Evolution and control of the DUO1 regulatory network in land plants

Thesis submitted for the degree of
Doctor of Philosophy
at the University of Leicester

by

Siti Nur Aishah Mohd Kamal

Department of Genetics and Genome Biology
University of Leicester

March, 2021

Evolution and control of the DUO1 regulatory network in land plants

Abstract

The sexual reproduction of land plants depends on the successful production and union of male and female gametes. The production of functional sperm cells involves the control of germ cell division and differentiation. These factors are dependent upon a key regulatory module formed by the MYB transcription factor DUO1 POLLEN 1 (DUO1), and DUO1-ACTIVATED ZINC FINGER 1 / DUO1-ACTIVATED ZINC FINGER 2 (DAZ1/DAZ2), C_2H_2 zinc finger proteins. The major aim of this thesis is to further understand the DUO1 network in land plants. The work is aligned as two separate strands, one in angiosperms and the other in bryophytes. *DAZ3* and its homolog, *DAZ3L*, are direct target genes of DUO1. The evolution and potential function of *DAZ3* and *DAZ3L* in Arabidopsis sperm cell development was investigated. Phylogenetic analysis showed that DAZ3 and DAZ3L are restricted to the eudicots superrosid-superasterid clade and most likely evolved from ancestral DAZ1 sequences. The analysis of Arabidopsis knockout mutants showed that neither *DAZ3* nor *DAZ3L* have essential functions in sperm cell development, fertilisation or seed development. The function of DUO1 in the model bryophyte *Physcomitrella patens* (moss) was also investigated. The absence of DUO1 in moss was shown to prevent the differentiation of spermatogenous cells. This resulted in a lack of flagella, rendering spermatogenous cells immotile, hence knockout mutants lacking DUO1 function were infertile. The potential conservation of the DUO1-DAZ1 network in moss was also analysed based on *in silico* analysis of expression data and the presence of DUO1 binding sites (DBS). The similarity in transcript profiles between *PpDUO1* and *PpDAZ1* in maturing antheridia and the presence of DUO1 binding sites in the *PpDAZ1* upstream region indicate that *PpDAZ1* is a direct target of PpDUO1. Hence, it is plausible that the wider DUO1-DAZ1 network is conserved in bryophytes.

ACKNOWLEDGEMENT

In the name of God, the most gracious the most merciful, I hereby would like to express my appreciation to all that have helped me and gave me advices throughout my PhD journey.

First and foremost, I especially would like to thank my supervisor, Professor David Twell for giving me the chance to carry out my PhD studies under his supervision. His advice and encouragement had given me motivation during the course of my project and the completion of this thesis. He had always been accessible and patiently willing to help all his students in their research. At the same level, to the love of my life, Shahidur Ashraf Saidon, for all the support and understanding while I was doing my research here in the UK. Many thanks to my sponsor, MARA Malaysia, for making my ambitions of pursuing research until PhD come true.

I would also like to thank Dr. Dieter Hackenberg and Dr. Celia Hansen for their supervision and guidance as well. Not to forget, the EM facility team Nat and Annia for their help with the EM work. With their help, I have learned a lot of new scientific concepts and lab skills regardless of having busy schedules themselves. They have a lot of experience in handling various laboratory equipment and knowledgeable enough to guide us in operating advanced machine that I have never used before. Not to forget all Lab 311 members that have made my university life here in Leicester a wonderful memory amidst the global COVID-19 pandemic.

Last but not least, I would like to thank my family, especially my parents who have always lent a listening ear, as well as provided me with additional financial support. Their help has made me easier to complete this project.

The Road Not Taken

"I shall be telling this with a sigh…Somewhere ages and ages hence:…Two roads diverged in a wood, and I—…I took the one less traveled by,…And that has made all the difference."

Poem by Robert Frost

List of abbreviation

°C	degrees Celsius
µg	microgram
µl	microlitre
µM	micromolar
µm	micrometre
a.a.	amino acid
A/Ala	alanine
ANOVA	analysis of variance
ATH1	GeneChip Arabidopsis ATH1 Genome Array (Affymetrix®)
BCP	bicellular pollen
bp	base pair
C / Cys	cysteine
cDNA	complementary DNA
CDS	coding sequence
cm	centimetre
Col-0	Columbia-0
D / Asp	aspartic acid
DAPI	4',6'-diamidino-2-phenylindole
DAT	DUO1-activated target
DAZ1	DUO1-ACTIVATED ZINC FINGER 1
DAZ2	DUO1-ACTIVATED ZINC FINGER 2
DAZ3	DUO1-ACTIVATED ZINC FINGER 3
DAZ3L	DAZ3-LIKE
dH$_2$O	distilled water

DNA	deoxyribonucleic acid
dNTP	deoxyribonucleotide triphosphate
DUO1	DUO POLLEN 1
E / Glu	glutamic acid
EAR motif	ERF-associated amphiphilic repression motif
EDTA	ethylenediaminetetraacetic acid
F / Phe	phenylalanine
fmol	femtamole
g	gram
g	gravity
G / Gly	glycine
GCS1	GENERATIVE CELL-SPECIFIC 1
gDNA	genomic DNA
GEX2	GAMETE-EXPRESSED PROTEIN 2
GFP	green fluorescent protein
GUS	β-glucuronidase
H / His	histidine
H2B	HISTONE H2B
HAP2	HAPLESS 2
HTR10	HISTONE3-RELATED 10
I / Ile	isoleucine
K / Lys	lysine
kb	kilobase
L / Leu	leucine
l	litre
LB	Luria Bertani broth

Ler-0	Landsberg erecta-0
M / Met	methionine
mCherry	monomeric cherry
MES	2-(N-Morpholino)ethanesulfonic acid
MGH3	MALE-GAMETE-SPECIFIC HISTONE 3
ml	millilitre
mM	millimolar
MPG	mature pollen grain
MS	Murashige-Skoog
N / Asn	asparagine
ng	nanogram
P / Pro	proline
PCR	polymerase chain reaction
PMI	pollen mitosis I
PMII	pollen mitosis II
pro	promoter
Q / Gln	glutamine
qRT-PCR	quantitative reverse transcription PCR
R/Arg	arginine
RNA	ribonucleic acid
RNAi	RNA interference
RNA-seq	RNA sequencing
rpm	revolutions per minute
RSAT	regulatory sequence analysis tools
RT-PCR	reverse transcription PCR
S / Ser	serine

SD	standard deviation
SpCs	spermatogenous cells
T / Thr	threonine
T1, T2, T3	first, second, third generation of transformed Arabidopsis
TAE	tris acetate EDTA
TCP	tricellular pollen
T-DNA	transfer DNA
Tm	melting temperature
Tris	tris(hydroxymethyl)aminomethane
Triton X-100	t-Octylphenoxypolyethoxyethanol
u	enzyme units
UNM	uninucleate microspore
UTR	untranslated region
V / Val	valine
v/v	volume per volume
W / Trp	tryptophan
WT	wild type
Y / Tyr	tyrosine
ZFP	zinc finger proteins
X^2	Chi-square statistic

Species abbreviations: abbreviated to either 2-letter initials or 3-letter including the second letter of the species name

At / *Ath*	*Arabidopsis thaliana*
Pp	*Physcomitrella patens*

Table of Contents

List of Figures

List of Tables

Chapter 1: Literature review

1.1 Evolution of plant sexual reproduction

Reproduction is important for generating new individuals and to ensure continuity of generations. Plant reproduction can take two forms, either asexual or sexual. In asexual reproduction, the offspring are genetically identical to the parent. In sexual reproduction, the progeny are new individuals that are not genetically identical to parents leading to genetic variation in the populations.

Land plants evolved from charophycean algae approximately 450 million years ago (Bowman et al., 2017). This process involved physiological adaptation to the new environment, while maintaining numerous ancestral developmental, biochemical and cell biological features (Delwiche and Cooper, 2015; Bowman et al., 2017). The movement of plants from fresh water to a terrestrial environment has impacted the life cycle and form of sexual reproduction. The oogamy type of sex observed in land plants has evolved from isogamy which is considered pleiomorphic (Mori et al., 2015). In oogamy, sexual reproduction involves the union of small motile or immotile sperm with a large immotile egg. Meanwhile isogamy involves the union of gametes of identical size and motility. In this context the evolution of plant sexual reproduction has long been an interesting biological topic to uncover.

1.2 The plant life cycle – an alternation of generations

The life cycle of land plants involves two distinct phases, sporophytic and gametophytic, through a process called alternation of generations (reviewed in Brownfield and Twell, 2016; Qiu et al., 2012; Niklas and Kutschera, 2010). Both stages produce cells that are involved in reproduction, spores and gametes, respectively. Although all plants undergo the alternate life stages, the dominant phase varies among different plant groups. In bryophytes, which are non-vascular land plants, such as liverworts, hornworts and mosses, the gametophyte as the dominant phase, which the diploid sporophyte depends upon. However, for vascular plants or tracheophytes such as seedless ferns and seeded angiosperms, the sporophyte is the dominant generation.

In general, the sporophyte stage is a multicellular diploid (2n) that starts from the zygote and proceeds to maturity. When conditions are appropriate, the sporophyte progresses through a reproductive phase and meiosis, which produces microspores and megaspores in structures called sporangia. In flowering plants, the micro- and megasporangia are located in the stamen and ovary, respectively (Figure 1.2). For bryophyte and seedless vascular plants, like ferns, this process occurs in specialised

structures called antheridia (male) and archegonia (female). The microspores and megaspores then give rise to male and female gametophytes through mitotic divisions. The gametophytes are multicellular haploid (n) structures and produce gametes in the form of sperm and/or eggs. The female egg cell is nonmotile such that the sperm cells need to be transported through or released to the environment in order to reach the female tissues. The sperm and egg fuse together upon fertilisation resulting in the formation of the diploid zygote. The growth of the new sporophytic generation starts with the mitotic division of the zygote to produce an embryo, subsequently leading to a mature sporophyte.
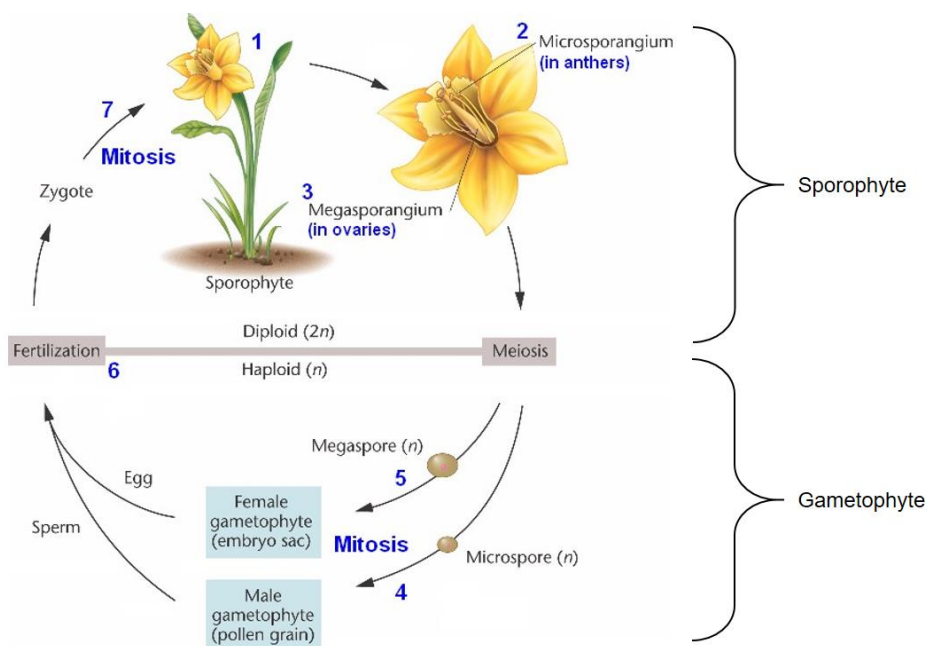
Figure 1.2. Overview of alternation of generations in flowering plants. The sporophytic diploid phase (2n) is the dominant stage, whereas the gametophytic haploid phase (1n) is short-lived. Sporophyte and gametophyte stages produce spores and gametes, respectively. Image adapted from bio3400.nicerweb.com/Locked/media/ch02/alternation_of_generations.html.

1.2.1 The bryophyte life cycle

Bryophytes show variation in the architecture of their vegetative (gametophyte) and reproductive (sporophyte) bodies (Vanderpoorten and Goffinet, 2009). Their vegetative body (gametophyte) are either dorsoventrally flattened thalli, such as in hornworts and liverworts, or branching stem-like tissues (caulids) with leaf-like structures (phyllids), like

in moss (Budke et al., 2018). Meanwhile, the sporophyte is unbranched and attached to the gametophyte. When the conditions in the environment are suitable, the spore germinates and develops into a filamentous network, the protonema. This marks the start of the multicellular stage of development in bryophytes. The protonema later develops and branches to form several upright gametophores. Bryophytes do not form roots, instead they possess a structure that is more similar to root hairs called filamentous rhizoids (Reski, 1998; Schaefer and Zrÿd, 2001; Menand et al., 2007).

Physcomitrium (*Physcomitrella*) *patens* is a well characterised model moss species which is the model adopted in the second and third part of this study (Rensing et al., 2020). *P. patens* is a monoecious moss where both sex organs are present on the same plant (Figure 1.2.1). When germinated, a haploid spore produces a polarised tip that continues to grow to form the protonema (Wu et al., 2018). The protonema is composed of a filamentous network of chloronemal cell that contain abundant chloroplasts, and fast growing caulonemal cell with fewer chloroplasts. Buds containing a single-celled apical meristem develop from caulonemal cell and grow into gametophores (Reski, 1998; Lang et al., 2018). The reproductive (gametangia) organs, antheridia and archegonia, are produced at the apex of the gametophores. When water is available, sperm cells are released from antheridia and swim into the open tip of archegonia ultimately to fertilise the egg. The sperm can either swim towards the archegonia on the same gametophore or to neighbouring gametophores. After fertilisation, the zygote develops into a diploid sporophyte. Meiosis soon take place producing spores.
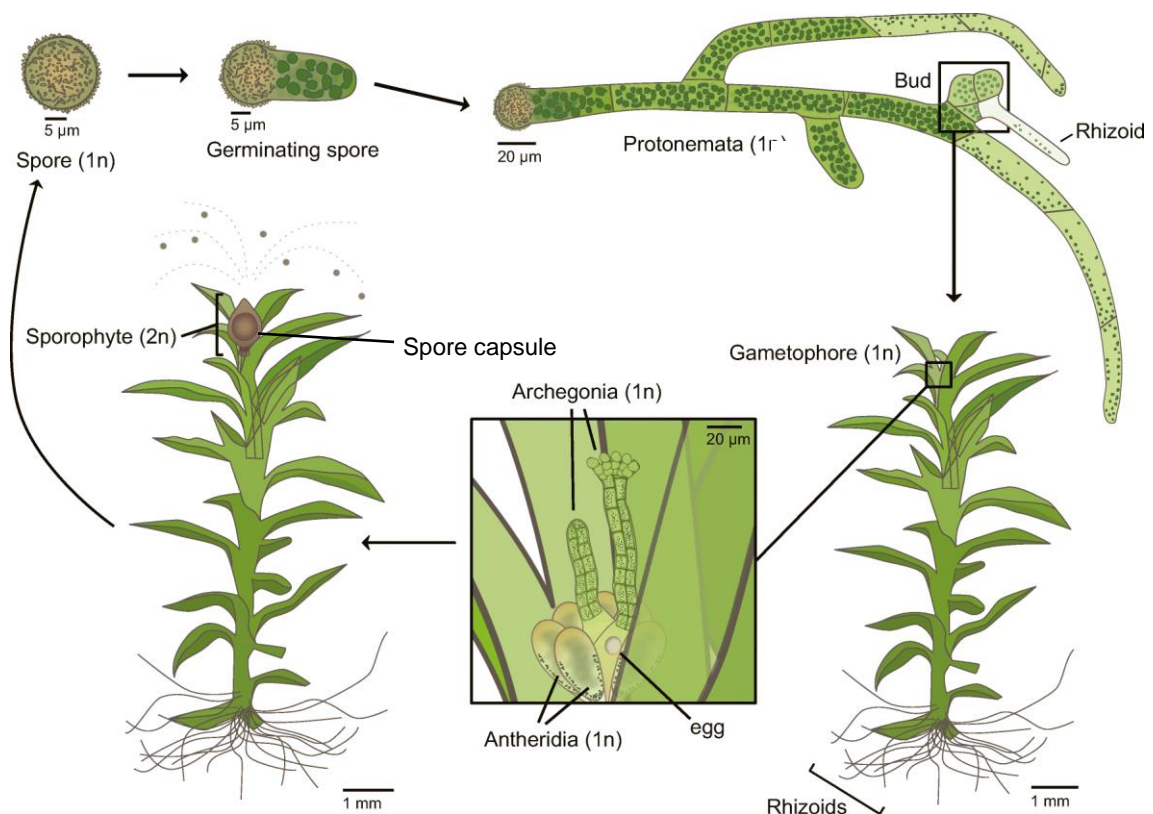
Figure 1.2.1. Life cycle of *Physcomitrella patens*. The gametophyte stage (1n) represents the majority of the life cycle of *P. patens*. The germinating spore develops to form the protonema followed by the gametophore on which the reproductive organs are formed. The sporophyte develops after fertilisation and produces haploid spores in the spore capsule. Image adapted from Wu et al. (2018).

1.3 Male gametogenesis in embryophytes

1.3.1 Angiosperms

Male gametophyte development takes place in the anther, a specialised structure forming part of the male reproductive organ of a flower (stamen) (reviewed in Brownfield and Twell, 2016; Hafidh et al., 2016; Twell, 2011). It comprises two consecutive stages, microsporogenesis and microgametogenesis, where the end products are unicellular microspores and male gametes within the mature male gametophyte respectively. The first stage is microsporogenesis (Figure 1.3.1). In this stage, diploid pollen mother cells (PMC) or microsporocytes, which differentiate from sporogenous cells, will undergo two meiotic divisions that produce four haploid microspores forming a tetrad. Following second meiotic division is the synthesis of cell wall consisting of callose between the individual microspores. An enzyme mixture secreted by tapetal cells, callase, degrades

5

the callose wall surrounding the tetrad (Lu et al., 2014). Unicellular haploid microspores are released from tetrads in a synchronized manner, thus marking the end of microsporogenesis.

The second stage (microgametogenesis) begins when free microspores increase in size and become polarized, due to the merging of small vacuoles into a single large vacuole which is associated with the movement of the nucleus to the cell periphery (Figure 1.3.1.1). Asymmetric division during pollen mitosis I (PMI) follows, resulting in two unequally sized daughter cells with distinct cell fates, a large vegetative cell and a small generative cell (or male germ cell). Then, engulfment of generative cell leads to its migration into the vegetative cell, creating a cell within a cell structure known as bicellular pollen (BCP). For tricellular pollen species such as *Arabidopsis thaliana*, the generative cells undergo another division (pollen mitosis II (PMII) before pollen is shed from the anther, to produce two sperm cells within tricellular pollen (TCP). For most species, generative cell division or PMII takes place within the pollen tube after pollination. Upon successful pollination, the vegetative cell forms an elongated pollen tube which facilitates the delivery of the sperm cells to the female gametophyte, or embryo sac, for a process called double fertilization.

Similar to pollen development, embryo sac development starts with the meiotic division of megaspore mother cell to produce four haploid megaspores by megasporogenesis. In polygonum type megagametogenesis, three megaspores undergo programmed cell death, leaving one surviving spore known as the functional megaspore (Yadegari and Drews, 2004). The functional megaspore undergoes three rounds of mitosis without cell division. The first round produces two nuclei, one of which migrates to the micropylar end while the other moves to the other opposite end, the chalazal end. Second and third rounds then follow producing three antipodal cells at the chalazal end and two synergid cells with egg cell between them at the micropylar end (Figure 1.3.1.2). One nucleus from each end of the embryo sac migrates to the centre to form the polar nuclei, which then fuse and becomes cellularises forming a diploid central cell with a single nucleus. While one sperm cell fertilises the egg cell to produce the embryo, the other sperm cell fuses with the central cell to form the endosperm (Yadegari and Drews, 2004).
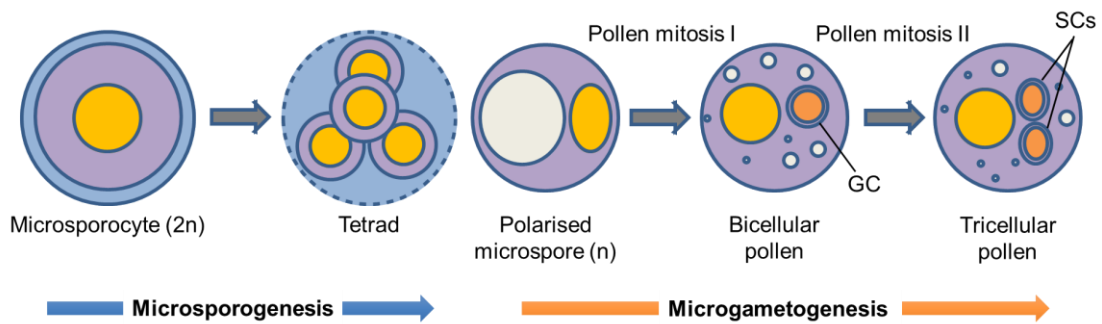
Figure 1.3.1.1. Male gametophyte (pollen) development in a typical tricellular pollen species. Diploid pollen mother cell (microsporocyte) undergoes meiotic division to produce a tetrad of haploid microspores. Microspore undergoes asymmetric division, pollen mitosis I, to produce bicellular pollen and subsequently pollen mitosis II, producing tricellular pollen.
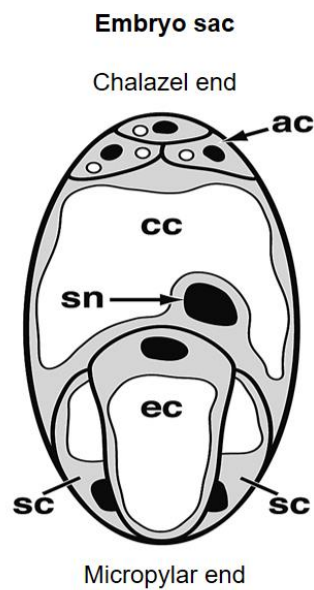
Figure 1.3.1.2. Schematic diagram of an embryo sac. The grey areas represent cytoplasm, the white areas represent vacuoles, and the black areas represent nuclei. Image taken from Yadegari and Drews (2004). ac, antipodal cells; cc, central cell; ec, egg cell; sc, synergid cell; sn, secondary nucleus.

1.3.2 Bryophytes

In bryophytes, male gametogenesis take place in a male gametangium called the antheridium (Figure 1.3.2.). The apical stem cell located at the tip of the gametophore produces the antheridium initial cell that give rise to an antheridium apical cell (Kofuji et al., 2018; Meyberg et al., 2020). The antheridium apical cell then divides to form proximally produced wedge-shaped cells (Kofuji et al., 2018). These cells divide periclinally to form the inner and outer cells. The outer cells form a sterile jacket cell layer which surround the inner primary spermatogenous cells (Renzaglia and Garbary, 2001; Hackenberg and Twell, 2019). Then the spermatogenous cells undergo proliferative cell division followed by cell differentiation to produce functional sperm (Hackenberg and Twell, 2019; Norizuki et al., 2020). In *Marchantia polymorpha*, the final mitotic division of spermatogenous cells is that of spermatid mother cells (SMC). Each SMC divides diagonally to form a pair of triangular sperm cells. The sperm cells differentiate to form motile sperm prior to their release from the antheridium (Hisanaga et al., 2019; Hackenberg and Twell, 2019).

Meanwhile, female gametogenesis occurs in an archegonium. In *P. patens*, archegonial apical stem cells divide to produce a central cell covered with peripheral cells (Kofuji et al., 2018). The central cell later divides to form an upper primary canal cell and a lower ventral cell (reviewed in Kofuji et al., 2018). The primary canal cell and its surrounding cell divide transversely to form neck canal cells and neck cells. Meanwhile, the ventral cell expands without dividing. Subsequently, the ventral cells divide asymmetrically, producing a smaller ventral canal cell and the egg cell (Kofuji et al., 2018). When conditions are suitable for fertilisation, the neck canal cells and the ventral canal cells are degraded to allow the movement of sperms to fertilise the egg.
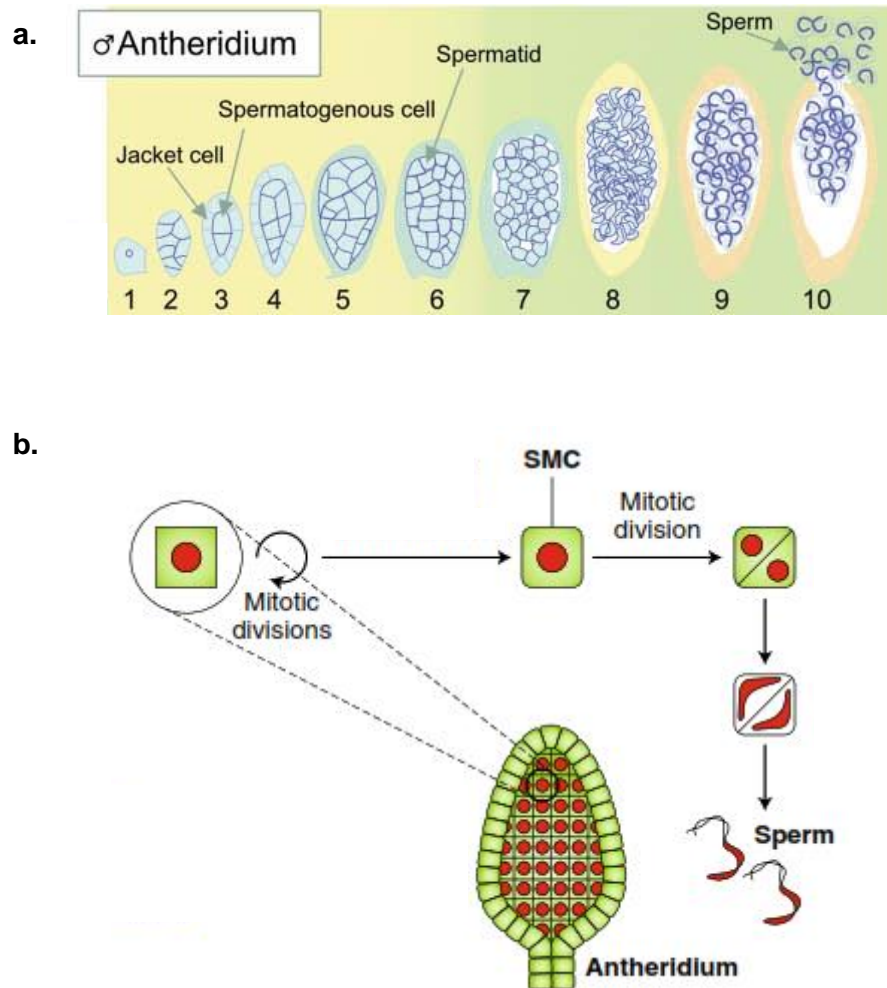
Figure 1.3.2. Male gametogenesis in bryophytes. Schematic representation of male gametogenesis in **(a)** *Physcomitrella patens* and **(b)** *Marchantia polymorpha*. In an antheridium, inner spermatogenous cells are surrounded by a single outer layer of sterile jacket cells. Spermatogenous cells proliferate through cell division that give rise to spermatid which then differentiate to produce functional sperm. In *Marchantia* however, spermatogenous cells undergo a mitotic cell division that subsequently produces spermatid mother cells (SMC). The SMCs divide diagonally and differentiate to produce motile sperm. Image adapted from Sanchez-Vera et al. (2017) and Hisanaga et al. (2019).

1.3.2.1 Moss a model plant

P. patens was first established as an experimental system used in laboratory to study plant development in 1920s by Fritz von Wettstein (1924). Since then, it has been widely used as a model organism to study the plant gene functions due to its exceptionally high rate of homologous recombination that facilitate gene knockout and allele replacement (Schaefer and Zrÿd, 1997; Reski, 1998; Müller et al., 2016). Various ecotypes were used for research study; however, the predominant ecotype is Gransden. It was based on cultures derived from a single spore of a sample collected from Gransden Wood (UK) in 1962 by H.L.K. Whitehouse (Engel, 1968; Rensing et al., 2020; Meyberg et al., 2020). Gransden is mainly propagated and distributed vegetatively across laboratorys worldwide and used as a standard laboratory strain. However, several laboratories have reported fertility issues likely due to periods of vegetative propagation that accumulates somatic mutations (Ashton and Raju, 2000; Perroud et al., 2011; Landberg et al., 2013; Hiss et al., 2017). Landberg et al. (2013) reported that sporophyte production for transgenic lines made using Gransden 2004 laboratory strain was insufficient for analysis. In another study, Gransden antherozoids were unable to fertilize eggs on the gametophores of Villersexel (Perroud et al., 2011). For a hermaphrodite population to acquire a male sterility mutation, the female fertility must be elevated (Charlesworth & Charlesworth, 1978). However, the loss of male fertility in Gransden strain was not compensated by a gain in female fertility when Gransden sporophyte production was not higher than in the Villersexel strain (Perroud et al., 2011).

In order to circumvent this problem, a new ecotype, Reute, has been introduced as Gransden alternative especially for studies involving sexual reproduction (Hiss et al., 2017; Meyberg et al., 2020). The Reute ecotype was collected in 2006 by Lüth from a moist, disturbed field close to Freiburg im Breisgau, Germany (Hiss et al., 2017). Comparative study of sexual reproduction between Reute and Gransden shows a significant difference in the number of their sporophytes (Hiss et al., 2017). The number of sporophytes for ecotype Reute was tremendously higher than in Gransden, signifying Reute was highly self-fertile. The gamete divergent between Reute and Gransden was also low as indicated by the low number of single nucleotide polymorphisms (SNPs) to Gransden (Hiss et al., 2017).

The fertility differences between ecotypes Gransden and Reute was utilised for functional characterisation of genes that affect male sterility in *P. patens*. Crossing analysis between Reute and Gransden revealed that Gransden archegonia were fully functional and can develop significant number of sporophyte, eliminating the assumption that

sterility was through defectivity in female reproductive apparatus (Meyberg et al., 2020). Mature antheridia for both ecotypes were then analysed for their spermatozoid number, which display comparability between them. The flagella of Gransden spermatozoids however, had coiled structure, indicating a connection to coiled-coil domain containing 39 (*CCDC39*) gene (Meyberg et al., 2020). In the same study, CCDC39 was found to be essential for proper flagella development. *ccdc39* mutant in Reute background had coiled flagella in comparison to the wild type Reute. The sporophyte also did not form in *ccdc39* mutant plant. Additionally, the expression analysis of *CCDC39* showed reduced expression level in the antheridia of Gransden compared to Reute (Meyberg et al., 2020). In other study, *HISTONE ACETYLTRANSFERASE 1* (*HAG1*) and *SWITCH/SUCROSE NONFERMENTING 3A/B* (*SWI3A/B*) were found as a part of network that controls the proper coordination of sexual reproduction (Genau et al., 2021). The loss of function for both genes resulted in the male infertility due to non-fertile spermatozoids. Both *hag1* and *swi3a/b* mutants showed a significant reduction in number of sporophytes compared to the wild type Reute (Genau et al., 2021). Double staining with DAPI and NAO revealed that *hag1* spermatozoid nucleus were round in shape and spermatozoids did not release from the antheridia in contrast to *swi3a/b* and wild type Reute. On the contrary, *swi3a/b* spermatozoid nucleus were slender in shape and spermatozoids were able to be released from the antheridia, similar to Reute control (Genau et al., 2021). However, an incomplete cytoplasmic reduction was observed surrounding the nucleus of the released spermatozoid. The *swi3a/b* spermatozoids were also unable to swim.

1.4 Regulation of gene expression in developing male germ cells of angiosperms

A dedicated male germ cell lineage in angiosperms is initiated following asymmetric division, of the microspore. The resulting vegetative and generative cells immediately enter different developmental pathways and have distinct transcriptional profiles. Numerous transcriptome studies have shown that vegetative cell and generative cells express different, but overlapping, sets of genes (reviewed in Rutley and Twell, 2015). The first sperm cell transcriptomic study was conducted in Arabidopsis using ATH1 Genome Array where 5829 and 7177 genes were detected in sperm cells and pollen, respectively (Borges et al., 2008). Given the advantage of RNA-seq, that enables more comprehensive transcriptome analysis compared with microarrays, 16985 and 18911 genes were reliably detected in sperm cells and vegetative cells of rice (Anderson et al., 2013). A common feature observed in Arabidopsis and rice is the reduced transcriptome complexity of sperm cells compared to pollen and the distinct transcriptomes of the

sperm cell and vegetative cell. Transcriptome analysis of male gametophyte development was first described in Honys and Twell (2004) where 13997 genes showed signal in at least one out of four stages, microspores (UNM), BCP, TCP and mature pollen (MP), of male gametophyte development. Co-expression clusters were also found in UNM-BCP and in TCP-MP with a sharp transcript reduction after BCP, leading to phase shift in gene expression between BCP and TCP. Pollen-enriched genes are involved in membrane transport, signalling and vesicle trafficking, whereas sperm cell-enriched transcripts are involved in DNA replication and repair, ubiquitin-mediated proteolysis and cell cycle (Borges et al., 2008; Rutley and Twell, 2015). Hence, due to the distinct and diverse transcriptome, transcriptional regulation is likely to be an important aspect of germline development (Borg et al., 2011; Borges et al., 2008; Engel et al., 2003).

Recent studies have established a regulatory framework for male germ cell division and differentiation. The discovery of *DUO POLLEN 1 (DUO1),* has linked male germ cell division and differentiation that leads to sperm cell specification (Rotman et al., 2005; Durbarry et al., 2005; Brownfield et al., 2009). DUO1 is a male germline-specific MYB transcription factor which regulates several germline-specific genes such as *MALE GAMETE-SPECIFIC HISTONE H3* (*MGH3*), *GAMETE-EXPRESSED 2* (*GEX2*) and *GENERATIVE CELL SPECIFIC 1* (*GSC1/HAP2*) (Brownfield et al., 2009; Mori et al., 2006; von Besser et al., 2006). Mutations in *DUO1* lead to cell a division defect, whereby pollen remains bicellular at anthesis, as well as an impairment of fertilisation (Durbarry et al., 2005; Rotman et al., 2005).

An analysis of seeds that was conducted to investigate genetic transmission of *DUO1* shows that *duo1* mutation exhibit strict male gametophytic control (Rotman et al., 2005). Crosses between *duo1* pollen with wild type plants showed *duo1* mutant lines were unable to fertilise wild type ovules, unlike pollination of *duo1* ovules with wild type pollen. *duo1* mutant pollen also shows single larger generative cell nucleus compared to two sperm cells nuclei of wild type pollen (Rotman et al., 2005). The generative cell identity of bicellular *duo1* pollen was confirmed through the expression of promAKV-H2B::YFP, a cell-identity reporter that was similarly expressed in sperm cells of mature pollen. Interestingly, the failure of *duo1*'s generative cell to undergo mitotic division does not prevent its entry into the S phase (Rotman et al., 2005). The nuclear DNA content increased to more than 2C, comparable to the sperm cell during the tricellular stages. The expression of the cell cycle regulatory protein *CYCLIN B1;1* (*CYCB1;1*) in *duo1* mutants was able to partially rescue the cell division defect, but rescued pollen was unable to undergo fertilisation (Brownfield et al., 2009). This is likely to be due to

incomplete germ cell differentiation, demonstrated by the lack of *GEX2* and *GCSI/HAP2* expression, which are important for gamete attachment and fusion during fertilisation (Mori et al., 2006; von Besser et al., 2006). Hence, the absence of *GEX2* and *GCSI/HAP2* expression is a key feature which explains the infertility of *duo1* mutant pollen (Brownfield et al., 2009).

The ectopic expression of *DUO1* in seedlings upregulates 61 candidate genes in addition to *MGH3*, *GEX2* and *GCS1/HAP2* (Borg et al., 2011). DUO1 was shown to transactivate the promoters of seven of these candidate genes in transient expression assays in tobacco leaves which were termed DUO1 activated target (DAT) genes. A further seven DAT gene promoters were shown to be DUO1-dependent in heterozygous *duo1* transgenic plants (Borg et al., 2011). Some DAT genes such as *DUO1-ACTIVATED ATPASE1* (*DAA*), *PLANT CADMIUM RESISTANCE 11* (*PCR11*) and *DUO1-ACTIVATED ZINC FINGER 3* (*DAZ3*) were revealed to be activated only late in pollen development based on their expression profiles. Interestingly, *DAA1*, *PCR11* and *DAZ3* transcripts were detected only after germ cell division and their promoters direct sperm-cell specific reporter activity (Borg et al., 2011; Borges et al., 2008; Taimur, 2014). DAT genes represent various gene families including transcription factors, where three of them, *DUO1-ACTIVATED ZINC FINGER* 1 (*DAZ1*), *DUO1-ACTIVATED ZINC FINGER* 2 (*DAZ2*) and *DAZ3* belong to the C_2H_2-type zinc finger family. Unlike *DAZ3*, *DAZ1* and *DAZ2* are early activated DAT genes in generative cells and they are found to be required for generative cell division and sperm cell differentiation (Borg et al., 2014).

1.5 DUO1 – a key player in male germ cell development

1.5.1 MYB transcription factors in plants

The MYB gene family is a super family of transcription factors present in all eukaryotes and massively expanded in plants (Riechmann et al., 2000; Dubos et al., 2010). The first MYB gene, *v-myb*, was found in avian myeloblastosis virus (AMV) that can cause myeloblastosis (myeloid leukemia) in chicken (Klempnauer et al., 1982; Klempnauer et al., 1984). In plants, the first MYB gene discovered was *COLORED1* (*C1*) that is required for anthocyanin synthesis in *Zea mays* (Paz-Ares et al., 1987). A MYB protein is characterised by having a highly conserved MYB domain which consists of up to four imperfect repeats (Ogata et al., 1996). Each repeat is about 52 amino acids in length and forms three alpha helices with a helix-helix-turn-helix structure (Figure 1.5.1). The second and third helices contain three conserved tryptophans, spaced 18 or 19 amino acids apart, that form a hydrophobic core (Ogata et al., 1992). Based on the number of

imperfect repats in a given protein, the MYB family has been classified into four groups namely 1R-MYB, 2R-MYB, 3R-MYB and 4R-MYB that contain 1, 2, 3 and 4 MYB repeats, respectively (Dubos et al., 2010).
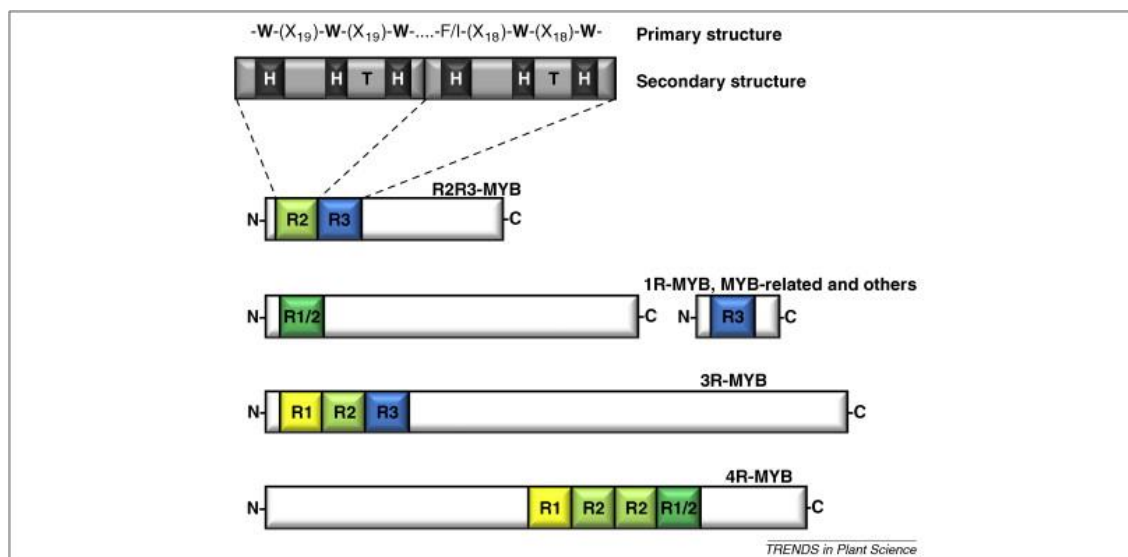
Figure 1.5.1. General overview of MYB protein in plants. Illustration from Dubos et al. (2010) showing the four groups of MYB family, 1R-, 2R-, 3R- and 4R-MYB with their respective repeats. Each repeat has a helix-helix-turn-helix structure and within the second and third helices, three regularly spaced tryptophan residues are present. H, helix; T, turn; W, tryptophan; X, amino acid (X).

Studies of MYB genes in plants have shown that the majority of MYB proteins belong to the R2R3-type MYB family (Figure 1.5.1). The first comprehensive description and classification of plant MYB genes was made possible owing to the Arabidopsis genome sequencing (Stracke et al., 2001; Dubos et al., 2010). Since then, the identification of R2R3 MYB genes has been tremendously expended in various plant species covering wide array of plant evolutionary divergent such as banana, soybean, orchids, tomato, rice, fern and moss (Katiyar et al., 2012; Du et al., 2012; Du et al., 2015; Li et al., 2016; He et al., 2019; Pucker et al., 2020; Hernández-Hernández et al., 2021). Members of this group have a very diverse functions encompassing plant defence, cell fate and identity, developmental processes and plant metabolism (Rotman et al., 2005; Borg et al., 2011; Lau et al., 2015; Schwinn et al., 2016; Ullah et al., 2020). An Arabidopsis MYB gene of particular interest is *AtDUO1*, encoding a novel R2R3 MYB protein, that was the first gene shown to control male gamete development in plants (Rotman et al., 2005).

1.5.2 Conservation of *DUO1* in flowering plants

Although DUO1 belongs to the large R2R3 MYB family, it was not included in the pioneer studies of R2R3 MYB analysis in plants (Martin and Paz-Ares, 1997; Stracke et al., 2001). However, it was not until the 2005 that the DUO1 gene was discovered as part of *R2R3 MYB* transcription factors by Rotman et al. (2005). DUO1's phylogenetic position in the MYB family was solidified in a review by Dubos et al. (2010). Sequence analysis of DUO1 orthologs in tobacco, rice and maize, reveals a unique feature exclusive to this R2R3 MYB subfamily (Figure 1.5.2). Detailed evaluation of the R2R3 MYB domain of DUO1 sequences revealed the presence of a uniquely conserved signature supernumerary lysine residue (K^{66} in AtDUO1) within the R3 repeat (Rotman et al., 2005). The DUO1 R2R3 MYB subfamily was later expended with the inclusion on DUO1 orthologs from castor bean, poplar, tobacco, rice, lily and other eudicot species (Brownfield et al., 2009; Peters et al., 2017).
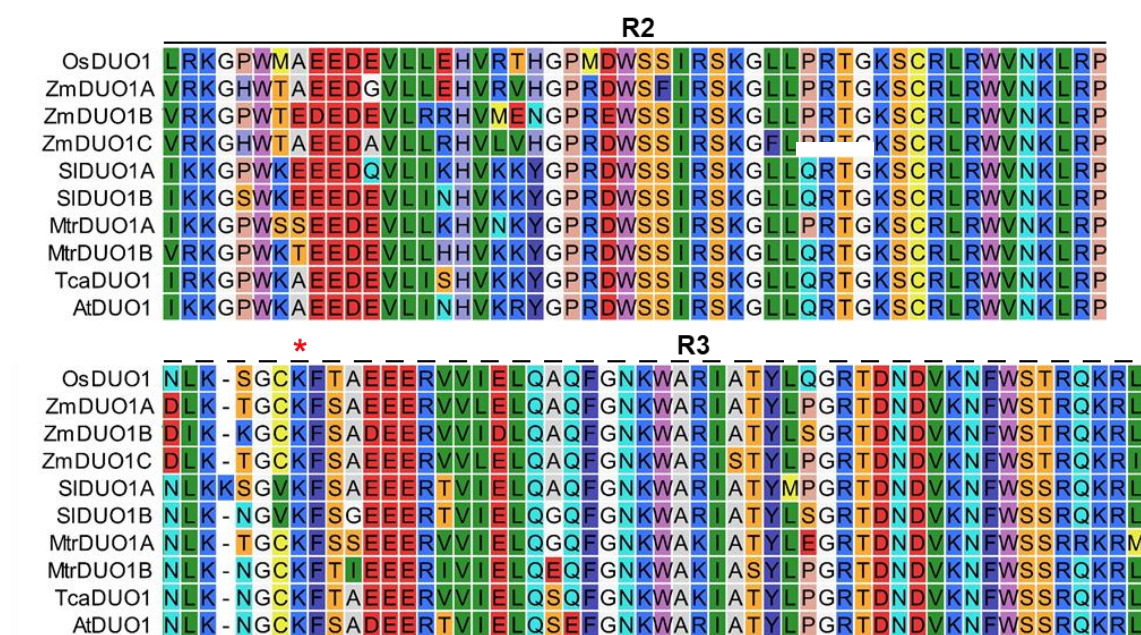
Figure 1.5.2. R2R3 MYB domain of DUO1 orthologs. The red asterisks indicate the position of K^{66} lysine residues unique to DUO1 R2R3 MYB subfamily.

The conservation of DUO1 in angiosperms is not limited to protein sequences but also includes expression pattern. *DUO1* has male germline-restricted expression in Arabidopsis and a similar pattern is also observed in other flowering plants. RT-PCR analysis showed that *DUO1* transcripts were only detected in the inflorescence of wild

type Arabidopsis and not in vegetative tissues (Rotman et al., 2005). The *DUO1* promoter fused to HISTONE2B::mRFP1 showed pollen-specific reporter protein expression in generative cells and in sperm cells. Furthermore, tobacco and maize *DUO1* transcripts were expressed in pollen, similar to what had been observed in Arabidopsis (Rotman et al., 2005). In tomato, *SlDUO1* expression was detected in anther, but absent from leaf, sepal and petal (Sari, 2015). qPCR analysis of rice *DUO1 (OsDUO1)* showed transcription in pollen and not seedling (Russell et al., 2012). *OsDUO1* expression was also enriched in sperm indicating selective transcription in the male germline of rice. In terms of protein expression, a DUO1:mRFP protein fusion construct driven by the Arabidopsis *DUO1* promoter, was expressed in the nucleus of sperm cells (Rotman et al., 2005; Brownfield et al., 2009).

The regulation of DUO1 is also conserved among eudicots at least in *A. thaliana* and *Medicago truncatula* (a legume). This was demonstrated when *Regulatory region of DUO1* (*ROD1*) from *M. truncatula* was shown to direct male germline-specific expression in Arabidopsis (Peters et al., 2017). *ROD1* was first characterised in Arabidopsis and is a cis-regulatory module that controlled the expression of DUO1. A similar regulatory module was also found in the *DUO1* promoter of soybean, grapes, tomato, rice, maize and banana which could potentially function as *ROD1* in these species (Peters et al., 2017).

The tomato and rice DUO1 orthologs have also been shown to transactivate the Arabidopsis *MGH3* promoter, that is a native DUO1 target (Sari 2015). In addition, when SlDUO1 and OsDUO1 were heterologously expressed using the AtDUO1 promoter, they were able to complement the generative cell division and genetic transmission defects in *duo1* mutants, illustrating the conservation of DUO1 protein function among flowering plants (Sari, 2015).

1.5.3 Conservation of *DUO1* in extant representatives of bryophytes

Based on the presence of the DUO1 R2R3 MYB signature sequence, the subfamily was widened by the addition of orthologs from spikemoss (*Selaginella moellendorffii*), moss (*P. patens*) and liverwort (*M. polymorpha*) (Brownfield et al., 2009; Higo et al., 2018). RNA-seq data for major developmental tissues of *M. polymorpha* showed that *MpDUO1* RNA is expressed in antheridiophores and antheridia, with enriched expression in the later (Higo et al., 2016; Higo et al., 2018). *MpDUO1* was shown to be expressed in the sperm cell lineage (in spermatid mother cells and spermatids) by in situ hybridisation (Higo et al., 2018). A similar general expression profile was also recorded for *HmnDUO1*

of basal liverwort (*Haplomitrium mnioides*) where the expression was detected in antheridiophores, but absent from vegetative gametophyte tissues (Higo et al., 2018).

The activity of the *DUO1* promoter in the male germline was shown to be conserved between Arabidopsis and *M. polymorpha*. The promoter of *MpDUO1* fused to H2B-Clover was shown to confer sperm cell-specific expression in mature pollen of Arabidopsis (Higo et al., 2018). In the same study, the *pro*At*DUO1:GUS* was shown to be expressed in spermatogenous cell tissue within antheridia, similar to the activity of the *MpDUO1* promoter construct *proMpDUO1:GUS*. With regard to conservation of *DUO1* protein expression, the MpDUO1-Citrine fusion protein was shown to localize to the nucleus of developing sperm (Higo et al., 2018).

The *DUO1-DAZ1* regulatory module was shown to be conserved in *Marchantia*. *MpDAZ1* expression was significantly reduced in *Mpduo1-1^{ko}* showing that *MpDUO1* controls the expression of *MpDAZ1* (Higo et al., 2018). However, the expression of *MpGEX1* and *MpGCS1* (orthologs of AtDUO1 target genes) were unaffected in *Mpduo1-1^{ko}*, which illustrates that the DUO1 target gene network is not fully conserved between *Marchantia* and Arabidopsis.

1.6 The DUO1-DAZ1 gene regulatory network

1.6.1 *DAZ* genes – direct target of DUO1

DAZ1, *DAZ2*, and *DAZ3* are zinc finger proteins that were discovered as DAT genes (Borg et al., 2011). The *DAZ1-*, *DAZ2-* and *DAZ3-* promoters each driving the H2B-GFP reporter showed sperm cell-specific expression in mature pollen. The dependency of the activity of these promoters on DUO1 was demonstrated by their reduced expression in *duo1* mutant pollen. In addition, the ability of DUO1 to transactivate the *DAZ1*, *DAZ2* and *DAZ3* promoters was also shown in dual luciferase assays in leaves of the heterologous host, tobacco.

DAZ1 and *DAZ2* are required for germ cell division and differentiation (Borg et al., 2014). Their protein expression was first detected in germ cell nucleus of bicellular pollen following asymmetric cell division and persists until mature pollen. RT-PCR analysis was used to validate the *DAZ1* and *DAZ2* transcripts where their expression was detected only in pollen and absent in other sporophytic tissues (Borg et al., 2014). In the same study, *DAZ1* and *DAZ2* are also found to be the direct targets of DUO1. This is evident by the decrease in *DAZ1* and *DAZ2* promoter activity in DUO1-dependent transactivation

assays when DUO1 binding site within *DAZ1* and *DAZ2* promoter region was mutated (Borg et al., 2014)

The *daz1 daz2* double knockout mutant was reported to have a single germ cell phenotype resembling that of *duo1* mutant (Figure 1.6.1.1) (Borg et al., 2014). This failed germ cell division phenotype was complimented with the expression of DAZ1 in *daz1 daz2* mutant carrying ProDAZ1:DAZ1-mCherry transgene. This demonstrates redundancy of *DAZ1* and *DAZ2* as well as highlighting their function in germ cell division. The localization of *DAZ1* protein orthologs are explored in a study by Darbar (2019). DAZ1 of rice, maize and tomato were reported to have nuclear enriched expression in wild type pollen similar to Arabidopsis DAZ1 (Darbar, 2019). In contrary, *DAZ1* of *Brassica rapa* had a nuclear-specific expression, like Arabidopsis DAZ2.
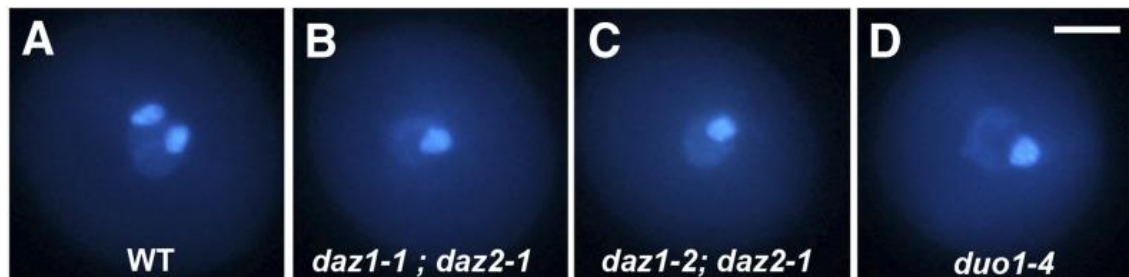
Figure 1.6.1.1. Phenotype of *daz1 daz2* double mutant. (A) Wild type (tricellular) and **(B-D)** mutant (bicellular) pollen of *daz1 daz2* and *duo1* stained with DAPI. Both *daz1 daz2* alleles have a phenotype similar to *duo1*. Scale bar = 10 μm. Image taken from Borg et al. (2014).

A homolog of *DAZ3*, *DUO1-ACTIVATED ZINC FINGER 3 LIKE* (*DAZ3L*), was found through in-silico analysis by Taimur (2014). *DAZ3* was preferentially expressed in Arabidopsis sperm cells and reported to have the highest expression level among the highly expressed genes (Borges et al., 2008). Meanwhile, transcriptome profiling using the AGRONOMICS1 platform detected *DAZ3L* expression in the flower tissue of Arabidopsis (Rehrauer et al., 2010). Further analysis has showed that both *DAZ3* and *DAZ3L* were expressed late in germline development and have sperm cell-specific expression (Borg et al., 2011; Taimur, 2014). Similar to *DAZ3*, the activity of the *DAZ3L* promoter was DUO1-dependent as shown in the dual luciferase assays conducted in tobacco leaves (Taimur, 2014). In addition, the expression of *DAZ3* and *DAZ3L* promoter

fused to H2B-GFP was also absent in *duo1* mutant pollen, illustrating their dependency on DUO1.

A study by Taimur (2014) also explored the localisation of *DAZ3* and *DAZ3L* proteins using ProDAZ3:DAZ3-mCherry and ProDAZ3L:DAZ3L-mCherry protein fusion constructs in sperm cells. DAZ3 was shown to localise to the cytoplasm based on the mCherry fluorescence signal, whereas DAZ3L localised to both nucleus and cytoplasm with a preference for the later (Figure 1.6.1.2). The protein localisation pattern was also maintained in sperm cells in pollen tubes grown in vitro.
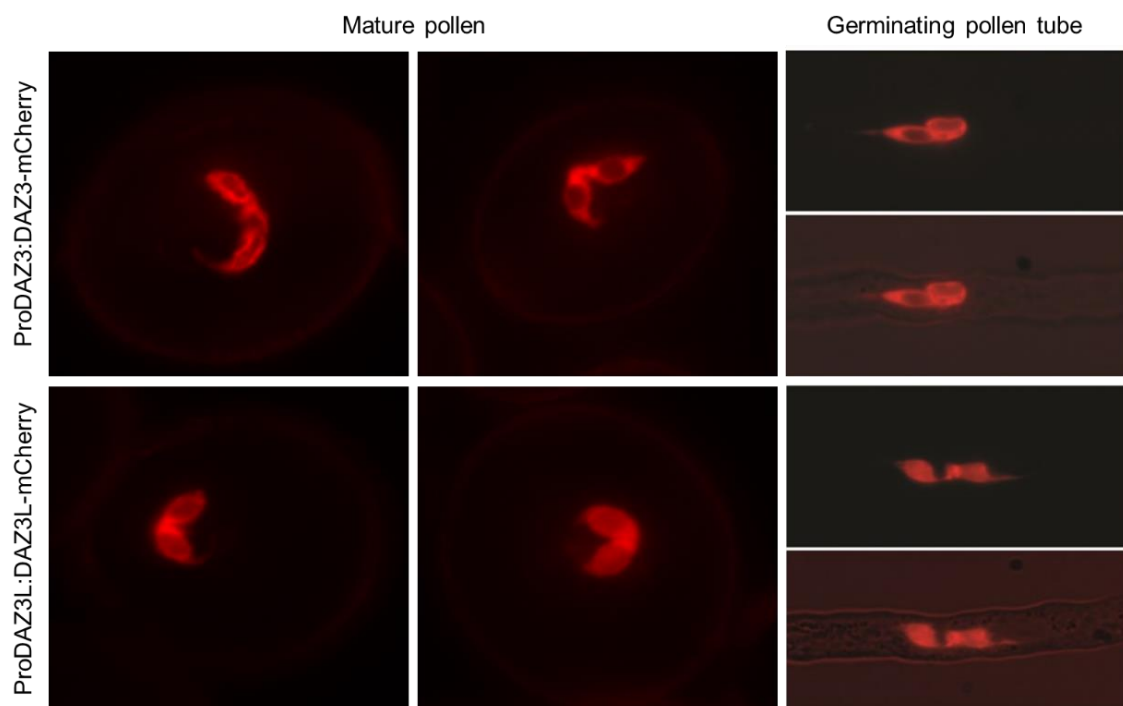
Figure 1.6.1.2. Protein localisation of *DAZ3* and *DAZ3L*. The mCherry fluorescence was detected in cytoplasm for DAZ3 and in both nucleus and cytoplasm for DAZ3L. Similar protein localisation pattern persist in developing pollen tube. Image adapted from Taimur (2014).

1.6.2 C_2H_2 zinc finger transcription factors

The zinc finger super family can be divided into several subfamilies according to the order and number of cysteine (Cys) and histidine (His) residues in their secondary structures (Ciftci-Yilmaz and Mittler, 2008; Klug, 2010). This Cys and His residues bind a zinc ion in a fashion originally described as resembling a finger (Miller et al., 1985).

Among examples of the subfamily are C_2H_2, C_3H, C_4 and C_2C_2 (Ciftci-Yilmaz and Mittler, 2008). Like DAZ1, DAZ2 and DAZ3, DAZ3L also belongs to the C_2H_2-type zinc finger family. The C_2H_2 zinc finger protein is a large family of transcription factors and accounts for 0.7 % of the predicted proteins in *Arabidopsis thaliana* (Englbrecht et al., 2004).

According to the classification of *A. thaliana* zinc fingers by Englbrecht et al. (2004), zinc finger proteins can be further divided into three sets namely set A, B and C. Set A and B are described as having tandem zinc finger in one or more than one array, respectively. Meanwhile set C comprise of proteins with single or dispersed zinc finger. Tandem arrangement is defined by having 0 to 10 amino acids in between the fingers whereas more than that, ≥ 11, is considered dispersed (Figure 1.6.2). The set A and B account for 20 % while set C consists of 80 % of the zinc finger proteins in *A. thaliana*. The sets are further divided into several subset according to the number of spacing between the zinc coordinating histidine residues. For example, subset 1 (C1) and 2 (C2) of set C have three (HX3H) and four (HX4H) amino acid residues in between the His, respectively.
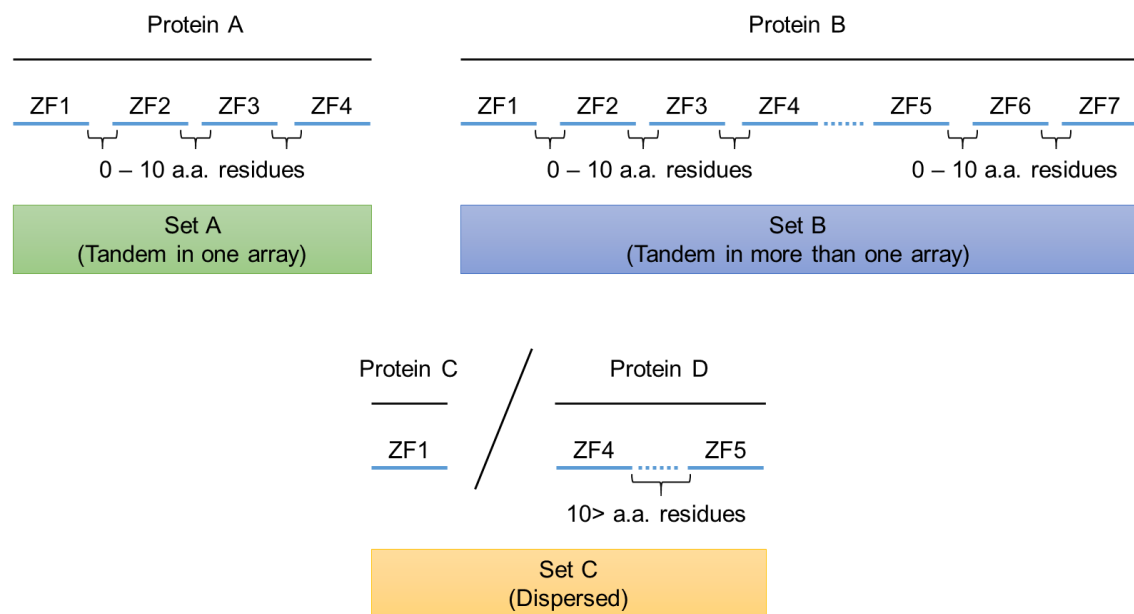
Figure 1.6.2. Classification of zinc finger proteins. In Englbrecht et al. (2004), zinc finger proteins are categorised into several sets. Zinc finger proteins that have tandem zinc fingers, separated by 0 to 10 amino acids, in one array or more are classified into set A and set B, respectively. Those with one or more zinc finger separated by more than 10 amino acids have a dispersed arrangement (set C). For example, protein A in the figure with four tandem zinc fingers in one array is grouped into set A.

C_2H_2 zinc finger protein consists of two antiparallel β-sheets and one α-helix structures (Omichinski et al., 1990). Zn^{2+} ion binds to conserved cys and his residues, located at the end of β-sheets and c-terminus of α-helix, respectively, to form a tetrahedral structure; hence stabilising the protein folding (Omichinski et al., 1990; Klug and Schwabe, 1995; Stubbs et al., 2011). The DNA binding motif of C_2H_2 zinc finger is also located in the α-helix structure. Most common C_2H_2 zinc finger motif in plants is the highly conserved sequence QALGGH, also called the Q-type motif (Englbrecht et al., 2004). This Q-type binding motif is found for example in SUPERMAN (SUP) of *A. thaliana* and ZPT2-1 and ZPT2-2 (previously EPF1 and EPF2) of petunia (Takatsuji and Matsumoto, 1996; Isernia et al., 2003; Han et al., 2020). Another type of zinc finger motif is the K-type which replaces the glutamine (Q) of QALGGH with lysine (K) or arginine (R) (Englbrecht et al., 2004).

The C_2H_2 zinc finger proteins not only can bind to DNA but also RNA and protein (Han et al., 2020). They are involved in many biological processes such as plant development, stress response and hormone signalling. One example of C_2H_2 zinc finger involvement in plant development is displayed by *DAZ1* and *DAZ2* which has been discussed previously. Another example is demonstrated by *HAIR* (*H*) that regulates the formation of multicellular trichomes in tomato (Chang et al., 2018). In their study, deletion of the entire coding region of *H* resulted in an absence of trichome phenotype in tomato. H was also shown to interact via pull down and yeast two-hybrid assays with another protein that regulates trichome formation called WOOLLY (WO) (Yang et al., 2011).

1.7 Aims and objectives

The overall aim of the project was to further understand the DUO1 network in land plants. This thesis has two major aims and work was aligned as two separate strands, one in angiosperms and the other in bryophytes.

1) To investigate potential function of *DAZ3 / DAZ3L* in sperm cell development and explore their evolutionary origin.

2) To explore DUO1 function in bryophytes.

The first major aim is discussed in chapter 3. In order to achieve this aim, two objectives were set up.

1) To explore sequence conservation of *DAZ3* in flowering plants.

   Sequence analysis was performed on DAZ3 and its orthologs and the analysis was extended to include DAZ1 and DAZ2 as related ancestral sequences.

2) To investigate the *in-vivo* role of *DAZ3* and *DAZ3L*.

   Knockout mutants were created using the CRISPR-Cas9 technique and the phenotypes of mutant plants were examined by analysis of fertility and seed production.

The second major aim is covered in chapter 4 and 5. To explore DUO1 function in bryophytes, two objectives were established.

1) To conduct a functional analysis of the two *DUO1* orthologs (*PpDUO1A* and *PpDUO1B*) identified in the moss, *Physcomitrella patens*.

   The phenotype of single and double PpDUO1 knockout mutants was observed by examining the sporophyte formation and spermatogenous cell development in the antheridium. In addition, promoter-marker constructs were designed to analyse the spatio-temporal expression of *PpDUO1*.

2) To study the conservation of the DUO1-DAZ1 regulatory module in *P. patens*.

   DAZ1 orthologs were identified in *P. patens* and transcript expression of *PpDAZ1* were analysed in relation to *PpDUO1* transcripts. The promoter region of *PpDAZ1* was also examined for the presence of DUO1 binding sites. Promoter-marker constructs were also created to investigate the *in-vivo* spatio-temporal expression of *PpDAZ1* in wild type and *Ppduo1* mutant plants.

Chapter 2: Material and methods

2.0 Material and methods

2.1 Purchase of materials

Chemicals used throughout this project were purchased from the following suppliers: Bioline, Duchefa Biochemie, Fisher Scientific UK, Melford and Sigma-Aldrich. Nucleic acid purification kits were supplied by Omega Bio-tek and New England Biolabs. Cloning kits, enzymes and other reagents were purchased from Bioline, Thermo Fisher Scientific Invitrogen™ and New England Biolabs. Sanger sequencing was done by GATC Services-Eurofins Genomics.

