



UNIVERSITY OF
BIRMINGHAM

DISSECTING SEED GENETIC NETWORKS THROUGH THE
EVOLUTION OF PLANT LIFE-CYCLE TRANSITIONS

by

KIRSTY MCCREADY

A thesis submitted to the University of Birmingham for the degree of
DOCTOR OF PHILOSOPHY

School of Biosciences
College of Life and Environmental Sciences
University of Birmingham
January 2025

UNIVERSITY OF
BIRMINGHAM

University of Birmingham Research Archive

e-theses repository

This unpublished thesis/dissertation is copyright of the author and/or third parties. The intellectual property rights of the author or third parties in respect of this work are as defined by The Copyright Designs and Patents Act 1988 or as modified by any successor legislation.

Any use made of information contained in this thesis/dissertation must be in accordance with that legislation and must be properly acknowledged. Further distribution or reproduction in any format is prohibited without the permission of the copyright holder.

© Copyright by KIRSTY MCCREADY, 2025

All Rights Reserved

Abstract

The evolution of seeds was a transformative event in land plant history, driving the diversification and ecological dominance of seed plants. While the genetic networks underpinning reproductive development are well-characterized in angiosperms, their conservation and function in earlier-diverging lineages remain poorly understood. This thesis investigates key reproductive regulators in the fern *Ceratopteris richardii*, a model system bridging the evolutionary gap between bryophytes and seed plants.

SPOROCTELESS/NOZZLE (SPL/NZZ) is a critical regulator of seed development in *Arabidopsis thaliana*, yet homologous genes are also found in seedless plants. In *C. richardii*, three *SPL*-like genes—*CrSPLA*, *CrSPLB*, and *CrSPLC*—were identified. Phylogenetic analysis placed *CrSPLB* and *CrSPLC* within the *SPL* clade, suggesting that *SPL* function predates the origin of seeds. Expression analyses revealed distinct patterns for each homolog, with *CrSPLB* showing enrichment in reproductive tissues, aligning with a potential ancestral role in sporangia development. Preliminary RNAi knockdown experiments targeting *CrSPLB* suggested phenotypic effects, such as increased sporangia number, implicating its role in regulating sporangia development. However, complementation assays in *Arabidopsis* were unsuccessful, leaving the question of functional conservation open for future research.

To explore broader genetic regulators of reproductive development in *C. richardii*, weighted gene co-expression network analysis (WGCNA) was employed. This analysis revealed modules of co-expressed genes associated with distinct stages of reproductive development. Modules linked to early reproductive fronds highlighted

processes such as transcriptional activation, flavonoid biosynthesis, and structural remodelling. In contrast, modules associated with later reproductive stages emphasized meiosis, DNA repair, and redox homeostasis. These findings suggest conservation of key regulatory pathways, with hub genes such as *KNAT3-related* and *MYB80* pointing to potential parallels with seed plant reproductive networks.

This thesis provides novel insights into the genetic regulation of reproduction in ferns and its evolutionary relevance to seed plant development. By integrating phylogenetic, expression, and network analyses, this work lays the groundwork for future studies on the conservation and divergence of reproductive regulatory networks across land plants.

Acknowledgements

This thesis would not have been possible without the unwavering support, encouragement and guidance of so many individuals along the way. First and foremost, I extend my heartfelt thanks to my supervisor, Dr. Juliet Coates, for her endless encouragement, patience and insight. Through her guidance, I have gained invaluable knowledge not only about science but also about myself - lessons I will carry with me for a lifetime.

To my lab partner and partner in crime, Dr. Derry Carr, I am deeply grateful to have shared this PhD journey with you. Thank you for your thoughtful gifts, kind words of encouragement during the writing process, and your friendship throughout.

I would like to express my sincere gratitude to the members of the Gibbs and Catoni labs for their support and camaraderie during my PhD. My thanks also extend to the central services and stores staff, whose efficiency and dedication were essential in enabling my work. I would also like to thank Dr. Andrew Plackett for his *Ceratopteris* knowledge and materials.

I am also incredibly grateful to the friends I made beyond the lab, especially those I met at Staff House. Thank you for welcoming me with open arms, celebrating the highs, and providing unwavering support through the lows.

Additionally, I would like to thank the BBSRC and University of Birmingham-funded MIBTP for supporting the work in this thesis. I would also like to thank the team at BEAR, especially Dr. Stephanie Thompson, for their indispensable training and resources, and the incredible opportunities they provided.

Finally, and most importantly, I would like to thank my family. To my partner, Stu Robinson, thank you for being my rock throughout this entire journey. Your constant support, patience and belief in me made this achievement possible, and I couldn't have done it without you. To my mum and dad, your unconditional love, encouragement and support have laid the foundation for everything I've achieved. Thank you for always believing in me and for being my greatest champions—I owe everything to you.

Contents

Abstract	i
Acknowledgements	iii
Contents	v
List of Figures	xiii
List of Tables	xvii
1 Introduction	1
1.1 Journey to the Seed	4
1.1.1 Embryophytes: a reproductive revolution	4
1.1.2 The gametophyte dominated early embryophyte life cycles	8
1.1.3 Sporophyte dominance in the tracheophytes	12
1.1.4 The rise of seed plants	18
1.1.5 Angiosperms: masters of reproductive innovation	20
1.1.6 Ovule development in Angiosperms: a key reproductive innovation	23
1.2 <i>SPOROCTELESS</i> : Master regulator of gametophyte development	28
1.2.1 The structure and functional domains of the SPL protein	29
1.2.2 <i>SPL</i> expression is highly localised in the developing ovule	32
1.2.3 <i>AGAMOUS</i> -directed <i>SPL</i> expression shapes floral reproductive organs	33

1.2.4	Tight regulation of <i>SPL</i> results in the specification of a single MMC precursor	35
1.2.5	<i>SPL</i> acts as an adaptor-like transcriptional co-repressor to regulate downstream ovule genes	40
1.3	<i>Ceratopteris richardii</i> : a model for evo-devo studies	44
1.4	Aims and Objectives	47
2	Materials and Methods	49
2.1	Plant growth and maintenance	50
2.1.1	<i>Ceratopteris richardii</i> sterile tissue culture	50
2.1.2	<i>C. richardii</i> sporophyte growth on soil	51
2.1.3	<i>Arabidopsis thaliana</i> growth and maintenance	51
2.2	Isolation of Plant Nucleic acids	53
2.2.1	Isolation of genomic DNA from <i>A. thaliana</i>	53
2.2.2	Isolation of RNA from <i>C. richardii</i> and <i>A. thaliana</i>	53
2.2.3	cDNA synthesis of extracted RNA	54
2.3	Expression analyses	55
2.3.1	PCR with GoTAQ polymerase	55
2.3.2	Quantitative RT-PCR	56
2.4	Cloning	58
2.4.1	Primer design	58
2.4.2	PCR with Phusion High Fidelity Polymerase	59
2.4.3	Agarose gel electrophoresis	59
2.4.4	Isolation of DNA fragments from agarose gels	60
2.4.5	Ligation	60
2.4.6	Transformation of <i>E. coli</i> DH5 α	65
2.4.7	Inoculation of Bacterial Liquid Cultures	67
2.4.8	Isolation of plasmid DNA from <i>E. coli</i>	67
2.4.9	Digestion	67

2.4.10	Sequencing	67
2.5	Genetic Transformation of <i>C. richardii</i>	69
2.5.1	Callogenesis	69
2.5.2	Large Scale preparation of plasmid DNA for transformation	69
2.5.3	Preparation of gold microparticles	70
2.5.4	Biolistic Bombardment of <i>C. richardii</i> callus	71
2.5.5	Regeneration of T ₀ transgenic sporophytes	72
2.5.6	Selection of T ₁ transgenic lines	72
2.5.7	Phenotypic analysis of T ₁ Transgenics	73
2.5.8	Expression analysis of T ₁ Transgenics	74
2.6	Phylogenetic analysis of embryophyte SPEAR proteins	76
2.6.1	SPEAR sequence acquisition	76
2.6.2	Comparison of similarity and identity within and between <i>C. richardii</i> and <i>A. thaliana</i> SPEAR protein sequences	78
2.6.3	Multiple sequence alignment and model selection	78
2.6.4	Phylogenetic inference	78
2.7	Processing and Analysis of RNA sequencing data	80
2.7.1	Sample and library preparation	80
2.7.2	Raw sequence quality control	80
2.7.3	Mapping reads to the genome with HiSAT2	81
2.7.4	Mapping reads to the transcriptome with Salmon	81
2.7.5	<i>de novo</i> assembly of unmapped reads	82
2.7.6	Read normalisation and filtering	83
2.7.7	Sample clustering and dimensionality reduction	84
2.8	Weighted Gene Correlatation Network Analysis	85
2.8.1	Choosing the soft threshold power	85
2.8.2	Topological Overlap Matrix (TOM) construction	85
2.8.3	Dynamic module detection	86

2.8.4	Module Eigengene calculation and merging	86
2.8.5	Module-trait correlation analysis	87
2.8.6	Hub Gene Identification	87
2.8.7	GO Term Enrichment Analysis	88
2.8.8	Gene Expression Patterns	88
2.8.9	Network Visualisation in Cytoscape	88
3	When did <i>SPL</i> evolve? Tracing the evolutionary history of the <i>SPEAR</i> family across the plant kingdom	91
3.1	Introduction	92
3.1.1	SPLs and SPEARS: structurally indistinguishable	92
3.1.2	SPL function: exclusively angiosperm?	93
3.1.3	<i>SPL</i> expression: a reproductive link across plant lineages	94
3.1.4	Where might <i>SPL</i> expression be found in a non-seed plant?	96
3.1.5	Aim and objectives	97
3.2	Results	98
3.2.1	Identifying <i>SPEAR</i> homologs in <i>Ceratopteris richardii</i>	98
3.2.2	<i>SPEAR</i> homologs exist throughout the embryophytes	100
3.2.3	<i>SPL</i> forms a clade distinct to the remaining <i>SPEARS</i>	107
3.2.4	Where are <i>CrSPLs</i> expressed?	113
3.3	Discussion	120
3.3.1	The <i>SPL</i> and <i>EAR</i> domains are highly conserved in land plant <i>SPEARS</i>	120
3.3.2	<i>SPL</i> function may predate the evolution of the seed	122
3.3.3	The three <i>CrSPLs</i> have different expression patterns	125
4	What are the functions of <i>CrSPLs</i>?	129
4.1	Introduction	130

4.1.1	SPL in flowering plants: conserved roles and context-specific adaptations	130
4.1.2	Investigating the evolution of <i>SPL</i> function using <i>Ceratopteris richardii</i>	131
4.1.3	Transformation methods in <i>C. richardii</i>	132
4.1.4	RNAi gene silencing in <i>C. richardii</i>	134
4.1.5	Aims and Objectives	135
4.2	Results	136
4.2.1	Generation of <i>CrSPL</i> RNAi knockdown constructs	136
4.2.2	Generation of T0 <i>CrSPL</i> -RNAi-pANDA transgenic plants	141
4.2.3	Screening T1 <i>CrSPLB</i> -RNAi transgenics	144
4.2.4	T1 <i>CrSPLB</i> -RNAi plants are not phenotypically different to CrWT .	148
4.2.5	Cross-complementation of the <i>Arabidopsis spl-4</i> mutant with <i>CrSPLs</i>	154
4.3	Discussion	166
4.3.1	<i>CrSPL</i> -RNAi transformation efficiency was low	166
4.3.2	<i>CrSPLB</i> -RNAi T1 survival and transgenic efficiency was low . . .	167
4.3.3	<i>CrSPLB</i> -RNAi were not phenotypically different to WT	168
4.3.4	Insertion of complementation cassettes into the plant transformation vector pART27 was not successful	170
5	Analysis of transcriptional changes during development of the <i>C. richardii</i> sporophyte by RNA sequencing	171
5.1	Introduction	172
5.1.1	Genomic and transcriptomic resources in <i>Ceratopteris richardii</i> .	172
5.1.2	Read mapping and quantification approaches: HISAT2 vs Salmon	174
5.1.3	From mapping to co-expression analysis	175
5.1.4	Key concepts of WGCNA	176
5.1.5	Aims and Objectives	178

5.2	Results	180
5.2.1	RNA preparation and sequencing	180
5.2.2	Comparison of genome-led and transcriptome-led mapping of RNAseq reads	183
5.2.3	Data Normalisation	188
5.2.4	Sample clustering and dimensionality reduction	189
5.2.5	Weighted Gene Co-expression Network Analysis (WGCNA)	194
5.2.5.1	Soft-threshold selection for scale-free network construction	194
5.2.5.2	Gene co-expression network construction	195
5.2.6	WGCNA identified gene modules associated with tissue type and plant age	199
5.2.6.1	Validation of module gene expression patterns	203
5.2.7	Exploring reproductive gene modules in the developing sporophyte	206
5.2.7.1	Ivory module: gene-trait correlation analysis	206
5.2.7.2	Ivory module: Gene Ontology enrichment analysis	209
5.2.7.3	Ivory module: Network visualisation	212
5.2.7.4	Darkorange module: gene-trait correlation analysis	214
5.2.7.5	Darkorange module: Gene Ontology enrichment analysis	216
5.2.7.6	Darkorange module: network visualisation	221
5.2.8	Co-expression network assignment of <i>SPL</i> homologs	224
5.3	Discussion	227
5.3.1	Mapping Performance Reveals a Well-Resolved <i>C. richardii</i> Genome but a Limited Transcriptome	227
5.3.2	Normalisation and filtering of mapped reads	229
5.3.3	Co-expression modules revealed distinct gene networks across developmental stages	231

5.3.4	Regulatory networks specific to early reproductive frond development: insights from the ivory module	232
5.3.5	Regulatory networks driving later reproductive frond development: insights from the darkorange module	234
5.3.6	Identifying future targets in reproductive development	237
6	General Discussion	239
6.1	Introduction: Uncovering the genetic basis of reproductive development	240
6.2	Chapter 3: Evolutionary insights into <i>C. richardii</i> SPL	242
6.2.1	Phylogenetic insights into SPL evolution	242
6.2.2	Expression patterns of <i>CrSPLs</i> in reproductive development	243
6.2.3	Conservation and divergence of <i>CrSPL</i> function in land plants . .	243
6.3	Chapter 4: Investigating the functional roles of <i>CrSPLs</i>	245
6.3.1	Challenges in microparticle bombardment and transformation efficiency	245
6.3.2	Preliminary insights from <i>CrSPLB-RNAi</i> lines	246
6.3.3	Future directions for functional analysis of <i>CrSPL</i> genes	246
6.4	Chapter 5: Exploring transcriptional networks in reproductive development	248
6.4.1	Evaluating the current transcriptomic resources	248
6.4.2	Advancing Co-expression network analysis in ferns	249
6.4.3	Insights into reproductive development from the ivory and darkorange modules	250
6.4.4	A framework for future investigations	251
6.5	Conclusion	252
	References	253
A	Growth media and buffer recipes	295
B	Plasmid cloning results	299

C	<i>De novo</i> assembly of Salmon unmapped reads	307
D	Module Gene Lists	311

List of Figures

1.1	The evolution of the land plants.	5
1.2	Comparative life cycles of haplontic and haplodiplontic plants.	7
1.3	The life cycle of <i>Physcomitrium patens</i>	11
1.4	The life cycle of <i>Ceratopteris richardii</i>	17
1.5	The life cycle of <i>Arabidopsis thaliana</i>	22
1.6	Five stages of ovule development in <i>Arabidopsis thaliana</i>	24
1.7	<i>Arabidopsis thaliana</i> SPL protein structure	30
1.8	Proposed model for the restriction of <i>SPL</i> expression to specify a single MMC precursor.	38
2.1	pJET1.2/blunt cloning vector.	62
2.2	pDONR207 Gateway entry vector.	63
2.3	pANDA vector map.	64
2.4	35S::OCS-pART7 donor vector for complementation	66
2.5	Phenotyping measurements taken from <i>C. richardii</i> T ₁ RNAi lines.	75
3.1	Pairwise comparisons of percentage similarity and identity between and amongst <i>C. richardii</i> and <i>Arabidopsis</i> SPEAR sequences.	99
3.2	Consensus sequence of conserved regions in embryophyte SPEAR alignment.	107
3.3	Maximum likelihood phylogenetic tree of embryophyte SPEAR amino acid sequences.	109

3.4	Bayesian Inference Phylogenetic tree of embryophyte SPEAR amino acid sequences.	111
3.5	Expression of <i>Ceratopteris</i> SPL homologs across developmentally young tissue types.	115
3.6	Fragments per Kilobase of Transcripts per Million (FPKM) counts of <i>CrSPLA</i> , <i>CrSPLB</i> and <i>CrSPLC</i> in sporophyte tissue of different ages and developmental stages.	119
4.1	Multiple sequence alignment of the first exon of <i>CrSPLA</i> , <i>CrSPLB</i> and <i>CrSPLC</i> , highlighting the regions selected for RNAi.	137
4.2	Diagnostic gels during cloning procedure of <i>CrSPLA</i> , <i>CrSPLB</i> and <i>CrSPLC</i> RNAi target fragments.	139
4.3	Transformation process of <i>Ceratopteris richardii</i>	142
4.4	Survival of <i>CrSPLB</i> -RNAi transformants in the T0 to T2 generations.	146
4.5	Survival of <i>CrpAND</i> Ae transformants in the T0 to T2 generations.	147
4.6	<i>CrSPLB</i> expression in the reproductive fiddleheads of T1 <i>CrSPLB</i> -RNAi lines relative to CrWT.	149
4.7	Phenotyping analysis of mature reproductive frond phenotypes of <i>CrSPLB</i> -RNAi lines batch 1 (A, C, E, G) and batch 2 (B, D, F, H).	152
4.8	Schematic of the restriction cloning steps taken to generate tagged pSPL::SPL-mCHERRY::SPL3'UTR-pART7 and untagged pSPL::SPL::SPL3'UTR-pART7 complementation vectors.	156
4.9	Cloning fragments for the complementation cassettes.	159
4.10	Step-wise construction of the tagged and untagged complementation vectors pSPL::mCHERRY::SPL3'UTR and pSPL::SPL3'UTR	162
4.11	Insertion of <i>AtSPL</i> and <i>AtSPL</i> ΔSTOP into pSPL::SPL3'UTR-pART7 and pSPL::mCHERRY::SPL3'UTR-pART7.	165
5.1	Scale-free network structure.	177

5.2	Developmental stages of <i>C. richardii</i> analysed by RNA sequencing. . . .	181
5.3	Overview of the RNAseq processing pipeline	184
5.4	Mapping proportion resulting from HISAT2 and Salmon alignment methods.	187
5.5	Violin plot showing the distribution of FPKM-normalised expression values across all samples.	189
5.6	Hierarchical clustering dendrogram (A) and sample distance matrix (B) of RNA sequencing data from <i>Ceratopteris richardii</i> sporophyte samples.	191
5.7	Principle Component Analysis (PCA) of normalised gene expression data.	193
5.8	Soft-threshold power selection for WGCNA.	195
5.9	Gene module detection and merging.	198
5.10	Gene modules exhibit distinct correlations with developmental stages and plant age in <i>Ceratopteris richardii</i>	201
5.11	Expression profiles of positively correlated genes within each module. .	204
5.12	Expression profiles of negatively correlated genes within each module. .	205
5.13	Gene-trait correlation heatmap for the ivory module.	208
5.14	Gene Ontology (GO) enrichment analysis of the ivory module.	211
5.15	Network visualisation of the ivory module.	214
5.16	Gene-trait correlation heatmap for the darkorange module.	216
5.17	Gene Ontology (GO) enrichment analysis of the darkorange module. . .	219
5.18	Network visualisation of the darkorange module.	222
C.1	<i>De novo</i> assembly of unmapped reads from quasi-alignment with Salmon yielded 822 “missing genes”.	308

List of Tables

2.1	Primers used in PCR expression analyses.	55
2.2	Primers, sequences, and targets for qPCR analysis.	56
2.3	Primers used to amplify <i>CrSPL</i> RNAi target regions with <i>attB</i> sites	58
2.4	Primers used to amplify complementation construct components (underline) and add restriction sites (bold).	59
2.5	Primers used to sequence plasmids.	68
2.6	Species, versions and sources of datasets searched for SPEAR homologs and included in subsequent alignments and phylogenies. . .	77
3.1	SPEAR homologs identified in the <i>Ceratopteris richardii</i> genome	99
3.2	Genome and transcriptome databases searched for SPEARS and the number of full-length, validated sequences found in each major group and how many were included in the final alignment.	101
3.3	Taxonomy and gene accessions of SPEAR sequences included in final phylogeny.	102
4.1	Restriction site occurrences for pART7 and pART27 plasmids.	155
5.1	Module membership and the corresponding number of genes in merged modules.	199
5.2	Annotation and putative function of hub genes in the ivory module. . . .	209
5.3	Transcription factors identified within the ivory module.	212
5.4	Annotation and putative functions of hub genes in the darkorange module.	217

5.5 Transcription factors identified in the darkorange module along with their
corresponding annotations and putative functions. 224

B.1 Table summarising the results of all plasmids attempted for cloning . . 299

D.1 GO annotation and best hits for genes in the ivory module. 312

D.2 GO annotation and best hits for genes in the darkorange module. Hub
genes are in bold. 319

Chapter 1

Introduction

Seeds nourish. Seeds unite. Seeds
endure. Seeds defend. Seeds travel.

Thor Hanson, 2015
The Triumph of Seeds

Seeds feed, clothe and fuel our civilisation. Just three major seed crops - rice, wheat and maize - provide approximately 70% of total global calorie consumption (Food and Agriculture Organization of the United Nations, [2023](#)). If we include the other roughly 250 plant species that we eat, as well as indirect sources such as seed-derived oils and seed-fed livestock, it is arguable that almost all of our human diet originates from seed. Global demand for food is increasing rapidly, with population and consumption growth not expected to slow until at least the middle of this century (Godfray et al., [2010](#)). Population growth, limited arable land and fresh water, and climate change present significant challenges for agriculture in meeting increased demand for food, feed, fuel and fibres (Fedoroff et al., [2010](#)). Producing better quality seeds with greater efficiency is critical to meeting this demand. To achieve this, we must first understand the genetics underpinning seed development and evolution: when and where did seeds come from, and how did they come about?

Seeds are a remarkable evolutionary innovation. Before the seed, plant sexual reproduction relied on free water to disperse, germinate and unite fragile gametes, risking the release of the next generation into adverse, uninhabitable environments (Bateman & DiMichele, [1994](#)). With the appearance of the seed habit, the female gametophyte is retained and nurtured on the mother plant, the male gamete is transported in protective pollen, and the fertilised embryo is enclosed within a nourishing and protective seed coat until favourable environmental conditions facilitate its release (Linkies et al., [2010](#)). This new protection of the sexual gametes, and their release from dependence on free water, enabled progeny to traverse distance and time into previously uninhabitable environments (Bateman & DiMichele,

1994; Prestianni & Gerrienne, 2010). Certainly, this fundamentally new kind of reproductive strategy paved the way for Darwin's "abominable mystery"; the rapid radiation and diversification of the flowering plants some 200 million years later (Darwin, 1879; Harris & Davies, 2016; Linkies et al., 2010; Willis et al., 2014).

Over 300 million years of functional gains and losses pre-dated the evolution of the seed. To uncover how the seed evolved, we must understand these gains and losses at a genetic, morphological and evolutionary level. Unfortunately, incomplete deposition and taphonomic degradation mean that we cannot rely on fossils of extinct transition species to provide a complete picture of seed evolution (Harrison & Morris, 2018). Furthermore, these snapshots in time, when available, cannot be experimented with in the same way that living plants can (Meade et al., 2021). Meanwhile, whilst extant representatives are amenable to experimentation, they are not direct representatives of their ancestors, and interpretation is required (McDaniel, 2021). Therefore, to understand the exact morphological and genetic transformations that led to the seed, we must combine knowledge from molecular biology, developmental genetics, palaeobotany and phylogenetics to generate a full, well-informed picture. Over the past two decades we have seen significant advances in plant genomics, a rise in the amenability of non-model organisms, and innovations in fossil reconstruction (Borrill, 2020; Delaux et al., 2019; Hetherington, 2021; Unamba et al., 2015). Armed with these tools, we are now in a position to begin deciphering how, when and from what the seed originated.

This review will take you on a journey through evolving reproductive strategies, from the emergence of land plants heavily dominated by the haploid reproductive phase, through to the success of the sporophyte-dominant seed plants. Focus will then shift to a pivotal gene - *SPOROCTELESS* - and its role in kick-starting the seed development program. Finally, the fern *Ceratopteris richardii* will be introduced as an emerging model organism, offering valuable insights into the evolutionary transition from seedless to seed plants.

1.1 Journey to the Seed

1.1.1 Embryophytes: a reproductive revolution

According to molecular clock time tree predictions, the earliest multicellular green organisms appeared on land between 515 and 469 million years ago (Ma) (Morris et al., 2018; Nie et al., 2019; Strother & Foster, 2021). The first unequivocal fossil evidence, consisting of spores and spore masses, aligns with the later end of this range, around 470 Ma, during the Mid-Ordovician period (Rubinstein et al., 2010). However, we cannot dismiss the possibility of an earlier emergence based on fossil evidence alone, as the plant fossil record is notoriously fragmented (Kenrick et al., 2012; Smith & McGowan, 2007). What can be said with more certainty is that the modern-day descendants of the land plants comprise a monophyletic group - the Embryophytes - that evolved from within freshwater streptophyte algae (Lewis & McCourt, 2004; Morris et al., 2018; Puttick et al., 2018) (Figure 1.1). Current phylogenomic analyses of both organellar and nuclear sequences resolve Zygnematophyceae, a predominantly unicellular class within the Charophyta, as the closest living sister lineage to the embryophytes (De Vries & Archibald, 2018; Donoghue & Paps, 2020; Harris et al., 2020; Karol et al., 2010; Lemieux et al., 2016; Puttick et al., 2018; Turmel et al., 2006, 2013).

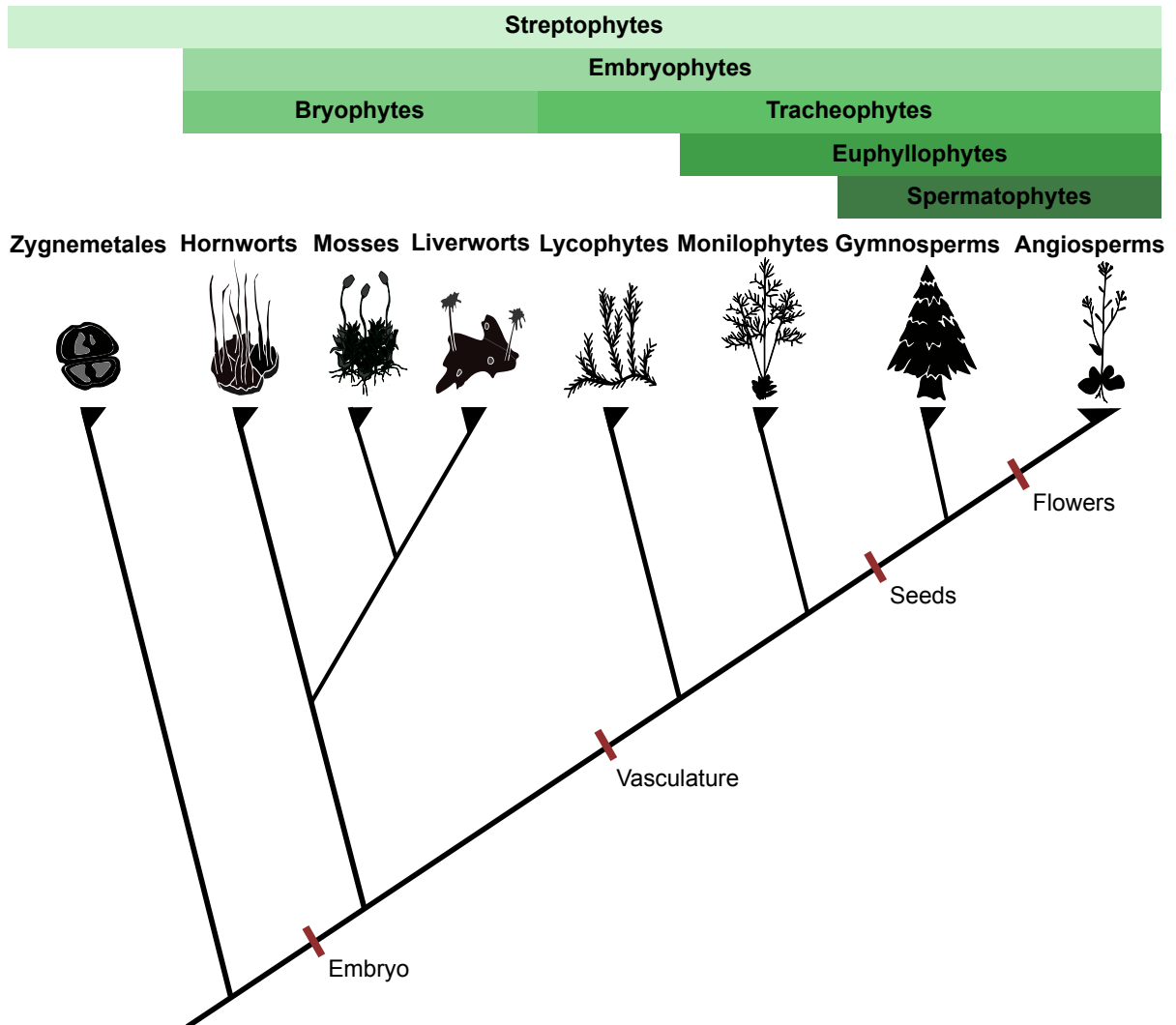


Figure 1.1: The evolution of the land plants. Simplified phylogenetic tree based on Puttick et al. (2018). Land plants (embryophytes) evolved from a green algal ancestor within the Zygnematales. Bryophytes, the earliest diverging embryophytes, are characterized by a gametophyte-dominant life cycle. The evolution of vascular tissue marked the emergence of tracheophytes, encompassing lycophytes, monilophytes, and seed plants. Lycophytes and monilophytes, as early diverging tracheophytes, transitioned to a sporophyte-dominant life cycle but retained free-living gametophytes. The origin of seeds occurred before the divergence of gymnosperms and angiosperms (spermatophytes), with the evolution of flowers defining the angiosperms.

The transition to land marked a fundamental change to reproductive strategy that distinguishes the embryophytes from their algal ancestors (Graham, 1993; Kenrick, 1994). Charophycean algae have a haplontic life history: there is no multicellular diploid phase (Figure 1.2A). Mitosis occurs in the haploid gametophyte to produce gametes, these unite to form the diploid zygote which immediately undergoes meiosis to begin the cycle again (Delwiche & Cooper, 2015; Lewis & McCourt, 2004). In contrast, the embryophyte life cycle consists of two distinct multicellular generations: a haploid gametophyte and a diploid sporophyte - termed a haplodiplontic lifestyle, or the alternation of generations (Figure 1.2B) (Hofmeister, 1851; Roe, 1975; Strasburger, 1894). Meiosis in the sporophyte produces sporopollenin-impregnated haploid spores (Edwards & Salter, 1982). These meiospores develop into gametophytes that form both male and female reproductive organs, the antheridia and archegonia in non-seed plants, and the micro- and mega-gametophyte in seed plants. These organs exhibit three-dimensional growth characteristic of land plants (Moody, 2019; Niklas, 2000). Mitosis within the reproductive organs then gives rise to male and female gametes. After fertilisation, a delay in zygotic meiosis in favour of mitotic divisions results in the multicellular diploid sporophyte, which is initially, or in some cases wholly (e.g. in mosses) nurtured by the gametophyte - a condition termed matrotrophy. The occurrence of a matrotrophic, multicellular embryo is a defining characteristic of the embryophytes, from which their name is derived (Graham & Wilcox, 2000; Kenrick & Crane, 1997; Niklas & Kutschera, 2010). These reproductive innovations collectively represent crucial adaptations to life on land, allowing embryophytes to thrive in terrestrial habitats where protection and support of the developing sporophyte are essential.

During land plant evolutionary history, the sporophyte has come to dominate the land plant life cycle, with the gametophyte reduced to just a few cells in the most recently evolved lineages (Ewing, 1977; Mable & Otto, 1998). It is widely accepted that delayed zygotic meiosis drove the evolution of progressively more complex

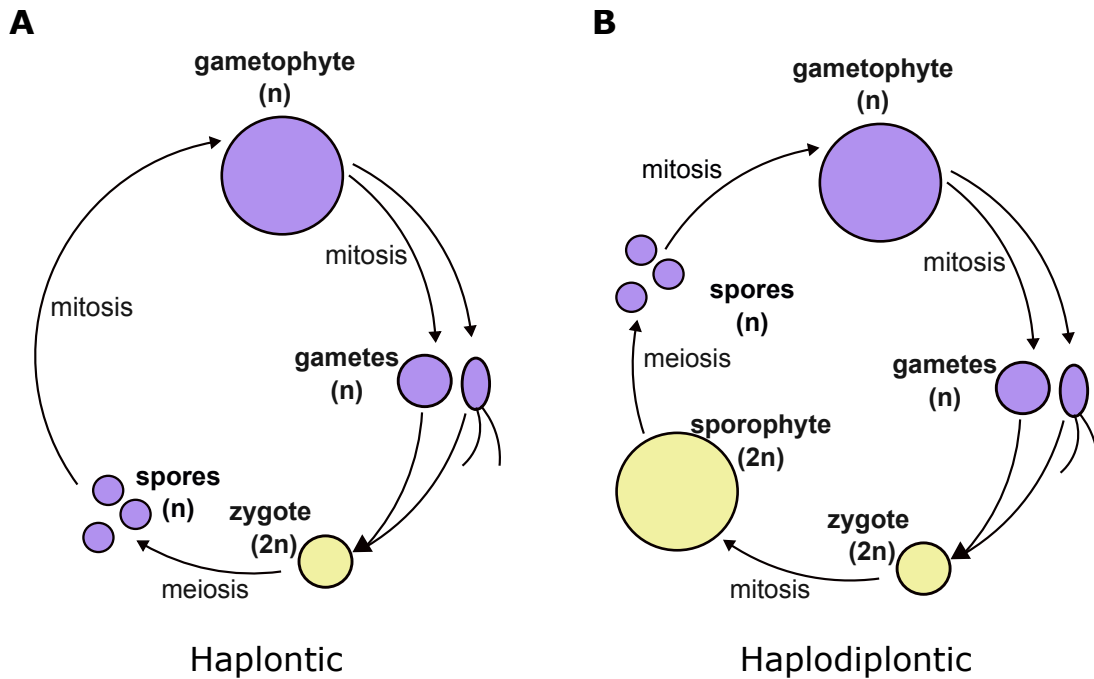


Figure 1.2: Comparative life cycles of haplontic and haplodiplontic plants.

The life cycles of plants alternate between haploid (n ; lemon yellow) and diploid ($2n$; lilac) generations. (A) The haplontic life cycle, exemplified by Charophycean algae, features a dominant haploid phase. The zygote is the only diploid phase, immediately undergoing meiosis to produce haploid spores, which germinate into gametophytes. (B) The haplodiplontic life cycle, characteristic of the land plants, alternates between multicellular haploid and diploid generations. Haploid spores germinate into gametophytes, which produce gametes through mitosis. Fusion of gametes results in a diploid zygote that grows into a multicellular sporophyte, completing the cycle.

sporophytes, which in turn allowed the production of more, genetically diverse haploid spores (Bower, 1908; Ligrone et al., 2012; Pires & Dolan, 2012). However, it has been suggested that the pivotal event in embryophyte evolution may predate the multicellular sporophyte: the heterochronic shift in timing of sporopollenin deposition from the diploid zygote to the haploid meiospore (Brown & Lemmon, 2011; Hemsley, 1994). This change in dispersal unit provided early land plants with enhanced protection from desiccation, allowing spores to survive in harsh terrestrial environments. This adaptation improved dispersal success, supported a shift in reproductive strategy better suited to land, and eventually promoted the development of a dominant sporophyte generation in the life cycle. Subsequent innovations that further enhanced the adaptation to the terrestrial environment include the evolution of stomata (reviewed in Raven (2002) and Beerling and Franks (2009)), specialised water and nutrient transport systems (vasculature) (reviewed in Lucas et al. (2013)), a sporophytic apical meristem that facilitates branching (discussed in Fouracre and Harrison (2022)), organs such as leaves (reviewed in Chitwood and Sinha (2016)) and roots (Kenrick & Strullu-Derrien, 2014) and, of course, seeds (Linkies et al., 2010) and flowers (Sauquet et al., 2017).

1.1.2 The gametophyte dominated early embryophyte life cycles

Bryophytes, comprising mosses, liverworts and hornworts, represent the earliest diverging group in the land plants and are considered living descendants of some of the very first non-vascular plants (Chang & Graham, 2011; Qiu et al., 2006; Wickett et al., 2014) (Figure 1.1). Bryophyte ancestors were among the earliest colonisers of terrestrial environments, and extant bryophytes continue to play a crucial role in ecological succession, often being the first to establish in disturbed or bare habitats, paving the way for more complex plant communities (Browning et al., 2010; Cong et al., 2023; Proctor, 2014; Takashima-Oleivera et al., 2020).

Whilst the role of early land plants in shaping terrestrial landscapes is not in

question, the timing and nature of the the move to land has been subject to much scrutiny. There has long been contention over the phylogenetic relationship of the bryophytes to the vascular plants (tracheophytes) with almost every possibility having been proposed at some stage (Donoghue et al., 2021). However, wider taxon sampling, facilitated by plant sequencing projects such as the 1000 plants (1KP) initiative (One Thousand Plant Transcriptomes Initiative, 2019), and improved phylogenetic methods now strongly support bryophyte monophyly, with the liverworts and mosses forming a clade (the setaphytes) within the bryophytes (de Sousa et al., 2019; Harris et al., 2020; Puttick et al., 2018; Wickett et al., 2014).

Wide acceptance of bryophyte monophyly fundamentally changes our understanding of how plants colonised land (Donoghue et al., 2021; Puttick et al., 2018). Until recently, bryophytes were considered paraphyletic to the vascular plants, with conflicting hypotheses focused on which of the liverworts (Clarke et al., 2011; Gao et al., 2010; Karol et al., 2001; Karol et al., 2010; Qiu et al., 2006), hornworts (Nishiyama & Kato, 1999; Wickett et al., 2014) or setaphytes (Karol et al., 2010) were sister to the tracheophytes. In this light, conclusions about the ancestral state of land plants were drawn based on individual bryophyte lineages, whereby sequential divergences produced bryophyte intermediates with increasing morphological complexity. Instead, bryophyte monophyly implies that the hornworts, liverworts and mosses are derived from a single ancestral land plant, which is also the last common ancestor to the tracheophytes (Donoghue et al., 2021). Shared characteristics between the three lineages likely represent ancestral traits, whilst differences can be attributed to gains or losses in specific lineages over the course of 400 million years of evolution (McDaniel, 2021). Extant moss (*Physcomitrium patens*) (Prigge & Bezanilla, 2010; Rensing et al., 2008), liverwort (*Marchantia polymorpha*) (Bowman et al., 2017; Kohchi et al., 2021) and hornwort (*Anthoceros agrestis*) (Li et al., 2020; Szövényi et al., 2015) models therefore represent a chimera of shared ancestral and unshared derived traits that together can help to uncover what the ancestor to the bryophytes and

tracheophytes may have looked like (Donoghue et al., 2021).

Bryophytes share a life history in which the haploid gametophyte is dominant and conspicuous, whilst the sporophyte is small, monosporangiate and completely dependent on the maternal gametophyte (Niklas & Kutschera, 2010; Shaw et al., 2011). Their life cycles are broadly similar, and parallels can be drawn at key stages, represented here by the life cycle of the canonical bryophyte model - *P. patens* (Figure 1.3). Briefly, the gametophyte germinates from a haploid sporopollenin-impregnated spore, and undergoes a period of vegetative growth before development of the archegonia and antheridia (Raven et al., 2012) (Figure 1.3A-E). The egg is housed at the base of the archegonia, in the flask-shaped venter, with an elongated neck canal whose central cells degenerate upon maturity to allow passage of the sperm (Hiss et al., 2017; Landberg et al., 2013; Sanchez-Vera et al., 2017) (Figure 1.3F-G). Motile sperm cells from the antheridia require an external water source to navigate to and fertilise the egg in the archegonia, a system retained from their algal ancestor in the move to land (Hackenberg & Twell, 2019). Their dependence on water for sexual reproduction explains the small size of the bryophytes relative to the tracheophytes (Shaw & Renzaglia, 2004). After fertilisation, the diploid zygote is retained and nurtured by the haploid gametophyte as it undergoes embryogenesis and develops into the mature sporophyte (Raven et al., 2012) (Figure 1.3H-J).

