsendii and Spartina anglica (Poaceae). New Phytol [Internet]. 2010 Apr;186(1):161–74. Available from: https://onlinelibrary.wiley.com/doi/10.1111/j.1469-8137.2010.03179.x
44. Edger PP, Smith RD, McKain MR, Cooley AM, Vallejo-Marin M, Yuan Y-WY, et al. Subgenome Dominance in an Interspecific Hybrid, Synthetic Allopolyploid, and a 140-Year-Old Naturally Established Neo-Allopolyploid Monkeyflower. Plant Cell [Internet]. 2017 Sep;29(9):2150–67. Available from: http://www.plantcell.org/lookup/doi/10.1105/tpc.17.00010
45. Ha M, Lu J, Tian L, Ramachandran V, Kasschau KD, Chapman EJ, et al. Small RNAs serve as a genetic buffer against genomic shock in Arabidopsis interspecific hybrids and allopolyploids. Proc Natl Acad Sci [Internet]. 2009 Oct 20;106(42):17835–40. Available from: https://pnas.org/doi/full/10.1073/pnas.0907003106
46. Shen Y, Zhao Q, Zou J, Wang W, Gao Y, Meng J, et al. Characterization and expression patterns of small RNAs in synthesized Brassica hexaploids. Plant Mol Biol [Internet]. 2014 Jun 2;85(3):287–99. Available from: http://link.springer.com/10.1007/s11103-014-0185-x
47. Song Q, Chen JZ. Epigenetic and developmental regulation in plant polyploids. Curr Opin Plant Biol [Internet]. 2015;24:101–9. Available from: http://dx.doi.org/10.1016/j.pbi.2015.02.007
48. Madlung A, Masuelli RW, Watson B, Reynolds SH, Davison J, Comai L. Remodeling of DNA Methylation and Phenotypic and Transcriptional Changes in Synthetic Arabidopsis Allotetraploids. Plant Physiol [Internet]. 2002 Jun 1;129(2):733–46. Available from: http://www.plantphysiol.org/lookup/doi/10.1104/pp.003095
49. Osborn TC, Butrulle D V, Sharpe AG, Pickering KJ, Parkin IAP, Parker JS, et al. Detection and Effects of a Homeologous Reciprocal Transposition in Brassica napus. Genetics [Internet]. 2003 Nov 1;165(3):1569–77. Available from: https://academic.oup.com/genetics/article/165/3/1569/6052947
50. Bird KA, VanBuren R, Puzey JR, Edger PP. The causes and consequences of subgenome dominance in hybrids and recent polyploids. New Phytol [Internet]. 2018 Oct 8;220(1):87–93. Available from: https://onlinelibrary.wiley.com/doi/10.1111/nph.15256
51. Hurgobin B, Golicz AA, Bayer PE, Chan CK, Tirnaz S, Dolatabadian A, et al. Homoeologous exchange is a major cause of gene presence/absence variation in the amphidiploid Brassica napus. Plant Biotechnol J [Internet]. 2018 Jul 10;16(7):1265–74. Available from: https://onlinelibrary.wiley.com/doi/10.1111/pbi.12867
52. Parisod C, Alix K, Just J, Petit M, Sarilar V, Mhiri C, et al. Impact of transposable elements on the organization and function of allopolyploid genomes. New Phytol [Internet]. 2010 Apr;186(1):37–45. Available from: https://onlinelibrary.wiley.com/doi/10.1111/j.1469-8137.2009.03096.x
53. Hollister JD, Smith LM, Guo Y-L, Ott F, Weigel D, Gaut BS. Transposable elements and small RNAs contribute to gene expression divergence between Arabidopsis thaliana and Arabidopsis lyrata. Proc Natl Acad Sci [Internet]. 2011 Feb 8;108(6):2322–7. Available from: https://pnas.org/doi/full/10.1073/pnas.1018222108
54. Diez CM, Roessler K, Gaut BS. Epigenetics and plant genome evolution. Curr Opin Plant Biol

[Internet]. 2014 Apr;18:1–8. Available from: https://linkinghub.elsevier.com/retrieve/pii/S1369526613001969
55. Freeling M, Scanlon MJ, Fowler JE. Fractionation and subfunctionalization following genome duplications: mechanisms that drive gene content and their consequences. Curr Opin Genet Dev [Internet]. 2015 Dec;35:110–8. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0959437X15001173
56. Tang H, Woodhouse MR, Cheng F, Schnable JC, Pedersen BS, Conant G, et al. Altered Patterns of Fractionation and Exon Deletions in Brassica rapa Support a Two-Step Model of Paleohexaploidy. Genetics [Internet]. 2012 Apr 1;190(4):1563–74. Available from: https://academic.oup.com/genetics/article/190/4/1563/6064110
57. Woodhouse MR, Schnable JC, Pedersen BS, Lyons E, Lisch D, Subramaniam S, et al. Following Tetraploidy in Maize, a Short Deletion Mechanism Removed Genes Preferentially from One of the Two Homeologs. Wolfe KH, editor. PLoS Biol [Internet]. 2010 Jun 29;8(6):e1000409. Available from: https://dx.plos.org/10.1371/journal.pbio.1000409
58. Lou RN, Jacobs A, Wilder AP, Therkildsen NO. A beginner's guide to low-coverage whole genome sequencing for population genomics. Mol Ecol. 2021;30(23):5966–93.
59. Ellegren H. Genome sequencing and population genomics in non-model organisms. Trends Ecol Evol [Internet]. 2014 Jan;29(1):51–63. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0169534713002310
60. Mayer KFX, Rogers J, Doležel J, Pozniak C, Eversole K, Feuillet C, et al. A chromosome-based draft sequence of the hexaploid bread wheat (Triticum aestivum) genome. Science (80-) [Internet]. 2014 Jul 18;345(6194). Available from: https://www.science.org/doi/10.1126/science.1251788
61. Zhou Y, Zhao X, Li Y, Xu J, Bi A, Kang L, et al. Triticum population sequencing provides insights into wheat adaptation. Nat Genet [Internet]. 2020 Dec 26;52(12):1412–22. Available from: https://www.nature.com/articles/s41588-020-00722-w
62. Chen ZJ, Sreedasyam A, Ando A, Song Q, De Santiago LM, Hulse-Kemp AM, et al. Genomic diversifications of five Gossypium allopolyploid species and their impact on cotton improvement. Nat Genet [Internet]. 2020 May 20;52(5):525–33. Available from: http://www.nature.com/articles/s41588-020-0614-5
63. Douglas GM, Gos G, Steige KA, Salcedo A, Holm K, Josephs EB, et al. Hybrid origins and the earliest stages of diploidization in the highly successful recent polyploid Capsella bursa-pastoris. Proc Natl Acad Sci [Internet]. 2015 Mar 3;112(9):2806–11. Available from: http://www.pnas.org/lookup/doi/10.1073/pnas.1412277112
64. Burns R, Mandáková T, Gunis J, Soto-Jiménez LM, Liu C, Lysak MA, et al. Gradual evolution of allopolyploidy in Arabidopsis suecica. Nat Ecol Evol [Internet]. 2021 Oct 19;5(10):1367–81. Available from: https://www.nature.com/articles/s41559-021-01525-w
65. Ramsey J, Schemske DW. PATHWAYS, MECHANISMS, AND RATES OF POLYPLOID FORMATION IN FLOWERING PLANTS. Annu Rev Ecol Syst [Internet]. 1998 Nov;29(1):467–501. Available from: https://www.annualreviews.org/doi/10.1146/annurev.ecolsys.29.1.467
66. Reyna-López GE, Simpson J, Ruiz-Herrera J. Differences in DNA methylation patterns are detectable during the dimorphic transition of fungi by amplification of restriction polymorphisms. Mol Gen Genet MGG [Internet]. 1997 Feb;253(6):703–10. Available from: http://link.springer.com/10.1007/s004380050374
67. Shaked H, Kashkush K, Ozkan H, Feldman M, Levy AA. Sequence Elimination and Cytosine Methylation Are Rapid and Reproducible Responses of the Genome to Wide Hybridization and Allopolyploidy in Wheat. Plant Cell [Internet]. 2001 Aug;13(8):1749–59. Available from: http://www.plantcell.org/lookup/doi/10.1105/TPC.010083
68. Xu Y, Zhong L, Wu X, Fang X, Wang J. Rapid alterations of gene expression and cytosine methylation in newly synthesized Brassica napus allopolyploids. Planta. 2009;229(3):471–83.
69. HEGARTY MJ, BATSTONE T, BARKER GL, EDWARDS KJ, ABBOTT RJ, HISCOCK SJ. Nonadditive changes to cytosine methylation as a consequence of hybridization and genome duplication in Senecio (Asteraceae). Mol Ecol [Internet]. 2011 Jan;20(1):105–13. Available from: http://doi.wiley.com/10.1111/j.1365-294X.2010.04926.x
70. Li E, Zhang Y. DNA Methylation in Mammals. Cold Spring Harb Perspect Biol [Internet]. 2014 May 1;6(5):a019133–a019133. Available from: http://cshperspectives.cshlp.org/lookup/doi/10.1101/cshperspect.a019133
71. Bewick AJ, Niederhuth CE, Ji L, Rohr NA, Griffin PT, Leebens-Mack J, et al. The evolution of CHROMOMETHYLASES and gene body DNA methylation in plants. Genome Biol. 2017;18(1):1–13.

72. Niederhuth CE, Bewick AJ, Ji L, Alabady MS, Kim K Do, Li Q, et al. Widespread natural variation of DNA methylation within angiosperms. Genome Biol [Internet]. 2016 Dec 27;17(1):194. Available from: http://dx.doi.org/10.1186/s13059-016-1059-0
73. Gaeta RT, Pires JC, Iniguez-Luy F, Leon E, Osborn TC. Genomic Changes in Resynthesized Brassica napus and Their Effect on Gene Expression and Phenotype. Plant Cell [Internet]. 2007 Nov;19(11):3403–17. Available from: http://www.plantcell.org/lookup/doi/10.1105/tpc.107.054346
74. Parisod C, Salmon A, Zerjal T, Tenaillon M, Grandbastien M-A, Ainouche M. Rapid structural and epigenetic reorganization near transposable elements in hybrid and allopolyploid genomes in Spartina. New Phytol [Internet]. 2009 Dec;184(4):1003–15. Available from: http://doi.wiley.com/10.1111/j.1469-8137.2009.03029.x
75. Liu B, Brubaker CL, Mergeai G, Cronn RC, Wendel JF. Polyploid formation in cotton is not accompanied by rapid genomic changes. Genome [Internet]. 2001 Jun 1;44(3):321–30. Available from: http://www.nrcresearchpress.com/doi/10.1139/g01-011
76. Wetterstrand KA. The Cost of Sequencing a Human Genome. 2021.
77. Kumar KR, Cowley MJ, Davis RL. Next-Generation Sequencing and Emerging Technologies. Semin Thromb Hemost [Internet]. 2019 Oct 16;45(07):661–73. Available from: http://www.thieme-connect.de/DOI/DOI?10.1055/s-0039-1688446
78. Sun Y, Shang L, Zhu Q-H, Fan L, Guo L. Twenty years of plant genome sequencing: achievements and challenges. Trends Plant Sci [Internet]. 2021 Nov; Available from: https://linkinghub.elsevier.com/retrieve/pii/S1360138521002818
79. Kyriakidou M, Tai HH, Anglin NL, Ellis D, Strömvik M V. Current Strategies of Polyploid Plant Genome Sequence Assembly. Front Plant Sci [Internet]. 2018 Nov 21;9. Available from: https://www.frontiersin.org/article/10.3389/fpls.2018.01660/full
80. Duchemin W, Dupont P-Y, Campbell MA, Ganley AR, Cox MP. HyLiTE: accurate and flexible analysis of gene expression in hybrid and allopolyploid species. BMC Bioinformatics [Internet]. 2015 Dec 16;16(1):8. Available from: https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-014-0433-8
81. Mithani A, Belfield EJ, Brown C, Jiang C, Leach LJ, Harberd NP. HANDS: a tool for genome-wide discovery of subgenome- specific base-identity in polyploids. BMC Genomics [Internet]. 2013 Dec 24;14(1):653. Available from: https://bmcgenomics.biomedcentral.com/articles/10.1186/1471-2164-14-653
82. Khan A, Belfield EJ, Harberd NP, Mithani A. HANDS2: accurate assignment of homoeallelic base-identity in allopolyploids despite missing data. Sci Rep [Internet]. 2016 Jul 5;6(1):29234. Available from: http://www.nature.com/articles/srep29234
83. Page JT, Gingle AR, Udall JA. PolyCat: A Resource for Genome Categorization of Sequencing Reads From Allopolyploid Organisms. G3: Genes|Genomes|Genetics [Internet]. 2013 Mar;3(3):517–25. Available from: http://g3journal.org/lookup/doi/10.1534/g3.112.005298
84. Page JT, Udall JA. Methods for mapping and categorization of DNA sequence reads from allopolyploid organisms. BMC Genet [Internet]. 2015;16(Suppl 2):S4. Available from: http://bmcgenet.biomedcentral.com/articles/10.1186/1471-2156-16-S2-S4
85. Peralta M, Combes M-C, Cenci A, Lashermes P, Dereeper A. SNiPloid: A Utility to Exploit High-Throughput SNP Data Derived from RNA-Seq in Allopolyploid Species. Int J Plant Genomics [Internet]. 2013 Sep 12;2013:1–6. Available from: https://www.hindawi.com/journals/ijpg/2013/890123/
86. Akama S, Shimizu-Inatsugi R, Shimizu KK, Sese J. Genome-wide quantification of homeolog expression ratio revealed nonstochastic gene regulation in synthetic allopolyploid Arabidopsis. Nucleic Acids Res [Internet]. 2014 Apr 1;42(6):e46–e46. Available from: https://academic.oup.com/nar/article/42/6/e46/2437554
87. Kuo T, Frith MC, Sese J, Horton P. EAGLE: Explicit Alternative Genome Likelihood Evaluator. BMC Med Genomics [Internet]. 2018 Apr 20;11(S2):28. Available from: https://bmcmedgenomics.biomedcentral.com/articles/10.1186/s12920-018-0342-1
88. Omony J, Nussbaumer T, Gutzat R. DNA methylation analysis in plants: review of computational tools and future perspectives. Brief Bioinform [Internet]. 2020 Apr 21;21(3):906–18. Available from: https://academic.oup.com/bib/article/21/3/906/5432309
89. Mohn F, Weber M, Schübeler D, Roloff T-C. Methylated DNA Immunoprecipitation (MeDIP). In: Tost J, editor. Totowa, NJ, NJ: Humana Press; 2009. p. 55–64. (Methods in Molecular Biology; vol. 507). Available from: http://link.springer.com/10.1007/978-1-59745-522-0_5
90. Lan X, Adams C, Landers M, Dudas M, Krissinger D, Marnellos G, et al. High Resolution Detection and Analysis of CpG Dinucleotides Methylation Using MBD-Seq Technology. Jothi

R, editor. PLoS One [Internet]. 2011 Jul 11;6(7):e22226. Available from: https://dx.plos.org/10.1371/journal.pone.0022226
91. Liu Y, Siejka-Zielińska P, Velikova G, Bi Y, Yuan F, Tomkova M, et al. Bisulfite-free direct detection of 5-methylcytosine and 5-hydroxymethylcytosine at base resolution. Nat Biotechnol. 2019 Apr;37(4):424–9.
92. Olova N, Krueger F, Andrews S, Oxley D, Berrens R V., Branco MR, et al. Comparison of whole-genome bisulfite sequencing library preparation strategies identifies sources of biases affecting DNA methylation data. Genome Biol [Internet]. 2018 Dec 15;19(1):33. Available from: https://genomebiology.biomedcentral.com/articles/10.1186/s13059-018-1408-2
93. Bock C. Analysing and interpreting DNA methylation data. Nat Rev Genet [Internet]. 2012 Oct 18;13(10):705–19. Available from: http://www.nature.com/articles/nrg3273
94. SHIMIZU-INATSUGI R, LIHOVÁ J, IWANAGA H, KUDOH H, MARHOLD K, SAVOLAINEN O, et al. The allopolyploid Arabidopsis kamchatica originated from multiple individuals of Arabidopsis lyrata and Arabidopsis halleri. Mol Ecol [Internet]. 2009 Oct;18(19):4024–48. Available from: http://doi.wiley.com/10.1111/j.1365-294X.2009.04329.x
95. Shimizu KK, Fuji S, Marhold K, Watanabe K, Kudoh H. Arabidopsis kamchatica (Fisch. ex DC.) K. Shimizu & Kudoh and A. kamchatica subsp. kawasakiana (Makino) K. Shimizu & Kudoh, New Combinations. Acta Phytotaxon Geobot [Internet]. 2005;56(2):163–72. Available from: https://doi.org/10.18942/apg.KJ00004623241
96. Vallejo-Marin M. Mimulus peregrinus (Phrymaceae): A new British allopolyploid species. PhytoKeys [Internet]. 2012 Jul 6;14:1–14. Available from: http://www.pensoft.net/journals/phytokeys/article/3305/abstract/mimulus-peregrinus-phrymaceae-a-new-british-allopolyploid-species
97. Briskine R V., Paape T, Shimizu-Inatsugi R, Nishiyama T, Akama S, Sese J, et al. Genome assembly and annotation of Arabidopsis halleri, a model for heavy metal hyperaccumulation and evolutionary ecology. Mol Ecol Resour [Internet]. 2016 Sep;17(5):1025–36. Available from: http://doi.wiley.com/10.1111/1755-0998.12604
98. Van de Peer Y, Ashman T-L, Soltis PS, Soltis DE. Polyploidy: an evolutionary and ecological force in stressful times. Plant Cell [Internet]. 2021 Mar 22;33(1):11–26. Available from: https://academic.oup.com/plcell/article/33/1/11/6015242
99. Fawcett JA, Maere S, Van de Peer Y. Plants with double genomes might have had a better chance to survive the Cretaceous-Tertiary extinction event. Proc Natl Acad Sci [Internet]. 2009 Apr 7;106(14):5737–42. Available from: http://www.pnas.org/cgi/doi/10.1073/pnas.0900906106
100. Vanneste K, Baele G, Maere S, Van de Peer Y. Analysis of 41 plant genomes supports a wave of successful genome duplications in association with the Cretaceous–Paleogene boundary. Genome Res [Internet]. 2014 Aug;24(8):1334–47. Available from: http://genome.cshlp.org/lookup/doi/10.1101/gr.168997.113
101. Gunn BF, Murphy DJ, Walsh NG, Conran JG, Pires JC, Macfarlane TD, et al. Evolution of Lomandroideae: Multiple origins of polyploidy and biome occupancy in Australia. Mol Phylogenet Evol [Internet]. 2020 Aug;149:106836. Available from: https://linkinghub.elsevier.com/retrieve/pii/S1055790320301081
102. Folk RA, Siniscalchi CM, Soltis DE. Angiosperms at the edge: Extremity, diversity, and phylogeny. Plant Cell Environ [Internet]. 2020 Dec 28;43(12):2871–93. Available from: https://onlinelibrary.wiley.com/doi/10.1111/pce.13887
103. Dong S, Adams KL. Differential contributions to the transcriptome of duplicated genes in response to abiotic stresses in natural and synthetic polyploids. New Phytol [Internet]. 2011 Jun;190(4):1045–57. Available from: https://onlinelibrary.wiley.com/doi/10.1111/j.1469-8137.2011.03650.x
104. Liu Z, Adams KL. Expression Partitioning between Genes Duplicated by Polyploidy under Abiotic Stress and during Organ Development. Curr Biol [Internet]. 2007 Oct;17(19):1669–74. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0960982207018477
105. Milosavljevic S, Kuo T, Decarli S, Mohn L, Sese J, Shimizu KK, et al. ARPEGGIO: Automated Reproducible Polyploid EpiGenetic GuIdance workflOw. BMC Genomics [Internet]. 2021 Dec 1;22(1):547. Available from: https://bmcgenomics.biomedcentral.com/articles/10.1186/s12864-021-07845-2

Chapter 1: ARPEGGIO: Automated Reproducible Polyploid EpiGenetic GuIdance workflOw

I conceptualized and designed ARPEGGIO's structure, I coded and tested the workflow and I documented ARPEGGIO's setup and usage. I wrote, revised and published the following manuscript.

Milosavljevic *et al. BMC Genomics* (2021) 22:547
https://doi.org/10.1186/s12864-021-07845-2

BMC Genomics

SOFTWARE Open Access

ARPEGGIO: Automated Reproducible Polyploid EpiGenetic GuIdance workflOw

Stefan Milosavljevic[1,2], Tony Kuo[3], Samuele Decarli[4], Lucas Mohn[1], Jun Sese[5,6], Kentaro K. Shimizu[1,7], Rie Shimizu-Inatsugi[1] and Mark D. Robinson[2,8*]

Abstract

Background: Whole genome duplication (WGD) events are common in the evolutionary history of many living organisms. For decades, researchers have been trying to understand the genetic and epigenetic impact of WGD and its underlying molecular mechanisms. Particular attention was given to allopolyploid study systems, species resulting from an hybridization event accompanied by WGD. Investigating the mechanisms behind the survival of a newly formed allopolyploid highlighted the key role of DNA methylation. With the improvement of high-throughput methods, such as whole genome bisulfite sequencing (WGBS), an opportunity opened to further understand the role of DNA methylation at a larger scale and higher resolution. However, only a few studies have applied WGBS to allopolyploids, which might be due to lack of genomic resources combined with a burdensome data analysis process. To overcome these problems, we developed the Automated Reproducible Polyploid EpiGenetic GuIdance workflOw (ARPEGGIO): the first workflow for the analysis of epigenetic data in polyploids. This workflow analyzes WGBS data from allopolyploid species via the genome assemblies of the allopolyploid's parent species. ARPEGGIO utilizes an updated read classification algorithm (EAGLE-RC), to tackle the challenge of sequence similarity amongst parental genomes. ARPEGGIO offers automation, but more importantly, a complete set of analyses including spot checks starting from raw WGBS data: quality checks, trimming, alignment, methylation extraction, statistical analyses and downstream analyses. A full run of ARPEGGIO outputs a list of genes showing differential methylation. ARPEGGIO was made simple to set up, run and interpret, and its implementation ensures reproducibility by including both package management and containerization.

Results: We evaluated ARPEGGIO in two ways. First, we tested EAGLE-RC's performance with publicly available datasets given a ground truth, and we show that EAGLE-RC decreases the error rate by 3 to 4 times compared to standard approaches. Second, using the same initial dataset, we show agreement between ARPEGGIO's output and published results. Compared to other similar workflows, ARPEGGIO is the only one supporting polyploid data.

Conclusions: The goal of ARPEGGIO is to promote, support and improve polyploid research with a reproducible and automated set of analyses in a convenient implementation. ARPEGGIO is available at https://github.com/supermaxiste/ARPEGGIO.

Keywords: Snakemake, Epigenetics, Bisulfite-sequencing, Polyploidy, Allopolyploids, Reproducibility, Automation, Workflow, Dna-methylation, Whole-genome-bisulfite-sequencing

* Correspondence: mark.robinson@mls.uzh.ch
[2]SIB Swiss Institute of Bioinformatics, University of Zurich, Zurich, Switzerland
[8]Department of Molecular Life Sciences, University of Zurich, Zurich, Switzerland
Full list of author information is available at the end of the article

Background

Polyploidy, also known as whole genome duplication (WGD), is a process leading to the formation of an organism with more than two sets of chromosomes. There are two types of polyploidy: autopolyploidy, the doubling of an entire genome in a single species, and allopolyploidy, the hybridization of two different species followed by whole genome duplication [1]. Both of these processes influenced the evolutionary history of many living organisms such as nematodes, arthropods, chordates, fungi, oomycetes and plants [1–3]. Of all these lineages, the most extensive research on polyploidy has been done on land plants [1–8], where about 35% of all species were estimated to be recent polyploids [7, 8] and at least one ancient WGD was inferred in the ancestry of every lineage [3].

To understand the successful prevalence of WGD and the underlying mechanisms, particular attention was given to early stages of polyploidy in allopolyploids [4, 9–11]. Among several observed genomic and epigenomic changes [4, 10, 12], DNA methylation was shown to play an important role to ensure the survival of a newly formed allopolyploid [13–19]. A well-studied example comes from Madlung and colleagues [13] in which they chemically treated synthetic *Arabidopsis suecica* allotetraploids to remove DNA methylation over the whole genome. With this treatment, they observed many phenotypic disorders such as abnormal branching or homeotic abnormalities in flowers, mostly leading to sterility. These abnormalities were not observed when treating the parent species or the natural allopolyploid, highlighting the importance of DNA methylation in the first generations after allopolyploidization. Follow-up studies focused on the epigenetic regulation in other resynthesized allopolyploid species with varying outcomes. In allopolyploid wheat, *Tragopogon*, *Spartina* and rice, DNA methylation changes indicated gene repression favoring one parental genome over the other [15–20]. This was not the case in *Arabidopsis*, where similar DNA methylation and expression changes were observed on both parental genomes [21]. In *Brassica*, both previously mentioned outcomes were reported [15, 22], while in cotton no changes were found [23]. All these studies proposed different mechanisms to clarify the role of methylation and its short and long term evolutionary impact, but the discussion remains open [4]. One reason that might complicate the grounds of such discussion, is the variety of tools and methods used to analyze DNA methylation data. To better control discrepancies between findings caused by methodological differences, a standardized set of tools would be ideal.

Despite the potential significance of DNA methylation in allopolyploid evolution, many of the previously mentioned findings were limited by low-throughput methods. These methods, such as methylation-sensitive amplified length polymorphisms (MSAP), were unable to capture changes at a whole genome level [24]. With advances in technology, new high-throughput methods such as whole genome bisulfite sequencing (WGBS) are able to obtain methylation information at individual nucleotides over the whole genome [25].

At the whole genome level, DNA methylation is separated into three different sequence contexts: CG, CHG and CHH (where H = A, T or C). Each context is regulated by different families of enzymes and depending on the species, some contexts might be more important than others [26]. For example, in mammals, methylation occurs mainly in CG context, while in plants it occurs in all three contexts [26].

Although WGBS is considered to be the gold standard in whole-genome DNA methylation studies [24, 27], research on allopolyploid species using WGBS is limited, with most of the studies coming from crop study systems [28–30]. On the one hand, these systems have excellent genomic resources to provide valuable insights, while on the other, it is unclear whether these insights can be extended to wild organisms in nature given their artificial selection [4].

In other polyploid study systems, two major challenges prevent the use of WGBS: limited genomic resources (i.e. genome assemblies) and a laborious data analysis process. The number of plant genome assemblies has been increasing exponentially in the last years [31], but polyploid genome assemblies are still an intensive, complex and expensive task [32, 33], preventing the development of genetic and epigenetic studies using polyploids. For allopolyploids, this obstacle can be avoided by using the genome assemblies of the two (known) parent species [34], usually diploid.

Besides limited genomic resources, another challenge in WGBS comes from a laborious and complex data analysis process [35–37]. In standard WGBS data analysis pipelines, complexities related to polyploids are often not taken into account. For example when mapping reads originating from an allopolyploid, high sequence similarity between parents can be challenging for read mapping algorithms [38, 39] and the outcome can have strong bias, especially when the quality of the assemblies is asymmetric [40]. To tackle this problem, several methods were developed to improve the categorization of allopolyploids' reads to the correct parental genome. HomeoRoq [41] and PolyDog [40] take into account alignment quality from both parental genomes to assign reads, while PolyCat [42] and EAGLE-RC [34] also use explicit genotype differences between parent genomes to classify reads. EAGLE-RC outperformed HomeoRoq in estimating homeolog expression with data from tetraploid *Arabidopsis* and hexaploid

wheat [34]. When comparing EAGLE-RC and PolyCat using *Gossypium* RNA-seq data, both tools outperformed other pipelines and had similar performance [43]. Among all the tools, only PolyCat supports bisulfite-treated WGBS data, but only with available variant information (i.e. SNPs) between subgenomes, which represents an additional obstacle for most allopolyploid systems [44].

To promote and support allopolyploid DNA methylation research, we developed the Automated Reproducible Polyploid EpiGenetic GuIdance workflOw (ARPEGGIO). ARPEGGIO is a specialized workflow to process raw WGBS data utilizing the assemblies of the allopolyploid's parent species (hereafter referred to as progenitors) or independently phased subgenomes of an allopolyploid. ARPEGGIO includes all the steps from raw WGBS data to a list of genes showing differential methylation: conversion check, quality check, trimming, alignment, read classification, methylation extraction, statistical analysis and downstream analysis. More details about the prerequisites, setup, tools and outputs are discussed in the implementation section.

To handle sequence similarity between two genomes, ARPEGGIO exploits an updated version of EAGLE-RC that supports bisulfite-treated reads and does not require variant information between subgenomes. This version of EAGLE-RC was evaluated using three WGBS datasets, and showed better performance compared to a genome concatenation approach.

ARPEGGIO's implementation combines the Snakemake workflow management system [45] with the Conda package manager [46] and Singularity containers [48] to ensure both ease of use and reproducibility. For ease of use, a centralized configuration file controls all parameters related to ARPEGGIO and through Conda, all the tools required by the workflow are automatically installed.

Implementation

Design, concepts and challenges

ARPEGGIO's design had three main objectives, each dealing with different aspects and challenges of the workflow: allopolyploid support, ease of use and reproducibility. These aspects will be discussed at high-level here and more details about their implementation can be found in the following sections.

To support allopolyploids, ARPEGGIO first needed to allow for different experimental designs (i.e. sample comparisons). For allopolyploids without a genome assembly, but progenitor assemblies available, there are two possible comparisons: allopolyploid against progenitors or allopolyploid against allopolyploid (Fig. 1a, b). The former compares the two allopolyploid's subgenomes to the progenitors, while the latter compares directly the two subgenomes in different experimental conditions. An additional third comparison allows two groups of individuals from a species with an available (phased) genome assembly (Fig. 1c), regardless of the ploidy level. After choosing a comparison, the next allopolyploid-specific step is read classification.

To analyze allopolyploid data with progenitor assemblies, we run two separate workflows in parallel, one per progenitor (Fig. 2). The separation occurs at the alignment and deduplication step, where two separate

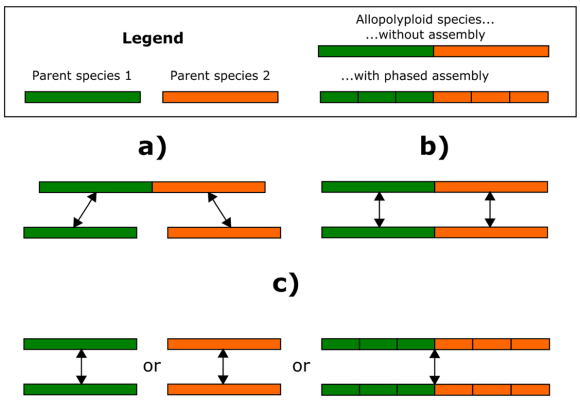


Fig. 1 Schematic view of the experimental designs supported by ARPEGGIO. There are 3 possible comparisons: a) polyploid species without assembly against its progenitors, b) same polyploid without assembly in two different experimental conditions and c) diploid species or polyploid species with an available phased assembly in two different experimental conditions. All comparisons are about whole genome DNA methylation patterns

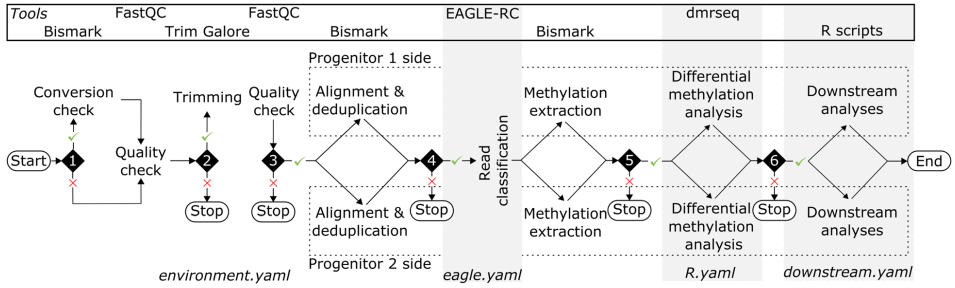


Fig. 2 Schematic overview of ARPEGGIO's structure. All the shapes follow the flowchart standardized symbols [47]. Ovals show the beginning and ends of the workflow. Diamonds represent conditional rules in ARPEGGIO's configuration file and those rules make ARPEGGIO more adaptable to the needs of the user. Each conditional rule can be set to "true" (tick) or "false" (cross). Besides the first conditional rule, all other rules stop the workflow at the given point when set to "false". The different grey backgrounds and the white background represent different Conda environments used by ARPEGGIO to carry out different steps of the analyses. In the scheme, the background of each step represents the environment that the step is part of. The bottom of each background shows the name of the file used to create the environment. At the top all the tools used by ARPEGGIO are shown and vertically aligned their corresponding step in the workflow. From the alignment and deduplication step, ARPEGGIO executes two workflows in parallel for each progenitor side, both highlighted by the dashed areas

alignments are performed for the same allopolyploid data, one for each progenitor. With each allopolyploid read being mapped twice, a read classification algorithm must choose one of the two progenitors; for the classification, ARPEGGIO uses EAGLE-RC. In short, EAGLE-RC applies a probabilistic method that compares the two mappings for each read and classifies its progenitor origin or deems it ambiguous (equal probabilities for both progenitors' sides). Two parameters were added to EAGLE-RC to deal with bisulfite data from allopolyploids. The first is called "no genotype information" (NGI) and allows EAGLE-RC to be used with no information about variants in the genome. This mode is especially useful to reduce prerequisites for using ARPEGGIO. The second parameter is called "bisulfite" (BS) and it causes bisulfite treatment to be taken into account when a bisulfite-treated read is mapped to a genome. This parameter considers C-T as a match (forward strand), G-A as a match (reverse strand) or both.

Both experimental design and EAGLE-RC's inclusion had a major impact on ARPEGGIO's structure and implementation, but other important aspects were also taken into account. For example, allopolyploids can be found in different lineages such as plants and mammals, meaning that different approaches should be considered for conversion efficiency checks and the selection of methylation contexts.

Once the general design of ARPEGGIO was established, the next challenge was to make the workflow easy to set up, run and interpret. ARPEGGIO requires the users to install the Conda package management system [46], then Snakemake [45] via Conda and, optionally, Singularity [48]. No other tools need to be installed as ARPEGGIO will take care of automatically installing what is needed. To prepare ARPEGGIO for a new dataset, input files have to be prepared and ARPEGGIO's settings have to be defined. Input files include raw data in FASTQ format and the progenitors' reference genome assemblies. To run downstream analyses, annotation files for both assemblies are also required. ARPEGGIO's settings are defined with a configuration file and a metadata file. The configuration file has different sections, each including parameters that define how ARPEGGIO will be run, while the metadata file contains information about samples such as filename, sequencing strategy, origin (allopolyploid or progenitor) and experimental condition (if present). A small dataset with its own configuration and metadata file is provided in ARPEGGIO's repository as an example. To run ARPEGGIO, only one command is needed and its main options are related to reproducibility (discussed below) and parallelization (i.e. multiple core usage). After ARPEGGIO is successfully run, the number of files in the output folder can be significant. For this reason, a map of the output is available in ARPEGGIO's user documentation: this map shows the general output structure with all the main folders and their contents. For each folder, there's a section describing the folder itself, sub-folders and all the files included in it.

Another key goal of ARPEGGIO was to ensure reproducibility. Considering the variety of tools and number of steps in the workflow, by letting users (or Conda) define the version of each tool, the outcome could be variable and lead to future reproducibility problems. To overcome this, we fixed all the versions of the tools and we combined ARPEGGIO with Conda and Singularity containers. The user can choose to use either only

Conda or Conda and Singularity together. The main difference between the two modes lies on potential issues between the user's system and Conda. When these issues happen, Singularity offers a containerized run of Conda. Both these options can be specified with one or two parameters respectively when running ARPEGGIO. Aside from tool version differences, which we addressed above, the configuration file specifies all parameters that were used in a workflow run. Associating results to a specific set of parameters further aids reproducibility. The configuration file may also be shared to other researchers aiming to reanalyze a given dataset.

Workflow overview

ARPEGGIO includes eight processes: conversion check, quality checks, trimming, alignment and deduplication, read classification, methylation extraction, differential methylation analysis and downstream analyses (Fig. 2). These processes are divided into six steps, each represented by a black diamond in Fig. 2. Step 1 includes conversion check, a quality check specific to WGBS data, where reads are aligned to an unmethylated control genome (usually plastid genome for plants and lambda genome for others) to assess the efficiency of the bisulfite conversion; the lower the mapping rate, the better the conversion [27]. This process is executed by Bismark [49]. The conversion check is followed by quality checks and trimming (step 1 and 2), executed by FastQC [50] and Trim Galore [51], respectively. Both processes are common procedures to assess read quality and remove noise. Step 3 performs read alignment to a reference genome, followed by deduplication, which removes duplicated reads. Both of these are carried out by the Bismark suite [49]. From this point of the workflow allopolyploid data is separated into two parallel workflows: one per progenitor side. These workflows intersect in the next, allopolyploid-specific read classification step (step 4), executed by the updated version of EAGLE-RC [34]. Here, EAGLE-RC will classify allopolyploid reads after comparing the read alignment on each progenitor's side. After read classification (from step 5 on), the two workflows are independent, but execute the same steps. During methylation extraction via Bismark, methylation information is extracted for each cytosine from classified reads to produce a methylation count table. This table is used for differential methylation analyses (step 5), performed by the R/Bioconductor package dmrseq [52], to output a list of tested differentially methylated regions (DMRs). Finally, downstream analyses (step 6) consist of a series of R scripts for computing overlaps between statistically significant DMRs and annotated gene regions provided by the user (if available). More specifically, by default ARPEGGIO uses q-value < 0.05 to define a significant DMR. With this cutoff, ARPEGGIO looks for overlaps of at least 1 base pair between significant regions and gene regions based on the annotations. Before ARPEGGIO finishes a run, all reports (conversion check, quality checks, trimming, alignment, deduplication and methylation extraction) are combined into one interactive HTML report with MultiQC [53].

Each part in ARPEGGIO is optional and the user can specify which parts of the workflow to execute in the configuration file. It must be noted that skipping some parts will stop the workflow at a specific step (Fig. 2). Assuming that all prerequisites are met, ARPEGGIO goes from raw sequencing data to a list of genes showing differential methylation. Some useful intermediate outputs are also produced: an interactive HTML report merging all quality, alignment and methylation reports and an Rdata file with the output from the dmrseq analysis, which can be used to visualize DMRs or for other custom analyses.

Implementation details

ARPEGGIO is written in Snakemake, a Python based language for workflow development [45]. With Snakemake, a workflow is broken down into a series of rules. One rule can be seen as one step in the workflow with a defined input and output. Rules are related to each other based on their input and output files. Once all the rules are set, to run a Snakemake workflow, a target file (or multiple) needs to be requested. Snakemake will automatically build the workflow to obtain the target file based on the input/output relationships between rules (dependencies). If the relationships are successfully established, the workflow will be run. To illustrate these principles, an example with ARPEGGIO's rules is given in Additional File 1. This figure shows all the input/output relationships between rules when running ARPEGGIO with single-end data, comparing an allopolyploid to its progenitor species (default experimental design).

In addition to the core features of Snakemake, ARPEGGIO takes advantage of the integrated Conda package management system [46]. Conda creates environments containing a specific set of software and users can switch between different environments depending on the software package(s) they need. An environment can be created in several ways. ARPEGGIO creates environments through YAML files, specifying all the packages to be included and the channels from which the packages are searched. The integration of Conda in Snakemake allows rules to be run within a specific environment and during the execution of a workflow, Snakemake takes care of switching between environments if different rules require different environments. From a user perspective, once Conda and Snakemake are installed, ARPEGGIO will take care of installing all the tools needed for the

analyses, running them and switching automatically between environments when needed (Fig. 2).

Making the workflow specific for allopolyploids presented major challenges with both Snakemake and Conda. Snakemake rules in ARPEGGIO had to be structured to allow for any combination between sequencing strategies and experimental designs. This meant combining rules for six workflows in one: three experimental designs, each with two sequencing strategies. In addition, since EAGLE-RC could not be installed as a Conda package, a Conda environment with a specific set of rules was created to take care of downloading, extracting and installing EAGLE-RC.

In practice, any user can take advantage of all the Conda and Snakemake features discussed above with a central configuration file. Here, we will discuss the first three sections of this file, that consist of parameters concerning the workflow as a whole: general parameters, conditional rules and experimental designs. All the other sections in the configuration file are related to tool-specific parameters for each of the main steps in ARPE GGIO. More details about these parameters can be found in ARPEGGIO's user documentation. General parameters include the location of the output folder, the location of the metadata file and a parameter to define the sequencing strategy. Conditional rules are shown as black diamonds on Fig. 2. Those rules are set to "True" or "False" to define which parts of the workflow to run. Practically, only the initial steps of ARPEGGIO, quality check and trimming, can be skipped; otherwise, the workflow will stop for any other step that is set to "False". Finally, experimental designs are implemented via special modes. By default, ARPEGGIO compares a polyploid species against its two progenitor species (Fig. 1a). With the special mode "POLYPLOID_ONLY", ARPEGGIO compares a polyploid species from two different experimental conditions (Fig. 1b), while the mode "DIPLOID_ONLY" compares a diploid species from two different conditions (or a polyploid species with an available phased assembly, Fig. 1c).

Results & discussion

Performance of read classification

A simple and common way to analyze polyploid datasets is to concatenate the genome assemblies of the two progenitor species and let the aligner assign a mapping position. The position would define the origin of the read depending on which of the two subgenomes the read was mapped to. We define this approach as the "concatenated" approach.

The performance of EAGLE-RC was assessed using ARPEGGIO v3.0.0 while shell scripts were used to evaluate the concatenated approach (see Availability of data and materials). In both cases, the same versions of tools as in ARPEGGIO were used.

For the evaluation, we used six datasets from three pairs of progenitor species that form an allopolyploid or a hybrid, and we compared EAGLE-RC's classification error to that of the concatenated approach in a similar fashion as [34]. In short, each progenitor dataset was treated as an allopolyploid dataset, meaning that all the reads were assigned to a progenitor's side. With datasets coming from progenitors, the true origin of the reads was known, thus reads assigned to the wrong progenitor's side were used to calculate a classification error rate.

Two datasets were from *Mimulus guttatus* and *Mimulus luteus*, obtained from [54], with four technical replicates each. Those two species are the progenitors of the allopolyploid *Mimulus peregrinus*. Data from *Gossypium arboreum* and *Gossypium raimondii* was obtained from [29] and consisted of two technical replicates each. Those two species are the progenitors of the hybrid *Gossypium arboreum x raimondii*. The last datasets were produced in-house (Additional File 3) from *Arabidopsis halleri* and *Arabidopsis lyrata* with two biological replicates each. Those two species are the progenitors of the allopolyploid *Arabidopsis kamchatica* [55].

EAGLE-RC showed a lower error rate in all datasets compared to the concatenated approach (Table 1). The error rate was consistently between 3 to 4 times less with EAGLE-RC. When looking at absolute values, the improvement from read classification varied: from changes below 0.1% in *Gossypium* to almost 20% when using *Mimulus* data. These differences could be attributed to many factors, such as divergence between species, quality of genome assembly, and sequence data quality. We assessed divergence for two out of three progenitor pairs using the average nuclear identity [56] and the two *Gossypium* genomes had lower similarity compared to the two *Arabidopsis* species (Table 1). This was consistent with the known divergence in genome size between *G. raimondii* (0.8Gb) and *G. arboreum* (1.7Gb) and contributed to make the read classification task easier (Additional File 4). From a qualitative point of view, *Mimulus* had lower quality assemblies compared to the other species, and this difference might also explain the higher error rates in both methods.

Overall, EAGLE-RC showed a lower error rate with minimal loss of reads classified as ambiguous (Additional File 4). On the one hand, EAGLE-RC showed a lower error rate, while on the other, the absolute number of correctly assigned reads was lower in EAGLE-RC compared to the concatenated approach (Additional File 4). This happened because the reads classified as "ambiguous" reduced the amount of the correctly classified reads (both true negative and true positive reads). When focusing on the difference in true positive reads between

Table 1 Overview of the read assignment accuracy of EAGLE-RC against the concatenation method with real datasets. The first part of the table provides details on each dataset such as the species of origin, the type of replication (biological or technical), the sequencing strategy and the divergence between the two progenitor species, represented by the two-way average nuclear identity (ANI). The sequencing strategy includes the sequencing layout (PE = paired-end, SE = single-end) followed by the read length in bp. The two-way ANI was obtained using the ANI calculator from [56] with default parameters. The ANI value for Mimulus could not be calculated because of excessive computation time requirements (> 6'000 CPU hours). The second part of the table shows the average number of uniquely mapped reads for each approach, which was used to calculate the average error rate on the third part of the table. The error rate was obtained by the number of reads assigned to the wrong genome divided by the total number of reads that were uniquely mapped and deduplicated

Datasets				Average number of uniquely mapped reads		Average error rate	
Species	**Type of replicate (#)**	**Sequencing layout and read length**	**Two-way average nuclear identity**	**Concatenated genome**	**Read classification**	**Concatenated genome**	**Read classification**
Arabidopsis halleri	Biological (2)	PE150	94.29 ± 3.94%	17'258'758	18'311'330	3.98%	1.16%
Arabidopsis lyrata	Biological (2)	PE150		22'204'342	23'301'056	5.94%	1.45%
Mimulus guttatus	Technical (4)	SE150	N.A.	1'420'116	1'288'800	26.78%	7.52%
Mimulus luteus	Technical (4)	SE150		3'889'458	3'760'614	9.80%	2.29%
Gossypium arboreum	Technical (2)	PE125	91.07 ± 4.68%	253'912'667	254'261'702	0.0044%	0.0013%
Gossipyium raimondii	Technical (2)	PE125		242'590'069	246'935'598	0.0039%	0.0019%

EAGLE-RC and concatenation, values are negligible for both *Arabidopsis* and *Gossypium* datasets, representing < 0.01% of uniquely mapped reads. In the case of *Mimulus*, the number of true positive reads is ~ 10% higher in the concatenated approach, but the error-rate is also 3 to 4 times higher compared to EAGLE-RC. Taken together, these results suggest that EAGLE-RC has a clear advantage when analyzing allopolyploid WGBS data, where higher accuracy in subgenome recognition is required.

In this evaluation, we have not examined in detail the effect of the genetic divergence between progenitor genomes and allopolyploid genomes. Divergence results from DNA mutations happening after polyploidization and leading to changes on both progenitor sides in the polyploid's genome. The magnitude of differences is proportional to the number of generations, i.e. time, since polyploidization. As an example, *M. peregrinus* is a 140-years old polyploid, and thus the changes in its genome might be very few. We speculate that ARPEGGIO should be tolerant for older allopolyploids, as both EAGLE-RC and HomeoRoq have shown good performance with both DNA and RNA-seq data of *A. kamchatica*, which is estimated to have originated around 20,000–250,000 years ago [41, 57, 58].

Example run with Mimulus data

To illustrate a full run of ARPEGGIO, we analyzed publicly available data coming from the natural allopolyploid *Mimulus peregrinus* and its progenitors *M. guttatus* and *M. luteus* [59].

First, we downloaded the raw WGBS data consisting of four technical replicates for each species, the genome assemblies of the progenitors with their annotation and a chloroplast genome to check conversion efficiency (details in Availability of data and materials). For WGBS data, genome assemblies and annotations we made sure that all files were formatted according to ARPEGGIO's user guidelines.

Second, we created a metadata file specifying for each sample the sequencing strategy, single end, and the origin of the samples, i.e. *M. guttatus* samples were labeled "parent1", *M. luteus* samples "parent2" and *M. peregrinus* samples "allopolyploid".

With the input files ready, the configuration file was set up in two rounds. In the first round the general parameters were configured with the locations of output folder and metadata file, and data was specified as single end. By default, ARPEGGIO compares allopolyploid to progenitors (Fig. 1a), meaning that no specific changes needed to be done to include the experimental design for this dataset. Then, all conditional rules were set to false and ARPEGGIO was run to only perform quality checks. With this round we were able to get more details for the trimming step. In the second round, all the parameters were set for all the different steps in the workflow and all conditional rules were set to true to perform a full run of ARPEGGIO with eight cores. The

configuration file, the MultiQC report and ARPEGGIO's output for the statistical and downstream analyses can be found in Availability of data and materials. The runtime of the full run on a Debian system, using eight CPU cores Intel(R) Xeon(R) CPU E5–4640 at 2.40GHz was approximately 24 h. The average times for each step can be found in Additional File 2, where for each step, a per-sample average over twelve samples in total is shown, with the exception of statistical analyses for which the average runtime is per methylation context over three contexts in total.

After comparing the methylation pattern of *M. peregrinus* to its progenitors, a total of 760 significant DMRs were found in the allopolyploid, most of them coming from the *M. luteus* side (Table 2). Downstream analyses found very few genes overlapping with these significant regions, suggesting that most of the methylation changes occur in intergenic rather than genic regions. For the *M. guttatus* side, 35 genes were found, mostly associated with changes in CG and CHG context, while for the *M. luteus* side only 2 genes were found in CG context. These genes represent a very small proportion of the total number of annotated genes in *M. guttatus*, almost 30′000, and *M. luteus*, almost 50′000. Taken all together, these results suggest almost no change in the global methylation pattern of genes in the natural allopolyploid compared to the two progenitors.

Our analyses use a different approach and different tools compared to [54], but Edger and colleagues also looked at changes in methylation pattern from progenitor to allopolyploid. The authors observed were similar methylation patterns within gene bodies, when comparing progenitors to natural allopolyploids. This is consistent with ARPEGGIO's downstream analyses showing few genes overlapping with DMRs. Additionally, further analyses in [54] showed that most of the methylation changes happened in transposable elements, another result in agreement with the number of intergenic DMRs found by ARPEGGIO.

Table 2 Summary of ARPEGGIO's downstream analyses on the dataset from Edger and colleagues. The table is divided in two parts, one per progenitor. For each progenitor, the table shows the number of differentially methylated regions (DMRs) for each context, the number of genes overlapping with DMRs and the total number of genes found over all contexts

Methylation context	*Mimulus guttatus*			*Mimulus luteus*		
	CG	CHG	CHH	CG	CHG	CHH
DMRs	65	126	23	277	211	58
Total DMRs	214			546		
Genes overlapping DMRs	13	20	2	2	0	0
Total genes	35			2		

User's experience and best practices

ARPEGGIO's user documentation, available through the GitHub Wiki, offers additional information for more and less experienced users. For less experienced users, the documentation offers a step-by-step guide of how to setup and run ARPEGGIO on a given dataset: data and system requirements, input files needed, configuration file instructions, commands to run the workflow and a map of the output structure. For experienced users, we tried to be as transparent as possible about ARPEGGIO's code and its architecture to make any customization of scripts and code easier.

As a whole, ARPEGGIO is meant to simplify reproducible data analysis, but best practices, such as data diagnostics and information sharing should be kept in mind. The complete ARPEGGIO pipeline should be run once data quality and potential sources of errors are assessed. To have more control over the analysis process, users also have the option to run ARPEGGIO steps one by one. By modifying the configuration file to add further steps, the workflow will rerun only the parts that need to be updated. To ensure reproducibility when using ARPEGGIO, there are three specifications that need to be included with the datasets: the configuration file settings, the metadata file and the version of ARPEGGIO.

Software choice

Many alternative tools exist to perform some of ARPEGGIO's steps. For example, several aligners exist for short-read bisulfite sequencing data such as bwa-meth [60], BSmap [61], BitMapperBS [62], SNAP [63] and gemBS [64]. The Bismark suite was selected because it included tools to perform alignment, deduplication and methylation extraction for any context all in one centralized package. Most if not all of the other aligners depend on external packages for downstream analyses of alignment files.