2.2 Plant materials

Arabidopsis thaliana plants were derived from the Columbia-0 (Col-0) accession. Transgenic mutant lines for single mutant *daz3* and *daz3l* were generated with help from Dr. Dieter Hackenberg who designed and transformed gene-targeted CRISPR constructs into T0 plants and kindly provided T1 seeds. Double mutant *daz3 daz3l* plants were generated by transforming *daz3* and *daz3l* single mutant plants with the paralog CRISPR construct.

All transgenic and wild type *Physcomitrella patens* plants used were generated from the Reute accession 2011 (Rt11) and 2016 (Rt16). *Ppduo1a^A*, *Ppduo1b^A* and *Ppduo1a^{A}b^A* knockout lines (refer to subchapter 4.3.1) were kindly generated and provided by Dr. Yasuko Kamisugi and Andrew Cumming University of Leeds. Wild type plants were also supplied by Cumming's laboratory alongside the knockout lines.

2.1.1 Arabidopsis plant growth conditions and selection

Arabidopsis seeds were sown on soil in a 3:1:1 ratio of compost (Levington), vermiculite and sand, stratified at 4 °C for 2 days and transferred to a growth room with continuous light at 22 °C. For selection of CRISPR transformants, ~100 mg of T1 seeds were plated on MS0 with 25 µg/ml hygromycin following the method described by Harrison et al. (2006). Prior to plating, seeds were surface sterilised with a mixture of 70 % ethanol and 0.01 % triton-X for 5 minutes, followed by a wash in 100 % ethanol for 5 minutes and drying on sterile filter paper for 15 minutes. The resistant seedlings were transplanted to soil.

2.1.2 Moss plant growth conditions

Physcomitrella patens Reute accessions were cultivated on solidified BCD media (BCD medium - 250 mg/l $MgSO_4.7H_2O$, 250 mg/l KH_2PO_4 pH 6.5, 1.01 g/l KNO_3, 12.5 mg/l $FeSO_4.7H_2O$, 1ml/l Trace element solution (see Appendix Table S2.1), 10 mM $CaCl_2$ – added after autoclave) supplied with 5 mM (di)ammonium tartrate for general propagation and on BCD medium with reduced potassium nitrate content (400 mM) for sporulation as described by Cove et al. (2009). All media were supplemented with 0.7% phyto agar. Plants were grown in a growth chamber at 25 °C with a 16 hour light/8 hour dark regime (light intensity 90 µmol/m^2/s). For sporulation, plants were transferred from 25 °C to 16 °C for sporophyte induction with an 8 hour light/16 hour dark regime (light intensity 20 µmol/m^2/s).

2.2.3 Long term storage of moss plant

Sterile 2.0 ml screw cap microtubes (Sarstedt), containing 1 ml to 1.5 ml of solid BCD medium, were prepared. A piece of gametophyte leaf or tissue was placed in the microtube and the screw cap was screwed in but not fully tightened to allow aeration. The culture was grown in a growth chamber for 3 – 4 weeks at 25 °C with a 16/8 h light/dark regime at intensities between 5 and 20 Wm–2. After the plant had grown, the screw cap was tightened and stored in a cold room with a 2 hour light/22 hour dark cycle.

2.2.4 Arabidopsis floral dip transformation

The Arabidopsis floral dip transformation protocol was modified and based on the protocol from Clough and Bent (1998). Plants were grown until bolting (approximately 4 weeks). The emerging bolts close to rosette leaves were clipped off to encourage growth of multiple secondary bolts and left to grow for another 1 week. Two days prior to floral dipping, 5 ml of *Agrobacterium* culture (GV3101) harbouring the appropriate CRISPR construct was set up using LB media and appropriate antibiotics. The culture was grown overnight with shaking (200 rpm) at 28 °C. 400 µl of the overnight culture was then added into 400 ml of fresh LB media supplemented with appropriate antibiotics and incubated for another 24 hours with good shaking (200 rpm) at 28 °C. The cells were centrifuged at 4600 rpm for 20 minutes and resuspended in standard infiltration medium (2.15 g/l MS salts, 3.16 g/l Gamborgs B5 medium, 0.5 g/l MES, 50 g/l sucrose, 10 µg/l 6-benzylaminopurine) supplemented with 400 µl of Silwet L-77. The siliques and fully open flowers were removed from plants before inflorescence were dipped into the medium for

approximately 45 seconds with gentle agitation. After dipping, the plants were placed under a propagator lid for 24 hours to recover and to maintain humidity and then grown until seed set.

2.2.5 Moss protoplast transformation and selection

This protocol was modified and based on the protocol from Cove et al. (2009). There were three stages involved: protoplast extraction, protoplast transformation and transformant selection. All centrifugation steps were carried out at 250 x g, without the use of the centrifuge brake. For protoplast extraction, 7-day old protonema harvested from 2 – 3 petri dishes using a spatula were added to a sterile petri dish containing 10 ml of 8.5 % D-mannitol. 10 ml of 1 % Driselase was added and the petri dish incubated at room temperature with gentle shaking for 1 hour. The protoplast suspension was filtered through a 100 μm pore size nylon mesh followed by 70 μm mesh before transfer to 15 ml centrifuge tubes and centrifuged for 5 minutes. The supernatant was removed and the protoplast pellet resuspended very gently with 10 ml of CaPW (8.5 % D-Mannitol, 10 mM $CaCl_2$). The protoplast solution was again centrifuged for 5 minutes and the supernatant discarded. The protoplast pellet was gently resuspended using 10 ml of 8.5 % D-Mannitol. 10 μl of the protoplast solution was sampled and placed upon a haemocytometer (1/400 mm^2 x 0.1 mm) for protoplast counting. The number of protoplasts in a 16 square area was multiplied by 10,000 to obtain the number of protoplasts per ml. After counting, the protoplast solution was centrifuged for 5 minutes and the supernatant discarded. The protoplasts were resuspended in MMM solution (9.1 % D-Mannitol, 15 mM $MgCl_2$, 0.1 % MES pH 5.6) at a density of 1.67 million protoplasts / ml.

For protoplast transformation, 30 μg of supercoiled plasmid DNA was added to a 15 ml centrifuge tube. 300 μl of protoplasts and 300 μl of PEGT (2g PEG-6000, 4.45 ml 8.5% D-mannitol, 500 μL 1 M $Ca(NO_3)_2$, 50 μL 1 M Tris HCL pH 8) were added to the tube containing the plasmid DNA. The tube was incubated for 5 minutes at room temperature. Then, the tube was heat shocked at 45 °C for 5 minutes before incubation in a water bath at room temperature for 10 min. 300 μL of 8.5 % D-Mannitol was added and the tube swirled; the same step was then repeated 4 times at 2 minute intervals. Next, 1 ml of 8.5 % D-Mannitol was added to the tube and swirled; the step was then repeated 4 times at 5 minutes intervals. The tube was centrifuged for 5 minutes and the supernatant removed. The pellet was resuspended in 1 ml 8.5 % D-Mannitol. 5 ml of molten PRMT media (BCD medium, 8 % D-Mannitol, 5 mM (di)ammonium tartrate, 0.4 % phyto agar,

10 mM CaCl$_2$ – added after autoclave) was added to the tube. 2 ml of the mixture was dispensed onto a plate containing PRMB media (BCD medium, 6 % D-Mannitol, 5 mM (di)ammonium tartrate, 0.7 % phyto agar, 10 mM CaCl$_2$ – added after autoclave) overlaid with cellophane. The plate was placed in a growth chamber at 25 °C with a 16 hour light/8 hour dark regime for 7 days.

For transformant selection, the cellophane was transferred onto a selection media (BCD media supplied with 5 mM (di)ammonium tartrate, 15 µg/ml hygromycin) and grown for 14 days. Then, the cellophane was transferred onto non-selective media (BCD media supplied with 5 mM (di)ammonium tartrate for 7 days for recovery of the transformant. The cellophane was again transferred onto a selection media for second round of selection for 14 days. Stable transformants were grown on non-selective media for 3 – 4 weeks.

2.3 Nucleic acid isolation

2.3.1 Small scale DNA extraction from plant tissue

1 to 2 pieces of leaf tissue were collected into 1.5 ml eppendorf tube containing ~100 glass beads (Sigma-Aldrich). Then, the tube was flash frozen in liquid nitrogen and ground for 20 seconds using Silamat amalgam mixer (Ivoclar Vivadent, UK) at room temperature. 250 µl extraction buffer (1.4 M NaCl, 3 % (w/v) CTAB, 20 mM EDTA, 100 mM Tris-HCl pH8.0) was added to the sample, vortexed briefly and incubated at room temperature for 15 to 20 minutes. In a fume hood, an equal volume (~250 µl) of chloroform:isoamyl alcohol (24:1) was added and mixed well. The sample was centrifuged for 12 minutes at 13,000 rpm and an aqueous layer was transferred to a fresh tube containing 0.7 volumes of isopropanol (~140 µl). Next, the sample was left at room temperature for 5 minutes before being mixed and centrifuged for 7 minutes at 13,000 rpm. The supernatant was removed and 1 ml of 70 % ethanol was added to the pellet followed by centrifugation at 13,000 rpm for 5 minutes. All ethanol was removed and the pellet was left to dry for approximately 30 minutes. The dried pellet was re-suspended in 100 µl of TE Buffer (10 mM Tris-Cl, 1 mM EDTA, pH 8.5), incubated for 5 minutes at 55 °C and stored at -20 °C.

2.3.2 High throughput DNA extraction from plant tissue

This protocol was modified from Edwards et al. (1991), adapted for high-throughput plant DNA extraction. Typically, this method is used to extract a large number of tissue samples in a 96-place rack format, using a TissueLyser II (Qiagen). A piece of leaf tissue was collected into 1.2 ml microtubes (T100 - Biotube™ System, Simport) arranged in a 96-place rack (Simport). Each microtubes contained one 5 mm stainless steel bead (Qiagen). Once all samples were collected and microtubes sealed with a 96 well rubber mat (Cole-Parmer), the tissue was flash frozen by placing the 96-place rack in liquid nitrogen for ~7 seconds. The 96-place rack was placed in the TissueLyser II adaptors and ground by shaking at 20 Hz for 20 seconds. 400 µl extraction buffer (200 mM Tris-HCl, pH7.5, 25 mM EDTA, 250 mM NaCl, 0.5 % SDS) was added using a 200 µl multi-channel pipette. The sample was centrifuged at 3,000 g for 30 minutes at 4 °C. 200 µl supernatant was transferred into a new microtube in a 96-place rack. Then, 200 µl of isopropanol was added to the supernatant and mixed by inverting 3 to 4 times before being centrifuged for 20 minutes at 3,000 g. The supernatant was removed and 200 µl of 70 % ethanol was added to the pellet followed by centrifugation at 3,000 g for 5 minutes at room temperature. The ethanol was poured off and the pellet left to dry in a sterile laminar flow hood. The dried pellet was re-suspended in 60 µl of TE Buffer (10 mM Tris-Cl, 1 mM EDTA, pH 8.5) and used immediately or stored at -20 ºC.

2.3.3 Plasmid DNA extraction from bacteria

Plasmid DNA extraction was carried out using E.Z.N.A.®Plasmid DNA Mini Kit I (Omega Bio-tek) according to manufacturer's spin protocol. All centrifugation steps were carried out at 13,000 x g. Around 1.5 ml – 3 ml of overnight culture was added to a microfuge tube before being spun for 1 minute. Supernatant was discarded and the cells were resuspended in 250 µL of Solution I/RNase A by pipetting up and down to mix thoroughly. Then, 250 µL of Solution II was added to the resuspended cells and the tube was gently inverted until clear lysate was seen. 350 µL of Solution III was added to the mixture and the tube was immediately inverted several times until white precipitate formed. The sample was centrifuged for 10 minutes and while waiting, a binding column was prepared by inserting HiBind® DNA Mini Column into a 2 ml Collection Tube. After centrifugation, the clear lysate was transferred into the column and then centrifuged for 1 minute. The filtrate was discarded and 500 µL of HBC Buffer was added to the column. After centrifugation for 1 minute, the filtrate was discarded and the column was washed with 700 µl DNA Wash Buffer. The sample was centrifuged for 1 minute and a second DNA

Wash Buffer wash step was performed. After the second wash, the empty column was centrifuged for 2 minutes to dry the column matrix. The column was transferred to a clean microfuge tube and 30 µl of Elution Buffer was added directly to the centre of the column membrane. The sample was left for 1 minute at room temperature and centrifuged for 1 minute to elute the DNA. Purified DNA was quantified using a NanoDrop™ UV-Vis 2000 spectrophotometer and stored at -20 °C until further needed.

2.3.4 Purification of PCR products

PCR product was purified using E.Z.N.A.® Cycle Pure Kit (Omega Bio-tek) according to the manufacturer's protocol. All centrifugation steps were carried out at 13,000 x g. The sample was transferred into a clean microfuge tube and 4 – 5 volumes of CP Buffer was added to the tube. The tube was vortexed to mix the solution thoroughly. A binding column was prepared by inserting HiBind® DNA Mini Column into a 2 ml Collection Tube. The sample was added to the HiBind® DNA Mini Column and centrifuged for 1 minute. The filtrate was discarded and 700 µl DNA Wash Buffer was added to the column. The sample was centrifuged for 1 minute and the washing step was repeated once. After the second wash, the empty column was centrifuged for 2 minutes to dry the column matrix. The column was transferred to a clean microfuge tube and 30 µl of Elution Buffer was added directly to the centre of the column membrane. The sample was left for 1 minute at room temperature and centrifuged for 1 minute to elute the DNA. Purified PCR product was quantified using NanoDrop™ UV-Vis 2000 spectrophotometers and stored at -20 °C until further needed.

2.3.5 Agarose gel electrophoresis for nucleic acid separation

Depending on the DNA fragments size, 1 % to 3 % agarose gels were made by adding 1x TAE (40 mM Tris, 20 mM acetic acid, 1 mM EDTA) to the agarose powder (Bioline). The solution was microwaved until the agarose was completely dissolved. Ethidium bromide (0.2 µg/ml final concentration) was added to the agarose solution and mixed well. The agarose solution was poured into a casting tray with the well comb in place and left to solidify. The solidified gel was placed in a gel tank containing 1x TAE buffer and depending on the application, HyperLadder™ 1 kb (Bioline) or 50 bp DNA Ladder (New England Biolabs) was loaded into the well. DNA samples were mixed with loading dye (New England Biolabs) prior loading into the subsequent well. The gel was typically run

at 100 mV until the loading buffer reached 3/4 of the total gel length. The DNA fragments were visualized using UV transilluminator (BioDoc-It™ Imaging System, UVP).

2.3.6 Purification of DNA from agarose gels

DNA fragments were purified from agarose gel pieces using Monarch® DNA Gel Extraction Kit (New England Biolabs) according to the manufacturer's instruction. All centrifugation steps were carried out at 16,000 x g (~13,000 RPM). The DNA fragment of interest was excised from the agarose gel, transferred to a microfuge tube and then weighed. Four volumes of Gel Dissolving Buffer, where 400 µl buffer per 100 mg agarose, was added to the gel slice. The sample was incubated at a temperature between 37 – 55 °C (typically 50 °C) and vortexed periodically until the gel slice was completely dissolved (generally 5 – 10 minutes). Column was inserted into the collection tube and the sample mixture was loaded onto the column. After centrifugation for 1 minute, the flow-through was discarded and the column was washed with 200 µl DNA Wash Buffer. The sample was centrifuged for 1 minute and the washing step was repeated once. After the second wash, the column was transferred to a clean microfuge tube. 30 µl of DNA Elution Buffer was added to the center of the matrix, incubated for 1 minute at room temperature and centrifuged for 1 minute to elute the DNA. Purified DNA was quantified using a NanoDrop™ UV-Vis 2000 spectrophotometers (Thermo Fisher Scientific) and stored at -20 °C.

2.3.7 Quantification of nucleic acids

Genomic and plasmid DNA concentration was measured using a NanoDrop™ UV-Vis 2000 spectrophotometer (Thermo Fisher Scientific) and NanoDrop™ 2000/2000c software. 1 µl of blank solution was pipetted onto the pedestal and the 'Blank' tab on the software was clicked. The pedestal was wiped with microscope lens tissue and 1 µL of sample was loaded onto the pedestal. 'Measure' button on the software was clicked to obtain the sample spectral measurement and the computed concentration of DNA based on the absorbance at 260 nm and the default extinction coefficient.

2.4 Polymerase Chain Reaction (PCR) and its application

2.4.1 Oligonucleotide primer design

Oligonucleotide primers used to amplify genomic DNA and plasmid DNA were designed using the Primer3 design tool (http://primer3.ut.ee/) Primers were designed to be between 18 bp to 30 bp with melting temperature (T_m) of 60 °C – 64 °C. The GC content was set between 35 – 65 %, with the optimum at 50 %. For genotyping *daz3* and *daz3l* mutant plants, the PCR product size was designed to be in the range 150 bp – 200 bp. Primers were purchased from Sigma-Aldrich. Additional primers used throughout the project were kindly provided by Dr. Dieter Hackenberg and Dr. Nadia Taimur, Department of Genetics and Genome Biology, University of Leicester, UK.

2.4.2 Genotyping *daz3* and *daz3l* mutant alleles by PCR

daz3 and *daz3l* CRISPR mutation were genotyped using PCR. PCR reactions were prepared using BioTaq DNA polymerase (Bioline) and relevant buffers supplied together with the enzyme, along with dNTPs (Invitrogen™). Forward and reverse primers from the sequences of *DAZ3* or *DAZ3L* gene were designed for PCR genotyping with primer target sites spanning the CRISPR guides target sites (Figure 3.4.1.5; Figure 3.4.1.6). The PCR reaction and conditions were listed below (Table 2.4.2 a, b). Lastly, the PCR fragments were confirmed by sequencing.

Table 2.4.2 (a) 20 µl PCR reaction for genotyping.

PCR reaction	Volume
10x NH_4 buffer	2.0 µl
50 mM $MgCl_2$	1.0 µl
10 mM dNTPs	0.5 µl
10 µM Forward primer	2.0 µl
10 µM Reverse primer	2.0 µl
Nuclease-free water	11.4 µl
Taq polymerase (5 u/µl)	0.1 µl
DNA template	1.0 µl
Total volume	**20.0 µl**

Table 2.4.2 (b) Standard PCR conditions.

Steps	Temperature (°C)	Time
Initial denaturation	94	2 min
35 cycles	94	30 sec
	57 - 60	30 sec
	72	1 min 30 sec
Final extension	72	5 min

2.4.3 PCR for TOPO® TA Cloning®

Promoter fragments were amplified for use in TOPO® TA Cloning® (Thermo Fisher Scientific Invitrogen™) using two-step PCR condition. Mixed polymerase PCR reactions were prepared using a combination of BioTaq DNA polymerase with relevant buffers (Bioline) and Phusion® High-Fidelity DNA Polymerase (New England Biolabs) along with dNTPs (Invitrogen™). The mixed polymerase was set to 10:3 – BioTaq:Phusion ratio with optimized PCR reaction condition (Table 2.4.4 a) to generate 3'-A overhang error free PCR product. High quality genomic DNA was used as a template in the 20 µl PCR reaction mixture and the thermocycler condition was programmed according to table below (Table 2.4.4 b).

Table 2.4.3 (a) 20 µl PCR reaction to amplify promoter fragment for TOPO® TA Cloning®.

PCR reaction	Volume
10x NH_4 buffer	2.0 µl
50 mM $MgCl_2$	1.2 µl
10 mM dNTPs	0.4 µl
10 µM Forward primer	0.8 µl
10 µM Reverse primer	0.8 µl
Nuclease-free water	13.1 µl
Taq polymerase (5 u/µl)	0.4 µl
Phusion polymerase (2 u/µl)	0.3 µl
DNA template (100 ng)	1.0 µl
Total volume	20.0 µl

Table 2.4.4 (b) Two-step PCR conditions.

Steps	Temperature (°C)	Time
Initial denaturation	98	1 min 30 sec
35 cycles	98	30 sec
	60 - 68	3 min
Final extension	60 - 68	7 min

2.4.4 Colony PCR

Positive colonies were screened by colony PCR after bacterial transformation, generally using a combination of insert specific primer and vector backbone specific primer. Single colonies were picked using sterile pipette tips, streaked onto antibiotic selection plate (Luria Agar) and grown as required. The remaining cells on the tips were dipped into 20 µl of PCR reaction mixture and gently swirled to release the cells (Table 2.4.2 a). Reaction tubes were then placed in a PCR thermal cycler and thermocycling was performed as described below (Table 2.4.5).

Table 2.4.5 Two-step colony PCR conditions.

Steps	Temperature (°C)	Time
Initial denaturation	94	5 min
35 cycles	94	1 min
	60 - 68	3 min
Final extension	60 - 68	7 min

2.5 Cloning promoter expression construct

2.5.1 TOPO® TA Cloning®

Promoter entry clone was constructed using pENTR™ 5'-TOPO™ TA Cloning™ kit (Invitrogen™) which allows the direct transfer of 5'-upstream sequence of interest into a plasmid vector in a defined order and orientation. The pENTR™5'-TOPO® promoter entry clone, containing *att*L4 and *att*R1 sites, was later recombined with an *att*L1 and *att*L2-flanked marker entry clone and pTHattR4-R2 destination vector in a MultiSite Gateway® LR recombination reaction to create a promoter expression construct. The reaction mixture was prepared according to the manufacturer's instruction and is listed in Table 2.5.1.

Table 2.5.1 6 µl reaction mixture for MultiSite LR cloning.

pENTR™ 5'-TOPO® TA reaction mixture	Quantity
PCR product (promoter fragment)	1 to 4 µl
Salt solution	1 µl
pENTR™5'-TOPO vector®	1 µl
Nuclease-free water	Accordingly
Total volume	**6 µl**

The pENTR™ 5'-TOPO® TA reaction mixture was prepared in a 1.5 ml microfuge tube. The mixture was vortexed briefly, spun and incubated at room temperature overnight.

3 µl of the LR reaction was added to 30 µl of α-Select Chemically Competent Cells (Bioline), gently mixed and incubated on ice for 30 minutes. The mixture was heat shocked at 42 ºC in a water bath for 45 seconds and transferred back on ice for another 5 minutes. 400 µl of LB media was added and the culture was incubated at 30 ºC for 2 hours with gentle shaking (200 rpm). The culture was divided into two volumes (100 µl and 300 µl) and spread on LA plates containing antibiotic selection. The plates were allowed to dry and then incubated for 64 hours (approximately 2 days and a half) at 30 ºC.

2.5.2 MultiSite Gateway® LR combination reaction

Verified promoter and marker entry clones were recombined into a destination vector through a MultiSite (Two-site) LR reaction. The reaction components were prepared in a standardised molar ratios such that each reaction contained 5 fmol of each entry clone and 10 fmol of the desired destination vector (Table 2.5.2). The amount of plasmid DNA needed to achieve the desired fmol was calculated according to the following formula:

$$\text{Amount needed (ng)} = \text{desired fmol} \times \text{size of vector (bp)} \times (660 \times 10^{-6})$$

Table 2.5.2 5 µl reaction mixture for MultiSite LR cloning.

Two-site LR reaction mixture	Quantity
Destination vector	10 fmol
Entry clone I (promoter)	5 fmol
Entry clone II (marker)	5 fmol
LR Clonase™ II Plus	1 µl
TE buffer	Accordingly
Total volume	**5 µl**

The LR reaction mixture was prepared in a 1.5 ml microfuge tube with the LR Clonase™ II Plus enzyme (Thermo Fisher Scientific Invitrogen™) added last. The mixture was then vortexed briefly, spun and incubated at room temperature overnight.

The reaction was terminated with 0.5 µl of proteinase K (Invitrogen™) at 37 °C for 10 mins. Then, 2.5 µl of the LR reaction was added to 25 µl of α-Select Chemically Competent Cells (Bioline), gently mixed and incubated on ice for 30 minutes. The mixture was heat shocked at 42 °C in a water bath for 45 seconds and transferred back on ice for another 5 minutes. 400 µl of LB media was added and the culture was incubated at 37 °C for 1 hours with gentle shaking (200 rpm). The culture was divided into two volumes (100 µl and 300 µl) and spread on LA plates containing antibiotic selection. The plates were allowed dry and then incubated overnight at 37 °C.

2.5.3 Restriction enzyme digestion

Restriction endonucleases were purchased from New England Biolabs (UK). Plasmid DNA was diluted to 1 μg of DNA with dH$_2$O. The enzyme reaction mixture was prepared according to the table below (Table 2.5.3). The enzyme reaction mixture was incubated at 37 °C, for 20 minutes for Time-Saver enzymes, or 1 hour for standard enzymes. The digestion reaction was analyzed by agarose gel electrophoresis.

Table 2.5.3 20 μl reaction mixture for restriction enzyme digest.

1x reaction mixture	Quantity
10x reaction buffer	2.0 μl
Enzyme 1	1 unit
Enzyme 2 (if needed)	1 unit
Plasmid DNA	1 μg
dH$_2$O	Accordingly
Total volume	**20.0 μl**

2.6 Sanger sequencing of DNA

Each DNA sample was prepared in a 1.5 ml microfuge tube according to the guidelines provided by GATC Services-Eurofins Genomics. Template DNA was prepared at a concentration of 80 – 100 ng/μl for purified plasmid DNA and 20 – 80 ng/μl for purified PCR product. 5 μl of primer with a concentration of 5 μM was added together with the template DNA in the same tube. When needed, TE buffer was added to make the total volume of sample to 10 μl per tube. A Barcode label (LightRun tube, GATC Services) was affixed to the tube for sample and sequencing result tracking.

2.7. Bioinformatic analysis

2.7.1 Upstream and coding sequence (CDS) retrieval

Upstream and CDS sequences were retrieved using BioMart tool integrated within EnsemblPlants and Phytozome12 website (Howe et al., 2020; Goodstein et al., 2012). Sequences upstream of ATG start codon were collected by clicking 'Flank-coding region (Gene)' in the 'Attributes' tab's 'Sequences' section. For collecting the CDS of a gene,

'Coding sequence' was selected in the 'Attributes' tab's 'Sequences' section. All sequences were downloaded in FASTA format for further use.

2.7.2 Binding motif analysis using RSAT

Regulatory Sequence Analysis Tools (RSAT) with default parameter was used for DUO1 binding sites searches. The retrieved upstream sequences from EnsemblPlants and Phytozome12 were used to scan for the presence of DBS. The 'dna-pattern' tool was used to pattern match the retrieved sequences to the DUO1 binding motif consensus RRCSGTT, described in Higo et al. (2018). Lastly, 'feature map' tool was used to visualize the location of DUO1 binding motif within the upstream sequences.

2.7.3 Sequence analysis and cladogram building

Sequence homologues were identified by performing BLASTP search in genome database of Phytozome 12, Ensembl Plants and CoNekT (Co-expression Network Toolkit). The E-value was set to 0.001 (1.0e-3) and default settings were maintained for other algorithm parameters. Species option and algorithm settings are not available in CoNekT as the search tool was much simpler and only requires entry of the protein sequence. Ortholog candidates that met the cutoff value of E <0.001 and bit score >50 were selected and exported to the BioMart tool, integrated in both Phytozome 12 and Ensembl Plants, for CDS sequence retrieval (Kinsella et al., 2011). Candidate sequences were saved in My Data Cart tool in Phytozome 12 for future use. For CoNekT, the BLAST hit result for the protein was selected for the gene description, expression profile and orthologous genes (OG in CoNekT). The OG was selected and CDS sequences for the orthologous genes were downloaded.

All nucleotide sequence retrieved were translated to protein sequences before being aligned using default MUSCLE parameters in MEGA 10 and CLC Sequence Viewer (CLC Bio, a QIAGEN Company, Aarhus, Denmark) software (Kumar et al., 2016). Protein sequences of DAZ3 homologues were used to build a cladogram using default Maximum Likelihood parameter with a bootstrap value of 1000.

2.8 Antheridia preparation and DAPI staining analysis

Three to five gametophores were placed in a 1.5 ml microcentrifuge tube containing fixative (50 % v/v ethanol, 3.7 % v/v formaldehyde and 5 % v/v acetic acid) and fixed for

30 minutes. Then, the fixative was removed by pipetting and replaced with water. Antheridia structures were carefully brushed into a drop of water on a polylysine-coated slide (Sigma-Aldrich, USA) and allowed to dry overnight. Antheridia were stained the following day with DAPI for visualisation. Images were viewed and captured using fluorescence microscopy on a Nikon ECLIPSE 80i (Nikon, Japan) instrument.

2.9 Microscopy and image processing

Antheridia were imaged using bright field, differential interference contrast (DIC) and fluorescence microscopy (Nikon ECLIPSE 80i). An LED-based excitation source (CoolLED, presicExcite) was used together with a Plan Fluor 40x / 1.3 NA oil immersion objective or a Plan Apo VC 60x / 1.4 oil immersion objective. Fluorescence images were captured with a DS-QiMc cooled CCD camera (Nikon, Japan). Images were previewed and captured using NES-Elements Basic Research version 4.13.04 software. Image processing was performed using Adobe Photoshop 7.0.1 and ImageJ software.

2.10 Data analysis

Statistical analysis was performed using Microsoft Excel 2013. Chi-square (χ^2) test was used to analyse the significant difference between the number of observed genotype compared to the number of expected genotype. The result was considered significantly different when the p-value was less than α of 0.05. Analysis of variance (ANOVA) test was used to determine which data sets differ in the multiple comparisons. Post hoc analysis was carried out with a Tukey-Kramer test following ANOVA analysis. All tests carried out were one-sided and a p-value less than an α of 0.05 was considered to be statistically significant.

Chapter 3: Functional analysis of *DAZ3* and *DAZ3L* in sperm cell development

Abstract

Background and Aims

Flowering plants produce two functional sperm cells during the gametophytic phase of the plant life cycle. However, there is limited knowledge of the molecular mechanisms governing sperm cell specification and development. Arabidopsis *DUO1-ACTIVATED ZINC FINGER 3* (*DAZ3*) and its homolog, *DAZ3-LIKE* (*DAZ3L*) encode C_2H_2 zinc finger proteins that are known to be specifically expressed in sperm cells of developing pollen. This study aims to investigate the evolutionary origin of *DAZ3*/*DAZ3L* and their potential role in sperm cell development.

Methods

To explore sequence conservation of *DAZ3* in flowering plants, sequence analysis was performed on *DAZ3* and its orthologs and subsequently extended to include DAZ1 and DAZ2 as related ancestral sequences. To investigate the *in-vivo* role of *DAZ3* and *DAZ3L*, the CRISPR-Cas9 technique was used to create knockout mutants. Fertility and seed production were analysed to examine the phenotypes of mutant plants.

Key Results

DAZ3 homologues were present in eudicots superrosid-superasterid clade, whereas *DAZ3L* was restricted to Brassicaceae. Phylogenetic analysis showed distinct DAZ3/DAZ3L clades resulting from gene duplication and the DAZ3/DAZ3L clade is proposed to be derived from ancestral *DAZ1* sequences. Analysis of RNA-seq data revealed that *DAZ3* and *DAZ3L* are highly expressed in mature pollen and sperm, which is maintained in growing pollen tubes. Plants harbouring single mutant knockout alleles for *daz3* and *daz3l* plants were fertile and genetic transmission for both genes was unaffected through pollen. Double mutant *daz3 daz3l* plants produced a normal number of viable seeds per silique compared with wild type plants and gave rise to fertile offspring.

Conclusions

The work supports the conclusion that *DAZ3* and *DAZ3L* do not have essential functions in sperm cells, fertilisation or seed development and may therefore have cryptic or conditional roles in sperm cell functions.

3.1 Introduction

The zinc finger protein gene, *DAZ3*, was identified as a putative direct target of the transcription factor DUO1 in Borg et al. (2011), whereas its homolog, *DAZ3L*, was identified through bioinformatic analysis (Taimur, 2014). Previous work by Taimur (2014) reported that *DAZ3* and *DAZ3L* promoters show sperm cell-specific expression which is restricted to tricellular and mature pollen (discussed in section 3.3). The expression profiles of *DAZ3* and *DAZ3L* were validated by RT-PCR analysis using samples from sporophytic tissues and pollen developmental stages. In addition, the expression of gene constructs in which the native *DAZ3* and *DAZ3L* promoters drive the reporter H2B-GFP was present specifically in Arabidopsis sperm (Borg et al., 2011; Taimur, 2014). The same expression pattern was also observed when *DAZ3* and *DAZ3L* native promoters were used to drive the expression of DAZ3 and DAZ3L mCherry fusion proteins. *DAZ3L* promoter activity is reduced in *duo1* mutant germ cells, while in transient expression assays in tobacco leaves, DUO1 is able to transactivate the *DAZ3L* promoter. Thus, similar to *DAZ3*, *DAZ3L* is also considered a direct target of DUO1. Protein localisation analysis using reporter constructs showed that DAZ3 was located predominantly in the cytoplasm while DAZ3L was present in both the nucleus and cytoplasm. The same localisation patterns are maintained for both proteins even when the sperm cells are present in pollen tubes grown *in vitro* (Taimur, 2014).

The aim of the work presented in this chapter was to investigate the evolutionary origin of *DAZ3* and *DAZ3L* sequences and their potential functional role in sperm cell development. C_2H_2 type zinc finger proteins are known to influence various developmental processes including floral organogenesis and seed development (Englbrecht et al., 2004). *DAZ3* is one of the most abundant transcripts in sperm cells, which suggest important functions in sperm cell differentiation including processes before and/or after fertilization. Previous work involving the generation of *DAZ3* knockdown plants by RNAi did not yield any abnormal phenotype (Taimur, 2014). This could suggest that the function of *DAZ3* is substituted by its homolog, *DAZ3L*, and that *DAZ3* and *DAZ3L* might have redundant functions. Alternatively, RNAi may not have been fully effective at reducing RNA levels of *DAZ3* since *DAZ3* transcript levels were not measured in RNAi knockdown plants (Taimur, 2014). Therefore, the hypothesis investigated was that *DAZ3* and *DAZ3L* have important, but redundant functions in sperm cell differentiation.

3.2 Sequence conservation of DAZ3 in flowering plants

3.2.1 DAZ3 and DAZ3L structure and domains

The *DAZ3* and *DAZ3L* proteins are characterised by possession of one K2-2 zinc finger domain and two DLNxxP EAR motifs at their C-terminal ends (Englbrecht et al., 2004; Kagale et al., 2010). Both *DAZ3* (At4g35700) and *DAZ3L* (At4g35610) are located on chromosome 4 in close proximity with just nine genes separating them. They are C_2H_2-type zinc finger proteins (ZFP) in which the C_2H_2-type zinc finger is stabilised by the interaction of pair of cysteine (C) and histidine (H) residues with a coordinating zinc ion (Figure 3.2.1.1.a.). *DAZ3* and *DAZ3L* zinc fingers also contain two short β-sheet forming regions and one α-helical region, the latter containing a signature K2-2 motif (Figure 3.2.1.1.b.).
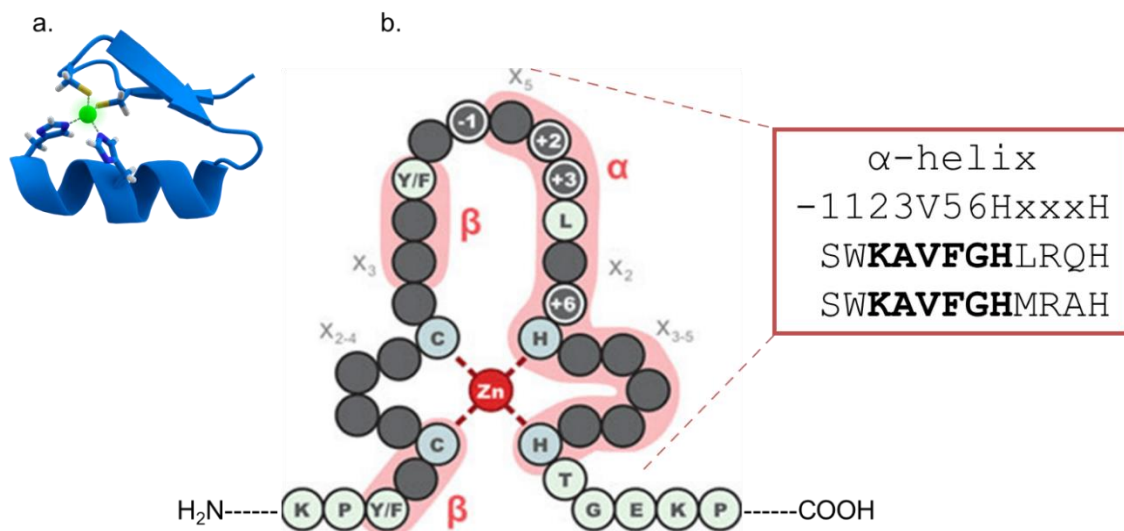
Figure 3.2.1.1. C2H2 zinc finger protein with folded protein domain. a. A typical C_2H_2-type zinc finger structure with coordinating zinc ion (green). **b.** Zinc finger structure displaying the DAZ3 and DAZ3L K2-2 DNA binding motif position in the α-helix structure. The two short β-sheet structures are also shown. Image (a) by Thomas Splettstoesser is released under a CC BY-SA 4.0 license (https://creativecommons.org/licenses/by-sa/4.0/) with no changes made and (b) is adapted from Stubbs et al. (2011).

Both DAZ3 and DAZ3L have H3XH spacing pattern between the histidine residues in the ZF domain. This places them in the C1 family of C_2H_2-type ZFP according to Englbrecht et al. (2004). They are also members of the C1-1i subfamily as both DAZ3 and DAZ3L only possess a single ZF domain. Like all C_2H_2 ZF proteins, the ZF motif recognises its target DNA sequence by binding to the major groove of DNA. DAZ3L has been identified to interact with DNA sequence with four base pairs core, AGCT (Franco-Zorilla et al., 2014; Wang et al., 2020). Likewise, DAZ3 binds to similar DNA sequence since both have identical K2-2 motif residues (Figure 3.2.1.2).

The domain structures of DAZ3 and DAZ3L proteins were annotated and drawn to scale based on sequence alignment (further described in subchapter 3.2.2) using MEGA 10 (Kumar et al., 2018). DAZ3 and DAZ1 are of similar length and the percentage identity between the two proteins is 49 % (Figure 3.2.1.2.).
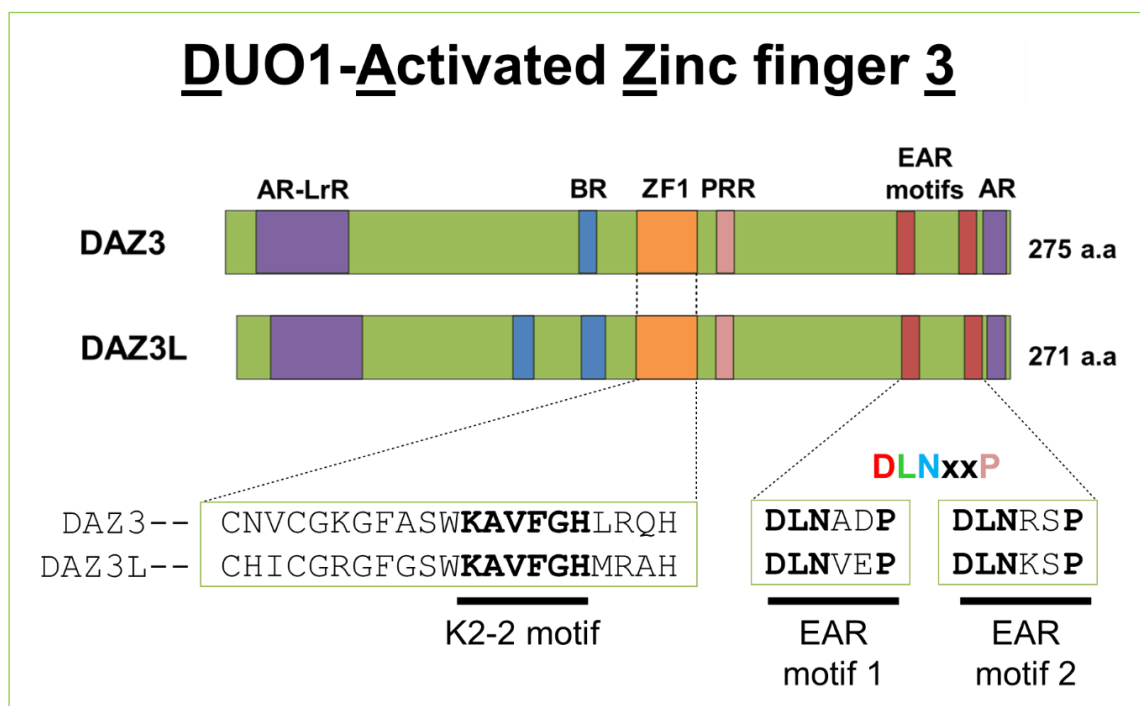
Figure 3.2.1.2. Schematic diagram of *DAZ3* and *DAZ3L* proteins for *Arabidopsis thaliana*. Updated version of *DAZ3* and *DAZ3L* proteins showing the K2-2 and EAR motif signature structure with other additional conserved regions. AR, acidic rich region; LrR, leucine rich region; BR, basic region; ZF1, zinc finger 1; PRR, proline rich region; EAR, Ethylene-responsive element binding factor-associated amphiphilic repression.

DAZ3 and DAZ3L sequences were re-analysed in detail to examine the presence of any additional conserved regions in addition what has been described (Englbrecht et al., 2004; Taimur, 2014). Acidic regions are present in DAZ3 and DAZ3L, near both N- and C-terminal ends (Figure 3.2.1.3). However, the acidic region near the N-terminal end is longer and followed by a leucine-rich region in both proteins. Similar to that previously described, a basic region is present in both proteins before the K2-2 ZF domain. However, an additional basic region (KKKSRPR) is present in DAZ3L, which could also be a nuclear localisation signal (NLS) (Chelsky et al., 1989). In addition, a region rich in proline (P) is present after the zinc finger domain in both DAZ3 and DAZ3L.

Figure 3.2.1.3. Protein sequence alignments of *DAZ3* and *DAZ3L* in Arabidopsis. Protein sequences were aligned using MEGA 10 and exported to CLC Sequence Viewer. The DAZ3 signature K2-2 zinc finger and EAR motifs are shown in black and red boxes and indicated with labels, respectively. Additional conserved regions are indicated above the sequence as follows: AR, acidic rich region; LrR, leucine rich region; PRR , proline rich region; Basic region, blue boxes.

3.2.2 Sequence analysis of DAZ3 homologs in flowering plants

To explore the extent of conservation of DAZ3-related sequences among flowering plants, the DAZ3 protein sequence (GenBank: ABK32160) was used in BLASTP searches of the genomes of all known angiosperm species using Phytozome 12 and Ensembl Plants (release 48). The E-value threshold was set to 0.001 (1.0e-3) and default settings were maintained for all other algorithm parameters (Figure 3.2.2.1). Ortholog candidates that met the cut-off value of E <0.001 and bit score >50 were selected for CDS sequence retrieval. The sequences were analysed using MEGA 10 and individually checked to exclude any repetitive sequence (different ID of the same gene used by different database) before being aligned using the MUSCLE algorithm with default parameters (Kumar et al., 2018). A total of 207 sequences with various numbers (up to 3) of zinc fingers were included as candidate orthologs of DAZ3.
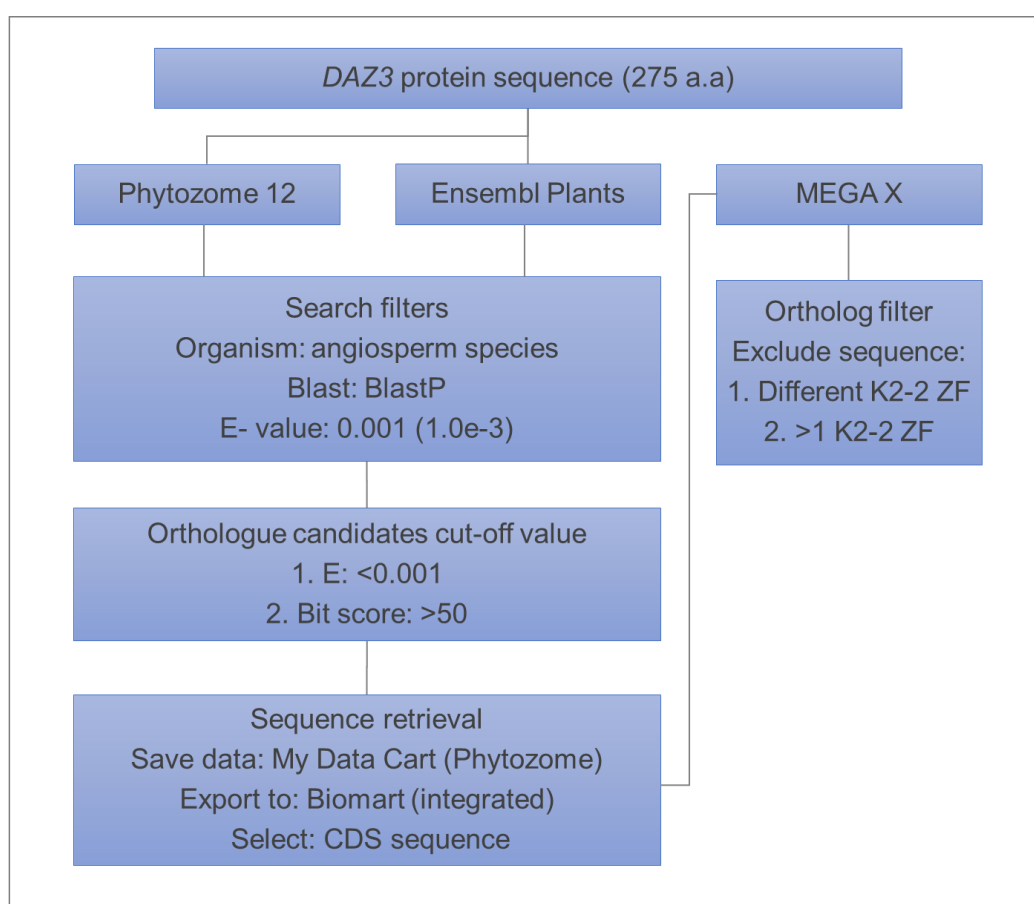
Figure 3.2.2.1. Bioinformatic flow diagram for identifying AtDAZ3 homologues. AtDAZ3 sequence was used in a BLASTP search in two different plant database platforms. CDS sequences were retrieved for candidate genes. Sequences that have more than one or different K2-2 zinc finger domain were excluded.

A filter was later applied to exclude sequences with different or more than one K2-2 zinc finger domain. This reduced the number of candidate orthologs sequences to 64. A pairwise comparison for overall amino acid identity was conducted to check the reliability of the multiple sequence alignment. The amino acid identity was 34 % which was above the 30 % threshold for reliable alignments. Protein alignment was then exported into CLC Sequence Viewer (QIAGEN Bioinformatics) to generate a superior graphical image.

Sequences with homology to DAZ3 are only present in the eudicot clade, with the majority in superrosids (Saxifragales and rosids), and fewer among asterids (Figure 3.2.2.3). The N-terminal acidic-rich region was conserved in Malvaceae and Brassicaceae orthologs in the Malvids clade (Figure 3.2.2.2). This region however has a greater proportion of acidic residues among DAZ3 homologues in Brassicaceae. It is noteworthy, that the acidic region is also present near the N-terminal end of DAZ3 sequences in Saxifragales (Kfe, *Kalanchoe fedtschenkoi*; Kla, *Kalanchoe laxiflora*) even though other clade in the rosids does not have it. Another conserved sequence high in leucine (L) residue is also observed after the stretch of acidic region. This feature however is present only in Brassicaceae with consensus sequence I/LVLG/ALPALR/N/SL. In addition, this region is highly comprised of non-polar/hydrophobic amino acid residues.

Stretch of basic region is present before the ZF domain in all DAZ3 homologs in superrosids (Figure 3.2.2.2). Nevertheless, the number of basic region is varying from one to two in the rosid clade. This basic region could serve as NLS region that would translocate the protein to the nucleus by nuclear transport.

The K2-2 type zinc finger was present in all aligned DAZ3 sequences and showed the characteristic KALFGH motif or the single residue variant KA<u>V</u>FGH (Figure 3.2.2.4). Interestingly, all Brassicaceae sequences possess a KAVFGH motif, whereas most other species have either KALFGH, KAVFGH or other variations. Such examples are observed in CcaDAZ3 (*Corcorus capsularis*), TcaDAZ3 (criollo) (*Theobroma cacao Belizian Criollo*), TcaDAZ3 (matina) (*Theobroma cacao Matina*) and GraDAZ3 (*Gossypium raimondii*) of Malvaceae. Among sequences from Malvids, the K2-2 motif is more varied in Malvaceae compared with Brassicaceae. GraDAZ3 possess the KAVFGH motif while in CcaDAZ3 and TcaDAZ3 other variations (KGVFGH and RGVFGH, respectively) are observed. Similarly, in Malpighiales, MesDAZ3 (*Manihot esculenta*) and RcoDAZ3 (*Ricinus communis*) have the KGVFGH variant, whereas other proteins have either KALFGH or KAVFGH (Figure 3.2.2.4). Outside Malvids clade, the KAVFGH motif can still be observed in sequences from Cucurbitales (Fabids) CusaDAZ3A (*Cucumis*

sativus) and ClaDAZ3B (*Citrullus lanatus*). Other motif variations in Malvids clade include KGAFGH, KAAFGH and RAAFGH. The majority of species in the Fabids clade possess the canonical K2-2 motif, KALFGH, which is also observed in all asterid sequences examined. This analysis indicated that the KAVFGH motif starts to appear in the rosids clade but is more fixed in Malvids clade compared to Fabids.

Another highly conserved region is the proline (P) rich region located after the ZF domain (Figure 3.2.2.4). Intriguingly, proline at position 3 is highly conserved and retain at the same position in all sequences analysed. This indicates the potential importance of this residue, which could be crucial for the function of DAZ3 and its orthologs.
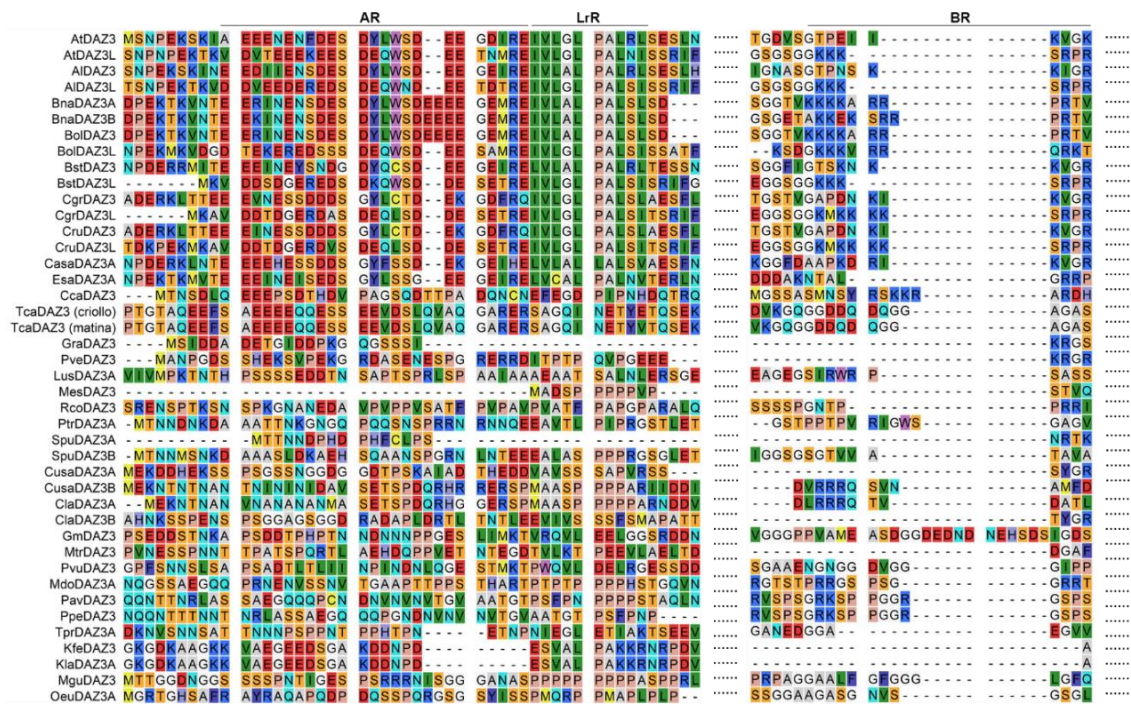
Figure 3.2.2.2. Part of multiple sequence alignment of DAZ3 homologs in flowering plants. Protein sequences were aligned using MUSCLE with default parameters in MEGA 10. AR, acidic rich region; LrR, leucine rich region; BR, basic region. Species: At, *Arabidopsis thaliana*; Al, *Arabidopsis lyrata*; Bna, *Brassica napus*; Bol, *Brassica oleracea*; Bst, *Boechera stricta*; Cgr, *Capsella grandiflora*; Cru, *Capsella rubella*; Casa, *Camelina sativa*; Esa, *Eutrema salsugineum*; Cca, *Corcorus capsularis*; Tca (criollo), *Theobroma cacao Belizian Criollo*; Tca (matina), *Theobroma cacao Matina*; Gra, *Gossypium raimondii*; Pve, *Pistacia vera*; Lus, *Linum usitatissimum*; Mes, *Manihot esculenta*; Rco, *Ricinus communis*; Ptr, *Populus trichocarpa*; Spu, *Salix purpurea*; Cusa, *Cucumis sativus*; Cla, *Citrullus lanatus*; Gm, *Glycine max*; Mtr, *Medicago truncatula*; Pvu, *Phaseolus vulgaris*; Mdo, *Malus domestica*; Pav, *Prunus avium*; Ppe, *Prunus persica*; Tpr, *Trifolium pratense*; Kfe, *Kalanchoe fedtschenkoi*; Kla, *Kalanchoe laxiflora*; Mgu, *Mimulus guttatus*; Oue, *Olea europaea var. sylvestris*.

Figure 3.2.2.3. Phylogenetic tree showing major groups of eudicots. Tree was drawn according to Angiosperm Phylogeny Group (APG) classification (Chase et al., 2016). Species related to DAZ3 sequence analysis were shown using abbreviations. Star represents triplication events (Panchy et al., 2016). Species: At, *Arabidopsis thaliana*; Al, *Arabidopsis lyrata*; Bna, *Brassica napus*; Bol, *Brassica oleracea*; Bst, *Boechera stricta*; Cgr, *Capsella grandiflora*; Cru, *Capsella rubella*; Casa, *Camelina sativa*; Esa, *Eutrema salsugineum*; Cca, *Corcorus capsularis*, Tca (criollo), *Theobroma cacao Belizian Criollo*; Tca (matina), *Theobroma cacao Matina*; Gra, *Gossypium raimondii*; Pve, *Pistacia vera*; Lus, *Linum usitatissimum*; Mes, *Manihot esculenta*; Rco, *Ricinus communis*; Ptr, *Populus trichocarpa*; Spu, *Salix purpurea*; Cusa, *Cucumis sativus*; Cla, *Citrullus lanatus*; Gm, *Glycine max*; Mtr, *Medicago truncatula*; Pvu, *Phaseolus vulgaris*; Mdo, *Malus domestica*; Pav, *Prunus avium*; Ppe, *Prunus persica*; Tpr, *Trifolium pratense*; Kfe, *Kalanchoe fedtschenkoi*; Kla, *Kalanchoe laxiflora*; Mgu, *Mimulus guttatus*; Oue, *Olea europaea var. sylvestris*.

Figure 3.2.2.4. Part of multiple sequence alignment of DAZ3 homologs in flowering plants. Protein sequences were aligned using MUSCLE with default parameters in MEGA 10. BR, basic region; ZF, zinc finger 1; PRR, proline rich region. Species: At, *Arabidopsis thaliana*; Al, *Arabidopsis lyrata*; Bna, *Brassica napus*; Bol, *Brassica oleracea*; Bst, *Boechera stricta*; Cgr, *Capsella grandiflora*; Cru, *Capsella rubella*; Casa, *Camelina sativa*; Esa, *Eutrema salsugineum*; Cca, *Corcorus capsularis*, Tca (criollo), *Theobroma cacao Belizian Criollo*; Tca (matina), *Theobroma cacao Matina*; Gra, *Gossypium raimondii*; Pve, *Pistacia vera*; Lus, *Linum usitatissimum*; Mes, *Manihot esculenta*; Rco, *Ricinus communis*; Ptr, *Populus trichocarpa*; Spu, *Salix purpurea*; Cusa, *Cucumis sativus*; Cla, *Citrullus lanatus*; Gm, *Glycine max*; Mtr, *Medicago truncatula*; Pvu, *Phaseolus vulgaris*; Mdo, *Malus domestica*; Pav, *Prunus avium*; Ppe, *Prunus persica*; Tpr, *Trifolium pratense*; Kfe, *Kalanchoe fedtschenkoi*; Kla, *Kalanchoe laxiflora*; Mgu, *Mimulus guttatus*; Oue, *Olea europaea var. sylvestris*.

Another characteristic of DAZ3 is the presence of two EAR motifs near the C terminal end. In Brassicaceae almost all DAZ3 orthologs possess two EAR motifs with the DLNxxP consensus (Figure 3.2.2.5). This characteristic is not present in individual sequences from BolDAZ3L and EsaDAZ3L, and EsaDAZ3B, where the first or both EAR motifs are absent, respectively. The number of residues separating DLN and P in both EAR motifs is constant in Brassicaceae with two amino acid residues in between. This number of residues however varies between one to two for other species in the ortholog list. For example, among Malvaceae species CcaDAZ3, TcaDAZ3 (Criollo) and GraDAZ3 all have a single residue between (Figure 3.2.2.5). The number of EAR motif also varies within Malvaceae; CcaDAZ3 has a single EAR motif while TcaDAZ3 and GraDAZ3 have three.

Still in the Malvids clade, a different number of residues between the EAR motifs is observed in PveDAZ3 of Sapindales order (Figure 3.2.2.5). The first EAR motif of PveDAZ3 has the DLNxxP consensus while the second EAR motif has the DLNxP consensus. The same pattern is also identified in Malpighiales order. In Fabids clade, majority of DAZ3 homologs possess one EAR motif (the second EAR in relation to AtDAZ3) with DLNxxP consensus. If the first EAR motif is present, the consensus tends to be DLNxP, such as in PvuDAZ3 (*Phaseolus vulgaris*). Only one EAR motif is present in asterid clade with either DLNxxP or DLNxP consensus.

A short acidic region that could serve as a potential transcriptional activation domain is conserved near the C-terminal end of DAZ3 orthologs and is markedly acidic in the Malvids clade (Brassicaceae and Malvaceae).

Cladistic analysis was performed to identify any potential clade formed from DAZ3 coding sequences. The analysis was performed using Maximum Likelihood in MEGA 10 with bootstrap value of 1000 (Kimura, 1980). The model and rates among sites were set to Kimura 2-parameter and G+I, respectively. Default settings were maintained for all other algorithm parameters.

The cladogram of DAZ3 orthologs revealed that DAZ3L formed a new clade in Brassicaceae (Figure 3.2.2.6). This DAZ3L clade is supported by a very high bootstrap value of 95 %. When cross-refer to the multiple sequence analysis, it is noticed that all species in Brassicaceae that belongs to the DAZ3L clade possesses two basic regions. The basic regions are both located in between the leucine rich region and the ZF domain (Figure 3.2.2.2; Figure 3.2.2.4).
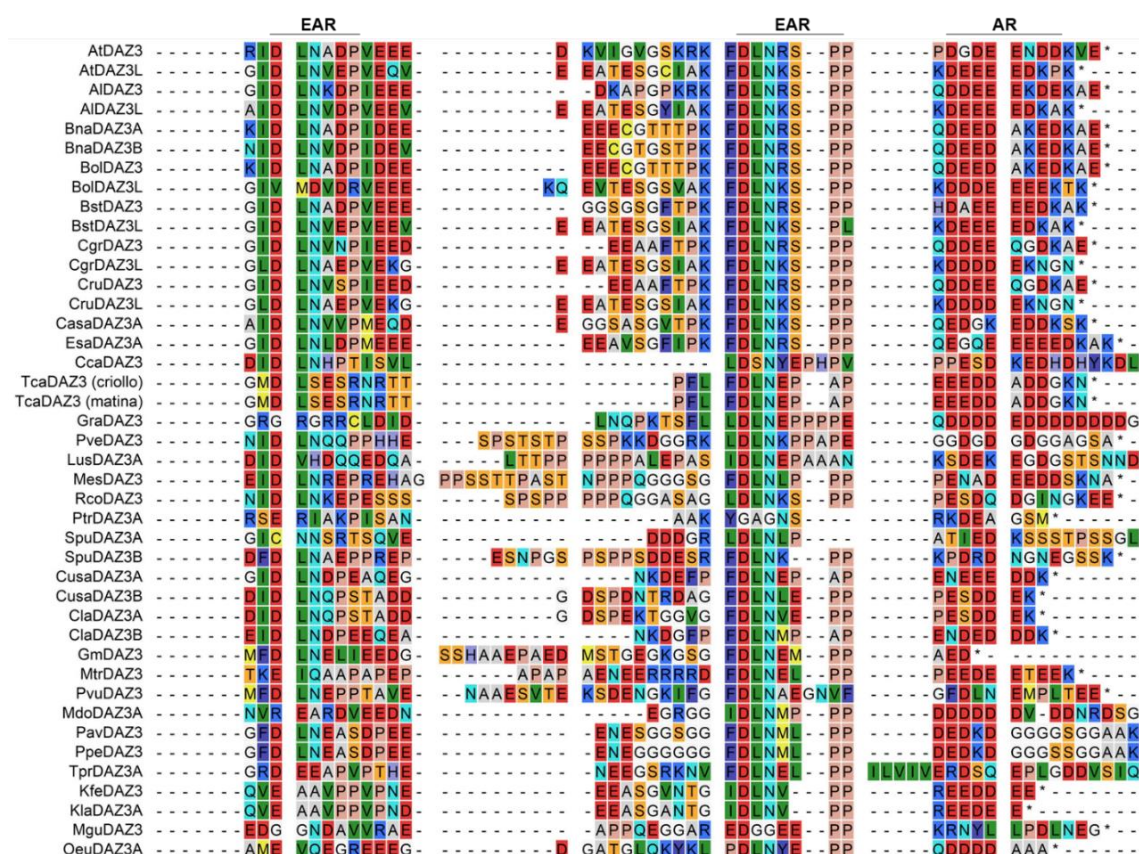
Figure 3.2.2.5. Part of multiple sequence alignment of DAZ3 homologs in flowering plants. Protein sequences were aligned using MUSCLE algorithm with default parameters in MEGA 10. BR, basic region; ZF, zinc finger 1 and PRR proline rich region. EAR, Ethylene-responsive element binding factor-associated amphiphilic repression; AR, acidic rich region. Species: At, *Arabidopsis thaliana*; Al, *Arabidopsis lyrata*; Bna, *Brassica napus*; Bol, *Brassica oleracea*; Bst, *Boechera stricta*; Cgr, *Capsella grandiflora*; Cru, *Capsella rubella*; Casa, *Camelina sativa*; Esa, *Eutrema salsugineum*; Cca, *Corcorus capsularis*, Tca (criollo), *Theobroma cacao Belizian Criollo*; Tca (matina), *Theobroma cacao Matina*; Gra, *Gossypium raimondii*; Pve, *Pistacia vera*; Lus, *Linum usitatissimum*; Mes, *Manihot esculenta*; Rco, *Ricinus communis*; Ptr, *Populus trichocarpa*; Spu, *Salix purpurea*; Cusa, *Cucumis sativus*; Cla, *Citrullus lanatus*; Gm, *Glycine max*; Mtr, *Medicago truncatula*; Pvu, *Phaseolus vulgaris*; Mdo, *Malus domestica*; Pav, *Prunus avium*; Ppe, *Prunus persica*; Tpr, *Trifolium pratense*; Kfe, *Kalanchoe fedtschenkoi*; Kla, *Kalanchoe laxiflora*; Mgu, *Mimulus guttatus*; Oue, *Olea europaea var. sylvestris*.

Figure 3.2.2.6. Rectangular cladogram of DAZ3 orthologs. Cladistic analysis showing of DAZ3 sequences inferred by using the Maximum Likelihood method with bootstrap of 1000 replicates. Branches corresponding to partitions reproduced in less than 50% bootstrap replicates are collapsed. Proteins labelled in red are the member of DAZ3L clade.

3.2.3 Evolutionary origin of *DAZ3*

To further investigate the ancestral origin of *DAZ3*, its full-length protein sequence was used in BLASTP searches with E (expect)-values equal to one and with default settings for other parameters in Phytozome 12 and Ensembl Plants. This enabled the retrieval of candidates for distant homologues that meet the cut-off value of E ≤1 and bit score ≥45 from all angiosperms with sequenced genomes. The sequences were individually checked to remove duplicates before being aligned using default MUSCLE parameter in MEGA 10 and CLC software (Kumar et al., 2016). Since DAZ3 is a K2-2 type zinc finger protein, sequence hits which did not contain either of the KALFGH or KAVFGH motifs in the K2-2 zinc finger domain were ignored. The presence of EAR motifs, which is another feature of DAZ3, at the C-terminal end of the proteins sequence were also used as an inclusive filter for selection.