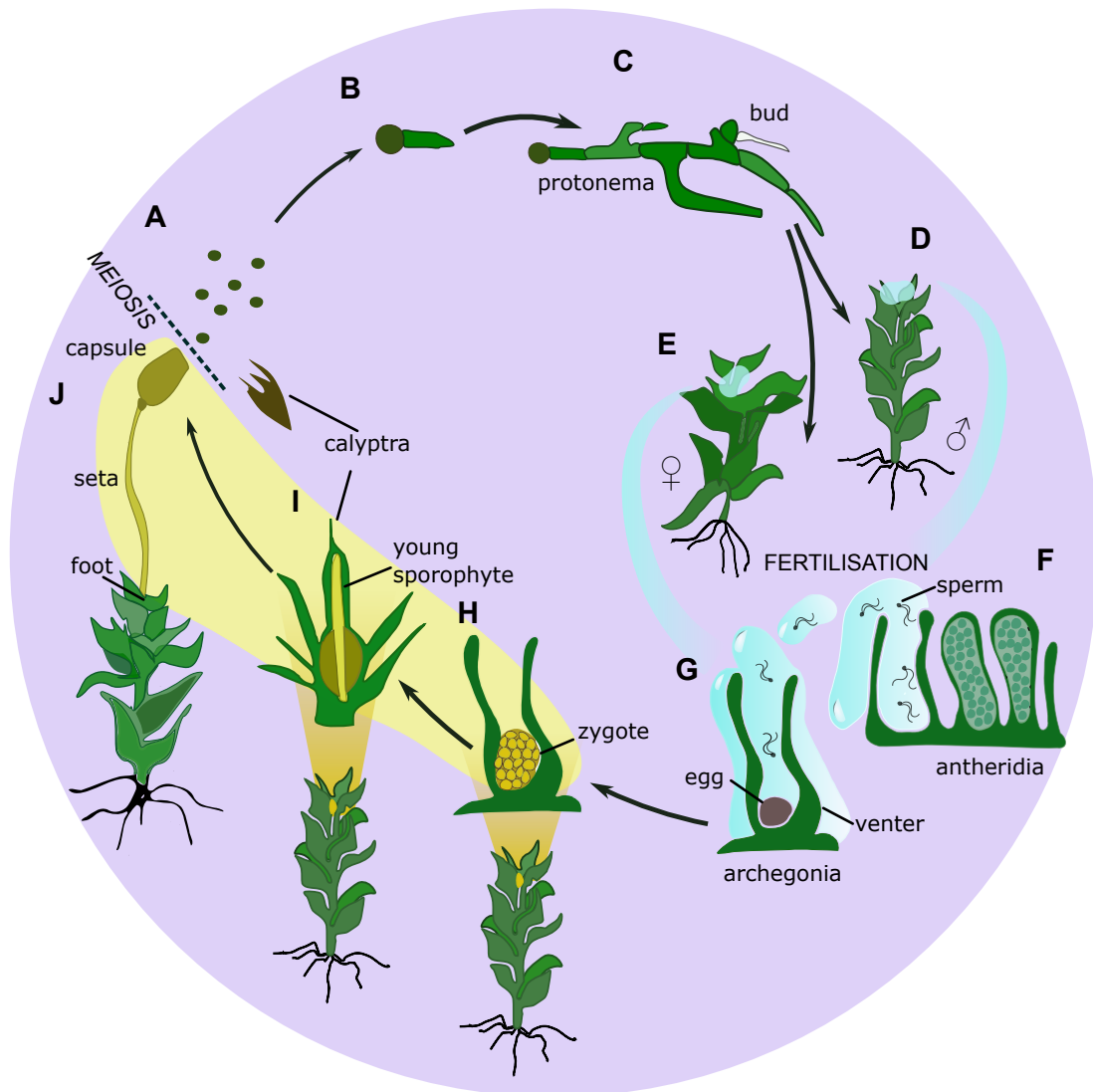


Figure 1.3: The life cycle of *Physcomitrium patens* is dominated by the gametophyte (lilac), with the sporophyte (yellow) entirely dependent on the gametophyte. (A) The gametophyte stage begins with the dispersal of sporopollenin-impregnated haploid spores. (B) Germinating spores grow as a 2D protonema, which transitions to 3D growth (C), initiating the development of male (D) and female (E) leafy gametophytes. Male antheridia (F) and female archegonia (G) develop on their respective gametophytes. (F) The antheridium, encased in sterile cells, produces biflagellated sperm requiring water for fertilization. (G) The flask-shaped archegonium houses a single egg in the venter, surrounded by a sterile jacket layer. Upon egg maturation, the neck canal disintegrates, enabling sperm to fertilize the egg. (H-I) The resulting diploid zygote, retained and nourished within the archegonium, divides mitotically to form an embryo, protected by the expanding calyptra. (J) The mature sporophyte consists of a foot anchoring it to the gametophyte, a seta elevating the sporangium, and a capsule where meiosis produces haploid spores.

The mature bryophyte sporophyte consists of a foot, seta, and sporangium (capsule), within which meiosis occurs to produce haploid spores (Figure 1.3J). Morphology-based studies of bryophyte evolution often interpret this structure as analogous to the dominant sporophyte of the tracheophytes (Kenrick & Crane, 1997; Mishler & Churchill, 1984, 1985). Despite its small and often transient nature, the bryophyte sporophyte plays a crucial role in enabling effective aerial dispersal of spores, thereby enhancing reproductive success. However, some argue that the diploid structure of bryophytes is essentially a footed sporangium with a stalk, whereas the tracheophyte sporophyte represents a novel vegetative organ (Kato & Akiyama, 2005). This highlights the need for caution when using morphology to infer evolutionary relationships, as traits observed in modern bryophytes and tracheophytes reflect over 400 million years of independent evolution and may represent more recent adaptations. Additionally, the fragility of bryophyte gametophytes limits their fossil record, further complicating our understanding of ancestral forms (Bowman, 2022).

As discussed above, the assumption of bryophyte paraphyly suggested that tracheophytes evolved from a gametophyte-dominant ancestor with a bryophyte-like sporophyte. In contrast, the now widely accepted view of bryophyte monophyly suggests that the common ancestor of bryophytes and tracheophytes could have been either gametophyte- or sporophyte-dominant, or isomorphic, with both generations of equal size (Cox et al., 2014). While there is no question that the bryophytes provide a window into the evolution of early land plant form, they represent only part of the story. To understand how seeds evolved, we must look further along the evolutionary tree.

1.1.3 Sporophyte dominance in the tracheophytes

Sister to the bryophytes, the tracheophytes represent the other major lineage of the embryophytes. The monophyly of this lineage is not in doubt, with the lycophytes (clubmosses, spikemosses and quillworts) resolving as sister to the euphyllophyte

clade, which comprises monilophytes (ferns and horsetails) and spermatophytes (seed plants) (Donoghue et al., 2021) (Figure 1.1). A defining characteristic of tracheophytes is their sporophyte-dominant life cycle, an innovation closely tied to the evolution of vascular tissue (Kenrick & Strullu-Derrien, 2014). This structural advancement enabled tracheophytes to grow taller and achieve greater complexity, enhancing their ability to exploit terrestrial environments (Niklas, 1997).

The shift to sporophyte dominance brought several advantages. Diploidy, a hallmark of the sporophyte generation, provided increased genetic diversity, reduced vulnerability to deleterious mutations, and enhanced adaptive potential (Bell, 1982; Bennici, 2008). The evolution of indeterminate proliferating organs, such as shoots and roots, driven by apical meristematic activity, allowed sporophytes to grow continuously and develop multiple sporangia (Harrison, 2017). This capability greatly enhanced access to water and minerals from the soil, as well as light and CO₂ from the atmosphere. It also enabled sporophytes to increase in size and produce more dispersal structures, in turn facilitating the production of greater quantities of spores (Graham, 1993). The emergence of traits such as branching, lignified cells, the cuticle, and specialized tissues further reinforced the dominance of the sporophyte, establishing it as the primary morphological and physiological phase in the life cycle (Bennici, 2008).

An important evolutionary milestone within the tracheophytes was the independent emergence of heterospory - the production of two distinct types of spores: larger female megaspores and smaller male microspores (Bateman & DiMichele, 1994). This shift was accompanied by the protection of developing gametophytes within the confines of the spore wall, a process known as endospory (Raven, 2002). Furthermore, gametophytes were obligately unisexual, with males arising from microspores and females from megaspores (Bateman & DiMichele, 1994). This transition from homospory - characterized by small, morphologically identical spores and bisexual gametophytes - was a crucial step toward the

development of seeds (Petersen & Burd, 2018). Heterospory, which arose independently as many as 11 times in the vascular plants, laid the groundwork for the seed habit by promoting reproductive specialisation and improving efficiency (Bateman & DiMichele, 1994). However, the selective advantages of heterospory when it first evolved, long before the appearance of seeds, remain a subject of debate (Petersen & Burd, 2017).

The earliest tracheophytes, such as *Cooksonia* and *Rhynia*, appeared in the Silurian period, approximately 425 Ma (Kenrick & Crane, 1997). These early vascular plants featured simple sporophyte structures characterised by dichotomous branching, rudimentary leaves (lycophylls) and terminal sporangia, representing a major evolutionary departure from their gametophyte-dominant bryophyte-like ancestors (Gerrienne et al., 2006; Kenrick, 1994; Lang, 1937). Today, the lycophytes, which comprise less than 1% of extant vascular plants, are the closest living relatives of these early tracheophytes (Doyle, 1998; Kenrick, 1994; Schuettpelz et al., 2016). Among them, *Selaginella moellendorffii* stands out as a valuable model organism for studying the evolution of vascular plants, providing insights into the biology and developmental pathways of these ancient lineages (Ferrari et al., 2020).

By the Early to Mid-Devonian, approximately 415 Ma, the lycophytes diverged from the euphyllophytes (Doyle, 1998; Kenrick, 1994; Nickrent et al., 2000; Pryer et al., 2001; Raubeson & Jansen, 1992) (Figure 1.1). Euphyllophytes are distinguished by the possession of megaphylls - true leaves of larger size with complex vasculature and distinct leaf gaps in the vascular stele (Kenrick, 1994; Stein, 1993). However, examples of megaphyll-like leaves in the lycophytes and microphyll-like leaves in the euphyllophytes blur the distinction between these two groups (Tomescu, 2009). Despite this overlap, the emergence of euphyllophytes represented a significant evolutionary step toward the development of seeds.

At the base of the euphyllophytes sit the monilophytes, which diverged from seed plants approximately 400 million years ago (Pryer et al., 2001) (Figure 1.1). As the

closest living relatives of seed plants, monilophytes are a key lineage for understanding the evolutionary innovations that laid the foundation for seed development. This ancient and remarkably diverse lineage includes tree-like ferns, aquatics, and epiphytes, showcasing their adaptation to diverse ecological niches over hundreds of millions of years (Schuettpeitz & Pryer, 2008; Watkins Jr & Cardelús, 2012). Among monilophytes, the leptosporangiate ferns dominate, accounting for approximately 80% of all non-flowering vascular plant species (Schuettpeitz & Pryer, 2009).

Three fern species have emerged as prominent models, with all three having sequenced genomes (Li et al., 2018; Marchant et al., 2022; Qiu & Yu, 2003). *Azolla filliculoides* and *Salvinia cucullata*, heterosporous water ferns from the Salviniales, represent an order comprising less than 1% of all fern species (Li et al., 2018). In contrast, *Ceratopteris richardii*, a homosporous semi-aquatic fern from the leptosporangiates, is better suited as a model for studying monilophyte biology (Kinosian & Wolf, 2022). Despite its homospory, *C. richardii* exemplifies the ancestral reproductive condition of vascular plants, making it invaluable for investigating evolutionary questions (Hickok et al., 1995). By examining the life cycle of *C. richardii*, key features of fern reproduction and their evolutionary significance can be explored.

Like all land plants, *C. richardii* exhibits an alternation of generations, with distinct haploid gametophyte and diploid sporophyte phases (Hickok et al., 1995) (Figure 1.4). Compared to the bryophytes, the gametophyte phase is much reduced, and completely independent of the sporophyte. The life cycle begins with the germination of the haploid spore to give a single-cell-layered thallus, initiating the gametophyte generation (Banks, 1999) (Figure 1.4A-B). Although *C. richardii* is a homosporous fern, its gametophytes differentiate into distinct male or hermaphroditic sexes through hormonal sex determination (Banks, 1999). Male gametophytes develop antheridia, while hermaphrodites produce both antheridia and archegonia (Figure 1.4C,D). Antheridia release motile sperm, and each archegonium houses a single egg. In the

presence of free water, fertilisation of one egg per gametophyte marks the beginning of the diploid sporophyte generation.

Following fertilisation, the sporophyte is initially dependent upon the gametophyte for nutrition (Banks, 1999) (Figure 1.4E). However this dependence is transient; as the sporophyte develops its first few fronds (Figure 1.4F), the gametophyte senesces and dies. This represents a significant departure from bryophytes, where the sporophyte remains permanently dependent on the gametophyte. The independence of the sporophyte in ferns highlights the transitional evolutionary position of *C. richardii* between gametophyte-dominant bryophytes and sporophyte-dominant seed plants, where the gametophyte is entirely dependent on the sporophyte (Reviewed in Qiu et al. (2012)). The *C. richardii* sporophyte develops as an indeterminate, continuously growing structure, producing a heteroblastic series of fronds that gradually increase in complexity, transitioning from simple juvenile forms to finely dissected reproductive fronds known as sporophylls (Johnson & Renzaglia, 2008) (Figure 1.4G-J). Meiosis occurs within sporangia located on the abaxial surface of these sporophylls, completing the cycle with the production of haploid spores (Figure 1.4J).

This alternation of generations, with independent gametophyte and sporophyte phases, provides a window into the reproductive strategies of early vascular plants. While the sporophyte is dominant in size and longevity, the free-living gametophyte remains critical for sexual reproduction. This duality highlights the evolutionary significance of *C. richardii* as a model for studying ancestral reproductive traits and their transitions toward the specialized strategies seen in seed plants. The homosporous condition, with its production of a single spore type and independent gametophyte and sporophyte generations, provides a window into ancestral traits that likely preceded the evolution of heterospory. Understanding these processes in *C. richardii* sheds light on how early vascular plants transitioned to the reproductive complexity seen in seed plants.

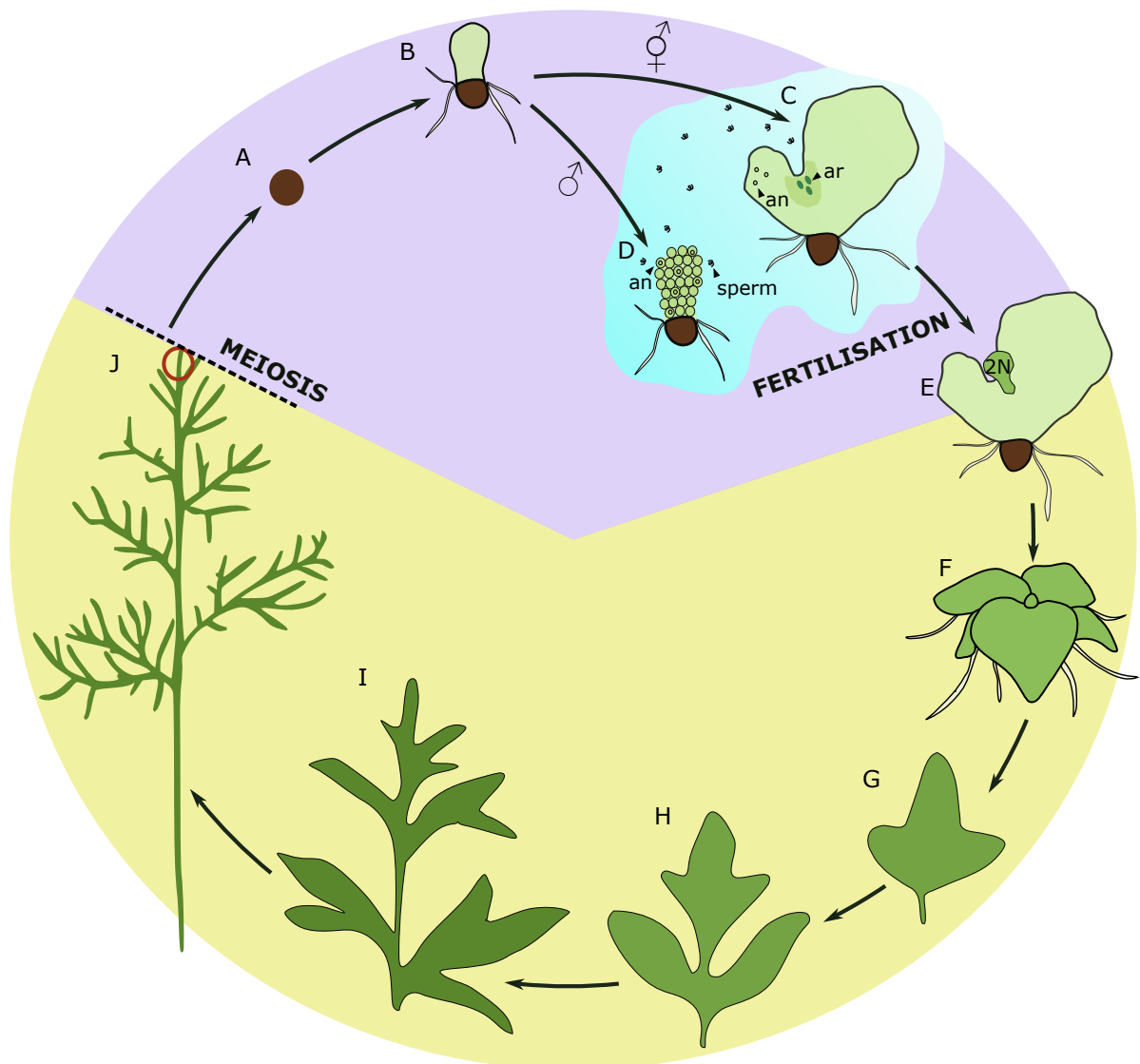


Figure 1.4: The life cycle of *Ceratopteris richardii*. The life cycle alternates between distinct independent haploid gametophyte (lilac) and diploid sporophyte (yellow) phases. The haploid, single celled spore (A) germinates, giving rise to a gametophyte with undetermined sex (B). The single-cell-layered photosynthetic thallus develops into one of two sexes: the default hermaphrodite with antheridia (an) and archegonia (ar) (each containing a single egg cell) (C) and the hormone induced male which forms antheridia (an) only (D). In the presence of water, motile sperm from the antheridia swim toward the archegonia. Fertilisation results in the development of a diploid zygote (2n) on the hermaphroditic thallus (E), marking the beginning of the sporophyte phase of the life cycle. The young sporophyte produces simple leaf-like fronds and roots from its apex (F). As the sporophyte matures, the fronds progressively increase in complexity, transitioning from lobed forms to finely dissected structures (G-I). The sporophyte becomes reproductive upon the development of specialised fronds known as sporophylls (J). Sporophylls bear sporangia on their abaxial surfaces, where meiosis occurs to produce haploid spores, completing the cycle.

1.1.4 The rise of seed plants

Seed plants (*Spermatophyta*) first emerged during the Late Devonian period, approximately 360 Ma, marking a transformative shift in reproductive strategies away from the evolutionary constraints of spore-based reproduction seen in the bryophytes, lycophytes and monilophytes (Bateman & DiMichele, 1994; Rothwell, 1986). This innovation addressed the inherent limitations of spore dispersal, such as the vulnerability of free-living gametophytes and their reliance on water for fertilisation (Bateman & DiMichele, 1994). Central to this transition were several key innovations: the retention of the megaspore within the parent sporophyte; the reduction to a single functional megaspore; and the development of integuments that surrounded the ovule (the pre-fertilised seed), ultimately forming the seed coat (Raven, 2002). These adaptations allowed seed plants to protect and nourish their embryos, significantly enhancing survival rates (Linkies et al., 2010). Additionally, the evolution of pollen eliminated the dependence on water for fertilization, while seed dormancy enabled delayed germination and long-distance dispersal (Bateman & DiMichele, 1994; Linkies et al., 2010; Prestianni & Gerrienne, 2010). Together, these innovations established the foundation for the ecological success and evolutionary diversification of seed plants.

The seeds plants comprise two major groups: the Acrogymnospermae (gymnosperms) and the Angiospermae (angiosperms). In both groups, seeds develop from ovules (Finch-Savage & Leubner-Metzger, 2006; Frohlich & Chase, 2007). An ovule consists of a stalk supporting the nucellus - a diploid tissue equivalent to the megasporangium - enveloped by one (gymnosperms) or two (angiosperms) layers of diploid maternal tissue, known as integuments. Within the nucellus, a megaspore develops into a haploid megagametophyte, which is functionally analogous to the female gametophytes of the non-seed plants but highly reduced in size (Linkies et al., 2010). For example, the mature angiosperm megagametophyte typically consists of only 7 cells with 8 nuclei (Baroux et al., 2002; Berger et al., 2008; Floyd & Friedman,

2000; Friedman & Ryerson, 2009; Friedman & Williams, 2004). After pollination, non-motile sperm reach and fertilise the egg cell through a pollen tube, leading to the development of the diploid embryo (Linkies et al., 2010). This siphonogamic mechanism of sperm transfer is characteristic of all extant seed plants except for two gymnosperm lineages, which produce multiflagellated, swimming sperm. As the resulting embryo matures, the integuments form the seed coat, enclosing the embryo and its nutrient reserves, ensuring protection, dormancy and efficient dispersal (Bergfeld & Schopfer, 1986; Debeaujon et al., 2000; Windsor et al., 2000).

The gymnosperms, with approximately 800 extant species, and angiosperms, with around 250,000 species, are widely regarded as monophyletic lineages, supported by extensive molecular and morphological analyses (Cantino et al., 2007; Doyle, 2006; Frohlich & Chase, 2007; Pennisi, 2009; Soltis et al., 2008). The common ancestor of gymnosperms and angiosperms is thought to lie within the extinct paraphyletic seed ferns (or the progymnosperms), a diverse group of early seed-bearing plants (Doyle, 1998; Frohlich & Chase, 2007; Soltis et al., 2008). These early fossil seeds featured a single functional megaspore enclosed within a nucellus, which was partially surrounded by a lobate integument to facilitate pollination (Prestianni & Gerrienne, 2010; Rothwell, 1986). These primitive seeds lacked the fully enclosed, multilayered integuments characteristic of modern seed plants, reflecting an intermediate stage in seed evolution (Linkies et al., 2010). However, while fossil evidence provides valuable insights into these ancestral forms, the lack of molecular data has made it challenging to infer how these early structures were regulated and how they transitioned into extant seed forms (Harrison, 2017; Meade et al., 2021).

Gymnosperms, comprising conifers, cycads, ginkgos, and Gnetales, are defined by their “naked” seeds, which are exposed on the surface of cones or modified sporophylls (Raven, 2002). These plants dominated terrestrial ecosystems for almost 200 million years between the Permian and the Late Cretaceous (260 - 70 Ma) (Pennisi, 2009; Pires & Dolan, 2012; Willis & McElwain, 2014). Unlike angiosperms,

gymnosperm seeds develop from ovules with a single integument, and their megagametophytes retain ancestral features, including multicellular structures with multiple archegonia (Baroux et al., 2002). Despite their simpler reproductive strategies, gymnosperms remain ecologically significant and provide critical context for understanding the innovations that define angiosperm reproductive success (De La Torre et al., 2019).

1.1.5 Angiosperms: masters of reproductive innovation

Angiosperms, or flowering plants, represent the most diverse lineage of seed plants and have become dominant across terrestrial ecosystems (Christenhusz & Byng, 2016). Fossil and molecular evidence suggests that basal angiosperms, magnolids, early monocots, and early eudicots emerged in rapid succession during the Early Cretaceous (100 - 145 Ma), marking the beginning of their evolutionary rise (Friis et al., 2006; Taylor et al., 2009). This initial diversification was followed by widespread radiation, enabling angiosperms to dominate the majority of terrestrial habitats from the Late Cretaceous (100 - 65 Ma) to the present day (Willis & McElwain, 2014). Central to this success was the evolution of flowers, a major functional innovation in angiosperms, which are thought to have arisen through the transformation of gymnosperms' separate male and female structures into an integrated hermaphrodite system (Pires & Dolan, 2012). These reproductive innovations underpinned the unparalleled ecological success and extraordinary diversification of angiosperms, a pivotal moment in plant evolution (Crane et al., 1995; Doyle, 1977; Stebbins, 1974; Wing & Boucher, 1998).

The angiosperm life cycle alternates between two distinct phases: a diploid sporophyte phase which dominates much of the life cycle and a haploid gametophyte phase comprising just a few cells, in which gametogenesis occurs (Hofmeister, 1851, 1862; Strasburger, 1894). A well-characterised example of this life cycle is provided by the canonical flowering plant model organism *Arabidopsis thaliana* (Figure 1.5).

Flowers, which house both male and female reproductive organs, play a central role in pollen transfer, often mediated by specialized relationships with animal pollinators (Crepet & Friis, 1987). Within the sporophyte tissue of the floral organs, the male gametophyte develops within the anthers, and the female gametophyte develops in the ovule enclosed within the carpel (Hong, 2005; Yang et al., 2010) (Figure 1.5A-C). Double fertilization, a process unique to angiosperms, gives rise to both the diploid zygote and the triploid endosperm, which nourishes the developing embryo (Baroux et al., 2002). The developing embryo is protected within the carpel, distinguishing angiosperms from gymnosperms with exposed ovules (Linkies et al., 2010) (Figure 1.5D). In addition, fruits derived from mature carpels enhance seed dispersal through diverse mechanisms involving wind, water, or animal vectors (Raven, 2002). This efficient resource allocation, combined with reduced gametophyte sizes, rapid life cycles, and specialised reproductive strategies, has enabled angiosperms to outcompete gymnosperms, ferns and bryophytes across a wide range of ecosystems (Raven, 2002).

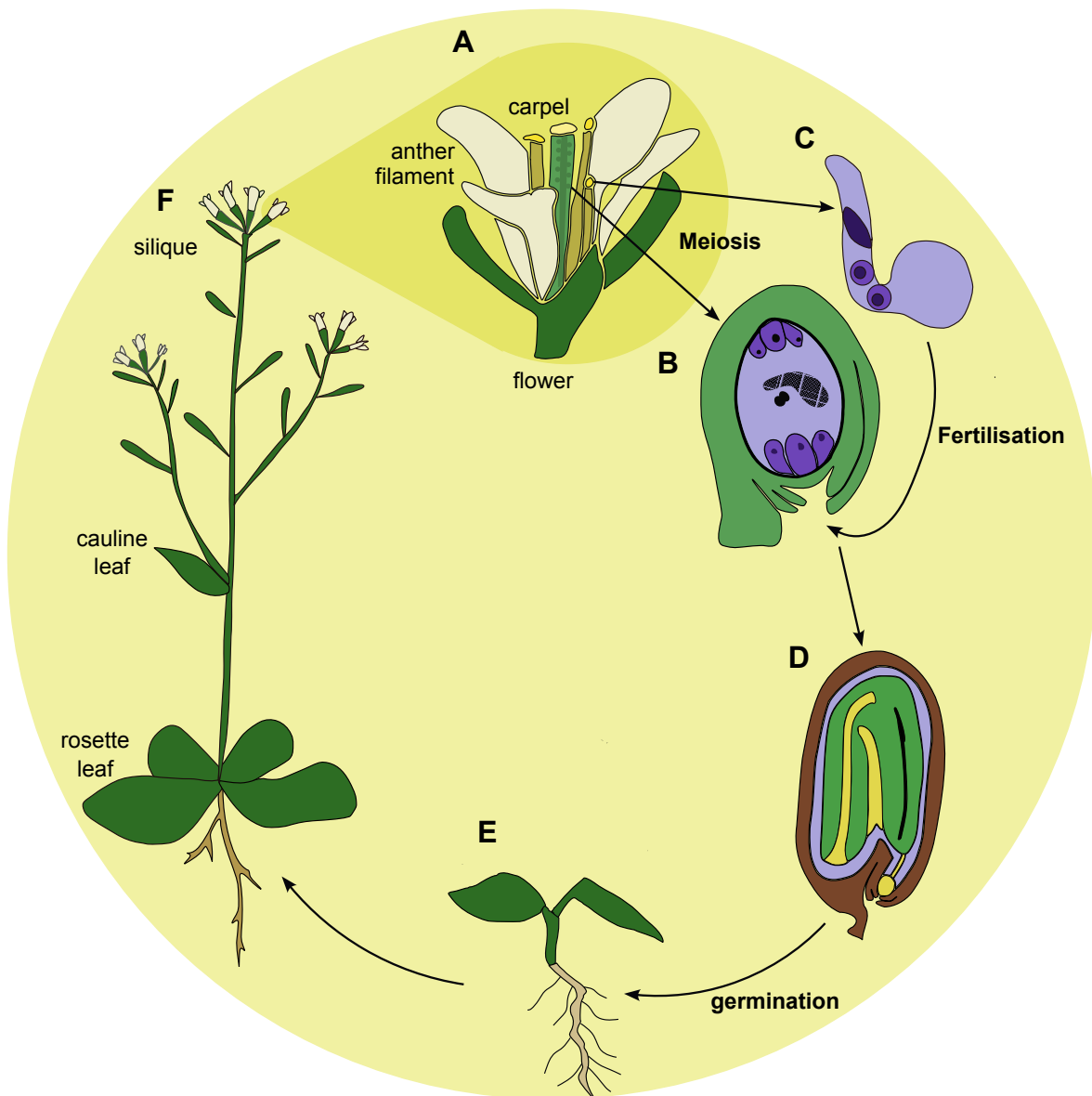


Figure 1.5: The life cycle of *Arabidopsis thaliana*. The *Arabidopsis* life cycle is dominated by the sporophyte (yellow), with the gametophyte reduced to a few cells (lilac). (A) The flower: petals and sepals surround the reproductive organs. The central carpel contains ovules, the site of female gametogenesis, while anthers house pollen sacs, where male gametogenesis occurs. Meiosis in these organs produces haploid spores that develop into pollen and ovules. (B) A mature ovule, showing the eight-celled female gametophyte (lilac) surrounded by integuments, which later form the seed coat. (C) A pollen grain with two sperm cells and a vegetative nucleus (lilac). After pollination, the pollen tube delivers the sperm cells to the ovule, where double fertilization forms a diploid zygote and triploid endosperm. (D) A developing seed, with gametophyte tissue (lilac) visible in the embryo sac, enclosed by the seed coat. (E) A seedling emerging from the seed, representing the early sporophyte stage. (F) The mature *Arabidopsis* plant, featuring a basal rosette of leaves, cauline leaves, flowers, and siliques containing developing seeds.

1.1.6 Ovule development in Angiosperms: a key reproductive innovation

The evolution of specialised structures to enclose, protect, and preserve the dormancy of developing embryos is a hallmark innovation of the seed plants (Linkies et al., 2010). At the heart of this innovation lies the ovule, a complex structure that underpins seed development and represents a key adaptation driving the success of the angiosperms. The development of the ovule has been well characterised (see review: Battaglia et al. (2018)) and is subdivided into five distinct stages of ovule development (Modrusan et al., 1994) (Figure 1.6). During ovule stage 1 (stage 8 of flower development (Smyth et al., 1990)), specialised meristematic regions at the internal surface of the carpels (termed the placenta) give rise to ovule primordia (Robinsonbeers et al., 1992) (Figure 1.6A). These ovule primordia are described as finger-like protrusions and have a radial symmetry (Battaglia et al., 2018). Along the proximal-distal axis of the ovule primordia, three distinct regions become distinguishable in late ovule stage 1 (Stage 9-10 of flower development); the funiculus, chalaza and nucellus (Figure 1.6B). At the proximal end, the funiculus describes the region connecting the ovule primordia to the placenta, whilst the chalaza is the central region from which the inner and outer integuments emerge. The nucellus is located at the distal end and is divided into two layers; the epidermal layer (L1) and the subepidermal layer (L2). Generally, the most distal cell in the L2 layer, just below the chalaza apex, undergoes a cell fate change to become the megaspore mother cell (MMC) (Figure 1.6B). The differentiation of the MMC occurs shortly after ovule primordia emergence and marks the initiation of gametophyte development.

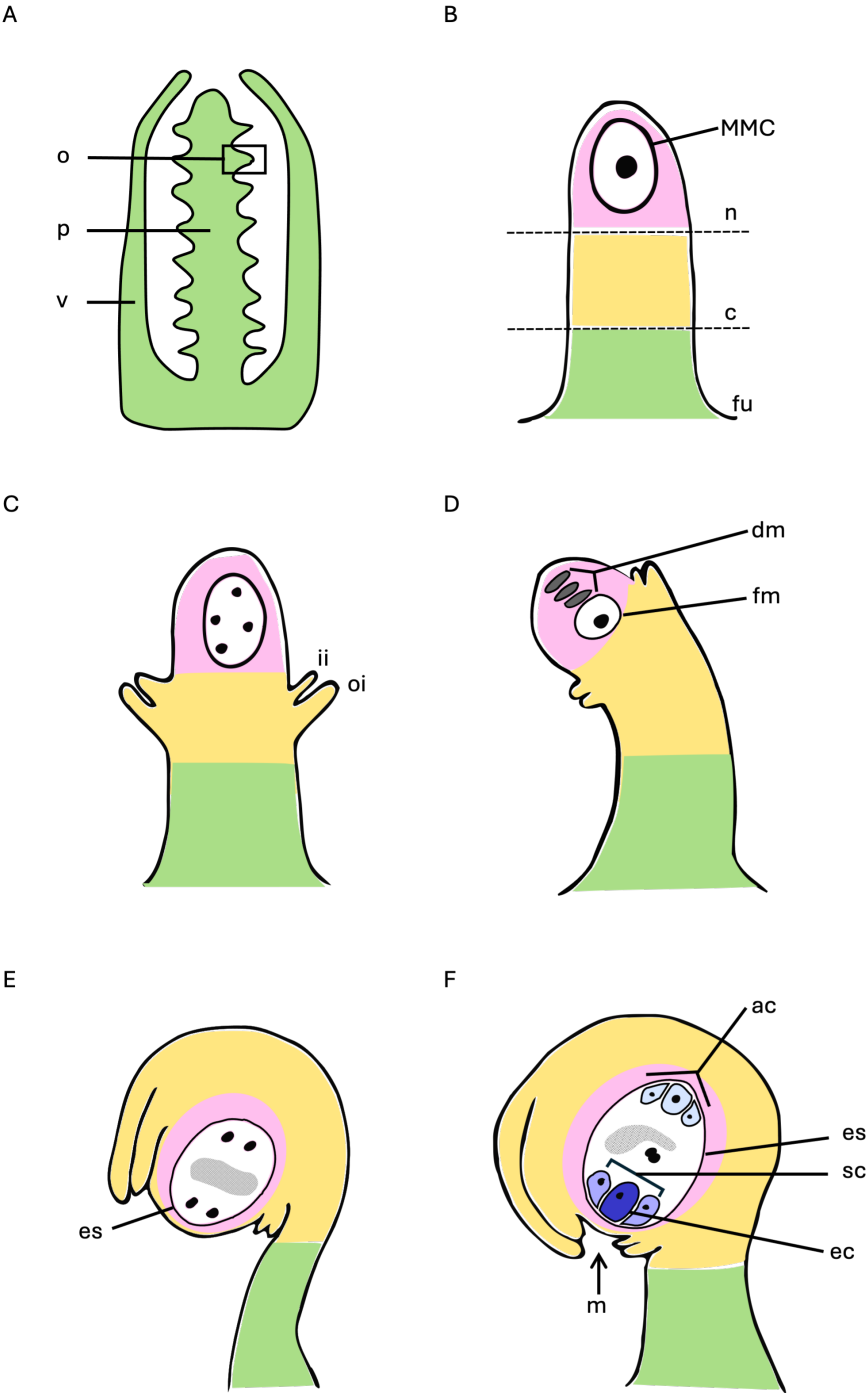


Figure 1.6: Caption overleaf.

Figure 1.6: Five stages of ovule development in *Arabidopsis thaliana*. (A – B) Stage 1. (A) Early Stage 1. Ovule primordia emerge as finger-like protrusions from the placenta on the internal surface of the carpel. (B) Late stage 1. A single ovule primordium is divided into three distinct regions along the proximal-distal axis; the funiculus is at the most proximal end (green), the chalaza is in the central region (yellow), and the nucellus forms the most distal region (pink). The megaspore mother cell (MMC) is visible in the nucellus. (C) Stage 2. Megasporogenesis and integument initiation occur. In the nucellus, the MMC divides into four haploid megaspores by meiosis. Meanwhile, the inner and outer integuments emerge from the chalaza. (D) Stage 3. The megaspore closest to the chalaza becomes the functional megaspore, whilst the three more distal megaspores degenerate. Increased cell division at the inner side of the outer integument leads to asymmetric growth. (E) Stage 4. The functional megaspore forms a syncytium with four nuclei, two at each pole, as a result of two mitoses without cytokinesis. The inner and outer integuments continue to grow and enclose the nucellus. (F) Stage 5: the mature ovule. The seven-celled, female gametophytic structure (embryo sac), consisting of an egg cell with two synergid cells at the micropylar pole, a large, 2-nucleate central cell and three antipodal cells at the chalazal end, is completely enveloped by the integuments. A small gap (the micropyle) remains, where the pollen will enter. *ac*, antipodal cells; *c*, chalaza; *dm*, degenerating megaspores; *ec*, egg cell; *es*, embryo sac; *fm*, functional megaspore; *fu*, funiculus; *ii*, inner integument; *m*, micropyle; *MMC*, megaspore mother cell; *n*, nucellus; *oi*, outer integument; *o*, ovule primordia; *p*, placenta; *sc*, synergid cells; *v*, valve.

During ovule stage 2, megasporogenesis and integument initiation occur (flower development stage 11) (Battaglia et al., 2018) (Figure 1.6C). In the chalaza, the inner and outer integuments emerge. Increased cell divisions at the inner side of the outer integument lead to asymmetric growth. Meanwhile in the nucellus, the diploid MMC undergoes meiosis to produce four haploid megaspores – a process termed megasporogenesis. Of the four megaspores, only one will become the functional megaspore (FM), the cell closest to the chalaza, whilst the other three undergo programmed cell death and degenerate at ovule stage 3 (flower development stage 12) (Figure 1.6D). The FM then proceeds through megagametogenesis as the inner and outer integuments continue to grow towards the distal end, to enclose the nucellus (Christensen et al., 1997; Schneitz et al., 1995). During megagametogenesis, the FM undergoes two mitoses without cytokinesis to form a syncytium with four nuclei, two at each pole (ovule stage 4) (Figure 1.6E). Following a third mitotic division resulting in 8 nuclei, one nucleus from each pole migrates towards the centre of the developing embryo sac and fuses to form the homodiploid secondary endosperm nucleus (ovule stage 5) (Figure 1.6F). At this stage, the embryo sac cellularises to form the mature female gametophytic structure. The seven-celled structure consists of an egg cell closely associated with two synergid cells at the micropylar pole, a large, diploid central cell which will form the endosperm after fertilization, and three antipodal cells at the chalazal end. In the mature ovule, the integuments completely envelope the embryo sac, with the outer integument overgrowing the inner integument to leave only a small gap where the pollen tube will later enter, termed the micropyle (Figure 1.6F).

The ovule represents one of the most significant evolutionary innovations in the history of land plants, serving as the developmental precursor to seeds and enabling the complex reproductive strategies that characterize gymnosperms and angiosperms (Baroux et al., 2002; Battaglia et al., 2018; Linkies et al., 2010). Through its intricate structure and tightly regulated development, the ovule ensures the

protection, nourishment, and successful propagation of the next generation, solidifying its role as a cornerstone of plant reproductive success. Central to ovule development are the genetic networks that regulate its formation and function. Among these, the transcription factor SPOROCTELESS (SPL) plays a pivotal role in specifying reproductive structures and ensuring the initiation of sporogenesis (Schiefthaler et al., [1999](#); Yang et al., [1999](#)). The next section explores the role of SPL in ovule development, highlighting its critical contributions to angiosperm reproduction and its evolutionary significance.

1.2 ***SPOROCYTELESS*: Master regulator of gametophyte development**

For more than two decades, *SPOROCYTELESS/NOZZLE* (*SPL/NZZ*) has been recognised as a key regulator of both male and female gametophyte initiation in the model flowering plant *Arabidopsis thaliana* (Schiefthaler et al., 1999; Yang et al., 1999). Mutant *spl/nzz* plants were first identified by two independent research groups; in a screen of *Ds* insertion lines in the Landsberg *erecta* ecotype of *Arabidopsis* and, coincidentally, in an *En-1/Spm*-transposon-induced mutant population of the *Arabidopsis* Colombia ecotype (Schiefthaler et al., 1999; Yang et al., 1999). Both groups identified that failure of pollen sac and nucellus formation in *spl/nzz* mutants resulted in male and female sterility. Thus, the *SPL/NZZ* gene (hereafter referred to as “*SPL*”) is necessary for proper pollen and egg development and is the earliest known gene regulating both male and female germline initiation. Homologs of *AtSPL* have since been identified in several other plant species, including stem mustard, (Yang et al., 2008), cucumber (Liu et al., 2018), rice (Ren et al., 2018) and tomato (Rojas-Gracia et al., 2017). These homologs share conserved functions in regulating germline specification (discussed further in section 4.1.1), highlighting the critical role of *SPL* across diverse plant lineages.

Mutant *Atspl* plants exhibit an overall wild-type morphology, with no obvious abnormalities in vegetative tissues (Schiefthaler et al., 1999; Yang et al., 1999). In contrast, elevated *AtSPL* expression, as observed in the T-DNA activation tagged mutant *spl-D*, leads to severe morphological defects in both vegetative and reproductive tissues (Li et al., 2008). The leaves of *spl-D* mutants were narrow, small and hyponastic (up-curved), with the severity of the phenotype increasing progressively with the production of each leaf. Furthermore, the shoot apical meristem (SAM) was aborted after the production of seven to eight rosette leaves in

homozygous *spl-D* plants. The SAM, destined for the reproductive transition in wild-type plants, was replaced by leaf-like structures in the homozygous *spl-D* mutants. Meanwhile, the SAM did not abort in heterozygous *spl-D* mutants, but fewer and smaller flowers, shorter and wrinkled siliques, and shrivelled seeds were produced. This reduction in severity of heterozygous *spl-D* mutants suggested these phenotypes were closely linked with *SPL* expression. This was confirmed by the generation of *35S::SPL* lines; lines with higher *SPL* expression displayed a concomitant increase in phenotypic severity (Li et al., 2008).

The contrasting phenotypes of *spl/nzz* loss-of-function and *spl-D* gain-of-function mutants suggest that *SPL* is a tightly regulated gene whose expression level is critical for maintaining normal plant development (Li et al., 2008; Schiefthaler et al., 1999; Yang et al., 1999). The fact that *spl/nzz* knockout mutants exhibit no vegetative phenotype while *spl-D* and *35S::SPL* lines show progressively more severe vegetative abnormalities highlights the importance of precise spatial and temporal regulation of *SPL* expression (Li et al., 2008; Schiefthaler et al., 1999; Yang et al., 1999). These observations indicate that *SPL* may act as a latent regulator in vegetative tissues, whose overexpression perturbs developmental programs, particularly in leaf morphology and SAM transitions.

1.2.1 The structure and functional domains of the *SPL* protein

Molecular cloning by Yang et al. (1999) showed that the *SPL* gene encodes a novel nuclear protein. Genomic sequences flanking the *Ds* element were used to probe an Arabidopsis flower cDNA library and a 1.3kb cDNA sequence encoding 314 amino acids representing the *SPL* sequence was obtained. Nuclear localisation of *SPL* was confirmed by the expression of an *SPL-GUS* fusion reporter gene in the nucleus. Sequence analysis revealed the presence of a helix region homologous to a region within the MADS domain of several MADS box transcription factors. It was therefore hypothesised that *SPL* may encode a MADS-related transcription factor (Yang et al.,

1999). However, detailed structural and functional analysis of SPL was not achieved for a further decade.

BLAST analysis against the Phytozome database revealed that *SPL* contains three conserved regions encoding functional motifs (Chen et al., 2014). These conserved domains include an N-terminal nuclear localisation signal (NLS), a short consensus sequence “RGMGVAKLE” (termed the SPL-motif) located just downstream, and a C-terminal consensus LxLxL sequence, designated as an ethylene-responsive element binding factor-associated amphiphilic repression (EAR) motif (Figure 1.7). The 239-residue sequence between the SPL-motif and the EAR domain is not well conserved. Four other SPL-domain, EAR-containing proteins - designated as SPEAR proteins - were identified in *A. thaliana*. Furthermore, SPEAR homologs were found throughout the embryophytes, highlighting the evolutionary conservation of this protein family (Chen et al., 2014).