Similarly, many tools exist for DMRs discovery in whole-genome bisulfite sequencing data for all methylation contexts: BSmooth [65], metilene [66], MOABS [67], BiSeq [68], MethylKit [69] and others [70].

In the case of dmrseq, the tool was chosen because of its two step approach: first selecting candidate regions and then evaluating their statistical significance by taking into account both biological variability and spatial correlation. This approach offers important advantages such as limited loss of power and better FDR control, both critical aspects when detecting DMRs [71].

The selection of an appropriate alignment or statistical tool for WGBS data would require an independent benchmark of such tools. An ideal benchmark should evaluate tools on a variety of conditions and provide some guidelines about their suitability and use. Currently, no such benchmarks exist, and a thorough

evaluation was out of the scope of this paper. ARPEGGIO provides a convenient implementation of the selected tools and its architecture allows future modifications as long as the input/output structure of the Snakemake rules is preserved.

This means that if any of the tools included in the workflow are shown to be underperforming compared to others, ARPEGGIO can be adapted accordingly.

Comparison to other workflows

To compare ARPEGGIO to other workflows, we selected key steps specifically related to WGBS data analysis (Table 3). The results included workflows able to work with raw bisulfite reads from WGBS and excluded highly specialized (i.e. alignment only or downstream only) and commercial workflows.

ARPEGGIO is the only workflow specifically targeted at polyploids, making it the main unique feature compared to other available workflows. Other features that were lacking in other workflows, but present in ARPEGGIO, were downstream analyses and reproducibility. Around half of the workflows investigated included downstream analyses [73, 74, 77, 79, 80]. The lack of this feature might be due to downstream analyses being highly variable according to biological context, question, and aim of the research. With ARPEGGIO, the aim was to consolidate performant tools into a common approach that could be used as a start for further investigation; in our case downstream analyses leading to a list of genes. Reproducibility was another main feature present in ARPEGGIO that was lacking in many workflows, but appeared to be more prevalent in more recent publications [73, 75, 80, 81]. Enhancing and promoting reproducibility is essential to ensure that discoveries stand the test of time [82]. Other features were very similar across workflows. All workflows support diploid data, which is considered the same as polyploid data with an available polyploid phased assembly. When comparing the presence of quality check, alignment and statistical analyses, most workflows included them all together, but some didn't include either quality check [75, 79, 80] or statistical analyses [76]. For methylation contexts, only two workflows focused on CpG context only [72, 81], while all the other allowed analyses for all contexts (CpG, CHG and CHH).

One feature not implemented in ARPEGGIO, but present in other workflows, is visualization of DMRs. This step, similar to downstream analyses, is highly context dependent. The dmrseq package offers ways to visualize DMRs, but this was not included in ARPE

Table 3 Comparison between ARPEGGIO and other available, non-commercial and general workflows able to work with raw WGBS data. There were a total of 12 workflows found and different features were selected for this comparison. The language indicates the main language(s) used to program the workflow. Polyploid support refers to support analysis of data from a polyploid with no official genome assembly available. Diploid support refers to analysis of data from a diploid or a polyploid with an available official genome assembly. Quality check, alignment, statistical and downstream analyses are all different steps in the data analysis process with downstream analyses being defined as follow-up analyses on DMRs found by the statistical analyses. Methylation contexts are 3 in total: CpG, CHG and CHH and this feature is sometimes limited to CpG only. Visualization represents any script or function allowing the user to visualize the DMRs found by the statistical analyses. Reproducibility is difficult to quantify and in this table a tool was considered reproducible if the corresponding paper mentioned reproducibility as one of their goals

	ARPEGGIO	**QUMA**	**MOABS**	**QuasR**	**MethPipe**	**bicycle**	**RUbioSeq**	**WBSA**	**P3BSseq**	**Methy-Pipe**	**MethFlow**	**snakePipes**
Language	Python, R	HTML, Perl, Javascript	C++, Perl	R	C++	Java	Perl	Perl, R	Python	Perl, R	Python, Perl, Java	Python, R
Polyploid support	✓	X	X	X	X	X	X	X	X	X	X	X
Diploid support	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Quality check	✓	✓	✓	✓	✓	X	✓	✓	✓	X	X	✓
Alignment	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Statistical analyses	✓	✓	✓	✓	✓	✓	X	✓	✓	✓	✓	✓
Methylation context	All	CpG only	All	All	All	All	All	All	All	All	?	CpG only
Downstream analyses	✓	X	X	✓	✓	X	X	✓	X	✓	✓	X
Visualization	X	X	X	✓	✓	X	X	✓	X	✓	✓	X
Reproducibility	✓	–	–	✓	–	✓	–	–	–	–	✓	✓
Paper	–	[72]	[67]	[73]	[74]	[75]	[76]	[77]	[78]	[79]	[80]	[81]

GGIO. Instead, the workflow outputs an Rdata file with all information concerning DMRs that users can use in their custom analyses. It is important to stress that visualization is essential for high-throughput data analysis, and should happen at any step in the data analysis process.

It is important to note that Table 3 focuses only on features related to WGBS data analysis, the only data type supported by ARPEGGIO. Some of the workflows support additional data types and analyses: QuasR supports ChIP-seq, RNA-seq, smRNA-seq and allele-specific data analyses, RUBioSeq supports single-nucleotide and copy number variants (SNVs and CNVs) analyses and snakePipes supports simple DNA-mapping, ChIP-seq, ATAC-seq, HiC, RNA-seq and scRNA-seq data.

Overall, ARPEGGIO was the only workflow supporting polyploid data, and among all the different aspects considered, one of the few workflows including downstream analyses that explicitly set reproducibility as one of its main goals.

Conclusions

Research on DNA methylation in allopolyploids at a whole genome level seems to be favoring established allopolyploid species (i.e. crops). This can be partially attributed to two factors: 1) challenges in generating allopolyploid genome assemblies; and, 2) a laborious data analysis process. Here we presented ARPEGGIO: the first workflow for the analysis of allopolyploid WGBS data. ARPEGGIO includes a read classification algorithm, EAGLE-RC, to assign allopolyploid reads to the correct progenitor's side. EAGLE-RC showed better performance against a common concatenation for six different WGBS datasets. Read classification is part of a full set of analyses included in ARPEGGIO, going from raw sequencing data up to a list of genes showing differential methylation. The implementation of ARPEGGIO aimed at ease of use and reproducibility, both essential factors to have an accessible yet up-to-standard tool.

With ARPEGGIO, we provide a first step towards a future of standardized tools and workflows in polyploid research.

Availability and requirements

Project name: ARPEGGIO

Project home page: https://github.com/supermaxiste/ARPEGGIO

Operating system: Linux

Programming language: Python and R

Other requirements: Python 3, Conda, [Singularity]

License: MIT

Abbreviations

WGBS: Whole genome bisulfite sequencing; DMRs: Differentially methylated regions

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12864-021-07845-2.

Additional file 1. Example of relationships between rules in ARPEGGIO. Description: A graph showing the input/output relationships between different rules in ARPEGGIO in an example "default" run with single end reads.

Additional file 2. Plot with average runtimes in ARPEGGIO with Mimulus data. Description: A plot with the average runtime for each main step in the ARPEGGIO pipeline: conversion check, quality check, trimming, alignment, deduplication, read classification, methylation extraction and statistical analyses. Each step shows a per sample average (12 samples in total), with the exception of the statistical analyses step where the average is per methylation context (3 contexts in total).

Additional file 3. Plant material and WGBS library synthesis. Description: Details about the plant conditions, sampling, DNA extraction, bisulfite treatment and sequencing strategies.

Additional file 4. Read statistics about datasets used to compare EAGLE-RC against concatenation method. Description: All the numbers related to the datasets used to compare EAGLE-RC to the concatenation method: total reads, uniquely mapped reads (and not), duplicated reads, correct, ambiguous and wrongly classified reads and error rate.

Acknowledgements

We thank A. Morishima and M. Wyler for all the support in the DNA extraction, bisulfite treatment and library preparation steps and the optimization of the protocol; the Functional Genomic Center Zurich and M. Hatakeyama for sequencing and data handling; the URPP Evolution in Action program for the opportunity to present ARPEGGIO; the Robinson and Shimizu group members for the feedback during different stages of the project, in particular R. Huang, K. Hembach, S. Orjuela and C. Soneson for the support with Snakemake and the workflow development process.

Authors' contributions

SM started the bisulfite treatment and library preparation protocol optimization, wrote most of the ARPEGGIO code and manuscript and tested EAGLE-RC. TK updated EAGLE-RC to a new version supporting bisulfite sequencing data and contributed to describe the model with its new features. TK, SD and MR tested ARPEGGIO on their own devices. SD helped with bug fixing in ARPEGGIO, optimized Conda support and implemented Singularity support. RSI managed, coordinated and optimized the bisulfite treatment protocol, library preparation and sequencing process of the *Arabidopsis* samples. LM executed the bisulfite treatment and library preparation. JS, MR, RSI and KS supervised the project and provided important insights at several stages of the project. All authors read and approved the final manuscript.

Funding

This work was supported by the University Research Priority Program (URPP) Evolution in Action of the University of Zurich, a JST CREST JPMJCR16O3, a Swiss National Science Foundation 31003A_182318 and MEXT KAKENHI 16H06469.

Availability of data and materials

Data from cotton taken from [29], available in the NCBI Nucleotide and Sequence Read Archive (SRA) under [SRA:SRP071640]. Data from *Mimulus* taken from [59], available in the NCBI Gene Expression Omnibus (GEO) under [GSE95799]. Data from *Arabidopsis* available in the DDBJ Sequence Read Archive (DRA) under [DRA009902].

The *Gossypium raimondii* v2.0 genome assembly [83] and *Mimulus guttatus* v2.0 [84] genome assembly and annotation were downloaded from Pythozome v12.1 [85]. The *Gossypium arboreum* v2_a1 [86] genome assembly was downloaded from CottonGen [87]. The *Mimulus luteus* [54] assembly and its annotation were downloaded from Dryad [59]. The *Arabidopsis halleri* v2.2 genome assembly was taken from [57] and the *Arabidopsis lyrata* v2.2 genome assembly was taken from [58].

The scripts used for the evaluation of EAGLE-RC and genome concatenation together with the details about the *Mimulus* example run can be found on: https://github.com/supermaxiste/ARPEGGIO_paperAnalyses

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Author details
[1]Department of Evolutionary Biology and Environmental Studies, University of Zurich, Zurich, Switzerland. [2]SIB Swiss Institute of Bioinformatics, University of Zurich, Zurich, Switzerland. [3]Centre for Biodiversity Genomics, University of Guelph, Guelph, Canada. [4]Department of Computer Science, ETH Zurich, Zurich, Switzerland. [5]AIST Artificial Intelligence Research Center, Tokyo, Japan. [6]Humanome Lab Inc., Chuo-ku, Tokyo, Japan. [7]Kihara Institute for Biological Research, Yokohama City University, Yokohama, Japan. [8]Department of Molecular Life Sciences, University of Zurich, Zurich, Switzerland.

Received: 6 August 2020 Accepted: 23 June 2021
Published online: 17 July 2021

References

1. Van de Peer Y, Mizrachi E, Marchal K. The evolutionary significance of polyploidy. Nat Rev Genet. 2017;18(7):411–24.
2. Blischak PD, Mabry ME, Conant GC, Pires JC. Integrating networks, Phylogenomics, and population genomics for the study of polyploidy. Annu Rev Ecol Evol Syst. 2018;49(1):253–78.
3. One thousand plant transcriptomes and the phylogenomics of green plants. Nature. 2019;574(7780):679–85.
4. Soltis DE, Visger CJ, Marchant DB, Soltis PS. Polyploidy: pitfalls and paths to a paradigm. Am J Bot. 2016;103(7):1146–66.
5. Soltis PS, Soltis DE. Ancient WGD events as drivers of key innovations in angiosperms. Curr Opin Plant Biol. 2016;30:159–65.
6. Clark JW, Donoghue PCJ. Whole-genome duplication and plant macroevolution. Trends Plant Sci. 2018;23(10):933–45.
7. Wood TE, Takebayashi N, Barker MS, Mayrose I, Greenspoon PB, Rieseberg LH. The frequency of polyploid speciation in vascular plants. Proc Natl Acad Sci. 2009;106(33):13875–9.
8. Mayrose I, Zhan SH, Rothfels CJ, Magnuson-Ford K, Barker MS, Rieseberg LH, et al. Recently formed polyploid plants diversify at lower rates. Science. 2011;333(6047):1257.
9. Soltis DE, Buggs RJA, Barbazuk WB, Chamala S, Chester M, Gallagher JP, et al. The early stages of polyploidy: rapid and repeated evolution in tragopogon. In: Polyploidy and genome evolution. Berlin: Springer Berlin Heidelberg; 2012. p. 271–92.
10. Chen ZJ. Genetic and epigenetic mechanisms for gene expression and phenotypic variation in plant Polyploids. Annu Rev Plant Biol. 2007;58(1):377–406.
11. Wendel JF, Lisch D, Hu G, Mason AS. The long and short of doubling down: polyploidy, epigenetics, and the temporal dynamics of genome fractionation. Curr Opin Genet Dev. 2018;49:1–7.
12. Wendel JF. Genome evolution in polyploids. In: Plant molecular evolution. Dordrecht: Springer Netherlands; 2000. p. 225–49.
13. Madlung A, Masuelli RW, Watson B, Reynolds SH, Davison J, Comai L. Remodeling of DNA methylation and phenotypic and transcriptional changes in synthetic arabidopsis allotetraploids. Plant Physiol. 2002;129(2):733–46.
14. Salmon A, Ainouche ML, Wendel JF. Genetic and epigenetic consequences of recent hybridization and polyploidy in Spartina (Poaceae). Mol Ecol. 2005;14(4):1163–75.
15. Xu Y, Zhong L, Wu X, Fang X, Wang J. Rapid alterations of gene expression and cytosine methylation in newly synthesized Brassica napus allopolyploids. Planta. 2009;229(3):471–83.
16. Shaked H, Kashkush K, Ozkan H, Feldman M, Levy AA. Sequence elimination and cytosine methylation are rapid and reproducible responses of the genome to wide hybridization and Allopolyploidy in wheat. Plant Cell. 2001;13(8):1749–59.
17. Sehrish T, Symonds VV, Soltis DE, Soltis PS, Tate JA. Gene silencing via DNA methylation in naturally occurring Tragopogon miscellus (Asteraceae) allopolyploids. BMC Genomics. 2014;15(1):1–7.
18. Ran L, Fang T, Rong H, Jiang J, Fang Y, Wang Y. Analysis of cytosine methylation in early generations of resynthesized Brassica napus. J Integr Agric. 2016;15(6):1228–38.
19. Bao Y, Xu Q. Extensive reprogramming of cytosine methylation in Oryza allotetraploids. Genes Genomics. 2015;37(6):517–24.
20. Parisod C, Salmon A, Zerjal T, Tenaillon M, Grandbastien M-A, Ainouche M. Rapid structural and epigenetic reorganization near transposable elements in hybrid and allopolyploid genomes in Spartina. New Phytol. 2009;184(4):1003–15.
21. Wang J, Tian L, Madlung A, Lee H-S, Chen M, Lee JJ, et al. Stochastic and epigenetic changes of gene expression in Arabidopsis Polyploids. Genetics. 2004;167(4):1961–73.
22. Gaeta RT, Pires JC, Iniguez-Luy F, Leon E, Osborn TC. Genomic changes in resynthesized Brassica napus and their effect on gene expression and phenotype. Plant Cell. 2007;19(11):3403–17.
23. Liu B, Brubaker CL, Mergeai G, Cronn RC, Wendel JF. Polyploid formation in cotton is not accompanied by rapid genomic changes. Genome. 2001;44(3):321–30.
24. Kurdyukov S, Bullock M. DNA methylation analysis: choosing the right method. Biology (Basel). 2016;5(1):3.
25. Laird PW. Principles and challenges of genome-wide DNA methylation analysis. Nat Rev Genet. 2010;11(3):191–203.
26. Law JA, Jacobsen SE. Establishing, maintaining and modifying DNA methylation patterns in plants and animals. Nat Rev Genet. 2010;11(3):204–20.
27. Urich MA, Nery JR, Lister R, Schmitz RJ, Ecker JR. MethylC-seq library preparation for base-resolution whole-genome bisulfite sequencing. Nat Protoc. 2015;10(3):475–83.
28. Li N, Xu C, Zhang A, Lv R, Meng X, Lin X, et al. DNA methylation repatterning accompanying hybridization, whole genome doubling and homoeolog exchange in nascent segmental rice allotetraploids. New Phytol. 2019;223(2):979–92.
29. Song Q, Zhang T, Stelly DM, Chen ZJ. Epigenomic and functional analyses reveal roles of epialleles in the loss of photoperiod sensitivity during domestication of allotetraploid cottons. Genome Biol. 2017;18(1):99.
30. Bird KA, Niederhuth C, Ou S, Gehan M, Chris Pires J, Xiong Z, et al. Replaying the evolutionary tape to investigate subgenome dominance in allopolyploid Brassica napus. bioRxiv. 2019:814491.
31. Kersey PJ. Plant genome sequences: past, present, future. Curr Opin Plant Biol. 2019;48:1–8. https://doi.org/10.1016/j.pbi.2018.11.001.
32. Claros MG, Bautista R, Guerrero-Fernández D, Benzerki H, Seoane P, Fernández-Pozo N. Why assembling plant genome sequences is so challenging. Biology (Basel). 2012;1(2):439–59.
33. Kyriakidou M, Tai HH, Anglin NL, Ellis D, Strömvik MV. Current strategies of polyploid plant genome sequence assembly. Front Plant Sci. 2018;9.
34. Kuo TCY, Hatakeyama M, Tameshige T, Shimizu KK, Sese J. Homeolog expression quantification methods for allopolyploids. Brief Bioinform. 2018.
35. Bock C. Analysing and interpreting DNA methylation data. Nat Rev Genet. 2012;13(10):705–19.
36. Yong W-S, Hsu F-M, Chen P-Y. Profiling genome-wide DNA methylation. Epigenetics Chromatin. 2016;9(1):26.
37. Wreczycka K, Gosdschan A, Yusuf D, Grüning B, Assenov Y, Akalin A. Strategies for analyzing bisulfite sequencing data. J Biotechnol. 2017;261:105–15.
38. Boatwright JL, McIntyre LM, Morse AM, Chen S, Yoo M-J, Koh J, et al. A robust methodology for assessing differential Homeolog contributions to the transcriptomes of allopolyploids. Genetics. 2018;210(3):883–94.
39. Gerard D, Ferrão LFV, Garcia AAF, Stephens M. Genotyping Polyploids from messy sequencing data. Genetics. 2018;210(3):789–807.
40. Page JT, Udall JA. Methods for mapping and categorization of DNA sequence reads from allopolyploid organisms. BMC Genet. 2015;16(Suppl 2):S4.
41. Akama S, Shimizu-Inatsugi R, Shimizu KK, Sese J. Genome-wide quantification of homeolog expression ratio revealed nonstochastic gene regulation in synthetic allopolyploid Arabidopsis. Nucleic Acids Res. 2014;42(6):e46–e46.
42. Page JT, Gingle AR, Udall JA. PolyCat: a resource for genome categorization of sequencing reads from allopolyploid organisms. G3. 2013;3(3):517–25.

43. Hu G, Grover CE, Arick MA, Liu M, Peterson DG, Wendel JF. Homoeologous gene expression and co-expression network analyses and evolutionary inference in allopolyploids. Brief Bioinform. 2020.
44. Garvin MR, Saitoh K, Gharrett AJ. Application of single nucleotide polymorphisms to non-model species: a technical review. Mol Ecol Resour. 2010;10(6):915–34.
45. Koster J, Rahmann S. Snakemake--a scalable bioinformatics workflow engine. Bioinformatics. 2012;28(19):2520–2.
46. Anaconda. Anaconda Software Distribution. 2014. Available from: https://anaconda.com.
47. International Organization for Standardization. Information processing — Documentation symbols and conventions for data, program and system flowcharts, program network charts and system resources charts; 1985. p. 25. Available from: https://www.iso.org/standard/11955.html. Cited 2019 Dec 19
48. Kurtzer GM, Sochat V, Bauer MW. Singularity: scientific containers for mobility of compute. PLoS One. 2017;12(5):e0177459.
49. Krueger F, Andrews SR. Bismark: a flexible aligner and methylation caller for bisulfite-Seq applications. Bioinformatics. 2011;27(11):1571–2.
50. Andrews S. FastQC: a quality control tool for high throughput sequence data. 2010. Available from: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/.
51. Krueger F. Trim galore. 2012. Available from: http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/.
52. Korthauer K, Chakraborty S, Benjamini Y, Irizarry RA. Detection and accurate false discovery rate control of differentially methylated regions from whole genome bisulfite sequencing. Biostatistics. 2019;20(3):367–83.
53. Ewels P, Magnusson M, Lundin S, Käller M. MultiQC: summarize analysis results for multiple tools and samples in a single report. Bioinformatics. 2016;32(19):3047–8.
54. Edger PP, Smith R, McKain MR, Cooley AM, Vallejo-Marin M, Yuan Y, et al. Subgenome dominance in an interspecific hybrid, synthetic allopolyploid, and a 140-year-old naturally established neo-allopolyploid monkeyflower. Plant Cell. 2017;29(9):2150–67.
55. Shimizu-Inatsugi R, Lihová J, Iwanaga h, kudoh h, Marhold K, Savolainen O, et al. The allopolyploid Arabidopsis kamchatica originated from multiple individuals of Arabidopsis lyrata and Arabidopsis halleri. Mol Ecol. 2009;18(19):4024–48.
56. Rodriguez-R L, Konstantinidis K. The enveomics collection: a toolbox for specialized analyses of microbial genomes and metagenomes. PeerJ Prepr. 2016;4:e1900v1.
57. Briskine RV, Paape T, Shimizu-Inatsugi R, Nishiyama T, Akama S, Sese J, et al. Genome assembly and annotation of Arabidopsis halleri, a model for heavy metal hyperaccumulation and evolutionary ecology. Mol Ecol Resour. 2017;17(5):1025–36.
58. Paape T, Briskine RV, Halstead-Nussloch G, Lischer HEL, Shimizu-Inatsugi R, Hatakeyama M, et al. Patterns of polymorphism and selection in the subgenomes of the allopolyploid Arabidopsis kamchatica. Nat Commun. 2018;9(1):3909.
59. Edger PP, Smith RD, McKain MR, Cooley AM, Vallejo-Marin M, Yuan Y-W, et al. Data from: subgenome dominance in an interspecific hybrid, synthetic allopolyploid, and a 140-year-old naturally established neo-allopolyploid monkeyflower: Dryad; 2017. Available from: https://datadryad.org/stash/dataset/doi:10.5061/dryad.d4vr0
60. Pedersen BS, Eyring K, De S, Yang I V., Schwartz DA. Fast and accurate alignment of long bisulfite-seq reads. 2014.
61. Xi Y, Li W. BSMAP: whole genome bisulfite sequence MAPping program. BMC Bioinformatics. 2009;10(1):232.
62. Cheng H, Xu Y. BitMapperBS: a fast and accurate read aligner for whole-genome bisulfite sequencing. bioRxiv. 2018;442798.
63. Zaharia M, Bolosky WJ, Curtis K, Fox A, Patterson D, Shenker S, et al. Faster and more accurate sequence alignment with SNAP. 2011.
64. Merkel A, Fernández-Callejo M, Casals E, Marco-Sola S, Schuyler R, Gut IG, et al. gemBS: high throughput processing for DNA methylation data from bisulfite sequencing. Bioinformatics. 2019;35(5):737–42.
65. Hansen KD, Langmead B, Irizarry RA. BSmooth: from whole genome bisulfite sequencing reads to differentially methylated regions. Genome Biol. 2012;13(10):R83.
66. Jühling F, Kretzmer H, Bernhart SH, Otto C, Stadler PF, Hoffmann S. metilene: fast and sensitive calling of differentially methylated regions from bisulfite sequencing data. Genome Res. 2016;26(2):256–62.
67. Sun D, Xi Y, Rodriguez B, Park H, Tong P, Meong M, et al. MOABS: model based analysis of bisulfite sequencing data. Genome Biol. 2014;15(2):R38.
68. Hebestreit K, Dugas M, Klein H-U. Detection of significantly differentially methylated regions in targeted bisulfite sequencing data. Bioinformatics. 2013;29(13):1647–53.
69. Akalin A, Kormaksson M, Li S, Garrett-Bakelman FE, Figueroa ME, Melnick A, et al. MethylKit: a comprehensive R package for the analysis of genome-wide DNA methylation profiles. Genome Biol. 2012;13(10):R87.
70. Shafi A, Mitrea C, Nguyen T, Draghici S. A survey of the approaches for identifying differential methylation using bisulfite sequencing data. Brief Bioinform. 2017;(January):1–17.
71. Robinson MD, Kahraman A, Law CW, Lindsay H, Nowicka M, Weber LM, et al. Statistical methods for detecting differentially methylated loci and regions. Front Genet. 2014;5.
72. Kumaki Y, Oda M, Okano M. QUMA: quantification tool for methylation analysis. Nucleic Acids Res. 2008;36(Web Server):W170–5.
73. Gaidatzis D, Lerch A, Hahne F, Stadler MB. QuasR: quantification and annotation of short reads in R. Bioinformatics. 2015;31(7):1130–2.
74. Song Q, Decato B, Hong EE, Zhou M, Fang F, Qu J, et al. A reference methylome database and analysis pipeline to facilitate integrative and comparative epigenomics. PLoS One. 2013;8(12):e81148.
75. Graña O, López-Fernández H, Fdez-Riverola F, González Pisano D, Glez-Peña D. Bicycle: a bioinformatics pipeline to analyze bisulfite sequencing data. Bioinformatics. 2018;34(8):1414–5.
76. Rubio-Camarillo M, Gomez-Lopez G, Fernandez JM, Valencia A, Pisano DG. RUbioSeq: a suite of parallelized pipelines to automate exome variation and bisulfite-seq analyses. Bioinformatics. 2013;29(13):1687–9.
77. Liang F, Tang B, Wang Y, Wang J, Yu C, Chen X, et al. WBSA: web service for bisulfite sequencing data analysis. PLoS One. 2014;9(1):e86707.
78. Luu P-L, Gerovska D, Arrospide-Elgarresta M, Retegi-Carrión S, Schöler HR, Araúzo-Bravo MJ. P3BSseq: parallel processing pipeline software for automatic analysis of bisulfite sequencing data. Bioinformatics. 2016:btw633.
79. Jiang P, Sun K, Lun FMF, Guo AM, Wang H, Chan KCA, et al. Methy-Pipe: an integrated bioinformatics pipeline for whole genome bisulfite sequencing data analysis. PLoS One. 2014;9(6):e100360.
80. Lebrón R, Barturen G, Gómez-Martín C, Oliver JL, Hackenberg M. MethFlowVM: a virtual machine for the integral analysis of bisulfite sequencing data. bioRxiv. 2016:66795.
81. Bhardwaj V, Heyne S, Sikora K, Rabbani L, Rauer M, Kilpert F, et al. snakePipes: facilitating flexible, scalable and integrative epigenomic analysis. Bioinformatics. 2019;35(22):4757–9.
82. Nekrutenko A, Taylor J. Next-generation sequencing data interpretation: enhancing reproducibility and accessibility. Nat Rev Genet. 2012;13(9):667–72.
83. Paterson AH, Wendel JF, Gundlach H, Guo H, Jenkins J, Jin D, et al. Repeated polyploidization of Gossypium genomes and the evolution of spinnable cotton fibres. Nature. 2012;492(7429):423–7.
84. Hellsten U, Wright KM, Jenkins J, Shu S, Yuan Y, Wessler SR, et al. Fine-scale variation in meiotic recombination in Mimulus inferred from population shotgun sequencing. Proc Natl Acad Sci. 2013;110(48):19478–82.
85. Goodstein DM, Shu S, Howson R, Neupane R, Hayes RD, Fazo J, et al. Phytozome: a comparative platform for green plant genomics. Nucleic Acids Res. 2012;40(D1):D1178–86.
86. Li F, Fan G, Wang K, Sun F, Yuan Y, Song G, et al. Genome sequence of the cultivated cotton Gossypium arboreum. Nat Genet. 2014;46(6):567–72.
87. Yu J, Jung S, Cheng C-H, Ficklin SP, Lee T, Zheng P, et al. CottonGen: a genomics, genetics and breeding database for cotton research. Nucleic Acids Res. 2014;42(D1):D1229–36.

Publisher's Note

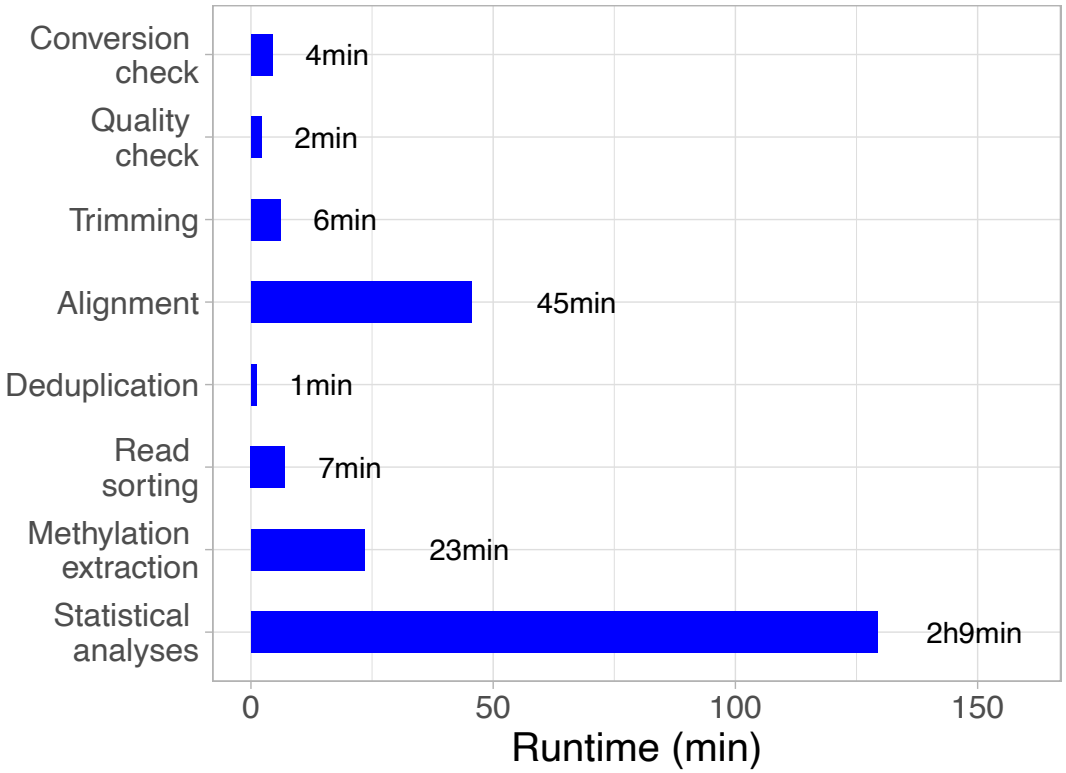
Supplementary Information

Additional file 1.

Example of relationships between rules in ARPEGGIO. Description: A graph showing the input/output relationships between different rules in ARPEGGIO in an example “default” run with single end reads.

Additional file 2.

Plot with average runtimes in ARPEGGIO with Mimulus data. Description: A plot with the average runtime for each main step in the ARPEGGIO pipeline: conversion check, quality check, trimming, alignment, deduplication, read classification, methylation extraction and statistical analyses. Each step shows a per sample average (12 samples in total), with the exception of the statistical analyses step where the average is per methylation context (3 contexts in total).



Additional file 3.

Plant material and WGBS library synthesis. Description: Details about the plant conditions, sampling, DNA extraction, bisulfite treatment and sequencing strategies.

Plant material and WGBS library synthesis

All plant individuals were incubated in a climate chamber set at 22˚C , with 60% relative humidity, 16 hours light/8hours dark cycles. Mature leaves were collected from each individual, and DNA samples were extracted by DNeasy Plant Mini Kit (Qiagen) from 3 individuals of each species. For the samples *A. halleri* 1 and *A. lyrata* 1, DNA was first treated by bisulfite (BS) using MethylEdge™ Bisulfite Conversion System (Promega), and sequencing library was synthesized by using TruSeq DNA Methylation Kit (Illumina). For the samples *A. halleri* G1 and *A. lyrata* G1, DNA was first synthesized to sequencing library using KAPA HyperPrep Kit with TruSeq DNA Single Indexes Set (Illumina), and BS-treated by EZ DNA Methylation-Gold Kit (ZYMO Research).
The libraries were paired-end sequenced by Illumina HiSeq 4000 (126bp x 2, *A. halleri* 1 and *A. lyrata* 1) and NovaSeq 6000 (150bp x 2, *A. halleri* G1 and *A. lyrata* G1).

Additional file 4.

Read statistics about datasets used to compare EAGLE-RC against concatenation method.
Description: All the numbers related to the datasets used to compare EAGLE-RC to the concatenation method: total reads, uniquely mapped reads (and not), duplicated reads, correct, ambiguous and wrongly classified reads and error rate.

Read statistics about datasets used to compare EAGLE-RC against concatenation method

The following tables provide all the numbers behind the comparison between EAGLE-RC against the concatenation method. Here's a list with all the samples used and their corresponding accession number:

- *A. halleri*: SAMD00208469
- *A. lyrata*: SAMD00208470
- *A. halleri* G1: SAMD00208471
- *A. lyrata* G1: SAMD00208472
- *M. luteus* 1, 2 , 3 , 4: SRX2618908, SRX2618909, SRX2618910, SRX2618911
- *M. guttatus* 1 , 2 , 3 , 4: SRX2618912, SRX2618913, SRX2618914, SRX2618915
- *G. arboreum* 1 , 2: SRR3219104, SRR3219105
- *G. raimondii* 1, 2: SRR3219088, SRR3219089

	A. halleri	*A. lyrata*	*A. halleri*	*A. lyrata*
Method	concatenated		EAGLE-RC based classification	
Total reads (TR)	40'160'266	44'895'570	40'160'266	44'895'570
Uniquely mapped (% from TR)	21'012'844 (52.3%)	23'067'388 (51.4%)	22'119'312 (55.3%)	24'433'418 (54.7%)
Not uniquely mapped	19'147'422	21'828'182	18'040'954	20'462'152
Duplicated reads (% from TR)	5'293'580 (13.2%)	7'585'268 (16.9%)	5'338'990 (24.1%)	7'897'614 (32.3%)
Uniquely mapped and deduplicated (UMD)	15'719'264	15'482'120	16'780'322	16'535'804
Correct reads (% from UMD)	15'013'508 (95.5%)	14'418'656 (91.0%)	14'953'178 (89.1%)	14'417'190 (87.2%)
Ambiguous	-	-	1'633'420	1'859'612
Wrong reads	705'756	1'063'464	193'724	259'002
Error %	4.49 %	6.87%	1.30 %	1.57 %

	A. halleri G1	*A. lyrata* G1	*A. halleri* G1	*A. lyrata G1*
Method	concatenated		EAGLE-RC based classification	
Total reads (TR)	80'445'492	124'215'824	80'445'492	124'215'824
Uniquely mapped (% from TR)	20'321'978 (25.3%)	31'692'092 (25.5%)	21'468'920 (26.7%)	32'969'572 (26.6%)
Not uniquely mapped	60'123'514	92'523'732	58'976'572	91'246'252
Duplicated reads (% from TR)	1'523'726 (1.9%)	2'765'528 (2.2%)	1'626'582 (7.6%)	2'903'264 (8.8%)
Uniquely mapped and deduplicated (UMD)	18'798'252	28'926'564	19'842'338	30'066'308
Correct reads (% from UMD)	18'147'370 (96.5%)	27'480'258 (95.0%)	18'625'582 (93.9%)	28'269'980 (94.0%)
Ambiguous	-	-	1'016'786	1'399'008
Wrong reads	650'882	1'446'306	199'970	397'320
Error %	3.46 %	5.00%	1.01 %	1.32 %

	Mimulus guttatus 1	*Mimulus luteus* 1	*Mimulus guttatus* 1	*Mimulus luteus* 1
Method	concatenated		EAGLE-RC based classification	
Total reads (TR)	11'198'784	14'540'898	11'198'784	14'540'898
Uniquely mapped (% from TR)	2'612'195 (23.3%)	6'312'383 (43.4%)	2'575'518 (23.0%)	5'908'489 (40.6%)
Not uniquely mapped	8'586'589	8'228'515	8'623'266	8'632'409
Duplicated reads (% from TR)	1'151'447 (44.1%)	2'300'413 (36.44%)	1'249'814 (48.5%)	2'027'318 (34.3%)
Uniquely mapped and deduplicated (UMD)	1'460'748	4'011'970	1'325'704	3'881'171
Correct reads (% from UMD)	1'069'658 (73.2%)	3'623'197 (90.3%)	976'317 (73.6%)	3'236'631 (83.4%)
Ambiguous	-	-	250'067	556'666
Wrong reads	391'090	388'773	99'320	87'874
Error %	26.77%	9.69%	7.49%	2.26%

	Mimulus guttatus 2	*Mimulus luteus* 2	*Mimulus guttatus* 2	*Mimulus luteus* 2
Method	concatenated		EAGLE-RC based classification	
Total reads (TR)	11'125'236	14'423'855	11'125'236	14'423'855
Uniquely mapped (% from TR)	2'608'418 (23.4%)	6'291'151 (43.6%)	2'571'997 (23.1%)	5'889'892 (40.8%)
Not uniquely mapped	8'516'818	8'132'704	8'553'239	8'533'963
Duplicated reads (% from TR)	1'149'555 (44.1%)	2'297'202 (36.5%)	1'248'440 (48.5%)	2'026'983 (34.4%)
Uniquely mapped and deduplicated (UMD)	1'458'863	3'993'949	1'323'557	3'862'909
Correct reads (% from UMD)	1'067'813 (73.2%)	3'606'521 (90.3%)	974'850 (73.7%)	3'219'702 (83.3%)
Ambiguous	-	-	249'601	555'449
Wrong reads	391'050	387'428	99'106	87'758
Error %	26.81%	9.70%	7.49%	2.27%

	Mimulus guttatus 3	*Mimulus luteus* 3	*Mimulus guttatus* 3	*Mimulus luteus* 3
Method	concatenated		EAGLE-RC based classification	
Total reads (TR)	10'955'877	14'136'379	10'955'877	14'136'379
Uniquely mapped (% from TR)	2'494'759 (22.8%)	6'016'962 (42.6%)	2'459'689 (22.5%)	5'629'101 (39.8%)
Not uniquely mapped	8'461'118	8'119'417	8'496'188	8'507'278
Duplicated reads (% from TR)	1'098'495 (44.0%)	2'193'358 (36.5%)	1'192'556 (48.5%)	1'933'785 (34.4%)
Uniquely mapped and deduplicated (UMD)	1'396'264	3'823'604	1'267'133	3'695'316
Correct reads (% from UMD)	1'022'678 (73.2%)	3'446'647 (90.1%)	932'804 (73.6%)	3'083'952 (83.5%)
Ambiguous	-	-	238'974	526'097
Wrong reads	373'586	376'957	95'355	85'267
Error %	26.76%	9.86%	7.53%	2.31%

	Mimulus guttatus 4	*Mimulus luteus* 4	*Mimulus guttatus* 4	*Mimulus luteus* 4
Method	concatenated		EAGLE-RC based classification	
Total reads (TR)	10'646'892	13'738'854	10'646'892	13'738'854
Uniquely mapped (% from TR)	2'433'041 (22.9%)	5'866'025 (42.7%)	2'398'688 (22.5%)	5'485'383 (39.9%)
Not uniquely mapped	8'213'851	7'872'829	8'247'915	8'253'471
Duplicated reads (% from TR)	1'068'453 (43.9%)	2'137'718 (36.4%)	1'159'882 (48.4%)	1'882'322 (34.3%)
Uniquely mapped and deduplicated (UMD)	1'364'588	3'728'307	1'238'806	3'603'061
Correct reads (% from UMD)	998'982 (73.2%)	3'357'807 (90.1%)	911'356 (73.6%)	3'003'804 (83.4%)
Ambiguous	-	-	233'955	515'291
Wrong reads	365'606	370'500	93'495	83'966
Error %	26.79%	9.94%	7.55%	2.33%

	Gossypium arboreum 1	*Gossypium raimondii* 1	*Gossypium arboreum* 1	*Gossypium raimondii* 1
Method	concatenated		EAGLE-RC based classification	
Total reads (TR)	432'844'852	356'699'260	432'844'852	356'699'260
Uniquely mapped (% from TR)	279'996'748 (64.7%)	273'814'708 (76.8%)	280'386'284 (64.8%)	278'732'236 (78.1%)
Not uniquely mapped	152'848'104	82'884'552	152'458'568	77'967'024
Duplicated reads (% from TR)	18'846'046 (6.7%)	13'284'372 (4.8%)	18'879'306 (6.7%)	13'697'054 (4.9%)
Uniquely mapped and deduplicated (UMD)	261'150'702	260'530'336	261'506'978	265'035'182
Correct reads (% from UMD)	260'132'278 (99.6%)	259'423'814 (99.6%)	260'022'724 (99.4%)	262'945'068 (99.2%)
Ambiguous	-		1'132'490	1'613'104
Wrong reads	1'018'424	1'106'522	351'764	477'010
Error %	0.00390%	0.00425%	0.00134%	0.00179%

	Gossypium arboreum 2	*Gossypium raimondii* 2	*Gossypium arboreum* 2	*Gossypium raimondii* 2
Method	concatenated		EAGLE-RC based classification	
Total reads (TR)	414’743’906	299’026’128	414’743’906	299’026’128
Uniquely mapped (% from TR)	264’247’068 (63.7%)	235’500’124	264’620’650 (63.8%)	240’038’096 (80.3%)
Not uniquely mapped	150’496’838	63’526’004	150’123’256	58’988’032
Duplicated reads (% from TR)	17’572’436 (4.2%)	10’850’322 (4.6%)	17’604’224 (4.2%)	11’202’082 (4.7%)
Uniquely mapped and deduplicated (UMD)	246’674’632	224’649’802	247’016’426	228’836’014
Correct reads (% from UMD)	245’718’300 (99.6%)	223’623’540 (99.5%)	245’626’594 (99.4%)	226’937’338 (99.2%)
Ambiguous	-	-	1’059’410	1’452’404
Wrong reads	956’332	1’026’262	330’422	446’272
Error %	0.00389%	0.00457%	0.00134%	0.00195%

Chapter 2: Environmental stress contributes to diverging DNA methylation response in early stages of polyploidy

Stefan Milosavljevic[1,2], Aki Morishima[1], Lucas Mohn[1], Jun Sese[3,4], Kentaro K. Shimizu[1,5], Mark D. Robinson[2,6] and Rie Shimizu Inatsugi[1]

[1] Department of Evolutionary Biology and Environmental Studies, University of Zurich, Zurich, Switzerland
[2] SIB Swiss Institute of Bioinformatics, University of Zurich, Zurich, Switzerland
[3] AIST Artificial Intelligence Research Center, Tokyo, Japan
[4] Humanome Lab Inc., Chuo-ku, Tokyo, Japan
[5] Kihara Institute for Biological Research, Yokohama City University, Yokohama, Japan
[6] Department of Molecular Life Sciences, University of Zurich, Zurich, Switzerland

Contributions

R.S.-I. conceived and designed the study; S.M. analyzed the data and wrote the manuscript; R.S.-I., A.M. and L.M. conducted and maintained the experiment; A.M. and L.M. performed all wet lab work including sample preparation for sequencing; R.S.-I., J.S, K.K.S. and M.D.R. provided feedback, suggestions and guidance throughout the project

Abstract

Polyploidy should be a major evolutionary force in the evolutionary history of plants even though the reasons of its prevalence are still under investigation. From a genomic point of view, polyploidy has been associated to many short term effects, with most studies focusing on expression changes and few of them exploring underlying mechanism potentially driving these changes, such as DNA methylation. Current knowledge on DNA methylation in early stages of polyploidy comes mostly from synthetic allopolyploid species, where studies with high-throughput sequencing technologies are sparse. In addition, synthetic species are always grown in conditions where environmental stress is excluded, although environmental conditions were hypothesized to play a critical role in the establishment and success of polyploids. To address all of these gaps, we used the *Arabidopsis kamchatica* study system to examine the combined effect of polyploidization and environment on DNA methylation in synthetic individuals with respect to their progenitor and natural species. In our experimental design, all plants were grown and propagated for several generations in two conditions, one mild (cold) and one stressful (hot). Our study showed how different environmental conditions lead to diverging DNA methylation patterns in synthetic polyploids with respect to their progenitors and to each other. Additionally we showed how synthetics and natural species converge in methylation pattern. Our findings emphasize the importance of environmental stress in the early stages of polyploidy and DNA methylation as a mechanism for polyploids to adapt to the environment.

Introduction

In the evolutionary history of land plants, whole genome duplication (WGD), also known as polyploidization, is ubiquitous. This prevalence is especially pronounced in flowering plants, one of the largest known clades in plants in terms of species (1), where on average 3-4 WGD were inferred in most lineages (2). Ancient WGD events were also associated to diversification of genes responsible for key traits such as seed and flower development (3). Polyploidy events are distinguished into two groups: autopolyploidy and allopolyploidy. Autopolyploid species are formed after a single species undergoes duplication of its own set of chromosomes, while allopolyploids are formed after hybridization between two different species followed by WGD (4). The proportion of allo- and autopolyploid species was estimated to be similar in plants, with most of the studies about polyploidy come from allopolyploids (5,6). These studies span genetic, genomic and ecological aspects of polyploids, forming a large body of literature (5).

In this study, we will focus on the short-term genomic effects of allopolyploidy right after polyploidization, specifically DNA methylation. One of the most extensively studied aspects in polyploids is gene expression change, where many newly formed polyploids tend to deviate from the sum of the expression patterns from their progenitors (7–9). These rapid expression changes in the short term can have an effect in the long-term as well (10), as shown in *Glycine dolichocarpa* where translational regulation was correlated with long-term retention of genes (11). With gene expression changes playing an important role from the formation to the establishment of polyploid species, increasing attention was given to epigenetics, particularly DNA methylation, as a candidate mechanism in the rapid regulation of expression. DNA methylation is a mechanism where a methyl group is added to a cytosine residue via the DNA methyltransferases enzyme family (12). In plants, cytosine methylation can happen on three different contexts, namely CG, CHG and CHH, each regulated by different enzymes and pathways (13).

Most studies on DNA methylation in newly formed polyploids highlighted the variability in terms of response across species, but technological limitations didn't allow to investigate changes at the whole genome level. These studies encompassed a variety of genera such as *Spartina* (14), *Senecio* (15), *Brassica* (16–19), *Tragopogon* (20), *Cotton* (21), *Arabidopsis* (22–24), *Triticum* (25,26) and *Mimulus* (27). In most cases, polyploidization was accompanied by rapid non-additive methylation changes in the polyploid compared to the progenitor species, with cotton being the only exception to this rule (21). The amount of non-additive changes was variable across species, 8.3% in *Arabidopsis suecica* (23), 7-9% in *Brassica napus* (17,19), 11.3% in wheat (25), 13.4% in *Senecio cambrensis* (15) and 28.6% in *Spartina anglica* (14), with levels potential related to clade effects (28). Such estimates were inferred via low-

throughput methods, analysing only an extremely small subset of cytosines and severely limiting functional analyses.

Few studies used high-throughput methods to explore genome-wide DNA methylation changes in synthetic polyploids and found evidence in regulation of expression patterns with differences across species. In allotetraploid *Mimulus peregrinus*, most of DNA methylation changes in synthetic individuals happened in CHH context and were associated to lower methylation within and around TE bodies (27). In synthetic *Arabidopsis suecica* changes in CG and CHG context, both associated to genes, were most prevalent (24). These differences in contexts involved was linked to different responses in synthetics, with *Mimulus* showing a rapid subgenome expression dominance, potentially associated with TE methylation, and *Arabidopsis* showing correlated methylation and expression changes in genes related to reproduction. Both results highlighted the importance of DNA methylation in guiding expression patterns after genome duplication, but additional work is required to extend these findings to settings closer to natural conditions, where both biotic and abiotic stresses happen.

A recent hypothesis suggested a critical role of stress response in the success and establishment of polyploids (29). Such hypothesis was supported by past signs, such as ancient WGD events overlapping with major global changes, and recent signs such as polyploid species showing adaptation to environments with strong abiotic stresses (30). Genetic variation resulting from gene duplication is considered an important factor providing higher adaptive potential in stressful conditions (31), but especially in the very early stages of polyploidy, variation in DNA methylation should be considered given the short timescale (32). For example, a small-scale study in *Ranunculus kuepferi* looked at methylation profiles of diploid and autotetraploid individuals grown in a condition, either hot or cold, later swapped to the other after flowering. Temperature changes in both directions induced a response in both cytotypes and polyploids showed higher variation among treatments compared to diploids (33). While environment could have an effect at the genomic level on polyploid species, no studies explored its direct effect on DNA methylation, particularly in essential periods such as the first few generations after WGD.

Arabidopsis kamchatica is an excellent study system to explore environmental stress in the early stages of polyploidy. *A. kamchatica* is a natural allotetraploid resulting from the crossing between *A. halleri* and *A. lyrata,* with a pan-pacific distribution wider in latitude compared to its progenitors including East Asia (Far East Russia, China, Korea, Taiwan and Japan) and North America (Alaska, Canada and Northwestern United States) (34). Genome assemblies for both progenitor species are available (35,36), allowing whole genome analyses. Another key aspect of *A. kamchatica* is the ability to generate synthetic individuals with colchicine treatments, making it ideal to control and analyse the progenitors directly

responsible for the polyploidization process. In addition, natural species can also be used as a reference for assessing a potential trajectory in the changes observed.

In this study, we investigated the change of DNA methylation in the early stages after polyploidization by comparing the methylation patterns of the progenitors to their descendant synthetic tetraploids. Our major questions are: 1) How much of the DNA methylation is affected by polyploidization? 2) How is the pattern changed/maintained in subsequent generations? 3) How does environment affect this change? For this purpose, we artificially synthesized a tetraploid *Arabidopsis kamchatica* from *A. lyrata* and *A. halleri*, mimicking a natural tetraploid species *A. kamchatica.* Progenitors, synthetics and two natural lines were cultivated and propagated in two different conditions, namely hot and cold (Fig. 1a). With this experimental design we compared the whole-genome DNA methylation patterns between progenitors and synthetics, between two generations of synthetics, between synthetics incubated in two different conditions, and between synthetic and natural lines (Fig. 1b).

