Investigating DNA methylation reprogramming in plant germlines

Judit Tálas

Department of Cell and Developmental Biology
John Innes Centre

University of East Anglia

This dissertation is submitted for the degree of

Doctor of Philosophy

September 2024

Abstract

Epigenetic reprogramming is an essential part of reproduction across both animal and plant kingdoms. Although the two kingdoms have diverged more than 1.5 billion years ago, they both have convergently evolved systems that protect the integrity of the genome throughout generations. This is in part due to the threat of mobile transposable elements, which when uncontrolled can lead to genomic instability through insertions and rearrangements. Whilst animals reset DNA methylation during germline development, plants largely maintain symmetric CG and CHG DNA methylation. However, asymmetric CHH methylation undergoes tightly regulated reprogramming during male sexual lineage development in plants.

This thesis investigates the dynamics of epigenetic regulation during plant reproduction in two distinct systems: *Arabidopsis thaliana* and *Marchantia polymorpha*. In *Arabidopsis*, the expression patterns of key RNA-directed DNA methylation (RdDM) components was examined, alongside exploring the small RNA (sRNA) profiles post-meiosis during the development of the male sexual lineage. The results reveal a complex spatiotemporal regulation of RdDM factors and a shift away from CLSY-dependent sRNA biogenesis after meiosis, potentially involving non-canonical RdDM pathways.

In the basal land plant *Marchantia polymorpha*, recent studies have revealed the presence of N4-methylcytosine (4mC) in the male germline, an epigenetic modification previously thought to exist only in prokaryotes. During spermiogenesis, 4mC is deposited globally at CG sites, yet after fertilisation, the mature embryo loses this mark, and the knockout of 4mC leads to accelerated embryonic development. This study further explores the fate of paternal 4mC in early *Marchantia* embryos and provides evidence that 4mC is actively removed before pronuclear fusion. Further investigation reveals that the loss of 4mC accelerates the onset of the first nuclear division and imposes a partial fertility cost, ultimately impacting long-term embryonic development.

Table of contents

List of figures

List of tables

Declaration

Throughout this thesis, I present both my original work and contributions from others. Where data from other sources is included, appropriate references are provided. A summary of the lines and libraries, either generated by myself and colleagues in the lab or sourced from published literature, can be found in Tables A1 and B2.

Judit Tálas
September 2024

Acknowledgements

Throughout the process of writing this thesis, I had the opportunity to reflect on these last 4 years and how thankful and privileged I am for having the people around me that made this all possible. First, I would like to thank the John Innes Centre and the NRPDTP programme for providing the perfect environment to carry out this research. I felt so spoilt by all the support and guidance offered by Lab Support, Bioimaging, Horticultural Sevices and Research Computing, thank you all!

I want to thank Christine Faulkner for offering insightful, impartial feedback on my PhD, whilst also being encouraging, realistic and empathetic. To my supervisor Xiaoqi Feng - thank you for always sharing my excitement for my project, allowing me to take the reign in exploring topics that excited me and for your unwavering faith in my abilities, both personal and scientific.

I am also deeply grateful to the members of the Feng lab, past and present, from whom I have learnt so much. To the lovely postdocs Yalin and Shaoli, it was so much fun to work with you and your work ethic and enthusiasm never fails to impress me. A big thank you to Shujuan and Chris, who always looked out for me after the lab moved abroad, your presence in the lab made all the difference. You were both always so approachable, easygoing and were always up for helping me or just having a much-needed pep talk or chat. I owe a special thanks to Jincheng Long who taught me everything I know about sRNAs and techniques in the lab - you were an amazing mentor in my first two years and so much fun to be around! I also want to thank Martin Vickers for making bioinformatics approachable and enjoyable — it became the highlight of my PhD, thanks to you.

To my adoptive lab, the Carella Lab - Hetty, Hyeonmin, Bea, Ellie, Sam and Kayla - you are all amazing! You made me feel so welcome in the lab, and I will miss our lunches and silly (and serious) conversations. I owe a massive thanks to Phil Carella - thank you so much for agreeing to be my supervisor for the final part of my PhD. Your kindness, support, and calm presence made all the difference.

I want to thank my friends, Sophie, Deirdre, Cat, Joe and especially Emily, you have been there through the ups and downs of these four years, I feel so privileged to have found friends like you. Your kindness, humour, enthusiasm and infectious passion for science has made living in Norwich around you amazing. I will miss our dinner parties, coffees and hangouts so much. I'm so proud of all of you for reaching the end of your PhDs!

Köszönöm a családomnak: Papa, Mama, Andris, Zsófi, Zsuzsi, Bogesz, Dorka, Balu, Zsuzsa, Nagyi és Nagypapa köszönöm a sok segítséget, bíztatást és támogatást, nélkületek nem sikerült volna! Nagyon szeretlek titeket!

To Johnny, thank you for being my cheerleader, you have been the most important rock throughout these 4 years. Thank you for believing in me and for moving to Norwich, so I could continue to pursue this degree. Thank you for cooking for me, and offering distractions when I needed them most.

Lastly, I want to give a huge thank you to Jimmy. You have been the most amazing, patient and invaluable mentor and friend throughout my PhD. The amount of dedication, kindness and enthusiasm you showed towards my project is absolutely incredible, even after you moved away to the other side of the pond. I am so grateful for your guidance and friendship. You made it possible for me to see this project through, and I truly couldn't have completed this degree without you.

Thank you all!

Chapter 1

Introduction

1.1 Introduction to epigenetics in prokaryotes, animals, and plants

Epigenetics refers to a "stably heritable phenotype resulting from changes in a chromosome without alterations in the DNA sequence" [1]. In eukaryotic organisms, chromatin organisation plays a critical role in various cellular processes, including differentiation, development, response to environmental stimuli, and the maintenance of genome integrity. Chromatin consists of DNA and specialised proteins called histones, which together form nucleosomes. Histones exist in multiple variants and can undergo various modifications that influence chromatin structure, accessibility, and gene expression [2, 3]. The addition of a methyl group to the 5th carbon of cytosine results 5-methylcytosine (5mC) which is an example of DNA methylation - another epigenetic modification. In prokaryotic organisms, DNA methylation was originally identified as a component in the restriction-modification pathway wherein methylated sequences are recognised as self, and unmethylated sequences are targeted for cleavage by restriction enzymes, thus safeguarding the genome against invading unmethylated phage DNA [4, 5].

In eukaryotic systems, including plants and animals, DNA methylation functions oppositely: invading transposable elements (TEs) are silenced by DNA methylation. In plants, 5mC is the most prevalent form of DNA methylation, primarily occurring in heterochromatic regions, though methylated loci can also be found in genic regions [6, 7]. DNA methylation can directly suppress transcription and gene expression by hindering the binding of transcription factors [8]. One of the earliest evidences for epigenetic modifications in plants was discovered in the *Arabidopsis thaliana* gene *SUPERMAN* (*SUP*). Mutations in *SUPERMAN* affect normal floral development. Interestingly, some alleles exhibited the mutant phenotype without any alterations in the DNA sequence. The *clark kent* epialleles were uncovered when it was discovered that the promoter sequence of *SUP* was methylated. In addition, SUP expression correlates with the extent of *SUP* promoter methylation [9]. This discovery was among the first to highlight the role of DNA methylation in controlling transcriptional expression in plants. Since then, the field of plant epigenetics has significantly expanded, uncovering a myriad of complex mechanisms by which DNA methylation and other epigenetic modifications regulate gene expression, development and protect genome integrity throughout generations.

1.2 Transposable elements

The size of different plant genomes is highly variable and not necessarily related to their complexity. Among eukaryotes, flowering plants display the most extreme genome size variation, with differences of up to 2400-fold; for instance, the genome of *Genlisea aurea* is just 63.6 megabases (Mb), while that of *Paris japonica* reaches a staggering 150 gigabases (Gb) [10, 11]. TEs, which are selfish and repetitive DNA sequences, persist in genomes by copying and inserting themselves into various genomic locations [12]. TEs are also essential for driving genome evolution: they can regulate gene expression, unequal crossover, recombination rate and in general contribute to genomic variation [13]. TEs have had some prominent stabilising impacts on chromosomes, playing a vital role in the evolution of repetitive structures including the centromere and telomeres [14].

However, the majority of TE activity tends to be harmful, posing risks to genome integrity and often being disadvantageous to the organism [13]. Transposable elements can cause heritable disruption to a vast range of cellular functions by insertion into promoter, regulatory or genic regions. For example, the mobilisation of the *CAC1* TE results in the disruption of the *DWARF4* gene and severe growth defects in *Arabidopsis* [15].

TEs are categorized into two classes: Class I (retrotransposons) and Class II (DNA transposons) TEs. Class I TEs are transcribed via an RNA intermediate, which then undergoes reverse transcription before insertion at another site in the genome. The reverse transcriptase required for this process is usually encoded within the transposable element itself. One of the largest families of Class I TEs are long-terminal-repeat (LTR) retrotransposons. Some of these LTRs (flanking the TE) have acquired transfer-RNA (tRNA) fragments, which can be used as binding sites for host tRNAs. Subsequently, the clover leaf structure of the tRNA can be recognised as a primer to initiate reverse transcription and transposition. To combat this, some hosts cleave tRNAs into tRNA fragments (tRFs) that compete for these LTR binding sites, thereby silencing these retrotransposons [16–18].

Conversely, Class II TEs encode transposase enzymes that excise the DNA segments and insert them elsewhere in the genome. Both Class I and Class II TEs can be further classified as either autonomous or non-autonomous elements. Autonomous elements encode all the necessary machinery for their own mobilisation, whereas non-autonomous elements lack some or all of these factors and rely on trans-acting transposases to facilitate their transposition [19, 20].

In addition to disrupting the function of individual genes, the excision and insertion of TEs can destabilise genome structure [13]. The repetitive nature of TEs can cause non-allelic homologous recombination during meiosis, leading to reproductive defects and the formation of dicentric or acentric chromosomes [21].

The contribution of TEs to genome size varies massively: while they comprise only around 21% of the *Arabidopsis* genome [22], crops with larger genomes have gone through significant TE amplification typically dominated by a few TE families [23, 24], resulting in an 82% and 85% genomic TE content of wheat and maize respectively [23, 25]. The importance of TEs in genome regulation is further exemplified in maize, where disruption of DDM1, a key chromatin remodeler required for the genome-wide maintenance of DNA methylation, results in embryo lethality [26]. Thus, whilst mutations of the components of the pathway for establishment and maintenance of DNA methylation have no apparent phenotypic effects in *Arabidopsis*, it offers an important model system for exploring the underlying mechanisms of genome-wide epigenetic regulation.

1.3 Chromatin Structure

The regulation of gene and TE transcription is essential for both prokaryotic and eukaryotic organisms. In eukaryotes, DNA is organised into chromatin, with the nucleosome serving as the basic unit. In a nucleosome, DNA is wrapped around histone octamers, which consist of two H2A-H2B heterodimers and one H3-H4 heterotetramer, along with the linker histone H1, altogether encompassing 147 base pairs of DNA [27]. Nucleosomes are further organised into dynamic higher-order chromatin structures, which mediate chromatin accessibility and arrange the genome into more open, gene-rich euchromatic regions and more condensed heterochromatic regions, typically located around chromosome ends or pericentromeric regions in *Arabidopsis* [28].

Chromatin structure is regulated through histone modifications, such as acetylation, methylation, phosphorylation, or ubiquitination, which convey chromatin states and interact with various pathways that promote or repress chromatin accessibility and condensation. For instance, trimethylation of lysine 27 of histone H3 (H3K27me3) and di- or trimethylation of lysine 9 on histone H3 (H3K9me2 or H3K9me3) are associated with repressed chromatin states. In contrast, H3K27 acetylation (H3K27ac) is associated with active promoters, as is histone H3 lysine 4 trimethylation (H3K4me3).

Additionally, trimethylation of lysine 36 on histone H3 (H3K36me3) is correlated with active transcription in gene bodies [29, 30].

There are also specific histone variants that are either highly conserved across eukaryotes or unique to particular systems. For instance, H2A.Z is conserved across eukaryotes and is enriched in euchromatic regions, where it facilitates the recruitment of RNA Polymerase II in yeast, plants, and animals [31, 32]. In contrast, H2B.8 is specific to the sperm cells of *Arabidopsis* and plays a key role in mediating sperm chromatin condensation [33].

1.4 DNA methylation maintenance and the RdDM pathway

The existence of epigenetic marks implies that there are mechanisms to read, write and erase them. Cytosine methylation exists in one of three sequence contexts: the symmetric CG and CHG and the asymmetric CHH (where H = A, T or C). This contrasts with mammals, in which most methylation is directed to CG sites and is not inherited by the next generation [6].

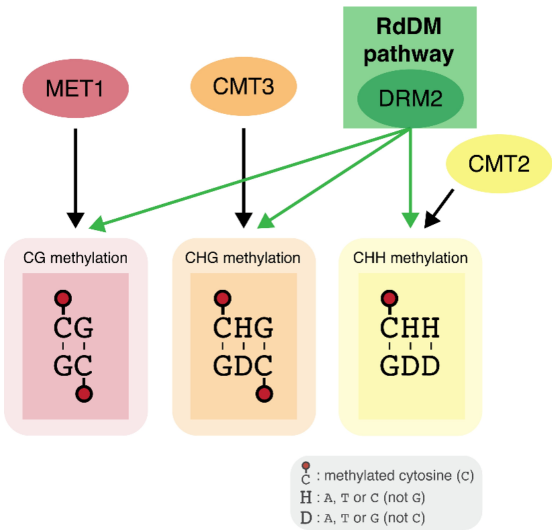


Fig. 1.1 DNA methylation sequence contexts and their associated DNA methyltransferases (Figure from [34])

RNA-directed DNA methylation (RdDM) is a pathway that can methylate cytosine in any sequence context [35]. In plants, it is also the only pathway responsible for the *de novo* methylation of previously unmethylated regions. During DNA replication, modifications at symmetric CG and CHG sites result in a hemimethylated state, where only one strand of the daughter DNA retains methylation. Full methylation can be restored by the highly conserved maintenance methyltransferases METHYLTRANSFERASE 1 (MET1, a homolog of DNMT1) and CHROMOMETHYLASE 3 (CMT3) (Figure 1.1) [34]. CHH sites are maintained by CMT2, however, due to the asymmetry of CHH methylation, methylation information may be lost during replication, necessitating the recruitment of RdDM for both *de novo* methylation and maintenance. Therefore, to sustainably combat the detrimental effects of TE mobilisation, plants have evolved ways to keep transposon activity tightly controlled through RdDM. Additionally, the robustness of the system is enhanced by the fact that CMT2/3 and SET domain histone methyltransferases (KYP, SUVH5/6) maintain non-CG methylation and H3K9me2 in an interdependent manner. Specifically, the chromo and BAH domains of CMT3 interact directly with H3K9me2, allowing it to recognise other heterochromatic marks and establish a positive feedback loop [35].

Following the sequencing of the *Arabidopsis* genome, two unexpected variants of RNA Polymerase II were discovered: RNA Polymerase IV (Pol IV) and RNA Polymerase V (Pol V) [36]. Through these two variants, the RdDM pathway generates small RNAs (sRNAs) in heterochromatic regions and is able to target *de novo* DNA methylation to these regions. The Pol IV-associated pathway initiates through recruitment via H3K9me2 associated chromatin remodeller SAWADEE HOMEODOMAIN HOMOLOG 1 (SHH1) [37–40]. Pol IV works in concert with chromatin remodellers CLASSY 1-4 (CLSY 1-4) to transcribe a short single-stranded RNA precursor [41], which is then extended into double-stranded RNA by RNA-dependent RNA polymerase 2 or 6 (RDR2 or RDR6) [34, 35]. This double-stranded RNA is processed by DICER-LIKE 3 (DCL3) into 24-nucleotide small interfering RNAs (24nt siRNAs), which are subsequently loaded onto specific ARGONAUTE proteins (AGO4/5/6/9) [35] (Figure 1.2). This complex interplay between DNA methylation and histone modifications ensures that patterns of DNA methylation are stably inherited through mitosis, maintaining consistent gene expression and genome stability across cell divisions.

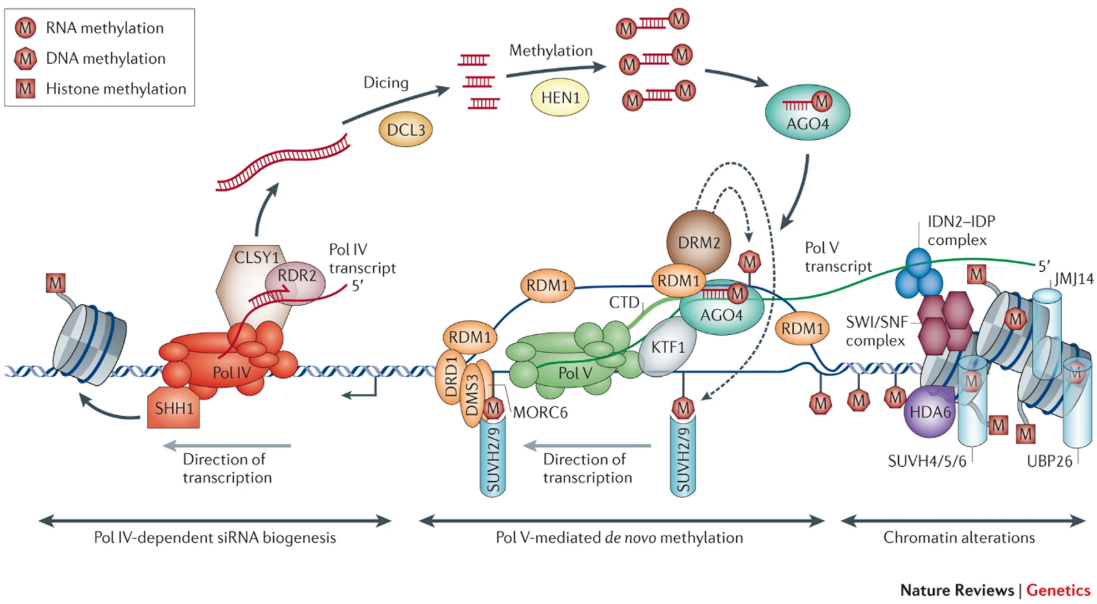


Fig. 1.2 An overview of the canonical RdDM pathway (Figure from [35])

Independently, Pol V is recruited by the DDR complex composed of chromatin remodellers DEFECTIVE IN MERISTEM SILENCING 3 (DMS3), DEFECTIVE IN RNA-DIRECTED DNA METHYLATION 1 (DRD1) and RNA-DIRECTED DNA METHYLATION (RDM1), along with SUPPRESSOR OF VARIEGATION 3–9 HOMOLOG 4/5/6 (SUVH4/5/6) methyl-DNA-binding proteins. The Pol V complex transcribes a single stranded RNA scaffold [35]. The siRNA-AGO complex binds to this complementary scaffold RNA, and the AGO hook motif of the Pol V-associated SPT5-LIKE/KOW DOMAIN-CONTAINING TRANSCRIPTION FACTOR 1 (SPT5L/KTF1) protein subsequently binds to AGO4. NUCLEAR RNA POLYMERASE E 1 (NRPE1), the largest subunit of Pol V, and the *de novo* methyltransferase DOMAINS REARRANGED METHYLTRANSFERASE 2 (DRM2) then associate with AGO4, catalysing DNA methylation (Figure 1.2) [6, 42, 43]. As discussed earlier, the maintenance of CHH sites depends on continuous *de novo* methylation mediated by DRM2.

Recent studies have shown that the CLASSY (CLSY) family of chromatin remodelling factors regulates tissue-specific DNA methylation in *Arabidopsis thaliana*, chiefly through the RdDM pathway. 5mC sequencing and sRNA sequencing across various somatic and germline tissues revealed distinct profiles of sRNAs and DNA methylation, driven by the tissue-specific expression of CLSY proteins. This provides a model where the unique expression profiles of chromatin remodellers shape tissue-specific DNA methylation [44].

1.5 Epigenetic reprogramming in plant sexual reproduction

In mammals, primordial germ cells are sequestered early in development and go through genome-wide demethylation during development and following fertilisation. This results in the resetting of global epigenetic marks as well as restoring pluripotency [45]. Subsequent *de novo* re-methylation is mediated by methyltransferase Dnmt3 as well as germline specific PIWI-interacting small RNAs (piRNAs). The loss of the piRNA pathway results in transposon mobilisation, erroneous transcription regulation and reduction of fertility [46–48].

In the male lineage of angiosperms, the germline initiates from totipotent cells in the shoot apical meristem, giving rise to a diploid meiocyte. The meiocyte, surrounded by a layer of somatic nurse cells (the tapetum), undergoes meiosis, giving rise to a tetrad of microspores. The four microspores subsequently undergo mitotic division, to each make a vegetative cell (the somatic companion) and a generative cell, which divides further to make two sperm cells (Figure 1.3) [49, 50] which forms the mature pollen grain.

AGO proteins play crucial roles in sRNA pathways during reproduction. Recent findings have shown that during male meiosis, AGO1, AGO2, and AGO5 — proteins

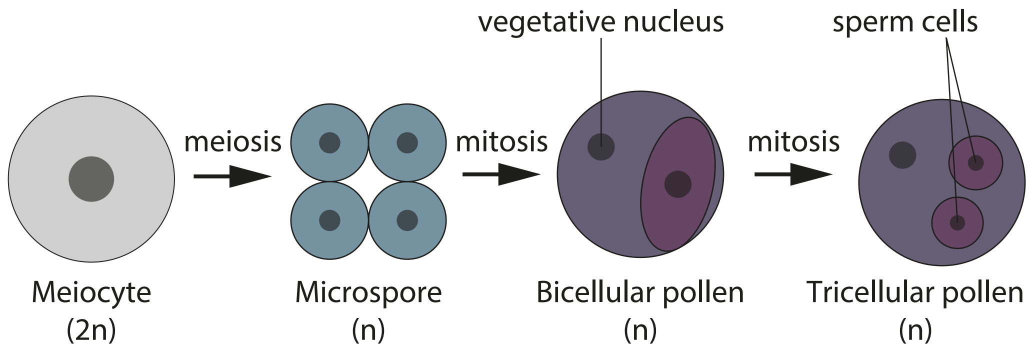


Fig. 1.3 An overview of the development of the male sexual lineage in *Arabidopsis*. Figure adapted from [50]

[50]

involved in post-transcriptional gene silencing (PTGS) — are localised in either the nucleus or cytoplasm, whereas AGO4 and AGO9 — associated with transcriptional gene silencing (TGS) — are found exclusively in the nucleus. [51]. This study also provided an overview of the expression and localisation of several AGOs throughout meiosis, underscoring their dynamic expression and critical roles in germline development. Additionally, the presence of AGO1 and AGO5-bound micro RNAs (miRNAs) in meiocytes suggests that the miRNA pathway may be involved during meiosis [51]. Furthermore, AGO5 and AGO9 have recently been shown to mark reproductive cells, similar to the early segregation of germline cells in animals [52].

The RdDM pathway and sRNAs play crucial roles not only in germline development in *Arabidopsis*, but also in other plant species. In maize, 21- and 24-nucleotide phased small interfering RNAs (phasiRNAs) are produced in the anthers and are essential for proper anther development [53, 54]. In sorghum, miRNAs were found to be differentially expressed before and after meiosis [55]. In *Arabidopsis*, cucumber, and soybean, several miRNAs are conserved and preferentially expressed in meiocytes, though their overall abundance is lower compared to somatic tissues [56]. In rice, the ARGONAUTE protein OsAGO18 is specifically expressed in meiocytes, where it facilitates the accumulation of miRNA-triggered secondary siRNAs and regulates genes involved in pollen grain development [57]. Additionally, the maintenance of CHH methylation through the RdDM pathway—particularly via RDR2, a critical component of RdDM—is necessary for both male and female sexual development in rice [58]. Moreover, rice sporogenesis depends on the degradation of the germline-specific ARGONAUTE protein MEL1, which prevents off-target effects from rogue phasiRNAs; disruption of this process results in a semi-sterile phenotype [59].

Recent work identified male sexual lineage specific DNA methylation reprogramming in the *Arabidopsis* germline, which will be discussed in detail in Chapter 2. Briefly, male meiocytes maintain high levels of CG and CHG methylation when compared to somatic canonical RdDM (cRdDM) loci levels, and a relatively low level of CHH methylation. However, in the CHH context, there is distinct hypermethylation at selected, sexual lineage specific loci (HyperTEs), including novel targets of RdDM activity. Additionally, *de novo* gene-associated methylated loci (MetGenes) were also identified, including a meiosis-specific gene that was mis-spliced when methylation was disrupted [50]. Recent evidence suggests that the biogenesis of these 24-nt sRNAs occurs in the somatic nurse cell layer surrounding the meiocytes, known as the tapetum [60].

Similarly, in the female germline, the surrounding somatic tissue synthesises RNA polymerase IV-dependent 24-nt sRNAs from siren (siRNA in the endosperm) loci [61, 62, 44]. Like the tapetal nurse cell-derived sRNAs that catalyse DNA methylation in male meiocytes, these siren loci-derived sRNAs are also dependent on CLSY3 and CLSY4 [44]. However, there is little overlap between the loci targeted by these two sets of sRNAs, and even less between siren loci across different species [62]. Despite their distinct origins, siren loci-derived sRNAs in *Brassica rapa* similarly catalyse DNA methylation, predominantly at protein-coding genes, thereby regulating gene expression in the ovule [63]. Furthermore, it has been recently shown that 24nt sRNAs accumulate in rice zygotes, which overlap gene rich, distinct loci from paternally derived sRNA loci [64], highlighting the conserved nature of 24 nucleotide sRNA driven DNA methylation reprogramming in both the male and female sexual lineage. The importance of RdDM in embryonic development is further underpinned by a recent study that showed embryos resulting from a maternal RdDM mutant cross result in incorrect gene expression in the endosperm and a reduction in seed viability [65].

1.5.1 Illumina library preparation and sequencing by synthesis

Over the past two decades, next-generation sequencing (NGS) technologies have transformed biological sequencing by dramatically improving speed, accuracy, and data output. These advancements enable both high-throughput and low-input sequencing while maintaining exceptional fidelity.

Illumina sequencing technology requires constructing a sequencing library tailored to the specific sample, by either converting genomic DNA (gDNA) or complementary DNA (cDNA) into a fragmented library for sequencing. Various workflows are available to accommodate different throughput requirements and starting materials. These include on-bead tagmentation [66], which reduces preparation time, AmpliSeq amplicon library preparation [67], and the widely used adapter ligation workflow [68].

The construction of an Illumina sequencing library using adapter ligation typically involves several key steps (an overview is presented in Figure 1.4 [69]). For DNA sequencing, the process begins with DNA extraction. In the case of plant material, this may involve a CTAB extraction [70] using chloroform to obtain ultra-pure DNA, free of secondary metabolites and proteins, though various other DNA extraction methods and kits can also be used. For RNA sequencing, an additional step is required after extraction, namely reverse transcription to convert RNA into complementary DNA (cDNA) [69, 71].

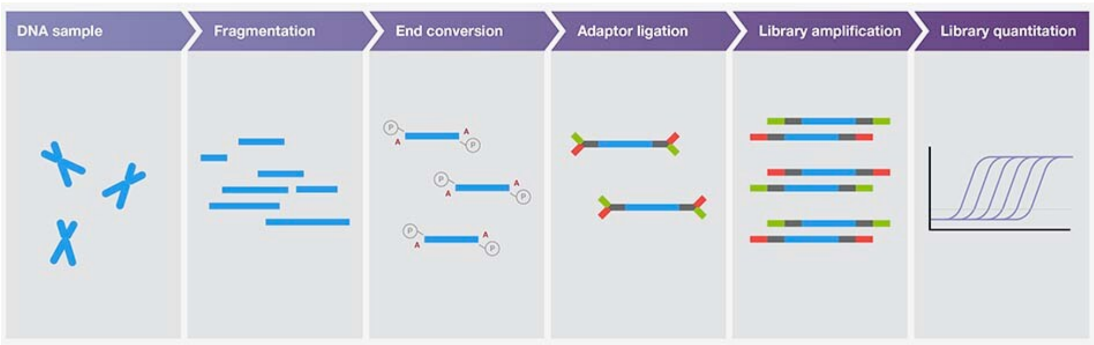


Fig. 1.4 An overview of the steps involved in library preparation for Illumina sequencing [69]

Once extracted, gDNA must be fragmented, as Illumina sequencing is optimised for relatively short sequences, typically in the 300–600 bp range. This fragmentation can be achieved through either mechanical shearing or enzymatic digestion. Mechanical shearing, often performed using high-frequency acoustic sonication, breaks the DNA's phosphodiester backbone. The optimal force and duration of sonication vary depending on the instrument and experimental needs and should ideally be determined experimentally [69].

Following an end repair step, where the sheared DNA fragments are repaired (polymerase activity) or blunted (exonuclease activity) and the ends phosphorylated, adapters are ligated to the ends of the fragments. Adapters are a pair of oligonucleotides which facilitate the clonal amplification step used during Illumina sequencing. The adapters are normally made up of 3 sections. Firstly, a complementary end which forms the adapter duplex, this is annealed to the ends of the insert. Secondly, the adapters contain unique index sequences which can be used to demultiplex several samples that were sequenced together on the same flow cell. Finally, the adapter sequences contain a non-complementary P5 and P7 end (so-called due to their binding site to the flow cell), which prevents self-ligation [69, 71].

After the adapter ligation step, PCR amplification may follow, which allows for a lower sample input, however it can introduce GC bias, polymerase errors, chimeric sequences or amplification bias, which might hinder SNP discovery and *de novo* genome assembly. Therefore, given an amount of starting material, the lowest possible cycles of PCR must be performed to prevent these effects [69, 71].

After fragmentation, adapter ligation and PCR amplification, a size- selection and clean-up step is usually performed, to ensure uniform library fragment size and remove excess unligated adapters or adapter dimers. The clean-up step may be performed using agarose or sometimes polyacrylamide gels, or more commonly using magnetic beads which can bind and elute DNA in a size-specific manner depending on the bead to sample DNA ratio. These steps help to increase data output, reduce noise and ensure the quality of the data output (as Illumina sequencing has an optimal insert length at which the sequencing is most accurate) [69, 71].

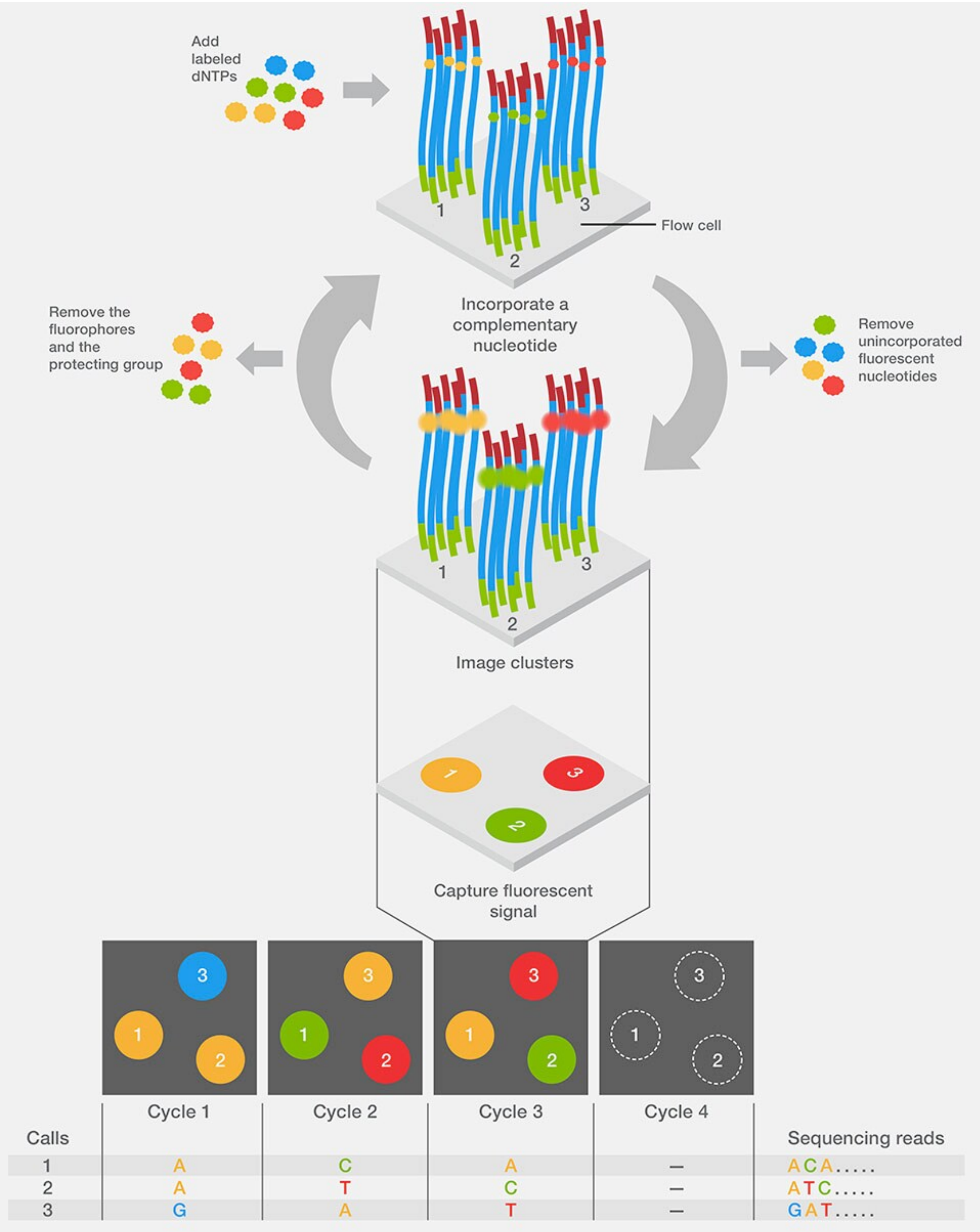


Fig. 1.5 An overview of sequencing by synthesis [72]

Following an accurate library quantification step, the Illumina libraries are ready to be sequenced. Briefly, the adapter ligated DNA fragments are flushed across the flow cell where they hybridise to immobilised complementary primer sequences on the flow cell. The fragments are clonally amplified using bridge amplification to form clusters of identical reads across the flow cell. The forward reads are sequenced by priming and extending the first read using fluorescently tagged dNTPs, which compete to be incorporated into the growing chain. After each cycle of the incorporation of the fluorescently tagged base, the clusters are excited by a light source and the fluorescent signal is recorded (an overview of how bases are sequenced is shown in Figure 1.5 [72]). In paired-end sequencing, following the sequencing of the forward read, the read product is washed away, the fragments are extended once again and the original forward read is cleaved off, followed by sequencing of the reverse read in a similar manner as before [72].

1.5.2 Detecting methylated cytosines in NGS sequencing

To characterise the methylation profile of a sample, various treatment protocols can be applied prior to Illumina library preparation to facilitate sequencing-based methylation analysis. These use specific chemistries or enzymatic activities to selectively alter or protect methylated bases.

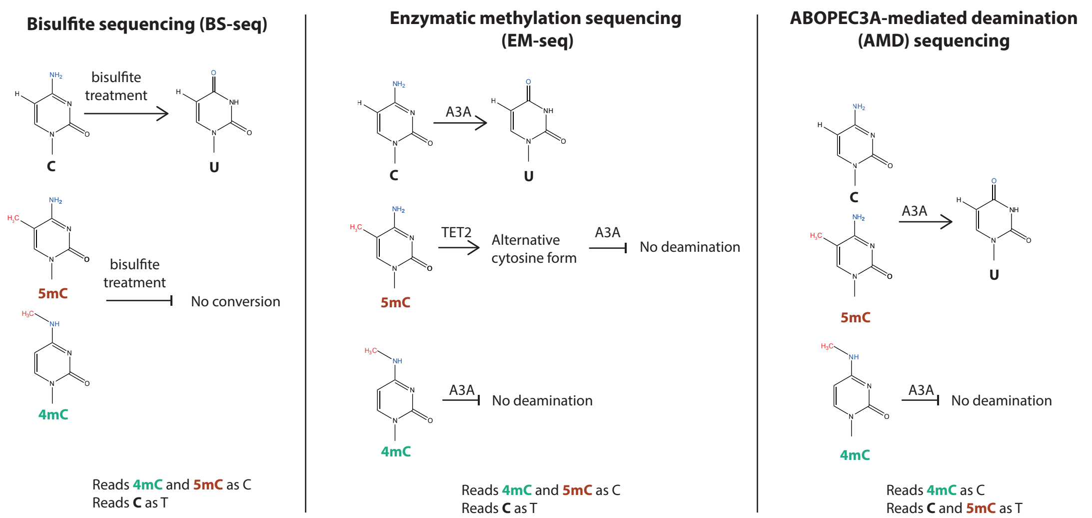


Fig. 1.6 An overview of the chemistries of modified base sequencing: BS-seq, EM-seq and AMD-seq.

The most widespread sequencing method utilised to sequence methylated DNA is bisulfite sequencing. The DNA fragments are treated with sodium bisulfite, which (through a series of intermediate steps) selectively deaminates unmethylated cytosines, resulting in uracil. At the same time, due to the structural changes introduced by the methyl group in 5-methylcytosine and N4-methylcytosine, these modified bases are protected from deamination (first panel, Figure 1.6). Consequently during sequencing, the unmethylated cytosines that were converted to uracils are read as thymines. In contrast, the methylated bases remain as cytosines, thereby enabling their identification by aligning bisulfite sequencing data to reference genomes using specialised bioinformatics tools, such as Bismark [73]. Bisulfite sequencing methods are robust and have also been developed for single-cell or very low input sequencing [74]. The downsides of bisulfite sequencing are that they do not distinguish between the methylated bases (4mC or 5mC) [75] and that bisulfite sequencing itself can introduce DNA damage due to the protocol requiring extremes of pH and temperature, resulting in depyrimidination of DNA [76].

Alternative enzymatic methods have been developed to combat some of the challenges of bisulfite sequencing, namely APOBEC3A-mediated deamination sequencing (AMD-seq) [77], which allows for reading 4mC as Cs specifically, as well as enzymatic methylation sequencing [78] (EM-seq) which is a direct alternative to bisulfite sequencing, reading both 5mC and 4mC as Cs during sequencing, but without significant DNA damage.

Both EM-seq and AMD-seq rely on the ability of APOBEC3A, which is an enzyme originally described in primates [79], to deaminate a variety of cytosine forms to uracil. Namely, APOBEC3A can deaminate both unmodified cytosines and 5mC. However, 4mC is protected from deamination by APOBEC3A, so following enzymatic treatment, an Illumina sequencing library is constructed and all unmethylated cytosines as well as 5mC will be read as Ts during sequencing, and only 4mC will be read as Cs [77] (Figure 1.6 third panel). EM-seq introduces an additional step, whereby prior to ABOBEC3A treatment, the Tet methylcytosine dioxygenase 2 (TET2) enzyme specifically converts 5mC sites to alternative cytosine forms which are not substrates of APOBEC3A. As a result, similar to traditional bisulfite sequencing, in EM-seq, both 5mC and 4mC are be read as Cs while unmethylated cytosines are be read as Ts (Figure 1.6 second panel).

1.6 Thesis outline

This thesis explores the dynamics of RdDM and DNA methylation after meiosis in the male sexual lineage of *Arabidopsis* and in the embryo of *Marchantia*.

In Chapter 2, the expression of CLSY chromatin remodellers, Pol IV, and Pol V in post-meiotic anthers of *Arabidopsis* is investigated. This is followed by the sequencing and comparison of sRNA profiles from microspores, sperm cells, and sperm nuclei with other published germline and somatic datasets. Potential reactivation of sRNA biogenesis at cRdDM loci is revealed, and non-canonical RdDM pathways are implicated in sRNA biogenesis post-meiosis. Additionally, sRNA clusters derived from MetGenes and ATGP2N TEs in sperm cells are identified, along with a cluster of sperm-specific heterochromatic loci.

In Chapter 3, the sRNA and DNA methylation profiles of *Marchantia* embryos are examined. MetGenes are defined in the sporophyte, and their connections to TE loci are identified. A method for isolating, imaging, and sequencing early embryos is tested, developed, and presented. Nuclear reporter lines are constructed to observe the phenotypic effects of knocking out paternal 4mC methylation in the embryo.

Each chapter contains a detailed materials and methods section, as well as appendices with supplemental information. In Chapter 4, the results are summarised in a broader context, and potential future directions, reflections, and limitations of the work are discussed.

Chapter 2

The spatiotemporal expression of RdDM components and sRNA landscapes in the germline of *Arabidopsis thaliana*

2.1 Abstract

This chapter explores the spatiotemporal expression of RdDM components and the sRNA landscapes in the male germline of *Arabidopsis thaliana*. By employing a combination of imaging fluorescent reporter lines, sRNA sequencing, and bioinformatics analysis, the dynamics of RdDM-associated factors during different stages of pollen development were investigated. It was demonstrated that the expression of key RdDM components such as NRPD1 and NRPE1 is tightly regulated, with NRPD1 showing tapetum-specific expression and NRPE1 being active in both germline and somatic tissues. Furthermore, CLSY3 expression was found to be confined to the tapetum and no other CLSY chromatin remodeller expression was detected after meiosis. Further analysis of the sRNA profiles reveals that the sperm cell harbours a distinct sRNA profile, uncoupled from known CLSY-dependent loci, suggesting a shift towards non-canonical RdDM pathways post-meiosis.

2.2 Introduction

It was previously thought that DNA methylation largely remained stable across reproductive cells in plants, due to its close association with the regulation of transposon activity [80] and therefore genome integrity. However, global demethylation does occur in companion cells of the germline, via passive or active demethylation. The former is due to decreased expression of MET1, leading to a lack of maintenance of methylation during DNA replication in the central cell. The latter is due to active demethylation by DEMETER (DME), which demethylates CG sites in the vegetative nucleus [81, 82]. This leads to the synthesis of 21nt epigenetically active small interfering RNAs (easiR-

NAs) in the vegetative cell from typically silenced TEs. As the vegetative cell is a terminal tissue rather than part of the germline, the destabilising effects to the genome are not of evolutionary concern. The easiRNAs produced from the vegetative cells are hypothesised to move into the sperm cells, where they accumulate and silence complementary target DNA sites through RdDM [49, 83].

The RdDM pathway provides some of the essential communication between different cell types in reproductive tissue. AGO6 and AGO9 are a close paralog of AGO4 expressed in reproductive cells, which binds to 24nt RNA to direct RdDM. AGO6 is expressed in the shoot apical meristem, and there is increasing evidence that 21-22 nt secondary siRNAs (deriving from the miRNA pathway) can associate with AGO6 and participate in RdDM [84, 34, 35]. *Arabidopsis* ago9 produces multiple megaspore mother cells (MMCs) in the female germline. Here, AGO9 is in the surrounding somatic nucellar cells, suggesting that sRNAs and corresponding essential RdDM machinery are transferred between cells. TEs that are constitutively silenced in wildtype plants were reactivated in ago9 embryo sacs, suggesting a pathway similar to the mammalian piRNA pathway previously discussed [49].

Recent work identified male sexual lineage specific DNA methylation reprogramming in the *Arabidopsis* germline. In general, male meiocytes maintain high levels of CG and CHG methylation when compared to somatic canonical RdDM loci levels, and a relatively low level of CHH methylation. However, in the CHH context, there is distinct hypermethylation at selected, sexual lineage specific loci (HyperTEs), including novel targets of RdDM activity. Further, *de novo* gene associated methylated loci were also identified (MetGenes), including a meiosis specific gene (Figure 2.1) [50].

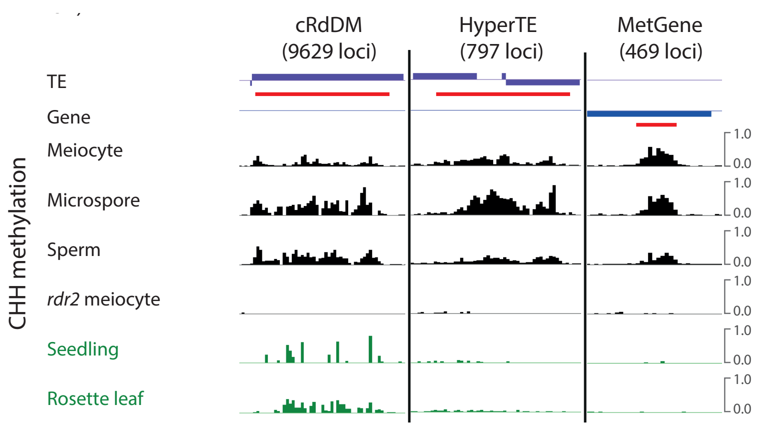


Fig. 2.1 Hypermethylated and *de novo* gene-associated methylated loci are identified in the germline of *Arabidopsis* (Figure from [50])

Further, it was discovered that meiocytes themselves are quiescent in sRNA biogenesis. The biogenesis of these germline specific siRNAs is in the somatic tapetal tissue and is associated with the chromatin remodeler CLASSY3. Nurse cell-derived siRNAs are then transported into the meiocyte from the tapetum, inducing DNA methylation at

HyperTE loci. MetGene loci can be uniquely targeted for methylation in the meiocyte by imperfectly matching siRNA derived from HyperTEs. These nurse-cell derived siRNAs inform the entire male sexual lineage specific methylation reprogramming, from meiocyte to sperm (Figure 2.2) [60].