Among all genomes examined, 120 proteins with homology to DAZ3 were found in various plant species. Within this set are DAZ1 and DAZ2 with lengths of 270 and 284 amino acid residues, respectively. *DAZ1* is located on chromosome 2 whereas *DAZ2* is located on chromosome 4 separated from *DAZ3* and *DAZ3L* by 42 and 33 genes.

DAZ1 and DAZ2 share similar signature features (K2-2 domain and EAR motifs) to DAZ3/DAZ3L but possess two further zinc finger domains (Figure 3.2.3.1). Pairwise sequence alignment was performed using MEGA 10, to identify the percentage identity between DAZ3, DAZ3L, DAZ1 and DAZ2 (Table 3.2.3.). Between DAZ1 and DAZ2, as well as DAZ3 and DAZ3L, a similarity percentage of around 50 % can be seen. However, comparing DAZ1/DAZ2 with DAZ3/DAZ3L reveals a 25-28 % similarity between these two groups.

Table 3.2.3. Protein percentage identity of DAZ1, DAZ2, DAZ3 and DAZ3L. Amino acids pairwise comparison was made to calculate the proportional (*p*) distance between two proteins. Protein identity = 1 – *p*-distance.

	Protein percentage identity (%)			
	AtDAZ1	AtDAZ2	AtDAZ3	AtDAZ3L
AtDAZ1				
AtDAZ2	56.4			
AtDAZ3	25. 1	26.8		
AtDAZ3L	28.5	27.4	49.2	

To evaluate the relatedness of all 120 protein sequences, the aligned protein sequences were retrieved and used to build a cladogram using MEGA 10 (Kumar et al., 2016). The Maximum Likelihood method with a bootstrap value of 1000 and other parameters set at default. A bootstrap value is the percentage in which the same diagram showed up when the test is replicated for certain number of times, in this case 1000.

Two major clades are formed, which are the DAZ3/DAZ3L and the DAZ1/DAZ2 clades (Figure 3.2.3.2). DAZ3/DAZ3L clade is first observed in eudicot species with a bootstrap value of 50 %. Meanwhile DAZ1/DAZ2 clade is present in all species representing all clades of flowering plants, bootstrap value 74 %. Lower bootstrap value observed in DAZ3/DAZ3L clade could be due to the lower species density for this clade. The cladogram also shows a sequence relation between members of the DAZ3 and DAZ1 clades. From the multiple sequence analysis, it can be suggested that *DAZ3* is most likely derived from *DAZ1* since they share similar features. Additionally, from the cladogram, it is clear that *DAZ3* is absent in monocot while *DAZ1*/*DAZ2* is present throughout the angiosperms. *DAZ3* could be duplicated from *DAZ1*/*DAZ2* before the eudicot diversification. The bootstrap confidence level between DAZ3/DAZ3L and DAZ1/DAZ2 clades might possibly be increased by performing a blast search using DAZ1 as a query. By combining the homologues from both DAZ1 and DAZ3, a more robust cladogram can be created.
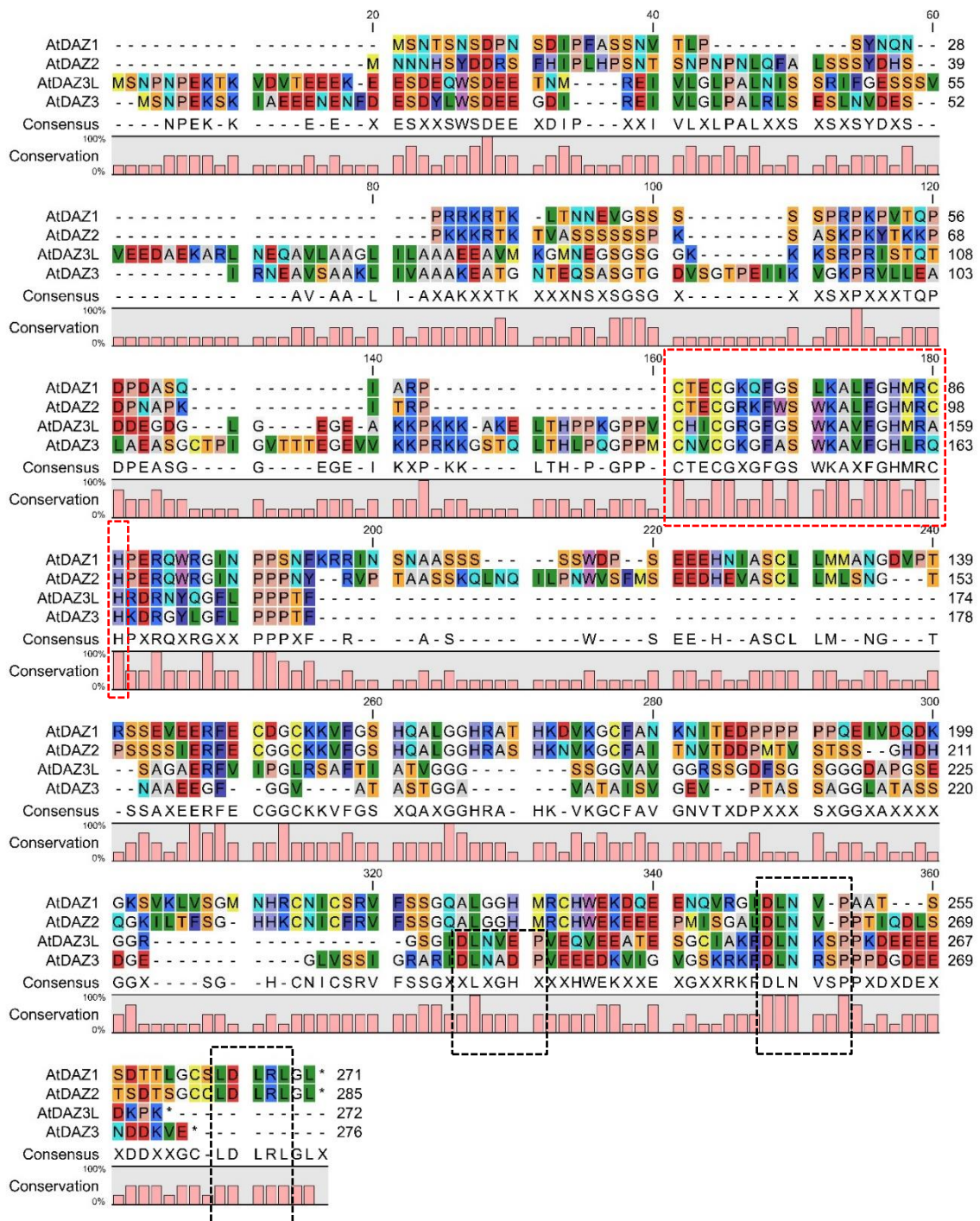
Figure 3.2.3.1. Sequence alignment of Arabidopsis DAZ proteins. Proteins were aligned using MUSCLE with default parameters in CLC Sequence Viewer. Conserved K2-2 zinc finger and EAR motifs are shown in boxes (red and black dotted box, respectively).

Figure 3.2.3.2. Circular cladogram of DAZ3/DAZ3L distance homologs. High E (expect)-value was used to retrieve distant homologs of DAZ3/DAZ3L using BLASTP. Maximum Likelihood cladistic tree with a bootstrap value of 1000 was generated using MEGA 10 based on MUSCLE alignment (default parameters) of protein sequences. Colours in cladogram represent the sequence clades. Arrow indicates point of divergence. Green, DAZ1/DAZ2 clades; blue, DAZ3/DAZ3L clade.

3.3 Analysis of *DAZ3* and *DAZ3L* expression in Arabidopsis

To understand the expression pattern of *DAZ3* and *DAZ3L* in developing pollen, RNA-seq data was analysed. The data analysed is publicly available via the RNA-seq resource web tool, Co-expression NetworkToolkit (CoNekT). The data for isolated pollen at different developmental stages from Columbia (Col-0) and Landsberg erecta (Ler-0) accessions were plotted (Figure 3.3.1; Figure 3.3.2). In addition, a comparative analysis was also made between the RNA-seq data with whole genome AGRONOMICS1 tiling array data (Twell group, unpublished) and published Affymetrix ATH1 microarray data, which support the findings based on RNA-seq data shown in figure 3.3.1 and 3.3.2 (Honys and Twell, 2004; Borges et al., 2008).

In the Col-0 accession low levels of *DAZ3* and *DAZ3L* transcripts are present in microspores (UNM) and in bicellular (BCP) stages (Figure 3.3.1). Expression increases as the bicellular stage (BCP) progresses, but peaks sharply in the tricellular pollen (TCP) and shows a further increase in sperm cells (SPC). The expression level of both genes is high in mature pollen (MP) and increases further in the pollen tubes (PT). In terms of the expression level in MP compared to SPC, *DAZ3* and *DAZ3L* expression are 32- and 38-fold higher in SPC which are consistent with the sperm cell-specific expression of these genes based on reporter gene analysis (Taimur, 2014). The dominant contribution of vegetative cell transcripts in the mature pollen RNA sample dilutes the transcript signal from sperm cells therefore making the expression appear less abundant. Although sperm cell and pollen tube data are not available for the Ler-0 accession, the same developmental expression profile is observed (Figure 3.3.2). *DAZ3* and *DAZ3L* expression is not detected in microspores and only low levels are present in BCP (Figure 3.3b). Again, the expression of both genes increases rapidly in TCP and is maintained at high levels in MP.
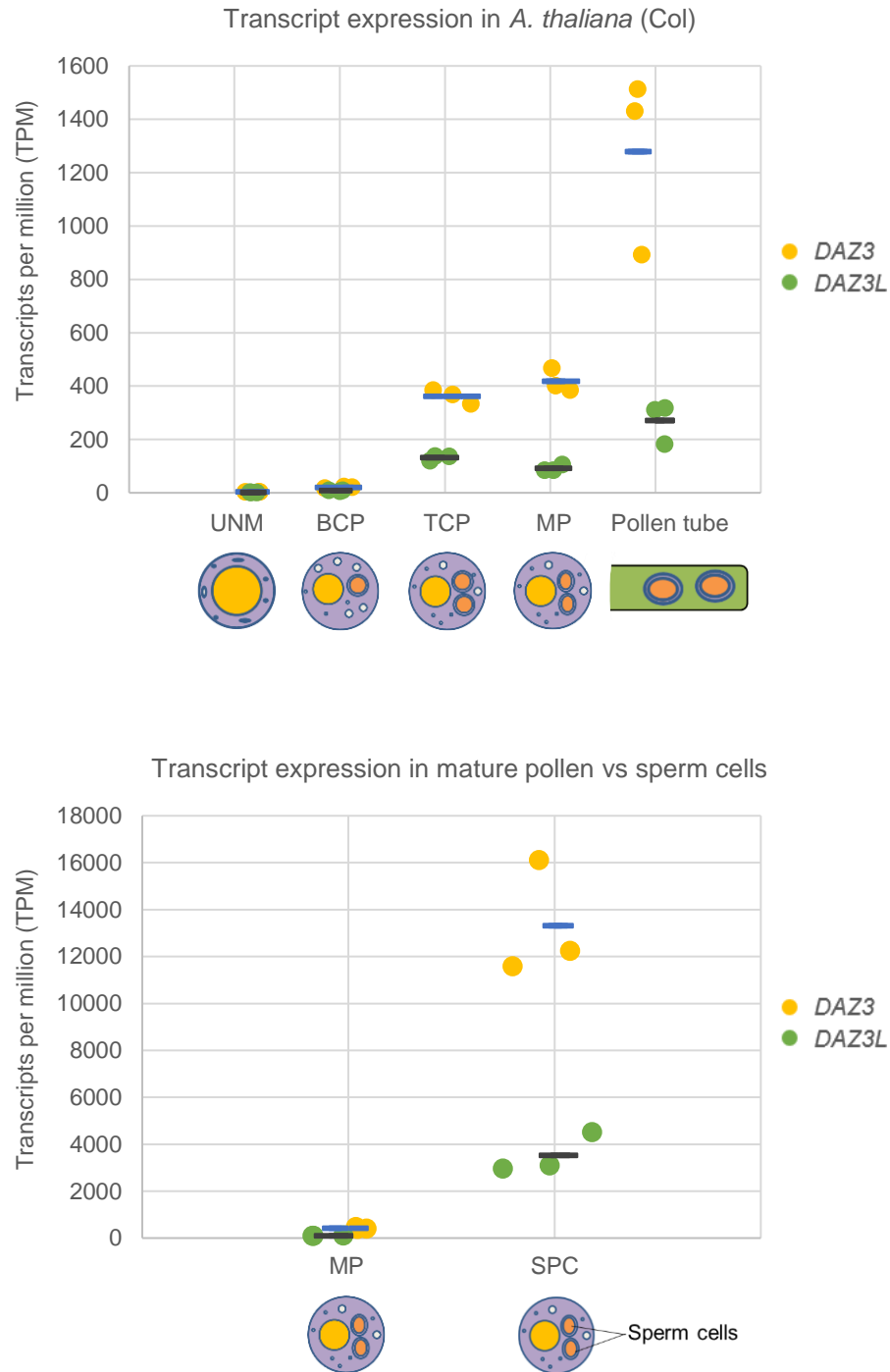
Figure 3.3.1. Transcript expression of *DAZ3* and *DAZ3L* in *A. thaliana* (Col-0). Data for three biological replicates are plotted as transcripts per million (TPM) from RNA-seq data in pollen developmental stages. UNM, uninucleate microspores; BCP, bicellular pollen; TCP, tricellular pollen; MP, mature pollen; SPC, sperm cells.
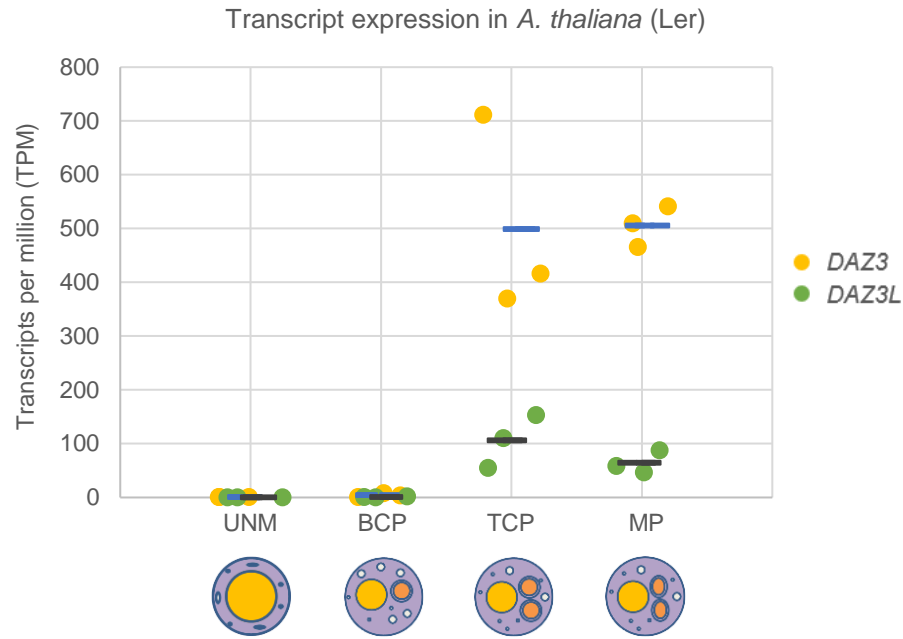
Figure 3.3.2. Transcript expression of *DAZ3* and *DAZ3L* in *A. thaliana* (Ler). Three biological replicates for Transcripts per million (TPM) from RNA-seq data in pollen developmental stages. UNM, uninucleate microspores; BCP, bicellular pollen; TCP, tricellular pollen; MP, mature pollen.

Data from RNA-seq is comparable to that obtained with the AGRONOMICS1 and ATH1 microarrays (Figure 3.3.3). On the AGRONOMICS1 platform, *DAZ3* and *DAZ3L* signals are absent or detected at low levels in UNM and BCP stages before increasing sharply in TCP and is maintained at high levels in MP. Meanwhile, only *DAZ3* signal is available in ATH1 data due to the unavailability of *DAZ3L* probes on the array, but the pattern is similar to that observed for RNA-seq data. A consistent observation is that *DAZ3* expression is always higher compared to *DAZ3L* on all platforms. This is consistent with ATH1 data from isolated sperm in which *DAZ3* is one of the most highly expressed genes in Arabidopsis sperm cells (Borges et al. 2008). Based on the results, there is a huge increase in expression of *DAZ3* and *DAZ3L* in tricellular pollen after division of the generative cell due to expression in sperm cells, with transcript abundance increasing further in pollen tubes.
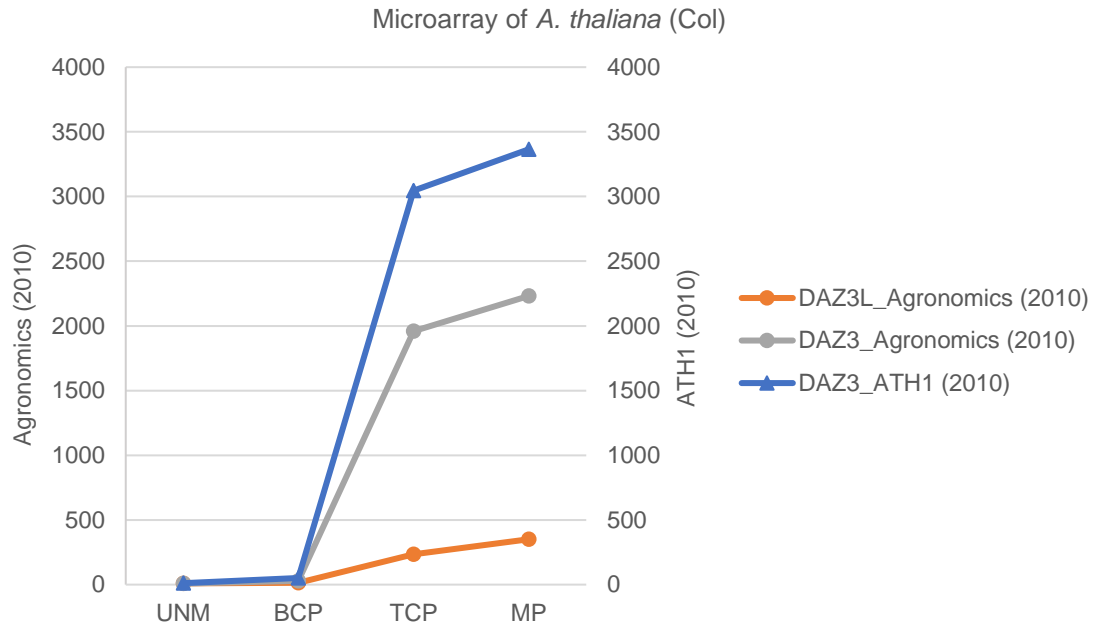
Figure 3.3.3. Microarray analysis of *DAZ3* and *DAZ3L* expression. Data are plotted for probes corresponding to each gene based on AGRONOMICS1 and ATH1 platform. Data for AGRONOMICS is a single replicate while ATH1 is an average of three replicates. UNM, uninucleate microspores; BCP, bicellular pollen; TCP, tricellular pollen; MP, mature pollen. *DAZ3L* probe is not available in the ATH1 array, which is why it is not shown.

3.4 Investigating the *in vivo* role of *DAZ3* and *DAZ3L* in *Arabidopsis thaliana*

3.4.1 Generation of *daz3* and *daz3l* CRISPR knockout mutants

To study the function of *DAZ3* and *DAZ3L* in Arabidopsis sperm cell development, a CRISPR/Cas9 based strategy was used to generate *daz3* and *daz3l* knockout mutants. This strategy was chosen for its efficiency as it allows a precise and specific DNA region to be mutated. Two sets of constructs were generated for *DAZ3* and *DAZ3L*; *DAZ3* sg8+69 (construct 1), *DAZ3* sg1+4 (construct 2), *DAZ3L* sg25+66 (construct 1) and *DAZ3L* sg14+39 (construct 2) (Figure 3.4.1.1). Each construct contains two single guide RNAs (sgRNAs) cloned into pHEE401, which harbours a *CAS9* gene driven by the EC1 egg cell-specific promoter (Xing et al., 2014; Wang et al., 2015). The promoter EC1-CAS9 cassette allows early and specific expression of CAS9 to facilitate the generation of homozygous or biallelic mutant plants in the T1 generation. The two sgRNAs were designed using the CRISPR-P v2.0 webtool by choosing sgRNAs that have minimal off targets with at least 4 mismatches. The two guides will target homologous sequences

located upstream of the K2-2 zinc finger sequence to generate severe mutations introduced through non-homologous end joining (NHEJ) repair mechanism of the double stranded break by CAS9 nuclease, predicted to disrupt the open reading frame. *Agrobacterium* mediated transformation in Arabidopsis was performed for all CRISPR constructs. T1 seeds containing CRISPR construct were selected by growing them on hygromycin selection plates. The resistant plants were screened for the presence of insertions or deletions (indels) by PCR genotyping and candidates showing differences in the expected product size were confirmed by sequencing.

In T1 generation *DAZ3* sg1+4 and *DAZ3* sg8+69 transformed plants, one homozygous and two chimeric mutants were generated from a total 43 and 24 plants screened, respectively (Table 3.4.1.1). Meanwhile, one homozygous and three chimeric mutants were produced for *DAZ3L* sg25+66 from a total 28 plants screened (Table 3.4.1.1). All mutant lines with homozygous mutations have deletions of 20 and 22 base pairs, which create frame shifts leading to an early stop codon. In the case of plants showing chimeric mutations, multiple alleles were generated from a single mutant line. All chimeric plants have three PCR products of different sizes (Figure 3.4.1.2). The sequencing result for each of the chimeric plants shows a deletion of 16 to 22 base pairs for the smallest band, creating an early stop codon due to frame shift.
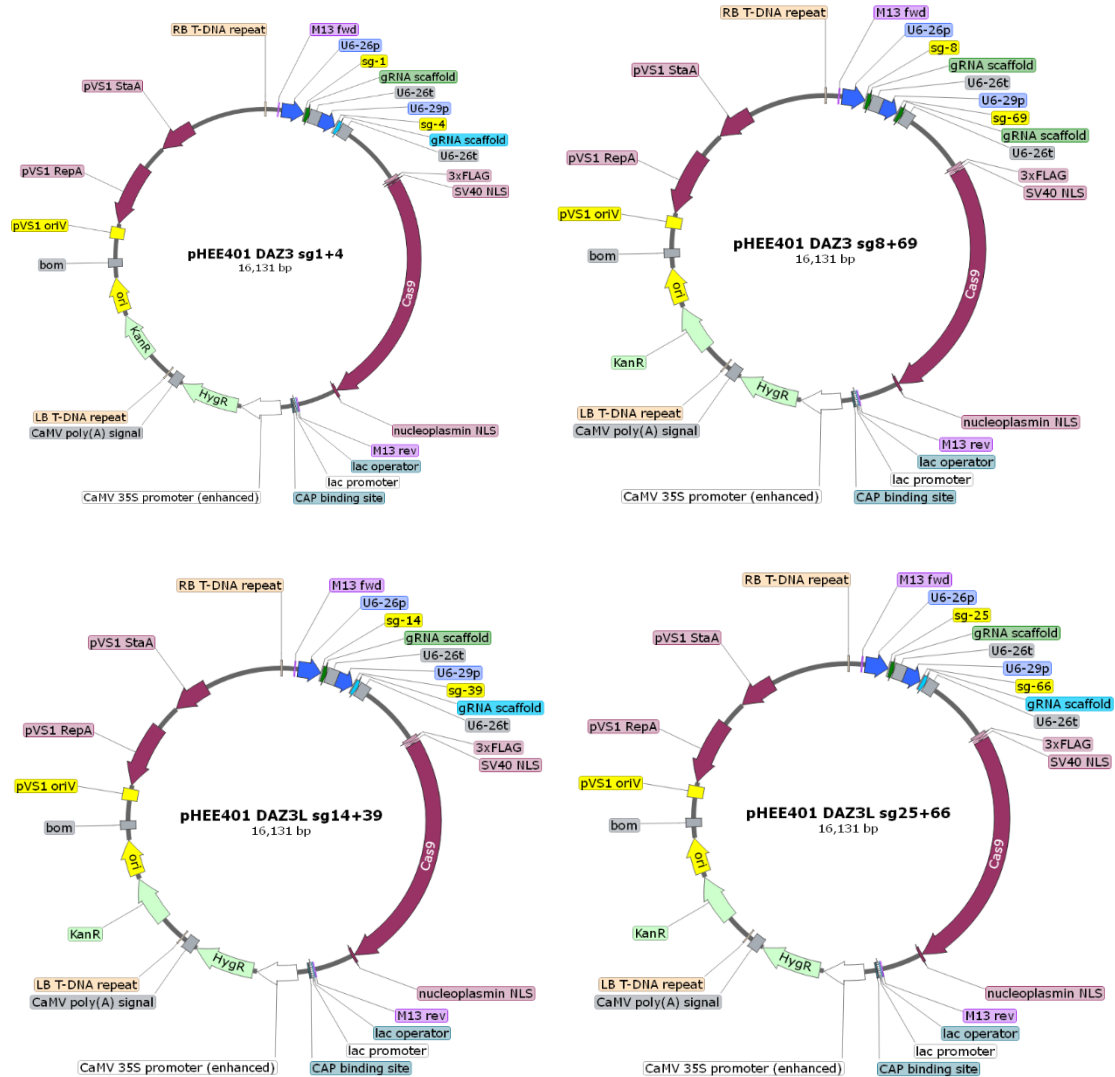
Figure 3.4.1.1. *DAZ3* and *DAZ3L* **CRISPR constructs.** Two single guides were cloned into pHEE401 plasmid containing CAS9 gene driven by EC1 egg cell-specific promoter (Xing et al., 2014; Wang et al., 2015). Labels on map: sg-1, single guide 1; sg-4, single guide 4; sg-8, single guide 8; sg-69, single guide 69; sg-14, single guide 14; sg-39, single guide 39; sg-25, single guide 25; sg-66, single guide 66; RB T-DNA repeat, right border repeat from nopaline C58 T-DNA; M13 fwd, M13 forward sequencing primer; M13rev, M13 reverse sequencing primer; U6-26p, U6 gene promoter; U6-26t, U6-26 terminator; U6-29p, U6 gene promoter; gRNA scaffold, guide RNA scaffold for the Streptococcus pyogenes CRISPR/Cas9 system; 3xFLAG, three tandem FLAG® epitope tags; SV40 NLS, nuclear localization signal of SV40 large T antigen; Cas9, Cas9 (Csn1) endonuclease from the *Streptococcus pyogenes* Type II CRISPR/Cas system; nucleoplasmin NLS, bipartite nuclear localization signal from nucleoplasmin; lac operator, lactose operator; lac promoter, promoter for *E. coli lac* operon; CAP binding site, catabolite activator protein binding site; CaMV 35S promoter (enhanced),

cauliflower mosaic virus 35S promoter with a duplicated enhancer region; KanR, Kanamycin Resistance gene; HygR, Hygromycin Resistance gene; CaMV poly(A) signal, cauliflower mosaic virus polyadenylation signal; LB T-DNA repeat, left border repeat from nopaline C58 T-DNA; ori, plasmid origin of replication; bom, basis of mobility region from pBR322; pVS1 oriV, origin of replication for the *Pseudomonas* plasmid pVS1; pVS1 RepA, replication protein from the *Pseudomonas* plasmid pVS1; pVS1 StaA, stability protein from the *Pseudomonas* plasmid pVS1

All *DAZ3* and *DAZ3L* mutant lines were screened for the presence of the T-DNA in the T2 generation (Table 3.4.1.2). The aim was to find a stable mutant without the T-DNA. It is important to remove the T-DNA once the desired mutant plant is obtained to eliminate the possibility of the generation of off target mutations from continued activity of Cas9. Only two lines (*daz3* B2_P5 and *daz3l* B2_P14) segregated T-DNA-free mutant plants in the T2 generation. Even though both lines generated non-transgenic mutant progenies, in general a high number of plants were screened for all lines to obtain a single T-DNA free mutant plant. This indicates that the T-DNA inserted at multiple loci during the transformation process.

Table 3.4.1.1. T1 generation of *DAZ3* and *DAZ3L* knockout mutants. Plants were screened for the presence of large mutation by genotyping using PCR. The mutation was then confirmed by sequencing.

T1				
Gene	CRISPR construct	Total screen (no. of T-DNA)	Line	Genotype
DAZ3	sg 1+4	43 (1)	B1_B20	Homozygous
	sg 8+69	24 (2)	B2_P5	Chimera
			B2_P9	Chimera
DAZ3L	sg 25+66	28 (4)	B1_P5	Homozygous
			B1_P6	Chimera
			B2_P5	Chimera
			B2_P14	Chimera
	sg 14+39	11 (0)	-	-
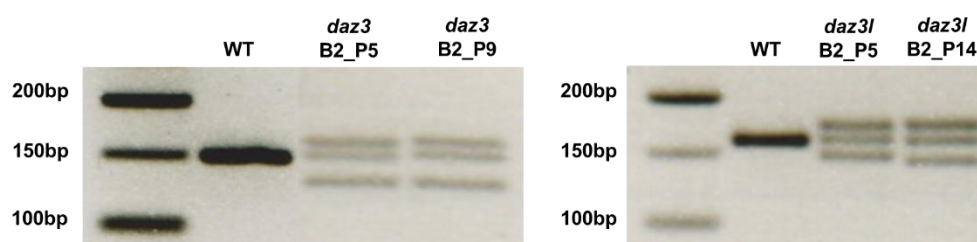
Figure 3.4.1.2. Genotyping of T1 generation of *DAZ3* and *DAZ3L* knockout line. *DAZ3* (sg 1+4 and sg 8+69) and *DAZ3L* (sg 14+39 and sg 25+66) transformed plants were screened for the presence of indels according to amplicon size. The agarose gel shows an example of genotyping by PCR screening, *daz3*, line B2_P5 and B2_P9 and *daz3l*, line B2_P5 and B2_P14, appear to be chimeras based on the presence of three bands with different sizes.

Table 3.4.1.2. *DAZ3* and *DAZ3L* non-transgenic mutant lines confirmation in T2 generation. T1 mutant lines was screened for the presence of T-DNA and for confirming germline transmission. Non-transgenic mutant line was sent for sequencing to confirm the genotype. X* = not send for sequencing.

Gene	Line (T1)	Total screen	T-DNA free	T2	
				Lines	Genotype
DAZ3	B1_B20	101	0	-	-
	B2_P5	51	5	B2_P5_P1H	x*
				B2_P5_P3B	Chimera
				B2_P5_P3E	Chimera
				B2_P5_P4D	x*
				B2_P5_P5H	x*
	B2_P9	33	0	-	-
DAZ3L	B1_P5	48	0	-	-
	B1_P6	48	0	-	-
	B2_P5	40	0	-	-
	B2_P14	34	3	B2_P14_P7	Homozygous
				B2_P14_P15	Chimera
				B2_P14_P26	x*

Primer pairs that bind to the promoter and terminator of sgRNAs were used to screen transformed lines for the presence of T-DNA. Only five plants were confirmed to lack the T-DNA out of 51 T2 plants for *daz3* B2_P5 line (Figure 3.4.1.3). Since *daz3* B2_P5 line was a chimeric plant, all alleles will segregate in the T2 generation, producing various allele combinations. All five *daz3* B2_P5 progeny appear to be biallelic based on agarose gel analysis. *DAZ3* PCR products for plants B2_P5_P3B and B2_P5_P3E were sequenced to confirm the genotype. Both B2_P5_P3B and B2_P5_P3E lower/smaller size bands have the same 22 base pairs deletion that creates early stop codon while their upper/bigger size bands have several nucleotide deletion, addition and change. The result for upper band is also a mix sequence by the presence of the double peaks in the chromatogram, suggesting that more than one allele is present. Hence, it is hard to confirm the allele's genotype. However, the double peaks did not continue until the end of the chromatogram but merged after approximately 45 base pairs. This shows that the mutation did not create any frame shift, therefore did not lead to any amino acid mutation or early codon termination. Both B2_P5_P3B and B2_P5_P3E are confirmed as a chimera by sequencing. Plant B2_P5_P3B was later grown and a homozygous deletion plant, B2_P5_P3B_3H, was generated in the T3 generation as a result of segregation of the chimeric allele (Figure 3.4.1.5). This line was free from T-DNA and was selected for *daz3* single mutant phenotypic analysis.

For *daz3l* B2_P14 line, three out of 34 T2 plants were confirmed to be non-transgenic (Figure 3.4.1.4). PCR product from one plant produced a small size single band while the other two produced a double band on the agarose gel. Plants B2_P14_P7 and B2_P14_P15 of *daz3l* B2_P14 line were genotyped for the mutations by sequencing. Plant B2_P14_P7 was confirmed as a homozygous with a 16 bp deletion which creates a frame shift leading to an early stop codon. Plant B2_P14_P15 has the same deletion as Plant B2_P14_P7 for its smaller sized band. However, the larger sized band has a mixed sequence indicated by double chromatogram peaks that might be due to nucleotide insertion. Hence, plant B2_P14_P15 is confirmed as a chimera by the sequencing result. Plant B2_P14_P7 was chosen for *daz3l* single mutant phenotypic analysis (Figure 3.4.1.6).
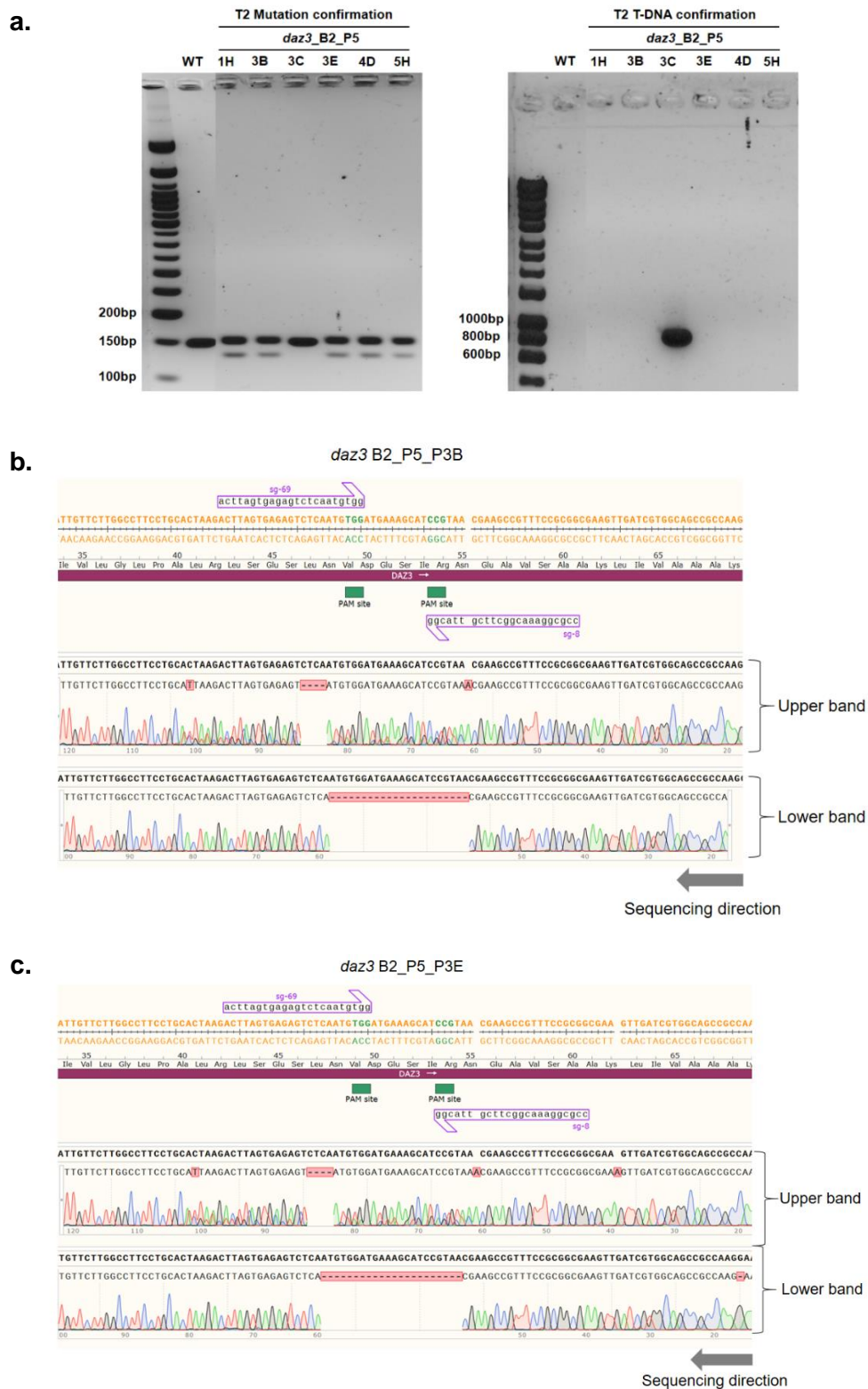
Figure 3.4.1.3. *daz3* B2_P5 non-transgenic mutant lines identification and germline transmission confirmation. (a) PCR products of *daz3* B2_P5 progenies show five plants were biallelic T-DNA-free mutants. PCR products from two progenies, **(b)** B2_P5_P3B and **(c)** B2_P5_P3E, were sequenced and both were confirmed as chimeric.

a.

T2 Mutation confirmation

*daz3l*_B2_P14

P7 P15 P26 WT + control

200bp
150bp
50bp

T2 T-DNA confirmation

*daz3l*_B2_P14

P7 P15 P26 WT + control

1000bp
800bp
600bp

b.

daz3-like B2_P14_P7

Sequencing direction

c.

daz3-like B2_P14_P15

Upper band

Lower band

Sequencing direction

Figure 3.4.1.4. *daz3l* **B2_P14 non-transgenic mutant lines identification and germline transmission confirmation. (a)** PCR products of *daz3l* B2_P14 progenies show three plants were T-DNA free mutants, one homozygous and two biallelic. PCR products from two progenies, **(b)** B2_P14_P7 and **(c)** B2_P14_P15, were sequenced. gDNA from plant B2_P14_P7 was confirmed to be a homozygous mutant while B2_P14_P15 was confirmed as a chimera.

Figure 3.4.1.5. Workflow for creating homozygous *daz3* mutant plant. Two single guide RNA, sg 8 and sg 69, guided CAS9 nuclease to its target sequences upstream of the zinc finger region. *daz3* homozygous plant was produced in the T3 generation where 22 bp deletion was introduced through NHEJ during the repair of double stranded break by CAS9 nuclease. The deletion created a frameshift which resulted an early stop codon, leading to *daz3* protein truncation. Red arrow, primers used for screening the presence of indels; blue region, single guide target sequence; highlighted region, protospacer adjacent motif.
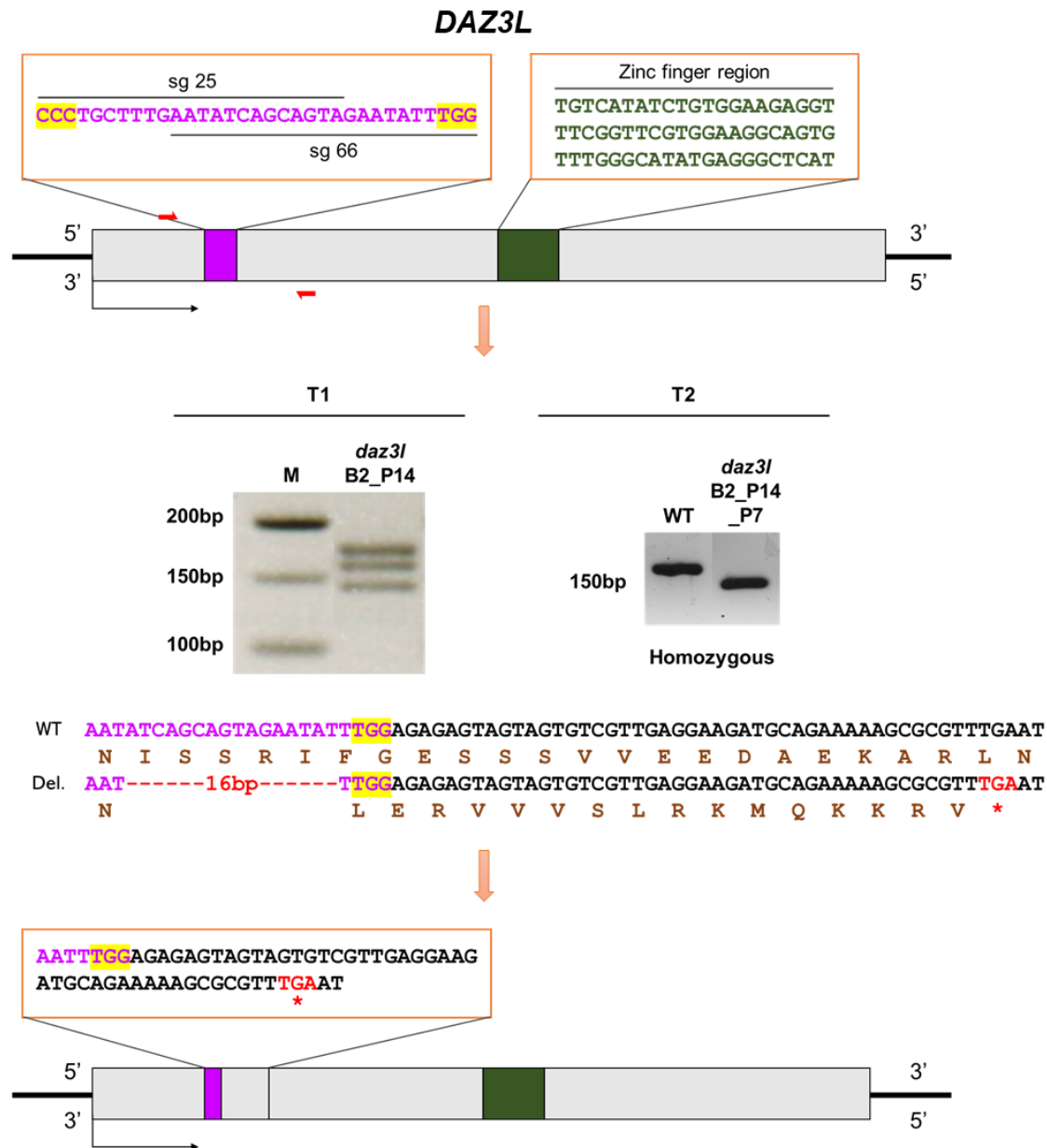
Figure 3.4.1.6. Workflow for creating homozygous *daz3l* mutant plant. Two single guide RNA, sg 25 and sg 66, guided CAS9 nuclease to its target sequences upstream of the zinc finger region. *daz3l* homozygous plant was produced in the T2 generation where 16 bp deletion was introduced through NHEJ during the repair of double stranded break by CAS9 nuclease. The deletion created a frameshift which resulted an early stop codon, leading to *daz3l* protein truncation. Red arrow, primers used for screening the presence of indels; pink region, single guide target sequence; highlighted region, protospacer adjacent motif.

3.4.2 Genetic analysis of *daz3* and *daz3l* single mutants

To investigate whether the mutation has an impact on the genetic transmission of the *daz3* allele, self and reciprocal crosses were carried out with wild type plants. The question to be answered was could the *daz3* allele successfully transmit to the next generation compared with the wild type allele? The prediction based on self-progeny was that if *daz3* was transmitted normally, the genotypic segregation ratio that would be expected to be 1:2:1 (Figure 3.4.2.1). Heterozygous *DAZ3*$^{+/-}$ mutants were allow to self and seeds were grown on soil. Genotypic analysis was performed on the progeny using PCR amplification to detect the genotype (*DAZ3*$^{-/-}$, *DAZ3*$^{+/-}$ and *DAZ3*$^{+/+}$). The same analysis was also performed on heterozygous *DAZ3*$^{+/-}$ mutant plants.
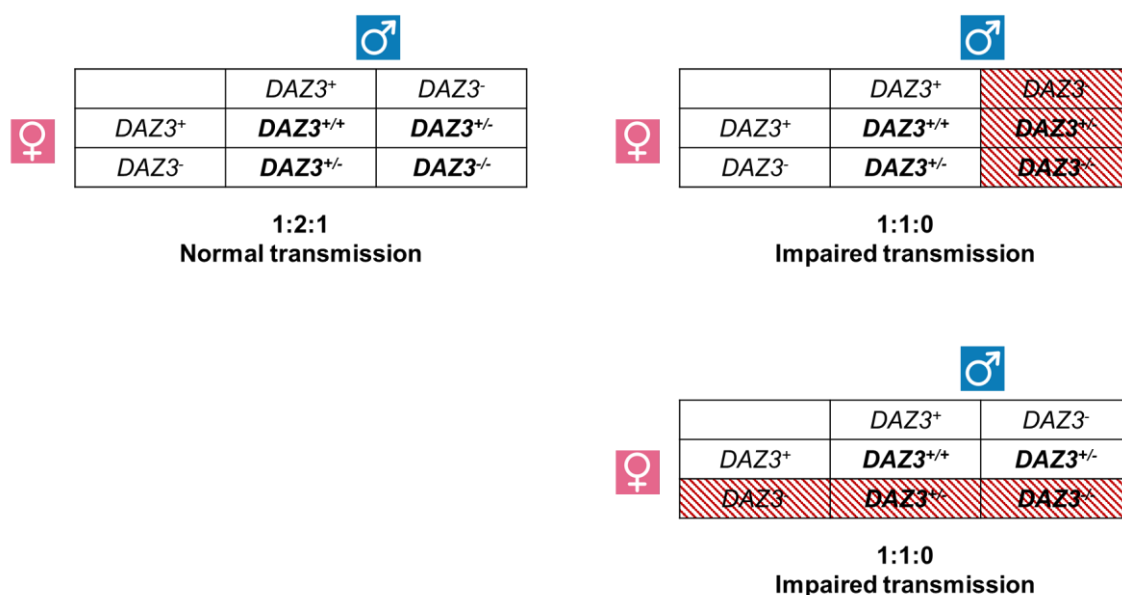
Figure 3.4.2.1. Punnet square showing transmission of the *daz3* mutant allele and the expected segregation ratios. Normal transmission with 1:2:1 ratio would be expected if *daz3* was successfully transmitted through male and female gametes. Impaired transmission with 1:1:0 ratio would be seen if *daz3* was unsuccessfully transmitted through either male or female gametes.

A total of 39 progeny plants were genotyped. PCR-based genotypic analysis on the plants found that 11 plants were wild type (*DAZ3*$^{+/+}$) and 20 and 8 plants were heterozygous (*DAZ3*$^{+/-}$) and homozygous (*DAZ3*$^{-/-}$) mutant, respectively. Segregation ratio was calculated for the observed genotyped compared to the total number of plant.

The observed ratio is 1.1 for wild type, 2.1 for *DAZ3+/-* mutant and 0.8 for *DAZ3-/-* mutant (Table 3.4.2.1).

Chi-square (χ^2) test was performed to analyse the significance difference between the number of observed genotype compared to the number of expected genotype (Table 3.4.4.2). The χ^2 value (χ^2 : 0.48) for the observed genotype has a p-value that is between 0.75 and 0.90. In order to be considered significantly different, the p-value has to be less than 0.05 critical value. The χ^2 p-value is more than the 0.05 critical value; hence, there is no significant difference between the number of observed genotype and the number of expected genotype for the *DAZ3+/-* progeny.

Table 3.4.2.1. *daz3* allele transmission analysis. Genotyping PCR was performed on *DAZ3+/-* progeny plants and genotype of each plants was recorded. Segregation ratio was calculated for the observed plant and compared with the predicted ratio.

Genotype	*DAZ3+/+*	*DAZ3+/-*	*DAZ3-/-*
(Observed/total)	11 / 39	20 / 39	8 / 39
Ratio (observed)	1.1	2.1	0.8

Table 3.4.2.2. Chi-square (χ^2) test for self-progeny of heterozygous *DAZ3+/-* plants. Deviation from the expected 1:2:1 Mendelian ratio for the genotypic classes was tested.

	DAZ3+/-		
Genotype	*DAZ3+/+*	*DAZ3+/-*	*DAZ3-/-*
Observed (O)	11	20	8
Expected (E) (1:2:1)	9.75	19.5	9.75
(O-E)	1.25	0.5	-1.75
(O-E)² / E	0.16	0.01	0.31
Conclusion	There is no significant difference between observed and expected frequency. Mutation does not affect *daz3* allele transmission, χ^2 (2, N = 39) = 0.48, p > .05		

PCR-based genotypic analysis of heterozygous *DAZ3L+/-* progeny identified 11, 15 and 13 plants were wild type, heterozygous (*DAZ3L+/-*) mutant and homozygous (*DAZ3L-/-*) mutant, respectively. Segregation ratio was calculated for the observed genotype against

the total number of progeny, which gives the ratio of 1.1 for wild type, 1.5 for $DAZ3L^{+/-}$ mutant and 1.3 for $DAZ3L^{-/-}$ mutant (Table 3.4.2.3).

A chi-square test was also conducted to see whether there is a significant difference between the number of observed genotyped and the number of expected genotype for $DAZ3L^{+/-}$ progeny (Table 3.4.2.4). From the chi-square distribution table, the calculated χ^2 value (χ^2 : 2.28), lies between the p-value of 0.25 and 0.5. This p-value is more than the 0.05 critical value. Hence, there is no evidence to support that there is a significant difference between the number of observed genotyped and the number of expected genotype for $DAZ3L^{+/-}$ progeny.

Table 3.4.2.3. *daz3l* allele transmission analysis. Genotyping PCR was performed on $DAZ3L^{+/-}$ progeny plants and genotype of each plants was recorded. Segregation ratio was calculated for the observed plant and compared with the predicted ratio.

Genotype	$DAZ3L^{+/+}$	$DAZ3L^{+/-}$	$DAZ3L^{-/-}$
(Observed/total)	11 / 39	15 / 39	13 / 39
Ratio (observed)	1.1	1.5	1.3

Table 3.4.2.4. Chi-square (χ^2) test for *daz3l*$^{+/-}$ progeny plants. Analysis was performed to test the significant difference between the number of observed genotype compared to the number of expected genotype (1:2:1) in the progeny plants.

	$DAZ3L^{+/-}$		
Genotype	$DAZ3L^{+/+}$	$DAZ3L^{+/-}$	$DAZ3L^{-/-}$
Observe (O) from experiment	11	15	13
Expected (E) from experiment; (1:2:1)	9.75	19.5	9.75
(O-E)	1.25	-4.5	3.25
(O-E)² / E	0.16	1.04	1.08
Conclusion	There is no significant difference between observed and expected frequency. Mutation does not affect *daz3l* allele transmission, χ^2 (2, $N = 39$) = 2.28, $p > .05$		

From the analysis, it could be seen that *daz3* and *daz3l* alleles have a normal genetic transmission. As a conclusion, the mutation has no impact on the transmission of *daz3* and *daz3l* allele and both alleles could be successfully transmitted to the next progeny.

3.4.3 Phenotypic analysis of *daz3* and *daz3l* single mutants

Based on the transcript expression data, *DAZ3* and *DAZ3L* are highly expressed in tricellular pollen and their expression peaks in sperm cells of mature pollen. Due to this, it is hypothesised that *DAZ3* and *DAZ3L* may have an important function in the development and function of sperm cells.

To study the functionality of sperm cells, the ability of $DAZ3^{-/-}$ and $DAZ3L^{-/-}$ plants to undergo fertilization and subsequently produce viable seeds was observed. Six siliques on the primary branch of seven weeks old plants were collected from one plant. The length of siliques was measured and later dissected to identify the viability of the seeds.

Viable seeds were observed in $DAZ3^{-/-}$ and $DAZ3L^{-/-}$ siliques, which the seeds were green, rounded and plump (Figure 3.4.3.1). Additionally, two categories of non-viable seeds were identified in mutant siliques. The first category was aborted seeds which was the result of embryo abortion. The seeds were either brownish and shrunken compared with the wild type, or white, translucent. The second category was undeveloped (or unfertilized) ovules which appear whitish, tiny and crumpled. Both viable and non-viable seeds (ovules) were also identified in wild type siliques.

Figure 3.4.3.1. Phenotype of seeds produced by *daz3* and *daz3l* mutants compared to those from wild type plants. Viable seeds were green, rounded and plump. Meanwhile, non-viable seeds were brownish and shrunken or white and translucent late aborting seeds. Scale bar = 0.3 mm.

The total number of seeds (viable and non-viable) for all siliques was recorded and the percentage of viable seeds per silique was presented as a scatter plot (Figure 3.4.3.2). In addition, the mean percentage of viable seeds were compared between $DAZ3^{-/-}$, $DAZ3L^{-/-}$, and wild type. On average, the percentage of viable seeds per silique for wild type is 96 %. Meanwhile, the average percentage for $DAZ3^{-/-}$ and $DAZ3L^{-/-}$ are 97 % and 98 %, respectively. One-way ANOVA was conducted to examine if the mean percentage of viable seeds in at least one of the test groups, which are WT, $DAZ3^{-/-}$ and $DAZ3L^{-/-}$, is different (Table 3.4.3.1; Appendix Table S3.3). The p-value from the test is 0.65 which is more than the 0.05 critical value, suggesting the mean percentage of viable seeds in all test groups are not different from each other ($F_{(2,15)} = 0.45$, $p = .65$). Therefore, the mean percentage of viable seeds for $DAZ3^{-/-}$ and $DAZ3L^{-/-}$ are comparable to each other as well to the wild type.
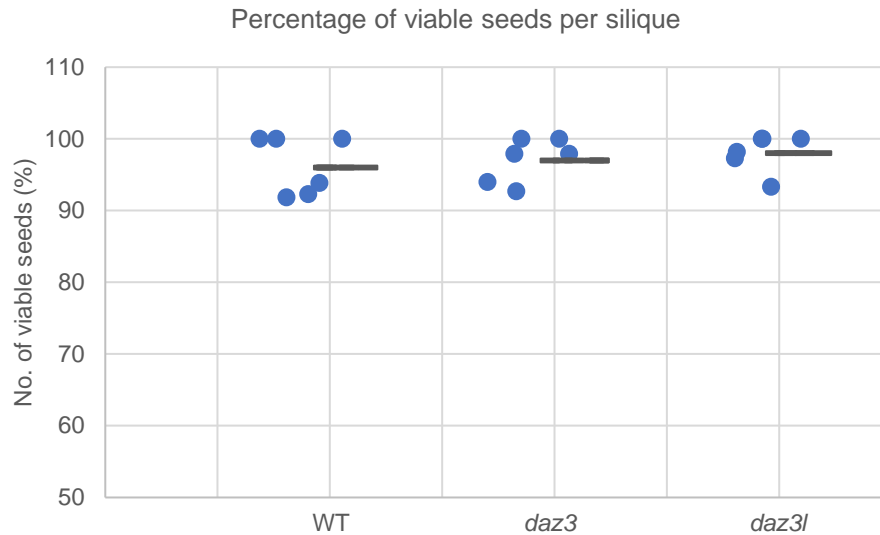
Figure 3.4.3.2. Scatter plot showing the percentage of viable seeds per silique in wild type and homozygous *daz3* and *daz3l* single mutant plants. Data for the number of viable seeds is plotted as percentage of viable seeds per silique (n = 6) for each genotype. The mean percentage of viable seeds from a total of six siliques is shown as a horizontal bar.

Table 3.4.3.1. One-way ANOVA for analysing any significant difference in the mean percentage of viable seeds per silique in three different test groups. The three test groups were wild type, $DAZ3^{-/-}$ and $DAZ3L^{-/-}$ single mutants. There were no statistically significant differences between group means as determined by one-way ANOVA (F (2,15) = 0.45, p = .65). ns = not significant.

	Total silique (n)	*F-statistic*	*P-value*	*F-critical*	Significance
Wild type	6				
$DAZ3^{-/-}$	6	0.45	0.65	3.68	ns
$DAZ3L^{-/-}$	6				

The seeds for $DAZ3^{-/-}$ and $DAZ3L^{-/-}$ were also examined for their germination potential. They were surface sterilized and grown on MS0 media (n = 45) for 3 days to investigate their ability to produce germinating seedlings. All seeds for both $DAZ3^{-/-}$ and $DAZ3L^{-/-}$ were able to germinate and produce seedlings (Figure 3.4.3.2). The size of the seedlings for both $DAZ3^{-/-}$ and $DAZ3L^{-/-}$ were also similar to the wild type.

Taken together, these data indicate that both *daz3* and *daz3l* single mutant plants were fertile and can produce viable progeny that could indicate potential functional redundancy.

Figure. 3.4.3.2. Germination test for wild type, *DAZ3*$^{-/-}$ and *DAZ3L*$^{-/-}$ mutant seeds. Wild type and single mutant seeds were grown on MS0 media for three days under constant light. Scale bar = 5 mm.

3.4.4 Generation of *daz3 daz3l* double knockout mutants

The single homozygous mutant lines *daz3* (B2_P5_P3E_P4E), and *daz3l* (B2_P14_P7), described in section 3.4.1, were used to generate double knockout lines and to study the role of *DAZ3* and *DAZ3L* in Arabidopsis sperm cell development. These lines harbour the knockout alleles that are referred to as *daz3-x* and *daz3l-y*, respectively. Observation of *daz3* and *daz3l* single mutant plants revealed that both were fertile, which is consistent with their potential functional redundancy if *DAZ3/DAZ3L* are required for fertility. Single mutants were transformed with the complementary CRISPR/Cas9 construct to generate *daz3 daz3l* double mutants. Transformed plants were screened by PCR for the presence of insertions or deletions (indels) at the regions targeted by the paired RNA guides for each construct and lines with candidate mutations were then confirmed by sequencing.

In the T1 generation of homozygous *daz3-x* plants transformed with the CRISPR construct *DAZ3L* sg25+66, 43 plants were screened for mutation events in *DAZ3L*. One plant, 4F, was found to be biallelic and produced two *DAZ3L* amplicons, one larger and one similar in size to that of the wild type (Table 3.4.4.1).

For *daz3l-y* plants transformed with the CRISPR construct *DAZ3* sg8+69, two out of 217 plants generated a smaller *DAZ3* amplicon. These two plants, 12H and 18E, were found to be homozygous for a 22 bp deletion in *DAZ3* (Table 3.4.4.1; Figure 3.4.4.1) resulting in an early stop codon and a predicted truncated DAZ3 protein (Figure 3.4.4.2; Appendix Figure S3.2). The *daz3* deletion allele sequenced from plant 12H resulting from transformation of *daz3l-y* plants with *DAZ3* sg8+69 is identical to *daz3-x* allele. Hence, plant 12H, also called *daz3-x' daz3l-y*, was confirmed as a homozygous *daz3 daz3l* double mutant and chosen for double mutant phenotypic analysis.

Plant 12H and 18E of *daz3l-ydaz3* homozygous double mutant lines were screened for T-DNA-free T2 plants (Table 3.4.4.2). A total of 100 and 50 T2 mutant plants were screened for plant 12H and 18E, respectively. From these, 46 lines were confirmed to be T-DNA-free for plant 12H, whereas 10 lines were T-DNA-free for plant 18E. Some of the lines were shown in Table 3.4.4.2.

Table 3.4.4.1. Screening T1 generation plants to generate *daz3 daz3l* double mutants. T1 plants were screened for the presence of indels by PCR- based genotyping. Amplicons that were larger or smaller than those of the wildtype locus were isolated and sequenced to confirm the mutations. * = not send for sequencing.

Mutant line	No. of T1 plants screened	Locus amplified	Plant	Type of mutation
daz3-x transformed with *DAZ3L* sg25+66	43	*DAZ3L*	4F	*Biallelic insertion
daz3l-y transformed with *DAZ3* sg8+69	217	*DAZ3*	12H	Homozygous deletion
			18E	Homozygous deletion
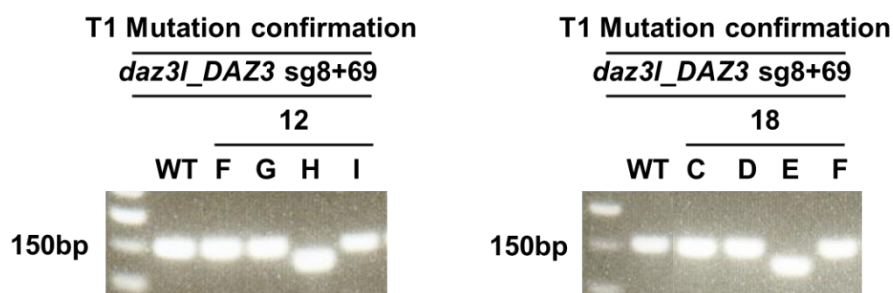
Figure 3.4.4.1. Genotyping of T1 generation of *daz3l* transformed with CRISPR construct *DAZ3* sg8+69. Plant 12H and 18E produce single band of smaller size than the wildtype plants, indicating a deletion mutation.

Sequencing direction
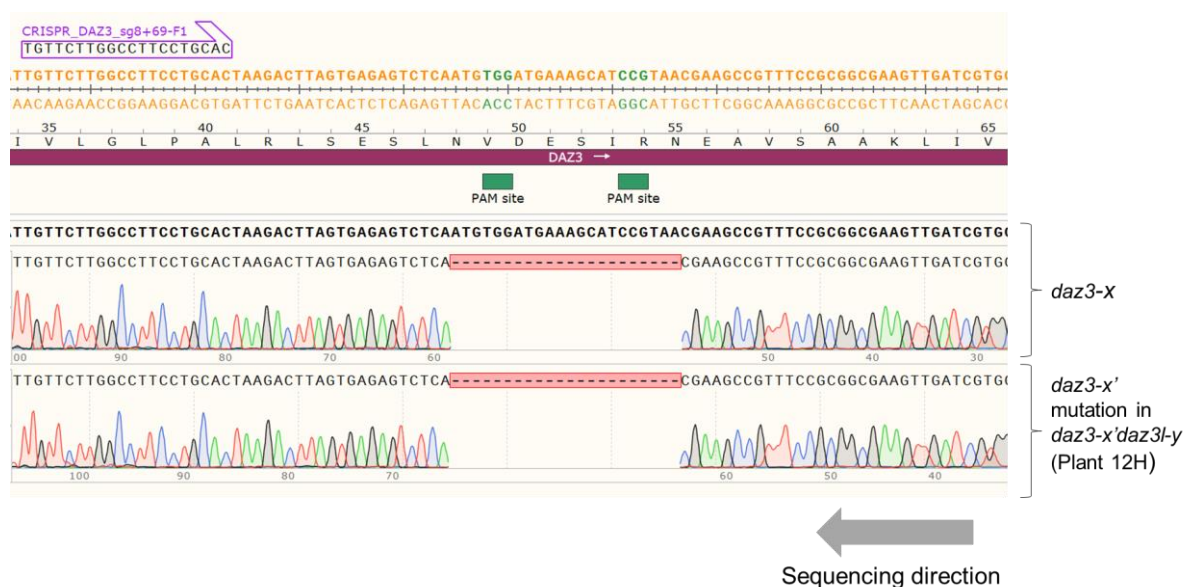
Figure 3.4.4.2 Chromatogram showing the DNA sequence of the *DAZ3* gene in *daz3-x'daz3l-y* double mutants. PCR products from plant 12H of *daz3l-y* transformed with CRISPR *DAZ3* sg 8+69 were sequenced and confirmed to be homozygous for 22 bp *daz3* deletion.

Table 3.4.4.2. *daz3 daz3l* **non-transgenic mutant lines confirmation in T2 generation.** T2 mutant plants was screened for the presence of T-DNA free stable mutant line.

Line (T1)	Total screen	T2		
		T-DNA free	Lines	Genotype
*daz3l-ydaz3_*12H	100	46	1F	Homozygous
			14A	Homozygous
*daz3l-ydaz3_*18E	50	10	2E	Homozygous

3.4.5 Phenotypic and genetic analysis of *daz3 daz3l* double mutants

To investigate potentially redundant function of *DAZ3* and *DAZ3L*, *DAZ3^{-/-} DAZ3L^{-/-}* double mutant plants were grown for seven weeks to observe any reduction in fertilization potential and seed set compared to wild type plants. Ten siliques were collected from the primary inflorescence of three different plants, starting from silique four apically from the base of the inflorescence. Silique lengths were measured and 5 siliques were dissected to observe seed morphology as well as to calculate seed set. Similar to phenotype observed in wild type plants and single mutants, two abnormal categories were observed in *DAZ3^{-/-} DAZ3L^{-/-}* mutant plants, undeveloped ovules and aborted seeds (Figure 3.4.5.1).