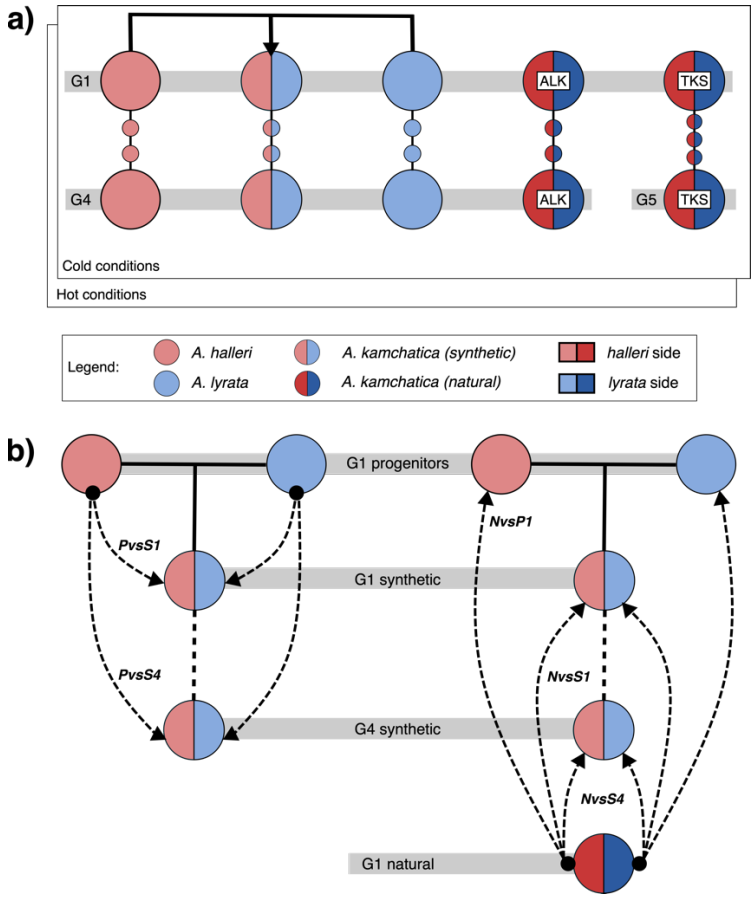


Figure 1: experimental design and comparisons of the A. kamchatica study system. In our experimental design (a), progenitor species A. halleri (red dot) and A. lyrata (blue dot) were grown together with a synthetic A. kamchatica (light red and blue dot) and two natural A. kamchatica lines (red and blue dots, ALK = Alaska line, TKS = Takashima line) in two different conditions, namely hot and cold. Our comparisons to explore DNA methylation changes (b) are divided in two groups. On the left progenitors species are used as reference and compared to first generation synthetics (PvsS1) and fourth generation synthetics (PvsS4). On the right natural lines are used as reference and compared to progenitors (NvsP1), first generation synthetics (NvsS1) and fourth generation synthetics (NvsS4).

Materials and methods

Scaffolding *A. halleri* and *A. lyrata* genome assemblies

To obtain informative spatial information about differentially methylated regions, we used the chromosome-level assembly of *Arabidopsis lyrata* v2.1 (37) as a reference to further scaffold the published genome assemblies of *A. halleri* (36) and *A. lyrata* (35). The mapping, tiling and scoring was done using *RagTag* v1.0.0 (38). To choose the aligner, a simple benchmark among *minimap2* v2.17 (39), *mummer4* v4.0.0beta2 (40) and *lastal* v1060 (41) with default parameters was done to compare the proportion of mapped scaffolds and the total length of mapped scaffolds. For *minimap2* and *mummer4*, the "scaffold" command line utility from *RagTag* was used to run all the necessary steps. For *lastal*, an alignment with default parameters was run independently from *RagTag* and the resulting output was converted to a format that could be used with the "scaffold" utility from *RagTag*.

Instructions and code to reproduce the scaffolding procedure is available at: https://github.com/supermaxiste/RemappingAssemblies . To visually assess synteny between the resulting assemblies and the reference assembly, we used *circlize* v0.4.13 (42) and to define, plot HE regions and compute DMR density we used ad-hoc scripts made in *R* v4.0.5 (43) with the packages *tidyverse* v1.3.1 (44) and *GenomicRanges* (45). To compute the density of DMRs in HE regions, we first defined a set of HE regions together with normal regions. For this purpose, we overlapped the *A. halleri* and *A. lyrata* coverage values and selected regions where *A. halleri* coverage was >9X and *A. lyrata* coverage was <2X (Supplementary Figure 1). With these regions, we counted the amount of overlapping DMRs and computed the density by dividing this value by the length of the regions. Additionally, we computed the ratio between hyper- and hypo-methylated DMRs for normal and HE regions. All of these analyses were done for each condition. The scripts to reproduce the plots and calculations are available at: https://github.com/supermaxiste/EarlyPolyploidDNAMethylation.

Plant material and sequencing

An artificial synthetic tetraploid line was constructed by the crossing of *A. halleri* subsp. *gemmifera* and *A. lyrata* subsp. *petraea*, which were also used for genome assemblies of *A. halleri* (36) and *A. lyrata* (35). One hybrid F1 individual polyploidized to produce tetraploid offspring. An individual from the offspring was employed as the mother of all individuals used to start this experiment. The condition settings of cold and hot incubators can be found in Supplementary Table 1. For each condition and generation, four individuals were incubated and propagated via self-fertilization. Seeds were collected from each individual, germinated

and only four individuals were incubated as next generation from one or two mother plants. We incubated one synthetic tetraploid line, two natural genetic lines of *A. kamchatica* originated from Alaska, U.S. and Takashima, Japan, and the two progenitor species, *A. helleri* subsp. *gemmifera* and *A. lyrata* subsp. *petraea*. As *A. halleri* is self-incompatible, the individuals were clonally propagated instead of undergoing sexual reproduction at each generation.

At each generation, for each group, three individuals were selected for sequencing as biological replicates. Leaf tissue was collected from each individual when the individual started flowering, and the tissue was used to extract DNA and RNA for WGBS-seq and RNA-seq respectively. Nucleic acid was first extracted by CTAB-method (46), and the solution was split into two parts, one for DNA extraction by DNeasy (QIAGEN) and another for RNA extraction by RNeasy (QIAGEN).

DNA libraries for WGBS-seq were synthesized as described in Chapter I, and RNA libraries were synthesized using Illumina TruSeq stranded mRNA kit, both of which were sequenced by Illumina NovaSeq 6000 in 150bp paired-end mode.

Sample quality assessment and basic WGBS data analysis workflow with ARPEGGIO

For a reproducible and automated WGBS data analysis, we used *ARPEGGIO v2.0.0* (47). In short, after a quality check with *FastQC v0.11.8* (48), reads were trimmed with *TrimGalore v0.6.5* (49) 10bp 3' and 5bp 5' to remove adapter sequences. Alignment and deduplication, both with *Bismark v0.22.3* (50), and read classification with *EAGLE-RC v1.1.2* (51) were run using the *Arabidopsis halleri* genome assembly *v2.2* (36), the *Arabidopsis lyrata* genome assembly *v2.2* (35) and their corresponding annotation. An additional alignment to the *A. halleri* chloroplast assembly (52) was executed for an *in silico* conversion check. Since we expected no methylation in the chloroplast genome (53,54), the proportion of reads uniquely mapped to it can provide an estimate of the bisulfite conversion error rate. This rate was used to calculate conversion efficiency. All alignments were analyzed with *QualiMap v2.2.2d* (55) to obtain coverage information. The quality reports from *FastQC*, *TrimGalore*, *Bismark* and *Qualimap* were combined in a single document via *MultiQC v1.8* (56).

The protocol for bisulfite conversion and library preparation affected the quality of the sequencing data from each species and conditions differently. Some samples displayed a significant proportion of overrepresented sequences, adapter sequences and/or duplication. Standard trimming procedures combined with deduplication after alignment, mitigated most of the initial issues, leading to a reduction in coverage in few samples. Quality checks, trimming, in silico conversion checks, alignment and deduplication results, coverage and bias and

confidence of methylation calls are all included in a sample quality report available in the repository below.

After read classification, methylation information was extracted from reads with *Bismark v0.22.3* (50). This information was used for two downstream analyses. First, the confidence of methylation calls was assessed, and then GML and methylation levels within and around gene bodies were computed. Second, we performed differential methylation analysis in *R v3.6.2* (43) with *dmrseq v1.6.0* (57) to obtain information about regions showing significant change in methylation. *ARPEGGIO* was run with Conda-only mode, using twelve CPU cores Intel(R) Xeon(R) CPU E5-4640 at 2.40GHz. ARPEGGIO's configuration and metadata files to reproduce the analyses (see Chapter I for details) can be found at https://github.com/supermaxiste/EarlyPolyploidDNAMethylation.

Global methylation level and average methylation levels around gene bodies

To estimate the global methylation level (GML) we used imputed cytosine calls from METHImpute *v1.10.0* (58). In short, *METHImpute* applies a Hidden Markov Model to WGBS data to impute methylation calls and levels for all cytosines in the genome. All of the predictions are accompanied by a posterior probability estimating their confidence. High-quality calls were defined as calls with posterior probability > 0.9. The proportion of high-quality cytosine calls was also included in the quality assessment of samples (see next section). The GML for each species was calculated with the average of the sum of all (imputed) methylated cytosines divided by the total number of cytosines. In the case of *A. kamchatica* synthetics, the imputation was done on masked input data where low coverage regions were excluded. For each condition separately, the coverage of G4 synthetics was used to mask all scaffolds with coverage <2X from both G1 and G4 synthetics.

To formalize the calculation of GML, let's assume that the total amount of cytosines calls for methylated cytosines is ${}^{m}C_i$ and for unmethylated cytosines is ${}^{u}C_i$, where *i* is the *i*th replicate for a given sample out of a total of *n* replicates. The total number of cytosines for each replicate is given by ${}^{m}C_i + {}^{u}C_i$. The GML for each sample was calculated as follows:

$$GML = \frac{1}{n}\sum_{i}^{n}\frac{{}^{m}C_i}{{}^{m}C_i + {}^{u}C_i} \times 100$$

The uncertainty for each sample was defined as the sample standard deviation (sd) of the proportion of methylated cytosines in each replicate. Formally:

$$sd(GML) = sd(\frac{{}^{m}\mathrm{C}_i}{{}^{m}\mathrm{C}_i + {}^{u}\mathrm{C}_i}) \times 100$$

The uncertainty defined above did not take into account the proportion of high and low confidence of methylation calls among Cs provided by *METHImpute*. For most samples, high confidence calls represented 80-90% of all cytosines in the genome and all of the samples had an average of 70% cytosines with high quality calls or more (see next section).

In addition to GML, *METHImpute* was also used to compute the average methylation rate within gene bodies and in their 500bp flanking regions for each methylation context: CG, CHG and CHH. Gene bodies were defined as the regions between the transcription start site (TSS) and the transcription termination site (TTS). The 500bp range was decided based on the theoretical proportion of the assembly represented by different ranges of flanking regions such as 100, 250, 500, 1000 and 2000bp. With 500bp, flanking regions represented 19.5% of the *A. halleri* assembly, together with 55.2% from gene bodies, leaving 25.3% to intergenic regions (Supplementary Material 1). For *A. lyrata*, flanking regions represented 20% of the assembly, gene bodies 56.7% and intergenic regions 23.3% (Supplementary Material 1).

MDS analyses

To construct MDS plots we used previous approaches (59–61). In short, coverage data from *Birsmark* was first filtered to only take into account overlapping Cs across replicates and samples with a certain amount of coverage (>=3). Next, methylation proportions were arcsin transformed to stabilize the variance (to prevent signal driven by the mean since higher mean leads to higher variance). The whole dataset was split into two conditions and for each condition there were two separate progenitors' sides (total of four plots). Since MDS depends on the amount of Cs selected we used several thresholds to assess if the relationship between samples would change significantly. For each threshold (10'000, 100'000, 1'000'000 and all), the "top" Cs with the maximum difference in transformed methylation proportion between samples were selected.

When applying the filtering step, the amount of resulting Cs for each side and condition was the following: for cold conditions, *halleri*-side: 6'227'144, *lyrata*-side: 3'052'057, for hot conditions, *halleri*-side: 15'591'169, *lyrata*-side: 8'044'839. Scripts to reproduce MDS plots can be found in https://github.com/supermaxiste/EarlyPolyploidDNAMethylation.

Differential methylation and downstream analyses

Differential methylation was done separately for each methylation context (CG, CHG, CHH) through *ARPEGGIO*. The output from each of these analyses provided a list of regions showing differential methylation, together with their coordinates in the genome. With these lists, *ARPEGGIO* ran downstream analyses to find and output all DMRs overlapping with annotated gene regions. *ARPEGGIO* was run with Conda-only mode, using twelve CPU cores Intel(R) Xeon(R) CPU E5-4640 at 2.40GHz.

Additional downstream analyses were done for differentially methylated regions (DMRs). First, general statistics were computed with a summary of the amount of DMRs, their average, median and total length for each context and progenitor's side (Supplementary Table 2). Second, we computed the proportion of hypo- and hypermethylated DMRs for each context and progenitor's side. Third, we checked the overlap between DMRs and genic, intergenic or flanking gene regions. The overlap was defined as at least 1bp in common between a functional region and a DMR. If a DMR would overlap with several functional regions, the priority was set first to genic regions, then flanking regions and last intergenic regions, meaning that a DMR overlapping with all three would be classified as overlapping with genic regions only. Fourth, we computed the proportion of DMRs in different functional regions of gene bodies (Supplementary Figure 2). Finally, we also plotted the spatial distribution of DMRs along chromosomes (Supplementary Figure 3-4).

All of these analyses were performed with *R v4.0.2* (43). To import large files, we used *data.table v1.13.0*. For data wrangling, barplots and donut charts, we used *tidyverse v1.3.0* (44) and *gridExtra v2.3* (62). For all analyses involving overlaps between regions, we used *GenomicRanges v1.40.0* (45). Chromosome plots were done with *karyoploteR v1.16.0* (63). The full reproducible code for the analyses discussed here is provided in https://github.com/supermaxiste/EarlyPolyploidDNAMethylation.

Gene overrepresentation analyses

Lists of differentially expressed methylated genes (DMG) were used for gene overrepresentation tests with PANTHER Overrepresentation Test (Released 20210224) (64) with *Arabidopsis thaliana* gene IDs and gene database. To obtain *A. thaliana* gene IDs from DMG with *A. halleri* and *A. lyrata* gene IDs, we checked for ortholog genes based on BLAST reciprocal best hit between genes of our reference species and *A. thaliana*. Since both *A. halleri* and *A. lyrata* genomes have a higher number of genes, a significant proportion did not match to *A. thaliana* genes, particularly in CHG and CHH context for DMG. These proportions together with the detailed results from overrepresentation tests can be found in https://github.com/supermaxiste/EarlyPolyploidDNAMethylation.

Results

Section 1 - Scaffolding *A. halleri* and *A. lyrata* assemblies and detection of homoeologous exchanges in synthetic polyploids

We used *A. halleri* subsp. *gemmifera* and *A. lyrata* subsp. *petraea* as progenitors (P), which were shown to be the closest progenitors of *A. kamchatica* (65), to synthesize an artificial tetraploid line (S). The genetically uniform offspring from the same mother plant was cultivated in two different conditions, named cold and hot, for four generations. The two progenitors and two genetically distinct natural tetraploids from different populations, Alaska (N_A) and Japan (N_J), were also cultivated together with the synthetic line.

Since the available genome assemblies of the diploid progenitors were not chromosome-level, an essential aspect for spatial analysis of whole-genome DNA methylation data, we used the chromosome-level assembly of *Arabidopsis lyrata* v2.1 (1) as a reference to scaffold our assemblies. To do so, we first had to compare and choose an appropriate mapper between *minimap2*, *mummer4* and *lastal*. For both *A. halleri* and *A. lyrata*, *minimap2* showed the highest mapping rate with the highest proportion of base pairs (bp) mapped (Supplementary Table 3). For *A. lyrata*, 753 out of 1'675 scaffolds (45%) were mapped, corresponding to 96% of the total bp length of the initial assembly. For *A. halleri*, 964 out of 2'239 scaffolds (43%) were mapped, corresponding to 95% of the total bp length of the initial assembly. For both species, only ~5% of the initial assemblies were unmapped, corresponding to numerous small scaffolds. For *mummer4*, the mapping rate was similar: 488 out of 1675 scaffolds (29%) mapped for *A. lyrata*, representing 95% of the total bp length of the initial assembly, and 538 out of 2'239 scaffolds (24%) mapped for *A. halleri*, representing 92% of the total initial bp length. *Lastal* showed the lowest mapping rate, 420 out of 1'675 scaffolds (25%) for *A. lyrata*, with 71% of the initial total bp length mapped and 362 out of 2'239 scaffolds (16%) for *A. halleri*, 59% of the initial total bp length. The scaffolding for *lastal* showed a large number of discrepancies compared to the other two methods (Supplementary Figure 5). On the other hand, *minimap2* and *mummer4* showed good consistency when compared against each other and against the reference *A. lyrata* assembly (Supplementary Figure 5-6). As *minimap2* and *mummer4* showed similar high scaffolding quality, *mummer4* was chosen as a reference because of its higher accuracy in previous work on polyploids (66).

With these new chromosome-level assemblies, we assessed sub-genome mapping coverage in the fourth generation (G4) synthetics from the BS-seq data. In both experimental

conditions, the *A. lyrata* side (L-subgenome) showed large stretches of chromosomes with lower coverage in G4 (Supplementary Figure 6). These stretches could represent either chromosome losses or homoeologous exchanges (HE), both previously reported in other polyploid systems (67). Flow cytometry did not show a significant decrease in the genome size of their further descendants in G6 or G7 (Supplementary Table 4), supporting HE rather than deletion. In addition, the coverage of the corresponding regions in the *A. halleri* side (H-subgenome) was higher than other regions (Supplementary Figure 1). Based on this evidence, we concluded that HE from L to H happened in some regions unexpectedly. To prevent potential biases of low coverage regions in some downstream analyses, we excluded these regions (see Materials and Methods for details).

Section 2 - Global whole genome methylation patterns reveal consistent patterns across environments with some key differences

Global methylation levels

As a first look into genome-wide methylation patterns, we computed global methylation levels (GMLs) for all samples and found that synthetic polyploids showed consistent yet different patterns across conditions (Fig. 2a). First generation (G1) synthetics showed similar GML levels in both H and L subgenomes with values consistently close to one progenitor, *A. halleri*, on both conditions, meaning that polyploidization increased the GML of the L subgenome. In G4 synthetics, GMLs of all subgenomes in both conditions decreased, getting closer to the other progenitor, *A. lyrata*, but GMLs were higher in hot condition for both subgenomes, suggesting a divergence in methylation pattern already after four generations. More specifically, in cold conditions synthetics had a GML of 22.42% *halleri*-side and 22.26% *lyrata*-side, both values close to 22.30% for first generation *A. halleri*. Similarly, in hot conditions synthetics had a GML of 21.86% *halleri*-side and 21.67% *lyrata*-side against 22.44% in *A. halleri*. Fourth generation synthetics still showed similar GMLs between subgenomes, but the values appeared closer to *A. lyrata* and showed some variation across conditions. For cold conditions, synthetics showed a GML of 20.01% and 19.68% for *halleri*- and *lyrata*-side respectively. These values were lower than the 21.33% and 20.74% for *halleri*- and *lyrata*-side obtained from the synthetics in hot conditions.

A similar difference between conditions could be found in both progenitors and two genetic lines of natural *A. kamchatica*, even though their GMLs did not change according to generation. In cold conditions, the largest progenitors' change in GML happened in *A. halleri,* going from 22.30% to 21.17% (-1.13%), a decrease not as strong as the one observed in the synthetics, where both progenitors' sides had ~2% less methylation. In hot conditions, both

diploid progenitors showed an increase in methylation (*A. halleri*: +0.68%, *A. lyrata*: +0.51%) compared to a decrease in synthetics (*A. halleri*-side: -0.88%, *A.lyrata*-side: -1.25%). In addition, natural lines showed closer GMLs between subgenomes than those of progenitor species, but the levels were different across lines.

Overall, differences in GMLs were not very pronounced over generations, suggesting relative stability in the GML, but this did not exclude extensive changes in methylation patterns. Since GML is a summary statistic focusing on the proportion of methylated cytosines, it cannot distinguish whether a cytosine changed its methylation state or not. Because of this, it would be possible to have the same GMLs between two individuals from the same species, even though they have completely different pattern of methylated cytosines. Nonetheless, differences in GMLs provide a baseline for methylation changes and offer a complementary view on local changes in methylation, i.e. differentially methylated regions.

Methylation levels within and around gene bodies

To further investigate genome-wide changes, we imputed average methylation levels within and around gene bodies for all of our samples, revealing variation in methylation response between conditions (Supplementary Figure 7-10). When comparing progenitors and synthetics, methylation level changes on the *lyrata*-side were more pronounced than the *halleri*-side (Supplementary Figure 7-8). On the *halleri*-side, average gene body methylation appeared relatively stable in cold conditions for CG and CHG context, with a decrease in CHH context. In hot conditions, the decrease in synthetics appeared in both CHG and CHH contexts. In contrast, the *lyrata*-side showed increase in CG and CHG methylation in both conditions and a condition-dependent change in CHH methylation, decrease in cold and increase in hot conditions. Taken together, these methylation changes suggested differences between conditions with a larger amount of changes in hot conditions compared to cold conditions.

Multidimensional scaling (MDS) analyses

To explore the relationship between samples, we performed MDS analyses and found consistent relations between samples in both conditions (Fig. 2b, Supplementary Figure 11-12). Three main clusters were found across a varying number of cytosines considered, across conditions and progenitors' sides. One cluster included all diploid and synthetic individuals from all generations, a second included all natural Alaska line individuals and the last all natural Takashima line individuals. This result suggested a notable difference between the methylation pattern of natural polyploids and that of diploid and synthetic samples.

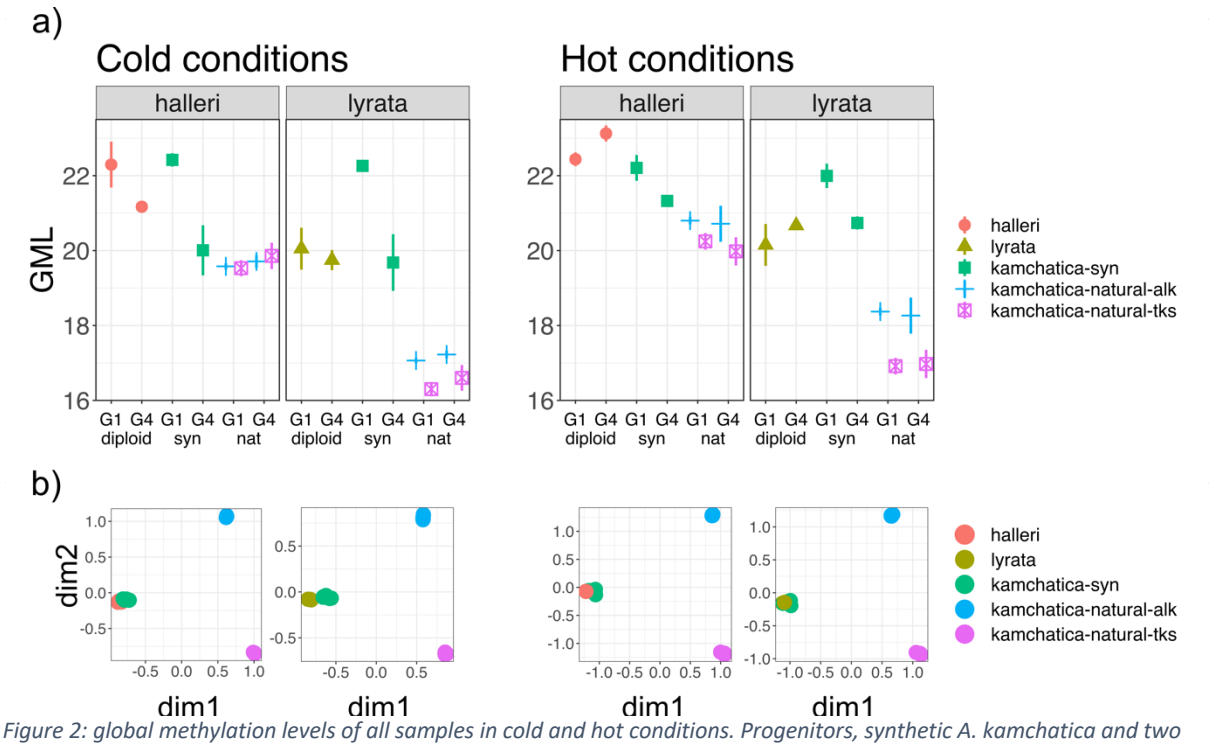


Figure 2: global methylation levels of all samples in cold and hot conditions. Progenitors, synthetic A. kamchatica and two natural A. kamchatica lines (ALK = Alaska, TKS = Takashima) are represented by a different shape. For all individuals, two GML values are shown, one in the first (G1) and one in the fourth generation (G4). GML values are shown for cold conditions (left plot) and hot conditions (right plot).

Section 3 – Methylation patterns in synthetic *A. kamchatica* show diverging patterns compared to its progenitors and converging patterns compared to natural species

Differentially methylated regions between synthetics and progenitors

To explore changes in DNA methylation in more detail, we computed differentially methylated regions (DMRs) between synthetics and G1 progenitors and found an increasingly diverging pattern. The amount of DMRs increased when going from PvsS1 (progenitor G1 vs. synthetic G1) to PvsS4 (progenitor G1 vs. synthetic G4) (Fig. 3). This increase happened across all contexts, with one exception, and in both conditions, further supporting a divergent pattern between diploids and synthetics over time. Also, between ~50% and 69% of all DMRs found in the first generation were also found in the fourth (Supplementary Table 5). These results suggest that DNA methylation changes in the synthetics occur right after polyploidization with limited reshuffling of methylation patterns at every generation, supporting the idea of partially conserved methylation changes over generations.

Even though both conditions showed a consistent pattern from PvsS1 to PvsS4, the amount of DMRs and their distribution across contexts showed notable differences between conditions, leading to different methylation patterns in synthetics. In PvsS4 cold conditions,

the number of DMRs was similar across contexts except for CHH on the *lyrata*-side, showing approximately twice the number of DMRs. This distribution was very distinct from the one in hot conditions, where CHG and CHH context on the *halleri*-side had almost double the amount of DMRs compared to other contexts on both sides, except for CHG context on the *lyrata*-side also showing high amounts of DMRs. Additionally, hot conditions, which were more stressful, led to an overall higher number of changes compared to cold conditions, with a longer stretch of the genome being affected (Supplementary Table 2). These different dynamics in two conditions suggested an effect of environmental stress in DNA methylation response. In support of this, comparing synthetics between conditions also showed an increasingly divergent methylation pattern. In the first generation no significant DMRs were detected (Supplementary Figure 13a), however several DMRs were found when comparing fourth generations (Supplementary Figure 13b). On the other hand, diploid progenitors and natural lines did not show drastic changes in DMRs between conditions or generations, suggesting stability of whole genome methylation status (Supplementary Figure 14). The largest amount of DMRs was detected in CHH contexts in hot conditions in both progenitors and natural lines.

Changes in methylation in synthetics were not extensive, spanning at most ~4% of the genome and showing a relatively uniform distribution across chromosomes. When considering the total length of DMRs in all comparisons (Supplementary Table 2), in cold conditions synthetics had modifications in ~1.3% of the genome for PvsS1 and ~2.3% for PvsS4, both lower values compared to hot conditions where ~2.3% and ~4% changes were found in PvsS1 and PvsS4 respectively, suggesting no drastic whole-genome changes in DNA methylation. When investigating distribution of DMRs along chromosomes, particularly around centromeres, no pronounced pattern was found (Supplementary Figure 3-4). Instead, both hypo- and hyper-methylated DMRs appeared uniformly distributed when excluding HE regions.

Functional analyses between synthetics and progenitors

We inspected the genomic context of all DMRs to find out the overlap with either gene bodies, flanking regions or intergenic regions and proportions were generally context dependent. For CG and CHG context 60-75% of the DMRs overlapped with gene bodies or flanking regions, while CHH context DMRs showed a larger overlap with intergenic regions from 36% up to 48% (Supplementary Figure 15).

We focused on the DMRs associated with genes and found overrepresented gene sets in broad functional categories when comparing synthetics to progenitors. For both conditions, differentially methylated genes in first generation synthetics showed some overrepresentation mostly in CG context (on both progenitors' sides) with broad functional categories related to

cellular (GO:0009987) and metabolic processes (GO:0071704, GO:0008152). For fourth generations, all contexts showed some overrepresentation with similar and wide-ranging categories such as cellular metabolic process (GO:0044237), response to stimulus (GO:0050896) and the same cellular and metabolic processes from the first generation. The only condition-specific enriched terms found were on the *lyrata*-side in CG context, namely ion transport (GO:0006811) in cold conditions and mitotic cytokinesis (GO:0000281) in hot conditions.

Differentially methylated regions between synthetics and natural lines

As a second set of comparisons, we used our natural lines as reference and tracked the amount of DMRs against first generation progenitors (NvsP1), first (NvsS1) and fourth generation synthetics (NvsS4), showing a consistent converging pattern (Fig. 4). Since low-coverage regions lead to a reduction in DMR detection, we included only DMRs falling within regions showing sufficient coverage in all our samples. For both conditions and for both natural lines the only consistent trend was a decrease in CG DMRs from NvsP1 to NvsS4. This decrease was more pronounced on the *lyrata*-side compared to the *halleri*-side. This suggested a converging methylation pattern between synthetics and naturals, but compared to the divergence from diploids, no notable environmental difference was found. For CHG and CHH context, the temporal trend didn't show a clear trajectory and the amount of DMRs fluctuated.

Patterns of differential methylation in HE regions

We also investigated DMR states in HE regions and compared them to representative regions where HE didn't occur, revealing fewer demethylation changes in HE regions. First, we computed DMR densities together with the proportion of hyper- and hypomethylated DMRs to assess if HE regions showed different methylation dynamics (Table 1). In both conditions, the overall DMR density was lower in HE compared to normal regions, with hot conditions showing the largest difference. DMR densities for hyper and hypomethylated DMRs showed a minor increase of hypermethylation density in HE together with a more pronounced decrease in hypomethylation density. Taken together, these results suggest less methylation changes, particularly decreasing methylation, in HE regions, but similar hypermethylated densities.

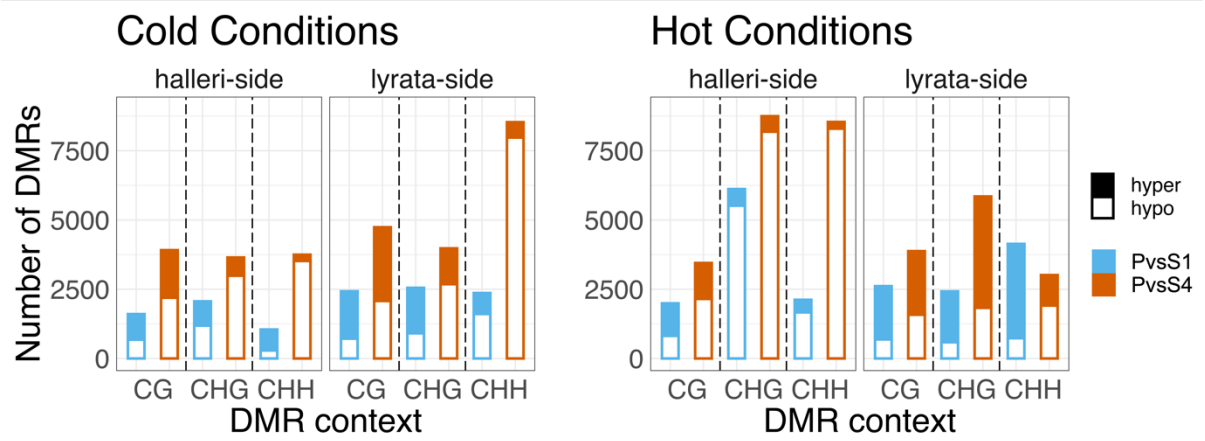


Figure 3: Total number of DMRs between first and fourth generation synthetics and their respective diploid progenitors (used as reference). On both sides the x-axis shows the generation of synthetics that was compared to the first generation of progenitors for each methylation context. The y-axis shows the amount of DMRs and the barplots distinguish between hyper DMRs, where a significant increase in methylation was observed in the synthetics, and hypo DMRs, where a significant decrease was found. On the left side is the result for samples grown in cold conditions and on the right is for samples grown in hot conditions.

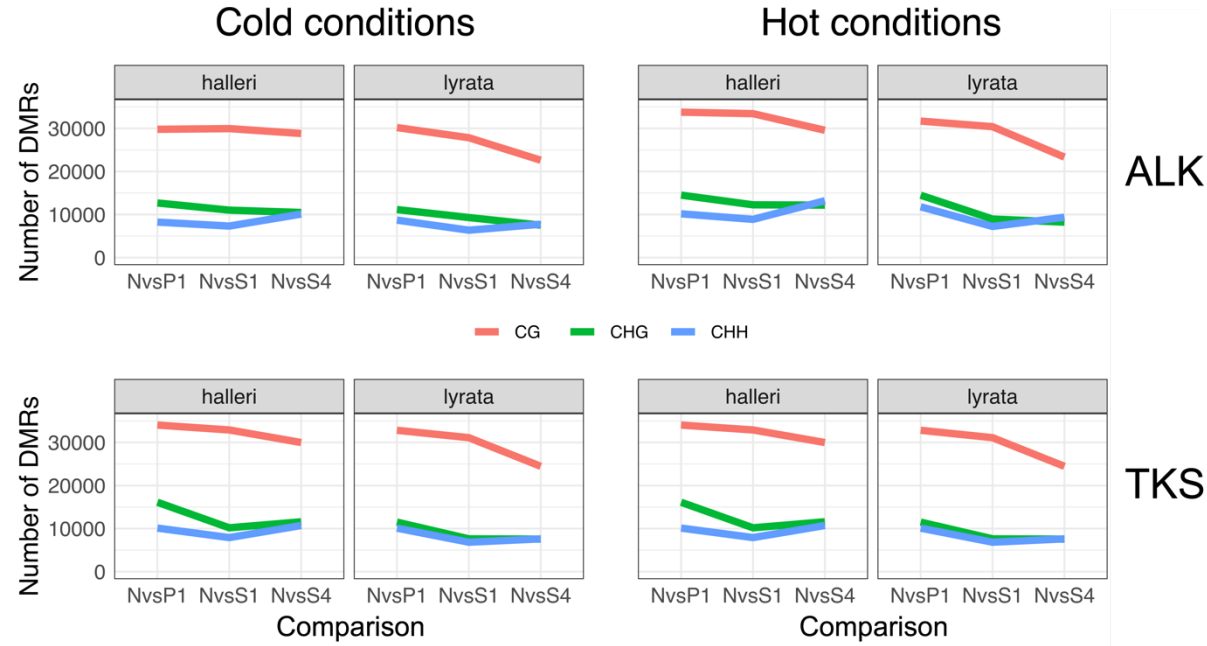


Figure 4: Total number of DMRs between natural lines (ALK = Alaska, TKS = Takashima) and first generation progenitors (NvsP1), first generation (NvsS1) and fourth generation synthetics (NvsS4). Each line represents the amount of DMRs for each context. The left and right side show DMR changes in cold and hot conditions respectively. The first row represents the results for the Alaska line as reference, while the second used Takashima as a reference.

Table 1: DMR density in homoelogous exchanged (HE) regions and normal regions. Table includes statistics about length of HE and normal regions, total amount of DMRs, DMRs density per Mb obtained by dividing total DMRs by region length. The same density weas computed for hypermethylated DMRs, regions in the synthetics showing an increase in methylation with respect to the progenitors, and hypomethylated DMRs.

Condition	Cold		Hot	
Region	HE	Normal	HE	Normal
Length	18.7Mb	27.5Mb	12.1Mb	30Mb
DMRs amount	980	1725	1026	3413
DMRs density	52.41	62.73	84.79	113.77
Hyper DMRs	274	370	138	313
Hypo DMRs	706	1355	888	3100
Hyper DMRs density	14.65	13.45	11.40	10.43
Hypo DMRs density	37.75	49.27	73.39	103.33

Discussion

First insight into HE regions and their DNA methylation dynamics

In our study, we reported the first case of HEs in synthetics *Arabidopsis kamchatica*. Such events have been previously observed in other synthetic polyploids such as *Brassica* (19,68–70) and rice (71), but also in natural polyploids including peanuts (72), tobacco (73), *Tragopogon* (74) and others (75,76). From a mechanistic point of view, the occurrence of HEs in synthetic polyploids has been associated to meiotic instability (77), but the evolutionary implications are still under discussion. In *Brassica* for example, HE events were associated to long-term gene copy number variation (70) and phenotypically to variation in flowering time (78) and disease resistance (79). For *A. kamchatica*, HE events might not have the same beneficial role as in *Brassica*, especially since the two *A. kamchatica* natural lines studied here did not show any evidence of large scale HEs such as in our fourth generation synthetics. More data from natural species would be needed to confirm this, but evidence from 15 natural accessions from the close relative *Arabidopsis suecica* was able to only find a single small-scale HE event (80).

To shed some light on the impact of HEs in the early stages of polyploidy, we looked at DNA methylation in HE regions and found a lower DMR density compared to normal regions, particularly in stressful conditions. Most of the reduction in DMR density was associated to a decrease in hypomethylated DMRs, while hypermethylated DMRs showed minor changes. Taken together, these results suggest different dynamics between HE regions and normal regions, with HE regions showing overall less changes. This result shows strong contrast to similar findings in rice, where HE regions showed two to five times more DMRs than normal regions (71). Reasons for such differences could be attributed to the type of polyploidy and genomic architecture. Rice in (71) was a segmental allopolyploid, an intermediate genetic state between auto- and allopolyploid (81). In addition, the rice genome has larger size, with almost 50% of it being TEs (82). Estimates of TE content for our progenitors were not available, but gene bodies represented already >50% of our assembly (Supplementary Material 1) and estimated GMLs were lower compared to rice, with low GML being associated to lower TE content (32).

While our study provided a unique opportunity to look at HEs in the early stages of polyploidy, there were limitations to our approach. The reference genome used for scaffolding originated from *A. lyrata subsp. lyrata* and the assemblies from this study were from *A. lyrata subsp. petraea*, another subspecies, and *A. halleri*, a different species. Although the species share the same genus, we cannot exclude the possibility of differences in genomic arrangements, especially between different species with a longer divergence time. Previous

work on *A. thaliana* showed an overall conserved synteny within angiosperms, but such comparisons were complicated by many gene deletions specifically in *A. thaliana* (83). Additionally, several improvements could be applied in future work to offer clearer genomic ranges of HE regions, such as chromosome conformation capture methods, and better track their development over time and across individuals by increasing the numbers of biological replicates and generations.

Epigenetic evolution of a newly formed polyploid

When analyzing DNA methylation changes between synthetics and progenitors more in depth, DMRs were found in all methylation contexts, independent from subgenome, condition or generation, implying links to both genic and intergenic responses. Similar results were obtained in synthetic *A. suecica*, where DMRs were also found in all contexts, but with CG showing most of the changes and stronger association to genes (24). In synthetic *A. kamchatica* both CG and CHG context showed an association to genes, but instead of CG, higher amounts of changes were found in CHG and CHH contexts. These differences in contexts involved might be caused by differences in genome architecture. Previous work looking at DNA methylation patterns in *A. thaliana*. *A. lyrata* and *C. rubella* found the expansion and reduction of repetitive sequences and transposable elements as main drivers (84). More specifically, a higher methylation rate was reported in introns in *A. lyrata* which correlated to a TE invasion and occurred only in CHG and CHH sites (84). In PvsS1 and PvsS4, we found an above average proportion of DMRs overlapping exons and a below average proportion of DMRs overlapping introns for all contexts and both subgenomes, showing no strong sign of intron or exon methylation bias. More analyses on gene expression and TEs are needed to clarify the link between genome architecture and methylation response in early stages of polyploidy.

From a functional perspective, DMRs between synthetics and progenitors showed analogous changes across conditions, suggesting a functionally similar divergence. The GO terms found were mostly related to broad metabolic and cellular processes and covered most contexts, with only few specific terms found in PvsS4. Such broad regulatory changes could be related to the dosage balance hypothesis, where imbalances in the amount of subunits of protein complexes can have deleterious effects (85). In our case, DNA methylation changes in synthetic *A. kamchatica* might help regulate the balance between subunits from both subgenomes via gene expression changes, but RNA-seq data would be needed to confirm this.

Spatial distribution of DMRs did not reveal strong patterns along chromosomes, but showed occurrence of DMRs close to centromeric regions that are known to be repeat-rich and can affect transcription patterns (22,86). Since in cold conditions PvsS4 showed genome-wide hypo-methylation on both subgenomes, these changes might result in activation of centromeric TEs and siRNA accumulation in the centromere, both known to affect transcription (86). For hot conditions, the same hypothesis could be subgenome specific, since only the *halleri*-side showed overall hypo-methylation, whereas the *lyrata*-side showed hyper-methylation.

Environmental stress drives DNA methylation patterns, guiding evolution in early stages of polyploidy

We showed for the first time how different environmental conditions affected the way synthetics diverge from progenitors in terms of methylation at the whole genome level, highlighting the importance of stress in the early stages of polyploidy. In our study, synthetic polyploids showed the highest variation in methylation across generations in both conditions, but we didn't observe major differences between natural polyploids and diploid progenitors, suggesting that polyploids could also differ greatly in their response depending on their age. This was expected considering how methylation patterns were optimized in natural polyploids after selection (87) while newly formed polyploids have extremely dynamic epigenetic activity (22,25,88), both increasing variation and possibly negatively affect the phenotype (89).

Furthermore, environmental conditions affected the distribution and amount of DMRs on each subgenome in *A. kamchatica,* potentially reflecting its ability to control its subgenomes to quickly adapt to novel environments. A majority of DMRs was found on the *lyrata*-side for PvsS4 in cold conditions, whereas in hot conditions it was the *halleri*-side. Additionally, in cold conditions both subgenomes exhibited mostly decreased methylation, while in hot conditions there was an overall decrease on the *halleri*-side and an increase on the *lyrata*-side. These trends could characterize *A. kamchatica*'s ability to take advantage of each subgenome's capability to respond to the environment. From an ecological context, *A. lyrata* is found in arctic-cycle latitudes with lower average yearly temperatures, while *A. halleri* is closer to the equator with milder average temperature and natural *A. kamchatica*'s is found in both its progenitors niches and beyond (65). Subgenome specific DNA methylation changes depending on the environmental condition could be a fast and plastic response to survive and thrive in early generations of *A. kamchatica*. The link between DNA methylation and plasticity has been already suggested in allotetraploid *Poa annua*, which was grown with two treatments (mowed and unmowed) for 8 months and increases in global methylation patterns for mowed plants were maintained in their progeny (90). In our study, this hypothesis was partly supported

by functional analyses, first in PvsS4, where some condition-specific categories were found, and second when comparing synthetics grown in two conditions, where we found several heat response genes. Further analyses, possibly with later generations of synthetics, would be required to establish a clearer link between subgenome-specific genes and environmental response.

Convergent evolution between synthetics and naturals

While the number of DMR between progenitors and synthetics increased over generations (Fig. 3), the number between synthetic and natural tetraploids decreased for DMRs, particularly in CG context which is strongly associated to genes (Fig. 4). This trend suggested that the methylation pattern of synthetic converged toward those of naturals. Considering that the methylation patterns of naturals were stable across different environmental conditions, the change in synthetics might suggest that there could be some stability point in the DNA methylation status after polyploidization. In support of this, the GMLs of G4 synthetics were lower than those at G1 in both conditions and subgenomes, getting closer to the GMLs of natural tetraploids which were the lowest level among all samples (Fig. 2a). Even though the exact genotypes of the progenitors of synthetics and naturals were different, the resulting stability point might exist, which we may refer to as epigenetic canalization after polyploidization (91).

The convergence in methylation between synthetics and both natural lines was too similar across conditions to distinguish a trajectory towards a specific line. Each natural line showed a distinct methylation pattern (Fig. 2b), possibly shaped by the different natural conditions from their original locations, one with higher average temperatures throughout the year (Takashima) compared to the other (Alaska). Given these differences between naturals, we expected synthetic *A. kamchatica* to approach a specific natural line depending on the conditions, but four generations were not enough to address this.

Conclusions

Whole genome studies in synthetic allopolyploids can be challenging and including environmental effects can increase complexity even more. Our study offered a first insight into this complex picture, highlighting the importance of environment in shaping polyploids right after formation. We showed how DNA methylation provided a rapid and plastic response mechanism towards both environment and polyploidy in newly formed polyploids. The response was characterized by novel methylation patterns compared to their progenitors and approached patterns found in their natural counterpart. Future research should also be cautious in the detection and analysis of HE, since these events are common in synthetic polyploids and their scale can considerably affect downstream bioinformatics analyses. Both epigenetics and HE are examples of how the aftermaths of polyploidy can be complex and variable, suggesting no common solution to this event in different plant species. Our research extended this picture to include environment as another important factor, offering a new path for future polyploid studies.

References

1. LUGHADHA EN, GOVAERTS R, BELYAEVA I, BLACK N, LINDON H, ALLKIN R, et al. Counting counts: revised estimates of numbers of accepted species of flowering plants, seed plants, vascular plants and land plants with a review of other recent estimates. Phytotaxa [Internet]. 2016 Aug 26;272(1):82. Available from: https://biotaxa.org/Phytotaxa/article/view/phytotaxa.272.1.5
2. One thousand plant transcriptomes and the phylogenomics of green plants. Nature [Internet]. 2019 Oct 23;574(7780):679–85. Available from: http://www.nature.com/articles/s41586-019-1693-2
3. Jiao Y, Wickett NJ, Ayyampalayam S, Chanderbali AS, Landherr L, Ralph PE, et al. Ancestral polyploidy in seed plants and angiosperms. Nature [Internet]. 2011 May 10;473(7345):97–100. Available from: http://www.nature.com/articles/nature09916
4. Soltis PS, Marchant DB, Van de Peer Y, Soltis DE. Polyploidy and genome evolution in plants. Curr Opin Genet Dev [Internet]. 2015 Dec;35:119–25. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0959437X15001185
5. Soltis DE, Visger CJ, Marchant DB, Soltis PS. Polyploidy: Pitfalls and paths to a paradigm. Am J Bot [Internet]. 2016 Jul;103(7):1146–66. Available from: http://doi.wiley.com/10.3732/ajb.1500501
6. Barker MS, Arrigo N, Baniaga AE, Li Z, Levin DA. On the relative abundance of autopolyploids and allopolyploids. New Phytol [Internet]. 2016 Apr 6;210(2):391–8. Available from: https://onlinelibrary.wiley.com/doi/10.1111/nph.13698
7. Grover CE, Gallagher JP, Szadkowski EP, Yoo MJ, Flagel LE, Wendel JF. Homoeolog expression bias and expression level dominance in allopolyploids. New Phytol [Internet]. 2012 Dec 4;196(4):966–71. Available from: https://onlinelibrary.wiley.com/doi/10.1111/j.1469-8137.2012.04365.x
8. ADAMS K, WENDEL J. Novel patterns of gene expression in polyploid plants. Trends Genet [Internet]. 2005 Oct;21(10):539–43. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0168952505002179
9. Bird KA, Niederhuth C, Ou S, Gehan M, Chris Pires J, Xiong Z, et al. Replaying the evolutionary tape to investigate subgenome dominance in allopolyploid Brassica napus bioRxiv [Internet]. 2019 Jan 1;814491. Available from: http://biorxiv.org/content/early/2019/10/22/814491.abstract
10. Freeling M, Woodhouse MR, Subramaniam S, Turco G, Lisch D, Schnable JC. Fractionation mutagenesis and similar consequences of mechanisms removing dispensable or less-expressed DNA in plants. Curr Opin Plant Biol [Internet]. 2012 Apr;15(2):131–9. Available from: https://linkinghub.elsevier.com/retrieve/pii/S1369526612000301
11. Coate JE, Bar H, Doyle JJ. Extensive Translational Regulation of Gene Expression in an Allopolyploid (Glycine dolichocarpa). Plant Cell [Internet]. 2014 Jan;26(1):136–50. Available from: https://academic.oup.com/plcell/article/26/1/136-150/6102313
12. Moore LD, Le T, Fan G. DNA Methylation and Its Basic Function. Neuropsychopharmacology [Internet]. 2013 Jan 11;38(1):23–38. Available from: http://www.nature.com/articles/npp2012112
13. Niederhuth CE, Schmitz RJ. Putting DNA methylation in context: from genomes to gene expression in plants. Biochim Biophys Acta - Gene Regul Mech [Internet]. 2017;1860(1):149–56. Available from: http://dx.doi.org/10.1016/j.bbagrm.2016.08.009
14. SALMON A, AINOUCHE ML, WENDEL JF. Genetic and epigenetic consequences of recent hybridization and polyploidy in Spartina (Poaceae). Mol Ecol [Internet]. 2005 Mar 16;14(4):1163–75. Available from: http://doi.wiley.com/10.1111/j.1365-294X.2005.02488.x
15. HEGARTY MJ, BATSTONE T, BARKER GL, EDWARDS KJ, ABBOTT RJ, HISCOCK SJ. Nonadditive changes to cytosine methylation as a consequence of hybridization and genome duplication in Senecio (Asteraceae). Mol Ecol [Internet]. 2011

Jan;20(1):105–13. Available from: http://doi.wiley.com/10.1111/j.1365-294X.2010.04926.x
16. RAN L, FANG T, RONG H, JIANG J, FANG Y, WANG Y. Analysis of cytosine methylation in early generations of resynthesized Brassica napus. J Integr Agric [Internet]. 2016 Jun;15(6):1228–38. Available from: https://linkinghub.elsevier.com/retrieve/pii/S2095311915612771
17. Xu Y, Zhong L, Wu X, Fang X, Wang J. Rapid alterations of gene expression and cytosine methylation in newly synthesized Brassica napus allopolyploids. Planta. 2009;229(3):471–83.
18. Bird KA, Niederhuth CE, Ou S, Gehan M, Pires JC, Xiong Z, et al. Replaying the evolutionary tape to investigate subgenome dominance in allopolyploid Brassica napus. New Phytol [Internet]. 2021 Apr 9;230(1):354–71. Available from: https://onlinelibrary.wiley.com/doi/10.1111/nph.17137
19. Gaeta RT, Pires JC, Iniguez-Luy F, Leon E, Osborn TC. Genomic Changes in Resynthesized Brassica napus and Their Effect on Gene Expression and Phenotype. Plant Cell [Internet]. 2007 Nov;19(11):3403–17. Available from: http://www.plantcell.org/lookup/doi/10.1105/tpc.107.054346
20. Sehrish T, Symonds VV, Soltis DE, Soltis PS, Tate JA. Gene silencing via DNA methylation in naturally occurring Tragopogon miscellus (Asteraceae) allopolyploids. BMC Genomics. 2014;15(1):1–7.
21. Liu B, Brubaker CL, Mergeai G, Cronn RC, Wendel JF. Polyploid formation in cotton is not accompanied by rapid genomic changes. Genome [Internet]. 2001 Jun 1;44(3):321–30. Available from: http://www.nrcresearchpress.com/doi/10.1139/g01-011
22. Wang J, Tian L, Madlung A, Lee H-S, Chen M, Lee JJ, et al. Stochastic and Epigenetic Changes of Gene Expression in Arabidopsis Polyploids. Genetics [Internet]. 2004 Aug;167(4):1961–73. Available from: http://www.genetics.org/lookup/doi/10.1534/genetics.104.027896
23. Madlung A, Masuelli RW, Watson B, Reynolds SH, Davison J, Comai L. Remodeling of DNA Methylation and Phenotypic and Transcriptional Changes in Synthetic Arabidopsis Allotetraploids. Plant Physiol [Internet]. 2002 Jun 1;129(2):733–46. Available from: http://www.plantphysiol.org/lookup/doi/10.1104/pp.003095
24. Jiang X, Song Q, Ye W, Chen ZJ. Concerted genomic and epigenomic changes accompany stabilization of Arabidopsis allopolyploids. Nat Ecol Evol [Internet]. 2021 Oct 19;5(10):1382–93. Available from: https://www.nature.com/articles/s41559-021-01523-y
25. Shaked H, Kashkush K, Ozkan H, Feldman M, Levy AA. Sequence Elimination and Cytosine Methylation Are Rapid and Reproducible Responses of the Genome to Wide Hybridization and Allopolyploidy in Wheat. Plant Cell [Internet]. 2001 Aug;13(8):1749–59. Available from: http://www.plantcell.org/lookup/doi/10.1105/TPC.010083
26. Dong YZ, Liu ZL, Shan XH, Qiu T, He MY, Liu B. Allopolyploidy in Wheat Induces Rapid and Heritable Alterations in DNA Methylation Patterns of Cellular Genes and Mobile Elements. Russ J Genet [Internet]. 2005 Aug;41(8):890–6. Available from: http://link.springer.com/10.1007/s11177-005-0177-7
27. Edger PP, Smith RD, McKain MR, Cooley AM, Vallejo-Marin M, Yuan Y-WY, et al. Subgenome Dominance in an Interspecific Hybrid, Synthetic Allopolyploid, and a 140-Year-Old Naturally Established Neo-Allopolyploid Monkeyflower. Plant Cell [Internet]. 2017 Sep;29(9):2150–67. Available from: http://www.plantcell.org/lookup/doi/10.1105/tpc.17.00010
28. Doyle JJ, Flagel LE, Paterson AH, Rapp RA, Soltis DE, Soltis PS, et al. Evolutionary Genetics of Genome Merger and Doubling in Plants. Annu Rev Genet [Internet]. 2008 Dec 1;42(1):443–61. Available from: https://www.annualreviews.org/doi/10.1146/annurev.genet.42.110807.091524
29. Van de Peer Y, Ashman T-L, Soltis PS, Soltis DE. Polyploidy: an evolutionary and ecological force in stressful times. Plant Cell [Internet]. 2021 Mar 22;33(1):11–26.