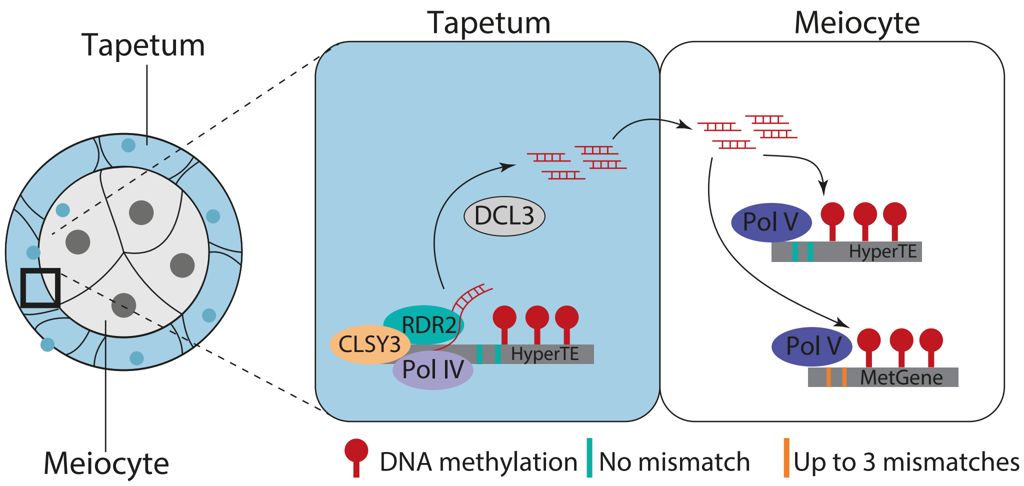


Fig. 2.2 Tapetum-derived sRNAs shape methylation in the germline of *Arabidopsis*

Tapetal sRNAs direct DNA methylation at HyperTEs and MetGenes in the meiocyte (Figure from [60])

The mechanism of movement of sRNA between companion and germline cells needs more evidence, however there is extensive evidence of cross-cellular RNA movement. Plasmodesmata provide a cytoplasmic network between cells (such as the tapetum and meiocyte) which allows nutrient transport but have also been shown to allow RNA movement between cells [85–87].

The aim of this chapter is to shed light on how this germline-specific sRNA and DNA methylation profile changes after meiosis. Namely, to investigate the sRNA profiles of the microspore and sperm cells, and to examine the expression of RdDM associated key factors involved during the development of male-sexual lineage.

2.3 NRPD1 is expressed in the tapetum, microspores but not in meiocytes

Since RNA polymerase IV and V preferentially associate with methylated DNA, RdDM is thought of as a self-reinforcing pathway. Therefore, the observation that there are few perfect matching 24nt sRNAs in MetGenes in *Arabidopsis* meiocytes, it was hypothesised that the Pol IV-dependent pathway is quiescent and that the sRNAs directing DNA methylation at MetGenes are produced from HyperTE loci in the tapetum that target HyperTE and MetGene methylation in the meiocyte [50, 60]. To confirm this expression pattern, pNRPD1::NRPD1-eGFP and pNRPE1::NRPE1-eGFP fluorescent reporter lines (of the largest subunits of RNA Polymerase IV and V respectively[35]) were constructed by dr Jincheng Long. Since the columella in the root tip of *Arabidopsis* is hypermethylated in the CHH context [88], the root tip was imaged to screen and select the best lines selected for further imaging (FigureA3).

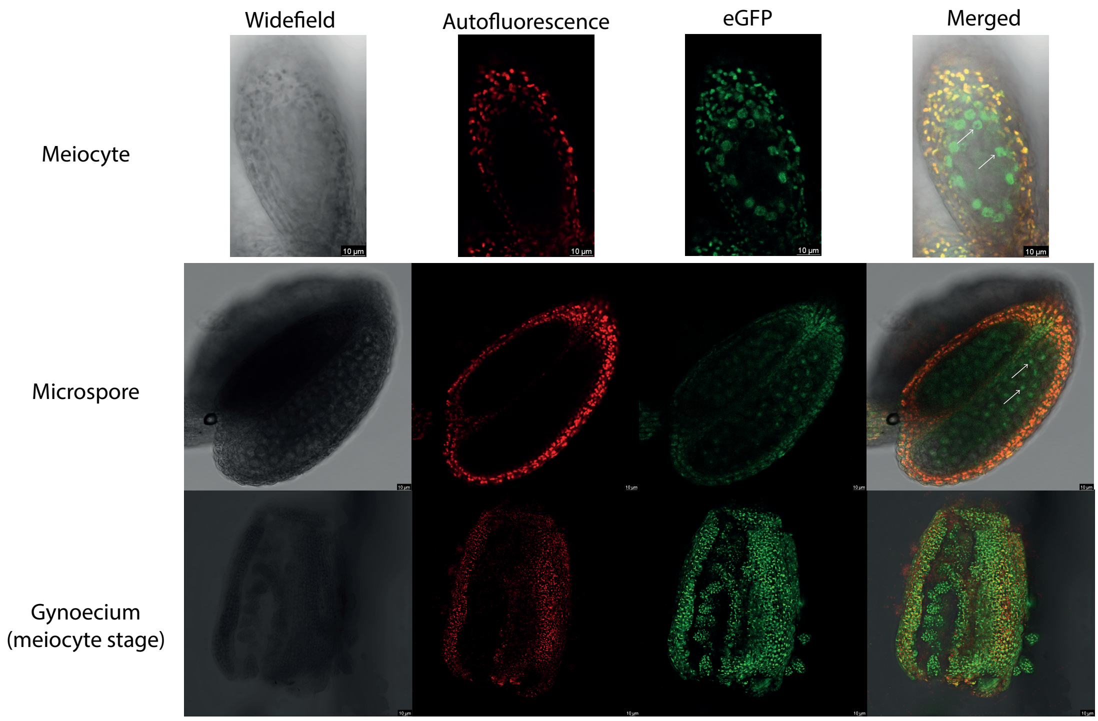


Fig. 2.3 NRPD1 is expressed in the tapetum, microspore and gynoecium.

NRPD1 expression in meiocyte stage anthers in the tapetum and surrounding somatic cells(first row, third column (GFP, green, white arrows), and expression in microspores (second row third column, white arrows). NRPD1 is also expressed in both the germ and somatic cells of the meiocyte stage gynoecium (third row (GFP, green)). Scale bar 10 μm. Reporter line constructed by dr Jincheng Long.

Meiocytes are surrounded by the tapetum - a somatic tissue layer of nurse cells with cytoplasmic connections to meiocytes through plasmodesmata. As expected, NRDP1 expression was confined to the tapetal layer and surrounding somatic anther tissue in meiocyte stage anthers (Figure 2.3). Once meiocytes undergo meiosis, the plasmodesmatal connection to the tapetum ceases and the meiocytes are surrounded by callose cell walls. At the microspore stage, the germline cells express NRPD1 (Figure 2.3 second row, 2.4 first and second rows). NRPD1 was also expected to be expressed strongly in the maternal reproductive tissue [63] which was confirmed in several independent reporter lines (Figure 2.3, row three).

Following a further two cell divisions, the bicellular and tricellular pollen grains are produced, neither of which express NRPD1 (Figure 2.4, rows 3 and 4). However, it must be added that the expression levels of the best NRPD1 reporter lines were a lot lower and more diffuse than the expression of NRPE1 or CLSY3 reporter lines (Figures 2.6, A3). Further, expression in germline tissues was not consistent even in lines that exhibited strong root signal (Figure 2.4, Figure A4). As these results are from screening 37 independent transformants from the same transformation event, it would be worth constructing new transformants perhaps with a stronger fluorescent reporter (such as tdTomato) to corroborate these results, and confidently confirm whether NRPD1 is expressed in the pollen.

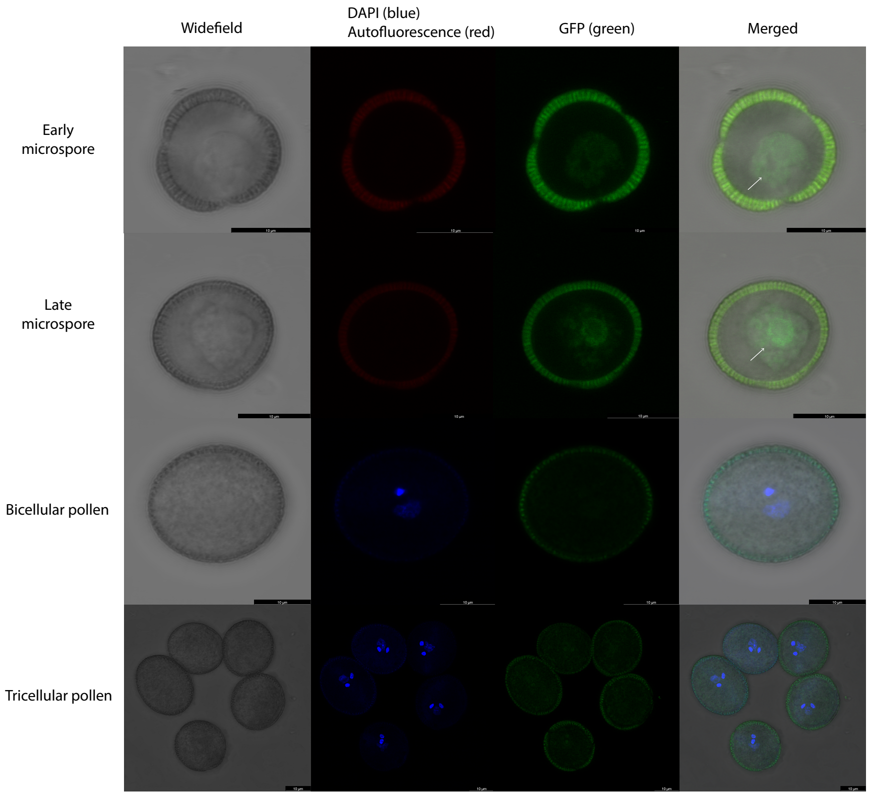


Fig. 2.4 NRPD1 is expressed in microspores, but its expression is absent in pollen

NRPD1 expression early and late microsproes (first and second rows, third column (GFP, green, white arrows), and expression in microspores (second row third column, white arrows). NRPD1 however us not expressed in bicellular or tricellular pollen (third and fourth rows (DAPI, blue GFP, green). Scale bar 10 μm. Reporter line constructed by dr Jincheng Long.

2.4 NRPE1 is expressed in both somatic and germline tissues in *Arabidopsis* anthers

According to methylation and 24nt sRNA datasets, the Pol V-dependent pathway of RdDM be active in the somatic tapetal layer and meiocytes, but it was not clear what happens after meiosis in microspores and pollen. As before, strong nuclear signal was confirmed in the root tip (Figure A3). NRPE1 is strongly expressed in meiocytes, tapetum and in the surrounding somatic tissues of the meiocyte stage anther (Figure 2.5, first row). This expression also persists into both soma and germline of microspore stage anthers and the female meiocyte stage gynoecium (Figure 2.5 second and third rows respectively), indicating that the Pol V-dependent pathway of RdDM is active in both somatic and germline cells of the anther. As we know RdDM is highly active in maternal reproductive tissue [62, 63], the expression of NRPE1 was confirmed in this tissue as well (Figure 2.5, third row).

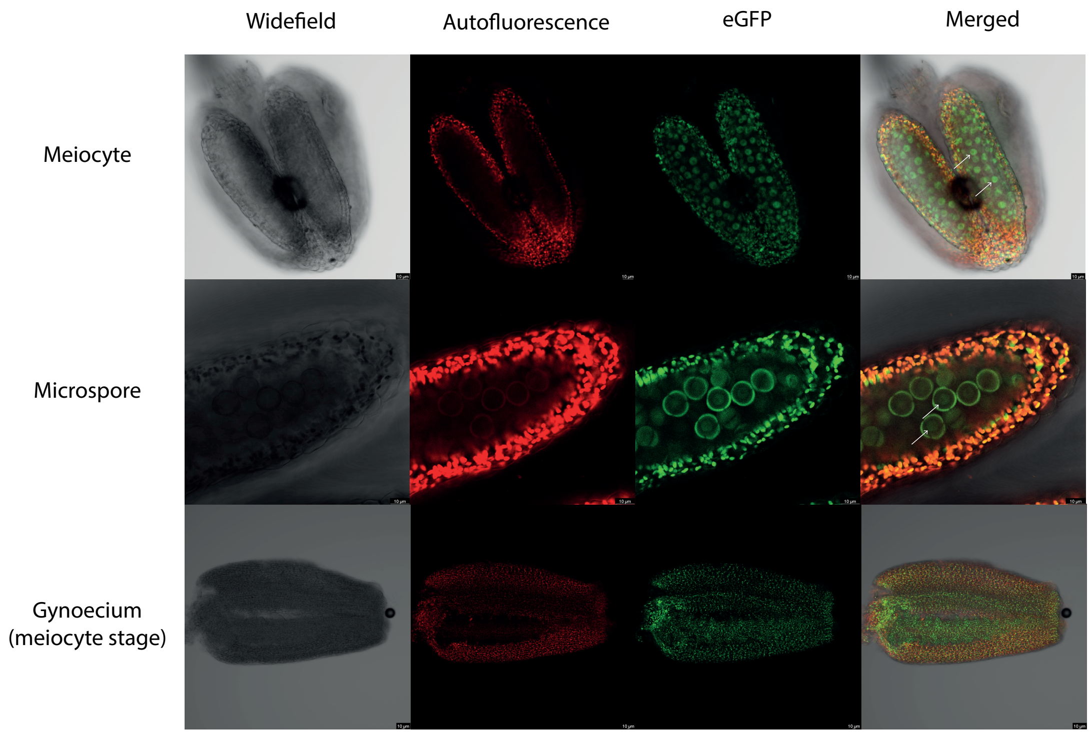


Fig. 2.5 NRPE1 is expressed in the tapetum, meiocyte, microspores and gynoecium

NRPE1 expression in meiocyte stage anthers in the tapetum, meiocyte and surrounding somatic cells(first row, third column (GFP, green, white arrows), and expression in microspores (second row third column, white arrows). NRPE1 is also expressed in both the germ and somatic cells of the meiocyte stage gynoecium (third row (GFP, green)). Scale bar 10 μm. Reporter line constructed by dr Jincheng Long.

2.5 CLSY3 is expressed in the tapetum of meiocyte stage anthers and in the gynoecium

The CLSY3 chromatin remodeler drives sRNA production in the tapetum [60] by recruiting Pol IV to discrete loci (including HyperTEs)[89]. In accordance with this, the expression of CLSY3 in meiocyte stage anthers is largely restricted to the tapetum (Figure 2.6 first row), with diminished expression in the tapetum after meiosis (Figure 2.6 second row). CLSY3 is also required for expression of abundant 24nt sRNA in the ovules, and therefore it is strongly expressed in the gynoecium and surrounding somatic tissues (Figure 2.6, third row. Reporter line constructed by dr James Walker/Sam Deans).

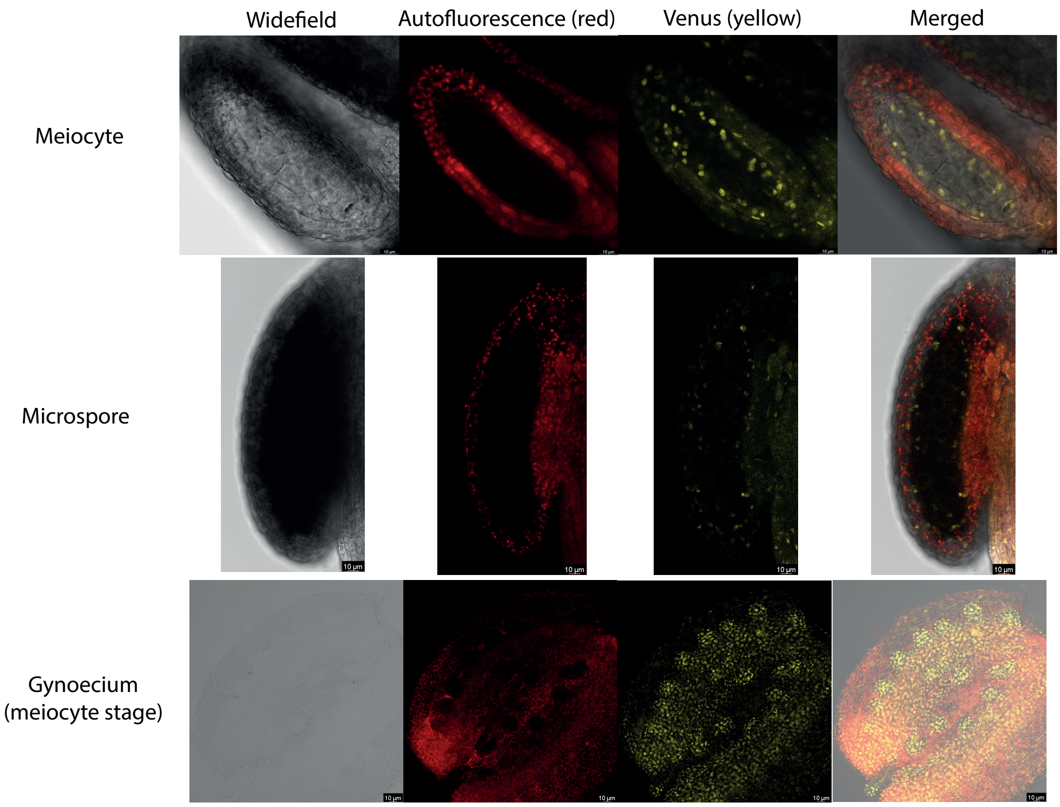


Fig. 2.6 CLSY3 is expressed in the tapetum of meiocyte stage anthers and in the gynoecium

CLSY3 is strongly expressed in the nucleus of meiocyte stage tapetal cells (first row (Venus, yellow)), which then diminishes in the senescing tapetum after meiosis in the microspore stage(second row, Venus (yellow)). CLSY3 is also expressed in both the germ and somatic cells of the meiocyte stage gynoecium (third row (Venus, yellow)). Scale bar 10 μm. Reporter line constructed by dr James Walker/Sam Deans.

However, following meiosis, CLSY3 is not expressed in microspore, bicellular or tricellular pollen stages (Figure 2.7) highlighting the restricted spatiotemporal expression of CLSY3 to produce 24nt sRNAs at HyperTEs and drive methylation of these loci, as well as MetGene loci in the meiocyte. Although endogenous sRNA production is active in the microspore (Figure 2.4), and potentially in pollen, CLSY3 does not seem to play a significant role in shaping the 24nt sRNA profiles after meiosis.

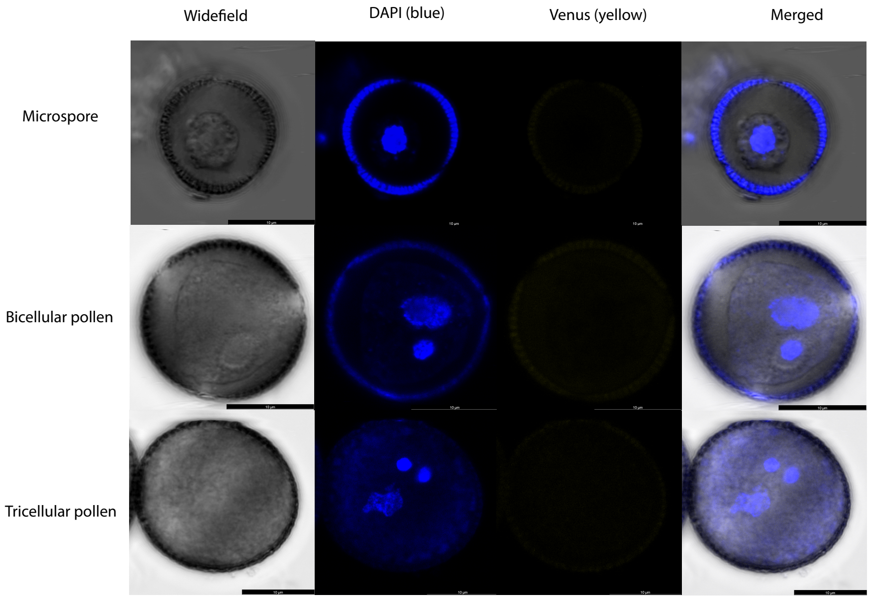


Fig. 2.7 CLSY3 is not expressed in the male germline cells

DAPI stained microspore, bicellular or tricellular pollen (second column (DAPI, blue). CLSY3 signal not detected in any germline cells (third row (Venus, yellow). Scale bar 10 μm. Reporter line constructed by dr James Walker/Sam Deans.

2.6 HyperTE sRNAs decline proportionally post-meiosis, despite elevated proportional dominance of 24nt sRNAs in sperm cells

To dissect the dynamics of the sRNA landscape in the germline, sRNA sequencing libraries were constructed for microspores, sperm cells, and sperm nuclei (2 replicates each). These were then compared to meiocyte and tapetum sRNA libraries [60], as well as to somatic sRNA sequencing libraries from the leaf, root and seedling [90–92].

All three somatic tissues had a strong 21nt and 24nt peak (Figure 2.8A). However, in the germline, except for microspores, the dominant sRNA length was 24nt, with sperm cells and sperm nuclei exhibiting the highest proportional levels of 24nt sRNA (Figure 2.8B).

When examining genomic features to capture a snapshot of 24nt sRNA distribution in germline tissues, it has been previously shown that HyperTEs dominate the 24nt sRNA landscape in meiocytes [60]. However, the proportion of 24nt sRNAs within HyperTE regions gradually declines through cell divisions following meiosis, from meiocytes to sperm cells, suggesting that endogenous HyperTE sRNA production does not continue or ceases in the germline cells after meiosis and the closure of the plasmodesmatal connections to microspores. (Figure 2.9A).

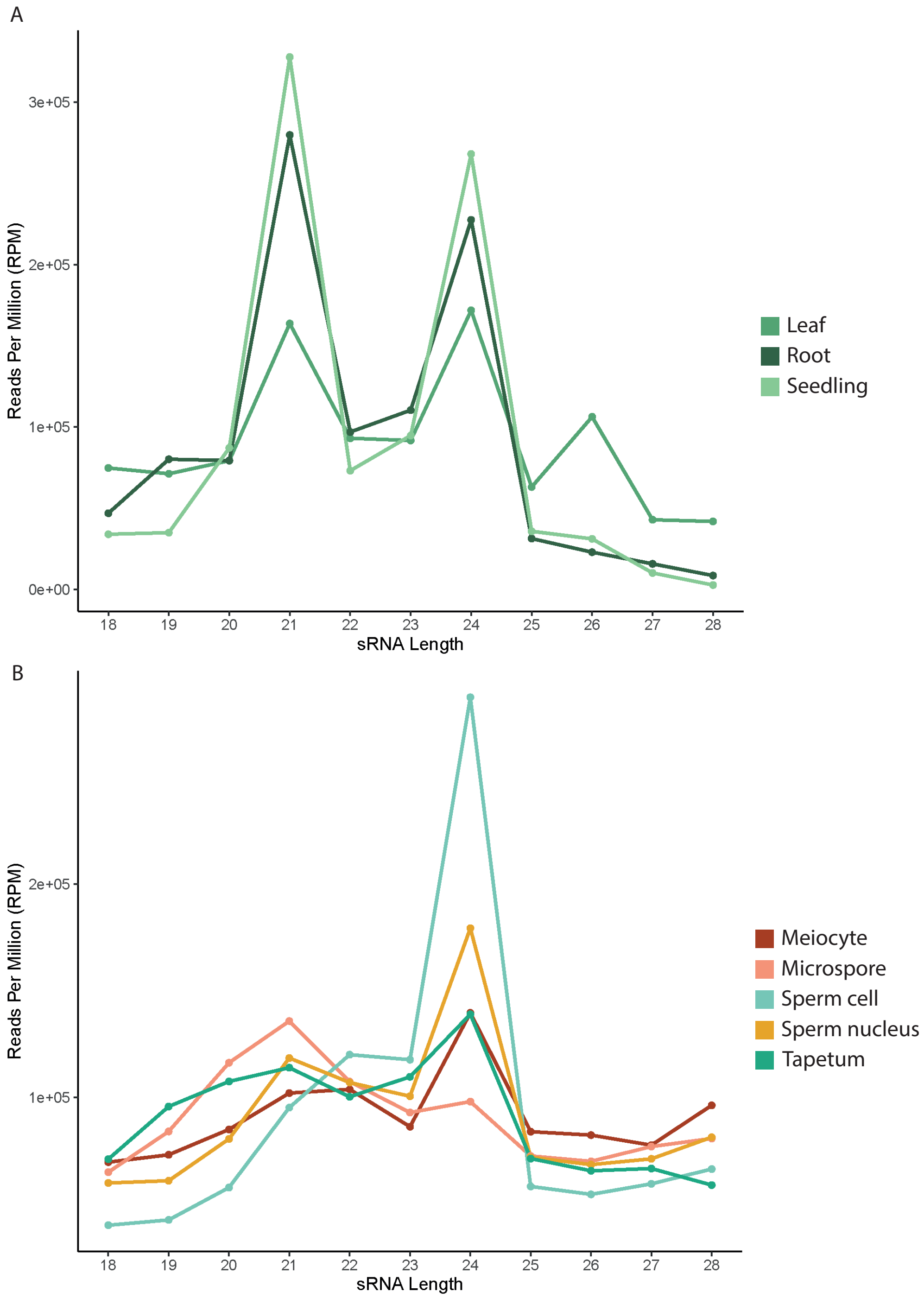


Fig. 2.8 In germline tissues, the dominant sRNA length is 24nt

(A) The size distribution of sRNAs in leaf, root and seedling. X axis: sRNA lengths, Y axis: reads per million (RPM) (B) The size distribution of sRNAs in meiocyte, microspore, sperm cell, sperm nucleus and tapetum. X axis: sRNA lengths, Y axis: reads per million (RPM)

In sperm cells, TE regions become the dominant source of 24nt sRNAs, mirroring the pattern observed in somatic tissues (Figure 2.9B). In somatic tissues, a large portion of these canonical RdDM loci are known to be associated with CLSY1 and CLSY2 [89]. However, in the pollen, neither CLSY1 nor CLSY2 were strongly expressed (Figure A2 suggesting (provided there is endogenous sRNA production in the pollen), these clusters are not strongly associated with the expression of any protein from the CLSY family (the CLSY1 and CLSY2 reporter lines were obtained from dr Jincheng Long).

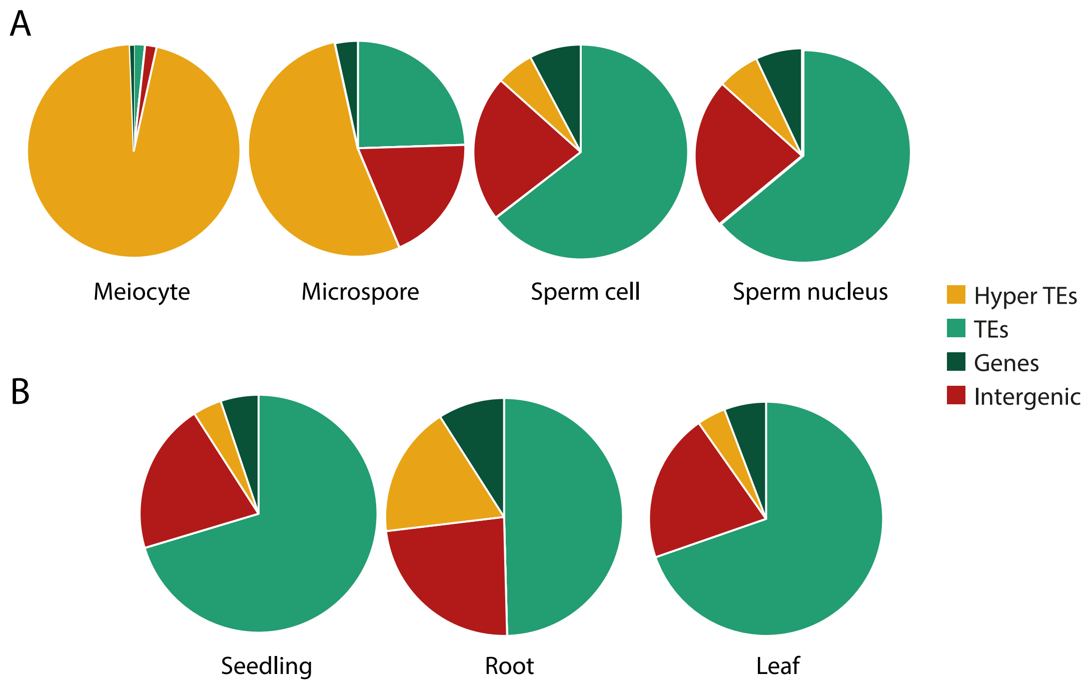


Fig. 2.9 HyperTE derived 24nt sRNAs decline in proportional abundance after meiosis and throughout pollen development

Pie chart showing the proportional mapping of 24nt sRNAs to genomic features of interest: HyperTEs, TEs, Genes, MetGenes and intergenic regions in (A) germline tissue (meiocyte, microspore, sperm cell and sperm nucleus) and (B) somatic tissue (seedling, root and leaf).

2.7 Sperm cell sRNA clusters are uncoupled from CLSY-dependent loci

Examining the expression levels of germline-associated CLSY3-dependent loci, HyperTEs, and somatic tissue-associated CLSY1/2-dependent loci, along with canonical RdDM (cRdDM) loci, we can first conclude that, as expected, CLSY1/2-dependent and cRdDM loci exhibit the highest overall abundance of sRNAs in somatic tissues (leaf, seedling, and root). In contrast, most cRdDM loci do not generate abundant sRNAs in the meiocyte, tapetum, or microspore. Interestingly, the sperm cell and sperm nucleus appear to represent an intermediate stage, where sRNA production at cRdDM loci is reactivated though not to the same extent as in somatic tissues (Figures 2.10B,D, A6).

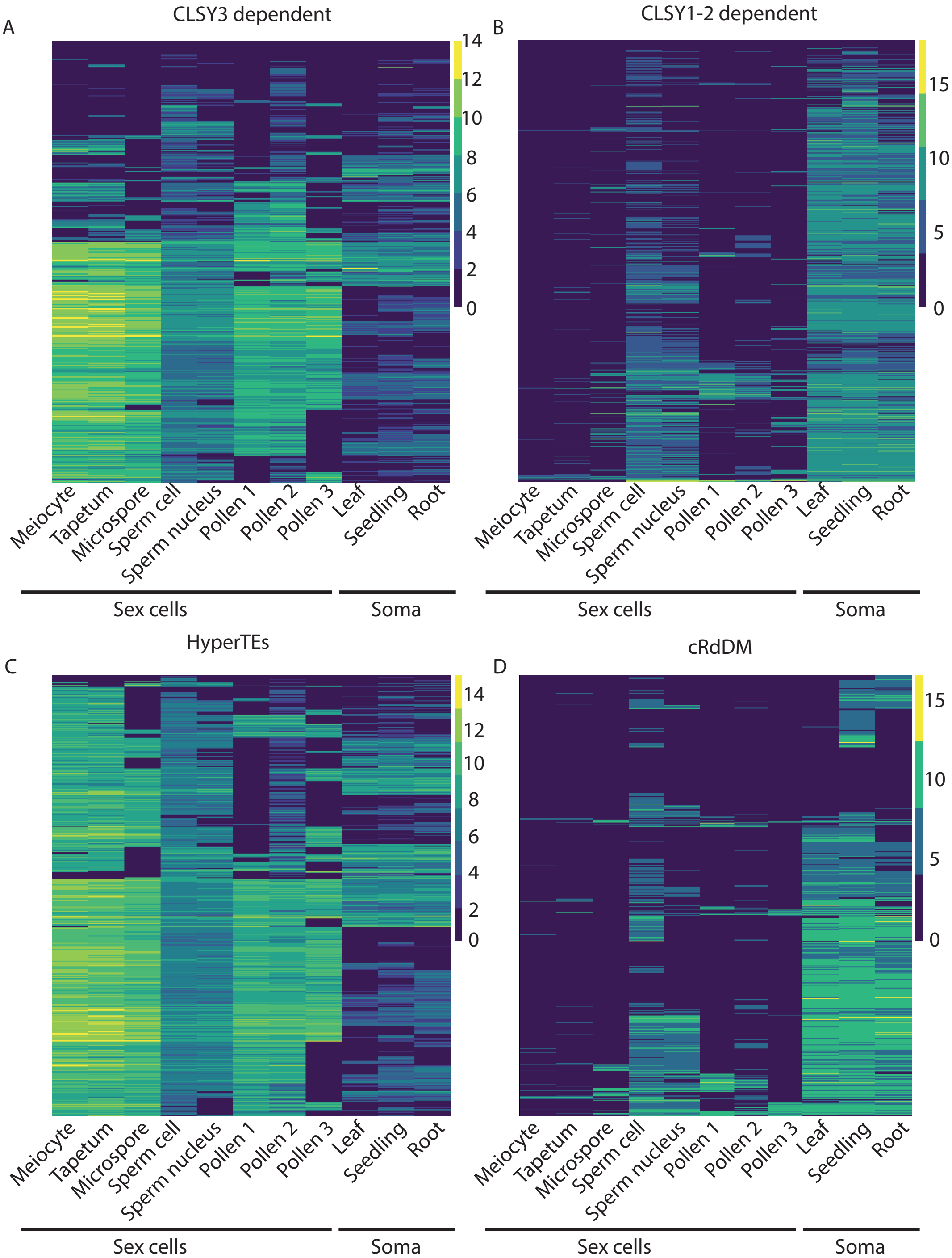


Fig. 2.10 CLSY3-dependent loci and HyperTEs generate abundant sRNAs in meiocytes and tapetum, as well as in microspores and sperm cells, though to a lesser degree. CLSY1&2-dependent loci and canonical RdDM loci primarily produce abundant sRNAs in somatic tissues, with intermediate levels in sperm cells.

Heatmaps showing the 24nt sRNA RPKM levels of (A) CLSY3 dependent, (B) CLSY1&2 dependent, (C) HyperTEs and (D) cRdDM loci in different germline (meiocyte, tapetum, sperm cell, sperm nucleus, pollen) and somatic tissues (leaf, seedling, root). Scale bar shows logRPKM values.

Conversely, in CLSY3-dependent and HyperTE loci, we observe an opposite pattern: 24nt sRNA abundance is highest in the meiocyte, tapetum, and microspore (though somewhat reduced in the microspore at certain loci), while somatic tissues exhibit lower sRNA levels. The sperm cell and sperm nucleus again show intermediate sRNA abundance. Pollen, in particular, displays higher overall levels of 24nt sRNAs from these loci, suggesting they are primarily produced in the vegetative cell. Notably, a subset of HyperTE loci continues to produce sRNAs even in somatic tissues (Figures 2.10A-C, A6).

Since the heatmaps show normalised levels of sRNA against total sRNA levels in each tissue, it must be ensured that these patterns are comparable. To achieve this, 24nt sRNA levels in the cRdDM loci were also normalized against 21nt miRNA levels, as these are assumed to remain constant across cell types. This approach yielded comparable sRNA levels across the tested tissues, validating the observed relative differences in sRNA patterns (Figure A5).

To further understand the sRNA landscape in sperm and microspore, the sRNA profiles were compared pairwise to somatic tissues as well as other germline tissues, namely the meiocyte. Genome wide, 24nt sRNA abundance in 24nt sRNA clusters correlate weakly between the meiocyte and microspore (Figure A7A), but less so than between tapetum and meiocyte (Figure 2.13A). This correlation is then further reduced between the meiocyte and sperm, indicating that global 24nt sRNA profiles are indeed vastly different between these cell types (Figure A7B).

Additionally, when comparing sperm cell-specific sRNA clusters between the meiocyte and sperm, and highlighting key genomic features, we clearly see that a significant subset of sperm-specific clusters producing sRNAs overlap with HyperTEs. These clusters have correspondingly high sRNA abundance in the meiocyte, with relatively lower abundance in the sperm cell (Figure A7B). In contrast, loci with low RPKM values in the meiocyte but high in sperm include a mixture of canonical RdDM loci and other sperm-specific loci that do not overlap with either HyperTEs or cRdDMs.

Similarly, highlighting CLSY-dependent loci in the same dataset yields results consistent with those in Figure A7B. In Figure 2.11, loci with abundant sRNAs in the meiocyte overlap CLSY3-dependent loci and some CLSY4 and CLSY2 dependent loci (Figure 2.11B, C, D). Surprisingly, loci that produce abundant sRNAs in sperm distinctly from the meiocyte do not exclusively overlap CLSY1 and CLSY2 dependent loci as would be expected in somatic tissues (Figure 2.11A,B). Instead, many of these loci do not overlap with any CLSY-dependent loci at all (Figure 2.11E).

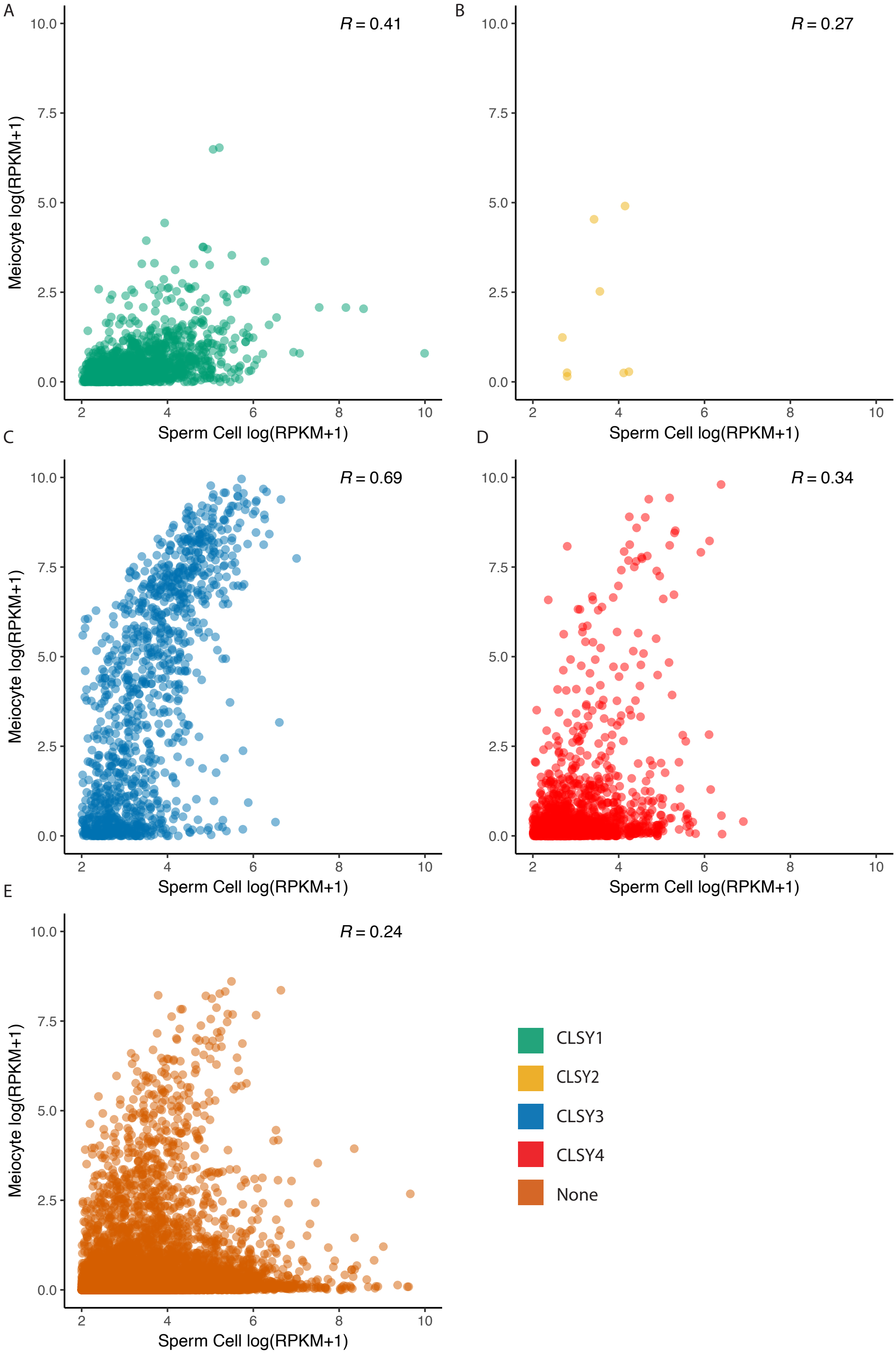


Fig. 2.11 The distribution of the abundance of 24nt sRNAs at CLSY associated loci compared between the meiocyte and sperm cell

Scatterplot shows the log(RPKM+1) values of 24nt sRNAs at sperm-specific clusters, split by CLSY-dependent loci A) CLSY 1 B) CLSY2 C) CLSY3 D) CLSY4 E) Loci that do not overlap any CLSY dependent loci. R = Spearman's rank

2.8 Sperm-cell specific loci do not completely share sRNA profiles with pollen

As illustrated in Figure 2.10 there is not a complete overlap in sRNA production between sperm and pollen. The expression of sperm-cell-specific clusters was compared with several published pollen datasets, as well as libraries derived from sperm cell and sperm nucleus tissues. While the sRNA profiles of the sperm cell and sperm nucleus are highly correlated (Figure A8) there are distinct sRNA expression patterns between pollen and sperm cells, despite their overall correlation (Figure A8) indicating the contribution of the vegetative cell.

Dissecting this further, we discovered that HyperTEs which continue to produce abundant sRNAs in the sperm cell, also do so in the entire pollen. This suggests that these sRNAs are either expressed in both the vegetative and sperm cells, or are able to cross into the sperm cell from the vegetative cell. Furthermore, cRdDM loci appear to fall into two categories: one group produces abundant sRNAs in both pollen and sperm, while another exhibits high sRNA levels in pollen but lower levels in sperm, indicating the presence of a vegetative cell-specific sRNA population. A similar pattern can be seen in the sRNA clusters that do not overlap these genomic features at all. Interestingly, the group of MetGenes that produce sRNAs in the sperm cell seem to be largely specific to this tissue (this is discussed later in the chapter). As the pollen encompasses both sperm cell and vegetative cell sRNA production, these results describe the contribution of the vegetative cell to overall sRNA production in the pollen (Figure 2.12).

2.9 Distinct Chromatin and DNA Methylation Patterns in Sperm-Specific sRNA Loci

As expected, loci with abundant sRNAs in the meiocyte and the tapetum are primarily HyperTEs (Figure 2.13A), which decline in abundance in the sperm cell (Figure 2.13B). Next, examining whether loci that produce abundant sRNAs in sperm compared to the seedling have different chromatin states, (based on H3K9me2 as an indicator of euchromatin or heterochromatin), we find that a subset of loci (n = 568) generating more sRNAs in sperm than in the seedling are predominantly heterochromatic (Figure 2.13C, red box). As expected, sperm cell-specific clusters show a strong correlation between the sperm cell and sperm nucleus (Figure 2.13D). Furthermore, loci synthesising the most abundant sRNA in both sperm cell and sperm nucleus tend to be heterochromatic.