Data collected were analysed in excel to compare the silique length and the percentage of viable seeds per silique in *DAZ3^{-/-} DAZ3L^{-/-}* versus wild type. The average length of the silique for the wild type is 15.5 mm and *DAZ3^{-/-} DAZ3L^{-/-}* is 16.3 mm (Figure 3.4.5.2). *DAZ3^{-/-} DAZ3L^{-/-}* had an average of 91 % viable seed compared to the wild type with 85 % (Figure 3.4.5.3). One-way ANOVA analysis was conducted to test for differences between the mean silique length of *DAZ3^{-/-} DAZ3L^{-/-}* and wild type (Table 3.4.5.1; Appendix Table S3.4) and for examining differences in the mean percentage of viable seeds per silique between the double mutant and the wild type (Table 3.4.5.2; Appendix Table S3.5). The p-value for both tests is more than 0.05 critical value, signifying that the mean silique length and the mean percentage of viable seeds per silique for *DAZ3^{-/-} DAZ3L^{-/-}* is not significantly different from the wild type (mean silique length: (F (1,58) = 3.60, p = .06) (mean percentage: (F (1,28) = 0.95, p = .34)).
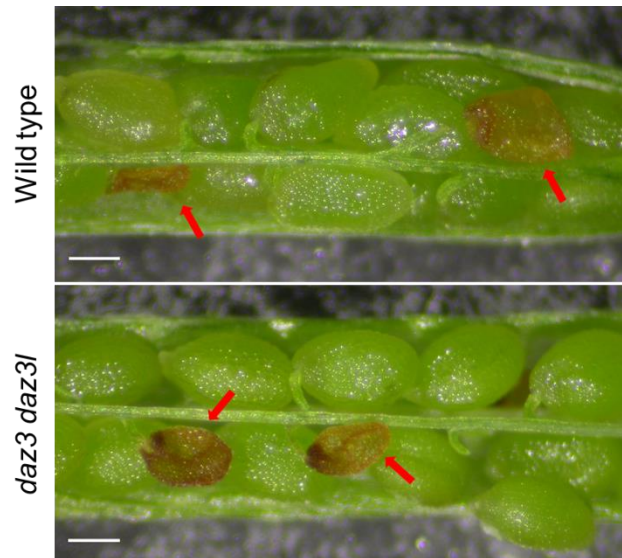
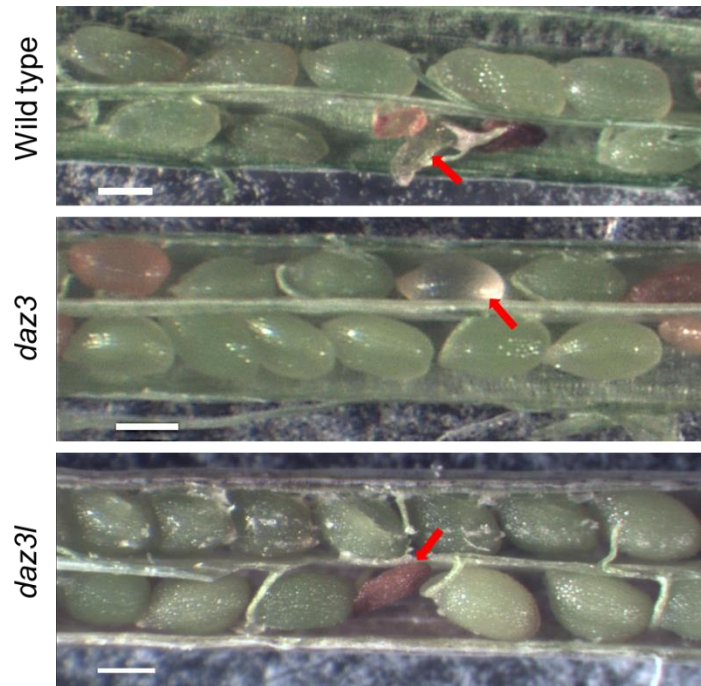**Figure 3.4.5.1. Phenotype of seeds produced by *daz3 daz3l* mutants compared to those from wild type plants.** Viable seeds were green, rounded and plump. Meanwhile, non-viable seeds were brownish and shrunken. Scale bar = 0.3 mm.
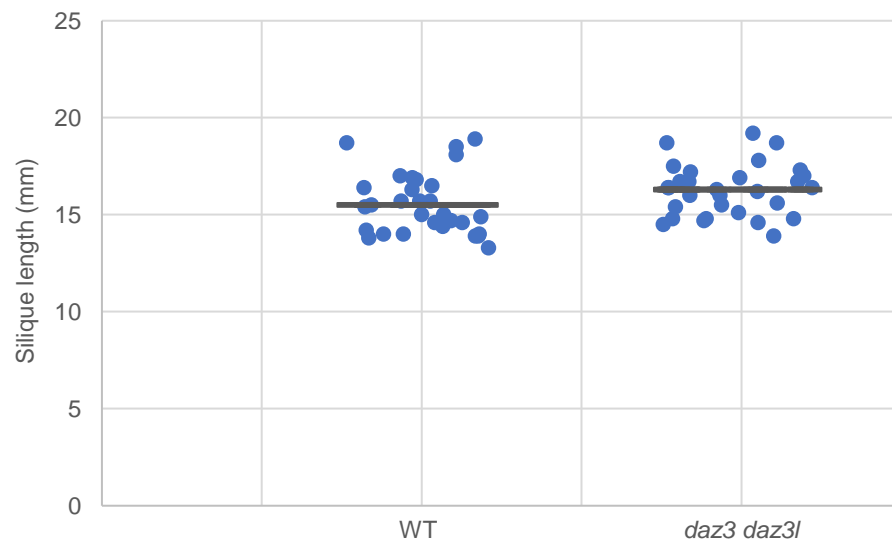
Figure 3.4.5.2. Silique length in wild type and homozygous *daz3 daz3l* double mutant plants. 10 siliques were collected from the primary branch of three different plants for both wild type and *DAZ3$^{-/-}$ DAZ3L$^{-/-}$* mutant plants. Their length was measured; the average silique length for wild type plant is 15.5 mm whereas for *DAZ3$^{-/-}$ DAZ3L$^{-/-}$* is 16.3 mm.

Table 3.4.5.1. One-way ANOVA for analysing any significant difference between the mean silique length of *DAZ3⁻ᐟ⁻ DAZ3L⁻ᐟ⁻* and wild type. There were no statistically significant differences between group means as determined by one-way ANOVA (F (1,58) = 3.60, p = .06). ns = not significant.

	Total silique (n)	*F-statistic*	*P-value*	*F-critical*	Significance
Wild type	30				
DAZ3⁻ᐟ⁻ DAZ3L⁻ᐟ⁻	30	3.60	0.06	4.01	ns

Percentage of viable seeds per silique
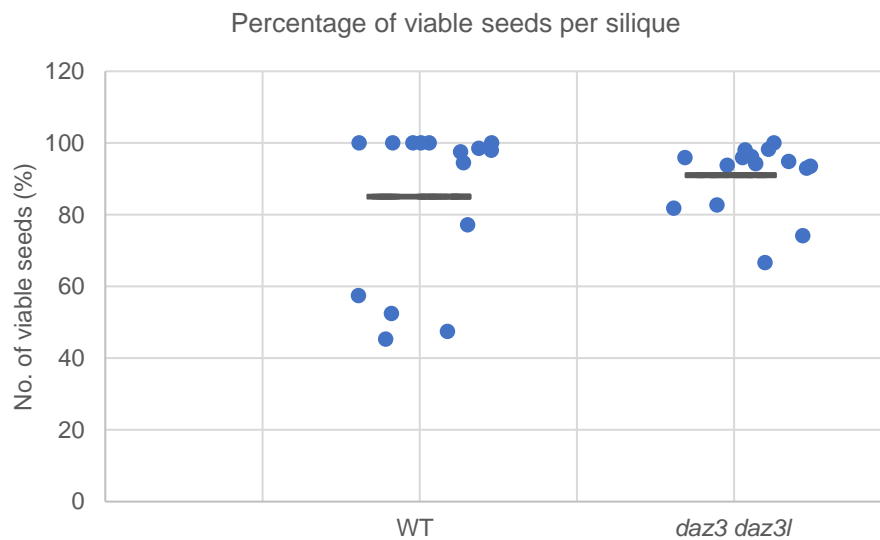
Figure 3.4.5.3. Scatter plot showing the percentage of viable seeds per silique in wild type and homozygous *daz3 daz3l* double mutant plants. 5 siliques of all three plants, for both wild type and *DAZ3⁻ᐟ⁻ DAZ3L⁻ᐟ⁻* mutant, were calculated for their number of viable seeds. Then, the percentage of viable seeds per silique were determined and plotted on the graph (n total = 15).

Table 3.4.5.2. One-way ANOVA for analysing any significant difference in the mean percentage of viable seeds per silique in wild type and *DAZ3^-/- DAZ3L^-/-* mutant. There were no statistically significant differences between group means as determined by one-way ANOVA ($F (1,28) = 0.95$, $p = .34$). ns = not significant.

	Total silique (n)	*F-statistic*	*P-value*	*F-critical*	Significance
Wild type	15				
DAZ3^-/- DAZ3L^-/-	15	0.95	0.34	4.2	ns

To test the germination potential, seeds for *DAZ3^-/- DAZ3L^-/-* was sown on MS0 agar (n=45) for three days under constant light. Similar to wild type, *DAZ3^-/- DAZ3L^-/-* double mutant seeds were able to germinate and produce healthy seedlings (Figure 3.4.5.4).
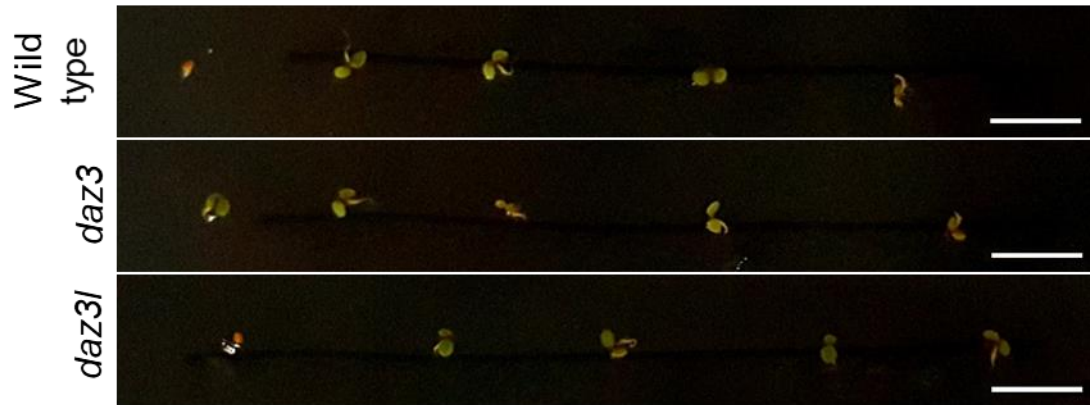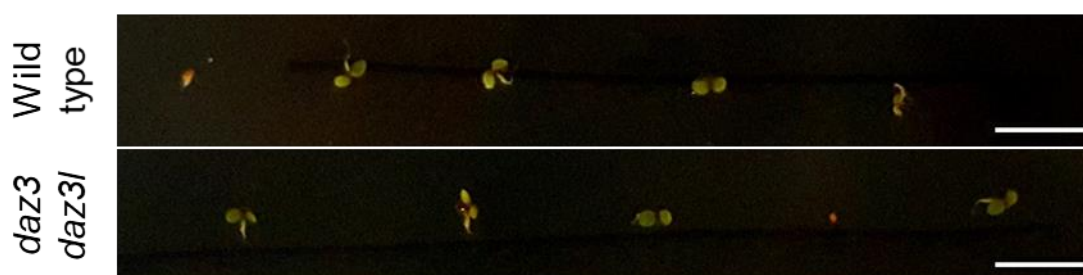

Figure 3.4.5.4. Germination test for wild type and *DAZ3^-/- DAZ3L^-/-* mutant seeds. Wild type and double mutant seeds were grown on MS0 media for three days under constant light. Scale bar = 5 mm.

From the experiment conducted, it clearly shown that *DAZ3^-/- DAZ3L^-/-* double mutant was able to produce viable seeds capable of germination. Therefore, it could be concluded that similar to the single mutants, *daz3 daz3l* double mutant was also fertile and can produce viable progeny.

3.5 Discussion

3.5.1 DAZ3 conservation in angiosperms

The *Arabidopsis thaliana* C_2H_2 zinc finger protein DAZ3 possesses one K2-2 zinc finger domain and two DLNxxP type EAR motifs. In this chapter, sequence analysis was conducted to examine the presence of DAZ3 in the genomes of other angiosperms. It has been found that orthologs of DAZ3 are only present in eudicots. The majority are found in the superrosid clade, which includes Saxifragales and rosids with fewer present in the asterid clade.

The C_2H_2 K2-2 zinc finger motif of DAZ3 is KAVFGH, which is not the main variant of the K2-2 motif (KALFGH) present in Englbrecht et al. (2004). Sequence conservation of DAZ3 homologs shows that the KAVFGH motif variant first appears in the rosid clade, becoming more prominent in Malvids and is highly conserved in Brassicaceae. The sequence conservation also highlights pattern changes in the DAZ3 K2-2 motif, from KALFGH to KAVFGH.

The difference in the residue at position 4 of the K2-2 DNA binding motif, L to V, shows that this position is prone to change. Previous studies on DNA binding specificities have found that both DAZ1, with KALFGH motif, and DAZ3L bind to the same AGCT motif (Franco-Zorilla et al., 2014; Wang et al., 2020). This suggests that the L to V change does not alter binding specificity and as L to V is a conservative substitution this could be the reason why KAVFGH motif is maintained in DAZ3 orthologues.

An acidic region is conserved near the C-terminal end of all DAZ3 homologs. Besides that, an additional acidic region is also present near the N-terminal end in the Malvids clade, particularly in Brassicaceae. Acidic regions are commonly found in a number of nuclear proteins, such as those involved in transcriptional activation. An acidic region at the N-terminal end of Arabidopsis *ZIM* gene for instance is found to function as transcriptional activator in a transient GAL4 fusion assay (Shikata et al., 2003). The same function was seen in *VP1* gene of maize, where the acidic domain functions as a transcriptional activator region (McCarty et al., 1991). Therefore, this region might serve as an activation domain and could potentially make DAZ3 as a transcriptional activator (Shikata et al., 2003).

A basic region, located before the zinc finger domain, is conserved in all DAZ3 homologs. This short stretch of basic amino acids has the monopartite nuclear localizing signal (NLS) consensus K(K/R)X(K/R) and could serve as an NLS (Chelsky et al., 1989; Lange et al., 2007). The ability of the basic region of DAZ1 to support nuclear protein import

was demonstrated by Darbar (2019). In this study, both DAZ1-GFP protein with mutated basic region and DAZ1-GFP protein with deleted basic region was unable to localise in the nucleus and appeared in the cytoplasm.

The presence of EAR motifs in DAZ3 and its homologs imply that they might be involved in transcriptional repression. EAR motifs are required for the ability of DAZ1/DAZ2 to act as transcriptional repressors and there is physical interaction between DAZ1/DAZ2 with the co-repressor TOPLESS (TPL) (Borg et al., 2014). Due to the presence of two EAR motifs, this suggests that DAZ3/DAZ3L may interact with TPL-family proteins. Additionally, the variation in number of residues separating DLN and P and reduction in EAR motifs of DAZ3 orthologs indicate the possibility of motifs drifting due to relaxed selection on the EAR motifs.

3.5.2 DAZ3L forms a distinct clade

Based on sequence analysis, DAZ3 is shown to be present and conserved in various eudicot species. Through cladistic analysis, DAZ3L was shown to form a distinct clade which is restricted to Brassicaceae. DAZ3L clade is distinguished from the DAZ3 by the presence of an additional basic region just before the conserved basic region in both proteins (Figure 3.2.1.3)

The analyses show that DAZ3 and DAZ3L clades evolved from a duplication event involving *DAZ3*. The most probable reason could be through proximal duplication of *DAZ3* based on the same chromosomal location and the proximity of the genes in all observed species (Qiao et al., 2019). The event therefore may have occurred before the divergence of Brassicaceae approximately 43.2 MYA (Panchy et al., 2016; Yu et al., 2017)

The additional basic region observed in DAZ3L clade suggest that there is a slight difference between the two genes. This could be seen from the difference in transcript expression level and the location of the two proteins (Taimur, 2014). The additional basic region in DAZ3L may improve nuclear localisation of DAZ3L compared with DAZ3, therefore this could account for the different in their protein localisation.

3.5.3 *DAZ3* is evolved from *DAZ1/DAZ2*

A deeper analysis was performed to search for more distant homologs of DAZ3 in all known angiosperms in an effort to trace the origin of DAZ3. DAZ1 and DAZ2 sequence

were present among these more distant homologs. Multiple sequence analysis of DAZ3/DAZ3L and DAZ1/DAZ2 shows that the signature features of DAZ3/DAZ3L, the K2-2 ZnF domain and EAR motifs, are also present in DAZ1/DAZ2. However, DAZ1 and DAZ2 also have two additional zinc finger domains after the K2-2 domain. Cladistic analysis of the homologs coding sequence shows distinct clades for DAZ3/DAZ3L and DAZ1/DAZ2. The analysis reveals that DAZ1/DAZ2 are present throughout angiosperms including basal species, while the DAZ3/DAZ3L clade only appears in superrosids and some asterids.

The evidence from sequence and cladistic analysis indicate that *DAZ3*/*DAZ3L* sequence could have evolved from *DAZ1*/*DAZ2* after the divergence of core eudicots. This could happen during the genome triplication event in eudicots before the divergence of superrosid-superasterid (Jiao et al., 2012; Panchy et al., 2016; Qiao et al., 2019). *DAZ3* may have emerged as a result of genome triplication involving *DAZ1*/*DAZ2* followed by gene loss that resulted in *DAZ3* as a single copy as seen in some species of Malpighiales and Fabids (De Smet et al., 2013). In addition, mutational events such as deletion, insertion and substitution could happen over the time course (Panchy et al., 2016). This resulted in the loss of two additional C_2H_2 zinc finger of DAZ3 and the substitution of amino acid L to V in the K2-2 motif. This could happen before the Fabids-Malvids divergence, hence explaining the occurrence of KAVFGH in these clades. Another explanation for the origin of *DAZ3* could be that it was the result of proximal or segmental duplication of *DAZ2* in the common ancestor of superrosid and superasterid (Qiao et al., 2019). This could explain why *DAZ2* and *DAZ3* are located on the same chromosome and in close proximity in *A. thaliana*.

The evolution of *DAZ3* from *DAZ1*/*DAZ2* suggests a new function evolved for the DAZ3 clade that is restricted to eudicots and could contribute to fitness with regard to biotic or abiotic stress responses.

3.5.4. *daz3 daz3l* double knockout mutants have no apparent phenotype

DAZ3 and *DAZ3L* are both shown in RNA-seq data to be highly expressed in sperm and this abundant expression continues in pollen tubes. In previous studies DAZ3 and DAZ3L proteins are predominantly located in the cytoplasm when linked to a reporter protein and this pattern is maintained even when the sperm cells are present within growing pollen tubes (Taimur, 2014). This suggest that these two proteins may have a role in the sperm development or in events taking place before or after fertilisation.

The fact *DAZ3L* and *DAZ3* are paralogs and have similar expression patterns suggests that these two genes may have redundant functions. Hence, a series of experiments were conducted to attempt to uncover the function of *DAZ3* and *DAZ3L*. First, CRISPR/Cas9 derived knock-out mutant plants were generated to produce T-DNA-free *daz3* and *daz3l* single mutant lines. Genetic transmission analysis showed that the mutant alleles have normal transmission and knockout mutations in *daz3* and *daz3l* have no effect on the success of the male and female gametes or plant fertility.

Dissected siliques of *daz3* and *daz3l* mutants showed green viable seeds, due to embryo greening, which is a product of successful fertilisation (ten Hove et al., 2015; O'Neill et al., 2019). The number of viable seeds were unaffected in *daz3* and *daz3l* mutants and comparable to the wild type. In addition, both *daz3* and *daz3l* mutant seeds were able to germinate normally in vitro and in soil. These lines of evidence support the hypothesis that *DAZ3* and *DAZ3L* are functionally redundant as the loss of one gene does not have an impact on the development of the mutant plant. Functional redundancy has been demonstrated in many studies and is clearly shown for the role of *DAZ1* and *DAZ2* in male gamete development (Borg et al., 2014). Therefore, the double knockout of *DAZ3* and *DAZ3L* would help to elucidate their potentially redundant functions encompassing several aspects of plant reproductive development, including sperm cell development and function in fertilisation. Defective *daz3 daz3l* male gametes would be expected to prevent or reduce male transmission, fertilisation and seed set.

daz3 daz3l double mutant plants were generated via CRISPR/Cas9 induced mutations. Homozygous mutations were produced in the T1 generation; therefore, phenotypic analysis was able to be conducted early. When *daz3 daz3l* double mutant siliques were dissected, green viable seeds were seen, a phenotype comparable to the wild type. The number of viable seeds were not affected and the double mutant seeds were able to germinate normally. From these experiments, it has been confirmed that *DAZ3* and *DAZ3L* have no essential function in the development of sperm cells. The ability of *daz3 daz3l* double mutant to produce viable seeds shows that they also do not have essential role in fertilization or post-fertilisation events including seed development.

The presence of two EAR motifs may suggest a potential interaction between DAZ3/DAZ3L and TPL. However, the fact that DAZ3/DAZ3L and TPL exist in different locations, cytoplasm and nucleus respectively could suggest otherwise (Taimur, 2014). Nonetheless, since TPL is synthesised in the cytoplasm, there is a possibility that they may bind to TPL or to TPL-family members, potentially limiting the accumulation of TPL-related proteins in the nucleus at some stage during sperm development or post-

fertilisation. In this sense, DAZ3 and DAZ3L could be seen as TPL regulators in the cytoplasm. It also can be suggested that a proportion of DAZ3 and DAZ3L together with TPL could be transported to the nucleus to modify the chromatin structure of regulatory regions through DNA binding involving the single zinc finger domain.

DAZ3 and *DAZ3L* could be important in plant biological processes such as environmental stress and plant hormone function. Many C_2H_2 zinc finger proteins have been reported to have a major role in abiotic and biotic stress response (Han et al., 2020). One such example is seen for *MaC2H2s* genes where they are involved in the cold stress response of banana fruit (Han and Fu, 2019). MaC2H2s were found to repress the transcription of *MaICE1*, a key component in the cold signaling pathway through dual-luciferase reporter assays. Additionally, they were also found to regulate ethylene production during banana fruit ripening by binding to the promoters of ethylene biosynthetic genes, *MaACS1* and *MaACO1*, and repressed their activity (Han et al., 2016).

Chapter 4: Analysis of DUO1 function in *Physcomitrella patens*

Abstract

Background and Aims

DUO POLLEN 1 (*DUO1*) is widely conserved in land plants, yet little is known about its functional conservation in extant representatives of the bryophytes. Recent work in the model liverwort *Marchantia polymorpha* has shown that *MpDUO1* has sperm cell-specific expression and is essential for sperm cell differentiation. This chapter describes functional conservation of DUO1 in the model moss *Physcomitrella patens*.

Methods

To study the conservation of DUO1 in bryophytes, sequence analysis was performed on DUO1 and its orthologs. Expression analysis using publicly available RNA-sequence data was conducted to investigate the expression profile of *PpDUO1* (*PpDUO1A* and *PpDUO1B*) genes. To conduct a functional analysis of *DUO1* in *P. patens*, single and double knockout mutants were generated and sporophyte formation examined. DAPI staining was performed on antheridia to analyse the development of spermatogenous cell nuclei and to enable spermatogenous cell counting. TEM was conducted for ultrastructural analysis of spermatogenous cells.

Key Results

Several conserved regions were found in the C-terminal region of bryophyte DUO1 orthologs that are absent from angiosperms. These regions could serve as bryophyte-specific transactivation domains. *PpDUO1A* and *PpDUO1B* show antheridia-specific and -preferential expression, respectively. *Ppduo1a$^\Delta$b$^\Delta$* double mutants were unable to form sporophytes unlike the wild type and the *Ppduo1a$^\Delta$* and *Ppduo1b$^\Delta$* single mutants. Spermatogenous cell nuclei of *Ppduo1a$^\Delta$b$^\Delta$* failed to show morphogenesis from the rounded nucleus to the crescent-shaped nucleus stage. *Ppduo1a$^\Delta$b$^\Delta$* spermatogenous cells also failed to show flagellar formation. The number of spermatogenous cells however was comparable to the wild type, showing that the sperm cell division is not affected in *Ppduo1a$^\Delta$b$^\Delta$*.

Conclusion

The evidence presented demonstrates that *PpDUO1* genes have an essential function in spermatogenous cell development. Their critical role in sperm cell morphogenesis and differentiation shows that the function of *DUO1* is largely conserved in *P. patens*.

4.1 Introduction

Early land plants such as bryophytes produce motile sperm in contrast to flowering plants that produce non-motile gametes. Different modes of spermatogenesis in the land plant lineage (embryophytes) may be associated with diversification in ancestral gene regulatory networks during land plant radiation (Higo et al., 2018). This raises the question of the conservation of germline development among land plants and whether sperm cell differentiation mechanisms share a common molecular origin.

Arabidopsis DUO POLLEN 1 (DUO1) is a male germline-specific MYB transcription factor that functions to control male germ cell division and differentiation (Rotman et al., 2005; Durbarry et al., 2005; Brownfield et al., 2009). Several genes are known to be part of the DUO1 regulatory network, such as the germline-specific genes DUO1 ACTIVATED ZINC FINGER 1 (DAZ1), DUO1 ACTIVATED ZINC FINGER 2 (DAZ2), GAMETE-EXPRESSED 2 (GEX2) and GENERATIVE CELL SPECIFIC 1 (GSC1/HAP2) (Borg et al., 2011; Brownfield et al., 2009; Mori et al., 2006; von Besser et al., 2006). DUO1 together with DAZ1 and DAZ2 regulate the expression of GEX2 and GCS1/HAP2, which function in gamete fusion and attachment during fertilization (Brownfield et al., 2009; Mori et al., 2006; von Besser et al., 2006).

A recent study has explored the expression and function of the single DUO1 homolog in *Marchantia polymorpha*, *MpDUO1* (Higo et al., 2018). MpDUO1 has sperm cell-specific expression and controls the expression of the *AtDAZ1* orthologues, *MpDAZ1*, but the expression of *MpGEX2* and *MpGCS1/HAP2* are not affected in *Mpduo1* mutants and so do not appear to be under MpDUO1 control (Higo et al., 2018). This shows that DUO1 function is conserved in *Marchantia* to some extent compared with its function in Arabidopsis. Given that bryophytes do not form a monophyletic clade (Shaw and Renzaglia, 2004) different bryophyte lineages may show functional differences in molecular mechanisms of male reproduction, which may include the role of DUO1.

A pair of DUO1 homologs, *PpDUO1A* and *PpDUO1B*, were first identified in *Physcomitrella patens* (Brownfield et al., 2009). The expression of these genes was first studied using RT-PCR by Sari (2015). *PpDUO1A* was shown to be expressed in reproductive and vegetative tissue whereas *PpDUO1B* was only expressed in reproductive tissues. Due to *PpDUO1B* expression being more restricted to reproductive tissue, the nomenclature used in Sari (2015) was different to that used in this chapter, such that the names of *PpDUO1A* (Gene ID: Pp3c8_16720 / Pp1s114_136V6) and *PpDUO1B* (Gene ID: Pp3c24_11770 / Pp1s16_281V6) have been swapped. In this

chapter, and the rest of the thesis, *PpDUO1A* is a gene with ID: Pp3c24_11770 / Pp1s16_281V6, while *PpDUO1B* is a gene with ID: Pp3c8_16720 / Pp1s114_136V6.

In the study by Sari (2015) functional complementation was carried out by transforming PpDUO1A-mCherry and PpDUO1B-mCherry gene constructs driven by the *DUO1* promoter in the heterozygous *duo1* mutant background of *Arabidopsis thaliana* (Sari, 2015). The PpDUO1A-mCherry fusion (PpDUO1B in this chapter) was not expressed, whereas the PpDUO1B-mCherry fusion (PpDUO1A in this chapter) was expressed in only some plants. Both PpDUO1A- and PpDUO1B-mCherry fusion constructs were unable to rescue failed germ cell division in *duo1* pollen.

The ability of PpDUO1A and PpDUO1B to bind and activate DUO1 target genes was also tested by using transient expression assays in tobacco plants (Sari, 2015). PpDUO1A and PpDUO1B expression were driven by the CaMV 35S promoter (effector construct) and the experiment was designed to test the transactivation of the target gene promoter, *HTR10* (also known as *MGH3*), that was fused to luciferase (reporter construct) (Sari, 2015). The luciferase expression in the presence of PpDUO1A and PpDUO1B effector constructs were not significantly different to the luciferase expression of the no effector negative control. The result of these assays indicated that neither PpDUO1A nor PpDUO1B were able to transactivate the *DUO1* promoter.

A further study of PpDUO1 activity investigated the ability of the PpDUO1 MYB domain to transactivate (& presumably bind) the *MGH3/HTR10* promoter (Zhao, 2017). A PpDUO1 chimeric protein, termed "PpChimera", was designed by combining the PpDUO1B N-terminal region which contains the MYB domain with the C-terminal region of AtDUO1. In transient expression assays in tobacco leaves it was shown that PpChimera was able to transactivate the target, *MGH3/HTR10* promoter. This illustrates the ability of PpDUO1B MYB domain to recognise and bind to the same target promoter sequence as DUO1. The difference in transactivation ability of AtDUO1 and PpDUO1 was mainly attributed to the differences in the C-terminal region of the proteins.

In this chapter, the conservation of DUO1 in bryophytes is investigated. This includes the conservation of protein and gene sequences, expression patterns and functional conservation in *Physcomitrella patens*.

4.2 DUO1 sequences in bryophytes

4.2.1 Sequence analysis of DUO1 homologs in bryophytes

To further examine sequence conservation of DUO1 in extant representatives of the bryophytes, the full-length protein sequence of MpDUO1 was obtained from GenBank (accession: LC172177) and used in BLASTP searches. The purpose of the analysis was to identify any novel DUO1 homologous sequences and also to further examine conserved regions, particularly those uniquely conserved in bryophytes.

The same pipeline described in chapter 3.2.2 (Figure 3.2.2.1) was applied and the results enabled the retrieval of candidate sequences of MpDUO1 orthologs from all land plants with sequenced genomes. Since MpDUO1 is an R2R3 MYB transcription factor, candidate sequences were analysed for the presence of R2 and R3 repeats as well as the signature lysine residue (K66 in AtDUO1), typical of the AtMYB125/DUO1 subfamily in clade 9 and sequences that did not contain an R2R3 MYB domain were removed (Qing et al., 2018).

The list of 90 sequences from various land plant species included PpDUO1A and PpDUO1B from *Physcomitrella patens*. The PpDUO1A sequence from public databases however was shorter and incomplete compared to the PpDUO1A sequence that was previously reported in Sari (2015) and Zhao (2017). Therefore, the PpDUO1A sequence from Sari and Zhao was used and incorporated into the ortholog list. The reliability of the multiple sequence alignments was checked by calculating the average amino acid identity in pairwise comparison. The result was 80 % identity which was far above the 30 % threshold (Rost, 1999). A cut-off of roughly 30% identity signifies that 90% of the pairs were homologous (Rost, 1999). Protein sequences of MpDUO1 orthologs were then exported into CLC Sequence Viewer (QIAGEN Bioinformatics) to generate a superior graphic image. Representative sequences were selected across major land plant clades and an alignment of these is presented in Figure 4.2.1 (also Appendix Figure S4.1).

Through the DUO1 homolog sequence search, a new ortholog was found for the bryophyte group. This is SfaDUO1A of *Sphagnum fallax* which was not found in previous study. The presence of the typical R2R3 MYB domain in PpDUO1A and PpDUO1B indicated functional conservation, similar to that for MpDUO1 (Higo et al., 2018). In addition, a conserved region (CR) consisting of mainly polar amino acids was present in both PpDUO1 proteins and conserved among all bryophyte DUO1 sequences. Two acidic rich regions are also present after this CR in all bryophyte DUO1 orthologs. A basic stretch of amino acids is present after the CR of PpDUO1A and PpDUO1B with

the consensus 'KR/QRPRSRR'. This basic region is also found in MpDUO1 and SfaDUO1B and could serve as a nuclear localization signal (NLS).
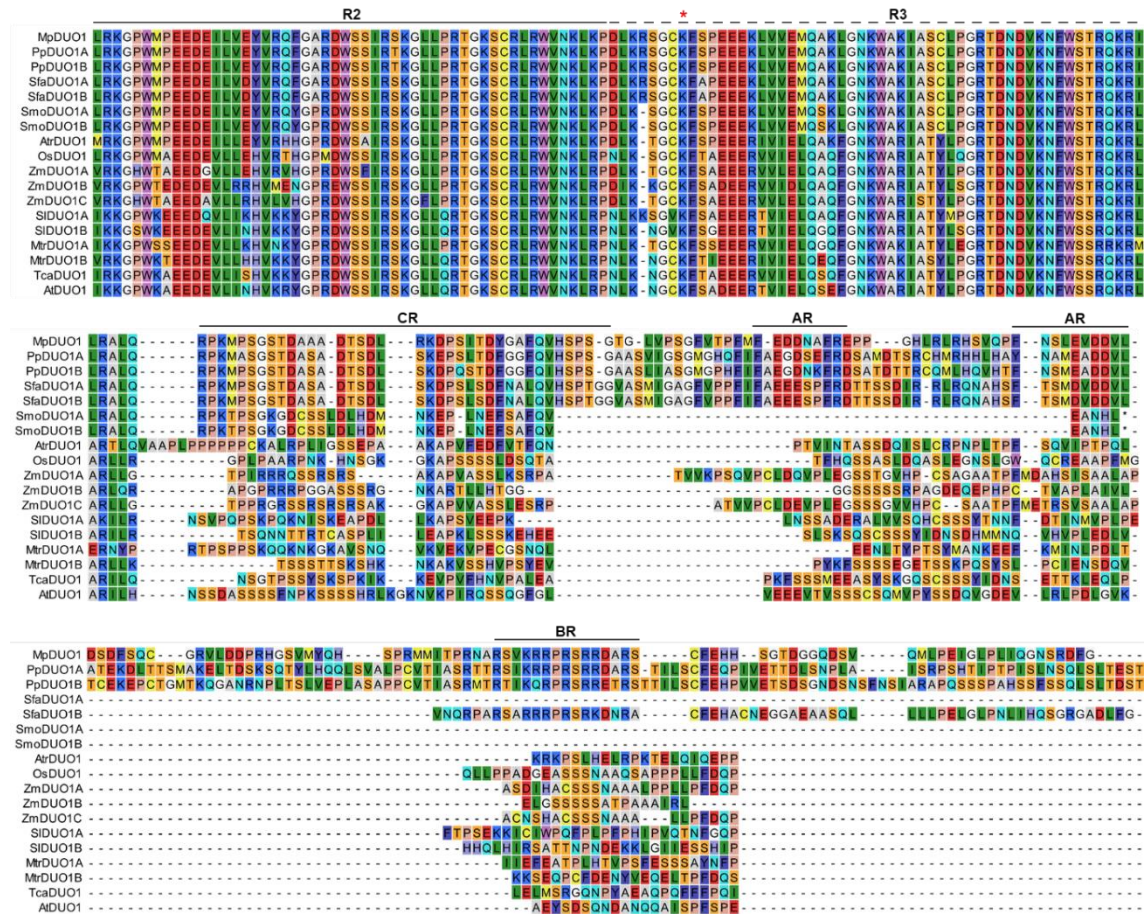
Figure 4.2.1. Part of a multiple sequence alignment of DUO1 homologs in evolutionary divergent species of major land plant clades. Proteins sequences were aligned with MUSCLE using default parameter in MEGA 10 and viewed in CLC Sequence Viewer. Signature lysine (K) residue is shown by the red arrow star. CR, conserve region; AR, acidic rich region; BR, basic region. Species: At, *Arabidopsis thaliana*; Tca, *Theobroma cacao*; Mtr, *Medicago truncatula*; Sl, *Solanum lycopersicum*; Os, *Oryza sativa*; Zm, *Zea mays*; Atr, *Amborella trichopoda*; Smo, *Selaginella moellendorffii*; Sfa, *Sphagnum fallax*; Pp, *Physcomitrella patens*; Mp, *Marchantia polymorpha*.

4.2.2 Structure and characterisation of *PpDUO1*

To further describe *PpDUO1* genes, characteristics such as chromosomal location, position and orientation and protein percentage identity are discussed. *PpDUO1A* and *PpDUO1B* proteins are 429 and 490 amino acids in length, respectively. They are located on chromosome 24 in forward orientation for *PpDUO1A* and chromosome 8 in reverse orientation for *PpDUO1B*. Amino acid pairwise comparison computed using MEGA 10 showed that they have a high percentage amino acid identity of 87 %. The percentage identity for PpDUO1A and PpDUO1B are also high when both proteins were compared to MpDUO1 at 78 % and 79 %, respectively. On the other hand, the protein identity drops to 36 % for both PpDUO1 when they were compared with AtDUO1 (Table 4.2.2.).

Table 4.2.2. Protein percentage identity of PpDUO1A, PpDUO1B, MpDUO1 and AtDUO1. Amino acids pairwise comparison was made using MEGA 10 to calculate the proportional (*p*) distance between two proteins. Protein identity = 1 – *p*-distance.

Protein percentage identity (%)			
	MpDUO1	PpDUO1A	PpDUO1B	AtDUO1
MpDUO1				
PpDUO1A	78			
PpDUO1B	79	87		
AtDUO1	37	36	36	

4.2.3 Expression of *PpDUO1* in *P. patens*

In order to investigate the functional conservation of DUO1 in *P. patens*, it is important to understand the expression pattern of the two paralogous genes, *PpDUO1A* and *PpDUO1B*. The CoNekT platform was utilised to obtain RNA-seq data on transcript expression and to simultaneously display the expression profiles of both genes for a direct comparison (Proost and Mutwil, 2018; Perroud et al., 2018; Julca et al., 2021). The platform also allows downloading of normalised RNA-seq data, which herein has been used to construct a graphical expression profile across different developmental stages and various tissues (Figure 4.2.3). The data set used in the platform was obtained from Gene Atlas dataset by Perroud et al. (2018) which comprises of 99 sequenced libraries of different *P. patens* developmental stages. The Gene Atlas data set was primarily

aimed to establish overall transcriptomic reference dataset for *P. patens*. In addition, the gene expression atlas by Julca et al. (2021) was also utilised in the platform. This atlas includes 38 sequenced libraries pertaining *P. patens* with the aim to provide a gene expression atlas for various organs and gametes of plant species including *P. patens* (Julca et al., 2021).

The dominant gametophyte stage in *Physcomitrella* starts from the protonema that emerge from haploid spores followed by development of the leafy gametophore. The reproductive organs (antheridia and archegonia) later develop and after the fertilisation of the egg inside the archegonium by sperm released from antheridia, a sporophyte capsule develops, marking transition to the sporophyte phase. From the graph it is shown that *PpDUO1A* expression is detectable at low levels by 9 days after induction (DAI) of antheridia involving transfer to inductive conditions. The expression level increases by 11 DAI and peaks at 9-fold expression on 14 and 15 DAI of the antheridia. Low levels of *PpDUO1A* transcripts are present in sperm cell packages but transcripts are otherwise absent from the other tissues. The expression of *PpDUO1B* however is detected at low level in the protonema and the leaflets. On 9 DAI of antheridia, *PpDUO1B* is still expressed at a low level and steadily increases in expression in 14 and 15 DAI antheridia. The 2-fold increase in expression of *PpDUO1B* relative to antheridia at 11 DAI is lower than that of *PpDUO1A*.

The expression of *PpDUO1A* and *PpDUO1B* was also explored in the antheridia from different *P. patens* accessions based on publicly available data. RNA-seq data of mature antheridia from Gransden and Reute accessions was obtained from a recent study by Meyberg et al. (2020). This RNA-seq data was plotted (Figure 4.2.3.2) and compared to data obtained from the CoNekT platform, the expression of *PpDUO1B* is 2-fold higher than *PpDUO1A* in mature stage of antheridia in both Gransden and Reute (Table 4.2.3).

All together, these data show that both *PpDUO1A* and *PpDUO1B* are highly expressed in the male reproductive organs of *P. patens*. This suggests the involvement of both genes in the process of male gametogenesis that takes place in the antheridia. The expression of *PpDUO1A* appears to be male-specific, while *PpDUO1B* is male-preferential or enriched rather than specific. This could suggest the dominance or importance of *PpDUO1A* relative to *PpDUO1B* in antheridia, or it might reflect drift or divergence in function for *PpDUO1B* to include roles in other tissues.
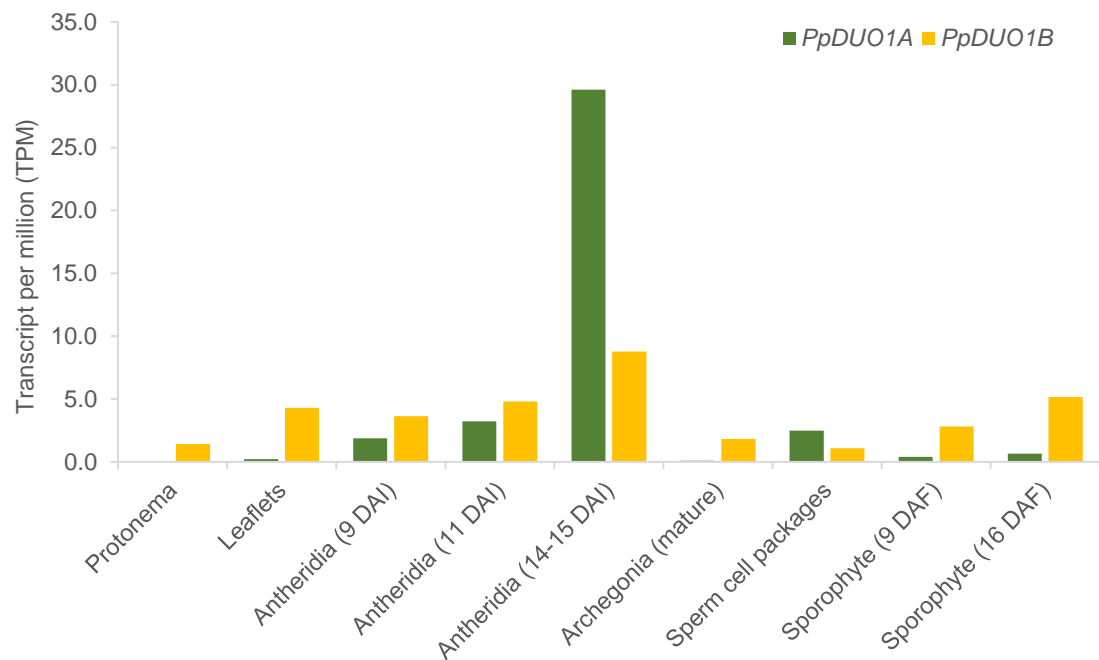
Figure 4.2.3.1. Expression profile of *PpDUO1A* and *PpDUO1B* in different developmental tissues from the Gransden accession. RNA-seq data of selected tissues representing major developmental stages downloaded from CoNekT platform. Gene expression was expressed in Transcripts per million (TPM).
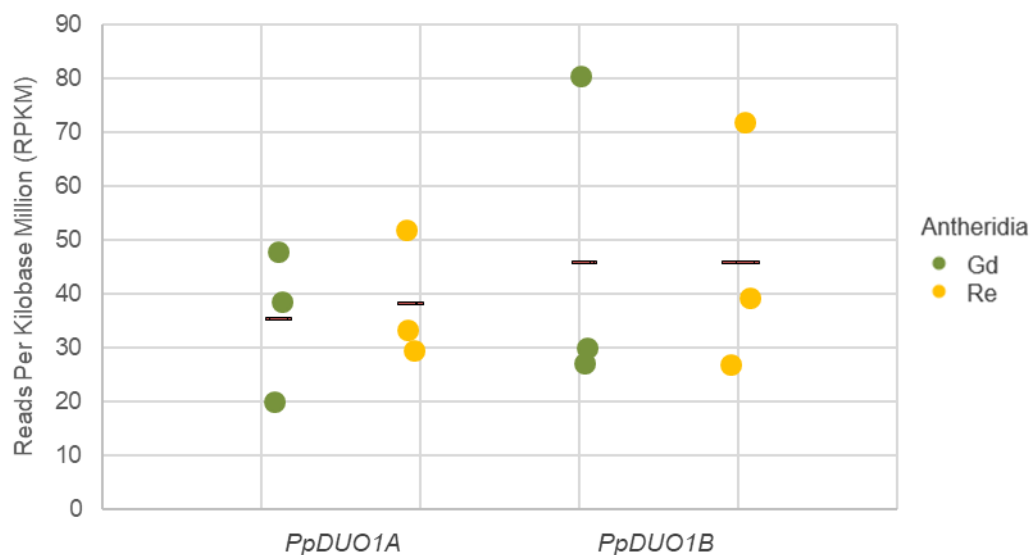
Figure 4.2.3.2. Expression profile of *PpDUO1A* and *PpDUO1B* in antheridia from Gransden and Reute. RNA-seq data of antheridia from Meyberg et al. (2020). Gene expression was expressed in Reads Per Kilobase Million (RPKM).

Table 4.2.3. The average transcripts (RPKM) of *PpDUO1A* and *PpDUO1B* in antheridia from Gransden and Reute. Average transcripts were calculated from three sample replicates (Meyberg et al., 2020). Fold difference was calculated by dividing average transcripts of *PpDUO1B* with *PpDUO1A*.

	Antheridia (RPKM)			
	Gransden (Gd)		Reute (Re)	
Gene	Average	Fold difference	Average	Fold difference
PpDUO1A	35.42	1.29	38.22	1.2
PpDUO1B	45.79		45.92	

4.3 Functional analysis of *PpDUO1A* and *PpDUO1B*

4.3.1 Generation of knockout mutants of *PpDUO1A* and *PpDUO1B*

To study the involvement of *PpDUO1A* and *PpDUO1B* in *Physcomitrella* male gametogenesis, single knockout lines for each gene as well as a double knockout were generated. All the knockout lines, *Ppduo1a$^\Delta$*, *Ppduo1b$^\Delta$* and *Ppduo1a$^\Delta$b$^\Delta$*, were generated and provided by Dr. Yasuko Kamisugi and Andrew Cumming University of Leeds. Single mutant *Ppduo1a$^\Delta$* and *Ppduo1b$^\Delta$* lines and a double mutant *Ppduo1a$^\Delta$b$^\Delta$* lines were designed by replacing the entire genomic sequence of each gene with a selection cassette (Figure 4.3.1.1). This cassette contains 5' end target and 3' end target sequences which are homologous to the sequences flanking the gene of interest. These target sequences flank two loxP sites and a selectable marker, which is the neomycin resistance (neo^R) gene for *Ppduo1a* and the hygromycin resistance (hyg^R) gene for *Ppduo1b*.

Two plasmids containing the selection cassette, namely pDUO1aKO and pDUO1bHKO2, were constructed in order to knockout *PpDUO1A* and *PpDUO1B* (Figure 4.3.1.2). The cassette was flanked by 5' and 3' targeting sequences that would target the Pp11770 locus on chromosome 24, and the Pp16720 locus on chromosome 8 for *PpDUO1A* and *PpDUO1B*, respectively. These targeting sequences were initially amplified from 5' upstream and 3' downstream regions of each genes and inserted into the multiple cloning site of the respective plasmids (Figure 4.3.1.3). The plasmids were linearised and transformed into moss protoplasts. Homologous recombination takes place at regions that are homologous 5' and 3' resulting in the replacement of *PpDUO1A* and *PpDUO1B* entire genomic sequences with the selection cassette. This method was used to generate single mutant *Ppduo1a* and *Ppduo1b* mutant strains. Southern blotting

was then conducted to confirm the knockout mutation (Appendix Figure S4.2). Cre-Lox recombination was then induced in each strain by transient expression of the Cre enzyme in protoplasts. Recombination occurs at the loxP sites which results in the deletion of the selection marker, generating 'marker-free' single $Ppduo1a^{\Delta}$ and $Ppduo1b^{\Delta}$ strains. In order to make the double mutant $Ppduo1a^{\Delta}b^{\Delta}$ strain, one of the 'marker-free' single mutants was transformed with a selection cassette corresponding to the unmutated gene locus and the same process and clean-up involving Cre-Lox removal of the selectable marker was applied.
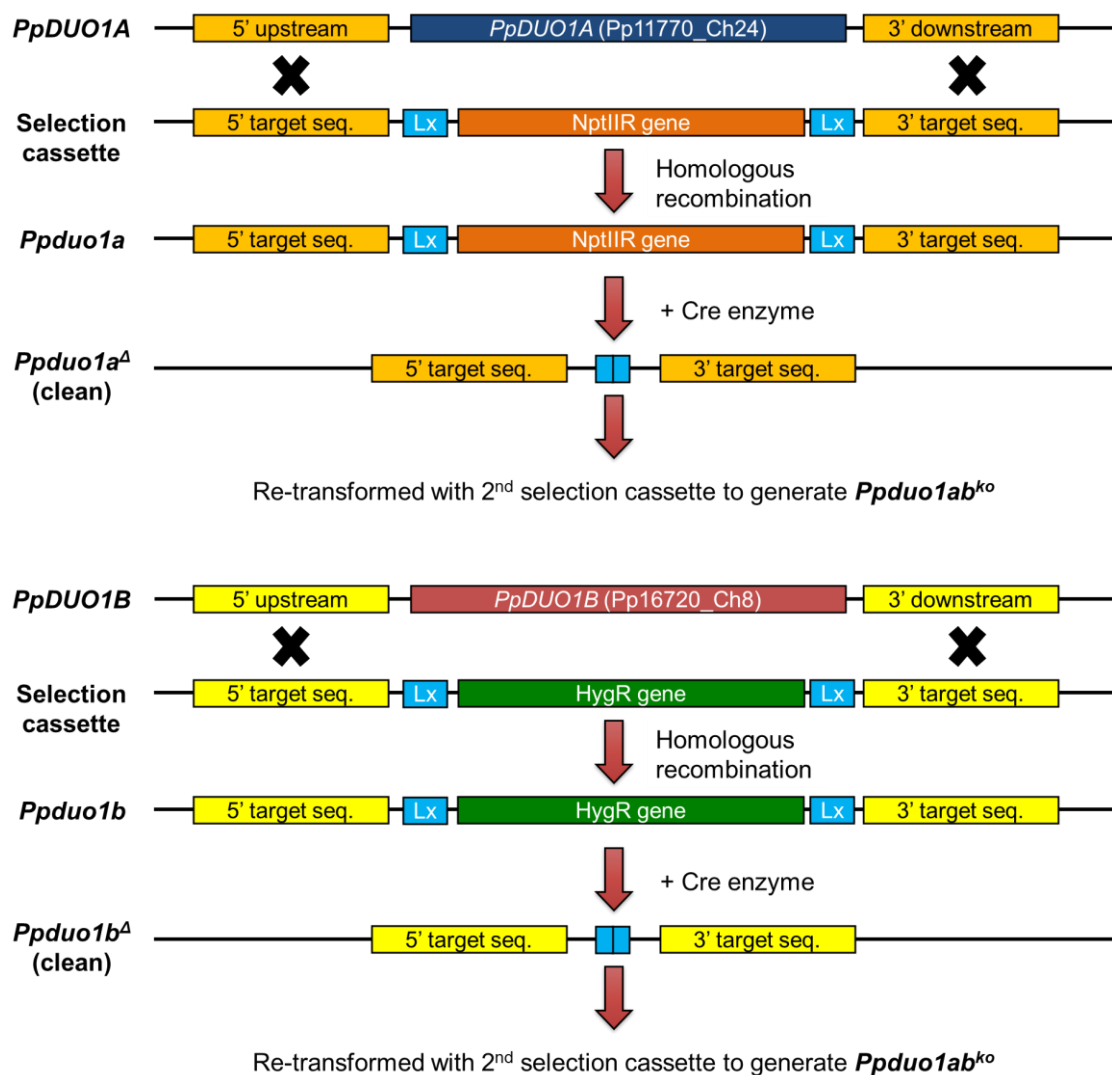
Figure 4.3.1.1. General diagram for the generation of knockout mutants. 5' and 3' target sequences flanking the loxP sites (Lx) of the selection cassette were homologous to the sequences flanking the *PpDUO1* genes. Entire *PpDUO1* genomic sequence is then shown being replaced by the selection marker.
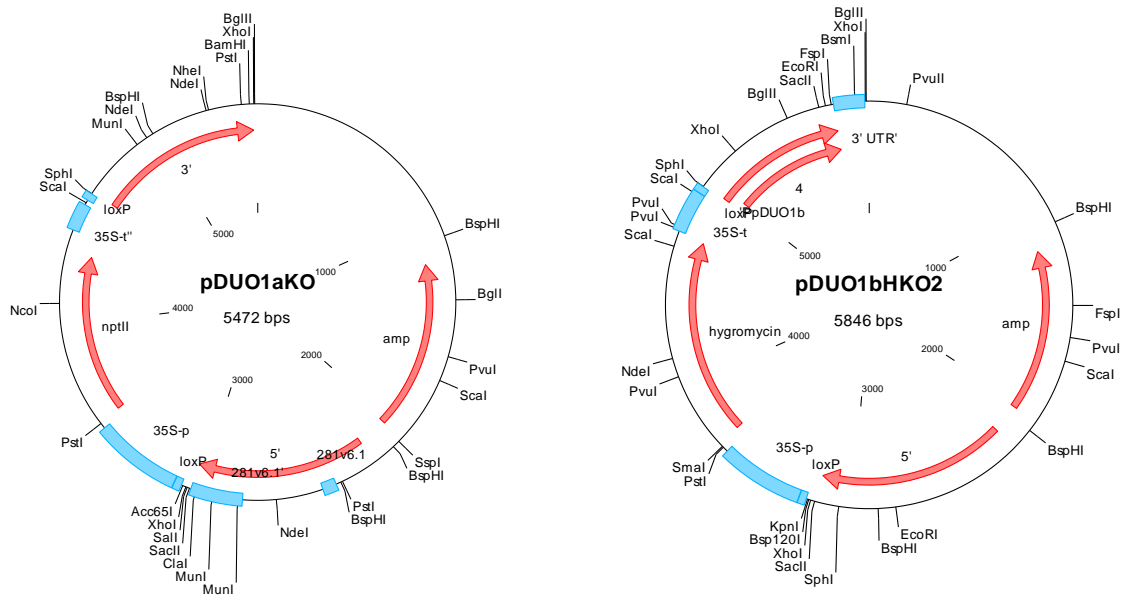
Figure 4.3.1.2. Plasmid constructs containing selection cassette for knocking out *PpDUO1A* and *PpDUO1B*. Map was provided by Dr. Yasuko Kamisugi. Labels on map: nptII, neomycin phosphotransferase II, hygromycin, hygromycin resistance gene, 5', 5'end target sequence, 3', 3'end target sequence, 4, 3'end target sequence, loxP, lox P site, 35S-p, 35S promoter; 35S-t, 35S terminator.
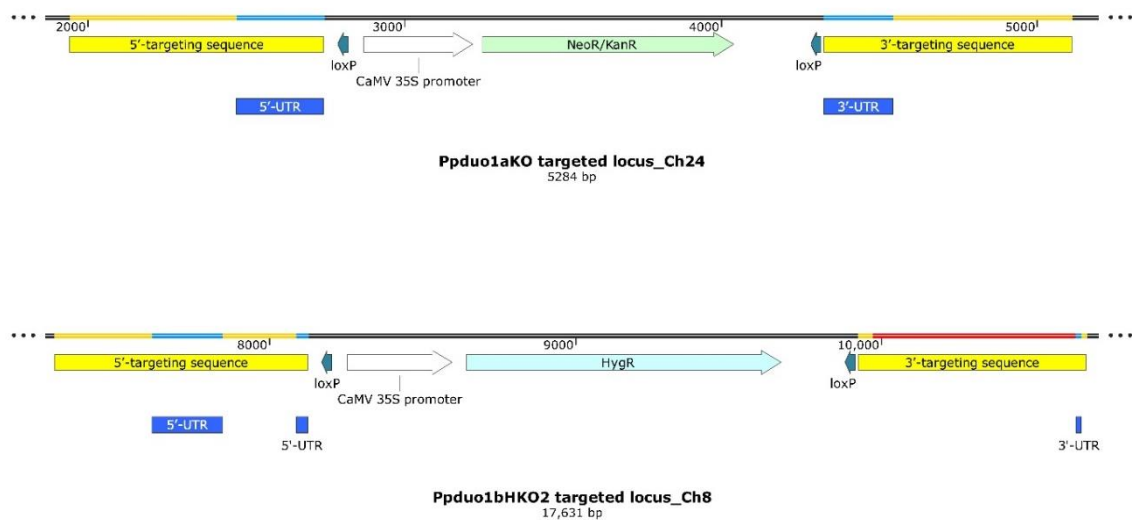
Figure 4.3.1.3. Schematic diagram of the *Ppduo1a* and *Ppduo1b* knockout targeted loci. *PpDUO1A* location is present on chromosome 24, while *PpDUO1B* is located on chromosome 8.

4.3.2 Sporophyte formation in single and double *PpDUO1A* and *PpDUO1B* knockout mutants

In order to study the role of *PpDUO1A* and *PpDUO1B* in male gametogenesis, the fertility of each of the *Ppduo1a$^\Delta$*, *Ppduo1b$^\Delta$* and *Ppduo1a$^\Delta$b$^\Delta$* knockout mutants was compared with that of the wild type. Sterile water was added to the culture container at the 14 to 15 DAI so that half of the gametophores were submerged. After two weeks, the gametophores were observed for the presence of sporophytes. A total of 25 gametophores were selected at random from a culture container for each genotype and the presence of sporophytes was recorded for each gametophore. In addition, the total number of sporophytes for each genotype was also counted. The experiment was repeated twice for all genotypes (total culture containers, n = 3) and the percentage of gametophores which produced at least one sporophyte was calculated.

Sporophyte capsules were formed at the top end of the gametophore of wild type, *Ppduo1a$^\Delta$* and *Ppduo1b$^\Delta$*. The capsules were round and either yellow (mature) or green (young) in colour, depending on the developmental stage of the sporophyte. Interestingly, none of the gametophores of *Ppduo1a$^\Delta$b$^\Delta$* produced even a single sporophyte (Figure 4.3.2.1.a.). Single mutant knockout lines *Ppduo1a$^\Delta$* and *Ppduo1b$^\Delta$* each produced an average of 100 % and 95 % for gametophores with at least one sporophyte. These values are close to the wild type gametophores that have an average of 97 %. Statistical analysis was performed using One-way ANOVA to examine differences in the mean percentage of gametophores with sporophytes in wild type, *Ppduo1a$^\Delta$*, *Ppduo1b$^\Delta$* and *Ppduo1a$^\Delta$b$^\Delta$* (Table 4.3.2.1). As expected, the result showed that there is a significant difference in the mean percentage of gametophores with sporophytes within the four tested genotypes (F (3,8) = 2669.83, p = 2.4E-12).
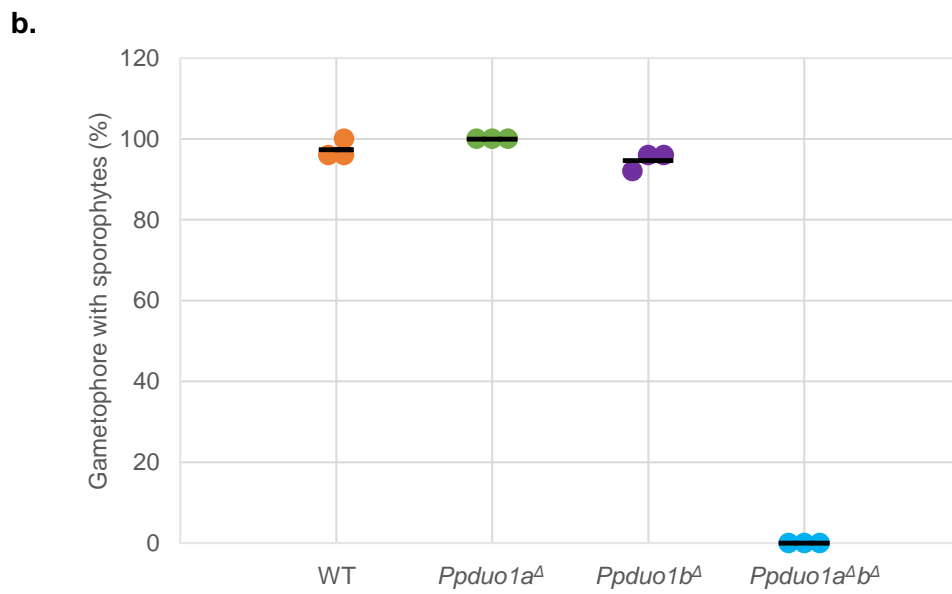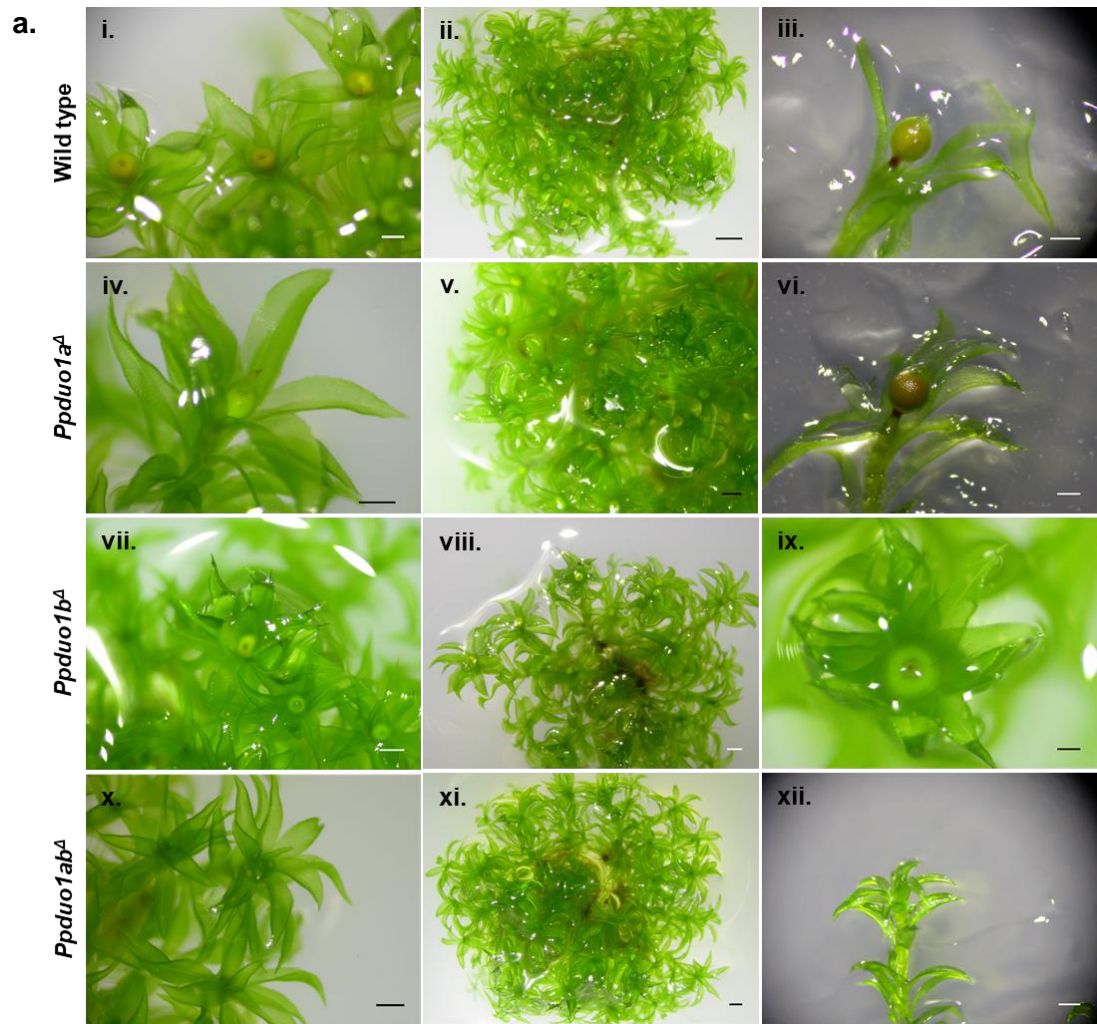
Figure 4.3.2.1. Sporophyte formation for *Ppduo1a^Δ*, *Ppduo1b^Δ* and *Ppduo1a^Δb^Δ*. a.
Sporophytes were observed on gametophores of *Ppduo1a^Δ* and *Ppduo1b^Δ* single

mutants but absent from gametophores of $Ppduo1a^{\Delta}b^{\Delta}$ double mutants. Scale bar = 0.2 mm (**ix**); scale bar = 0.5 mm (**i, iii, iv, vi, vii, x, xii**); scale bar = 1 mm (**v, viii, xi**); scale bar = 2 mm (**ii**) **b.** Percentage of *P. patens* gametophores which produced at least one sporophyte. Each data point represents three biological replicates and the bar shows the average percentage of gametophore with sporophytes.

Table 4.3.2.1. One-way ANOVA for analysing differences in the proportion of gametophores with sporophytes in four different test groups. The four test groups were wild type, $Ppduo1a^{\Delta}$, $Ppduo1b^{\Delta}$ and $Ppduo1a^{\Delta}b^{\Delta}$ mutants. There were statistically significant differences between group means as determined by one-way ANOVA ($F_{(3,8)}$ = 2669.83, p = 2.4E-12). * = significant.