Available from: https://academic.oup.com/plcell/article/33/1/11/6015242
30. Rice A, Šmarda P, Novosolov M, Drori M, Glick L, Sabath N, et al. The global biogeography of polyploid plants. Nat Ecol Evol [Internet]. 2019 Feb;3(2):265–73. Available from: http://www.nature.com/articles/s41559-018-0787-9
31. Van de Peer Y, Mizrachi E, Marchal K. The evolutionary significance of polyploidy. Nat Rev Genet [Internet]. 2017 Jul 15;18(7):411–24. Available from: http://www.nature.com/articles/nrg.2017.26
32. Vidalis A, Živković D, Wardenaar R, Roquis D, Tellier A, Johannes F. Methylome evolution in plants. Genome Biol. 2016;17(1):1–14.
33. Syngelaki E, Schinkel CCF, Klatt S, Hörandl E. Effects of Temperature Treatments on Cytosine-Methylation Profiles of Diploid and Autotetraploid Plants of the Alpine Species Ranunculus kuepferi (Ranunculaceae). Front Plant Sci [Internet]. 2020 Apr 8;11. Available from: https://www.frontiersin.org/article/10.3389/fpls.2020.00435/full
34. Shimizu KK, Fuji S, Marhold K, Watanabe K, Kudoh H. Arabidopsis kamchatica (Fisch. ex DC.) K. Shimizu & Kudoh and A. kamchatica subsp. kawasakiana (Makino) K. Shimizu & Kudoh, New Combinations. Acta Phytotaxon Geobot [Internet]. 2005;56(2):163–72. Available from: https://doi.org/10.18942/apg.KJ00004623241
35. Paape T, Briskine R V., Halstead-Nussloch G, Lischer HEL, Shimizu-Inatsugi R, Hatakeyama M, et al. Patterns of polymorphism and selection in the subgenomes of the allopolyploid Arabidopsis kamchatica. Nat Commun [Internet]. 2018 Dec 25;9(1):3909. Available from: http://www.nature.com/articles/s41467-018-06108-1
36. Briskine R V., Paape T, Shimizu-Inatsugi R, Nishiyama T, Akama S, Sese J, et al. Genome assembly and annotation of Arabidopsis halleri, a model for heavy metal hyperaccumulation and evolutionary ecology. Mol Ecol Resour [Internet]. 2016 Sep;17(5):1025–36. Available from: http://doi.wiley.com/10.1111/1755-0998.12604
37. Goodstein DM, Shu S, Howson R, Neupane R, Hayes RD, Fazo J, et al. Phytozome: a comparative platform for green plant genomics. Nucleic Acids Res [Internet]. 2012 Jan;40(D1):D1178–86. Available from: https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkr944
38. Alonge M, Soyk S, Ramakrishnan S, Wang X, Goodwin S, Sedlazeck FJ, et al. RaGOO: fast and accurate reference-guided scaffolding of draft genomes. Genome Biol [Internet]. 2019 Dec 28;20(1):224. Available from: https://genomebiology.biomedcentral.com/articles/10.1186/s13059-019-1829-6
39. Li H. Minimap2: pairwise alignment for nucleotide sequences. Birol I, editor. Bioinformatics [Internet]. 2018 Sep 15;34(18):3094–100. Available from: https://academic.oup.com/bioinformatics/article/34/18/3094/4994778
40. Marçais G, Delcher AL, Phillippy AM, Coston R, Salzberg SL, Zimin A. MUMmer4: A fast and versatile genome alignment system. Darling AE, editor. PLOS Comput Biol [Internet]. 2018 Jan 26;14(1):e1005944. Available from: https://dx.plos.org/10.1371/journal.pcbi.1005944
41. Kiełbasa SM, Wan R, Sato K, Horton P, Frith MC. Adaptive seeds tame genomic sequence comparison. Genome Res [Internet]. 2011 Mar;21(3):487–93. Available from: http://genome.cshlp.org/lookup/doi/10.1101/gr.113985.110
42. Gu Z, Gu L, Eils R, Schlesner M, Brors B. circlize implements and enhances circular visualization in R. Bioinformatics [Internet]. 2014 Oct;30(19):2811–2. Available from: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btu393
43. R Foundation for Statistical Computing. R: A language and environment for statistical computing [Internet]. Vienna, Austria; 2020. Available from: https://www.r-project.org/
44. Wickham H, Averick M, Bryan J, Chang W, McGowan L, François R, et al. Welcome to the Tidyverse. J Open Source Softw [Internet]. 2019 Nov 21;4(43):1686. Available from: https://joss.theoj.org/papers/10.21105/joss.01686
45. Lawrence M, Huber W, Pagès H, Aboyoun P, Carlson M, Gentleman R, et al. Software for Computing and Annotating Genomic Ranges. Prlic A, editor. PLoS Comput Biol [Internet]. 2013 Aug 8;9(8):e1003118. Available from:

https://dx.plos.org/10.1371/journal.pcbi.1003118
46. Murray MG, Thompson WF. Rapid isolation of high molecular weight plant DNA. Nucleic Acids Res [Internet]. 1980;8(19):4321–6. Available from: https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/8.19.4321
47. Milosavljevic S, Kuo T, Decarli S, Mohn L, Sese J, Shimizu KK, et al. ARPEGGIO: Automated Reproducible Polyploid EpiGenetic GuIdance workflOw. BMC Genomics [Internet]. 2021 Dec 1;22(1):547. Available from: https://bmcgenomics.biomedcentral.com/articles/10.1186/s12864-021-07845-2
48. Andrews S. FastQC: a quality control tool for high throughput sequence data [Internet]. 2010. Available from: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/
49. Krueger F. Trim Galore [Internet]. 2012. Available from: http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/
50. Krueger F, Andrews SR. Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. Bioinformatics [Internet]. 2011 Jun 1;27(11):1571–2. Available from: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btr167
51. Kuo T, Frith MC, Sese J, Horton P. EAGLE: Explicit Alternative Genome Likelihood Evaluator. BMC Med Genomics [Internet]. 2018 Apr 20;11(S2):28. Available from: https://bmcmedgenomics.biomedcentral.com/articles/10.1186/s12920-018-0342-1
52. Asaf S, Khan AL, Khan MA, Waqas M, Kang S-M, Yun B-W, et al. Chloroplast genomes of Arabidopsis halleri ssp. gemmifera and Arabidopsis lyrata ssp. petraea: Structures and comparative analysis. Sci Rep [Internet]. 2017 Dec 8;7(1):7556. Available from: http://www.nature.com/articles/s41598-017-07891-5
53. Ahlert D, Stegemann S, Kahlau S, Ruf S, Bock R. Insensitivity of chloroplast gene expression to DNA methylation. Mol Genet Genomics [Internet]. 2009 Jul 17;282(1):17–24. Available from: http://link.springer.com/10.1007/s00438-009-0440-z
54. Fojtová M, Kovařík A, Matyášek R. Cytosine methylation of plastid genome in higher plants. Fact or artefact? Plant Sci [Internet]. 2001 Mar;160(4):585–93. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0168945200004118
55. García-Alcalde F, Okonechnikov K, Carbonell J, Cruz LM, Götz S, Tarazona S, et al. Qualimap: evaluating next-generation sequencing alignment data. Bioinformatics [Internet]. 2012 Oct 15;28(20):2678–9. Available from: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/bts503
56. Ewels P, Magnusson M, Lundin S, Käller M. MultiQC: summarize analysis results for multiple tools and samples in a single report. Bioinformatics [Internet]. 2016 Oct 1;32(19):3047–8. Available from: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btw354
57. Korthauer K, Chakraborty S, Benjamini Y, Irizarry RA. Detection and accurate false discovery rate control of differentially methylated regions from whole genome bisulfite sequencing. Biostatistics [Internet]. 2019 Jul 1;20(3):367–83. Available from: https://academic.oup.com/biostatistics/article/20/3/367/4899074
58. Taudt AS, Roquis D, Vidalis A, Wardenaar R, Johannes F, Tatche MC. METHimpute: Imputation-guided construction of complete methylomes from WGBS data. bioRxiv [Internet]. 2017;190223. Available from: http://www.biorxiv.org/content/early/2017/09/18/190223.article-info
59. Parker HR, Orjuela S, Martinho Oliveira A, Cereatti F, Sauter M, Heinrich H, et al. The proto CpG island methylator phenotype of sessile serrated adenomas/polyps. Epigenetics [Internet]. 2018 Nov 2;13(10–11):1088–105. Available from: https://www.tandfonline.com/doi/full/10.1080/15592294.2018.1543504
60. Yu G. Variance stabilizing transformations of Poisson, binomial and negative binomial distributions. Stat Probab Lett [Internet]. 2009 Jul;79(14):1621–9. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0167715209001473
61. Park Y, Wu H. Differential methylation analysis for BS-seq data under general

experimental design. Bioinformatics [Internet]. 2016 May 15;32(10):1446–53. Available from: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btw026
62. Auguie B. gridExtra: Miscellaneous Functions for “Grid” Graphics. 2017.
63. Gel B, Serra E. karyoploteR: an R/Bioconductor package to plot customizable genomes displaying arbitrary data. Hancock J, editor. Bioinformatics [Internet]. 2017 Oct 1;33(19):3088–90. Available from: https://academic.oup.com/bioinformatics/article/33/19/3088/3857734
64. Mi H, Ebert D, Muruganujan A, Mills C, Albou L-P, Mushayamaha T, et al. PANTHER version 16: a revised family classification, tree-based classification tool, enhancer regions and extensive API. Nucleic Acids Res [Internet]. 2021 Jan 8;49(D1):D394–403. Available from: https://academic.oup.com/nar/article/49/D1/D394/6027812
65. SHIMIZU-INATSUGI R, LIHOVÁ J, IWANAGA H, KUDOH H, MARHOLD K, SAVOLAINEN O, et al. The allopolyploid Arabidopsis kamchatica originated from multiple individuals of Arabidopsis lyrata and Arabidopsis halleri. Mol Ecol [Internet]. 2009 Oct;18(19):4024–48. Available from: http://doi.wiley.com/10.1111/j.1365-294X.2009.04329.x
66. Alonge M, Shumate A, Puiu D, Zimin A V, Salzberg SL. Chromosome-Scale Assembly of the Bread Wheat Genome Reveals Thousands of Additional Gene Copies. Genetics [Internet]. 2020 Oct 1;216(2):599–608. Available from: https://academic.oup.com/genetics/article/216/2/599/6066189
67. Mason AS, Wendel JF. Homoeologous Exchanges, Segmental Allopolyploidy, and Polyploid Genome Evolution. Front Genet [Internet]. 2020 Aug 28;11. Available from: https://www.frontiersin.org/article/10.3389/fgene.2020.01014/full
68. Xiong Z, Gaeta RT, Pires JC. Homoeologous shuffling and chromosome compensation maintain genome balance in resynthesized allopolyploid Brassica napus. Proc Natl Acad Sci [Internet]. 2011 May 10;108(19):7908–13. Available from: http://www.pnas.org/cgi/doi/10.1073/pnas.1014138108
69. Lloyd A, Blary A, Charif D, Charpentier C, Tran J, Balzergue S, et al. Homoeologous exchanges cause extensive dosage-dependent gene expression changes in an allopolyploid crop. New Phytol [Internet]. 2018 Jan;217(1):367–77. Available from: https://onlinelibrary.wiley.com/doi/10.1111/nph.14836
70. Rousseau-Gueutin M, Morice J, Coriton O, Huteau V, Trotoux G, Nègre S, et al. The Impact of Open Pollination on the Structural Evolutionary Dynamics, Meiotic Behavior, and Fertility of Resynthesized Allotetraploid Brassica napus L. G3 Genes|Genomes|Genetics [Internet]. 2017 Feb 1;7(2):705–17. Available from: https://academic.oup.com/g3journal/article/7/2/705/6027675
71. Li N, Xu C, Zhang A, Lv R, Meng X, Lin X, et al. DNA methylation repatterning accompanying hybridization, whole genome doubling and homoeolog exchange in nascent segmental rice allotetraploids. New Phytol [Internet]. 2019 Jul 30;223(2):979–92. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1111/nph.15820
72. Bertioli DJ, Jenkins J, Clevenger J, Dudchenko O, Gao D, Seijo G, et al. The genome sequence of segmental allotetraploid peanut Arachis hypogaea. Nat Genet [Internet]. 2019 May 1;51(5):877–84. Available from: http://www.nature.com/articles/s41588-019-0405-z
73. LIM KY, MATYASEK R, KOVARIK A, LEITCH AR. Genome evolution in allotetraploid Nicotiana. Biol J Linn Soc [Internet]. 2004 Aug 9;82(4):599–606. Available from: https://academic.oup.com/biolinnean/article-lookup/doi/10.1111/j.1095-8312.2004.00344.x
74. Chester M, Gallagher JP, Symonds V V., Cruz da Silva A V., Mavrodiev E V., Leitch AR, et al. Extensive chromosomal variation in a recently formed natural allopolyploid species, Tragopogon miscellus (Asteraceae). Proc Natl Acad Sci [Internet]. 2012 Jan 24;109(4):1176–81. Available from: http://www.pnas.org/cgi/doi/10.1073/pnas.1112041109
75. Jarvis DE, Ho YS, Lightfoot DJ, Schmöckel SM, Li B, Borm TJA, et al. The genome of

Chenopodium quinoa. Nature [Internet]. 2017 Feb 16;542(7641):307–12. Available from: http://www.nature.com/articles/nature21370
76. Lashermes P, Combes M-C, Hueber Y, Severac D, Dereeper A. Genome rearrangements derived from homoeologous recombination following allopolyploidy speciation in coffee. Plant J [Internet]. 2014 May;78(4):674–85. Available from: https://onlinelibrary.wiley.com/doi/10.1111/tpj.12505
77. Pelé A, Rousseau-Gueutin M, Chèvre A-M. Speciation Success of Polyploid Plants Closely Relates to the Regulation of Meiotic Recombination. Front Plant Sci [Internet]. 2018 Jun 28;9. Available from: https://www.frontiersin.org/article/10.3389/fpls.2018.00907/full
78. PIRES JC, ZHAO J, SCHRANZ ME, LEON EJ, QUIJADA PA, LUKENS LN, et al. Flowering time divergence and genomic rearrangements in resynthesized Brassica polyploids (Brassicaceae). Biol J Linn Soc [Internet]. 2004 Aug 9;82(4):675–88. Available from: https://academic.oup.com/biolinnean/article-lookup/doi/10.1111/j.1095-8312.2004.00350.x
79. Zhao J, Udall JA, Quijada PA, Grau CR, Meng J, Osborn TC. Quantitative trait loci for resistance to Sclerotinia sclerotiorum and its association with a homeologous non-reciprocal transposition in Brassica napus L. Theor Appl Genet [Internet]. 2006 Feb 7;112(3):509–16. Available from: http://link.springer.com/10.1007/s00122-005-0154-5
80. Burns R, Mandáková T, Gunis J, Soto-Jiménez LM, Liu C, Lysak MA, et al. Gradual evolution of allopolyploidy in Arabidopsis suecica. Nat Ecol Evol [Internet]. 2021 Oct 19;5(10):1367–81. Available from: https://www.nature.com/articles/s41559-021-01525-w
81. Doyle JJ, Sherman-Broyles S. Double trouble: taxonomy and definitions of polyploidy. New Phytol [Internet]. 2017 Jan 7;213(2):487–93. Available from: https://onlinelibrary.wiley.com/doi/10.1111/nph.14276
82. Ou S, Su W, Liao Y, Chougule K, Agda JRA, Hellinga AJ, et al. Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. Genome Biol [Internet]. 2019 Dec 16;20(1):275. Available from: https://genomebiology.biomedcentral.com/articles/10.1186/s13059-019-1905-y
83. Barnes S. Comparing Arabidopsis to other flowering plants. Curr Opin Plant Biol [Internet]. 2002 Apr;5(2):128–34. Available from: https://linkinghub.elsevier.com/retrieve/pii/S136952660200239X
84. Seymour DK, Koenig D, Hagmann J, Becker C, Weigel D. Evolution of DNA Methylation Patterns in the Brassicaceae is Driven by Differences in Genome Organization. PLoS Genet. 2014;10(11).
85. Papp B, Pál C, Hurst LD. Dosage sensitivity and the evolution of gene families in yeast. Nature [Internet]. 2003 Jul;424(6945):194–7. Available from: http://www.nature.com/articles/nature01771
86. Chen M, Ha M, Lackey E, Wang J, Chen ZJ. RNAi of met1 Reduces DNA Methylation and Induces Genome-Specific Changes in Gene Expression and Centromeric Small RNA Accumulation in Arabidopsis Allopolyploids. Genetics [Internet]. 2008 Apr 1;178(4):1845–58. Available from: https://academic.oup.com/genetics/article/178/4/1845/6073875
87. Comai L. The advantages and disadvantages of being polyploid. Nat Rev Genet [Internet]. 2005 Nov 11;6(11):836–46. Available from: http://www.nature.com/articles/nrg1711
88. Qiu T, Dong YZZ, Yu XMM, Zhao N, Yang YFF. Analysis of allopolyploidy-induced rapid genetic and epigenetic changes and their relationship in wheat. Genet Mol Res [Internet]. 2017;16(2):1–14. Available from: http://www.funpecrp.com.br/gmr/year2017/vol16-2/pdf/gmr-16-02-gmr.16029303.pdf
89. Comai L, Tyagi AP, Winter K, Holmes-Davis R, Reynolds SH, Stevens Y, et al. Phenotypic Instability and Rapid Gene Silencing in Newly Formed Arabidopsis Allotetraploids. Plant Cell [Internet]. 2000 Sep;12(9):1551–67. Available from: https://academic.oup.com/plcell/article/12/9/1551-1567/6009349

90. Benson CW, Mao Q, Huff DR. Global DNA methylation predicts epigenetic reprogramming and transgenerational plasticity in Poa annua L. Crop Sci [Internet]. 2021 Sep 11;61(5):3011–22. Available from: https://onlinelibrary.wiley.com/doi/10.1002/csc2.20337
91. Gibson G, Wagner G. Canalization in evolutionary genetics: a stabilizing theory? BioEssays [Internet]. 2000 Mar 17;22(4):372–80. Available from: https://onlinelibrary.wiley.com/doi/10.1002/(SICI)1521-1878(200004)22:4%3C372::AID-BIES7%3E3.0.CO;2-J
92. Cabanettes F, Klopp C. D-GENIES: dot plot large genomes in an interactive, efficient and simple way. PeerJ [Internet]. 2018 Jun 4;6:e4958. Available from: https://peerj.com/articles/4958

Supplementary Information

Supplementary material 1

Arabidopsis halleri assembly

Total bp length	196'243'198 bp → 196 Mb
Total number of genes	38'289
Total bp length of genes	108'399'147 bp
Proportion of genes in the assembly	55.2%

Flanking region size and theoretical total bp length*:

Size	Theoretical length	Theoretical % of the assembly
100 bp	7'657'800 bp	3.9%
250 bp	19'144'500 bp	9.8%
500 bp	38'289'000 bp	19.5%
1000 bp	76'578'000 bp	39.0%
2000 bp	153'156'000 bp	78.0%

Number of cytosines for each context:

Total	60’158’068	100%
CG context	8’375’020	14%
CHG context	8’570’006	14%
CHH context	43’213’042	72%

Arabidopsis lyrata assembly

Total bp length	175'182'717bp → 175 Mb
Total number of genes	34'967
Total bp length of genes	99'282'044 bp
Proportion of genes in the assembly	56.7%

Flanking region size and theoretical total bp length*:

Size	Theoretical length	Theoretical % of the assembly
100 bp	6'993'400 bp	4.0%
250 bp	17'483'500 bp	10.0%
500 bp	34'967'000 bp	20.0%
1000 bp	69'934'000 bp	39.9%
2000 bp	139'868'000 bp	79.8%

Number of cytosines for each context:

Total	55’013’262	100%
CG context	7’593’028	14%
CHG context	7’830’903	14%
CHH context	39’589’331	72%

* formula: (size * 2 * total genes) / total assembly length

Supplementary Table 1

Supplementary Table 1: incubator settings for cold and hot conditions.

	Cold conditions	Hot conditions
Max day temperature	22 °C	26 °C
Light hours	16h	16h
Dark hours	8h	8h
Winter duration	8 weeks	4 weeks
Winter temperature	4 °C	
Relative humidity	60%	

Supplementary Table 2

				Cold conditions							Hot conditions						
	side	context	type	total	avg	sd	median	max bp	length	totC	total	avg	sd	median	max bp	length	totC
PvsS1	halleri	CG	all	1602	501 bp	413 bp	415 bp	5394 bp	0.8 Mb	33'694	1990	484 bp	365 bp	424 bp	5192 bp	0.9 Mb	50'129
			hyper	923	585 bp	458 bp	498 bp	5394 bp	0.5 Mb	20'740	1175	536 bp	402 bp	468 bp	5192 bp	0.6 Mb	30'960
			hypo	679	388 bp	308 bp	346 bp	3785 bp	0.3 Mb	12'954	815	408 bp	288 bp	373 bp	3786 bp	0.3 Mb	19'169
		CHG	all	2063	537 bp	423 bp	462 bp	5256 bp	1.1 Mb	46'396	6115	723 bp	508 bp	614 bp	5595 bp	4.4 Mb	259'599
			hyper	878	593 bp	490 bp	490 bp	5256 bp	0.5 Mb	19'041	601	558 bp	494 bp	464 bp	5595 bp	0.3 Mb	15'892
			hypo	1185	496 bp	360 bp	446 bp	4841 bp	0.6 Mb	27'355	5514	741 bp	506 bp	633 bp	5335 bp	4.1 Mb	243'707
		CHH	all	1046	450 bp	269 bp	416 bp	2412 bp	0.5 Mb	95'035	2118	408 bp	264 bp	388 bp	2489 bp	0.9 Mb	203'583
			hyper	757	457 bp	271 bp	431 bp	2412 bp	0.4 Mb	68'158	455	428 bp	264 bp	416 bp	2489 bp	0.2 Mb	44'626
			hypo	289	433 bp	265 bp	397 bp	1754 bp	0.1 Mb	26'877	1663	402 bp	264 bp	380 bp	2360 bp	0.7 Mb	158'957
	lyrata	CG	all	2425	521 bp	492 bp	419 bp	8952 bp	1.3 Mb	54'352	2611	532 bp	486 bp	448 bp	8893 bp	1.4 Mb	70'005
			hyper	1710	586 bp	547 bp	474 bp	8952 bp	1 MB	42'068	1923	578 bp	528 bp	479 bp	8893 bp	1.1 Mb	56'045
			hypo	715	367 bp	270 bp	318 bp	2747 bp	0.3 Mb	12'284	688	405 bp	307 bp	374 bp	2938 bp	0.3 Mb	13'960
		CHG	all	2550	555 bp	560 bp	448 bp	8141 bp	1.4 Mb	63'197	2421	632 bp	558 bp	514 bp	5952 bp	1.5 Mb	81'080
			hyper	1642	641 bp	639 bp	502 bp	8141 bp	1.1 Mb	47'703	1837	682 bp	587 bp	557 bp	5952 bp	1.3 Mb	68'156
			hypo	908	397 bp	323 bp	346 bp	5039 bp	0.3 Mb	15'494	584	475 bp	422 bp	416 bp	4665 bp	0.3 Mb	12'924
		CHH	all	2362	435 bp	295 bp	412 bp	5442 bp	1 Mb	219'176	4136	475 bp	331 bp	442 bp	5544 bp	2 Mb	424'724
			hyper	756	453 bp	375 bp	408 bp	5442 bp	0.3 Mb	69'645	3404	490 bp	348 bp	451 bp	5544 bp	1.7 Mb	355'109
			hypo	1606	427bp	249 bp	414 bp	3533 bp	0.7 Mb	149'531	732	409 bp	219 bp	410 bp	1523 bp	0.3 Mb	69'615
PvsS4	halleri	CG	all	3910	429 bp	316 bp	380 bp	5460 bp	1.6 Mb	72'602	3447	448 bp	331 bp	393 bp	5334 bp	1.6 Mb	73'384
			hyper	1717	489 bp	360 bp	433 bp	5460 bp	0.8 Mb	33'265	1297	504 bp	397 bp	431 bp	5334 bp	0.7 Mb	28'426
			hypo	2193	382 bp	269 bp	341 bp	2810 bp	0.8 Mb	39'337	2150	414 bp	279 bp	372 bp	2838 bp	0.9 Mb	44'958
		CHG	all	3646	603 bp	433 bp	530 bp	4475 bp	2.2 Mb	90'497	8746	653 bp	492 bp	551 bp	6160 bp	5.7 Mb	239'804
			hyper	659	601 bp	489 bp	518 bp	4224 bp	0.4 Mb	15'139	560	570 bp	473 bp	466 bp	5328 bp	0.3 Mb	13'946
			hypo	2987	603 bp	420 bp	532 bp	4475 bp	1.8 Mb	75'358	8186	658 bp	493 bp	556 bp	6160 bp	5.4 Mb	225'858
		CHH	all	3746	486 bp	306 bp	449 bp	4231 bp	1.8 Mb	363'659	8538	509 bp	368 bp	455 bp	6046 bp	4.3 Mb	846'627
			hyper	227	399 bp	240 bp	403 bp	1441 bp	0.1 Mb	18'166	242	423 bp	295 bp	370 bp	1655 bp	0.1 Mb	22'768
			hypo	3519	492 bp	308 bp	452 bp	4231 bp	1.7 Mb	345'493	8296	512 bp	370 bp	457 bp	6046 bp	4.2 Mb	823'859
	lyrata	CG	all	4737	459 bp	439 bp	379 bp	9102 bp	2.2 Mb	93'107	3874	501 bp	507 bp	412 bp	10854 bp	1.9 Mb	85'814
			hyper	2662	525 bp	525 bp	422 bp	9102 bp	1.4 Mb	58'794	2295	565 bp	602 bp	458 bp	10854 bp	1.3 Mb	56'763
			hypo	2075	376 bp	273 bp	329 bp	3836 bp	0.8 Mb	34'313	1579	407 bp	300 bp	359 bp	3662 bp	0.6 Mb	29'051
		CHG	all	3974	544 bp	521 bp	446 bp	11140 bp	2.2 Mb	89'725	5848	810 bp	743 bp	612 bp	8564 bp	4.8 Mb	249'784
			hyper	1292	615 bp	663 bp	476 bp	8144 bp	0.8 Mb	35'325	4020	942 bp	818 bp	717 bp	8564 bp	3.8 Mb	210'387
			hypo	2682	510 bp	433 bp	436 bp	11140 bp	1.4 Mb	54'400	1828	521 bp	418 bp	444 bp	4465 bp	1 Mb	39'397
		CHH	all	8530	510 bp	392 bp	455 bp	5008 bp	4.3 Mb	889'850	3011	430 bp	360 bp	387 bp	6014 bp	1.3 Mb	270'570
			hyper	556	399 bp	386 bp	320 bp	3417 bp	0.2 Mb	42'448	1101	429 bp	461 bp	347 bp	6014 bp	0.5 Mb	100'682
			hypo	7974	518 bp	391 bp	465 bp	5008 bp	4.1 Mb	847'402	1910	430 bp	285 bp	406 bp	3404 bp	0.8 Mb	169'888

Supplementary Table 2: general statistics on DMRs found in the comparisons between progenitors and first (PvsS1) and fourth (PvsS4) generation synthetics. For each condition, subgenome, context and direction of methylation (hyper = increase of methylation in the synthetic, hypo = decrease of methylation in the synthetic), several statistics were computed: total amont of DMRs (total), average (avg), median (median), standard deviation (sd) and maximum (max bp) of DMR lengths, the total length of all DMRs (length) with the total amont of cytosines covered (totC).

Supplementary Table 3

Supplementary Table 3: general mapping statistics for the three aligners minimap2, mummer4 and lastal when aligning the A. lyrata and A. halleri *genome to be scaffolded to the A. lyrata chromosome-level reference assembly. The first three rows provide statistics on scaffolds and the last three rows complement this information with base-pairs statistics.*

	Arabidopsis halleri			*Arabidopsis lyrata*		
	minimap2	mummer4	lastal	minimap2	mummer4	Lastal
Total scaffolds (%)	2’239 (100%)			1’675 (100%)		
Unmapped scaffolds (%)	1275 (67%)	1’701 (76%)	1’877 (84%)	922 (55%)	1’187 (71%)	1’255 (75%)
Mapped scaffolds (%)	964 (43%)	538 (24%)	362 (16%)	753 (45%)	488 (29%)	420 (25%)
Total base pairs (%)	196'243'198 (100%)			175'182’717 (100%)		
Unmapped base pairs (%)	10’059’804 (5%)	16’577’860 (8%)	79’786’837 (41%)	6’908’420 (4%)	8’797’161 (5%)	50’320’313 (29%)
Mapped base pairs (%)	186’183’394 (95%)	179’665’338 (92%)	116’456’361 (59%)	168’274’297 (96%)	166’385’556 (95%)	124’862’404 (71%)

Supplementary Table 4

Supplementary Table 4: expected genome sizes with and without expected chromosome losses. Expected chromosome losses were computed by measuring the length of all regions with less than 2x coverage.

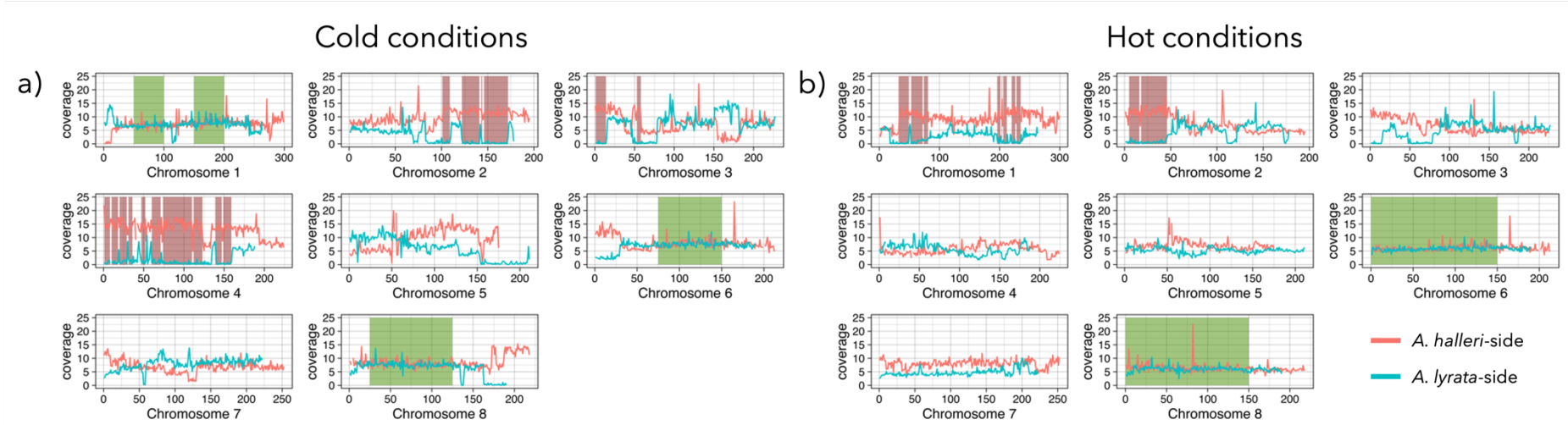
	A. kamchatica synthetics (Generation 4)			
Condition	Cold		Hot	
Expected genome size	475Mbp			
Sub-genome	H	L	H	L
Expected chromosome loss	11Mb	55Mb	1.5Mb	35Mb
Expected genome size with chromosome losses	409Mbp		438.5Mbp	
Genome size measured by flow cytometry	479Mbp (Generation 6)		473Mbp (Generation 7)	

Supplementary Table 5

Supplementary Table 5: Overview of overlapping DMRs found when comparing first and fourth generation synthetics against first generation progenitors.

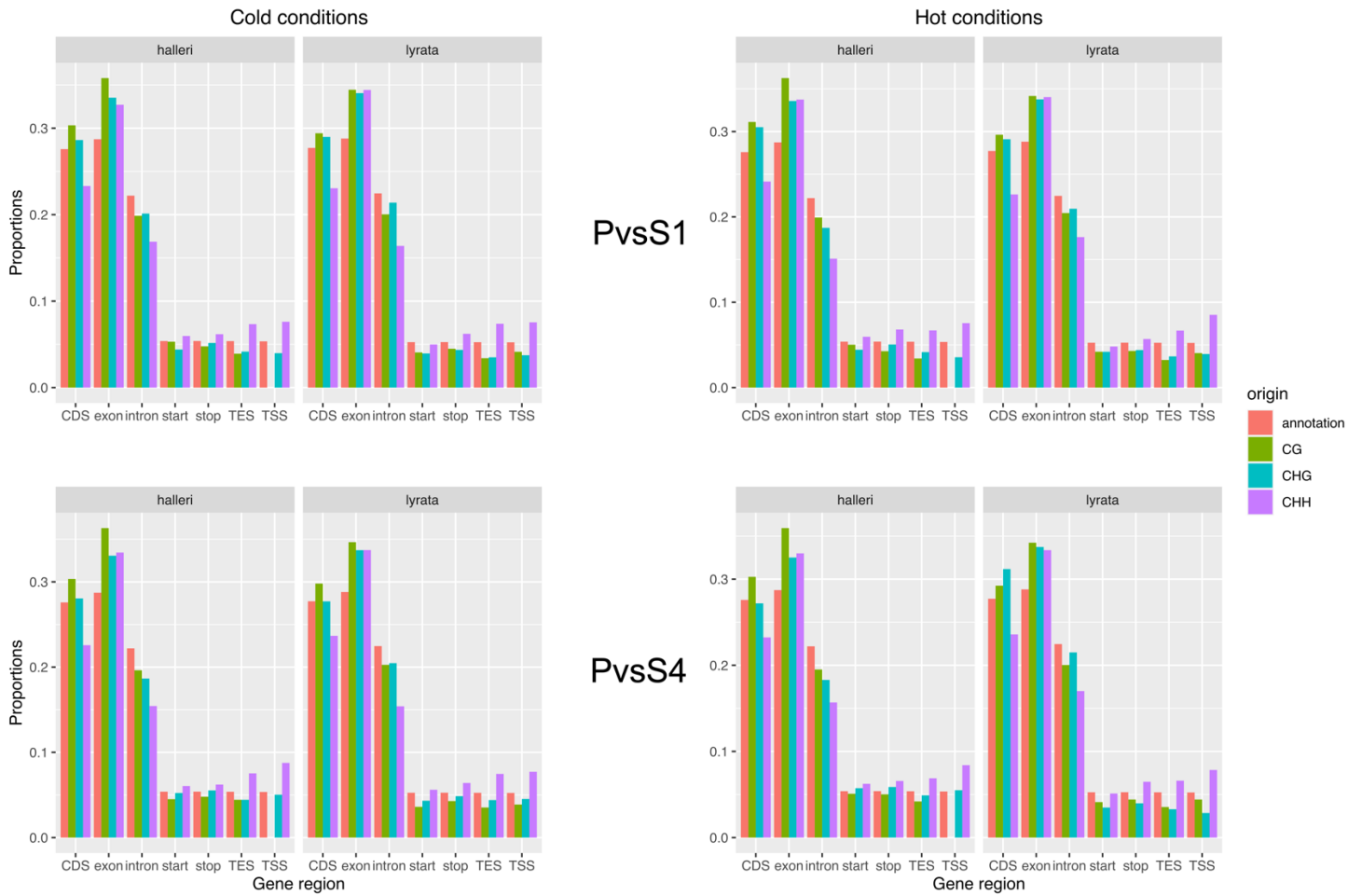
	Cold conditions						Hot conditions					
	CG context		CHG context		CHH context		CG context		CHG context		CHH context	
	hal	*lyr*	*hal*	*lyr*	*hal*	*lyr*	*hal*	*lyr*	*hal*	*lyr*	*hal*	*lyr*
Total DMRs G1	1602	2425	2063	2550	1046	2362	1990	2611	6115	2421	2118	4136
Overlapping DMRs	1144	1846	1252	1474	376	1569	1270	1886	2986	1564	1458	1290
Overlap % G1	71.4 %	76.1 %	60.7 %	57.8 %	35.9 %	66.4 %	63.8 %	72.2%	48.8 %	64.6%	68.8 %	31.2%
Total DMRs G4	3910	4737	3646	3974	3746	8530	3447	3874	8746	5848	8538	3011
Overlap % G4	29.3 %	39.0 %	34.3 %	37.1 %	10.0 %	18.4 %	36.8 %	48.7 %	34.1 %	26.7 %	17.1 %	42.8 %

Supplementary Figure 1



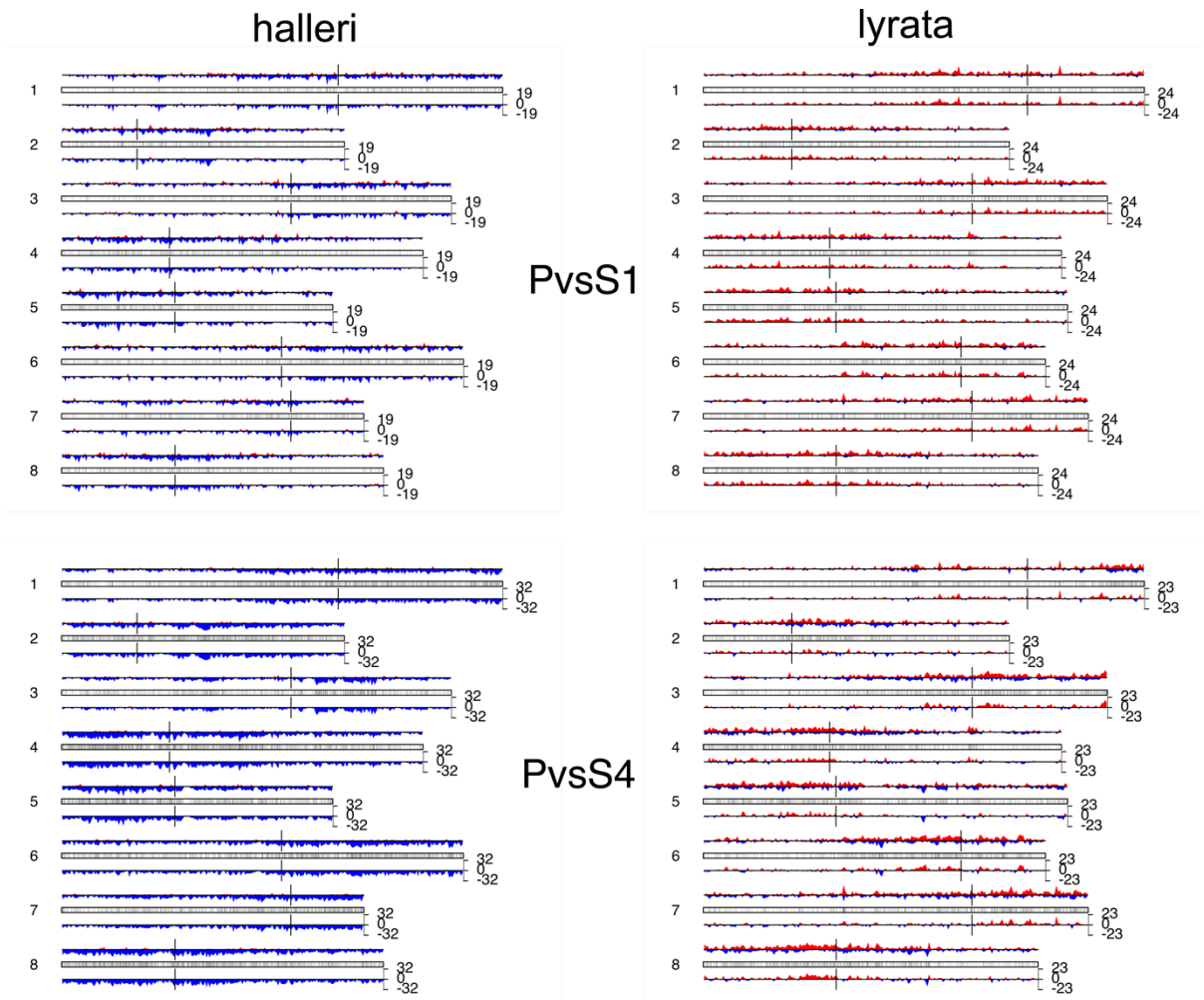
Supplementary Figure 1: overlapped coverage values for A. halleri and A. lyrata subgenomes in A. kamchatica synthetics (fourth generation) used to define HE regions. HE regions (red areas) were defined as regions where the A. halleri coverage values were >9x and A. lyrata coverage values were <2x for cold conditions (a) and hot conditions (b). For downstream analyses we also defined normal regions qualitatively by selecting ranges where A. halleri and A. lyrata coverages were similar.

Supplementary Figure 2



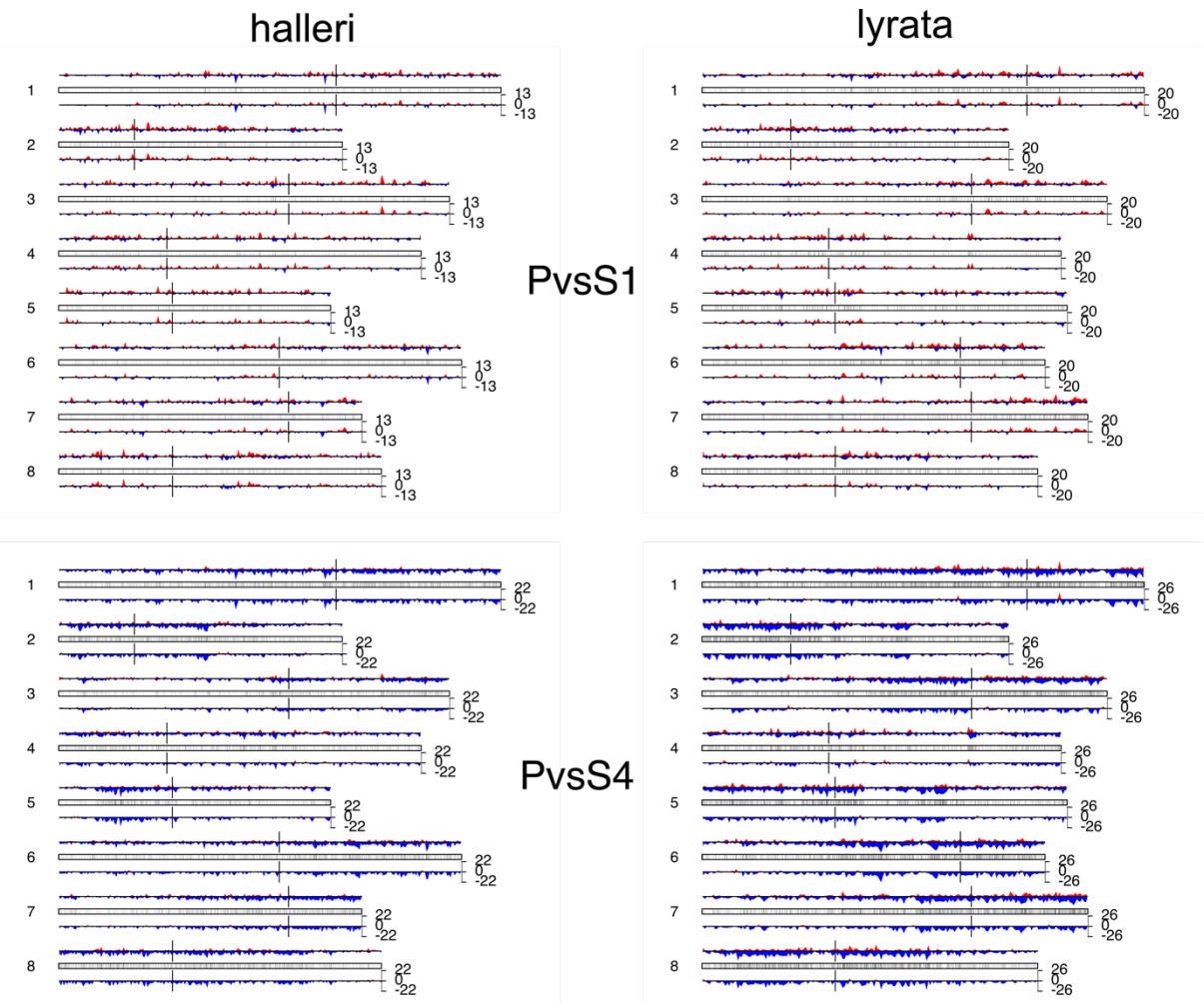
Supplementary Figure 2: proportions of overlaps between DMRs and genic functional regions. Left and right panels show results for cold and hot conditions respectively, while first and second row represent progenitors against first (PvsS1) and fourth (PvsS4) generation synthetics. Proportions were obtained by counting all overlaps between a DMRs and any functional region (i.e.a DMR was counted multiple times if it overlapped with different functional regions) and divided by the total amount of overlaps. These proportions were computed for each context. The genome annotation was used as a reference to compute expected proportions for all functional regions.

Supplementary Figure 3



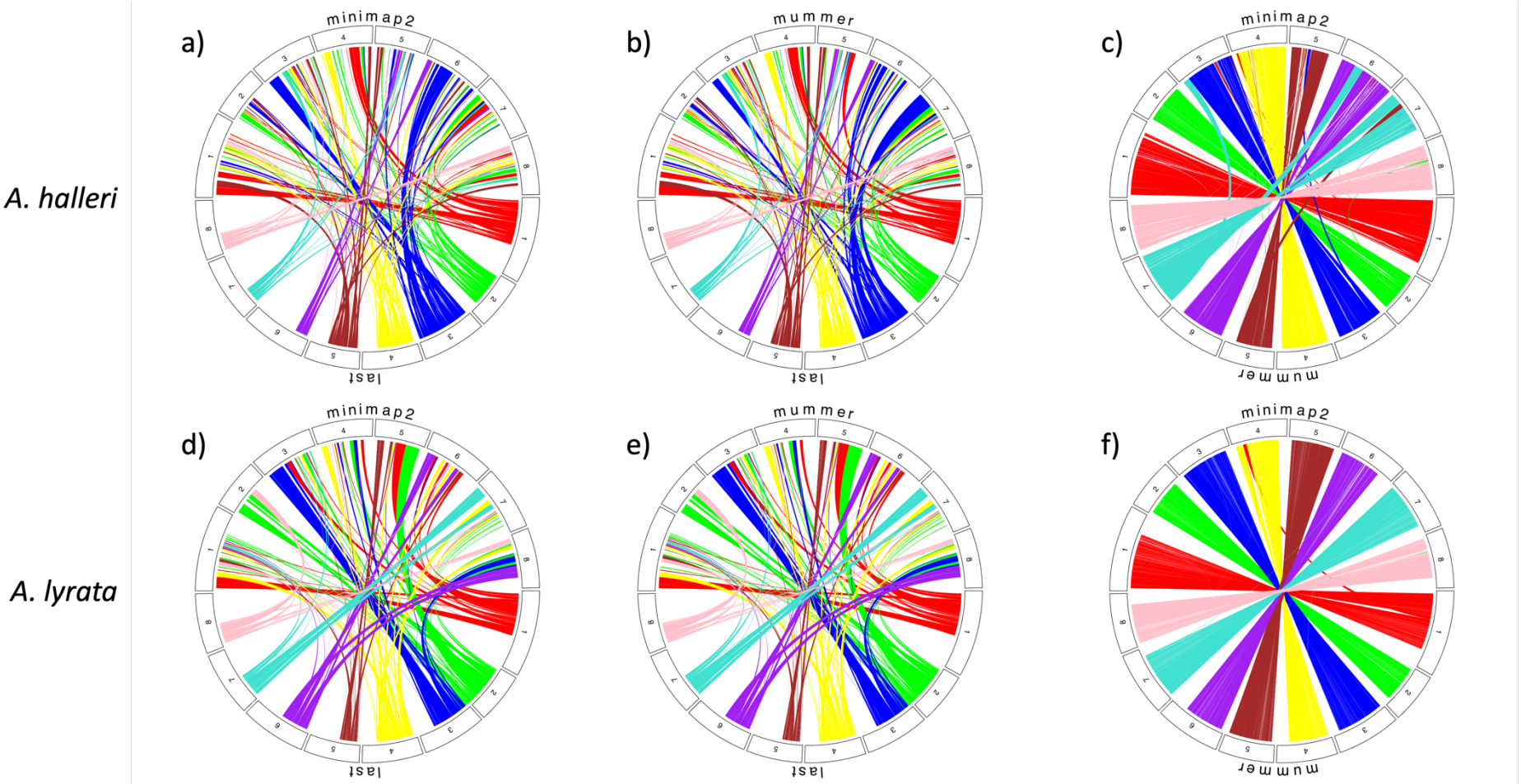
Supplementary Figure 3: spatial distributions of DMRs along chromosomes in hot conditions. Each plot shows eight numbered chromosomes represented by white rectangles, with each DMR shown as a black vertical line. On top of each chromosome the density of DMRs is shown (red for hypermethylated DMRs and blue for hypomethylated DMRs), computed as the amount of DMRs in 10kb windows. At the bottom of each chromosome the difference between two densities is shown (red for excess hypermethylated and blue for excess hypomethylated regions). At both top and bottom of each chromosome, vertical bars indicate approximate centromere coordinates.