It is crucial to investigate whether the heterochromatic sperm-specific n568 subcluster has a specific methylation pattern distinct from the whole set. We can conclude that generally, methylation in the sperm-specific sRNA clusters is consistently high, intermediate, and low in the CG CHG and CHH context throughout germline development

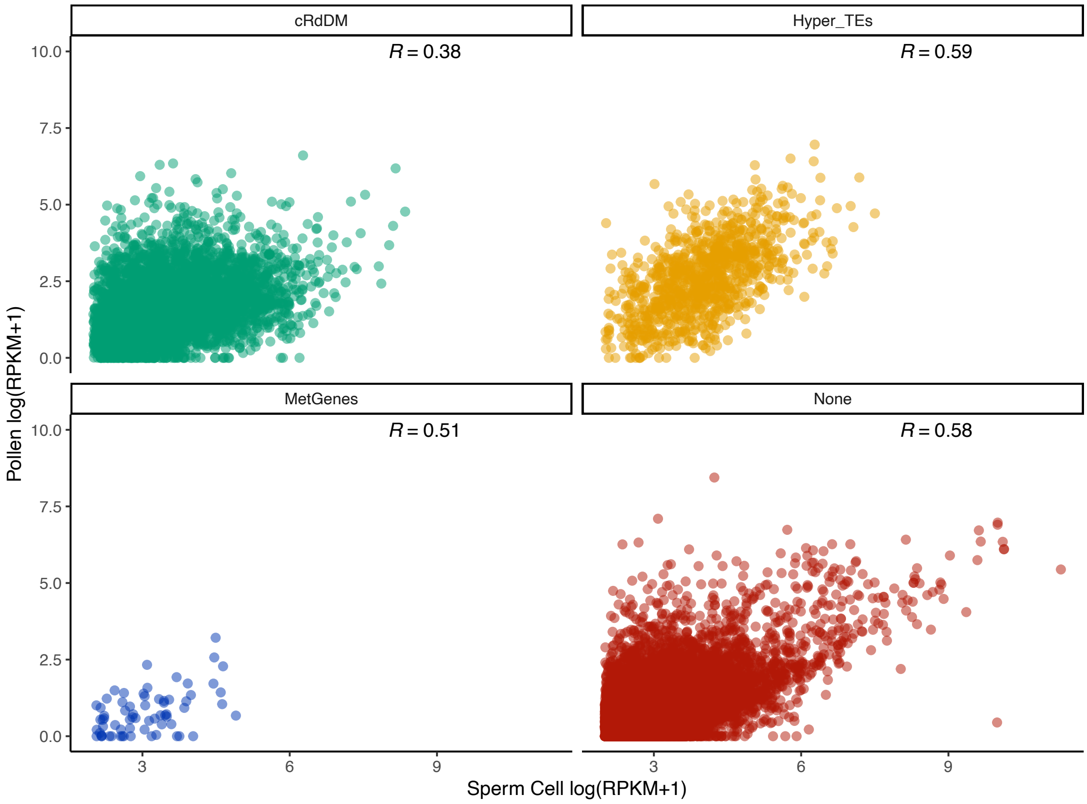


Fig. 2.12 The distribution of the abundance of 24nt sRNAs overlapping genomic features of interest, compared between the pollen and sperm cell

Scatterplot shows the log(RPKM+1) values of 24nt sRNAs at sperm-specific clusters, split by cRdDM (green), HyperTE (yellow), MetGene (blue) and non-overlapping (red) loci. R = Spearman's rank

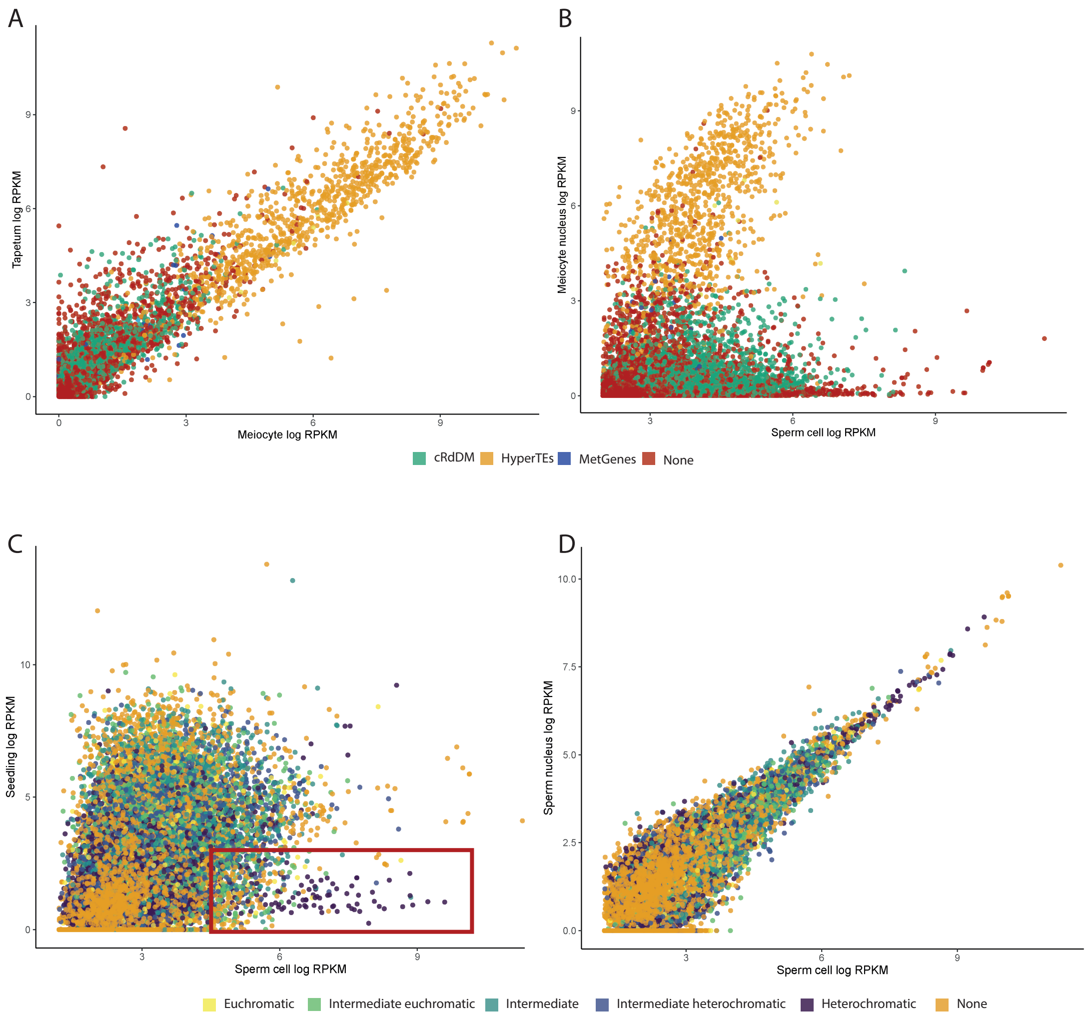


Fig. 2.13 HyperTEs produce highly abundant 24nt sRNAs specifically in the meiocyte and tapetum, which tapers off in the sperm cell. Loci with highly abundant 24nt sRNAs specifically in the sperm cell tend to be more heterochromatic.

Correlations between 24nt sRNA abundance and the indicated cell types, highlighting (A-B) genomic features of interest and (C-D) chromatin structure. (C) Sperm cell specific heterochromatic cluster highlighed with red box.

respectively (Figure 2.14, dark green bars). In contrast, the n568 subcluster exhibits generally lower methylation levels compared to the full set (Figure 2.14, light green bars). However, when examining methylation patterns in other tissues, it becomes clear that this pattern is not exclusive to the sperm cell.

2.10 A subset of MetGenes produce 24nt sRNAs and gain CHH methylation specifically in the sperm cell

In contrast to other sex cells, about 20% of MetGenes generate 24nt sRNAs in the sperm cell, with a subset producing sRNAs exclusively in sperm (Figure 2.15, clusters 1 and 2, Figure A9B). Since previous work has identified that MetGenes produce no perfect matching sRNAs in the meiocyte and tapetum[60], it is likely that the clusters which produce sRNAs in meiocyte and tapetum are either at HyperTE/MetGenes boundaries, or are HyperTE derived. However, the majority of sRNA producing MetGenes in the sperm cell and sperm nucleus do not overlap the same loci (Figure 2.15, clusters 1 and 2: sperm specific reactivation, cluster 3: HyperTE/boundary derived meiocyte and tapetum specific clusters).

Next, it was important to determine whether the increased sRNA production actually leads to increased DNA methylation in the sperm cell. Overall, the sperm cell-specific MetGenes exhibited higher CG and CHG methylation compared to the full set in meiocytes and microspores (Figure 2.16A,B,D,E), and higher methylation in the CHG context in sperm. Interestingly, the only tissue where this subset showed higher CHH methylation was in the sperm cell, suggesting that renewed sRNA production at these loci also drives methylation in the sperm cell (Figure 2.16C,F).

2.11 ATGP2N TEs produce highly abundant 24nt sRNAs in sperm cells, the sperm nucleus, and pollen.

As previously mentioned, after meiosis, HyperTE-derived sRNAs cease to be the dominant source of 24nt sRNAs in the sperm cell. Therefore, it is important to analyse the landscape of 24nt sRNAs from sperm clusters overlapping TEs (Figure 2.17A). This revealed a cluster of TE families conserved across pollen, sperm cells, and sperm nuclei, which produce uniquely abundant sRNAs (Figure 2.17A, cluster 8, n=24). Within this subset, approximately 90% of transposons belong to the LTR/Gypsy superfamily, and about 70% are from the ATGP2 and ATGP2N families. Of these, about 40% overlap CLSY3/4 dependent loci specifically.

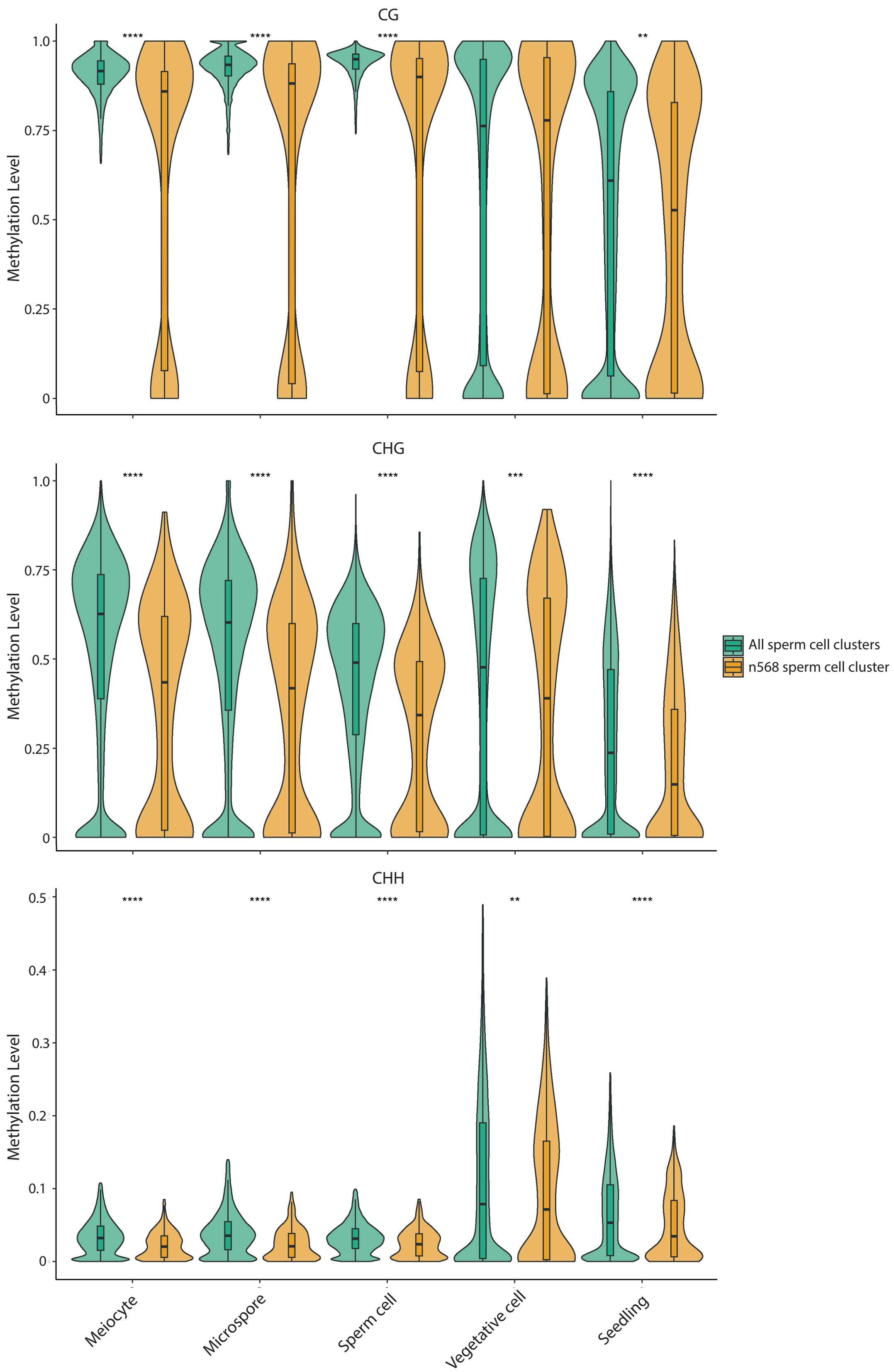


Fig. 2.14 Sperm-specific heterochromatic loci exhibit significantly lower overall methylation levels compared to the entire sperm dataset

CG, CHG and CHH methylation levels of all sperm clusters (dark green) and the heterochromatic 568 loci subset (light green). ***: p-value < 0.001, **: p-value < 0.01,*: p-value < 0.05

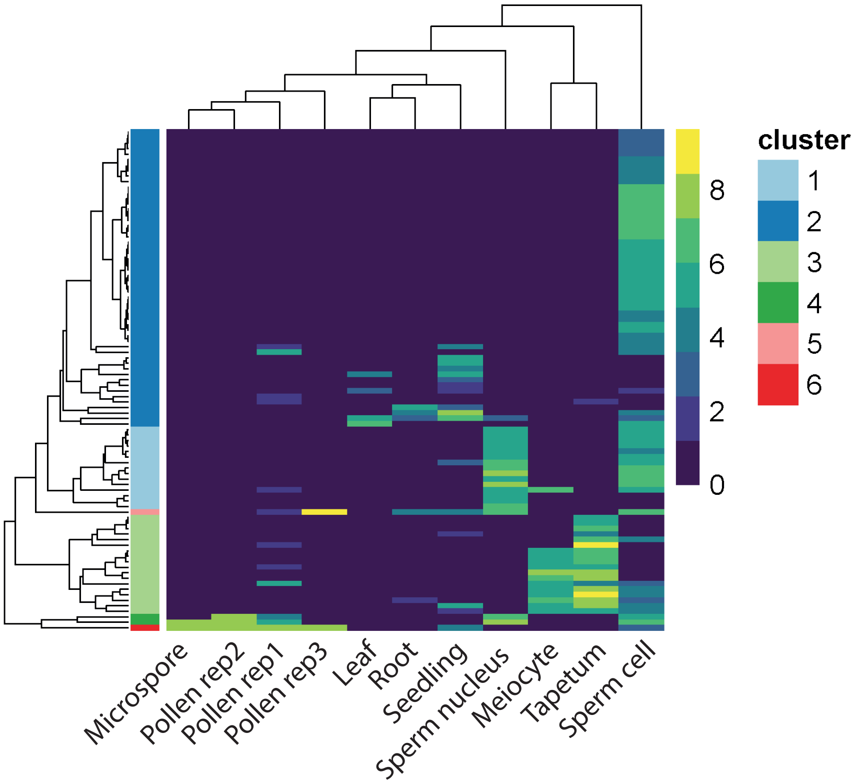


Fig. 2.15 Approximately 20% of MetGenes produce sRNAs in the sperm

Clusters 1,2: sperm specific sRNA production. Cluster 3: HyperTE/MetGene boundary derived meiocyte and tapetum specific clusters. Scale bar shows logRPKM values.

Notably, another LTR/Gypsy transposon ATGP1 has been reported to be a direct target of RdDM in meiocytes, targeted by sRNAs produced in the tapetum, and gaining release from repression in RdDM defective mutants [60], specifically in the male germline.

When comparing ATGP2N methylation between the soma and germline, both showed consistently high methylation in the CG and CHG contexts (Figure 2.18A), with generally low methylation in the CHH context. However, compared to other TE loci in the genome, ATGP2N retained higher methylation levels in the germline than in the soma in the CG and CHG contexts. Additionally, ATGP2N exhibited low CHH methylation in meiocytes and sperm cells compared to other TE loci but gained CHH methylation in vegetative cells, mirroring the pattern observed in seedlings (Figure 2.18B).

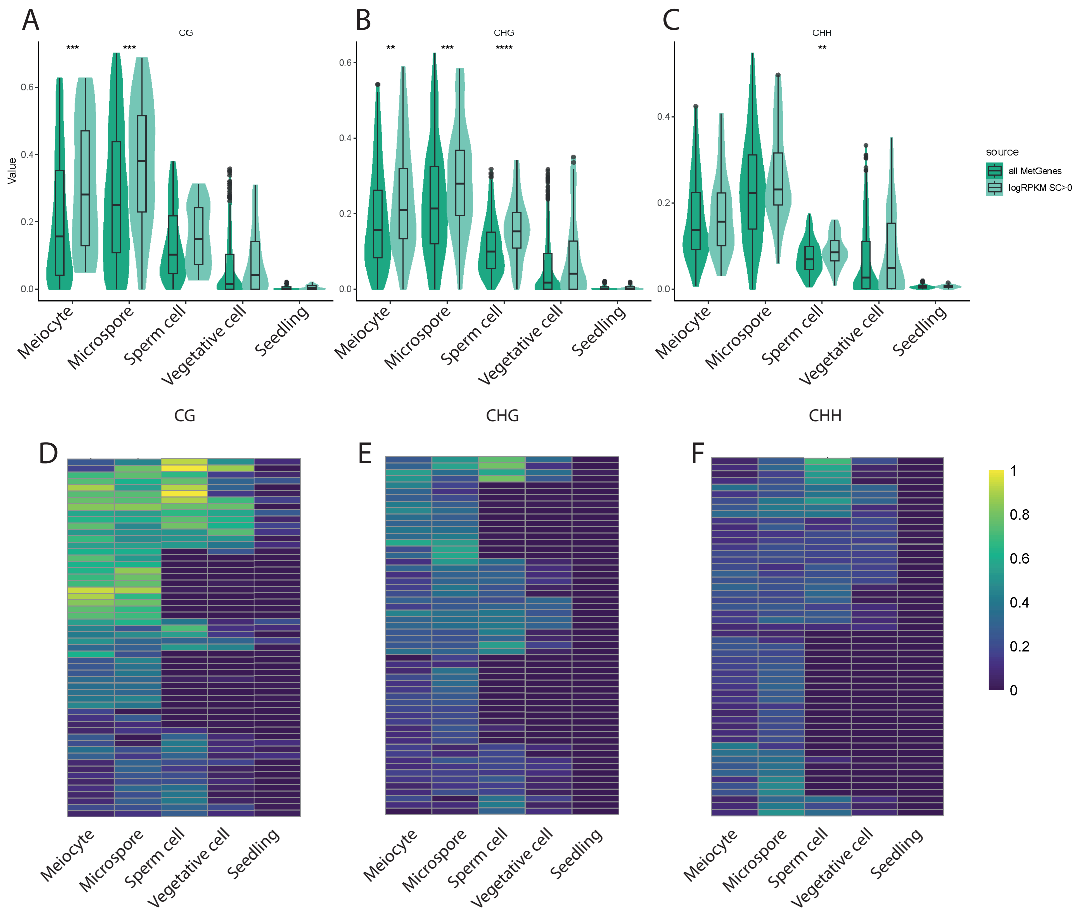


Fig. 2.16 The reactivated MetGenes gain CHH methylation in the sperm cell.

(Box plots of CG (A), CHG (B) and CHH (C) methylation of all MetGenes compared to the reactivated subset, in germline and somatic tissues. Heatmaps of the CG (D) CHG (E) and CHH (F) methylation MetGenes that produce sRNAs in the sperm cell. ***: p-value < 0.001, **: p-value < 0.01,*: p-value < 0.05.

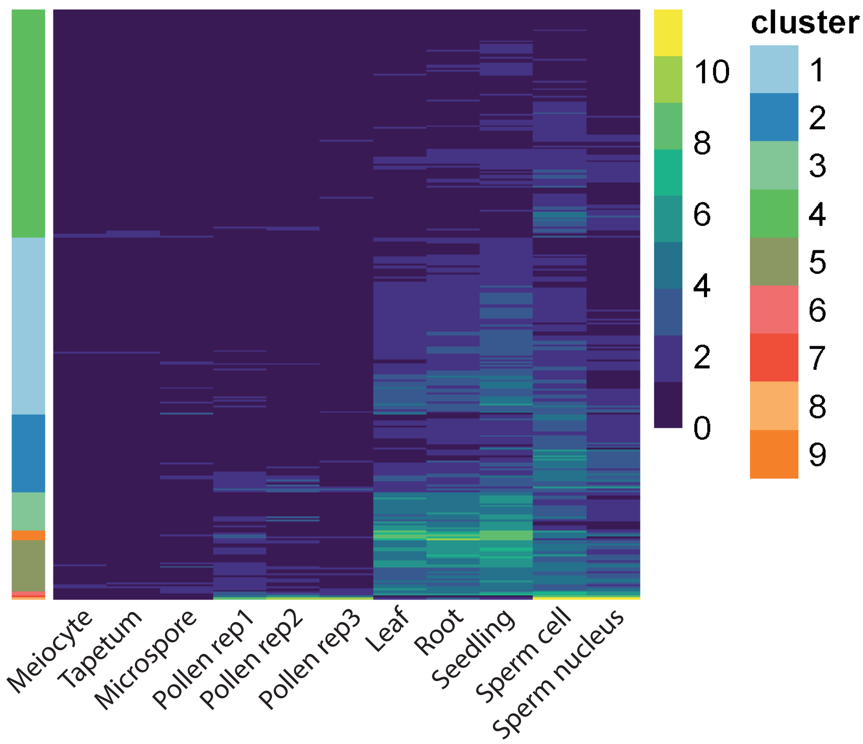


Fig. 2.17 ATGP2N TEs produce highly abundant 24nt sRNAs in the sperm cell, sperm nucleus and pollen

(A) Heatmap showing the 24nt sRNA production of TE families in germline and somatic tissues. Cluster 9 contains the ATGP2N families. Scale bar shows logRPKM values.

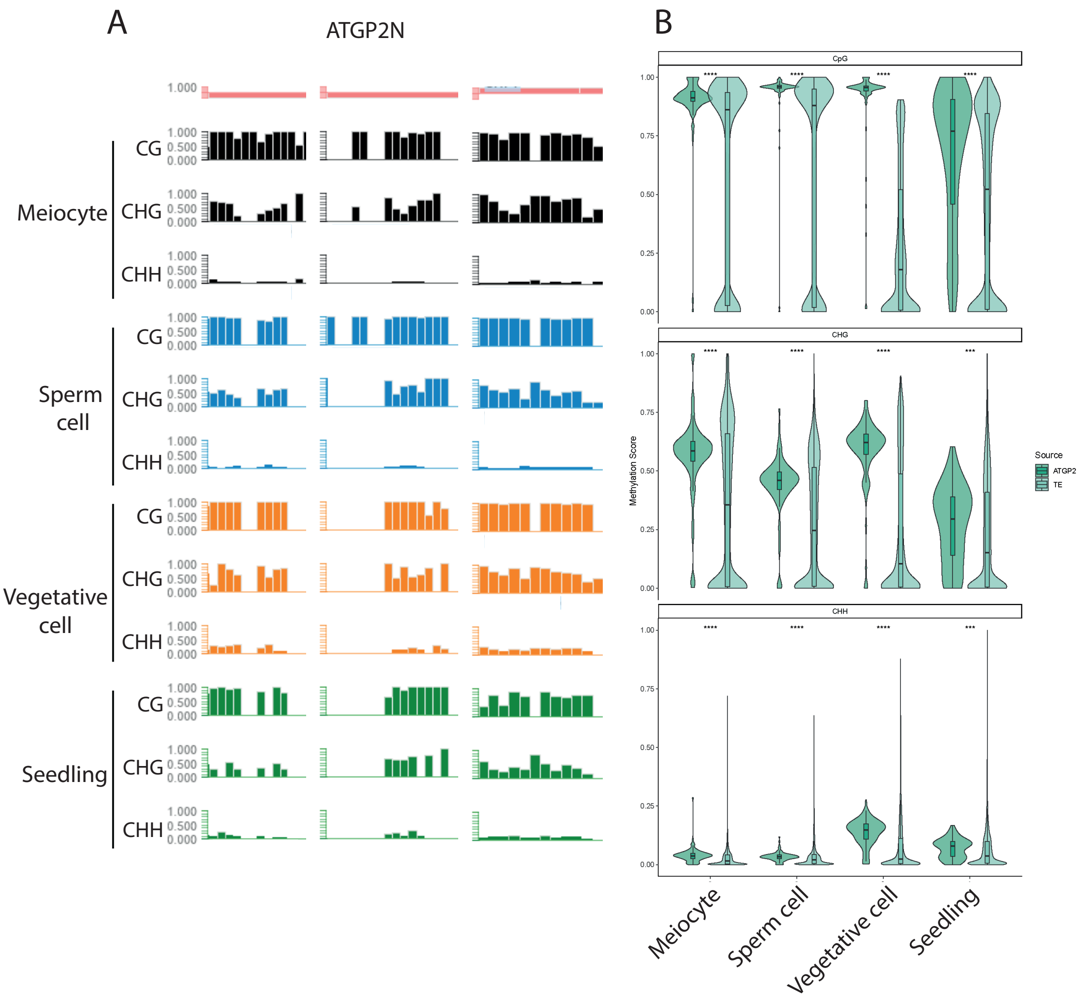


Fig. 2.18 ATGP2N TEs produce highly abundant 24nt sRNAs in the sperm cell, sperm nucleus and pollen

(A) Snapshot CHH methylation levels of three ATGP2N TEs which produces abundant sRNAs in the sperm. (A) Boxplots of CG, CHG and CHH methylation levels of the ATGP2 subset compared to all TEs in germline and somatic tissues. ***: p-value < 0.001, **: p-value < 0.01,*: p-value < 0.05

2.12 Discussion

The NRPD1 reporter line confirms the expected expression of RNA Polymerase IV in meiocyte stage anthers, namely that RNA Polymerase IV expression is restricted to the somatic tissue surrounding the sex cells and therefore 24nt sRNA production is quiescent in meiocytes (Figure 2.3). At the same time, the NRPE1 reporter line also confirms the expected expression; RNA Polymerase V is expressed throughout the somatic and germline tissue, facilitating DNA methylation. This pattern aligns with the data showing similar sRNA and DNA methylation profiles in the tapetum and meiocyte, including strong germline specific methylation at 797 HyperTE loci. At the same time, specifically in the Pol IV quiescent meiocytes, HyperTE-derived 24nt sRNAs from the tapetum can target HyperTEs for methylation in the meiocytes as well as 435 *de novo* genic targets (MetGenes) for methylation [60, 50].

As mentioned earlier, pNRPD1::NRPD1-eGFP expression was inconsistent in the germline of several independently transformed lines (Figures 2.4 and A4). This could be due to a variety of factors.

First, we need to consider the site of integration of the transgene during T-DNA insertion mediated by *Agrobacterium*-mediated transformation. If the T-DNA inserts into heterochromatic regions, which are typically transcriptionally inactive, or into dynamic chromatin regions that undergo changes during development, it can lead to variable expression patterns of the reporter gene [93]. Second, gene silencing mechanisms may also play a crucial role in transgene expression patterns. High transgene copy numbers can trigger RNA interference pathways, leading to PTGS and eventually may trigger TGS via RdDM, a phenomenon observed in *de novo* silencing of TEs [94]. Third, the choice of promoter driving the reporter gene is another critical factor. Promoters lacking necessary cis-regulatory elements or not interacting effectively with trans-acting factors may fail to achieve desired expression levels in specific tissues or developmental stages [95]. As the lines in question were constructed with native promoters, this effect could explain their inconsistent expression patterns. Finally, the use of fluorescent tags such as eGFP, fused to large proteins like NRPD1 (the largest subunit of RNA polymerase IV) can impede proper protein folding, localisation, or function due to the tag's size, leading to aberrant expression patterns.

Though generally DNA methylation levels in all sequence contexts declines through germline development overall, HyperTEs and MetGenes stay highly methylated from meiocyte to sperm [96, 50]. After meiosis, these results suggest that the Pol IV pathway becomes active in the germline once more, through endogenous sRNA production and DNA methylation. The unreliable expression observed by the several NRPD1 expression lines would need further validation. This could be achieved by constructing NRPD1/NRPE1 reporter lines using a stronger fluorophore, such as tdTomato, or by confirming that the RNA levels of the reporter constructs align with the expected transcript levels of the endogenous NRPD1/NRPE1, as tissue-specific silencing may affect the transgenic reporter lines.

CLSY chromatin remodelers recruit RNA polymerase IV in a locus-specific manner to drive 24nt sRNA production in both somatic and germline tissues. CLSY1 and CLSY2 are required for the association of SHH1 with the RNA polymerase IV complex, which recognises H3K9me2 histone modifications, linking CLSY1/2-dependent loci to this histone modification [89]. In contrast, CLSY3 and CLSY4 can recognise CG methylation to drive sRNA production at specific loci, independently of H3K9me2 [89]. CLSY1 and CLSY2 are specifically required to produce sRNAs at cRdDM loci in leaf tissue, while CLSY3 crucial for sRNA production at siren loci in ovules and HyperTE loci in the tapetum [44, 89, 60]. This tapetum-driven sRNA production also induces CHH methylation at HyperTE and MetGene loci in meiocytes, which persists throughout the germline. In fact, the production of sRNAs in the tapetum alone is sufficient to maintain HyperTE methylation in all germline tissues [60]. In accordance with this, CLSY3 is not expressed in the meiocytes or any of the later stage sex cells.

The CLSY1 and CLSY2 reporter lines indicate that these CLSY chromatin remodellers are not strongly expressed in maturing pollen either (Figure A2). This raises the question of how endogenous 24nt sRNA production is regulated in these germline cells. One possibility is that the reporter lines may be silenced or inaccurately expressed. To verify this, CLSY-eGFP transcript levels should be compared to RNA-seq data. RNA-seq data from our lab indicates that CLSY2 may be expressed in the vegetative cell and CLSY4 in the sperm cell (Table A2), though these findings need further confirmation, including checks on corresponding protein expression. Additionally, some CLSY1- and CLSY4-dependent loci still appear to produce sRNAs in the sperm cell (Figure 2.11). To conclusively determine whether any CLSY chromatin remodellers are important for sRNA production and CHH methylation in the pollen, clsy mutant sRNA data and methylation sequencing should be generated and compared with wild-type data.

In line with these findings, data analysis from the sRNA datasets seem to suggest that after meiosis, endogenous 24nt sRNA production shifts away from CLSY-mediated loci. Two hypotheses could explain this shift. First, DME-mediated hypomethylation in the vegetative cell leads to upregulated sRNA biogenesis from reactivated TEs, which in turn directs DNA methylation in the sperm cell [82, 97] via 21 and 22nt epigenetically active small interfering RNAs (easiRNAs). These easiRNAs are produced either by RNA polymerases II or IV and processed RDR2 or RDR6 and by DCL2/4 [98]. Though easiRNAs are primarily 21 and 22nt in length, the discovery that RNA polymerase IV can also generate them hints at the complexity of the interplay between various RdDM variant pathways. In fact, it is already known that several other non-canonical RdDM pathways can produce and direct 24nt sRNAs into the canonical RdDM pathway. One such pathway bypasses Pol IV-RDRs entirely, where double-stranded Pol II-derived transcripts (such as inverted repeats or microRNA precursors) are directly cleaved by DCL proteins to produce sRNAs [99]. The DCLs compete for these transcripts and produce 21-24nt sRNAs [100]), and the 24nt sRNAs can feed into the Pol V pathway via AGO4 either in *cis* or at other similar loci in *trans* [101, 99]. Alternatively, post-transcriptional gene silencing (PTGS) has also been linked to the RdDM pathway. In this pathway, dsRNA is first produced by PTGS via Pol II and RDR6 which is then cleaved by either DCL2/4, or DCL3 to produce 21,22 or 24nt sRNAs respectively, in a hierarchical manner [94], which can be loaded onto AGO proteins and feed into the Pol V pathway this way [84].

Several accessory proteins are involved in the RdDM pathway [35], many of which have paralogs in *Arabidopsis*. While some of these proteins may have redundant functions, RNA-seq data generated by members of our lab shows that specific Pol IV and Pol V accessory proteins are expressed at specific developmental stages in the male sexual lineage. I would like to highlight three specific examples and discuss their roles.

First, a paralog of SPT5L/KTF1, which recruits AGO proteins through an AGO hook motif [42, 102] and interacts with NRPE1 [103, 104], is notably upregulated in meiocytes and pollen. Second, RNA polymerase V transcription is facilitated by the DDR complex, a key component of which, a DRD1 homolog, is specifically upregulated in the meiocyte (Table A2). Finally, the MORC protein family, which consists of ATPases involved in chromatin compaction and transcriptional gene silencing in mammals, also plays a crucial role in the canonical RdDM pathway in plants. MORC proteins co-localise to sites of RdDM and are important for the establishment of RdDM. One paralog, MORC5 is specifically upregulated in the meiocyte (Table A2) and has been previously shown to be specifically expressed in the pollen [105], while MORC4 is also upregulated in the meiocyte, pollen and embryo (Tables A2 and A3). These observations hint at possible candidates that could contribute to tissue-specific non-canonical RdDM pathways in the male sexual lineage.

Given the critical importance of genome integrity in reproductive cells, the precise regulation and silencing of TEs are essential. In accordance with this, 568 loci with abundant 24nt sRNAs also have high H3K9me2 methylation, which is associated with heterochromatic regions, however, these loci are not associated with increased CG, CHG or CHH methylation levels when compared to the whole set (Figures 2.13 2.14. This may suggest that this subset represent TEs that are activated in the vegetative cell, which are situated in pericentromeric regions, enriched in H3K9me2, and produce 24nt sRNAs that target DNA methylation in sperm cells [106]. Nevertheless, further careful study is necessary to determine the specific significance of this group of TEs and to clarify whether they overlap with the heterochromatic TEs activated in the vegetative cell.

The LTR/Gypsy TE family ATGP1 has previously been shown to be specifically active in reproductive cells of RdDM mutants [60]. Investigating the 24nt sRNA landscape at TE loci identified another LTR/Gypsy TE family, ATGP2N, which produce abundant sRNAs specifically in the sperm cell (Figure 2.17). ATGP2N has previously been reported to gain chromatin accessibility and be transcriptionally upregulated in *met1* mutants. In this study, ATGP2N was part of a subset of TEs that gained chromatin accessibility and increased in transcription as well as an increase in the production of 21nt sRNAs in the inflorescences of *met1* mutants. The other examined groups of TEs gained accessibility but had no change in transcription or sRNA production, one already being highly expressed in wild type plants, and the other remaining repressed with respect to transcription and sRNA production even with increased chromatin accessibility[107]. Although CHH methylation levels of TEs both germline and somatic tissues were generally low, the ATGP2N family had slightly increased CHH methylation in all tissues, most notably in the vegetative cell (Figure 2.18B and C). Several explanations could account for these findings.

Firstly, as around 40% of these ATGP2N loci overlap CLSY3/4 dependent loci, it is possible that these chromatin remodellers are responsible for the cell-specific silencing of these TEs in the sperm cell. An important question to address empirically is whether, in addition to the 21 and 22nt easiRNAs, 24nt sRNAs produced in the hypomethylated vegetative cells can direct methylation in the sperm cells, similar to how tapetal cells mediate methylation at specific loci. Secondly, another potential explanation for this sRNA and methylation pattern is related to the general observation that most TEs are more hypomethylated in the vegetative cell compared to the sperm cell, which leads to the production of abundant easiRNAs that direct RdDM in sperm. However, since ATGP2N exhibits increased CHH methylation in the vegetative cell compared to the sperm cell, it is plausible that the abundant 24nt sRNAs detected in the sperm cell are produced through the non-canonical PolII-RDR6 or PolII-DCL3 RdDM pathways, due to ATGP2N reactivation specifically in the sperm cell. To verify these hypotheses, it will be necessary to examine ATGP2N expression levels in the germline of wild-type and RdDM mutants, as well as analyse 24nt sRNA production and CHH methylation in clsy3/4 double mutants.

Notably, around 20% of MetGenes produce perfect-matching 24nt sRNAs uniquely in the sperm cell (Figures 2.12, 2.15) and these genes specifically gain CHH methylation in the sperm cell (Figure 2.16C). This could be an off-target effect of sRNAs produced in the vegetative cell due to hypomethylation of TEs, a relationship that warrants further investigation. Alternatively, this CHH methylation gain may have biological significance. While CHH methylation is often linked to transcriptional repression, this is not always the case. For example, gene body methylation near transposon-gene boundaries is associated with transcriptionally active genes in maize [108]. Additionally, the loss of RdDM-directed gene body methylation of MetGene MPS1 has been proven to cause mis-splicing and disrupt meiosis [50].Among the MetGenes producing 24nt sRNAs in the sperm cell are VIM2 and VIM4. The VIM proteins are required for maintaining CG methylation by recruiting MET1 to hemimethylated CG sites [109]. In addition, VIM5 was identified as one of few paternally imprinted genes in *Arabidopsis* and is involved in mediating CG methylation in the endosperm [110]. RNA-seq data provides evidence that VIM proteins are specifically upregulated in the sperm cell and VIM1 specifically upregulated in the vegetative cell (Table A4), suggesting a potential connection between tissue-specific gene body CHH methylation and transcriptional regulation. This differential expression may play a crucial role in maintaining epigenetic control and ensuring correct gene expression and TE silencing in these important tissues.

Together, these observations highlight that, in addition to producing 24nt sRNAs from cRdDM loci, sperm cells display a unique sRNA profile distinct not only from earlier germline cells and somatic tissues but also, to some extent, from vegetative cells. This difference may be driven by non-canonical RdDM pathways and may shift away from CLSY chromatin remodellers regulating RdDM after meiosis. The tissue-specific expression of several key RdDM components in the male sexual lineage suggests potential mechanisms that could allow mismatch-targeted DNA methylation in meiocytes and underscores the significance and complexity of non-canonical RdDM pathways active during post-meiotic male sexual development.

2.13 Materials and Methods

2.13.1 Plant materials, plasmid construction and growth conditions

Arabidopsis thaliana plants were grown in 16h light/8h dark conditions, in 21°C, 70% humidity in a controlled environment chamber on germination medium (GM) without glucose supplementation. The following lines have been grown: Col-0 wild type plant for the microspore and pollen sRNA sequencing libraries. For the anther imaging experiments, the transgenic reporter lines pCLSY1::CLSY1-eGFP (clsy1), pCLSY2::CLSY2-eGFP (WT), pCLSY3::-CLSY3-Venus (clsy3), pNRPD1::NRPD1-eGFP (WT) and pNRPE1::NRPE1-eGFP (WT) were used with pAGO4::AGO4-Venus or pCLSY3::-CLSY3-Venus (clsy3) for a positive control.

Plants transformed with pCLSY3::-CLSY3-Venus (clsy3) and pAGO4::AGO4-Venus (WT) vectors were obtained from dr James Walker and Sam Deans (Feng lab) who constructed the plasmids and reporter lines. The plasmids for the CLSY1 and CLSY2 reporter lines as well as the NRPD1 and NRPE1 reporter lines were designed by and the transformation performed by dr Jincheng Long (Feng Lab). The plasmids were constructed using restriction cloning using pCambia1300 as a backbone. The *Arabidopsis thaliana* plants transformed with pNRPD1::NRPD1-eGFP and pNRPE1::NRPE1-eGFP were genotyped and selected (both based on antibiotic resistance and fluorescence), as referenced in 2.3 and Figure A3. Please find a summary of all the plant materials and where they were obtained in Table A1.

2.13.2 Microscopy

Meiocyte and microspore stage anthers were dissected using a Leica dissecting microscope on 0.8% agar and vacuum infiltrated in DAPI buffer (Galbraith's buffer (45 mM $MgCl_2$-$6H_2O$, 30 mM trisodium citrate, 20 mM MOPS, pH 7.0), 1 µg/mL DAPI, and 0.01% [v/v] Triton X-100) briefly, then imaged using imaging spacers (SecureSealTM Grace BioLabs). Individual microspores and pollen were imaged after isolation from unopened flower buds and opened flowers respectively, by collecting 500 µL of flower buds into DAPI buffer, briefly vortexing and pelleting cells. All samples were imaged using Leica Stellaris 8 and Leica SP8X confocal microscopes.

2.13.3 Isolation of microspores and pollen using fluorescence activated cell sorting

Arabidopsis thaliana microspores were isolated by manually collecting microspore stage flower buds, using the morphology described previously [111]. The buds were collected into 1.5mL microcentrifuge tubes and suspended in PEB (10 mM $CaCl_2$, 2 mM MES, 1 mM KCl, 1% H_3BO_3, 10% Sucrose, pH 7.5) buffer. The microspore cells were isolated as described previously [112]. Briefly, the buds were gently ground in a clean pestle and mortar in PEB buffer and filtered through Miracloth (Merck-Millipore 475855) into a 1.5mL Eppendorf tube. This crude fraction was centrifuged for 5 minutes at 800g to gently pellet the microspore cells. The pellet was resuspended in 500µL PEB and sequentially filtered through 30µm and 20µm mesh filters (CellTrics®).

The microspores were isolated using fluorescence-activated cell sorting (FACS) on a BD FACSMelody cell sorter (Beckton Dickinson) into TRI reagent (Zymo Research). The purity of the sorted cells was inspected using a widefield microscope, and FACS purification was performed for the whole fraction when the purity of the cell extract was above approximately 80% (judged visually). The quality of the cell extract was further confirmed by the presence of a microspore cell specific population cluster by the cell sorter as described in Figure 4 of [112]

Sperm cells were isolated as previously described [112, 113] except the sperm cell release step was repeated 4 times to maximise sperm cell release from pollen grain. Sperm cells were stained using SYBR green and SITOX orange and sperm cells and nuclei were sorted into TRI reagent (Zymo Research). The quality of the cell extract was confirmed according to the protocol described in [112, 113] and the sperm cells were isolated using a FACS cell sorter according to the gating parameters described in Figure 2 of [112].

2.13.4 sRNA sequencing

RNA was released from isolated microspores by vortexing the samples with RNase free glass beads (Sigma-Aldrich) for 4 minutes before proceeding with RNA isolation. RNA was isolated from all samples using the Direct-zol™ RNA kit (Zymo Research, CN: R2061).

2 replicates of sRNA sequencing libraries were constructed per cell type (microspore, sperm cell, sperm nuclei) using the RealSeq®-biofluids NGS Library Preparation Kit for miRNAs and sRNAs. For the three library types, the number of cells used for each replicate were as follows: 300 000 cells for microspore, 600 000 cells for sperm cells, and 1 500 000 cells for sperm nuclei.

Briefly, the RealSeq® adapters were ligated and blocked, then the sample was circularised, and the adapter dimers were removed. Then reverse transcription was performed as well as PCR amplification and size selection to end up with the final library.

Further, gel separation and purification was performed to enrich the sRNA libraries for the desired fragment lengths (20-30 nucleotides long) as described previously [60]. The libraries were run on Novex TBE 6% gels (Invitrogen, CN: EC6265BOX) along with a custom RNA ladder consisting of bands that were 146bp and 161bp in length.

The gel was run for 45 mins with 120V and then stained using ethidium bromide. The gel was illuminated with UV light and the area between the two bands were then cut out using a clean razor blade and the gel slice was weighed. 2 volumes of elution buffer (0.5 M ammonium acetate, 10 mM magnesium sulfate, 1 mM EDTA (pH 8.0), 0.1% SDS) were added to the gel slice. The sample was incubated at 37°C on a rotary platform (1000 RPM) for 4 hours.

The sample was then centrifuged at 12000g for 1 minute at 4°C in a microcentrifuge. The supernatant was transferred into a fresh 1.5mL Eppendorf tube. An additional 200µL of elution buffer was added to the polyacrilamide gel fragment, vortexed, re-centrifuged and the two supernatants were combined. Any remaining fragments of polyacrilamide were removed by passing the supernatant through a plastic column containng cellulose acetate filters. Two volumes of cold absolute ethanol were added, and the solution was incubated on ice for 30 minutes. The DNA was recovered by centrifugation at 12000g got 10 minutes at 4°C. The pellet was dried and resuspended in 200µL TE buffer (10mM Tris-HCl, 1mM EDTA, pH 7.6) and 25µL of 3M sodium acetate (pH 5.2) was added. The precipitation step was repeated, the pellet was rinsed with 70% ethanol and thoroughly dried before resuspending in TE buffer to a final volume of 30µL. The sample was incubated at 4°C overnight to redissolve the DNA.

The concentration and size distribution of purified sRNA libraries were determined using High Sensitivity DNA Chips (Agilent Technologies, CN: 5067-5594).

The sRNA libraries were sequenced using a NextSeq 500 (Illumina).

2.13.5 Bioinformatics and data analysis

Adaptor trimming and filtering of reads that are 18nt to 28nt in length was performed on the raw sequencing data using Cutadapt [114] v1.9.1). These reads were mapped to the *Arabidopsis thaliana* reference genome TAIR10 using Bowtie [115]. 24nt clusters were extracted and loci were kept that overlapped previously determined non-overlapping genomic features of interest such as canocical RdDM loci (9629 loci), HyperTEs (797 loci), MetGenes (435 loci), transposons (TEs), genes17,20, and H3K9me2 enriched loci. These overlapping feature files were produced using BEDtools (v2.27.0) [116]. 24nt sRNA clusters were extracted using ShortStack [117] (n = 45639). To firstly identify true 24 nucleotide clusters, the ShortStack clusters from the two replicates were filtered to keep clusters where the predominant read length was 24 nucleotides long, where log2(RPKM+1) 24nt sRNA > log2(RPKM+1) 25nt sRNA +2 and that were >99bp (n = 34032).

Following this general filtering step, sperm cell specific clusters were identified where: log2(RPKM+1) 24nt sperm cell sRNA > 0 & where: log2(RPKM+1) 24nt sperm cell sRNA - log2(RPKM+1) 24nt (other tissues: tapetum, meiocyte, microspore, root seedling and leaf) sRNA > 3 which resulted in 513 loci.

The euchromatic/heterochromatic annotation was allocated based on data from reference [118] and classified into 5 chromatin states based on H3K9me2/H3: euchromatic (less than 0.6), intermediate euchromatic (bw. 0.6 and 0.9), intermediate (bw 0.9 and 1.4), intermediate heterochromatic (bw. 1.4 and 2.6), heterochromatic (above 2.6).

Figures were plot using ggplot2 and pheatmap packages in R (v4.0.3). Please find all scripts used on my personal GitHub page.

Chapter 3

The dynamic sRNA profile and inheritance of paternal epigenetic modifications post-fertilisation in the embryo of *Marchantia polymorpha*

3.1 Abstract

Global cytosine methylation reprogramming has recently been reported in the male sexual lineage of *Marchantia polymorpha*. During spermiogenesis, global 5mC expansion occurs, followed by the expansion of 4mC onto CG sites. Following fertilisation, the mature embryo loses this epigenetic mark. Therefore, the fate of paternal N4-methylcytosine methylation is explored in early *Marchantia* embryos. The results suggest that 4mC is actively removed prior to pronuclear fusion. This is further corroborated by exploring the phenotype of embryos fertilised by 4mC knockout sperm using fluorescent reporter lines, which exhibit accelerated embryonic development and a partial fertility cost.