Figure 1.7: *Arabidopsis thaliana* SPL protein structure. Nuclear localisation signal (NLS) at the 5' end. SPOROCTELESS (SPL) motif just downstream. Ethylene-responsive element binding factor-associated amphiphilic repression (EAR) domain at the 3' end. There are two CARG boxes; one in the 5' regulatory region, one in the 3'. Adapted from Wei et al. (2015).

The EAR domain, defined by consensus sequences such as LxLxL, DLNxP or DLNxxP, has been identified in a wide range of transcription factors with repressor activity, including members of the TOPLESS/TOPLESS-Related (TPL/TPRs) family (Kagale & Rozwadowski, 2011; Kagale et al., 2010). To investigate whether the *SPL* EAR motif confers repression activity, Chen et al. (2014) performed a UAS/GAL4-based transcriptional repression assay in *Nicotiana benthamiana* leaves (Szemenyei et al., 2008). Either *SPL* or *SPL* without the EAR domain (*SPL*Δ*EAR*) were fused to the 35S promoter-driven GAL4 DNA binding domain (BD) to give 35S::BD-*SPL* and 35S::BD-*SPL*Δ*EAR* (Chen et al., 2014). These constructs were then co-infiltrated

with the *35S-UAS::GUS* construct, which contains six copies of the GAL4 binding sequence *UAS* inserted upstream of the *35S* promoter. When tobacco leaves were co-infiltrated with *35S-UAS::GUS* and *35S::BD-SPL*, significant repression of GUS activity was observed, whereas *35S::SPL Δ EAR* could not repress GUS activity. This indicates that the EAR motif is necessary for the repressive function of *SPL*. These findings were confirmed by Wei et al. (2015) in a similar repression assay.

Having confirmed that the EAR domain conferred transcription repressor activity, Chen et al. (2014) then asked whether *SPL* was able to form dimers, as is the case in other EAR-containing repressors (such as *IAA*; Kim et al., 1997). In a yeast two-hybrid (Y2H) assay, the activating domain bound to *SPL* (AD-*SPL*) protein was able to interact with both *BD-SPL* and *BD-SPL Δ EAR*, but not the EAR motif alone (Chen et al., 2014). This was confirmed with an *in vitro* GST pull-down assay (Chen et al., 2014). This result suggests that *SPL* can form homodimers through its N-terminal region both in yeast and *in vitro*. Taken together, it was hypothesised that *SPL* likely belongs to a novel transcription repressor family, with repression activity conferred through the EAR domain, and dimerisation to form more complex repressors made possible through the *SPL* domain (Chen et al., 2014).

In addition to its conserved motifs, two CArG boxes - DNA binding regions recognised by the MADS-box transcription factors - were detected in the regulatory regions of *SPL*; one upstream from the coding sequence, in the 5' untranslated region (UTR) and one downstream of the coding sequence, in the 3'UTR (Ito et al., 2004; Mendes et al., 2020) (Figure 1.7). The presence of these CArG boxes raises the possibility that *SPL* expression may be regulated by MADS-box transcription factors, potentially linking its expression to developmental cues controlled by these regulatory proteins. This possibility will be discussed in later sections.

Since its initial discovery in 1999, understanding of *SPL* function in plant reproductive development has significantly advanced. While research over the past two decades has explored *SPL* function in both ovule and anther development, with

significant and interesting discoveries in each area, the remainder of this section will primarily examine SPL's role in the ovule. This focus is driven by the overarching question of this thesis, which concerns seed development, as the ovule serves as the direct precursor to the seed (Battaglia et al., 2018; Linkies et al., 2010).

1.2.2 *SPL* expression is highly localised in the developing ovule

Early research hypothesised that *SPL* is required after ovule initiation, to specify MMC differentiation. This is because in *spl* mutants, ~99% ovules do not develop an MMC (Schiefthaler et al., 1999; Yang et al., 1999). *In situ* hybridization experiments detected *SPL* transcript throughout the early stages of ovule development (Schiefthaler et al., 1999). Transcript was detected throughout the young ovule primordia (stage 1), and in the MMC and emerging inner and outer integuments at stage 2. The detection of expression in the nucellus, chalaza and funiculus of the ovule primordia suggested a wider role for *SPL* function in ovule development. More recently, Mendes et al. (2020) generated an *SPL-GFP* (Green Fluorescent Protein) fusion protein expressed using both the 5' and 3' regulatory elements of the *SPL* gene. After confirming that the *pSPL5'::SPL-GFP::3'* construct was able to complement the fertility defects in the *spl* mutant fully, expression of *SPL-GFP* was analysed in three independent lines. *SPL-GFP* expression was restricted to the nuclei of a few cells in the L1 layer of the extreme nucellar tip of premeiotic ovules (stage 1). Interestingly, *SPL-GFP* signal was not detected in the MMC, nor in its products or precursors. This observation contradicts the earlier reports from *in situ* hybridisation studies which suggested a broader expression pattern, including within the MMC (Schiefthaler et al., 1999).

To address concerns about their contradictory findings, Mendes et al. (2020) employed Renaissance staining alongside confocal imaging to more precisely determine the localization of *SPL-GFP* expression during ovule development. Renaissance staining confirmed that *SPL-GFP* signal remains outside of the MMC,

providing further evidence that *SPL* expression is restricted to the L1 layer. However, faint GFP signal is visible in nuclei throughout the primordium, suggesting low-level broader expression. Furthermore, Mendes et al. (2020) generated an *SPL:GUS* reporter construct, which revealed expression restricted to the nucellar tip of ovule primordia. However, unlike the more precise *SPL:GFP* localization to the L1 layer, the *SPL:GUS* staining appeared broader, making it difficult to discern whether expression was confined to a specific cell layer. This is reiterated in an independent study by Bezodis et al. (2022) where broad *SPL:GUS* expression in the nucellar tip was likewise observed. This broader expression pattern is consistent with the observations from Yang et al. (1999) and Schiefthaler et al. (1999), suggesting that differences in reporter sensitivity or construct design may influence the apparent breadth of expression.

Taken together, these studies suggest that while *SPL* expression is clearly enriched in the nucellar tip, the precise spatial localization may vary depending on the detection method used (Mendes et al., 2020; Schiefthaler et al., 1999; Yang et al., 1999). Such observations highlight the need for complementary approaches to confirm expression domains, particularly in complex and dynamic tissues like ovule primordia.

1.2.3 AGAMOUS-directed *SPL* expression shapes floral reproductive organs

The regulatory regions of *AtSPL* contain two CArG boxes, which are DNA-binding motifs recognised by MADS-box transcription factors, suggesting that *SPL* expression may be regulated by such factors (Shore & Sharrocks, 1995). MADS-box transcription factors play essential roles in floral development, functioning in the ABC model to specify floral organ identity across four concentric whorls (Bowman et al., 1991; Coen & Meyerowitz, 1991; Haughn & Somerville, 1988; Ma, 1994; Meyerowitz et al., 1991; Weigel & Meyerowitz, 1994). In this model, class A genes specify sepal identity in the outermost whorl (whorl 4); class A and B genes specify petal identity in whorl 3; class B and C genes specify stamen identity in whorl 2; and class C genes alone specify

carpel identity in the innermost whorl (whorl 1). The class C gene *AGAMOUS* (AG) is critical for stamen and carpel specification, making it a strong candidate for regulating *SPL* expression (Honma & Goto, 2001; Huang et al., 1993; Mizukami & Ma, 1992; Shiraishi et al., 1993).

Ito et al. (2004) investigated the potential regulation of *SPL* by AG and demonstrated that the 3' CArG sequence contains a single mismatch to the consensus AG-binding CArG box. They confirmed *in vitro* binding of AG to this sequence and evaluated its functional significance. Using reverse transcriptase polymerase chain reaction (RT-PCR), *SPL* expression was shown to be dependent on AG in an AG-inducible mutant. Moreover, the application of the protein synthesis inhibitor cycloheximide suggested that this induction was likely direct.

To evaluate whether the 3' AG-binding site is essential for AG-dependent *SPL* expression *in vivo*, Ito et al. (2004) compared the expression of an *SPL-GUS* reporter fusion in wild-type plants and transgenic lines carrying a mutated 3' AG-binding site incapable of binding AG. In wild-type plants, GUS staining recapitulated endogenous *SPL* expression patterns, with strong expression in the stamen primordia and throughout the developing stamen, and in the ovule primordia, nucellus, integuments, and funiculus of developing ovules (Schiefthaler et al., 1999; Yang et al., 1999). In contrast, transgenic plants with a mutated AG binding site exhibited significantly weaker GUS activity and reduced expression domains. GUS staining in the stamen primordia was restricted to the lateral sides, while in ovules, staining was absent in the outer integuments and funiculus. This pattern closely corresponds to AG expression domains, supporting the conclusion that the 3' AG box is necessary for normal *SPL* expression in both the anther and ovule (Reiser et al., 1995). Further evidence of the regulatory importance of the 3' region was provided by the *nzz-3* null allele, caused by a transposon insertion 20 bp downstream of the stop codon, which disrupted normal *SPL* function (Schiefthaler et al., 1999).

Although these findings demonstrated the importance of the AG-binding site in the

3' region of *SPL* for its expression in both stamens and ovules, the authors did not pursue further investigation into the role of *SPL* in ovule development (Ito et al., 2004). Instead, subsequent analyses focused on the regulation and function of *SPL* in stamen development. During early stages of stamen development, expression of the *SPL*-GUS fusion protein was entirely within the expression domain of *AG*, and this expression was abolished in the *ag* mutant (Ito et al., 2004). To assess whether *SPL* could compensate for the absence of *AG* in microsporogenesis, an inducible 35S::*SPL*-GLUCOCORTICOD RECEPTOR (GR) construct was employed. In the mutant *ag* background, dexamethasone (DEX) treatment induced locule-like structures in whorl 3 organs, and pollen grains were detected within these structures. These results suggest that *SPL* can partially induce microsporogenesis in the absence of *AG*, reinforcing its role as a central mediator of sporogenesis. However, ectopic locule formation was restricted to whorl 3 and inward, with no induction in whorl 2, suggesting the involvement of additional whorl-specific factors in microsporogenesis.

Interestingly, there was no evidence of *SPL* rescuing ovule development in the absence of *AG*, indicating that other *AG*-induced factors are necessary for proper ovule formation (Ito et al., 2004). Overall, these findings underscore the complexity of *AG* regulation of *SPL* expression in the developing anther and ovule, reflecting the intricate genetic interactions required to form these highly specialized reproductive structures.

1.2.4 Tight regulation of *SPL* results in the specification of a single MMC precursor

At the molecular level, numerous transcription factors, hormones and small RNAs have been shown to control the early stages of ovule development (see review: (Pinto et al., 2019)). The transcription factor *SPL* plays an early and essential role. The expression data described above further corroborates with this suggestion; *SPL* expression is enriched in the nucellar tip of the ovule primordia, making it a strong

candidate for the specification of the MMC. However, recent research suggests that whilst *SPL* is necessary to specify the MMC precursor, it is not sufficient to confer MMC identity (Mendes et al., 2020). When expressed under the control of the 35S constitutive promoter, eight independent *SPL* lines produced multiple enlarged cells in the L2 layer of the nucellus. Therefore, broadening the expression of *SPL* leads to the formation of supernumerary enlarged cells that resemble the MMC germline precursor. However, further investigation found that these multiple enlarged cells resulting from the upregulation of *SPL* did not possess MMC function.

A network involving ovule identity genes and the RNA-dependent DNA methylation (RdDM) pathway has been proposed to restrict the expression of *SPL*, resulting in the specification of a single MMC precursor (Mendes et al., 2020). The RdDM pathway regulates DNA expression via either mRNA degradation mediated by ARGONAUTE (AGO) or via DOMAINS REARRANGED METHYLASEs (DRM1 and DRM2)-mediated cytosine DNA methylation (Matzke & Mosher, 2014). Furthermore, AGO9 has been linked to the trans-acting small interfering RNA biogenesis gene *RNA-DEPENDENT RNA POLYMERASE 6* (*RDR6*) (Olmedo-Monfil et al., 2010). *ago9-2* and *rdr6-11* single mutant and *drm1drm2* double mutant lines all developed two or more MMC-like cells in 47%, 46% and 65% of premeiotic ovules, respectively (Mendes et al., 2020; Olmedo-Monfil et al., 2010). Furthermore, when *SPL-GFP* was crossed into the *ago9-2* and *drm1drm2* mutant backgrounds, ectopic expression of *GFP* throughout the L1 cell layer of the nucellus was observed. Taken together, the multiple MMC phenotype and the broadening of *SPL-GFP* expression in the RdDM mutants implies that the full RdDM pathway is necessary to prevent multiple MMCs and is likely to be acting upstream of *SPL*. To determine how this pathway may be linked to MMC specification, the potential involvement of ovule identity genes was investigated.

The MADS-box transcription factor *SEEDSTICK* (*STK*) controls ovule identity (redundantly with *SHATTERPROOF1* and 2 (*SHP1* and *SHP2*)) and is expressed in the nucellus, chalaza and integuments of the developing ovule (Brambilla et al., 2007;

Ezquer et al., 2016; Favaro et al., 2003; Pinyopich et al., 2003). In *stk* mutants, two or more enlarged MMC-like cells were produced in 46% of ovules, a broadly similar level to that recorded in *ago9-2*, *rdr6-11* and *drm1drm2* mutant ovules (Mendes et al., 2020; Olmedo-Monfil et al., 2010). Furthermore, quantitative RT-PCR (qRT-PCR) confirmed that *SPL* expression was upregulated in *stk* mutants, whilst *AGO9*, *RDR6* and *DRM1* were significantly downregulated compared to wild type. To investigate whether this regulation by *STK* may be direct, Chromatin Immunoprecipitation (ChIP) assays using an *STK-GFP* line were performed (Mendes et al., 2020; Mizzotti et al., 2020). Mendes et al. (2020) searched specifically for CArG boxes (with a maximum of two mismatches in the consensus sequence) in the regulatory regions of *SPL*, *AGO9*, *RDR6*, *DRM1* and *DRM2*. Two CArG boxes were detected in the *SPL* sequence – one upstream of the coding region and one downstream in the 3'UTR – but *STK-GFP* immunoprecipitated chromatin was not enriched in either of these regions (Mendes et al. (2020), Figure 1.7). Meanwhile, seven CArG boxes were identified in four regions of the promoter and 5'UTR of *RDR6*. Two of these regions were significantly enriched in immunoprecipitated *STK-GFP* chromatin when compared to wild type. Of the six CArG boxes (divided into four regions) identified in the *AGO9* promoter and 5' UTR, one region was significantly enriched in *STK-GFP* chromatin. Meanwhile, no such regions were found in the promoter or 5'UTR of *DRM1* or *DRM2*. Taken together, this data suggests that *STK* may be responsible for the indirect regulation of *SPL* expression via the direct regulation of the RdDM pathway components *AGO9* and *RDR6*, to prevent the formation of multiple MMC-like cells (Mendes et al. (2020), Figure 1.8). This strict molecular control is of the utmost importance to ensure the proper progression of female gametogenesis.

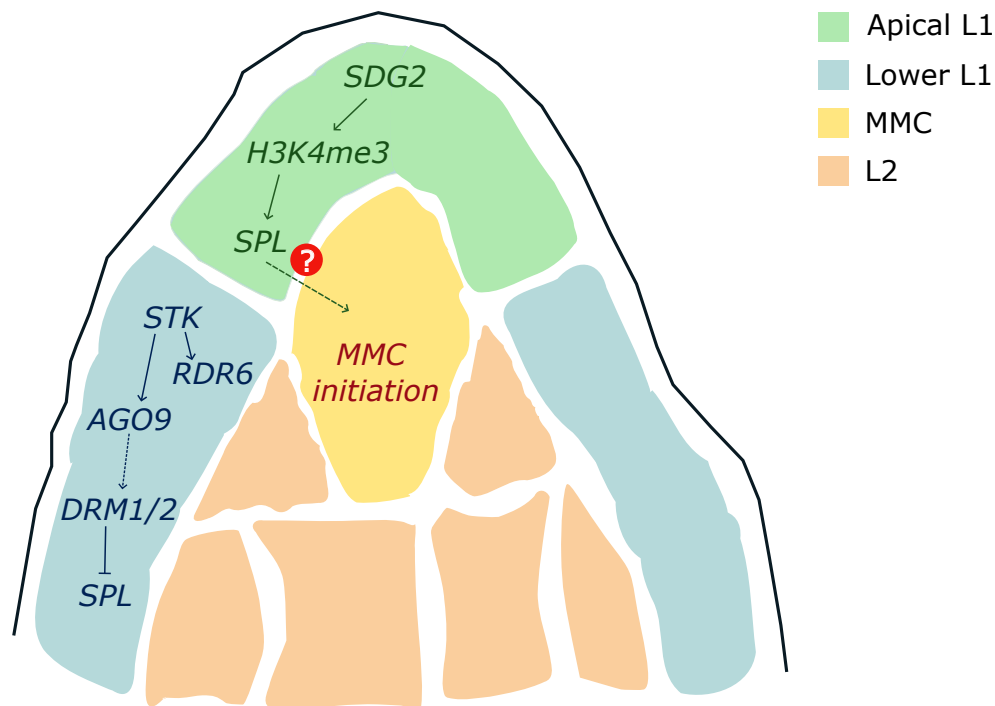


Figure 1.8: Proposed model for the restriction of *SPL* expression to specify a single MMC precursor. In the apical L1 cell layer, *SET DOMAIN GROUP 2* (*SDG2*) is proposed to mediate Histone H3 lysine 4 trimethylation (*H3K4me3*) deposition at *SPL*, thereby activating transcription. *SPL* then specifies MMC identity in the L2 layer via the activation/repression of an unknown factor. Meanwhile, in the lower part of the L1 cell layer, *SEEDSTICK* (*STK*) is proposed to indirectly negatively regulate *SPL* via the direct regulation of the RNA-dependent DNA methylation (RdDM) pathway components *ARGONAUTE9* (*AGO9*), *RNA-DEPENDENT RNA POLYMERASE 6* (*RDR6*) and *DOMAINS REARRANGED METHYLASEs* (*DRM1* and *DRM2*). *STK* binds directly to *AGO9* and *RDR6*, activating the RdDM pathway, which in turn represses *SPL* expression via methylation of the *SPL* genomic region by *DRM1* and *DRM2*. This restriction of *SPL* expression to the apical L1 cell layer ensures only a single MMC is specified in the L2. Adapted from Mendes et al. (2020).

Whilst disrupting the expression of *STK* and/or the RdDM pathway components – and thereby upregulating *SPL* expression – leads to the development of multiple enlarged cells, it was not clear whether these MMC-like cells possessed MMC function (Mendes et al., 2020). To determine the identity of the supernumerary cells,

the transcriptional marker *pKNU::nlsYFP*, which specifically marks cells with MMC identity (Tucker et al., 2003), was crossed into the *stk* and RdDM mutant backgrounds (Mendes et al., 2020). Consistently across *stk* and RdDM mutant ovules, one or fewer of the enlarged MMC-like cells expressed the *YFP* marker. Therefore, although there were multiple MMC-like cells, only one was able to function as MMC. After confirming that the additional MMC-like cells did not possess MMC function, Mendes et al. (2020) investigated whether they shared physical features with the MMC. During wild type megagametogenesis, callose accumulates in the cell wall of the MMC before meiosis, and in the walls of the four megaspores undergoing meiosis in a distinct pattern (Lersten, 2004). Therefore, the pattern of callose deposition in *stk*, *ago9*, *rdr6* and *drm1drm2* mutant ovules before and after meiosis was analysed by Aniline Blue staining (Mendes et al., 2020; Rodkiewicz, 1970). In the premeiotic ovule, callose accumulation in the MMC walls of *stk*, *ago9*, *rdr6* and *drm1drm2* mutants closely resembled the wild type pattern, with callose only ever accumulating in one enlarged MMC-like cell (Mendes et al., 2020). Furthermore, callose accumulation during the second meiotic division was only ever observed in one of the two (or more) enlarged MMC-like cells in the mutant backgrounds, thus meiosis was only ever progressing in one cell. Therefore, whilst indirectly overexpressing *SPL* leads to the formation of multiple MMC-like cells, *SPL* alone cannot confer full MMC identity – there must be other factors involved. Since the ectopic expression of *SPL* throughout the ovule primordia leads to the formation of multiple MMC-like cells, this suggests that *SPL* itself cannot inhibit MMC-like cell formation. Rather, it is more likely that, following the specification of the MMC precursor by *SPL*, a specialised transcriptome unique to the L1 cell layer at the extreme nucellar tip is necessary to confer full MMC function.

In contrast to the negative regulation of *SPL* by the RdDM methylation pathway, it has been proposed that Histone H3 lysine 4 trimethylation (H3K4me3) deposition is responsible for the transcriptional activation of *SPL* via *SET DOMAIN GROUP 2* (*SDG2*) (Berr et al., 2010). Microarray analysis of an *sdg2-1* mutant found that *SPL* transcripts

were significantly downregulated compared to wild type, which was confirmed by RT-PCR. Furthermore, *in situ* hybridization analysis found that *SPL* expression was clearly reduced in the ovule primordia of *sdg2-1* mutants. Since protein immunoblot analysis showed a dramatic reduction in the level of H3K4me3 in *sdg2-1* plants, ChIP analysis of H3K4me3 was performed at the *SPL* locus. In wild type plants, relatively high levels of H3K4me3 were detected at *SPL*, whilst they were dramatically reduced in the *sdg2-1* mutant. Taken together, this data indicates that *SDG2* mediates H3K4me3 deposition at *SPL*, thereby potentially activating transcription (Figure 1.8).

1.2.5 *SPL* acts as an adaptor-like transcriptional co-repressor to regulate downstream ovule genes

Following its specific activation, *SPL* interacts with several factors to regulate downstream expression in the ovule. Via its EAR motif, *SPL* binds to the known transcriptional co-repressors TOPLESS (TPL)/TPL-RELATED (TPRs), thereby inhibiting CINCINNATA (CIN)-like TEOSINTE BRANCHED1/CYCLOIDEA/PCF (TCP) transcription factors (Wei et al., 2015). Y2H assays demonstrated that *SPL* interacts with TPL, while *SPL*ΔEAR and *SPL* with mutations in the EAR domain (*SPL*mEAR) do not. Furthermore, an *SPL*ΔEAR construct fused to the C-terminus of TPL (TPLC) and driven by the *SPL* promoter was able to rescue the sterile phenotype of the T-DNA insertion mutant *spl-2* (Wei et al., 2015). In contrast, *SPL*ΔEAR without TPLC could not functionally complement sterility. These findings demonstrate that TPL is required for *SPL* function in yeast, and the *SPL* EAR motif is essential for its recruitment.

To confirm the role of TPL during ovule development, its expression and function were examined (Wei et al., 2015). Dynamic *GUS* expression driven by the *TPL* promoter was detected throughout ovule development, and qRT-PCR confirmed overlapping expression of *SPL* and *TPL*. In the dominant-negative mutant *tpl-1*, 43.8% of ovules were aborted, and the mutants arrested at the same stage of ovule development as *spl-2* (Wei et al., 2015). This data provides strong evidence that *SPL* and *TPL* interact

to drive early ovule development, with the SPL EAR domain being crucial for this interaction.

A Y2H screen using SPL as bait identified all eight CIN-like TCP transcription factors as potential interactors (Wei et al., 2015). This interaction was found to occur with the N-terminal region of SPL and did not require the EAR motif. Among these, TCP2, TCP3, TCP5, TCP10, TCP17 and TCP24 were confirmed to interact with SPL *in vivo* through firefly luciferase complementation assays. Given that SPL interacts with TPL/TPRs at its C-terminal EAR motif, and with CIN-like TCPs at its N-terminal, Wei et al. (2015) hypothesised that SPL may be acting as a bridge linking TPL/TPRs to TCPs. Supporting this idea, qRT-PCR of the SPL gain-of-function mutant *spl-D* revealed downregulation of TCP target genes, mirroring findings in the *jaw-5D* mutant, where five CIN-like TCPs are inhibited. This suggests that excessive SPL represses TCP activity. Additionally, disrupting TPL/TPRs largely reversed this repression, implying that TPL/TPRs are required for the inhibitory activity of SPL. To further validate this interaction, TCP17 fused to the C-terminal end of SPL and driven by the 35S promoter was compared to the *spl-D* overexpression phenotype (Li et al., 2008; Wei et al., 2015). Both lines exhibited the same mutant phenotype of up-curved leaves and fewer, smaller flowers (Li et al., 2008; Wei et al., 2015), confirming that SPL links CIN-like TCPs with TPL/TPRs.

To explore the role of this interaction in ovule development, the expression and function of CIN-like TCPs were analysed (Wei et al., 2015). CIN-like TCPs were expressed at all stages where SPL expression was detected in the ovule. Further analysis of the *tcp3 tcp4 tcp5 tcp10* quadruple mutant and *jaw-5D* mutant revealed a shared phenotype with the *spl-D* overexpression line: seeds protruding from siliques due to disrupted ovule arrangement (Li et al., 2008; Wei et al., 2015). Furthermore, the double *jaw-5D/spl-D* mutant displayed a more enhanced phenotype. Conversely, overexpression of TCP3 and TCP5 led to aborted ovules, phenocopying the *spl* mutant. Taken together, these results suggest that SPL regulates ovule development

by recruiting TPL/TPR co-repressors to suppress TCP activity. By controlling TCP-mediated gene expression, SPL ensures proper ovule development. This mechanism may explain how SPL modulates the expression of its downstream targets, either repressing or relieving their suppression as required.

SPL is also known to regulate pollen development in *Arabidopsis* (Wei et al., 2015). It remains unclear whether SPL regulates pollen development through a similar mechanism, such as recruiting TPL/TPR co-repressors to specific transcription factors. However, evidence suggests this may be the case. TPL was found to be highly expressed in stamens and pollen, and *tpl-1* mutants produced defective pollen. Furthermore, the pSPL-SPL Δ EAR-TPLC construct not only rescued the female sterility of *spl-2* but also complemented its male sterility (Wei et al., 2015). This strongly indicates that the SPL-TPL/TPR interaction also functions in pollen development. Recent studies support this hypothesis, showing that TPL/TPR co-repressors directly interact with EAR motif-containing transcriptional repressors DAZ1 and DAZ2 during microgametogenesis and male germ cell division (Wei et al., 2015). These findings highlight the pivotal role of TPL/TPR-mediated gene repression in both ovule and pollen development.

Although it has not been directly investigated, there is substantial evidence to suggest that SPL uses the same mechanism - recruiting TPL/TPR co-repressors to specific transcription factors - to regulate pollen development (Wei et al., 2015). For example, *TPL* expression is highly localized in stamens and pollen, and *tpl-1* mutants exhibit pollen defects. Additionally, pSPL-SPL Δ EAR-TPLC was able to restore both male and female fertility in *spl-2* mutants (Wei et al., 2015). This strongly suggests that SPL-TPL/TPR interactions play a significant role in pollen development. Supporting this, TPL/TPRs have been shown to directly interact with the EAR motif-containing transcriptional repressors DAZ1 and DAZ2, which are critical for microgametogenesis and male germ cell division (Borg et al., 2014). These findings underscore the importance of TPL/TPR-mediated repression not only in ovule

development but also in the regulation of pollen formation and function.

In summary, SPL is a critical component in a complex regulatory network that ensures proper ovule and pollen development, which are essential for seed formation (Schiefthaler et al., [1999](#); Yang et al., [1999](#)). While interactions between SPL and AGAMOUS, RdDM pathways, and downstream factors like TPL/TPRs and CIN-like TCPs highlight a pivotal role for this transcription factor, it is likely only one part of a broader system of interconnected pathways (Chen et al., [2014](#); Ito et al., [2004](#); Mendes et al., [2020](#); Wei et al., [2015](#)). These findings underscore the intricate genetic and molecular processes required to coordinate reproductive development in *Arabidopsis* and pave the way for future studies to uncover additional regulatory players in this network. How these complex networks first evolved to make seeds is not currently known.

1.3 *Ceratopteris richardii*: a model for evo-devo studies

Ferns occupy a unique position in the plant evolutionary tree, bridging features of both non-seed and seed plants (Pryer et al., 2001). Like bryophytes, they produce spores that germinate into free-living haploid gametophytes, which exhibit distinct patterns of cell division and differentiation (Raven, 2002). Simultaneously, they share key traits with vascular plants, such as a three-dimensional, long-lived sporophyte generation with vascular tissue, true roots and leaves, and a permanent shoot apical meristem (Harrison, 2017; Johnson & Renzaglia, 2008). This combination of ancestral and derived features, coupled with their phylogenetic position as the closest extant relatives to seed plants, makes ferns an ideal system for evolutionary development (evo-devo) studies (Pryer et al., 2001).

Ceratopteris richardii, a homosporous leptosporangiate fern in the Pteridaceae (introduced in Section 1.1.3), has been a cornerstone of fern research for over three decades (Hickok et al., 1987, 1995; Schuettpehl et al., 2007). This species possesses numerous advantages as a model organism: gametophytes are easily cultured *in vitro*, allowing for the rapid screening of thousands of individuals within weeks (Hickok et al., 1995). Hermaphroditic gametophytes are capable of self-fertilization, enabling the production of stable homozygous sporophytes in a single generation (Schedlbauer & Klekowski, 2008). Its complete life cycle - from spore to mature sporophyte - can be completed within approximately four months under laboratory conditions, making *C. richardii* both accessible and efficient for experimental studies (Hickok et al., 1995).

Despite these advantages, *C. richardii* is not entirely representative of all ferns (Kinosian & Wolf, 2022). Its semi-aquatic habit is rare among the approximately 12,000 extant fern species (Schuettpehl et al., 2016), and its remarkably short life cycle, while ideal for laboratory studies, is atypical for ferns (Stein, 1971). Additionally, *Ceratopteris* produces fewer spores per sporangium - 32 or 16 compared to the typical 64 in other leptosporangiate ferns (Lloyd, 1974). These unique traits enhance its

experimental utility but highlight its limitations as a representative fern species (Kinosian & Wolf, 2022).

Nonetheless, *C. richardii* remains a highly valuable model for fern evo-devo research. The availability of an ontogenic framework, with detailed description of gametophyte and sporophyte development provides an important reference to ensure reproducibility across the *Ceratopteris* community (Conway & Di Stilio, 2020). The species also benefits from advanced experimental tools, including reliable methods for genetic transformation, such as microparticle bombardment and *Agrobacterium*-mediated transformation (Bui et al., 2015; Plackett et al., 2014). These techniques have facilitated studies of molecular function through RNA interference (RNAi) (Plackett et al., 2018; Withers et al., 2023), gene editing (Jiang et al., 2024; Xiang & Li, 2024) and gene overexpression (Youngstrom et al., 2022). Furthermore, the publication of a high-quality reference genome has paved the way for genome-wide studies of gene function and regulation, enabling comparisons with seed plants and other vascular lineages (Marchant et al., 2022; Marchant et al., 2019). These resources, combined with its rapid life cycle and ease of propagation, make *C. richardii* an indispensable tool for investigating the molecular and developmental underpinnings of plant evolution.

Taken together, the tractability of its life cycle, the availability of advanced genetic and genomic resources, and its phylogenetic position make *C. richardii* an indispensable model for exploring the molecular and developmental underpinnings of plant evolution (Bui et al., 2015; Conway & Di Stilio, 2020; Hickok et al., 1995; Plackett et al., 2014). These resources have facilitated studies uncovering conserved roles of key developmental genes, bridging the evolutionary gap between non-seed and seed plants. For instance, research on the *WUSCHEL*-related *HOMEBOX B* (*CrWOXB*) gene revealed its conserved role in promoting cell division and organ development in both gametophyte and sporophyte generations (Youngstrom et al., 2022). Similarly, a study of the transcription factor *LEAFY* (*LFY*) in *C. richardii*

demonstrated its intermediate evolutionary function: maintaining cell divisions in apical stem cells of both haploid and diploid phases, thereby bridging its roles in mosses (zygotic cell division) and angiosperms (floral meristem identity) (Plackett et al., [2018](#)). Additionally, work on *EXCESS MICROSPOROCYTES 1* (*EMS1*), a gene downstream of *SPOROCYTELESS* (*SPL*) in *Arabidopsis*, highlighted its role in restricting sporangia formation to reproductive fronds, offering insights into the regulation of reproductive organ development in ferns (Withers et al., [2023](#)). Together, these studies underscore the value of *C. richardii* as a model for investigating conserved genetic pathways and their evolutionary adaptations across the plant lineage.

1.4 Aims and Objectives

This thesis aims to elucidate the evolutionary and functional roles of key reproductive regulators in early-diverging plant lineages, with a specific focus on *Ceratopteris richardii* as a model system. By integrating phylogenetic, molecular, and transcriptomic approaches, this work seeks to uncover the genetic networks underpinning reproductive development, their evolutionary conservation across land plants, and their contribution to reproductive innovation. Chapters 3 and 4 specifically investigate the *SPL* gene family, while Chapter 5 adopts a broader approach to identify co-expression networks and candidate regulators of reproductive development. The specific objectives of this thesis are:

1. **To investigate the evolutionary relationships of *SPL* and *SPEAR* homologs across diverse plant lineages:**
 - Conduct a comprehensive phylogenetic analysis to elucidate patterns of divergence and conservation.
 - Identify how these genes may have transitioned from ancestral roles in non-seed plants to specialised roles in seed plants.
2. **To characterise the spatial and temporal expression of *SPL*-like genes in *C. richardii*:**
 - Examine expression patterns in reproductive and non-reproductive tissues to explore their potential roles in reproductive development.
 - Integrate expression data with phylogenetic results to hypothesise how *SPL* functions evolved.
3. **To functionally investigate *CrSPL* genes in *C. richardii* using molecular tools:**
 - Develop RNAi constructs and establish effective transformation methods to generate *CrSPL* knockdown lines.

- Characterise phenotypes associated with *CrSPL* knockdown, focusing on the key homolog *CrSPLB*.
- Attempt functional complementation of the *Atspl* mutant with *CrSPL* homologs.

4. To uncover co-expression networks and key regulators of reproductive development in *C. richardii*:

- Evaluate the effectiveness of genome-guided and transcriptome-based read-mapping strategies using the recently published *C. richardii* genome and transcriptome.
- Perform a weighted gene co-expression network analysis (WGCNA) to identify modules correlated with reproductive tissues and stages.
- Identify and characterise hub genes within these modules, particularly those linked to key developmental processes.

Chapter 2

Materials and Methods

2.1 Plant growth and maintenance

Wild-type (WT) spores of *Ceratopteris richardii* strain *Hn-n* were obtained from Dr Andrew Plackett (University of Birmingham). WT Col-0 and *spl-4* seeds of *Arabidopsis thaliana* were provided by Dr Andrew Plackett (University of Birmingham).

2.1.1 *Ceratopteris richardii* sterile tissue culture

Ceratopteris richardii was grown in sterile culture under long day (LD) conditions (16 h light / 8 h dark, 28 °C / 28 °C) with 50 % humidity, according to the protocol described in Plackett et al., 2015. All steps were carried out in a sterile laminar flow cabinet (ESCO). *C. richardii* spores were sterilised in 20 % bleach solution (20 % Sodium Hypochlorite, 0.1 % Tween-20) for 15 minutes at room temperature with gentle mixing. Sterilised spores were washed with sterile dH₂O four times before resuspension in 1 mL dH₂O. Suspended spores were incubated at room temperature in the dark for 48-72 hours to imbibe.

Imbided spores were sown on 90 mm tissue culture plates with 25 mL 1X C-fern nutrient 1 % agar medium, pH 6.0 (Plackett et al., 2015) (Appendix 6.5). A 90 mm cellophane disc, sterilised between filter paper discs (Whatman plc) in a glass petri dish for 20 minutes at 121°C, was placed between the media and spores to prevent penetration of rhizoids into the media. Cellophane discs were submerged in sterilised dH₂O and laid flat on the media using watchmakers forceps, without introducing air bubbles. Germinating spores were kept wet for 7 days by the addition of sterile dH₂O as required. Water was then restricted for a further 7 days to ensure synchronised fertilisation of gametophytes. At 14 days, plates were flooded with 7 mL dH₂O to induce fertilisation by free-mating. Sporophytes became visible growing on the hermaphrodite gametophyte thallus approximately 7 days after fertilisation. Sporophytes developed a visible shoot and root approximately 14 days after fertilisation. At this stage, sporophytes were maintained in tissue culture for

subsequent RNA extraction, callogenesis or transfer to soil.

To maintain sporophytes in sterile culture, sterile watchmakers forceps were used to isolate individual sporophytes and transfer to a new 90 mm C-fern nutrient media plate without cellophane. High humidity was maintained by addition of dH₂O as required. Sporophytes can be maintained in this manner for one month, after which they must be transferred to soil, harvested or discarded.

2.1.2 *C. richardii* sporophyte growth on soil

One month old *C. richardii* sporophytes were transplanted to Levington Peat Free Multi Purpose Compost (Surrey, UK). Plants were grown under long-day (LD) conditions at 28 °C, 85 % humidity in a MLR-352H-PE Climate Chamber (PHCbi, Europe). Soil was maintained in a fully waterlogged state, with trays filled to the rim every 2-3 days. To mitigate the spread of mildew and prevent obscuring of emerging fronds, vegetative fronds were regularly pruned as new fronds emerged. A minimum of 4 fronds were retained on a plant at any given time. Approximately 4 months after transplantation to soil, mature reproductive fronds were ready for harvesting. Reproductive fronds were harvested when the sporangia were visibly red. Whole fronds were excised at the base and sealed in bags.

2.1.3 *Arabidopsis thaliana* growth and maintenance

A. thaliana seeds were surface sterilised in 70 % ethanol for 5 minutes, then 10 % bleach solution (10 % sodium hypochlorite, 0.1 % Tween-20) for a further 15 minutes with gentle mixing. Seeds were then rinsed four times with sterile dH₂O, resuspended in dH₂O and stratified by incubation at 5 °C in the dark for 48 hours.

Stratified seeds were sown on half-strength Murishage and Skoog (MS; Sigma M0404) medium (1/2 MS), 2 % sucrose, 0.7 % agar nutrient medium (Appendix 6.5) pH 5.8 in 90 mm plates, and kept in LD conditions at 25 °C. Antibiotic was added to the media at the appropriate concentration where required. Approximately 7 to 14 days

after sowing, when four true leaves were present on germinated seeds, seedlings were transplanted to Levington Peat Free Multi Purpose Compost (Surrey, UK). Plants were grown under LD conditions at 25 °C in a MLR-352-PE Climatic Test Chamber (PHCbi, Europe).

2.2 Isolation of Plant Nucleic acids

2.2.1 Isolation of genomic DNA from *A. thaliana*

Arabidopsis genomic DNA was isolated via a crude Hexadecyltrimethylammonium bromide (CTAB) extraction. 200 mg plant tissue was ground in 200 μ L 2 % CTAB extraction buffer (Appendix 6.5; 100 mM Tris-HCl pH9, 2 % CTAB, 1.4 M NaCl, 20 mM EDTA, 5 % RNAase A) and incubated at 65 °C for 20 minutes. 200 μ L Chloroform isoamyl alcohol (24:1) was then added, mixed and centrifuged at 13000 RCF for 2 minutes. The resulting supernatant was transferred to a new 1.5 ml microcentrifuge tube and 200 μ L Propan-2-ol added. The solution was inverted to mix, incubated at -20 °C for 5 minutes, then centrifuged at 13000 RCF for 5 minutes. The pellet was washed with 70 % Ethanol and centrifuged at 13000 RCF for 1 minute. Ethanol was removed and the DNA resuspended in 1X TE buffer (pH 8). DNA concentration was quantified using a NanoDrop 1000 Spectrophotometer (Thermofisher Scientific).

2.2.2 Isolation of RNA from *C. richardii* and *A. thaliana*

Total RNA was extracted from 30-50 mg plant tissue using the RNeasy Plant Mini Kit (Qiagen) following the manufacturer's protocol. Plant tissue was flash-frozen at harvest and transported in liquid nitrogen. Frozen tissue was ground in RLT buffer with 10 μ L β -Mercaptoethanol added per mL of buffer to denature proteins. Homogenised solutions were immediately transferred to a QIAshredder column and centrifuged at 13000 RCF for 2 minutes to remove cell debris. The supernatant of the flow-through was directly transferred to a new 1.5 mL RNase-free microcentrifuge tube with 200 μ L 100 % ethanol, mixed and transferred to an RNeasy Mini spin column, where the RNA was absorbed onto a silica membrane. Samples were washed once with buffer RW1 and centrifuged for 1 minute at 13000 RCF. DNase solution was directly applied to the membrane and incubated at room temperature for 30 minutes to degrade any DNA

present in the sample. The membrane was then subjected to a second wash with RW1 and a further 2 washes with buffer RPE. Before elution, the column was transferred to a new collection tube and centrifuged for 2 minutes at 13000 RCF to dry the membrane. The column was then transferred to a clean, RNase-free 1.5 mL microcentrifuge tube, and 35 μ L RNase-free water added directly to the column membrane, incubated for 1 minute, then centrifuged for 1 minute at 13000 RCF. The eluted RNA was then replaced onto the column membrane, incubated and spun again to collect any residual RNA still bound to the membrane. Samples were then immediately quantified using the NanoDrop (Thermofisher Scientific).

2.2.3 cDNA synthesis of extracted RNA

First-strand cDNA synthesis was performed using SuperScript™ II Reverse Transcriptase (RT) (Invitrogen). A total of 500 ng of RNA was used for each reaction. Where the RNA concentration was sufficient, the required volume was calculated and adjusted with RNase-free water to achieve a final volume of 11 μ L. For samples with low RNA concentrations, the RNA was not diluted, and the final concentration was recorded. All surfaces were treated with RNaseZap (Invitrogen) and 70% ethanol before beginning the protocol to minimize RNA degradation.

1 μ L Oligo(dT) primer (500 ng μ L⁻¹) and 1 μ L dNTP (10 mM) (Thermo Scientific) was added to the RNA and incubated at 65 °C for 5 minutes, followed by rapid cooling on ice. 4 μ L of 5X First-Strand Buffer and 2 μ L of 0.1 M DTT was added to each reaction. The tubes were gently mixed, centrifuged, and pre-incubated at 42 °C for 2 minutes. Subsequently, 1 μ L of SuperScript™ II RT was added directly to each reaction, and the mixture was incubated at 42 °C for 50 minutes to synthesize the cDNA. The reaction was terminated by heating to 70 °C for 15 minutes.

Residual RNA was removed by adding 0.5 μ L of RNase H to each reaction, followed by incubation at 37 °C for 30 minutes. The resulting cDNA was diluted to a 1:20 working stock and stored at -20 °C for subsequent use.