Supplementary Figure 4



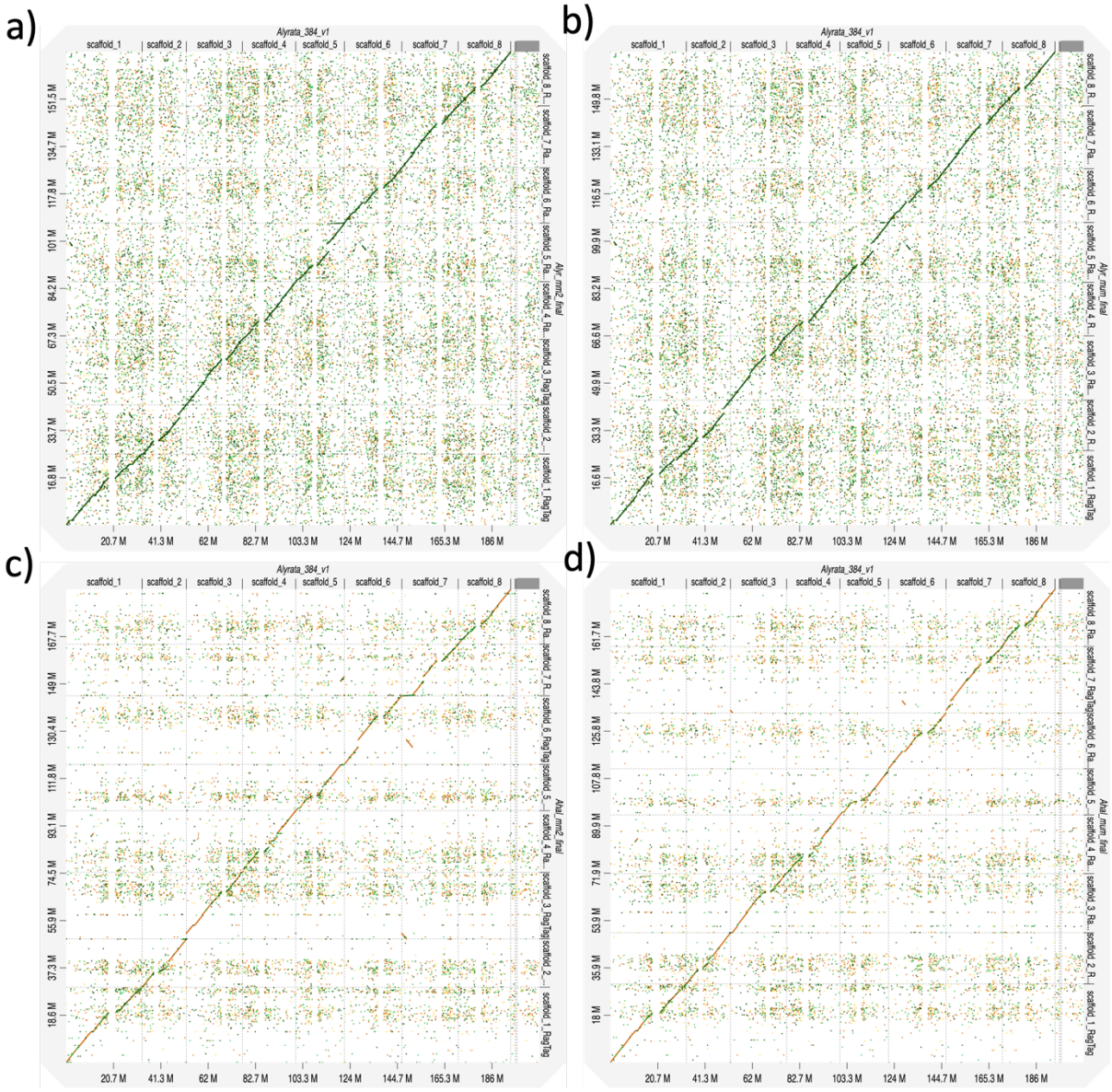
Supplementary Figure 4: spatial distributions of DMRs along chromosomes in cold conditions. Each plot shows eight numbered chromosomes represented by white rectangles, with each DMR shown as a black vertical line. On top of each chromosome the density of DMRs is shown (red for hypermethylated DMRs and blue for hypomethylated DMRs), computed as the amount of DMRs in 10kb windows. At the bottom of each chromosome the difference between two densities is shown (red for excess hypermethylated and blue for excess hypomethylated regions). At both top and bottom of each chromosome, vertical bars indicate approximate centromere coordinates.

Supplementary Figure 5



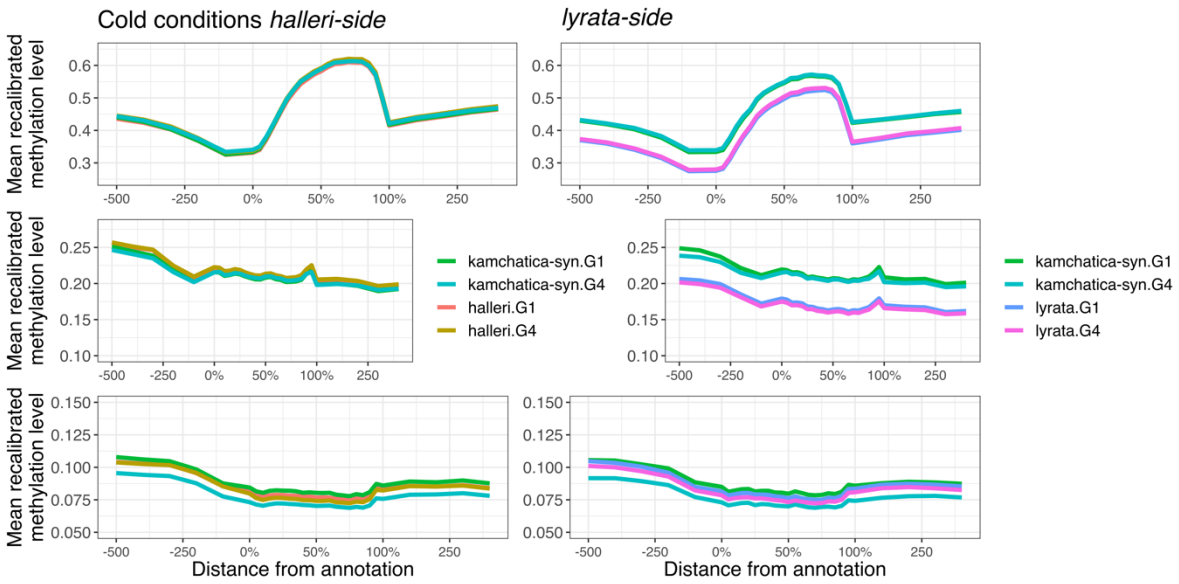
Supplementary Figure 5: Circular plot was used to qualitatively assess the consistency between mappers. Plots show three pairwise comparisons: last vs minimap2 (a, d), last vs mummer4 (b, e) and minimap2 vs mummer4 (c, f). All of these comparisons were done for both A. halleri (a-c) and A. lyrata (d-f). In each plot the chromosome sizes used were always from the reference A. lyrata genome assembly. For each plot, the reference mapper was always the one with the lowest mapping rate. Every scaffold of the reference mapper was matched against the other mapper and, in case of a match, a line was drawn to "link" the scaffold's position in the reference mapper and the other mapper. Lines were colored according to scaffold of origin, for a total of eight colors.

Supplementary Figure 6



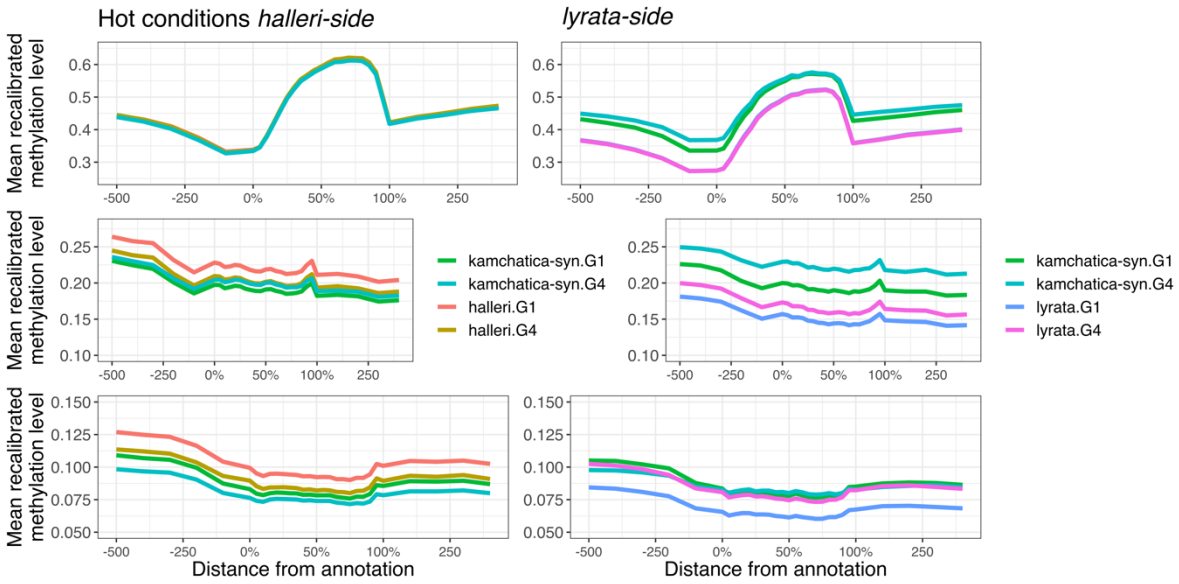
Supplementary Figure 6: dot plot view of the scaffolded A. lyrata (a-b) and A. halleri (c-d) genomes against the reference A. lyrata chromosome level assembly. The two left panels use the minimap2 scaffolded assembly (a,c) while the right panels use the output from mummer4 (b,d). Dots and lines in the plot represent scaffolds. Plots were generated using D-GENIES (92).

Supplementary Figure 7



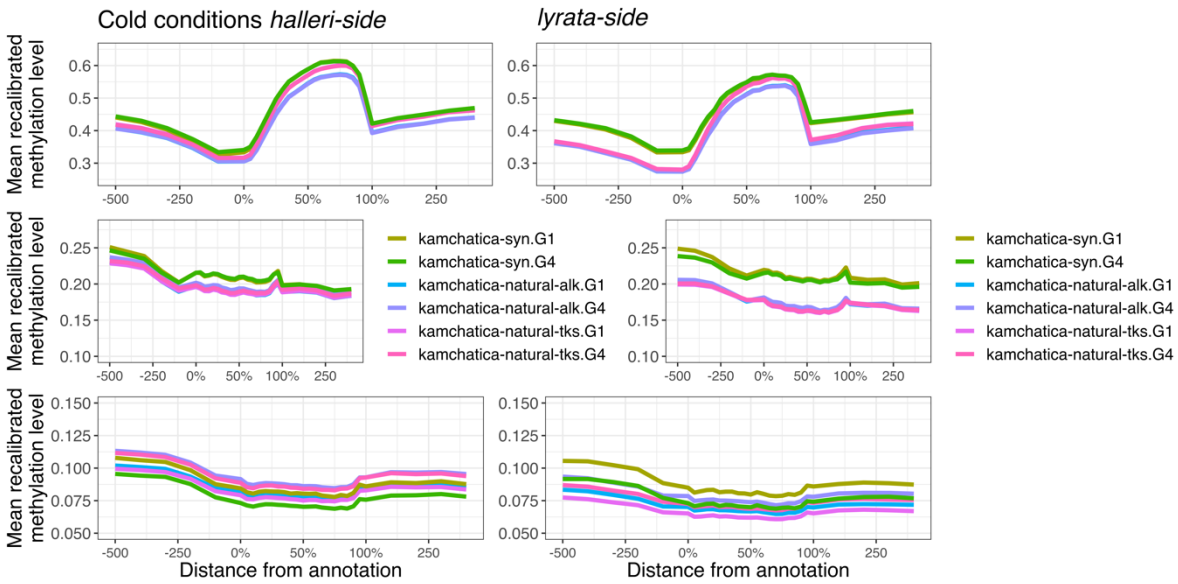
Supplementary Figure 7: average mean recalibrated methylation level within and around gene bodies for diploid and synthetic individuals in cold conditions. The x-axis shows a range within a gene body (0-100%) and 500bp upstream and downstream. Plots on the left and right side show halleri-side and lyrate-side methylation levels respectively.

Supplementary Figure 8



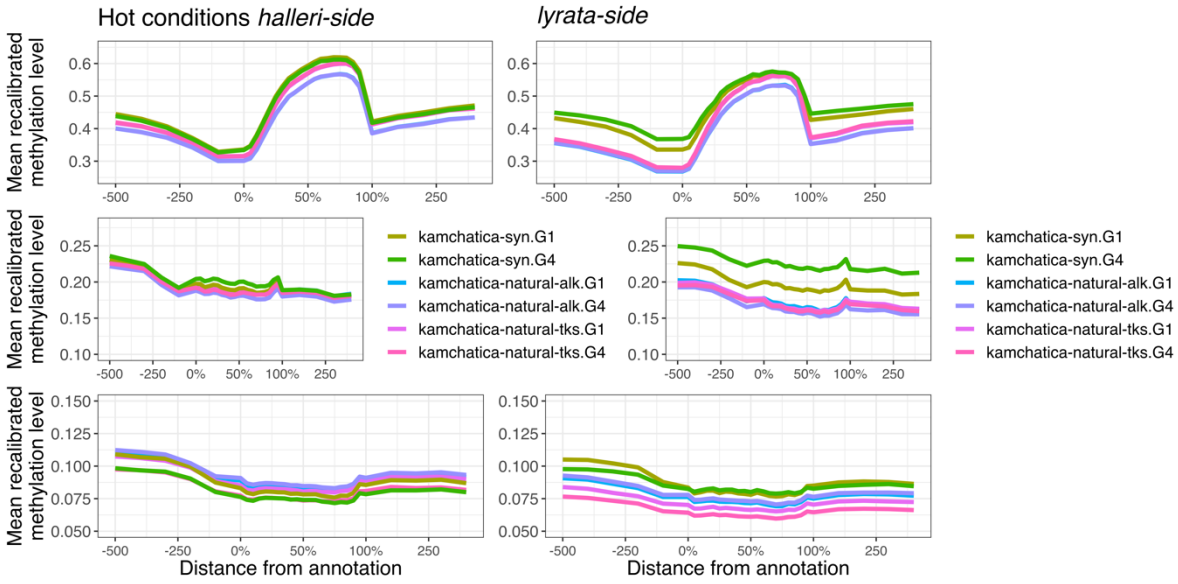
Supplementary Figure 8: average mean recalibrated methylation level within and around gene bodies for diploid and synthetic individuals in hot conditions. The x-axis shows a range within a gene body (0-100%) and 500bp upstream and downstream. Plots on the left and right side show halleri-side and lyrate-side methylation levels respectively.

Supplementary Figure 9



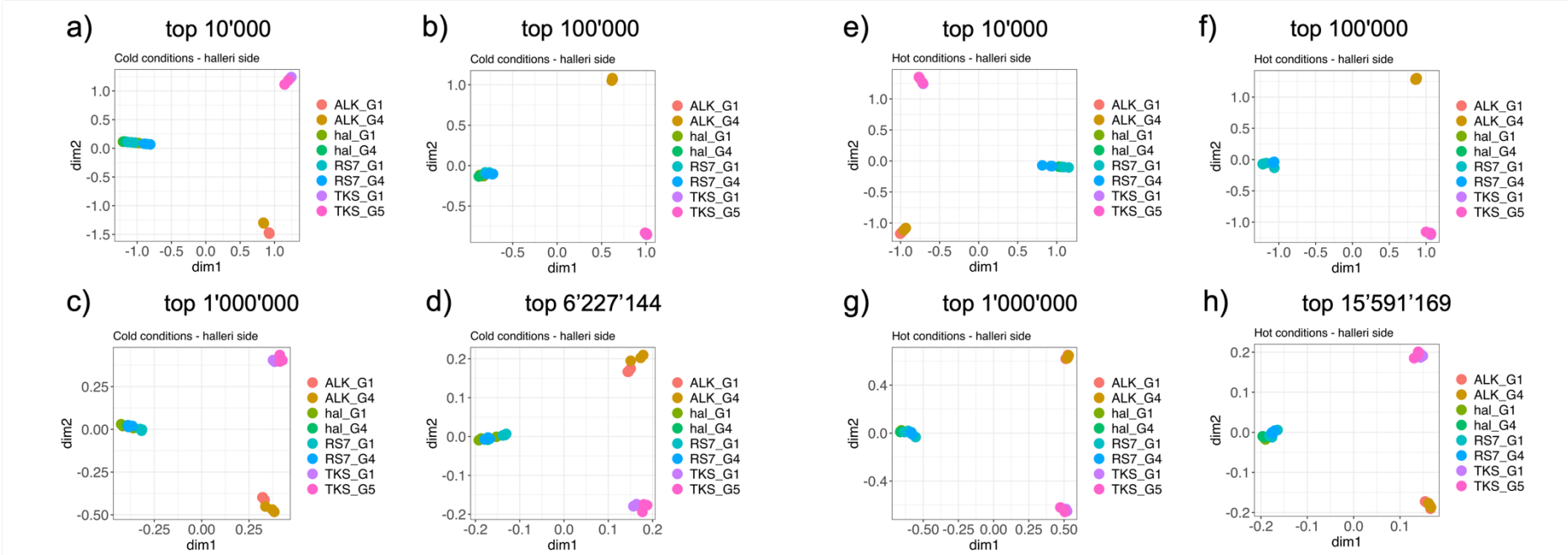
Supplementary Figure 9: average mean recalibrated methylation level within and around gene bodies for synthetic and natural individuals in cold conditions. The x-axis shows a range within a gene body (0-100%) and 500bp upstream and downstream. Plots on the left and right side show halleri-side and lyrate-side methylation levels respectively.

Supplementary Figure 10



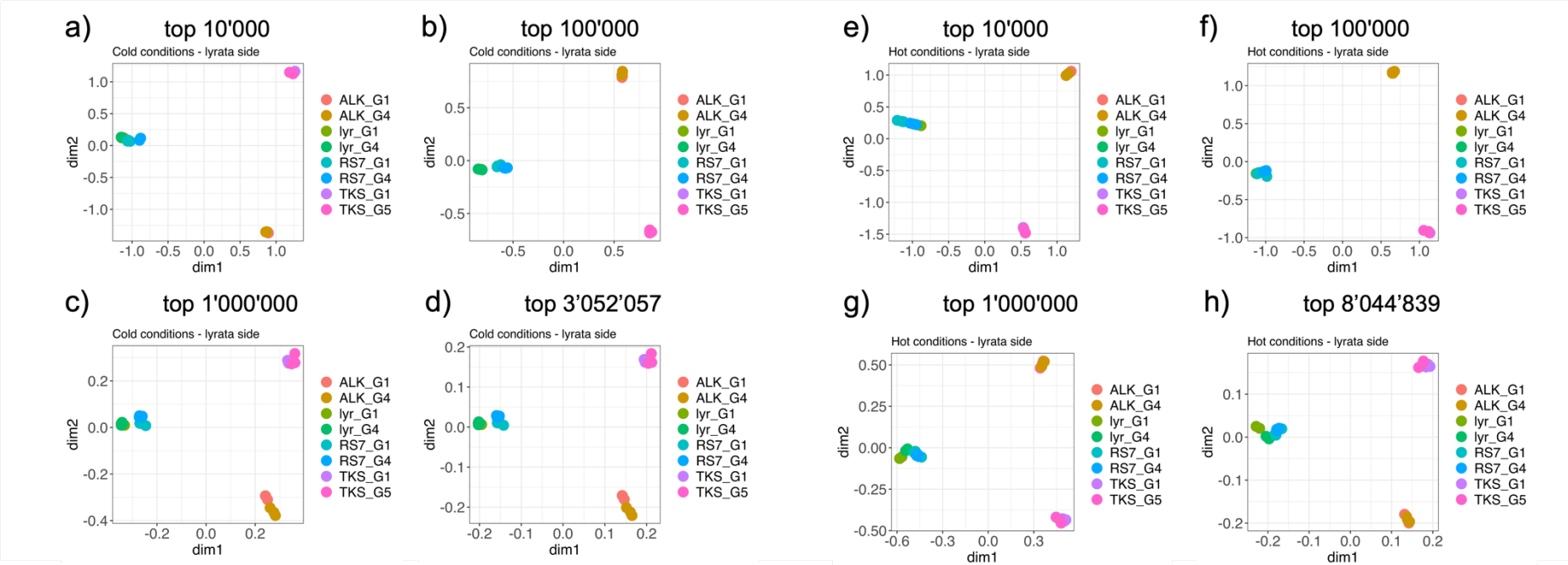
Supplementary Figure 10: average mean recalibrated methylation level within and around gene bodies for synthetic and natural individuals in hot conditions. The x-axis shows a range within a gene body (0-100%) and 500bp upstream and downstream. Plots on the left and right side show halleri-side and lyrate-side methylation levels respectively.

Supplementary Figure 11



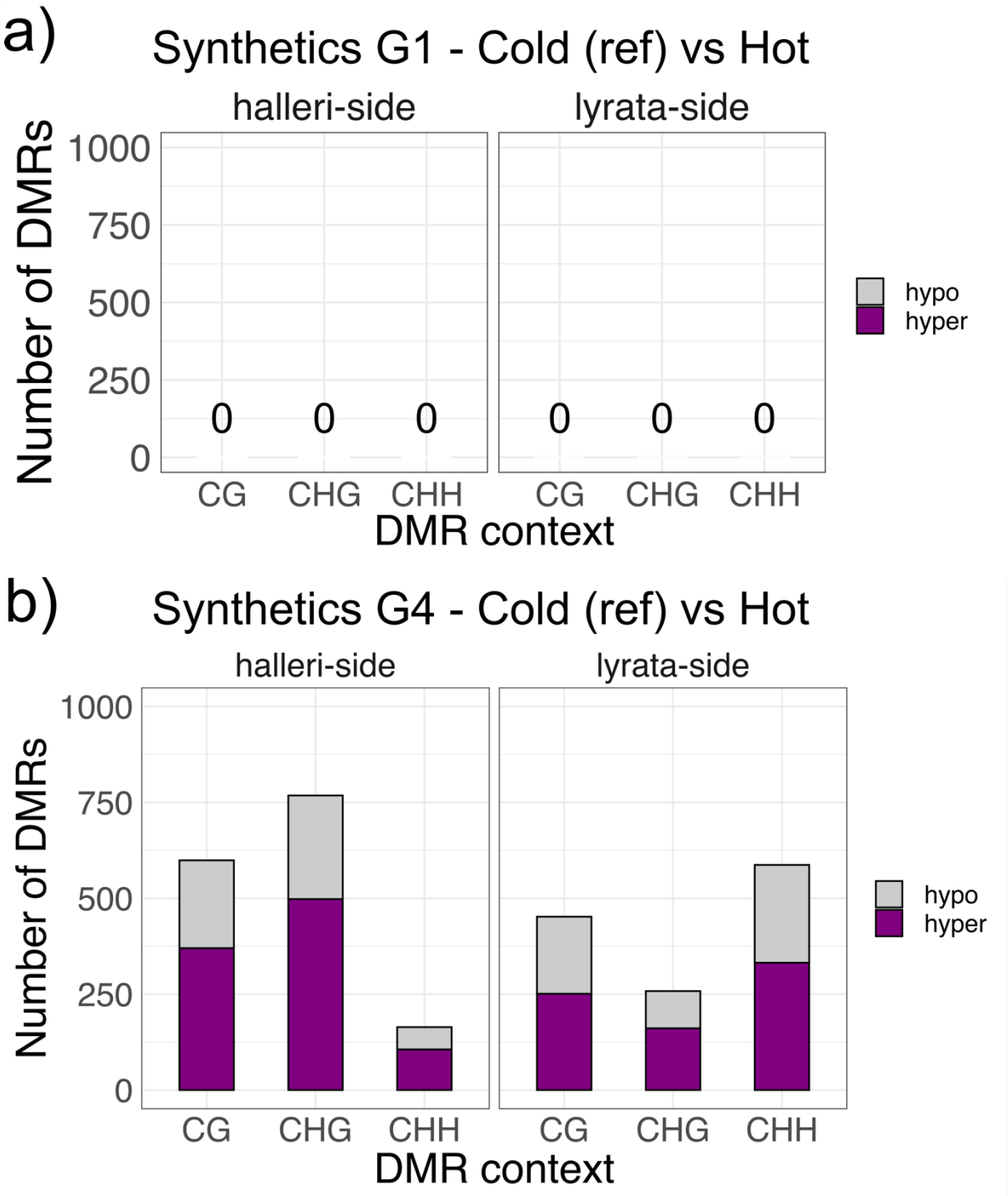
Supplementary Figure 11: MDS plots for the halleri side of all samples in cold conditions (a-d) and hot conditions (e-h). Four threshold were selected to assess the relationship across samples: top 10'000 (a,e), 100'000 (b,f), 1'000'000 (c,g) and all cytosines (d,h).

Supplementary Figure 12



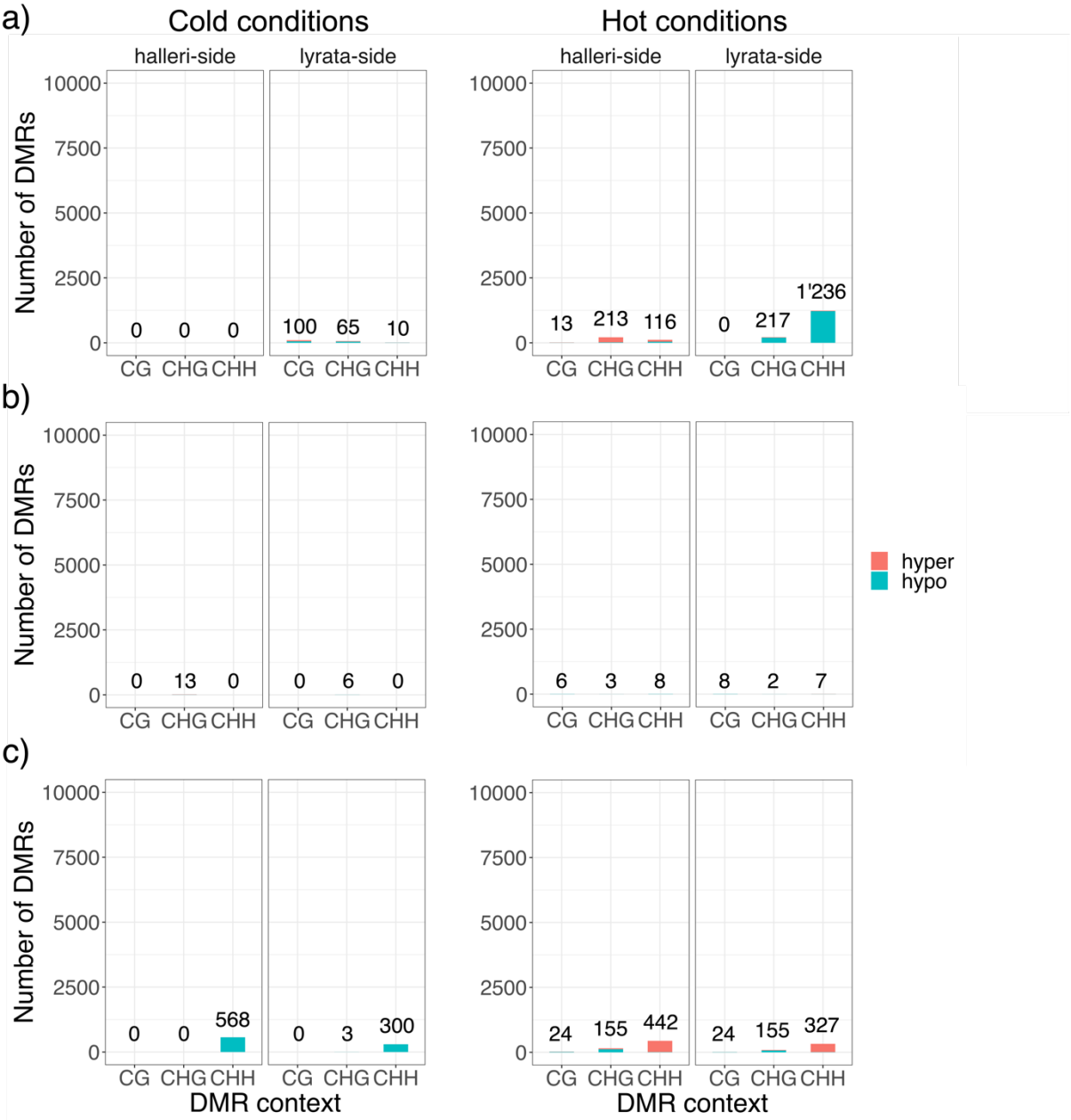
Supplementary Figure 12: MDS plots for the lyrata side of all samples in cold conditions (a-d) and hot conditions (e-h). Four threshold were selected to assess the relationship across samples: top 10'000 (a,e), 100'000 (b,f), 1'000'000 (c,g) and all cytosines (d,h).

Supplementary Figure 13



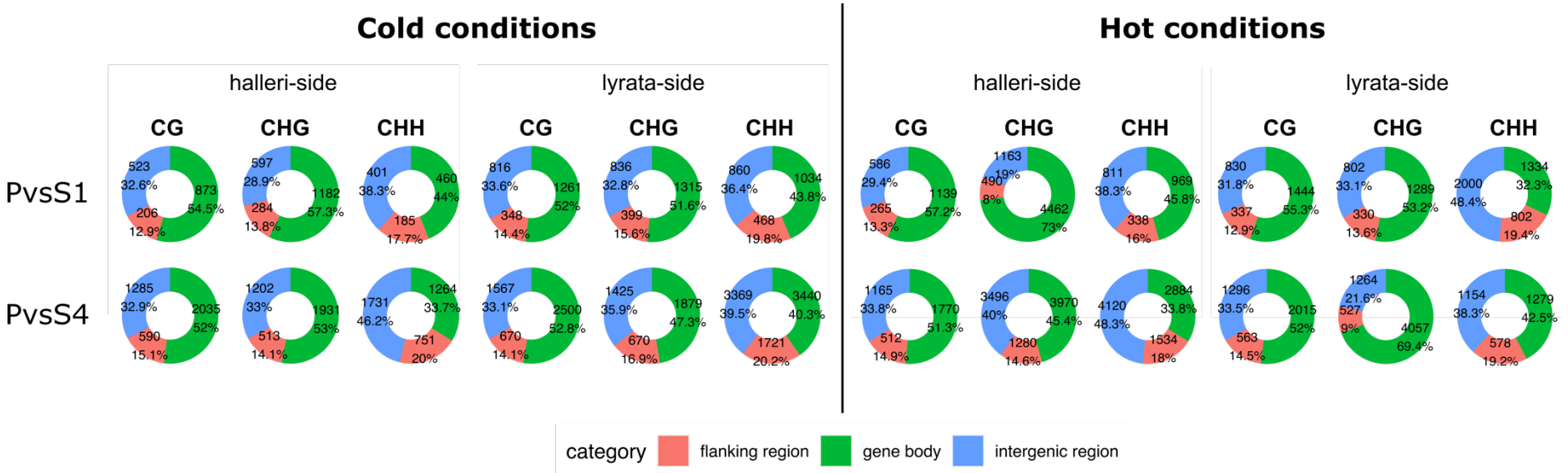
Supplementary Figure 13: total DMRs for each methylation context and subgenome when comparing first (a) and fourth generation synthetics (b) in cold conditions (reference) against hot conditions.

Supplementary Figure 14



Supplementary Figure 14: Total DMRs found when comparing fourth and first generation progenitors (a), first and fourth generation natural Alaska (b) and Takashima (c) lines. The exact number of DMRs is shown at the top of each bar, for each context and subgenome. Hyper- or hypo-methylated DMRs represent methylation increases or decreases respectively in the fourth generation compared to the first.

Supplementary Figure 15



Supplementary Figure 15: proportions of DMRs overlapping with different genomic functional regions. Proportions were computed for each condition (hot and cold), subgenome (halleri and lyrata side) and methylation context (CG, CHG, CHH). The first and second row are for DMRs found when comparing progenitors to first (PvsS1) and fourth (PvsS4) generation synthetics. Gene bodies were defined as regions between the transcriptional start and end site of a gene, flanking regions as 500bp regions around gene bodies and intergenic regions as the remaining part of the genome.

Chapter 3: Large transcriptomic changes in newly formed polyploid partially overlap with DNA methylation patterns

Stefan Milosavljevic[1,2], Aki Morishima[1], Lucas Mohn[1], Jun Sese[3,4], Kentaro K. Shimizu[1,5], Mark D. Robinson[2,6] and Rie Shimizu Inatsugi[1]

[1] Department of Evolutionary Biology and Environmental Studies, University of Zurich, Zurich, Switzerland
[2] SIB Swiss Institute of Bioinformatics, University of Zurich, Zurich, Switzerland
[3] AIST Artificial Intelligence Research Center, Tokyo, Japan
[4] Humanome Lab Inc., Chuo-ku, Tokyo, Japan
[5] Kihara Institute for Biological Research, Yokohama City University, Yokohama, Japan
[6] Department of Molecular Life Sciences, University of Zurich, Zurich, Switzerland

Contributions

R.S.-I. conceived and designed the study; S.M. analyzed the data and wrote the manuscript; R.S.-I., A.M. and L.M. conducted and maintained the experiment; A.M. and L.M. performed all wet lab work including sample preparation for sequencing; R.S.-I., J.S, K.K.S. and M.D.R. provided feedback, suggestions and guidance throughout the project.

Abstract

Large scale gene expression changes occurring right after formation of a plant polyploid species are known to have important effects, possibly explaining the prevalence and success of polyploidy in the evolutionary history of land plants. Studies in synthetic polyploids observed deviations in expression patterns compared to their progenitor species, showing variability in the impact of genome duplication across different species. With the advent of novel approaches to more accurately quantify expression changes, and the increasing knowledge on critical factors affecting expression, it becomes possible to offer a more comprehensive picture on the early stages of polyploidy. In this study, we analyzed expression changes in a synthetic *Arabidopsis kamchatica* allotetraploid grown in two different conditions (hot and cold) with the same experimental design from Chapter 2. These results were examined with the results obtained for DNA methylation changes to assess the consistency and association between the two.

Our results showed partial consistency between DNA methylation and expression, with a much more extensive and rapid amount of changes in expression. Nevertheless, genes showing both DNA methylation and expression changes showed functional enrichment for polyploidy and environmental-related responses. We also highlighted a considerable environmental effect that affected expression patterns in synthetics with similarities to DNA methylation. Additionally, we investigated genes of interest for future studies in newly formed polyploids.

Introduction

Expression changes can have short-term effects in newly formed plant polyploid species that can result in favorable long-term effects (1), ultimately leading to the establishment and success of a polyploid (2). Synthetic polyploids generated in the laboratory provide an excellent model to look at expression changes right after polyploidization and can be compared to their known progenitors (3).

Studies in *Arabidopsis*, wheat, cotton, *Nicotiana*, *Senecio* and *Brassica* synthetics showed a variable number of genes showing non-additive expression with respect to progenitor species (4–12). Non-additivity in these studies was defined as expression levels in the synthetic being significantly different from a mid-parental value (MPV), represented by the average expression between the two progenitors (Figure 1a) (4). This approach focused on homoeologous genes, i.e. genes that were separated by speciation and reunited under the same organism by polyploidization (13). With this approach, the majority of genes in newly formed allopolyploid were found to be additive, and the proportion of non-additive ones was highest in synthetic wheat (7%), followed by *Arabidopsis* (5.5%), *Senecio* (5%) and cotton showed a variable range (1-6.1%) together with *Brassica* (1.6-32%) (4,6,8,9,11,12). Taken together, these results emphasize different expression responses to polyploidy depending on the species. Additional factors were also found to contribute to expression pattern differences. In *Brassica*, genetically different lines showed different non-additive expression patterns with limited overlap (8). In wheat, non-additive genes showed stable expression over generations (11). Finally, in cotton, expression patterns showed differences across organs (5).

To offer a more comprehensive view on gene expression in allopolyploids, several improvements can be made to further break down expression patterns and link them to other important factors. First, instead of applying the MPV approach, by distinguishing the contributions to expression from each subgenome in an allopolyploid, a direct comparison to the expression in progenitors can be made (Figure 1b). With this new approach, all genes in the allopolyploid can be compared to their progenitors and no assumptions are made towards an expected expression pattern. The main requirement of such an approach is the ability to classify reads coming from an allopolyploid to either of the progenitor species. Recent methodological advances provided several tools to classify allopolyploid reads, such as Homeoroq (14), its more accurate successor EAGLE-RC (15,16) and others (17–21). As a next step, expression pattern changes can be compared to other changes happening in the genome at the same time. As seen in Chapter 2, DNA methylation could be a candidate control mechanism explaining some of the observed expression patterns.

Another critical gap in expression studies in synthetic allopolyploids, similarly to Chapter 2, is the absence of environmental stress. Several studies in established

allopolyploids acknowledged the effect of environment on expression patterns. In the allopolyploid *Coffea arabica*, differences in growth temperature lead to not only divergence in expression with respect to the progenitors, but also a strong difference in expression across conditions (22). In cotton, homoeologous gene expression of the alcohol dehydrogenase gene was altered under different environmental stresses, favoring a specific homeolog under specific conditions (23). These examples showed how environment can shape expression in established polyploids, but its role right after polyploid formation is unclear.

In this study, we will analyze expression changes with the same experimental design and study system from Chapter 2 in order to answer the following questions: 1) What is the magnitude of expression changes with respect to methylation changes? 2) Are expression patterns consistent with methylation patterns found in Chapter 2? 3) Is there an association between changes in methylation and expression for genes exhibiting both? 4) Can we find candidate genes that might be of interest in the early stages of polyploidy and their response to the environment?

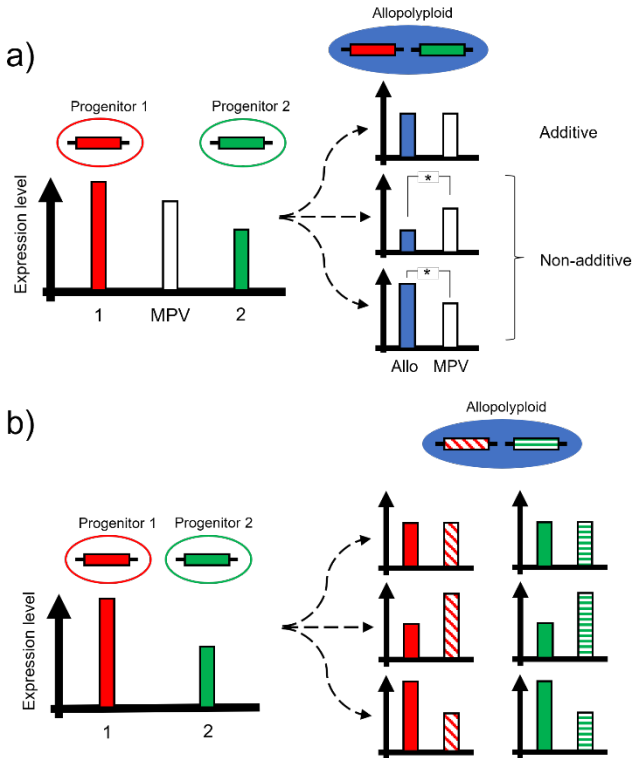


Figure 1: schematic view of the different approaches to look at expression changes in allopolyploids compared to their progenitors. One approach is the mid-parental value (MPV), where the average expression of a homeologous gene in progenitor species is used as the expected expression value in the polyploid (a, left side). Gene expression in the allopolyploid is considered additive when no significant difference is found between MPV and the expression in the polyploid, while in non-additive expression the polyploid has significantly lower or higher expression compared to the MPV (a, right side). A second approach is directly using the expression levels of the progenitors (b, left side). With the appropriate tools, the contribution in expression of each allele in the allopolyploid can be quantified and directly compared to the progenitors' expression (b, right side).

Materials and methods

Plant material and sequencing

See “Materials and Methods” section in Chapter 2, specifically subsection “Plant material and sequencing” for a detailed description of the procedure.

RNA-seq workflow

To analyze RNA-seq data, reads were first quality checked with *FastQC* v0.11.8 (24) and all quality reports were merged with *MultiQC* v1.8 (25). Next, reads were aligned with *STAR* v2.7.3a (26) to the published genome assemblies of either *A. halleri* (27) and *A. lyrata* (28). For synthetic and natural *A. kamchatica* data, reads were mapped to both genome assemblies and classified to progenitors’ sugenomes with *EAGLE-RC* v1.1.2 (15) using the parameters –ngi (no genotype information) and –paired (for paired end reads). As a final step, counts were obtained with *featureCounts* from *Subread* v2.0.1 (29). An overview of read numbers throughout the alignment and read classification steps are shown in Supplementary Table 1. The scripts for each step are available at https://github.com/supermaxiste/EarlyPolyploidDNAMethylation

Differential expression analyses

For differential expression analyses we used *edgeR* v3.32.1 (30). All samples were analyzed together, and reads were filtered with default parameters resulting in 21’518 genes passing filters for *A. halleri* and *halleri*-side (H-side) genes (out of 38’289) and 19’595 genes passing filters for *A. lyrata* and *lyrata*-side (L-side) genes (out of 34’967). Library sizes were scaled and the relationship between samples was investigated through the biological coefficient of variation (Figure 2). Differential expression (DE) was computed for all pairwise comparisons planned in Chapter 2 and between generations from the same sample. Genewise exact tests were used to compute DE and p-values were adjusted with false discovery rate to control for multiple testing. For all pairwise comparisons, a volcano plot was used to visualize differentially expressed genes (Supplementary Figure 1) and barplots were used to assess DEGs between synthetics and progenitors (Supplementary Figure 2), between natural species and all other samples (Figure 4) and across generations for all samples (Supplementary Figure 3). Given the low coverage regions found in Chapter 2, most probably caused by homoeologous exchanges (HE), we removed all DE genes falling in these regions for all pairwise comparisons involving synthetic polyploids. These genes appeared as DE because their counts were only low or zero on low coverage samples. Because of this, low coverage

regions could show high number of (false positive) DEGs, as seen in Supplementary Figures 4-5. Additionally, for Figure 4, low coverage genes were also excluded for NvsP1 comparisons. We also assessed the normalized read coverage over all samples and we found evidence of HE (Supplementary Figure 6). The full reproducible code can be found at https://github.com/supermaxiste/EarlyPolyploidDNAMethylation

Heatmaps with expression data

To visualize expression profiles for all samples and their relationship, we generated a heatmap for the expression of the most variable genes for each progenitor's side and condition (Figure 3). To do this, we imported raw counts for all samples and we filtered out genes falling in low coverage regions defined in Chapter 2 with BS-seq data (coverage <2X). Next, we applied a variance stabilizing transformation to have the expression variance approximately independent from the mean expression (31). As a final step, the expression variance of all genes across all samples was computed and the 1'000 genes with the highest variance were selected to generate a heatmap and cluster the expression profiles of all the samples. Code and files can be found at https://github.com/supermaxiste/EarlyPolyploidDNAMethylation.

BS-seq workflow and differential methylation analyses

Details on the data analysis process can be found in Chapter 2. In short, we used *ARPEGGIO v2.0.0* (32) to perform quality checks, alignments, read classification, methylation extraction and differential methylation analyses. For differential methylation, *dmrseq v1.6.0* was used, a method applying a two-step approach where first a smoothed pooled difference is used to select candidate regions showing a difference over a certain threshold. Second, the significance of candidate regions is assessed with generalized least squares models (33). Differentially methylated genes were then defined as genes with differentially methylated regions overlapping their body (from the transcriptional start site until the transcriptional end site).

Combining transcription and methylation data

To compare methylation and expression changes we first computed raw methylation changes for all DMRs found between synthetic *A. kamchatica* and progenitors (PvsS1 and PvsS4). To do so we created a command line R script requiring as arguments 1) the RData file resulting from a given comparison from ARPEGGIO, 2) the output name, 3) the set of Bismark coverage files from a given progenitor and 4) the coverage files for a given synthetic. This script was

executed for all conditions, comparisons, progenitors' sides and contexts. With the raw methylation values, we assigned significant gene expression changes (log fold-change values) to genes showing differential methylation (if significant expression changes occurred). Next, we plotted the association between raw methylation change and logFC for all contexts, comparisons and subgenomes (Supplementary Figure 7-8). For each plot we computed a linear regression to assess the relationship between methylation and expression and also tested for correlation between the two with Pearson's product moment correlation coefficient (cor.test function with Rv4.0.2 (34)).

To visualize the overlap between DEGs and DMGs, we plotted Venn diagrams for PvsS1 and PvsS4 comparisons for both progenitors' sides and conditions (Figure 5). For these diagrams, differentially methylated genes were defined as genes showing differential methylation in at least one context. Data wrangling was performed *with tidyverse* v1.3.0 (35) and *VennDiagram* v1.7.0 was used to plot diagrams.

The full reproducible code for the analyses discussed here is provided in https://github.com/supermaxiste/EarlyPolyploidDNAMethylation.

Gene overrepresentation analyses

Lists of both differentially expressed and methylated genes (DEMGs) were used for gene overrepresentation tests with PANTHER Overrepresentation Test (Released 20210224) (36) with *Arabidopsis thaliana* gene IDs and gene database. Differentially methylated genes were defined as genes showing differential methylation in at least one context. To obtain *A. thaliana* gene IDs from DEMGs with *A. halleri* and *A. lyrata* gene IDs, we checked for ortholog genes based on BLAST reciprocal best hit between genes of our reference species and *A. thaliana*. The Overrepresentation Test (OT) was performed for DEMGs for each condition, progenitors' side and generation, for a total of 8 tests (Figure 5). To complement these analyses we also performed OTs on genes showing only differential methylation (Figure 5).

The results of the OTs are provided in https://github.com/supermaxiste/EarlyPolyploidDNAMethylation.

Exploring methylation and expression state of genes of interest

To find genes of interest, we looked at genes showing consistent methylation and expression changes for both PvsS1 and PvsS4 in both conditions. This procedure was repeated for each progenitor's side. On the H-side a total of 80 genes were found and 36 (45%) had an orthologous *A. thaliana* gene (reciprocal best hit). On the L-side a total of 123

genes were found and 51 (42%) had an orthologous *A. thaliana* gene. Based on these lists, we selected genes of interest, plotted their corresponding DMR (Supplementary Figures 9-15) and reported both methylation and expression changes (Table 1, Supplementary Table 2).

The script for these analyses, a summary of all genes with their methylation change for all contexts, comparisons and conditions, together with their expression change is available at https://github.com/supermaxiste/EarlyPolyploidDNAMethylation.

Results

Section 1 – Transcriptomic profiles and trajectory of differential expression patterns in different conditions and across generations

BCV and expression heatmaps

Exploratory analyses with BCV on transcriptomic data revealed a consistent relationship found with BS-seq data, where samples from each natural line form a separate cluster and all other samples clustered together (Figure 2). This relationship was consistent across both conditions. Expression patterns between progenitors and synthetic samples were not as close as with BS-seq data, instead synthetics appeared closer to both natural species, without showing more proximity to a specific natural line. This suggested a more pronounced change in expression between synthetics and progenitors compared to methylation changes.

Heatmaps confirmed the clustering observed with BCV and emphasized the distinctive expression patterns within natural species and the variability in synthetics (Figure 3). Across conditions, each natural line displayed a unique set of genes either lowly or highly expressed compared to all other samples. Also, natural lines showed distinctive patterns for each condition that were consistent across replicates and over generations. Synthetics on the other hand showed the highest variability between replicates from the first generation (G1). This variability seemed to increase in the fourth generation (G4) with replicates either all clustering separately from G1 synthetics in cold conditions or only one replicate clustering with G1 synthetics in hot conditions (Figure 3).

Although coverage varied across samples and replicates, particularly for regions where homoeologous exchanges (HE) happened (Supplementary Figure 6), relationships between samples were unaffected. For BCV analyses, no filtering was applied for low coverage regions genes, while heatmaps had genes falling within low coverage regions removed from all samples. The consistency in results between the two approaches suggested that the major drivers of variation in gene expression did not result from HE.

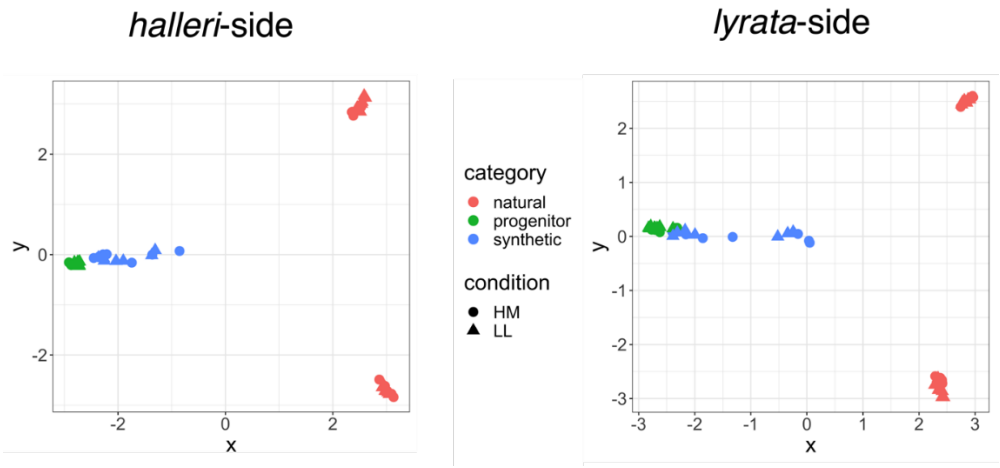


Figure 2: plot of the Biological Coefficient of Variation (BCV) from RNA-seq data for all samples. Samples are colored based on their origin: progenitors (diploid), synthetic or natural polyploids. Shapes are based on the environmental conditions, either cold (HM) or hot (LL). The plot on the left shows the halleri-side and right shows lyrata-side.

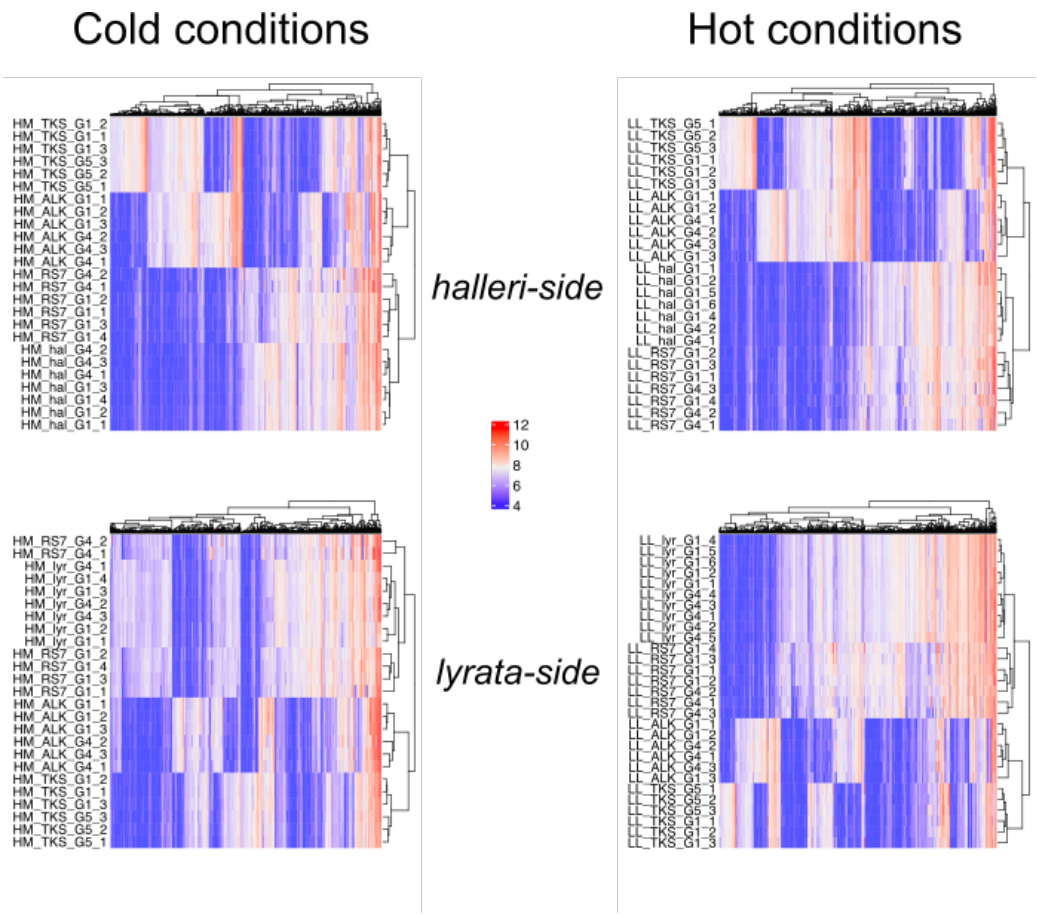


Figure 3: heatmap of the expression of the most variable genes in all of our analyzed samples in cold and hot conditions. Each heatmap shows the adjusted expression level (high = red, medium = white, low = blue) for the top 1000 genes (columns) with the highest variance in expression across all samples. Each row in the heatmap shows the expression profile of each analyzed sample and replicates (hal = A. halleri, lyr = A. lyrata, RS7 = synthetic A. kamchatica, ALK = Alaska line of natural A. kamchatica, TKS = Takashima line of natural A. kamchatica, G1 = first generation, G4 = fourth generation, numbers from 1 to 4 represent replicate number). A heatmap was generated for each progenitor's side, halleri-side (top) and lyrata-side (bottom), and each condition, cold (left) and hot (right).