3.2 Introduction

As discussed in Chapter 1, DNA methylation is an important epigenetic modification in both prokaryotes and eukaryotes. In eukaryotes the most prevalent DNA modification is 5mC, important in gene regulation, imprinting, silencing TEs and responding to environmental stimuli. In prokaryotes, in addition to 5mC, additional base modifications, such as N6-methyladenine (6mA) and 4mC are also present and play key roles in restriction-modification systems and in marking self DNA to distinguish it from foreign DNA (Figure 3.1) [119].

DNA methylation reprogramming during sexual development is a common feature in eukaryotes. As discussed in previous chapters, the epigenetic reprogramming in

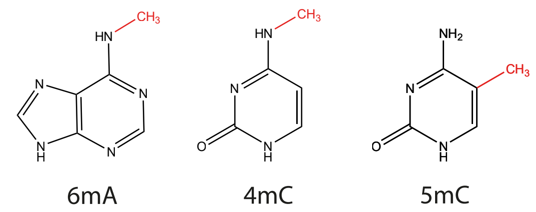


Fig. 3.1 The structure of eukaryotic and prokaryotic base modifications N6-methyladenine, N4-methylcytosine and 6-methyladenine.

the male sexual lineage of *Arabidopsis* is extensive and is a characteristic shared by other flowering plants. However, given that male sexual lineage development and meiosis in *Arabidopsis* occur within a few cell divisions (Figure 1.3), our group aimed to investigate whether gene-targeted methylation is specifically associated with meiosis or spermiogenesis.

The lifecycle of the basal land plant *Marchantia polymorpha* alternates between haploid and diploid phases. In the haploid vegetative phase (thallus), they can reproduce asexually or be induced by far-red light to produce gametophytes. These gametophytes form antheridiophores (which produce antheridia and sperm in males) and archegoniophores (which produce archegonia and egg cells in females). Once the egg cell is fertilised by sperm, *Marchantia* proceeds to go through a rapid diploid sporophytic phase, which culminates in meiosis and the dispersal of haploid spores (Figure 3.2) [120].

Recently, our group demonstrated that during spermiogenesis in *Marchantia*, there is extensive DNA methylation reprogramming. In addition to the deposition of 5mC, a second wave of DNA methylation reprogramming occurs, involving the deposition of N4-methylcytosine (4mC) — an epigenetic modification previously thought to exist only in prokaryotes — at the majority of CG sites across the genome (Figure 3.3). A 4mC methyltransferase, MpDN4MT1, was identified as the enzyme catalysing genome-wide 4mC methylation. Knockout of the *dn4mt1* gene results in the loss of 4mC, partial fertility defects, and impaired sperm motility [121].

Paternal 4mC methylation is absent in the mature embryo (which contains over 1000 cells). The exact timing of when 4mC methylation is lost post-fertilisation remains unknown, as does whether the paternal genome retains any 4mC marks during the early stages of zygotic development. In *Marchantia*, the first division of the zygote occurs approximately 3–4 days after fertilisation (DAF) [122], suggesting that 4mC may need to be actively removed for pronuclear fusion to take place. This hypothesis aligns with the observed phenotype where embryos fertilised by *dn4mt1* sperm exhibit accelerated development (Figure 3.4, data collected by dr Yalin Liu), though this phenomenon still requires confirmation during early embryonic stages. Alternatively, 4mC methylation could be lost passively following fertilisation through dilution by cell division, or

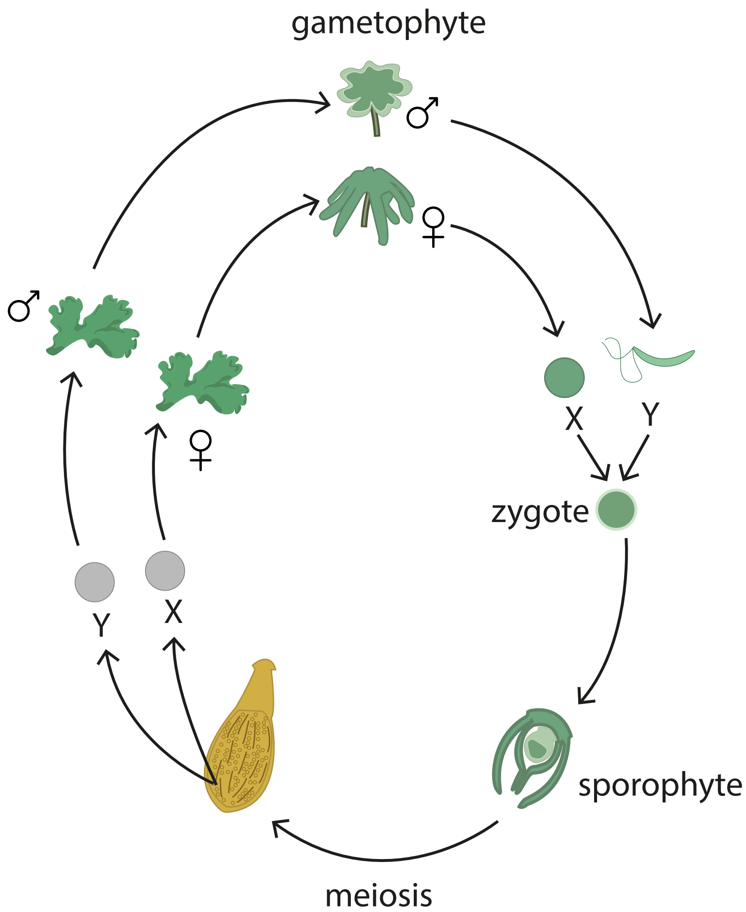


Fig. 3.2 The lifecycle of *Marchantia polymorpha*

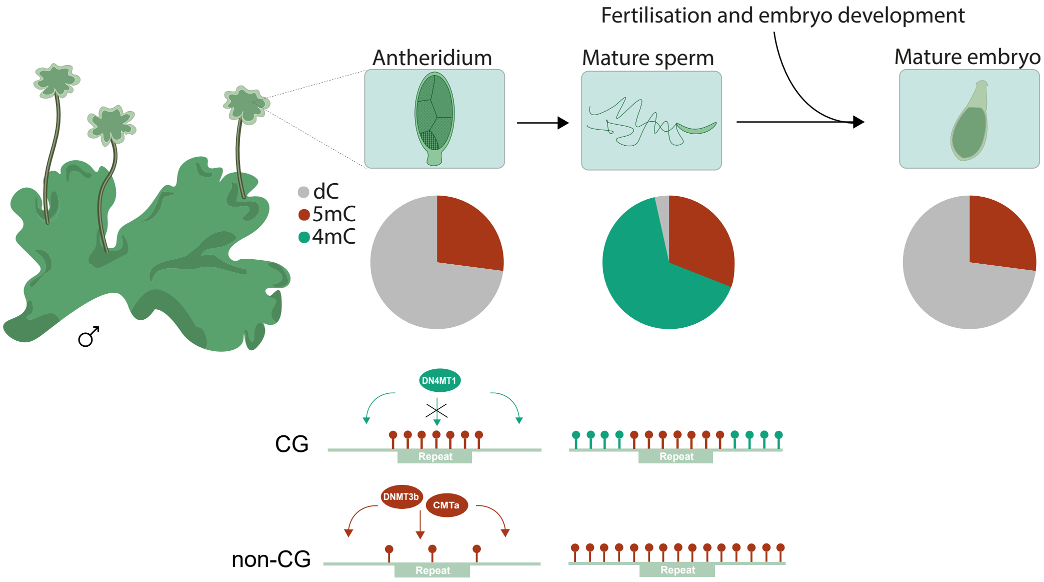


Fig. 3.3 The landscape of DNA methylation during the spermiogenesis and embryogenesis of *Marchantia*

In the liverwort *Marchantia polymorpha* it was recently reported [121] that during spermiogenesis, two waves of DNA methylation reprogramming occur. First 5mC is deposited (brown) facilitated by DNA methyltransferases DNMT3b and CMTa followed by a dramatic second wave, wereby 4mC expands to genic regions to cover the majority of CG sites in sperm (green), catalysed by a novel methyltransferase MpDN4MT1. Following fertilisation, 4mC methylation is lost in the embryo. Figure adapted from [121].

actively removed post-fertilisation for DNA replication or cell division to function normally. Additionally, 4mC may play a role in ensuring correct genome dosage and transcriptional activity during early embryonic development.

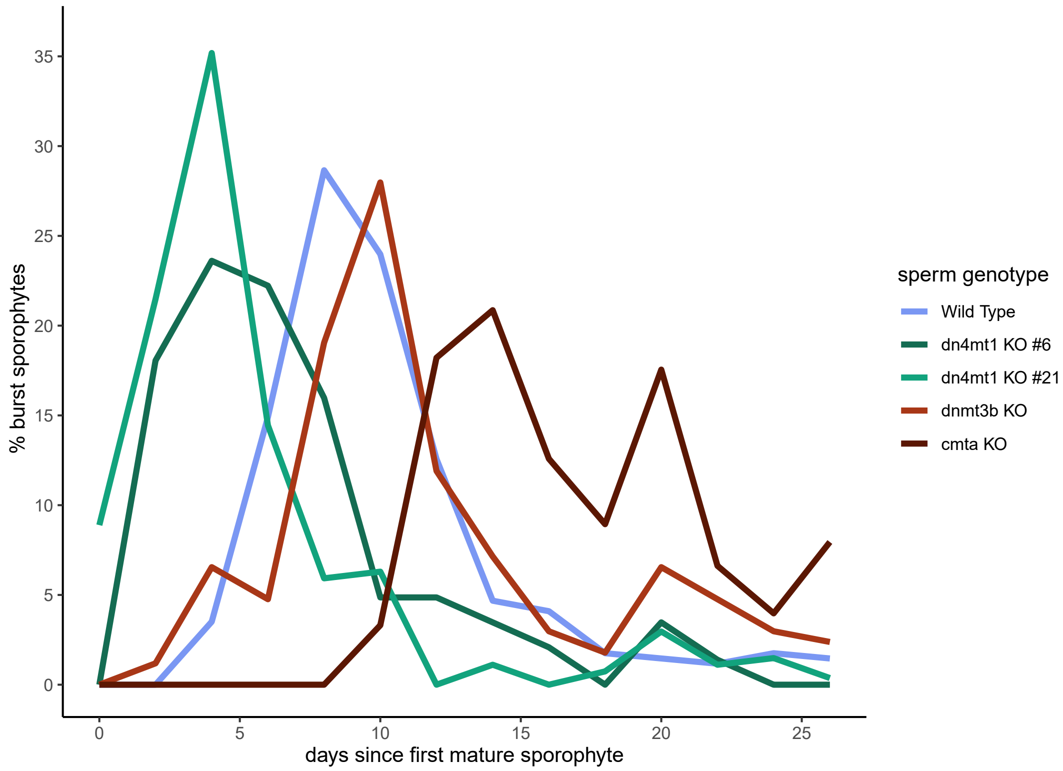


Fig. 3.4 Embryos fertilised by *dn4mt1* knockout sperm develop more rapidly than WT

Total percentage of burst sporophytes every day since first mature sporophyte, fertilised with wild type sperm (blue), two independent lines of *dn4mt1* knockout sperm (#6 dark green, #21 light green) or 5mC (*cmta* (dark brown) or *dnmt3B*) knockout sperm (light brown). Data collected by dr Yalin Liu.

In *Arabidopsis*, 5mC methylation is actively removed to maintain homeostasis through the action of several DNA glycosylases. *Arabidopsis* encodes four DNA demethylases: DEMETER (DME), REPRESSOR OF SILENCING 1 (ROS1), and the DEMETER-LIKE proteins DML2 and DML3. While DME is involved in vegetative cell demethylation in pollen [82] (as discussed in Chapter 2), it also plays a critical role in the allele-specific expression of maternally imprinted genes in the central cell and endosperm [81]. The remaining DNA glycosylases appear to be functionally redundant, primarily preventing the spread of transposon-derived 5mC methylation into genic regions[123], including ROS1 [124]. Therefore, the orthologs of ROS1 in *Marchantia*, MpROS1a and MpROS1x (located on the female sex chromosome) offer possible candidates for erasing paternal 4mC methylation as they are expressed in the sperm [125] and mature eggs/embryos [126] respectively [127, 126, 121].

Recent findings have shown that the heterochromatic histone modification H3K27me3 is exclusively deposited on the paternal genome during the pronuclear stage of the embryogenesis of *Marchantia*, leading to the repression of the paternal genome, which persists throughout embryogenesis [128]. In contrast, in mammalian germlines, H3K27me3-mediated imprinting is limited to a few loci within the female gamete [129]. 4mC therefore offers a possible mechanism through which the paternal genome could be marked during early development.

Hence, investigating the methylation profiles of early *Marchantia* embryos during the first zygotic divisions and examining the early developmental phenotype of *dn4mt1* mutants complement our previous work, expanding our understanding of how paternal DNA methylation affects early embryonic development.

3.3 Paternal 4mC is lost in the early embryo of *Marchantia*

To establish an effective method for staging early *Marchantia* embryos and identify the earliest feasible window for embryo extraction for AMD sequencing (AMD-seq), several approaches were explored. Despite incorporating detergents, vacuum-assisted staining, and testing various stains (DAPI, Hoechst 33342, and ethidium bromide), live-cell staining of developing embryos within the archegonium remained inconsistent. This inconsistency likely arises from the embryo's position within a cavity (the venter) surrounded by somatic cells and cell walls (Figure 3.5 rows one and two). Moreover, variations in environmental conditions during embryo development significantly impacted growth, as illustrated in Figure B1 which shows the same genotype embryos 10 days post-fertilisation under different growth conditions. To ensure consistency and reproducibility, a controlled crossing method was adopted, wherein archegoniophores were exposed to sperm in a tube for an hour, ensuring synchronised fertilisation and uniform embryo development. This approach is particularly suited for studying early embryo development, as embryos can be cultured under these conditions for up to two weeks [122]. Under these conditions, the earliest developing embryos could be manually dissected was around 7–8 days after fertilisation (Figure 3.5).

As previously established, 4mC methylation is deposited in the CG context over gene bodies and outside TEs during spermiogenesis by MpDN4MT1 (Figure 3.6A)[121]. Although 5mC methylation across all sequence contexts over TEs increases during germline development, non-CG methylation is specifically deposited during spermiogenesis by MpCMTa and MpDNMT3B (Figure 3.6B,D and F)[121]. Additionally, this methylation pattern is deposited and maintained over gene bodies (excluding the transcription start site, TSS), with 4mC serving as the primary methylation mark at the TSS.

As early embryos are difficult to harvest, first AMD-seq (sequencing specifically 4mC) and EM-seq (sequencing of 5mC and 4mC) was tested on wild type sperm and compared to data previously collected from wild type sperm and *dn4mt1* sperm (Figure 3.6 WT sperm rep2 AMD-seq (light green) and *dn4mt1* sperm AMD-seq (grey)). The expected methylation patterns over genes and TEs in the 4mC and 5mC contexts were confirmed by the sperm AMD-seq and EM-seq libraries (Figure 3.6 wild type sperm rep2 AMD-seq (light green), wild type sperm EM-seq (dark brown).)

Thereafter, pre-meiotic embryos were dissected and sequenced using AMD-seq to establish a baseline for comparing early embryonic methylation patterns (Figure 3.6, 14 DAF embryos (orange)). The results confirmed the absence of 4mC methylation over genes or outside TEs (Table 3.1). Subsequently, 300 embryos at 7–8 DAF were collected, with 200 used for AMD-seq library construction and the remaining 100 embryos used for bisulfite sequencing libraries, pre-treated with APOBEC3A for an

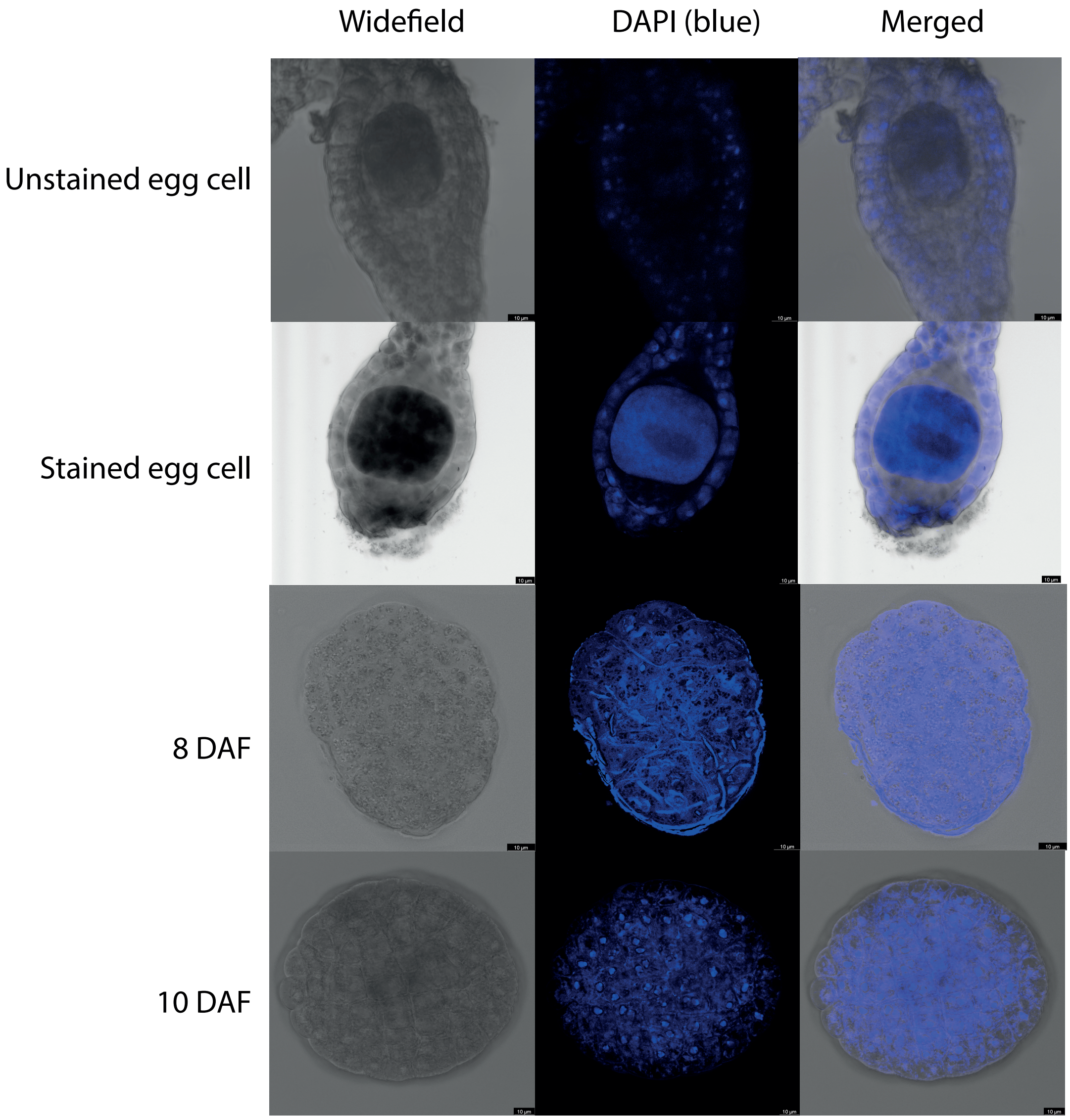


Fig. 3.5 Live cell imaging of the developmental stages of *M. polymorpha* embryos

Unstained egg cell (first row), DAPI stained egg cell (second row), and embryos 8 days (third row), and 10 days (fourth row) after fertilisation. Scale bar 10μm

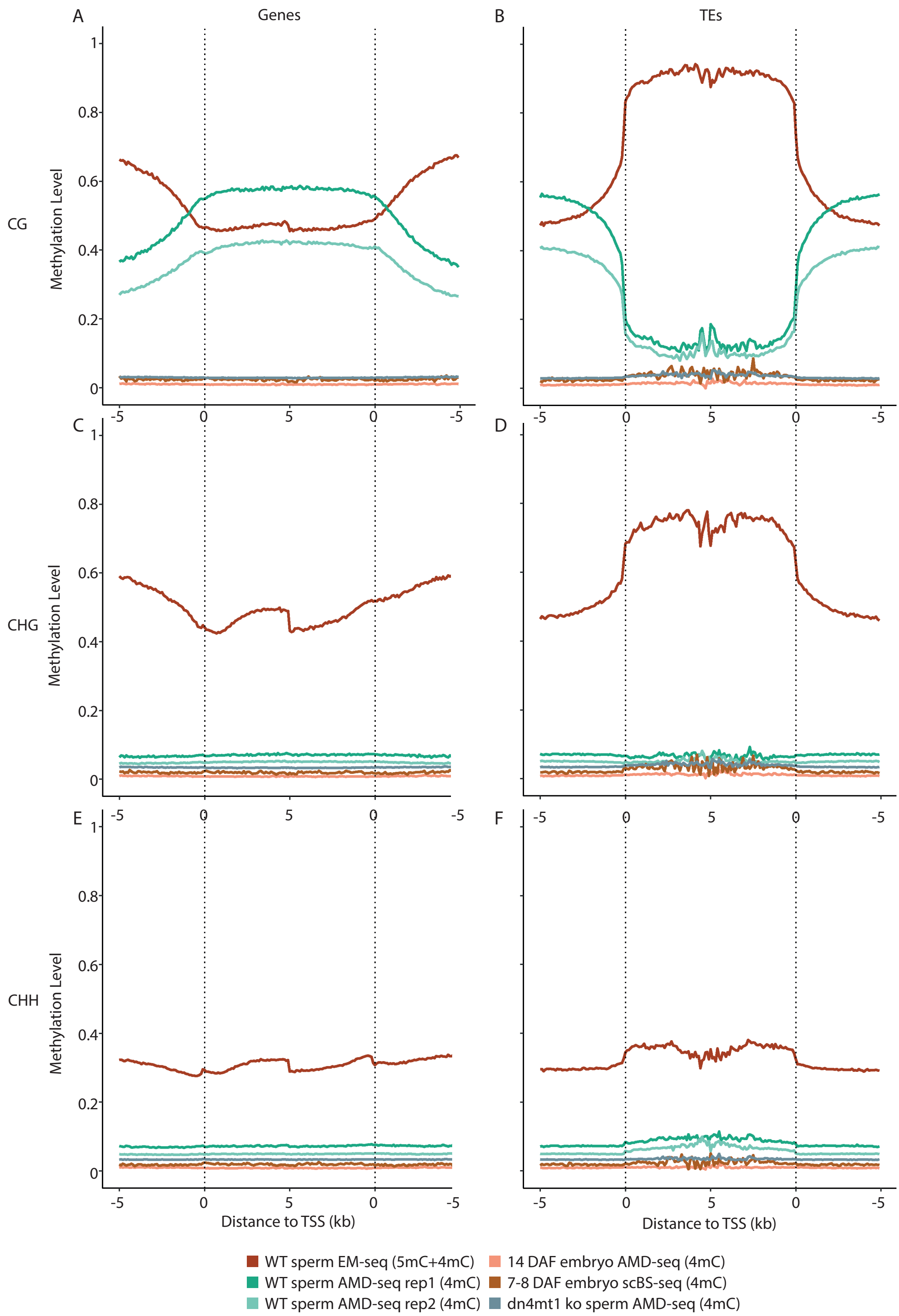


Fig. 3.6 4mC methylation occurs in the CG context and is enriched in genic regions and outside TEs, while 5mC methylation is largely confined to TEs

Panel showing 4mC and 5mC methylation across TEs (A) and Genes (B) in CG (first row), CHG (second row) and CHH (third row) sequence contexts in wild type sperm *dn4mt1 sperm*, the embryo 7-8 and 14 days after fertilisation.

alternative 4mC sequencing method. Although AMD-seq is theoretically capable of handling low DNA input down to picogram levels [130], this was not tested prior to sequencing the early embryos. The bisulfite sequencing protocol, on the other hand, was developed for low-input or single-cell sequencing and this has been robustly tested in our lab. Due to conversion issues, the AMD-seq library was not usable in this study.

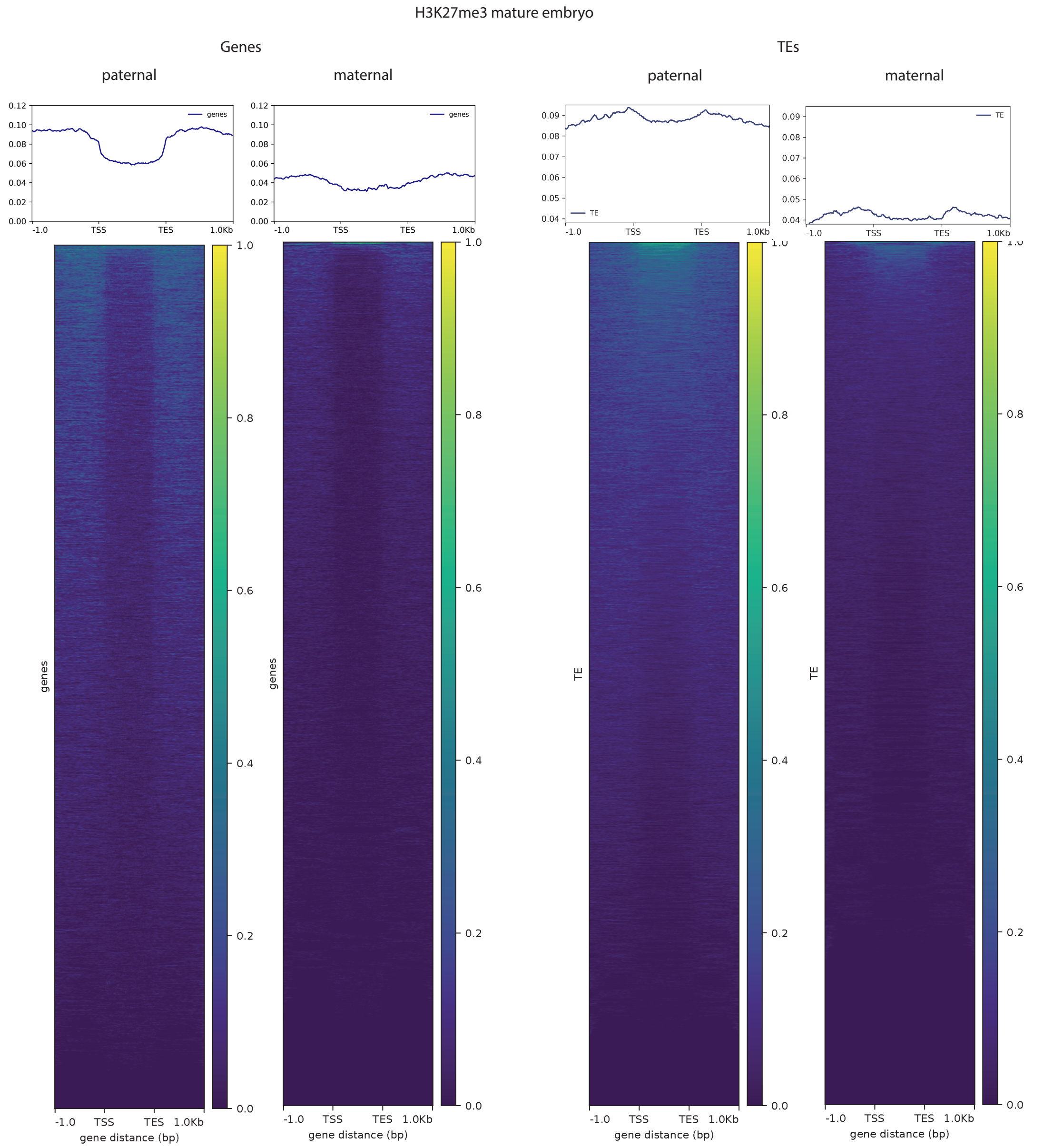


Fig. 3.7 H3K27me3 levels are elevated in the paternal genome and correlate with the presence of 5mC over genes

Heatmaps showing the relative H3K27me3 levels in *Marchantia* embryos over the genes and TEs in the paternal and maternal genomes. Data presented is from [128]

At a surface level, the early embryo AMD-seq data indicated marginally higher 4mC methylation across genes compared to the mature AMD-seq library (Table 3.1). However, this increase fell within the error range of the experimental limitations (Table 3.1, "CHG and CHH levels of 4mC"). Moreover, the low levels of residual 4mC methylation did not display the expected distribution over genes and TEs, with no relative increase outside TEs or over genes. This suggests that the removal of paternal 4mC methylation may be essential for embryo development. However, starting with a methylation level of around 20% in the zygote (based on the sperm AMD-seq levels), passive loss of 4mC by the 16–32 cell stage (7–8 days after fertilisation in our conditions) would result in very low residual 4mC methylation levels (Table 3.1).

It has been previously demonstrated that during embryonic development the paternal genome is transcriptionally silenced and is marked with the repressive chromatin mark H3K27me3 mediated by maternally expressed Polycomb Repressive Complex 2 (PRC2) in the male pronucleus before the first zygotic division, approximately 3 days after fertilisation [128]. This suggests the existence of a mechanism for paternal genome recognition, with blanket 4mC methylation being a potential candidate. To this end, visualizing H3K27me3 data across genes and TEs from this study confirms a paternal bias in the deposition of H3K27me3 (Figure 3.7 right side). However, the pattern of H3K27me3 occupancy around gene and transposon transcription start sites does not align with the distribution of 4mC (Figure 3.7 left side).

Methylation context	Overall			Genes		
	CG	CHG	CHH	CG	CHG	CHH
WT sperm AMD seq rep1	0.467	0.071	0.075	0.559	0.071	0.072
WT sperm AMD seq rep2	0.341	0.050	0.051	0.404	0.051	0.049
dn4mt1 ko sperm	0.031	0.035	0.033	0.029	0.034	0.033
14 DAF embryo	0.011	0.009	0.009	0.009	0.008	0.008
7-8 DAF embryo	0.027	0.021	0.019	0.025	0.020	0.020
Theoretical methylation levels						
Zygote	0.202	0.030	0.031	0.241	0.030	0.030
16 cell embryo	0.013	0.016	0.015	0.015	0.015	0.015
32 cell embryo	0.006	0.007	0.007	0.008	0.007	0.007

Table 3.1 Methylation levels across WT sperm, *dn4mt1* sperm and the early embryo

3.4 *De novo* methylation is deposited in genic regions specifically in the sporophyte of *Marchantia*, targeted for methylation 24nt sRNAs through mismatches

As described previously, *de novo* methylation of genes in *Arabidopsis* male meiocytes is mediated by 24nt sRNAs produced by HyperTE loci in the tapetum, the biogenesis of which is mediated by CLSY3 chromatin remodeler. Gene body methylation also exists specifically in the embryo and sporophyte of *Marchantia* (Figure 3.8, data from Dr. James Walker). Sporophytic non-CG DMRs were determined as described before[131] and filtered to yield 221 methylated genic loci (MetGenes).

With access to sRNA sequencing libraries from thallus, embryo, and sporophyte, the next step was to determine whether MetGenes could be targeted for methylation through

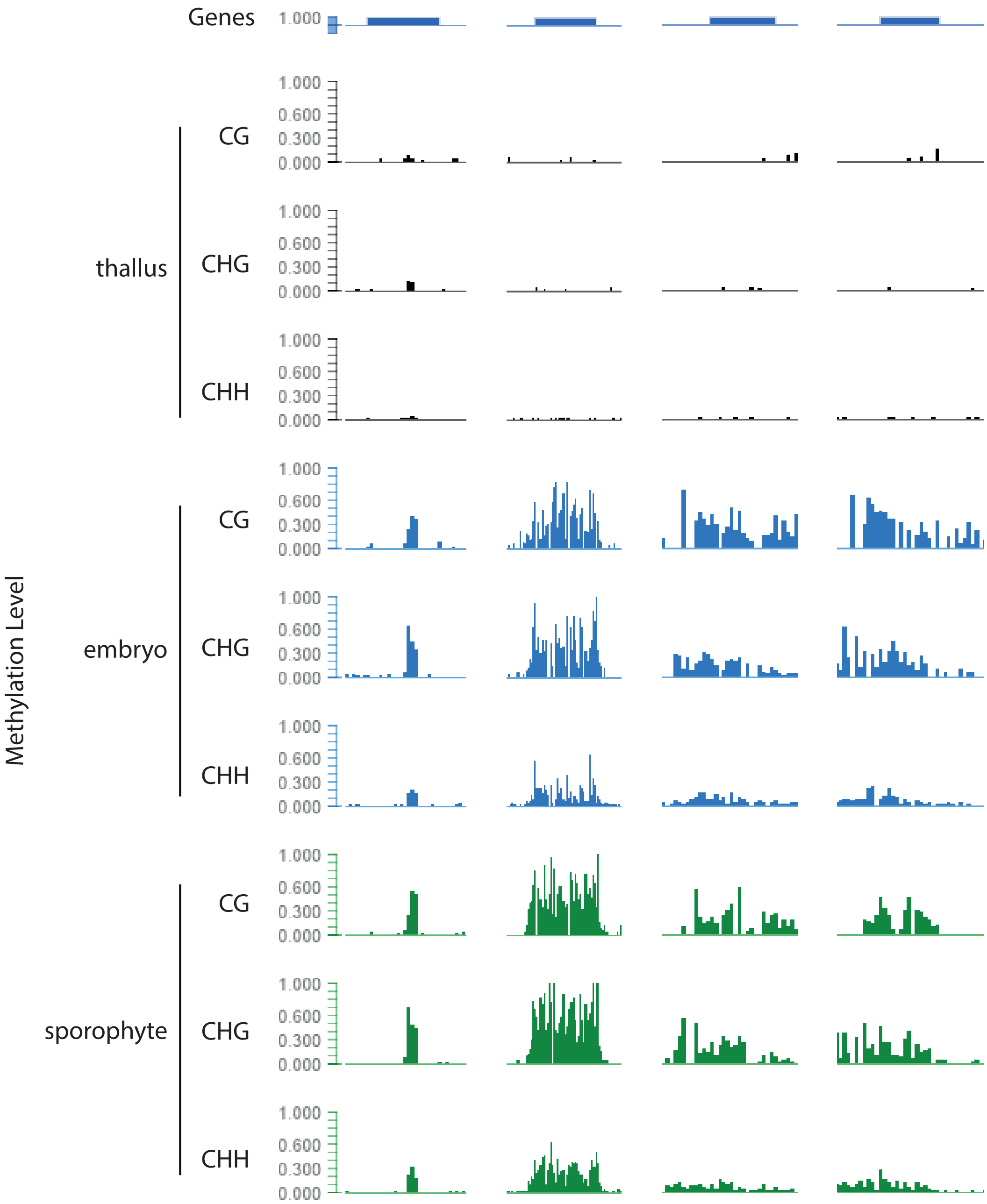


Fig. 3.8 Sporophyte specific methylation exists in the embryos of *Marchantia*

CH, CHG and CHH methylation levels of genes which gain *de novo* methylation specifically in the embryo and sporophyte of *Marchantia*. Thallus (black), embryo (blue) and sporophyte (green) mismatch targeting. To investigate this, the sRNA libraries were mapped to the reference genome, allowing for 0 and 3 mismatches. The sRNAs from the 3 mismatch dataset were then remapped onto the 0 mismatch dataset to identify identical sRNA sequences capable of targeting both TEs and genes with up to 3 mismatches. Reads overlapping MetGenes with 3 mismatches and TEs with 0 mismatches were then extracted. The resulting TE-MetGene pairs are visualised in Figure 3.9A. It must be noted that one limitation of this method is the challenge of obtaining a comprehensive TE annotation for the gene model used in the study, as not all MetGenes could be paired with annotated TEs.

Indeed, when aligning the sequence of a source TE to its corresponding MetGene, an almost perfectly matching $\sim$160bp sequence was found (Figure 3.9B). Notably, this alignment corresponds to the exact location where sporophyte-specific methylation is deposited within the MetGene (Figure 3.10 red bar).

3.5 Investigating the parental bias of sRNA production in *Marchantia* embryos, sporophyte and thallus

In the developing embryo, the paternal genome is deposited with H3K27me3, resulting in tight heterochromatic foci and transcriptional silencing [128]. This raises the question whether sRNA production in the embryo and sporophyte is biased toward the maternal genome. To explore this, single nucleotide polymorphisms (SNPs) between Tak-1 (male) and Tak-2 (female) were analysed to calculate a SNP ratio. The sRNA sequencing libraries utilised were from the thallus, archegoniophore, and antheridiophore [132] in addition to the libraries previously mentioned (Table B2).

The sRNA sequencing libraries were mapped to the reference genome with both no mismatches, and allowing one mismatch. Reads with mismatches at SNP positions were retained, as well as perfect matching sRNAs at SNP positions. A SNP ratio was calculated between the alternative and reference alleles for further analysis. Although several thousand SNP regions were initially covered, only those with at least 5 reads were retained to ensure robust results, which substantially reduced the number of SNP locations.

Reads from archegoniophores displayed a SNP ratio near 1, as expected (Figure 3.11A,B archegoniophore). Interestingly, in the early embryo, the data suggest a paternal bias in sRNA production, which shifts to a maternal bias in the sporophyte (Figure 2.8A,B embryo and sporophyte). As expected, the SNP ratio distribution in the embryo and sporophyte was less bimodal than in other tissues, indicating concurrent sRNA production from both genomes at multiple loci in the diploid phase. The thallus data

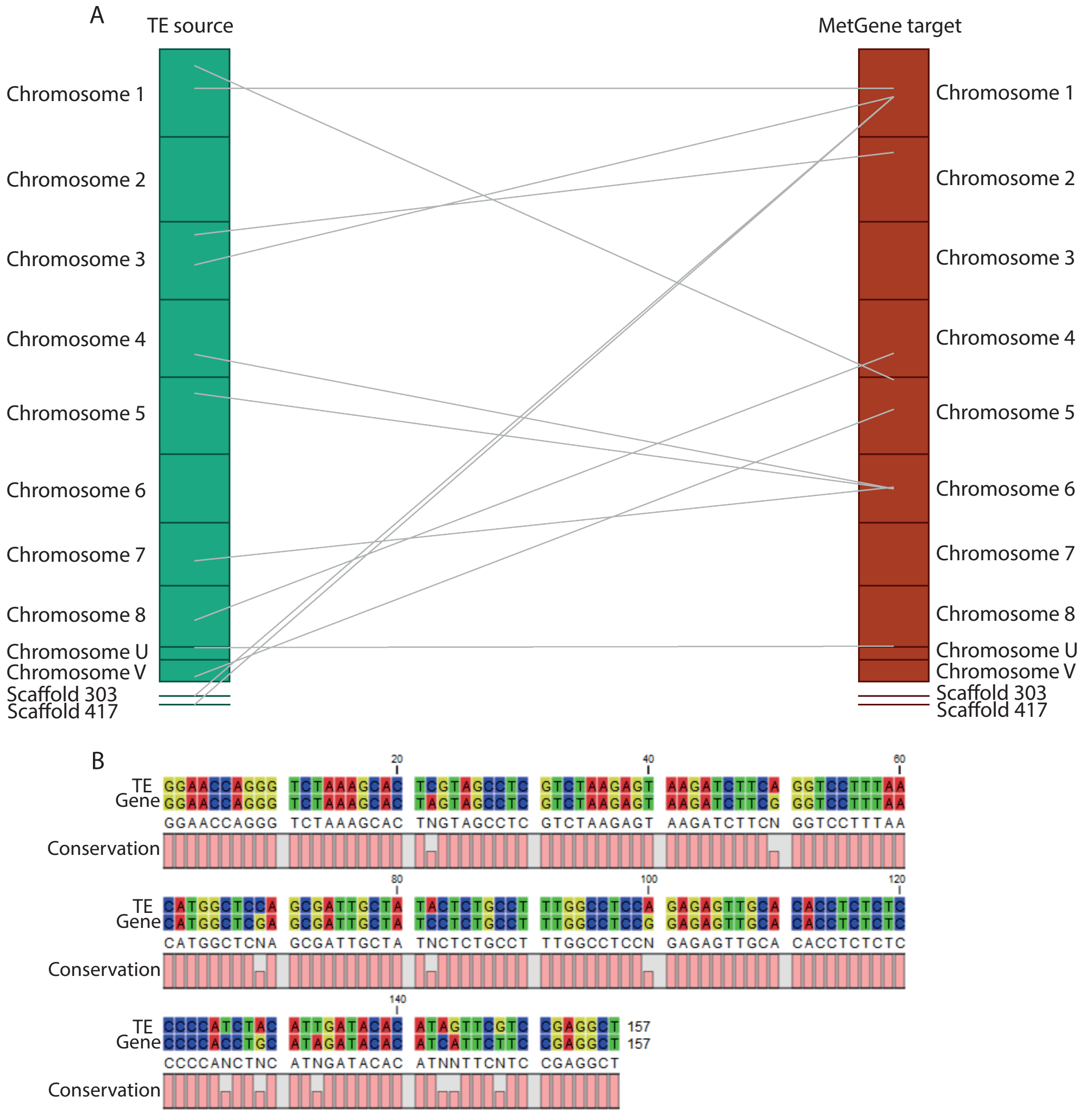


Fig. 3.9 TE loci produce 24nt sRNA that target genic loci for methylation with mismatch targeting

A)Locations of TE source loci (left, green) connected to genic target loci (right, brown) B) Alignment showing the sequence similarity between a TE-MetGene pair.

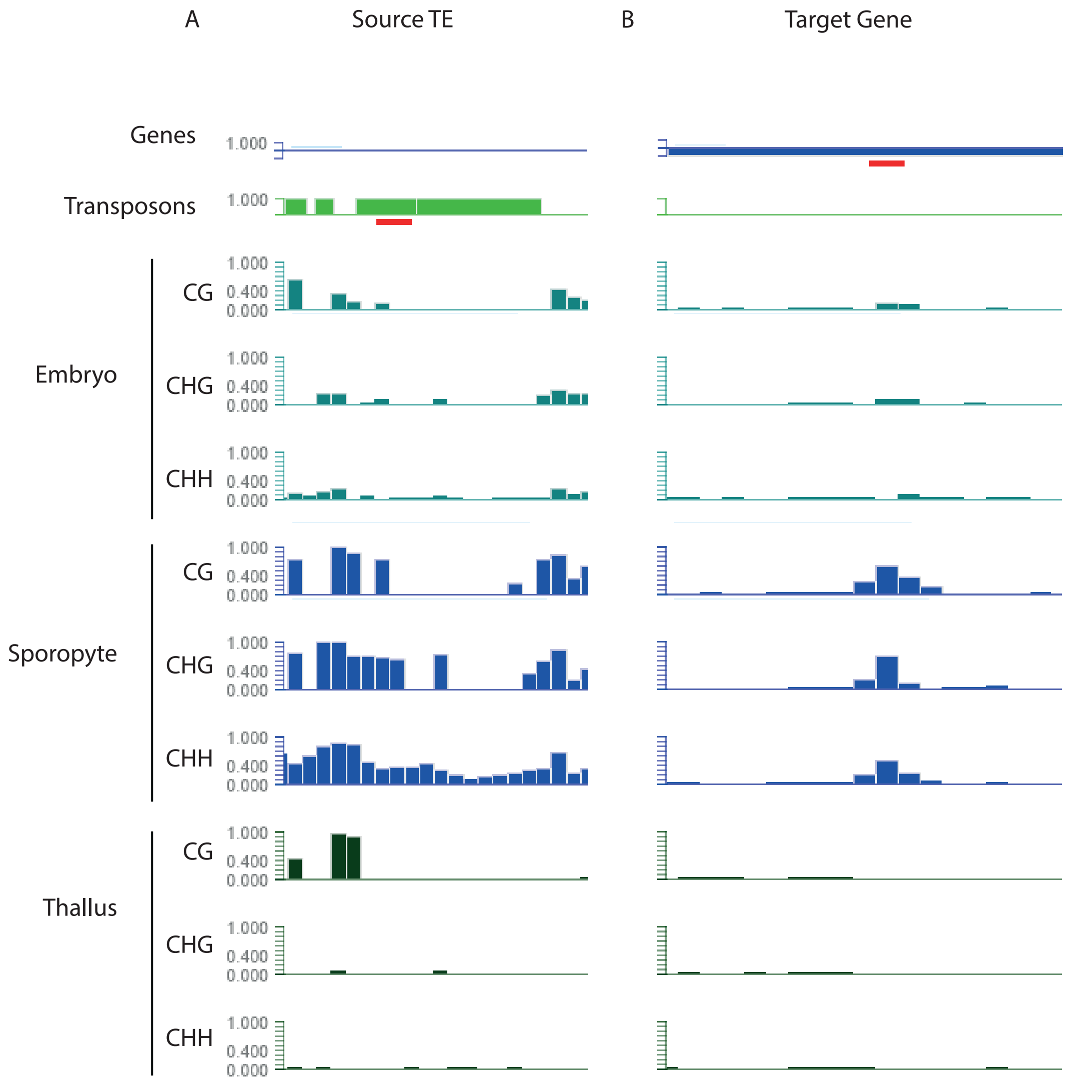


Fig. 3.10 Example of a TE source and MetGene target that gain methylation in the sporophyte

Methylation levels in (A) TE source and (B) MetGene target pair in CG, CHG and CHH sequence contexts in embryo, (teal) sporophyte (blue) and thallus (green). Overlapping sequence highlighted with a red bar.

seemed unreliable, with a surprising maternal bias despite originating from Tak-1 males [132].