2.3 Expression analyses

2.3.1 PCR with GoTAQ polymerase

cDNA generated in the above steps was used as template in a Polymerase Chain Reaction (PCR) with GoTaq DNA polymerase (Promega). Primers were designed to amplify the full coding regions of *CrSPLA*, *CrSPLB*, *CrSPLC* and *AtSPL*. Primers were designed manually with the following criteria: melting temperature (T_m) of $55 \pm 2^\circ\text{C}$, 40-60 % GC content and length 18-30 bp. Additionally, primers had a G or C at the 5' and 3' ends, had no more than 3 Gs or Cs in a row, and the T_m of forward and reverse primers were within 1°C of each other. Primers were synthesised by Integrated DNA Technologies (IDTDNA, Belgium). Primers used in RT-PCR are detailed in Table 2.1.

Table 2.1: Primers used in PCR expression analyses.

Primer	Sequence	Target
CrSPL_A-F CrSPL_A-R	GTCTTCATCGCAGAGTCTTG CACCAAGTTGTCCAACAGC	CrSPLA
CrSPL_B-F CrSPL_B-R	GATACACGCACAGTG GAATTTGCAACAATTAGC	CrSPLB
CrSPL_C2-F CrSPL_C2-R	GATTATGGCGTATCTTTCCC CCATTATACAGAGACATTCTGG	CrSPLC
AtSPL-F AtSPL-R	CGTCTTTACCTCTTCCCTTC GACTCAGCTCTCAGATGC	AtSPL

PCR with GoTAQ DNA polymerase was carried out according to the manufacturer's instructions. Reaction mixture consisted of 50 ng cDNA, 0.46 μL 10 mM dNTP, 1 μL 10 μM forward and reverse primer, 4 μL 5X GoTAQ buffer, 1.2 μL 25 mM MgCl_2 and 0.09 μL GoTAQ enzyme in a 20 μL volume. Thermal cycling conditions followed those recommended in the GoTAQ polymerase protocol: 2 minutes initial denaturation at 94°C , followed by 30 cycles of denaturation at 94°C for 15 s, annealing for 30 s with temperature adjusted according to primer T_m , extension at 72°C adjusted according to template size (1 Kb per minute), then a final extension at 72°C for 6 minutes. Samples

were then run immediately on an agarose gel. In addition, bands of the expected size were excised, cloned and sequenced according to the protocols outlined in sections 2.4.4 to 2.4.10, to confirm the correct sequences had been amplified.

2.3.2 Quantitative RT-PCR

Primers were designed to analyse the expression of *CrSPLA*, *CrSPLB* and *CrSPLC* (Table 2.2) in WT and Transgenic tissues according to the following criteria: primers must span an exon-exon boundary, have a T_m of $60 \pm 1^\circ\text{C}$ and produce a product of size 140-160 bp. Quantitative RT-PCR (qPCR) was performed with an AriaMx machine (Agilent) using Brilliant III ultra-fast SYBR®Green low ROX qPCR master mix (Agilent). 2.5 μL SYBR®Green Supermix, 0.25 μL 10 μM forward and reverse primer, and 2 μL 5 $\text{ng } \mu\text{L}^{-1}$ cDNA were mixed in individual wells of a 384 well qPCR plate. A minimum of 3 biological and 4 technical replicates was performed per sample. The cycling parameters were followed according to the manufacturer's instructions: initial denaturation at 95°C for 10 minutes, followed by 40 cycles of 95°C denaturation for 15 s and 60°C annealing for 20 s. A melt curve was produced for each reaction to confirm primer specificity.

Table 2.2: Primers, sequences, and targets for qPCR analysis.

Primer	Sequence	Target
qCRSPLA-F qCrSPLA-R	CGTGTGCTCAACAGCTTGCC GTCACACAAGAGGTTCCAATCATCC	CrSPLA
qCrSPLB-F qCrSPLB-R	CTCCAGTTGCTGCACAGAGAGATG GCAACCCTGCCATGCTTGTAC	CrSPLB
qCrSPLC-F qCrSPLC-R	CGAAGTACCAGGCAAAGGAAACC GCTTCTCAGAAGCGAAGAGAGAGTC	CrSPLC
qM08-F qM08-R	TCAATCTGCTCCCTGCTCAC GTCCAGTTCCTTGCTGTTGC	putative Ubiquitin9
qActin-F qActin-R	GAGAGAGGCTACTCTTTCACAACC AGGAAGTTCGTAACCTCTCTCCAA	Actin B

The comparative C_q ($\Delta\Delta C_q$) method was used to analyse relative expression,

according to Taylor et al., [2019](#). First, the mean quantitative cycle (Cq) of the technical replicates was determined. The relative difference (ΔCq) was then calculated by taking the average Cq of all samples in the control group from the mean Cq of each individual sample. Relative quantities ($2^{\Delta Cq}$) representing the fold change between the biological groups before reference gene normalisation were determined. The relative quantities of each biological group were then divided by the geometric mean of the reference gene relative quantities and log-transformed to give relative normalised expression, from which standard error was calculated. The geometric mean of the relative normalised expression for each biological group was calculated to give the average relative normalised fold expression.

Statistical analysis was performed using the relative quantities of biological groups before log transformation. Analysis of Variance (ANOVA) followed by multiple comparisons with Tukey's Honest Significant Differences (HSD), adjusted by Generalised Linear Hypothesis Testing (GLHT) was carried out to determine if there was any significant difference between relative expression values.

2.4 Cloning

Two different cloning methods were used to generate the constructs for RNAi knockdown of *C. richardii* genes in *C. richardii* and for complementation of the *Arabidopsis spl-4* mutant. RNAi knockdown constructs were generated by Gateway cloning with BP and LR reactions. Complementation constructs were built using traditional restriction cloning. The majority of molecular cloning steps are shared by both, with the main differences occurring at ligation steps.

2.4.1 Primer design

To amplify the RNAi targets for use in knockdown of *CrSPLs*, primers were designed to amplify distinct regions of exon 1 of *CrSPLA*, *CrSPLB* and *CrSPLC* (Table 2.3). Primers were designed following the same criteria as in Section 2.3.1, with *AttB1* and *AttB2* sites added to the 5' ends of the forward and reverse primers, respectively.

Table 2.3: Primers used to amplify *CrSPL* RNAi target regions (bold) with *attB* sites (underline).

Primer	Sequence
CrSPL_A_attB1-F	GGGGACAAGTTTGTACAAAAAAGCAGGCT <u>AAGGCTAGCTTTACAGACC</u>
CrSPL_A_attB2-R	GGGGACCACTTTGTACAAGAAAGCTGGGT <u>TCCACTAAAGAAGCAGGAG</u>
CrSPL_B_attB1-F	GGGGACAAGTTTGTACAAAAAAGCAGGCT <u>CATCTCCTCTACTCATTCTCTG</u>
CrSPL_B_attB2-R	GGGGACCACTTTGTACAAGAAAGCTGGGT <u>AGTGAGTTCTGCAAGGAATTC</u>
CrSPL_C_attB1-F	GGGGACAAGTTTGTACAAAAAAGCAGGCT <u>CACATAGCTTCTTGTATGC</u>
CrSPL_C_attB2-R	GGGGACCACTTTGTACAAGAAAGCTGGGT <u>GCAAAGAACATTCCACATTC</u>

To amplify the modules necessary for the complementation cassette, primers were similarly designed to capture the desired regions and add the appropriate restriction sites to the 5' and 3' ends (Table 2.4).

Table 2.4: Primers used to amplify complementation construct components (underline) and add restriction sites (bold).

Name	Sequence
SPLpro-NotI-F	CAG CGGCCGCT AGAAATGCAATACATGGCTG
SPLpro-KpnI-R	CATG GGTACC TGATGATGATCTTCTTCTCGGA
SPL3'UTR-XbaI-F	CCAT CTAGA ATGTTTTATCTTTCTATATTGATTAAAC
SPL3'UTR-NotI-R	CAG CGGCCGCG GATCGTCTTCTTCTTCTTCTCTGC
mCHERRY-linker-F	ggagcagctgctggctgccgctgcggcagcgccggatccATGGTGAGCAAGGGCGAG
mCHERRY-R	TTACTTGACAGCTCGTCCATGC
mCHERRY-KpnI-F	CTA GGTACC GGAGCAGCTGCGGCTG
mCHERRY-XbaI-R	CATGT CTAGA TTACTTGTACAGCTCGTCCATGC
AtSPL-KpnI-F	CAG GTACC ATGGCGACTTCTCTTCTTCTTC
AtSPL-KpnI-R	GT GGTACC TTAAAGCTTCAAGGACAAATCAATGG
AtSPL-KpnI-NoSTOP-R	GT GGTACCA AGCTTCAAGGACAAATCAATGG

2.4.2 PCR with Phusion High Fidelity Polymerase

Cloned *CrSPLA*, *CrSPLB* and *CrSPLC* coding sequences from expression analyses in section 2.3.1 were used as template in a PCR reaction to amplify the RNAi target regions with *attB* sites. Meanwhile, PCR to generate the complementation modules used *Arabidopsis* gDNA as template. Amplification was achieved using the Phusion High Fidelity DNA polymerase (New England Biolabs). Reaction mixture was as follows: 50 ng DNA template, 0.75 μ L 10 μ M forward and reverse primer, 0.5 μ L 10 μ M dNTP, 5 μ L GC buffer and 0.5 μ L Phusion DNA polymerase in a 25 μ L volume. Thermal cycling conditions recommended by the Phusion protocol were followed: initial denaturation at 98 °C for 30 s, followed by 34 cycles of denaturation at 98 °C for 10 s, annealing for 20 s, temperature adjusted according to primer T_M , extension at 72 °C according to target size (2 Kb/minute), and a final extension at 72 °C for 2 minutes. Products were then immediately run on an agarose gel.

2.4.3 Agarose gel electrophoresis

0.8 % agarose gels were used for all procedures in this thesis. Agarose was dissolved in 1X TAE solution (Appendix 6.5) by microwave heating, cooled to 70 °C,

poured into an appropriately sized mould with a comb, and allowed to set for 20 minutes. Gels were then submerged in a gel tank containing 1X TAE solution. The DNA dye GelRed (Biotium) was mixed with 6X loading buffer (3 μ L in 1 mL loading buffer). The loading buffer with GelRed was then mixed with samples and the 10 Kb ladder before loading, and samples were run for 60 minutes at 80 V. Gels were imaged using a BioRad Gel Doc viewer.

2.4.4 Isolation of DNA fragments from agarose gels

Individual fragments of interest were separated by agarose gel electrophoresis. Band patterns were visualised with a UV transilluminator, and target bands were directly excised with a clean razor blade. Nucleic acids in the gel slice were isolated using the Zymoclean gel DNA recovery kit (Zymo Research) according to the manufacturer's protocol. Excised gel slices were weighed and dissolved in 3 volumes of agarose dissolving buffer (ADB) at 55 °C for 10 minutes. Once fully dissolved, the melted agarose solution was transferred to a Zymo-Spin column in a collection tube and centrifuged for 1 minute at 11000 RCF to bind the DNA. Bound DNA was then washed twice with 200 μ L DNA wash buffer and centrifuged at 11000 RCF for 30 s. DNA was eluted using the DNA elution buffer provided with the kit: 6 μ L was added directly to the column membrane and centrifuged at 11000 RCF for 1 minute. Eluted DNA was then immediately ligated into pJET for propagation.

2.4.5 Ligation

To confirm that the intended fragments had been PCR amplified and gel extracted, eluates from the above steps were cloned with the CloneJET PCR cloning kit (Thermo Scientific) before proceeding with vector building. Propagation in pJET_{1.2} (Figure 2.1) enabled diagnostic digests and sequencing to be carried out. 1.5 μ L PCR gel elution was mixed with 2.5 μ L 2X reaction buffer, 0.5 μ L pJET_{1.2}/Blunt cloning vector and 0.5 μ L T4 DNA ligase and incubated at room temperature for 15 minutes. The ligation product

was then immediately transformed into *E. coli* DH5 α .

To generate *CrSPL* RNAi vectors, a two-step Gateway cloning process was followed (Invitrogen). RNAi targets were first ligated into the donor vector pDONR207 *via* site-specific recombination of *attB* sites at the ends of the target fragments with *attP* sites present in pDONR207, generating *attL* sites (Figure 2.2). This was achieved *via* the BP reaction detailed in the Gateway cloning manual: 150 ng RNAi target, 150 ng pDONR207 and 2 μ L BP enzyme was mixed with 1X TE buffer (pH 8) up to a volume of 15 μ L and incubated at 25 °C overnight. 1 μ L Proteinase K was then added to the ligation and incubated at 37 °C for 10 minutes. The ligation product was then transformed into DH5 α *E. coli* and verified before the second step.

In the second step, the LR reaction was carried out to recombine two copies of the target fragment with *attL* sites in opposite orientations into the pANDA vector containing AttR sites (Miki & Shimamoto, 2004) (Figure 2.3). 150 ng RNAi-pDONR207, 150 ng pANDA and 2 μ L LR clonase was mixed with 1X TE buffer (pH 8) up to a volume of 15 μ L and incubated at 25 °C overnight. 1 μ L Proteinase K was then added to the ligation mix and incubated at 37 °C for 10 minutes. The resulting ligation mix was transformed into DH5 α for propagation.

The *Arabidopsis* complementation vector was built using traditional restriction cloning. Modules were added step-wise to the 35S::OCS-pART7 (Figure 2.4) intermediate vector *via* sequential digestion and ligation reactions. Modules and the pART7 vector were digested in parallel with the appropriate restriction enzyme/s in the following quantities: 1.5 μ g plasmid DNA, 0.5 μ L restriction enzyme/s, 4 μ L CutSmart buffer in a 30 μ L reaction. Where plasmids contained only one restriction site, reaction mixtures were incubated at 37 °C for 4 hours. Should multiple restriction sites be present, a partial digest was carried out with 1, 1:5 and 1:25 restriction enzyme incubated at room temperature for 10 minutes then immediately run on an agarose gel. Target fragments were excised, eluted and ligated together. The ratio of insert:vector included in the ligation was calculated according to the following

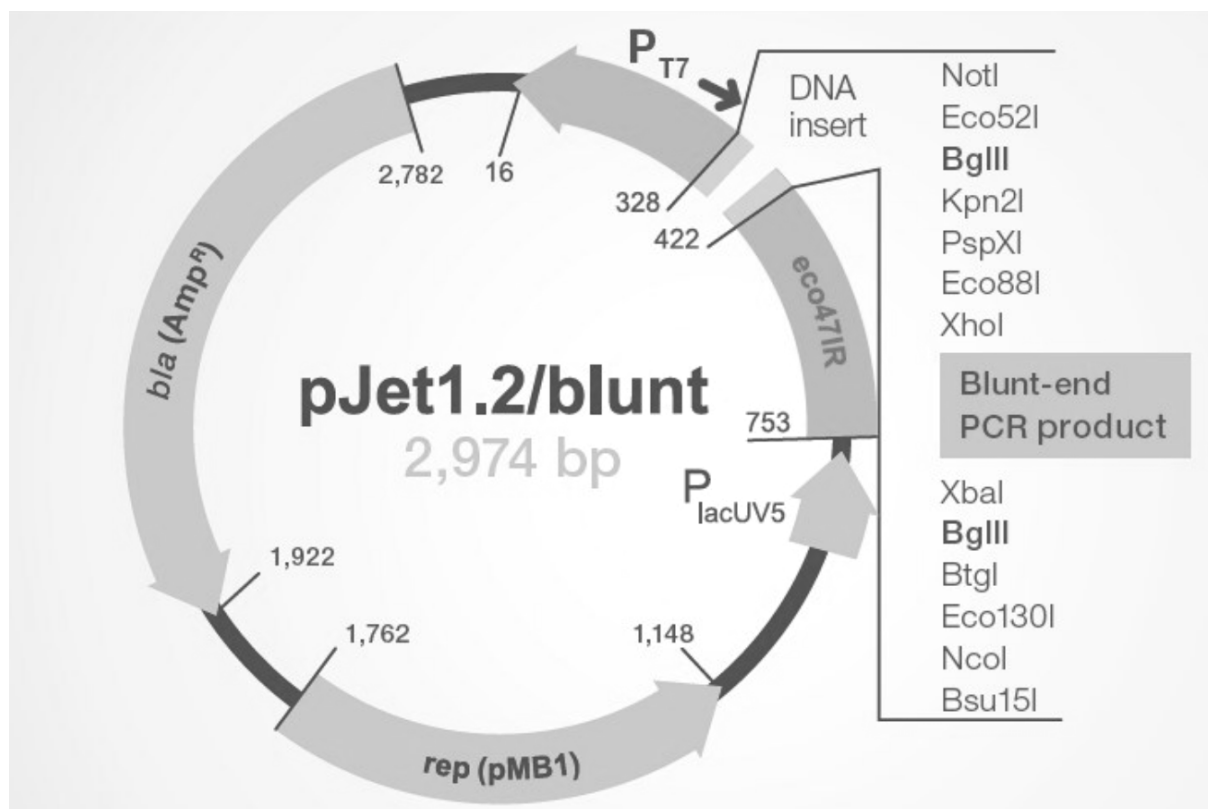


Figure 2.1: pJET1.2/blunt cloning vector. The pJET1.2/blunt cloning vector is a high-efficiency plasmid vector designed for cloning PCR products. Derived from a pUC origin of replication (pMB1), pJET1.2 is a high-copy-number plasmid, allowing for the recovery of large quantities of recombinant DNA. The vector contains an *eco47IR* gene, which is lethal to *E. coli* when expressed. Successful insertion of the DNA fragment into the multiple cloning site disrupts the *eco47IR* gene, allowing only recombinant colonies to survive. *BglII* sites flanking the insertion region allow for diagnostic digestion. In addition, the vector contains a β -lactamase (*bla*) gene that confers ampicillin resistance (Amp^R, allowing for selection of recombinant cells. Vector map adapted from Thermofisher Scientific (thermofisher.com/tstopicks)

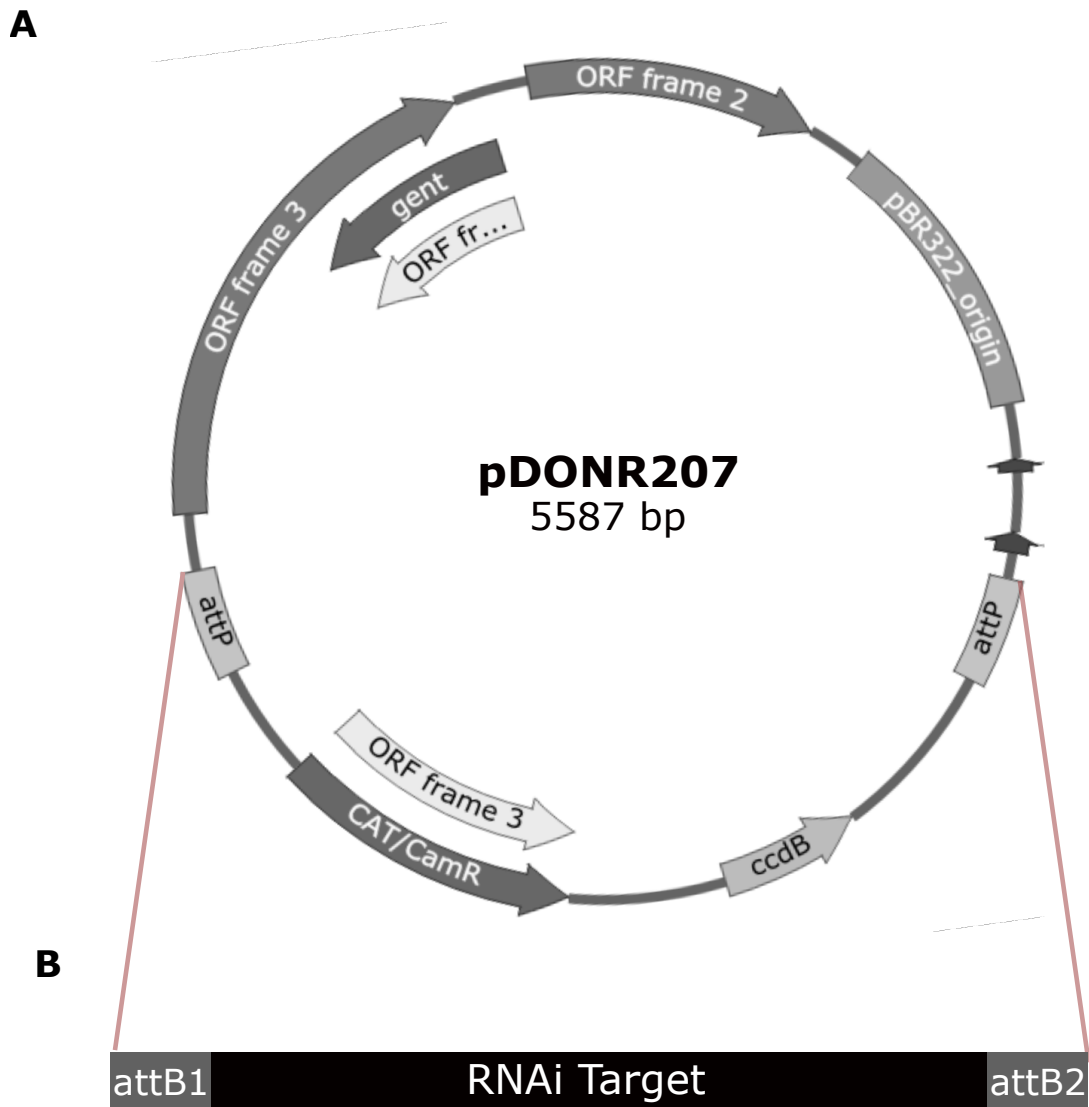


Figure 2.2: pDONR207 Gateway entry vector. (A) The pDONR207 Gateway vector is constructed specifically for the creation of entry clones in the Gateway cloning system. The vector contains a pUC origin of replication and a gentamycin resistance gene (*gent*), allowing selection of transformed bacterial colonies. The multiple cloning site is flanked by *attP* recombination sites, facilitating site specific recombination with *attB* sites on DNA fragments, mediated by BP clonase. A *ccdB* gene located between the *attP* sites enables negative selection against non-recombinant clones. (B) The RNAi target is flanked by *attB1* and *attB2* sites, facilitating recombination with the *attP* sites in pDONR207 to form *attL* restriction sites.



Figure 2.3: pANDA vector map. (A) The pANDA vector was designed specifically for the expression of RNA interference (RNAi) hairpins in rice. Expression is driven by the maize Ubiquitin promoter. Two pairs of attR sites either side of a non-functional GUS linker facilitate the insertion of inverted repeat target regions. When expressed *in vivo* these will form a hairpin leading to downregulation of expression. The plasmid additionally confers bacterial resistance to Kanamycin and plant resistance to Hygromycin. The full sequence of the backbone is not known (grey dotted line). (B) Cassette that results from recombination of RNAi targets plus attL sites with the attR sites in pANDA.

equations:

For inserts <3 Kb:

$$\text{Insert (Z ng)} = \frac{50 \text{ ng vector DNA} \times X \text{ Kb insert size}}{Y \text{ Kb vector size}} \times 3$$

For inserts >3 Kb:

$$\text{Insert (Z ng)} = \frac{50 \text{ ng vector DNA} \times X \text{ Kb insert size}}{Y \text{ Kb vector size}} \times 5$$

Ligation mixes included the calculated ratio of insert and vector, with 1 µL 10X DNA ligase buffer and 1 µL T4 DNA, made up to a volume of 10 µL with dH₂O. Ligation mixes were incubated at 37 °C for 1-2 hours. To quantify the success of a ligation, a Vector Only control (ligation mix without insert) and a No Ligase control (Ligation mix without T4 DNA ligase) were incubated alongside. After incubation, ligation mixes were transformed into DH5α *E. coli*.

2.4.6 Transformation of *E. coli* DH5α

DH5α competent *E. coli* cells (SupE44, Δ*lac*U169 (Φ80/*lacZ*ΔM15), *hsdR*17, *recA*1, *endA*1, *gyrA*96, *thi*-1, *relA*1) were used in transformation. Whole ligation mixtures were added to 50 µL *E. coli* DH5α cells and incubated on ice for 30 minutes. The mixture was then heat shocked at 45 °C for 45 s, after which cells were immediately transferred to ice for 2 minutes. 500 µL Super Optimal Broth with Catabolite repression (SOC) media was then added to the heat-shocked cells and the mixture incubated at 37 °C whilst shaking for 2 hours. The incubated mixture was spread on LB medium (Appendix 6.5) with 1 % agar and the appropriate antibiotic for selection of transformed cells in 90 mm petri dishes. Plates were incubated at 37 °C overnight.

A

Figure 2.4: 35S::OCS-pART7 donor vector for complementation. (A) The 35S::OCS-pART7 consists of a 35S promoter and an OCS terminator with multiple restriction sites in between. *NotI* restriction sites flank the promoter and terminator, facilitating release of the cassette for ligation into a plant transformation vector. pART7 additionally contains an Ampicillin resistance marker (Amp^R). (B) The untagged *SPL* complementation cassette will replace the 35S::OCS cassette through multiple restriction cloning steps. (C) The *SPL* cassette tagged with *mCHERRY* will replace the 35S::OCS cassette through multiple restriction cloning steps.

2.4.7 Inoculation of Bacterial Liquid Cultures

Individual colonies that grew overnight on LB agar + selection were scooped and used to inoculate universal tubes containing 10 mL of LB with the appropriate antibiotic. Inoculated cultures were incubated at 37 °C whilst shaking overnight.

2.4.8 Isolation of plasmid DNA from *E. coli*

Plasmid DNA was extracted using the ISOLATE II Plasmid Mini Kit (Bioline). 7.5 mL bacterial overnight culture was used for extraction according to the manufacturer's protocol for high copy plasmids. Briefly, 7.5 mL bacterial culture was pelleted by centrifugation at 11000 RCF for 30 s. The pellet was resuspended in resuspension buffer P1 containing RNaseA. Bacteria were lysed by the addition of Lysis buffer P2 and incubation at room temperature for 5 minutes. Neutralisation buffer P3 was then added to terminate the lysis reaction and precipitate DNA. Samples were centrifuged at 11000 RCF and the supernatants transferred to a spin column. Plasmid DNA bound to the silica membrane of the spin column was washed with buffers PW1 and PW2, then eluted in Tris-EDTA (TE) buffer.

2.4.9 Digestion

To confirm that the correct plasmid had been transformed into *E. coli*, diagnostic digests were performed on the miniprep DNA. Restriction sites within the targets of interest were selected for restriction digest, and the resulting patterns analysed by agarose gel electrophoresis.

2.4.10 Sequencing

To further verify that plasmid DNA was correct, sequencing was performed using primers upstream of the the 5' and downstream of the 3' ends of the insert. For DNA inserted into pJET_{1,2}, standard primers pJET-F and pJET-R were used. When the insert

was larger than 1 Kb, internal primers were designed to ensure the entire sequence was correct. All primers used for sequencing are listed in Table 2.5. DNA was sequenced using the appropriate primers by Source Bioscience (Cambridge, UK).

Table 2.5: Primers used to sequence plasmids.

Primer	Sequence
AtSPL-3	CTGACCTGTCTTCGATCC
AtSPL-4	GAGGTGTGAAGGAGTACG
CrSPL_A-3	GACTCCATCTACCTCG
CrSPL_A-4	GTTAGGCTTCACTTCG
CrSPL_B-3	GGATTGAGTGAGTTCTGC
CrSPL_B-4	CTCCTCCACATCTGTTG
AtSPL_3'pro-2-F	CAGGAGGTGTGAAGGAGTACG
AtSPL_5'pro-F2	CACCTACTCTGCCTCC
AtSPL_5'pro-F3	CTCGCTAACCGCAGAGTG
AtSPL_5'pro-F4	CTACTCTACTCCGTAGCTC
AtSPL_5'pro-F5	CCACGATCCAATGCATGG
AtSPL_5'pro-R2	CCTCCTTTTGGACCACC
AtSPL_5'pro-R3	GCAAAGGACGAAAGAGG
AtSPL_5'pro-R4	CCAGATGAGATGTTGTACATC
AtSPL_5'pro-R5	CCTACCATGGAGACAGC
AtSPL_3'UTR-3	GTTTCGGCTAAAATTAGTTAG
AtSPL_3'UTR-4	GAAAACGAAGCTGTACATCC
pJETF	CGACTCACTATAGGGAGAGCGGC
pJETR	AAGAACATCGATTTTCCATGGCAG

2.5 Genetic Transformation of *C. richardii*

Callus induced in the *C. richardii* sporophyte was transformed by bombardment with microparticles coated in the plasmid of interest.

2.5.1 Callogenesis

Callus was induced in the *C. richardii* sporophyte 14 days after fertilisation in sterile culture. Sporophytes with a visible shoot and root were transplanted to 90 mm tissue culture plates with 1X C-fern medium (Appendix 6.5) plus 5 μ M 6-Benzylaminopurine (BAP), spaced 1 cm apart. Plates were incubated under 28 °C LD conditions, with 2 mL sterile dH₂O added every 2-3 days as required. Approximately 10-14 days after transplantation, callus tissue was visible at the shoot apex in place of new fronds. Callus was excised using sterile watchmakers forceps and transplanted to 1X MS (Sigma M0404), 2 % sucrose, 0.7 % agar plates (pH 5.8) (Appendix 6.5) with 5 μ M BAP (MS-BAP medium). Plates were sealed with Parafilm™ and incubated at 28 °C LD conditions. Callus was sub-cultured by transfer to fresh MS-BAP every two weeks for a maximum of 1 month.

3 days before transformation, callus was transferred to 50mm 1X MS 2 % sucrose 0.7 % agar plates (pH 5.8) with 5 μ M Kinetin (MS-Kinetin). Callus was densely packed in a 30 mm diameter circle at the centre of the plate to form a target. Plates were returned to 28 °C LD conditions until bombardment, to promote growth of the target.

2.5.2 Large Scale preparation of plasmid DNA for transformation

A minimum of 1 μ g μ L⁻¹ of DNA is required for particle coating. Therefore, 2.5 mL bacterial liquid cultures of RNAi-pANDA plasmids that had been verified by digestion and sequencing were used to inoculate 500 mL LB. Inoculated cultures were incubated at 37 °C whilst shaking overnight. The resulting large culture was

maxi-prepped and precipitated to maximise plasmid DNA concentration.

500 mL bacterial overnight culture was maxiprepmed using the GenElute™ Plasmid Maxiprep Kit (Sigma) following the manufacturer's instructions. Briefly, the 500 mL overnight culture was pelleted by centrifugation for 10 minutes at 4000 RCF. Pelleted cells were resuspended in resuspension buffer, lysed with lysis buffer and DNA precipitated with neutralisation buffer. The cleared lysate was loaded onto a prepared DNA binding column and centrifuged at 4000 RCF for 2 minutes. Bound DNA was washed with wash buffer and spun at 4000 RCF for 2 minutes. Elution buffer heated to 65 °C was added to the bound DNA and incubated at room temperature for 10 minutes, then centrifuged at 4000 RCF.

3 M Sodium Acetate solution pH 5.2 was added to eluted DNA at a volume ratio of 1:9. 2.5 volumes of ice cold 100 % ethanol was then added and the solution was incubated on ice for 30 minutes to precipitate the DNA. The solution was centrifuged at 13000 RCF for 20 minutes to recover precipitated DNA in the pellet. 70 % ethanol was added to the pellet, ensuring the entire pellet and sides of tube were covered, and centrifuged for 1 minute at 13000 RCF. Ethanol was then removed and the pellet resuspended in TE buffer.

2.5.3 Preparation of gold microparticles

1.0 µM Gold microparticles (Microcarriers, Bio-Rad) were used in bombardment. 20-40 mg were sufficient for one bombardment round (4 firings). The microparticles were incubated in 1 mL 70 % ethanol for 15 minutes at room temperature. Sterilised microparticles were then washed four times with sterile dH₂O before being suspended in 50 % glycerol to achieve a concentration of 60 mg/ml, assuming no loss during sterilisation.

Following sterilisation, microparticles were coated with the DNA construct of interest. Each bombardment round was accompanied with a No DNA control, therefore 25 µL microparticle-glycerol suspension was aliquoted into two 1.5 mL

microcentrifuge tubes. Tubes were continuously vortexed whilst 5 μ L plasmid DNA or dH₂O (No DNA control), 50 μ L filter-sterilised 2.5 M CaCl₂ and 20 μ L filter-sterilised 0.1 M Spermidine were added in quick succession. Tubes were then vortexed for a further 3 minutes, then left to stand for 1 minute before washing twice with 140 μ L 70 % ethanol. Finally, microparticles were suspended in 50 μ L 100 % ethanol ready for bombardment.

2.5.4 Biolistic Bombardment of *C. richardii* callus

Callus tissue was bombarded using a PDS-1000/HeTM system (Bio-rad) in an ESCO Laminar Flow Cabinet. Set-up and operation was performed according to the manufacturer's instructions. The biolistic gun, microcarrier launch assembly and rupture disc retaining cap were sterilised in 100 % ethanol along with two pairs of bayonet forceps (Sigma) and watchmaker's forceps, Biolistic Macrocarrier holders (Bio-Rad), Biolistic Macrocarrier discs (Bio-Rad) and stopping screens. Biolistic rupture disks (P/N 9202966 "C", 650 PSI) were incubated in 100 % Isopropanol. Macrocarrier discs were inserted into Macrocarrier holders using bayonet forceps, pressing down until the disk lay flat. 15 μ L microparticles suspended in 100 % ethanol was then spread evenly over the centre of the Macrocarrier disk, and the ethanol allowed to evaporate.

At the start of each new bombardment, a sterilised Biolistic rupture disk was placed into the Rupture Disk Retaining Cap using watchmaker's forceps to ensure the disk lay flat. The cap was then screwed into the gun and tightened using a torque wrench. Next, a sterilised stopping screen was placed centrally over the hole in the firing stage, and a loaded Macrocarrier holder was placed over the top, microparticles facing down. The Microcarrier launch was assembled and positioned directly below the rupture disc retaining cap. Callus was positioned on the target plate shelf at a distance of 6 cm beneath the Microcarrier launch assembly. Once all equipment was in place, the PDS-1000/HeTM bombardment chamber was evacuated to 28 inches of

mercury (inHg). The fire switch was then held ON continuously to achieve a helium pressure of approximately 900 psi until the rupture disk burst.

Each fresh plate of callus was similarly bombarded, replacing the Macrocarrier holder and disk, and rupture disk each time. The stopping screen was re-used for up to 4 bombardments with the same construct. The no-DNA control was always bombarded first at the start of a new bombardment round with a new construct. Following bombardment, callus plates were immediately sealed with Parafilm and incubated under 28 °C LD conditions for 3 days before transfer to selection.

2.5.5 Regeneration of T₀ transgenic sporophytes

3 days after bombardment, callus was transferred to 90 mm MS-Kinetin plus 20 µg mL⁻¹ hygromycin-B (Hyg), spaced 1 cm apart. Plates were sealed with Parafilm and returned to 28 °C LD conditions for two weeks. Following this, callus was subjected to continuous antibiotic selection for a further 6 weeks by transfer to new MS plus 20 µg mL⁻¹ Hyg without Kinetin every two weeks. After 8 weeks of continuous selection, transformed callus should have regenerated sporophyte shoots and roots. Regenerated sporophytes were transplanted to 1X C-fern medium for a further two weeks to promote sporophyte regeneration before transplantation to soil. Sporophytes on soil were grown under 28 °C LD conditions at 85 % humidity. Approximately 4 months after transplantation to soil, mature reproductive fronds were harvested to collect T₁ spores.

2.5.6 Selection of T₁ transgenic lines

Due to the chimeric nature of regenerated T₀ sporophytes observed by Plackett et al., 2015, T₁ lines were screened for antibiotic resistance. CrWT and T₁ spores were sterilised, imbibed, sown and germinated on C-fern nutrient media with 20 µg mL⁻¹ hygromycin as in Section 2.1.1. The presence of hygromycin in the media slows the development of the gametophytes. At approximately 14-21 days, T₁ hermaphrodite

gametophytes could be distinguished, and control *CrWT* spores should not have survived selection. T_1 hermaphrodite gametophytes were isolated using two pairs of watchmakers forceps under a bright field microscope, ensuring no male gametophytes were stuck to the hermaphrodite, and transferred to individual wells of a 24 well tissue culture plate containing C-fern nutrient medium without antibiotic. Plates were returned to 28 °C LD conditions until gametophytes were mature.

Gametophytes were self-fertilised by the addition of sterile dH_2O to each individual well, to ensure homozygosity of the transgene. 14 days later, established homozygous T_1 sporophytes were returned to 90 mm C-fern nutrient media plates with $20\text{ }\mu\text{g mL}^{-1}$ hygromycin, sealed with micropore tape and returned to 28 °C LD conditions. 2 mL sterile dH_2O was added every 2-3 days. After 14 days on selection, surviving transgenic sporophytes were transplanted to soil as in Section 2.1.2. Due to space restrictions, a maximum of 4 sporophytes per transgenic line - defined as the progeny of a T_0 sporophyte regenerated from an independent callus - were transplanted to soil.

2.5.7 Phenotypic analysis of T_1 Transgenics

Mature reproductive T_1 sporophytes were phenotyped at 130 days post-fertilisation (Figure 2.5A). One frond per T_1 plant was selected for further analysis, equating to 4 replicates per line. Selected fronds were of a similar developmental age, with mature sporangia but no evidence of senescence. The selected frond was removed from the plant at the base, and its height measured from base to tip (Figure 2.5B). The number of branches emerging from the main stem, and the number of pinnae emerging from the branches, was counted (Figure 2.5B). 3 pinnae per frond were randomly selected for further analysis. Pinnae were excised from the plant and examined under a dissecting GX Microscope (GT Vision Ltd ultraZOOM-2) (Figure 2.5C). Pinnae were photographed, then unfurled using watchmakers forceps to reveal the sporangia inside (Figure 2.5D). Unfurled pinnae were photographed again and the number of sporangia counted. To allow comparison between pinnae and lines, the number of sporangia was divided by

the length of the pinnae.

Statistical analysis was performed in R version 4.2.2 (R Core Team, [2024](#)). Frond length, number of branches and number of pinnae were analysed by Analysis of Variance (ANOVA) followed by post-hoc pairwise comparisons with Tukey's Honest Significant Difference (HSD) adjusted for multiple comparisons by Generalised Linear Hypothesis Testing (GLHT). Number of sporangia per pinnae was analysed with a non-parametric Kruskal-Wallis test followed by a Wilcoxon Rank Sum Test adjusted for multiple comparisons with Bonferroni. Results were analysed and visualised using the R packages ggplot2 (Wickham, [2016](#)), dplyr (Wickham et al., [2023](#)) and multcompView (Graves et al., [2024](#)).

2.5.8 Expression analysis of T₁ Transgenics

1-3 reproductive fiddleheads were harvested at the time of phenotyping, and RNA extracted following the protocol outlined in Section [2.2.2](#). cDNA was synthesised according to the protocol in Section [2.2.3](#) and qPCR performed (Section [2.3.2](#)) to determine whether *CrSPLB* expression was downregulated in the T₁ transgenics compared to *CrWT*.

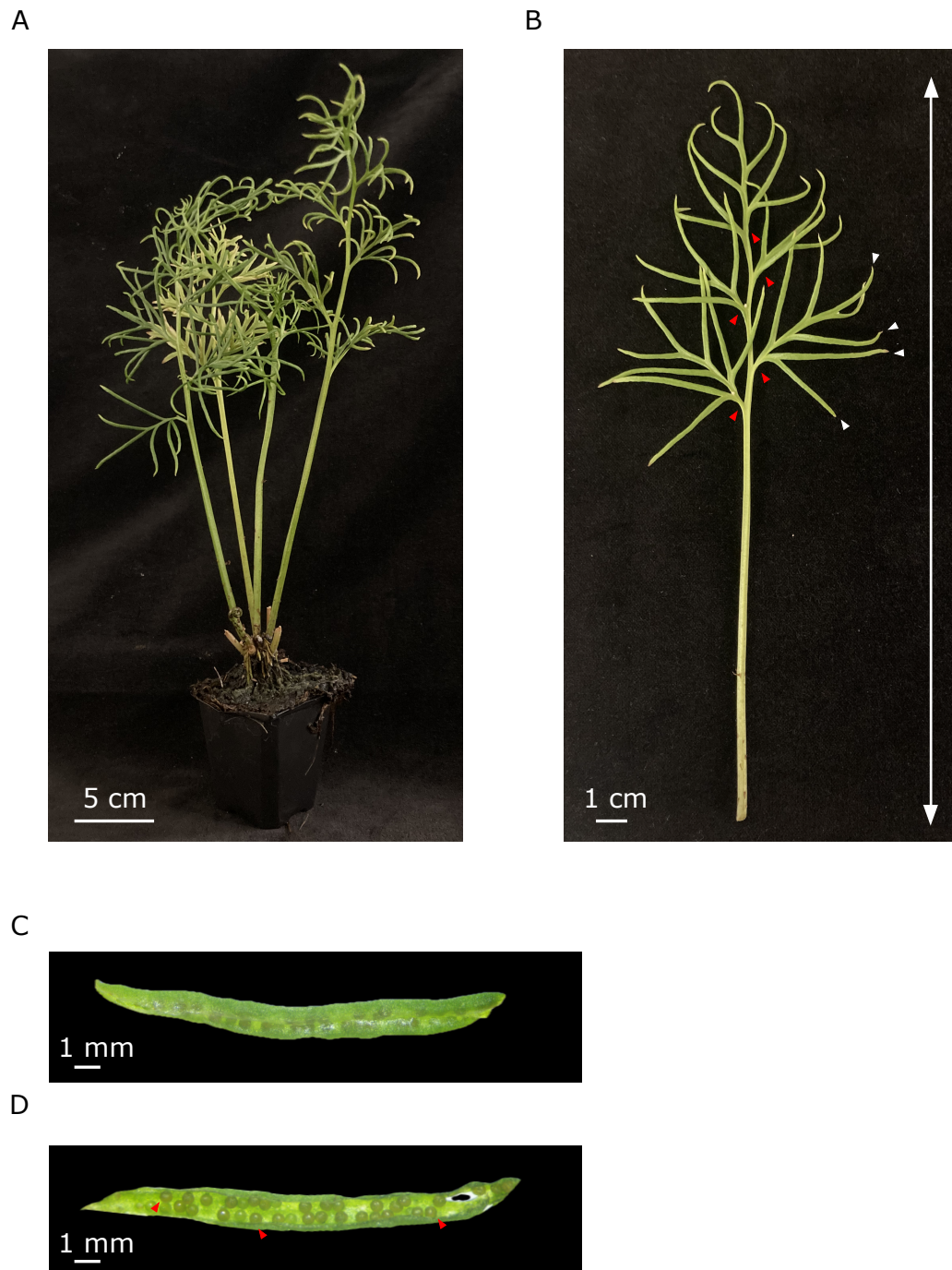


Figure 2.5: Phenotyping measurements taken from *C. richardii* T₁ RNAi lines. (A) 130 day old sporophyte. Phenotyping was carried out on sporophytes of approximately this age. (B) One frond per transgenic plant was measured from tip to base (white arrow). The number of branches emerging from the main stem (red arrowheads) and the total number of pinnae (white arrowheads) was counted. Three pinnae were selected at random from the selected frond. (C) Representative single pinna. (D) Pinna tissue was pulled back to enable counting of the sporangia (red arrowheads) inside.