Trajectory of differentially expression patterns

Differential expression analyses between synthetics and diploid progenitors (PvsS1 and PvsS4) in both conditions showed a large number of significant genes (up to almost half of all detected genes), with different trends for each condition (Supplementary Figure 1). The number of DEGs from PvsS1 to PvsS4 showed an increasing trend for cold conditions and similar numbers in hot conditions. This trend was the same in both subgenomes. The helleri-side H-side showed a consistently higher number of DEGs and this effect was probably driven by a higher proportion of low coverage regions on the L-side (see Chapter 2) that were masked to prevent false positive DEGs (see Materials and Methods and Supplementary Figure 6). When looking at the number of up- and down-regulated genes, in both conditions G4 synthetics showed a higher proportion of upregulated genes on the H-side and a higher proportion of downregulated genes on the L-side (Supplementary Figure 2). This is in contrast with proportions in G1 synthetics, except for H-side in cold conditions, where numbers were similar.

We repeated differential expression analyses with natural species as reference against progenitors and synthetics (NvsP1, NvsS1 and NvsS4) and observed condition-specific trends (Figure 4). The number of DEGs showed a consistently decreasing trend only in hot conditions for both natural lines while cold conditions didn't show a clear trend. This pattern was similar across subgenomes. Proportions of up- and down-regulated genes were similar across conditions and natural lines on the H-side. On the other hand, the L-side exhibited a higher proportion of up-regulated genes with respect to Alaska lines and vice-versa for Takashima lines, independent from the condition.

When assessing expression changes over generations for all samples, synthetics showed by far the largest number of changes (Supplementary Figure 3). Compared to all other samples, synthetics showed between 2000 and 6000 DEGs, with more DEGs in cold conditions. Progenitors and natural Alaska line had the second highest amount of changes, mostly happening in cold conditions, while natural Takashima line showed the least amount of changes in both conditions.

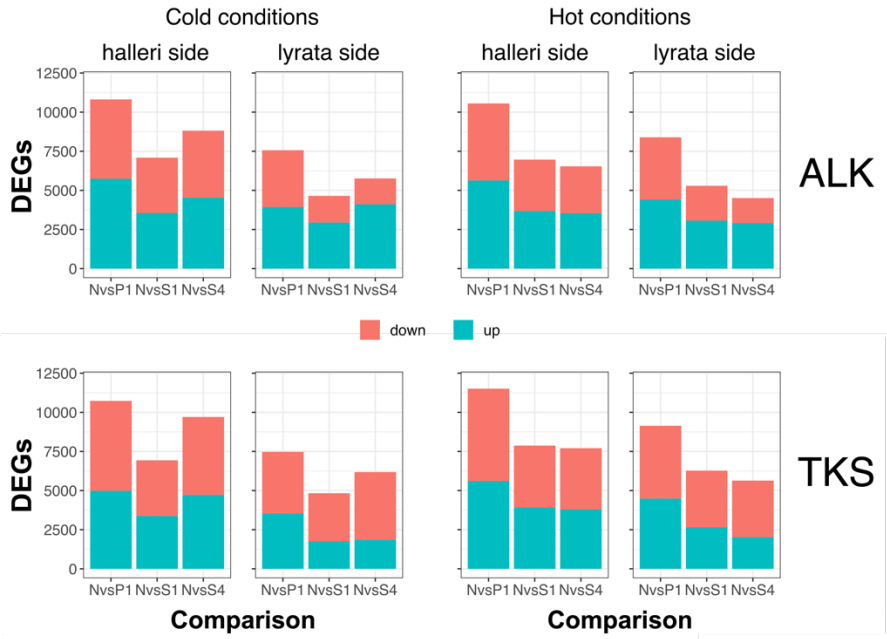


Figure 4: differentially expressed genes between natural A. kamchatica lines (reference) and progenitor species (NvsP1), synthetics generation one (NvsS1) and four (NvsS4). Barplots on the left side are for cold conditions, while right side ones are for hot conditions. Top barplots have Alaska as reference natural line and bottom ones have Takashima as reference. For both conditions, results for two progenitors' sides are shown with the proportion of genes being significantly upregulated (up) and downregulated (down).

Section 2 – Association between differential expression and DNA methylation in genes

We then analysed the relationship between differential methylation and expression for PvsS1 and PvsS4, and found a positive correlation between the two, particularly for CG and CHG context (Supplementary Figures 7-8). Although PvsS1 included fewer genes showing both differential methylation and expression, the relationships between raw methylation change and expression change were analogous to PvsS4. Specifically, for CG and CHG context, there was a general positive significant correlation and linear relationship, suggesting that an increase in methylation corresponded to an increase in expression and vice-versa. In the case of CHH methylation, this relationship was not as pronounced, with several conditions and subgenomes showing non-significant correlation.

When looking at overlaps between genes showing significant changes for both expression and methylation, we found a high number of conserved patterns between PvsS1 and PvsS4, together with smaller proportions of differentially methylated and expressed genes then expected (Figure 5, Supplementary Figure 9). For all comparisons across all conditions, the number of DEGs was consistently higher than the number of DMGs. To assess whether the DEGs and DMGs were conserved between PvsS1 and PvsS4, we computed the number of DEGs and DMGs from PvsS1 still present in PvsS4. Except for H-side under hot conditions,

60-70% of all DEGs in PvsS1 were found again in PvsS4. These proportions were higher for DMGs, with the same previous exception, showing 70-80% of all DMGs from PvsS1 still present in PvsS4. Taken together, these results suggest a high proportion of conserved DEGs and DMGs between first and fourth generation synthetics with respect to the progenitors. We also assessed the overlaps between DEGs and DMGs across all conditions and found between 4% and 10% of genes showing changes in both expression and methylation. We tested for an association between differential methylation and expression through a chi-square test (Supplementary Figure 9). All tests were significant, supporting the hypothesis of a significant association with overlaps significantly smaller than expected.

Section 3 – Functional analyses of differentially expressed and methylated genes

We compared the biological function of genes showing both expression and methylation changes to genes showing differential methylation only and we detected overrepresentation mostly in the former (Figure 5). For overlaps, functional enrichment was found in all comparisons except for the overlap in PvsS1 (cold conditions) on the H-side, which was also the overlap with the lowest number of genes. Genes showing differential methylation did not show any enrichment except for PvsS1 (hot conditions) H-side, where only one GO term was found. These results suggested a functional implication of DNA methylation changes in gene bodies, but with an unclear causal relationship between these changes and expression changes.

When examining the GO terms in the overlaps, we found the same broad metabolic and cellular GO terms found in methylation data (see Chapter 2), together with GO terms in PvsS4 potentially related to polyploid-specific and environment-specific responses. For overlaps in PvsS1 and PvsS4 (both conditions), we found overrepresentation for broad functional categories related to cellular (GO:0009987) and metabolic processes (GO:0071704, GO:0008152), the same ones obtained with methylation data. Additional terms were found only for PvsS4 overlaps. In cold conditions, we detected environment-related terms such as response to stress (GO:0006950) and response to abiotic stimulus (GO:0009628), and polyploid-related terms such as post-embryonic (GO:00099791) and anatomical structure development (GO:0048856). In hot conditions polyploid-related terms were found such as protein ubiquitination (GO:0016567) and both a polyploid- and environment-related term, cell growth (GO:0016049). Finally, in both conditions RNA related functions were detected such as mRNA export from nucleus (GO:0006406) and RNA splicing (GO:0008380), implying links to expression control in polyploids.

Section 4 – Exploring genes of interest related to early stages of polyploidy

To explore genes of interest involved in the early stages of polyploidy, we looked for genes showing a consistent change in methylation and expression in PvsS1 and PvsS4 on one progenitor's side in both conditions (Table 1 and Supplementary Table 1). All of the genes obtained showed the same change in expression and methylation across conditions, supporting a polyploid-specific response rather than an environment-specific response. The only gene found changing expression and methylation on both progenitors' sides was a Glucose-6-phosphate isomerase (AT5G42740). Its DNA methylation and expression levels were significantly downregulated for both PvsS1 and PvsS4. Specific to the H-side, a polyubiquitin gene (AT4G02890) was found together with the Flowering locus C gene (AT5G10140). Specific to the L-side, four genes were found. Two genes related to epigenetic control, Factor of DNA methylation 1 (AT1G15910) and a Histone Deacetylase gene (AT3G18520). The two other genes were the development-related gene MADS-box transcription factor ANR1 (AT2G14210) and the stress related gene Heavy Metal ATPase 3 (AT4G30120).

The relationship between expression and methylation patterns of these genes of interest was variable both in terms of directions and level of changes, and methylation contexts involved (Supplementary Figures 10-16). For three out of seven genes, we found the same positive relationship between methylation and expression changes found in Section 2, while for the other four genes, an increase in methylation resulted in a decrease in expression and vice-versa. Regarding methylation contexts, genes showed different numbers of contexts involved, with occasionally two differentially methylated regions overlapping with a gene in the same context. In four genes, only one methylation context, specifically CG, was involved. For the other three genes, two or even all three contexts were involved with CG being always part of those.

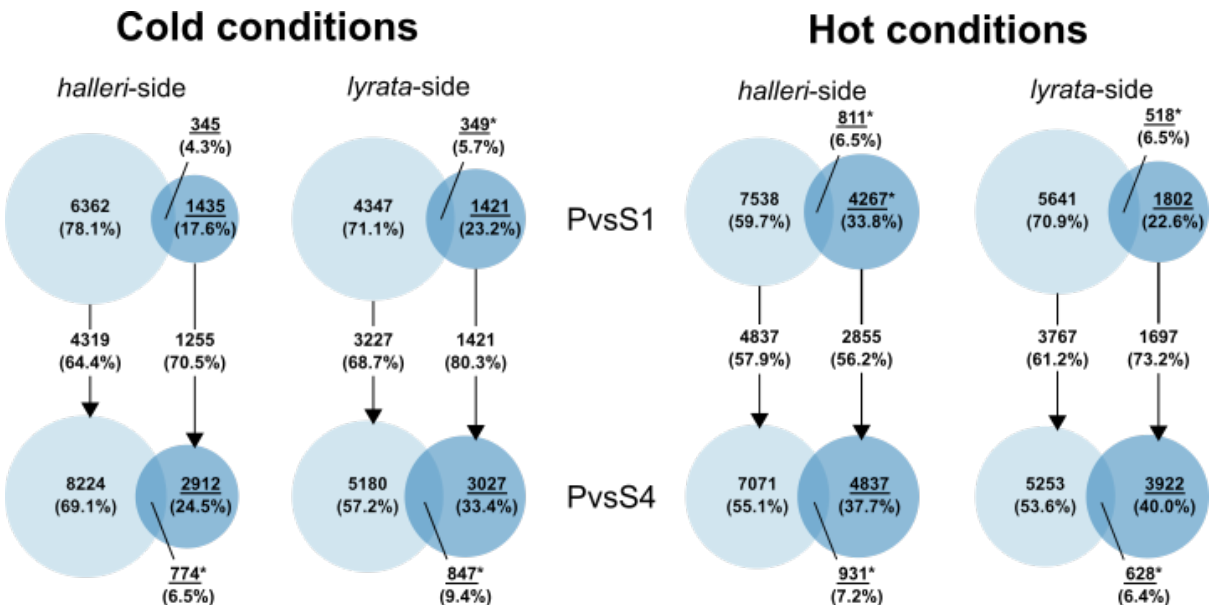


Figure 5: overlap between differentially expressed (DEG) and differentially methylated genes (DMG) across comparisons between synthetics and first generation progenitors. Each row represents a specific comparison, either first generation (PvsS1) or fourth generation (PvsS4) synthetics against progenitors. Overlaps for cold conditions on the left side and hot conditions on the right side. The size of all Venn diagrams is proportional to the value shown. Together with gene number, a percentage is shown to specify the proportion of the value with respect to the whole set including DMG and DEG. Arrows going from the first to the second row show the proportions of genes present in the comparison between first generations that are still found in the comparison with fourth generation synthetics. Underlined values represent gene sets tested for overrepresentation of biological functions and significant results are represented by an asterisk.

Table 1: candidate genes showing consistent methylation and expression changes across conditions and comparisons. The first column shows the gene names, the second shows the A. thaliana gene ID, the third column specifies the progenitors side where the gene was found, the fourth and fifth columns show the direction of methylation and expression changes (increase means increase in the synthetic with respect to the progenitors and likewise for decrease).

Gene name	Gene ID	Consistent methylation and expression changes on…		Methylation change	Expression change
		H-side	L-side		
Glucose-6-phosphate isomerase	AT5G42740	Yes	Yes	Decrease	Decrease
Polyubiquitin (UBQ14)	AT4G02890	Yes	No	Decrease	Increase
Flowering Locus C (FLC)	AT5G10140	Yes	No	Increase	Decrease
Arabidopsis Nitrate Regulated 1 (ANR1)	AT2G14210	No	Yes	Increase	Decrease
Factor of DNA methylation 1 (FDM1)	AT1G15910	No	Yes	Increase	Increase
Heavy Metal ATPase 3 (HMA3)	AT4G30120	No	Yes	Increase	Increase
Histone Deacetylase 15 (HDA15)	AT3G18520	No	Yes	Increase	Decrease

Discussion

Partial correspondence between methylation and expression changes

Genome-wide expression and methylation profiles showed consistent relationships between progenitors, synthetics and natural species, but comparisons between synthetics and progenitors or natural species were only partially consistent, implying different changes in methylation and expression. Genes with the highest variation in expression across species showed a stable pattern over generations for natural lines and progenitor species, but not synthetics. This was in concordance with methylation patterns, specifically global methylation levels and differentially methylated regions, both supporting a more stable methylation pattern in natural lines and progenitors compared to synthetics (see Chapter 2). On the other hand, although methylation patterns in synthetics showed a clear diverging course from progenitors (see Chapter 2), expression showed such pattern for only one environmental condition, namely cold. Similarly, synthetics showed increasingly similar expression pattern with natural lines only for hot, stressful conditions.

Expression changes were also more extensive than methylation changes when comparing synthetics to progenitors. The proportion of DEGs was remarkably high on both sub-genomes, with a third up to more than half of all detected genes being significant. On the other hand methylation changes, even though many, covered only 1-5% of the whole genome. Although there are mechanisms for smaller amounts of methylation to affect the expression of multiple genes, such as targeting regulatory regions (37), this difference emphasized the potential involvement of other (epi)genetic systems in the regulation of expression. Potential candidates include non-coding RNAs, which are increasingly characterized and recognized (38), histone modifications and chromatin structure (39). Another epigenetic trait that was not investigated in this study was transposable elements (TEs), often associated with larger, repeat rich genomes such as wheat (40), but recently found to be playing an important role in subgenome dominance in the early stages of allotetraploid *Mimulus peregrinus* (41).

Expression patterns between synthetics from different environmental conditions showed increasing divergence, which was consistent with DNA methylation patterns, but with differences emerging already from the first generation. In particular, no differences in DNA methylation were found between first generation synthetics grown in hot and cold conditions, but several were found when comparing fourth generation synthetics. In the case of expression, >1’000 genes were differentially expressed on each progenitor side from the first generation and increased to >7’000 on each side in the fourth generation. These results suggest that expression patterns in synthetics are more dynamic compared to methylation patterns, leading to diverging patterns as early as in the first generation. There could be

several drivers behind this dynamic behavior in addition to environmental stress. For example, in the *Spartina* study system transcriptomic changes after polyploidization were attributed to hybridization rather than polyploidization *per se* (42), while in *Arabidopsis suecica* HE events contributed to variation in gene expression (43). While it is unclear how polyploidization and hybridization have contributed to expression changes in synthetic *A. kamchatica*, HE events did not affect relationships across samples when excluded, suggesting a limited effect.

Despite methylation and expression having very stable diverging patterns over time and expression showing broad changes, we found less genes than statistically expected exhibiting alterations in both. Our analyses focused on DNA methylation changes within gene bodies, which represent a subset of all DNA methylation changes that could affect expression. Such changes include TEs and repeat silencing (see above), activation or repression of transcription through regulatory regions (see above) and regulation of RNA processing (44). This means that we might be underestimating the number of genes showing both expression and methylation changes. By including the epigenetic mechanisms outlined above, the overlap in terms of genes between expression and methylation changes could be larger.

The effect of environmental stress on expression in the early stages of polyploidy

Expression changes in synthetics with respect to progenitors and natural species showed environment-dependent patterns that highlighted distinct effects of stress on expression profiles of newly formed polyploids. These results were consistent with studies in natural allopolyploid *Coffea arabica*, where expression pattern divergence with respect to diploid progenitors was greatly influenced by environmental conditions (22), but there are several differences compared to this study. First, our focus was on a synthetic allopolyploid and the effect of environment on its expression pattern was the largest over generations compared to both progenitors and natural lines. Second, *C. arabica* grown in mild conditions showed similar proportions of DEG on both progenitors' sides, but in warmer conditions a strong expression dominance pattern appeared with one progenitor side being responsible for most expression changes in the polyploid. In synthetic *A. kamchatica,* no clear expression dominance was found in either condition. Third, our analyses were complicated by HE events, resulting in an uneven number of genes across subgenomes and requiring masking to get HE-independent results, but also restricting our view on the whole genome.

Additionally, results suggested that environment influenced expression divergence between a newly formed polyploid and its progenitors, but further analyses would be needed to assess if differential expression between homoeologous genes might be driving this pattern in *A. kamchatica*. A study on *A. thaliana* highlighted how environmental stress led to

expression divergence in duplicated genes in the short-term, possibly facilitating sub- or neo-functionalization of genes on the long-term and lead to adaptive mechanisms to deal with environmental stress (45). In *A. kamchatica*, homoeologs' expression in the short-term could affect their expression in the longer-term, potentially influencing the fate of these genes. Expression analyses on later generations of synthetics could be useful to explore the stability of expression patterns over time and could support an involvement over longer time scales.

The functional importance of differentially expressed and methylated genes in newly formed polyploids

In synthetics, genes displaying both methylation and expression change showed functional enrichment revealing links with polyploidy-related and environmental-related functions, while no enrichment was detected for genes only differentially methylated. For polyploidy, we detected developmental functions, possibly linked to developmental regulation and also observed in other allopolyploid species (4,46). Similarly, RNA-related terms in both conditions suggested a response to control the broad expression changes observed in the first generation of synthetics, a very common trait found in other polyploids (5,47–49). This could also be a factor to further generate differentiation in methylation and expression. Related to environmental stress, we found stress and abiotic response in cold conditions and in hot conditions detected a function related to cell growth, potentially linked to cell size increase, a common phenotype in polyploids with connections to stress response (50,51). It remains unclear whether differential methylation in genes with no expression change might still have a functional role such as in maize, where methylation in intron-exon junctions affected splicing efficiency and led to reduced alternative splicing (52). This effect was only found in CHG methylation context, but not in CHH context. Alternatively, these methylation changes might also be a temporary state resulting from a noisy hybrid methylation machinery.

Candidate genes for future studies in novel polyploids

We investigated several candidate genes exhibiting consistent transcription and methylation changes that could be of interest in the early stages of polyploidy, and found only one house-keeping gene on both progenitors' sides (Table 1). The gene was a Glucose-6-Isomerase (G6I), coding for an enzyme interconverting D-glucose-6-phosphate and D-fructose-6-phosphate, a reaction leading to either glycolysis, sugar breakdown into energy for the plant, or gluconeogenesis, sugar formation (53,54). Since perturbations in parts of the carbohydrate

metabolism can lead to negative effects on growth and development (55,56), G6I downregulation could be essential for polyploid survival in the early stages.

On the H-side, the Flowering Locus C (FLC) gene, usually expressed in leaves, root tips and shoot apex (57), is well known for its repression of flowering (58) and could be linked to different flowering time regulation and phenotype in *A. kamchatica*. In the case of synthetic *A. suecica*, FLC expression was found to be higher on one progenitor side compared to the other, leading to a late flowering phenotype (4). A complementary genome-wide study in synthetic *A. suecica* highlighted the importance of copy number variation in FLC, correlating strongly with flowering time variation (49). For *A. kamchatica* copy number might also play a role since in *A. lyrata*'s FLC underwent a tandem duplication, leading to two copies of this gene (FLC1 and FLC2) (59). Additionally, flowering of synthetic *A. kamchatica* appears less synchronized than its natural counterpart (personal observation), suggesting variability in FLC methylation among individuals. To better understand the link between downregulation in methylation and expression of FLC in *A. kamchatica* and different flowering time, additional data from the field on flowering time would be needed, together with comparisons between H- and L-side FLC expression and methylation states. The last gene found on the H-side was a polyubiquitin gene related to the regulation of protein degradation (60) and potentially linked to protein dosage balance in polyploids (61,62).

On the L-side, Factor of DNA Methylation 1 (FDM1) is a key component of the RNA-directed DNA methylation pathway (RdDM), a major epigenetic pathway related to small interfering RNAs contributing to control of transposons and stress response (63). Decrease in methylation and increase in expression in *A. kamchatica* might suggest an increased activity of FDM1, leading to increased methylation of transposable elements after polyploidization and environmental stress. Another gene found on the L-side was Histone Deacetylase 15 (HD15), responsible for repressing plant response to high temperatures in *A. thaliana* (64). Since this gene was consistently downregulated in both expression and methylation on the L-side, it might indicate a possible activation of thermal responsive genes in this particular subgenome, supporting *A. kamchatica*'s environmental tolerance. The role of the last two genes found in the L-side could be related to metabolism control, with Heavy Metal ATPase 3 (HMA3) associated to heavy metal homeostasis and accumulation in tissues, particularly leaves (65,66) and the MADS-box transcription factor ANR1, associated to nitrogen-associated metabolism in roots (67) and seed germination (68).

Conclusions

Expression changes in newly formed polyploids are rapid and extensive. We showed how global patterns of expression and methylation in novel polyploids were coherent, but changes in methylation were not always accompanied by changes in expression. A positive correlation was found for genes exhibiting changes in both. These genes were also associated to polyploidy-related and environment-related responses, supporting DNA methylation as a decisive mechanism in shaping polyploid expression in their early stages.

Our results also highlighted the strong influence of environmental stress on the expression patterns in synthetic polyploids, similarly to methylation patterns. We also investigated a list of candidate genes that could be interesting for future studies in synthetic polyploids.

Our study aimed to provide a first expanded view on the early stages of polyploidy combining transcriptomics, epigenomics and stress. Similar efforts in other species, considering additional epigenetic mechanisms and environmental stresses will help shed light on the complex, yet prevalent successful formation and establishment of polyploids.

References

1. Adams KL. Evolution of Duplicate Gene Expression in Polyploid and Hybrid Plants. J Hered [Internet]. 2007 Mar 1;98(2):136–41. Available from: http://academic.oup.com/jhered/article/98/2/136/2187871/Evolution-of-Duplicate-Gene-Expression-in
2. Otto SP. The Evolutionary Consequences of Polyploidy. Cell [Internet]. 2007 Nov;131(3):452–62. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0092867407013402
3. Chen ZJ, Ni Z. Mechanisms of genomic rearrangements and gene expression changes in plant polyploids. BioEssays [Internet]. 2006 Mar;28(3):240–52. Available from: https://onlinelibrary.wiley.com/doi/10.1002/bies.20374
4. Wang J, Tian L, Lee H-S, Wei NE, Jiang H, Watson B, et al. Genomewide Nonadditive Gene Regulation in Arabidopsis Allotetraploids. Genetics [Internet]. 2006 Jan 1;172(1):507–17. Available from: https://academic.oup.com/genetics/article/172/1/507/6065206
5. Adams KL, Percifield R, Wendel JF. Organ-Specific Silencing of Duplicated Genes in a Newly Synthesized Cotton Allotetraploid. Genetics [Internet]. 2004 Dec 1;168(4):2217–26. Available from: https://academic.oup.com/genetics/article/168/4/2217/6059384
6. Yoo M-J, Szadkowski E, Wendel JF. Homoeolog expression bias and expression level dominance in allopolyploid cotton. Heredity (Edinb) [Internet]. 2013 Feb 21;110(2):171–80. Available from: http://www.nature.com/articles/hdy201294
7. Anssour S, Baldwin IT. Variation in Antiherbivore Defense Responses in Synthetic Nicotiana Allopolyploids Correlates with Changes in Uniparental Patterns of Gene Expression. Plant Physiol [Internet]. 2010 Aug 3;153(4):1907–18. Available from: https://academic.oup.com/plphys/article/153/4/1907/6111405
8. Gaeta RT, Yoo S-Y, Pires JC, Doerge RW, Chen ZJ, Osborn TC. Analysis of Gene Expression in Resynthesized Brassica napus Allopolyploids Using Arabidopsis 70mer Oligo Microarrays. Hazen SP, editor. PLoS One [Internet]. 2009 Mar 10;4(3):e4760. Available from: https://dx.plos.org/10.1371/journal.pone.0004760
9. Xu Y, Zhong L, Wu X, Fang X, Wang J. Rapid alterations of gene expression and cytosine methylation in newly synthesized Brassica napus allopolyploids. Planta. 2009;229(3):471–83.
10. Kashkush K, Feldman M, Levy AA. Gene Loss, Silencing and Activation in a Newly Synthesized Wheat Allotetraploid. Genetics [Internet]. 2002 Apr 1;160(4):1651–9. Available from: https://academic.oup.com/genetics/article/160/4/1651/6049775
11. Chagué V, Just J, Mestiri I, Balzergue S, Tanguy A-M, Huneau C, et al. Genome-wide gene expression changes in genetically stable synthetic and natural wheat allohexaploids. New Phytol [Internet]. 2010 Sep;187(4):1181–94. Available from: http://doi.wiley.com/10.1111/j.1469-8137.2010.03339.x
12. HEGARTY MJ, JONES JM, WILSON ID, BARKER GL, COGHILL JA, SANCHEZ-BARACALDO P, et al. Development of anonymous cDNA microarrays to study changes to the Senecio floral transcriptome during hybrid speciation. Mol Ecol [Internet]. 2005 Jul;14(8):2493–510. Available from: https://onlinelibrary.wiley.com/doi/10.1111/j.1365-294x.2005.02608.x
13. Glover NM, Redestig H, Dessimoz C. Homoeologs: What Are They and How Do We Infer Them? Trends Plant Sci [Internet]. 2016 Jul;21(7):609–21. Available from: https://linkinghub.elsevier.com/retrieve/pii/S1360138516000595
14. Akama S, Shimizu-Inatsugi R, Shimizu KK, Sese J. Genome-wide quantification of homeolog expression ratio revealed nonstochastic gene regulation in synthetic allopolyploid Arabidopsis. Nucleic Acids Res [Internet]. 2014 Apr 1;42(6):e46–e46. Available from: https://academic.oup.com/nar/article/42/6/e46/2437554
15. Kuo T, Frith MC, Sese J, Horton P. EAGLE: Explicit Alternative Genome Likelihood Evaluator. BMC Med Genomics [Internet]. 2018 Apr 20;11(S2):28. Available from: https://bmcmedgenomics.biomedcentral.com/articles/10.1186/s12920-018-0342-1
16. Hu G, Grover CE, Arick MA, Liu M, Peterson DG, Wendel JF. Homoeologous gene expression and co-expression network analyses and evolutionary inference in allopolyploids. Brief Bioinform [Internet]. 2020 Mar 27; Available from: https://academic.oup.com/bib/advance-article/doi/10.1093/bib/bbaa035/5811916
17. Mithani A, Belfield EJ, Brown C, Jiang C, Leach LJ, Harberd NP. HANDS: a tool for genome-wide discovery of subgenome- specific base-identity in polyploids. BMC Genomics [Internet]. 2013 Dec 24;14(1):653. Available from: https://bmcgenomics.biomedcentral.com/articles/10.1186/1471-2164-14-653

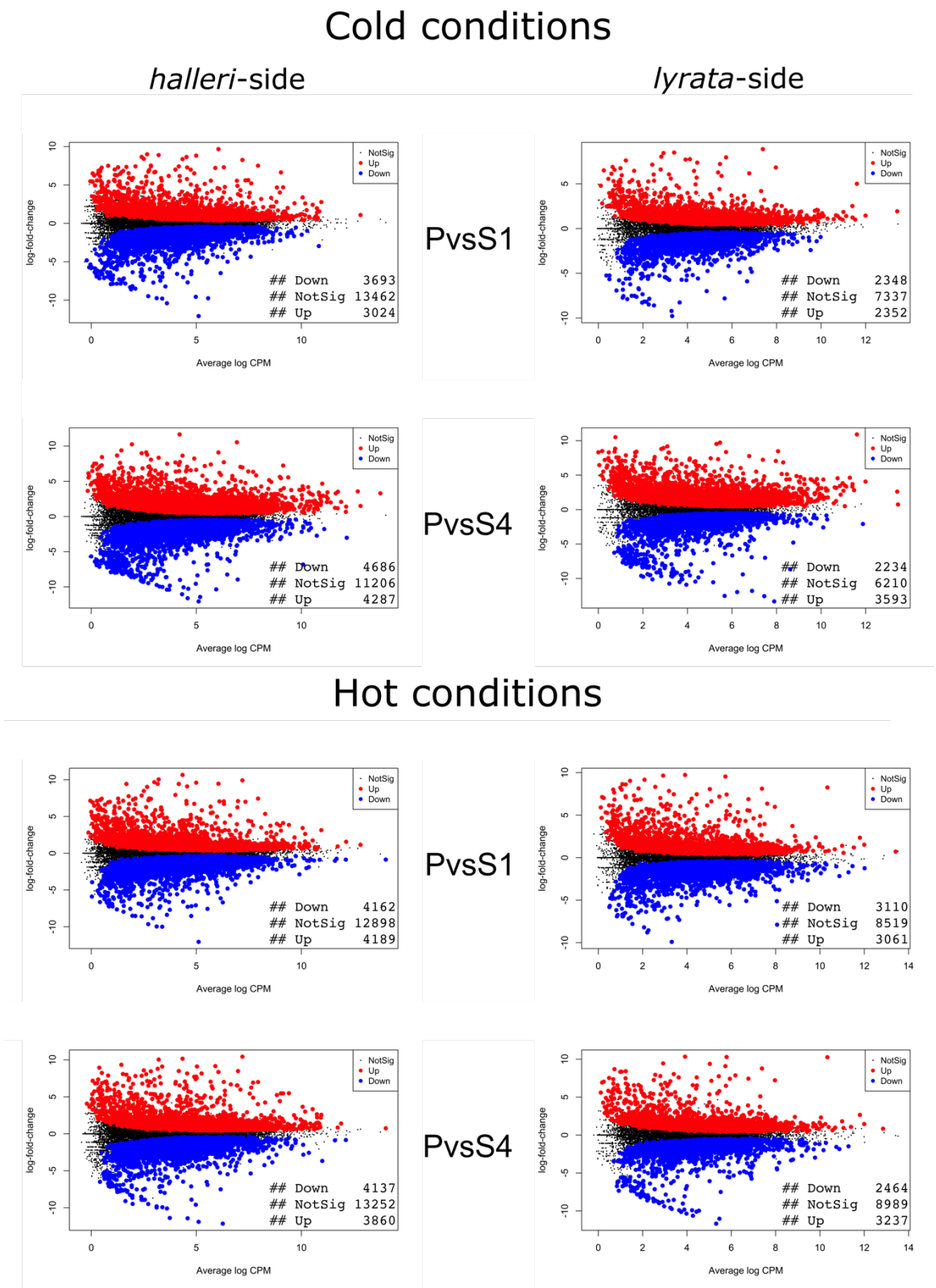
18. Khan A, Belfield EJ, Harberd NP, Mithani A. HANDS2: accurate assignment of homoeallelic base-identity in allopolyploids despite missing data. Sci Rep [Internet]. 2016 Jul 5;6(1):29234. Available from: http://www.nature.com/articles/srep29234
19. Page JT, Gingle AR, Udall JA. PolyCat: A Resource for Genome Categorization of Sequencing Reads From Allopolyploid Organisms. G3: Genes|Genomes|Genetics [Internet]. 2013 Mar;3(3):517–25. Available from: http://g3journal.org/lookup/doi/10.1534/g3.112.005298
20. Peralta M, Combes M-C, Cenci A, Lashermes P, Dereeper A. SNiPloid: A Utility to Exploit High-Throughput SNP Data Derived from RNA-Seq in Allopolyploid Species. Int J Plant Genomics [Internet]. 2013 Sep 12;2013:1–6. Available from: https://www.hindawi.com/journals/ijpg/2013/890123/
21. Page JT, Udall JA. Methods for mapping and categorization of DNA sequence reads from allopolyploid organisms. BMC Genet [Internet]. 2015;16(Suppl 2):S4. Available from: http://bmcgenet.biomedcentral.com/articles/10.1186/1471-2156-16-S2-S4
22. Bardil A, de Almeida JD, Combes MC, Lashermes P, Bertrand B. Genomic expression dominance in the natural allopolyploid Coffea arabica is massively affected by growth temperature. New Phytol [Internet]. 2011 Nov;192(3):760–74. Available from: https://onlinelibrary.wiley.com/doi/10.1111/j.1469-8137.2011.03833.x
23. Liu Z, Adams KL. Expression Partitioning between Genes Duplicated by Polyploidy under Abiotic Stress and during Organ Development. Curr Biol [Internet]. 2007 Oct;17(19):1669–74. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0960982207018477
24. Andrews S. FastQC: a quality control tool for high throughput sequence data [Internet]. 2010. Available from: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/
25. Ewels P, Magnusson M, Lundin S, Käller M. MultiQC: summarize analysis results for multiple tools and samples in a single report. Bioinformatics [Internet]. 2016 Oct 1;32(19):3047–8. Available from: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btw354
26. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics [Internet]. 2013 Jan;29(1):15–21. Available from: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/bts635
27. Briskine R V., Paape T, Shimizu-Inatsugi R, Nishiyama T, Akama S, Sese J, et al. Genome assembly and annotation of Arabidopsis halleri, a model for heavy metal hyperaccumulation and evolutionary ecology. Mol Ecol Resour [Internet]. 2016 Sep;17(5):1025–36. Available from: http://doi.wiley.com/10.1111/1755-0998.12604
28. Paape T, Briskine R V., Halstead-Nussloch G, Lischer HEL, Shimizu-Inatsugi R, Hatakeyama M, et al. Patterns of polymorphism and selection in the subgenomes of the allopolyploid Arabidopsis kamchatica. Nat Commun [Internet]. 2018 Dec 25;9(1):3909. Available from: http://www.nature.com/articles/s41467-018-06108-1
29. Liao Y, Smyth GK, Shi W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics [Internet]. 2014 Apr 1;30(7):923–30. Available from: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btt656
30. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics [Internet]. 2010 Jan 1;26(1):139–40. Available from: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btp616
31. Anders S, Huber W. Differential expression analysis for sequence count data. Genome Biol [Internet]. 2010 Oct 27;11(10):R106. Available from: https://genomebiology.biomedcentral.com/articles/10.1186/gb-2010-11-10-r106
32. Milosavljevic S, Kuo T, Decarli S, Mohn L, Sese J, Shimizu KK, et al. ARPEGGIO: Automated Reproducible Polyploid EpiGenetic GuIdance workflOw. BMC Genomics [Internet]. 2021 Dec 1;22(1):547. Available from: https://bmcgenomics.biomedcentral.com/articles/10.1186/s12864-021-07845-2
33. Korthauer K, Chakraborty S, Benjamini Y, Irizarry RA. Detection and accurate false discovery rate control of differentially methylated regions from whole genome bisulfite sequencing. Biostatistics [Internet]. 2019 Jul 1;20(3):367–83. Available from: https://academic.oup.com/biostatistics/article/20/3/367/4899074
34. R Foundation for Statistical Computing. R: A language and environment for statistical computing [Internet]. Vienna, Austria; 2020. Available from: https://www.r-project.org/
35. Wickham H, Averick M, Bryan J, Chang W, McGowan L, François R, et al. Welcome to the Tidyverse. J Open Source Softw [Internet]. 2019 Nov 21;4(43):1686. Available from:

https://joss.theoj.org/papers/10.21105/joss.01686
36. Mi H, Ebert D, Muruganujan A, Mills C, Albou L-P, Mushayamaha T, et al. PANTHER version 16: a revised family classification, tree-based classification tool, enhancer regions and extensive API. Nucleic Acids Res [Internet]. 2021 Jan 8;49(D1):D394–403. Available from: https://academic.oup.com/nar/article/49/D1/D394/6027812
37. Niederhuth CE, Schmitz RJ. Putting DNA methylation in context: from genomes to gene expression in plants. Biochim Biophys Acta - Gene Regul Mech [Internet]. 2017;1860(1):149–56. Available from: http://dx.doi.org/10.1016/j.bbagrm.2016.08.009
38. Hou J, Lu D, Mason AS, Li B, Xiao M, An S, et al. Non-coding RNAs and transposable elements in plant genomes: emergence, regulatory mechanisms and roles in plant development and stress responses. Planta [Internet]. 2019 Jul 16;250(1):23–40. Available from: http://link.springer.com/10.1007/s00425-019-03166-7
39. ADAMS K, WENDEL J. Novel patterns of gene expression in polyploid plants. Trends Genet [Internet]. 2005 Oct;21(10):539–43. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0168952505002179
40. Bariah I, Keidar-Friedman D, Kashkush K. Where the Wild Things Are: Transposable Elements as Drivers of Structural and Functional Variations in the Wheat Genome. Front Plant Sci [Internet]. 2020 Sep 18;11. Available from: https://www.frontiersin.org/article/10.3389/fpls.2020.585515/full
41. Edger PP, Smith RD, McKain MR, Cooley AM, Vallejo-Marin M, Yuan Y-WY, et al. Subgenome Dominance in an Interspecific Hybrid, Synthetic Allopolyploid, and a 140-Year-Old Naturally Established Neo-Allopolyploid Monkeyflower. Plant Cell [Internet]. 2017 Sep;29(9):2150–67. Available from: http://www.plantcell.org/lookup/doi/10.1105/tpc.17.00010
42. Chelaifa H, Monnier A, Ainouche M. Transcriptomic changes following recent natural hybridization and allopolyploidy in the salt marsh species Spartina × townsendii and Spartina anglica (Poaceae). New Phytol [Internet]. 2010 Apr;186(1):161–74. Available from: https://onlinelibrary.wiley.com/doi/10.1111/j.1469-8137.2010.03179.x
43. Burns R, Mandáková T, Gunis J, Soto-Jiménez LM, Liu C, Lysak MA, et al. Gradual evolution of allopolyploidy in Arabidopsis suecica. Nat Ecol Evol [Internet]. 2021 Oct 19;5(10):1367–81. Available from: https://www.nature.com/articles/s41559-021-01525-w
44. Zhang H, Lang Z, Zhu J-K. Dynamics and function of DNA methylation in plants. Nat Rev Mol Cell Biol [Internet]. 2018 Aug 21;19(8):489–506. Available from: http://www.nature.com/articles/s41580-018-0016-z
45. Ha M, Li W-H, Chen ZJ. External factors accelerate expression divergence between duplicate genes. Trends Genet [Internet]. 2007 Apr;23(4):162–6. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0168952507000583
46. Song Q, Chen JZ. Epigenetic and developmental regulation in plant polyploids. Curr Opin Plant Biol [Internet]. 2015;24:101–9. Available from: http://dx.doi.org/10.1016/j.pbi.2015.02.007
47. Ramírez-González RH, Borrill P, Lang D, Harrington SA, Brinton J, Venturini L, et al. The transcriptional landscape of polyploid wheat. Science (80-) [Internet]. 2018 Aug 17;361(6403):eaar6089. Available from: https://www.sciencemag.org/lookup/doi/10.1126/science.aar6089
48. Wang J, Tian L, Madlung A, Lee H-S, Chen M, Lee JJ, et al. Stochastic and Epigenetic Changes of Gene Expression in Arabidopsis Polyploids. Genetics [Internet]. 2004 Aug;167(4):1961–73. Available from: http://www.genetics.org/lookup/doi/10.1534/genetics.104.027896
49. Jiang X, Song Q, Ye W, Chen ZJ. Concerted genomic and epigenomic changes accompany stabilization of Arabidopsis allopolyploids. Nat Ecol Evol [Internet]. 2021 Oct 19;5(10):1382–93. Available from: https://www.nature.com/articles/s41559-021-01523-y
50. Van de Peer Y, Ashman T-L, Soltis PS, Soltis DE. Polyploidy: an evolutionary and ecological force in stressful times. Plant Cell [Internet]. 2021 Mar 22;33(1):11–26. Available from: https://academic.oup.com/plcell/article/33/1/11/6015242
51. Schoenfelder KP, Fox DT. The expanding implications of polyploidy. J Cell Biol [Internet]. 2015 May 25;209(4):485–91. Available from: https://rupress.org/jcb/article/209/4/485/38097/The-expanding-implications-of-polyploidyThe
52. Regulski M, Lu Z, Kendall J, Donoghue MTA, Reinders J, Llaca V, et al. The maize methylome influences mRNA splice sites and reveals widespread paramutation-like switches guided by small RNA. Genome Res [Internet]. 2013 Oct;23(10):1651–62. Available from: http://genome.cshlp.org/lookup/doi/10.1101/gr.153510.112
53. Glucose - 6 - phosphate isomerase. Philos Trans R Soc London B, Biol Sci [Internet]. 1981

Jun 26;293(1063):145–57. Available from: https://royalsocietypublishing.org/doi/10.1098/rstb.1981.0068
54. Sung S-JS, Xu D-P, Galloway CM, Black CC. A reassessment of glycolysis and gluconeogenesis in higher plants. Physiol Plant [Internet]. 1988 Mar;72(3):650–4. Available from: https://onlinelibrary.wiley.com/doi/10.1111/j.1399-3054.1988.tb09177.x
55. Fernie A, Tauberger E, Roessner U, Willmitzer L, Trethewey R, Lytovchenko A. Antisense repression of cytosolic phosphoglucomutase in potato (Solanum tuberosum) results in severe growth retardation, reduction in tuber number and altered carbon metabolism. Planta [Internet]. 2002 Feb 1;214(4):510–20. Available from: http://link.springer.com/10.1007/s004250100644
56. Sturm A, Tang G-Q. The sucrose-cleaving enzymes of plants are crucial for development, growth and carbon partitioning. Trends Plant Sci [Internet]. 1999 Oct;4(10):401–7. Available from: https://linkinghub.elsevier.com/retrieve/pii/S1360138599014703
57. Kemi U, Niittyvuopio A, Toivainen T, Pasanen A, Quilot-Turion B, Holm K, et al. Role of vernalization and of duplicated FLOWERING LOCUS C in the perennial Arabidopsis lyrata. New Phytol [Internet]. 2013 Jan 26;197(1):323–35. Available from: https://onlinelibrary.wiley.com/doi/10.1111/j.1469-8137.2012.04378.x
58. Sheldon CC. The molecular basis of vernalization: The central role of FLOWERING LOCUS C (FLC). Proc Natl Acad Sci [Internet]. 2000 Mar 28;97(7):3753–8. Available from: http://www.pnas.org/cgi/doi/10.1073/pnas.060023597
59. Soppe WJJ, Viñegra de la Torre N, Albani MC. The Diverse Roles of FLOWERING LOCUS C in Annual and Perennial Brassicaceae Species. Front Plant Sci [Internet]. 2021 Feb 15;12. Available from: https://www.frontiersin.org/articles/10.3389/fpls.2021.627258/full
60. Devoto A, Muskett PR, Shirasu K. Role of ubiquitination in the regulation of plant defence against pathogens. Curr Opin Plant Biol [Internet]. 2003 Aug;6(4):307–11. Available from: https://linkinghub.elsevier.com/retrieve/pii/S1369526603000608
61. Papp B, Pál C, Hurst LD. Dosage sensitivity and the evolution of gene families in yeast. Nature [Internet]. 2003 Jul;424(6945):194–7. Available from: http://www.nature.com/articles/nature01771
62. Bekaert M, Edger PP, Pires JC, Conant GC. Two-Phase Resolution of Polyploidy in the Arabidopsis Metabolic Network Gives Rise to Relative and Absolute Dosage Constraints. Plant Cell [Internet]. 2011 May 1;23(5):1719–28. Available from: https://academic.oup.com/plcell/article/23/5/1719/6097060
63. Matzke MA, Mosher RA. RNA-directed DNA methylation: an epigenetic pathway of increasing complexity. Nat Rev Genet [Internet]. 2014 Jun 8;15(6):394–408. Available from: http://www.nature.com/articles/nrg3683
64. Shen Y, Lei T, Cui X, Liu X, Zhou S, Zheng Y, et al. Arabidopsis histone deacetylase HDA15 directly represses plant response to elevated ambient temperature. Plant J [Internet]. 2019 Dec 12;100(5):991–1006. Available from: https://onlinelibrary.wiley.com/doi/10.1111/tpj.14492
65. Chao D-Y, Silva A, Baxter I, Huang YS, Nordborg M, Danku J, et al. Genome-Wide Association Studies Identify Heavy Metal ATPase3 as the Primary Determinant of Natural Variation in Leaf Cadmium in Arabidopsis thaliana. Bomblies K, editor. PLoS Genet [Internet]. 2012 Sep 6;8(9):e1002923. Available from: https://dx.plos.org/10.1371/journal.pgen.1002923
66. Morel M, Crouzet J, Gravot A, Auroy P, Leonhardt N, Vavasseur A, et al. AtHMA3, a P1B-ATPase Allowing Cd/Zn/Co/Pb Vacuolar Storage in Arabidopsis. Plant Physiol [Internet]. 2009 Feb 6;149(2):894–904. Available from: https://academic.oup.com/plphys/article/149/2/894/6108046
67. Liu L, Gao H, Li S, Han Z, Li B. Calcium signaling networks mediate nitrate sensing and responses in Arabidopsis. Plant Signal Behav [Internet]. 2021 Oct 3;16(10):1938441. Available from: https://www.tandfonline.com/doi/full/10.1080/15592324.2021.1938441
68. Lin J, Yu L, Xiang C. ARABIDOPSIS NITRATE REGULATED 1 acts as a negative modulator of seed germination by activating ABI3 expression. New Phytol [Internet]. 2020 Jan 14;225(2):835–47. Available from: https://onlinelibrary.wiley.com/doi/10.1111/nph.16172

Supplementary Information

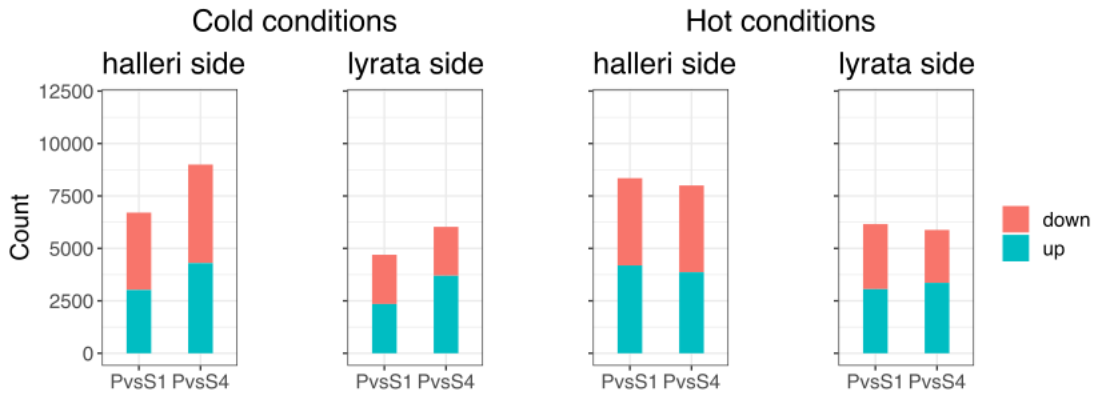
Supplementary Figure 1



Supplementary Figure 1: Differential expression plot for first generation progenitors (reference) against first and fourth generation synthetics in cold and hot conditions. Red and blue dots represent significantly up or down regulated genes,

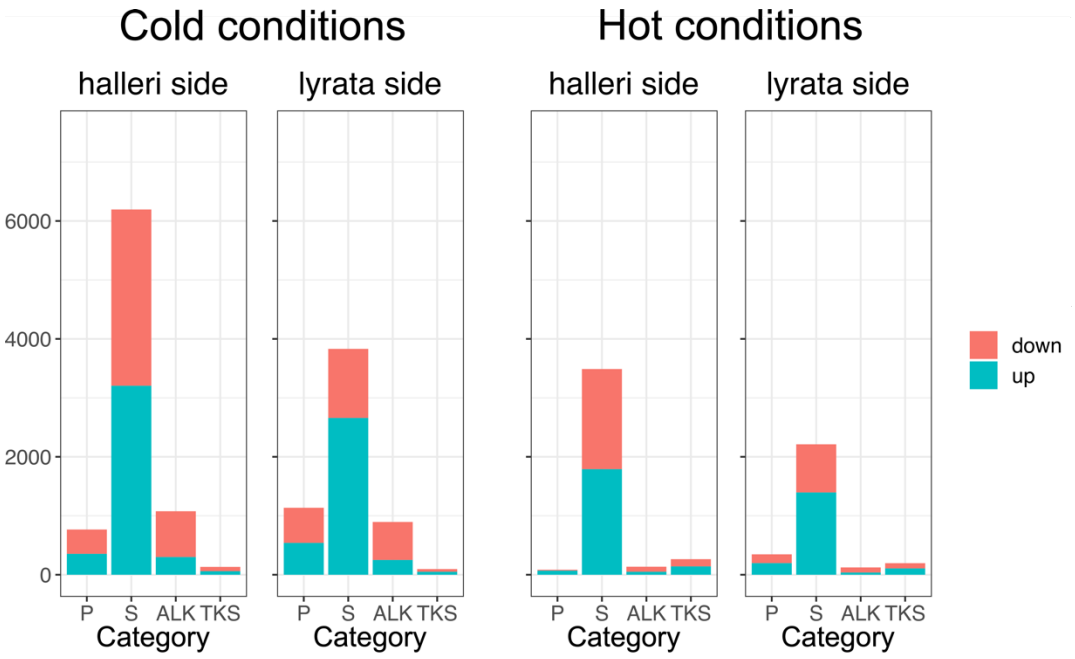
while black dots were genes showing no expression change. Numbers for each one of these categories are shown at the bottom right of each plot.

Supplementary Figure 2



Supplementary Figure 2: Total number of differentially expressed genes for first and fourth generation synthetics when compared to first generation progenitors (PvsS1 and PvsS4 respectively). Two left-side plots are for cold conditions, while right-side plots are for hot conditions.