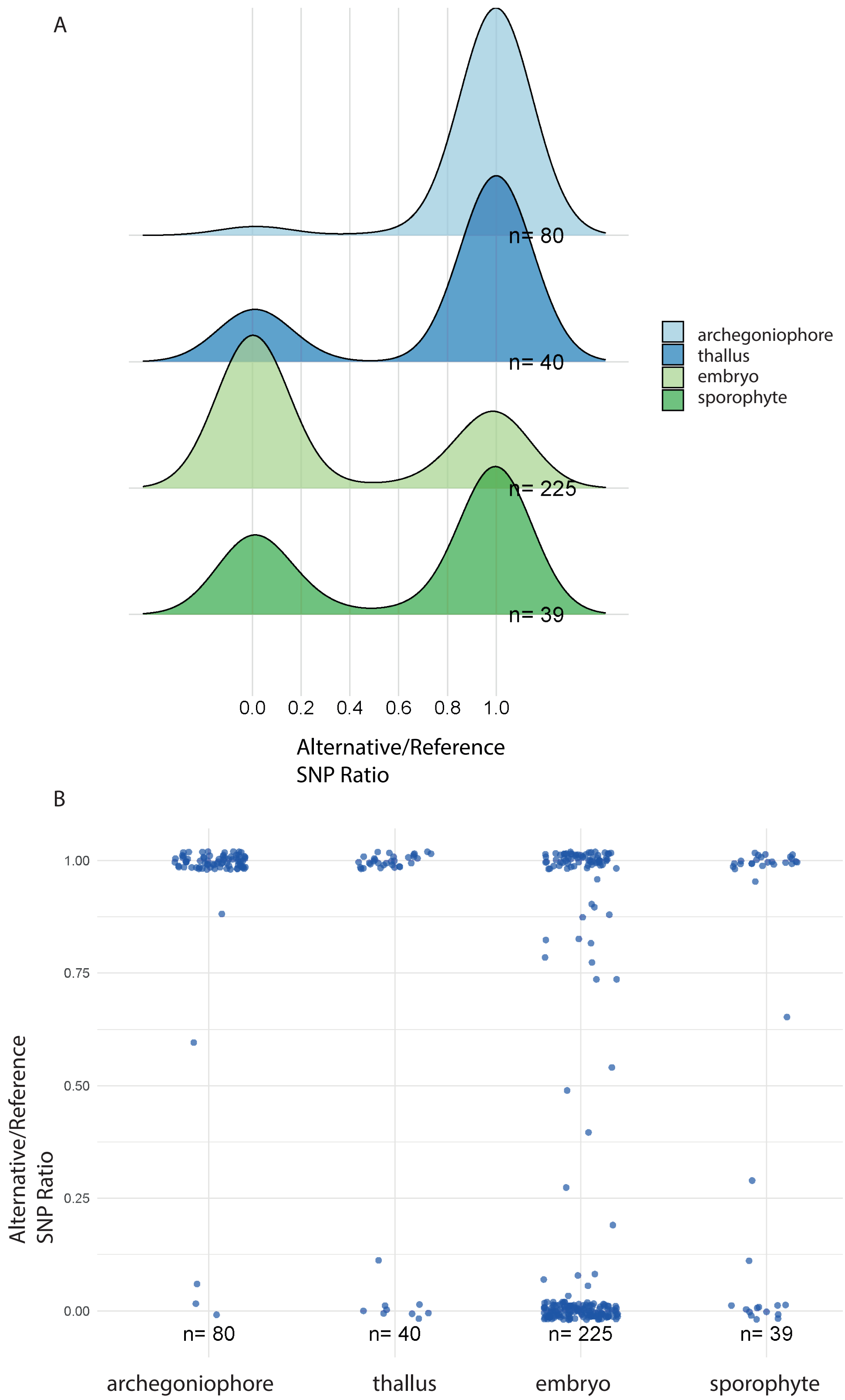


Fig. 3.11 The association of sRNA sequences with parental genomes in different tissues

A) Ridge plot to show the SNP ratio of sRNAs in archegoniophores, thallus, embryo and sporophyte. Since the reference genome is Tak-1 a ratio close to 0 signifies a paternal bias, whilst a ratio close to 1 signifies a maternal bias. B) A density plot showing the SNP ratios at remaining SNP positions after stringent filtering conditions.

3.6 Developing an effective method for visualising early embryonic development in *Marchantia*

The next objective was to determine whether the observed phenotype - where embryos fertilised with 4mC methylation mutants reach maturity earlier than wild type (Figure 3.4) - was due to accelerated pronuclear fusion in the zygote, potentially caused by the absence of paternal 4mC. This led to the hypothesis that paternal 4mC may need to be actively removed before pronuclear fusion can occur.

As described earlier, staining live *Marchantia* embryos proved challenging. A variety of fixing and staining methods were also trialled, including DAPI, Hoechst 33342 and ethidium bromide as staining dyes (the latter was used due to its small molecule size, which was expected to penetrate the fixed embryo), as well as different triton and formaldehyde concentrations and fixation times. Additionally, an egg cell reporter line was also tested (ECpro:MpSUN-GFP [122]) showing strong expression; this faded after the first zygotic division, making it insufficient for visualizing early embryonic development (Figure B2).

Next, cell wall digestion enzymes such as cellulase and pectinase were trialled to improve dye penetration. While these enzymes increased dye uptake, they also compromised cell ultrastructure, hindering proper imaging of early embryonic development and cellular ultrastructure (Figure B3). As a result, four constructs were designed to create constitutively expressing nuclear reporter lines. These constructs were driven by either the constitutive *35S* promoter or the native *elongation factor 1*α (*EF1*α) promoter, fused with citrine or tdTomato, and tagged with a nuclear localisation signal (NLS) (Figures B10, B11, B12, B13).

The constructs were initially transformed into Tak-1 thalli and selected in gemmae to avoid chimeric expression, resulting in the recovery of several independent lines from three of the four constructs (Figure 3.12). The constructs were then also transformed into sporelings from WTxWT, WTx*dn4mt1* #6 knockout and WTx*dn4mt1* #21 knockout crosses. Transformants were successfully recovered and screened for fluorescence expression in gemmae and antheridia (Figure 3.13). Interestingly, while citrine reporter lines showed strong expression in gemmae, the strongest and most consistent expression was observed in transformants with the *35S*::tdTomato-NLS and *EF1*α::tdTomato-NLS constructs. These lines were selected for crosses with wild-type female archegonia.

Unfortunately, despite strong expression in the antheridia, no nuclear reporter expression was detected in the developing zygote or embryo. This may be attributed to the previously mentioned shutdown of the paternal genome (Figure 3.14, first and second rows). In response to this, wild-type female plants were recovered from the WTxWT *35S*::tdTomato-NLS and *EF1*α::tdTomato-NLS sporeling crosses (Figure 3.14, rows

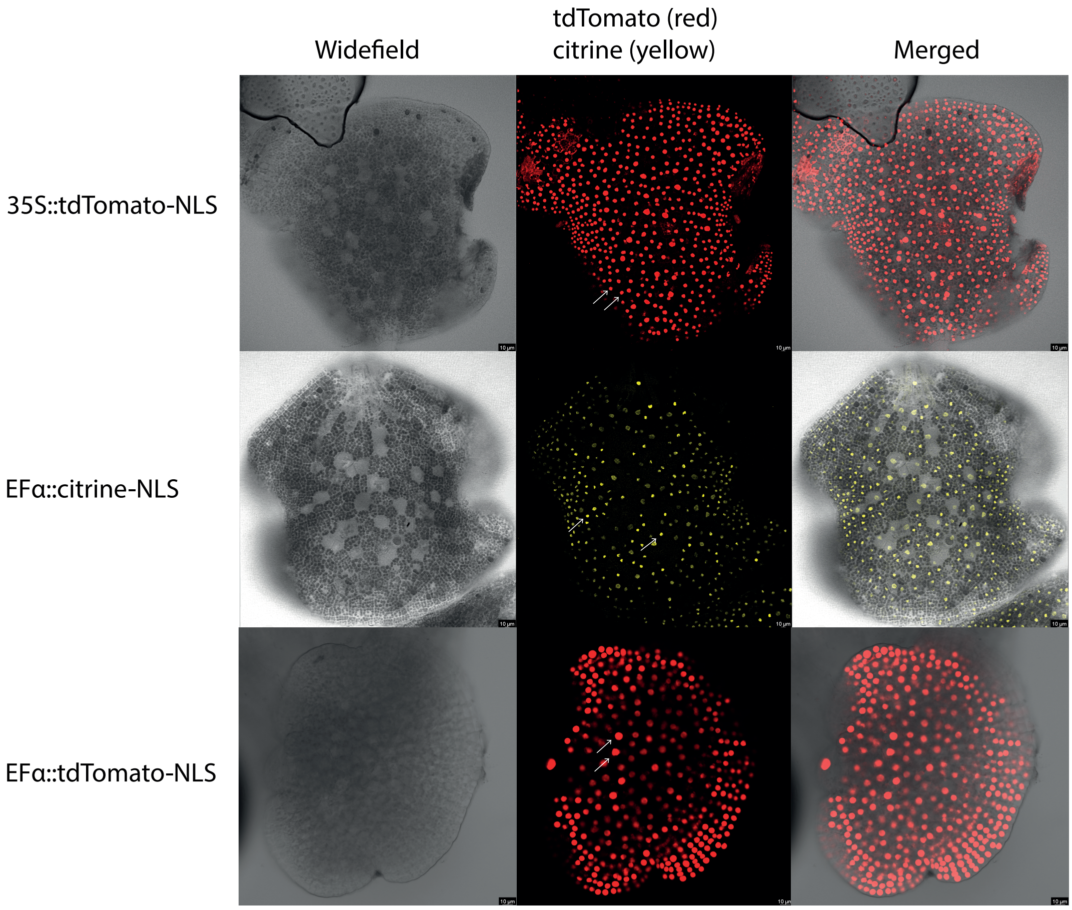


Fig. 3.12 Gemmae screening of constitutively expressed nuclear reporter lines

Nuclear signal of tdTomato or citrine (second column, red or yellow respectively, white arrows) in gemmae, driven by either *35S* constitutive promoter (first row) native constitutive promoter *EF1*α (second and third rows). Scale bar 10μm.

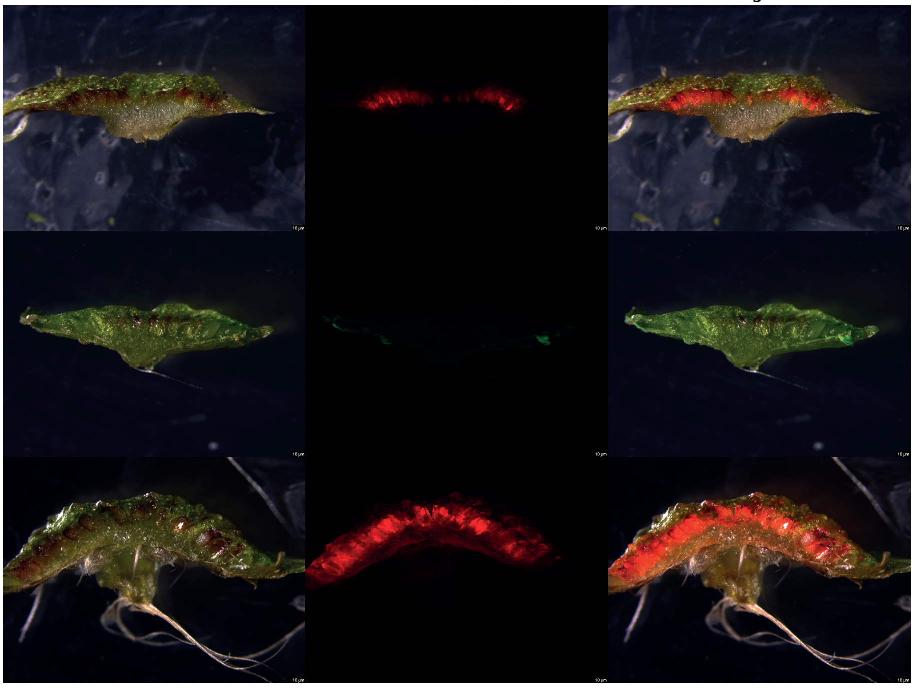


Fig. 3.13 The tdTomato based nuclear reporter lines are expressed in the antheridia

Nuclear signal of tdTomato or citrine (second column, red or green respectively) in the antheridiophore, driven by either *35S* constitutive promoter (first row) native constitutive promoter *EF1*α (second and third rows). Scale bar 10μm.

3 and 4) with strong expression of their respective nuclear reporter lines. To improve clarity and fluorescence signal, a different clearing and staining method was employed, namely iTOMEI [133]. In this protocol, the fixing component caprylyl sulfobetaine is able to clear chlorophyll without significantly quenching GFP (or other) fluorophores. After testing several environmental conditions, the best images were collected from 1% FA fixation in PBS (pH 7.4) buffer for an hour, followed by clearing with caprylyl sulfobetaine for 24 hours (please see Materials and Methods for further details).

3.7 Embryos crossed with 4mC methylation mutant sperm develop more rapidly than wild type during the first 7 days post-fertilisation

Once mature archegoniophore-producing female plants were available from the selected *35S*::tdTomato-NLS and *EF1*α::tdTomato-NLS lines, they were crossed with WT, *dn4mt1* #6 and *dn4mt1* #21 sperm. As before, the sperm was co-cultured with the archegoniophores for an hour, after which the developing embryos were dissected carefully, fixed and cleared daily, 1 to 7 days post-fertilisation. Due to the use of immature archegoniophores, embryos at 3 DAF were unfortunately unsuitable for imaging. In total, 338 embryos were imaged during the 7 days post-fertilisation and the developmental stages manually determined.

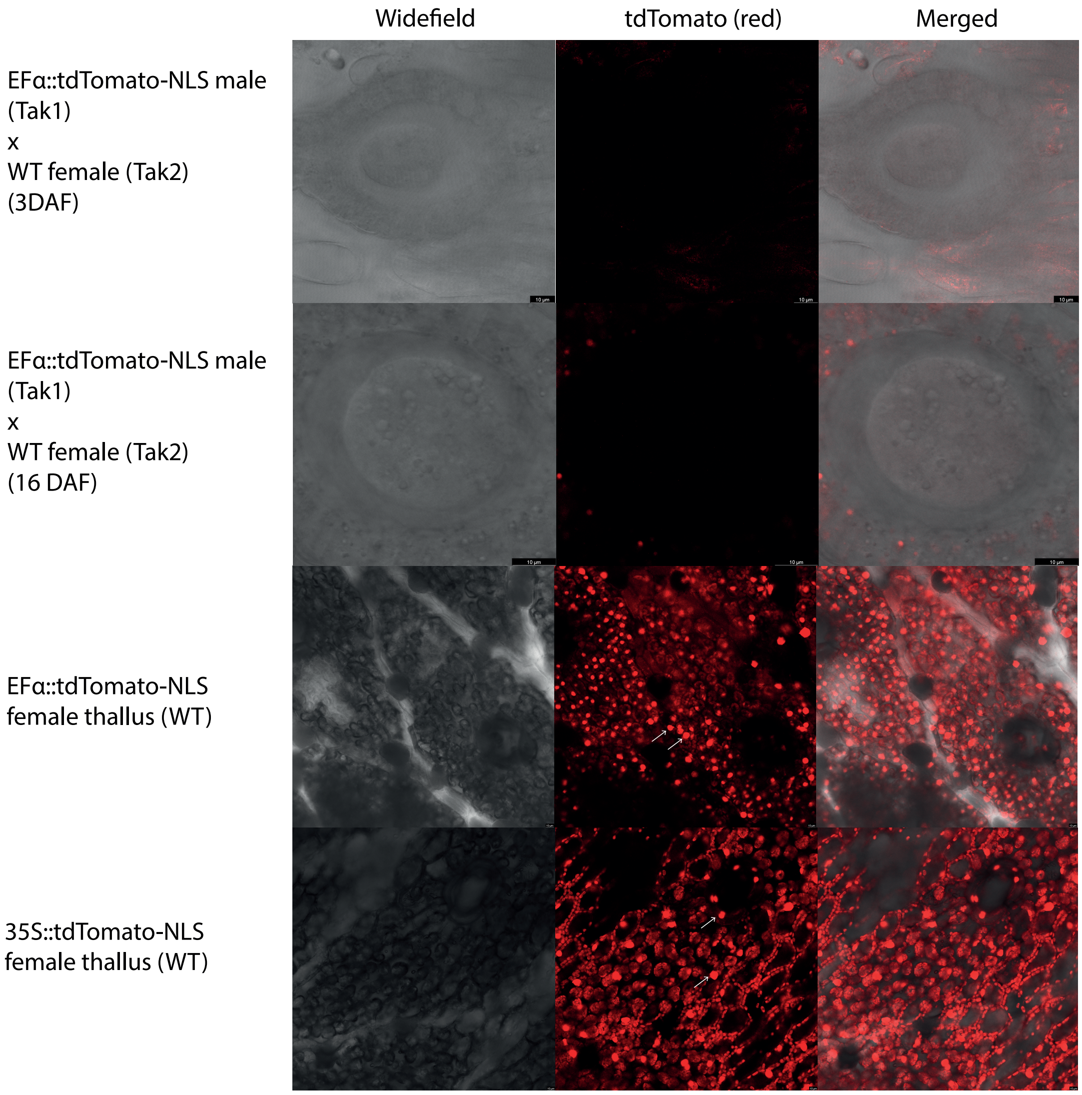


Fig. 3.14 Tak1 male nuclear reporter lines crossed to Tak2 females have no nuclear expression in the embryo

Nuclear signal of tdTomato (second column, red, white arrows) in embryos crossing *EF1α*::tdTomato-NLS Tak1 male to WT Tak2 female, 3 days (first row) and 16 days (second row) after fertilisation. Expression of *EF1α*::tdTomato-NLS (third row) and *35S*::tdTomato-NLS (fourth row) in WT female thallus. Scale bar 10μm.

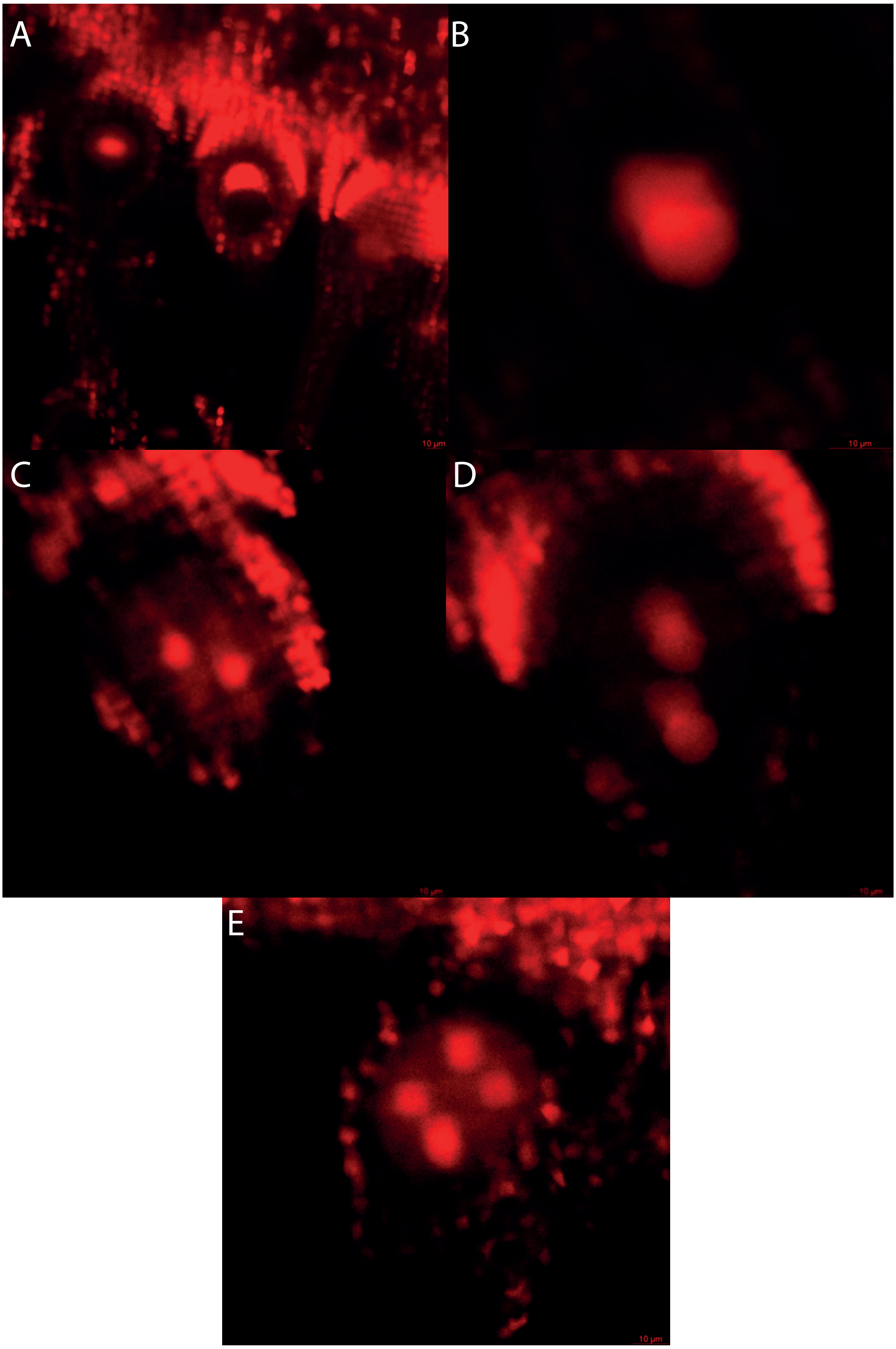


Fig. 3.15 Developmental stages of the early embryo (Tak1 male x *EF1*α::tdTomato-NLS WT female)

(A) Zygote (stage 1) (B) Zygote dividing (stage 2) (C) Two-cell stage (stage 3) (D) Two-cell embryo dividing (stage 4) (E) Four-cell stage (stage 5). Scale bar 10μm.

The developing embryo initially elongates through a transverse division, perpendicular to the archegonial axis, followed by a second vertical division to form four cells. The third division occurs at right angles to the previous ones, creating an octant of equal cells [120, 134]. Subsequent cell divisions progressively alter the embryo's initially circular shape. To further assess early cellular development, in addition to counting the nuclei, the transverse diameter of the embryonic venter cavity was measured. Interestingly, the mean transverse diameter of the embryos remained consistent and increased at a similar rate across all lines all until day 6, and between WT and *dn4mt1* k.o. #6 between days 6 and 7 (Figure B7).

During the first two zygotic divisions, cells were frequently observed with multiple nuclei that had not yet undergone cytokinesis (Figure 3.15 B and D). To facilitate tracking of cell development, these different stages were classified into distinct developmental phases:1. 1 nucleus 2. 2 nuclei, 1 cell, 3. 2 nuclei, 2 cells 4. 4 nuclei, 2 cells, 5. 4 nuclei, 4 cells, 6. 8 nuclei, 4 cells, 7. 8 nuclei, 8 cells, 8. 9–16 nuclei, 9. 16–25 nuclei, 10. 26–35 nuclei, 11. 36–45 nuclei, 12. 46–55 nuclei, 13. more than 55 nuclei.

Between days 1 and 5 post-fertilisation, embryos fertilised by either of the independent *dn4mt1* knockout lines appeared to initiate division earlier than those fertilised by wild-type sperm. For example, by day 2, nearly all imaged cells in the *dn4mt1* knockout lines had 2 nuclei, compared to fewer than a quarter in the wild type (Figure 3.16 B,C).

This trend persisted through day 4, where approximately a quarter of cells in the *dn4mt1* knockout line #6 had reached stage 4, while none of the wild-type cells had progressed that far. By day 5, the proportion of cells in various developmental stages became more synchronized across the genotypes.

After day 5, the developmental differences between the wild type and the *dn4mt1* knockout line #6 were slight, but present: Some embryos from a *dn4mt1* #6 background reached stage 12 by day 7, whereas the most developed wild type embryos were slightly behind, at stage 11. It is also noteworthy that embryos from a *dn4mt1* #6 knockout background displayed a proportion of cells arrested in stage 1 or 2, even at day 6 and day 7 (3.16 B,C). In contrast, embryos from the *dn4mt1* knockout line #21 continued to develop more rapidly than the wild type, the most advanced embryos reaching stage 13 by day 6 compared to stage 10 in wild-type (Figure 3.16 B,C).

Together, this data provides strong support that 4mC delays the onset of the first nuclear division, impacting long term developmental duration.

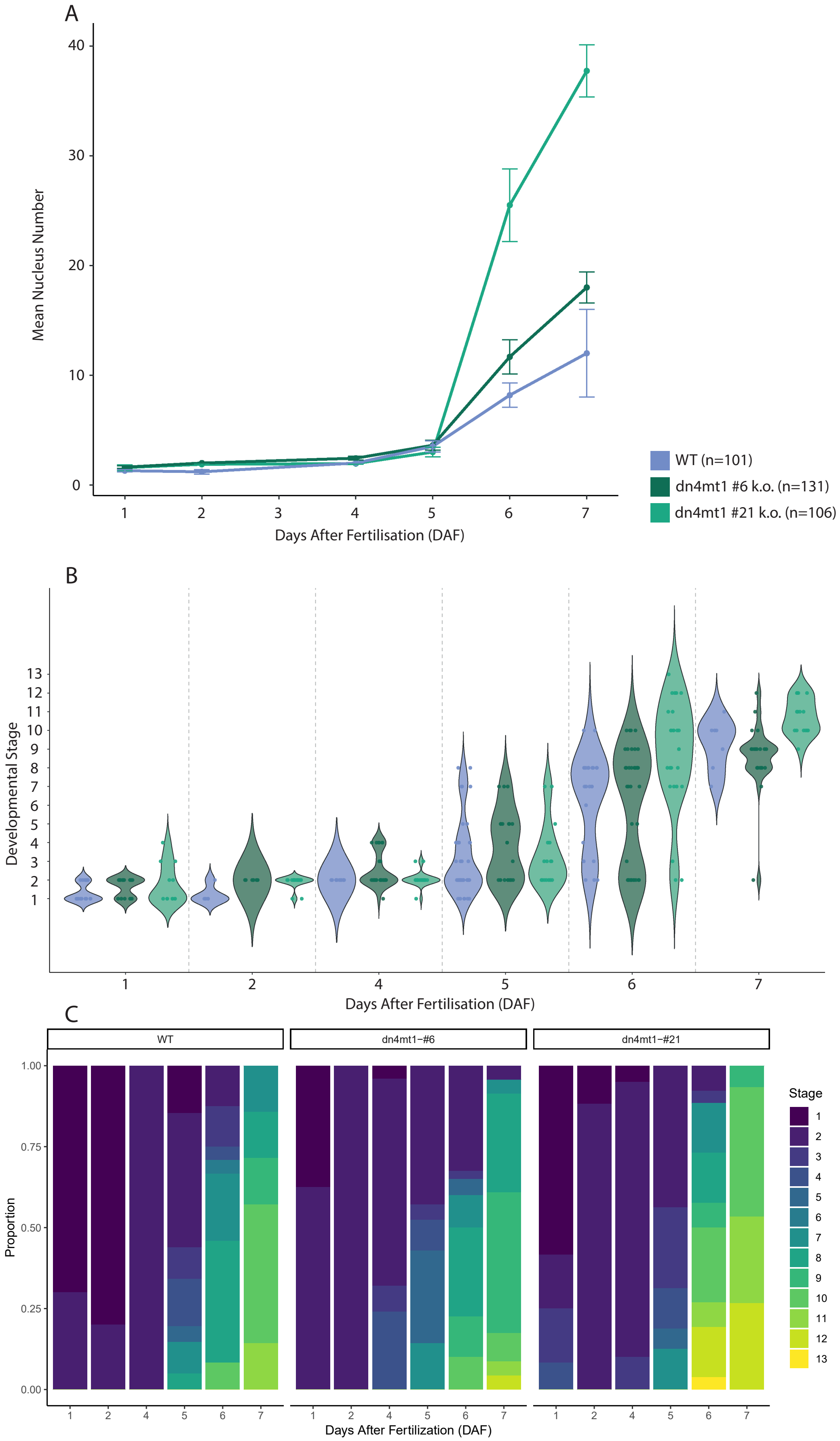


Fig. 3.16 Embryos fertilised by *dn4mt1* knockout sperm develop more rapidly than WT
See next page for full caption.

Full caption for Figure 3.16:

A) Mean nucleus number of embryos from 1 to 7 days after fertilisation, fertilised by wild type (blue), *dn4mt1* knockout #6 (dark green), or *dn4mt1* knockout #21 (light green) sperm. B) Distribution of imaged embryos in different developmental stages from 1 to 7 days after fertilisation, categorised into distinct stages based on nuclei and cell numbers. Developmental stages: 1. 1 nucleus 2. 2 nuclei, 1 cell, 3. 2 nuclei, 2 cells 4. 4 nuclei, 2 cells, 5. 4 nuclei, 4 cells, 6. 8 nuclei, 4 cells, 7. 8 nuclei, 8 cells, 8. 9–16 nuclei, 9. 16–25 nuclei, 10. 26–35 nuclei, 11. 36–45 nuclei, 12. 46–55 nuclei, 13. 55< nuclei. C) Proportion of embryos in developmental stages 1–7 days after fertilisation fertilised by wild type, *dn4mt1* knockout #6, or *dn4mt1* knockout #21 sperm.

3.8 Discussion

In angiosperms and other land plants, spermiogenesis involves chromatin compaction through the replacement of histones with sperm-specific histone variants [135, 3, 33]. However, in *Marchantia*, this process differs, as nucleosomes are replaced by protamines rather than histone variants. Consequently, DNA methylation becomes the primary heritable epigenetic mark in the paternal genome. This resembles the process in mammalian sperm, where the majority of histones are replaced by protamines, although some nucleosomes are retained [136]. Following fertilisation, mammalian zygotes undergo global DNA demethylation, driven by Ten-eleven translocation (TET) enzymes. This demethylation is crucial for transcriptional regulation, imprinting, and maintaining pluripotency in the early stages of development [137, 138].

In *Marchantia*, following fertilisation, protamines in the paternal pronucleus are replaced by histones, and H3K27me3 is deposited by a maternally expressed PRC2 complex [128] specifically on paternal chromatin. This suggests a potential mechanism for paternal genome recognition, with paternal N4-methylcytosine blanket methylation emerging as a potential candidate.

However, when examining the distribution of H3K27me3 marks over genes and transposons, there is no clear alignment with 4mC occupancy (Figure 3.7). Furthermore, results from single cell bisulfite sequencing of early *Marchantia* embryos revealed no evidence of paternal 4mC (Figure 3.6), Table 3.1). Additionally, Dr. Shujuan Xu confirmed that paternal H3K27me3 foci still form in embryos fertilised by *dn4mt1* knockout sperm, 13 days after fertilisation (Figure B9). These findings suggest that paternal genome recognition is not directly linked to blanket genome-wide 4mC methylation. However, it is important to consider that theoretical paternal 4mC trace levels in the embryo would likely be very low by 7-8 days post-fertilisation, assuming passive loss through cell division (Table 3.1). Further work is needed to optimise the protocol for low-input AMD-sequencing of *Marchantia* embryos, as previous attempts were unsuccessful despite prior high-quality control datasets. To obtain a definitive answer, developing reliable 4mC immunostaining and/or single-cell AMD-seq methods for zygotes or embryos at the 2-4 cell stage is essential, provided successful dissection at this early developmental stage can be achieved.

Considering the shut-down of the paternal genome in the sporophyte, the possibility of sRNA production from the paternal genome became an intriguing question. Investigating whether there is a difference in sRNA production between the two parental genomes in the sporophyte, particularly in light of this potential suppression, revealed intriguing results. Initially, the sRNA data suggested a paternal bias in the embryo, which appeared to shift toward a maternal bias in the sporophyte, though substantial sRNA production was observed from both parental genomes (Figure 3.11A,B).

Despite these findings, it is essential to acknowledge the limitations of this analysis. The male and female accessions used as the parental lines (Tak-1 and Tak-2) are asexually maintained in lab conditions, which means accumulated point mutations over time may significantly alter their genomic sequences. Indeed, at several SNP positions, bases other than the reference or alternative were observed across all datasets. Additionally, if Tak-1 had been crossed with other lines at any point, the resulting hybrid would exhibit a different SNP distribution from either parent.The unequal SNP distribution along different chromosomes further complicates the analysis, making these cultivars less suitable for complete genome coverage (Figure B4). Moreover, the filtering steps applied to the sRNA data resulted in a limited number of data points, which restricts the ability to draw definitive conclusions. Ensuring the precise genetic background of the parental lines is therefore essential for accurate results, as is obtaining sufficient sequencing depths to definitively answer whether 24nt sRNAs are produced from both parental genomes in *Marchantia* embryos.

Furthermore, exploring the distribution of H3K9me1 histone modifications on TEs in the mature embryo (as a proxy for locations that may be active in RdDM, as H3K9me2 has a direct interaction with SHH1, which recruits the Pol IV complex in *Arabidopsis* [40]) does not reveal a strong parental bias (Figure B8) [128]. Nevertheless, addressing the limitations of the current dataset and producing a high-confidence, genetically distinct parental cross would provide valuable insights into the sRNA landscape of *Marchantia* embryos. These results could then be cross-referenced with bisulfite sequencing data to draw more robust conclusions.

It has been previously noted that similar to genes hypermethylated in the male germline of *Arabidopsis* (MetGenes), *de novo* methylated genes are also present in the sporophyte of *Marchantia* (Figure 3.8 [131]). In *Arabidopsis*, few perfectly matching sRNAs map directly to methylated MetGene regions. This observation led to the discovery that HyperTE-derived sRNAs target MetGene locations in meiocytes for methylation with up to three mismatches [60] (Figure B5).

Building on this, once MetGene loci were identified in *Marchantia* sporophytes, it was explored whether CHH methylation at these genes could be targeted by sRNAs produced by TEs with high sequence homology. Although sequence homology between TE sources and target MetGenes was identified (Figure 3.10), a clear relationship between sRNA mismatch targeting and CHH methylation was not established. Mapping sRNAs to MetGene locations with up to 3 mismatches did not significantly increase the abundance of sRNAs at these locations (Figure B6).

Nevertheless, a mechanism must allow either the production of 24nt sRNAs and/or the targeting of TE-derived sRNAs for methylation at MetGenes specifically during the sporophyte stage. Further investigation is needed to verify this hypothetical causal relationship, including examining RdDM-associated protein variants uniquely expressed in the sporophyte and/or creating CRISPR/Cas9 knockout lines to mutate homologous sequences from the putative TE source to determine whether CHH methylation is abolished at the corresponding MetGene.

Previous studies have shown that the mature pre-meiotic embryo loses all traces of paternal 4mC methylation [121]. The removal of 4mC may be necessary for embryo development, as there is evidence that 4mC suppresses chromatin accessibility and transcription [121]. There are two possible pathways through which paternal blanket 4mC methylation could be absent in the mature embryo or sporophyte. One possibility is that after fertilisation, MpDN4MT1 is not expressed, or 4mC methylation is not actively maintained, leading to its gradual dilution through cell division. Alternatively, we can hypothesise that paternal 4mC methylation needs to be removed to allow pronuclear fusion. This hypothesis could explain the unusually long period required for *Marchantia* embryos to undergo pronuclear fusion and complete the first zygotic division (up to 3–4 days after fertilisation) [122]. Additionally, the accelerated sporophyte maturation observed in embryos fertilised by *dn4mt1* mutant sperm, maturing up to 4 days earlier than wild-type embryos (Figure 3.4), could support this explanation. In this scenario, MpROS1X (a DNA glycosylase encoded on the female chromosome) emerges as a potential candidate that could facilitate the removal of paternal blanket 4mC prior to pronuclear fusion.

When comparing early embryonic development between the two 4mC mutants and wild type, it is evident that zygotic divisions began earlier in both 4mC mutants than in the wild type (Figure 3.16B,C). Following this initial difference, all genotypes developed at a similar rate until day 5, with comparable nucleus numbers. However, after day 5, the 4mC mutants exhibited a sharp exponential increase in cell numbers over the next two days, while wild type embryos also underwent rapid cell division, but with a more gradual increase (Figure 3.16). Interestingly, despite the expectation that the wild type embryo's delayed early embryonic development would lead to a delayed entry into exponential growth compared to the 4mC mutants, both genotypes entered this phase simultaneously. The primary difference lay in the rate of cell proliferation, rather than the timing of its onset.

These findings suggest that 4mC removal may be required prior to pronuclear fusion. However, more detailed immunostaining or single-cell AMD-seq in the zygote is necessary, along with an investigation into the phenotypic effects of knocking out DNA demethylases, particularly MpROS1x. Additionally, it would be important to explore whether ROS1x possesses N4-methylcytosine glycosylase activity. Furthermore, despite accelerated embryonic development, the knockout of paternal 4mC appears to have a partial fertility cost, as some embryos in these knockouts exhibit delayed development or arrest (Figure 3.16B,C), a phenomenon observed in previous studies [121].

3.9 Materials and Methods

3.9.1 Plant materials and growth conditions

Male and female *Marchantia polymorpha, L.* subsp. *rudealis* accessions Takaragaike-1 (Tak-1, male) Takaragaike-2 (Tak-2, female) and Cam-2 (female) was used. The plants were grown on 1% agar (Sigma-Aldrich), supplemented with ½ strength Gamborg's B5 medium. They were grown in a controlled environment chamber (Conviron) at 21°C with 70% humidity under constant light with far-red light for induction of sexual reproduction, as described previously [125, 139].

3.9.2 Plasmid construction

Four constructs were generated in total using a Gateway®cloning system with either constitutive promoter *35S* or the endogenous promoter *EF1*α, driving either citrine or tdTomato fluorophores with a nuclear localisation signal (NLS).

The insert for the BP reaction was amplified from plasmids pMpGWB115 (plasmid #68569, Addgene) and pMpGWB116 (plasmid #68570, Addgene) [140] to get the citrine-NLS and tdTomato-NLS fragments respectively. The BP reactions were performed using the donor plasmid pDONR™221 (Thermo Fisher (Invitrogen)). The LR reactions were performed using plasmids pMpGWB102 (plasmid #68556, Addgene) and pMpGWB103 (plasmid #68557, Addgene)[140] as the backbones for the *35S* and *EF1*α promoter driven expression clones respectively. These constructs yielded the final 4 expression clones: MpGWB102(p*35S*)-citrine-NLS, MpGWB102(p*35S*)-tdTomato-NLS, MpGWB103(p*EF1*α)-citrine-NLS, MpGWB103(p*EF1*α)-tdTomato-NLS. The constructs were transformed into *Escherichia coli*, purified and the sequence confirmed, followed by transformation into *Agrobacterium tumefaciens* strain MP90.

The primers used and plasmid maps are available in Appendix B Table B1 and Figures B10, B11, B12 and B13.

3.9.3 Thallus and sporeling transformation

For thallus transformation, Tak-1 gemmae were grown for 18-20 days in a controlled environment chamber (Conviron) at 21°C with 70% humidity under constant light on 1.2% agar (Sigma-Aldrich), supplemented with ½ strength Gamborg's B5 medium. Following 18-20 days, thalli were cut and the basal fragments regenerated on plates supplemented with 1% sucrose for 3 days. The thalli were then co-cultured with *Agrobacteria* (strain MP90) in 0M51C liquid media for 3 days and transformants selected on 1% agar (Sigma-Aldrich) supplemented with 10 μg/mL hygromycin and 120 μg/mL cefotaxime for several weeks. Fluorescent gemmae were selected from transformant thalli and regenerated again to avoid chimeric plants[141].

For sporeling transformation, sporangia of backgrounds WT x WT, WT x Mp*dn4mt1* #6 and T x Mp*dn4mt1* #21 were sterilised in 1mL of Milton's sterilising solution on a rotating shaker for 20 minutes. The spores were pelleted and washed with sterile water twice. The sterilised spore suspension was added into 25mL 0M51C liquid medium and the flasks were incubated in a growth chamber with constant light at 1500-2000 lux, 23 °C for 5-7 days. Induced *Agrobacterial* cultures (strain MP90) containing each vector was co-cultivated with the sporelings for 2 days on a rotating shaker. the spore culture was strained and washed with sterile water and transformants were selected on 1% agar (Sigma-Aldrich) supplemented with 10 μg/mL hygromycin and 120 μg/mL cefotaxime for several weeks[142].

3.9.4 Semi in vitro culture and crossing

Semi in vitro culture and crossing was performed as described previously[122]. Briefly, mature female archegonia were collected into 5mL tubes containing 3mL water and co-cultured with male antheridia for 1 hour under white light at 22°C. The fertilised archegonia were then washed and incubated in 5mL tubes containing 3mL of fresh water under white light at 22°C until observation.

3.9.5 Microscopy

The archegoniophores were collected for dissection and imaging and were either observed live or were fixed for imaging. For live cell imaging, rows of archegonia were manually dissected from the base of digitate rays and the embryo dissected out using micro knives or double lancet sapphire surgical knives (Fine Science Tools, World Precision Instruments) collected on cavity slides (BRAND®) and washed twice with 1x PBS to remove any maternal tissue. They were then stained with DAPI for 5 mins and briefly vacuum infiltrated and observed using slides with imaging spacers (Thermo Scientific™ Gene Frame) or on cavity slides [122].

For observing developmental synchronicity and the Mp*dn4mt1* early developmental phenotype, rows of archegonia were manually dissected from the base of digitate rays removing involucres. The fluorescent samples were fixed and stained using the iTOMEI protocol [143] using 1% FA fixation for an hour in 1X PBS (pH 7.4) followed by clearning (20% caprylyl sulfobetaine for 24 hours) and mounted in iohexol with imaging spacers (Thermo Scientific™ Gene Frame). The embryo/egg cells were imaged using a Leica SP8X confocal microscope.

3.9.6 AMD-sequencing and bisulfite-sequencing library construction

Dissected Tak-1 x Cam-2 embryos were used to construct 4mC AMD-seq libraries using the xGen™ Methyl-Sequencing DNA Library Preparation Kit (IDT, 10009860) after treatment of sheared DNA with APOBEC3A (NEB #E7120S), following manufacturer's instructions. In parallel, a bisulfite sequencing library was also contructed after APOBEC3A treatment using the Imprint®DNA Modification Kit (Sigma-Aldrich, MOD50)

3.9.7 Imaging analysis

The collected images were analysed in Fiji (ImageJ) [144]. The macro was created to automate the image analysis steps, whereby a maximum Z projection was applied to the 3D image stack and a Gaussian blur was applied. The selected region on interest (the embryo) was then masked and a local contrast was then enhanced using the CLAHE plugin, and nuclei were counted manually in the resulting images using the ROI Manager. The resulting data was visualised in R using the ggplot2 package.

3.9.8 Determination of MetGenes

MetGene locations were defined using parameters previously described [131]. Briefly, bisulfite sequencing reads were mapped using a custom mapping script based on Bismark [73], and fractional methylation (50bp windows) was determined for each sample in the CG, CHG and CHH contexts using MethylDackel v0.4.0 and custom scripts. Windows were selected where C-methylation was greater than 0.05 and adjacent windows merged if they occured within 200bp. The methylation difference in each context was determined between the sporophyte and thallus, and retained according to these filtering criteria: Islands greater than 99bp; significantly different non-CG metylation between sporophyte and thallus (Fisher's exact test $p<0.001$); CG, CHG, CHH and non-CG difference between sporophyte and thallus bigger than 0, 0.05, 0.1 and 0.3 respectively. The final filtering step kept islands that had less than 0.1 CG methylation in the thallus and overlapped genic regions. The final list yielded 221 DMRs.

3.9.9 Histone analysis

CUT&RUN reads were were processed using SAMtools v1.9 [145], BEDtools v2.27.1 [116] and Picard v2.18.27 [146] then mapped to the Takv6.1 genome [147] using bwa v0.7.17 [148]. To distinguish between maternal and paternal sequencing reads, a SNP analysis was conducted. SNPs were called using gatk v4.0.1.2 [149] wherein the SNPs between the parental Tak-1 and Cam-2 were replaced with Ns. The parental reads were assigned using SNPSplit v0.3.4 [150] and counts for each parental genome were calculated using SAMtools v1.9 and BEDTools v2.27.1. Please see reference [128] for a detailed description of the pipeline.

3.9.10 sRNA analysis

The sRNA reads were processed using Trim Galore! v0.4.2 [151], and mapped using bowtie v1.0.1 [115] allowing for 0 or up to 3 mismatches with the -v option. For the TE-MetGene matching analysis, the resulting BAM files were converted to fasta files using custom awk scripts. The reads were once again mapped using bowtie, with no mismatches. Finally the resulting reads in the BAM files were queried against original genome locations using custom awk scripts, and the resulting file was filtered by MetGene locations with BEDtools v2.27.0.