2.6 Phylogenetic analysis of embryophyte SPEAR proteins

2.6.1 SPEAR sequence acquisition

AtSPL, AtSPEAR1, AtSPEAR2, AtSPEAR3 and AtSPEAR4 peptide sequences from the *A. thaliana* genome (TAIR10) were obtained from Phytozome 13 (Goodstein et al., 2012). With the exception of *Quercus robur*, all angiosperm SPEAR homolog sequences were obtained from Phytozome 13 via a BLASTp search with AtSPEAR sequences as query. Sources and versions of included species are detailed in table 2.6. SPEAR homologs of *C. richardii*, *Selaginella moellendorffii*, *Diphasiastrum complanatum*, *Physcomitrium patens*, *Sphagnum fallax*, *Ceratodon purpureus* and *Marchantia polymorpha* were likewise obtained from Phytozome 13. All other sequences were obtained from specific databases and publications as listed in table 2.6. Due to the inconsistency of BLAST programs in these databases, individual genomes and/or transcriptomes were manually searched for the presence of an SPL motif using the regular expression G[LMP]GV[AE][QE]LE[KR] in BBEdit.

All putative SPEAR sequences identified by BLAST or keyword search were manually checked for the presence of an SPL domain in the C-terminal region and an EAR domain at the N terminus. All algae, bryophyte, lycophyte, fern and gymnosperm species with a publicly available genome or transcriptome were searched via BLAST (Camacho et al., 2009), OrthoFinder (Emms & Kelly, 2019) and by keyword search with the regular expression G[LMP]GV[AE][QE]LE[KR] for potential SPEAR homologs. Those that yielded only a partial or ambiguous sequence were subsequently removed from further analysis.

Table 2.6: Species, versions and sources of datasets searched for SPEAR homologs and included in subsequent alignments and phylogenies.

Species	Version	Source
<i>Amborella trichopoda</i>	1.0	Phytozome
<i>Nymphaea colorata</i>	1.2	Phytozome
<i>Acorus americanus</i>	1.1	Phytozome
<i>Spirodela polyrhiza</i>	2.0	Phytozome
<i>Oryza sativa</i>	7.0	Phytozome
<i>Asparagus officinalis</i>	1.1	Phytozome
<i>Cinnamomum kanehirae</i>	3.0	Phytozome
<i>Aquilegia coerulea</i>	3.1	Phytozome
<i>Beta vulgaris</i>	EL10_1.0	Phytozome
<i>Solanum lycopersicum</i>	ITAG3.2	Phytozome
<i>Kalanchoe fedtschenkoi</i>	1.1	Phytozome
<i>Arabidopsis thaliana</i>	TAIR10	Phytozome
<i>Populus trichocarpa</i>	4.1	Phytozome
<i>Vitis vinifera</i>	2.1	Phytozome
<i>Quercus robur</i>	2_2N	Plomion et al. (2018)
<i>Thuja plicata</i>	3.1	Phytozome
<i>Picea abies</i>	1.0	PlantGenIE.org
<i>Pinus taeda</i>	2.01	PineRefSeq
<i>Ginkgo biloba</i>	PRJNA307642	Guan et al. (2016)
<i>Gnetum montanum</i>	2018	Wan et al. (2018)
<i>Torreya grandis</i>	2022	Sun (2022)
<i>Pseudotsuga menziesii</i>	1.0	PineRefSeq
<i>Ceratopteris richardii</i>	2.1	Phytozome
<i>Salvinia cucullata</i>	1.2	Fernbase
<i>Azolla filiculoides</i>	1.1	Fernbase
<i>Lygodium japonicum</i>	2014	Aya et al. (2014)
<i>Adiantum capillus-veneris</i>	2024	Fernbase
<i>Alsophila spinulosa</i>	3.1	Fernbase
<i>Marsilea vestita</i>	3.0	Fernbase
<i>Selaginella moellendorffii</i>	1.0	Phytozome
<i>Lycopodium clavatum</i>	2022	Yu (2022)
<i>Diphasiastrum complanatum</i>	3.1	Phytozome
<i>Selaginella cf. pallescens</i>	2019	1KP
<i>Huperzia asiatica</i>	2023	Li (2023)
<i>Anthoceros agrestis</i>	PRJNA574453	Li et al. (2020)
<i>Anthoceros punctatus</i>	PRJNA574424	Li et al. (2020)
<i>Physcomitrium patens</i>	3.3	Phytozome
<i>Ceratodon purpureus</i>	1.1	Phytozome
<i>Sphagnum fallax</i>	1.1	Phytozome
<i>Marchantia polymorpha</i>	3.1	Phytozome

2.6.2 Comparison of similarity and identity within and between *C. richardii* and *A. thaliana* SPEAR protein sequences

Pairwise global alignments of the full SPEAR sequences from *A. thaliana* and *C. richardii* were conducted with Needle in the EMBL-EBI job dispatcher (Madeira et al., 2024). The substitution scoring matrix PAM250, with a gap open penalty of 10 and a gap extend penalty of 0.1 was used to allow fewer but longer gaps. The default gap end open and extend values of 10 and 0.5, respectively, were used as all SPEAR proteins should end with a conserved EAR domain. It was not necessary to align the conserved SPL Motif of the *A. thaliana* and *C. richardii* SPEARs. A 19 amino acid sequence, 6 residues upstream and 4 residues downstream of the conserved motif GXGVXXLEX was extracted from each SPEAR and pairwise identity and similarity calculated in the online Sequence Manipulation Suite (Stothard, 2000).

2.6.3 Multiple sequence alignment and model selection

SPEAR sequences were aligned by MAFFT L-INS-i (`mafft -reorder -leavegappyregion -maxiterate 2 -retree 1 -localpair`) (Katoh et al., 2005). The alignment was visualised in Jalview v2.11.4.1 (Waterhouse et al., 2009). Sequences that did not align in the conserved SPL motif and/or EAR domain regions were removed. In addition, sequences displaying very close homology, particularly moss and pine sequences, were removed except for a single representative. The remaining sequences were realigned by MAFFT L-INS-i and no further curation was made. Consensus sequence logos of SPL and EAR domains were generated with ggseqlogo (Wagih, 2017).

2.6.4 Phylogenetic inference

Maximum Likelihood (ML) and Bayesian Inference (BI) methods were used to conduct phylogenetic analyses. The ML phylogenetic tree was constructed using the

Randomized Axelerated Maximum Likelihood Next Generation (RAXML-NG v. 1.0.0) inference method (Kozlov et al., 2019). ML analysis was conducted using the Jones-Taylor-Thornton substitution matrix with invariant sites, gamma-distributed rate variation across sites with four categories, and median-based rate categories (JTT+I+G4m), selected by ModelTest-NG (Darriba et al., 2020; Flouri et al., 2014). 500 starting trees were specified, 250 random and 250 maximum parsimony, with 1000 bootstrap replicates.

The BI analyses were performed using MrBayes v3.2.7a with iterative parameter tuning to achieve convergence (Ronquist et al., 2012). A mixed substitution model with an inverse gamma distribution for rate variation among sites was specified (`lset nst=mixed rates=invgamma`), with a mixed amino acid substitution model (`prset aamodelpr=mixed`). Two independent runs of four Markov Chain Monte Carlo chains each were performed, with the total generations extended iteratively up to 15 million to achieve acceptable convergence. The chain temperature was progressively lowered to 0.01 to improve mixing among chains (`mcmc nruns=2 nchains=4 ngen=15000000 temp=0.01`), ensuring that the average standard deviation of split frequencies fell below 0.1. Parameter and tree samples were taken every 100 generations (`samplefreq=100`). Posterior distributions were summarized using a 30% burn-in for both parameter (`sump burnin=30`) and tree (`sumt burnin=30`) analyses. The consensus tree was constructed using the “allcompat” setting to include all compatible splits (`sumt cotype=allcompat`), reflecting a comprehensive summary of the topological results.

Trees were visualised in ITOL, and further edited in Inkscape (Letunic & Bork, 2021).

2.7 Processing and Analysis of RNA sequencing data

2.7.1 Sample and library preparation

Samples were collected and RNA extracted by Dr. Andrew Plackett according to the protocol outlined in Section [2.2.2](#). Three biological replicates per tissue type were prepared. mRNA library preparation and sequencing was performed by Novogene (Novogene Bioinformatics Technology co. Ltd., Beijing, China), using their standard RNA sequencing pipeline. Sample quality control verified that the samples met RNAseq requirements. Libraries were then prepared by Poly(A) capture to enrich mRNA, fragmentation, reverse transcription into cDNA and adapter ligation. Following a second quality assessment, libraries were sequenced using an Illumina platform, strategy PE150. Raw data was transformed into sequence reads via traditional base calling, producing FASTQ format files with sequence information and associated quality.

2.7.2 Raw sequence quality control

Sequence quality was assessed by fastqc (Andrews, [2010](#)) and summarised by multiqc (Ewels et al., [2016](#)). Sequences were then trimmed to remove adapter sequences, and filtered to remove low quality sequences with fastp (Chen et al., [2018](#)). Quality assessment with fastqc and multiqc was repeated to confirm that trimming and filtering had improved overall quality.

Sequences were then processed using two alternative methods: read mapping to the published reference transcriptome with Salmon, and read mapping to the published reference genome with HISAT2.

2.7.3 Mapping reads to the genome with HiSAT2

RNA-Seq reads were mapped to the *C. richardii* reference genome v2.1 (Marchant et al., 2022) using HISAT2 v2.2.1 (Hierarchical Indexing for Spliced Alignment of Transcripts 2) (Kim et al., 2019), a fast and efficient splice-aware aligner. The reference genome was pre-indexed using HISAT2's `hisat2-build` function to facilitate rapid alignment. Paired-end reads were then aligned to the indexed genome using the `hisat2` command with default parameters, enabling accurate detection of splice junctions. The resulting sam files were converted to bam, sorted and indexed with samtools v1.10 (Danecek et al., 2021). At this stage, paired reads that failed to map to the genome (unmapped reads) were extracted from the bam files using samtools and converted to fastq format using BEDtools v2.29.2 for additional analysis (Quinlan & Hall, 2010).

Raw gene counts were generated from the sorted and indexed BAM files using HTSeq v0.13.5 (Anders et al., 2014), a Python-based tool for feature counting. Gene-level quantification was performed using the `htseq-count` script in unstranded mode (`-stranded=no`) to account for the non-stranded library preparation. Reads were assigned to features annotated as `gene` in the *C. richardii* v2.0 reference annotation (GFF3 format) using the `Name` attribute as the gene identifier. The union mode (`-mode=union`) with `non-unique=None` was applied for conservative assignment of reads overlapping multiple features, i.e. if paired reads mapped to different exons, they were removed from the alignment. Output files containing raw counts for each gene were used for downstream differential expression analysis.

2.7.4 Mapping reads to the transcriptome with Salmon

Transcript quantification was performed using Salmon v1.10.2 (Patro et al., 2017), a fast and accurate tool for RNA-Seq quantification. The *C. richardii* v2.1 transcriptome and v2.1 genome were downloaded from Phytozome and used to build a decoy-aware

index. Decoy sequences were extracted from the genome using `grep`, and the transcriptome and genome were concatenated into a “gentrome” file with transcript sequences listed first. The Salmon index was then created using the `salmon index` command, incorporating the gentrome and decoy sequences.

Paired-end reads were quantified using the `salmon quant` command with automatic library type inference (`-l A`). Bias corrections for GC content and sequence bias were applied using the `-gcBias` and `-seqBias` options, while the `-validateMappings` option ensured accurate read mapping. Unmapped reads were saved for downstream analysis using the `-writeUnmappedNames` option. Output quantification files were generated for each sample and stored in the designated directory.

Unmapped reads were subsequently extracted using a custom script, which retrieved the names and sequences of unmapped and decoy-mapped reads from Salmon’s output. These reads were further analysed to investigate sequences that failed to map to the reference transcriptome.

2.7.5 *de novo* assembly of unmapped reads

Unmapped reads extracted from mapping with HISAT2 (Kim et al., 2019), and unmapped and decoy reads extracted from mapping with Salmon (Patro et al., 2017) were processed and assembled *de novo* to identify novel transcripts and sequences not present in the reference genome or transcriptome. Initial quality control of the unmapped reads was performed using FastQC (Andrews, 2010) and summarized with MultiQC (Ewels et al., 2016). Errors in RNA-Seq reads were corrected using rCorrector v1.0.5 (Song & Florea, 2015), and unfixable reads were removed with a Python3-compatible script from the Harvard Informatics Transcriptome Assembly Tools repository. Ribosomal RNA (rRNA) contamination was filtered out using Bowtie2 v2.5.1 (Langmead & Salzberg, 2012; Langmead et al., 2018), with custom rRNA databases constructed from Silva SSUParc and LSUParc FASTA files (Quast et al., 2013). After preprocessing, FastQC was rerun to confirm the quality of the

cleaned reads.

De novo transcriptome assembly was performed using Trinity v2.15.1 with default parameters (Grabherr et al., 2011; Haas et al., 2013). Cleaned reads were specified using a sample table text file provided via Trinity's `-samples file` option. Post-processing steps included redundancy removal with CD-HIT v4.8.1 (Li & Godzik, 2006) and computation of assembly statistics using the Trinity `TrinityStats.pl` script (Grabherr et al., 2011; Haas et al., 2013). Read representation was assessed by mapping reads back to the assembly using Bowtie2 v2.5.1, and a gene-transcript map was created for the non-redundant assembly. Gene expression matrices were generated using Kallisto and assembly quality was further evaluated by calculating the ExN50 metric within the `TrinityStats.pl` script. This comprehensive workflow ensured a robust *de novo* assembly of unmapped reads for further analyses.

2.7.6 Read normalisation and filtering

Read-level counts resulting from alignment to the genome with HISAT2 and extraction with HTseq were selected for downstream analysis in R. Gene expression levels were normalised for both sequencing depth and gene length by calculation of FPKM (Fragments Per Kilobase of transcript per Million mapped reads). Gene sizes were determined by summing the lengths of exonic regions for each annotated gene. The total number of reads mapped for each sample was calculated to account for sequencing depth. FPKM values for each gene were then computed using the formula:

$$\text{FPKM} = \frac{\text{Raw Counts}}{\left(\frac{\text{Gene Length (bp)}}{1000} \right) \times \left(\frac{\text{Total Reads}}{10^6} \right)}$$

where gene length refers to the cumulative length of the exonic regions for each gene, measured in kilobases, and total reads represent the number of mapped reads for each sample, scaled to millions.

Normalised counts were filtered to only retain genes with expression values ≥ 0.5

FPKM in at least three samples. The filtered FPKM counts were then log-transformed using a $\log_2(FPKM + 1)$ transformation to stabilize variance and normalize the data for subsequent analyses.

2.7.7 Sample clustering and dimensionality reduction

To assess the relationships and potential outliers among the samples, hierarchical clustering and a sample distance matrix were calculated using R's built-in functions for distance calculation (`dist`) and hierarchical clustering (`hclust`). The dendrogram resulting from clustering was visualized as a sample tree. The sample distance matrix was directly visualized as a heatmap using the `pheatmap` package (Kolde, 2018). This dual approach enabled both clustering visualization and confirmation of relative sample distances.

Principal Component Analysis (PCA) was performed to explore variation amongst sample types and between replicates. PCA was computed using the base R function `prcomp` with scaling enabled. PCA results were visualized using `ggplot2`, with PC1 and PC2 plotted to capture the maximum variance (Wickham, 2016).

2.8 Weighted Gene Correlation Network Analysis

Weighted Gene Correlation Network Analysis (WGCNA) was performed to identify clusters of co-expressed genes (referred to as modules) that exhibit similar expression patterns across the sampled developmental stages of *Ceratopteris richardii* sporophytes (Zhang & Horvath, 2005). The analysis involved multiple key steps, including the selection of an appropriate soft-thresholding power, construction of a topological overlap matrix (TOM), dynamic module detection, and subsequent merging of closely related modules. All analyses were performed using the WGCNA R package (Langfelder & Horvath, 2008).

2.8.1 Choosing the soft threshold power

The $\log_2$ -transformed FPKM values were used as input for WGCNA. The `goodSamplesGenes()` function confirmed that all samples and genes were suitable for downstream analysis (Langfelder & Horvath, 2008). A range of soft-thresholding powers (1 to 30) was evaluated to identify the optimal power for constructing a scale-free network, a fundamental assumption of WGCNA. The scale-free topology model fit index and mean connectivity were calculated for each power. The chosen power was 14, which achieved a scale-free topology fit index of ≥ 0.8 , balancing scale independence and network connectivity.

2.8.2 Topological Overlap Matrix (TOM) construction

Using the selected soft-threshold power, a topological overlap matrix (TOM) was constructed using the `TOMsimilarityFromExpression()` function with default parameters (Langfelder & Horvath, 2008). The TOM quantifies the similarity between genes based on their shared connections in the network. The dissimilarity matrix ($1 - \text{TOM}$) was then calculated and used to cluster genes hierarchically.

Hierarchical clustering was performed with the Base R `hclust` function using average linkage clustering to create a dendrogram of genes. This dendrogram represents the hierarchical relationship between genes based on their expression profiles across the samples.

2.8.3 Dynamic module detection

Gene modules were identified using the dynamic tree cut method, which adaptively detects modules by cutting the branches of the gene dendrogram (Langfelder & Horvath, 2008; Langfelder et al., 2008). The `cutreeDynamic()` function was used with the parameters `deepSplit = 2` to allow the detection of more, smaller modules, and `minClusterSize = 30` to discard modules with fewer than 30 genes. Each module was assigned a unique colour for visualization purposes. Modules represent groups of genes that are likely to be co-regulated and may share common biological functions.

2.8.4 Module Eigengene calculation and merging

The expression pattern of each module was summarized by calculating its module eigengene (ME) using the `moduleEigengenes()` function, which is the first principal component of the expression data for all genes within the module (Langfelder & Horvath, 2008). MEs represent the overall expression profile of each module across all samples. The MEs were clustered using hierarchical clustering, and a dissimilarity measure (1 - correlation) between the MEs was calculated. Closely related modules were merged with the `mergeCloseModules()` function using a merge threshold of 0.25. This merging step reduces redundancy by combining modules with highly similar expression patterns.

2.8.5 Module-trait correlation analysis

To investigate the relationship between identified gene modules and tissue-specific developmental stages of the *Ceratopteris richardii* sporophyte, module-trait associations were calculated. A sample metadata table, containing categorical variables describing tissue types and developmental stages, was used to create binary trait matrices. The categorical variables were binarised into columns representing the presence or absence of specific tissue types using the `binarizeCategoricalColumns()` function from the WGCNA package (Langfelder & Horvath, 2008). The binary trait matrix was structured by developmental age to allow the visualisation of gene expression changes over time.

Pearson correlations between the MEs and the binary trait matrices were calculated using the base R function `cor()`. To assess the significance of these correlations, p-values were calculated using the `corPvalueStudent()` function, which accounts for the sample size in determining the significance of module-trait associations. To visualize the results, a correlation heatmap was generated using the `CorLevelPlot` package to cluster samples by their ME profiles and highlight significant correlations between gene modules and traits (Blighe, 2018).

Modules of interest were selected based on their correlations with specific tissues and developmental stages. For each module, the corresponding genes were extracted from the WGCNA results, and their expression profiles were analysed to explore module-specific biological patterns.

2.8.6 Hub Gene Identification

Hub genes were identified as highly connected genes within each module, based on their intramodular connectivity (`kWithin`) and correlation with the module eigengene (`kME`) (Langfelder & Horvath, 2008). Genes with an absolute `kME` value above 0.9 were classified as hub genes. The top ten hub genes for each module were

selected for further analysis. These genes represent key regulatory components of each co-expression network.

2.8.7 GO Term Enrichment Analysis

To assess the biological relevance of each module, Gene Ontology (GO) enrichment analysis was performed using an annotation file adapted from Marchant et al., [2022](#). GO terms associated with module genes were compared to a genome-wide background using a hypergeometric test. GO terms with adjusted p-values ≤ 0.05 were considered significantly enriched. The results were visualized using bar plots and dot plots, showing the top enriched GO terms and their significance.

2.8.8 Gene Expression Patterns

Gene expression patterns within each module were visualized using line plots across all samples. Separate plots were generated for positively and negatively correlated genes, with each module displayed in its own facet. All visualizations, including GO enrichment plots, module-specific expression plots, and correlation heatmaps, were generated using ggplot2 in R (Wickham, [2016](#)).

2.8.9 Network Visualisation in Cytoscape

To visualize co-expression networks, the topological overlap matrix (TOM) from the WGCNA analysis was used to define the network structure, while edges displayed pairwise gene-gene correlations. For each selected module, the TOM was exported, representing the network topology. Edges were weighted using Pearson correlation values to reflect the strength and direction of gene co-expression relationships.

To simplify the networks and focus on biologically relevant connections, a correlation threshold was applied to filter edges. For the MEivory module, only gene

pairs with a Pearson correlation ≥ 0.8 were retained, while for the MEdarkorange module, a stricter threshold of ≥ 0.9 was applied. This filtering ensured that only the strongest co-expression relationships were visualized, resulting in clearer and more interpretable networks.

The filtered data was imported into Cytoscape (v3.9.1) (Shannon et al., [2003](#)), where nodes represented individual genes, and edges displayed correlation strengths above the specified thresholds. Nodes were sized and coloured based on attributes such as module membership (kME), functional annotations, or hub gene status.

Chapter 3

When did *SPL* evolve? Tracing the evolutionary history of the *SPEAR* family across the plant kingdom

3.1 Introduction

3.1.1 SPLs and SPEARS: structurally indistinguishable

The *Arabidopsis* *SPOROXYTELESS* gene (*AtSPL*) encodes a putative transcription factor that is necessary for sporogenesis (Schieffthaler et al., 1999; Yang et al., 1999). *AtSPL* belongs to a conserved transcription repressor family specific to the embryophytes - the SPL-like, EAR-containing (SPEAR) family (Chen et al., 2014). SPEARs are so-called due to their possession of both a short consensus sequence (RGXGVXXLEXX) at the N-terminal end, termed the SPL motif, and a C-terminal EAR domain (LXLXL). Furthermore, previous studies have identified a Nuclear Localisation Signal (NLS) (Dingwall & Laskey, 1991) both up- (RGRK; (Chen et al., 2014)) and down-stream (KKKR, (Rojas-Gracia et al., 2017)) of the *Arabidopsis* SPL motif. Identified by Chen et al., 2014, the NLS upstream of the SPL motif was conserved in 28 of the 29 embryophyte SPEAR sequences analysed, though the sequence varied ([KR]X[KR][KR]) and the characters did not consistently align. More recently, Rojas-Gracia et al., 2017 identified a conserved NLS downstream of the SPL motif with a stronger affinity to a traditional NLS (KKKR). This sequence was conserved amongst the five dicotyledonous angiosperms sampled: *Arabidopsis*, *Solanum tuberosum*, *S. lycopersicum*, *Nicotiana sylvestris* and *N. tomentosiformis*. The presence of a downstream NLS has not been investigated in non-flowering land plant species. Furthermore, it has not been proven experimentally which, if either, of the proposed NLS sequences function as a true NLS.

Whilst the structure of SPEARs is well characterised (Chen et al., 2014), no distinguishing features have been identified that differentiate SPEARs with an SPL function from other members of the SPEAR family. Of the five known *Arabidopsis* SPEARs, only one has been shown to possess SPL function (Chen et al., 2014). Similarly, in other species such as rice (*O. sativa*), tomato (*S. lycopersicum*) and

cucumber (*Cucumis sativus*) only one SPEAR has been identified with SPL-like function (Hao et al., 2017; Liu et al., 2018; Ren et al., 2018; Rojas-Gracia et al., 2017). Interestingly, these functionally validated SPL homologs exhibit low overall sequence homology with AtSPL. For example cucumber SPL (CsSPL) shares only 20.47% identity with AtSPL, and tomato SPL (S/SPL/HYDRA (HYD)) shares 26.9% identity (Hao et al., 2017; Liu et al., 2018; Rojas-Gracia et al., 2017). However, conservation is markedly higher in the SPL, NLS and EAR domains, which are crucial for their functional roles, further complicating efforts to distinguish SPL-like SPEARs by sequence alone. In all organisms studied to date, SPEARs with SPL-like functions have been identified experimentally, as sequence data alone cannot reliably predict their functional role (Hao et al., 2017; Liu et al., 2018; Ren et al., 2018; Schiefthaler et al., 1999; Yang et al., 1999). Phylogenetic analysis, however, can provide valuable insights, highlighting SPEARs that may be more likely to possess SPL-like functions and guiding further experimental validation (Eisen & Wu, 2002).

3.1.2 SPL function: exclusively angiosperm?

Phylogenetic analysis of embryophyte SPEARs conducted by Chen et al., 2014 indicated that the SPEARs fall into two distinct clades: an SPL clade and a SPEAR clade. The SPL clade exclusively contained SPEARs from angiosperms, whilst the SPEAR clade included SPEARs from all embryophyte lineages. Based on this analysis, it was concluded that SPL function is exclusive to the angiosperms (Chen et al., 2014). However, the SPL clade did not exclusively consist of SPEARs with SPL function, as several non-SPL SPEARs from *Arabidopsis* and rice were also present in this clade (Chen et al., 2014; Ren et al., 2018). Furthermore, this phylogeny was derived from a limited subset of embryophyte sequences, which did not adequately represent the full taxonomic or phylogenetic diversity of the land plants (One Thousand Plant Transcriptomes Initiative, 2019). The dataset included sequences from only three angiosperms (*Arabidopsis*, *Z. mays*, and *O. sativa*), one gymnosperm (*Picea abies*),

and one bryophyte (*Physcomitrium patens*), providing a narrow view of embryophyte diversity (Chen et al., 2014). A more recent phylogenetic analysis, which incorporated two eudicots (*Arabidopsis* and *S. lycopersicum*) and two monocots (*O. sativa* and *Z. mays*) converged on a similar topology to that of Chen et al. (2014), with a distinct SPL and SPEAR clade (Ren et al., 2018). However, this analysis also suffered from limited taxon sampling.

To confirm that SPL is flowering plant-specific, a phylogenetic analysis incorporating a broader and more representative set of taxa across embryophytes is required. Such an approach would provide a more comprehensive understanding of the evolutionary history and functional diversification of the SPEARs.

3.1.3 *SPL* expression: a reproductive link across plant lineages

While phylogenetic analyses offer valuable insights into the evolutionary relationships of SPL and SPEAR proteins, expression analysis provides a critical functional dimension. In *Arabidopsis*, SPL expression is well-characterized. Reverse transcriptase polymerase chain reaction (RT-PCR) detected SPL expression in the flower, rosette leaves, stem tissue and seedlings of *Arabidopsis* (Schiefthaler et al., 1999). Surprisingly, given the highly specific floral phenotype of *spl* mutants, SPL expression was observed in vegetative tissues. The authors suggested that SPL may have no function or redundant roles in these tissues.

The spatial and temporal expression of SPL during early anther and ovule development was further investigated using *in situ* hybridization (Schiefthaler et al., 1999). In the ovule, SPL is expressed throughout the ovule primordium, in the megaspore mother cell (MMC) and the initiating and developing integuments. A more recent study reported that SPL expression is restricted to the L1 cell layer of the ovule primordium (Mendes et al., 2020), however this may be due to differences in experimental procedure (discussed in Section 1.2.2). In the anther, SPL is expressed broadly in the primordium before meiosis, and becomes restricted to the pollen

mother cells (PMC) and tapetum during and after meiosis (Schiefthaler et al., 1999). Expression persists in the tapetum and PMCs until pollen release, highlighting the critical role of *SPL* in both ovule and anther development.

Comparative studies in species such as rice, tomato, and cucumber reveal both conserved and divergent patterns of *SPL* expression (Hao et al., 2017; Liu et al., 2018; Ren et al., 2018; Rojas-Gracia et al., 2017). Expression of the cucumber homolog *CsSPL* closely resembles the *AtSPL* expression pattern (Liu et al., 2018). *In situ* hybridization revealed specific expression in male flowers within the anther primordium, locules, tapetum, and pollen mother cells (Liu et al., 2018). In female flowers, *CsSPL* was expressed in the placenta and ovule primordia, later becoming localized to the nucellus and integuments during ovule development. Likewise, in rice, *OsSPL* transcript levels peaked in young floral structures prior to meiosis and decreased as development progressed (Ren et al., 2018). *In situ* hybridization and reporter assays identified *OsSPL* expression in precursor cells involved in meiosis initiation and reproductive tissue differentiation (Ren et al., 2018). In tomato, *SISPL/HYD* transcripts are most abundant in young flower buds and decline as flowers mature (Rojas-Gracia et al., 2017). *In situ* hybridization localized *SISPL/HYD* expression to sporogenous tissues in the anther, pollen, and developing ovules, both before and after meiosis, but not in the integuments.

Together, these studies demonstrate that *SPL* genes play critical roles in early reproductive development across diverse plant species, with conserved expression patterns in sporogenous tissues and ovule primordia (Liu et al., 2018; Ren et al., 2018; Rojas-Gracia et al., 2017; Schiefthaler et al., 1999). However, species-specific differences, such as the absence of *SISPL/HYD* in tomato integuments, highlight the evolutionary divergence of *SPL* regulation and function (Rojas-Gracia et al., 2017). Furthermore, these comparative studies largely focus on reproductive tissues and do not extensively address *SPL* expression in vegetative tissues. For instance, in stem mustard, an *SPL*-like gene is expressed in leaves as well as buds, suggesting that *SPL*

homologs may also have functional roles beyond reproductive development (Yang et al., 2008). This raises the possibility that the regulatory roles of *SPL* genes extend to vegetative tissues in some species, underscoring the need for broader investigations into their tissue-specific functions across diverse plant lineages.

3.1.4 Where might *SPL* expression be found in a non-seed plant?

Whilst a considerable amount is known about *SPL* expression in flowering plants, little attention has been given to how these expression patterns evolved. What was *SPL* expression like before the evolution of seeds? Did *SPL* have a similar function in earlier land plants, or did its role diverge over time? Was it also associated with reproductive structures in these ancestral lineages?

As the closest living seedless relatives to the seed plants, species within the fern lineage provide an ideal starting point for exploring these questions (Pryer et al., 2001). The homosporous fern *Ceratopteris richardii* serves as a well-established model for investigating the ancestral roles of *SPL*-like genes (Hickok et al., 1995). Unlike seed plants, ferns reproduce *via* spores, and their life cycle alternates between independent gametophyte and sporophyte generations, offering a unique opportunity to examine *SPL* expression across these temporally and spatially separated phases.

Given the conserved reproductive roles of *SPL* in flowering plants, an *SPL*-like gene in *C. richardii* might be expected to be expressed in tissues associated with meiosis, sporogenesis and gametogenesis (Liu et al., 2018; Ren et al., 2018; Rojas-Gracia et al., 2017; Schiefthaler et al., 1999; Yang et al., 1999). Expression could occur in the expanding reproductive fronds, where sporangia initiate, develop, and undergo meiosis (Conway & Di Stilio, 2020; Withers et al., 2023). Additionally, the free-living gametophytes may also exhibit *SPL* expression, particularly during the formation of the antheridia and archegonia. Since *SPL* is expressed in both the male and female reproductive organs of the flower, similar expression patterns would be expected in both the male and hermaphrodite gametophytes (Schiefthaler et al., 1999). *AtSPL*

expression has also been observed in rosette leaves, stem tissue and seedlings, though this expression is not linked to any obvious function (Schiefthaler et al., 1999). Consequently, expression may also be observed in analogous tissues of *C. richardii*, such as the young sporophyte, vegetative fronds and stem tissue.

Studying *SPL* expression in *C. richardii* provides an opportunity to investigate the evolutionary trajectory of these genes. By comparing *SPL* expression patterns in ferns with those in seed plants, it may be possible to infer how their roles have shifted or diversified in response to changes in plant reproductive strategies, offering deeper insights into the molecular evolution of reproduction in land plants.

3.1.5 Aim and objectives

The primary aim of this chapter is to investigate the evolutionary and functional roles of *SPL* and *SPEAR* homologs, focusing on their phylogenetic relationships and expression patterns in the non-seed plant *C. richardii*. This chapter begins by performing a comprehensive phylogenetic analysis of *SPL* and *SPEAR* homologs across diverse plant lineages to elucidate their evolutionary relationships. Following this, the expression patterns of *SPL*-like genes are examined in *C. richardii*, focusing on their potential roles in reproductive development. By integrating phylogenetic and expression data, this chapter aims to provide insights into the ancestral functions of *SPL* and *SPEAR* genes. The objectives of this chapter are to:

1. Conduct a comprehensive phylogenetic analysis of *SPL* and *SPEAR* homologs across diverse plant lineages to investigate their evolutionary relationships.
2. Characterise the expression patterns of *SPL*-like genes in *C. richardii*, particularly in tissues associated with reproduction.
3. Integrate phylogenetic and expression data to hypothesize how *SPL* gene functions may have evolved from ancestral roles in non-seed plants to specialised functions in seed plants.

3.2 Results

3.2.1 Identifying *SPEAR* homologs in *Ceratopteris richardii*

SPEAR homologs have not been previously annotated in the *C. richardii* genome (v2.1) (Marchant et al., 2022). To identify potential homologs, the full length amino acid sequences of all five *Arabidopsis* *SPEAR*s (AT4G27330 (AtSPL), AT2G20080 (AtSPEAR1), AT2G34010 (AtSPEAR2), AT4G28840 (AtSPEAR3) and AT1G29010 (AtSPEAR4)) were used as query in a BLAST search of the *C. richardii* genome (Phytozome 13, v2.1) (Goodstein et al., 2012; Marchant et al., 2022). Three potential homologs were identified: Ceric.23G076500, Ceric.29G040900 and Ceric.21G042000, hereafter referred to as *CrSPLA*, *CrSPLB* and *CrSPLC*, respectively (Table 3.1). The three homologs encode proteins of a similar size, however their nucleotide sequences vary considerably in length due to large intronic regions. For example, in *CrSPLC* one intron is greater than 17.8k bp, dwarfing the transcribed regions (Table 3.1).

To identify candidates with *SPL* function, pairwise similarity and identity within and between *C. richardii* and *Arabidopsis* *SPEAR* homologs in the *SPL* motif alone and in the full sequence was calculated (Figure 3.1). Within *Ceratopteris*, *CrSPLA*, *CrSPLB* and *CrSPLC* share high similarity in the *SPL* motif region and N terminal EAR domain, but are less well conserved in the remainder of the protein (Figure 3.1A,C). *CrSPLB* and *CrSPLC* share a higher degree of similarity (54.5%) and identity (36%) across the full sequence than with *CrSPLA* (Figure 3.1C,D). In *Arabidopsis*, AtSPEAR1 and AtSPEAR3, and AtSPEAR2 and AtSPEAR4 share high degrees of similarity (74.5% and 59.3%, respectively), suggesting that the two pairs may have arisen from duplication events.

Table 3.1: SPEAR homologs identified in the *Ceratopteris richardii* genome

Gene name	Gene Identifier	Protein length (aa)	Gene length (bp)	Location
<i>CrSPLA</i>	Ceric.23G076500	585	5125	Chr23
<i>CrSPLB</i>	Ceric.29G040900	494	10,464	Chr29
<i>CrSPLC</i>	Ceric.21G042000	451	24,247	Chr21

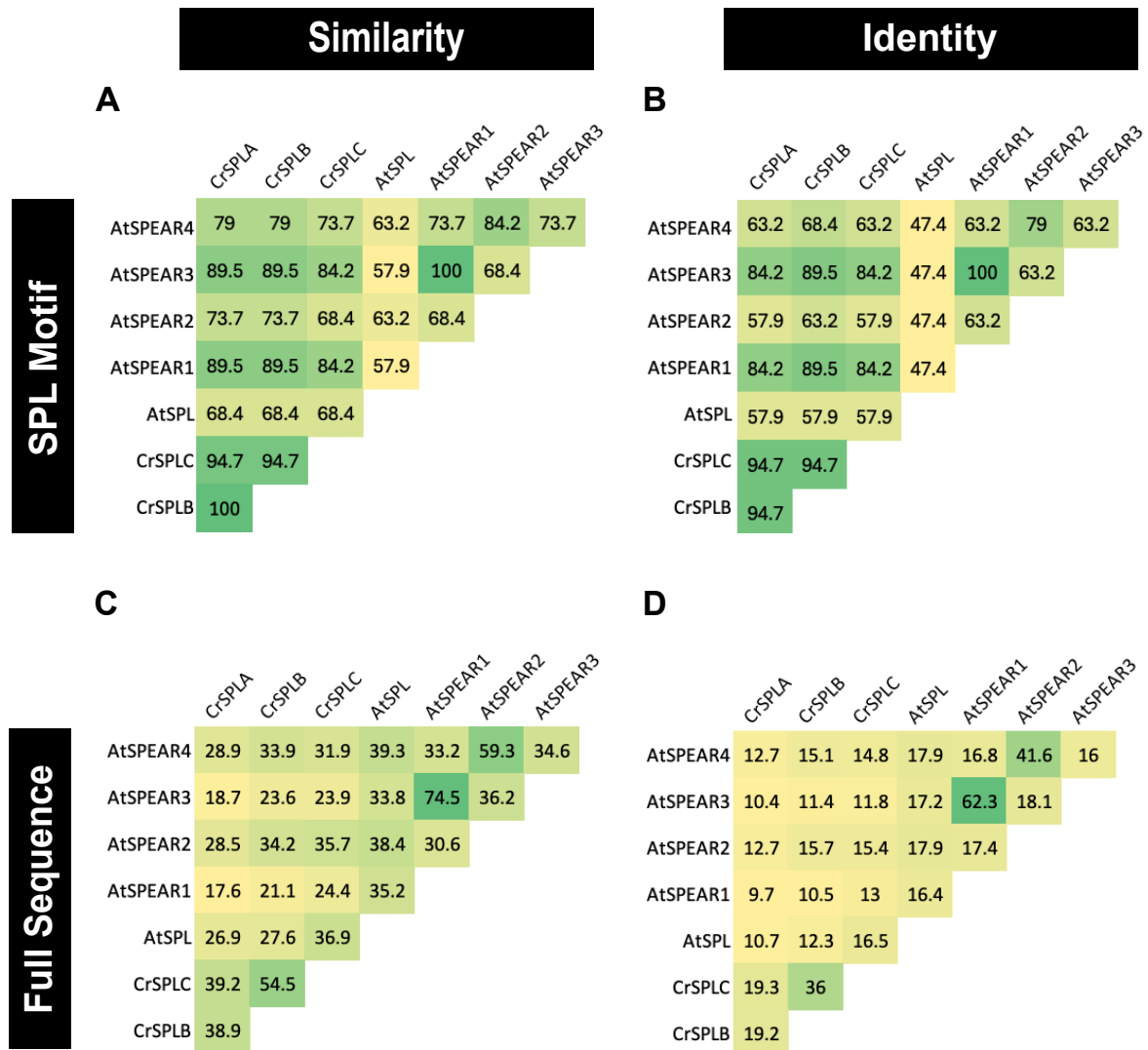


Figure 3.1: Pairwise comparisons of percentage similarity and identity between and amongst *C. richardii* and *Arabidopsis* SPEAR sequences. (A-B) In the SPL motif only and (C-D) across the whole sequence. Panels A and C (left) display pairwise percentage similarity. Panels B and D (right) display pairwise percentage identity.

The three *Ceratopteris* homologs share the same similarity and identity with the *AtSPL* SPL motif, whilst *CrSPLC* has highest similarity across the full length of the *AtSPL* protein. Meanwhile *CrSPLA* is most similar to *AtSPEAR2* and *CrSPLB* is most similar to *AtSPEAR4*. It is not clear from the sequences alone which, if any, of the *CrSPLs* may have SPL function. To achieve this, a phylogeny of *SPEARs* in the embryophytes was generated.

3.2.2 *SPEAR* homologs exist throughout the embryophytes

To identify *SPEAR* homologs in all major groups of the embryophytes, a total of 83 genome (Aya et al., 2014; Goodstein et al., 2012; Guan et al., 2016; Li, 2023; Li et al., 2020; Plomion et al., 2018; Sun, 2022; Wan et al., 2018; Yu, 2022) and 390 transcriptome (One Thousand Plant Transcriptomes Initiative, 2019) databases were searched, yielding 259 and 23 *SPEAR* homolog sequences, respectively (Table 3.2). Due to a disproportionate representation of the angiosperms in available sequenced genomes, the search for *SPEARs* in this group was limited to genomes available at Phytozome v2.1 (Goodstein et al., 2012). Furthermore, to prevent over-representation in the final phylogeny, the initial 46 angiosperm species were reduced to 15, covering 15 major orders (Table 3.3).

In contrast to the angiosperms, the non-flowering embryophytes are represented by fewer sequenced genomes. The search for non-flowering *SPEARs* was therefore expanded to include available transcriptomes (One Thousand Plant Transcriptomes Initiative, 2019). An additional 390 transcriptomes from the bryophytes, lycophytes, ferns and gymnosperms were searched. The sequences found in transcriptomes were often partial, consisting of only the SPL domain, without a C and/or an N terminus. All partial sequences were ultimately removed from the analysis after their inclusion led to skewed phylogenies. In addition, conservation between *SPEAR*-like genes within the pines and within the mosses contributed to skewing, therefore closely related species were removed with the exception of a single representative. A full list of all

species and sequences included in the final alignment is provided in Table 3.3. Of note, no SPEAR homologs, full or partial, were identified in the 9 available genomes and 203 available transcriptomes of algal species (Goodstein et al., 2012; One Thousand Plant Transcriptomes Initiative, 2019).

Table 3.2: Genome and transcriptome databases searched for SPEARS and the number of full-length, validated sequences found in each major group and how many were included in the final alignment.

	Genome		Transcriptome		Incl. in alignment	
	Species	Sequences	Species	Sequences	Species	Sequences
Algae	9	0	203	0	0	0
Hornworts	2	2	7	0	2	2
Liverworts	1	1	21	0	1	1
Mosses	4	9	17	21	3	7
Lycophytes	4	11	17	2	5	13
Ferns	7	20	67	0	7	19
Gymnosperms	10	16	58	0	7	15
Angiosperms	46	200	NA	NA	15	51
Totals	83	259	390	23	40	108

Table 3.3: Taxonomy and gene accessions of SPEAR sequences included in final phylogeny.