Supplementary Figure 3

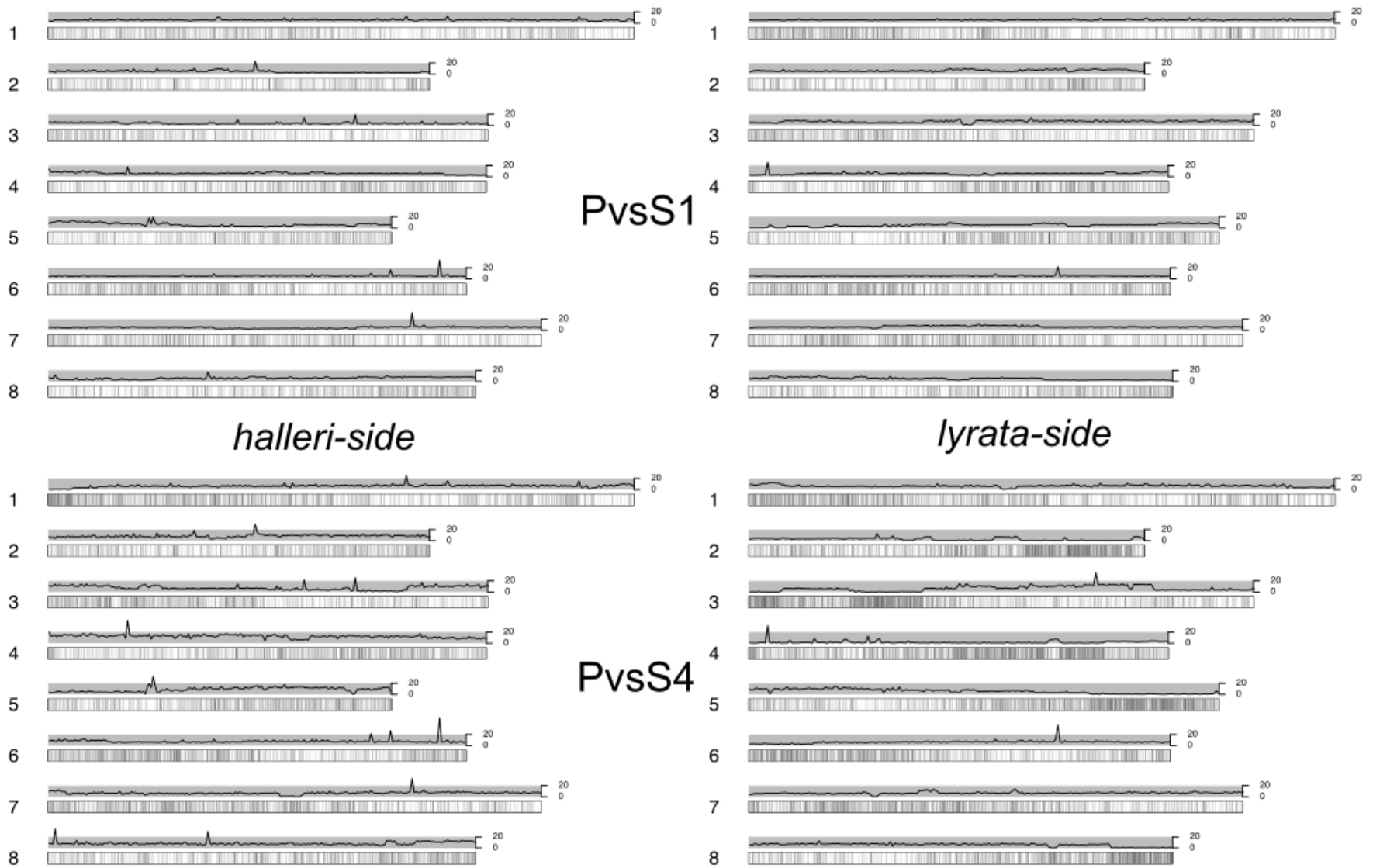


Supplementary Figure 3: Total number of differentially expressed genes across generations for progenitors (P), synthetics (S) and natural lines (ALK and TKS). The fourth generation was used as reference for all samples. Barplots are shown for both conditions and progenitors' sides. Each barplot shows the proportion of upregulated (up) and downregulated (down) genes.

Supplementary Figure 4 (next page)

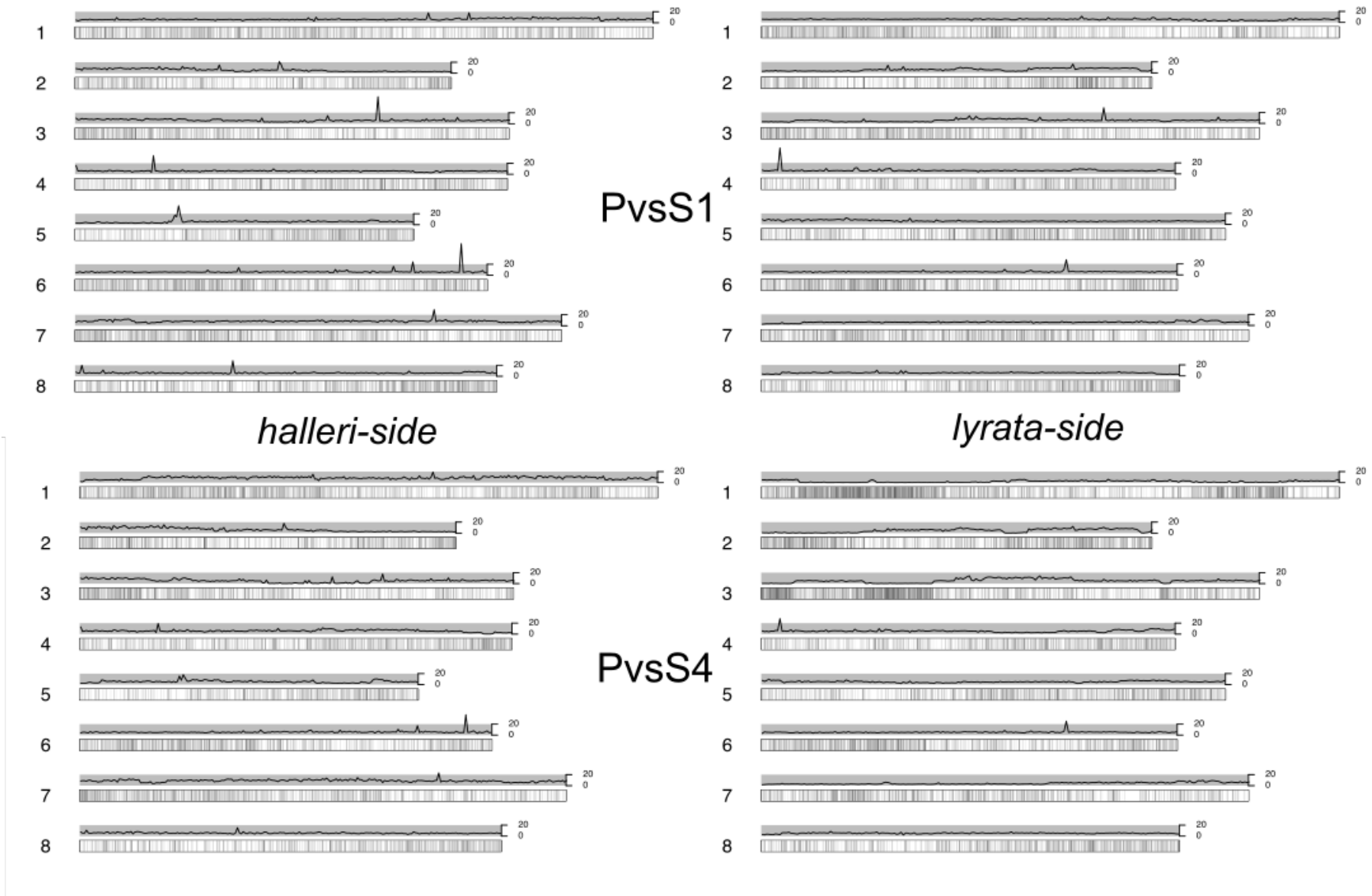
Supplementary Figure 4: distribution of DEGs along chromosomes for PvsS1 and PvsS4 in cold conditions with respect to coverage. Each plot shows eight white rectangles, each representing a chromosome, with a black line showing the position of detected DEGs. At the top of each chromosome the coverage value from whole genome bisulfite sequencing data of a synthetic individual is shown. The first row shows DEGs for PvsS1 with coverage values from G1 synthetics, while the second

row shows DEGs for PvsS4 with coverage from G4 synthetics. The right and left sides are for halleri-side and lyrata-side comparisons respectively.

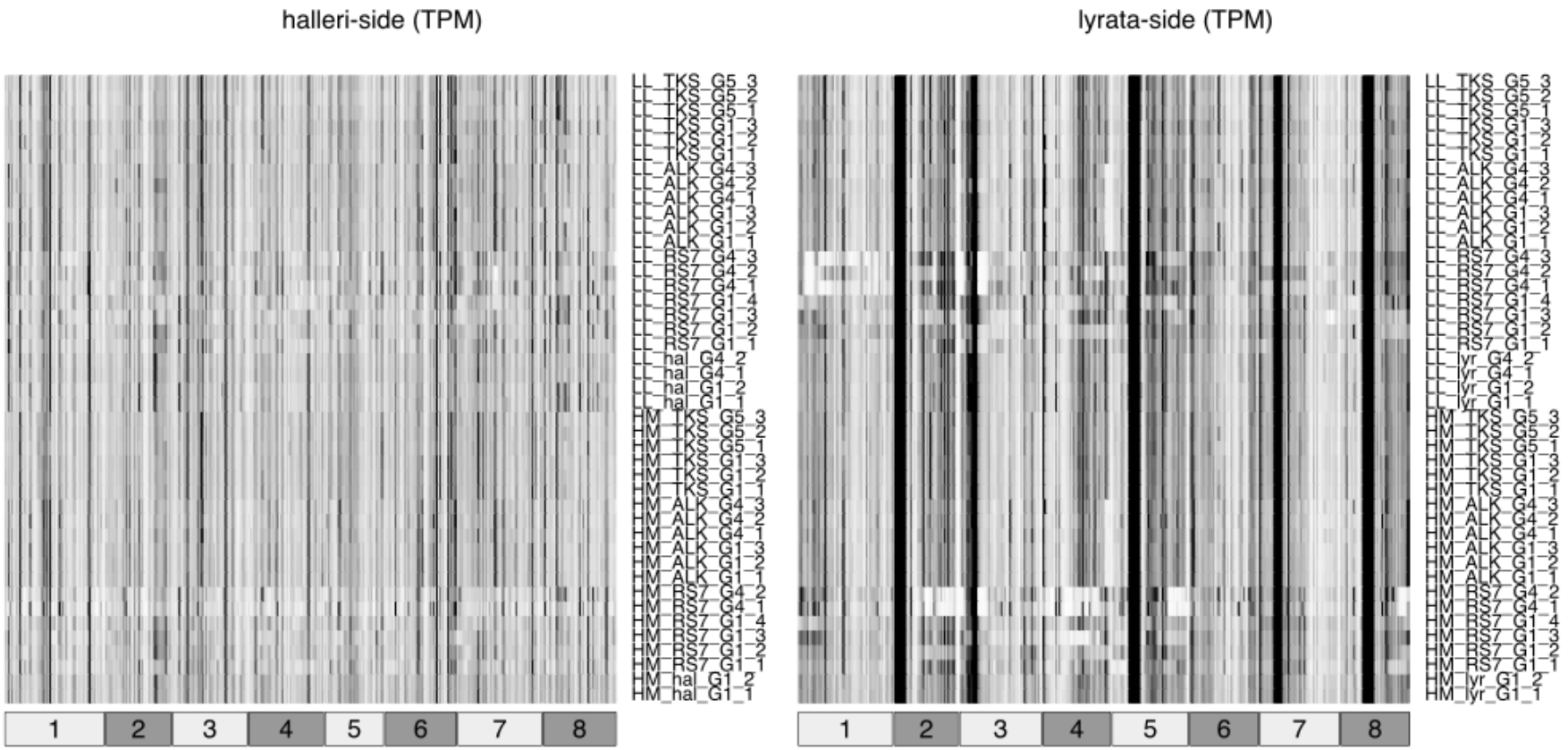

PvsS1
PvsS4
halleri-side
lyrata-side
1
2
3
4
5
6
7
8
20
0

Supplementary Figure 5 (next page)

Supplementary Figure 5: distribution of DEGs along chromosomes for PvsS1 and PvsS4 in hot conditions with respect to coverage. Each plot shows eight white rectangles, each representing a chromosome, with a black line showing the position of detected DEGs. At the top of each chromosome the coverage value from whole genome bisulfite sequencing data of a synthetic individual is shown. The first row shows DEGs for PvsS1 with coverage values from G1 synthetics, while the second row shows DEGs for PvsS4 with coverage from G4 synthetics. The right and left sides are for halleri-side and lyrata-side comparisons respectively.

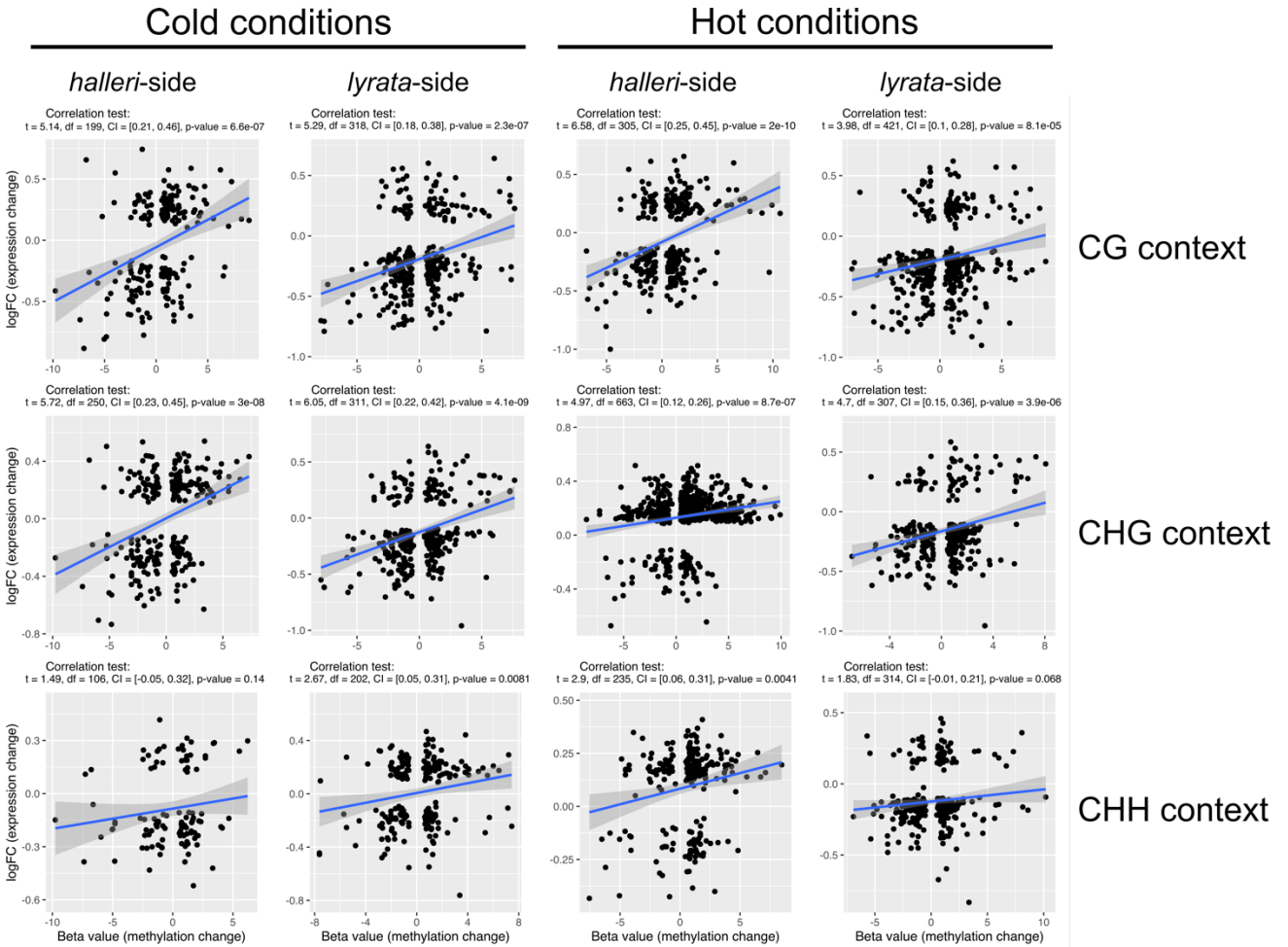

PvsS1
PvsS4
halleri-side
lyrata-side
1
2
3
4
5
6
7
8
20
0

Supplementary Figure 6



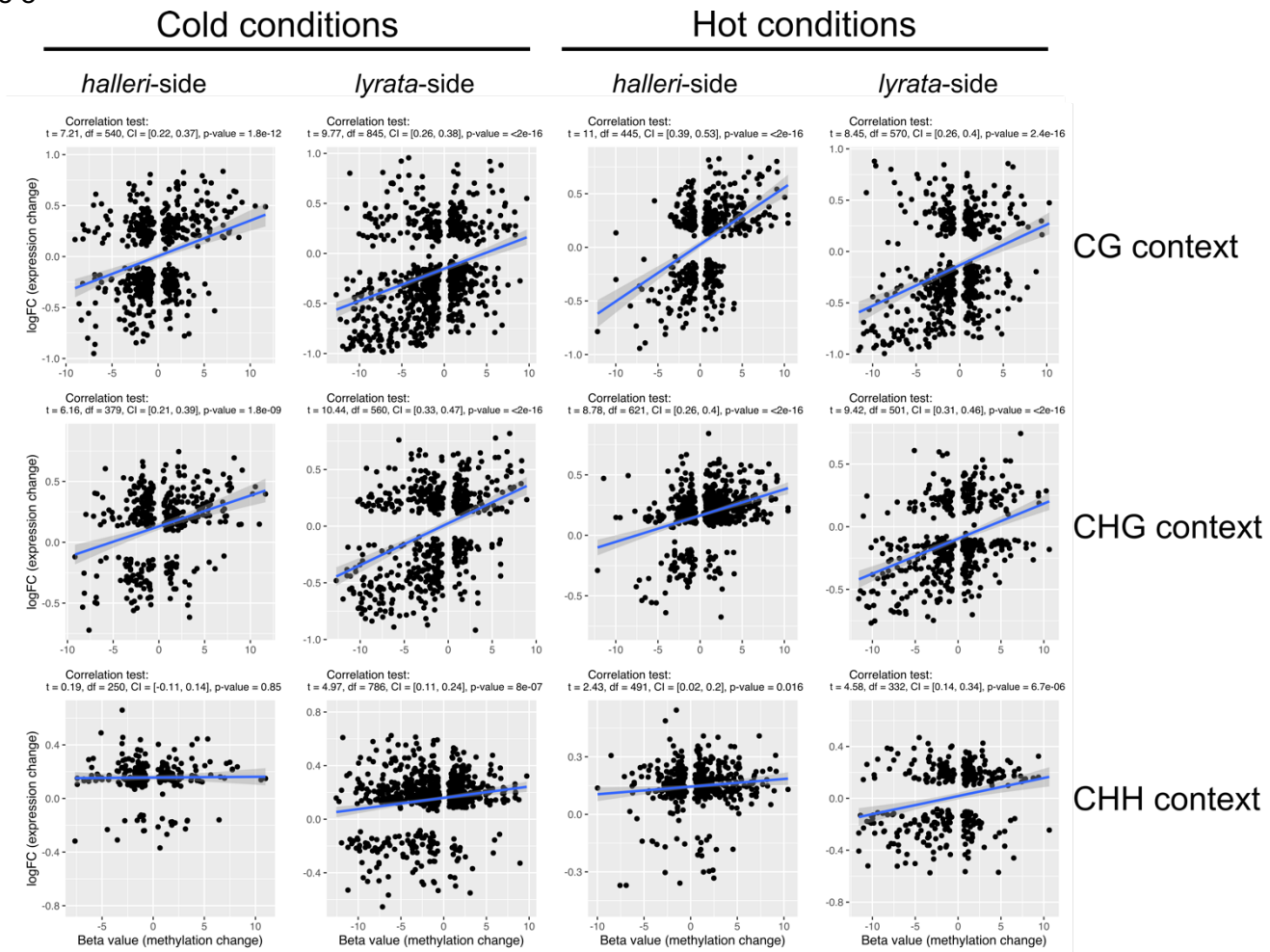
Supplementary Figure 6: Heatmap of the normalized coverage for all samples analyzed over the whole genome. Each heatmap shows coverage (0X = white, >=100X = black) for 500kb stretches windows along the genome for all samples analyzed on both conditions (labels on the right side: HM = cold conditions, LL = hot conditions). Chromosomes are shown at the bottom. Left and right heatmaps are for halleri-side and lyrata-side respectively. For G4 synthetics, particularly on lyrate-side, long white bands corresponding to homoeologous exchanges events can be recognized.

Supplementary Figure 7



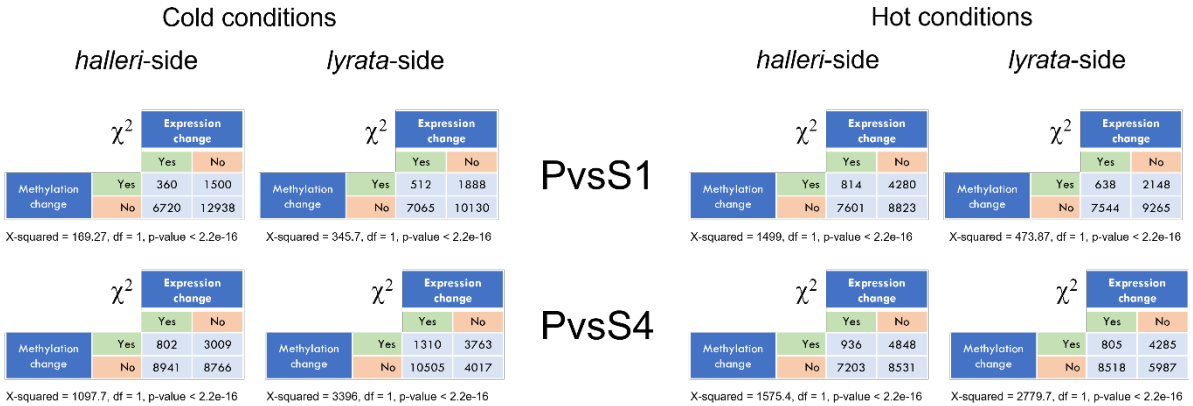
Supplementary Figure 7: relationship and correlation between expression and methylation changes in PvsS1 across conditions, progenitors' side and methylation context. Each scatter plot shows raw methylation changes (beta values) on the x-axis and corresponding log fold changes on the y-axis with each dot representing a gene. For each scatter plot a linear regression was computed with its confidence interval (shaded area around the regression line) together with a Pearson's correlation test with results shown at the top of each plot. A significant p-value indicates significant correlation between methylation and expression changes. Each row represents a specific methylation context (CG, CHG or CHH). Each column represents a progenitors' side (halleri or lyrata) at a given condition (hot or cold).

Supplementary Figure 8



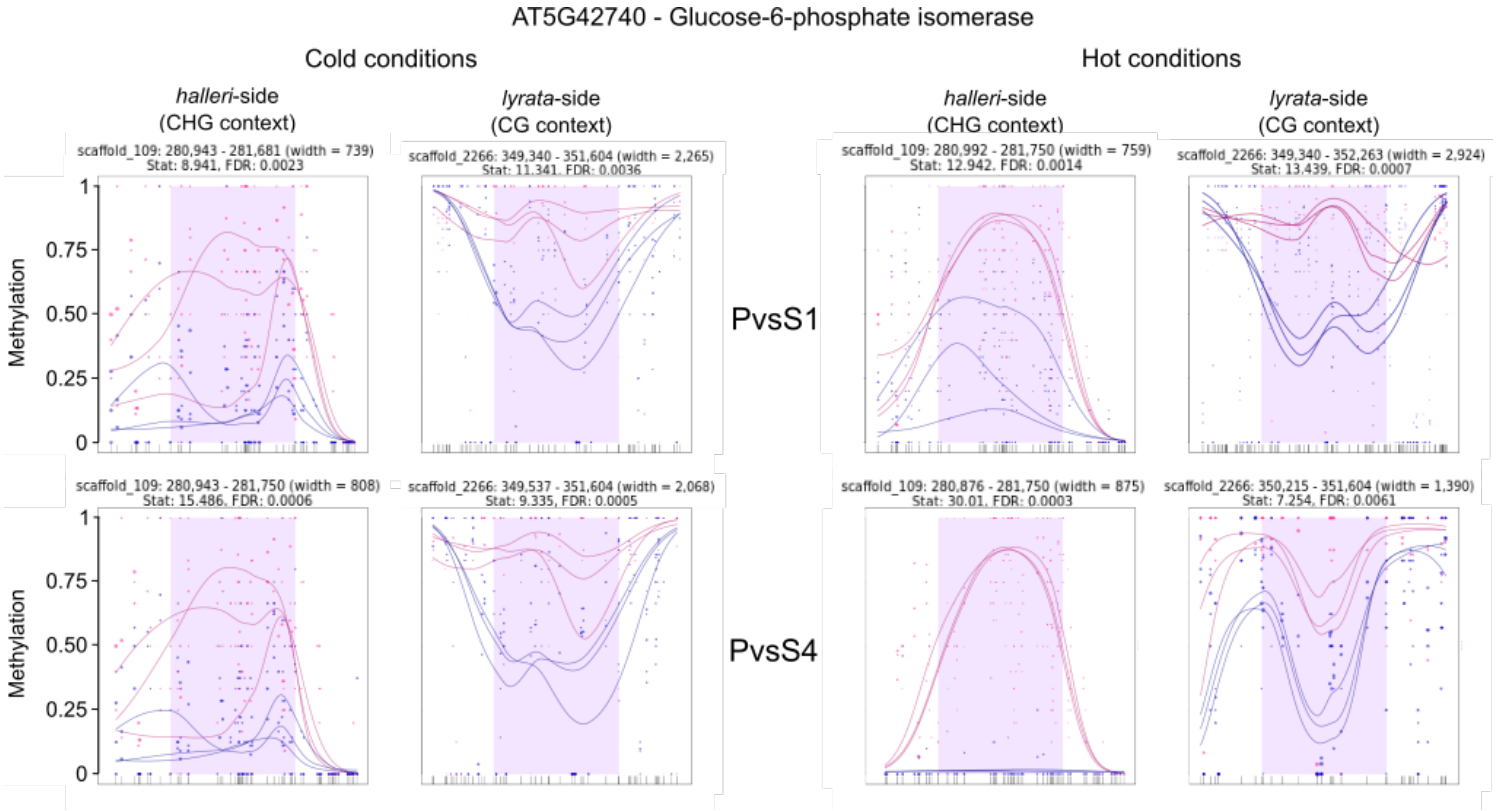
Supplementary Figure 8: relationship and correlation between expression and methylation changes in PvsS4 across conditions, progenitors' side and methylation context. Each scatter plot shows raw methylation changes (beta values) on the x-axis and corresponding log fold changes on the y-axis with each dot representing a gene. For each scatter plot a linear regression was computed with its confidence interval (shaded area around the regression line) together with a Pearson's correlation test with results shown at the top of each plot. A significant p-value indicates significant correlation between methylation and expression changes. Each row represents a specific methylation context (CG, CHG or CHH). Each column represents a progenitors' side (halleri or lyrata) at a given condition (hot or cold).

Supplementary Figure 9



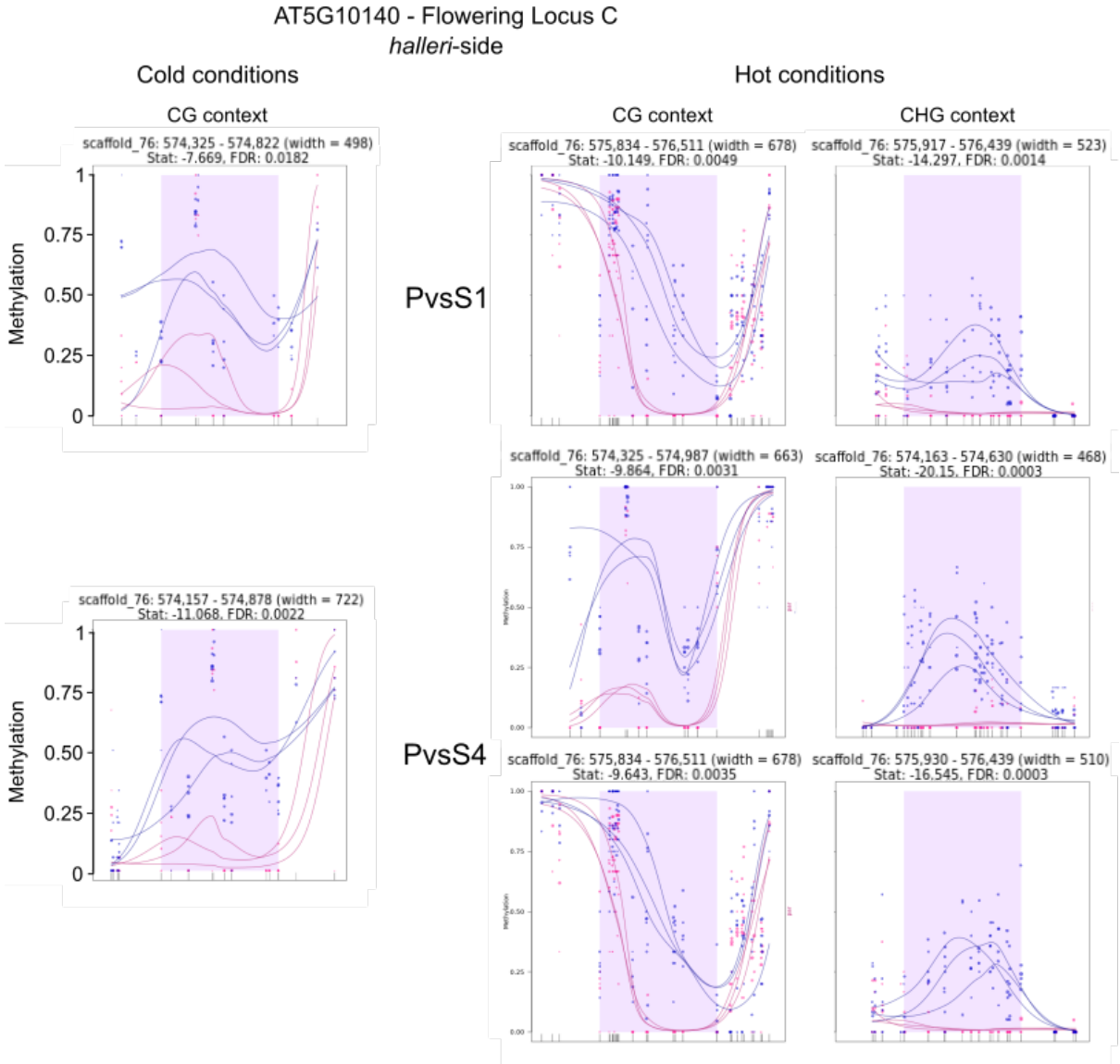
Supplementary Figure 9: Chi-square test for overlaps between DEGs and DMGs for all our comparisons between synthetics and first generation progenitors. Each contingency table shows the number of genes showing changes in both methylation and expression, either of the two and neither of the two. At the bottom of each table details of the chi-square test can be found. Tables on the left side are for comparisons in cold conditions, while tables on the right are for comparisons in hot conditions. Each condition shows results for helleri-side (left) and lyrate-side (right).

Supplementary Figure 10



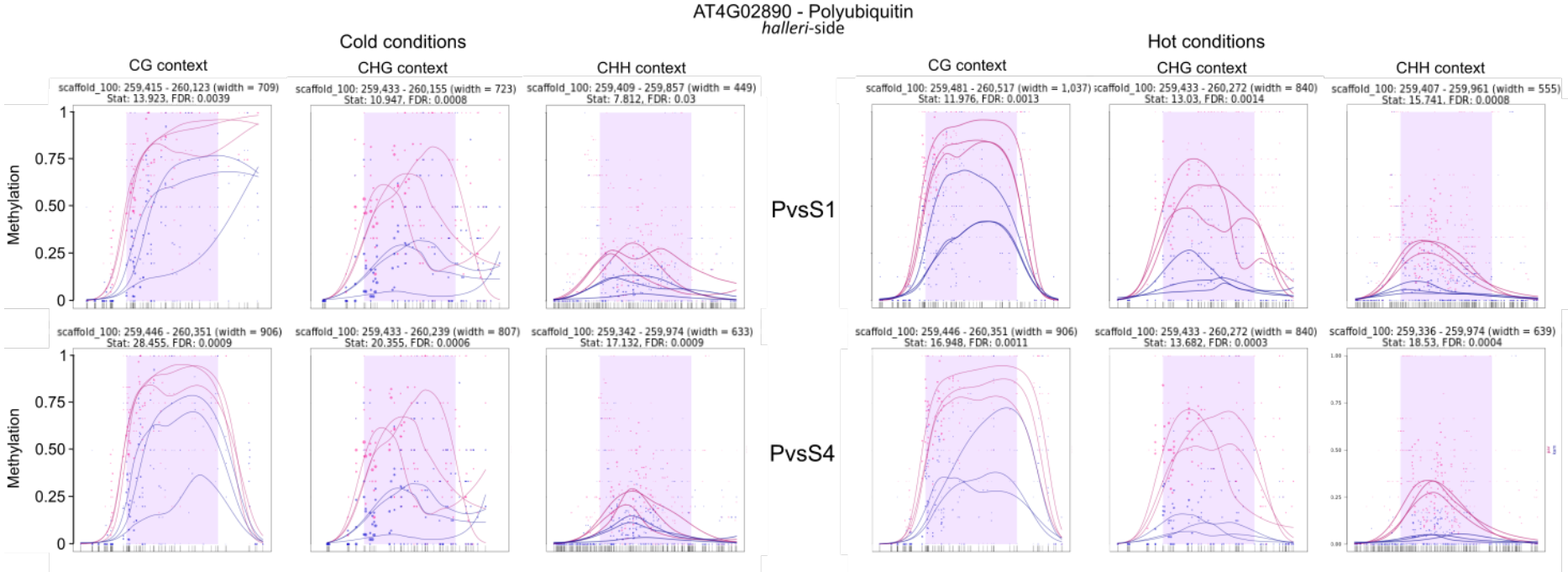
Supplementary Figure 10: differentially methylated regions overlapping with the Glucose-6-phosphate isomerase gene (AT5G42740). Each plot shows methylation levels on the y-axis and genomic position on the x-axis with each bar representing a cytosine. Blue lines and dots are for progenitors' samples while pink lines are from synthetic A. kamchatica samples. Dots represent methylation levels for a given cytosine. Each line shows the smoothed methylation pattern for each replicate (three per sample). The rectangular shade within each plot shows the exact range of the DMR. At the top of each plot, genomic coordinates and statistics for the DMR are given. Each row represents DMRs resulting from a specific comparison: either progenitor vs synthetic G1 (PvsS1) or progenitor vs synthetic G4 (PvsS4).

Supplementary Figure 11



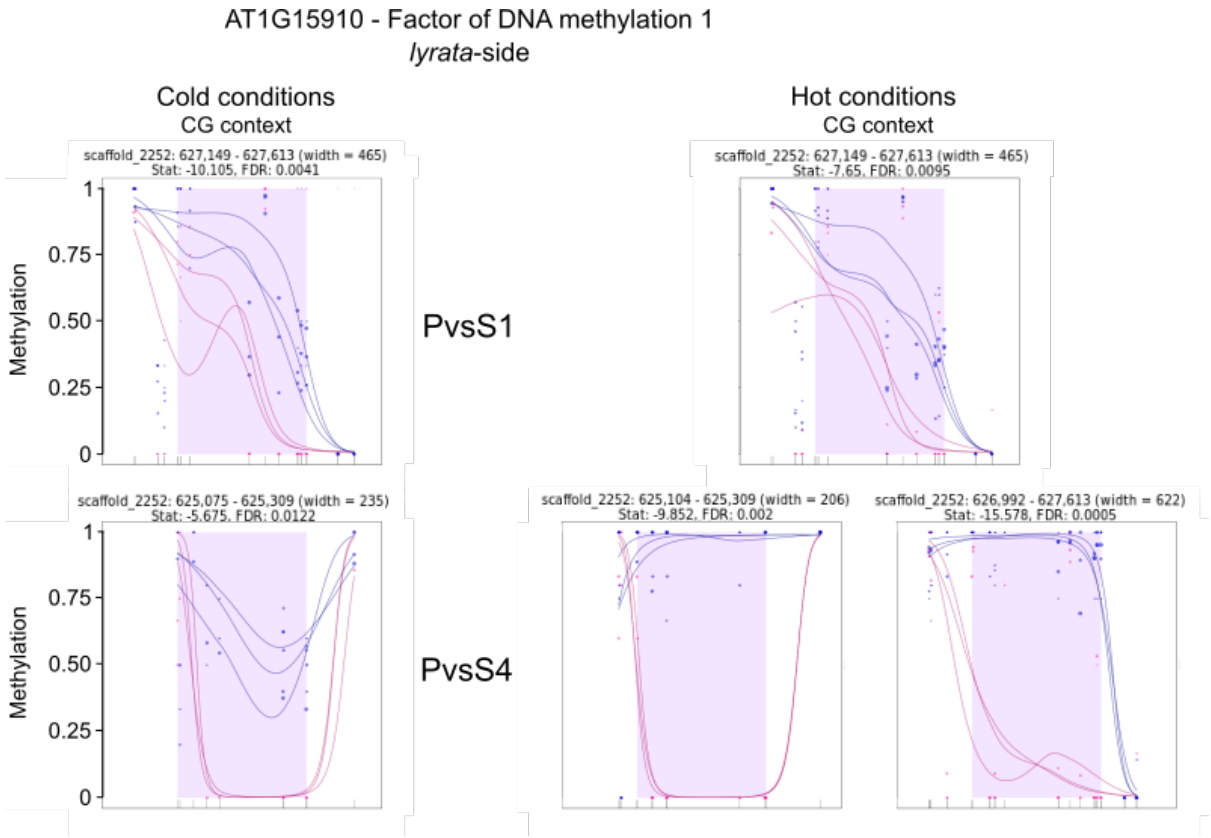
Supplementary Figure 11: differentially methylated regions overlapping with the Flowering Locus C gene (AT5G10140). Each plot shows methylation levels on the y-axis and genomic position on the x-axis with each bar representing a cytosine. Blue lines and dots are for progenitors' samples while pink lines are from synthetic A. kamchatica samples. Dots represent methylation levels for a given cytosine. Each line shows the smoothed methylation pattern for each replicate (three per sample). The rectangular shade within each plot shows the exact range of the DMR. At the top of each plot, genomic coordinates and statistics for the DMR are given. Each row represents DMRs resulting from a specific comparison: either progenitor vs synthetic G1 (PvsS1) or progenitor vs synthetic G4 (PvsS4).

Supplementary Figure 12



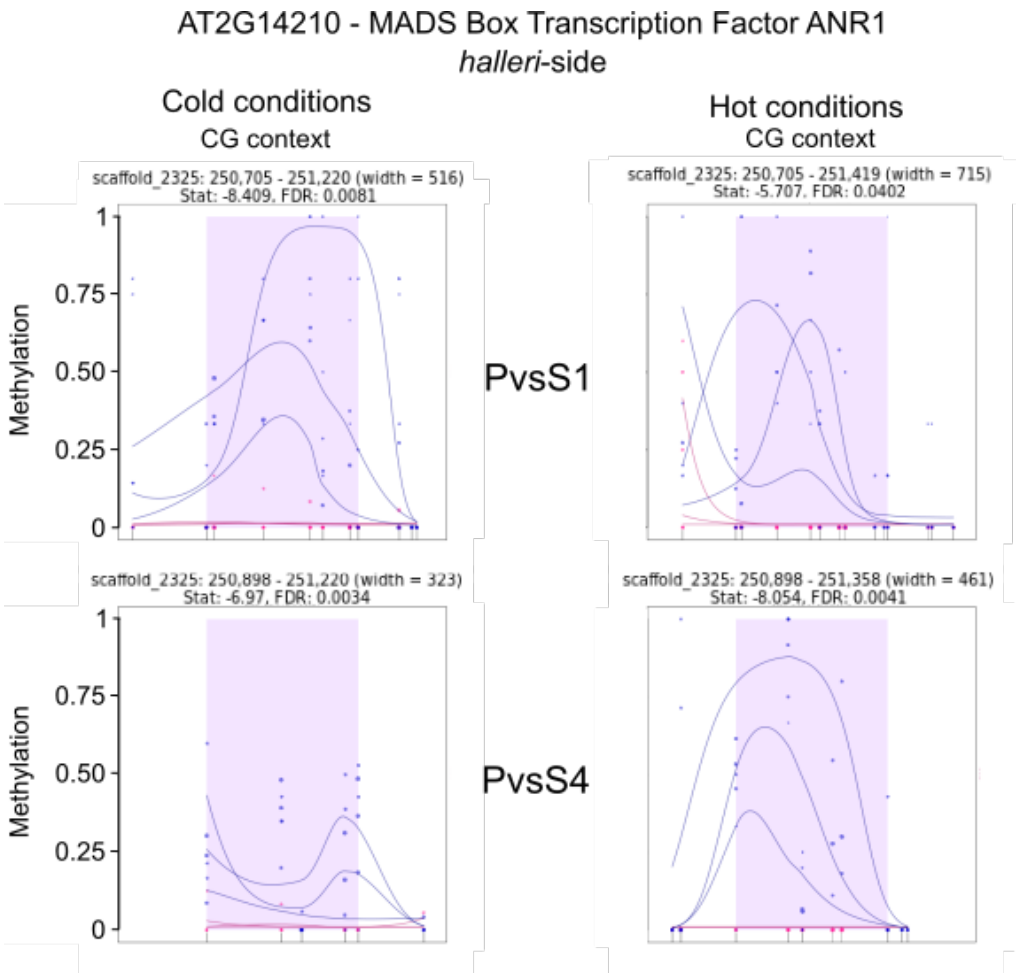
Supplementary Figure 12: differentially methylated regions overlapping with the polyubiquitin gene (AT4G02890). Each plot shows methylation levels on the y-axis and genomic position on the x-axis with each bar representing a cytosine. Blue lines and dots are for progenitors' samples while pink lines are from synthetic A. kamchatica samples. Dots represent methylation levels for a given cytosine. Each line shows the smoothed methylation pattern for each replicate (three per sample). The rectangular shade within each plot shows the exact range of the DMR. At the top of each plot, genomic coordinates and statistics for the DMR are given. Each row represents DMRs resulting from a specific comparison: either progenitor vs synthetic G1 (PvsS1) or progenitor vs synthetic G4 (PvsS4).

Supplementary Figure 13



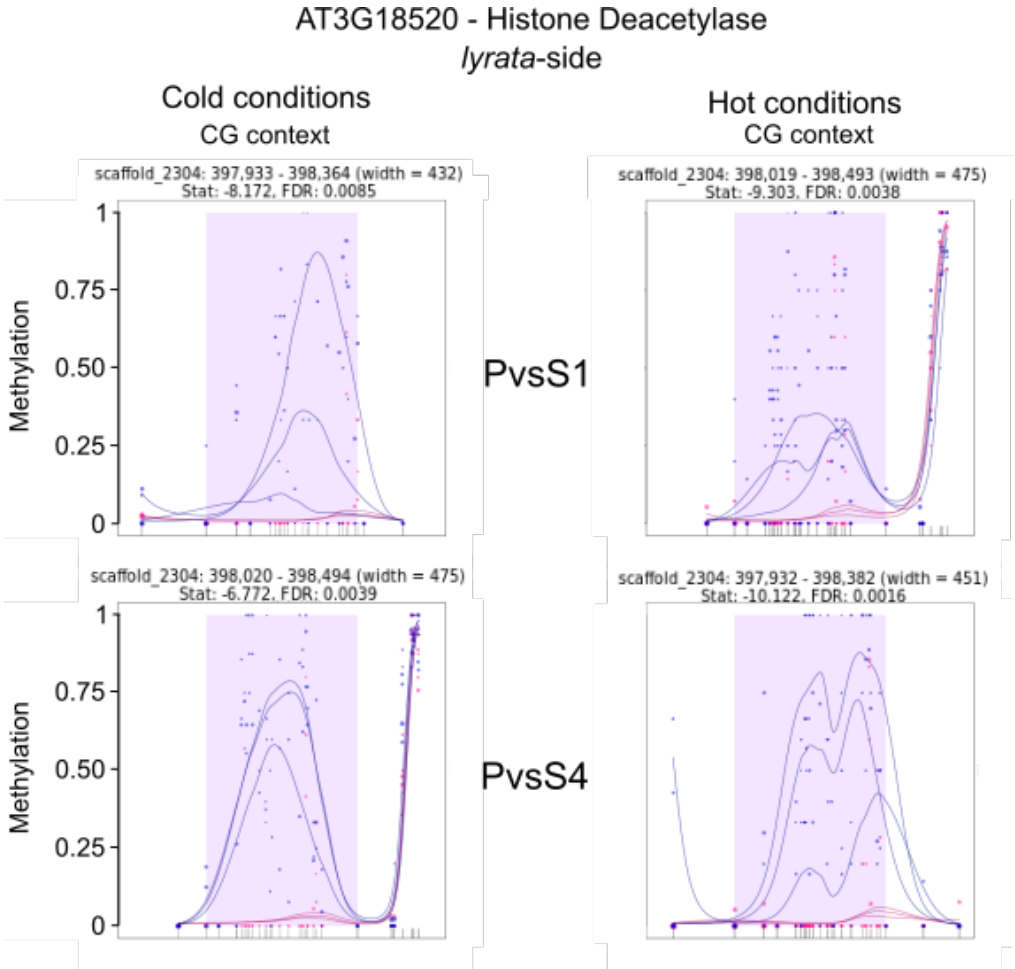
Supplementary Figure 13: differentially methylated regions overlapping with the Factor of DNA Methylation 1 gene (AT1G15910). Each plot shows methylation levels on the y-axis and genomic position on the x-axis with each bar representing a cytosine. Blue lines and dots are for progenitors' samples while pink lines are from synthetic A. kamchatica samples. Dots represent methylation levels for a given cytosine. Each line shows the smoothed methylation pattern for each replicate (three per sample). The rectangular shade within each plot shows the exact range of the DMR. At the top of each plot, genomic coordinates and statistics for the DMR are given. Each row represents DMRs resulting from a specific comparison: either progenitor vs synthetic G1 (PvsS1) or progenitor vs synthetic G4 (PvsS4).

Supplementary Figure 14



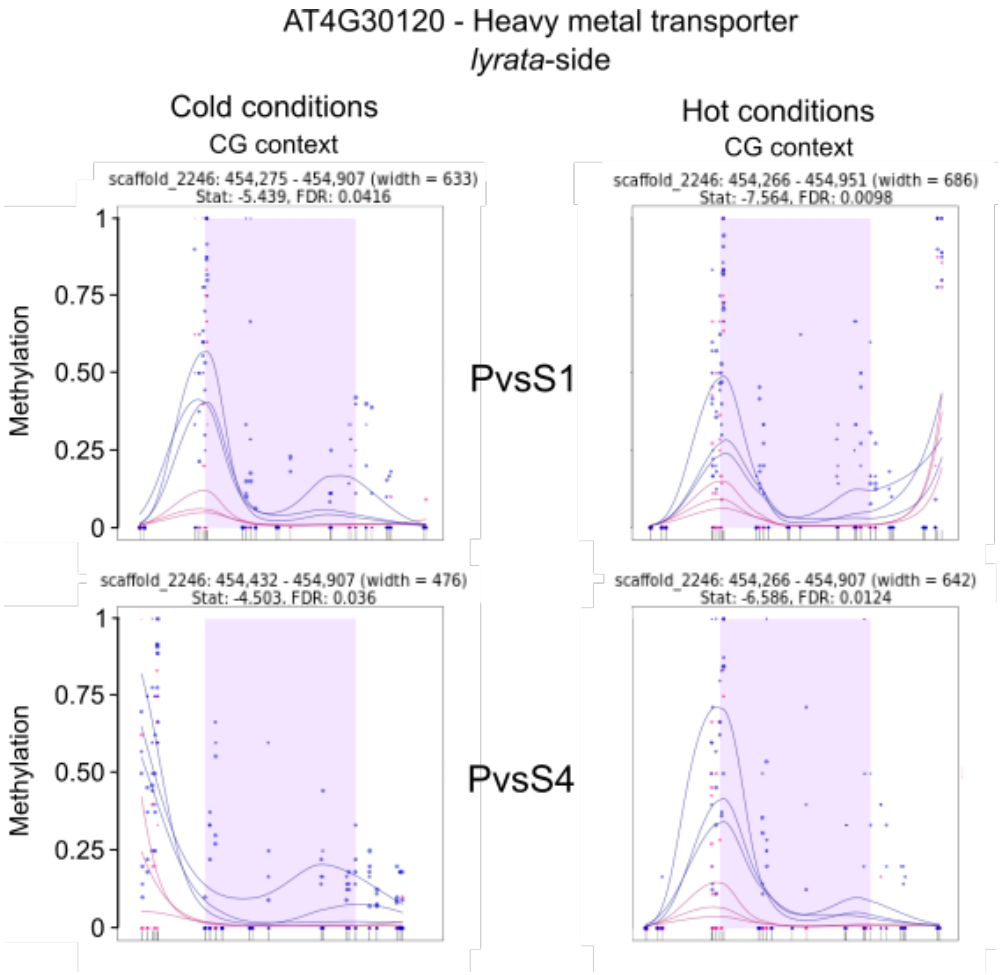
Supplementary Figure 14: differentially methylated regions overlapping with MADS Box Transcription Factor ANR1 gene (AT2G14210). Each plot shows methylation levels on the y-axis and genomic position on the x-axis with each bar representing a cytosine. Blue lines and dots are for progenitors' samples while pink lines are from synthetic A. kamchatica samples. Dots represent methylation levels for a given cytosine. Each line shows the smoothed methylation pattern for each replicate (three per sample). The rectangular shade within each plot shows the exact range of the DMR. At the top of each plot, genomic coordinates and statistics for the DMR are given. Each row represents DMRs resulting from a specific comparison: either progenitor vs synthetic G1 (PvsS1) or progenitor vs synthetic G4 (PvsS4).

Supplementary Figure 15



Supplementary Figure 15: differentially methylated regions overlapping with histone deacetylase gene (AT3G18520). Each plot shows methylation levels on the y-axis and genomic position on the x-axis with each bar representing a cytosine. Blue lines and dots are for progenitors' samples while pink lines are from synthetic A. kamchatica samples. Dots represent methylation levels for a given cytosine. Each line shows the smoothed methylation pattern for each replicate (three per sample). The rectangular shade within each plot shows the exact range of the DMR. At the top of each plot, genomic coordinates and statistics for the DMR are given. Each row represents DMRs resulting from a specific comparison: either progenitor vs synthetic G1 (PvsS1) or progenitor vs synthetic G4 (PvsS4).

Supplementary Figure 16



Supplementary Figure 16: differentially methylated regions overlapping with heavy metal transporter gene (AT4G30120). Each plot shows methylation levels on the y-axis and genomic position on the x-axis with each bar representing a cytosine. Blue lines and dots are for progenitors' samples while pink lines are from synthetic A. kamchatica samples. Dots represent methylation levels for a given cytosine. Each line shows the smoothed methylation pattern for each replicate (three per sample). The rectangular shade within each plot shows the exact range of the DMR. At the top of each plot, genomic coordinates and statistics for the DMR are given. Each row represents DMRs resulting from a specific comparison: either progenitor vs synthetic G1 (PvsS1) or progenitor vs synthetic G4 (PvsS4).