For the sRNA SNP analysis, a SNP file was created using the method described above but this time using Tak-1 and Tak-2 as the parental genomes. 24nt reads overlapping MetGene locations with 0 or 1 mismatches were filtered using BEDtools v2.27.0 and were kept if they overlapped SNP positions. From the 1 mismatches dataset, only reads that mismatched exactly at the SNP positions were kept to prevent reads that mismatch elsewhere but also covering SNP positions from biasing the data. The perfect matching dataset was downsampled to prevent this much larger dataset from biasing the data. The reference, alternative or other bases were counted at each SNP position using SAMtools mpileup and this dataset was imported to R for visualisation in R using the ggplot, ggpubr and ggridges packages. Please find all scripts used at my personal GitHub page.

Chapter 4

Main Discussion

This thesis aimed to investigate the epigenetic mechanisms regulating gene expression and genome stability during plant reproduction, focusing on two plant systems: *Arabidopsis thaliana* and *Marchantia polymorpha*. Specifically, I examined the expression patterns of key RNA-directed DNA methylation (RdDM) components in *Arabidopsis* in the development of the male sexual lineage, described the 24nt sRNA profiles after meiosis in microspores and sperm. I also explored the dynamics of DNA methylation, particularly the inheritance of the paternal epigenetic mark N4-methylcytosine (4mC), during early embryo development in *Marchantia*. To address these aims, I employed a combination of sRNA sequencing, bisulfite and APOBEC3A-mediated methylation sequencing, fluorescent reporter line construction and imaging and developing bioinformatics pipelines and data analysis of novel and existing data. The findings reveal novel insights into the epigenetic regulation of plant reproduction and highlight the diverse mechanisms employed by different plant lineages.

4.1 Tissue-Specific Regulation of RdDM

The analysis of NRPD1 and NRPE1 expression patterns in *Arabidopsis* revealed a dynamic regulation of these key RdDM components during pollen development. NRPD1, the largest subunit of RNA Polymerase IV, showed confined expression in the tapetal layer and surrounding somatic anther tissue during the meiocyte stage, with subsequent expression in microspores but absence in mature pollen. In contrast, NRPE1, the largest subunit of RNA Polymerase V, exhibited broader expression in both somatic and germline tissues throughout anther development.

Investigation of CLSY chromatin remodeller expression in anthers provided further insights into RdDM regulation. CLSY3 expression was confirmed to be confined to the tapetum and absent in later developmental stages. Similarly, CLSY1 and CLSY2 were not detected in maturing pollen. However, some discrepancies were noted when these findings were cross-referenced with RNA-sequencing data. Despite these ambiguities, the sRNA profiles largely corroborated the observed expression patterns of CLSY proteins. This suggests a potential shift in the regulation of sRNA production after meiosis, possibly involving alternative mechanisms for sRNA biogenesis, and highlighting the importance of non-canonical RdDM pathways in the male sexual lineage.

These findings collectively indicate a complex temporal and spatial regulation of the RdDM pathway during the development of the male sexual lineage, with potential implications for the establishment and maintenance of DNA methylation patterns. The observed changes in CLSY expression patterns may reflect the dynamic nature of epigenetic regulation during pollen development and highlight the need for further investigation into the mechanisms governing sRNA production in mature pollen.

4.2 Sperm cell-specific sRNA clusters are uncoupled from CLSY-dependent loci

The analysis of sRNA profiles in *Arabidopsis* sperm cells revealed a unique landscape distinct from both earlier germline stages and somatic tissues. Firstly, a family of LTR/Gypsy TEs was identified as producing sperm-specific 24nt sRNAs which may hint at ATGP2N reactivation in the sperm cell, or the 24nt sRNAs may originate from the vegetative cell based on CHH profiles. Secondly, around 20% of MetGenes produce 24nt sRNAs specifically in the sperm cell, including genes important in mediating CG methylation. Finally, while CLSY1 and CLSY2-dependent loci produce abundant sRNAs in somatic tissues, sperm cells exhibit a different pattern. Many sperm-specific sRNA clusters do not overlap with known CLSY-dependent loci, further hinting at the involvement of non-canonical RdDM pathways in the biogenesis or sRNAs in this tissue. These findings indicate a shift in the regulation of epigenetic pathways during the final stages of the development of the male sexual lineage, which may be crucial for genome integrity.

4.3 A novel dissection and sequencing protocol showed that paternal 4mC is lost in early *Marchantia* embryos and that the knockout of paternal 4mC results in accelerated early embryonic development

A novel method of dissecting early *Marchantia* embryos was trialled and perfected, allowing the successful trial of low input AMD and bisulfite sequencing protocols. Furthermore, a fixing and clearing protocol was optimised to explore early embryonic development in *Marchantia*. Nuclear reporter lines of wild type male, female and *dn4mt1* male *Marchantia* plants were constructed screened and used to study the effect of 4mC knockouts in early embryonic development.

I presented evidence that paternal 4mC methylation, which is deposited during spermiogenesis, is lost in the early embryo. Our analysis of early embryos (7-8 days after fertilisation) showed no detectable levels of 4mC methylation. This, together with the accelerated developmental phenotype of embryos fertilised by *dn4mt1* mutant sperm, hints at the requirement of 4mC removal prior to pronuclear fusion.

Our live cell imaging experiments revealed that embryos fertilised by sperm lacking 4mC methylation (*dn4mt1* mutants) initiate cell division earlier and develop more rapidly during the first 7 days post-fertilization compared to wild-type embryos. This finding provides insight into the functional significance of 4mC in regulating the timing of early embryonic events in *Marchantia*.

4.4 Sporophyte-specific genic methylation is likely targeted by RdDM and a novel method for assigning reads from parental origins for sRNA data is developed

Regions with *de novo* genic methylation in the sporophyte of *Marchantia* were identified and their relationship with TE-derived sRNA clusters was explored. While sequence homology between TEs and MetGenes was identified, no clear relationship was established between sRNA mismatch targeting and increased CHH methylation at MetGenes. Mapping sRNAs with up to three mismatches did not result in a significant increase in sRNA abundance at these MetGene loci.

The study also explored the potential differences in sRNA production between the paternal and maternal genomes in the sporophyte, especially given the shutdown of the paternal genome. Initial data suggested a paternal bias in sRNA production during the embryo stage, which shifted towards a maternal bias in the sporophyte, although both parental genomes continued to produce substantial amounts of sRNAs.

4.5 Future Directions

4.5.1 Understanding cross-cellular transport of sRNAs

This thesis raises important questions about the mechanisms underlying cross-cellular sRNA movement in reproductive tissues. It seems that the production of sRNAs in somatic tissues adjacent to germline cells with plasmodesmatal or other cytoplasmic connections seems to be a common feature in several plant sexual developmental pathways [60, 54, 82, 81]. While it is clear that plasmodesmata facilitate the transfer of sRNAs between the tapetum and meiocytes in *Arabidopsis*, the exact molecular players involved in this process are still poorly understood. Future studies should focus on elucidating the pathways that regulate sRNA movement during germline development.

4.5.2 Investigating non-canonical RdDM pathways post-meiosis

Our discovery of unique sRNA clusters in *Arabidopsis* sperm cells that are uncoupled from known CLSY-dependent loci opens up new avenues for research. Future studies should aim to identify the mechanisms responsible for generating these sperm-specific sRNAs and their potential targets. Investigating the functional significance of these sRNAs in sperm cell development, fertilisation, and early embryogenesis would provide valuable insights into male sexual lineage-specific epigenetic regulation.

4.5.3 Investigating the mechanisms of paternal 4mC removal in *Marchantia* and the embryonic epigenetic landscape

While we have observed the rapid loss of paternal 4mC in early *Marchantia* embryos, the exact mechanism remains unclear. Future research should focus on determining whether this loss occurs through passive dilution or active removal. Developing single-cell AMD sequencing methods for zygotes or 2-4 cell stage embryos would provide valuable insights into the dynamics of 4mC loss. Additionally, investigating the potential role of DNA demethylases, particularly MpROS1x, in active 4mC removal would be crucial. Functional studies of MpROS1x, including its potential N4-methylcytosine glycosylase activity, could shed light on the mechanisms of epigenetic reprogramming in early *Marchantia* embryos.

The accelerated development observed in embryos fertilised by 4mC-deficient sperm suggests a regulatory role for this epigenetic mark in early embryogenesis. Future research should focus on elucidating the molecular mechanisms by which 4mC influences cell cycle progression and gene expression in early embryos. This could involve transcriptome analysis of wild-type and *dn4mt1* mutant embryos at various early developmental stages, as well as identifying potential 4mC readers that may influence its effects on embryonic development. Additionally, investigating the long-term consequences of accelerated early development on sporophyte growth and fitness would provide insights into the evolutionary significance of 4mC-mediated embryonic development.

Further work should be done investigating both the significance of and the mechanism by which sporophytic *de novo* methylation is established in *Marchantia*. This could include examining RdDM-associated protein variants specific to the sporophyte and employing CRISPR/Cas9 to mutate homologous sequences from potential TE sRNA sources, which would help determine whether CHH methylation at the corresponding MetGenes is abolished.

4.6 Concluding remarks

To conclude, this thesis has uncovered novel aspects of epigenetic regulation during plant reproduction in both *Arabidopsis* and *Marchantia*. It highlights the increasing complexity and dynamic nature of epigenetic reprogramming during germline and embryonic development, conserved across basal land plants and flowering plants. These findings offer new insight into our understanding of plant reproductive epigenetics and hope that it will form the foundation of further studies.

Appendix A - Chapter 2

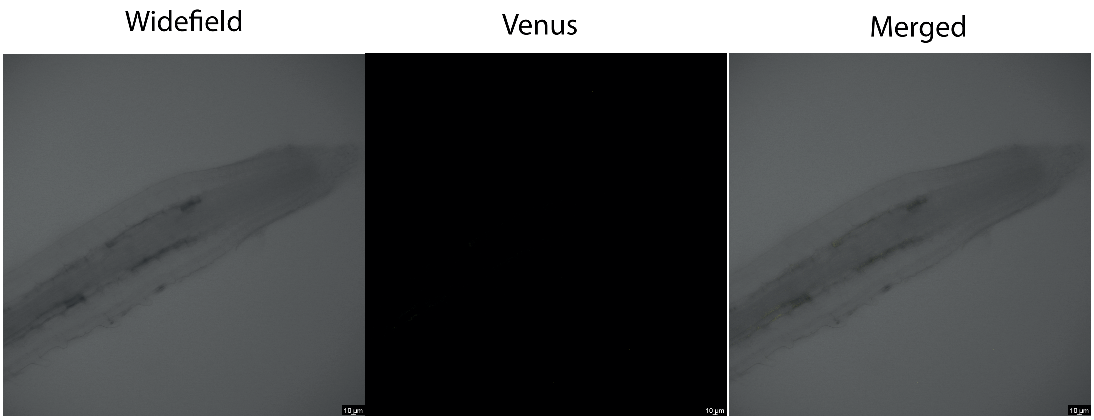


Fig. A1 pCLSY3::CLSY3-Venus is not expressed in the root tip

Scale bar 10 μm.

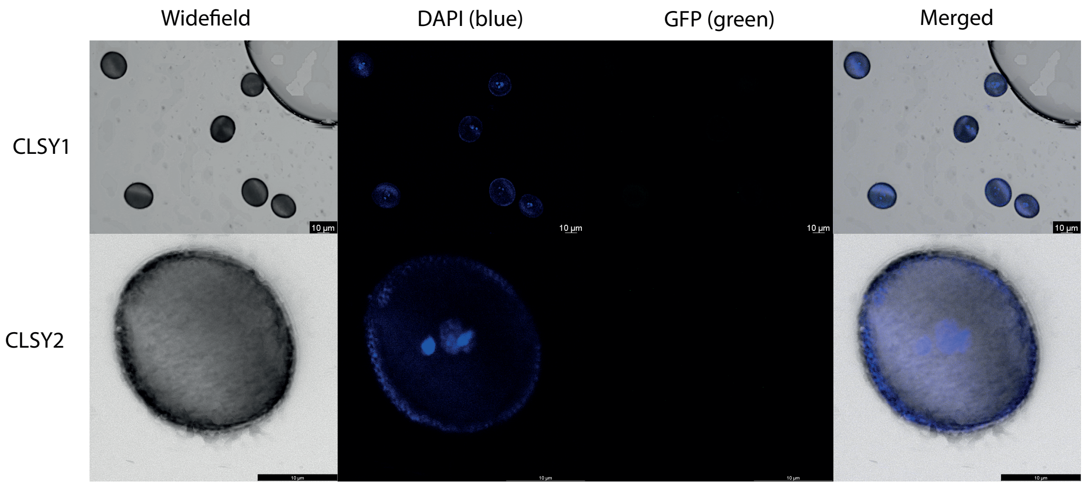


Fig. A2 CLSY1 and CLSY2 are not specifically expressed in the pollen

DAPI (blue, second row) staining of pCLSY1::CLY1-eGFP (first row) and pCLSY2::CLY2-eGFP (second row) showing no expression of CLSYs (third column, GFP, green) in mature pollen. Scale bar 10 μm. Reporter lines constructed by dr Jincheng Long.

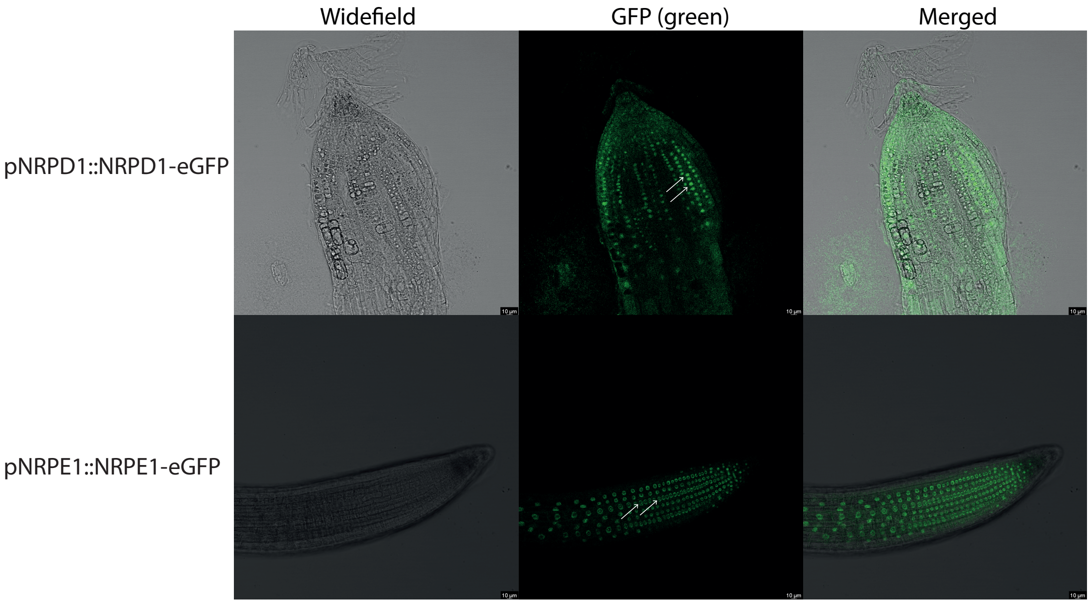


Fig. A3 pNRPD1::NRPD1-eGFP and pNRPE1::NRPE1-eGFP are strongly expressed in the root

eGFP (green, white arrows) nuclear signal of NRPD1 (first row) and NRPE1 (second row) in the root tip of *Arabidopsis*. Scale bar 10 μm. Reporter lines constructed by dr Jincheng Long.

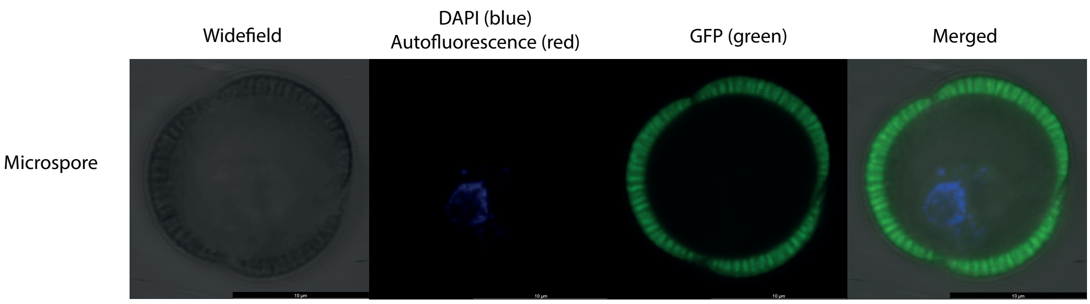


Fig. A4 pNRPD1 is not consistently expressed in germline tissues

NRPD1 expression in microspores. eGFP (green) DAPI (blue). Scale bar 10 μm. Reporter lines constructed by dr Jincheng Long.

Line name	Background	**Source**
pCLSY1::CLSY1-eGFP	clsy1	dr Jincheng Long (Feng Lab)
pCLSY2::CLSY2-eGFP	clsy2	dr Jincheng Long (Feng Lab)
pCLSY::CLSY3-Venus	clsy3	dr James Walker (Feng Lab)
pNRPD1::NRPD1-eGFP	WT	dr Jincheng Long (Feng Lab)
pNRPE1::NRPE1-eGFP	WT	dr Jincheng Long (Feng Lab)
Library name	**Library type**	**Source**
Seedling	sRNA	[152]
Root	sRNA	[153]
Leaf	sRNA	[154]
Tapetum	sRNA	Feng Lab
Meiocyte	sRNA	Feng Lab
Microspore	sRNA	Feng Lab (constructed by JT)
Sperm cell	sRNA	Feng Lab (constructed by JT)
Sperm nucleus	sRNA	Feng Lab (constructed by JT)
Pollen rep1	sRNA	[155]
Pollen rep2	sRNA	[156]
Pollen rep3	sRNA	[157]
Meiocyte	BS-seq	Feng Lab
Microspore	BS-seq	[158]
Sperm cell	BS-seq	Feng Lab
Vegetative cell	BS-seq	Feng Lab
Seedling	BS-seq	[159]

Table A1 Table depicting the line names, library types, and sources of the data presented in this chapter.

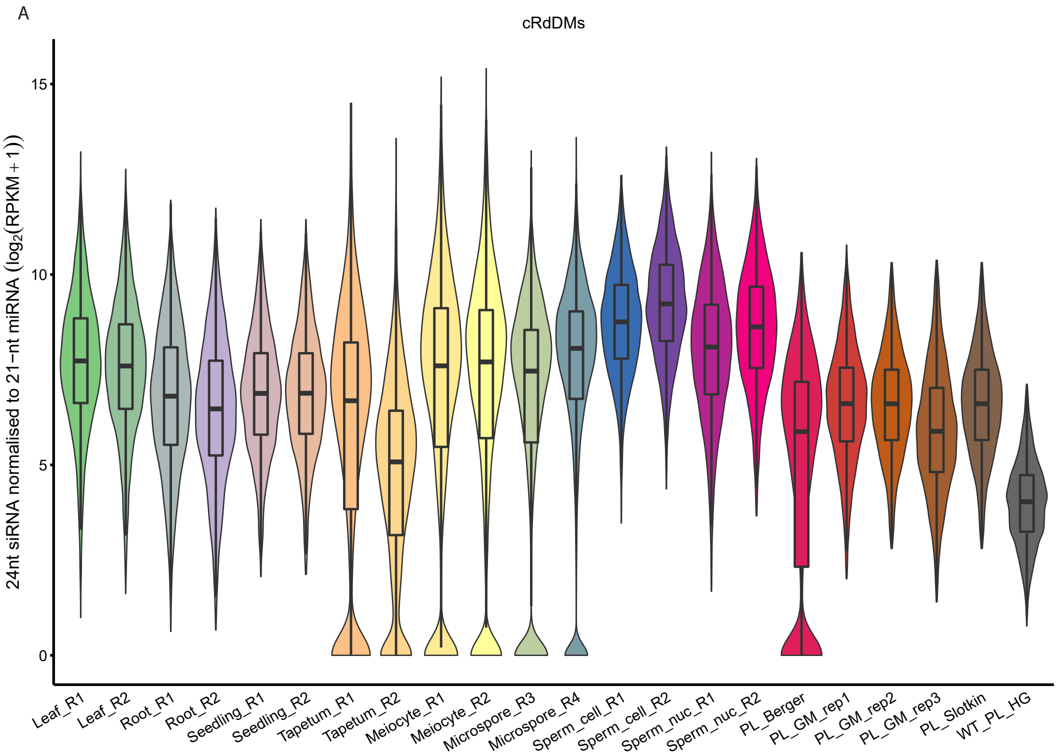


Fig. A5 Violin/box plots depicting 24nt sRNA abundance normalised against total 21nt miRNAs in different somatic and germline tissues

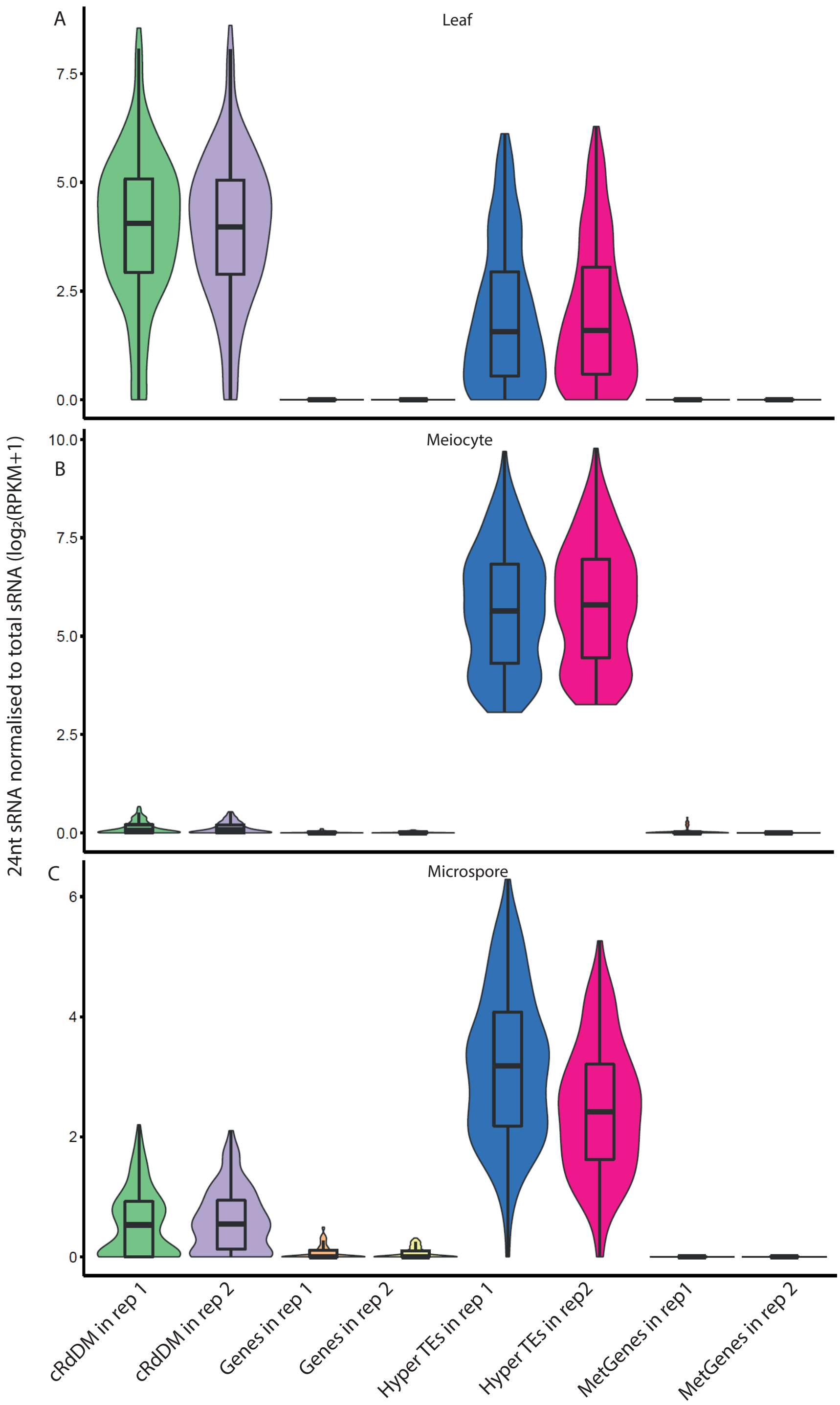
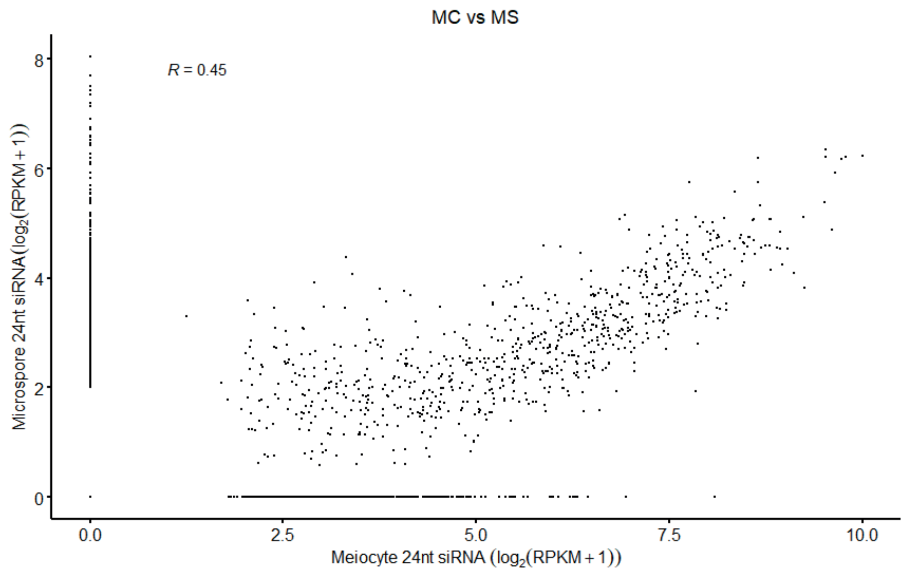


Fig. A6 Violin/box plots depicting 24nt sRNA abundance normalised against total sRNA of cRdDM, genic, HyperTE and MetGene loci in the leaf, meiocyte and microspore

A



B

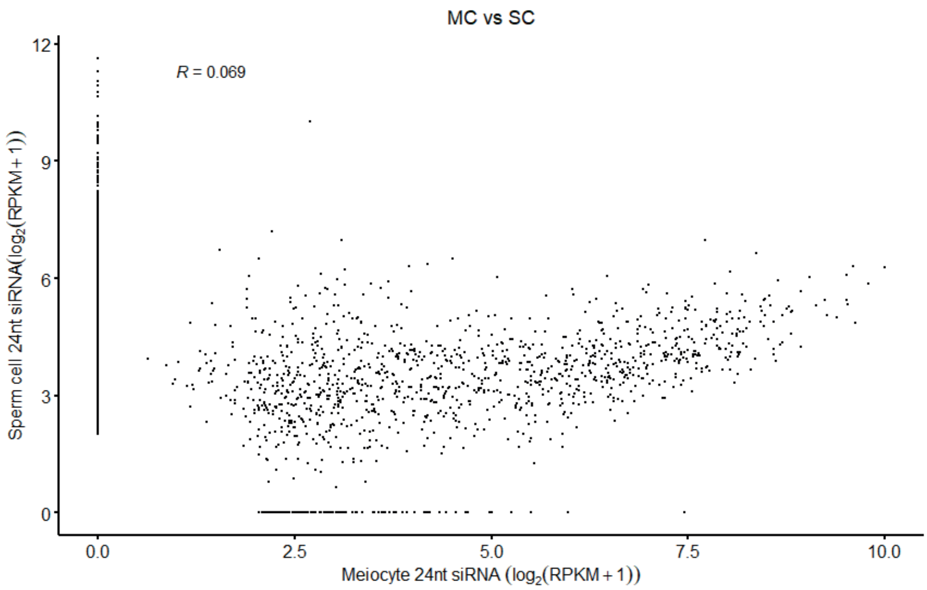


Fig. A7 Scatterplot comparing the abundance of 24nt sRNAs between A) the meiocyte and microspore and B) the meiocyte and the sperm cell

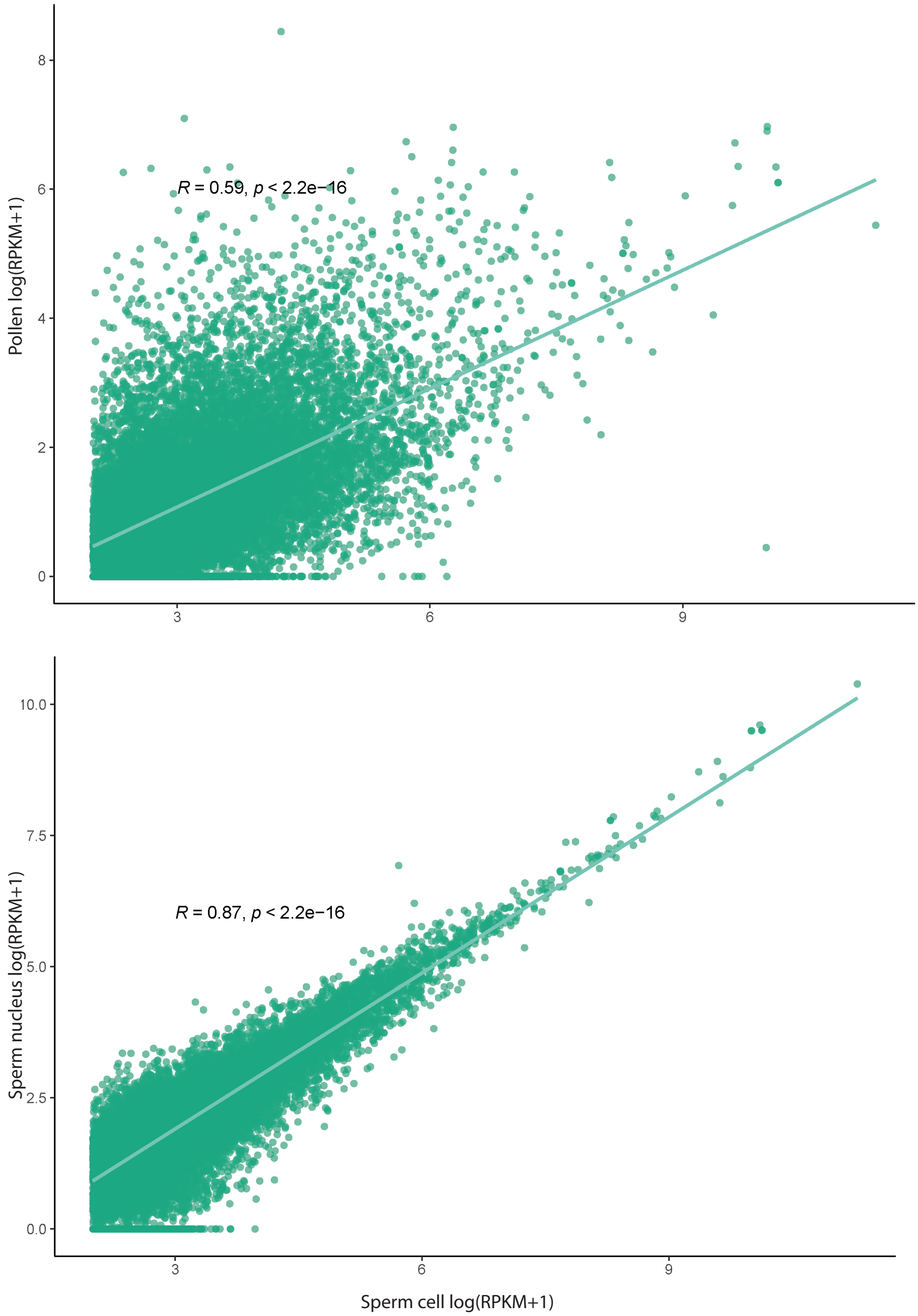


Fig. A8 Scatterplot comparing the abundance of 24nt sRNAs at sperm specific clusters between the pollen and the sperm cell (top graph) and the sperm nucleus and the sperm cell (bottom graph)

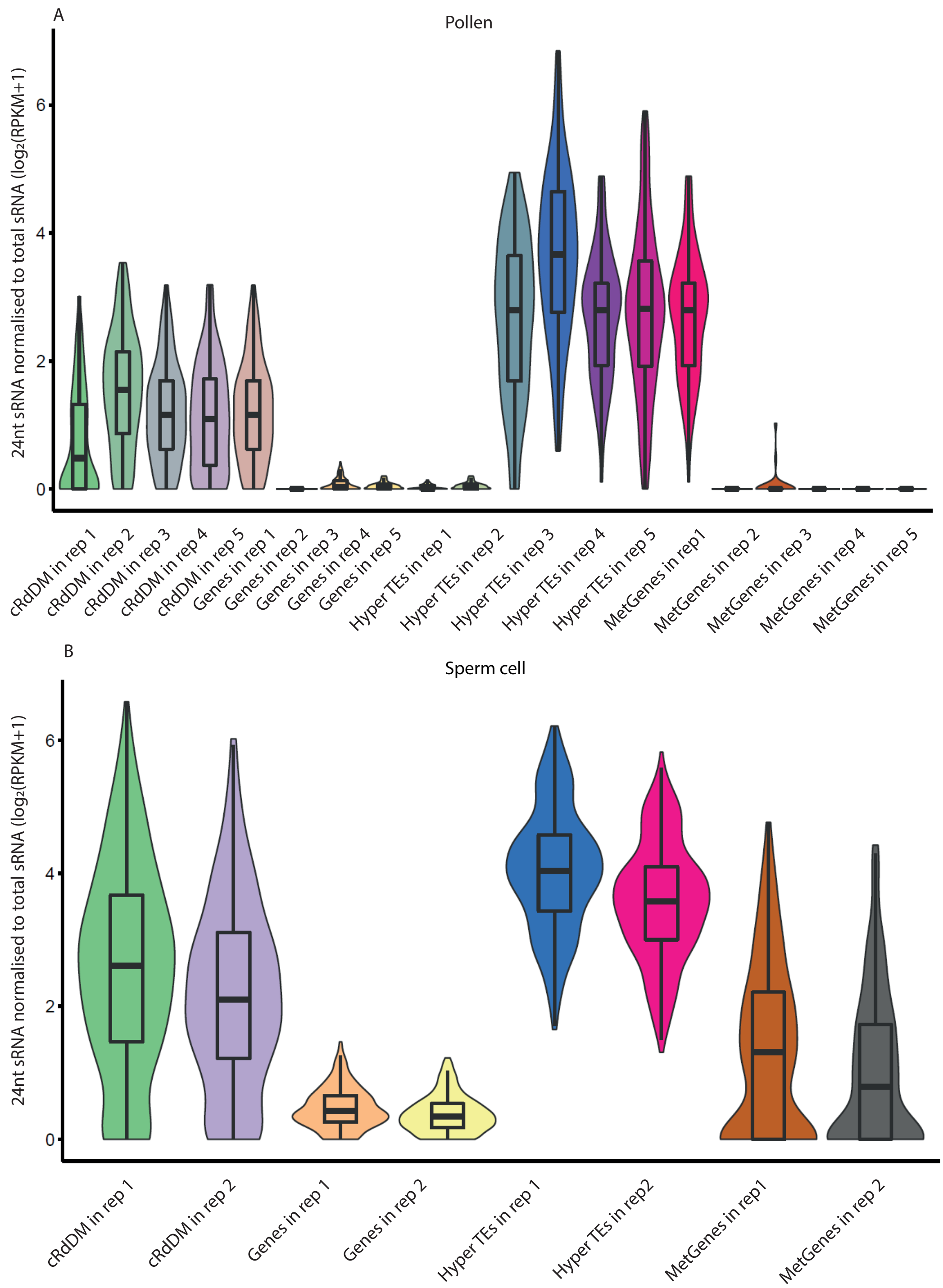


Fig. A9 Violin/box plots depicting 24nt sRNA abundance normalised against total sRNA of cRdDM, genic, HyperTE and MetGene loci in replicates of the A) pollen and B) sperm cell

	Source	Dr. James Walker			Dr. Toby Buttress		Dr. Shaoli Zhou		
Gene ID	Gene Name	Tapetum	Meiocyte	Pollen	Sperm	Vegetative	Meiocyte	Microspore	Sperm
AT3G42670	CLSY1	0.069	2.362	2.425	0.305	2.407	1.882	7.154	1.886
AT5G20420	CLSY2	0.124	1.030	1.634	1.014	8.164	0.752	2.074	0.000
AT1G05490	CLSY3	11.412	0.534	0.000	0.039	0.154	0.439	1.339	0.000
AT3G24340	CLSY4	0.245	22.269	3.276	16.006	1.530	20.507	5.112	143.407
AT2G16390	DRD1	6.335	11.622	1.706	0.214	0.553	10.439	5.727	0.000
AT1G15910	FDM1	12.840	14.499	1.132	0.268	0.574	15.091	11.453	0.513
AT4G00380	FDM2	2.826	10.899	0.000	0.018	0.011	10.419	4.151	0.000
AT3G12550	FDM3	2.521	5.639	0.000	0.159	0.273	4.889	5.072	0.056
AT1G13790	FDM4	7.197	11.961	42.011	87.597	3.240	9.969	10.073	1306.790
AT1G80780	FDM5	87.837	20.310	1517.160	3.485	32.138	14.331	26.673	6.443
AT5G15380	DRM1	1.083	8.019	0.000	0.000	0.164	8.406	1.952	0.614
AT5G14620	DRM2	17.582	0.774	0.000	0.080	0.039	0.953	5.015	4.406
AT3G17310	DRM3	5.015	1.266	0.260	0.000	0.000	1.818	5.923	32.001
AT3G22680	RDM1	28.999	6.994	3.681	0.000	0.000	7.715	2.573	0.256
AT5G04290	KTF1	2.233	35.857	10.397	1.171	1.340	37.257	17.184	0.450
AT3G49250	KTF1 hom	16.080	46.676	25.977	0.000	0.016	5.805	3.582	0.239
AT2G30280	KTF1 hom	0.602	44.270	8.030	1.831	2.452	0.731	9.876	29.358
AT4G36290	MORC1	3.172	2.488	0.000	0.131	0.096	3.051	4.647	0.066
AT4G36270	MORC2	0.485	16.507	0.000	0.000	2438.620	10.101	1.316	3.145
AT4G36270	MORC3	0.097	16.216	0.000	0.017	0.000	11.087	3.810	1.964
AT5G50780	MORC4	0.679	52.595	17.862	35.680	1.265	53.671	9.817	495.904
AT5G13130	MORC5	0.000	3.191	0.261	0.053	0.046	3.169	19.304	0.218
AT1G19100	MORC6	4.610	22.521	8.392	0.000	50.949	23.561	5.809	0.000

AT4G24970	MORC7	0.051	0.251	0.806	0.017	0.253	0.335	3.584	0.756
AT5G04940	SUVH1	3.269	4.217	0.000	0.281	0.159	4.097	6.219	0.708
AT2G33290	SUVH2	1.546	9.811	0.369	0.314	0.659	13.030	1.673	2.822
AT1G73100	SUVH3	3.373	0.659	3.395	1.077	0.491	0.526	4.473	10.688
AT5G13960	SUVH4	5.464	8.346	1.073	4.163	0.355	3.007	6.048	32.060
AT2G35160	SUVH5	8.072	24.464	7.386	16.224	2.036	21.982	17.397	250.727
AT2G22740	SUVH6	3.692	22.468	97.319	20.015	52.259	23.143	6.751	34.313
AT1G17770	SUVH7	0.000	0.318	0.838	0.320	0.168	0.295	0.000	6.145
AT2G24740	SUVH8	0.000	1.004	0.000	0.000	0.022	0.848	1.833	0.000
AT4G13460	SUVH9	4.288	1.660	5.555	0.960	0.996	1.520	3.572	1.321
AT2G05900	SUVH10	0.000	0.000	0.000	0.000	0.169	0.000	0.000	0.000
AT5G47150	SUVH hom	0.911	3.590	0.000	0.000	0.000	4.070	0.856	0.000
AT5G47160	SUVH hom	0.006	4.307	0.000	3.233	0.000	3.451	1.243	25.517
AT1G04050	SUVR1	30.801	20.204	13.628	7.794	0.156	3.913	7.442	46.306
AT5G43990	SUVR2	4.326	22.124	0.000	0.022	0.028	20.202	8.224	0.000
AT3G03750	SUVR3	9.598	1.024	7.831	1.528	1.647	1.600	4.982	3.445
AT3G04380	SUVR4	0.935	14.174	0.000	0.858	0.053	11.715	6.492	12.778
AT2G23740	SUVR5	4.494	1.785	4.050	1.021	3.892	1.802	20.790	0.178
AT4G08350	DMS3	18.867	6.984	0.000	4.929	8.594	59.555	23.154	39.541
AT2G34210	DMS4	15.176	0.640	0.000	1.831	2.452	0.731	9.876	29.358
AT4G35800	NRPB1	9.760	124.307	53.801	16.013	24.952	179.741	54.232	22.260
AT5G09920	NRPB4	33.043	3.486	10.789	2.164	0.803	3.303	23.280	154.460
AT3G22320	NRPB5A	44.975	5.953	5.603	3.139	0.925	4.689	26.076	23.576
AT5G57980	NRPB5L1	0.727	0.721	6.832	0.000	1134.040	0.511	12.250	0.000

AT5G45140	NRPC2	1.005	41.613	1.371	0.357	0.129	44.598	20.155	4.112
AT1G63020	NRPD1	0.000	0.769	0.000	0.558	0.016	0.880	5.191	1.894
AT3G23780	NRPD2/ NRPE2	14.986	52.199	0.247	0.200	0.063	62.401	13.488	0.000
AT3G18090	NRPD2B	0.033	4.055	0.000	0.006	0.005	5.035	1.692	0.000
AT4G15950	NRPD4	19.473	1.273	0.000	0.478	0.567	0.552	4.757	3.143
AT2G41340	NRPD5B	0.000	0.046	0.000	0.000	0.000	0.072	0.000	0.000
AT2G40030	NRPE1	1.082	24.894	0.000	0.000	0.000	24.162	8.895	0.000
AT3G57070	NRPE5A	29.421	0.216	5.750	0.000	0.000	0.113	5.075	0.000
AT3G54490	NRPE5C	17.226	4.276	0.000	0.000	0.000	2.725	11.066	0.000
AT3G16980	NRPE9A	6.392	1.796	0.000	0.135	0.610	2.602	3.470	6.034
AT4G16265	NRPE9B	0.379	2.491	0.000	0.000	0.000	1.092	6.801	0.000
AT4G21710	NRPPB2	12.819	48.481	7.926	3.972	6.622	49.733	23.705	0.608
AT1G48405	AGO1	92.316	62.973	0.320	0.399	0.752	84.635	45.999	0.721
AT1G31280	AGO2	1.897	3.440	0.243	0.099	0.296	4.270	8.758	0.000
AT1G31290	AGO3	0.000	0.040	0.000	0.047	1.069	0.023	0.138	0.038
AT2G27040	AGO4	4.670	13.424	0.315	0.127	0.070	13.311	62.406	0.000
AT2G27880	AGO5	29.327	23.341	5.436	0.000	527.075	26.161	11.675	41.270
AT2G32940	AGO6	0.271	9.677	0.481	0.567	0.014	9.734	3.238	2.067
AT1G69440	AGO7	0.000	0.821	0.000	0.113	0.432	0.659	0.693	0.000
AT5G21030	AGO8	0.000	0.011	0.000	4.486	15.940	0.016	0.000	0.601
AT5G21150	AGO9	53.737	110.374	49.874	75.117	1.354	128.692	28.993	739.026
AT5G43810	AGO10	3.353	2.119	0.991	0.467	0.034	2.410	2.815	0.575
AT1G01040	DCL1	5.333	1.900	2.926	3.523	1.612	1.612	17.886	17.858

AT3G03300	DCL2	7.214	15.916	1.286	0.171	0.353	5.327	14.302	0.000
AT3G43920	DCL3	6.432	2.550	0.000	0.066	0.719	2.545	15.484	3.168
AT5G20320	DCL4	2.318	26.588	0.000	0.013	0.107	28.061	2.743	0.071
AT4G11130	RDR2	0.368	3.019	0.946	0.399	0.599	2.796	2.608	11.298
AT3G49500	RDR6	3.296	18.003	0.000	0.131	0.048	18.123	6.580	0.000
AT5G23570	SGS3	8.552	73.316	0.000	0.100	0.160	83.520	13.197	1.387

Table A2 FPKM values of RdDM components in germline tissues (data from Dr. James Walker, Dr Shaoli Zhou and Dr. Toby Buttress)