Class	Order	Species	SPEAR sequences
Angiosperms			
Magnoliopsida	Nymphaeales	<i>Nymphaea colorata</i>	Nycol02570, Nycol03275, Nycol02241
	Amborellales	<i>Amborella trichopoda</i>	AmTr0007224, AmTr00021118
	Laurales	<i>Cinnamomum kanehirae</i>	Ckan02006900, Ckan01161600, Ckan00636200
Liliopsida	Acorales	<i>Acorus americanus</i>	Acora08G107000, Acora01G117200, Acora07G061600, Acora02G101700
	Alismatales	<i>Spirodela polyrhiza</i>	Spipo0G0163300
	Poales	<i>Oryza sativa</i>	Os01g11430, Os05g48570, Os01g48530, Os01g13040, Os05g42240
	Asparagales	<i>Asparagus officinalis</i>	Aspof051941, Aspof012536, Aspof03882
Eudicotyledons	Ranunculales	<i>Aquilegia coerulea</i>	Aqcoe1G240500, Aqcoe4G061100, Aqcoe7G164000
	Caryophyllales	<i>Beta vulgaris</i>	Bevul18428
Continued on next page			

Class	Order	Species	SPEAR sequences
	Solanales	<i>Solanum lycopersicum</i>	Solyc07g063670 (SlSES), Solyc07240, Solyc018370, Solyc07156, Solyc07152, Solyc62685
	Saxifragales	<i>Kalanchoe fedtschenkoi</i>	Kaladp1291, Kaladp1345, Kaladp0016, Kaladp0173, Kaladp0021
	Brassicales	<i>Arabidopsis thaliana</i>	AT4G27330 (AtSPL), AT2G20080 (AtSPEAR1), AT2G34010 (AtSPEAR2), AT4G28840 (AtSPEAR3), AT1G29010 (AtSPEAR4)
	Malpighiales	<i>Populus trichocarpa</i>	Potri12560, Potri40900, Potri37560, Potri16110, Potri08410, Potri05290
	Vitales	<i>Vitis vinifera</i>	Vivi03940, Vivi00840, Vivi03500
	Fagales	<i>Quercus robur</i>	Qrob2316
Gymnosperms			
Pinopsida	Cupressales	<i>Thuja plicata</i>	Thupl8202, Thupl9208
	Cupressales	<i>Torreya grandis</i>	TG9g00137, TG6g05771, TG1g03545
	Pinales	<i>Picea abies</i>	Picab12809
	Pinales	<i>Pinus taeda</i>	Pita_12604, Pita_20850
			<i>Continued on next page</i>

Class	Order	Species	SPEAR sequences
	Pinales	<i>Pseudotsuga menziesii</i>	Pme0010727, Pme0102804, Pme0024501
Ginkgoopsida	Ginkgoales	<i>Ginkgo biloba</i>	Gb31711, Gb21350
Gnetopsida	Gnetales	<i>Gnetum montanum</i>	Gnmo996777, Gnmo57613
Ferns			
Polypodiopsida	Polypodiales	<i>Ceratopteris richardii</i>	Ceric23G076500 (CrSPEARA), Ceric29G040900 (CrSPEARB), Ceric21G042000 (CrSPEARC)
	Polypodiales	<i>Adiantum capillus-veneris</i>	Adc02091, Adc24181, Adc29526
	Salviniales	<i>Salvinia cucullata</i>	Sacu12255, Sacu02137
	Salviniales	<i>Azolla filiculoides</i>	Azfi_s0139g051205, Azfi_s0002g001430
	Salviniales	<i>Marsilea vestita</i>	Mvest_S11g13912
	Schizaeales	<i>Lygodium japonicum</i>	Lyja07421, Lyja24119, Lyja24434
	Cyatheaales	<i>Alsophila spinulosa</i>	Aspi0116214, Aspi0143184, Aspi0144245, Aspi0148637, Aspi0169971
Lycophytes			
Lycopodiopsida	Selaginellales	<i>Selaginella moellendorffii</i>	Smo438336, Smo411605, Smo410500
Continued on next page			

Class	Order	Species	SPEAR sequences
	Lycopodiales	<i>Lycopodium clavatum</i>	EVM0022355
	Lycopodiales	<i>Diphasiastrum complanatum</i>	Dicom13G054300, Dicom18G060600, Dicom17G024700
	Selaginellales	<i>Selaginella cf. pallescens</i>	Sepal10255, Sepal4611
	Lycopodiales	<i>Huperzia asiatica</i>	HaA09G0732, HaA21G0432, HaA40G0238, HaA58G0516
Bryophytes			
Anthocerotopsida	Anthocerotidales	<i>Anthoceros agrestis</i>	Aagr3402441
	Anthocerotidales	<i>Anthoceros punctatus</i>	Apun000045
Bryopsida	Funariales	<i>Physcomitrium patens</i>	Pp3c226200, Pp3c431500, Pp3c113420
	Dicranales	<i>Ceratodon purpureus</i>	Cepur4G024, Cepur9G046
	Sphagnales	<i>Sphagnum fallax</i>	Sphfalx06G, Sphfalx09G
Marchantiopsida	Marchantiales	<i>Marchantia polymorpha</i>	

Following retrieval and curation, 108 amino acid sequences from 40 species (Table 3.3) were aligned with MAFFT L-INS-i (Katoh et al., 2005). The resulting alignment was of length 1367 due to the introduction of several long gaps. Sequences were highly conserved in the N-terminal SPL motif and the C-terminal EAR domain, with little conservation through the rest of the proteins (Figure 3.2A). Furthermore the NLS sequence upstream of the SPL motif (RGRK) was conserved across the majority of the land plants, albeit less consistently than the SPL and EAR domains. Notably, *CrSPLB* and *CrSPLC* both had convincing NLS sequences (KPRR and RTRR, respectively), whilst *CrSPLA* was less similar to an NLS (RSHR). A conserved NLS downstream of the SPL motif was not detected. The three conserved domains are highlighted in red in the alignment summary, whilst the yellow highlight of the majority of the alignment indicates low conservation (Figure 3.2B). It is clear that the majority of the alignment is not well conserved and is not aligned with high confidence. However, the decision was taken to continue with the full alignment without further curation, so as not to remove potentially informative regions.

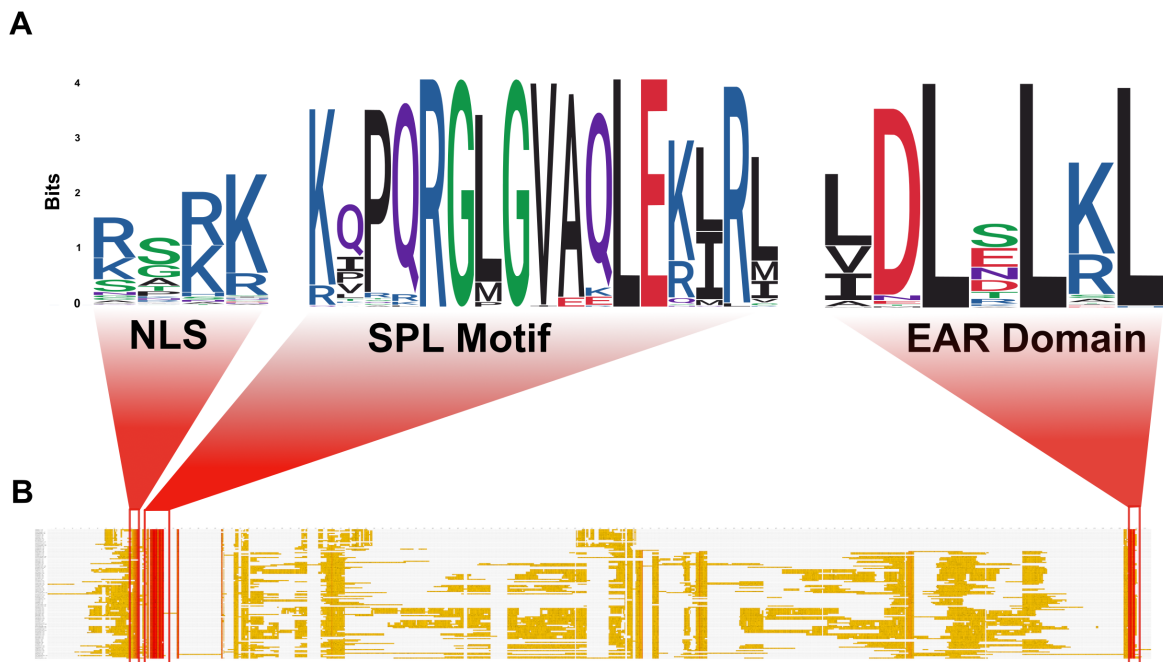


Figure 3.2: Consensus sequence of conserved regions in embryophyte SPEAR alignment. (A) Consensus sequences of the conserved NLS, SPL and EAR domains. Larger letters are more highly conserved. Amino acids coloured by chemistry: red; acidic, blue; basic, black; hydrophobic, purple; neutral, green; polar. (B) Summary of full SPEAR alignment in Jalview. More conserved regions are highlighted in red, less conserved are highlighted in yellow. The conserved NLS, SPL (residues 119-135) and EAR (residues 1343-1348) motifs are surrounded by red boxes.

3.2.3 SPL forms a clade distinct to the remaining SPEARs

Phylogenetic analysis was conducted to determine whether any of the *C. richardii* SPEAR homolog sequences are related to *Arabidopsis* SPL using Maximum Likelihood (ML; Figure 3.3) and Bayesian Inference (BI; Figure 3.4) methods. Both methods converged on a similar phylogenetic topology. Furthermore, the topology is generally in line with the currently accepted topology of the land plants (Puttick et al., 2018). The embryophyte SPEARs formed two distinct monophyletic clades (ML BS support 0.76, BI support <50%); one clade containing AtSPL - the “SPL clade” and another, more recently derived “SPEAR clade” containing the four AtSPEARs without SPL function.

The topology of the SPL clade differed between ML and BI analyses, offering

deeper insight into evolutionary relationships. In the ML tree, the SPL clade is structured with three fern sequences at the base, followed by *Selaginella* sequences, then another clade of ferns, including *CrSPLB* and *CrSPLC* (0.73 TBE support). This is followed by two gymnosperm sequences and a well-supported angiosperm clade, with basal angiosperms forming a clade with monocots and eudicots at the crown. Furthermore, experimentally confirmed SPLs from *Oryza sativa* (Os05g42240/OsSPL) and *Solanum lycopersicum* (Soly63670/SISPL) fall into the SPL clade. In contrast, the BI tree excluded *Selaginella* from the SPL clade, placing it as the outgroup to both SPL and SPEAR clades. Here, a clade of ferns containing *CrSPLB* and *CrSPLC* sits at the base of the SPL clade (99 % BI support), followed by the three more distantly related fern sequences from the ML tree, which now form the base to the gymnosperm and angiosperm clades. As in the ML tree, gymnosperm sequences are sister to the angiosperm clade, with basal angiosperms forming the foundation of the latter. Notably, *OsSPL* was excluded from the SPL clade in the BI tree, appearing instead as an outgroup to the SPEAR clade.

The SPEAR clade contains angiosperm, gymnosperm and fern sequences, but no lycophytes. AtSPEAR1 and 3 and AtSPEAR2 and 4 group together tightly but do not form distinct clades in either topology. The third *Ceratopteris* SPEAR homolog, *CrSPLA* falls outside of the SPL and SPEAR clades within a monophyletic group of fern sequences. The ML tree is rooted by a monophyly containing all bryophyte sequences and all non-*Selaginella* lycophyte sequences. Meanwhile, the BI tree is rooted by moss and liverwort sequences, followed by the non-*Selaginella* lycophytes, with the hornworts diverging later (46% support). Support for the outgroups to the SPL and SPEAR clades is generally low, suggesting that both algorithms were unable to converge on a likely topology.

From this data, it was concluded that *CrSPLB* and *CrSPLC* were more likely to have an SPL-like function. The next step was to see where they are expressed and whether they are specifically expressed in reproductive tissues.

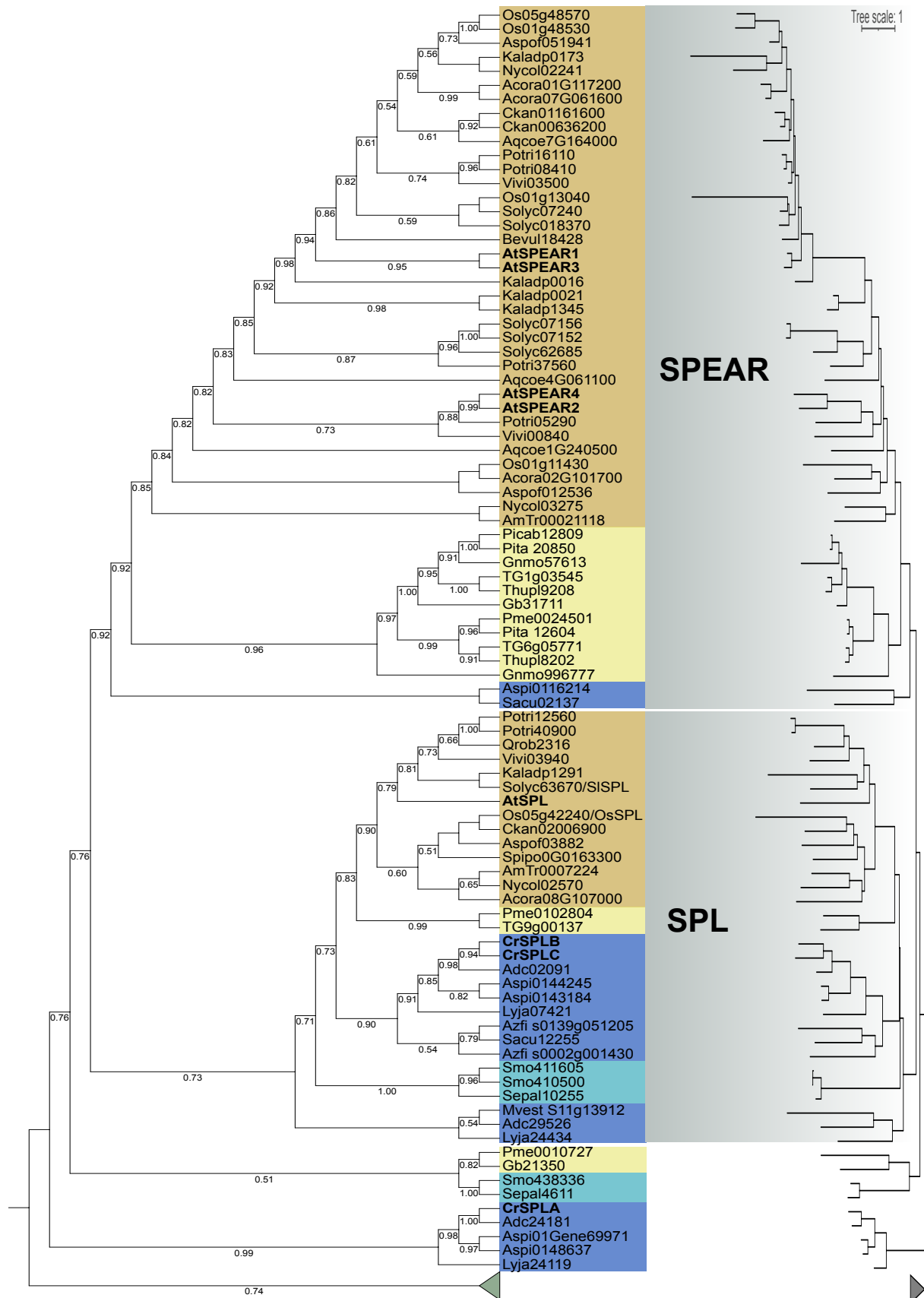


Figure 3.3: Maximum likelihood phylogenetic tree of embryophyte SPEAR amino acid sequences.

Figure 3.3: Maximum likelihood phylogenetic tree of embryophyte SPEAR amino acid sequences. Maximum likelihood analysis was conducted with RAxML-NG, and the resulting tree was visualised in ITOL. *Left* Branching order displayed as rectangular cladogram. The values at the midpoint of each node indicate Transfer Bootstrap Expectation. Values below 0.5 are not shown. Major plant groups are highlighted by colours: Gold; Angiosperms, Yellow; Gymnosperms, Dark Blue; Ferns, Light Blue; Lycophytes. Collapsed clade contains rooted outgroup of Bryophytes and all non-*Selaginella* lycophytes. *Right* Phylogram depicting branch lengths. Tree scale top right.

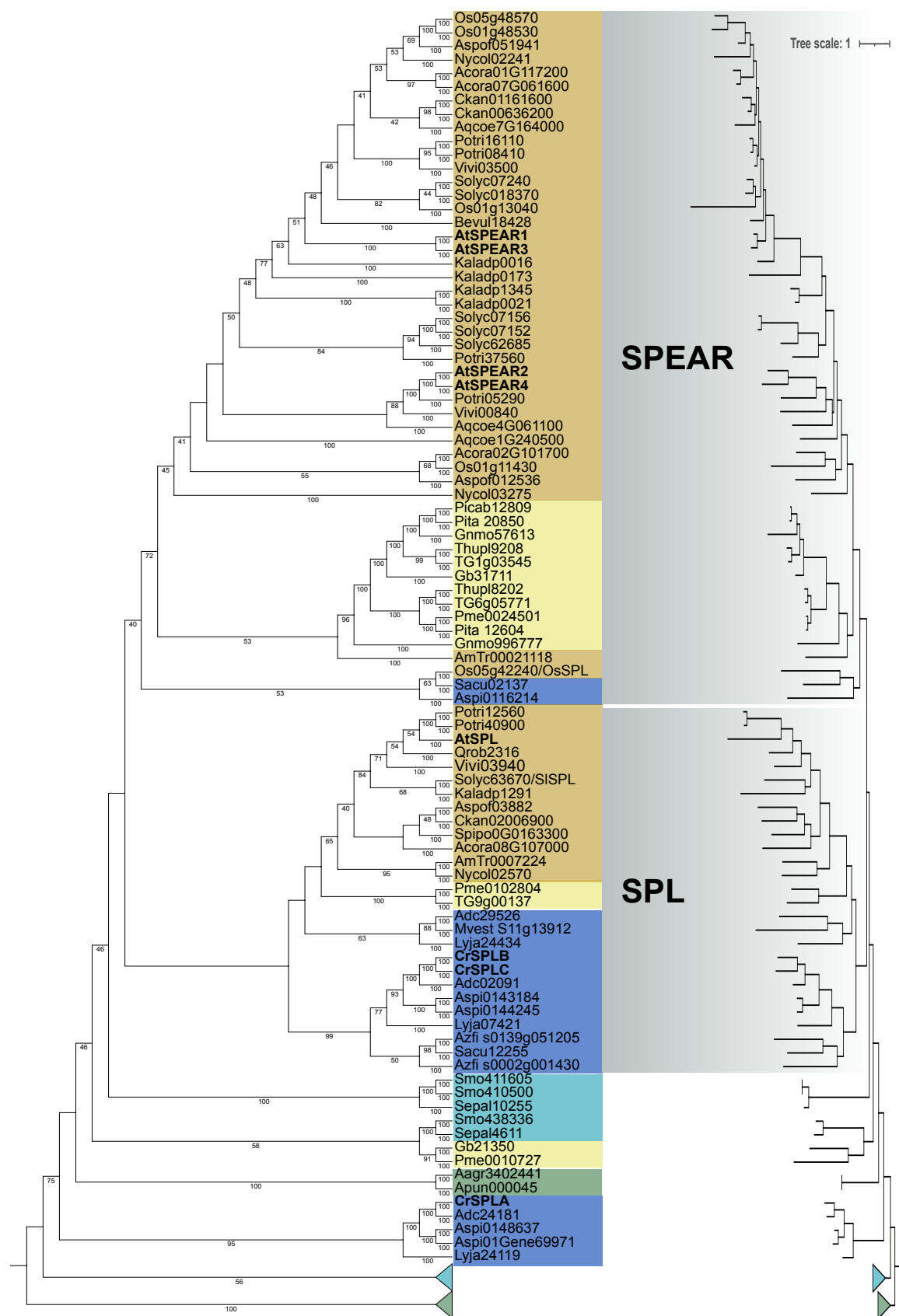


Figure 3.4: Bayesian Inference Phylogenetic tree of embryophyte SPEAR amino acid sequences.

Figure 3.4: Bayesian inference phylogenetic tree of Embryophyte SPEAR amino acid sequences. Bayesian inference analysis was conducted with MrBayes, and the resulting tree visualised in ITOL. *Left* Branching order displayed as rectangular cladogram. The values at the midpoint of each node represent percentage posterior probability. Values below 50% are not displayed. Major plant groups are highlighted by colours: Gold; Angiosperm, Yellow; Gymnosperms, Dark Blue; Ferns, Light Blue; Lycophytes, Green; Bryophytes. Green collapsed clade contains rooted outgroup of mosses and liverworts, light blue collapsed clade contains all non-*Selaginella* lycophytes. *Right* Phylogram depicting branch lengths. Tree scale top right.

3.2.4 Where are *CrSPLs* expressed?

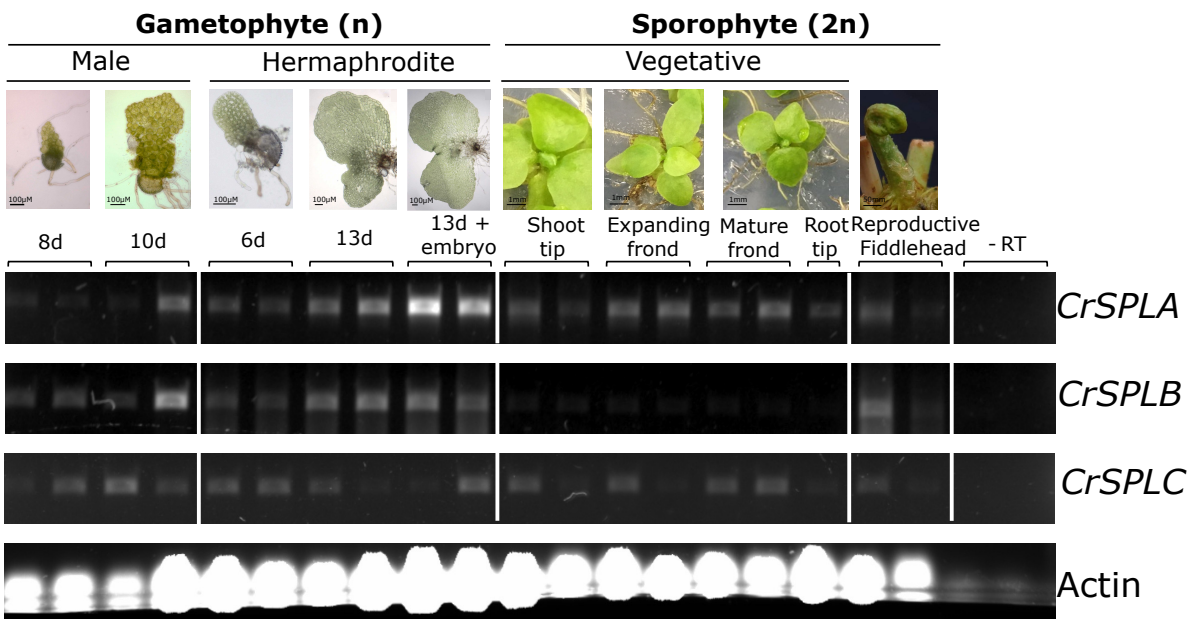
To further investigate which, if any, of the *CrSPLs* may have SPL function, their expression was analysed by semi-quantitative Reverse Transcription PCR (RT-PCR) and quantitative RT-PCR (qPCR) (Fig 3.5). Primers were designed to amplify the full coding regions of the three homologs for use in RT-PCR, and a 100 bp non-redundant region for use in qPCR. The tissues analysed differ slightly between the two methods. The tissue used in RT-PCR was as follows: 8 and 10 day old male gametophyte, 6 and 13 day old hermaphrodite gametophyte - with and without embryo, shoot tip, expanding frond and mature frond of 25 day old sporophyte, and 130 day old reproductive fiddlehead (Fig 3.5A). Meanwhile, samples analysed in qPCR were more limited: 6 day and 13 day old hermaphrodite without embryo, expanding and mature frond of 25 day sporophyte and 60 day old vegetative and 130 day old reproductive fiddlehead (Fig 3.5B). 6 day old gametophyte samples were removed from the analysis of *CrSPLA* and *CrSPLC* due to poor quality amplification by qPCR.

Expression patterns differed between RT-PCR and qPCR. The three genes had different expression patterns across the different tissue types. In the RT-PCR, expression of *CrSPLA* was fairly uniform across all tissue types, with slightly higher expression in the 13 day hermaphrodite gametophyte with embryo (Fig 3.5A). Expression appeared relatively reduced in the male gametophyte samples, however expression of the housekeeping gene *actin* was also reduced, suggesting this may be due to lower cDNA concentration. This expression pattern is in stark contrast to the qPCR data, where expression was significantly higher in the 13 day hermaphrodite gametophyte by almost 20-fold (ANOVA, $F = 64.54$, $p = 1.26 \times 10^{-6}$, $p_{\text{adj}} < 1 \times 10^{-4}$), whilst expression was consistently lower across the remaining tissue, with no significant differences between them ($p_{\text{adj}} > 0.9$)(Fig 3.5B).

In contrast to *CrSPLA*, *CrSPLB* expression was restricted to reproductive tissues in the RT-PCR analysis, with stronger bands in mature gametophyte tissues (Fig 3.5A).

Expression did not significantly differ across tissue types when analysed by qPCR (ANOVA, $F = 1.448, p = 0.316$) (Fig 3.5B). Meanwhile, *CrSPLC* had relatively uniform, low expression across all samples (Fig 3.5A). Whilst it is not clear from the figure, it must be noted that *CrSPLC* bands were markedly fainter than *CrSPLA* or *CrSPLB*, so much so that it was difficult to PCR-amplify *CrSPLC* at all. This is reflected in the qPCR, where fold change is much smaller (with a y axis of 0-3) than the other two genes (Fig 3.5B). Fold change did not significantly differ between tissue types (ANOVA, $F = 2.167, p = 0.147$).

A



B

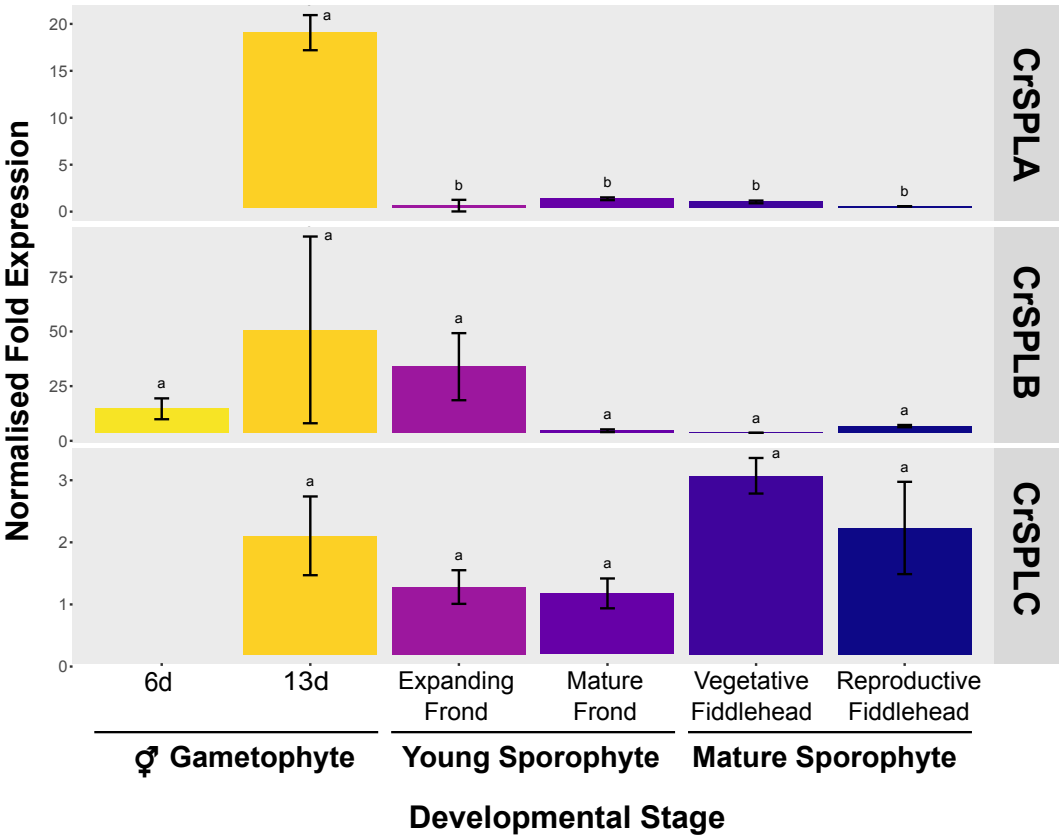


Figure 3.5: *Caption overleaf*

Figure 3.5: Expression of *Ceratopteris* SPL homologs across developmentally young tissue types.(A) RT-PCR analysis of *CrSPL* expression across gametophyte and sporophyte tissues, two replicates per tissue (except root). Samples were amplified with *actin* as a control. (B) qPCR analysis of *CrSPL* expression in the hermaphrodite gametophyte, young sporophyte and mature sporophyte. Error bars represent the standard error of Log2 normalised expression. Compact letter display represents ANOVA with Tukey's Honest Significant Differences (HSD) pairwise comparisons, adjusted for multiple comparisons with Generalised Linear Hypothesis Testing (GLHT). *CrSPLA*: ANOVA, $F = 64.54, p = 1.26 \times 10^{-6}$. 13d Gametophyte is significantly higher than all other tissues ($p_{\text{adj}} < 1 \times 10^{-4}$). *CrSPLB*: ANOVA, $F = 1.448, p = 0.316$. *CrSPLC*: ANOVA, $F = 2.167, p = 0.147$. Gametophyte images provided by Dr Andrew Plackett (personal communication), reproductive fiddlehead image provided by Dr Derry Carr (personal communication).

In addition to qPCR and RT-PCR analysis, Fragments per Kilobase of Transcripts per Million (FPKM) values of *CrSPLA-C* were calculated by RNA sequencing (RNAseq) analysis (Figure 3.6). This analysis took a broader view of *C. richardii* sporophyte development, with various developmental stages from young (25 day (25d)) (Figure 3.6D), mature vegetative (60 day (60d)) (Figure 3.6E) and mature reproductive (130 day (130d)) (Figure 3.6F) sporophytic tissues. Tissues included the shoot apex and frond primordia of 25d, 60d and 130d sporophytes, along with fiddlehead, expanding frond and mature frond of 60d vegetative and 130d reproductive sporophytes.

A kruskal-wallis test revealed significant differences in gene expression across tissues for all three genes ($p = 0.000757$ (*CrSPLA*), 0.000659 (*CrSPLB*), 0.000811 (*CrSPLC*)). However, subsequent pairwise Wilcoxon tests, adjusted for multiple comparisons, did not identify significant differences between any specific tissue pairs ($p_{\text{adj}} > 0.05$). The three *CrSPLs* had markedly different expression profiles. *CrSPLA* followed a pattern of age dependent decline, with high expression values in the developmentally young shoot apex tissues, and then a regular decrease in expression through to the oldest tissue type - the mature frond (Figure 3.6A). This pattern occurred regardless of plant age. From this, it was concluded that there was no obvious reproductive link. In contrast, *CrSPLB* appeared to follow a pattern of age-dependent increase, though this pattern was relatively flat compared to *CrSPLA* and *CrSPLC* (Figure 3.6B). Expression was lowest in the frond primordia of all three tissue ages, then a gradual increase was observed as fronds developed in 60d and 130d plants. The mature 130d reproductive frond had the highest level of *CrSPLB* expression. Once again, the expression profile appeared to be linked to age rather than reproduction.

Meanwhile, *CrSPLC* expression did not appear to be directly linked to tissue age (Figure 3.6C). Expression decreased from shoot apex to frond primordia in the 25d and 60d sporophyte. In the 60d plant, this decline continued until the expanding frond, when there was a large increase in expression in the mature frond. This higher level of

expression was then observed in the 130d shoot apex and frond primordia, followed by a dip in expression in the fiddlehead and then a rise in the expanding and mature fronds. Expression was highest in the 130d mature frond. This apparent link to the reproductive sporophyte could suggest that *CrSPLC* has a reproductive function. However, high expression in the mature fronds of both the 60 and 130d sporophyte could suggest the link is with maturity.

Overall the three expression analyses, coupled with the phylogenies, indicate that *CrSPLB* and *CrSPLC* are the best candidates for a reproductive SPL-like function.

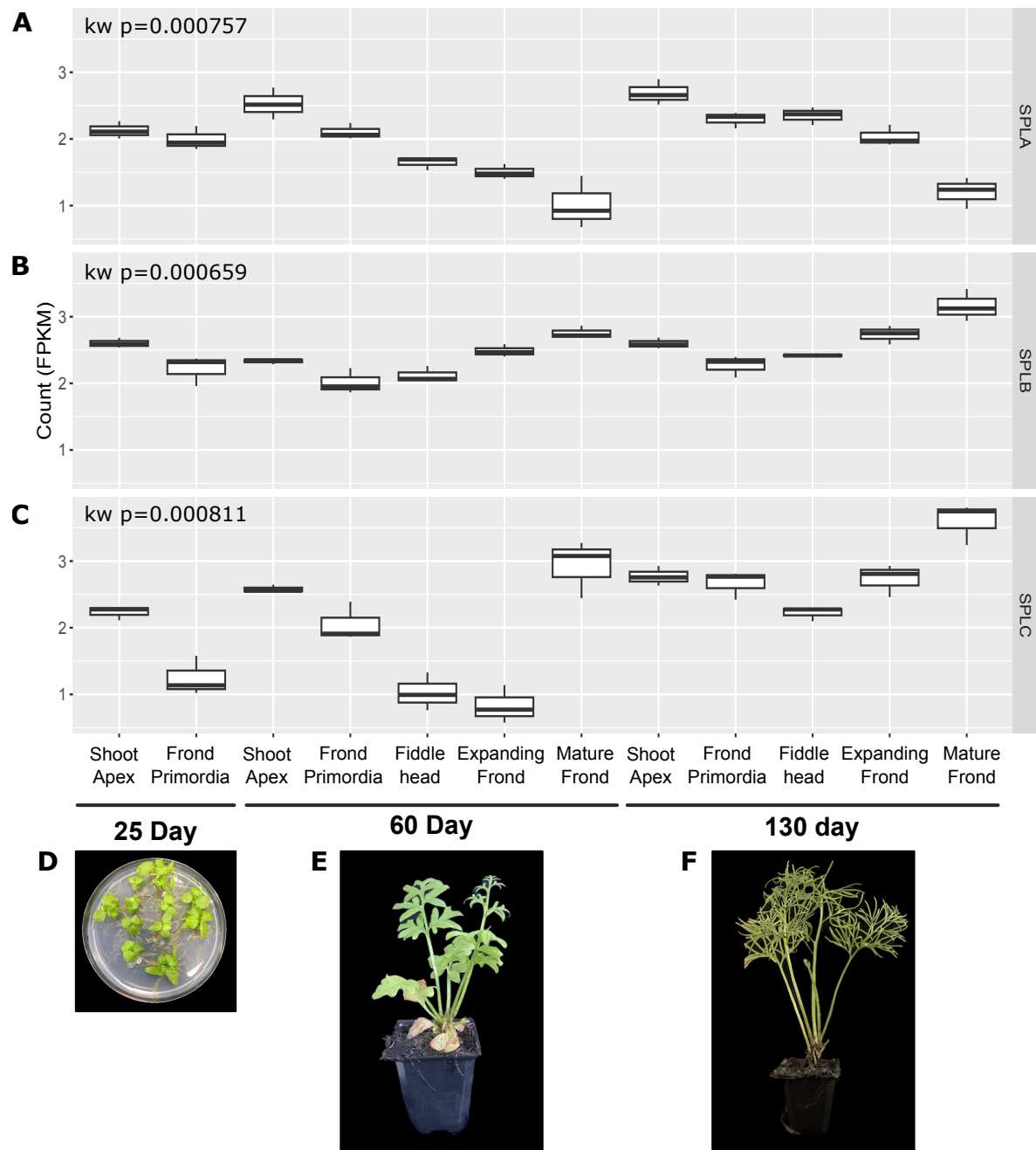


Figure 3.6: Fragments per Kilobase of Transcripts per Million (FPKM) counts of *CrSPLA*, *CrSPLB* and *CrSPLC* in sporophyte tissue of different ages and developmental stages. (A) Expression profile of *CrSPLA*. (B) Expression profile of *CrSPLB*. (C) Expression profile of *CrSPLC*. (D) Tissue culture plate with 25 day sporophytes. (E) 60 day vegetative sporophyte. (F) 130 day reproductive sporophyte. kw = kruskal-wallis.

3.3 Discussion

3.3.1 The SPL and EAR domains are highly conserved in land plant SPEARs

Three genes encoding proteins with SPL and EAR domains were identified in the *C. richardii* genome: *CrSPLA*, *CrSPLB* and *CrSPLC* (Table 3.1). Akin to *Arabidopsis* SPEAR protein sequences, the *CrSPLs* shared high levels of conservation in the N-terminal SPL motif and C-terminal EAR domain, while the intervening regions were not well conserved (Chen et al., 2014). All three genes shared similar levels of similarity and identity with the full *AtSPL* sequence, with similarity levels of 26.9 %, 27.6 % and 36.9 % and identity levels of 10.7 %, 12.3 % and 16.5 % in *CrSPLA*, *CrSPLB* and *CrSPLC*, respectively (Figure 3.1). These values are broadly consistent with the levels of conservation observed in other species, such as tomato and cucumber (Liu et al., 2018; Rojas-Gracia et al., 2017), though overall similarity and identity are lower in *Ceratopteris*, likely reflecting its earlier divergence from flowering plants.

However, conservation and identity of potential SPL-like sequences with *AtSPL* is not a reliable indicator of whether they possess SPL function. For instance, the four *Arabidopsis* SPEAR proteins display approximately 17 % identity and 30 - 40 % similarity with *AtSPL* - levels comparable to those of confirmed tomato and cucumber SPLs - yet none of these *AtSPEARs* have SPL function (Figure 3.1) (Liu et al., 2018; Rojas-Gracia et al., 2017). Therefore, it is not possible to infer SPL function from sequence data alone. To address this, a phylogeny of SPEARs across land plants was constructed to identify candidate *C. richardii* SPLs.

Numerous studies have demonstrated that greater and denser taxon sampling improve phylogenetic accuracy (Heath et al., 2008; Hedtke et al., 2006; Hillis et al., 2003; Zwickl & Hillis, 2002). Building on this principle, new phylogenetic analyses with increased taxonomic sampling and resolution were conducted to expand upon the

previous phylogenetic analyses discussed in Section 3.1. SPEAR sequences were obtained from all major embryophyte groups. In addition, a search of 9 algal genome and 203 algal transcriptome databases did not yield any SPEAR sequences, either full or partial. This corroborates the hypothesis that SPEARs are embryophyte-specific (Chen et al., 2014).

To achieve comprehensive representation across major plant lineages, the selection of sequences balanced available data with taxonomic breadth, despite the limited availability of sequenced genomes from non-flowering embryophytes. At the time of writing, the plant genomics resource Phytozome (v2.1) provided access to 161 land plant genomes (excluding additional versions and varieties), of which only 10 of were non-flowering (Goodstein et al., 2012). To maintain a balanced representation across lineages, the number of included angiosperm species was intentionally limited to 15, representing 15 major orders. Additional non-flowering embryophyte transcriptomes available through OneKP were also searched for SPEARs, however the majority of obtained sequences were partial (One Thousand Plant Transcriptomes Initiative, 2019). Partial sequences are known to strongly bias estimates of topology and branch lengths in phylogenetic analyses (Lemmon et al., 2009). Therefore, only full length sequences, complete with start codon (M), SPL motif and N-terminal EAR domain were accepted into the final alignment.

A multiple sequence alignment of the selected sequences was carried out with MAFFT L-INS-i, to match characters (amino acids) with shared homology, i.e. characters with a common ancestry (Kemena & Notredame, 2009). In line with previous studies, the sequences were highly conserved in the SPL and EAR domains, partially conserved in the N-terminal NLS motif and less conserved throughout the remainder of the protein (Figure 3.2) (Chen et al., 2014). A conserved NLS motif was not detected downstream of the SPL motif, however this has only previously been identified in a small alignment of 5 dicot angiosperms, and could therefore be specific to the dicots (Rojas-Gracia et al., 2017).

The SPEAR sequences included in this study were of varying length, and were not well conserved between the SPL and EAR domains. As a consequence, alignment introduced many gaps and ambiguously aligned regions. It has been suggested that while gap regions may contain potential phylogenetic signal, their inclusion—due to their typically greater difficulty in alignment compared to conserved regions—can lead to an overall reduction in the signal-to-noise ratio of the alignment (Talavera & Castresana, 2007). Exclusion of gaps and ambiguously aligned regions from phylogenetic analyses is therefore a common recommendation (Lemmon et al., 2009). On the other hand, several studies have stressed the importance of gap retention (Aagesen, 2004; Dessimoz & Gil, 2010; Simmons et al., 2008; Wang, Leebens-Mack, et al., 2009). Given that the SPEAR alignment in this study contained only short, highly conserved regions, the decision was taken to retain all gaps and ambiguous regions for phylogenetic analysis so as not to remove potential phylogenetic signal. To mitigate poor topological placement due to ambiguous sites, two phylogenies were built with different methods for direct comparison - Maximum Likelihood (ML) and Bayesian Inference (BI).

3.3.2 SPL function may predate the evolution of the seed

The phylogenetic analysis conducted in this study provides valuable insights into the evolutionary history of SPL and SPEAR proteins. Gene duplication is widely recognized as a key driver of functional innovation in plants, enabling regulatory networks to diversify and adapt to changing environmental and developmental needs (Conant & Wolfe, 2008; Soltis et al., 2015; Van de Peer et al., 2009). Such duplications have often coincided with major evolutionary transitions, including the transition to land. For example, the duplication of *AINTEGUMENTA* (*ANT*) genes in early embryophytes facilitated desiccation tolerance and multicellularity (Dipp-Álvarez & Cruz-Ramírez, 2019). The results presented here align with this concept, suggesting that early duplication events led to the divergence of SPL and SPEAR proteins, with

subsequent lineage-specific expansions shaping their functional diversity across land plants.

The phylogeny reconstructed in this study reveals two monophyletic clades within the embryophyte SPEAR family: the SPL clade, containing AtSPL and experimentally validated SPLs from species such as *Oryza sativa* and *Solanum lycopersicum* (Ren et al., 2018; Rojas-Gracia et al., 2017; Schiefthaler et al., 1999; Yang et al., 1999), and the SPEAR clade, which includes AtSPEARs without SPL function (Chen et al., 2014). This separation highlights the subfunctionalization of these proteins, with the SPL clade retaining key roles in reproductive development. This topology aligns with that seen in (Chen et al., 2014), but with an important distinction: non-angiosperm sequences, including gymnosperms, ferns and lycophytes, also fall within the SPL clade. This finding suggests that SPL function may predate the evolution of seeds.