Supplementary Table 1

Sample	Total reads	Aligned reads (H)	Aligned reads (L)	Classified reads (H)	Classified reads (L)	Final reads
HM_hal_G1_1	11’175’233	8252990 (73.85%)	-	-	-	8’252’990 (73.85%)
HM_hal_G1_2	7’268’968	5113962 (70.35%)	-	-	-	5’113’962 (70.35%)
HM_hal_G1_3	10’643’417	5507190 (51.74%)	-	-	-	5’507’190 (51.74%)
HM_hal_G1_4	13’249’389	6796594 (51.30%)	-	-	-	6’796’594 (51.30%)
HM_hal_G4_1	10’345’270	5’211’714 (50.38%)	-	-	-	5’211’714 (50.38%)
HM_hal_G4_2	11’456’683	6’315’691 (55.13%)	-	-	-	6’315’691 (55.13%)
HM_hal_G4_3	9’158’897	3’980’174 (43.46%)	-	-	-	3’980’174 (43.46%)
HM_lyr_G1_1	6’834’772	-	5’000’435 (73.16%)	-	-	5’000’435 (73.16%)
HM_lyr_G1_2	12’027’193	-	8’004’646 (66.55%)	-	-	8’004’646 (66.55%)
HM_lyr_G1_3	13’134’073	-	7’719’690 (58.78%)	-	-	7’719’690 (58.78%)
HM_lyr_G1_4	9’888’915	-	5’830’735 (58.96%)	-	-	5’830’735 (58.96%)
HM_lyr_G4_1	10’172’327	-	5’809’313 (57.11%)	-	-	5’809’313 (57.11%)
HM_lyr_G4_2	10’725’366	-	5’370’047 (50.07%)	-	-	5’370’047 (50.07%)
HM_lyr_G4_3	9’462’701	-	5’901’353 (62.36%)	-	-	5’901’353 (62.36%)
HM_RS7_G1_1	12’280’415	8’077’335 (65.77%)	8’105’014 (66.00%)	4’936’496 (40.20%)	4’930’810 (40.15%)	9’867’306 (80.35%)
HM_RS7_G1_2	12’178’827	7’750’113 (63.64%)	7’652’582 (62.84%)	5’362’206 (44.03%)	4’241’788 (34.83%)	9’603’994 (78.86%)
HM_RS7_G1_3	14’074’377	9’033’379 (64.18%)	8’870’169 (63.02%)	6’259’338 (44.47%)	4’499’056 (31.97%)	10’758’394 (76.44%)
HM_RS7_G1_4	15’093’816	10’106’710 (66.96%)	10’143’460 (67.20%)	6’197’444 (41.06%)	6’359’416 (42.13%)	12’556’860 (83.19%)
HM_RS7_G4_1	19’684’850	8’515’727 (43.26%)	8’394’075 (42.64%)	4’732’804 (24.04%)	2’558’240 (13.00%)	7’291’044 (37.04%)
HM_RS7_G4_2	20’968’910	11’325’885 (54.01%)	10’810’968 (51.56%)	6’794’118 (32.40%)	3’790’594 (18.08%)	10’584’712 (50.48%)
HM_RS7_G4_3	16’687’837	9’258’946 (55.48%)	9’062’941 (54.31%)	5’637’828 (33.78%)	3’256’514 (19.51%)	8’894’342 (53.29%)
HM_ALK_G1_1	21’982’015	12’554’980 (57.11%)	12’370’083 (56.27%)	7’008’918 (31.88%)	5’745’834 (26.14%)	12’754’752 (58.02%)
HM_ALK_G1_2	19’961’647	10’220’943 (51.20%)	10’073’491 (50.46%)	5’360’548 (26.85%)	4’377’242 (21.93%)	9’737’790 (48.78%)
HM_ALK_G1_3	22’362’100	11’569’410 (51.74%)	11’425’735 51.09%)	6’054’720 (27.08%)	4’951’886 (22.14%)	11’006’606 (49.22%)
HM_ALK_G4_1	8’642’373	4’273’163 (49.44%)	4’188’392 (48.46%)	2’129’472 (24.64%)	1’740’568 (20.14%)	3’870’040 (44.78%)
HM_ALK_G4_2	12’530’026	6’600’686 (52.68%)	6’516’705 (52.01%)	3’446’972 (27.51%)	2’784’050 (22.22%)	6’231’022 (49.73%)
HM_ALK_G4_3	13’380’453	6’080’050 (45.44%)	6’057’016 (45.27%)	2’951’302 (22.06%)	2’512’268 (18.78%)	5’463’570 (40.84%)
HM_TKS_G1_1	20’361’317	10’217’059 (50.18%)	9’981’282 (49.02%)	5’264’738 (25.86%)	3’915’638 (19.23%)	9’180’376 (45.09%)
HM_TKS_G1_2	19’212’307	10’400’897 (54.14%)	10’133’817 (52.75%)	5’707’244 (29.71%)	4’247’380 (22.11%)	9’954’624 (51.82%)
HM_TKS_G1_3	24’015’434	12’755’123 (53.11%)	12’461’662 (51.89%)	6’665’576 (27.76%)	4’981’832 (20.74%)	11’647’408 (48.5%)
HM_TKS_G4_1	15’268’453	6’545’422 (42.87%)	6’376’243 (41.76%)	3’302’458 (21.63%)	2’451’388 (16.06%)	5’753’846 (37.69%)
HM_TKS_G4_2	17’407’969	9’472’456 (54.41%)	9’244’727 (53.11%)	5’321’182 (30.57%)	3’975’296 (22.84%)	9’296’478 (53.41%)
HM_TKS_G4_3	10’777’350	5’468’979 (50.75%)	5’326’203 (49.42%)	2’975’104 (27.61%)	2’189’714 (20.32%)	5’164’818 (47.93%)
LL_hal_G1_1	7’892’717	5’551’113 (70.33%)	-	-	-	5’551’113 (70.33%)

LL_hal_G1_2	7’772’078	5’451’322 (70.14%)	-	-	-	5’451’322 (70.14%)
LL_hal_G1_4	11446188	5’935’865 (51.86%)	-	-	-	5’935’865 (51.86%)
LL_hal_G1_5	10059624	5’362’316 (53.31%)	-	-	-	5’362’316 (53.31%)
LL_hal_G1_6	9773696	5’294’641 (54.17%)	-	-	-	5’294’641 (54.17%)
LL_hal_G4_1	13’902’801	9’527’267 (68.53%)	-	-	-	9’527’267 (68.53%)
LL_hal_G4_2	17’884’148	12’146’374 (67.92%)	-	-	-	12’146’374 (67.92%)
LL_lyr_G1_1	8’612’533	-	6’215’178 (72.16%)	-	-	6’215’178 (72.16%)
LL_lyr_G1_2	8’584’999	-	6’213’437 (72.38%)	-	-	6’213’437 (72.38%)
LL_lyr_G1_4	11240706	-	6’017’103 (53.53%)	-	-	6’017’103 (53.53%)
LL_lyr_G1_5	11789598	-	7’309’194 (62.00%)	-	-	7’309’194 (62.00%)
LL_lyr_G1_6	6590717	-	3’612’892 (54.82%)	-	-	3’612’892 (54.82%)
LL_lyr_G4_1	8’941’486	-	6’727’620 (75.24%)	-	-	6’727’620 (75.24%)
LL_lyr_G4_2	16’566’471	-	12’344’140 (74.51%)	-	-	12’344’140 (74.51%)
LL_lyr_G4_3	14’140’740	-	10’727’065 (75.86%)	-	-	10’727’065 (75.86%)
LL_lyr_G4_4	7’351’336	-	5’132’892 (69.82%)	-	-	5’132’892 (69.82%)
LL_lyr_G4_5	10626467	-	5’287’435 (49.76%)	-	-	5’287’435 (49.76%)
LL_RS7_G1_1	16’024’251	11’161’564 (69.65%)	11’040’587 (68.90%)	7’618’282 (47.54%)	6’346’926 (39.61%)	13'965'208 (87.15%)
LL_RS7_G1_2	16’009’272	11’013’736 (68.80%)	10’799’070 (67.46%)	7’788’750 (48.65%)	5’749’172 (35.91%)	13'537'922 (84.56%)
LL_RS7_G1_3	19’599’966	12’708’301 (64.84%)	12’482’257 (63.69%)	8’364’702 (42.68%)	7’158’174 (36.52%)	15'522'876 (79.20%)
LL_RS7_G1_4	19’568’096	12’777’958 (65.30%)	12’676’718 (64.78%)	8’101’818 (41.40%)	7’768’620 (39.70%)	15'870'438 (81.10%)
LL_RS7_G4_1	10290772	6’921’520 (67.26%)	6’763’435 (65.72%)	5’010’412 (48.69%)	2’965’404 (28.82%)	7'975'816 (77.51%)
LL_RS7_G4_2	11381329	7’706’712 (67.71%)	7’520’623 (66.08%)	5’369’380 (47.18%)	3’491’398 (30.68%)	8'860'778 (77.86%)
LL_RS7_G4_3	7914248	4’951’577 (62.57%)	4’771’492 (60.29%)	3’280’990 (41.46%)	1’718’720 (21.72%)	4'999'710 (63.18%)
LL_ALK_G1_1	11481835	6’855’497 (59.71%)	6’787’027 (59.11%)	3’816’996 (33.24%)	3’149’612 (27.43%)	6'966'608 (60.67%)
LL_ALK_G1_2	9972538	6’525’728 (65.44%)	6’476’979 (64.95%)	4’064’976 (40.76%)	3’396’154 (34.06%)	7'461'130 (74.82%)
LL_ALK_G1_3	8262207	5’213’185 (63.10%)	5’127’505 (62.06%)	2’924’008 (35.39%)	2’347’242 (28.41%)	5'271'250 (63.80%)
LL_ALK_G4_1	18243615	11’692’755 (64.09%)	11’530’397 (63.20%)	7’106’812 (38.96%)	5’831’596 (31.97%)	12'938'408 (70.93%)
LL_ALK_G4_2	14149697	8’914’171 (63.00%)	8’742’717 (61.79%)	5’385’348 (38.06%)	4’218’378 (29.81%)	9'603'726 (67.87%)
LL_ALK_G4_3	10457741	6’351’068 (60.73%)	6’236’766 (59.64%)	3’442’796 (32.93%)	2’836’602 (27.12%)	6'279'398 (60.05%)
LL_TKS_G1_1	23370905	12’442’050 (53.24%)	12’165’842 (52.06%)	7’246’058 (31.00%)	5’486’846 (23.48%)	12'732'904 (54.48%)
LL_TKS_G1_2	22334232	11’563’294 (51.77%)	11’699’255 (52.38%)	6’015’372 (26.93%)	5’066’192 (22.68%)	11'081'564 (49.61%)
LL_TKS_G1_3	23803740	12’647’772 (53.13%)	12’292’821 (51.64%)	7’025’412 (29.51%)	5’307’604 (22.30%)	12'333'016 (51.81%)
LL_TKS_G4_1	22810839	11’244’047 (49.29%)	10’960’421 (48.05%)	6’164’942 (27.03%)	4’471’282 (19.60%)	10'636'224 (46.63%)
LL_TKS_G4_2	22122294	11’704’740 (52.91%)	11’394’421 (51.51%)	6’869’786 (31.05%)	5’009’260 (22.64%)	11'879'046 (53.69%)
LL_TKS_G4_3	21240079	10’890’021 (51.27%)	10’630’913 (50.05%)	6’454’126 (30.39%)	4’786’974 (22.54%)	11'241'100 (52.93%)

Supplementary Table 2

				Raw methylation change								Expression change (logFC)							
		Consistent methylation and expression changes on…		*halleri*-side				*lyrata*-side				*halleri*-side				*lyrata*-side			
Gene name	**Gene ID**	**H-side**	**L-side**	**PvsS1**		**PvsS4**		**PvsS1**		**PvsS4**		**PvsS1**		**PvsS4**		**PvsS1**		**PvsS4**	
				C	**H**	**C**	**H**	**C**	**H**	**C**	**H**	**C**	**H**	**C**	**H**	**C**	**H**	**C**	**H**
Glucose-6-phosphate isomerase	AT5G42740	Yes	Yes	-0.32 (CHG)	-0.38 (CHG)	-0.48 (CHG)	-0.67 (CHG)	-0.21 (CG)	-0.18 (CG)	-0.18 (CG)	-0.25 (CG)	-1.17	-0.88	-0.89	-1.62	-2.00	-1.98	-0.48	-1.95
Polyubiquitin (UBQ14)	AT4G02890	Yes	No	-0.32 (CG)	-0.32 (CG)	-0.55 (CG)	-0.35 (CG)	-	-	-	-	1.03	1.00	2.05	0.98	-	-	-	-
Flowering Locus C (FLC)	AT5G10140	Yes	No	0.25 (CG)	0.29 (CG)	0.36 (CG)	0.25 (CG)	-	-	-	-	-0.80	-0.81	-1.38	-0.80	-	-	-	-
Arabidopsis Nitrate Regulated 1 (ANR1)	AT2G14210	No	Yes	-	-	-	-	0.40 (CG)	0.29 (CG)	0.24 (CG)	0.41 (CG)	-	-	-	-	-7.34	-4.82	-7.29	-7.41
Factor of DNA methylation 1 (FDM1)	AT1G15910	No	Yes	-	-	-	-	0.23 (CG)	0.18 (CG)	0.33 (CG)	0.75 (CG)	-	-	-	-	1.85	1.41	1.76	2.18
Heavy Metal ATPase 3 (HMA3)	AT4G30120	No	Yes	-	-	-	-	0.16 (CG)	0.13 (CG)	0.12 (CG)	0.18 (CG)	-	-	-	-	3.56	4.23	3.75	4.45
Histone Deacetylase 15 (HDA15)	AT3G18520	No	Yes	-	-	-	-	0.26 (CG)	0.27 (CG)	0.42 (CG)	0.38 (CG)	-	-	-	-	-1.75	-1.35	-0.95	-1.91

Supplementary table 2: genes of interest with consistent methylation and expression changes in both experimental conditions and for each progenitor side. The first column shows the gene names, the second shows the A. thaliana gene ID, the third column specifies the progenitors side where the gene was found, the fourth and fifth columns show the raw methylation and expression changes. For both methylation and expression changes, each progenitor side is shown, the comparison (PvsS1: progenitor vs Synthetic G1 or PvsS4: progenitor vs Synthetic G1) and the condition (C for cold and H for hot). For raw methylation changes, in the case of multiple contexts showing differential methylation, CG was taken as reference. All raw methylation values represent the change with respect to the progenitor species, i.e. positive values indicate an increase in the synthetic with respect to the progenitors and negative values a decrease.

General discussion

a. Improving reproducibility in polyploid studies to set better grounds for discussion

With the increasing amount of genomic data from polyploids and computational tools to analyze it, reproducibility in the data analysis process in polyploid studies must be discussed and promoted more to offer better and solid grounds for discussion. Reproducibility here refers to methods reproducibility, defined as the completeness and clarity in sharing the tools and procedures to analyze a dataset (1). Reproducibility is a moral responsibility towards science, it allows findings to stand the test of time, it ensures effective application of previous methods on new data and incentivizes reusing code or results for new projects (2). In the specific case of polyploid studies with genomic data, considering the large amount of variation in findings for different species (3), it is worth thinking whether some amount of variation might be attributed to methodological differences. To determine this, we will look at publications on DNA methylation in synthetic polyploids at the whole-genome level and evaluate their reproducibility compared to our study. Four different criteria will be used: disclosure of the version of tools, disclosure of parameters for each tool, availability of code and presence of a workflow. The first three criteria address transparency and clarity in sharing methods, while the fourth is related to easing reproducibility. When examining the three publications available (4–6), it is clear that reproducibility is not properly considered in the case of whole genome DNA methylation (Table 1). In the study from Edger *et al.* only the versions of the tools used were provided without any code or command and no workflow. In the paper, other analyses were described with their parameters. For example specific parameters were provided for expression analyses. It is unclear why such difference in transparency between methylation and expression analyses exists. In the study from Jiang and colleagues, parameters of the tools are provided, but no code or workflow was made available. Our study addressed all of the criteria mentioned, with all versions, tools, code and workflow available. It is important to note that several other criteria besides the ones in Table 1 exist to assess reproducibility such as clearly documenting and describing all the analysis steps in publications, literate programming to focus on explaining to users what we'd like the computer to do in our code (7), refraining from manual manipulation of data, recording and making available intermediate results, sharing seeds for analyses including randomness and others (2,8). These examples highlight the numerous improvements that could be applied to future polyploid studies with the ultimate goal to acknowledge and add reproducibility to the

scientific discussion, ensuring the validity of results and allow findings to stand the test of time.

Table 2: summary of whole genome studies on DNA methylation in synthetic polyploid species. The table outlines the studies (first column), the organism studied (second column), the presence of versions and parameters for tools (third and fourth column), code and workflow availability (fifth and sixth column respectively).

Study (year)	Organism	Tools		Code available	Workflow
		Versions	Parameters		
Edger *et al.* (2017)	*Mimulus peregrinus*	Yes	No	No	No
Jiang *et al.* (2021)	*Arabidopsis suecica*	Yes	Yes	No	No
Our study	*Arabidopsis kamchatica*	Yes	Yes	Yes	Yes

In practice, several tools exist to ensure and ease reproducibility that could be used in polyploid studies, meaning that educating the community about them becomes an essential step. In order to achieve this, we will provide definitions and examples of software for literate programming, virtual machines, containers and workflow managers, all important reproducibility instruments. To support literate programming (see definition above), two common tools are Jupyter (9) and knitr (10). The former offers a web-based interface to create interactive notebooks with text, data, code, equations and plots in a variety of programming languages such as python, R, Julia and Bash (11). The latter can also be used to create interactive documents through the open-source tool RStudio (12), including the same elements as Jupyter and supporting additional programming languages (13). Both tools allow users to export their notebooks to formats such as HTML and PDF that can be easily read, shared and included in publications (8). In our study, differential expression analyses are available as an R Markdown file compatible with knitr. To further ensure reproducibility, containers and virtual machines offer a supplementary layer for software and operating system control, preventing potential incompatibility issues caused by software versions and local operating systems. Virtual machines (VMs) are emulators of computer operating systems including all code, data and software needed to execute a set of analyses (8). Examples of VMs include VMWare, XenProject and VirtualBox (8,14). Containers can be seen as a lighter and shareable version of virtual machines, still including operating system, code, data and dependencies. Examples of commonly used container platforms are Docker and LinuxContainers (15,16). When the amount of tools, steps and lines of code increases, workflow management systems can help facilitate the process of making analyses

reproducible. These workflow managers allow to chain together, automate and share all of the steps required in an often laborious data analysis process, limiting the amount of manual error-prone steps (17). There are over 300 workflow managers being used or developed (18), but common ones in bioinformatics include Snakemake (19), Nextflow (20), GenPipes (21) and Galaxy (22). The first three are designed for researchers familiar with programming, while Galaxy works through a graphical user interface, allowing people with less programming experience to create complex workflows (17). Several comparisons between those managers were done, with many differences reported in ease of use, community involvement, learning resources and other criteria (17,23,24). For polyploid research, the main message is that a variety of managers exist that can suit a variety of needs from researchers.

In Chapter 1, ARPEGGIO followed and used some of the approaches and software outlined above, offering a first step towards a more reproducible future for analyses in polyploid studies, but additional features could further improve this workflow. First, ARPEGGIO offers a defined selection of tools, specifically for alignment and differential methylation analysis, even though many alternatives exist. Adding the possibility to choose different tools for some of the steps would improve ARPEGGIO's flexibility, meaning that researchers could choose tools they are familiar with or they know work best. Additionally, ARPEGGIO could offer a benchmarking system to test the combinations of tools and the consistency or accuracy of their output. Such implementation would be easier for steps such as alignment, where the amount of tools available is limited and the input format is the same, and more difficult for differential methylation analyses, where the amount of tools is large and both input and output format would need to be standardized for compatibility with other parts of the workflow. Another important feature is the addition of downstream analyses, particularly for visualizing data. For example in Chapter 2 the package METHImpute was used to compute and visualize both global methylation levels and methylation levels within and around gene bodies. Including the METHImpute pipeline in ARPEGGIO would offer a complementary view to the current differential methylation in regions approach. Other useful visualization options could be MDS plots for all samples analyzed (similar to Supplementary Figure 12), barplots showing the amount of DMRs per context, with proportion of hyper and hypo-methylation (similar to Figure 3, Chapter 2), upset plots to assess the overlap of DMRs across contexts, histograms to check the distribution of DMR lengths and raw methylation differences and donut charts to show the proportion of DMRs overlapping with different genomic functional regions (similar to Supplementary Figure 15). In addition to visualization, as seen in both Chapter 2 and 3, there's a lot of interest in homoeolog genes and their state right after polyploidization. To explore homoeologs, ARPEGGIO could include analyses summarizing the methylation state (both in the gene body and in the promoter region) for all

homoeolog pairs and the direction of change (increase or decrease in methylation). Beyond methylation, as seen in Chapter 3, ARPEGGIO could also include a workflow for transcriptomic data and a series of analyses to explore methylation and expression data together.

b. Further investigation on the role of DNA methylation in early stages of polyploidy

Current studies on DNA methylation in synthetic polyploids suffer from methodological gaps, but differences between species, tissues and ploidy level need to be also considered. An important methodological gap, not specific to polyploid studies, is related to methods to find differentially methylated regions at the whole genome level. An overview of different methods exists (25,26), but few independent comparisons of these methods has been done to assess not only the accuracy, but also specific use-cases for methods (27). Several reasons were outlined in Chapter 1 on why *dmrseq* was selected as a tool to find DMRs, but there are also limitations for this tool as well. Default parameters in *dmrseq*, used throughout Chapter 2, were optimized based on human and mouse data, focusing only on CG context, meaning that additional exploration of the parameter space would be needed to optimize *dmrseq* for plant data and different contexts. Two sets of parameters would be the focus, smoothing parameters and parameters for candidate region construction. The first set defines a bandwidth on which smoothing is applied and this bandwidth is defaulted to span at least 1000bp and covering at least 30 cytosines. The bandwidth definition might have implications for genome assemblies with numerous short scaffolds, preventing smoothing on scaffolds that are smaller than 1000bp, which is the case for both of our assemblies. The second set of parameters to construct candidate regions partitions the genome into groups with at least 5 cytosines spaced apart by at most 1000bp, with each cytosine having an amount of smoothed difference over a certain threshold (default: 0.1). For these parameters, it is unclear whether they should be optimized for each methylation context, since for example CHH context includes a much higher number of cytosines and the smoothed difference is usually smaller than CG and CHG context. To address this, plant datasets with a known ground truth in terms of methylation should be used to evaluate the effect of parameter selection. Scalability is another potential limitation of *dmrseq*. The computational time for our analyses was also relatively high considering that our two progenitors had a genome size of around 200Mb; analyses on 16 threads required ~5h for CG and CHG context and ~12h for CHH context for a total of ~22h for all three contexts for one

comparison. Considering that the estimated median genome size in flowering plants is 1.6Gb (28), *dmrseq*'s scalability could be improved to prevent analyses being possible only with considerable computational resources.

Differences in response to polyploidy levels across species are one of the main open questions in the field (3) and genome architecture could be an interesting underlying factor worth investigating, particularly transposable elements (TEs). It has been reported that a larger genome size in plants correlates with a higher proportion of TEs and a higher global methylation level (29,30). It follows that species with a larger genome and a higher proportion of TEs might have a different DNA methylation response at the whole genome level right after polyploidization compared to species with smaller genomes and fewer TEs. This hypothesis is consistent with the difference in DNA methylation changes between *A. thaliana*, with a small genome and fewer TEs, showing changes mostly associated to CG context and gene regions, *A. kamchatica*, with a larger genome than *A. thaliana*, probably more TEs and changes happening in all contexts, suggesting the involvement of both genic and intergenic elements, and *M. peregrinus* with a large genome size, more TEs and showing changes mainly in CHH context, associated to methylation changes in TE bodies. It is hypothesized that bursts of TEs right after polyploidization could be the main responsible of DNA methylation changes, leading to changes in gene expression which in the case of homoeologs genes could have a long term effect on genome fractionation (31). Additional analyses in DNA methylation in TEs in *A. kamchatica*, together with findings from other species with varying genome size such as *Brassica napus* (1.1GB), *Senecio cambrensis* (1.5GB), *Spartina anglica* (5.5Gb) and *Triticum aestivum* (17.3Gb) (32) would help shed light on the role of TEs. At lower organizational levels, tissue types, type and level of ploidy should also be taken into account. Studies in diploid *A. thaliana* revealed high tissue-specificity in methylation patterns (33,34), suggesting that synthetic polyploids might also have tissue specific responses after polyploidization. As for ploidy, mixed-ploidy species provide excellent systems to understand whether ploidy-level and type of ploidy (auto- or allopolyploidy) could lead to different genomic responses (35). Findings in autopolyploid maize indirectly supported this by investigating varying ploidy levels founding that, at the phenotypic level, higher ploidy was associated to slower growth, later flowering, reduced height and fertility (36).

Throughout Chapter 2 and 3, we've highlighted the potential of *A. kamchatica* as an exemplary system for studies in early polyploidy and additional ideas and experimental design could contribute to further our understanding on the role of DNA methylation. Of particular interest would be the variability in DNA methylation changes across different lines of *A. kamchatica.* This would include both genetically identical individuals with independent origin (different crossings from the same progenitors) and genetically distinct individuals

(different crossings from different progenitors), since Chapter 2 explored methylation patterns in one line only. No studies investigated methylation in this context, but expression studies in synthetic *B. rapa* showed a very small overlap in expression across different lines (37). Small overlaps in DNA methylation and expression across lines of *A. kamchatica* would further strengthen the link between expression and methylation and it would emphasize the importance of multiple polyploid formation as a source of variation, increasing the chances of survival and establishment. Another unique aspect in *A. kamchatica* that could investigated in future studies is its distribution in wider latitude than progenitors. With such a large ecological niche, it would be interesting to grow synthetics across an altitude gradient or different geographical locations and measure methylation changes to find some common patterns (if any), together with some phenotypic traits of interest such as flowering. Related to the experimental design from Chapter 2 (ongoing), exploring DNA methylation changes further generations down the line could also help outline the dynamics of the methylome over generations. For example, it would interesting to know whether the diverging patterns with respect to progenitors and convergence towards natural species are maintained and if yes, which parts of the genome are involved and what is their functional role.

c. Additional approaches and novel technologies to explore DNA methylation

In our study we used bisulfite treatment followed by WGBS to explore DNA methylation, providing results with high resolution, but with several shortcomings. First, bisulfite treatment is known to be harsh towards DNA and leading to large decrease in its initial amount (38), meaning that to have enough DNA, enough plant tissue needs to be available. In the case of synthetic *A. kamchatica*, leaf tissue is the most prevalent, but for other types of tissue, particularly in the case of roots or flower tissue, it could be difficult to obtain sufficient amounts without severely affecting the plant's survival. From a bioinformatic point of view, since bisulfite treatment leads to conversion of unmethylated cytosines to thymines, the resulting sequences become harder to map, leading to low mapping rates and requiring high sequencing coverage to compensate for it, leading to higher costs (39).

Several novel or alternative approaches could address or complement the limits of WGBS in the future. Methylated DNA immunoprecipitation sequencing (MeDIP-seq) offers a good trade-off between resolution, between 100bp and 300bp, and cost, approximately 50 times less than WGBS, while requiring 100 times less DNA (40,41). Few polyploid studies on whole genome methylation patterns with MeDIP-seq were done with rice (42,43), but this

method could provide an affordable opportunity to explore non-model polyploid species. Another approach that improves on many aspects related to WGBS is TET-assisted pyridine borane sequencing (TAPS). TAPS offers a bisulfite-free high-resolution detection method to improve mapping rates, increase coverage and decrease costs (44). The central part of this method relies on a borane reaction converting methylated cytosines into thymines, leading to sequences easier to map since the proportion of methylated cytosines is usually lower than unmethylated ones (44). The only limitation of this method is the reduced conversion rates for CHG and CHH context compared to CG context, over 10% less, which make it less reliable for studying these contexts in plants.

d. Environmental stress as a key component for success in newly formed polyploids

As seen in Chapter 2 and 3, environmental stress shaped methylome and transcriptome of synthetic *A. kamchatica* in different ways. The stressful conditions from our experimental design, even though relatively severe, included only few of the many abiotic stress factors found in natural conditions such as soil, water availability, wind, variable weather conditions, sunlight and others (45). Additionally, no biotic stresses were considered such as insects, animals and pathogens. Since the subset of stressful conditions from our study was already enough to observe a clear effect on methylome and transcriptome, we speculate that adding stressful elements and approaching natural conditions could strengthen the effect on transcriptome and methylome even further, leading to strong diverging patterns in polyploids formed in different natural conditions. This is consistent with the exploratory analyses for both methylation and transcription data in our two natural lines, showing very distinct patterns and leading to clearly separate clusters. Since in nature very different conditions are associated to very different geographical locations, it is worth considering whether newly formed polyploids would show meaningful differences in their genomic response even in close geographical location and similar natural conditions. We would expect this to be the case given the large variability found in our study, but further investigation is needed. Another potential consequence of additional stress would be a very strong selection on new polyploids where they would either die or adapt rapidly. Several questions arise from this scenario, particularly on how the term ‘rapid’ is defined in terms of generations. With survival studies on synthetic polyploids over generations it would be possible to assess whether many polyploids are able to propagate for several generations or there is a strong bottleneck

from the first generation(s) already. This is of importance for DNA methylation response, since it was shown to be slower than transcriptomic response. Other questions would be related to the existence of a 'core set' of rapid genomic responses, defined as a collection of responses that is constant across any situation, mostly leading to adaptation and the individual contribution of biotic and abiotic factors to this adaptation.

From an evolutionary perspective, we support the hypothesis from Van de Peer and colleagues of environmental stress as a critical element in the establishment and success of polyploids (46). We have shown how environment influenced synthetic polyploids in the short-term and, as seen above, this influence could be even stronger in natural settings. Incubation of synthetic polyploids in natural conditions would provide further insights for this hypothesis. Further examination of other -omics such as proteomics and metabolomics could widen and better frame the overall short-term genomic response and an essential future step will be to combine this genomic view to phenotype and ecology (47). Finally, with a more comprehensive understanding of short-term responses, establishing a bridge with long-term responses might reveal the reasons and mechanisms behind the omnipresence and the triumph of polyploidy in the evolutionary history of land plants.

References

1. Goodman SN, Fanelli D, Ioannidis JPA. What does research reproducibility mean? Sci Transl Med [Internet]. 2016 Jun;8(341). Available from: https://www.science.org/doi/10.1126/scitranslmed.aaf5027
2. Sandve GK, Nekrutenko A, Taylor J, Hovig E. Ten Simple Rules for Reproducible Computational Research. Bourne PE, editor. PLoS Comput Biol [Internet]. 2013 Oct 24;9(10):e1003285. Available from: https://dx.plos.org/10.1371/journal.pcbi.1003285
3. Soltis DE, Visger CJ, Marchant DB, Soltis PS. Polyploidy: Pitfalls and paths to a paradigm. Am J Bot [Internet]. 2016 Jul;103(7):1146–66. Available from: http://doi.wiley.com/10.3732/ajb.1500501
4. Edger PP, Smith RD, McKain MR, Cooley AM, Vallejo-Marin M, Yuan Y-WY, et al. Subgenome Dominance in an Interspecific Hybrid, Synthetic Allopolyploid, and a 140-Year-Old Naturally Established Neo-Allopolyploid Monkeyflower. Plant Cell [Internet]. 2017 Sep;29(9):2150–67. Available from: http://www.plantcell.org/lookup/doi/10.1105/tpc.17.00010
5. Ramírez-González RH, Borrill P, Lang D, Harrington SA, Brinton J, Venturini L, et al. The transcriptional landscape of polyploid wheat. Science (80-) [Internet]. 2018 Aug 17;361(6403):eaar6089. Available from: https://www.sciencemag.org/lookup/doi/10.1126/science.aar6089
6. Jiang X, Song Q, Ye W, Chen ZJ. Concerted genomic and epigenomic changes accompany stabilization of Arabidopsis allopolyploids. Nat Ecol Evol [Internet]. 2021 Oct 19;5(10):1382–93. Available from: https://www.nature.com/articles/s41559-021-01523-y
7. Knuth DE. Literate Programming. Comput J [Internet]. 1984 Feb 1;27(2):97–111. Available from: https://academic.oup.com/comjnl/article-lookup/doi/10.1093/comjnl/27.2.97
8. Piccolo SR, Frampton MB. Tools and techniques for computational reproducibility. Gigascience [Internet]. 2016 Dec 11;5(1):30. Available from: https://academic.oup.com/gigascience/article-lookup/doi/10.1186/s13742-016-0135-4
9. Perez F, Granger BE. IPython: A System for Interactive Scientific Computing. Comput Sci Eng [Internet]. 2007;9(3):21–9. Available from: http://ieeexplore.ieee.org/document/4160251/

10. Xie Y. Dynamic Documents with R and knitr. Chapman and Hall/CRC; 2017.
11. Team J. Jupyter for Data Science [Internet]. Available from: https://docs.jupyter.org/en/latest/use/use-cases/data_science.html
12. RStudio, PBC, Boston M. RStudio: Integrated Development for R. 2022.
13. Xie Y. Language engines - use other languages in knitr [Internet]. Available from: https://yihui.org/knitr/demo/engines/
14. Hurley DG, Budden DM, Crampin EJ. Virtual Reference Environments: a simple way to make research reproducible. Brief Bioinform [Internet]. 2015 Sep;16(5):901–3. Available from: https://academic.oup.com/bib/article-lookup/doi/10.1093/bib/bbu043
15. Merkel D. Docker: lightweight linux containers for consistent development and deployment. Linux J. 2014;2014(239):2.
16. Bernstein D. Containers and Cloud: From LXC to Docker to Kubernetes. IEEE Cloud Comput [Internet]. 2014 Sep;1(3):81–4. Available from: http://ieeexplore.ieee.org/document/7036275/
17. Wratten L, Wilm A, Göke J. Reproducible, scalable, and shareable analysis pipelines with bioinformatics workflow managers. Nat Methods [Internet]. 2021 Oct 23;18(10):1161–8. Available from: https://www.nature.com/articles/s41592-021-01254-9
18. Amstutz P, Mikheev M, Crusoe MR, Tijanić N, Lampa S. Existing Workflow systems [Internet]. 2022. Available from: https://s.apache.org/existing-workflow-systems
19. Koster J, Rahmann S. Snakemake--a scalable bioinformatics workflow engine. Bioinformatics [Internet]. 2012 Oct 1;28(19):2520–2. Available from: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/bts480
20. Di Tommaso P, Chatzou M, Floden EW, Barja PP, Palumbo E, Notredame C. Nextflow enables reproducible computational workflows. Nat Biotechnol [Internet]. 2017 Apr 11;35(4):316–9. Available from: http://www.nature.com/articles/nbt.3820
21. Bourgey M, Dali R, Eveleigh R, Chen KC, Letourneau L, Fillon J, et al. GenPipes: an open-source framework for distributed and scalable genomic analyses. Gigascience [Internet]. 2019 Jun 1;8(6). Available from: https://academic.oup.com/gigascience/article/doi/10.1093/gigascience/giz037/5513895
22. Afgan E, Baker D, Batut B, van den Beek M, Bouvier D, Čech M, et al. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. Nucleic Acids Res [Internet]. 2018 Jul 2;46(W1):W537–44. Available from: https://academic.oup.com/nar/article/46/W1/W537/5001157
23. Jackson M, Kavoussanakis K, Wallace EWJ. Using prototyping to choose a bioinformatics workflow management system. Ouellette F, editor. PLOS Comput Biol [Internet]. 2021 Feb 25;17(2):e1008622. Available from: https://dx.plos.org/10.1371/journal.pcbi.1008622
24. Larsonneur E, Mercier J, Wiart N, Floch E Le, Delhomme O, Meyer V. Evaluating Workflow Management Systems: A Bioinformatics Use Case. In: 2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) [Internet]. IEEE; 2018. p. 2773–5. Available from: https://ieeexplore.ieee.org/document/8621141/
25. Shafi A, Mitrea C, Nguyen T, Draghici S. A survey of the approaches for identifying differential methylation using bisulfite sequencing data. Brief Bioinform [Internet]. 2017;(January):1–17. Available from: https://academic.oup.com/bib/article/3064341/A
26. Robinson MD, Kahraman A, Law CW, Lindsay H, Nowicka M, Weber LM, et al. Statistical methods for detecting differentially methylated loci and regions. Front Genet [Internet]. 2014 Sep 16;5. Available from: http://journal.frontiersin.org/article/10.3389/fgene.2014.00324/abstract
27. Kreutz C, Can NS, Bruening RS, Meyberg R, Mérai Z, Fernandez-Pozo N, et al. A blind and independent benchmark study for detecting differentially methylated regions in plants. Valencia A, editor. Bioinformatics [Internet]. 2020 Jun 1;36(11):3314–21. Available from: https://academic.oup.com/bioinformatics/article/36/11/3314/5809142
28. Pellicer J, Hidalgo O, Dodsworth S, Leitch I. Genome Size Diversity and Its Impact on the Evolution of Land Plants. Genes (Basel) [Internet]. 2018 Feb 14;9(2):88. Available from: https://www.mdpi.com/2073-4425/9/2/88
29. Vidalis A, Živković D, Wardenaar R, Roquis D, Tellier A, Johannes F. Methylome evolution in plants. Genome Biol. 2016;17(1):1–14.
30. Wendel JF, Jackson SA, Meyers BC, Wing RA. Evolution of plant genome architecture. Genome Biol. 2016 Dec;17(1):37.
31. Wendel JF, Lisch D, Hu G, Mason AS. The long and short of doubling down: polyploidy, epigenetics, and the temporal dynamics of genome fractionation. Curr Opin Genet Dev [Internet]. 2018 Apr;49:1–7. Available from:

https://linkinghub.elsevier.com/retrieve/pii/S0959437X17301557
32. Pellicer J, Leitch IJ. The Plant DNA C-values database (release 7.1): an updated online repository of plant genome size data for comparative studies. New Phytol. 2020 Apr;226(2):301–5.
33. Zhang X, Yazaki J, Sundaresan A, Cokus S, Chan SW-L, Chen H, et al. Genome-wide High-Resolution Mapping and Functional Analysis of DNA Methylation in Arabidopsis. Cell. 2006 Sep;126(6):1189–201.
34. Widman N, Feng S, Jacobsen SE, Pellegrini M. Epigenetic differences between shoots and roots in Arabidopsis reveals tissue-specific regulation. Epigenetics. 2014 Feb;9(2):236–42.
35. Kolář F, Čertner M, Suda J, Schönswetter P, Husband BC. Mixed-Ploidy Species: Progress and Opportunities in Polyploid Research. Trends Plant Sci. 2017 Dec;22(12):1041–55.
36. Yao H, Kato A, Mooney B, Birchler JA. Phenotypic and gene expression analyses of a ploidy series of maize inbred Oh43. Plant Mol Biol. 2011 Feb;75(3):237–51.
37. Gaeta RT, Yoo S-Y, Pires JC, Doerge RW, Chen ZJ, Osborn TC. Analysis of Gene Expression in Resynthesized Brassica napus Allopolyploids Using Arabidopsis 70mer Oligo Microarrays. Hazen SP, editor. PLoS One [Internet]. 2009 Mar 10;4(3):e4760. Available from: https://dx.plos.org/10.1371/journal.pone.0004760
38. Tanaka K, Okamoto A. Degradation of DNA by bisulfite treatment. Bioorg Med Chem Lett. 2007 Apr;17(7):1912–5.
39. Laird PW. Principles and challenges of genome-wide DNA methylation analysis. Nat Rev Genet [Internet]. 2010 Mar 2;11(3):191–203. Available from: http://www.nature.com/articles/nrg2732
40. Taiwo O, Wilson GA, Morris T, Seisenberger S, Reik W, Pearce D, et al. Methylome analysis using MeDIP-seq with low DNA concentrations. Nat Protoc [Internet]. 2012 Apr 8;7(4):617–36. Available from: http://www.nature.com/articles/nprot.2012.012
41. Beck S. Taking the measure of the methylome. Nat Biotechnol [Internet]. 2010 Oct 13;28(10):1026–8. Available from: http://www.nature.com/articles/nbt1010-1026
42. Li X, Yu H, Jiao Y, Shahid MQ, Wu J, Liu X. Genome-wide analysis of DNA polymorphisms, the methylome and transcriptome revealed that multiple factors are associated with low pollen fertility in autotetraploid rice. Sun M, editor. PLoS One [Internet]. 2018 Aug 6;13(8):e0201854. Available from: https://dx.plos.org/10.1371/journal.pone.0201854
43. Zhang H-Y, Zhao H-X, Wu S-H, Huang F, Wu K-T, Zeng X-F, et al. Global Methylation Patterns and Their Relationship with Gene Expression and Small RNA in Rice Lines with Different Ploidy. Front Plant Sci [Internet]. 2016 Jul 21;7. Available from: http://journal.frontiersin.org/Article/10.3389/fpls.2016.01002/abstract
44. Liu Y, Siejka-Zielińska P, Velikova G, Bi Y, Yuan F, Tomkova M, et al. Bisulfite-free direct detection of 5-methylcytosine and 5-hydroxymethylcytosine at base resolution. Nat Biotechnol. 2019 Apr;37(4):424–9.
45. Gull A, Ahmad Lone A, Ul Islam Wani N. Biotic and Abiotic Stresses in Plants. In: Abiotic and Biotic Stress in Plants [Internet]. IntechOpen; 2019. Available from: https://www.intechopen.com/books/abiotic-and-biotic-stress-in-plants/biotic-and-abiotic-stresses-in-plants
46. Van de Peer Y, Ashman T-L, Soltis PS, Soltis DE. Polyploidy: an evolutionary and ecological force in stressful times. Plant Cell [Internet]. 2021 Mar 22;33(1):11–26. Available from: https://academic.oup.com/plcell/article/33/1/11/6015242
47. Fox DT, Soltis DE, Soltis PS, Ashman T-L, Van de Peer Y. Polyploidy: A Biological Force From Cells to Ecosystems. Trends Cell Biol [Internet]. 2020 Sep;30(9):688–94. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0962892420301240

Acknowledgements

A PhD without people is like a sky without stars. Everyone around me during this period made this experience unique, extremely enjoyable and truly fun. The biggest gratitude goes towards my two supervisors Dr. Rie Shimizu-Inatsugi and Prof. Mark Robinson. Their complementary view on my project was essential for its inception, its development and its blossoming. Rie has been close in supporting at all stages of the project, considering my doubts, addressing my issues and answering all the wet-lab and background questions I had. I'm also grateful for her efforts to make sure the samples and the data for this project were of the best quality and I can't even state how impressed I was with her long term planning. Mark gave me all the independence I needed to not be overwhelmed by being shared by two groups, but still managed to provide all the pros of having an extra research group: bioinformatic specific conferences and seminars, methodological lab meetings and journal clubs, social and scientific lab retreats and news from the Swiss Institute of Bioinformatics. Mark's transparent way of managing the group has inspired me several times and had me transferring some of his approaches to the Shimizu group. I'm also a big supporter of reproducibility now. Besides science, I'm very grateful for all the emotional and IT support I got during the PhD, specifically during the pandemic. Both Mark and Rie made sure I was able to work remotely (also abroad!) for a prolonged period of time without issues. I would also like to thank all members of the Shimizu group for all the nice moments we had together, especially our retreat to Lugano (never seen so much rain in my home town) and our outdoor BBQs and aperos. Thanks to Gwyneth for your company throughout my PhD, all the beers, chats and the unforgettable organization of the SOLA together with all the sport activities we did. Thanks to Naoto for being the catalyst of a great atmosphere between junior lab members, for letting me discover Lichtenstein and of course all the epic moments where beer was involved. Thanks to Epi, Kathi and Chongmeng for providing a breath of fresh air to the lab in terms of personalities, presence and humor. Thanks to Reiko for all the great scientific exchanges and all the opportunities where I learnt something new about presenting, writing and time management. Thank you for Misako for the giving me the opportunity to organize an international workshop and taking care of the onboarding and well-being of all lab members, me included. Thank you to Masa for the unvaluable bioinformatics support, the collaboration for IT tasks and your great spirit (and sake) during teaching and aperos. Thanks to Chiara for all the open exchanges about science and life, your Italian-ness was always appreciated and your motivation, dedication and determination have been inspiring. Thanks to Dario for teaching me the philosophy behind each slide in a presentation and all the scientific tips for my project. Thanks to Lucas and Aki for all the lab work behind my data, your efforts and accuracy lead to a huge amount of high-quality results and without you this project would have probably been much smaller scale. Thanks to Marcel for all the laughs, chats and Youtube suggestions. Thanks to Nangsa and Judith for all the administrative support which was always impeccable. Thanks also to Tim, Ang, Reini, Yasu, Ingrid (thank you for the tip on writing the acknowledgements as a warm up for writing other sections, greatest suggestion ever), Moeko (your company before and during the pandemic has been amazing!) all master's, bachelor's and apprenticeship students that made my stay special. Finally special thanks to Ken for providing feedback, suggestions and guidance throughout several stages of my project. Another big shutout to Isabel, the greatest institute secretary I know, dealing with issues from me without batting an eye, always ready and helpful, especially during the pandemic.

I would also like to thank all the people from the Robinson's group. Thanks to Stephany, my DNA methylation colleague from another (scientific) universe that helped me all the way throughout the project and beyond, making sure I got social interactions and venting opportunities. Thanks to Anthony, long time colleague from Lausanne and now colleague in Zurich, for all the chats, for keeping my dying French alive, all the gym sessions, all the suggestions on how to train properly, and all the beers drank, pizzas eaten and game sessions along the way. Thanks to Lukas and Fiona for their participation to our SOLA team which led to great moments. Thanks to Almut and Iza for the good discussions both in real life and on Slack, your good mood and funny, interesting posts and chats eased my days. Thanks to Stephan for all the company throughout URPP and all the funny chats. Thanks to Kathi, Charlotte, Simone, Vladimir, Reto, Elyas, Will, Helena and Pierre-Luc as well for the nice moments. Big shutout to Sabine for her impeccable work as a Robinson administrator to deal with my double institute identity.

I'm very grateful to everyone involved in the 2nd phase of URPP Evolution in Action. Shutout first to the coordination: Mira, Yvonne, Annegret and Cornelia. Your organization of events allowed an active, yet relaxed atmosphere promoting chats, networking and great scientific exchanges. I cannot express how grateful I am to have been part of such a well-organized program where I was able to interact with scientists from all over the University of Zurich, have fun, visit places both around Switzerland and Zurich, share and discuss my research and finally grow both as a person and as a scientist. Thank you to prof. Ueli Grossniklaus and prof. Beat Keller for their role as Co-Directors and bringing this program into existence. Thank you to Dr. Gregor Rot, Dr. Carla Bello, Dr. Stefan Wyder and Dr. Heidi Tschanz-Lischer for the potential bioinformatics support. Big thank you to Huyen for your open-mindedness, your support, your hilarious and adventurous stories, your suggestions and for showing me a good example on how to live life to the fullest. Thank you to Alex, Enrique, Yagmur and Felix for fueling interactions with jokes, plenty of scientific curiosity, jokes, coffee breaks full of laughs, jokes, memes and also for all the jokes. Thank you to Luca especially for that one time where we were the only one showing up for a URPP get together and ended up sharing deep life thoughts, I really liked that. Thank you Jana for the cool moments, particularly during BIO144 TA sessions where we worked as a team to help students. Thank you also to Alessia, Alexandros, Marion, Hoda and Xeniya for all the amazing URPP moments at our retreats.

Thank you to all the people from the Keller group for your company in our shared corridors. Particular gratitude to Simon and Thomas for creating and organizing our inter-group journal club sessions which contributed quite a lot in keeping my scientific critical spirit sharp. Thanks to both Thomas and Mathieu for the nice discussions about PhD life and the nice party moments. Also thanks to Glauco for bringing joy and curiosity in each one of our interactions.

There are other people from outside our corridors that I would also like to thank. All the people that collaborated with me to organize the Zurich Interaction Seminar, so Justus, Joe and Dilsad: thank you for this special opportunity and the beers and discussions (scientific and not) shared with ETH students. Thank you to Valérien for also keeping my French alive, making me laugh so much with your stories and hilarious sense of humor and the lively lunches. Thanks a lot to Tony for your helpful role in coordinating and helping evolutionary PhDs, allowing me to organize a retreat which turned out to be amazing, allowing me to eat a lot of pizza once and helping me with all the struggles related to the DissGo-StudentAdmin transition. Thank you Jana for the extremely pleasant beer sessions and encounters, your free spirit, your humor and your beer skills are a great mix that makes me always feel

welcome. Thank you to my long time friend, and now also co-author, Samuele, for all of the nice dinners together, talking about life, watching TV shows, playing a lot of entertaining videogames and all of the help to correct, optimize and extent ARPEGGIO's code.
My final thank you goes to my family and Lucia. Your support throughout my studies has been unvaluable and you've always believed in my career goals and dreams. You reminded me that growing as a person is as important as growing as a scientist, if not more. Thanks to you I made sure to be kind and helpful to everyone around me, to be a good colleague, a good friend and a good human being.

Stefan Milosavljevic
14.02.2022
Zurich, Switzerland

Curriculum Vitae

Stefan Milosavljevic

Swiss and Serbian
24.07.1993
stefan.milos.srb.ch@gmail.com | +41 76 592 24 07
Weststrasse 82, CH-8620 Wetzikon | linkedin.com/in/stefan-milosavljevic

Work experience

10.2017 – 03.2022 **Scientific Researcher and Bioinformatician**
University of Zurich, Institute of Evolutionary Biology and Environmental studies

- Analyse big high-throughput sequencing data
- Develop automated data analysis workflows in Python
- Interpret and visualize data in R
- Present advanced research to public with various expertise

10.2017 – 12.2021 **Teaching Assistant in Programming, Statistics and Bioinformatics**
University of Zurich

- Manage, explain and summarize data in R
- Prepare practical bioinformatics exercises on cutting edge research
- Teach basic and advanced statistics applied to biological research
- Critical thinking on use and limitations of statistical approaches

Education

10.2017 – 06.2021 **PhD candidate in Evolutionary Biology**
University of Zurich, Institute of Evolutionary Biology and Environmental studies

- Standardize and streamline (big) data analysis processes
- Keep updated with method advances and scientific discoveries
- Publish on peer-reviewed scientific journals
- Collaborate across disciplines and continents
- Optimize independent time and project management

09.2015 – 09.2017 **MSc in Computational Biology and Bioinformatics**
ETH Zurich, Department of Biosystems Science and Engineering

- Programming in C++, R and Python
- Learn quantitative modelling, simulations and optimizations
- Use computational statistics, data mining and machine learning
- Thesis about machine learning applied to time gene expression data

09.2012 – 08.2015 **BSc in Biology**
University of Lausanne, Faculty of Biology and Medicine

Extracurricular activities

01.2019 – 12.2021 **Project initiator**, ETH Library
In collaboration with a colleague:
- Define and successfully pitch an idea in the field of knowledge
- Evaluate user needs and potential implementations
- Transfer project management to ETH Library after promising results

09.2019 – 09.2020 **Organizer for the Zurich Interaction Seminar (ZIS)**, University of Zurich
- Find and recruit speakers across disciplines and departments
- Schedule and manage dates, speakers and location

05.2019 **Public speaker for Pint of Science**, Zurich
- Present a complex topic to the public
- Engage pro-actively with the audience

01.2018 **Organizer of the Zurich-Kyoto International Workshop: Plant Development and Evolution**, Zurich
- Manage and guide international speakers and guests
- Organize venue, program, hotels, catering and sessions

Awards

05.2014 **Finalist at the Intel Science and Engineering Fair 2014**, Los Angeles
- High-school project on medical uses of an invasive plant
- Represent Switzerland as a young scientist

Languages

Serbian and Italian: native English and French: fluent German: conversational

Computer skills

Programming in Python (advanced), R (advanced), C++ (basic), bash (basic). Familiar with descriptive, computational and inferential statistics. Experience with unsupervised machine learning (Self-Organizing Maps), workflow development, big data analysis, cluster computing. Windows, MacOS and Linux operative systems.

Hobbies

I like quiet spaces and the sound of nature. I enjoy reading books. Especially fond of thrillers (Dan Brown), sci-fi (Asiimov) and humor (Douglas Adams). Fan of cooperative or team-based videogames to spend time with friends. I love cooking and experimenting with Eastern recipes, especially Balkan, Indian and Japanese. Sport enthusiast, 10k-20k runner (next goal: marathon), road biking to explore new places, bouldering, swimming beginner, badminton and fitness.