	Source	Dr. James Walker				Dr. Shaoli Zhou		
Gene ID	Gene name	Seedling	Rosette	Columella	Embryo	Seedling	Leaf	Root
AT3G42670	CLSY1	0.961	1.208	13.150	2.733	1.659	1.316	1.733
AT5G20420	CLSY2	0.079	0.061	1.083	0.000	0.165	0.000	0.120
AT1G05490	CLSY3	0.092	0.022	0.027	20.036	0.057	0.037	0.203
AT3G24340	CLSY4	0.109	0.040	0.126	0.256	0.216	0.150	0.209
AT2G16390	DRD1	1.626	2.243	2.916	1.204	2.263	3.352	1.793
AT1G15910	FDM1	5.530	4.794	6.337	15.276	5.959	7.078	5.874
AT4G00380	FDM2	1.134	0.716	1.026	9.490	2.840	1.409	0.595
AT3G12550	FDM3	2.176	2.500	3.021	1.139	1.355	4.527	2.760
AT1G13790	FDM4	0.894	0.887	1.007	0.124	0.832	2.610	0.254
AT1G80780	FDM5	54.647	67.899	62.413	107.629	8.427	25.398	14.350
AT5G15380	DRM1	0.019	0.019	0.004	4.552	0.000	0.016	0.033
AT5G14620	DRM2	2.704	3.862	5.379	4.888	4.719	2.662	3.300
AT3G17310	DRM3	4.623	5.665	5.569	3.255	7.510	6.246	4.070
AT3G22680	RDM1	17.900	14.269	22.518	366.562	11.155	4.951	16.888
AT5G04290	KTF1	3.203	3.124	3.715	7.349	6.860	15.337	5.311
AT3G49250	KTF1 hom	9.321	11.530	17.212	25.813	2.319	2.388	2.370
AT2G30280	KTF1 hom	0.214	0.082	0.053	0.120	4.624	7.242	5.133
AT4G36290	MORC1	3.733	4.219	5.787	2.180	3.048	8.723	5.013
AT4G36270	MORC2	1.298	6.368	1.383	0.032	0.965	3.306	0.057
AT4G36270	MORC3	0.645	0.924	0.780	0.000	0.068	0.829	0.738
AT5G50780	MORC4	2.468	2.601	2.204	3.015	2.191	6.396	4.145

AT5G13130	MORC5	0.017	0.161	0.000	0.060	0.307	0.263	0.047
AT1G19100	MORC6	2.030	2.228	5.795	0.409	2.092	2.747	2.192
AT4G24970	MORC7	1.223	3.298	2.819	0.000	1.512	4.043	1.315
AT5G04940	SUVH1	3.438	2.470	4.917	6.052	3.990	4.242	4.225
AT2G33290	SUVH2	2.245	2.998	1.156	5.306	4.415	4.289	2.697
AT1G73100	SUVH3	3.982	3.510	11.081	13.420	6.664	6.552	4.452
AT5G13960	SUVH4	2.410	2.030	3.627	47.075	3.758	1.401	5.955
AT2G35160	SUVH5	0.910	0.324	0.894	4.801	0.545	0.406	1.354
AT2G22740	SUVH6	2.887	4.517	5.086	2.496	4.822	7.003	4.470
AT1G17770	SUVH7	0.000	0.000	0.000	0.000	0.000	0.000	0.000
AT2G24740	SUVH8	0.015	0.000	0.007	0.558	0.000	0.000	0.027
AT4G13460	SUVH9	7.398	8.519	6.302	4.992	7.239	11.555	7.267
AT2G05900	SUVH10	0.000	0.000	0.000	0.000	0.103	0.000	0.000
AT5G47150	SUVH hom	0.033	0.022	0.000	0.000	0.065	0.000	0.188
AT5G47160	SUVH hom	0.354	0.131	1.544	0.000	0.293	0.105	0.695
AT1G04050	SUVR1	18.759	6.641	11.463	631.832	0.603	0.632	1.471
AT5G43990	SUVR2	1.374	1.293	3.316	4.269	1.004	1.233	2.264
AT3G03750	SUVR3	4.171	5.450	6.394	51.586	6.049	3.844	4.321
AT3G04380	SUVR4	1.302	1.193	2.034	1.404	0.791	2.723	1.687
AT2G23740	SUVR5	2.568	2.736	15.758	0.041	2.324	5.222	3.804
AT4G08350	DMS3	2.643	2.858	4.623	12.359	11.099	18.797	18.205
AT2G34210	DMS4	4.667	4.279	2.068	6.444	4.624	7.242	5.133
AT4G35800	NRPB1	8.264	11.717	27.087	4.111	15.824	42.750	16.185

AT5G09920	NRPB4	34.759	36.097	52.051	71.478	29.417	22.003	26.689
AT3G22320	NRPB5A	36.980	38.321	45.945	36.660	48.069	26.584	40.246
AT5G57980	NRPB5L1	4.188	0.049	4.656	0.000	0.662	0.000	10.906
AT5G45140	NRPC2	3.125	2.325	3.281	1.950	2.685	4.914	5.913
AT1G63020	NRPD1	0.542	0.847	7.558	0.337	0.915	1.334	0.975
AT3G23780	NRPD2/ NRPE2	2.514	3.646	8.900	21.602	4.794	6.592	4.763
AT3G18090	NRPD2B	1.075	0.643	0.740	0.931	0.470	1.057	0.715
AT4G15950	NRPD4	6.346	8.052	8.840	3.641	7.663	5.590	4.387
AT2G41340	NRPD5B	3.469	0.685	0.045	0.000	3.643	0.053	0.328
AT2G40030	NRPE1	0.820	0.753	2.561	9.645	0.914	1.668	0.818
AT3G57070	NRPE5A	19.086	24.535	20.167	18.248	17.301	4.845	18.518
AT3G54490	NRPE5C	0.056	0.256	0.352	7.184	0.000	0.000	0.000
AT3G16980	NRPE9A	1.164	0.689	3.998	35.167	1.109	1.156	1.084
AT4G16265	NRPE9B	14.513	10.429	5.805	130.605	7.653	2.782	13.837
AT4G21710	NRPPB2	7.853	9.516	21.464	7.309	8.877	13.461	15.022
AT1G48405	AGO1	18.825	22.312	46.941	94.760	20.245	19.846	42.322
AT1G31280	AGO2	3.051	2.911	8.118	45.176	2.932	4.019	5.520
AT1G31290	AGO3	0.564	0.105	0.185	0.582	1.101	0.273	0.033
AT2G27040	AGO4	5.175	5.423	11.535	118.524	5.237	7.882	8.325
AT2G27880	AGO5	0.023	0.000	0.000	43.974	0.019	0.018	0.019
AT2G32940	AGO6	0.469	0.514	0.775	1.389	0.322	0.203	0.658
AT1G69440	AGO7	0.886	1.762	0.479	2.809	1.853	3.459	0.849

AT5G21030	AGO8	0.015	0.042	0.181	0.000	0.000	0.000	0.023
AT5G21150	AGO9	0.105	0.023	0.022	231.763	0.068	0.370	0.000
AT5G43810	AGO10	5.387	4.099	4.786	24.060	5.745	10.163	4.985
AT1G01040	DCL1	9.284	12.103	4.714	0.287	1.368	9.010	3.317
AT3G03300	DCL2	6.672	7.583	9.472	1.559	1.049	3.269	3.074
AT3G43920	DCL3	0.745	1.353	2.398	0.385	0.680	1.969	0.864
AT5G20320	DCL4	1.892	3.328	4.931	0.435	1.517	3.871	2.199
AT4G11130	RDR2	0.842	0.976	10.446	0.446	1.477	1.104	1.943
AT3G49500	RDR6	6.338	8.269	6.807	65.028	6.535	7.052	5.326
AT5G23570	SGS3	3.240	2.973	6.134	4.181	3.217	9.856	4.575

Table A3 FPKM levels of somatic RdDM components (data from Dr. James Walker, Dr Shaoli Zhou and Dr. Toby Buttress)

Gene name	Gene ID	Tapetum	Meiocyte	Microspore	Sperm rep1	Sperm rep2	Vegetative	Pollen	Embryo	Seedling	Leaf
VIM4	AT1G66040	0.000	0.208	0.278	129.821	4.492	0.341	9.202	0.328	0.011	0.000
VIM2	AT1G66050	0.099	2.658	7.049	35.078	0.622	0.055	1.011	5.699	0.290	0.393
VIM1	AT1G57820	14.205	38.148	9.812	317.158	0.000	1320.830	4.708	67.593	6.529	3.569
VIM3	AT5G39550	3.176	41.704	23.178	188.381	21.844	0.510	11.350	20.306	5.242	1.057
VIM5	AT1G57800	0.000	0.037	1.877	63.004	4.596	0.300	2.614	0.697	0.005	0.000

Table A4 VIM FPKM expression in soma and germline (data from Dr. James Walker, Dr Shaoli Zhou and Dr. Toby Buttress)

Appendix B - Chapter 3

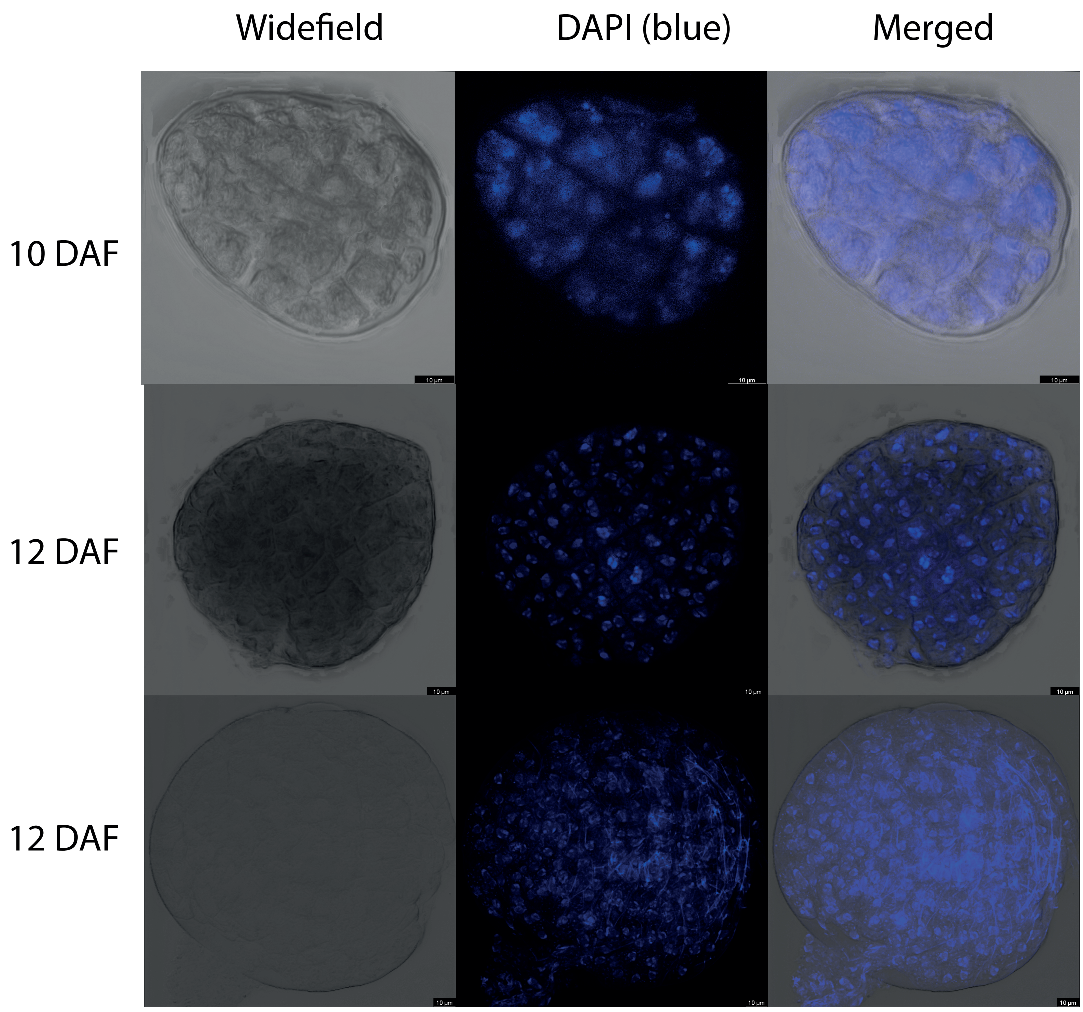


Fig. B1 *Marchantia* embryos develop at different rates in different environmental conditions

DAPI stained *Marchantia* embryos at 10 days after fertilisation (first row), and 12 days after fertilisation (second and third rows) showing different developmental stages. Scale bar 10 μm.

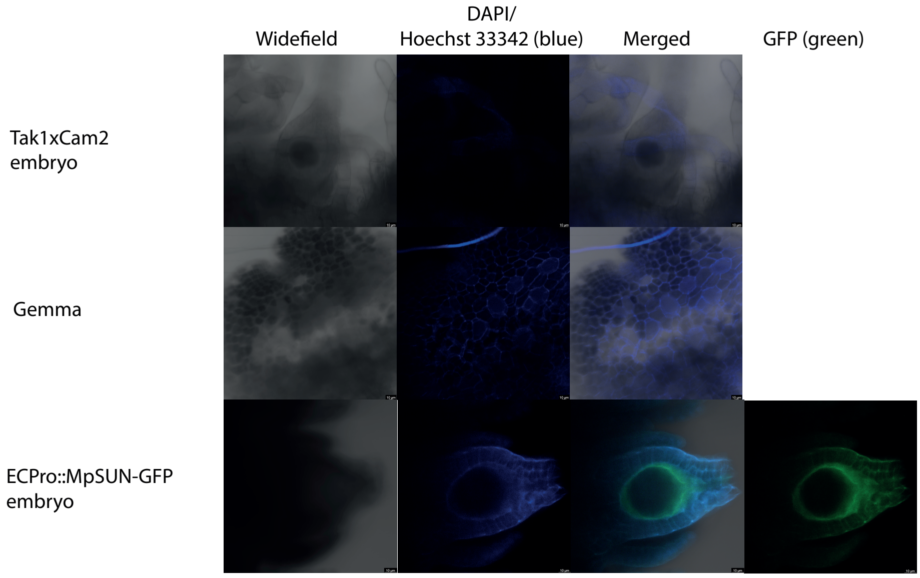


Fig. B2 Fixing and staining results of wild type (first row) ECPro::MpSUN-GFP embryos (third row) and gemmae (second row). Scale bar 10 μm.

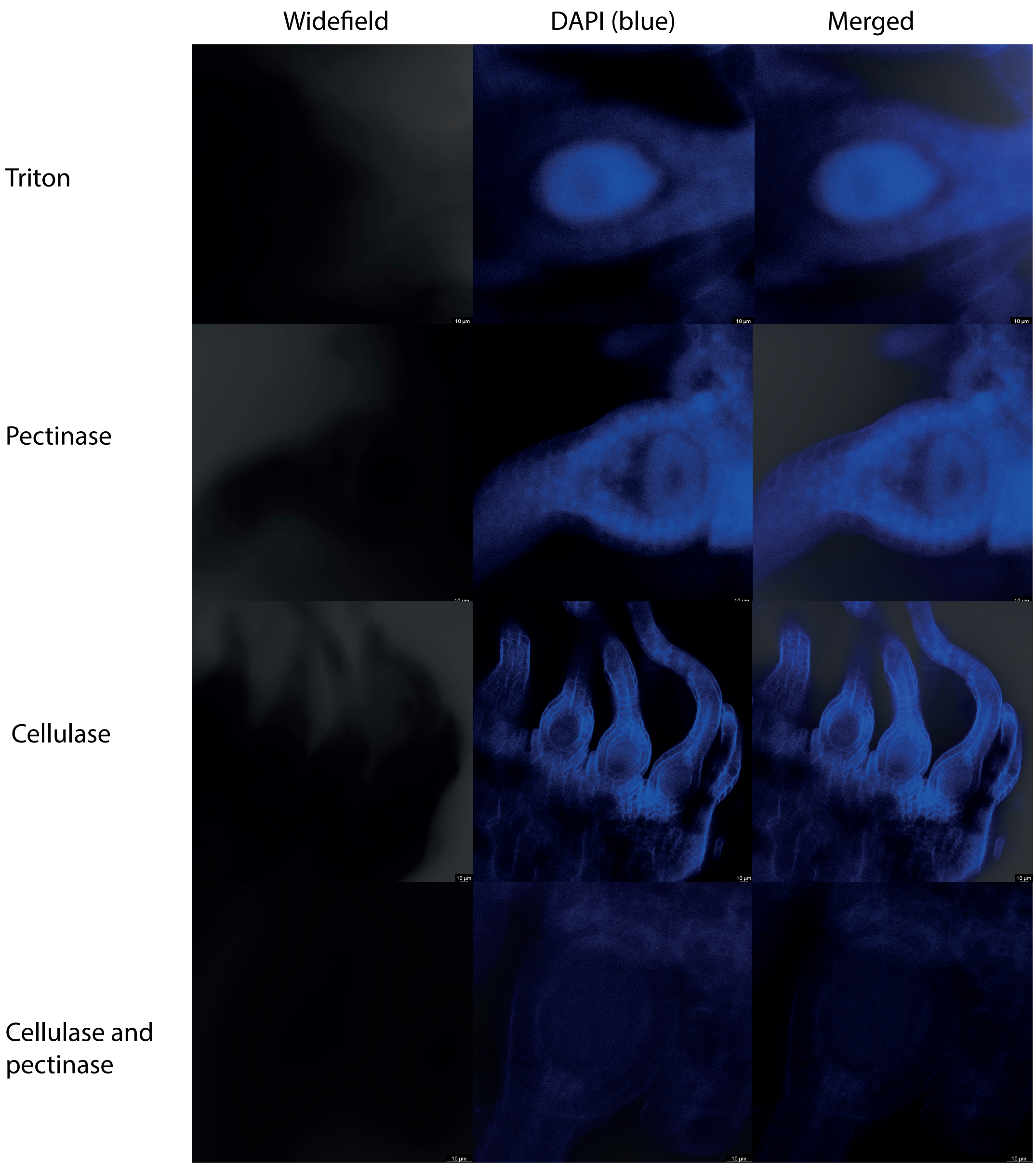


Fig. B3 Fixing and staining results of embryos after treatment with cell wall digesting enzymes

Scale bar 10 μm.

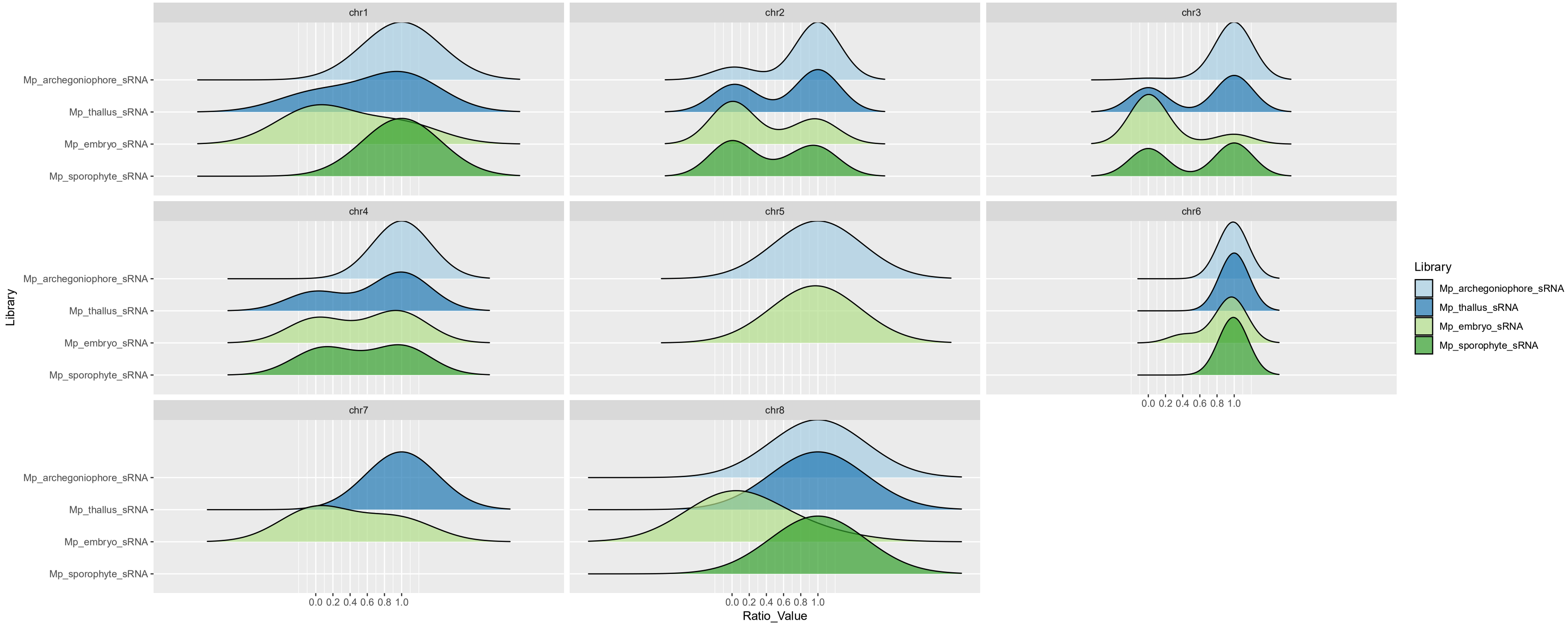


Fig. B4 Ridge plot to show the SNP ratio of sRNAs in archegoniophores, thallus, embryo and sporophyte, faceted by chromosome.

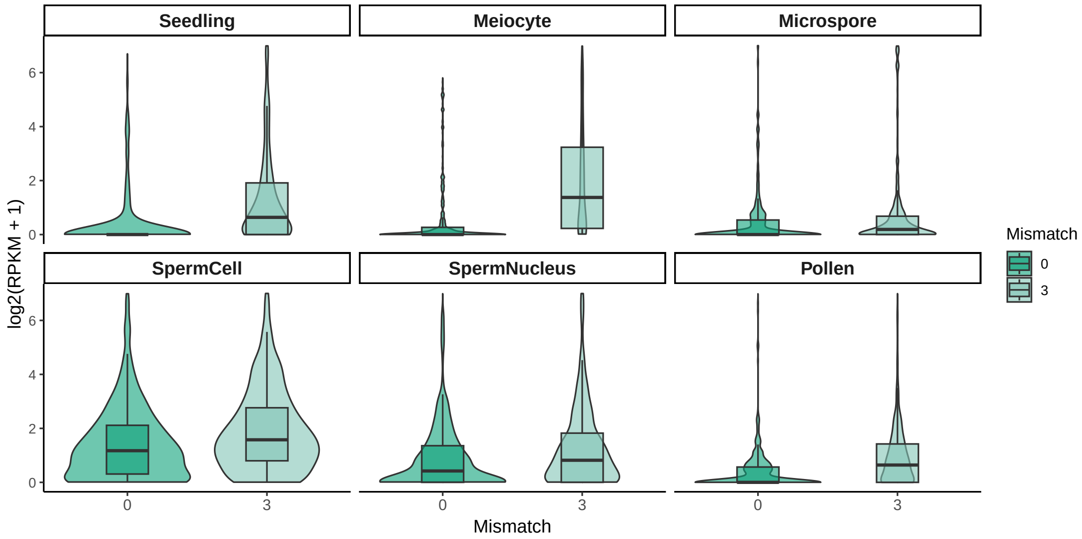


Fig. B5 Violin/box plots of the abundance of 24nt sRNAs in seedling, meiocyte, sperm cell, sperm nucleus and pollen in *Arabidopsis* mapped with 0 (dark green) and 3 (light green) mismatches

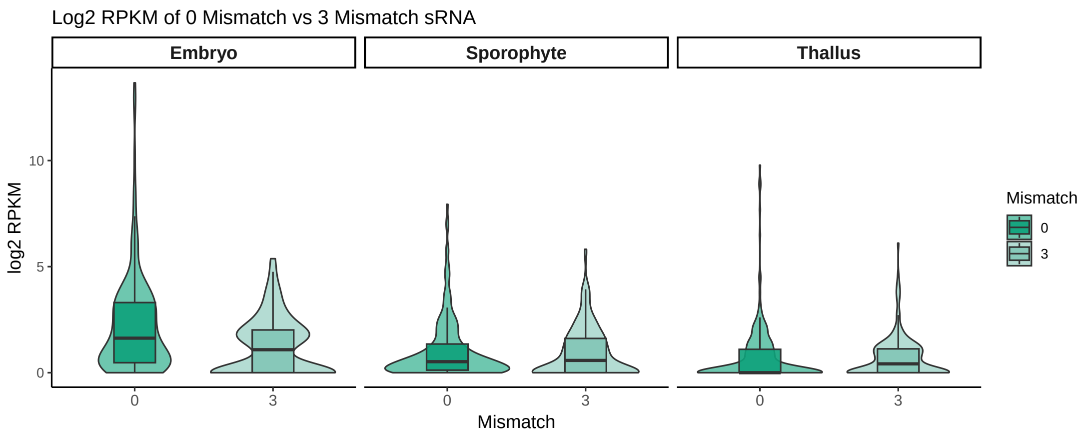


Fig. B6 Violin/box plots of the abundance of 24nt sRNAs in the thallus, embryo and sporophyte in *Marchantia* mapped with 0 (dark green) and 3 (light green) mismatches

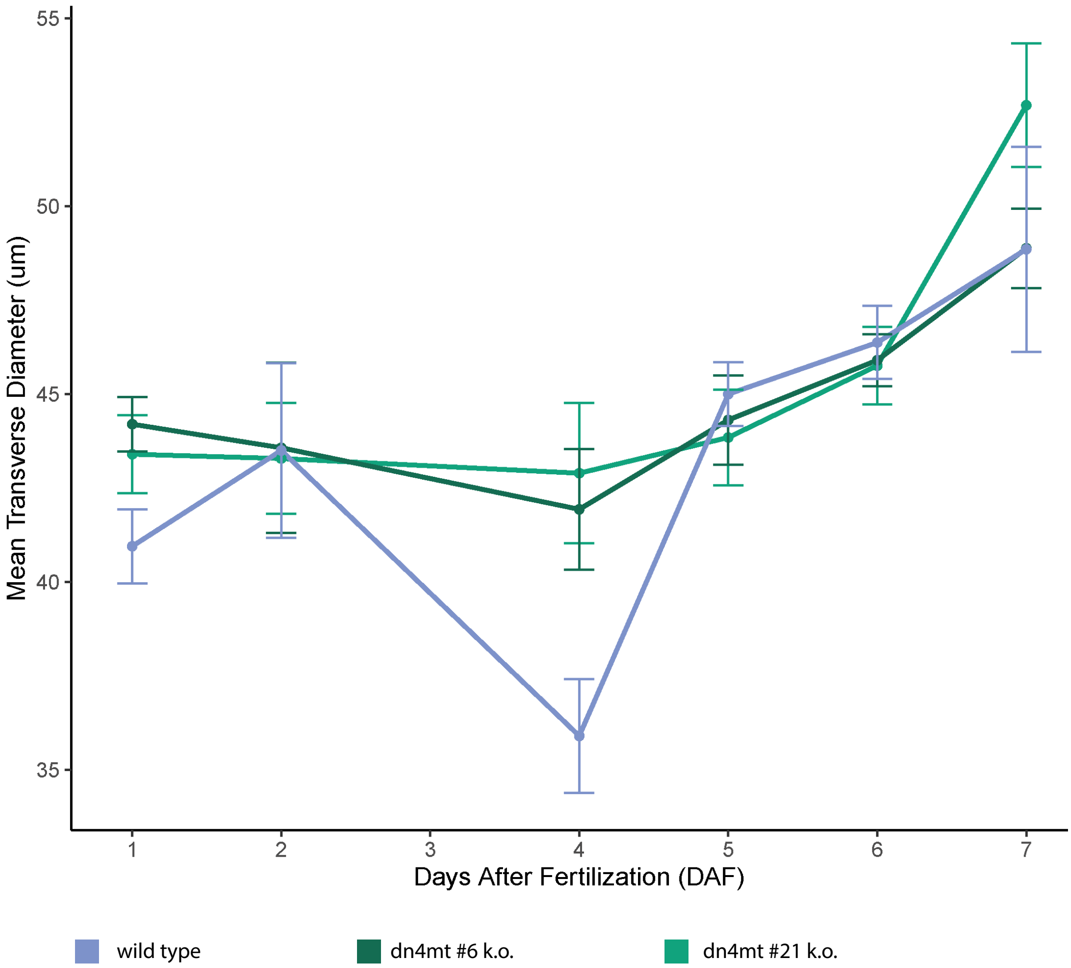


Fig. B7 Line graph that shows the transverse diameter of embryos fertilised by wild type (blue) or 2 independent 4mC mutant (green) sperm.

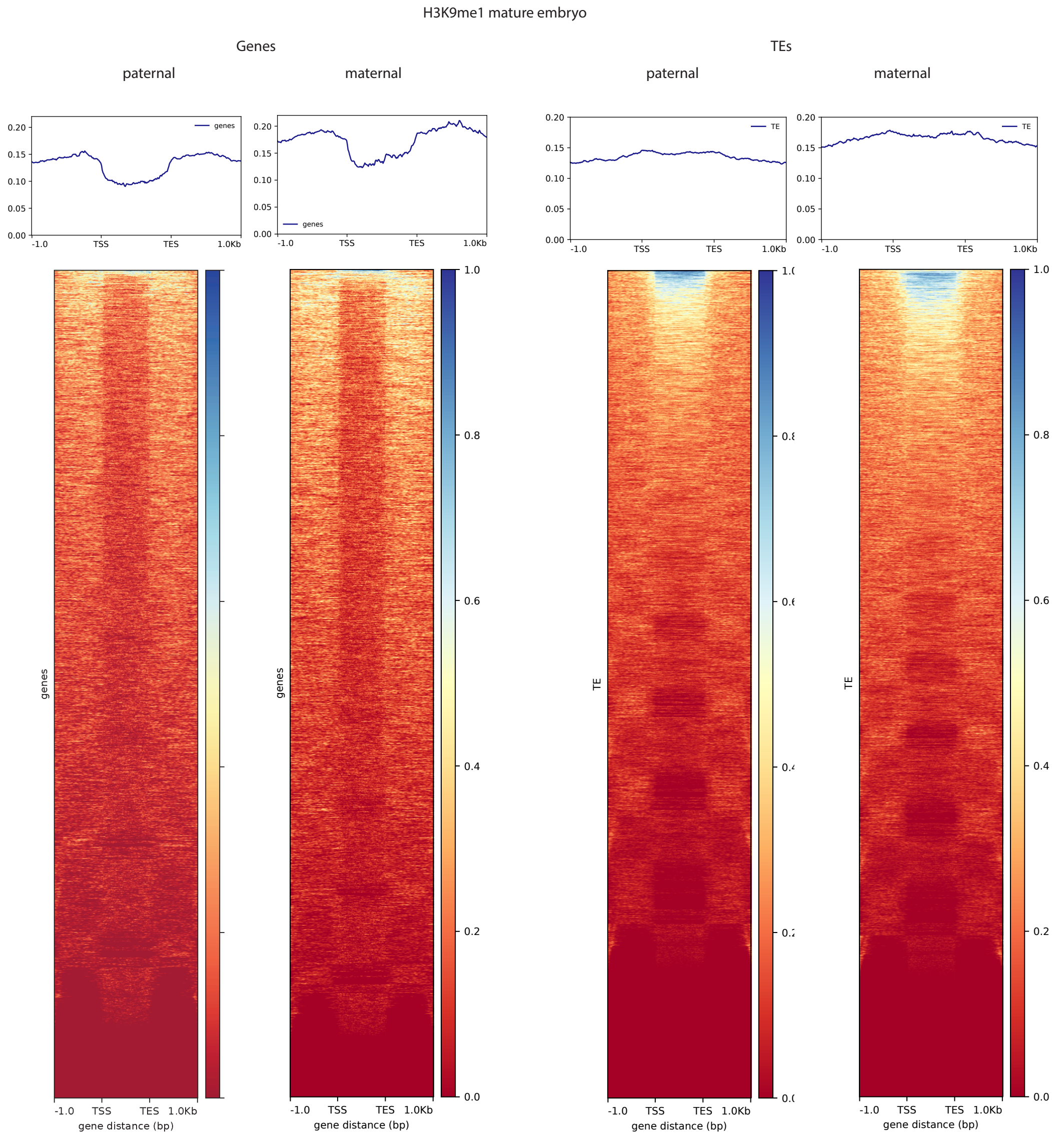


Fig. B8 H3K9me1 over genes and TEs is not associated with the paternal genome

Heatmap showing the relative distribution of H3K9me1 over genes and TEs associated with paternal or maternal genomes. Data presented is from [128]

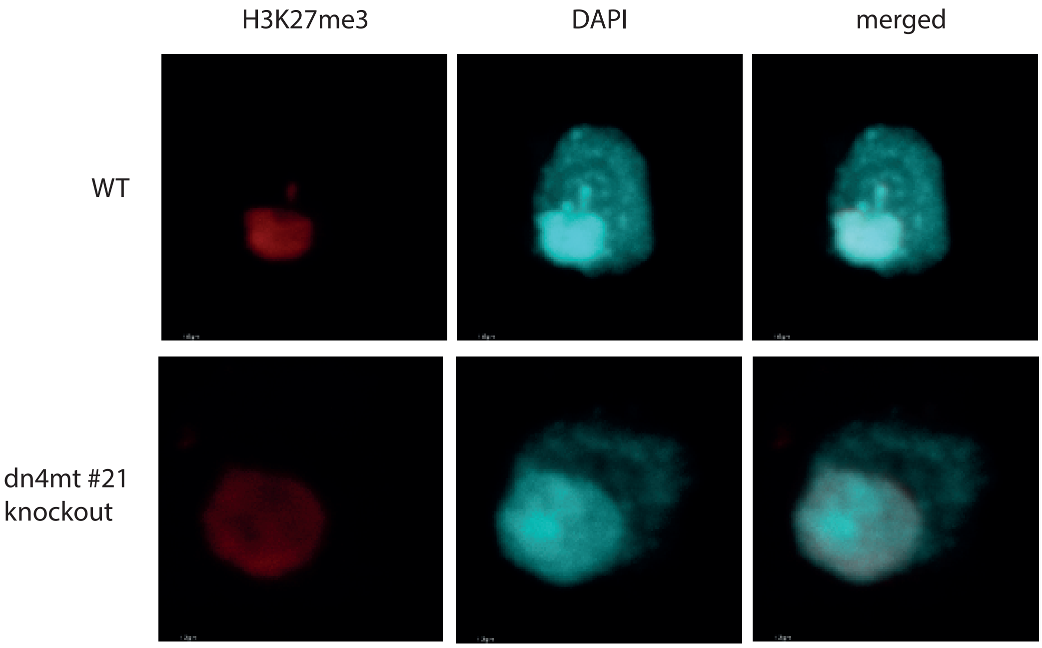


Fig. B9 Immunostaining of H3K27me3 in 13 DAF embryos fertilised by wild type and dn4mt1 #21 knockout sperm

Fig. B10 Plasmid map of pMpGWB102(35S)-citrine-NLS

Table B1 Primers used for nuclear reporter line cloning

Construct Target	Name	Sequence	Purpose
pMpGWB115 or pMpGWB116	attB1-fluo-NLS-attB2-f1	GGGGACAAGTTTGTACAAAAAAG CAGGCTTAATGGTGAGCAAGGGCG AG	citrine/tdTom insert + attB sites
pMpGWB115 or pMpGWB117	attB1-fluo-NLS-attB2-r1	GGGGACCACTTTGTACAAGAAAGC TGGGTTCTATCCTCCAACCTTTCTC TTCTTCTTAGG	citrine/tdTom insert + attB sites
pMpGWB102-Citrine-NLS & pMpGWB103-Citrine-NLS entry clones (pDONR221)	citrine-ent-seq-f1	TTAATGGTGAGCAAGGGC	Entry clone sequencing
	citrine-ent-seq-r1	TGGGTTCTATCCTCCAACC	Entry clone sequencing
pMpGWB102-tdTom-NLS & pMpGWB103-tdTom-NLS entry clones (pDONR221)	tdTom-ent-seq-f1	AGGCTTAATGGTGAGCAAG	Entry clone sequencing
	tdTom-ent-seq-r1	GGGTTCTATCCTCCAACC	Entry clone sequencing
pMpGWB102-Citrine-NLS	102-citrine-exp-seq-f1	TTGGAGAGAACACGGG	Expression clone sequencing
	102-citrine-exp-seq-r1	CTGGGTTCTATCCTCCAAC	Expression clone sequencing
pMpGWB102-tdTom-NLS	102-tdTom-exp-seq-f1	AGAGAACACGGGGGACTC	Expression clone sequencing
	102-tdTom-exp-seq-r2	TCGGAGGAGGCGGTG	Expression clone sequencing
	102-tdTom-exp-seq-r1	ACAAGAAAGCTGGGTTCTATCCTC	Expression clone sequencing
pMpGWB103-Citrine-NLS	103-citrine-exp-seq-f1	CCTCGAGCGAGTGATTTTTTAGG	Expression clone sequencing
	103-citrine-exp-seq-r1	GGGTTCTATCCTCCAACCTTTC	Expression clone sequencing
pMpGWB103-tdTom-NLS	103-tdTom-exp-seq-f1	ATTTCGCTAATCATTCCCTAATTTC	Expression clone sequencing
	103-tdTom-exp-seq-r2	CGGCCATGTTGTTGTC	Expression clone sequencing
	103-tdTom-exp-seq-r1	AAGAAAGCTGGGTTCTATCC	Expression clone sequencing

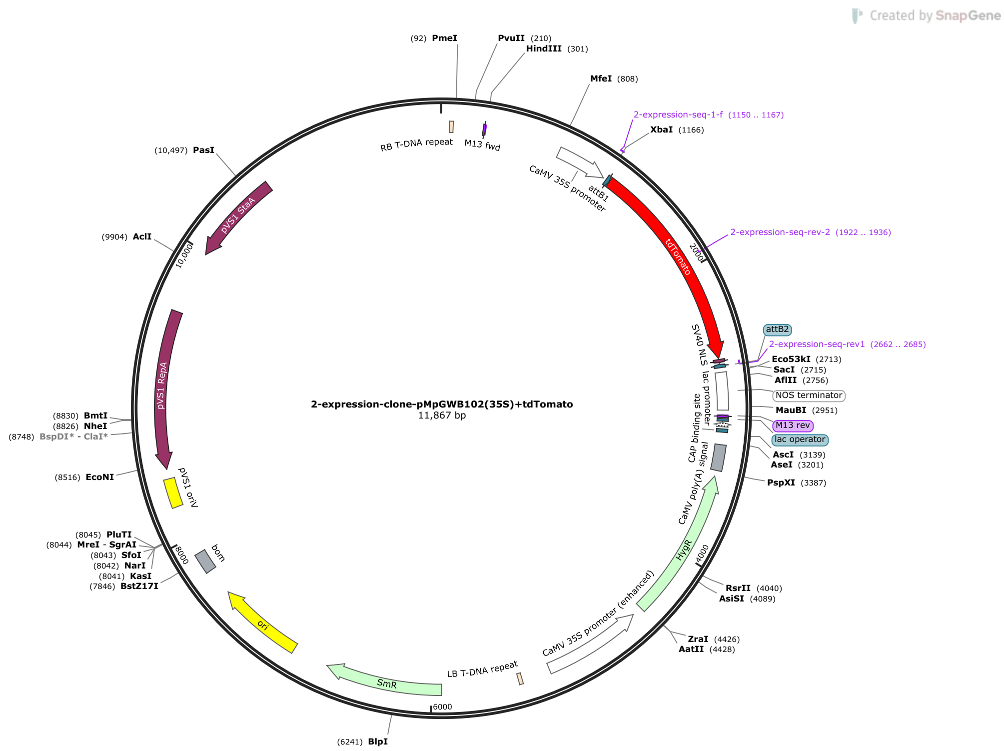


Fig. B11 Plasmid map of pMpGWB102(35S)-tdTomato-NLS

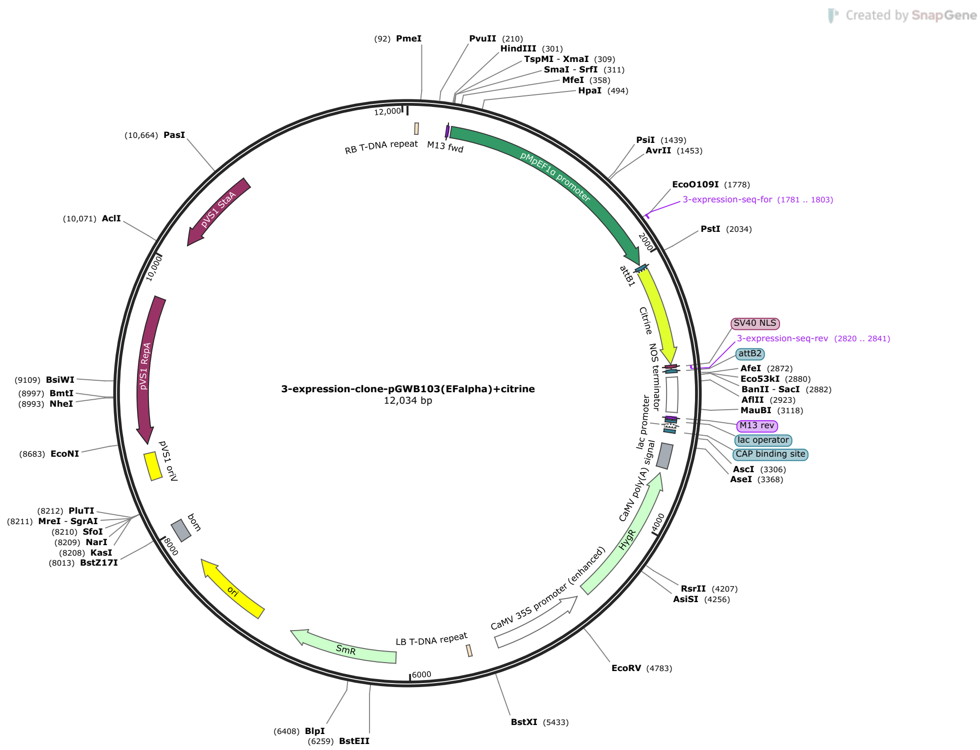


Fig. B12 Plasmid map of pMpGWB103(*EF1α*)-citrine-NLS

Line name		Source
Tak-1		Feng Lab
Tak-2		Feng Lab
Cam-2		Feng Lab
35s::tdTomato-NLS (Tak-1 male, WT female backgrounds)		Feng Lab (constructed by JT)
EF1α::citrine-NLS (Tak-1 male, WT female backgrounds)		Feng Lab (constructed by JT)
EF1α::tdTomato-NLS (Ta-k1 male, WT female backgrounds)		Feng Lab (constructed by JT)
dn4mt1 k.o. #6		Feng Lab [121]
dn4mt1 k.o. #21		Feng Lab [121]
dnmt3b k.o.		Feng Lab [121]
cmta k.o.		Feng Lab [121]
Library name	**Library type**	**Source**
WT sperm	EM-seq	Feng Lab (constructed by JT)
WT sperm	AMD-seq	Feng Lab (constructed by JT)
WT sperm	AMD-seq	Feng Lab
dn4mt1 k.o. sperm	AMD-seq	Feng Lab
14 DAF embryo	AMD-seq	Feng Lab (constructed by JT)
7-8 DAF embryo	A3A-sc-BS-seq	Feng Lab (constructed by JT)
Archegoniophore	sRNA	[160]
Embryo	sRNA	Feng Lab
Sporophyte	sRNA	Feng Lab
Thallus	sRNA	Feng Lab
Embryo	BS-seq	Feng Lab
Sporophyte	BS-seq	Feng Lab
Thallus	BS-seq	Feng Lab
Embryo	H3K27me3 CUT&RUN	[128]
Embryo	H3K9me1 CUT&RUN	[128]

Table B2 Table depicting the line names, library types, and sources of the data presented in this chapter.