The detailed topology provides further insights into SPL evolution. The ML SPL clade includes angiosperm, gymnosperm, fern, and lycophyte sequences, while the BI SPL clade contains angiosperm, gymnosperm, and fern sequences but excludes lycophytes, placing *Selaginella* as an outgroup to both SPL and SPEAR clades. In both trees, two *C. richardii* sequences, *CrSPLB* and *CrSPLC*, sit within the SPL clade with moderate to high support (73% and 99%, respectively). Their placement within a clade of fern sequences at the base of the SPL clade in the BI tree may reflect the conservation of ancestral SPL functions within ferns. Conversely, their closer association with gymnosperms and angiosperms in the ML tree hints at functional diversification, consistent with lineage-specific adaptations observed in other transcription factor families (Lang et al., 2010).

The absence of SPEAR sequences in algae and their restriction to embryophytes supports the hypothesis that the SPEAR family emerged early in land plant evolution (Chen et al., 2014). Within the liverworts and hornworts, only a single SPEAR sequence was identified per lineage, whereas mosses exhibited two to three SPEAR sequences, suggesting an early duplication event followed by further expansion in specific lineages. This pattern aligns with the evolution of MADS-box genes, where repeated

duplications and neofunctionalisation coincide with major evolutionary transitions such as the elaboration of the sporophyte and differentiation of male and female reproductive organs (Qiu et al., 2024). Notably, the duplication and diversification of SPEAR sequences coincide with the evolution of vascular plants, which are characterized by an elaborated sporophyte phase and a reduced gametophyte phase (Kenrick & Strullu-Derrien, 2014). This correlation suggests that SPL function may have evolved to regulate developmental processes required for sporophyte dominance, a key innovation in land plant evolution. SPL may have then been co-opted in the seed plants to specifically regulate meiosis.

Despite these advances, challenges remain in reconstructing the evolutionary history of SPEAR proteins. The ML and BI methods produced similar topologies, but support for outgroups and basal relationships was generally low, highlighting difficulties in resolving deep evolutionary splits. This issue may stem from limited taxonomic sampling of basal lineages, particularly non-flowering embryophytes, which is known to bias phylogenetic estimates (Lemmon et al., 2009). Although efforts were made to include as many representative taxa as possible, non-flowering species remain under-represented due to the limited availability of genomic and transcriptomic data (Xie et al., 2024). Expanding sampling to include newly sequenced taxa as they become available will be crucial to improving phylogenetic resolution.

Additionally, the alignment used for tree construction contained many gaps due to low conservation between domains, which may have contributed to the lack of resolution in some branches. A complementary approach could involve excluding poorly conserved regions and rerunning the phylogenetic analyses to assess the impact on tree topology. Structural-based alignments, which often outperform sequence-based alignments in resolving evolutionary relationships, could also be employed (Carpentier & Chomilier, 2019; Malik et al., 2020; Rajapaksa et al., 2023). However, the lack of experimentally determined or confidently predicted structures for SPEAR proteins currently limits this approach. Addressing these challenges will

require integrating novel data and methods, paving the way for a deeper understanding of SPEAR evolution and function.

Together, these findings provide a robust framework for exploring the functional roles of *CrSPLB* and *CrSPLC* in *C. richardii*. Their placement within the SPL clade suggests a conservation of SPL-like functions, and further expression analyses will be essential to determine whether these genes are specifically expressed in reproductive tissues, as observed for SPLs in flowering plants (Schieffthaler et al., 1999; Yang et al., 1999). Such investigations will provide critical insights into the evolutionary conservation and divergence of SPL functions in vascular plants.

3.3.3 The three *CrSPLs* have different expression patterns

The combined results of RT-PCR, qPCR, and RNA-seq analyses provide a multifaceted view of *CrSPLA*, *CrSPLB* and *CrSPLC* expression across diverse tissue type and developmental stages in *C. richardii*. By integrating these complementary approaches, we can discern potential roles for these genes in sporophyte and gametophyte development, while also addressing inconsistencies across methods.

The expression pattern of *CrSPLA* varies notably across the three analyses. RT-PCR results suggest relatively uniform expression across tissues, with slight enrichment in 13-day hermaphrodite gametophytes containing embryos. However, qPCR data reveal a striking 20-fold increase in expression in the same tissue, with significantly lower levels in other tissues. RNAseq adds further context by demonstrating an age-dependent decline in *CrSPLA* expression across sporophyte tissues, with the highest levels in the youngest shoot apex and a consistent decrease through to mature fronds. These discrepancies highlight the strengths and limitations of each method. While RT-PCR provides qualitative insights, it may lack sensitivity to detect finer differences captured by qPCR and RNAseq (Aguilar et al., 2023; Jalali et al., 2017). The pronounced expression in gametophyte tissues by qPCR suggests a role for *CrSPLA* in early developmental processes, possibly related to gametogenesis

or embryo development. Conversely, the RNAseq pattern implies a broader role in early sporophyte development, potentially regulating growth and differentiation during juvenile stages.

The expression of *CrSPLB* is consistently associated with reproductive tissues across all analyses, though the degree of variation differs. RT-PCR shows restricted expression to reproductive tissues, with strong bands in gametophyte and reproductive sporophyte samples. However, qPCR data suggest no significant differences across tissue types, indicating relatively uniform expression. RNAseq further refines this profile, showing a gradual increase in expression with tissue age, peaking in the 130-day reproductive mature fronds. These results collectively point to *CrSPLB* having a potential role in supporting reproductive structures or processes. The stronger expression in reproductive tissues observed in RT-PCR may reflect qualitative differences not captured quantitatively in qPCR. The RNAseq pattern suggests that *CrSPLB* could contribute to processes associated with sporophyte maturation, with its peak in reproductive fronds hinting at a secondary specialization in reproduction.

In both RT-PCR and qPCR, *CrSPLC* exhibits consistently low expression levels. RT-PCR bands were faint across all tissues, and qPCR revealed no significant differences in fold change between tissue types. Meanwhile, RNAseq of sporophyte stages highlights a more complex expression pattern. Expression peaks in the mature fronds of both vegetative and reproductive sporophytes, suggesting a possible association with tissue maturity rather than strictly reproductive functions. However, expression was consistently higher in all reproductive stages, suggesting that *CrSPLC* may play a specific role in reproductive tissues. This was not reflected in the RT-PCR and qPCR analysis, suggesting that *CrSPLC* may have a sporophyte-specific role.

The combined expression data, coupled with phylogenetic analyses, strongly support *CrSPLB* and *CrSPLC* as the most likely candidates for SPL-like reproductive functions in *C. richardii*. The clear association of *CrSPLB* with reproductive tissues in

RT-PCR and its age-dependent increase in RNAseq align well with the roles of SPL genes in reproductive development observed in flowering plants (Schiefthaler et al., 1999; Yang et al., 1999). Similarly, the higher expression of *CrSPLC* in the reproductive sporophyte suggests a potential role in initiating reproductive development in the sporophyte.

The discrepancies between methods underscore the need for complementary approaches to validate expression patterns. For instance, while RNAseq provides broad coverage, its reliance on FPKM values can mask subtle tissue-specific differences (Conesa et al., 2016; Garber et al., 2011; Zhao et al., 2020). On the other hand, RT-PCR and qPCR, despite being more targeted, may be influenced by primer efficiency, cDNA quality, and amplification biases (Jalali et al., 2017). Notably, replicates in RT-PCR did not consistently show identical expression patterns, raising concerns about reproducibility. Additionally, the technical limitations of the qPCR data rendered it less reliable for drawing definitive conclusions. To address these challenges, future studies should prioritise replication using freshly synthesised cDNA, ensuring consistent stages across analyses to enable direct comparisons. Moreover, spatial expression analyses - such as *in situ* hybridization or promoter-reporter constructs - are critical for confirming the localization of *CrSPLs* in reproductive tissues (Jackson, 1991; Jefferson et al., 1987). These approaches will provide robust validation of expression patterns and shed light on the potential functional roles of these genes in *C. richardii*.

The combined insights from phylogenetic and expression analyses provide a compelling framework for understanding the evolutionary and functional trajectories of *CrSPLs* in *C. richardii*. Phylogenetic reconstruction positioned both *CrSPLB* and *CrSPLC* within the SPL clade, suggesting the conservation of SPL-like functions across vascular plants and reinforcing the hypothesis that SPL function predates the evolution of seeds. Expression patterns further substantiate these phylogenetic findings, with *CrSPLB* and *CrSPLC* exhibiting distinct, tissue-specific expression

profiles consistent with roles in sporophyte development and reproduction. Notably, *CrSPLB* was preferentially expressed in gametophytes and reproductive sporophytes, while *CrSPLC* demonstrated a potential association with tissue maturity and reproductive sporophyte-specific processes. These findings highlight the potential for *CrSPLB* and *CrSPLC* to retain ancestral SPL functions while adapting to the unique developmental and reproductive strategies of ferns.

Chapter 4

What are the functions of CrSPLs?

4.1 Introduction

4.1.1 SPL in flowering plants: conserved roles and context-specific adaptations

SPOROCTELESS (*SPL*) is a conserved regulator of sporogenesis in the flowering plants, with essential roles in both male and female gametophyte initiation across diverse species (Liu et al., 2018; Ren et al., 2018; Rojas-Gracia et al., 2017; Schiefthaler et al., 1999; Yang et al., 1999). In *Arabidopsis thaliana*, *SPL* is necessary to specify the archesporial cells, the precursors to meiotic cells, in anthers and ovules (discussed in Section 1.2). Loss of *AtSPL* function results in sterility due to the complete failure of sporogenesis. Similarly, in rice and tomato, knockout of *SPL* homologs results in complete male and female sterility (Ren et al., 2018; Rojas-Gracia et al., 2017). Furthermore, the tomato gene, *SISPL/HYDRA* (*HYD*), was able to replace *AtSPL* function in the *Atspl* mutant and recover fertility, demonstrating conservation of function in the eudicots.

While *SPL* function is highly conserved across species, there are notable differences in its roles and phenotypic effects. In tomato, *SISPL/HYD* not only regulates sporogenesis but also mediates fruit set, a role that highlights species-specific adaptations of the *SPL* gene (Rojas-Gracia et al., 2017). In cucumber (*Cucumis sativus*), knockdown of *CsSPL* results in severe reproductive phenotypes, but it does not completely abolish ovule and pollen formation (Liu et al., 2018). The severity of phenotypes correlates with *CsSPL* expression levels, suggesting a dosage-dependent effect. These species-specific differences in *SPL* function raise questions about how its regulatory roles are balanced across developmental contexts, particularly in light of findings from *AtSPL* overexpression studies (Li et al., 2008) (discussed in Section 1.2).

The contrasting phenotypes of *Atspl* loss-of-function mutants, which exhibit no

vegetative abnormalities, and *spl-D* gain-of-function mutants, which display severe defects in both vegetative and reproductive development, highlight the importance of tightly regulated *SPL* expression (Li et al., 2008; Schiefthaler et al., 1999). Overexpression of *AtSPL*, as seen in *spl-D* and *35S::SPL* lines, disrupts normal development, leading to progressively severe vegetative phenotypes such as narrow, curled leaves, and aborted shoot apical meristems (SAMs) (Li et al., 2008). The phenotypic severity correlates directly with *SPL* expression levels, indicating a dosage-dependent effect. Interestingly, overexpression studies of *SPL* homologs have not been reported in other species, leaving open questions about whether similar phenotypes or dosage-dependent effects occur outside of *Arabidopsis*.

These findings suggest that *SPL* functions as a latent regulator in vegetative tissues, with its misregulation leading to developmental disruption. The necessity of precise spatial and temporal control of *SPL* expression underscores its critical role not only in reproductive development but also in maintaining normal vegetative growth. This has broader implications for understanding how *SPL* homologs may have evolved diverse regulatory roles across plant species while maintaining the need for stringent control to avoid deleterious effects.

4.1.2 Investigating the evolution of *SPL* function using *Ceratopteris richardii*

To investigate the evolutionary origins of *SPL* function, *Ceratopteris richardii* serves as an ideal model (Hickok et al., 1995). As a homosporous fern, *C. richardii* represents a lineage where reproduction is dependent on spores rather than seeds, providing an opportunity to study whether *SPL* function in reproductive development was co-opted during evolution. In flowering plants, *SPL* plays a critical role in specifying archesporial cells, initiating meiosis, and supporting gametophyte development (Liu et al., 2018; Ren et al., 2018; Rojas-Gracia et al., 2017; Schiefthaler et al., 1999; Sieber et al., 2004; Yang et al., 1999). If *SPL* function is conserved, one might expect an *SPL* homolog in *C. richardii*

to regulate similar processes, such as sporangium development, spore production, or early gametophyte formation.

A knockdown of *SPL* in *C. richardii* could provide valuable insights into its ancestral roles. Potential phenotypes might include disrupted sporangium formation, reduced spore viability, or abnormalities in gametophyte development. In *Ceratopteris*, meiosis occurs within the sporangia on reproductive fronds, and gametophyte development begins post-spore germination (Conway & Di Stilio, 2020; Withers et al., 2023). If *SPL* expression is required for these processes, its absence would likely result in defects at key reproductive stages, shedding light on how *SPL* functions were adapted and expanded in seed plants.

Attempting to functionally complement the *Atspl* mutant with *CrSPL* homologs could further elucidate its evolutionary conservation. Given the essential role of *AtSPL* in sporogenesis, successful complementation would restore fertility in the mutant, indicating a shared function between the fern and angiosperm homologs. Conversely, failure to complement would suggest divergence in *SPL* function between the two lineages, possibly reflecting adaptations to their distinct reproductive strategies.

4.1.3 Transformation methods in *C. richardii*

The transformation of *C. richardii* has been explored using various methods, each with distinct advantages and limitations (Bui et al., 2015; Plackett et al., 2014; Rutherford et al., 2004; Stout et al., 2003). One approach involves microparticle bombardment of callus induced in the sporophyte (Plackett et al., 2015, 2014). This method has been used to successfully introduce stable reporter constructs and induce stable gene silencing, providing a platform for functional studies (Geng et al., 2022; Plackett et al., 2018; Xiang & Li, 2024). However, microparticle bombardment has notable drawbacks that limit its efficiency and precision.

Shortly after the publication of Plackett et al., 2014, an alternative transformation method utilizing *Agrobacterium tumefaciens* was developed, enabling stable

transformation of *C. richardii* gametophytes (Bui et al., 2015). This approach bypasses the need for callus induction in the sporophyte, a labour-intensive step inherent to biolistic transformation (Plackett et al., 2015). In *Agrobacterium*-mediated transformation, the haploid gametophyte is directly transformed, and self-fertilization of the transformed gametophyte produces stable T1 homozygous mutants in significantly less time compared to microparticle bombardment (Bui et al., 2015). While stable transformants *via* biolistic methods can take at least 10 months to generate, *Agrobacterium*-mediated transformation achieves this outcome rapidly, making it more efficient and practical (Bui et al., 2015; Plackett et al., 2015).

In addition to time constraints, microparticle bombardment introduces several challenges. The method randomly integrates multiple transgene copies into the genome, often resulting in transgene silencing, aberrant expression, or loss (Agrawal et al., 2005; Ismagul et al., 2018; Kohli et al., 1999; Sparks & Jones, 2004; Srivastava et al., 1996; Stoger et al., 1998). For example, Plackett et al., 2014 recorded over eight copies of the transgene in nearly all T0 individuals, making single-copy mutants exceedingly rare. In contrast, *Agrobacterium*-mediated transformation is well-documented for its ability to produce single-copy mutants, reducing the risk of unintended transgene behaviour and offering a more reliable tool for functional studies (Gao & Nielsen, 2013).

Despite these limitations, microparticle bombardment remains a commonly used method, particularly in laboratories with established expertise and protocols (Plackett et al., 2018). In this study, the pANDA vector was employed for RNA interference (RNAi) knockdown experiments using microparticle bombardment (Miki & Shimamoto, 2004; Plackett et al., 2014). The following section details the functionality of the pANDA system, including its design, mechanism of action, and application in *C. richardii*.

4.1.4 RNAi gene silencing in *C. richardii*

RNA interference (RNAi) is a powerful tool for gene silencing in a variety of organisms, leveraging the organism's own immune system to reduce or eliminate the expression of a target gene (Hannon, 2002). The principle behind RNAi involves mimicking a viral attack by introducing double-stranded RNA (dsRNA) specific to the target gene, tricking the organism into suppressing transcription at all genomic loci containing the dsRNA sequence.

In plants, RNAi is typically achieved by introducing a hairpin RNA molecule specific to the target gene (Chuang & Meyerowitz, 2000; Fusaro et al., 2006; Johansen & Carrington, 2001; Schweizer et al., 2000). The hairpin RNA contains two inverted repeats of the target sequence separated by a linker region, and it is transcribed under the control of a constitutive promoter (Fusaro et al., 2006). Once transcribed, the RNA folds into a secondary structure, with complementary regions binding to form dsRNA connected by a loop. This dsRNA triggers recognition by the DICER protein, which cleaves it into small interfering RNAs (siRNAs) (Bernstein et al., 2001; Elbashir et al., 2001). These siRNAs guide the RNA-induced silencing complex (RISC) to complementary mRNA sequences, ultimately repressing transcription or degrading the target mRNA (Hannon, 2002; Liu et al., 2003).

For *C. richardii*, a pre-existing RNAi vector originally developed for rice, pANDA (Miki & Shimamoto, 2004), has been adapted for use (Plackett et al., 2018). The pANDA vector simplifies the creation of hairpin RNA constructs by including a framework with a constitutive promoter, linker region, and two cloning sites for inserting the target sequence as an inverted repeat (Miki & Shimamoto, 2004). This approach has proven effective in *C. richardii* for exploring the function of LEAFY in the maintenance of apical stem cells (Plackett et al., 2018).

4.1.5 Aims and Objectives

This chapter focuses on investigating the function of *SPL* genes in *C. richardii* using RNAi and complementation approaches. By transforming *C. richardii* with a pANDA construct containing targets to *CrSPL* homologs, this work aims to uncover the role of *CrSPLs*, thereby hopefully providing clues as to the ancestral roles of *SPL* and their conservation across plant lineages.

The primary aims of this chapter are to:

1. Generate *CrSPL* RNAi targets using the pANDA vector.
2. Introduce the resulting vectors to *C. richardii* by microparticle bombardment transformation.
3. Characterise the resulting phenotypes of *CrSPL* knockdown, with a focus on *CrSPLB*, one of two most likely candidates identified in Chapter 3.
4. Construct a complementation vector to attempt functional complementation of the *Atspl* mutant.

4.2 Results

4.2.1 Generation of *CrSPL* RNAi knockdown constructs

RNAi knockdown in *C. richardii* was achieved using the pANDA vector (Miki & Shimamoto, 2004; Plackett et al., 2015). The pANDA vector is configured such that two copies of a target DNA fragment are inserted either side of a linking sequence (the GUS linker) in opposite orientations (Miki & Shimamoto, 2004). When expressed under the maize Ubiquitin promoter (*pUbi*) *in planta*, these complementary sequences will form a double stranded RNA (dsRNA) hairpin which is targeted for degradation, resulting in knockdown of expression of the target gene. Target sequences for the three *CrSPL*s were inserted into the pANDA vector *via* Gateway cloning (Thermo Fisher Scientific, 2003). This involved several steps: amplification of a target region with *attB1* and *attB2* sites at the 5' and 3' ends, respectively, transfer into the Gateway entry vector pDONR207 via the Gateway BP reaction and finally transfer into the destination pANDA vector via the Gateway LR reaction.

To ensure targeted knockdown of *CrSPLA*, *B* and *C*, divergent regions of each gene sequence were selected. The coding sequences of the *CrSPL* homologs were aligned with ClustalW to find regions with low similarity (Thompson et al., 1994). Exon 1 was deemed satisfactorily dissimilar, and primers were designed to amplify a 281 bp region of *CrSPLA*, a 379 bp region of *CrSPLB* and a 371 bp region of *CrSPLC*, hereon referred to as *CrSPLA-RNAi*, *CrSPLB-RNAi* and *CrSPLC-RNAi* (Figure 4.1). The selected regions were further subjected to BLAST2n and pairwise alignments against the full-length coding sequence of their homologs to confirm there was no significant similarity (Camacho et al., 2009).

The full coding sequences of *CrSPLA*, *CrSPLB* and *CrSPLC* were amplified and cloned into pJET. Correct amplification was validated by digestion with *Bgl*II (Figure 4.2, Panel A). The validated *CrSPL* coding sequences were then used as template for the RNAi

CrSPLA	---ATGCTAGATCGCGATCCGAATGCTGA-----	26
CrSPLB	ATGATGGCCCATCTCATTTGCCAATAATCCAGGCCAGCAGGATGCAGGCGAGATTTTGCTC	60
CrSPLC	---ATGGTCCATCTGTTAGCTAATAAACCAAGCCAGCAAGATGCAGGTGAATATTTCTC	57
	*** ** *	
CrSPLA	-----TCGTCTCCGGACCCC-----CCACTATTGCGCACACGT	60
CrSPLB	GCTGGATTACCCGACATGGCCCTCCGAATGGAAACTCAAAGAAGGTTGACGAAGCCAGGA	120
CrSPLC	TCTGGATTACGCCATTTAGCTCTTCAAATGGAAGCCCAAGGAAGGTTGATGAATATATCA	117
	* * *** *	
CrSPLA	TCTCGATCACACAGGCAGAAGAATGCTCCCAAGAAGCAGCCTCAGCGAGGGCTCGGTGTT	120
CrSPLB	TCCAAACCCAGAAGAAGCAAAAAGGCCCAAAAGCAGCCACAGAGAGGACTTGGCGTC	180
CrSPLC	CCCAGGACGAGAAGAAAGAAAGGCCCAAGAAGCAGCCGACGAGGGCTTGAGATT	177
	* * * * *	
CrSPLA	GCACAGTTGGAGAAGCTTCGCCTAGAAGAACAAGCAACAGGAGGCCCTCTGCTTT	180
CrSPLB	GCTCAGCTGGAGAAGATCAGGCTTGAGGAGCAAGAAGAGCTTCGGAGCATGCACACAG	240
CrSPLC	GCTCAGCTAGAGAAGATGAGGCTGGAAGAGCAGAGCAGCATCACATCAGGCGCAGCAG	237
	* * * * *	
CrSPLA	GCTTCTGTACATCATCAGAACCTTCATCCGCACAATCAATTCGTCCAATTGCAACAGGAT	240
CrSPLB	ATTTTAGCTCTCCGTCAGCTTCAGTCTGCATCTCCTCTACTC-----	282
CrSPLC	CTCATAGCTCTACGTCAACTACACTCTACACATAGCTTCTTG-----	279
	* * * * *	
CrSPLA	CGGCAAGCTTCGTATGTTCTTTATCCGCATCAGAATTTCCAACATCAACACAATCAGGGT	300
CrSPLB	-----ATTCTTGCTA--GCGGAT	298
CrSPLC	-----TATGCAAACG--AAC--	292
	*	
CrSPLA	CCGCACCCCTCTCCGCTATTCTTGCCAAGACATTTCCAGCGTATGGAAAA--TTCATA	356
CrSPLB	CCATGCAAGGCG--CTGCTTCGGTCATTCCCCGCTGGATAGGGAAGATTATTCAAAA	355
CrSPLC	ACGCTCCTA-----GTGTTAGTGTATCCCGCAACTGGAGCGAAATAGATCAGTCACTG	346
	* * * * *	
CrSPLA	TGAAAAGGCTAGCTTTACA-----GACCCTCGCTTTCGCCTACCTTGT	399
CrSPLB	CTAGCAGCTTTGGGATACATGCGGATCCCTCTTCAAAGACCTCCTCCGCAATTTCACTTTG	415
CrSPLC	TTAATGGCCGC--ATGCTTCGTATCCATCA--ACCAGCCAGCTTTGCATGAACCTTC	400
	* * * * *	
CrSPLA	GTCAAGCTCAATGCTCTGCAGAAATGATTACTCCCGCTCGCGTCTGTCAACAGCCAAAGCG	459
CrSPLB	CCTGCGCTGATAGCAATGGAGGCCAAGCTATTCCGAACATGCTCTCTCCGCA-----	468
CrSPLC	TTGCGCTCGAGGGCTCTCCGAGCTCGGTTCTTTTGAACATCTTTTCTAAGGCC-----	453
	* * * * *	
CrSPLA	CAAACCAGGCCGGTGCCTTGTTCATACCTCAATGGCAATTTCCCGCTGAGCAGCAGGCG	519
CrSPLB	-----TGTCGTAATGCTTCTTCTGCAACATTC	496
CrSPLC	-----TCGACTGCGTGTGCTGCTGCAAGCTCTG	481
	* * *	
CrSPLA	TCTACTTCTCTGTATCACAAGCTCATGATCTATGCAGTCCCGCCACACAATTGAGACCG	579
CrSPLB	TCGAACCTCTTCCAC-----	512
CrSPLC	CCTGC-----	486
	*	
CrSPLA	TCATCTGCCGAGAGTCGCCTTTGTCTCTACTGAAAGGTGCATCTCCTGCTTCTTTAGTG	639
CrSPLB	-----GGAAAGTTCCACAGGAGCAAGCTCAGTGAACCTTTCTGGTAGTTTCTTGCT	564
CrSPLC	-----TTCTGCTAGGTTCTGATGGTGAAGTCTCTC	516
	*** * *	
CrSPLA	GACAAGTCATGGCTCAATCATCTTTATGACAGGGATTCGGACA-----AGGAA	687
CrSPLB	GAAAATCCCTGGCTGCATCAACGGGAAGCCTCGGATGAGGGGAGATTGCAATTTCTGCAAT	624
CrSPLC	GACAGGCCACTT-----CAACTACAGTTGGGTAAACAAAGCGACGTTTTCGGGAG	567
	** * * * *	
CrSPLA	AAAGATTCTTTAATCATTTCTCTCTCTGTCATGTTCCACGACATTGATGTACCAATT	747
CrSPLB	GCGAATTCTTGCAGAACTCACTCAATCCACCATGCGGAA-----GAAGCAA--	671
CrSPLC	AAGAATTCTCTTTACCTCACCTTCAGACTGATGAACGCC-----CTGACTA--	614
	*** * * *	
CrSPLA	CAGTGCCAAGATTGCAGTGCCACACCCGGTGTTAACACAGGAAATGCGAATTTCCACGAG	807
CrSPLB	-----TGGACATGCGGGATCTCTCTCTTCCACA	700
CrSPLC	-----TTGGAATGGAATGTTCTTTGCTTCTCA	643
	* * * * *	

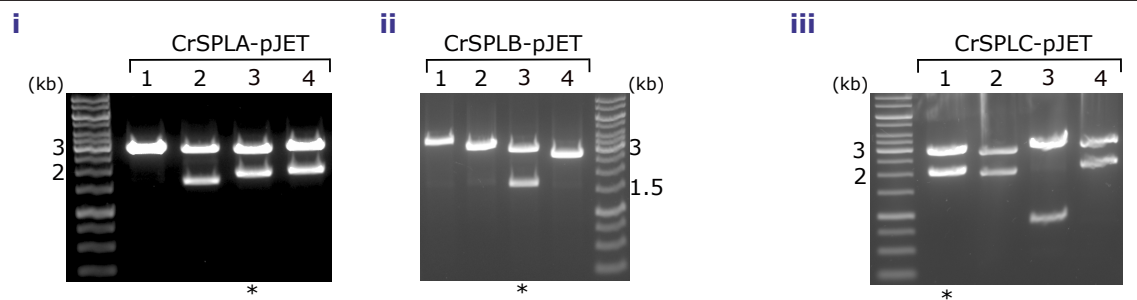
Figure 4.1: Multiple sequence alignment of the first exon of *CrSPLA*, *CrSPLB* and *CrSPLC*. Target regions selected for RNAi are highlighted in black (*CrSPLA*), dark grey (*CrSPLB*) and light grey (*CrSPLC*). Alignment was performed with ClustalW.

fragments. Primers were designed to amplify the target fragments and add the *attB1* and *attB2* sites necessary for Gateway cloning in one step (Thermo Fisher Scientific, 2003). Bands of the expected size were obtained for all three fragments (Figure 4.2, Panel B). To confirm that the correct sequence had been obtained, diagnostic digestion and sequencing was performed. Fragments were gel-extracted, ligated into pJET1.2 and transformed into *E. coli DH5 α* . Plasmid DNA was extracted and a diagnostic digest with *Bgl*II performed. The correct digestion pattern was obtained for *CrSPLC-RNAi* and confirmed by sequencing (Figure 4.2, Panel Ci). However, several attempts to ligate *CrSPLA-RNAi* and *CrSPLB-RNAi* into pJET1.2 were not successful. As a result, the PCR products were transferred directly into pDONR207 without prior validation by digestion or sequencing.

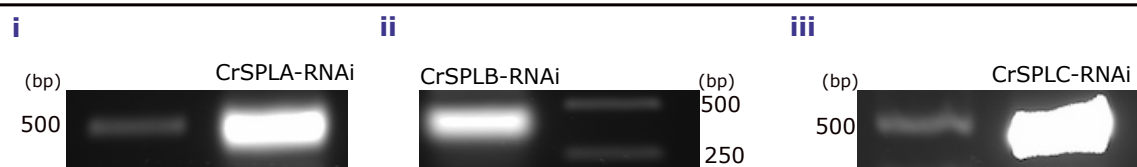
After confirmation that the correct sequence had been obtained, *CrSPLC-RNAi* was transferred into the intermediate pDONR207 entry vector via Gateway BP recombination using the linearized *CrSPLC-RNAi-pJET* plasmid, and propagated in *DH5 α* (Thermo Fisher Scientific, 2003). Insertion was confirmed via diagnostic digest with *Apal* and *EcoRI* (Figure 4.2 Panel Diii). Meanwhile, the *CrSPLA-RNAi* and *CrSPLB-RNAi* PCR products were transferred directly into pDONR207 via the same BP reaction modified for PCR products, followed by propagation in *DH5 α* . Since at this stage, the sequences of neither the *CrSPLA-RNAi* nor *CrSPLB-RNAi* had been validated, they were confirmed by diagnostic digest with *Apal* and *EcoRI* (Figure 4.2 Panel Di-ii), followed by sequencing (data not shown).

The target fragments were transferred from pDONR207 into the destination pANDA vector via the gateway LR reaction. Correct insertion of the target fragments into pANDA was confirmed by digestion with *NotI* (Figure 4.2 Panel Ei-iii). Whilst the bands observed were larger than expected, they did display the correct pattern compared to the vector only control. This was confirmed with additional digests and sequencing (data not shown).

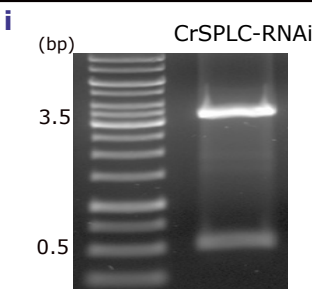
Panel A: Full CDS / *Bg*/II



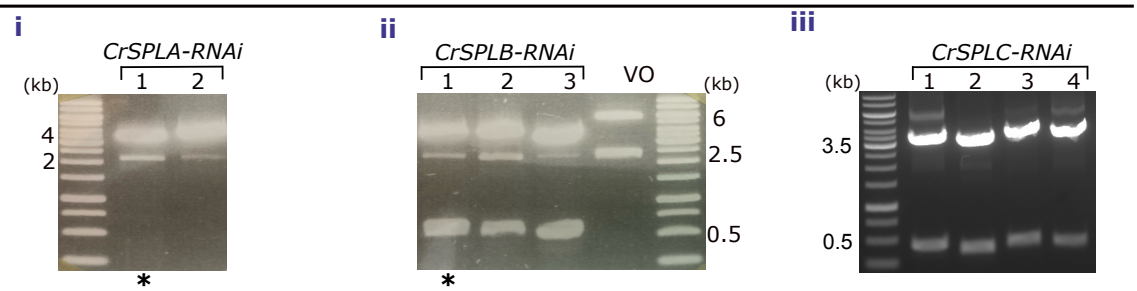
Panel B: Fragment + *attB* sites



Panel C: RNAi-pJET / *Bg*/II



Panel D: RNAi-pDONR207 / *Apa*I/*Eco*RI



Panel E: RNAi-pANDA / *Not*I

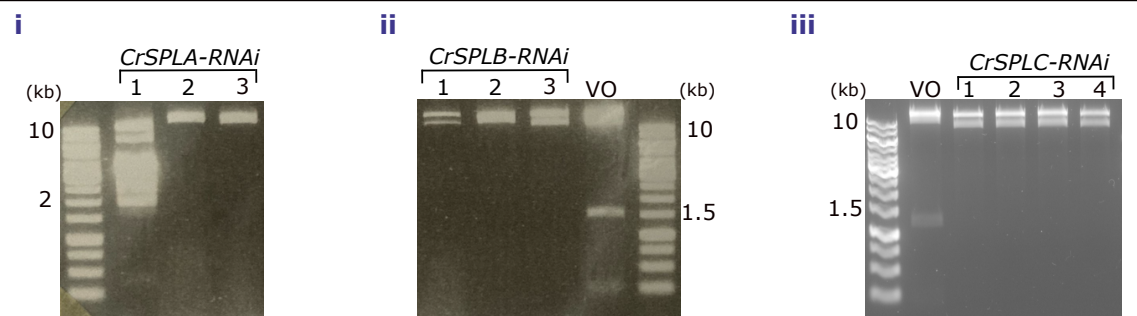


Figure 4.2: Caption overleaf.

Figure 4.2 Diagnostic gels during cloning procedure of *CrSPLA*, *CrSPLB* and *CrSPLC* RNAi target fragments. Panel A: Diagnostic digests with *Bgl*II to confirm ligation of full coding sequences of *CrSPLA* (i; 1755 bp), *CrSPLB* (ii; 1526 bp) and *CrSPLC* (iii; 1353 bp) into pJET1.2 vector (2974 bp). Panel B: PCR amplification of RNAi fragments with *attB* sites. Expected fragment sizes: (i) *CrSPLA*: 339 bp, (ii) *CrSPLB*: 439 bp, (iii) *CrSPLC*: 400b bp. Panel C: Diagnostic digestion of RNAi-pJET vector with *Bgl*II to confirm ligation of fragment for *CrSPLC* only. Target fragment size: 475 bp. Panel D: Diagnostic digestion of RNAi-pDONR207 plasmids with *Apal* and *EcoRI*. Expected fragment sizes: (i) *SPLA-RNAi*: 3651 bp, (ii) *SPLB-RNAi*: 3285 bp, 464 bp, (ii) *SPLC-RNAi*: 3332 bp, 409 bp. Vector only (VO): 3863 bp, 1722 bp. Panel E: Diagnostic digestion of RNAi-pANDA plasmids with *NotI*. Expected fragment sizes: (i) *SPLA-RNAi*: 3980 bp, (ii) *SPLB-RNAi*: 4176 bp, (iii) *SPLC-RNAi*: 4162 bp. VO: 5433 bp, 1127 bp, 168 bp.

4.2.2 Generation of T0 *CrSPL-RNAi-pANDA* transgenic plants

Following the validation of the *CrSPL-RNAi-pANDA* vectors, the constructs were transformed into *C. richardii* following the protocol established in Plackett et al., 2015. Callus was generated from young sporophytes, which was then transformed by microparticle bombardment with a biolistics gun. Callus was induced in young sporophytes 14 days after fertilisation by application of BAP, then excised and sub-cultured for a further 4 weeks to ensure enough callus growth for several bombardments (Figure 4.3A-D). Sub-cultured callus was transferred to 50 mm bombardment plates, and densely packed to ensure maximum bombardment (Figure 4.3E-F). Due to limitations of the biolistics gun, a maximum of 4 callus plates could be transformed per bombardment round; 1 no DNA control (unlabelled microparticles) and 3 construct replicates. The no DNA control was always bombarded at the start of a new bombardment round to enable screening for antibiotic selection and confirm no cross-contamination from previous bombardment rounds. Two rounds of bombardment were carried out per construct, resulting in 6 independent plates of transformed callus each for *CrSPLA-RNAi*, *CrSPLB-RNAi* and *CrSPLC-RNAi*. Each plate was deemed an independent transgenic line. In addition, 3 plates of callus were bombarded with the empty pANDA vector (*CrPANDAe*) to control for phenotype differences as a result of bombardment and/or vector insertion.

Three days after transformation, the bombarded callus was transferred to antibiotic selection. All tissue from one bombardment plate was transferred to a single 90 mm plate. Individual calli were spaced separately to allow independent regenerated shoots to be easily distinguished (Plackett et al., 2015). Calli were subjected to 8 weeks continuous antibiotic selection after which resistant regenerated sporophytes were transplanted to soil (Figure 4.3Gii). No DNA control plates showed no growth, confirming effective selection and absence of cross-contamination from earlier bombardments (Figure 4.3Gi).

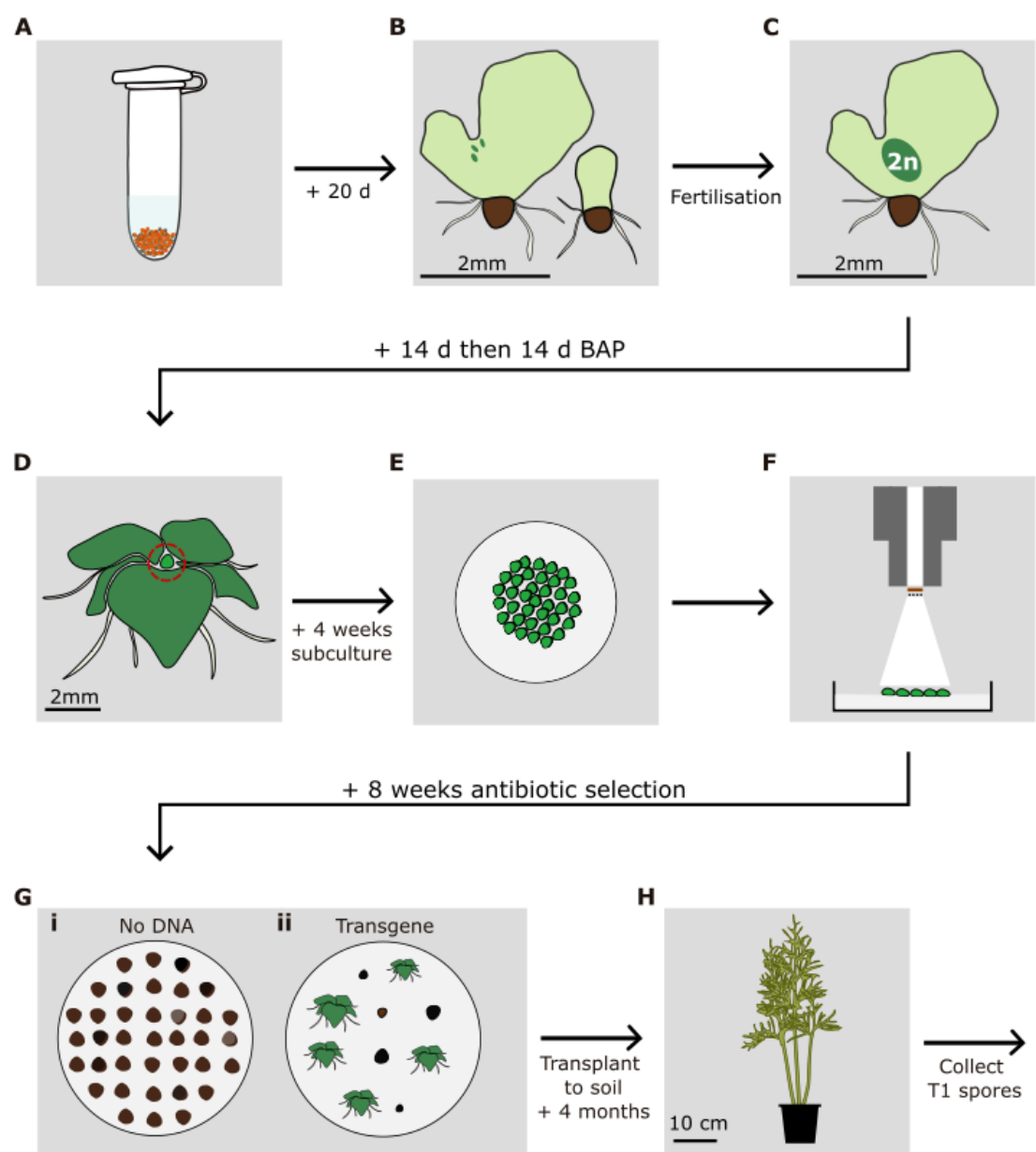


Figure 4.3: Caption overleaf.

Figure 4.3 Transformation process of *Ceratopteris richardii*. *C. richardii* WT spores are imbibed (A) and germinated on C-fern media. The germinated spore develops a two-dimensional thallus which assumes either hermaphrodite or male characteristics by sexual maturity, approximately 20 days post-imbibition (B). Fertilisation by free-mating results in the development of a diploid sporophyte on the hermaphrodite gametophyte thallus (C). The sporophyte continues to develop independently of the gametophyte and produces 3-5 simple fronds by 14 days after fertilisation. Young sporophytes are then treated with 6-Benzylaminopurine (BAP) to induce callus growth at the shoot apex in place of new fronds (D). Red dashed circle indicates callus. The resulting callus is excised and sub-cultured for 4 weeks on media containing BAP before transformation. Callus is densely packed to form a target on 50 mm plates (E) and transformed via microparticle bombardment (F) with plasmid DNA carrying the transgene of interest and a linked antibiotic selection marker, alongside unlabelled microparticles as a No DNA control. Bombarded callus is then subjected to 8 weeks of antibiotic selection to identify stable transgenics. After 8 weeks of continuous selection, all No DNA-transformed callus should be dead (Gi) and only stably transformed callus should have regenerated sporophytes (Gii). Resistant regenerated sporophytes are transplanted to soil, developing fronds of increasing size and complexity until reproductive fronds bearing T1 spores are produced at around 4 months (H).