	Population (replicate)	*F-statistic*	*P-value*	*F-critical*	Significance
Wild type	3				
$Ppduo1a^{\Delta}$	3	2669.83	2.4E-12	4.07	*
$Ppduo1b^{\Delta}$	3				
$Ppduo1a^{\Delta}b^{\Delta}$	3				

A Tukey-Kramer post hoc test was then performed to detecting genotypes that had significant differences (Table 4.3.2.2). Unsurprisingly, $Ppduo1a^{\Delta}b^{\Delta}$ is confirmed to be significantly different from wild type, $Ppduo1a^{\Delta}$ and $Ppduo1b^{\Delta}$ (q-stat > 4.53, p < 0.05; q-stat > 6.20, p < 0.01), whereas $Ppduo1a^{\Delta}$ and $Ppduo1b^{\Delta}$ are not significantly different to the wild type (q-value < 4.53, p > 0.05).

The total number of sporophytes was recorded for all three replicates of wild type, $Ppduo1a^{\Delta}$ and $Ppduo1b^{\Delta}$ (Table 4.3.2.3). An average of 44 and 43 sporophytes were observed for $Ppduo1a^{\Delta}$ and $Ppduo1b^{\Delta}$, respectively. These values were almost identical to the wildtype value of 42 sporophytes. The mean number of sporophytes per gametophore was also calculated for each replicate and these data are presented in Figure 4.3.2.2. Both $Ppduo1a^{\Delta}$ and $Ppduo1b$ had an average of two sporophytes per gametophore which not significantly different from that of the wild type ($F_{(2,6)}$ = 0.08, p = .93) (Appendix Table S4.1).

Table 4.3.2.2. Tukey-Kramer test for detecting test group that were significantly different in the proportion of gametophores with sporophytes in four different test groups. Tukey-Kramer post hoc test was performed after one-way ANOVA test by comparing two test groups. Asterisk (*) and (**) refers to $P < 0.05$ and $P < 0.01$, respectively. ns = not significant; * / ** = significant.

		q-stat	q-value, α: 0.05	q-value, α: 0.01	Significance
WT	$Ppduo1a^\Delta$	2.83			ns
$Ppduo1a^\Delta$	$Ppduo1b^\Delta$	5.66			*
$Ppduo1b^\Delta$	$Ppduo1a^\Delta b^\Delta$	100.41	4.53	6.20	*, **
$Ppduo1a^\Delta b^\Delta$	WT	103.24			*, **
$Ppduo1a^\Delta b^\Delta$	$Ppduo1a^\Delta$	106.07			*, **
$Ppduo1b^\Delta$	WT	2.83			ns

Table 4.3.2.3. Number of sporophytes for wild type, $Ppduo1a^\Delta$, $Ppduo1b^\Delta$ and $Ppduo1a^\Delta b^\Delta$. Sporophytes were counted in 25 gametophores and the experiment was repeated three times. The average number of sporophytes was calculated from all replicates data that was pooled.

Population (replicate)	WT	$Ppduo1a^\Delta$	$Ppduo1b^\Delta$	$Ppduo1a^\Delta b^\Delta$
1	46	52	42	0
2	42	32	52	0
3	38	48	36	0
Average	42	44	43	0
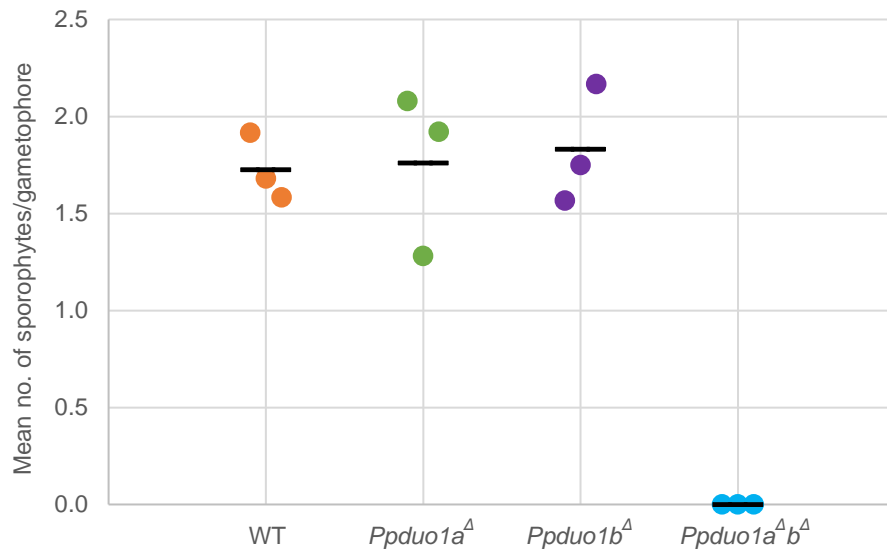
Figure 4.3.2.2. Mean number of sporophytes per gametophore. The mean number of sporophytes are shown (y-axis) for each genotype (x-axis). Each spot represents average number of sporophytes per gametophore (n = 25). *Ppduo1a^Δ^*, *Ppduo1b^Δ^* and wild type had an average of two sporophytes per gametophore. No sporophytes were observed on *Ppduo1a^Δ^b^Δ^* gametophores.

Since the mean percentage of gametophores with sporophytes for *Ppduo1a^Δ^* is significantly different to *Ppduo1b^Δ^* at critical value 0.05, an additional analysis was conducted where *Ppduo1a* and *Ppduo1b* were compared to each other. A total of 100 gametophores were collected randomly from three culture containers for each genotypes and the percentage of gametophores which produced at least one sporophyte was calculated. The experiment was repeated once and a One-way ANOVA test was performed (Table 4.3.2.4). From the analysis, *Ppduo1a* does not show a significantly different proportion of gametophore with sporophytes to *Ppduo1b* (F (1,28) = 0.04, p = .85).

Table 4.3.2.4. One-way ANOVA for analysing differences in the mean percentage of gametophores with sporophytes in *Ppduo1a* and *Ppduo1b*. There were no statistically significant differences between group means as determined by one-way ANOVA ($F(1,28) = 0.04$, $p = .85$). ns = not significant.

SUMMARY				
Groups	*Count*	*Sum*	*Average*	*Variance*
Ppduo1a (T314/2-8)	15	1400	93.33333	77.38095
Ppduo1b (T311/4-5)	15	1390	92.66667	103.0952

	F-statistic	*P-value*	*F-critical*	Significance
Ppduo1a				
Ppduo1b	0.04	0.85	4.20	ns

In summary, these analyses described show no significant differences in the percentage of *Ppduo1a$^\Delta$* and *Ppduo1b$^\Delta$* gametophores which produced at least one sporophyte as well as in the mean number of sporophytes per gametophore. The absence of sporophytes for *Ppduo1a$^\Delta$b$^\Delta$* illustrate the requirement for both *PpDUO1A* and *PpDUO1B* in the formation of sporophytes. Collectively, these data indicate a critical role for *PpDUO1A* and *PpDUO1B* in the fertility of *P. patens*. The normal fertility of the single mutants further suggests that *PpDUO1A* and *PpDUO1B* show functional redundancy. Thus, *PpDUO1A* or *PpDUO1B* are sufficient to maintain fertility in *P. patens* in the absence of the other paralogue, such that the loss of function of either is compensated by the other.

4.3.3 Morphological analysis of germ cell nuclei in *Ppduo1a$^\Delta$b$^\Delta$* double mutants

The enriched expression and likely function of *PpDUO1A* and *PpDUO1B* in antheridia may explain the infertility observed in the *Ppduo1a$^\Delta$b$^\Delta$* double mutant. To investigate the potential role of these genes in the development of spermatogenous cells (SpCs), developing antheridia of wild type and *Ppduo1a$^\Delta$b$^\Delta$* plants were examined after fixation and 4′,6-diamidino-2-phenylindole (DAPI) staining. This allowed the morphology and the DNA compaction of SpCs nuclei to be observed.

Six gametophores at 12 DAI were fixed in a formalin-acetic acid-alcohol (FAA) for wild type and *Ppduo1a$^\Delta$b$^\Delta$*. The leaves were removed from gametophore tips and antheridia

were brushed into a water droplet on a polylysine slide (Horst and Reski, 2017). The antheridia were later stained with DAPI and viewed using differential interference contrast (DIC) and fluorescence microscopy. The experiment was repeated twice and similar results were observed each time.

A cluster of antheridia representing several stages of development was examined from gametophores 12 DAI. Although the majority of antheridia in the cluster were at the mature stage, there were a few younger antheridia present. The mature antheridia in both genotypes were larger and had a swollen apical tip compared to younger ones when viewed under DIC. When viewed using a DAPI filter, the shape of the SpCs nucleus in wild type antheridia changed from round in the younger antheridia, to crescent-shaped in mature antheridia. This change in shape of SpCs nuclei however, was not observed in the mature antheridia of $Ppduo1a^\Delta b^\Delta$ and SpCs nuclei remained rounded in mature antheridia. The inability of $Ppduo1a^\Delta b^\Delta$ SpCs to undergo this aspect of cellular morphogenesis shows impairment in male germ cell development. The failure of SpC morphogenesis in the double mutant demonstrates the requirement of *PpDUO1A* and *PpDUO1B* for germ cell development.
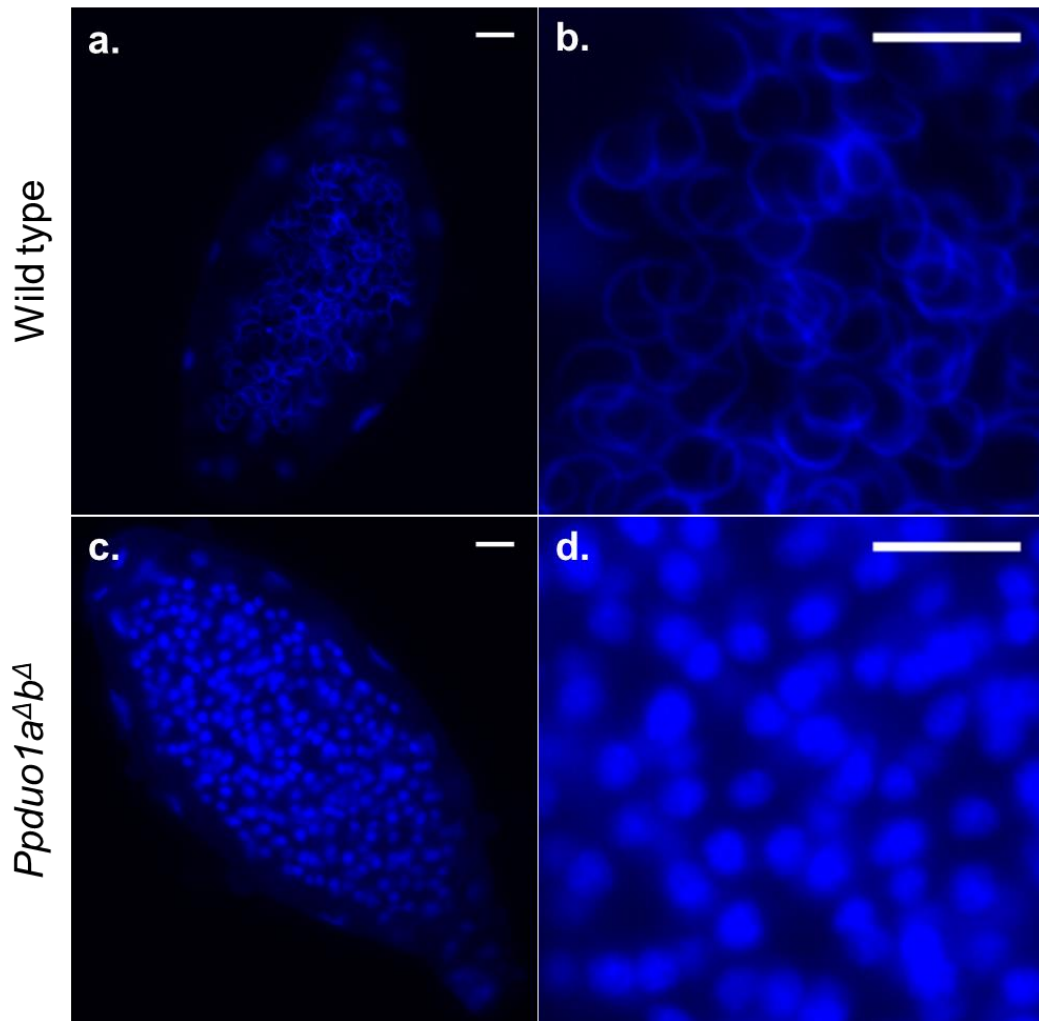
Figure 4.3.3. Antheridia at day 12 post-induction. DAPI fluorescence images of **(a, b)** wild type antheridia and **(c, d)** *Ppduo1a^{A}b^A* antheridia. Nuclei in SpCs of *Ppduo1a^{A}b^A* failed to transition from the spherical form to crescent-shape unlike the wild type. Scale bar = 10 μm.

4.3.4 Sperm cell number in antheridia of *Ppduo1a^{A}b^A* double mutants

To examine the potential effect of *Ppduo1a^{A}b^A* mutation on germ cell proliferation, the number of sperm cells in mature antheridia of mutant and wild type plants were compared. The samples used were the same as those used for morphological analysis. Six mature antheridia (n=6) were selected and the number of SpCs for each were counted from captured images using ImageJ software (Schneider et al., 2012).

The average number of SpCs counted in wild type antheridia was 119 ± 14 SD, while *Ppduo1a^{A}b^A* produced 120 ± 21 SD (Figure 4.3.4; Appendix Figure S4.3). One-way ANOVA was conducted to examine whether the average number of SpCs in *Ppduo1a^{A}b^A*

antheridia is different from the wild type (Table 4.3.4; Appendix Table S4.2). The p-value from the test is 0.9 which is above the 0.05, which suggest that the average number of sperm cells in $Ppduo1a^{\Delta}b^{\Delta}$ antheridia is not significantly different from that of the wild type ($F_{(1,10)}$ = 10.017, p = .90). These result shows that there is no difference in the overall number of undifferentiated SpCs in $Ppduo1a^{\Delta}b^{\Delta}$ compared with the number of morphologically normal SpCs present in wild type antheridia.
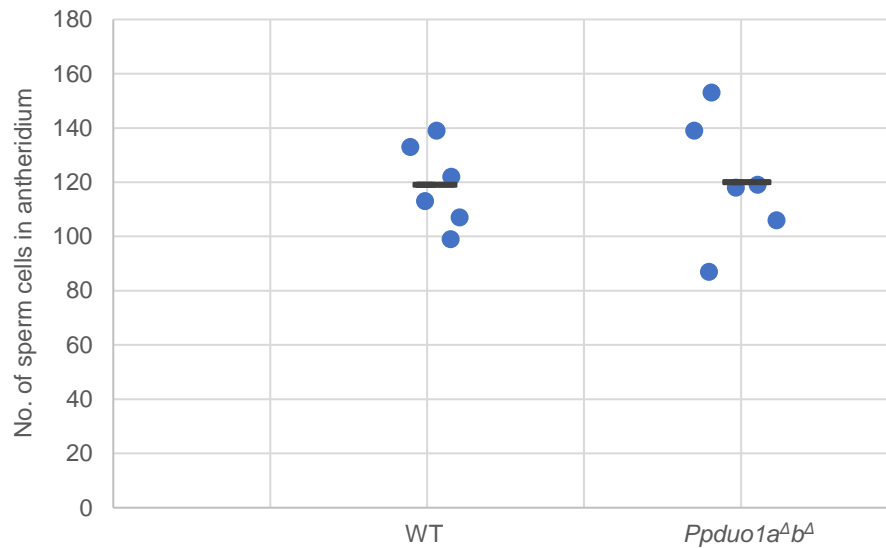
Figure 4.3.4. The number of sperm cells in mature antheridia at day 12 post-induction. The number of sperm in each of six different antheridia were counted for wild type and $Ppduo1a^{\Delta}b^{\Delta}$ mutants. Then the mean number of sperms per antheridium was determined. Bar shows the mean of the means.

Table 4.3.4. One-way ANOVA for analysing any significant difference in the average number of sperm cells in mature antheridia of wild type and $Ppduo1a^{\Delta}b^{\Delta}$ mutants. There were no statistically significant differences between group means as determined by one-way ANOVA ($F_{(1,10)}$ = 10.017, p = .90). ns = not significant.

	Antheridium (n)	F-statistic	P-value	F-critical	Significance
Wild type	6	0.017	0.90	4.96	ns
$Ppduo1a^{\Delta}b^{\Delta}$	6				

4.3.5 Ultrastructural analysis of wild type moss and *Ppduo1a$^\Delta$b$^\Delta$*

In order to understand the effect of *Ppduo1a$^\Delta$b$^\Delta$* mutation on the development of SpCs, fixed antheridia samples were sectioned, processed and examined by transmission electron microscopy (TEM) with the help of Ania Straatman-Iwanowska from the Electron Microscopy Facility, University of Leicester. Antheridia were sampled at different times after the gametophores were transferred to inductive conditions to provide different stages of antheridia development. The exact developmental stages of individual antheridia were determined by examining the size, overall morphology and ultrastructural features of the jacket cells (JCs) surrounding the inner SpCs. Wild-type antheridia were classified as early, mid and late developmental stages and antheridia from *Ppduo1a$^\Delta$b$^\Delta$* mutants were similarly classified in comparison with established features in the wild-type.

In early-stage antheridia, JCs were rectangular cells, with electron dense cell boundaries (Figure 4.3.5.1.a). The nucleus was round, centrally positioned and largely euchromatic with small patches of heterochromatin. Multiple vacuoles irregular in size were present in the cytoplasm, along with a large, elongated chloroplast. Small, rounded mitochondria were also scattered in the jacket cell cytoplasm. SpCs at the early stage, were generally quadrilateral (four-sided) in profile. The plasma membrane and the cell wall were very close or in contact. In addition, short projections were present on the plasma membrane as the SpCs progress further (Figure 4.3.5.1.b). The nucleus in each SpC was round and euchromatic with small patches of heterochromatin. Vacuoles were few and small, while the single plastid was highly elongated. Smaller circular mitochondria were scattered inside the cell cytoplasm and their inner membranes (which form the cristae) were not strongly folded. The SpC cytoplasm was dense and packed with ribosomes. Golgi with folded cisternae were also present. Small vesicles were seen near the golgi, while rough endoplasmic reticulum (ER) were located near to the nucleus and plasma membrane. Similar characteristics were also observed among the *Ppduo1a$^\Delta$b$^\Delta$* antheridia samples. The jacket cells and spermatogenous cells in early-stage antheridia of *Ppduo1a$^\Delta$b$^\Delta$* possessed the same features as the wild type, with no clear differences between them (Figure 4.3.5.1.a. and b.).

In mid-stage antheridia, JCs retain several features that were observed during the early stage. JCs were larger and more elongated at this stage and possessed larger vacuoles with irregular profiles. The SpCs undergo a major shift in morphology at this stage from quadrilateral to circular in profile. The plasma membrane of SpCs were no longer in contact with the cell wall and an electron dense fibrillar matrix separated SpCs. The plasma membrane also had an irregular, wavy profile compared to the short plasma

membrane projections seen in early-stage SpCs. The nucleus in SpCs was round and largely euchromatic with patches of heterochromatin. Some larger vacuoles were present in SpCs, whereas the plastid was elongated and thicker than at earlier stages. SpCs mitochondria were elongated, with folded cristae present. The cytoplasm of SpCs was denser and packed with ribosomes compared with than at the early-stage, rough ER was also present. Locomotary apparatus structures were first observed in SpCs during the mid-stage, comprising basal bodies and the multiple layer structure (MLS). The basal bodies showed triplet microtubule orientation while the MLS comprised the spline and lamellar strips.

Multiple differences were seen in the mid-stage of $Ppduo1a^{\Delta}b^{\Delta}$ antheridia samples, in comparison with wild type samples. While the jacket cells were comparable to the wild type, the $Ppduo1a^{\Delta}b^{\Delta}$ SpCs exhibited some striking differences. The $Ppduo1a^{\Delta}b^{\Delta}$ SpCs plasma membrane had a gently undulating plasma membrane profile compared to the distinctive, wavy profile of wild type SpCs. Fewer larger vacuoles or a single large vacuole were present in the $Ppduo1a^{\Delta}b^{\Delta}$ SpCs. Golgi were abundant with many associated small vesicles present near the plasma membrane. Similar to the wild type, basal bodies and the MLS were present in $Ppduo1a^{\Delta}b^{\Delta}$ SpCs (n=25).

In late-stage antheridia, JCs were larger and elongated compared to the mid-stage. Rupture of the apical cap cell was commonly observed in mature antheridia. The JCs vacuoles were large and smooth in outline. The SpCs were now irregular in profile and the fibrillar matrix between SpCs had become less dense than its composition during the mid-stage antheridia. The size of SpCs was also highly reduced. The SpCs cell wall boundary was intact, although less visible than at the mid-stage. The SpCs plasma membrane had smooth and irregular profile compared to the mid stage which was irregular and wavy in profile. The nucleus had a crescent shape outline, strikingly heterochromatic and was elongated with the spline adjacent to the plasma membrane. SpCs had fewer larger vacuoles or single large vacuole that were reduced or absent as SpCs matured. The single plastid contained starch granules whereas mitochondria were elongated with folded cristae and more rounded at later stages. The SpCs cytoplasm was denser than the mid stage and packed with ribosomes. Small vesicles were present near the cell membrane. The MLS, comprised of spline and lamellar strips, was present and flagella were also visible at this stage. In addition, motile sperm cells were observed being released from the antheridia using DIC microscopy.

In late stage of $Ppduo1a^{\Delta}b^{\Delta}$ antheridia samples, JCs showed similar features to those observed for wild type antheridia. In terms of the SpCs, the plasma membrane was now

smooth yet irregular in profile, similar to that of wild type SpCs. The nucleus of SpCs was partially elongated and euchromatic with larger patches of heterochromatin. Large single vacuoles were present in *Ppduo1a$^\Delta$b$^\Delta$* SpCs at the late stage. In stark contrast to the differentiated spermatozoids of wild type, no flagella were detected in the late stage of *Ppduo1a$^\Delta$b$^\Delta$* SpCs. When viewed using DIC microscopy, immotile SpCs were seen being released from antheridia in *Ppduo1a$^\Delta$b$^\Delta$* mutants as opposed to the wild type. The absence of flagella structures demonstrates the requirement for *PpDUO1A* and *PpDUO1B* in the development of the locomotary apparatus of SpCs.
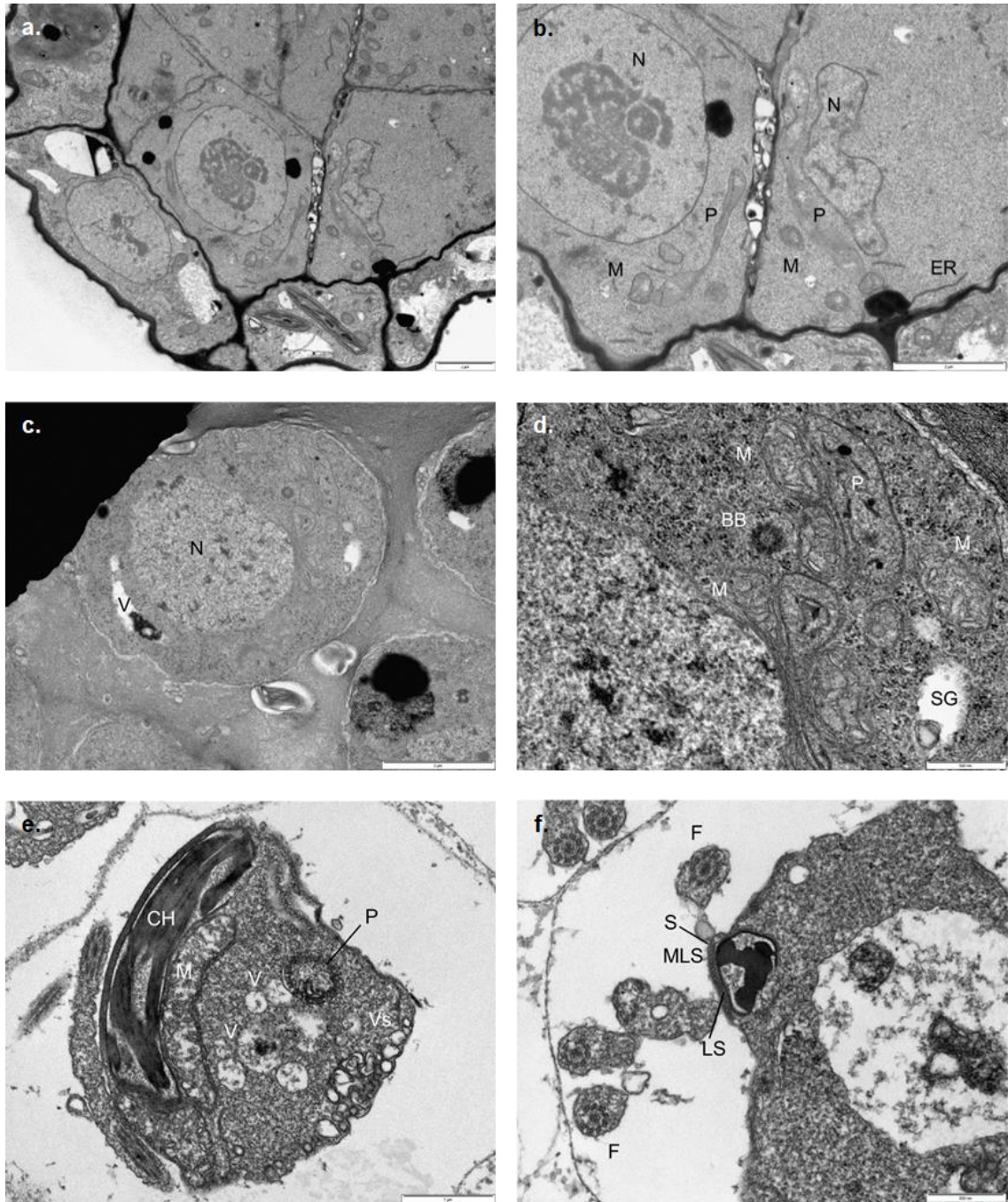
4.3.5.1. Ultrastructural analysis of SpCs of wild type antheridia. (a and b) SpCs during the early-stage of development, **(c and d)** SpCs during the mid-stage of development, **(e and f)** SpCs during the late stage of development. Flagella were visible in the late stage of SpCs. N, nucleus; BB, basal body; M, mitochondria; P, plastid; SG, starch vacuole; V, vacuole; CH, compact chromatin; CV, central vacuole; ER, endoplasmic reticulum; Vs, vesicles; F, flagella; LS, lamellar Strip; MLS, multiple layer structure; S, spline.
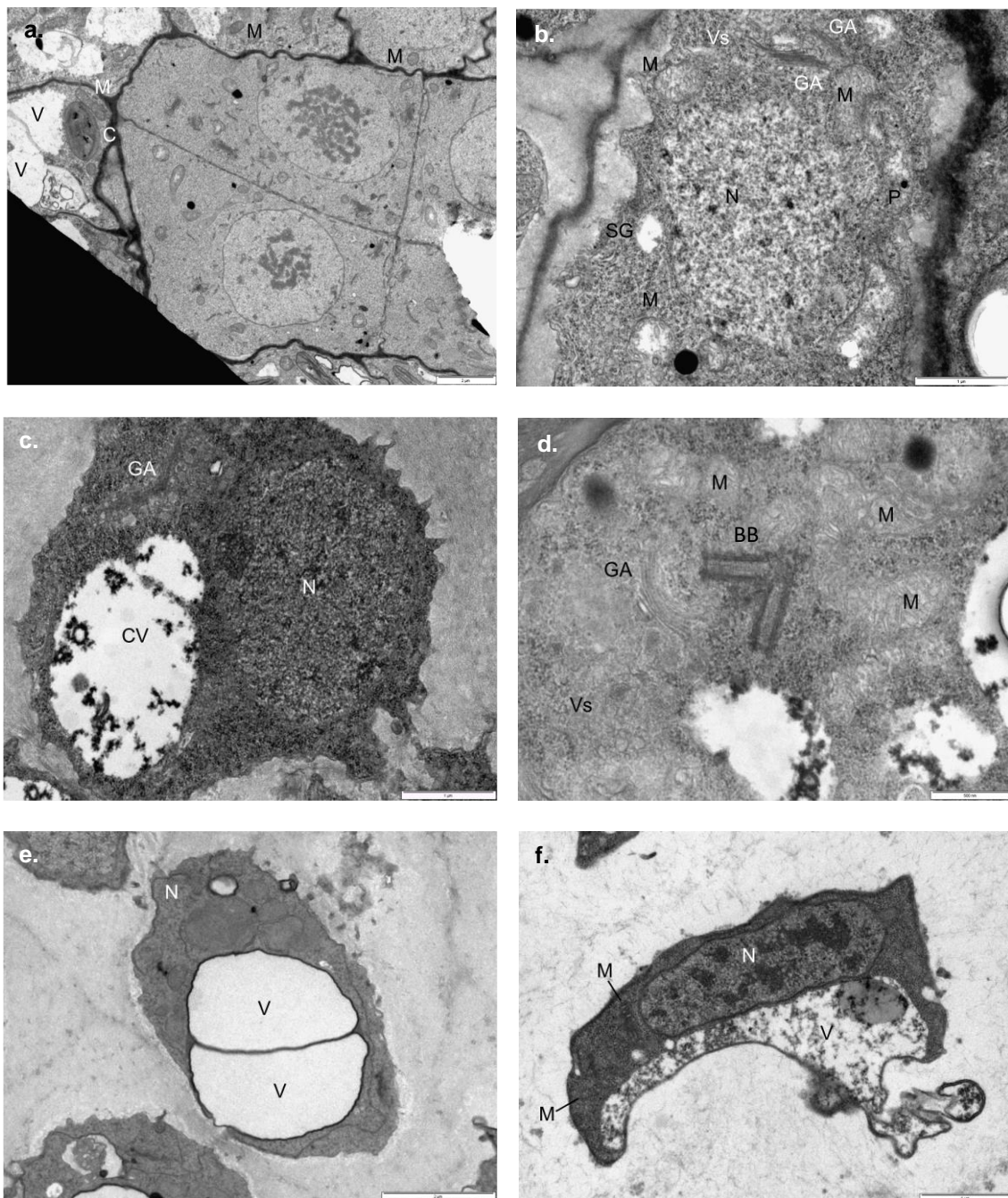
4.3.5.2. Ultrastructural analysis of SpCs of _Ppduo1a^Δb^Δ_ antheridia. (a and b) SpCs during the early-stage of development, **(c and d)** SpCs during the mid-stage of development, **(e and f)** SpCs during the late stage of development. The flagella were absent in the late stage of SpCs. N, nucleus; GA, golgi apparatus; C, chloroplast; BB, basal body; M, mitochondria; P, plastid; SG, starch vacuole; V, vacuole; CH, compact chromatin; CV, central vacuole; ER, endoplasmic reticulum; Vs, vesicles.

4.4 Analysis of *PpDUO1A* and *PpDUO1B* promoter activity

4.4.1 Generation of *PpDUO1A* and *PpDUO1B* promoter-reporter gene constructs

To investigate the spatial and temporal expression patterns of *PpDUO1A* and *PpDUO1B* transcripts, promoter-reporter gene constructs were designed for *PpDUO1A* and *PpDUO1B*. The promoters were fused to the *GUS* reporter gene and transformed into wild type *P. patens* (Reute). The transformed plants will be used to analyse the level of promoter activity through histochemical staining for GUS activity. Two additional constructs where both promoters were fused to a *GFP* reporter with a nuclear localisation signal were also constructed. Through this approach the activity of the promoters can be investigated in developing tissues vivo together with the localisation of DUO1-fusion proteins in expressing cells.

Promoter fragments of 1.5 kb and 1.7 kb were selected for *PpDUO1A* and *PpDUO1B*, respectively. These were amplified from genomic DNA of the Reute accession using two step PCR with annealing and extension temperatures optimised to 65°C for *PpDUO1A* and 60°C for *PpDUO1B*. The fragments were later cloned into the pENTR5' TOPO-TA vector to generate *PromPpDUO1A* and *PromPpDUO1B* entry clones (Figure 4.4.1.1). The *GUS* coding sequence entry clone was made and provided by Dr. Dieter Hackenberg who amplified the *GUS* coding sequence from pLAT52-7 (Twell et al., 1989) and subsequently cloned it into the pDONR221 entry vector (Figure 4.4.1.2) The *NLS-GFP* entry clone was obtained from the Tol2kit (tol2kit.genetics.utah.edu).

The empty gateway compatible destination vector, pTHattR4R2, was constructed and provided by Dr. Dieter Hackenberg. This plasmid is a derivative of pTH-UBI-Gate and confers resistance to hygromycin for plant selection (Figure 4.4.1.3). Two sequence fragments termed 108 5' and 108 3' were designed to target the Pp108 neutral locus on chromosome 20. They were part of the T-DNA cassette by flanking the promoter-reporter sequence as well as the hygromycin resistance gene. The 108 fragments were homologous to Pp108 locus and serve as a location for homologous recombination, therefore facilitating the insertion of the T-DNA cassette at this region. Since Pp108 is a neutral genomic locus, the insertion at this location will not create any unwanted or off target gene mutation.

Promoter entry clones were cloned into the empty pTHattR4R2 destination vector together with the GUS entry clone in Gateway LR reactions to generate *PromPpDUO1A:GUS* and *PromPpDUO1B:GUS* expression clones (Figure 4.4.1.4). The same approach was applied to create *PromPpDUO1A:NLS-GFP* and *PromPpDUO1B:NLS-GFP* expression clones from promoter and *NLS-GFP* entry clones

115

(Figure 4.4.1.5). All promoter constructs and plasmid structures were confirmed by diagnostic PCR and restriction enzyme analysis before being introduced into the wild type *P. patens* by protoplast transformation.
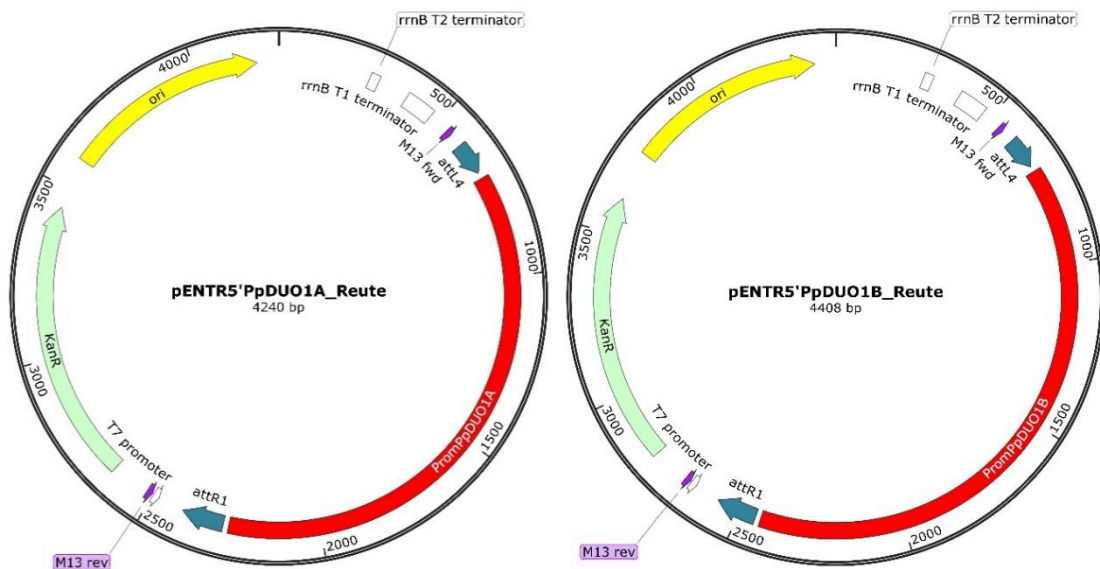
Figure 4.4.1.1. Promoter *PpDUO1A* and *PpDUO1B* entry clone. *PpDUO1A* and *PpDUO1B* promoter fragments were cloned into pENTR5' TOPO-TA vector using TOPO TA cloning. Labels on map: KanR, Kanamycin Resistance gene; ori, plasmid origin of replication; M13 fwd, M13 forward sequencing primer; M13rev, M13 reverse sequencing primer; attL4/attR1, recombination site for the Gateway® LR reaction; rrnB T1 terminator, transcription terminator T1 from the E. coli rrnB gene; rrnB T2 terminator, transcription terminator T2 from the E. coli rrnB gene.
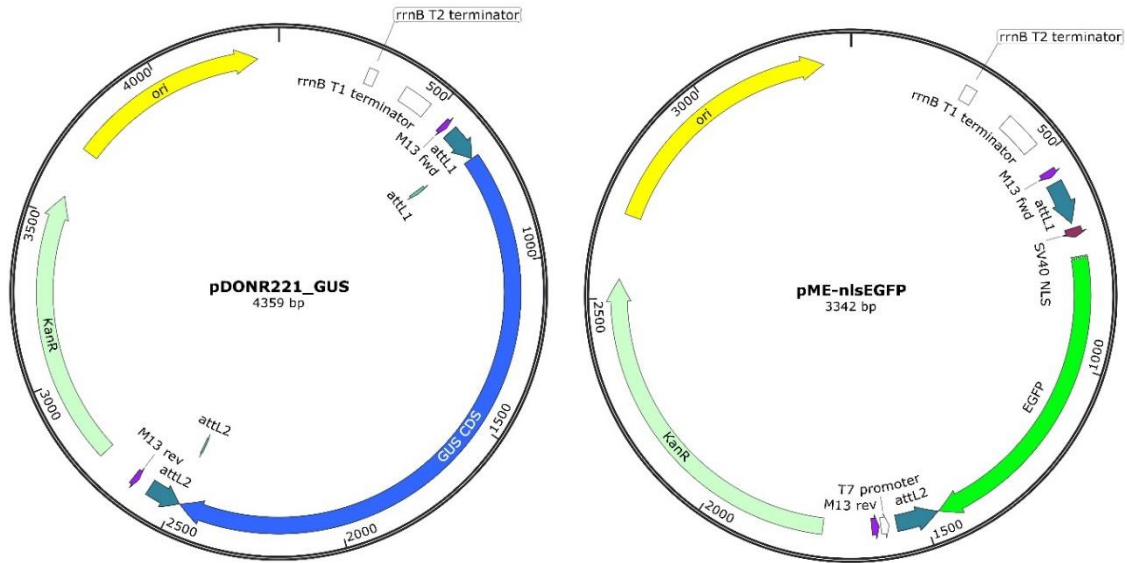
Figure 4.4.1.2. *GUS* and *NLS-GFP* **entry clones.** *GUS* **and** *NLS-GFP* **were cloned into pDONR221 and pME vector, respectively.** Labels on map: KanR, Kanamycin Resistance gene; ori, plasmid origin of replication; M13 fwd, M13 forward sequencing primer; M13rev, M13 reverse sequencing primer; attL1/attL2, recombination site for the Gateway® LR reaction; rrnB T1 terminator, transcription terminator T1 from the E. coli; rrnB gene, rrnB T2 terminator, transcription terminator T2 from the E. coli rrnB gene; SV40 NLS, Nuclear Localization Signal of SV40 (simian virus 40) large T antigen.
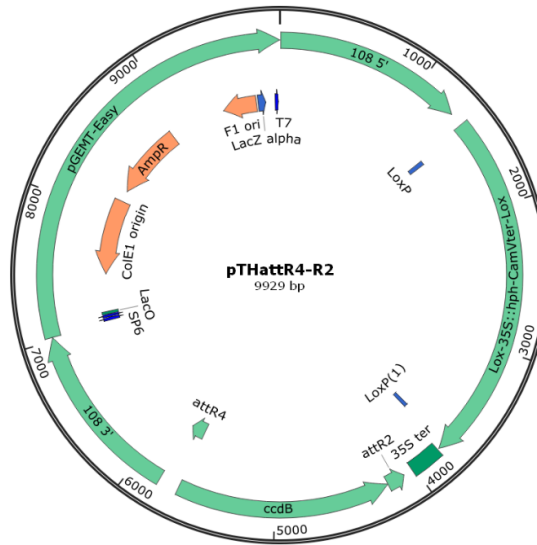
Figure 4.4.1.3. pTHattR4-R1 expression vector. pTHattR4-R2 was used in the Gateway Cloning LR reactions. Labels on map: AmpR, Ampicillin Resistance gene; 108 5', 5'end target sequence at neutral locus 108; 108 3', 3'end target sequence at neutral locus 108; LacO, lactose operon; SP6, promoter for bacteriophage SP6 RNA polymerase; T7, promoter for bacteriophage T7 RNA polymerase; ColE1 origin, ColE1 plasmid origin of replication; F1 ori, F1 phage origin of replication; attR4/attR2, recombination site for the Gateway® LR reaction; 35S ter, 35S terminator; LoxP/LoxP(1), locus of crossover on bacteriophage P1.
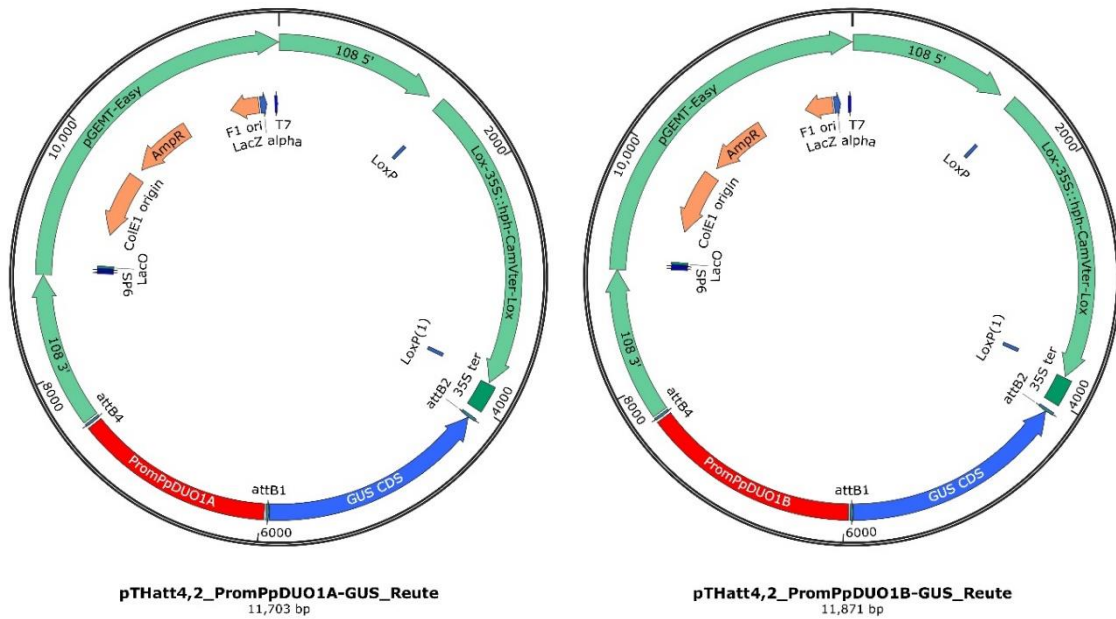
Figure 4.4.1.4. *PromPpDUO1A:GUS* and *PromPpDUO1B:GUS* **expression clones.** Promoter *PpDUO1A* and *PpDUO1B* were recombined with *GUS* coding sequence in pTHattR4-R2 vector using Gateway Cloning LR reactions. Labels on map: AmpR, Ampicillin Resistance gene; 108 5', 5'end target sequence at neutral locus 108; 108 3', 3'end target sequence at neutral locus 108; LacO, lactose operon; SP6, promoter for bacteriophage SP6 RNA polymerase; T7, promoter for bacteriophage T7 RNA polymerase; ColE1 origin, ColE1 plasmid origin of replication; F1 ori, F1 phage origin of replication; attB4/attB2, recombination site for the Gateway® BP reaction; 35S ter, 35S terminator; LoxP/LoxP(1), locus of crossover on bacteriophage P1.

Figure 4.4.1.5. *PromPpDUO1A:NLS-GFP* and *PromPpDUO1B:NLS-GFP* expression **clones.** Promoter *PpDUO1A* and *PpDUO1B* were recombined with *GFP* coding sequence with nuclear localizing signal in pTHattR4-R2 vector using Gateway Cloning LR reactions. Labels on map: AmpR, Ampicillin Resistance gene; 108 5', 5'end target sequence at neutral locus 108; 108 3', 3'end target sequence at neutral locus 108; LacO, lactose operon; SP6, promoter for bacteriophage SP6 RNA polymerase; T7, promoter for bacteriophage T7 RNA polymerase; ColE1 origin, ColE1 plasmid origin of replication; F1 ori, F1 phage origin of replication; SV40 NLS, Nuclear Localization Signal of SV40 (simian virus 40) large T antigen; attB4/attB2, recombination site for the Gateway® BP reaction; 35S ter, 35S terminator; LoxP/LoxP(1), locus of crossover on bacteriophage P1.

4.5 Discussion

4.5.1. Conservation of *DUO1* in bryophytes

In this chapter, the conservation of DUO1 in *P. patens* was explored using different approaches. First, through sequence analysis of DUO1 orthologs present in all known land plant genomes. From this analysis, new DUO1 orthologues were identified in bryophytes such as SfaDUO1A of *Sphagnum fallax*. This enabled a more thorough analysis compared with earlier studies (Sari, 2015 and Zhao, 2017) eventually highlighting conserved region between PpDUO1A and PpDUO1B and that were also conserved in other bryophyte DUO1 sequences. A long stretch, 37 amino acids, conserved region after the R3 MYB domain was found exclusively in bryophyte DUO1 sequences (Figure 4.2.1). Short acidic regions were also present after the conserved region only in bryophyte DUO1 orthologs, which could function as activation domains. Previous work in MpDUO1 has highlighted the functional importance of the C-terminal region (Zhao, 2017). A chimeric DUO1 protein (MpChimera), which combines the MpDUO1 MYB domain with the C-terminal region of AtDUO1, was able to transactivate the target of AtDUO1, *HTR10*. The transactivation signal however was very low when MpDUO1 was used to transactivate *HTR10* (Zhao, 2017; Higo et al., 2018). This difference in activation ability is primarily due to the C-terminal region. Moreover, sequences that mediate transactivation are yet to be identified in MpDUO1 (Zhao, 2017) and conserved regions in bryophyte DUO1 may be responsible.

The second approach that was used to explore *DUO1* conservation in bryophytes was by examining the expression profiles of *PpDUO1A* and *PpDUO1B* using RNA-seq data. This clearly showed that *PpDUO1A* is specifically expressed whereas *PpDUO1B* is preferentially expressed in antheridia. *PpDUO1B*, showed a more diverse expression which was relatively low in other developmental stages and highest in antheridia. Analysis of RNA-seq data for *MpDUO1* revealed antheridia-specific expression in *M. polymorpha* (Higo et al., 2016; Higo et al., 2018). In addition, *HmnDUO1* from *Haplomitrium mnioides*, a representative of the most basal liverworts, showed expression in antheridia (Higo et al., 2018). The similarity in protein sequence and expression pattern between *PpDUO1* and *MpDUO1* illustrate broad conservation among two major bryophyte lineages and support the hypothesis that their function might also be conserved.

4.5.2. PpDUO1 is essential for sporophyte formation in *P. patens*

The function of *DUO1* in *P. patens* was elucidated through the generation and analysis of *PpDUO1A* and *PpDUO1B* knockout mutants. Single mutant lines were created by introducing a selection cassette into *PpDUO1* locus by replacing the entire genomic sequence of the gene through homologous recombination. The selection marker was then removed by Cre-Lox recombination thereby creating *Ppduo1a$^\Delta$* and *Ppduo1b$^\Delta$* single mutant (Gilbertson, 2003). The co-expression in antheridia and similarity between *PpDUO1A* and *PpDUO1B* indicated a potentially redundant function, therefore, the single mutant lines were used to generate *Ppduo1a$^\Delta$b$^\Delta$* double mutant lines.

Phenotypic analysis showed that *Ppduo1a$^\Delta$* and *Ppduo1b$^\Delta$* single mutants were fertile and formed sporophytes, while the *Ppduo1a$^\Delta$b$^\Delta$* double mutant failed to do so. The number of sporophytes formed *by Ppduo1a$^\Delta$* and *Ppduo1b$^\Delta$* mutants was comparable to the wild type. This implies that *Ppduo1a$^\Delta$b$^\Delta$* double mutant was sterile in due to the disruption of both *PpDUO1* paralogues. Similarly, male sterility was also observed in *MpDUO1* mutants where Mp*duo1-1^{ko}* antheridiophores showed defective male gametogenesis (Higo et al., 2018).

RNA-seq expression data have shown that *PpDUO1A* is exclusively expressed in the antheridia while *PpDUO1B* is preferentially expressed in the same tissue. The presence of low levels of *PpDUO1B* in other tissues suggest it may have additional roles. Any defect in male gametogenesis would affect the fertility of *P. patens.* Collectively, the expression data together with evidence seen in *Mpduo1* strongly support that the inability to form sporophyte in *Ppduo1a$^\Delta$b$^\Delta$* mutant could be due to impaired male gametogenesis.

4.5.3. *PpDUO1* function in sperm cell differentiation

The possibility of disruption to male gametogenesis in *Ppduo1a$^\Delta$b$^\Delta$* was explored by analysing the morphology of spermatogenous cells in the mature antheridia. DAPI stained SpCs showed a dramatic change in morphology in the mature antheridia of *Ppduo1a$^\Delta$b$^\Delta$* double mutants. The nuclei in *Ppduo1a$^\Delta$b$^\Delta$* SpCs failed to transition from the spherical form in earlier stages to crescent-shaped. In addition, TEM analysis revealed an absence of flagella formation in the late stage of SpCs of *Ppduo1a$^\Delta$b$^\Delta$*. The basal body and multiple layer structure (MLS) present in the mid stage of SpCs failed to elongate to form flagella and spline structure during the late stage of SpCs. The absence of spline which serves as the backbone structure of SpCs is also linked to the lack of elongation of the nucleus and dense chromatin compaction was not achieved. This

shows that *PpDUO1* genes in *P. patens* have an important role in male germ cell morphogenesis and key aspects of spermatogenesis including cellular differentiation. In addition, the number of SpCs per antheridium was also calculated to investigate the effect of *Ppduo1a$^\Delta$b$^\Delta$* in male germ cell division. The number of SpCs was shown to be comparable to the wild type, which clearly demonstrates that despite the striking effect on germ cell morphogenesis, germ cell division is unaffected in the absence of functional copies of *PpDUO1*.

These experiments demonstrate that *PpDUO1* genes have a critical function in male gametogenesis in *P. patens*. The function however only affects morphogenesis and not the division of spermatogenous cells. A comparable role has been deduced from studies of *MpDUO1* in *M. polymorpha* (Higo et al., 2018). The Mp*duo1-1^{ko}* fails to undergo sperm differentiation, therefore preventing the rounded sperm cell nucleus from acquiring a crescent shape. No flagella and spline were formed in the Mp*duo1-1^{ko}* and the nucleus was unable to elongate and condense. The final sperm cell division in Mp*duo1-1^{ko}* was reported to occur normally as in wild type. Although no specific sperm cell counts were reported for Mp*duo1-1^{ko}*, it is clearly shown that the cells complete the characteristic diagonal sperm mother cell division. MpDUO1 was shown to be expressed just before the diagonal division of the sperm mother cell, yet the cell division does not appear to be affected in Mp*duo1-1^{ko}*. This is in accordance with the results seen in *Ppduo1a$^\Delta$b$^\Delta$*.

In conclusion, the role of DUO1 in bryophytes is that of a regulator of morphogenesis and differentiation and not of germ cell proliferation. This indicates that the role of DUO1 in the control of germ cell division is a derived trait of DUO1 that has evolved in angiosperms.

Chapter 5: Functional conservation of the DUO1-DAZ1 network in *Physcomitrella patens*

Abstract

Background and Aims

The MYB protein DUO1 controls the expression of *Arabidopsis* male germline-specific C_2H_2 zinc finger proteins *DAZ1* and *DAZ2* by directly binding to their promoter regions. DAZ1/DAZ2 are required for germ cell division and DUO1-dependent germ cell differentiation. The DUO1-DAZ1 regulatory module is conserved throughout flowering plants but is not thoroughly explored in bryophytes. In this chapter, the conservation of DUO1-DAZ1 network in *P. patens*, was investigated.

Methods

BLAST queries and sequence analyses were used to investigate sequence conservation of DAZ1 in bryophytes. Public RNA-sequence data was analysed to investigate the expression pattern of *PpDAZ1* in relation to *PpDUO1* transcripts. The promoter sequence of *PpDAZ1* was examined for the presence of DUO1 binding sites. In silico comparative analysis of antheridia-specific genes between moss and Marchantia was executed to identify other potential targets of PpDUO1. Orthologs of *MpDUO1* targets were identified using BLAST searches in various databases. Promoter sequence analysis was also performed to analyse the presence of DUO1 binding sites in other potential PpDUO1 target genes.

Key Results

Four DAZ1 paralogs are present in *P. patens*, *PpDAZ1A*-*PpDAZ1D*. Multiple sequence alignment showed that the PpDAZ1 family possesses the regions that are conserved in DAZ1 orthologs present in vascular plants. The PpDAZ1 family also shares additional features only with DAZ1 orthologs of other bryophytes, namely, an additional zinc finger, two additional basic regions and one additional acidic region. *PpDAZ1* genes are expressed in developing antheridia with peak expression at mature stage. DUO1 binding sites was present in the upstream regions of *PpDAZ1A*, *PpDAZ1B* and *PpDAZ1C*. The potential target genes of PpDUO1 possess DUO1 binding sites in their upstream sequence and are mostly related to microtubule and flagellar formation.

Conclusion

The work presented supports the conservation of the DUO1-DAZ1 network in *P. patens* and indicates that *PpDAZ1* genes are directly regulated by PpDUO1.

5.1 Introduction

DUO1-ACTIVATED ZINC FINGER1 (*DAZ1*) gene and its paralog, *DAZ2*, are C_2H_2-type zinc finger transcription factor genes, that were identified as DUO1-activated target (DAT) genes (Borg et al., 2011). They were shown to be induced by ectopic expression of DUO1 in seedlings and exhibited DUO1-dependent expression in a *duo1* background. The DUO1-DAZ1 network was first described in Borg et al. (2014). Through RT-PCR analysis, *DAZ1* and *DAZ2* were shown to be specifically expressed in pollen and absent in other sporophytic tissue. Both *DAZ1* and *DAZ2* had peak expression in tricellular pollen. Protein expression of DAZ1 and DAZ2 fused to mCherry was also detected in early sperm cell development, located primarily in nucleus for DAZ1 and exclusively in nucleus for DAZ2.

The conservation of the DUO1-DAZ1 network in *Marchantia polymorpha* was recently described in Higo et al. (2018). The qPCR analysis of *MpDAZ1*, the *M. polymorpha* ortholog of *DAZ1*, showed that its expression was significantly reduced in antheridia of *Mpduo1* mutants. This demonstrated the requirement of MpDUO1 for the expression of *MpDAZ1*, therefore suggesting the conservation of DUO1-DAZ1 regulatory module among land plants (Higo et al., 2018).

The conservation of the DUO1-DAZ1 network is further explored here in *P. patens*. In this chapter, *DAZ1* homologues in bryophytes, specifically *PpDAZ1* genes, were identified and characterised. Multiple sequence analysis was performed to investigate the level of sequence conservation between the homolog sequence. The expression of *PpDAZ1* transcripts were explored at major developmental stages in *P. patens* and their expression profiles compared that of the two *PpDUO1* paralogs. The presence of DUO1 binding sites (DBS) was also examined in *PpDAZ1* genes promoters to identify potential direct targets of PpDUO1. The spatio-temporal expression of *PpDAZ1* genes were initiated by building promoter-reporter gene fusions and plant transformation.

5.2 *PpDAZ1A-D* – a family of PpDUO1 target genes in *P. patens*

5.2.1 Analysis of *PpDAZ1* homologs in bryophytes

To identify *DAZ1* orthologs in bryophyte species, the AtDAZ1 sequence (NCBI Reference Sequence: NP_179309.1; Uniprot: Q9SIJO) was used in BLASTP search in three different plant databases, as mentioned in Chapter 4. The search was conducted according to the method described in subchapter 2.7.3 and a quick outline was presented in Figure 5.2.1.1.
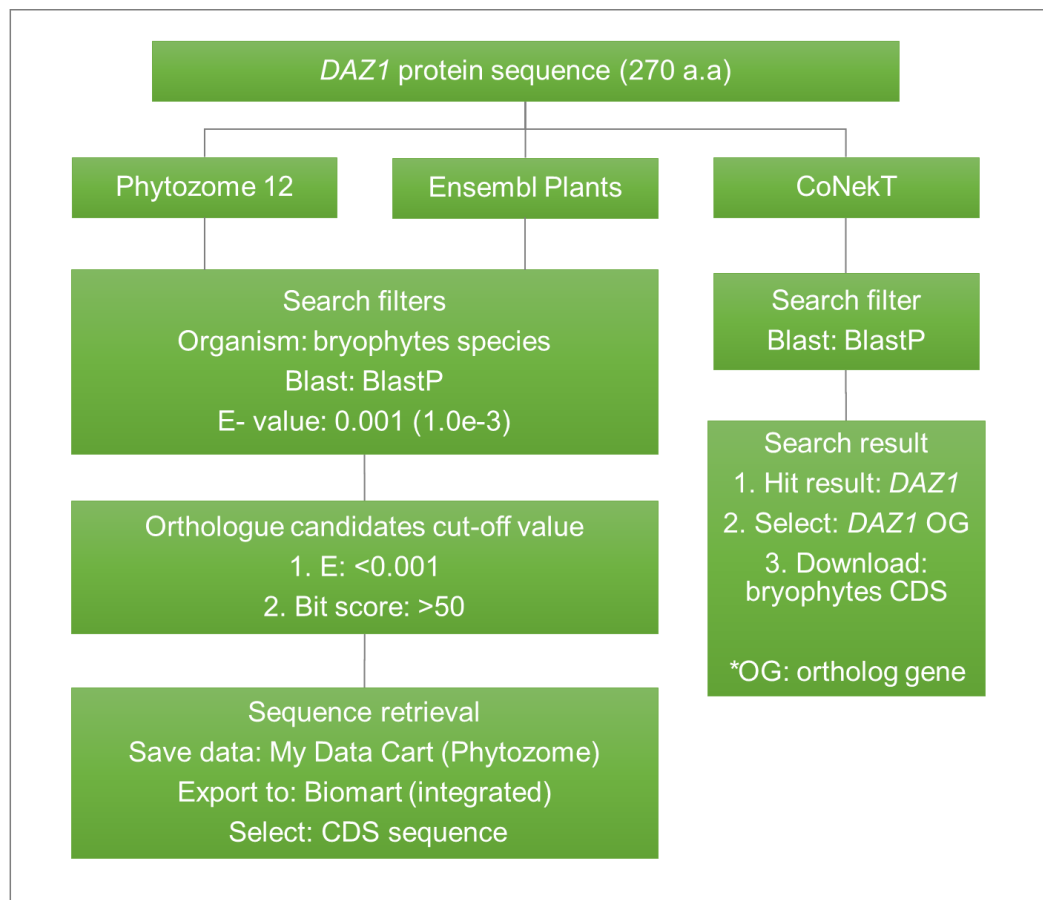
Figure 5.2.1.1. Bioinformatic flow diagram for identifying *PpDAZ1* homologues. AtDAZ1 sequence was used in a BLASTP search in three different plant database platforms. CDS sequences were retrieved for the candidate of the orthologous genes.

Similar to the sequence analysis described in Chapter 3 and 4, all retrieved sequences were combined into a list using MEGA 10 (Kumar et al., 2016). The list was manually checked and purged of any replicated sequence entries resulting from annotation of the same sequence by different databases. In addition, sequences of few representative

species for each major group of land plants were included as a point of comparison between homologues in bryophytes and tracheophytes. Among them were *SmoDAZ1A* (*Selaginella moellendorffii*), *PabDAZ1* (*Picea abies*), *AtrDAZ1* (*Amborella trichopoda*), *ZmDAZ1A* (*Zea mays*) and *AtDAZ1* (*Arabidopsis thaliana*) (Figure 5.2.1.2). Multiple sequence alignment was made using MUSCLE by selecting 'align codons' and maintaining all default algorithm parameters (Edgar, 2004). In order to view the protein alignment and perform protein sequence analysis, translated protein sequences tab was selected. The reliability of the multiple sequence alignment generated was checked by performing pairwise comparison for average amino acid identity. The percentage amino acid identity was 46.3 % which was above the 30 % threshold for reliable alignments. For generating a superior graphical image, the protein alignment was then exported into CLC Sequence Viewer (QIAGEN Bioinformatics https://digitalinsights.qiagen.com/).

Among several bryophyte species ten orthologs of *AtDAZ1* were identified, including nine moss and one liverwort sequence. In mosses, there were four genes in *P. patens* (*PpDAZ1A-D*) and five genes in *Sphagnum fallax*. The PpDAZ1A-D protein organisation was similar to that of SfaDAZ1A-C, SmoDAZ1A, AtrDAZ1 and AtDAZ1 structures where three zinc finger domains were present (ZF2, ZF3 and ZF4) as well as one EAR motif (EAR) near the C-terminal end (Figure 5.2.1.3 and Figure 5.2.1.4). In addition, one basic region (BR1) adjacent to ZF2 was also present in PpDAZ1A-D and conserved among other bryophyte DAZ1 sequences, SmoDAZ1A, AtrDAZ1, SlDAZ1 and AtDAZ1 (refer Appendix, Figure S5.1). There was also a conserved region (CR2), highly comprised of hydrophobic amino acids, between ZF2 and ZF3 which was present in all aligned species. However, some features were conserved in all PpDAZ1 orthologs and shared among bryophytes (Figure 5.2.1.3). These features were an additional zinc finger domain (ZF1) before basic region 1 (BR1) and the presence of two additional basic regions (BR2 and BR3) between ZF2 and CR2. In addition, a stretch of conserved, primarily acidic amino acids (CR1), was present between BR2 and BR3.