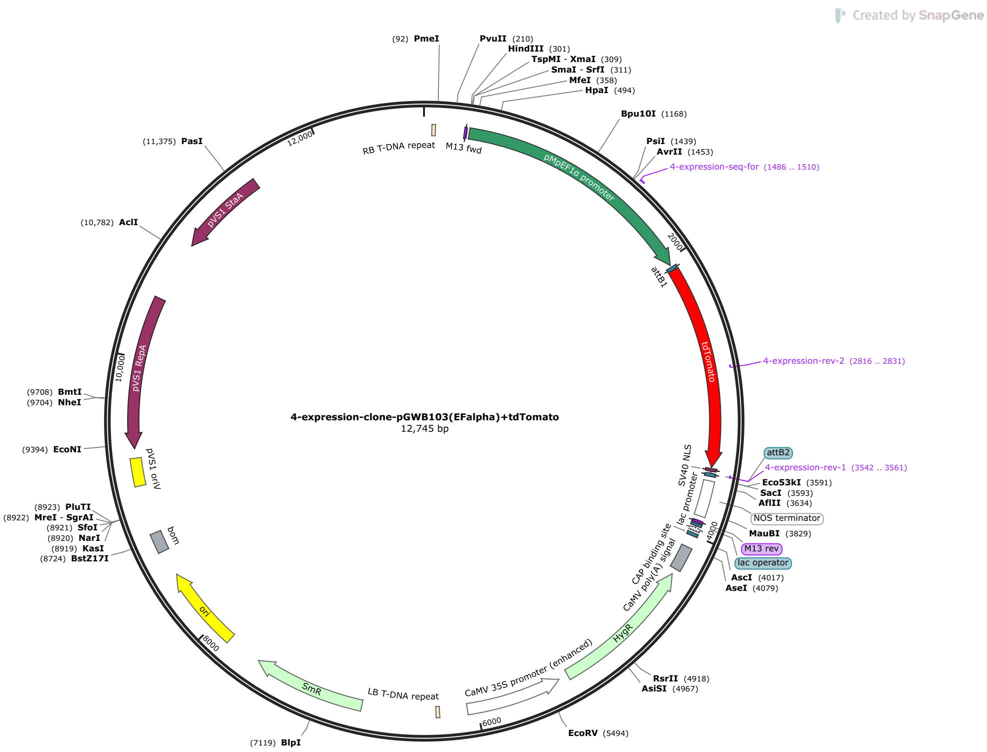


Fig. B13 Plasmid map of pMpGWB102(*EF1α*)-tdTomato-NLS

Glossary

4mC	N4-methylcytosine
5mC	5-methylcytosine
AGO	ARGONAUTE
AMD-seq	APOBEC3A-mediated deamination sequencing
BS-seq	bisulfite sequencing
CLSY	CLASSY
CMT	CHROMOMETHYLASE
cRdDM	canonical RdDM
DAF	days after fertilisation
DME	DEMETER
DML	DEMETER-LIKE
DMS	DEFECTIVE IN MERISTEM SILENCING
DNA	deoxyribonucleic acid
DNMT	DECREASED DNA METHYLATION
DRD	DEFECTIVE IN RNA-DIRECTED DNA METHYLATION
DRM	DOMAINS REARRANGED METHYLTRANSFERASE
easiRNA	epigenetically active small interfering RNA
EFα	elongation factor 1α
EM-seq	Enzymatic methylation sequencing
FPKM	fragments per kilobase of transcript per million mapped reads
Gb	gigabases

gDNA	genomic DNA
KTF	KOW DOMAIN-CONTAINING TRANSCRIPTION FACTOR
LTR	long-terminal-repeat
Mb	megabases
MET1	METHYLTRANSFERASE 1
miRNA	micro RNA
MORC	MICRORCHIDIA
NERD	NEEDED FOR RDR2-INDEPENDENT DNA METHYLATION
NRPD	NUCLEAR RNA POLYMERASE D
NRPE	NUCLEAR RNA POLYMERASE E
nt	nucleotide
PGTS	post-transcriptional gene silencing
piRNA	PIWI-interacting small RNA
Pol IV	RNA Polymerase IV
Pol V	RNA Polymerase V
PTGS	post-transcriptional gene silencing
RdDM	RNA-directed DNA methylation
RDM1	RNA-DIRECTED DNA METHYLATION1
RDR	RNA-DEPENDENT RNA POLYMERASE
RNA	ribonucleic acid
ROS	REPRESSOR OF SILENCING
RPKM	reads per kilobase per million mapped reads
RPM	reads per million
SHH	SAWADEE HOMEODOMAIN HOMOLOGUE
SPT5L	SUPPRESSOR OF TY INSERTION 5 LIKE
sRNA	small RNA
siRNA	small interfering RNA

SUVH	SUPPRESSOR OF VARIEGATION 3–9 HOMOLOG
TE	transposable element
TET	Ten-eleven translocation
TGS	transcriptional gene silencing
TSS	transcriptional start site
tRFs	tRNA fragments
tRNA	transfer RNA

References

[1] Shelley L Berger, Tony Kouzarides, Ramin Shiekhattar, and Ali Shilatifard. An operational definition of epigenetics. *Genes & development*, 23(7):781–783, 2009.

[2] Tony Kouzarides. Chromatin modifications and their function. *Cell*, 128(4):693–705, 2007.

[3] B Jamge, ZJ Lorković, E Axelsson, A Osakabe, V Shukla, R Yelagandula, S Akimcheva, AL Kuehn, and F Berger. Histone variants shape chromatin states in *Arabidopsis*. elife 12: Rp87714. *DOI: https://doi. org/10.7554/eLife*, 87714, 2023.

[4] Wil AM Loenen, David TF Dryden, Elisabeth A Raleigh, Geoffrey G Wilson, and Noreen E Murray. Highlights of the DNA cutters: a short history of the restriction enzymes. *Nucleic acids research*, 42(1):3–19, 2014.

[5] María A. Sánchez-Romero, Ignacio Cota, and Josep Casadesús. DNA methylation in bacteria: from the methyl group to the methylome. *Current Opinion in Microbiology*, 25:9–16, 2015.

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

[7] Assaf Zemach, Ivy E. McDaniel, Pedro Silva, and Daniel Zilberman. Genome-wide evolutionary analysis of eukaryotic DNA methylation. *Science*, 328(5980):916–919, 2010.

[8] Heng Zhu, Guohua Wang, and Jiang Qian. Transcription factors as readers and effectors of DNA methylation. *Nature Reviews Genetics*, 17(9):551–565, 2016.

[9] S. E. Jacobsen and E. M. Meyerowitz. Hypermethylated SUPERMAN epigenetic alleles in *Arabidopsis*. *Science*, 277(5329):1100–3, 1997.

[10] T Ryan Gregory. The C-value enigma in plants and animals: a review of parallels and an appeal for partnership. *Annals of botany*, 95(1):133–146, 2005.

[11] Jaume Pellicer and Ilia J Leitch. The plant DNA C-values database (release 7.1): an updated online repository of plant genome size data for comparative studies. *New Phytologist*, 226(2):301–305, 2020.

[12] D. A. Hickey. Selfish DNA: a sexually-transmitted nuclear parasite. *Genetics*, 101(3-4):519–31, 1982.

[13] S. Ayarpadikannan and H. S. Kim. The impact of transposable elements in genome evolution and genetic instability and their implications in various diseases. *Genomics Inform*, 12(3):98–104, 2014.

[14] Z. Joly-Lopez and T. E. Bureau. Diversity and evolution of transposable elements in *Arabidopsis*. *Chromosome Res*, 22(2):203–16, 2014.

[15] A. Miura, S. Yonebayashi, K. Watanabe, T. Toyama, H. Shimada, and T. Kakutani. Mobilization of transposons by a mutation abolishing full DNA methylation in *Arabidopsis*. *Nature*, 411(6834):212–4, 2001.

[16] German Martinez, Sarah G Choudury, and R Keith Slotkin. tRNA-derived small RNAs target transposable element transcripts. *Nucleic acids research*, 45(9):5142–5152, 2017.

[17] Andrea J Schorn, Michael J Gutbrod, Chantal LeBlanc, and Rob Martienssen. LTR-retrotransposon control by tRNA-derived small RNAs. *Cell*, 170(1):61–71. e11, 2017.

[18] Upasna Sharma, Colin C Conine, Jeremy M Shea, Ana Boskovic, Alan G Derr, Xin Y Bing, Clemence Belleannee, Alper Kucukural, Ryan W Serra, and Fengyun Sun. Biogenesis and function of tRNA fragments during sperm maturation and fertilization in mammals. *Science*, 351(6271):391–396, 2016.

[19] Martín Muñoz-López and José L. García-Pérez. DNA transposons: nature and applications in genomics. *Current genomics*, 11(2):115–128, 2010.

[20] Thomas Wicker, François Sabot, Aurélie Hua-Van, Jeffrey L Bennetzen, Pierre Capy, Boulos Chalhoub, Andrew Flavell, Philippe Leroy, Michele Morgante, and Olivier Panaud. A unified classification system for eukaryotic transposable elements. *Nature Reviews Genetics*, 8(12):973–982, 2007.

[21] M. Sasaki, J. Lange, and S. Keeney. Genome destabilization by homologous recombination in the germ line. *Nat Rev Mol Cell Biol*, 11(3):182–95, 2010.

[22] Nicolas Buisine, Hadi Quesneville, and Vincent Colot. Improved detection and annotation of transposable elements in sequenced genomes using multiple reference sequence sets. *Genomics*, 91(5):467–475, 2008.

[23] Clémentine Vitte, Margaux-Alison Fustier, Karine Alix, and Maud I Tenaillon. The bright side of transposons in crop evolution. *Briefings in Functional Genomics*, 13(4):276–295, 2014.

[24] Cédric Feschotte. Transposable elements and the evolution of regulatory networks. *Nature Reviews Genetics*, 9(5):397–405, 2008.

[25] Clémentine Vitte and Jeffrey L Bennetzen. Analysis of retrotransposon structural diversity uncovers properties and propensities in angiosperm genome evolution. *Proceedings of the National Academy of Sciences*, 103(47):17638–17643, 2006.

[26] Jin Cheng Long, Ai Ai Xia, Jing Han Liu, Ju Li Jing, Ya Zhong Wang, Chuang Ye Qi, and Yan He. Decrease in DNA methylation 1 (DDM1) is required for the formation of mCHH islands in maize. *Journal of Integrative Plant Biology*, 61(6):749–764, 2019.

[27] Karolin Luger, Armin W Mäder, Robin K Richmond, David F Sargent, and Timothy J Richmond. Crystal structure of the nucleosome core particle at 2.8 Å resolution. *Nature*, 389(6648):251–260, 1997.

[28] Wei Feng and Scott D Michaels. Accessing the inaccessible: the organization, transcription, replication, and repair of heterochromatin in plants. *Annual Review of Genetics*, 49(1):439–459, 2015.

[29] Andrew J Bannister and Tony Kouzarides. Regulation of chromatin by histone modifications. *Cell research*, 21(3):381–395, 2011.

[30] Gabriel E Zentner and Steven Henikoff. Regulation of nucleosome dynamics by histone modifications. *Nature structural & molecular biology*, 20(3):259–266, 2013.

[31] Vidya Subramanian, Paul A Fields, and Laurie A Boyer. H2A.Z: a molecular rheostat for transcriptional control. *F1000prime reports*, 7, 2015.

[32] Jordanka Zlatanova and Amit Thakar. H2A.Z: view from the top. *Structure*, 16(2):166–179, 2008.

[33] Toby Buttress, Shengbo He, Liang Wang, Shaoli Zhou, Gerhard Saalbach, Martin Vickers, Guohong Li, Pilong Li, and Xiaoqi Feng. Histone H2B.8 compacts flowering plant sperm through chromatin phase separation. *Nature*, 611(7936):614–622, 2022.

[34] Robert M. Erdmann and Colette L. Picard. RNA-directed DNA methylation. *PLOS Genetics*, 16(10):e1009034, 2020.

[35] M. A. Matzke and R. A. Mosher. RNA-directed DNA methylation: an epigenetic pathway of increasing complexity (vol 15, 394, 2014). *Nature Reviews Genetics*, 15(8):1, 2014.

[36] Samir Kaul, Hean L Koo, Jennifer Jenkins, Michael Rizzo, Timothy Rooney, Luke J Tallon, Tamara Feldblyum, William Nierman, Maria Ines Benito, and Xiaoying Lin. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *nature*, 408(6814):796–815, 2000.

[37] H. Stroud, T. Do, J. Du, X. Zhong, S. Feng, L. Johnson, D. J. Patel, and S. E. Jacobsen. Non-CG methylation patterns shape the epigenetic landscape in *Arabidopsis*. *Nat. Struct. Mol. Biol.*, 21(1):64–72, 2014.

[38] Hume Stroud, Maxim V. C. Greenberg, Suhua Feng, Yana V. Bernatavichute, and Steven E. Jacobsen. Comprehensive analysis of silencing mutants reveals complex regulation of the *Arabidopsis* methylome. *Cell*, 152(1-2):352–364, 2013.

[39] J. Du, X. Zhong, Y. V. Bernatavichute, H. Stroud, S. Feng, E. Caro, A. A. Vashisht, J. Terragni, H. G. Chin, A. Tu, J. Hetzel, J. A. Wohlschlegel, S. Pradhan, D. J. Patel, and S. E. Jacobsen. Dual binding of chromomethylase domains to H3K9me2-containing nucleosomes directs DNA methylation in plants. *Cell*, 151(1):167–80, 2012.

[40] J. A. Law, J. Du, C. J. Hale, S. Feng, K. Krajewski, A. M. Palanca, B. D. Strahl, D. J. Patel, and S. E. Jacobsen. Polymerase IV occupancy at RNA-directed DNA methylation sites requires SHH1. *Nature*, 498(7454):385–9, 2013.

[41] Jixian Zhai, Sylvain Bischof, Haifeng Wang, Suhua Feng, Tzuu-fen Lee, Chong Teng, Xinyuan Chen, Soo Young Park, Linshan Liu, and Javier Gallego-Bartolome. A one precursor one siRNA model for Pol IV-dependent siRNA biogenesis. *Cell*, 163(2):445–455, 2015.

[42] Marjori A Matzke, Tatsuo Kanno, and Antonius JM Matzke. RNA-directed DNA methylation: the evolution of a complex epigenetic pathway in flowering plants. *Annual review of plant biology*, 66:243–267, 2015.

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

[44] Ming Zhou, Ceyda Coruh, Guanghui Xu, Laura M Martins, Clara Bourbousse, Alice Lambolez, and Julie A Law. The CLASSY family controls tissue-specific DNA methylation patterns in *Arabidopsis*. *Nature Communications*, 13(1):244, 2022.

[45] S. Seisenberger, J. R. Peat, T. A. Hore, F. Santos, W. Dean, and W. Reik. Reprogramming DNA methylation in the mammalian life cycle: building and breaking epigenetic barriers. *Philos. Trans. R. Soc. Lond., B, Biol. Sci.*, 368(1609):20110330, 2013.

[46] Christina Ernst, Duncan T. Odom, and Claudia Kutter. The emergence of piRNAs against transposon invasion to preserve mammalian genome integrity. *Nature Communications*, 8(1):1411, 2017.

[47] Katalin Fejes Tóth, Dubravka Pezic, Evelyn Stuwe, and Alexandre Webster. The piRNA pathway guards the germline genome against transposable elements. *Advances in experimental medicine and biology*, 886:51–77, 2016.

[48] Eva-Maria Weick and Eric A. Miska. piRNAs: from biogenesis to function. *Development*, 141(18):3458–3471, 2014.

[49] X. Q. Feng, D. Zilberman, and H. Dickinson. A conversation across generations: Soma-germ cell crosstalk in plants. *Developmental Cell*, 24(3):215–225, 2013.

[50] James Walker, Hongbo Gao, Jingyi Zhang, Billy Aldridge, Martin Vickers, James D. Higgins, and Xiaoqi Feng. Sexual-lineage-specific DNA methylation regulates meiosis in *Arabidopsis*. *Nat. Genet.*, 50(1):130–137, 2018.

[51] C. Oliver and G. Martinez. Accumulation dynamics of ARGONAUTE proteins during meiosis in *Arabidopsis*. *Plant Reprod*, 35(2):153–160, 2022.

[52] Gabriele Bradamante, Vu Hoang Nguyen, Marco Incarbone, Zohar Meir, Heinrich Bente, Mattia Donà, Nicole Lettner, Ortrun Mittelsten Scheid, and Ruben Gutzat. Two ARGONAUTE proteins loaded with transposon-derived small RNAs are associated with the reproductive cell lineage in *Arabidopsis*. *The Plant Cell*, 36(4):863–880, 2024.

[53] Xue Zhou, Kun Huang, Chong Teng, Ahmed Abdelgawad, Mona Batish, Blake C Meyers, and Virginia Walbot. 24-nt phasiRNAs move from tapetal to meiotic cells in maize anthers. *bioRxiv*, pages 2021–08, 2021.

[54] Jixian Zhai, Han Zhang, Siwaret Arikit, Kun Huang, Guo-Ling Nan, Virginia Walbot, and Blake C Meyers. Spatiotemporally dynamic, cell-type–dependent premeiotic and meiotic phasiRNAs in maize anthers. *Proceedings of the National Academy of Sciences*, 112(10):3146–3151, 2015.

[55] Namrata Dhaka, Shalini Sharma, Ira Vashisht, Manu Kandpal, Manoj Kumar Sharma, and Rita Sharma. Small RNA profiling from meiotic and post-meiotic anthers reveals prospective miRNA-target modules for engineering male fertility in sorghum. *Genomics*, 112(2):1598–1610, 2020.

[56] Jiyue Huang, Cong Wang, Xiang Li, Xiaolong Fang, Ning Huang, Ying Wang, Hong Ma, Yingxiang Wang, and Gregory P Copenhaver. Conservation and divergence in the meiocyte sRNAomes of *Arabidopsis*, soybean, and cucumber. *Plant physiology*, 182(1):301–317, 2020.

[57] Soumita Das, Chenna Swetha, Kannan Pachamuthu, Ashwin Nair, and PV Shivaprasad. Loss of function of *Oryza sativa* Argonaute 18 induces male sterility and reduction in phased small RNAs. *Plant reproduction*, 33(1):59–73, 2020.

[58] Lili Wang, Kezhi Zheng, Longjun Zeng, Dachao Xu, Tianxin Zhu, Yumeng Yin, Huadong Zhan, Yufeng Wu, and Dong-Lei Yang. Reinforcement of CHH methylation through RNA-directed DNA methylation ensures sexual reproduction in rice. *Plant Physiology*, 188(2):1189–1209, 2022.

[59] Jian-Ping Lian, Yu-Wei Yang, Rui-Rui He, Lu Yang, Yan-Fei Zhou, Meng-Qi Lei, Zhi Zhang, Jia-Hui Huang, Yu Cheng, Yu-Wei Liu, Yu-Chan Zhang, and Yue-Qin Chen. Ubiquitin-dependent Argonauteprotein MEL1 degradation is essential for rice sporogenesis and phasiRNA target regulation. *The Plant Cell*, 33(8):2685–2700, 2021.

[60] J. Long, J. Walker, W. She, B. Aldridge, H. Gao, S. Deans, M. Vickers, and X. Feng. Nurse cell–derived small RNAs define paternal epigenetic inheritance in *Arabidopsis*. *Science*, 373(6550), 2021.

[61] Jessica A Rodrigues, Randy Ruan, Toshiro Nishimura, Manoj K Sharma, Rita Sharma, Pamela C Ronald, Robert L Fischer, and Daniel Zilberman. Imprinted expression of genes and small RNA is associated with localized hypomethylation of the maternal genome in rice endosperm. *Proceedings of the National Academy of Sciences*, 110(19):7934–7939, 2013.

[62] Jeffrey W Grover, Diane Burgess, Timmy Kendall, Abdul Baten, Suresh Pokhrel, Graham J King, Blake C Meyers, Michael Freeling, and Rebecca A Mosher. Abundant expression of maternal siRNAs is a conserved feature of seed development. *Proceedings of the National Academy of Sciences*, 117(26):15305–15315, 2020.

[63] Diane Burgess, Hiu Tung Chow, Jeffrey W Grover, Michael Freeling, and Rebecca A Mosher. Ovule siRNAs methylate protein-coding genes in trans. *The Plant Cell*, 34(10):3647–3664, 2022.

[64] Chenxin Li, Jonathan I Gent, Hengping Xu, Hong Fu, Scott D Russell, and Venkatesan Sundaresan. Resetting of the 24-nt siRNA landscape in rice zygotes. *Genome research*, 32(2):309–323, 2022.

[65] Prasad RV Satyaki and Mary Gehring. RNA Pol IV induces antagonistic parent-of-origin effects on *Arabidopsis* endosperm. *PLoS Biology*, 20(4):e3001602, 2022.

[66] Illumina Inc. Reduce your library prep time with on-bead tagmentation.

[67] Illumina Inc. Ampliseq for illumina – a comprehensive targeted resequencing solution.

[68] Illumina Inc. Uniformity, precision and reliability in library preparation.

[69] Thermo Fisher Scientific Inc. Preparation of DNA sequencing libraries for illumina systems—6 key steps in the workflow.

[70] OPS Diagnostics. Ctab protocol for isolating DNA from plant tissues.

[71] Steven R. Head, H. Kiyomi Komori, Sarah A. LaMere, Thomas Whisenant, Filip Van Nieuwerburgh, Daniel R. Salomon, and Phillip Ordoukhanian. Library construction for next-generation sequencing: Overviews and challenges. *BioTechniques*, 56(2):61–77, 2014.

[72] Thermo Fisher Scientific Inc. Next-generation sequencing illumina workflow–4 key steps.

[73] F. Krueger and S. R. Andrews. Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics*, 27(11):1571–2, 2011.

[74] Sébastien A. Smallwood, Heather J. Lee, Christof Angermueller, Felix Krueger, Heba Saadeh, Julian Peat, Simon R. Andrews, Oliver Stegle, Wolf Reik, and Gavin Kelsey. Single-cell genome-wide bisulfite sequencing for assessing epigenetic heterogeneity. *Nat. Methods*, 11(8):817–820, 2014.

[75] G. Vilkaitis and S. Klimasauskas. Bisulfite sequencing protocol displays both 5-methylcytosine and N4-methylcytosine. *Anal. Biochem.*, 271(1):116–9, 1999.

[76] Kazuo Tanaka and Akimitsu Okamoto. Degradation of DNA by bisulfite treatment. *Bioorganic & medicinal chemistry letters*, 17(7):1912–1915, 2007.

[77] Jun Xiong, Ping Wang, Wen-Xuan Shao, Gaojie Li, Jiang-Hui Ding, Neng-Bin Xie, Min Wang, Qing-Yun Cheng, Conghua Xie, and Yu-Qi Feng. Genome-wide mapping of N4-methylcytosine at single-base resolution by APOBEC3A-mediated deamination sequencing. *Chemical Science*, 13(34):9960–9972, 2022.

[78] Romualdas Vaisvila, VK Chaithanya Ponnaluri, Zhiyi Sun, Bradley W Langhorst, Lana Saleh, Shengxi Guan, Nan Dai, Matthew A Campbell, Brittany S Sexton, Katherine Marks, et al. Enzymatic methyl sequencing detects DNA methylation at single-base resolution from picograms of DNA. *Genome research*, 31(7):1280–1289, 2021.

[79] Michel Henry, Christophe Terzian, Martine Peeters, Simon Wain-Hobson, and Jean-Pierre Vartanian. Evolution of the primate apobec3a cytidine deaminase gene and identification of related coding regions. *PLoS One*, 7(1):e30036, 2012.

[80] Suhua Feng, Steven E. Jacobsen, and Wolf Reik. Epigenetic reprogramming in plant and animal development. *Science*, 330(6004):622–627, 2010.

[81] Christian A. Ibarra, Xiaoqi Feng, Vera K. Schoft, Tzung-Fu Hsieh, Rie Uzawa, Jessica A. Rodrigues, Assaf Zemach, Nina Chumak, Adriana Machlicova, Toshiro Nishimura, Denisse Rojas, Robert L. Fischer, Hisashi Tamaru, and Daniel Zilberman. Active DNA demethylation in plant companion cells reinforces transposon methylation in gametes. *Science*, 337(6100):1360–1364, 2012.

[82] R. Keith Slotkin, Matthew Vaughn, Filipe Borges, Miloš Tanurdžić, Jörg D. Becker, José A. Feijó, and Robert A. Martienssen. Epigenetic reprogramming and small RNA silencing of transposable elements in pollen. *Cell*, 136(3):461–472, 2009.

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

[84] D. Cuerda-Gil and R. K. Slotkin. Non-canonical RNA-directed DNA methylation. *Nat Plants*, 2(11):16163, 2016.

[85] Kuan-Ju Lu, Nien-Chen Huang, Yu-Shan Liu, Chung-An Lu, and Tien-Shin Yu. Long-distance movement of *Arabidopsis* FLOWERING LOCUS T RNA participates in systemic floral regulation. *RNA biology*, 9(5):653–662, 2012.

[86] Julia Kehr and Friedrich Kragler. Long distance RNA movement. *New Phytologist*, 218(1):29–40, 2018.

[87] Roberto Ruiz-Medrano, Beatriz Xoconostle-Cazares, and Friedrich Kragler. The plasmodesmatal transport pathway for homeotic proteins, silencing signals and viruses. *Current opinion in plant biology*, 7(6):641–650, 2004.

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

[89] M. Zhou, A. M. S. Palanca, and J. A. Law. Locus-specific control of the de novo DNA methylation pathway in *Arabidopsis* by the CLASSY family. *Nature Genetics*, 50(6):865–+, 2018.

[90] Hongxing Yang, Pingli Lu, Yingxiang Wang, and Hong Ma. The transcriptome landscape of *Arabidopsis* male meiocytes from high-throughput sequencing: the complexity and evolution of the meiotic process. *The Plant Journal*, 65(4):503–516, 2011.

[91] T. Hewezi, T. Lane, S. Piya, A. Rambani, J. H. Rice, and M. Staton. Cyst nematode parasitism induces dynamic changes in the root epigenome. *Plant Physiol*, 174(1):405–420, 2017.

[92] Z. Ma, C. Castillo-González, Z. Wang, D. Sun, X. Hu, X. Shen, M. E. Potok, and X. Zhang. *Arabidopsis* serrate coordinates histone methyltransferases ATXR5/6 and RNA processing factor RDR6 to regulate transposon expression. *Dev Cell*, 45(6):769–784.e6, 2018.

[93] Florian Jupe, Angeline C. Rivkin, Todd P. Michael, Mark Zander, S. Timothy Motley, Justin P. Sandoval, R. Keith Slotkin, Huaming Chen, Rosa Castanon, Joseph R. Nery, and Joseph R. Ecker. The complex architecture and epigenomic impact of plant T-DNA insertions. *PLOS Genetics*, 15:1–25, 01 2019.

[94] Arturo Marí-Ordóñez, Antonin Marchais, Mathilde Etcheverry, Antoine Martin, Vincent Colot, and Olivier Voinnet. Reconstructing *de novo* silencing of an active plant retrotransposon. *Nature Genetics*, 45(9):1029–1039, 2013.

[95] Palak Kathiria, Corinne Sidler, Rafal Woycicki, Youli Yao, and Igor Kovalchuk. Effect of external and internal factors on the expression of reporter genes driven by the n resistance gene promoter. *Plant Signaling & Behavior*, 8(7):e24760, 2013.

[96] P. H. Hsieh, S. B. He, T. Buttress, H. B. Gao, M. Couchman, R. L. Fischer, D. Zilberman, and X. Q. Feng. *Arabidopsis* male sexual lineage exhibits more robust maintenance of cg methylation than somatic tissues. *Proceedings of the National Academy of Sciences of the United States of America*, 113(52):15132–15137, 2016.

[97] G. Martinez, K. Panda, C. Kohler, and R. K. Slotkin. Silencing in sperm cells is directed by RNA movement from the surrounding nurse cell. *Nature Plants*, 2(4):8, 2016.

[98] German Martinez, Philip Wolff, Zhenxing Wang, Jordi Moreno-Romero, Juan Santos-González, Lei Liu Conze, Christopher DeFraia, R Keith Slotkin, and Claudia Köhler. Paternal easiRNAs regulate parental genome dosage in *Arabidopsis*. *Nature genetics*, 50(2):193–198, 2018.

[99] Kaushik Panda, Lexiang Ji, Drexel A Neumann, Josquin Daron, Robert J Schmitz, and R Keith Slotkin. Full-length autonomous transposable elements are preferentially targeted by expression-dependent forms of RNA-directed DNA methylation. *Genome biology*, 17:1–19, 2016.

[100] Ian R Henderson, Xiaoyu Zhang, Cheng Lu, Lianna Johnson, Blake C Meyers, Pamela J Green, and Steven E Jacobsen. Dissecting *Arabidopsis thaliana* DICER function in small rna processing, gene silencing and DNA methylation patterning. *Nature genetics*, 38(6):721–725, 2006.

[101] Liang Wu, Huanyu Zhou, Qingqing Zhang, Jianguang Zhang, Fangrui Ni, Chang Liu, and Yijun Qi. DNA methylation mediated by a microRNA pathway. *Molecular cell*, 38(3):465–475, 2010.

[102] Xin-Jian He, Yi-Feng Hsu, Shihua Zhu, Andrzej T Wierzbicki, Olga Pontes, Craig S Pikaard, Hai-Liang Liu, Co-Shine Wang, Hailing Jin, and Jian-Kang Zhu. An effector of RNA-directed DNA methylation in *Arabidopsis* is an ARGONAUTE 4-and RNA-binding protein. *Cell*, 137(3):498–508, 2009.

[103] Natacha Bies-Etheve, Dominique Pontier, Sylvie Lahmy, Claire Picart, Danielle Vega, Richard Cooke, and Thierry Lagrange. RNA-directed DNA methylation requires an AGO4-interacting member of the SPT5 elongation factor family. *EMBO reports*, 10(6):649–654, 2009.

[104] Linfeng Huang, Alexandra ME Jones, Iain Searle, Kanu Patel, Hannes Vogler, Nina C Hubner, and David C Baulcombe. An atypical RNA polymerase involved in RNA silencing shares small subunits with RNA polymerase II. *Nature structural & molecular biology*, 16(1):91–93, 2009.

[105] Markus Schmid, Timothy S Davison, Stefan R Henz, Utz J Pape, Monika Demar, Martin Vingron, Bernhard Schölkopf, Detlef Weigel, and Jan U Lohmann. A gene expression map of *Arabidopsis thaliana* development. *Nature genetics*, 37(5):501–506, 2005.

[106] Shengbo He, Martin Vickers, Jingyi Zhang, and Xiaoqi Feng. Natural depletion of histone H1 in sex cells causes DNA demethylation, heterochromatin decondensation and transposon activation. *eLife*, 8:e42530, may 2019.

[107] Z. Zhong, S. Feng, S. H. Duttke, M. E. Potok, Y. Zhang, J. Gallego-Bartolomé, W. Liu, and S. E. Jacobsen. DNA methylation-linked chromatin accessibility affects genomic architecture in *Arabidopsis*. *Proc Natl Acad Sci U S A*, 118(5), 2021.

[108] Jonathan I Gent, Nathanael A Ellis, Lin Guo, Alex E Harkess, Yingyin Yao, Xiaoyu Zhang, and R Kelly Dawe. CHH islands: *de novo* DNA methylation in near-gene chromatin regulation in maize. *Genome research*, 23(4):628–637, 2013.

[109] Jeongsik Kim, Jin Hee Kim, Eric J Richards, Kyung Min Chung, and Hye Ryun Woo. *Arabidopsis* VIM proteins regulate epigenetic silencing by modulating DNA methylation and histone modification in cooperation with MET1. *Molecular plant*, 7(9):1470–1485, 2014.

[110] Tzung-Fu Hsieh, Juhyun Shin, Rie Uzawa, Pedro Silva, Stephanie Cohen, Matthew J. Bauer, Meryl Hashimoto, Ryan C. Kirkbride, John J. Harada, Daniel Zilberman, and Robert L. Fischer. Regulation of imprinted gene expression in *Arabidopsis* endosperm. *Proceedings of the National Academy of Sciences*, 108(5):1755–1762, 2011.

[111] D. R. Smyth, J. L. Bowman, and E. M. Meyerowitz. Early flower development in *Arabidopsis*. *The Plant cell*, 2(8):755–767, 1990.

[112] Filipe Borges, Rui Gardner, Telma Lopes, Joseph P. Calarco, Leonor C. Boavida, R. Keith Slotkin, Robert A. Martienssen, and Jörg D. Becker. FACS-based purification of *Arabidopsis* microspores, sperm cells and vegetative nuclei. *Plant Methods*, 8(1):44, 2012.

[113] Paul Kron, João Loureiro, Sílvia Castro, and Martin Čertner. Flow cytometric analysis of pollen and spores: An overview of applications and methodology. *Cytometry Part A*, 99(4):348–358, 2021.

[114] Marcel Martin. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. journal*, 17(1):10–12, 2011.

[115] B Langmead. Aligning short sequencing reads with bowtie. *Current Protocols in Bioinformatics*, 2010.

[116] Aaron R Quinlan and Ira M Hall. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6):841–842, 2010.

[117] M. J. Axtell. ShortStack: comprehensive annotation and quantification of small RNA genes. *Rna*, 19(6):740–51, 2013.

[118] R. H. Bloomer, C. E. Hutchison, I. Bäurle, J. Walker, X. Fang, P. Perera, C. N. Velanis, S. Gümüs, C. Spanos, J. Rappsilber, X. Feng, J. Goodrich, and C. Dean. The *Arabidopsis* epigenetic regulator ICU11 as an accessory protein of Polycomb Repressive Complex 2. *Proc Natl Acad Sci U S A*, 117(28):16660–16666, 2020.

[119] Wei-Ying Meng, Zi-Xin Wang, Yunfang Zhang, Yujun Hou, and Jian-Huang Xue. Epigenetic marks or not? the discovery of novel DNA modifications in eukaryotes. *Journal of Biological Chemistry*, page 106791, 2024.

[120] Masaki Shimamura. *Marchantia polymorpha* : Taxonomy, phylogeny and morphology of a model system. *Plant and Cell Physiology*, 57(2):230–256, 2015.

[121] James Walker, Jingyi Zhang, Yalin Liu, Martin Vickers, Liam Dolan, Keiji Nakajima, and Xiaoqi Feng. Extensive N4 cytosine methylation is essential for *Marchantia* sperm function. *bioRxiv*, page 2021.02.12.428880, 2021.

[122] Tetsuya Hisanaga, Shota Fujimoto, Yihui Cui, Katsutoshi Sato, Ryosuke Sano, Shohei Yamaoka, Takayuki Kohchi, Frédéric Berger, and Keiji Nakajima. Deep evolutionary origin of gamete-directed zygote activation by KNOX/BELL transcription factors in green plants. *eLife*, 10:e57090, 2021.

[123] Yan Li, Suresh Kumar, and Weiqiang Qian. Active DNA demethylation: mechanism and role in plant development. *Plant cell reports*, 37:77–85, 2018.

[124] Huiming Zhang, Zhaobo Lang, and Jian-Kang Zhu. Dynamics and function of DNA methylation in plants. *Nature Reviews Molecular Cell Biology*, 19(8):489–506, 2018.

[125] Marc W. Schmid, Alejandro Giraldo-Fonseca, Moritz Rövekamp, Dmitry Smetanin, John L. Bowman, and Ueli Grossniklaus. Extensive epigenetic reprogramming during the life cycle of *Marchantia polymorpha*. *Genome Biol.*, 19(1):9, 2018.

[126] John L. Bowman, Takayuki Kohchi, Katsuyuki T. Yamato, Jerry Jenkins, Shengqiang Shu, Kimitsune Ishizaki, Shohei Yamaoka, Ryuichi Nishihama, Yasukazu Nakamura, Frédéric Berger, Catherine Adam, Shiori Sugamata Aki, Felix Althoff, Takashi Araki, Mario A. Arteaga-Vazquez, Sureshkumar Balasubrmanian, Kerrie Barry, Diane Bauer, Christian R. Boehm, Liam Briginshaw, Juan

Caballero-Perez, Bruno Catarino, Feng Chen, Shota Chiyoda, Mansi Chovatia, Kevin M. Davies, Mihails Delmans, Taku Demura, Tom Dierschke, Liam Dolan, Ana E. Dorantes-Acosta, D. Magnus Eklund, Stevie N. Florent, Eduardo Flores-Sandoval, Asao Fujiyama, Hideya Fukuzawa, Bence Galik, Daniel Grimanelli, Jane Grimwood, Ueli Grossniklaus, Takahiro Hamada, Jim Haseloff, Alexander J. Hetherington, Asuka Higo, Yuki Hirakawa, Hope N. Hundley, Yoko Ikeda, Keisuke Inoue, Shin-ichiro Inoue, Sakiko Ishida, Qidong Jia, Mitsuru Kakita, Takehiko Kanazawa, Yosuke Kawai, Tomokazu Kawashima, Megan Kennedy, Keita Kinose, Toshinori Kinoshita, Yuji Kohara, Eri Koide, Kenji Komatsu, Sarah Kopischke, Minoru Kubo, Junko Kyozuka, Ulf Lagercrantz, Shih-Shun Lin, Erika Lindquist, Anna M. Lipzen, Chia-Wei Lu, Efraín De Luna, Robert A. Martienssen, Naoki Minamino, Masaharu Mizutani, Miya Mizutani, Nobuyoshi Mochizuki, Isabel Monte, Rebecca Mosher, Hideki Nagasaki, Hirofumi Nakagami, Satoshi Naramoto, Kazuhiko Nishitani, Misato Ohtani, Takashi Okamoto, Masaki Okumura, Jeremy Phillips, Bernardo Pollak, Anke Reinders, Moritz Rövekamp, Ryosuke Sano, Shinichiro Sawa, Marc W. Schmid, Makoto Shirakawa, Roberto Solano, Alexander Spunde, Noriyuki Suetsugu, Sumio Sugano, Akifumi Sugiyama, Rui Sun, Yutaka Suzuki, Mizuki Takenaka, et al. Insights into land plant evolution garnered from the *Marchantia polymorpha* genome. *Cell*, 171(2):287–304.e15, 2017.

[127] M. H. Frank and M. J. Scanlon. Transcriptomic evidence for the evolution of shoot meristem function in sporophyte-dominant land plants through concerted selection of ancestral gametophytic and sporophytic genetic programs. *Mol Biol Evol*, 32(2):355–67, 2015.

[128] Sean Akira Montgomery, Tetsuya Hisanaga, Nan Wang, Elin Axelsson, Svetlana Akimcheva, Milos Sramek, Chang Liu, and Frédéric Berger. Polycomb-mediated repression of paternal chromosomes maintains haploid dosage in diploid embryos of *Marchantia*. *eLife*, 11:e79258, 2022.

[129] A. Inoue, L. Jiang, F. Lu, T. Suzuki, and Y. Zhang. Maternal H3K27me3 controls DNA methylation-independent imprinting. *Nature*, 547(7664):419–424, 2017.

[130] Integrated DNA Technologies. xgen methyl-seq dna library kit. https://www.idtdna.com/pages/products/next-generation-sequencing/workflow/xgen-ngs-library-preparation/methyl-seq-dna-library-kit#references, 2023.

[131] James Walker. *DNA methylation reprogramming during sexual reproduction in land plants*. PhD thesis, University of East Anglia, 2019.

[132] Masayuki Tsuzuki, Ryuichi Nishihama, Kimitsune Ishizaki, Yukio Kurihara, Minami Matsui, John L. Bowman, Takayuki Kohchi, Takahiro Hamada, and Yuichiro Watanabe. Profiling and characterization of small RNAs in the liverwort, *Marchantia polymorpha*, belonging to the first diverged land plants. *Plant and Cell Physiology*, 57(2):359–372, 11 2015.

[133] Yuki Sakamoto, Anna Ishimoto, Yuuki Sakai, Moeko Sato, Ryuichi Nishihama, Konami Abe, Yoshitake Sano, Teiichi Furuichi, Hiroyuki Tsuji, Takayuki Kohchi, et al. Improved clearing method contributes to deep imaging of plant organs. *Communications Biology*, 5(1):12, 2022.

[134] Elias J Durand. The development of the sexual organs and sporogonium of *Marchantia polymorpha*. *Bulletin of the Torrey Botanical Club*, 35(7):321–335, 1908.

[135] Yang Yu, Sihan Wang, Ziqin Wang, Renwei Gao, and Joohyun Lee. *Arabidopsis thaliana*: a powerful model organism to explore histone modifications and their upstream regulations. *Epigenetics*, 18(1):2211362, 2023.

[136] Ulises Torres-Flores and Abrahan Hernández-Hernández. The interplay between replacement and retention of histones in the sperm genome. *Frontiers in genetics*, 11:780, 2020.

[137] Xiaoji Wu and Yi Zhang. Tet-mediated active DNA demethylation: mechanism, function and beyond. *Nature Reviews Genetics*, 18(9):517–534, 2017.

[138] Kenji Ichiyama, Tingting Chen, Xiaohu Wang, Xiaowei Yan, Byung-Seok Kim, Shinya Tanaka, Delphine Ndiaye-Lobry, Yuhua Deng, Yanli Zou, Pan Zheng, et al. The methylcytosine dioxygenase Tet2 promotes DNA demethylation and activation of cytokine gene expression in t cells. *Immunity*, 42(4):613–626, 2015.

[139] Shota Chiyoda, Kimitsune Ishizaki, Hideo Kataoka, Katsuyuki T. Yamato, and Takayuki Kohchi. Direct transformation of the liverwort *Marchantia polymorpha* l. by particle bombardment using immature thalli developing from spores. *Plant Cell Rep.*, 27(9):1467, 2008.

[140] K. Ishizaki, R. Nishihama, M. Ueda, K. Inoue, S. Ishida, Y. Nishimura, T. Shikanai, and T. Kohchi. Development of gateway binary vector series with four different selection markers for the liverwort *Marchantia polymorpha*. *PLoS One*, 10(9):e0138876, 2015.

[141] A. Kubota, K. Ishizaki, M. Hosaka, and T. Kohchi. Efficient agrobacterium-mediated transformation of the liverwort *Marchantia polymorpha* using regenerating thalli. *Biosci Biotechnol Biochem*, 77(1):167–72, 2013.

[142] Shota Chiyoda, Kimitsune Ishizaki, Hideo Kataoka, Katsuyuki T. Yamato, and Takayuki Kohchi. Direct transformation of the liverwort *Marchantia polymorpha L.* by particle bombardment using immature thalli developing from spores. *Plant Cell Rep.*, 27(9):1467, 2008.

[143] Yuki Sakamoto, Anna Ishimoto, Yuuki Sakai, Moeko Sato, Ryuichi Nishihama, Konami Abe, Yoshitake Sano, Teiichi Furuichi, Hiroyuki Tsuji, Takayuki Kohchi, and Sachihiro Matsunaga. Improved clearing method contributes to deep imaging of plant organs. *Communications Biology*, 5(1):12, 2022.

[144] Johannes Schindelin, Ignacio Arganda-Carreras, Erwin Frise, Verena Kaynig, Mark Longair, Tobias Pietzsch, Stephan Preibisch, Curtis Rueden, Stephan Saalfeld, Benjamin Schmid, et al. Fiji: an open-source platform for biological-image analysis. *Nature Methods*, 9:676–682, 2012.

[145] Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, Richard Durbin, and Genome Project Data Processing Subgroup. The sequence alignment/map format and SAMtools. *bioinformatics*, 25(16):2078–2079, 2009.

[146] I Soifer. Picard GitHub, 2022.

[147] Miyuki Iwasaki, Tomoaki Kajiwara, Yukiko Yasui, Yoshihiro Yoshitake, Motoki Miyazaki, Shogo Kawamura, Noriyuki Suetsugu, Ryuichi Nishihama, Shohei Yamaoka, and Dierk Wanke. Identification of the sex-determining factor in the liverwort *Marchantia polymorpha* reveals unique evolution of sex chromosomes in a haploid system. *Current Biology*, 31(24):5522–5532. e7, 2021.

[148] Heng Li and Richard Durbin. Fast and accurate short read alignment with burrows–wheeler transform. *Bioinformatics*, 25(14):1754–1760, 2009.

[149] Aaron McKenna, Matthew Hanna, Eric Banks, Andrey Sivachenko, Kristian Cibulskis, Andrew Kernytsky, Kiran Garimella, David Altshuler, Stacey Gabriel, and Mark Daly. The genome analysis toolkit: a mapreduce framework for analyzing next-generation DNA sequencing data. *Genome research*, 20(9):1297–1303, 2010.

[150] Felix Krueger and Simon R Andrews. SNPsplit: Allele-specific splitting of alignments between genomes with known snp genotypes. *F1000Research*, 5, 2016.

[151] Felix Krueger. Trim galore!, 2012. Babraham Bioinformatics.

[152] Zeyang Ma, Claudia Castillo-González, Zhiye Wang, Di Sun, Xiaomei Hu, Xuefeng Shen, Magdalena E Potok, and Xiuren Zhang. *Arabidopsis* serrate coordinates histone methyltransferases atxr5/6 and rna processing factor rdr6 to regulate transposon expression. *Developmental cell*, 45(6):769–784, 2018.

[153] Tarek Hewezi, Thomas Lane, Sarbottam Piya, Aditi Rambani, J Hollis Rice, and Meg Staton. Cyst nematode parasitism induces dynamic changes in the root epigenome. *Plant Physiology*, 174(1):405–420, 2017.

[154] Hongxing Yang, Pingli Lu, Yingxiang Wang, and Hong Ma. The transcriptome landscape of *Arabidopsis* male meiocytes from high-throughput sequencing: the complexity and evolution of the meiotic process. *The Plant Journal*, 65(4):503–516, 2011.

[155] R Keith Slotkin, Matthew Vaughn, Filipe Borges, Miloš Tanurdžić, Jörg D Becker, José A Feijó, and Robert A Martienssen. Epigenetic reprogramming and small RNA silencing of transposable elements in pollen. *Cell*, 136(3):461–472, 2009.

[156] German Martinez, Sarah G Choudury, and R Keith Slotkin. tRNA-derived small RNAs target transposable element transcripts. *Nucleic acids research*, 45(9):5142–5152, 2017.

[157] Filipe Borges, Jean-Sébastien Parent, Frédéric Van Ex, Philip Wolff, German Martínez, Claudia Köhler, and Robert A Martienssen. Transposon-derived small RNAs triggered by miR845 mediate genome dosage response in *Arabidopsis*. *Nature genetics*, 50(2):186–192, 2018.

[158] Joseph P Calarco, Filipe Borges, Mark TA Donoghue, Frédéric Van Ex, Pauline E Jullien, Telma Lopes, Rui Gardner, Frédéric Berger, José A Feijó, Jörg D Becker, et al. Reprogramming of DNA methylation in pollen guides epigenetic inheritance via small RNA. *Cell*, 151(1):194–205, 2012.

[159] Assaf Zemach, M Yvonne Kim, Ping-Hung Hsieh, Devin Coleman-Derr, Leor Eshed-Williams, Ka Thao, Stacey L Harmer, and Daniel Zilberman. The *Arabidopsis* nucleosome remodeler DDM1 allows DNA methyltransferases to access H1-containing heterochromatin. *Cell*, 153(1):193–205, 2013.

[160] Masayuki Tsuzuki, Ryuichi Nishihama, Kimitsune Ishizaki, Yukio Kurihara, Minami Matsui, John L Bowman, Takayuki Kohchi, Takahiro Hamada, and Yuichiro Watanabe. Profiling and characterization of small RNAs in the liverwort, *Marchantia polymorpha*, belonging to the first diverged land plants. *Plant and Cell Physiology*, 57(2):359–372, 2016.