PpDAZ1A-D are C_2H_2-type zinc finger proteins (ZFP) that possess the H3XH spacing pattern between the histidine residues in all four ZF domains. Therefore, they belong to the C1 family according to the criteria used by Englbrecht et al. (2004). Since all PpDAZ1 paralogues had four ZFs, they were further classified in the C1-4i subfamily where '4' represents the number of ZF and 'i' is the acronym for ZF. Each ZF domain was then assigned according to the motif variant in the alpha helix position. ZF2 has a KALFGH motif, characteristic of K2-2 type ZFs (Table 5.2.1.1). ZF3 and ZF4 both possess a QALGGH motif but different amino acid residues in the rest of the helix positions, which categorizes as Q2-2 and Q2-3 helix signature types, respectively. On the other hand, the

ZF1 motif does not belong to any known helix signature type. There is slight variation between ZF1 motifs of PpDAZ1 paralogs. PpDAZ1B and PpDAZ1C have KSLNLH, whereas PpDAZ1D show a single amino acid difference (KSLNFH). Meanwhile, PpDAZ1A show three amino acid differences in ZF1 (KRLHSH) compared to PpDAZ1B and PpDAZ1C (Figure 5.2.1.3).
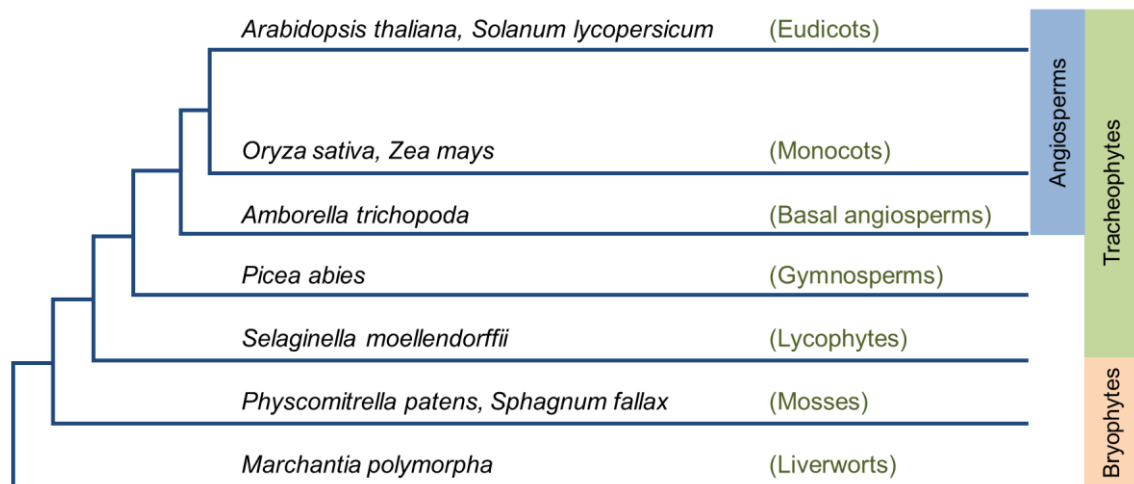
Figure 5.2.1.2. Representative species for each major group of land plants. Sequences from representative species were included in the PpDAZ1 sequence analysis as a point of comparison between homologues in bryophytes and tracheophytes.
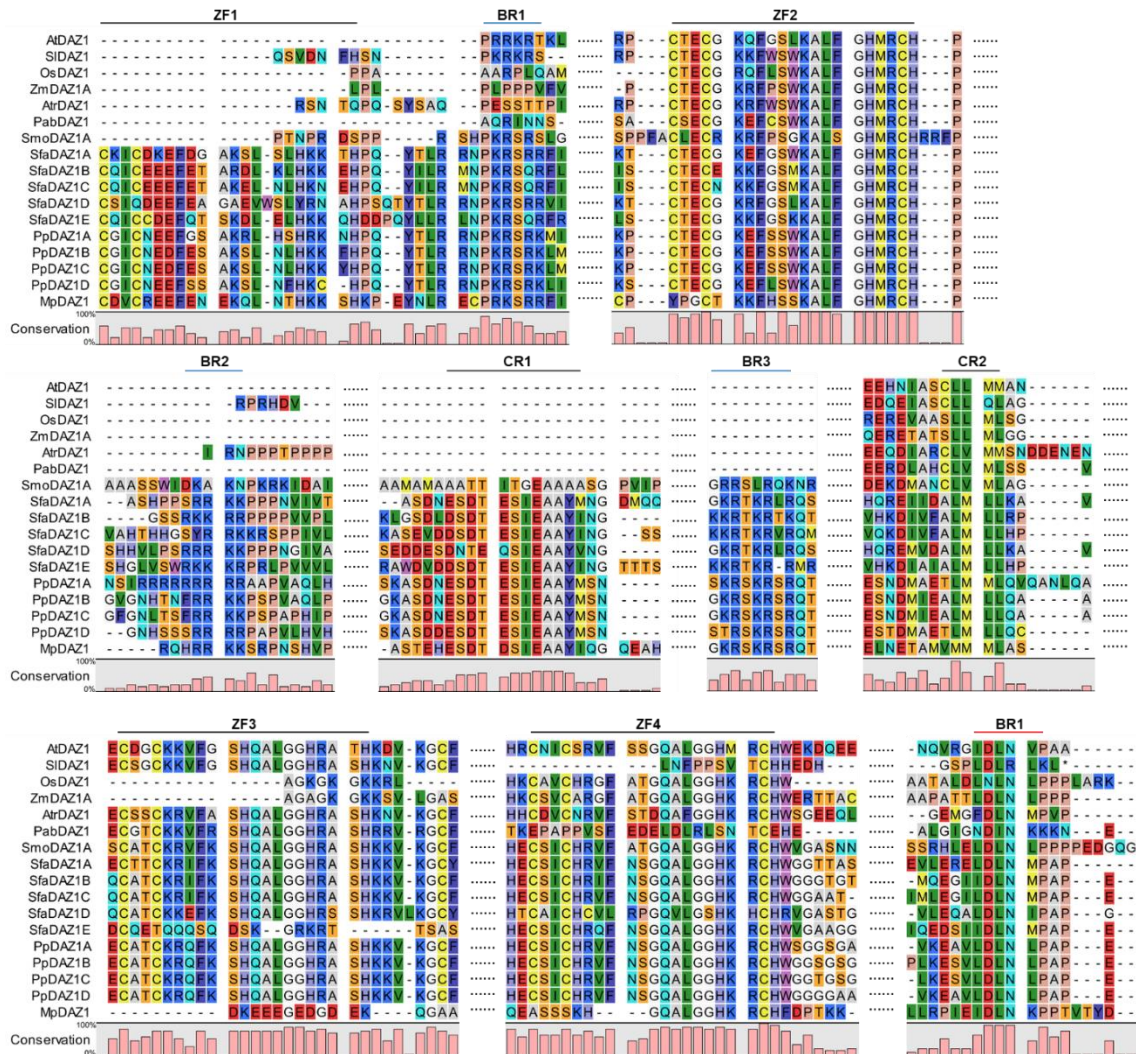
Figure 5.2.1.3. Sequence alignment of PpDAZ1 homologues in bryophytes and representative species of other major land plant groups. Protein sequences were aligned in MEGA 10 with MUSCLE tool using default parameter and later transported to CLC Sequence Viewer for superior image. ZF, zinc finger; BR, basic region; EAR, Ethylene-responsive element binding factor-associated amphiphilic repression. Species: At, *Arabidopsis thaliana*; Sl, *Solanum lycopersicum*; Os, *Oryza sativa*; Zm, *Zea mays*; Atr, *Amborella trichopoda*; Pab, *Picea abies*; Smo, *Selaginella moellendorffii*; Sfa, *Sphagnum fallax*; Pp, *Physcomitrella patens*; Mp, *Marchantia polymorpha*.

Table 5.2.1.1. C$_2$H$_2$ zinc finger domains alpha helix signature types.

Protein	Amino acids	α-helix signature type				EAR Type
		ZF1	ZF2	ZF3	ZF4	
PpDAZ1A	715	Other	K2-2	Q2-2	Q2-3	DLNXP
PpDAZ1B	849	Other	K2-2	Q2-2	Q2-3	DLNXP
PpDAZ1C	842	Other	K2-2	Q2-2	Q2-3	DLNXP
PpDAZ1D	760	Other	K2-2	Q2-2	Q2-3	DLNXP

Protein domain maps for PpDAZ1A-D indicating the conserved regions are shown in Figure 5.2.1.3. PpDAZ1A (715 a.a) was the shortest among the four paralogues, followed by PpDAZ1D (760 a.a), PpDAZ1C (842 a.a) and PpDAZ1B (849 a.a). Overall, the predicted proteins were closely similar in length. The PpDAZ1 family had a long region prior to the ZF1, representing on average 29 % of the whole structure. The region however was shorter in PpDAZ1A compared to the other paralogues, causing the coordinates (location) of the ZF1 more upfront than the other three proteins (Appendix, Table S5.1). Meanwhile, the ZF1 coordinates for PpDAZ1B-D were near, or in a close range to each other. The coordinates of ZF1 for PpDAZ1C and PpDAZ1D were so close, that they seem to be located at a similar position (Figure 5.2.1.3). The region after ZF4 to C-terminal end, which contains the EAR motif, was also relatively short in PpDAZ1A and PpDAZ1D, 75 and 80 amino acids, respectively. By comparison, PpDAZ1B and PpDAZ1C had a long amino acid residues for the said region, two times longer than the other two PpDAZ1. Interestingly, regardless of the long residues, the coordinates of the EAR motif for PpDAZ1B and PpDAZ1C was still relatively similar to PpDAZ1D. The long region after ZF4 for PpDAZ1B and PpDAZ1C did not make the EAR motif to be positioned further down the protein structure. From protein structure and alignment, PpDAZ1A clearly had a longer BR2 than PpDAZ1B-D by 5 amino acids (Figure 5.2.1.3; Figure 5.2.1.4).
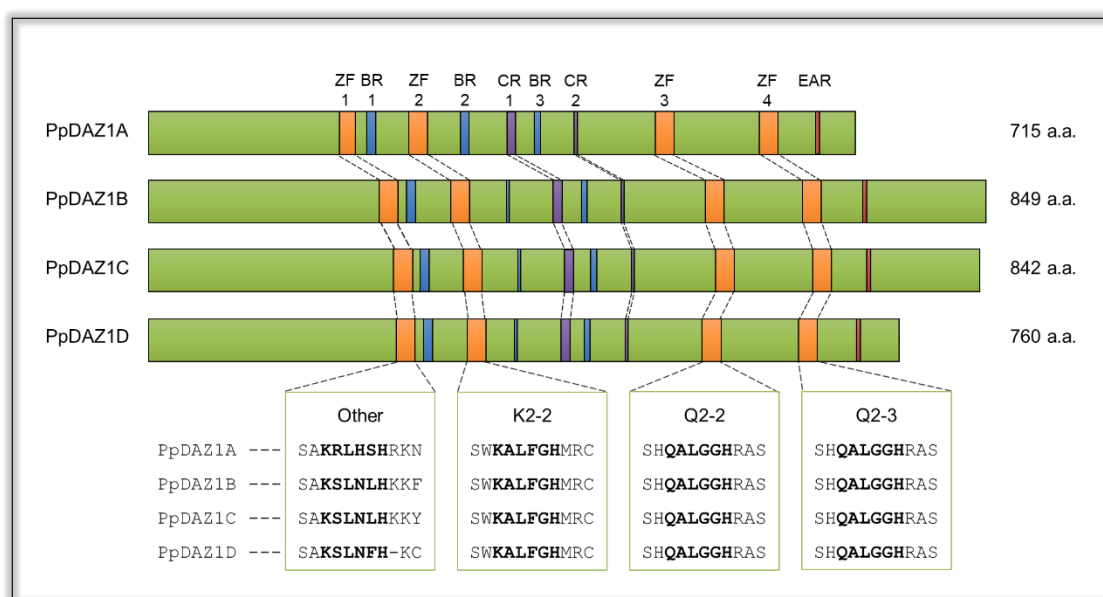
Figure 5.2.1.4. Protein domain maps for the PpDAZ1 family. Zinc finger domains with signature amino acid sequences are shown. Proteins are drawn to scale. Conserved domains are colour coded. Orange, zinc finger domain; blue, basic region; purple, conserved region; red, EAR motif.

The percentage identity between pairs of PpDAZ1 paralogues was calculated from MEGA 10 alignments (Table 5.2.1.2). The highest percentage identity (87.3 %) was between PpDAZ1B and PpDAZ1C, followed by 74.3 % for PpDAZ1A and PpDAZ1D, while other pairwise comparisons were in the range of 70 - 71 % identity. From the analysis of protein domain organisation and percentage identity, it is suggested that all PpDAZ1 proteins may have a similar function. The conserved regions shared between the PpDAZ1 family with other bryophytes and tracheophytes species, in particular AtDAZ1, also indicates that the protein function could be conserved. However, the presence of an additional ZF (ZF1), an acidic conserved region, CR1, and two additional basic regions only within the bryophytes group highlights potential differences in DAZ1 function between bryophytes and tracheophytes.

Table 5.2.1.2. Percentage protein identity between PpDAZ1 family. Amino acids pairwise comparison was generated using MEGA 10 to calculate the proportional (*p*) distance between two PpDAZ1 proteins. Protein identity = 1 – *p*-distance.

	Pairwise amino acid identity (%)			
	PpDAZ1A	PpDAZ1B	PpDAZ1C	PpDAZ1D
PpDAZ1A	100.0			
PpDAZ1B	71.2	100.0		
PpDAZ1C	70.7	87.8	100.0	
PpDAZ1D	74.3	70.7	70.3	100.0

5.2.2 *PpDAZ1* gene structure and characterisation

The conservation of gene structure between *PpDAZ1* paralogues was studied by examining exon-intron structure as well as the presence of potential conserved regulatory regions (further analysis for upstream sequence in section 5.3). Other features such as chromosome location, orientation and surrounding intergenic regions are also discussed. This was performed briefly as follows. My Data Cart tool in Phytozome 12 was utilised to gather information related to *PpDAZ1A-D*, for example current gene annotation, exons-introns length, chromosome number and gene coordinates (Figure 5.2.1.1). This information was cross checked with the *PpDAZ1A-D* information obtained from Ensembl Plants. In addition, JBrowse tool was used to view the physical map location of *PpDAZ1* genes and flanking intergenic regions (Figure 5.2.2.1).
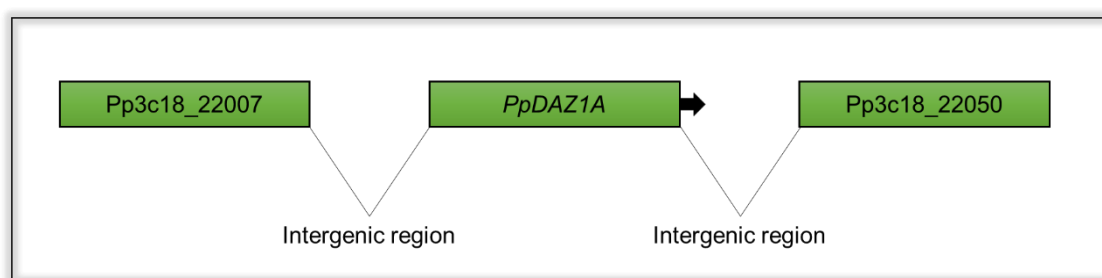
Figure 5.2.2.1. General schematic diagram for gene structure. *PpDAZ1A* and the neighbouring genes are shown as green boxes. The arrow indicates the direction of the gene on the chromosome, either forward or reverse strand.

PpDAZ1A and *PpDAZ1C* were located in the forward strand (strand: 1) of chromosome 18 and 19, respectively (Table 5.2.2.1). As for *PpDAZ1B* and *PpDAZ1D*, their location was on the reverse strand (strand: -1) of chromosome 22 and 21. The upstream intergenic region of *PpDAZ1* genes ranges from 175 bp to 25.6 kb, while the downstream intergenic region ranges from 75 bp to 33.75 kb (Table 5.2.2.1).

All *PpDAZ1* genes had three exons and two introns split in a similar pattern (Figure 5.2.2.2). The ZF1 and ZF2 coding sequences were found in exon 2 while ZF3 and ZF4 were in exon 3. The length of non-coding exon varies between the paralogues, such that *PpDAZ1C* had the longest 5' and 3' UTR whereas *PpDAZ1B* had the shortest for both. In summary, conservation of all intron/exon boundaries was observed in all *PpDAZ1* paralogues although there was some slight variation in the length of the UTR region. To examine potential differences in expression between *PpDAZ1* paralogues an upstream sequence analysis was conducted to attempt to discover the presence of any regulatory element or sequence that could influence their transcriptional profile.

Table 5.2.2.1. General information of *PpDAZ1* family with the latest gene annotation.

Gene	Gene annotation	Chromosome	Strand	Coordinates
PpDAZ1A	Pp3c18_22010V3.1	Chr18	1	15331889 – 15337574
PpDAZ1B	Pp3c22_2200V3.1	Chr22	-1	1464704 – 1470279
PpDAZ1C	Pp3c19_21390V3.1	Chr19	1	14191211 – 14200926
PpDAZ1D	Pp3c21_440V3.1	Chr21	-1	323755 – 331222

Table 5.2.2.2. Intergenic regions between *PpDAZ1* family and neighbouring genes.
Length was expressed in kilo base pairs (kb) for more than 1 kb and base pairs (bp) for less than 1 kb. '~' refers to approximately.

Left gene	Intergenic region	Gene	Intergenic region	Right gene
Pp3c18_22007	~ 175 bp	*PpDAZ1A*	20 kb	Pp3c18_22050
Pp3c22_2260	12.5 kb	*PpDAZ1B*	33.7 kb	Pp3c22_2140
Pp3c19_21290	~ 25.6 kb	*PpDAZ1C*	~ 1.8 kb	Pp3c19_21410
Pp3c21_520	3.7 kb	*PpDAZ1D*	~ 75 bp	Pp3c21_430
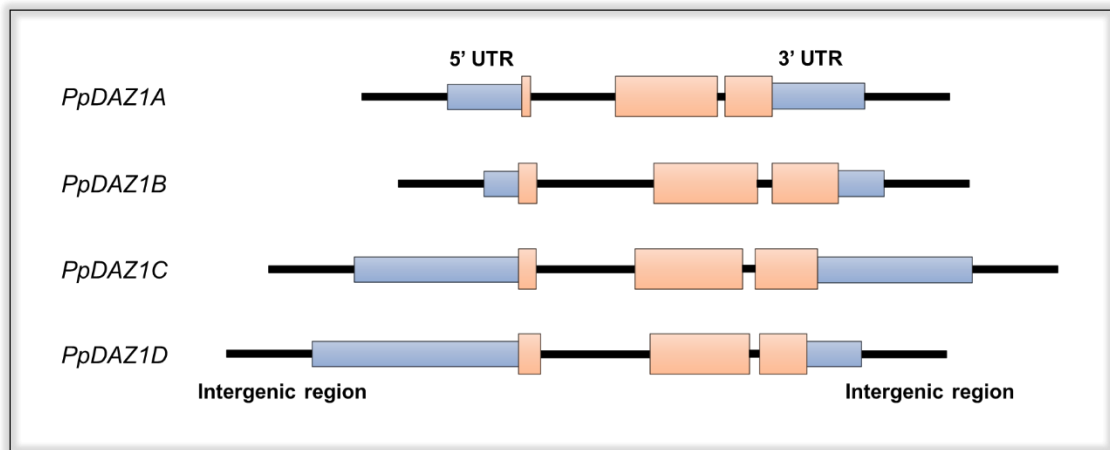
Figure 5.2.2.2. *PpDAZ1A-D* gene structure. All *PpDAZ1* genes possess three exons and two introns split in a similar pattern. Zinc finger 1 and 2 were in exon 2 while zinc finger 3 and 4 were in exon 3.

5.2.3 Expression of the *PpDAZ1* family in *P. patens*

The level of conservation among the *PpDAZ1* paralogues was further investigated by exploring their transcript expression pattern. RNA-seq data from various *P. patens* major developmental stages were downloaded from CoNekT platform, https://evorepro.sbs.ntu.edu.sg/, and used to construct a graphical expression profile (Figure 5.2.3.1). The expression threshold was set at three transcripts per million (TPM). From the graph, only *PpDAZ1B* transcript was detected in the protonema and leaflet stages. Significant expression was not detected in the archegonia and the younger stage of antheridia (9 to 11 DAI). However, transcript expression was detected for *PpDAZ1B* in the mature stage of antheridia (14 to 15 DAI), approximately 1.3 to 2-fold higher than protonema and leaflet stages. The *PpDAZ1A-D* transcript was not significantly expressed in the young sporophyte (9 DAF). Nevertheless, in the mature stage of sporophyte (16 DAF), three of *PpDAZ1* genes, *PpDAZ1B-D*, were significantly expressed. *PpDAZ1D* has the highest transcript expression in this stage, followed by *PpDAZ1B* and *PpDAZ1C*.

Additional RNA-seq data was also collected from a recently published study that compare the expression profiles of the mature antheridia between the Reute and Gransden accessions (Meyberg et al., 2020). For the *PpDAZ1* genes, transcript expression was detected in *PpDAZ1A*, *PpDAZ1B* and *PpDAZ1C* for both accessions (Table 5.2.3.1). *PpDAZ1B* had the highest expression among the three *PpDAZ1* genes.

This was followed by *PpDAZ1C* which expression was slightly higher than *PpDAZ1A* in Reute. This pattern however was reversed in Gransden.

Since *DAZ1* is a direct target of DUO1 in Arabidopsis*,* the expression profiles of the *PpDUO1* and *PpDAZ1* genes were compared in *P. patens*. As mentioned in Chapter 4, *PpDUO1A* was exclusively expressed, whereas *PpDUO1B* was preferentially expressed, in the male reproductive organ. Therefore, the best way to explore the expression profile of both of the gene families was by looking at antheridia RNA-seq data (Meyberg et al., 2020). Based on the RNA-seq dot plot graph, the transcript expression of the *PpDUO1* genes was much higher compared to the transcript expression of the *PpDAZ1* genes for both accessions (Figure 5.2.3.2). The expression profile for *PpDUO1A* and *PpDUO1B* was more than 6 and 7-fold, compared to *PpDAZ1* genes in Reute (Table 5.2.3.2).

Based on their expression profiles in *P. patens*, the *PpDAZ1* genes are expressed in various stage of development. At least one paralog, *PpDAZ1B*, was expressed during early (protonema and leafy) stage. Most of the paralogs, *PpDAZ1B-D*, were expressed in the mature (antheridia) to late (sporophyte) stage of development. *PpDUO1* genes were expressed at high levels in the antheridia stage of development. This coincided with the switching on of most *PpDAZ1* genes in antheridia. This could suggest the potential regulation of *PpDAZ1* genes by *PpDUO1*, similar to that reported for *MpDAZ1* and *MpDUO1* (Higo et al., 2018).
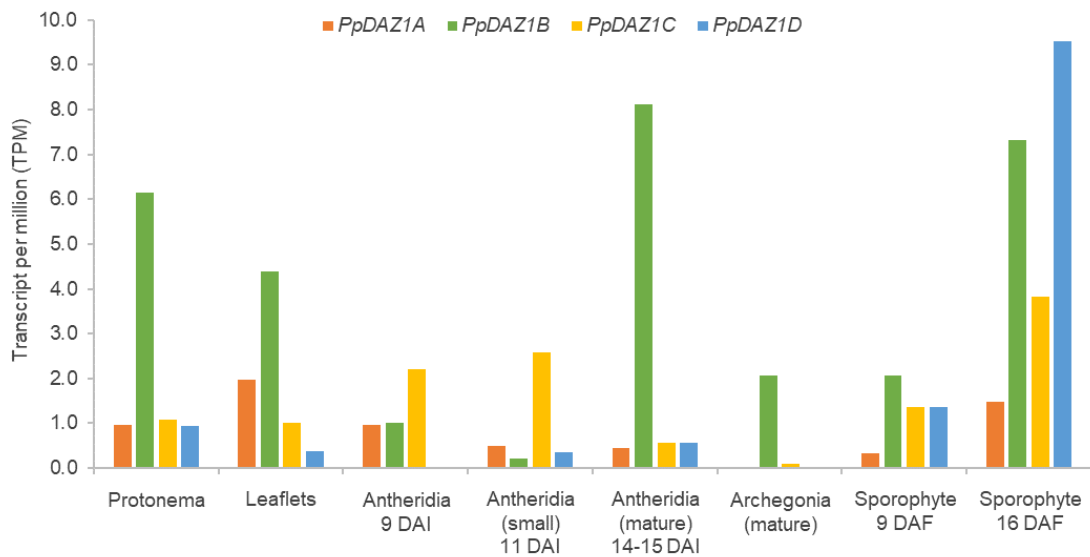
Figure 5.2.3.1. Developmental expression profiles of *PpDAZ1* paralogs in *P. patens* (Gransden). RNA-seq data representing major developmental stages of development was collected from CoNekT (Proost and Mutwil, 2018). The expression data is in transcripts per million (TPM).
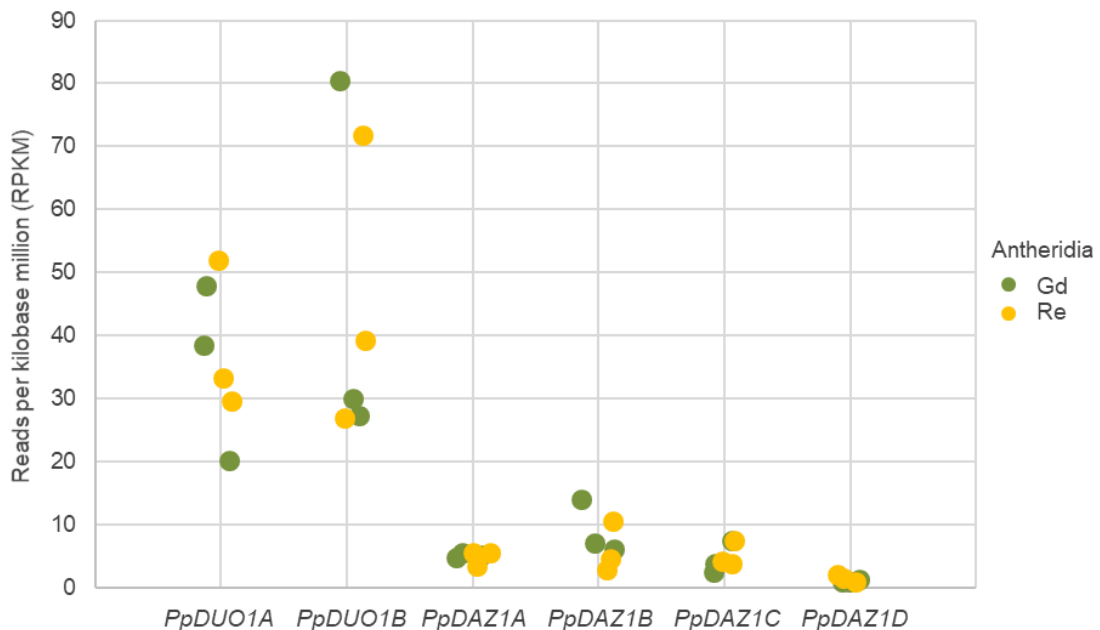
Figure 5.2.3.2. RNA-seq data for *PpDAZ1* in antheridia tissue from two different ecotypes. The expression data was collected from recently published study that focus on the difference in transcript expression pattern of the antheridia tissue from Reute and Gransden accessions (Meyberg et al., 2020). In addition, the *PpDUO1* transcript was also included to infer a *PpDUO1-PpDAZ1* relationship. Gd, Gransden and Re, Reute.

Table 5.2.3.1. Transcript expression (mean) in antheridia and fold difference.

Antheridia (RPKM)			
Gene	Gransden (Gd)	Reute (Re)	Fold (Gd/Re)
PpDUO1A	35.42	38.22	0.93
PpDUO1B	45.79	45.92	1.00
PpDAZ1A	5.10	4.71	1.08
PpDAZ1B	8.99	5.86	1.53
PpDAZ1C	4.45	5.06	0.88
PpDAZ1D	1.00	1.41	0.71

Table 5.2.3.2. Fold change between *PpDUO1* and *PpDAZ1* genes in Reute ecotype.

Fold_Reute (*PpDUO1*/*PpDAZ1*)		
Gene	*PpDUO1A*	*PpDUO1B*
PpDAZ1A	8.11	9.75
PpDAZ1B	6.52	7.83
PpDAZ1C	7.55	9.07
PpDAZ1D	27.03	32.48

5.3 In-silico analysis of DUO1 binding sites (DBS) in *PpDAZ1* promoters

5.3.1 Conserved motifs in putative upstream regions of *PpDAZ1*

The presence of conserved motifs and putative cis-regulatory sequence in *PpDAZ1* family were further explored by comparing their upstream sequence with other *DAZ1* orthologs. Sequences from orthologs of species that represent the major plant clades such as *Selaginella* for lycophytes, *Amborella* for basal angiosperms, rice and maize for monocots, Arabidopsis and tomato for eudicots, along with other bryophyte, *Marchantia*, were included in the analysis. Up to 1 kb of sequence upstream of the translation start codon was analysed using the Motif Discovery tool in MEME Suite 5.1.1 (Bailey et al., 2009). The tool allows the user to determine the number of motifs and the analysis was conducted with 10 motifs, to maximise the number of outcomes.

According to the parameters 10 consensus motifs were found among the analysed *DAZ1* sequences. The number of sites however varied, from the lowest, 4, to the highest, 13. This means a motif consensus with 13 sites was present in all sequences analysed,

whereas a motif with 4 sites was only present in 4 sequences. Those with more than 5 sites were selected for further analysis (Figure 5.3.1). Only motif 5 was present in all upstream sequences tested, while motifs 1 and 2 were found in 8 and 7 upstream sequences, respectively. Meanwhile motif 3, 4 and 6 were found in 6 upstream sequences. Motifs 2 and 6 were present in bryophytes (*MpDAZ1*, *PpDAZ1*) and *Selaginella* (*SmoDAZ1*), while motifs 3 and 4 were only present in *PpDAZ1* and *SmoDAZ1*.
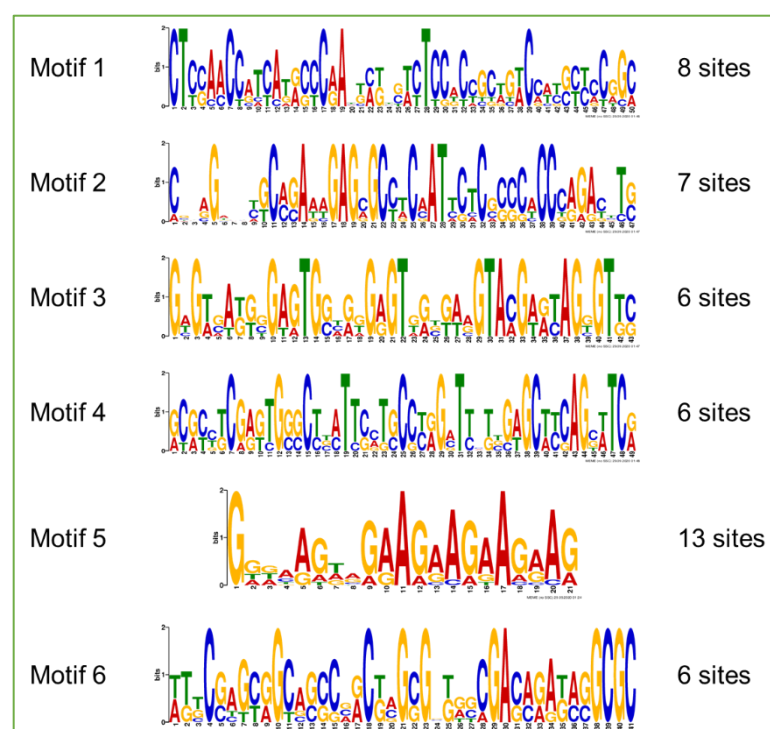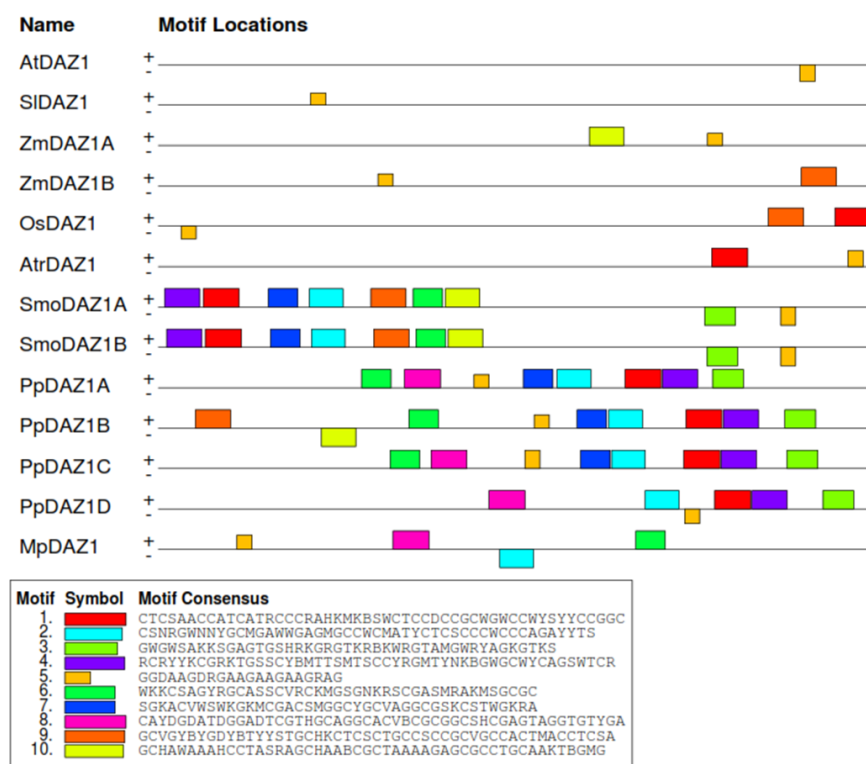
Figure 5.3.1. Conserved motifs in *PpDAZ1* genes upstream sequences. 1000 bp sequence upstream of the start codon were analysed for the presence of conserved motifs in *PpDAZ1* and few species representing the major land plants groups. The analysis was performed using Motif Discovery tool in MEME Suite 5.1.1 (Bailey et al., 2009). 10 motif consensus are shown in the motif location map and 6 motif consensus

that present in more than 5 sites were shown with their PSSM logo. The number of sites detected are shown on the right of the PSSM logo.

Motif 5 which is present in all sequences analysed, and motif 2 and 6 which conserved in bryophytes and lycophytes, were selected for further analysis using TOMTOM tool in order to identify any putative regulators (Gupta et al., 2007). The motifs were compared against database, JASPAR (non-redundant) DNA Core (2018) for plant (Khan et al., 2018). Match with the lowest p-value was chosen to avoid the probability of random motif from the database to align with our query motif. Motif 2 was significantly aligned with the binding site of ERF/AP2 transcription factor, *ERF027* (p-value=1.22e-03). Meanwhile for motif 5, the motif was matched with the binding site of C_2H_2 zinc finger transcription factor, DOF3.4 (p-value= 1.12e-05). Motif 6 was hit matched with the binding site of transcriptional activator *BPC5* (p-value= 5.11e-04). In summary, this analysis had identified the conserved motif in the upstream sequence of some *PpDAZ1* homologue where one of the motifs was conserved in all species while others in more closely related species. The analysis however had failed to identify the presence of the DUO1 binding motif (Higo et al., 2018).

5.3.2 DUO1 binding motifs in *PpDAZ1* promoters

To investigate the potential direct regulation of *PpDAZ1* family by *PpDUO1*, the promoter region of *PpDAZ1A-D* were analysed for the presence of DUO1 binding sites (DBS). Upstream region of 2000 bp were retrieved for all *PpDAZ1* genes using the BioMart tool integrated in Phytozome 12 (Figure 5.3.2.1). 'Flank-coding region (Gene)' was chosen in the sequence section as this option includes the sequence upstream of the ATG translation start site and includes the 5'UTR. The retrieved sequences were used to scan for the presence of DBS using Regulatory Sequence Analysis Tools (RSAT) (Nguyen et al., 2018). The dna-pattern tool in RSAT was used to pattern match the retrieved sequences to the DBS consensus RRCSGTT as described in Higo et al. (2018). Lastly, feature map tool was used to visualize the location of DUO1 binding sites (Figure 5.3.2.2).
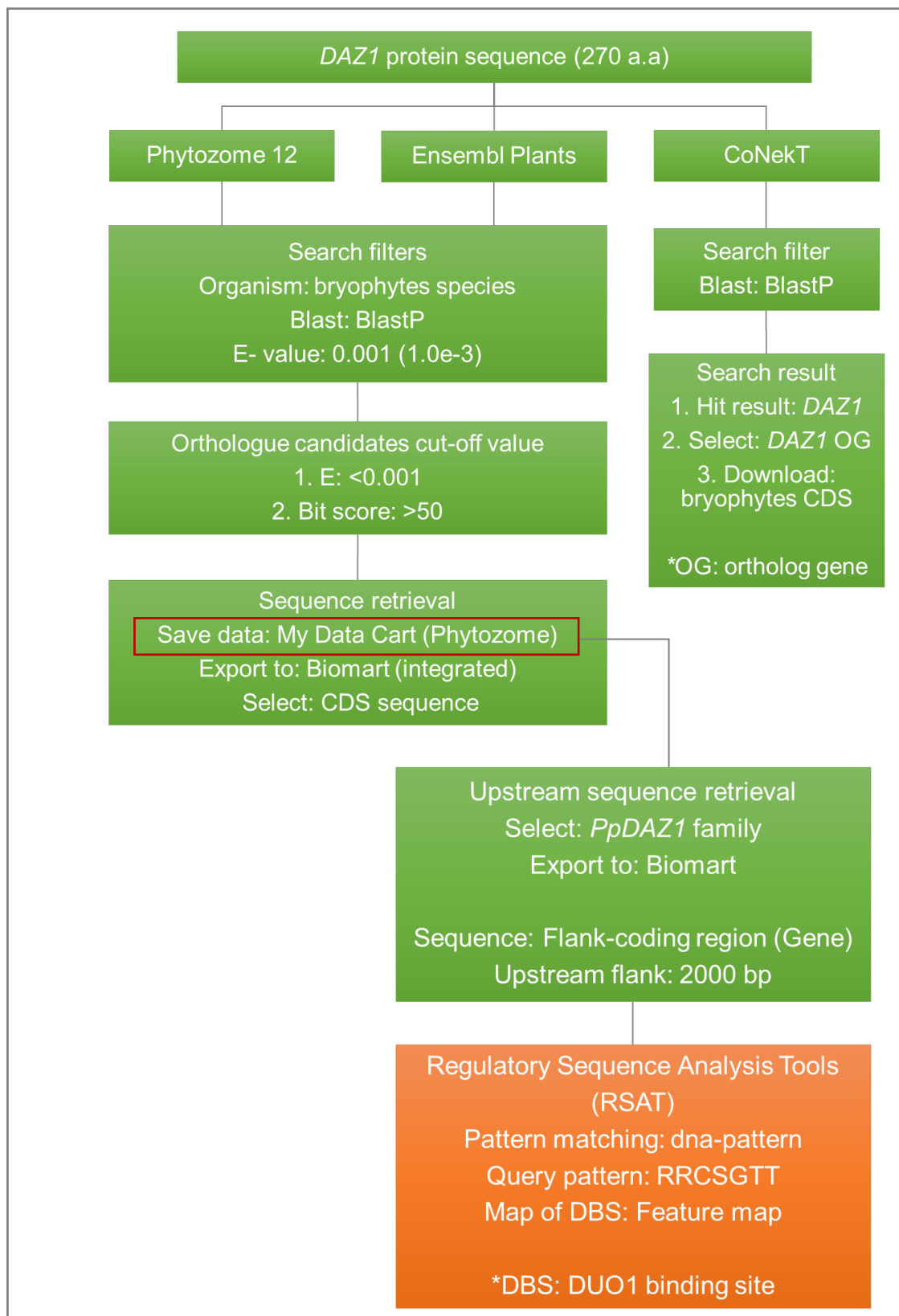
Figure 5.3.2.1. Workflow to identify DUO1 Binding Site (DBS) in upstream of *PpDAZ1* genes. Sequence upstream of ATG start codon was retrieved using Biomart tool and uploaded in RSAT.

One DBS was found in each *PpDAZ1A* and *PpDAZ1C* upstream region at coordinates -521 to -515 and -438 to -432, respectively (Table 5.3.2.1). Meanwhile, two DBSs were found in *PpDAZ1B* at positions -949 to -943 and at -440 to -434. All DBS found were in close proximity, less than 600 bp, to the codon start site except for the second DBS in *PpDAZ1B*. No DBS was found in the 2000 bp upstream region of *PpDAZ1D*. The presence of DBS in the upstream region of *PpDAZ1A-C* supports the hypothesis that they are direct targets of *PpDUO1* genes. The presence of two DBS upstream of *PpDAZ1B* also place it as a highly probable target of *PpDUO1*. The absence of a DBS in *PpDAZ1D* suggests that this gene is not directly regulated by *PpDUO1*.
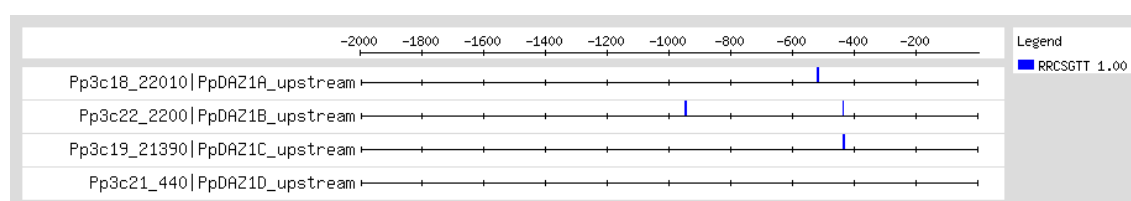
Figure 5.3.2.2. DUO1 binding sites (DBS) in *PpDAZ1* family upstream sequence. 2000 bp sequence upstream of ATG start codon were pattern match with 'RRCSGTT' motif from Higo et al. (2018) using Regulatory Sequence Analysis Tools.

Table 5.3.2.1. Analysis result of DBS in *PpDAZ1* family upstream sequence. D refers to direct strand.

Strand	Pattern	SeqID	Start	End	Matching_seq
D	RRCSGTT	PpDAZ1A_upstream	-521	-515	tactGACGGTTcgtg
D	RRCSGTT	PpDAZ1B_upstream	-949	-943	agggGGCGGTTttgt
D	RRCSGTT	PpDAZ1B_upstream	-440	-434	gactGACGGTTcgcg
D	RRCSGTT	PpDAZ1C_upstream	-438	-432	gactGACGGTTccga

5.4 Understanding the relationship of *PpDAZ1* and *PpDUO1*

5.4.1 Generation of *PpDAZ1* promoter-reporter gene constructs

The presence of DBS in the upstream region of several *PpDAZ1* paralogs (*PpDAZ1A*, *PpDAZ1B* and *PpDAZ1C*) highlights these genes as strong candidates as the target genes of DUO1. In order to understand this *DUO1-DAZ1* interaction in *P. patens*, *PpDAZ1A-D* promoter-reporter gene constructs were generated by fusing the promotor

of each gene with CDS of GUS reporter gene. The constructs were then transformed into wild type and *Ppduo1a^{A}b^A* mutant plants of Reute accession. The successful transformants will be used in GUS histological staining which will reflect the promotor activity of the *PpDAZ1* genes. By doing this, the regulation of *PpDAZ1* by PpDUO1 could be visualised through comparing the promotor activity of *PpDAZ1* with and without the presence of *PpDUO1* gene in vivo.

Upstream fragments of ~1.7 kb from the codon start site were amplified from the genomic DNA of *PpDAZ1A-D* (Reute accession). The routine PCR where the extension temperature was set at 72 °C failed to produce any fragment for all *PpDAZ1* genes. Hence, optimisation was made to the protocol by lowering the extension temperature to 68 °C and consequently, the expected promoter fragments were produced for all genes. TA cloning was performed where the fragments were cloned into the pENTR5' TOPO-TA vector to create *PromPpDAZ1A*, *PromPpDAZ1B*, *PromPpDAZ1C* and *PromPpDAZ1D* entry clones (Figure 5.4.1.1). Promoter fragment of the promoter entry clones were recombined with GUS CDS fragment from the GUS entry clone in pTHattR4-R2 expression vector through the Gateway LR reaction to produce *PromPpDAZ1A-D:GUS* expression clones (Figure 5.4.1.2). As mentioned previously in Chapter 4, the GUS entry clone and pTHattR4-R2 were made and provided by Dr. Dieter Hackenberg (Figure 4.4.1.2 and Figure 4.4.1.3). The structure and orientation of all the expression clones were confirmed using diagnostic PCR and restriction enzyme analysis before being introduced into the wild type and *Ppduo1a^{A}b^A* mutant through protoplast transformation.
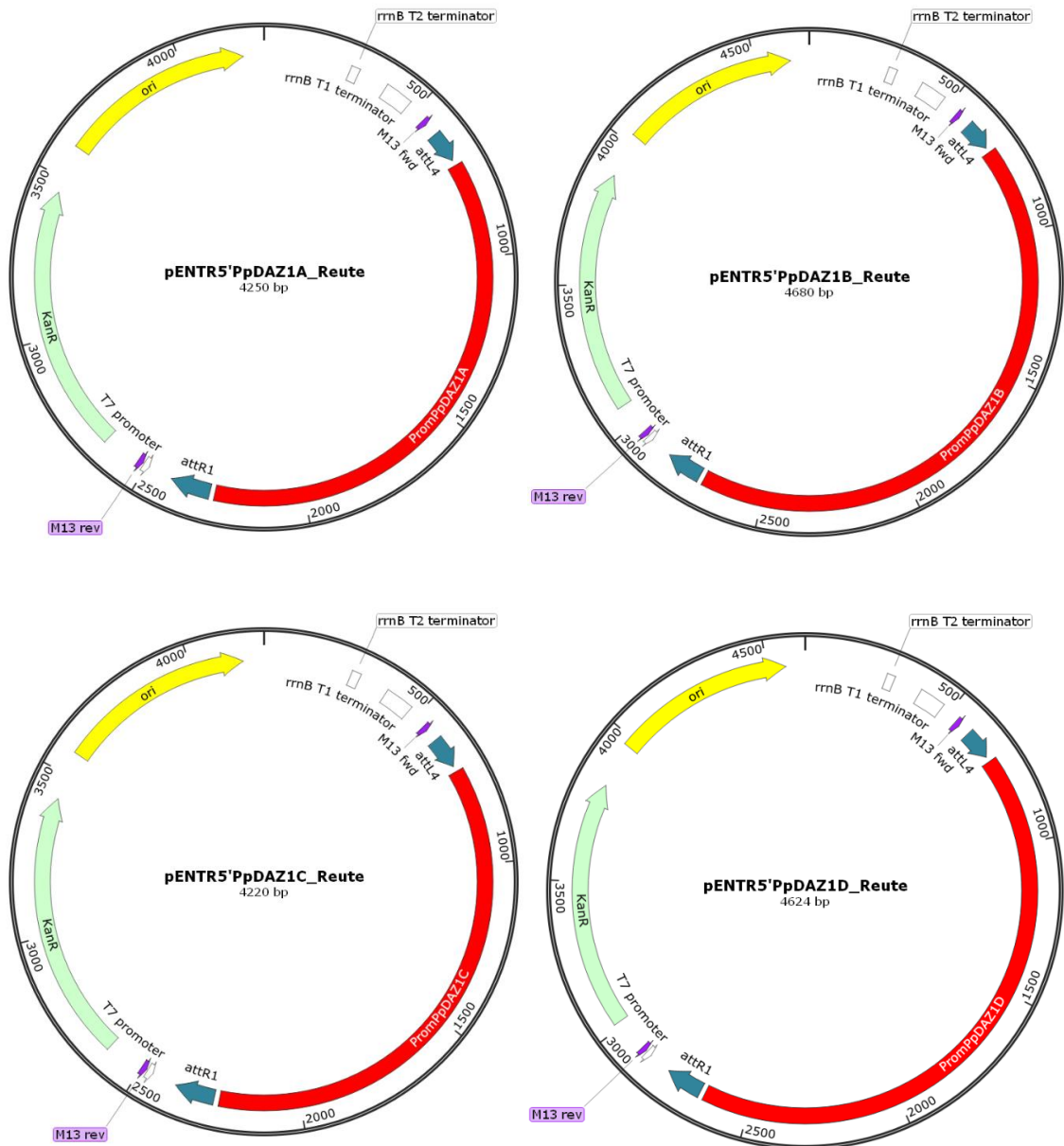
Figure 5.4.1.1 Promoter *PpDAZ1A-D* entry clones. *PpDUO1A-D* promoter fragments were cloned into pENTR5' TOPO-TA vector using TOPO TA cloning. Labels on map: KanR, Kanamycin Resistance gene; ori, plasmid origin of replication; M13 fwd, M13 forward sequencing primer; M13rev, M13 reverse sequencing primer; attL4/attR1, recombination site for the Gateway® LR reaction; rrnB T1 terminator, transcription terminator T1 from the *E. coli* rrnB gene; rrnB T2 terminator, transcription terminator T2 from the *E. coli* rrnB gene.
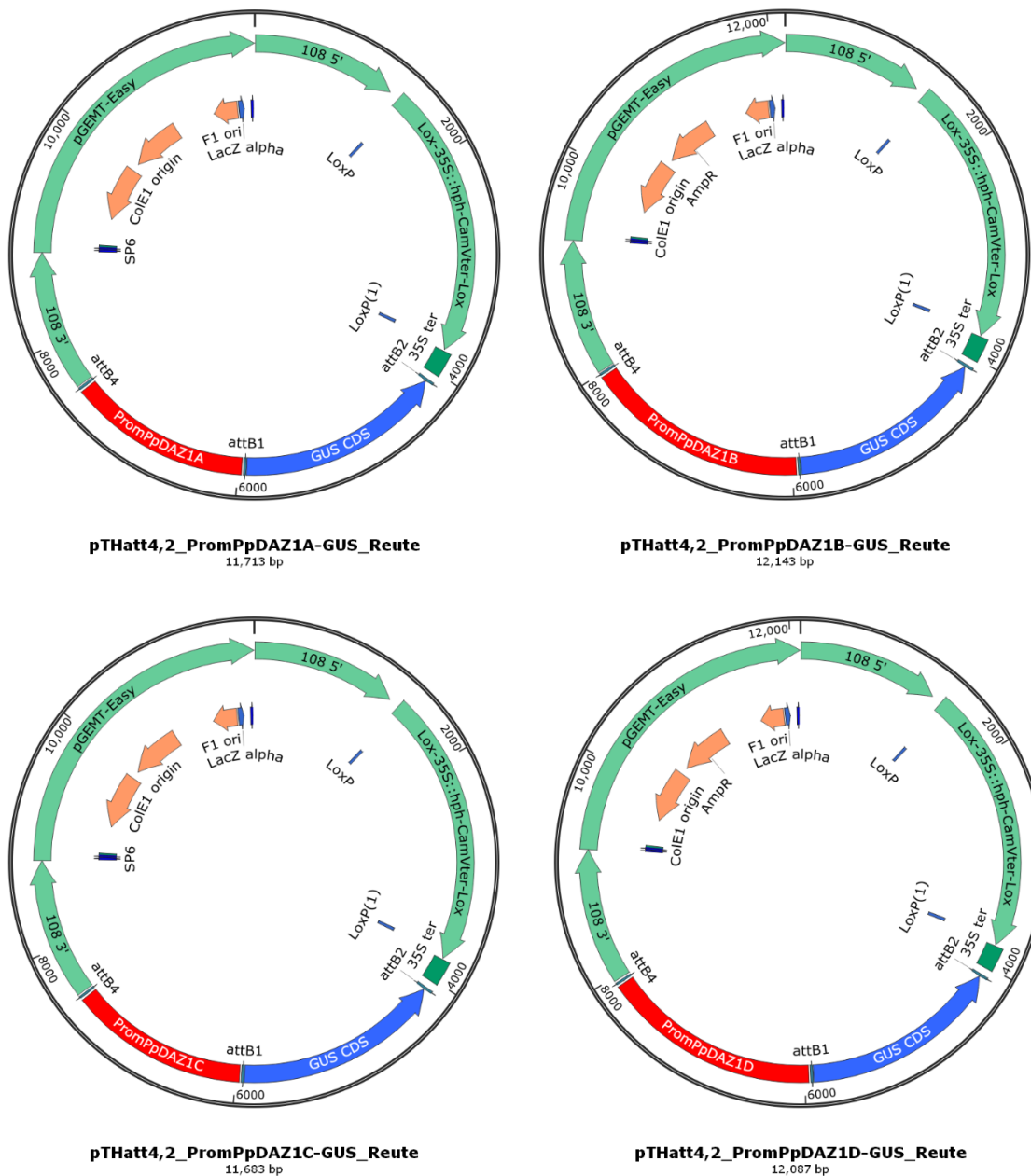
Figure 5.4.1.2. *PromPpDAZ1A-D:GUS* expression clones. Promoter *PpDAZ1A-D* were recombined with *GUS* coding sequence in pTHattR4-R2 vector using Gateway Cloning LR reactions. Labels on map: AmpR, Ampicillin Resistance gene; 108 5', 5'end target sequence at neutral locus 108; 108 3', 3'end target sequence at neutral locus 108; LacO, lactose operon; SP6, promoter for bacteriophage SP6 RNA polymerase; T7, promoter for bacteriophage T7 RNA polymerase; ColE1 origin, ColE1 plasmid origin of replication; F1 ori, F1 phage origin of replication; attB4/attB2, recombination site for the Gateway® BP reaction; 35S ter, 35S terminator; LoxP/LoxP(1), locus of crossover on bacteriophage P1.

5.4.2 Analysis of *PpDAZ1* promoter activity

PpDAZ1A-D promoter-GUS transformants were generated to investigate the spatio-temporal expression of *PpDAZ1* genes in *P. patens*. In addition, the regulation of each *PpDAZ1* genes by PpDUO1 also could be analysed by introduction of the *PpDAZ1A-D* promoter-GUS constructs into *Ppduo1a$^\Delta$b$^\Delta$* mutants. On average, around 30 stable transformants were generated for *PromPpDAZ1B-D:GUS* constructs in wild type *P. patens* and the *Ppduo1a$^\Delta$b$^\Delta$* double mutant after two rounds of hygromycin selection (Figure 5.4.2). The genotyping of all the stable transformants was incomplete. Therefore, the full analysis of *PpDAZ1A-D* promoter activity could not be completed.
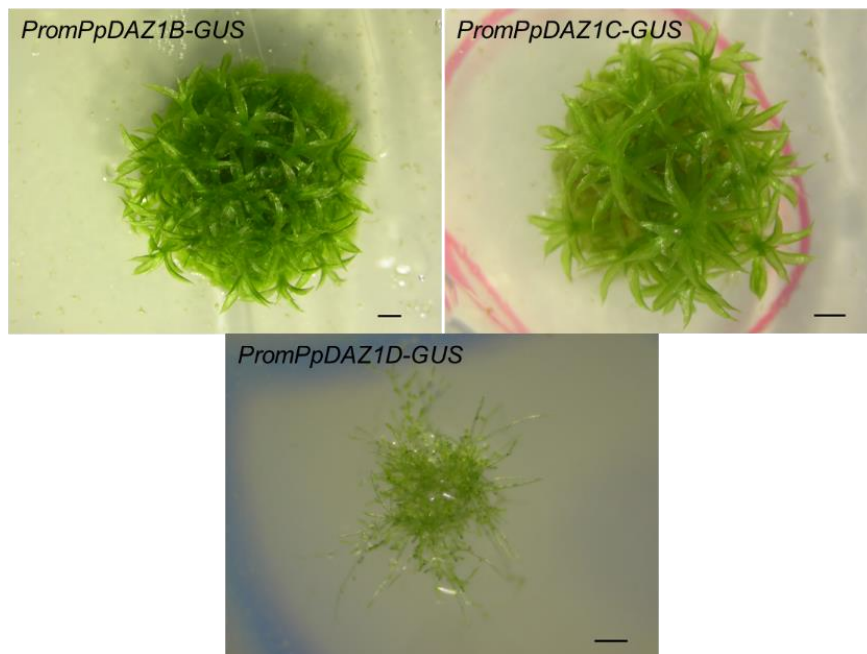
Figure 5.4.2. *PpDAZ1B/C/D-GUS* transformant in wild type *P. patens*. Image showing examples of stable *PromPpDAZ1B-GUS* and *PromPpDAZ1C-GUS* transformation after two rounds of hygromycin selection. *PromPpDAZ1D-GUS* at the first stage of hygromycin selection. Scale bar = 1 mm.

5.5 Potential target genes of PpDUO1 in *P. patens*

5.5.1 Exploring the candidates of PpDUO1 target genes

Including *DAZ1* and *DAZ2*, 63 genes were found to be under the regulation of DUO1 (Borg et al., 2011). These genes included well known male germline-specific genes, *MGH3* and *GEX2* (Brownfield et al., 2009; Mori et al., 2006). To identify potential DUO1

target genes in *P. patens*, two approaches were undertaken. The first approach was by comparing the antheridia specific genes between *P. patens* with *M. polymorpha*. The Compare Specificities tool in CoNekT was used to identify the genes that are specifically expressed in antheridia of both species (Proost and Mutwil, 2018). The specificity measure (SPM) cutoff was set to recommended value, 0.85, which signifies the level of contribution from one tissue to the entire expression profile (Xiao et al., 2010).

A Venn diagram was generated which shows the number of genes that are specifically expressed in the antheridia of *M. polymorpha* and *P. patens* (Figure 5.5.1.1). 114 ortholog groups (orthogroups) were found belong to the 'intersection' of the Venn diagram. The genes in the orthogroups were expressed specifically in the antheridia of *P. patens* or *M. polymorpha* and have their ortholog in the reciprocal species. Other than that, 1517 and 869 genes were found to have antheridia-specific expression exclusive to *M. polymorpha* and *P. patens*, respectively.
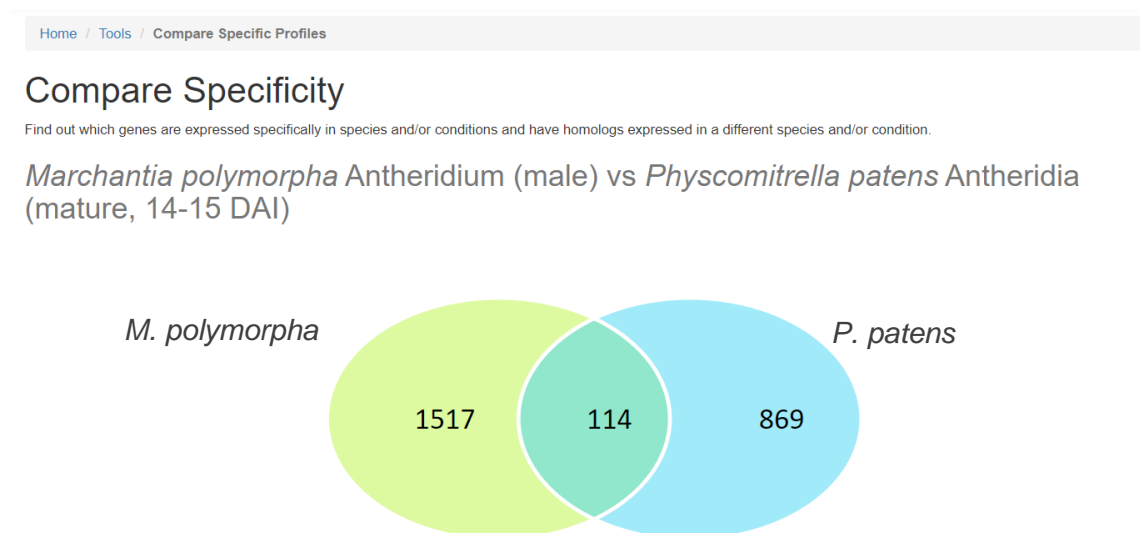
Figure 5.5.1.1. Venn diagram showing the comparison of antheridia specifically expressed genes in *P. patens* and *M. polymorpha*. Compare specificity tools in CoNekT was used to retrieve genes that were specifically expressed in antheridia of both species. 114 group of genes, called orthogroup, were shared between the two species.

In order to search for potential targets of DUO1 in *P. patens*, DBS analysis (described in subchapter 5.3.2) was performed on all genes belonging to the 114 orthogroups. When the list of genes for these orthogroups were downloaded, the total number of genes involved was more than 114. This was due to paralogues that some of the genes in the

orthogroups possess. In total, there were 135 and 133 genes for *P. patens* and *M. polymorpha*, respectively. The genes were later classified into four categories – orthogroups with DBS, only orthogenes in *P. patens* with DBS, only orthogenes in *M. polymorpha* with DBS and orthogroups without DBS (example in Table 5.5.1.1). Out of 114 orthogroups, 36 were under 'orthogroups with DBS' category, 20 were for 'only orthogenes in *P. patens* with DBS' category, 37 were for 'only orthogenes in *M. polymorpha* with DBS' category and 21 were for 'orthogroups without DBS' category.

Table 5.5.1.1. Example of orthogroups of antheridia specifically expressed genes. List showing example of 12 orthogroups from total of 114. Based on the presence of DBS in their 2 kb upstream sequence, the orthogroups were further categorised into four categories. Yellow, orthogroups with DBS; blue, orthogroups where only orthogenes in *P. patens* with DBS; green, orthogroups where only orthogenes in *M. polymorpha* with DBS.

Orthogroups		
M. polymorpha Antheridium (male)	*P. patens* Antheridia (mature, 14-15 DAI)	Category
Mapoly0003s0010	Pp3c4_29840	Orthogroups with DBS
Mapoly0009s0120	Pp3c12_4290	
Mapoly0091s0065	Pp3c9_15350	
Mapoly0001s0515	Pp3c1_110	Only orthogenes in *P. patens* with DBS
Mapoly0016s0047	Pp3c1_13140	
Mapoly0119s0051	Pp3c10_2420	
Mapoly0005s0276	Pp3c13_10870	Only orthogenes in *M. polymorpha* with DBS
Mapoly0006s0030	Pp3c7_7010	
Mapoly0006s0221	Pp3c6_29330	
Mapoly0005s0228	Pp3c3_32810	Orthogroups without DBS
Mapoly0007s0199	Pp3c13_13660	
Mapoly0008s0156	Pp3c12_14140	

A total of 43 % (58) genes for *P. patens* and 63 % (84) genes for *M. polymorpha* were found to have the DBS in their 2 kb upstream sequence. In all genes that have a DBS, the number of DBS per gene ranged from one to five and one to seven for *P. patens* and

M. polymorpha, respectively, with the majority genes possessing one or two DBS (81 % and 69 % in *P. patens* and *M. polymorpha*, respectively). A DBS distribution graph for *P. patens* was constructed, showing the frequency of DBS detected in 2 kb upstream region in all 58 genes (Figure 5.5.1.2). The highest frequency of DBS was detected at the 500 bp region upstream of the ATG start site. This was followed by the region between 501 bp to 1000 bp upstream region. There is a clear trend of a reduced frequency of DBS with greater distance from the ATG start site.

DBS distribution
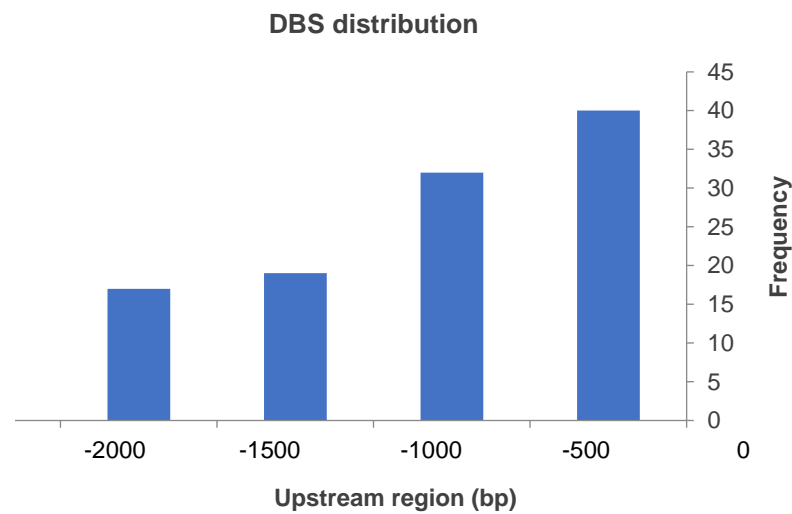
Figure 5.5.1.2. DBS frequency detected in 2 kb upstream region in all 58 genes of *P. patens*. The highest frequency of DBS was present within 500 bp upstream of the ATG start site.

Since a higher frequency of DBS was detected in proximal regions, two types of filter were applied to pull out the potential target genes of DUO1 in *P. patens*. The first filter was more specific, aiming to capture a list of genes with more than one DBS in 1.5 kb upstream sequence in 'orthogroups with DBS' category. Orthogroup with gene members that met the filter requirement for both *P. patens* and *M. polymorpha* were selected. This filter pulled out seven orthogroups with one gene member each, in both *P. patens* and *M. polymorpha*. The second filter was more inclusive, involving all 58 genes in *P. patens* that have a DBS. The objective was to find *P. patens* genes that have DBS in 1.5 kb upstream that show the trend of increasing transcript expression in developing antheridia. The transcript profile for all 58 genes were retrieved from CoNekT involving

150

antheridia at 9 DAI, 11 DAI and 14 – 15 DAI. This second filter managed to extract 22 *P. patens* genes.

The second approach to identify DUO1 potential target genes was through exploration based on Higo et al. (2018). DUO1 target genes in *M. polymorpha* was discussed in the study; hence, *P. patens* orthologs of those target genes were collected from CoNekT, Ensembl Plants and Phytozome12 using method described in subchapter 2.7.3. Among the analysed *M. polymorpha* genes were *MpTUA5*, *MpTUB4*, *MpPACRG*, *MpPRM*, *MpLC7* and *MpCEN1*, which are required for flagella formation. Orthologs of these genes in *P. patens* were found and presented in Table 5.5.1.2. DBS analysis was performed and histogram graph with bins was constructed to analyse the frequency distribution of the DBS within the 2 kb upstream region.

Table 5.5.1.2. Potential DUO1 target genes in *P. patens* based on *M. polymorpha* genes in Higo et al. (2018). Orthologous genes were downloaded from CoNekT, Ensembl Plants and Phytozome12 based on parameter described in subchapter 2.7.3.

Genebank accession	*M. polymorpha* gene	*P. patens* ortholog gene ID
LC172181	*MpTUA5*	Pp3c14_17800
		Pp3c23_21870
		Pp3c4_1960
		Pp3c4_2000
		Pp3c9_24320
		Pp3c3_5290
		Pp3c3_17990
KM096548	*MpTUB4*	Pp3c5_20340
		Pp3c6_7400
LC102460	*MpPACRG*	Pp3c9_23450
LC102462	*MpPRM*	Pp3c26_7260
		Pp3c19_7650
		Pp3c5_3110
		Pp3c6_25560
LC102461	*MpLC7*	Pp3c14_5710
LC379265	*MpCEN1*	Pp3c12_14390

From the DBS analysis of candidate genes using approach 2, the number of DBS per gene was in the range of one to four (Figure 5.5.1.3). The majority of the genes had two DBS which occur in 50 % of the genes, followed by one DBS in 31 % of genes. Histogram describing the distribution of DBS in 2 kb upstream sequence showed that 84 % of DBS lie within 1.5 kb upstream of the ATG start codon (Figure 5.5.1.4). Then, 42 % of DBS was found within 0.5 kb range of upstream sequence, followed by 32 % in the 1.0 kb range such that there was a trend of reduced frequency of DBS in more distal promoter regions.
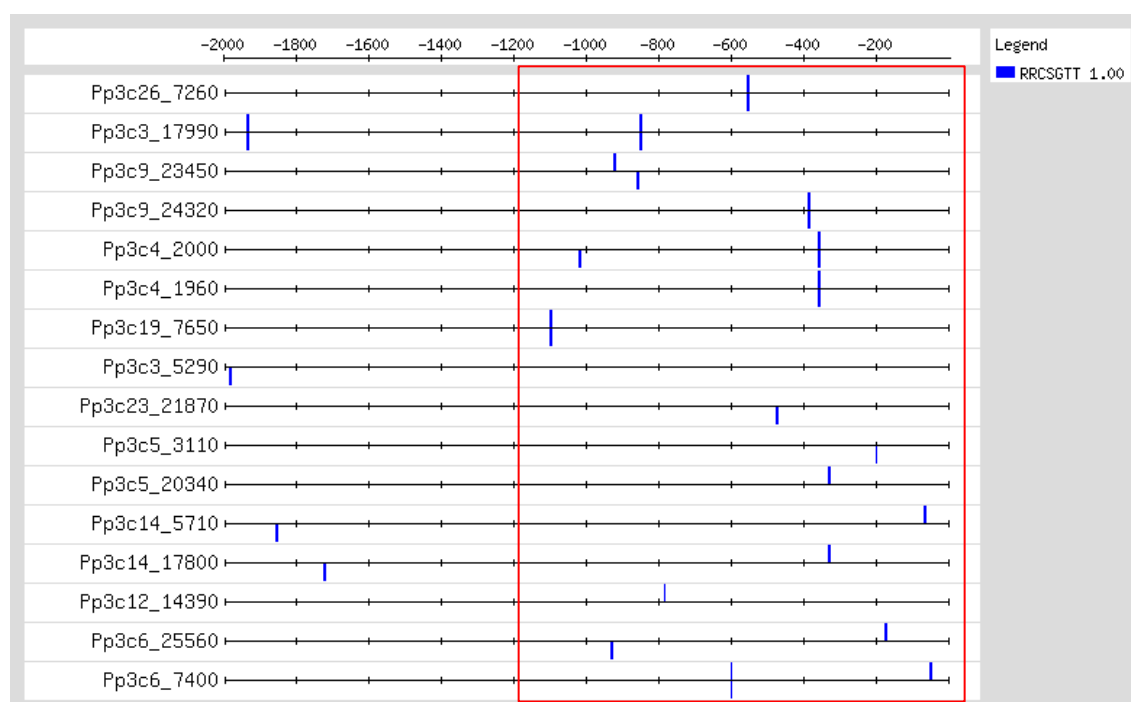
Figure 5.5.1.3. DBS analysis of potential DUO1 target genes in *P. patens* based on *M. polymorpha* genes in Higo et al. (2018). 2000 bp sequence upstream of ATG start codon were pattern match with 'RRCSGTT' motif from Higo et al. (2018) using Regulatory Sequence Analysis Tools. Red box was showing the location of majority DBS situated.
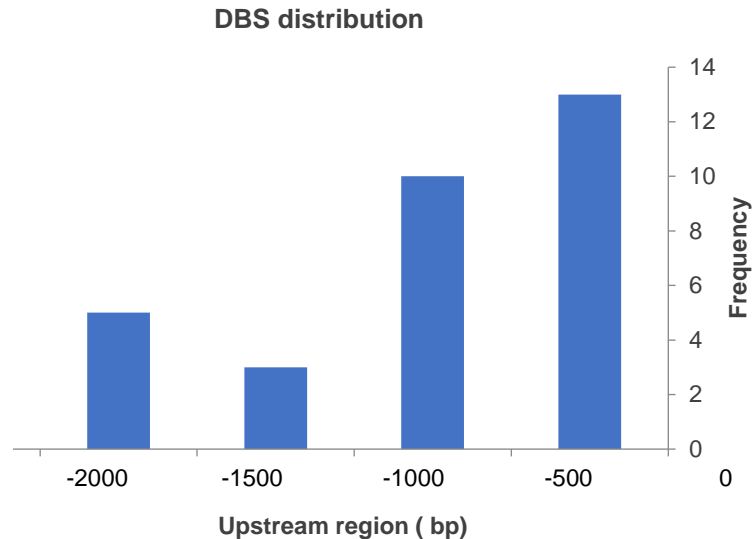
DBS distribution

Figure 5.5.1.4. DBS distribution in putative DUO1 target genes in *P. patens* based on *M. polymorpha* genes in Higo et al. (2018). Most of the DBS was located within 500 bp upstream of the ATG start site. The further the upstream sequence from ATG start site, the lower the frequency of DBS.

The results of all analyses were compared to identify the presence of overlapping potential DUO1 target genes (Figure 5.5.1.5). Six genes with gene ID Pp3c4_29840, Pp3c12_4290, Pp3c1_35370, Pp3c7_6230, Pp3c9_15350 and Pp3c18_5690 were found to be shared between filter 1 and 2 in approach 1. Meanwhile, one candidate gene, Pp3c9_23450, the ortholog of *MpPACRG*, was found in both approach 1 (filter 2) and approach 2. Information search using PANTHER database (http://www.pantherdb.org/) had found description for Pp3c1_35370 (ID: PTHR24073:SF209) and Pp3c7_6230 (ID: PTHR12509:SF9). Pp3c1_35370 belongs to Ras-related protein RAB-23 subfamily and involves in biological processes autophagosome assembly, intracellular protein transport and proteolysis. Meanwhile, Pp3c7_6230 is a gene in the spermatogenesis-associated protein 4 family. It has a molecular function in microtubule binding and biological processes like cytoskeleton organization and regulation of cytoskeleton organization. UniProt search for Pp3c4_29840 (UniProt ID: A0A2K1KQJ2) showed that it encodes for enkurin domain-containing protein which is required for mouse sperm motility (Jungnickel et al., 2018; Sutton et al., 2004). Information in InterPro (https://www.ebi.ac.uk/interpro/) reported that Pp3c12_4290 (ID: IPR000210) belongs to the BTB/POZ domain-containing protein and has a molecular function in protein binding. Meanwhile, the gene description in the NCBI database stated that Pp3c9_15350 (Gene ID: 112286768) is a radial spoke head 1 homolog. According to the UniProt database,

153

RSPH1 (*Radial Spoke Head Component 1*) (UniProt ID: Q8WYR4 (Human); Q8VIG3 (Mouse)) is involved in several biological processes such as axoneme assembly, meiotic cell cycle and spermatid development in humans and mice. Data from PANTHER database showed Pp3c18_5690 (ID: PTHR33649:SF2) belongs to the PAR1 protein subfamily. The gene ID Pp3c4_1800 of filter 1 was also of particular interest as it is annotated as centrin2 in CoNekT and Centrin-2 (Caltractin isoform 1) in UniProt.

In conclusion, candidate DUO1 target genes in *P. patens* were discovered through several in silico approaches. The presence of DBS in the putative target genes imply that they might be direct targets of PpDUO1.
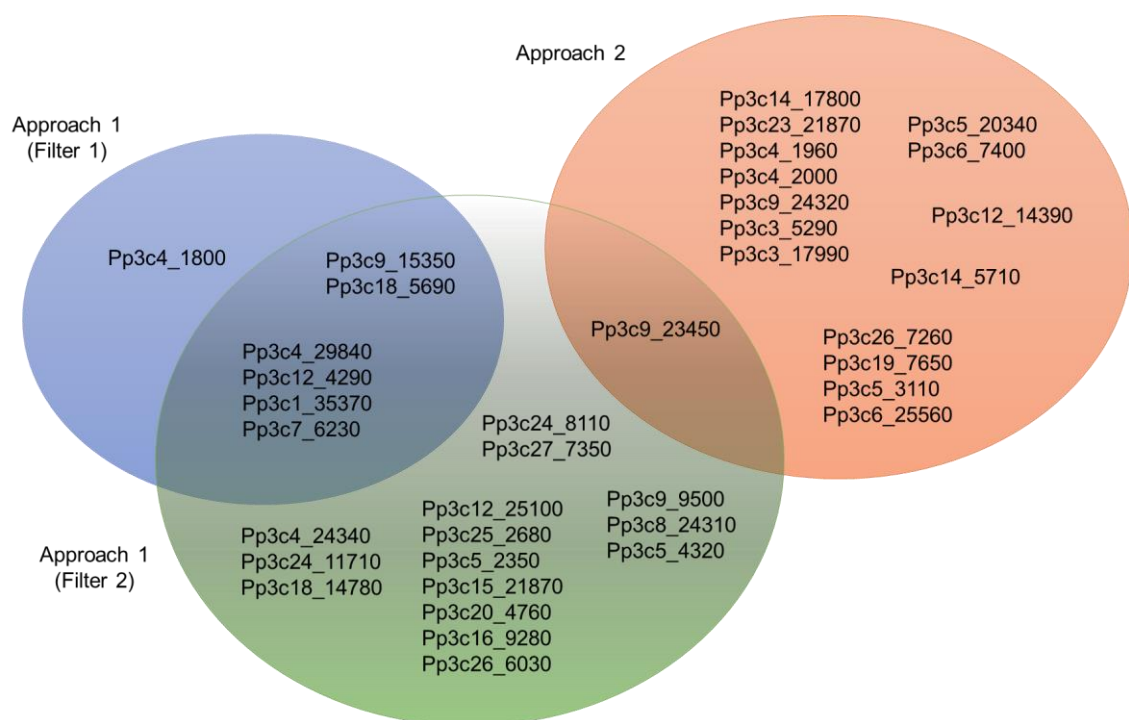
Figure 5.5.1.5. Venn diagram showing DUO1 potential target genes in *P. patens*. Seven and 22 potential target genes were found using approach 1 with filter 1 and filter 2, respectively, with six genes were shared between them. 16 candidate genes were found using approach 2 based on potential DUO1 target genes in *M. polymorpha* (Higo et al., 2018). One gene was shared between approach 2 with approach 1 filter 2.

5.6 Discussion

5.6.1 Sequence conservation of DAZ1 in bryophytes

Sequence conservation of DAZ1 was explored through analysis of orthologs in land plants, in particular the bryophytes. A clear difference was displayed in DAZ1 of bryophytes where each ortholog possess an additional zinc finger near the N-terminal region, unlike the DAZ1 of tracheophytes (Figure 5.2.1.3). The additional zinc finger has a zinc finger motif K/RXLXXH which resembles the K2-1 (RALGGH) and K2-2 (KALFGH) motifs. This suggests that the zinc finger motif could be initially belong to either K2-1 or K2-2 motifs. The motif accumulates mutation and subsequently majority of the amino acid residues in the zinc finger domain could be replaced, as seen in SfaDAZ1D (Figure 5.2.1.3). The substitution experienced by the zinc finger motif perhaps indicates that the extra zinc finger domain at the N-terminal is no longer important for the DAZ1 to function and might lost over time.

Alternatively, the extra zinc finger observed in bryophytes is most likely to have been present in earlier land plants and then lost in more recently divergent lineages like *A. thaliana*. Therefore, the variations observed in bryophytes DAZ1 are different to those reported in Englbrecht et al. (2004). Due to this, the additional zinc finger could contribute to DNA sequence binding differences and subsequently affect the targets of DAZ1 in bryophytes. The target of DAZ1, therefore could be different in bryophytes compared to the tracheophytes species like *A. thaliana*. This difference in targets of DAZ1 are likely to involve in processes related to sperm motility in which this characteristic is lost in angiosperms.

DAZ1 in bryophytes also has an acidic-rich conserved region and additional basic regions. In contrast to AtDAZ1, bryophytes DAZ1 possess only a single EAR motif. EAR motif was the first repression motif reported in plants (Kagale and Rozwadowski, 2011; Ohta et al., 2001). Hence, bryophytes DAZ1 may also function as a transcriptional repressor. This is supported by the repression activity displayed by AtDAZ1 through EAR motif in Borg et al. (2014). Bryophytes DAZ1 could also function as transcriptional activator due to the presence of acidic-rich region similar to that reported for APETALA1 (AP1) (Cho et al., 1999). Similar to the basic region in maize zinc finger protein Dof1, the basic regions in bryophytes DAZ1 could serve as nuclear localisation signal (NLS) (Yanagisawa, 2001).

5.6.2 Expression conservation of *DUO1-DAZ1* in *P. patens*

It was shown in the RNA-seq data that *PpDAZ1* genes are expressed in various developmental stages. Such example is *PpDAZ1B* which has high transcript expression in the vegetative tissues. Nevertheless, *PpDAZ1* genes are collectively expressed during the mature stage of antheridia to the late stage of sporophyte development. The collective onset of *PpDAZ1* genes during mature stage of antheridia coincides with high expression *PpDUO1* genes in antheridia. *PpDUO1* expression starts to increase during 11 DAI which was a few days before antheridia maturation at 14 to 15 DAI. The expression of *PpDAZ1* genes could be dependent on the expression of *PpDUO1* genes. This pattern is similar to the *DUO1-DAZ1* expression observed in *A. thaliana* where *AtDAZ1* and *AtDAZ2* had their peak developmental expression in tricellular pollen (Borg et al., 2014). Meanwhile, *AtDUO1* was highly expressed in bicellular pollen, in advance of tricellular pollen stage.

Based on the known regulatory relationship between DUO1 and DAZ1/DAZ2 in *A. thaliana*, RNA expression analysis of *P. patens* suggests that *PpDAZ1* could be under the regulation of *PpDUO1*. The early onset of *PpDUO1* transcription could indicate an early PpDUO1 protein expression. Thus, PpDUO1 expression could potentially regulate the *PpDAZ1* genes directly or indirectly.

5.6.3 DUO1 binding site in *PpDAZ1* promoter region

The upstream sequence of *DAZ1* genes were analysed for the presence of conserved motifs and putative cis-regulatory sequence. Three motifs were found which match with the binding site of DOF3.4, ERF027 and BCP5. The motif for DOF3.4 was found in the upstream sequence of all species analysed, whereas for ERF027 and BCP were only found in bryophytes and lycophytes (Figure 5.3.1). DOF3.4 is a zinc finger transcription factor that has a role in cell cycle regulation (Noguero et al., 2013). This suggest that *PpDAZ1* could also have role in cell cycle progression. Borg et al. (2014) had shown that DAZ1/DAZ2 are also required for germ cell division and for the proper accumulation of mitotic cyclins. In the absence of DUO1, DAZ1/DAZ2 are sufficient to promote G2 to mitotic phase and germ cell division.

The ERF027 is a transcription factor belongs to APETALA 2/ethylene-responsive element binding factor (AP2/ERF) that played crucial roles in regulating plant growth, development and response to stress (Cui et al., 2016). The BPC5 belongs to BASIC PENTACYSTEINE (BPC) transcription factor that function in cytokinin signalling

response (Shanks et al., 2018). However, *BCP5* is a pseudogene that has in-frame stop codon and is unlikely to produce an active protein (Monfared et al., 2011).

Interestingly, DUO1 binding sites are also found in the upstream region of *PpDAZ1A*, *PpDAZ1B* and *PpDAZ1C*. The presence of DUO1 binding site strongly suggest that these *PpDAZ1* genes are directly regulated by PpDUO1. As for *PpDAZ1D*, there is no DUO1 binding site detected in its upstream region. However, there is also a possibility that it could still be under the *PpDUO1* regulation network, through the activation of other *PpDUO1* target genes. Therefore, PpDUO1-*PpDAZ1* promoter expression analysis would be helpful to elucidate these interactions in vivo.

5.6.3 Candidates for DUO1 target genes in *P. patens*

Apart from *PpDAZ1* genes, other potential targets of PpDUO1 were also identified *in silico*. Two approaches were applied to select candidate of PpDUO1 target genes. The first one was by extracting genes that have similar expression pattern with *PpDUO1*. Since *PpDUO1* genes are highly enriched in antheridia, the genes that are specifically expressed in the same tissue would most likely be under the regulation of PpDUO1. In addition, if the ortholog of the candidate gene shares the same expression profile, they are likely to be under the same regulatory network. Such an example is Pp3c7_6230 (and all genes in approach 1) where this gene is specifically expressed in antheridia of *P. patens* and has an ortholog that is also specifically expressed in the antheridia of *M. polymorpha*.

DUO1 binding site analyses were used to retrieve candidate genes that could be the direct target of PpDUO1. This resulted in two groups of candidate genes, filter 1 where candidate genes and their orthologs both have DBS in their upstream regions, and filter 2 where only the candidate genes in *P. patens* have DBS in their upstream regions. The candidate genes in filter 1 would be of particular interest due to the presence and conservation of DBS in the upstream regions of orthologs. This could also suggest a conserved gene regulatory network of DUO1 among bryophytes. Meanwhile, the candidate genes from filter 2 might highlight a DUO1 network that is not conserved among bryophytes.

In a third straightforward approach orthologs of potential DUO1 target genes in *M. polymorpha* were retrieved and subjected to DUO1 binding site analysis (all genes in approach 2). These candidate genes encode proteins likely to be important for flagella formation (Higo et al. 2018). *PpDUO1* was found to be involved in flagella formation (see

subchapter 4.3.5). Furthermore, the DBS analysis shows that all the candidate genes identified in this approach possesses DUO1 binding site in their upstream sequence. This imply that they could be the direct target of PpDUO1 and together they may cooperate to regulate flagella formation.

Chapter 6: General discussion

6.1 DAZ3 – a DUO1 target with an unknown function

DAZ1, *DAZ2* and *DAZ3* are C_2H_2 zinc finger transcription factors that are among DUO1-activated target (DAT) genes (Borg et al., 2011). *DAZ3L*, a paralog of *DAZ3* was also identified through *in-silico* analysis (Taimur, 2014). In this thesis, the study of sequence conservation has shown that DAZ3/DAZ3L could have evolved from DAZ1/DAZ2 based on the similarity of their protein domains. This evolution may have happened after a genome triplication event in eudicots before the divergence of the superrosid-superasterid clade (Jiao et al., 2012; Panchy et al., 2016; Qiao et al., 2019). Following this event, DAZ3/DAZ3L are only present in eudicots, specifically in superrosids and some asterids. In contrast, DAZ1/DAZ2 is present is all known angiosperm genomes.

DAZ3 and *DAZ3L* are proposed to have a function in sperm cell development. This is due to their highly abundant transcript and specific expression in sperm cells (Borges et al., 2008; Taimur 2014). Since they contain a C_2H_2 DNA binding motif, DAZ3 and DAZ3L are expected to be localised to sperm cell nuclei (Franco-Zorilla et al., 2014; Wang et al., 2020). However, previous analysis of fusion proteins with a reporter protein did not fully support this prediction as they are predominantly localised in the cytoplasm (Taimur, 2014). In addition, DAZ3L expression also present in the nucleus. The same pattern continues even when the sperm cells are present in growing pollen tubes. Based on this finding, DAZ3 and DAZ3L are also thought to have a role in fertilisation.

In this thesis, experiments are carried out to investigate whether *DAZ3* and *DAZ3L* have a role in sperm cell development and fertilisation. Single knockout mutations in both genes produced plants with normal fertility, potentially indicating an essential but redundant function. Intriguingly, their double knockout mutant plants generated viable sperm cells. Homozygous double knockout mutants produce viable seeds capable of germination and the production of viable progeny. These results indicate that *DAZ3* and *DAZ3L* are not required for sperm cell development and do not have essential functions in ensuring successful fertilisation.

This study demonstrates that that transcript abundance of a gene does not directly translates to its perceived function. *DAZ3*/*DAZ3L* may be important in a more general cellular process in sperm or there could be other gene(s) that could compensate for the absence of DAZ3/DAZ3L leading to the absence of an obvious phenotype.

6.2 DAZ3 and ethylene response protein needed for transcriptional repression?

In a recent study by Wang et al. (2020) DAZ3L, which they called EIN3-dependent ETHYLENE-RESPONSE 1 (TREE1), was shown to interact with ETHYLENE INSENSITIVE 3 (EIN3) in shoots. In the presence of ethylene, ETHYLENE INSENSITIVE2 (EIN2) in the cytoplasm enters the nucleus to activate EIN3, which then acts as a central transcriptional regulator in the ethylene response (Munné-Bosch et al., 2018). EIN3 is a short-lived protein that undergoes ubiquitination and proteasomal degradation by EIN3 BINDING F-BOX1 (EBF1) and EBF2 (Dolgikh et al., 2019). In order to ensure EIN3 stability, EIN2 inhibits EBF1/EBF2 through translational repression and proteasomal degradation (Li et al., 2015; Merchante et al., 2015).

DAZ3L/TREE1 is shown to be a transcriptional repressor and is enhanced by EIN3 (Wang et al., 2020). In coexpression assays in *N. benthamiana*, EIN3 as an effector directs the expression of a construct in which the luciferase gene is driven by the 35S promoter containing both DAZ3L/TREE and EIN3 binding motifs. Luciferase expression however is repressed when DAZ3L/TREE is the effector. Further, the expression of luciferase is even lower when both DAZ3L/TREE and EIN3 were effectors. This study also showed that DAZ3L/TREE and EIN3 interact strongly in in yeast two-hybrid assays, while the interaction between DAZ3 and EIN3 is relatively weak.

Based the recent findings of Wang et al. (2020) and the results presented in this thesis, a model is proposed where DAZ3/DAZ3L interact with EIN3 in sperm cells to regulate transcriptional repression in the absence of ethylene (Figure 6.2.). In the current study, no significant change in phenotype is observed when DAZ3 and DAZ3L are absent, suggesting EIN3 can compensate for their loss of function. The expression of *EIN3* is observed to be high based on RNA-seq data (in CoNekT, https://evorepro.sbs.ntu.edu.sg/), similar to that of *DAZ3L* in sperm cells and significantly higher (enrich) than its expression in other tissue. However, more study is required to identify the activation pathway for EIN3 as its known activator, *EIN2*, transcript expression is absent from sperm cells based on the data found in CoNekT.

In addition, DAZ3L was found to be one of the RNA-binding proteins (RBPs) in leaf by Bach-Pages et al. (2020). As DAZ3 and DAZ3L are predominantly located in the cytoplasm, they may ensure the stability of EIN3 by carrying out post transcriptional control by binding to RNA, similar to EIN2. They also could be involved in other RNA regulation in the cytoplasm, hence playing a crucial role in regulating RNA function and fate (Bach-Pages et al., 2020). DAZ3/DAZ3L and EIN3 together with its paralog, *ETHYLENE-INSENSITIVE3-LIKE 1* (*EIL1*), could be collectively important for

161

transcriptional repression in sperm cells, yet their role can be compensated by each other.
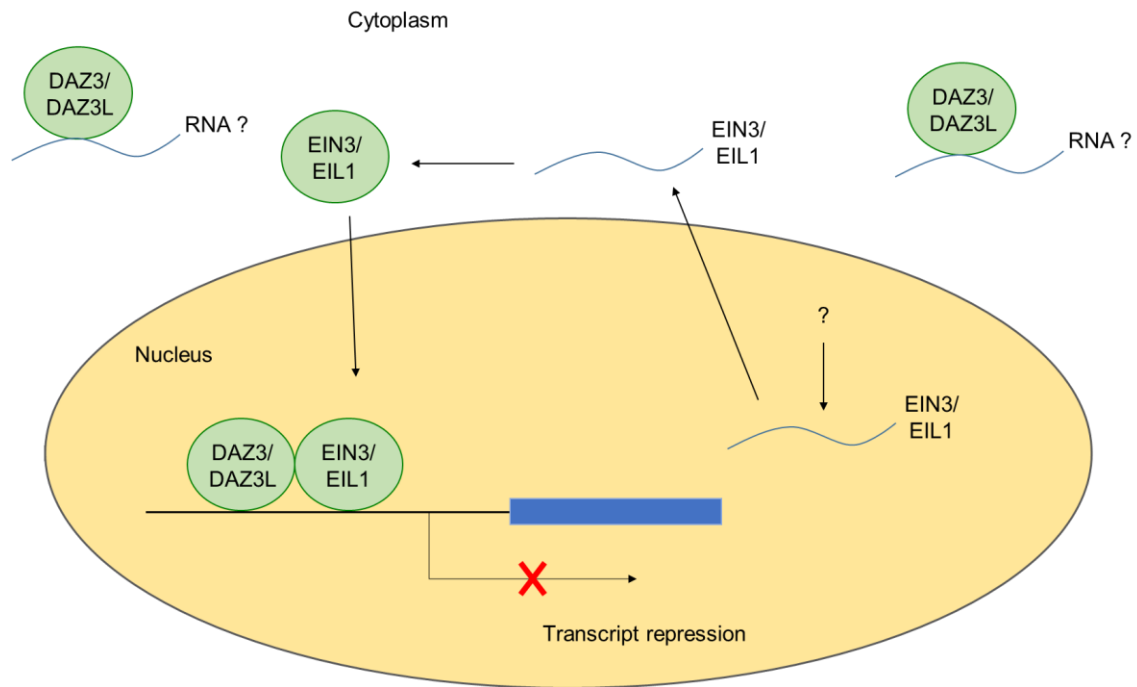
Figure 6.2. Proposed model of DAZ3/DAZ3L-EIN3 transcriptional repression in sperm. Either DAZ3/DAZ3L or EIN3 is needed for transcript repression. If both are absent transcript repression does not occur. EIN3 is activated by unknown pathway. In addition, DAZ3/DAZ3L may also be involved in RNA regulation in cytoplasm indicated by the question mark.

6.3 DUO1 function is conserved in bryophytes

DUO1 is known to control germ cell division and differentiation in Arabidopsis (Borg et al., 2011). Previous work in *M. polymorpha*, has shown partial conservation of DUO1 function between angiosperms and bryophytes (Higo et al., 2018). In particular, the role in sperm cell differentiation is conserved, but not the control of germ cell division. In this thesis, the conservation of DUO1 function in bryophytes is further explored and expanded to include another bryophyte lineage. In the moss, *P. patens*, DUO1 is also shown to be crucial for sperm cell differentiation, highlighting the functional conservation of DUO1 between bryophytes. The spermatogenous cell division is not affected in the absence of PpDUO1, however the cells are not able to differentiate to form flagella, therefore preventing the generation of functional sperm. The inability to generate fully

162

differentiated sperm results in male sterility in *P. patens* as mutant sperm are non-motile, thereby blocking sperm movement and fertilisation. The findings from studies on DUO1 function in *M. polymorpha* and *P. patens* are consistent with a conserved and specific role for DUO1 in bryophyte sperm differentiation.

There is partial conservation of the DUO1 regulatory network in *M. polymorpha*. While MpDAZ1 was found to be regulated by DUO1, *M. polymorpha* orthologs of other well established DUO1 target genes involved in fertilisation in Arabidopsis, such as GEX2 and GCS1, were not under MpDUO1 control (Higo et al., 2018). GEX2 is a membrane associated protein involved in sperm cell adhesion to the egg cell, while and GCS1 is required for successful membrane fusion fertilisation (Engel et al., 2005; Mori et al., 2006; von Besser et al., 2006; Mori et al., 2014).

6.4 The DUO1-DAZ1 network could be partially conserved in *P. patens*

The conservation of the DUO1-DAZ1 regulatory network is further explored in bryophytes. In *M. polymorpha* there is a single *DAZ1* orthologue, which was shown to be under control of MpDUO1 (Higo et al., 2018). In *P. patens*, four *DAZ1* paralogs were identified and three of these possess DUO1 binding site in their upstream promoter regions. The same binding site is also found in the promoter of *MpDAZ1* (Higo et al., 2018). In Arabidopsis, DUO1 directly activates *DAZ1*/*DAZ2* by binding to the DUO1 binding sites present in their promoter regions (Borg et al., 2014). Hence, the same mechanism is likely to be conserved in bryophytes.

Another interesting aspect of DUO1-DAZ1 regulatory module is the tight expression profile. Similar to DUO1, DAZ1/DAZ2 are specifically expressed in the developing male germ cells in pollen (Borg et al., 2014). *PpDUO1* showed pronounced antheridia-enriched or specific expression, while the expression profiles of *PpDAZ1* genes is more diverse, with enhanced expression in antheridia, but significant expression in sporophyte and vegetative gametophyte tissue. While *DUO1-DAZ1* regulation may be conserved in *P. patens*, the function of *DAZ1* may be more diverse. Therefore, it can be proposed that *DAZ1* function in *P. patens* is not restricted to male germline development but has further biological roles at other developmental stages in the sporophyte and vegetative gametophyte stages.

6.4 Future works

The work presented in this thesis portrays the ambiguous function of DAZ3/DAZ3L in sperm cell. As proposed in the DAZ3-EIN3 model, DAZ3/DAZ3L and EIN3/EIL might have a similar transcriptional repression function in sperm cells and the loss of function of either may be compensated by the other. Previous research has reported that *ein3-1 eil1-1* double mutants exhibit ethylene insensitive phenotype such as inhibition of root growth in the presence of salt and enhanced freezing tolerance, but no infertility phenotype has been reported (Lin et al., 2013; Shi et al., 2012). A quadruple mutant of *daz3 daz3l ein3 eil1* could be generated and analysed for their phenotype in germ cell development and fertilisation. If the model is true, complementation of the mutant with either DAZ3 and DAZ3L or EIN3 and EIL1 would rescue the phenotype seen in the *daz3 daz3l ein3 eil1* mutant. Additionally, the effect of their mutation could be compared in the absence and presence of ethylene.
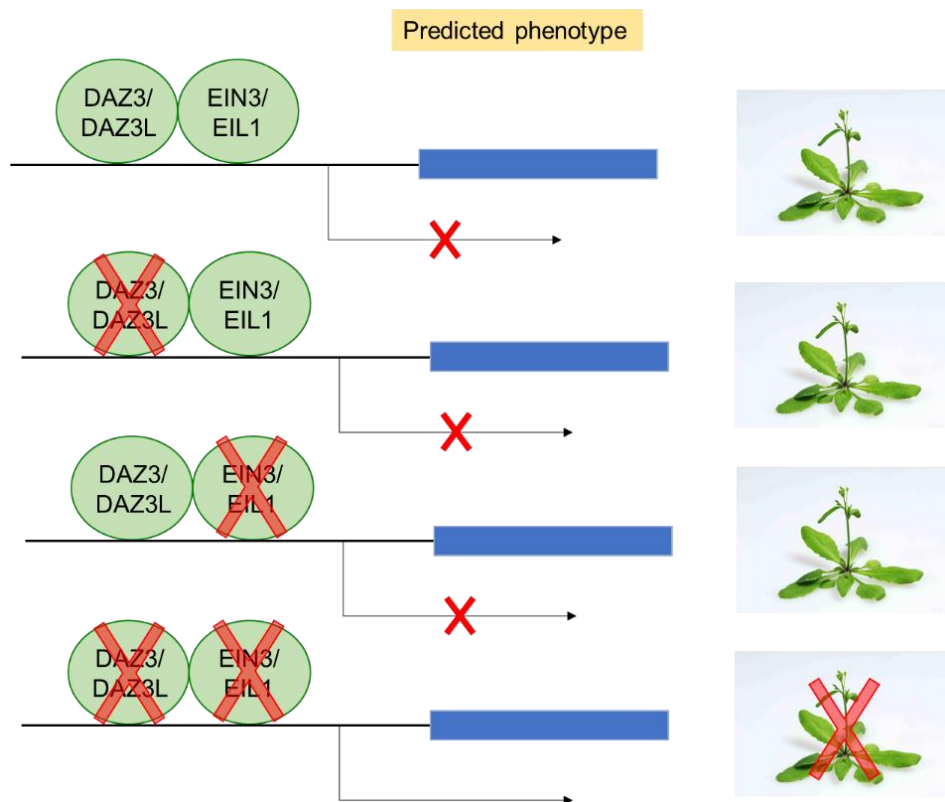
Figure 6.4. Proposed experiment for analysing the DAZ3/DAZ3L-EIN3. *daz3 daz3l ein3 eil1* mutant is expected to affect germ cell development and fertility. X indicates no transcripts.

The understanding of conservation of DUO1 in bryophytes could be expanded by analysing *DUO1* expression in *P. patens*. This could be done through *PpDUO1* promoter expression analysis in wild type plants. Although the transformation of promoter-marker constructs has been performed in this thesis, the transformants screening and expression analysis is yet to be completed. Based on RNA-expression profile, the *PpDUO1* expression is expected to be detected in early stage of antheridia development and accumulate as antheridia development progresses to later stages. In addition, the PpDUO1 protein expression could also be analysed by expressing native PpDUO1 gene fused to a marker gene in wild type plants. The conservation of the DUO1 network in bryophytes could be further explored by analysing the expression level of orthologs of AtDUO1 target genes such as GEX2 and GCS1 in *Ppduo1* double knockout plants. Since PpDUO1 is essential for sperm cell differentiation, the expression level of putative PpDUO1 target genes involved in flagella formation could also be analysed in *Ppduo1* mutant to validate their control by PpDUO1.

Components of the DUO1-DAZ1 network could be visualised by analysing *PpDAZ1* promoter activity in wild type and *Ppduo1* mutant plants. Transformation of *PpDAZ1* promoter-marker constructs have been carried out in this thesis, however further analysis is not yet complete. Meanwhile, the dependence of the expression of each of the four *PpDAZ1* genes on PpDUO1 could be examined through their expression in *Ppduo1* mutants. It is expected that there should be reduced or no expression of *PpDAZ1* genes that are under the control of PpDUO1. Further, mutations of the putative DUO1 binding sites of *PpDAZ1* promoters and examination of their in vivo expression in *P. patens*, or in PpDUO1 transactivation assays in tobacco leaves. A further interesting avenue to explore is the functional analysis of *PpDAZ1* genes in *P. patens* through gene knockout studies. Then, based on the results of *PpDAZ1* promoter analysis in *Ppduo1* mutant, *Ppdaz1* mutant combination could be generated to investigate the conservation of *DAZ1* function in *P. patens*.

The knowledge obtained in this thesis may be used as a steppingstone towards the advancement of sperm cell development research, stimulating subsequent research of the regulatory events in gamete development and function. In agriculture, the reproduction of crops is considered the crucial point of its life cycle. Therefore, this project can assist in understanding how to maintain and/or manipulate fertility to produce high quality seeds and improve crop production.

Appendix

Table S2.1 Composition of 1L stock of trace element solution.

Solution	Chemical	1 L stock
	H_3BO_3	614 mg
	$AlK(SO_4)_2.12\ H_2O$	110 mg
	$CuSO_4.5H_2O$	55 mg
	KBr	28 mg
	LiCl	28 mg
	$Na_2.MoO_4.2H_2O$	25 mg
Trace Element Solution	$MnCl_2.4H_2O$	389 mg
	$CoCl_2.6H_2O$	55 mg
	$ZnSO_4.7H_2O$	55 mg
	KI	28 mg
	$SnCl_2.2H_2O$	28 mg
	$NiCl_2.6H_2O$	59 mg
	Distilled H_2O	to 1 L

Figure S3.1. Cladogram of DAZ3 homologues in flowering plants. Cladistic analysis showing of DAZ3 sequences inferred by using the Maximum Likelihood method with bootstrap of 1000 replicates. Branches corresponding to partitions reproduced in less than 50% bootstrap replicates are collapsed. Proteins labelled in red are the member of DAZ3L clade. Gene ID is shown after the gene name.
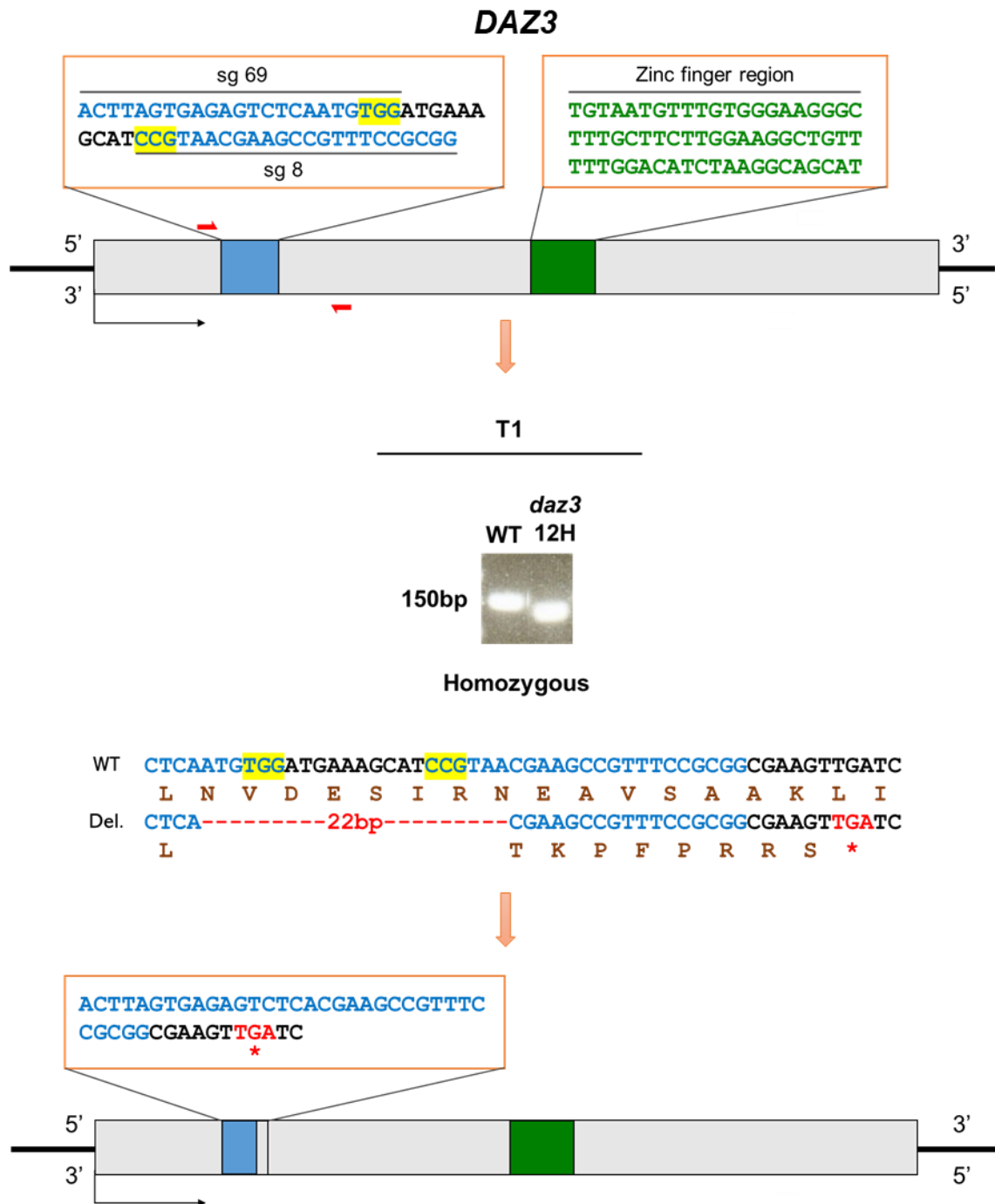
Figure S3.2. Workflow for creating homozygous *daz3* mutation in plant *daz3l* (B2_P14_P7) (also known as *daz3l-y*) background. Two single guide RNA, sg 8 and sg 69, guided CAS9 nuclease to its target sequences upstream of the zinc finger region. *daz3* homozygous plant was produced in the T1 generation where 22 bp deletion was introduced through NHEJ during the repair of double stranded break by CAS9 nuclease. The deletion created a frameshift which resulted an early stop codon, leading to *daz3* protein truncation. Red arrow, primers used for screening the presence of indels; blue region, single guide target sequence; highlighted region, protospacer adjacent motif.

Table S3.1. RNA-seq data in transcripts per million (TPM) for *DAZ3* and *DAZ3L* in pollen developmental stages for *A. thaliana* (Col-0). UNM, uninucleate microspores; BCP, bicellular pollen; TCP, tricellular pollen; MP, mature pollen; SPC, sperm cells.

Developmental stages	Repeats	Average
UNM_*DAZ3*	3.03	3.09
UNM_*DAZ3*	2.92	3.09
UNM_*DAZ3*	3.33	3.09
UNM_*DAZ3L*	1.31	0.97
UNM_*DAZ3L*	0.46	0.97
UNM_*DAZ3L*	1.13	0.97
BCP_*DAZ3*	20.18	20.04
BCP_*DAZ3*	16.63	20.04
BCP_*DAZ3*	23.32	20.04
BCP_*DAZ3L*	8.37	7.44
BCP_*DAZ3L*	4.90	7.44
BCP_*DAZ3L*	9.06	7.44
TCP_*DAZ3*	332.40	361.49
TCP_*DAZ3*	367.95	361.49
TCP_*DAZ3*	384.11	361.49
TCP_*DAZ3L*	137.79	131.27
TCP_*DAZ3L*	135.99	131.27
TCP_*DAZ3L*	120.02	131.27
MP_*DAZ3*	467.09	417.63
MP_*DAZ3*	401.17	417.63
MP_*DAZ3*	384.65	417.63
MP_*DAZ3L*	105.54	91.47
MP_*DAZ3L*	84.33	91.47
MP_*DAZ3L*	84.54	91.47
PT_*DAZ3*	1430.42	1278.18
PT_*DAZ3*	891.56	1278.18
PT_*DAZ3*	1512.54	1278.18
PT_*DAZ3L*	310.10	270.11
PT_*DAZ3L*	182.29	270.11
PT_*DAZ3L*	317.95	270.11

Developmental stages	Repeats	Average
MP_*DAZ3*	467.09	417.63
MP_*DAZ3*	401.17	417.63
MP_*DAZ3*	384.65	417.63
MP_*DAZ3L*	105.54	91.47
MP_*DAZ3L*	84.33	91.47
MP_*DAZ3L*	84.54	91.47
SPC_*DAZ3*	16111.46	13310.72
SPC_*DAZ3*	11578.32	13310.72
SPC_*DAZ3*	12242.38	13310.72
SPC_*DAZ3L*	4514.74	3518.34
SPC_*DAZ3L*	2954.92	3518.34
SPC_*DAZ3L*	3085.35	3518.34

Gene	MP	SPC	Fold change
DAZ3	417.63	13310.72	32
DAZ3L	91.47	3518.34	38

Table S3.2. Transcript per million (TPM) reads for _DAZ3_ and _DAZ3L_ in pollen developmental stages for _A. thaliana_ (Landsberg). UNM, uninucleate microspores; BCP, bicellular pollen; TCP, tricellular pollen; MP, mature pollen.

Developmental stages	Repeats	Average
UNM_*DAZ3*	0.51	0.69
UNM_*DAZ3*	0.75	0.69
UNM_*DAZ3*	0.81	0.69
UNM_*DAZ3L*	0.25	0.11
UNM_*DAZ3L*	0.08	0.11
UNM_*DAZ3L*	0.00	0.11
BCP_*DAZ3*	3.66	4.14
BCP_*DAZ3*	8.03	4.14
BCP_*DAZ3*	0.72	4.14
BCP_*DAZ3L*	1.67	0.85
BCP_*DAZ3L*	0.76	0.85
BCP_*DAZ3L*	0.12	0.85
TCP_*DAZ3*	369.57	498.99
TCP_*DAZ3*	415.95	498.99
TCP_*DAZ3*	711.44	498.99
TCP_*DAZ3L*	110.37	105.88
TCP_*DAZ3L*	54.54	105.88
TCP_*DAZ3L*	152.74	105.88
MP_*DAZ3*	540.75	505.22
MP_*DAZ3*	465.53	505.22
MP_*DAZ3*	509.36	505.22
MP_*DAZ3L*	87.44	64.17
MP_*DAZ3L*	58.48	64.17
MP_*DAZ3L*	46.59	64.17

Table S3.3. One-way ANOVA of seed viability data per silique in wild type, *daz3* and *daz3l* genotypes. For significantly different, p<0.05. *SS*=sum of squares; *df*=degrees of freedom; *MS*=mean square; *F*=F-value and *F crit*=F-critical value.

Percentage of viable seeds per silique (%)			
Silique no.	WT	*daz3-/-*	*daz3l-/-*
1	92	93	100
2	100	94	100
3	92	98	98
4	100	98	100
5	100	100	93
6	94	100	97

SUMMARY				
Groups	*Count*	*Sum*	*Average*	*Variance*
WT	6	578.022	96.337	16.55786
daz3-/-	6	582.5163	97.08604	9.454651
daz3l-/-	6	588.8124	98.13541	6.838923

ANOVA						
Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	9.793046	2	4.896523	0.45	0.65	3.68
Within Groups	164.2572	15	10.95048			
Total	174.0502	17				

Table S3.4. One-way ANOVA for analysing any significant difference between the mean silique length of *daz3-/-daz3l-/-* and wild type. For significantly different, p<0.05. *SS*=sum of squares; *df*=degrees of freedom; *MS*=mean square; *F*=F-value and *F crit*=F-critical value.

Silique no	Silique length (mm)	
	WT	***daz3-/-daz3l-/-***
1	14.4	14.7
2	13.9	14.6
3	14	13.9
4	14	14.8
5	13.8	14.8
6	14.9	15.4
7	13.9	14.8
8	14	15.5
9	14.6	16.4
10	16.5	15.6
11	15.7	16
12	16.8	16
13	15.5	17
14	15	16.4
15	15.4	18.7
16	16.4	16.4
17	17	14.5
18	15.7	16.7
19	16.9	19.2
20	16.3	18.7
21	15	15.1
22	14.2	16.7
23	14.7	16.3
24	13.3	16.9
25	18.5	17.5
26	18.1	17.2
27	15.7	16.2
28	14.6	16.7
29	18.9	17.3

30	18.7	17.8

Table S3.4.continue. One-way ANOVA for analysing any significant difference between the mean silique length of *daz3-/-daz3l-/-* and wild type. For significantly different, p<0.05. *SS*=sum of squares; *df*=degrees of freedom; *MS*=mean square; *F*=F-value and *F crit*=F-critical value.

SUMMARY				
Groups	*Count*	*Sum*	*Average*	*Variance*
WT	30	466.4	15.54667	2.468782
daz3-/-daz3l-/-	30	487.8	16.26	1.776276

ANOVA						
Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	7.632667	1	7.632667	3.60	0.06	4.01
Within Groups	123.1067	58	2.122529			
Total	130.7393	59				

Table S3.5. One-way ANOVA for analysing any significant difference in the mean percentage of viable seeds per silique in wild type and *daz3-/-daz3l-/-*. For significantly different, p<0.05. *SS*=sum of squares; *df*=degrees of freedom; *MS*=mean square; *F*=F-value and *F crit*=F-critical value.

	Percentage of viable seeds per silique (%)	
Silique no.	WT	*daz3-/-daz3l-/-*
1	100	96
2	77	96
3	100	94
4	100	67
5	94	93
6	100	93
7	98	82
8	98	83
9	100	74
10	100	98
11	45	94
12	52	98
13	57	96
14	47	95
15	98	100

SUMMARY				
Groups	*Count*	*Sum*	*Average*	*Variance*
WT	15	1268.156	84.54373	486.5291
daz3-/-daz3l-/-	15	1359	90.60001	94.98599

ANOVA						
Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	275.0889	1	275.0889	0.95	0.34	4.20
Within Groups	8141.211	28	290.7576			
Total	8416.3	29				
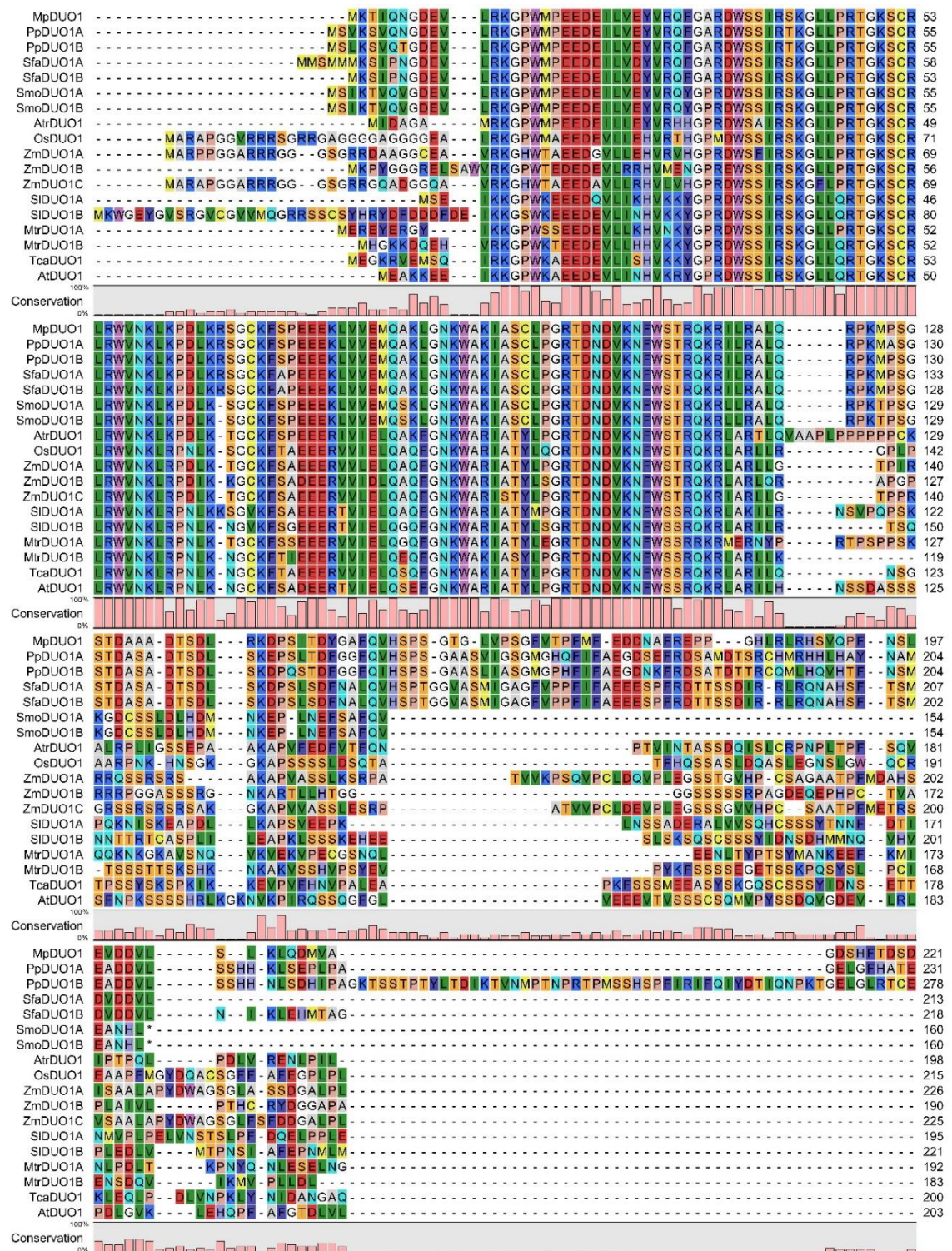
Figure S4.1. DUO1 orthologs sequence alignment from selected species. Proteins were aligned with MUSCLE using default parameter and viewed in CLC Sequence Viewer. Species: At, *Arabidopsis thaliana*; Tca, *Theobroma cacao*; Mtr, *Medicago truncatula*; Sl, *Solanum lycopersicum*; Os, *Oryza sativa*; Zm, *Zea mays*; Atr, *Amborella trichopoda*; Smo, *Selaginella moellendorffii*; Sfa, *Sphagnum fallax*; Pp, *Physcomitrella patens*; Mp, *Marchantia polymorpha*.

Figure S4.1.continue.

Table S4.1. One-way ANOVA for analysing any significant difference in the average number of sperm cells in mature antheridia of wild type and *Ppduo1a$^\Delta$b$^\Delta$*. For significantly different, p<0.05. *SS*=sum of squares; *df*=degrees of freedom; *MS*=mean square; *F*=F-value and *F crit*=F-critical value.

Replicate	Mean no. of sporophyte / gametophore		
	WT	***Ppduo1a$^\Delta$***	***Ppduo1b$^\Delta$***
1	1.9	2.1	1.8
2	1.7	1.3	2.2
3	1.6	1.9	1.6

SUMMARY				
Groups	*Count*	*Sum*	*Average*	*Variance*
WT	3	5.18	1.727	0.029411111
Ppduo1a$^\Delta$	3	5.28	1.76	0.1792
Ppduo1b$^\Delta$	3	5.482	1.827	0.094916159

ANOVA						
Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	0.015766	2	0.007883	0.08	0.925966	5.143253
Within Groups	0.607055	6	0.101176			
Total	0.62282	8				
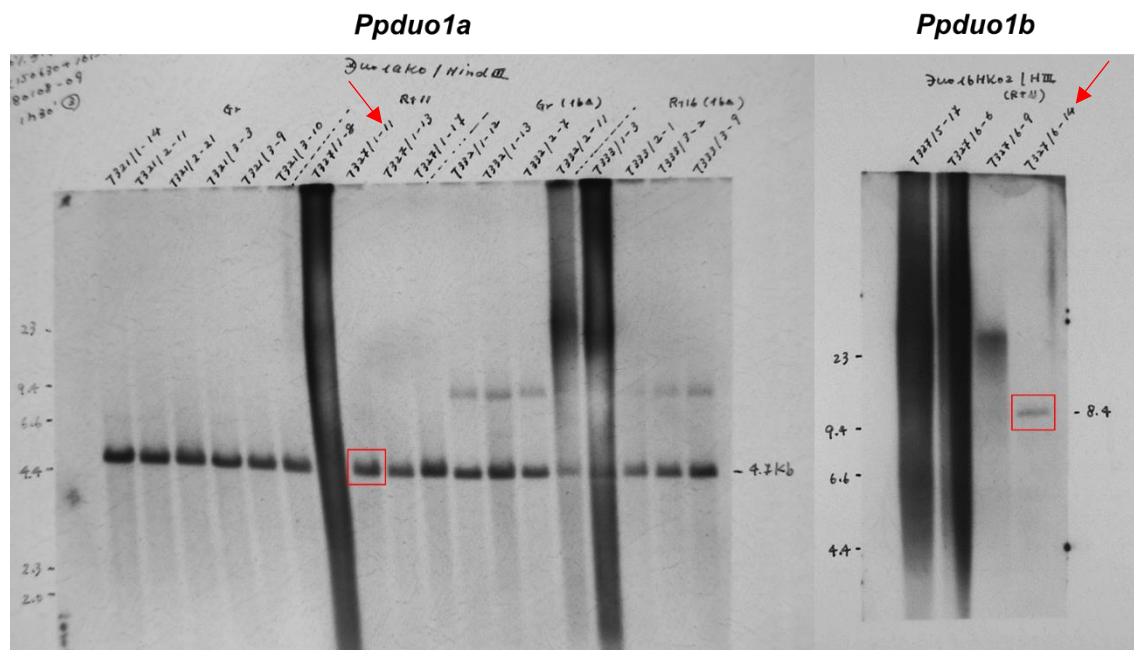
Figure S4.2. Confirmation of *Ppduo1a* and *Ppduo1b* knockout lines using Southern blotting. Lines T327/1-11 and T327/6-14 in background Reute was chosen for *Ppduo1a* and *Ppduo1b*, respectively.
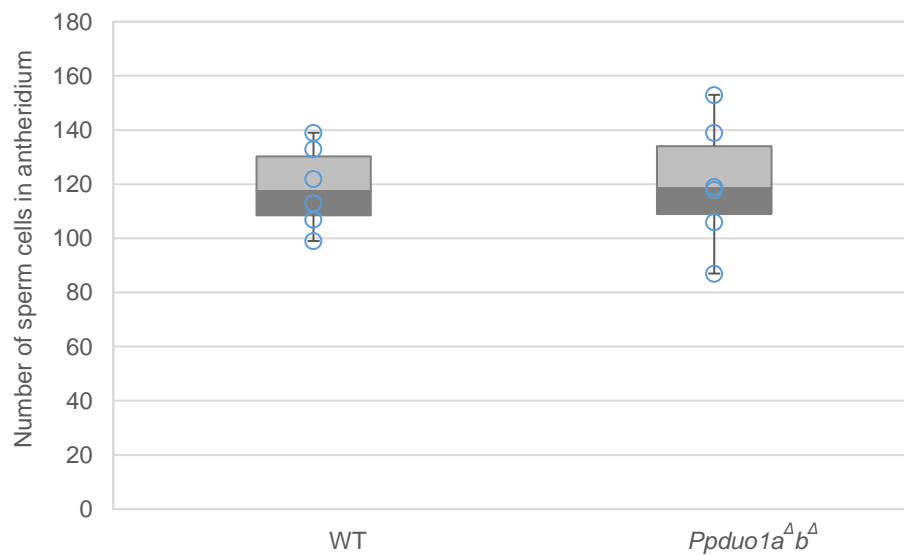
Figure S4.3. Number of SpCs counted in mature antheridia at day 12 post-induction. The average number of SpCs for wild type was 119 ± 14 SD while for *Ppduo1a$^\Delta$b$^\Delta$* was 120 ± 21 SD.
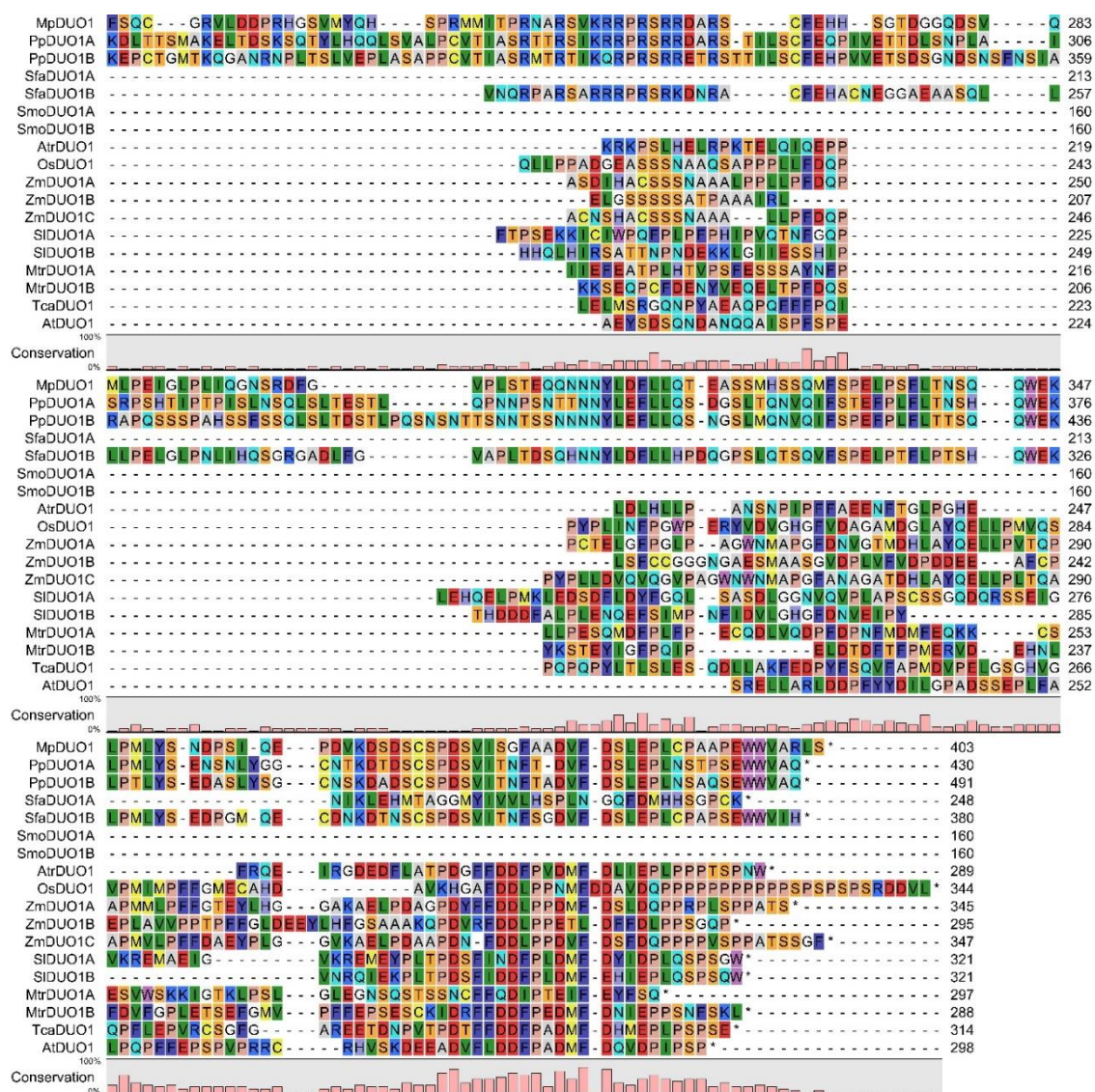
Table S4.2. One-way ANOVA for analysing any significant difference in the average number of sperm cells in mature antheridia of wild type and *Ppduo1a$^\Delta$b$^\Delta$*. For significantly different, p<0.05. *SS*=sum of squares; *df*=degrees of freedom; *MS*=mean square; *F*=F-value and *F crit*=F-critical value.

	No. of sperm cells in mature antheridia	
Antheridium	**WT**	***Ppduo1a$^\Delta$b$^\Delta$***
1	99	139
2	122	118
3	133	87
4	107	106
5	139	119
6	113	153

SUMMARY				
Groups	*Count*	*Sum*	*Average*	*Variance*
WT	6	713	118.8333	236.9667
Ppduo1a$^\Delta$b$^\Delta$	6	722	120.3333	547.8667

ANOVA						
Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	6.75	1	6.75	0.017	0.90	4.96
Within Groups	3924.167	10	392.4167			
Total	3930.917	11				
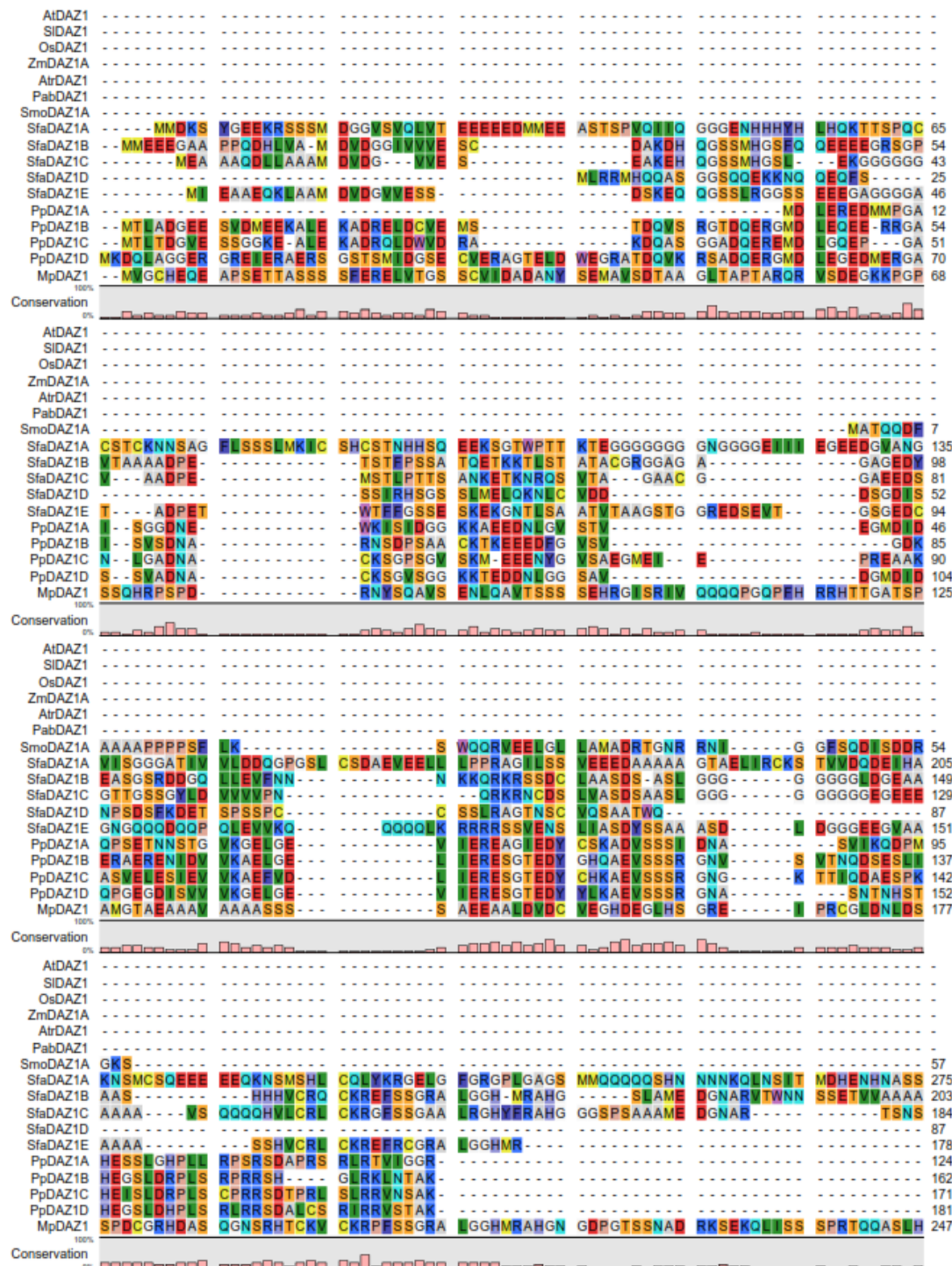
Figure S5.1. Multiple sequence alignment for PpDAZ1 homologue sequences from the bryophytes and species representing major groups of plant development. Image showed alignment for the whole sequence from start to stop codon. Basic region 1 (BR1) for AtrDAZ1 was located few amino acids ahead of the conserved BR1 alignment. Red box, BR1. Species: At, *Arabidopsis thaliana*; Sl, *Solanum lycopersicum*; Os, *Oryza sativa*; Zm, *Zea mays*; Atr, *Amborella trichopoda*; Pab, *Picea abies*; Smo,

Selaginella moellendorffii; Sfa, *Sphagnum fallax*; Pp, *Physcomitrella patens*; Mp, *Marchantia polymorpha*.
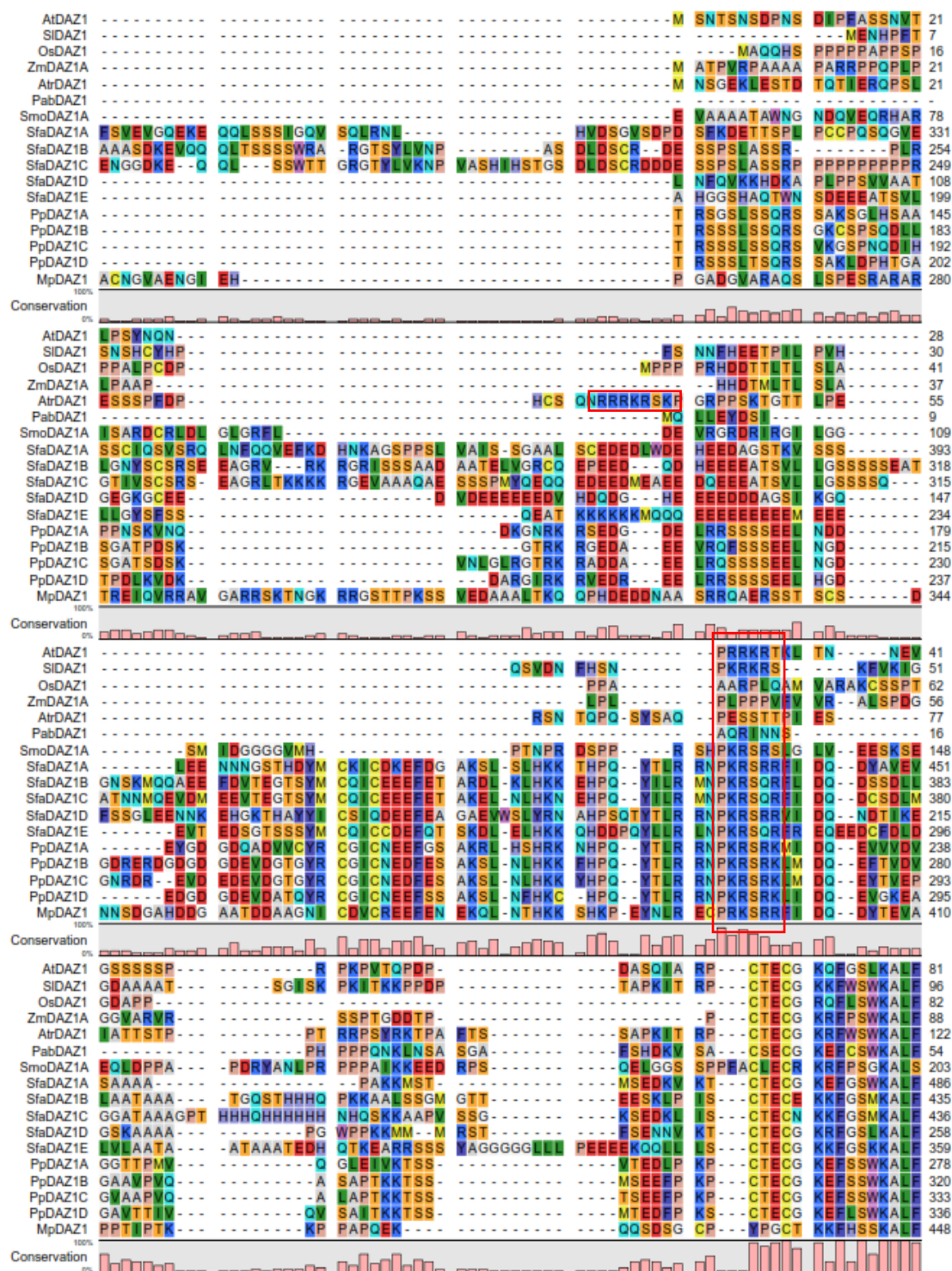
Figure S5.1.continue.
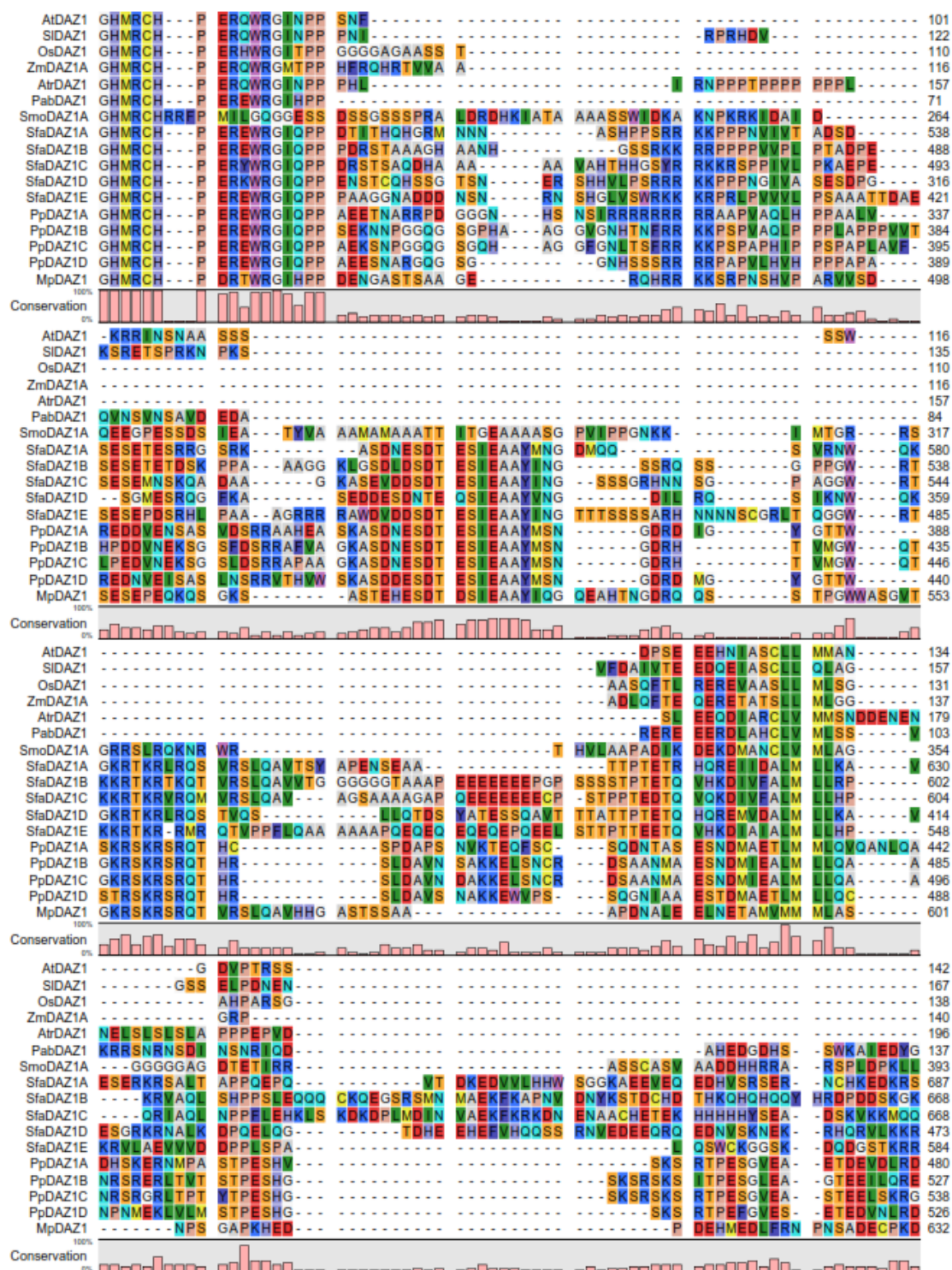
Figure S5.1.continue.
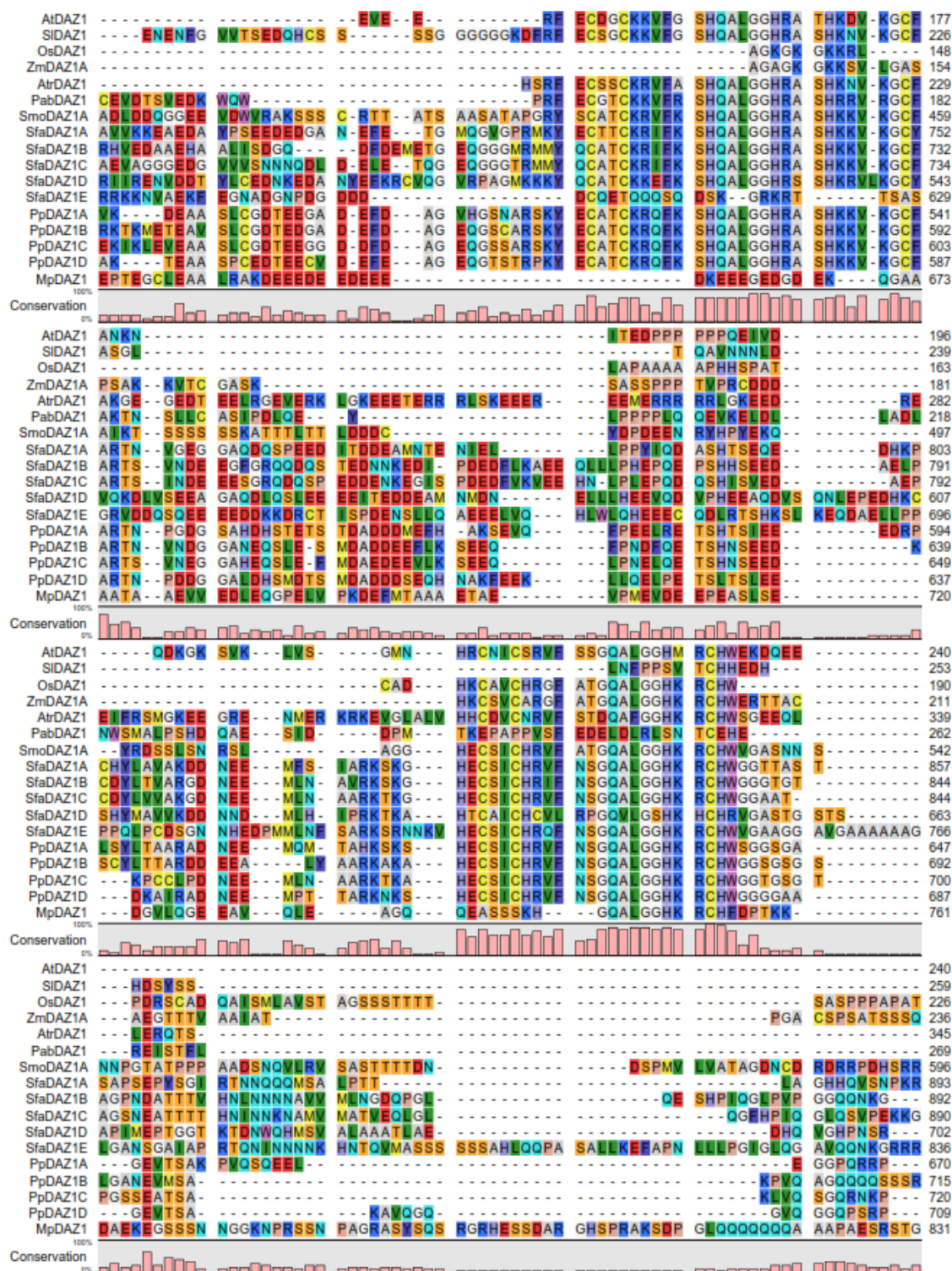
Figure S5.1.continue.
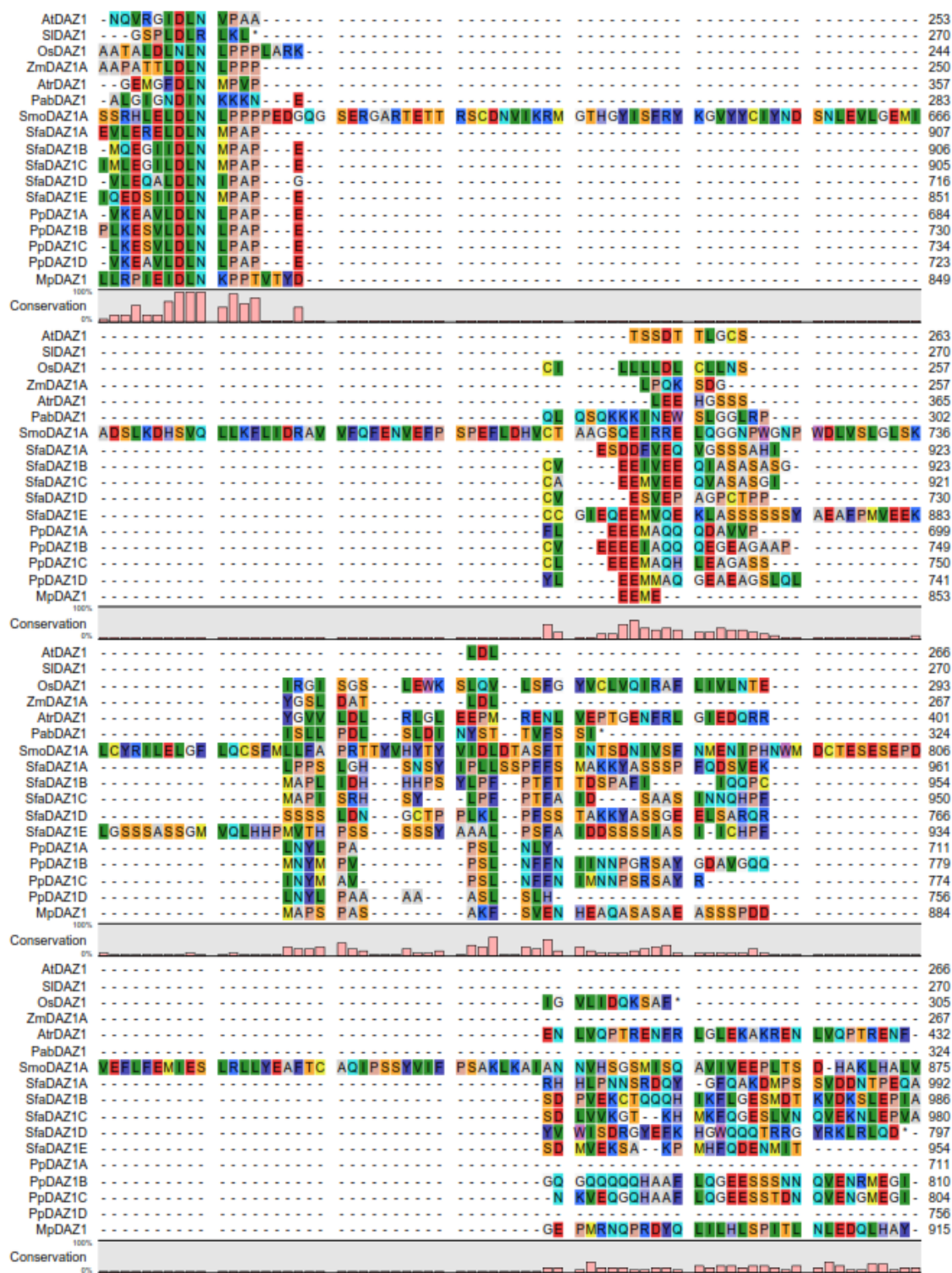
Figure S5.1.continue.
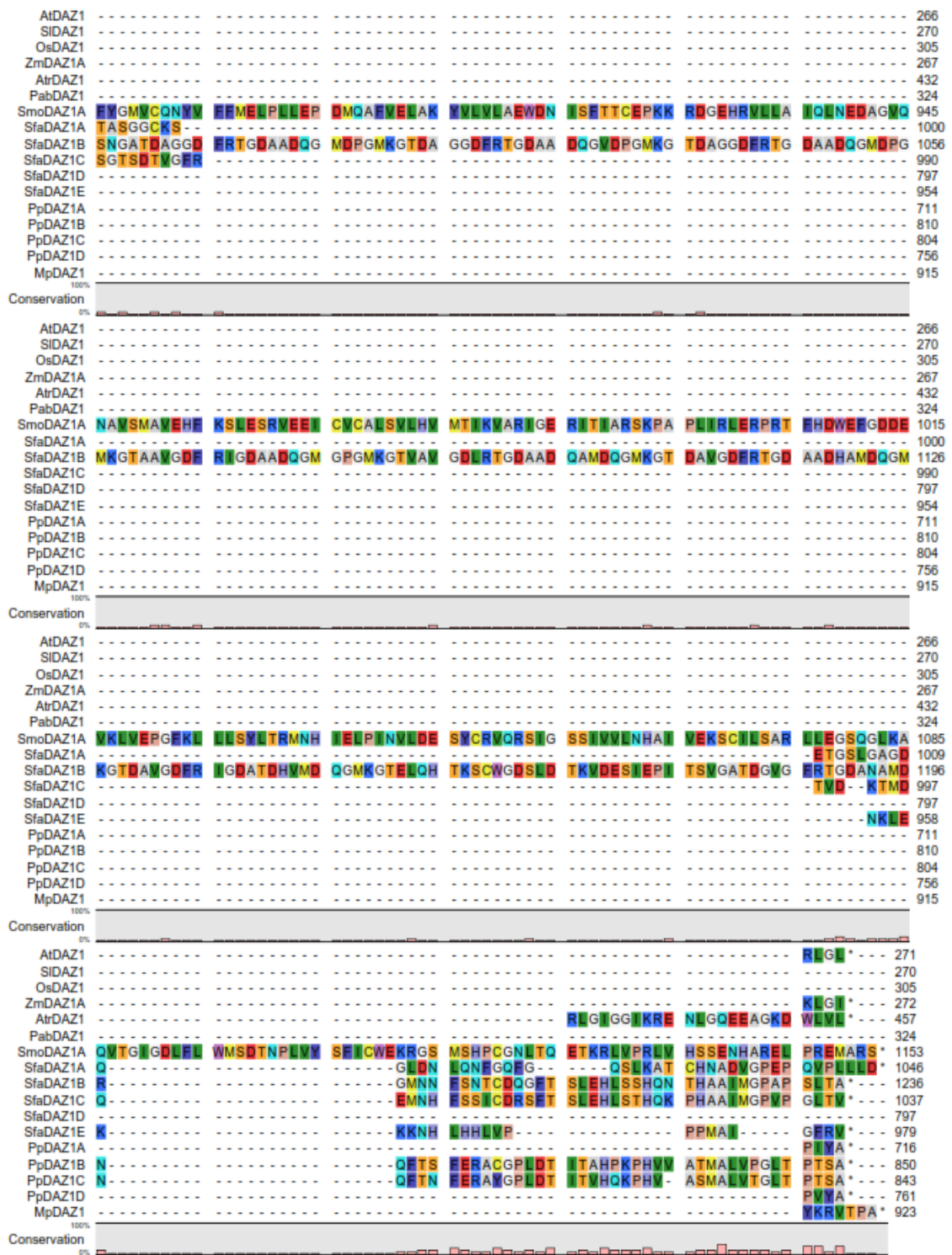
Figure S5.1.continue.

Table S5.1. Coordinates (position) of the conserved regions for PpDAZ1 family.

Position (amino acid)										
Protein	ZF1	BR1	ZF2	BR2	CR1	BR3	CR2	ZF3	ZF4	EAR
PpDAZ 1A	194 - 214	220 - 228	264 - 284	315 - 323	365 - 374	390 - 396	431 - 434	514 - 534	620 - 640	677 - 681
PpDAZ 1B	236 - 256	262 - 270	306 - 326	363 - 366	412 - 421	437 - 443	479 - 482	565 - 585	664 - 684	723 - 727
PpDAZ 1C	249 - 269	275 - 283	319 - 339	375 - 378	423 - 432	448 - 454	490 - 493	576 - 596	672 - 692	727 - 731
PpDAZ 1D	252 - 271	277 - 285	322 - 342	372 - 375	417 - 426	443 - 448	483 - 486	560 - 580	660 - 680	716 - 720

Table S5.2. Coordinates of *PpDAZ1* gene structure.

Gene	5' UTR	Exon 1	Intron 1	Exon 2	Intron 2	Exon 3	3' UTR
PpDAZ1A	1008	124	1160	1380	116	644	1254
PpDAZ1B	496	250	1607	1407	202	893	631
PpDAZ1C	2218	238	1361	1452	181	839	2092
PpDAZ1D	2798	298	1500	1344	138	641	749

Table S5.3. Individual replicates for antheridia RNA-seq data from Meyberg et al. (2020).

Antheridia (RPKM)									
	Gransden (Gd)				Reute (Re)				
Gene	I	II	III	Mean	I	II	III	Mean	Fold
PpDUO1 A	38.42	20.00	47.85	35.42	29.53	33.25	51.89	38.22	0.93
PpDUO1 B	29.86	27.13	80.38	45.79	39.17	26.90	71.70	45.92	1.00
PpDAZ1 A	5.42	5.16	4.72	5.10	3.30	5.35	5.48	4.71	1.08
PpDAZ1 B	6.93	6.01	14.02	8.99	4.47	2.69	10.44	5.86	1.53
PpDAZ1 C	2.31	3.70	7.33	4.45	4.01	3.79	7.38	5.06	0.88
PpDAZ1 D	0.88	1.28	0.85	1.00	1.89	1.48	0.87	1.41	0.71

Figure S5.2. DBS analysis of 58 potential DUO1 target genes in *P. patens*. The candidate genes were found using Compare Specificities tool in CoNekT and specifically expressed in antheridia. 2000 bp sequence upstream of ATG start codon were pattern
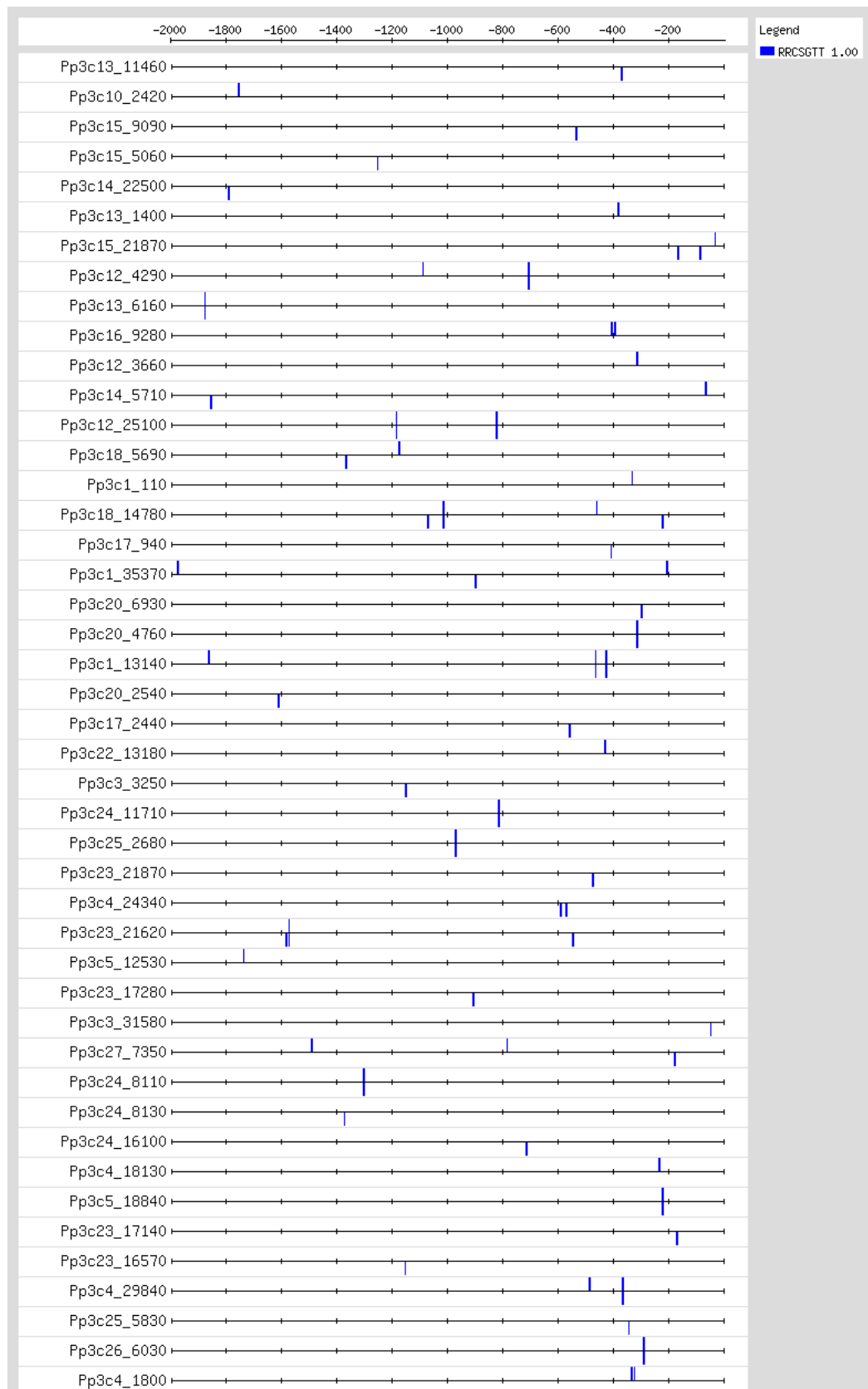
match with 'RRCSGTT' motif from Higo et al. (2018) using Regulatory Sequence Analysis Tools.
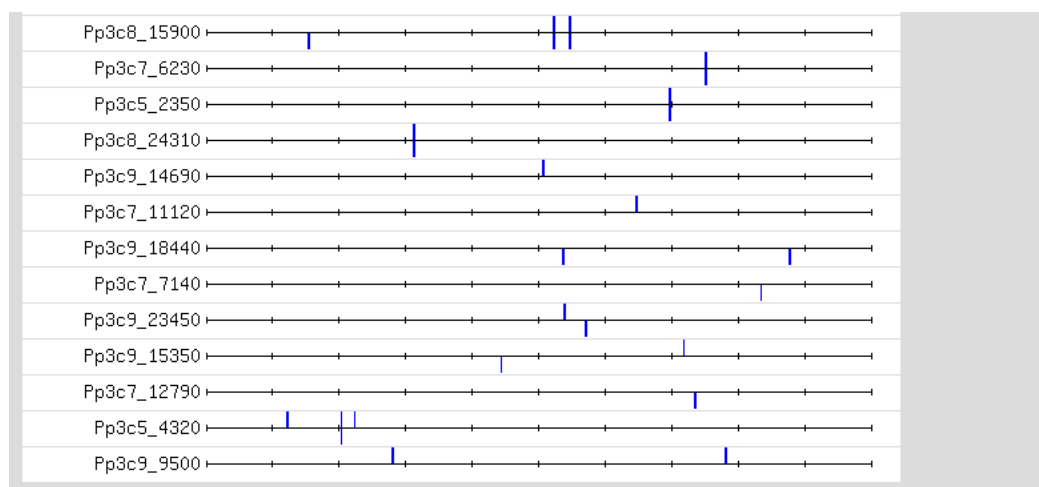

Figure S5.2.continue.

References

Anderson, S.N., Johnson, C.S., Jones, D.S., Conrad, L.J., Gou, X., Russell, S.D. and Sundaresan, V. (2013) Transcriptomes of isolated *Oryza sativa* gametes characterized by deep sequencing: evidence for distinct sex-dependent chromatin and epigenetic states before fertilization. *The Plant Journal*, 76(5), pp.729-741.

Ashton, N.W. and Raju, M.V.S. (2000) The distribution of gametangia on gametophores of *Physcomitrella* (*Aphanoregma*) patens in culture. *Journal of Bryology*, *22*(1), pp.9-12.

Bach-Pages, M., Homma, F., Kourelis, J., Kaschani, F., Mohammed, S., Kaiser, M., van der Hoorn, R.A., Castello, A. and Preston, G.M. (2020) Discovering the RNA-binding proteome of plant leaves with an improved RNA interactome capture method. *Biomolecules*, *10*(4), p.661.

Bailey, T.L., Boden, M., Buske, F.A., Frith, M., Grant, C.E., Clementi, L., Ren, J., Li, W.W. and Noble, W.S. (2009) MEME SUITE: tools for motif discovery and searching. *Nucleic acids research*, *37*(suppl_2), pp.W202-W208.

Borg, M., Brownfield, L., Khatab, H., Sidorova, A., Lingaya, M. and Twell, D. (2011) The R2R3 MYB transcription factor DUO1 activates a male germline-specific regulon essential for sperm cell differentiation in Arabidopsis. *The Plant Cell*, *23*(2), pp.534-549.

Borg, M., Rutley, N., Kagale, S., Hamamura, Y., Gherghinoiu, M., Kumar, S., Sari, U., Esparza-Franco, M.A., Sakamoto, W., Rozwadowski, K. and Higashiyama, T. (2014) An EAR-dependent regulatory module promotes male germ cell division and sperm fertility in Arabidopsis. *The Plant Cell*, 26(5), pp.2098-2113.

Borges, F., Gomes, G., Gardner, R., Moreno, N., McCormick, S., Feijó, J.A. and Becker, J.D. (2008) Comparative transcriptomics of Arabidopsis sperm cells. *Plant physiology*, *148*(2), pp.1168-1181.

Bowman, J.L., Kohchi, T., Yamato, K.T., Jenkins, J., Shu, S., Ishizaki, K., Yamaoka, S., Nishihama, R., Nakamura, Y., Berger, F. and Adam, C. (2017) Insights into land plant evolution garnered from the Marchantia polymorpha genome. *Cell*, *171*(2), pp.287-304.

Brownfield, L. and Twell, D. (2016) Plant reproduction. *eLS*, pp.1-12.

Brownfield, L., Hafidh, S., Borg, M., Sidorova, A., Mori, T. and Twell, D. (2009) A plant germline-specific integrator of sperm specification and cell cycle progression. *PLoS genetics*, *5*(3), p.e1000430.

Budke, J.M., Bernard, E.C., Gray, D.J., Huttunen, S., Piechulla, B. and Trigiano, R.N. (2018) Introduction to the special issue on bryophytes. *Critical Reviews in Plant Sciences*, *37*(2-3), pp.102-112

Chang, J., Yu, T., Yang, Q., Li, C., Xiong, C., Gao, S., Xie, Q., Zheng, F., Li, H., Tian, Z. and Yang, C. (2018) Hair, encoding a single C2H2 zinc-finger protein, regulates multicellular trichome formation in tomato. *The Plant Journal*, *96*(1), pp.90-102.

Charlesworth, B. and Charlesworth, D. (1978) A model for the evolution of dioecy and gynodioecy. *The American Naturalist*, *112*(988), pp.975-997.

Chase, M.W., Christenhusz, M.J.M., Fay, M.F., Byng, J.W., Judd, W.S., Soltis, D.E., Mabberley, D.J., Sennikov, A.N., Soltis, P.S. and Stevens, P.F. (2016) An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG IV. *Botanical Journal of the Linnean Society*, *181*(1), pp.1-20.

Chelsky, D., Ralph, R. and Jonak, G. (1989) Sequence requirements for synthetic peptide-mediated translocation to the nucleus. *Molecular and cellular biology*, *9*(6), pp.2487-2492.

Cho, S., Jang, S., Chae, S., Chung, K.M., Moon, Y.H., An, G. and Jang, S.K. (1999) Analysis of the C-terminal region of *Arabidopsis thaliana APETALA1* as a transcription activation domain. *Plant molecular biology*, *40*(3), pp.419-429.

Ciftci-Yilmaz, S. and Mittler, R. (2008) The zinc finger network of plants. *Cellular and Molecular Life Sciences*, *65*(7), pp.1150-1160.

Clough, S.J. and Bent, A.F. (1998) Floral dip: a simplified method for Agrobacterium-mediated transformation of Arabidopsis thaliana. *The plant journal*, *16*(6), pp.735-743.

Cove, D.J., Perroud, P.F., Charron, A.J., McDaniel, S.F., Khandelwal, A. and Quatrano, R.S. (2009) The moss *Physcomitrella patens*: a novel model system for plant development and genomic studies. *Cold Spring Harbor Protocols*, 2009(2), pp.pdb-emo115.

Cui, L., Feng, K., Wang, M., Wang, M., Deng, P., Song, W. and Nie, X. (2016) Genome-wide identification, phylogeny and expression analysis of AP2/ERF transcription factors family in Brachypodium distachyon. *BMC genomics*, *17*(1), pp.1-19.

Darbar, J.N. (2019) *Genetic analysis of the DAZ1 transcription factor in Arabidopsis thaliana* (Doctoral dissertation, University of Leicester).

De Smet, R., Adams, K.L., Vandepoele, K., Van Montagu, M.C., Maere, S. and Van de Peer, Y. (2013) Convergent gene loss following gene and genome duplications creates single-copy families in flowering plants. *Proceedings of the National Academy of Sciences*, *110*(8), pp.2898-2903.

Delwiche, C.F. and Cooper, E.D. (2015) The evolutionary origin of a terrestrial flora. *Current Biology*, *25*(19), pp.R899-R910.

Dolgikh, V.A., Pukhovaya, E.M. and Zemlyanskaya, E.V. (2019) Shaping ethylene response: the role of EIN3/EIL1 transcription factors. *Frontiers in plant science*, *10*, p.1030.

Du, H., Liang, Z., Zhao, S., Nan, M.G., Tran, L.S.P., Lu, K., Huang, Y.B. and Li, J.N. (2015) The evolutionary history of R2R3-MYB proteins across 50 eukaryotes: new insights into subfamily classification and expansion. *Scientific reports*, *5*(1), pp.1-16.

Du, H., Yang, S.S., Liang, Z., Feng, B.R., Liu, L., Huang, Y.B. and Tang, Y.X. (2012) Genome-wide analysis of the MYB transcription factor superfamily in soybean. *BMC plant biology*, *12*(1), pp.1-22.

Dubos, C., Stracke, R., Grotewold, E., Weisshaar, B., Martin, C. and Lepiniec, L. (2010) MYB transcription factors in Arabidopsis. *Trends in plant science*, *15*(10), pp.573-581.

Durbarry, A., Vizir, I. and Twell, D. (2005) Male germ line development in Arabidopsis. *duo* pollen mutants reveal gametophytic regulators of generative cell cycle progression. *Plant physiology*, *137*(1), pp.297-307.

Edgar, R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic acids research*, *32*(5), pp.1792-1797.

Edwards, K., Johnstone, C. and Thompson, C. (1991) A simple and rapid method for the preparation of plant genomic DNA for PCR analysis. *Nucleic acids research*, 19(6), p.1349.

Engel, M.L., Chaboud, A., Dumas, C. and McCormick, S. (2003) Sperm cells of *Zea mays* have a complex complement of mRNAs. *The Plant Journal*, 34(5), pp.697-707.

Engel, M.L., Holmes-Davis, R. and McCormick, S. (2005) Green sperm. Identification of male gamete promoters in Arabidopsis. *Plant physiology*, *138*(4), pp.2124-2133.

Engel, P.P. (1968) The induction of biochemical and morphological mutants in the moss Physcomitrella patens. *American Journal of Botany*, *55*(4), pp.438-446.

Englbrecht, C.C., Schoof, H. and Böhm, S. (2004) Conservation, diversification and expansion of C2H2 zinc finger proteins in the *Arabidopsis thaliana* genome. *BMC genomics*, *5*(1), pp.1-17.

Franco-Zorrilla, J.M., López-Vidriero, I., Carrasco, J.L., Godoy, M., Vera, P. and Solano, R. (2014) DNA-binding specificities of plant transcription factors and their potential to define target genes. *Proceedings of the National Academy of Sciences*, *111*(6), pp.2367-2372.

Genau, A.C., Li, Z., Renzaglia, K.S., Pozo, N.F., Nogué, F., Haas, F.B., Wilhelmsson, P.K., Ullrich, K.K., Schreiber, M., Meyberg, R., Grosche, C. and Rensing, S. A. (2021) HAG1 and SWI3A/B control of male germ line development in *P. patens* suggests conservation of epigenetic reproductive control across land plants. *Plant reproduction*, *34*(2), pp.149-173.

Gilbertson, L. (2003) Cre–lox recombination: Cre-ative tools for plant biotechnology. *TRENDS in Biotechnology*, *21*(12), pp.550-555.

Goodstein, D.M., Shu, S., Howson, R., Neupane, R., Hayes, R.D., Fazo, J., Mitros, T., Dirks, W., Hellsten, U., Putnam, N. and Rokhsar, D.S. (2012) Phytozome: a comparative platform for green plant genomics. *Nucleic acids research*, *40*(D1), pp.D1178-D1186.

Gupta, S., Stamatoyannopoulos, J.A., Bailey, T.L. and Noble, W.S. (2007) Quantifying similarity between motifs. *Genome biology*, *8*(2), pp.1-9.

Hackenberg, D. and Twell, D. (2019) The evolution and patterning of male gametophyte development. *Current topics in developmental biology*, *131*, pp.257-298.

Hafidh, S., Fíla, J. and Honys, D. (2016) Male gametophyte development and function in angiosperms: a general concept. *Plant reproduction*, 29(1-2), pp.31-51.

Han, G., Lu, C., Guo, J., Qiao, Z., Sui, N., Qiu, N. and Wang, B. (2020) C2H2 zinc finger proteins: master regulators of abiotic stress responses in plants. *Frontiers in plant science*, *11*, p.115.

Han, Y.C. and Fu, C.C. (2019) Cold-inducible MaC2H2s are associated with cold stress response of banana fruit via regulating MaICE1. *Plant cell reports*, *38*(5), pp.673-680.

Han, Y.C., Fu, C.C., Kuang, J.F., Chen, J.Y. and Lu, W.J. (2016) Two banana fruit ripening-related C2H2 zinc finger proteins are transcriptional repressors of ethylene biosynthetic genes. *Postharvest Biology and Technology*, *116*, pp.8-15.

He, C., da Silva, J.A.T., Wang, H., Si, C., Zhang, M., Zhang, X., Li, M., Tan, J. and Duan, J. (2019) Mining MYB transcription factors from the genomes of orchids (Phalaenopsis and Dendrobium) and characterization of an orchid R2R3-MYB gene involved in water-soluble polysaccharide biosynthesis. *Scientific reports*, *9*(1), pp.1-19.

Hernández-Hernández, B., Tapia-López, R., Ambrose, B.A. and Vasco, A. (2021) R2R3-MYB gene evolution in plants, incorporating ferns into the story. *International Journal of Plant Sciences*, *182*(1), pp.1-8.

Higo, A., Kawashima, T., Borg, M., Zhao, M., López-Vidriero, I., Sakayama, H., Montgomery, S.A., Sekimoto, H., Hackenberg, D., Shimamura, M. and Nishiyama, T. (2018) Transcription factor DUO1 generated by neo-functionalization is associated with evolution of sperm differentiation in plants. *Nature communications*, *9*(1), p.5283.

Higo, A., Niwa, M., Yamato, K.T., Yamada, L., Sawada, H., Sakamoto, T., Kurata, T., Shirakawa, M., Endo, M., Shigenobu, S. and Yamaguchi, K. (2016) Transcriptional Framework of Male Gametogenesis in the Liverwort *Marchantia polymorpha* L. *Plant and Cell Physiology*, *57*(2), pp.325-338.

Hisanaga, T., Yamaoka, S., Kawashima, T., Higo, A., Nakajima, K., Araki, T., Kohchi, T. and Berger, F. (2019) Building new insights in plant gametogenesis from an evolutionary perspective. *Nature plants*, *5*(7), pp.663-669.

Hiss, M., Meyberg, R., Westermann, J., Haas, F.B., Schneider, L., Schallenberg-Rüdinger, M., Ullrich, K.K. and Rensing, S.A. (2017) Sexual reproduction, sporophyte development and molecular variation in the model moss *Physcomitrella patens*: introducing the ecotype Reute. *The Plant Journal*, *90*(3), pp.606-620.

Honys, D. and Twell, D. (2004) Transcriptome analysis of haploid male gametophyte development in Arabidopsis. *Genome biology*, 5(11), p.R85.

Horst, N.A. and Reski, R. (2017) Microscopy of *Physcomitrella patens* sperm cells. *Plant Methods*, *13*(1), p.33.

Howe, K.L., Contreras-Moreira, B., De Silva, N., Maslen, G., Akanni, W., Allen, J., Alvarez-Jarreta, J., Barba, M., Bolser, D.M., Cambell, L. and Carbajo, M. (2020) Ensembl Genomes 2020—enabling non-vertebrate genomic research. *Nucleic acids research*, *48*(D1), pp.D689-D695.

Isernia, C., Bucci, E., Leone, M., Zaccaro, L., Di Lello, P., Digilio, G., Esposito, S., Saviano, M., Di Blasio, B., Pedone, C. and Pedone, P.V. (2003) NMR structure of the

single QALGGH zinc finger domain from the *Arabidopsis thaliana* SUPERMAN protein. *Chembiochem*, *4*(2-3), pp.171-180.

Jiao, Y., Leebens-Mack, J., Ayyampalayam, S., Bowers, J.E., McKain, M.R., McNeal, J., Rolf, M., Ruzicka, D.R., Wafula, E., Wickett, N.J. and Wu, X. (2012) A genome triplication associated with early diversification of the core eudicots. *Genome biology*, *13*(1), pp.1-14.

Julca, I., Ferrari, C., Flores-Tornero, M., Proost, S., Lindner, A.C., Hackenberg, D., Steinbachová, L., Michaelidis, C., Gomes Pereira, S., Misra, C.S., Kawashima, T., Borg, M., Berger, F., Goldberg, J., Johnson, M., Honys, D., Twell, D., Sprunck, S., Dresselhaus, T., Becker, J.D. and Mutwil, M. (2021) Comparative transcriptomic analysis reveals conserved programmes underpinning organogenesis and reproduction in land plants. *Nature Plants*, pp.1-17.

Jungnickel, M.K., Sutton, K.A., Baker, M.A., Cohen, M.G., Sanderson, M.J. and Florman, H.M. (2018) The flagellar protein Enkurin is required for mouse sperm motility and for transport through the female reproductive tract. *Biology of reproduction*, *99*(4), pp.789-797.

Kagale, S. and Rozwadowski, K. (2011) EAR motif-mediated transcriptional repression in plants: an underlying mechanism for epigenetic regulation of gene expression. *Epigenetics*, *6*(2), pp.141-146.

Kagale, S., Links, M.G. and Rozwadowski, K. (2010) Genome-wide analysis of ethylene-responsive element binding factor-associated amphiphilic repression motif-containing transcriptional regulators in Arabidopsis. *Plant physiology*, *152*(3), pp.1109-1134.

Katiyar, A., Smita, S., Lenka, S.K., Rajwanshi, R., Chinnusamy, V. and Bansal, K.C. (2012) Genome-wide classification and expression analysis of MYB transcription factor families in rice and Arabidopsis. *BMC genomics*, *13*(1), pp.1-19.

Khan, A., Fornes, O., Stigliani, A., Gheorghe, M., Castro-Mondragon, J.A., Van Der Lee, R., Bessy, A., Cheneby, J., Kulkarni, S.R., Tan, G. and Baranasic, D. (2018) JASPAR 2018: update of the open-access database of transcription factor binding profiles and its web framework. *Nucleic acids research*, *46*(D1), pp.D260-D266.

Kimura, M. (1980) A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *Journal of molecular evolution*, *16*(2), pp.111-120.

Kinsella, R.J., Kähäri, A., Haider, S., Zamora, J., Proctor, G., Spudich, G., Almeida-King, J., Staines, D., Derwent, P., Kerhornou, A. and Kersey, P. (2011) Ensembl BioMarts: a hub for data retrieval across taxonomic space. *Database*, *2011*.

Klempnauer, K.H., Gonda, T.J. and Bishop, J.M. (1982) Nucleotide sequence of the retroviral leukemia gene v-myb and its cellular progenitor c-myb: the architecture of a transduced oncogene. *Cell*, *31*(2), pp.453-463.

Klempnauer, K.H., Symonds, G., Evan, G.I. and Bishop, J.M. (1984) Subcellular localization of proteins encoded by oncogenes of avian myeloblastosis virus and avian leukemia virus E26 and by the chicken c-myb gene. *Cell*, *37*(2), pp.537-547.

Klug, A. (2010) The discovery of zinc fingers and their applications in gene regulation and genome manipulation. *Annual review of biochemistry*, *79*, pp.213-231.

Klug, A. and Schwabe, J.W. (1995) Zinc fingers. *The FASEB journal*, *9*(8), pp.597-604.

Kofuji, R., Yagita, Y., Murata, T. and Hasebe, M. (2018) Antheridial development in the moss *Physcomitrella patens*: implications for understanding stem cells in mosses. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *373*(1739), p.20160494.

Kofuji, R., Yoshimura, T., Inoue, H., Sakakibara, K., Hiwatashi, Y., Kurata, T., Aoyama, T., Ueda, K. and Hasebe, M. (2018) Gametangia development in the moss *Physcomitrella patens*. *Annual Plant Reviews online*, pp.167-181.

Kumar, S., Stecher, G. and Tamura, K. (2016) MEGA7: molecular evolutionary genetics analysis version 7.0 for bigger datasets. *Molecular biology and evolution*, *33*(7), pp.1870-1874.

Kumar, S., Stecher, G., Li, M., Knyaz, C. and Tamura, K. (2018) MEGA X: molecular evolutionary genetics analysis across computing platforms. *Molecular biology and evolution*, *35*(6), pp.1547-1549.

Landberg, K., Pederson, E.R., Viaene, T., Bozorg, B., Friml, J., Jönsson, H., Thelander, M. and Sundberg, E. (2013) The moss *Physcomitrella patens* reproductive organ development is highly organized, affected by the two *SHI/STY* genes and by the level of active auxin in the *SHI/STY* expression domain. *Plant physiology*, *162*(3), pp.1406-1419.

Lang, D., Ullrich, K.K., Murat, F., Fuchs, J., Jenkins, J., Haas, F.B., Piednoel, M., Gundlach, H., Van Bel, M., Meyberg, R. and Vives, C. (2018) The *Physcomitrella patens*

chromosome-scale assembly reveals moss genome structure and evolution. *The Plant Journal*, *93*(3), pp.515-533.

Lange, A., Mills, R.E., Lange, C.J., Stewart, M., Devine, S.E. and Corbett, A.H. (2007) Classical nuclear localization signals: definition, function, and interaction with importin α. *Journal of Biological Chemistry*, *282*(8), pp.5101-5105.

Lau, S.E., Schwarzacher, T., Othman, R.Y. and Harikrishna, J.A. (2015) dsRNA silencing of an R2R3-MYB transcription factor affects flower cell shape in a Dendrobium hybrid. *BMC plant biology*, *15*(1), pp.1-14.

Li, W., Ma, M., Feng, Y., Li, H., Wang, Y., Ma, Y., Li, M., An, F. and Guo, H. (2015) EIN2-directed translational regulation of ethylene signaling in *Arabidopsis*. *Cell*, *163*(3), pp.670-683.

Li, Z., Peng, R., Tian, Y., Han, H., Xu, J. and Yao, Q. (2016) Genome-wide identification and analysis of the MYB transcription factor superfamily in *Solanum lycopersicum*. *Plant and Cell Physiology*, *57*(8), pp.1657-1677.

Lin, Y., Yang, L., Chen, D., Zu, Y. and Tang, Z. (2013) A role for Ethylene-Insensitive3 in the regulation of hydrogen peroxide production during seed germination under high salinity in *Arabidopsis*. *Acta physiologiae plantarum*, *35*(5), pp.1701-1706.

Lu, P., Chai, M., Yang, J., Ning, G., Wang, G. and Ma, H. (2014) The Arabidopsis CALLOSE DEFECTIVE MICROSPORE1 gene is required for male fertility through regulating callose metabolism during microsporogenesis. *Plant physiology*, 164(4), pp.1893-1904.

Martin, C. and Paz-Ares, J. (1997) MYB transcription factors in plants. *Trends in Genetics*, *13*(2), pp.67-73.

McCarty, D.R., Hattori, T., Carson, C.B., Vasil, V., Lazar, M. and Vasil, I.K. (1991) The Viviparous-1 developmental gene of maize encodes a novel transcriptional activator. *Cell*, *66*(5), pp.895-905.

Menand, B., Yi, K., Jouannic, S., Hoffmann, L., Ryan, E., Linstead, P., Schaefer, D.G. and Dolan, L. (2007) An ancient mechanism controls the development of cells with a rooting function in land plants. *Science*, *316*(5830), pp.1477-1480.

Merchante, C., Brumos, J., Yun, J., Hu, Q., Spencer, K.R., Enríquez, P., Binder, B.M., Heber, S., Stepanova, A.N. and Alonso, J.M. (2015) Gene-specific translation regulation mediated by the hormone-signaling molecule EIN2. *Cell*, *163*(3), pp.684-697.

Meyberg, R., Perroud, P.F., Haas, F.B., Schneider, L., Heimerl, T., Renzaglia, K.S. and Rensing, S.A. (2020) Characterisation of evolutionarily conserved key players affecting eukaryotic flagellar motility and fertility using a moss model. *New Phytologist*, *227*(2), pp.440-454.

Miller, J., McLachlan, A.D. and Klug, A. (1985) Repetitive zinc-binding domains in the protein transcription factor IIIA from Xenopus oocytes. *The EMBO journal*, *4*(6), pp.1609-1614.

Monfared, M.M., Simon, M.K., Meister, R.J., Roig-Villanova, I., Kooiker, M., Colombo, L., Fletcher, J.C. and Gasser, C.S. (2011) Overlapping and antagonistic activities of *BASIC PENTACYSTEINE* genes affect a range of developmental processes in Arabidopsis. *The Plant Journal*, *66*(6), pp.1020-1031.

Mori, T., Igawa, T., Tamiya, G., Miyagishima, S.Y. and Berger, F. (2014) Gamete attachment requires GEX2 for successful fertilization in Arabidopsis. *Current Biology*, *24*(2), pp.170-175.

Mori, T., Kawai-Toyooka, H., Igawa, T. and Nozaki, H. (2015) Gamete dialogs in green lineages. *Molecular plant*, *8*(10), pp.1442-1454.

Mori, T., Kuroiwa, H., Higashiyama, T. and Kuroiwa, T. (2006) GENERATIVE CELL SPECIFIC 1 is essential for angiosperm fertilization. *Nature cell biology*, *8*(1), pp.64-72.

Müller, S.J., Gütle, D.D., Jacquot, J.P. and Reski, R. (2016) Can mosses serve as model organisms for forest research?. *Annals of forest Science*, *73*(1), pp.135-146.

Munné-Bosch, S., Simancas, B. and Müller, M. (2018) Ethylene signaling cross-talk with other hormones in *Arabidopsis thaliana* exposed to contrasting phosphate availability: Differential effects in roots, leaves and fruits. *Journal of plant physiology*, *226*, pp.114-122.

Nguyen, N.T.T., Contreras-Moreira, B., Castro-Mondragon, J.A., Santana-Garcia, W., Ossio, R., Robles-Espinoza, C.D., Bahin, M., Collombet, S., Vincens, P., Thieffry, D. and van Helden, J. (2018) RSAT 2018: regulatory sequence analysis tools 20th anniversary. *Nucleic acids research*, *46*(W1), pp.W209-W214.

Niklas, K.J. and Kutschera, U. (2010) The evolution of the land plant life cycle. *New Phytologist*, 185(1), pp.27-41.

Noguero, M., Atif, R.M., Ochatt, S. and Thompson, R.D. (2013) The role of the DNA-binding One Zinc Finger (DOF) transcription factor family in plants. *Plant Science*, *209*, pp.32-45.

Norizuki, T., Minamino, N. and Ueda, T. (2020) Role of autophagy in male reproductive processes in land plants. *Frontiers in Plant Science*, *11*, p.756.

Ogata, K., Hojo, H., Aimoto, S., Nakai, T., Nakamura, H., Sarai, A., Ishii, S. and Nishimura, Y. (1992) Solution structure of a DNA-binding unit of Myb: a helix-turn-helix-related motif with conserved tryptophans forming a hydrophobic core. *Proceedings of the National Academy of Sciences*, *89*(14), pp.6428-6432.

Ogata, K., Kanei-Ishii, C., Sasaki, M., Hatanaka, H., Nagadoi, A., Enari, M., Nakamura, H., Nishimura, Y., Ishii, S. and Sarai, A. (1996) The cavity in the hydrophobic core of Myb DNA-binding domain is reserved for DNA recognition and trans-activation. *Nature structural biology*, *3*(2), pp.178-187.

Ohta, M., Matsui, K., Hiratsu, K., Shinshi, H. and Ohme-Takagi, M. (2001) Repression domains of class II ERF transcriptional repressors share an essential motif for active repression. *The plant cell*, *13*(8), pp.1959-1968.

Omichinski, J.G., Clore, G.M., Appella, E., Sakaguchi, K. and Gronenborn, A.M. (1990) High-resolution three-dimensional structure of a single zinc finger from a human enhancer binding protein in solution. *Biochemistry*, *29*(40), pp.9324-9334.

O'Neill, J.P., Colon, K.T. and Jenik, P.D. (2019) The onset of embryo maturation in Arabidopsis is determined by its developmental stage and does not depend on endosperm cellularization. *The Plant Journal*, *99*(2), pp.286-301.

Panchy, N., Lehti-Shiu, M. and Shiu, S.H. (2016) Evolution of gene duplication in plants. *Plant physiology*, *171*(4), pp.2294-2316.

Paz-Ares, J., Ghosal, D., Wienand, U., Peterson, P.A. and Saedler, H. (1987) The regulatory *c1* locus of *Zea mays* encodes a protein with homology to *myb* proto-oncogene products and with structural similarities to transcriptional activators. *The EMBO journal*, *6*(12), pp.3553-3558.

Perroud, P.F., Cove, D.J., Quatrano, R.S. and McDaniel, S.F. (2011) An experimental method to facilitate the identification of hybrid sporophytes in the moss *Physcomitrella patens* using fluorescent tagged lines. *New Phytologist*, *191*(1), pp.301-306.

Perroud, P.F., Haas, F.B., Hiss, M., Ullrich, K.K., Alboresi, A., Amirebrahimi, M., Barry, K., Bassi, R., Bonhomme, S., Chen, H., Coates, J.C., Fujita, T., Guyon-Debast, A., Lang, D., Lin, J., Lipzen, A., Nogué, F., Oliver, M.J., Ponce de León, I., Quatrano, R.S., Rameau, C., Reiss, B., Reski, R., Ricca, M., Saidi, Y., Sun, N., Szövényi, P., Sreedasyam, A., Grimwood, J., Stacey, G., Schmutz, J. and Rensing, S.A. (2018) The Physcomitrella patens gene atlas project: large-scale RNA-seq based expression data. *The Plant Journal*, *95*(1), pp.168-182.

Peters, B., Casey, J., Aidley, J., Zohrab, S., Borg, M., Twell, D. and Brownfield, L. (2017) A conserved cis-regulatory module determines germline fate through activation of the transcription factor DUO1 promoter. *Plant physiology*, *173*(1), pp.280-293.

Proost, S. and Mutwil, M. (2018) CoNekT: an open-source framework for comparative genomic and transcriptomic network analyses. *Nucleic acids research*, *46*(W1), pp.W133-W140.

Pucker, B., Pandey, A., Weisshaar, B. and Stracke, R. (2020) The R2R3-MYB gene family in banana (*Musa acuminata*): Genome-wide identification, classification and expression patterns. *PloS one*, *15*(10), p.e0239275.

Qiao, X., Li, Q., Yin, H., Qi, K., Li, L., Wang, R., Zhang, S. and Paterson, A.H. (2019) Gene duplication and evolution in recurring polyploidization–diploidization cycles in plants. *Genome biology*, *20*(1), pp.1-23.

Qing, X.U., Jie, H.E., Jianhui, D.O.N.G., Xiaojin, H.O.U. and Zhang, X. (2018) Genomic survey and expression profiling of the MYB gene family in watermelon. *Horticultural Plant Journal*, *4*(1), pp.1-15.

Qiu, Y.L., Taylor, A.B. and McManus, H.A. (2012) Evolution of the life cycle in land plants. *Journal of Systematics and Evolution*, 50(3), pp.171-194.

Rehrauer, H., Aquino, C., Gruissem, W., Henz, S.R., Hilson, P., Laubinger, S., Naouar, N., Patrignani, A., Rombauts, S., Shu, H. and Van de Peer, Y. (2010) AGRONOMICS1: a new resource for Arabidopsis transcriptome profiling. *Plant physiology*, *152*(2), pp.487-499.

Rensing, S.A., Goffinet, B., Meyberg, R., Wu, S.Z. and Bezanilla, M. (2020) The moss Physcomitrium (Physcomitrella) patens: a model organism for non-seed plants. *The Plant Cell*, *32*(5), pp.1361-1376.

Renzaglia, K.S. and Garbary, D.J. (2001) Motile gametes of land plants: diversity, development, and evolution. *Critical reviews in plant sciences*, *20*(2), pp.107-213.

Reski, R. (1998) Development, genetics and molecular biology of mosses. *Botanica Acta*, *111*(1), pp.1-15.

Reski, R. (1998) *Physcomitrella* and *Arabidopsis*: the David and Goliath of reverse genetics. *Trends in Plant Science*, *6*(3), pp.209-210.

Riechmann, J.L., Heard, J., Martin, G., Reuber, L., Jiang, C.Z., Keddie, J., Adam, L., Pineda, O., Ratcliffe, O.J., Samaha, R.R. and Creelman, R. (2000) Arabidopsis transcription factors: genome-wide comparative analysis among eukaryotes. *Science*, *290*(5499), pp.2105-2110.

Rost, B. (1999) Twilight zone of protein sequence alignments. *Protein engineering*, *12*(2), pp.85-94.

Rotman, N., Durbarry, A., Wardle, A., Yang, W.C., Chaboud, A., Faure, J.E., Berger, F. and Twell, D. (2005) A novel class of MYB factors controls sperm-cell formation in plants. *Current Biology*, *15*(3), pp.244-248.

Russell, S.D., Gou, X., Wong, C.E., Wang, X., Yuan, T., Wei, X., Bhalla, P.L. and Singh, M.B. (2012) Genomic profiling of rice sperm cell transcripts reveals conserved and distinct elements in the flowering plant male germ lineage. *New Phytologist*, *195*(3), pp.560-573.

Rutley, N. and Twell, D. (2015) A decade of pollen transcriptomics. *Plant reproduction*, 28(2), pp.73-89.

Sanchez-Vera, V., Kenchappa, C.S., Landberg, K., Bressendorff, S., Schwarzbach, S., Martin, T., Mundy, J., Petersen, M., Thelander, M. and Sundberg, E. (2017) Autophagy is required for gamete differentiation in the moss *Physcomitrella patens*. *Autophagy*, *13*(11), pp.1939-1951.

Sari, U. (2015) *Pollen polymorphism in Magnoliaceae and conservation of plant male germline regulators* (Doctoral dissertation, University of Leicester).

Schaefer, D.G. and Zrÿd, J.P. (1997) Efficient gene targeting in the moss Physcomitrella patens. *The Plant Journal*, *11*(6), pp.1195-1206.

Schaefer, D.G. and Zrÿd, J.P. (2001) The moss *Physcomitrella patens*, now and then. *Plant physiology*, *127*(4), pp.1430-1438.

Schneider, C.A., Rasband, W.S. and Eliceiri, K.W. (2012) NIH Image to ImageJ: 25 years of image analysis. *Nature methods*, *9*(7), pp.671-675.

Schwinn, K.E., Ngo, H., Kenel, F., Brummell, D.A., Albert, N.W., McCallum, J.A., Pither-Joyce, M., Crowhurst, R.N., Eady, C. and Davies, K.M. (2016) The onion (*Allium cepa L.*) R2R3-MYB gene MYB1 regulates anthocyanin biosynthesis. *Frontiers in plant science*, *7*, p.1865.

Shanks, C.M., Hecker, A., Cheng, C.Y., Brand, L., Collani, S., Schmid, M., Schaller, G.E., Wanke, D., Harter, K. and Kieber, J.J. (2018) Role of BASIC PENTACYSTEINE transcription factors in a subset of cytokinin signaling responses. *The Plant Journal*, *95*(3), pp.458-473.

Shaw, J. and Renzaglia, K. (2004) Phylogeny and diversification of bryophytes. *American journal of botany*, *91*(10), pp.1557-1581.

Shi, Y., Tian, S., Hou, L., Huang, X., Zhang, X., Guo, H. and Yang, S. (2012) Ethylene signaling negatively regulates freezing tolerance by repressing expression of CBF and type-A ARR genes in Arabidopsis. *The Plant Cell*, *24*(6), pp.2578-2595.

Shikata, M., Takemura, M., Yokota, A. and Kohchi, T. (2003) Arabidopsis ZIM, a plant-specific GATA factor, can function as a transcriptional activator. *Bioscience, biotechnology, and biochemistry*, *67*(11), pp.2495-2497.

Stracke, R., Werber, M. and Weisshaar, B. (2001) The R2R3-MYB gene family in *Arabidopsis thaliana*. *Current opinion in plant biology*, *4*(5), pp.447-456.

Stubbs, L., Sun, Y. and Caetano-Anolles, D. (2011) Function and evolution of C2H2 zinc finger arrays. *A handbook of transcription factors*, pp.75-94.

Sutton, K.A., Jungnickel, M.K., Wang, Y., Cullen, K., Lambert, S. and Florman, H.M. (2004) Enkurin is a novel calmodulin and TRPC channel binding protein in sperm. *Developmental biology*, *274*(2), pp.426-435.

Taimur, N. (2014) *A study of molecular and genetic mechanisms mediating the formation of twin sperm cells in Arabidopsis thaliana* (Doctoral dissertation, University of Leicester).

Takatsuji, H. and Matsumoto, T. (1996) Target-sequence recognition by separate-type Cys2/His2 zinc finger proteins in plants. *Journal of Biological Chemistry*, *271*(38), pp.23368-23373.

ten Hove, C.A., Lu, K.J. and Weijers, D. (2015) Building a plant: cell fate specification in the early Arabidopsis embryo. *Development*, *142*(3), pp.420-430.

Twell, D. (2011) Male gametogenesis and germline specification in flowering plants. *Sexual Plant Reproduction*, 24(2), pp.149-160.

Twell, D., Klein, T.M., Fromm, M.E. and McCormick, S. (1989) Transient expression of chimeric genes delivered into pollen by microprojectile bombardment. *Plant physiology*, *91*(4), pp.1270-1274.

Ullah, A., Qamar, M.T.U., Nisar, M., Hazrat, A., Rahim, G., Khan, A.H., Hayat, K., Ahmed, S., Ali, W. and Yang, X. (2020) Characterization of a novel cotton MYB gene, GhMYB108-like responsive to abiotic stresses. *Molecular biology reports*, *47*(3), pp.1573-1581.

Vanderpoorten, A. and Goffinet, B. (2009) *Introduction to bryophytes.* Cambridge University Press.

von Besser, K., Frank, A.C., Johnson, M.A. and Preuss, D. (2006) *Arabidopsis HAP2* (*GCS1*) is a sperm-specific gene required for pollen tube guidance and fertilization. *Development*, *133*(23), pp.4761-4769.

Wang, L., Ko, E.E., Tran, J. and Qiao, H. (2020) TREE1-EIN3–mediated transcriptional repression inhibits shoot growth in response to ethylene. *Proceedings of the National Academy of Sciences*, *117*(46), pp.29178-29189.

Wang, Z.P., Xing, H.L., Dong, L., Zhang, H.Y., Han, C.Y., Wang, X.C. and Chen, Q.J. (2015) Egg cell-specific promoter-controlled CRISPR/Cas9 efficiently generates homozygous mutants for multiple target genes in Arabidopsis in a single generation. *Genome biology*, *16*(1), pp.1-12.

Wettstein, F.V. (1924) Morphologie und Physiologie des Formwechsels der Moose auf genetischer Grundlage. I. *Zeitschrift für induktive Abstammungs-und Vererbungslehre*, *33*(1), pp.1-236.

Wu, S.Z., Yamada, M., Mallett, D.R. and Bezanilla, M. (2018) Cytoskeletal discoveries in the plant lineage using the moss *Physcomitrella patens*. *Biophysical reviews*, *10*(6), pp.1683-1693.

Xiao, S.J., Zhang, C., Zou, Q. and Ji, Z.L. (2010) TiSGeD: a database for tissue-specific genes. *Bioinformatics*, *26*(9), pp.1273-1275.

Xing, H.L., Dong, L., Wang, Z.P., Zhang, H.Y., Han, C.Y., Liu, B., Wang, X.C. and Chen, Q.J. (2014) A CRISPR/Cas9 toolkit for multiplex genome editing in plants. *BMC plant biology*, *14*(1), pp.1-12.

Yadegari, R. and Drews, G.N. (2004) Female gametophyte development. *The Plant Cell, 16*(suppl 1), pp.S133-S141.

Yanagisawa, S. (2001) The transcriptional activation domain of the plant-specific Dof1 factor functions in plant, animal, and yeast cells. *Plant and Cell Physiology*, *42*(8), pp.813-822.

Yang, C., Li, H., Zhang, J., Luo, Z., Gong, P., Zhang, C., Li, J., Wang, T., Zhang, Y. and Ye, Z. (2011) A regulatory gene induces trichome formation and embryo lethality in tomato. *Proceedings of the National Academy of Sciences*, *108*(29), pp.11836-11841.

Yu, J., Tehrim, S., Wang, L., Dossa, K., Zhang, X., Ke, T. and Liao, B. (2017) Evolutionary history and functional divergence of the cytochrome P450 gene superfamily between *Arabidopsis thaliana* and Brassica species uncover effects of whole genome and tandem duplications. *BMC genomics*, *18*(1), pp.1-21.

Zhao, M. (2017) *Evolution of Plant Male Germline-Specific Transcription Factor DUO POLLEN 1* (Doctoral dissertation, University of Leicester).