Resistant regenerated sporophytes from six independent bombardments per transgene and 2 independent bombardments for *pANDAE* were obtained. The number of plants per line varied due to survival on selection, as well as contamination. Due to space limitations, a maximum of 11 sporophytes per bombardment replicate were transplanted to soil. Sporophyte survival on soil varied amongst lines: 95 % *CrPANDAE*, 100 % *CrSPLA-RNAi*, 58 % *CrSPLB-RNAi* and 76 % *CrSPLC-RNAi* transplanted T0 sporophytes survived to maturity. This resulted in a total of 22 mature *CrpANDAE* T0 sporophytes (Line 1: 11 plants, 91 % survival; Line 2: 11 plants, 100 % survival), 45 *CrSPLA-RNAi* mature T0 sporophytes (Lines 1, 2 and 4: 8 plants, 100 % survival; lines 3, 5 and 6: 7 plants, 100 % survival), 22 *CrSPLB-RNAi* mature T0 sporophytes (Line 1: 3 plants, 50 % survival; Line 3: 7 plants, 70 % survival; Line 4: 2 plants, 40 % survival; Line 5: 1 plant, 50 % survival; Line 6: 4 plants, 44 % survival; Line 8: 5 plants, 83 % survival), and 35 *CrSPLC-RNAi* mature T0 sporophytes (Lines 1, 2 and 4: 7 plants, 70 % survival; Line 3: 3 plants, 100 % survival; Line 5: 2 plants, 67 % survival and Line 6: 9 plants, 90 % survival).

Sporophytes began producing reproductive fronds approximately 3 months after transfer to soil, which were harvested from 4-9 months after transfer. Due to space and time limitations, generation of *CrSPLA-RNAi* and *CrSPLC-RNAi* lines terminated here with the collection of T0 reproductive fronds containing T1 spores. Work continued with *CrSPLB-RNAi* and *CrpANDAE*. *CrSPLB* was chosen based on the the expression patterns observed in Chapter 3, figure 3.5.

4.2.3 Screening T1 *CrSPLB-RNAi* transgenics

It has been suggested that newly harvested spores retain dormancy for a period of 1-2 months (Plackett et al., 2015). Therefore, the harvested T0 reproductive fronds were ripened for 2 months before collection of spores for imbibition. Not all T1 spores carry the desired transgene due to the chimeric nature of T0 sporophytes, therefore T1 individuals were screened to identify transgenics (Plackett et al., 2015). *CrSPLB-RNAi*

and *CrpANDAE* T1 spores were imbibed alongside WT spores and sowed on plates with and without antibiotic selection. WT spores did not survive germination on antibiotic plates, confirming the selection process was working. Gametophyte survival varied between transgenes. Of the 25 *CrSPLB-RNAi* stocks imbibed, 7 were not viable at the gametophyte stage i.e. did not germinate on selection or control (Figure 4.4). Of the 18 viable spore stocks, 11 were confirmed transgenic, i.e. they successfully germinated on selection. Meanwhile, all 12 of the *CrpANDAE* spore stocks imbibed were viable, but 4 did not grow on selection (Figure 4.5).

Previous studies have demonstrated that antibiotic selection renders gametophytes infertile, and therefore fertilisation was induced in the absence of selection (Plackett et al., 2014). 10-14 days after sowing, resistant hermaphrodite gametophytes and control WT hermaphrodite gametophytes were isolated in individual wells of a 24-well plate for fertilisation. Forcing self-fertilisation at this stage ensures that the resulting sporophytes will be homozygous for the transgene (Plackett et al., 2015). Once established, sporophytes were returned to selection plates for a minimum of 14 days. WT sporophytes were transferred to plates with and without selection. Resistant transgenic sporophytes were transplanted to soil alongside self-fertilised WT sporophytes. Of the 11 *CrSPLB-RNAi* spore stocks with confirmed transgenic gametophytes, 8 survived selection in the sporophyte generation (32 % transformation efficiency); one line from bombardment replicates 1 (1.6) and 5 (5.2) and two lines from bombardment replicates 3 (3.1 and 3.3), 4 (4.2 and 4.5) and 6 (6.1 and 6.9) (Figure 4.4). Of the 8 *CrpANDAE* spore stocks with confirmed transgenic gametophytes, 5 were transgenic in the sporophyte generation and subsequently transplanted to soil; four lines from bombardment replicate 1 (1.2, 1.4, 1.6 and 1.7) and one line from bombardment replicate 2 (2.4) (Figure 4.5). A maximum of 4 sporophytes per spore stock were transplanted to soil. Of those that survived to maturity, all produced spores.

		T0	T1					T2		
Line	Plant	Survived selection	Survived in soil	Spores Collected	T1 Spores inhibited	Genetophyte viable	Genetophyte transgenic	T1 sporophytes on soil	Survived to harvest?	Spores produced?
1	1									
	2									
	3									
	4									
	5									
	6									
3	1									
	2									
	3									
	4									
	5									
	6									
	7									
	8									
	9									
	10									
4	1									
	2									
	3									
	4									
	5									
5	1									
	2									
6	1									
	2									
	3									
	4									
	5									
	6									
	7									
	8									
	9									
8	1									
	2									
	3									
	4									
	5									
	6									

Figure 4.4: Survival of *CrSPLB-RNAi* transformants in the T0 to T2 generations. All transformants that survived antibiotic selection after bombardment were tracked through to the generation of T2 spores. Each line represents one independent bombardment replicate. Dark grey = positive, Light grey = negative, White = Untested due to inviability.

		T0		T1							T2		
		Line	Plant	Survived selection	Survived on soil	Spores Collected	T1 Spores imbibed	Gametophyte viable	Gametophyte transgenic	T1 sporophytes on soil	Survived to harvest?	Spores produced?	
1	1	1											
	2												
	3												
	4												
	5												
	6												
	7												
	8												
	9												
	10												
	11												
	12												
2	1												
	2												
	3												
	4												
	5												
	6												
	7												
	8												
	9												
	10												
	11												

Figure 4.5: Survival of *CrpAND*Ae transformants in the T0 to T2 generations. All transformants that survived antibiotic selection after bombardment were tracked through to the generation of T2 spores. Each line represents an independent bombardment replicate. Dark grey = positive, Light grey = negative, White = Untested due to inviability. T1 spores from Line 1, plants 10-12 and Line 2, plants 6-11 were not imbibed.

4.2.4 T1 *CrSPLB-RNAi* plants are not phenotypically different to CrWT

To investigate whether *CrSPLB-RNAi* had successfully downregulated native expression of *CrSPLB* and whether this conferred a mutant phenotype, phenotyping and expression analysis was carried out. Owing to limitations of space, plants were grown sequentially in two batches with different conditions (Batch 1 were planted immediately next to each other whilst Batch 2 were spaced with at least one empty cell between them). Therefore, separate WT controls subjected to the same conditions were analysed alongside each batch.

To confirm that *CrSPLB* expression was knocked down in *CrSPLB-RNAi* lines compared to CrWT, qPCR was performed. Two to three 130 d reproductive fiddleheads were harvested from each surviving sporophyte and RNA was extracted. RNA quality was assessed by gel electrophoresis (Figure 4.6A-B). Quality varied within and between lines, and several had very poor quality. Unfortunately, due to tissue availability, it was not possible to re-extract. Therefore, qPCR was performed using the available RNA. RNA extracted from line 6.9 was not of sufficient quality for detection by qPCR. Across all lines, *CrSPLB* expression was not significantly different to WT (ANOVA, Batch 1: $F = 2.56$, $p = 0.084$, Batch 2: $F = 3.36$, $p = 0.164$) (Figure 4.6C-D). *CrSPLB* expression was indistinguishable from WT in lines 1.6 ($p_{\text{adj}} = 0.99$), 3.1 ($p_{\text{adj}} = 0.99$) and 4.2 ($p_{\text{adj}} = 0.99$). Lines 3.3, 5.2 (Figure 4.6C) and 6.1 (Figure 4.6D) appeared to have the most convincing down-regulation of expression, however this was not statistically significant ($p_{\text{adj}} = 0.89, 0.98, 0.16$, respectively). Meanwhile line 4.5 had 2.5-fold higher expression than WT, though this was also not statistically significant ($p_{\text{adj}} = 0.164$) (Figure 4.6A).

AtSPL expression is necessary for meiosis to occur in the *Arabidopsis* anther and ovule (Mendes et al., 2020). In WT *C. richardii*, meiosis occurs in the sporangia of reproductive fronds to produce haploid spores. Phenotyping effort was therefore focussed on the mature reproductive fronds of *CrSPLB-RNAi* plants. Due to limited

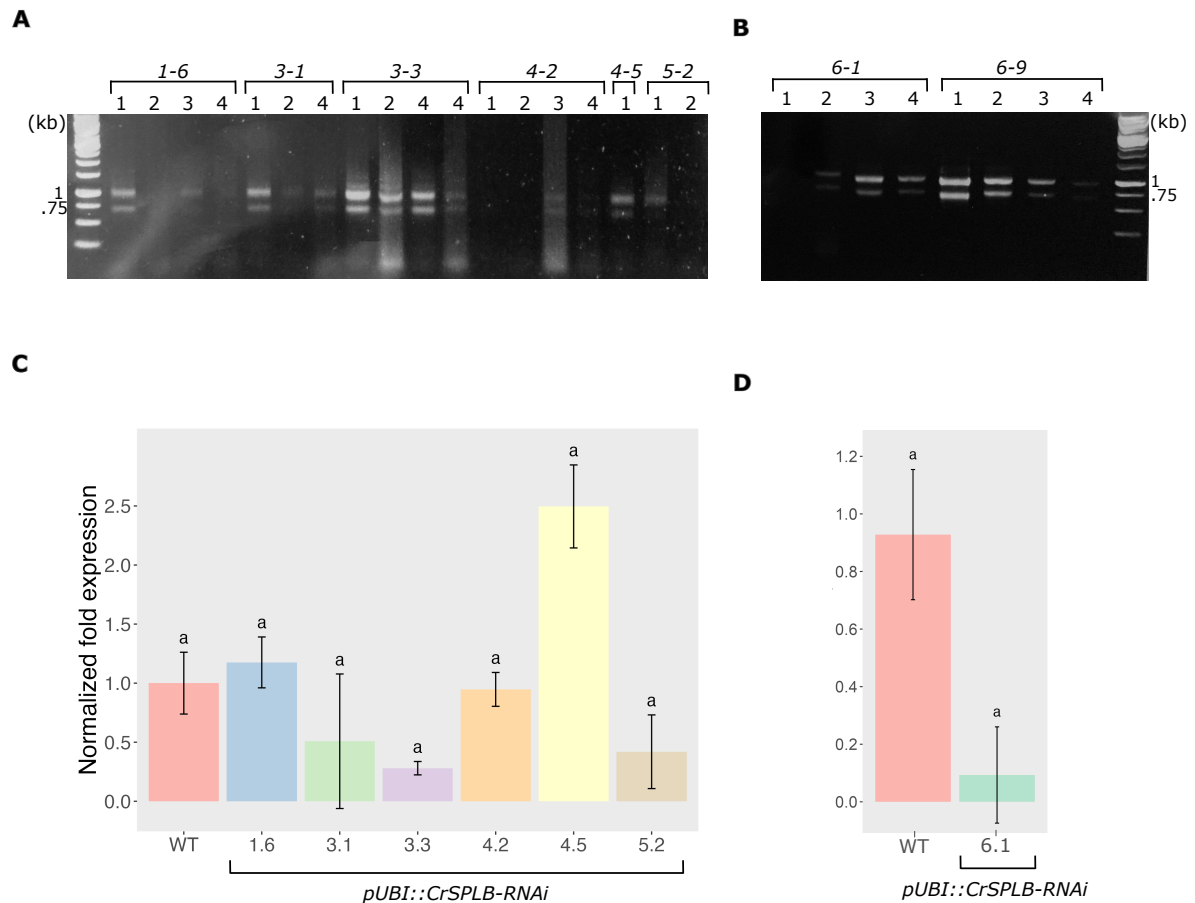


Figure 4.6: *CrSPLB* expression in the reproductive fiddleheads of T1 *CrSPLB-RNAi* lines relative to *CrWT*. (A-B) Quality of RNA extracted from the fiddleheads of individual plants of batch 1 (A) and batch 2 (B) transgenics for use in qPCR expression analysis. (C-D) Relative expression of *CrSPLB* across *CrSPLB-RNAi* transgenic lines detected by qPCR. Independent qPCR analyses were carried out for each batch. Ubiquitin was used as housekeeping gene. Error bars represent standard error of the geometric mean. ANOVA (C: $F = 2.56$, $p = 0.084$; D: $F = 3.36$, $p = 0.164$) with Tukey's HSD post-hoc pairwise comparisons, adjusted for multiple comparisons with Generalised Linear Hypothesis Testing (GLHT) ($p > 0.05$ in all comparisons).

tissue availability, one reproductive frond per plant was selected for further analysis. Frond length was measured and the number of branches and pinnae counted (See Figure 2.5).

Across all lines, frond length did not significantly differ to WT (ANOVA (Batch 1, 2): $F = 4, 0.906; p = 0.0172, 0.438$) (Figure 4.7 A-B). ANOVA found that the number of branches per frond was significantly different between fronds of Batch 1 (Figure 4.7C; $F = 5.274, p = 0.00587$) but not Batch 2 ($F = 0.115, p = 0.892$). Post-hoc pairwise comparisons with Tukey's Honest Significant Difference (HSD) adjusted for multiple comparisons (Generalised Linear Hypothesis Test; GLHT) found that line 3.3 only was significantly different to WT. Line 3.3 had significantly more branches than WT ($p_{\text{adj}} = 0.0141$) (Figure 4.7C). The number of pinnae per frond was also significant between lines of Batch 1 (Figure 4.7E; ANOVA: $F = 3.73, p = 0.0216$) but not Batch 2 (Figure 4.7F; ANOVA: $F = 1.409; p = 0.294$). Post-hoc pairwise comparisons with Tukey's HSD (GLHT adjusted) again found that Line 3.3 only was significantly higher than WT ($p_{\text{adj}} = 0.038$). Of note, Line 4.5 was consistently lower than WT and other lines, though this was not statistically significant.

To investigate whether *CrSPLB-RNAi* insertion had an effect on sporangia development, 3 representative pinnae per frond were microscopically analysed. Each selected pinnae was measured, and the number of mature sporangia inside was counted. To account for variation in pinnae length, the number of sporangia was divided by pinnae length to allow comparison to be made between lines. A Kruskal-Wallis test found that the number of sporangia per pinnae significantly differed amongst lines of Batch 1 ($\chi^2 = 30.922, p = 2.623 \times 10^{-5}$) but not Batch 2 ($\chi^2 = 2.3967, p = 0.3017$). Post-hoc pairwise comparisons with Wilcoxon rank sum test adjusted for multiple comparisons with Bonferroni found that Line 5.2 only was significantly different to WT ($p = 0.0158$). Line 5.2 had significantly more sporangia per pinnae than WT.

Whilst largely not statistically significant, some general patterns can be drawn from

the expression and phenotyping data when taken together. Lines 3.3 and 5.2 had that most convincing downregulation of *CrSPLB* expression compared to WT (Figure 4.6C), and both had more sporangia per pinnae than WT (Figure 4.7G). However, whilst 3.3 had significantly more branches and pinnae, 5.2 was not significantly different to *CrWT* (Figure 4.7C, E). Meanwhile, Line 4.5 had the highest fold change in *CrSPLB* expression (not significant) and interestingly displayed the opposite phenotype to lines 3.3 and 5.2: consistently lowest frond length (Figure 4.7A), branches (Figure 4.7C), pinnae (Figure 4.7E) and sporangia per pinnae (Figure 4.7G) (NS).

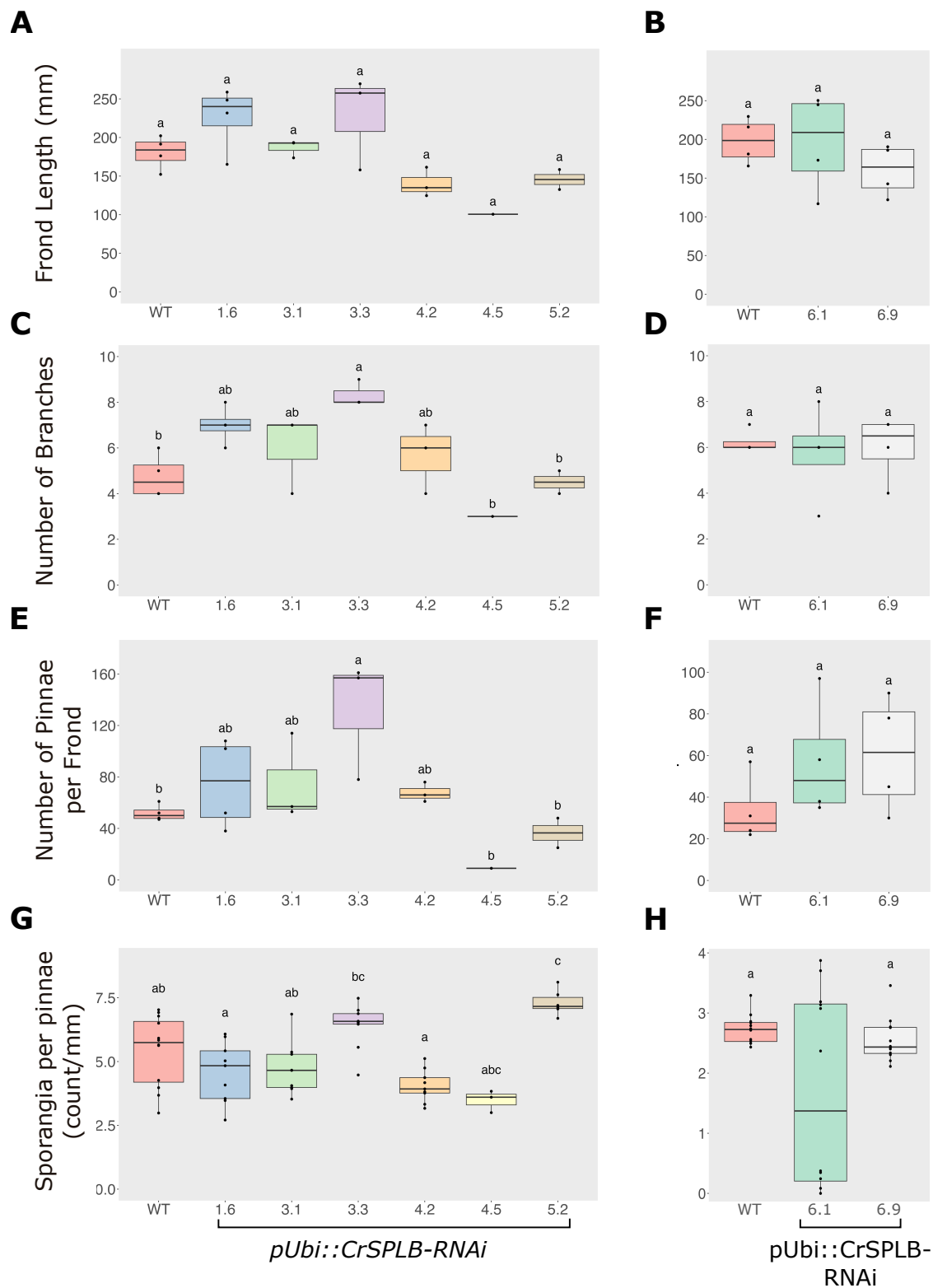


Figure 4.7: Caption overleaf.

Figure 4.7 Phenotyping analysis of mature reproductive fronds (130d) of *CrSPLB-RNAi* lines batch 1 (A, C, E, G) and batch 2 (B, D, F, H). One frond per transgenic plant was harvested for analysis, equating to 4 replicates per line. (A-B) Length of chosen frond from base to tip measured in mm. ANOVA (F = 4 (A), 0.906 (B); P = 0.0172 (A), 0.438 (B)) with Tukey's HSD adjusted for multiple comparisons with GLHT (p > 0.05 in all comparisons to WT). (C-D) Number of branches emerging from the main stem of chosen fronds. ANOVA (F = 5.274 (C), 0.115 (D); p = 0.00587 (C), 0.892 (D) with Tukey's HSD (GLHT) (p > 0.05 in all comparisons to WT except line 3.3 (p = 0.0141)). (E-F) Total number of pinnae on chosen frond. ANOVA (F = 3.73 (E), 1.409 (F), p = 0.0216 (E), 0.294 (F)) with Tukey's HSD (GLHT) (p > 0.05 in all comparisons to WT except 3.3 (p = 0.038)). (G-H) Three pinnae per frond were chosen for further analysis. The number of mature sporangia was counted and pinnae length recorded. Number of sporangia/ pinnae length is displayed. Kruskal-Wallis ($p = 2.623 \times 10^{-5}$ (G), 0.317 (H)) with a pairwise Wilcoxon rank-sum test with Bonferroni correction (p > 0.05 in all comparisons to WT except line 5.2 (p = 0.0158)).

4.2.5 Cross-complementation of the *Arabidopsis spl-4* mutant with *CrSPLs*

To further investigate whether *CrSPLs* have a similar reproductive function to *Arabidopsis* SPL, functional cross-complementation was attempted. The aim of this section was to build complementation constructs containing each of the *CrSPLs* as well as *AtSPL* control, and subsequently express them in the *spl-4* knockout line (Andrew Plackett, personal communication) to assess rescue of SPL function. Mendes et al., 2020 previously achieved full complementation of *spl* fertility defects with their *pSPL5'::SPL-GFP::3'* construct. The construct in this thesis was therefore designed to include the same 3906 bp region upstream of *SPL* and 1609 bp region downstream of *SPL*. Two constructs were designed, one tagged with mCHERRY (*pSPL::mCHERRY::SPL3'UTR*) and one untagged (*pSPL::SPL3'UTR*).

Constructs were assembled using the *35S::OCS-pART7* vector as the starting template (Figure 4.8A). The *35S::OCS* cassette is flanked by *NotI* restriction sites at the 5' and 3' ends, allowing the entire cassette to be excised and inserted into the *pART27* vector for expression *in planta*. A series of additional restriction sites between the *35S* promoter and *OCS* terminator enable modular modification of the cassette (Figure 4.8A). To preserve these internal sites, the assembly strategy was designed to work from the outermost sites inward, minimizing the risk of losing critical restriction sites during cloning. The choice of restriction sites and the order of construct assembly was further influenced by the distribution of restriction sites in the DNA fragments (Table 4.1). *KpnI* was specifically reserved for inserting the coding sequence (CDS) because none of the fragments contained this site, making it ideal for ensuring compatibility without unintended cuts.

Table 4.1: Restriction site occurrences for pART7 and pART27 plasmids.

	BamHI	EcoRI	HindIII	KpnI	XbaI	XhoI	NotI
pSPL	1	2	3	0	2	0	0
SPL3'UTR	0	0	0	0	0	0	0
AtSPL	0	0	1	0	0	1	0
CrSPLA	0	1	4	0	0	0	0
CrSPLB	3	1	0	0	0	0	0
CrSPLC	0	1	0	0	0	0	0
mCHERRY	1	0	0	0	0	0	0

The assembly process is outlined in Figure 4.8. First, the OCS terminator was replaced with the *AtSPL* 3' UTR using *XbaI* and *NotI* restriction sites (Figure 4.8B). Then, the *mCHERRY* tag was inserted between the 35S promoter and *AtSPL* 3'UTR using *KpnI* and *XbaI* restriction sites (Figure 4.8C). 35S was replaced by *pSPL* in both the tagged (Figure 4.8D) and untagged (Figure 4.8F) constructs using *NotI* and *KpnI* sites. Finally, the *SPL* CDS without a stop codon (*SPL*ΔSTOP) was inserted between *pSPL* and *mCHERRY* in the tagged construct via *KpnI* sites (Figure 4.8E). *SPL* with a stop codon was inserted into the untagged vector using the *KpnI* restriction site (Figure 4.8G).

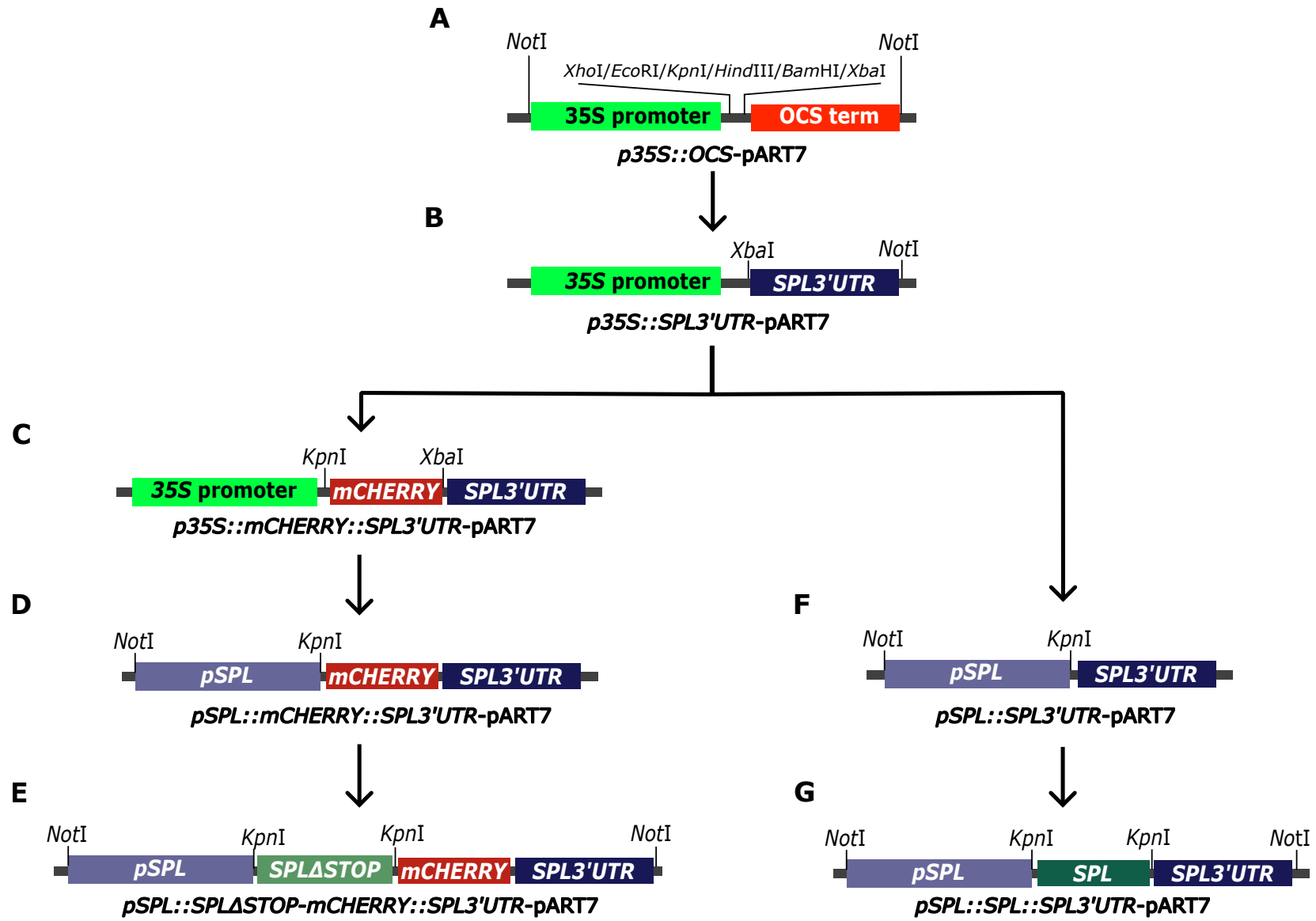


Figure 4.8: Caption overleaf

Figure 4.8 Schematic of the restriction cloning steps undertaken to generate tagged *pSPL::SPL-mCHERRY::SPL3'UTR-pART7* and untagged *pSPL::SPL::SPL3'UTR-pART7* complementation vectors. (A) The pART7 vector contains a 35S promoter and OCS terminator, flanked by *NotI* sites at the 5' and 3' ends, facilitating cassette excision for ligation into pART27. A series of restriction sites is positioned between the promoter and terminator. (B) The OCS terminator was replaced with the *SPL* 3'UTR by digestion with *XbaI* and the 3' *NotI* site. (C) *mCHERRY* was inserted upstream of the *SPL* 3'UTR via digestion with *KpnI* and *XbaI*. (D and F) The 35S promoter was replaced by *pSPL* by digestion with the 5' *NotI* site and *KpnI* in both the tagged (C) and untagged (F) cassettes. (E) *SPL* without a STOP codon (*SPL*ΔSTOP) was inserted between *pSPL* and *mCHERRY* via *KpnI* digestion. Removal of the STOP codon enables transcription to continue into *mCHERRY*. (G) *SPL* with a STOP codon was inserted into the untagged cassette via *KpnI* digestion.

Before the complementation constructs could be assembled, the promoter, terminator, tag and CDS fragments were amplified with the relevant restriction sites attached (Figure 4.9). Each fragment was first amplified and ligated into pJET1.2, and correct insertion confirmed via diagnostic digests with *Bgl*II (Figure 4.9A, D, G, J) and subsequent sequencing. The confirmed fragments were then used as template in a second PCR to add the relevant restriction sites at the 5' and 3' ends and again ligated into pJET1.2. Addition of complete restriction sites was confirmed with digestion using the respective restriction enzymes (Figure 4.9B, E, H, K). The resulting fragments with their 5' and 3' restriction sites are depicted in Figure 4.9C, F, I, L and M.

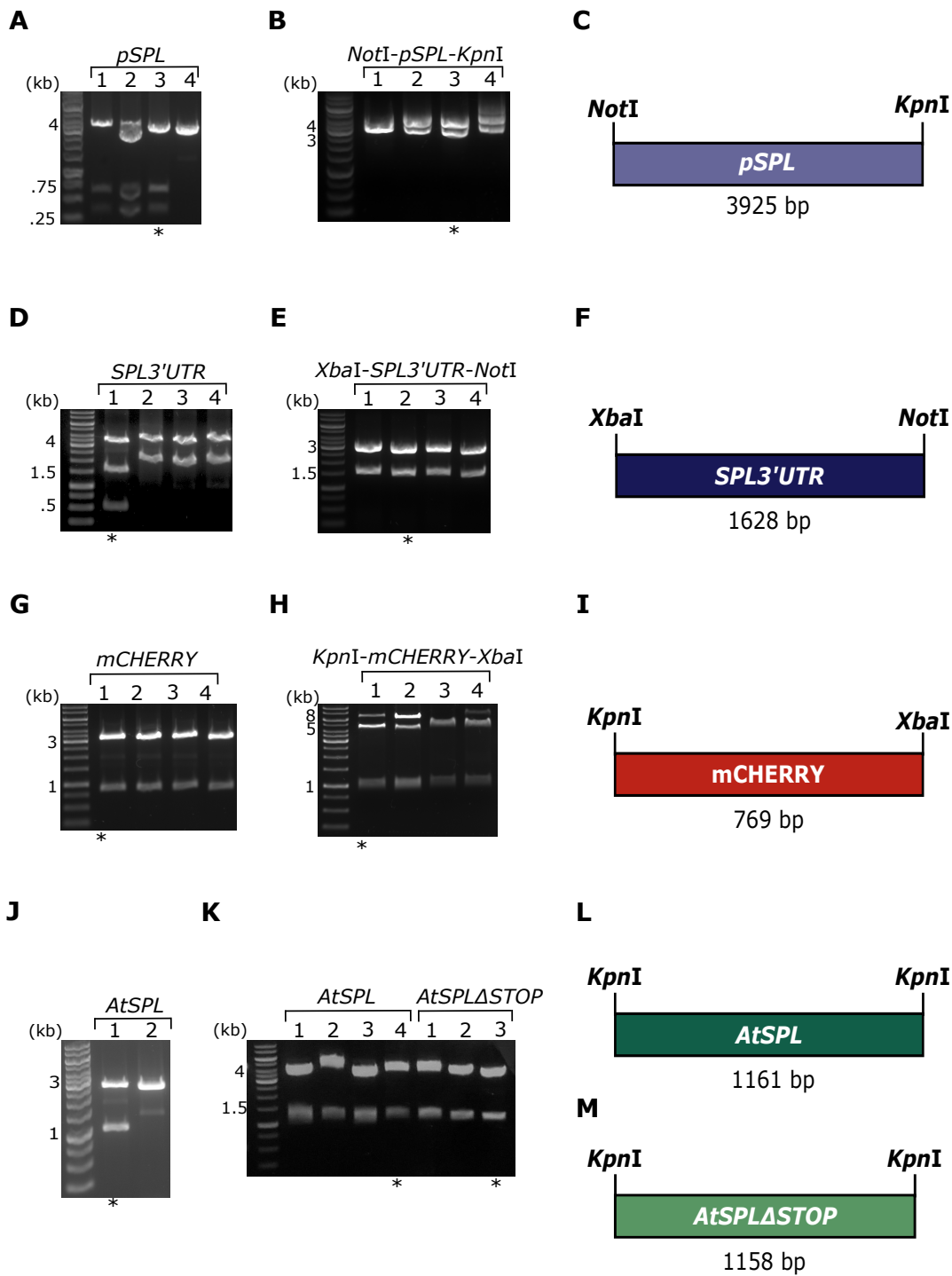
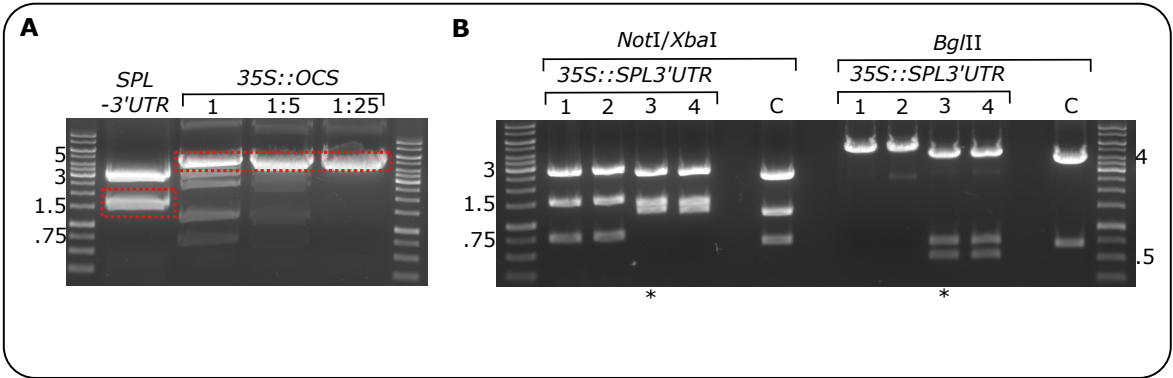


Figure 4.9: Caption overleaf

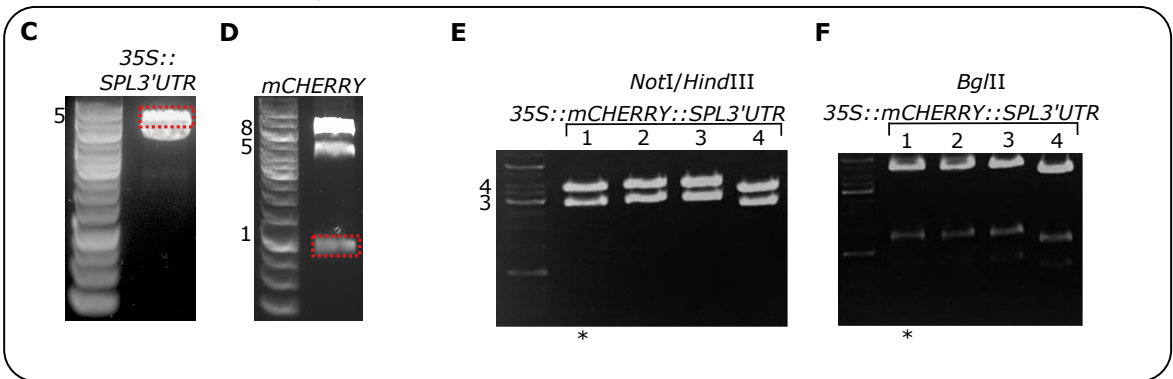
Figure 4.9 Cloning fragments for the complementation cassettes. (A-C) *SPL* promoter (*pSPL*) with 5' *NotI* and 3' *KpnI* sites. (A) 3906 bp of the 5' region directly upstream of the *SPL* start codon was amplified to capture *pSPL* and ligated into pJET. Digestion with *BglII* confirmed that *pSPL*-3 was successfully inserted (Correct band pattern: 2942 bp, 610 bp, 354 bp). (B) *pSPL*-3 was re-amplified with primers to add 5' *NotI* and 3' *KpnI* sites, then ligated into pJET. Digestion with *NotI* and *KpnI* confirmed that the sites were present in minis 2-4 (Correct band pattern: 3825 bp, 2975 bp). (C) *pSPL* with restriction sites. (D-F) *SPL*3'UTR with 5' *XbaI* and 3' *NotI* sites. (D) 1609 bp directly downstream of the *SPL* stop codon was amplified and ligated into pJET. Digestion with *BglII* confirmed that *SPL*3'UTR-1 was successfully inserted (Correct band pattern: 2928, 1343, 440). (E) *SPL*3'UTR-1 was re-amplified to add 5' *XbaI* and 3' *NotI* sites. Digestion with *XbaI* and *NotI* confirmed that the sites were present in all minipreps (Correct band pattern: 2925, 1786). (F) *SPL*3-UTR with restriction sites. (G-I) *mCHERRY* with 5' *KpnI* and 3' *XbaI* sites (G) *mCHERRY* was amplified and ligated into pJET. Digestion with *BglII* confirmed that *mCHERRY* was successfully inserted across all minipreps (Expected band sizes: 2928, 792). (H) *mCHERRY* was re-amplified to add *KpnI* and *XbaI* sites. Digestion with *KpnI* and *XbaI* confirmed that the sites were present in all minipreps (Correct band pattern: 2972, 756, 10, 5). (I) *mCHERRY* with restriction sites. (J-M) *AtSPL* and *AtSPL* without a stop codon (*AtSPL*ΔSTOP) with 5' and 3' *KpnI* sites. (J) The full *AtSPL* coding sequence (CDS) was amplified and ligated into pJET. Digestion with *BglII* confirmed that *AtSPL* was successfully inserted in mini 1 (Expected band sizes: 2974, 1141). (K) *AtSPL* was re-amplified to add *KpnI* sites to the full CDS and the CDS without a stop codon. Digestion with *KpnI* confirmed that sites were present (Expected band sizes for *AtSPL*/*AtSPL*ΔSTOP: 2972, 1161/1158). (L) *AtSPL* with restriction sites. (M) *AtSPL*ΔSTOP with restriction sites.

Once the fragments required were obtained, assembly of the tagged and untagged complementation cassettes began. First, the *OCS* terminator was replaced with *SPL3'UTR* via restriction digest with *NotI* and *KpnI* (Figure 4.10A-B). *35S::OCS* contains two *NotI* restriction sites, and therefore a partial digest was carried out to recover all possible fragments (Figure 4.10A). Partial digest simulation indicated that the desired band was the largest of five bands, and therefore the topmost band was gel excised for ligation with *SPL3'UTR*. The product was transformed into *DH5a* and miniprep. Diagnostic digests with *NotI/XbaI* and *BglII* confirmed that 2 out of 4 minipreps were correct (Figure 4.10B). Next, *35S::SPL3'UTR* miniprep 3 and *mCHERRY* were digested with *KpnI* and *XbaI* and the desired fragments excised, ligated together and transformed into *DH5a* (Figure 4.10C, D). Diagnostic digests with *NotI/HindIII* and *BglII* confirmed that the intended fragments had been successfully ligated together (Figure 4.10E, F). In the final step, the *35S* promoter was substituted with *pSPL* in both the tagged and untagged constructs. *35S::mCHERRY::SPL3'UTR* from the previous step and *35S::SPL3'UTR* from step one were partially digested with *NotI* and *KpnI* (Figure 4.10G). The topmost fragments containing the full backbone without *35S* were gel excised and ligated with digested *pSPL*. Confirmation that the correct constructs had been obtained was achieved by diagnostic digest with *XbaI* and *EcoRI* (Figure 4.10H,I).

Substitution of *OCS* terminator with *SPL3'UTR* to generate *35S::SPL3'UTR* construct



Insertion of *mCHERRY* to generate *35S::mCHERRY::SPL3'UTR* construct



Substitution of *35S* promoter with *pSPL* to generate tagged *pSPL::mCHERRY::SPL3'UTR* and untagged *pSPL::SPL3'UTR* constructs

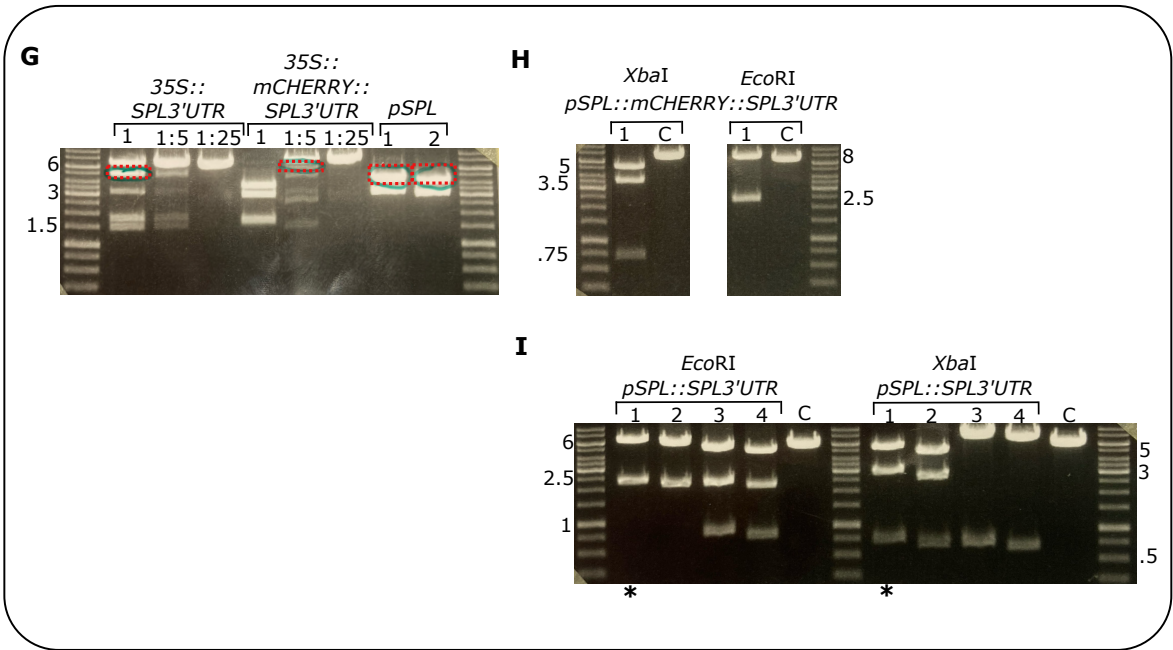


Figure 4.10: Caption overleaf

Figure 4.10 Step-wise construction of the tagged and untagged complementation vectors *pSPL::mCHERRY::SPL3'UTR* and *pSPL::SPL3'UTR*.

(A-B) Substitution of OCS terminator in 35S::OCS-pART7 for *SPL3'UTR*. (A) *SPL3'UTR*-pJET was digested and 35S::OCS-pART7 was partially digested with *NotI* and *KpnI*. Serial dilution of *NotI* was used in partial digest. Target fragments of size 1615 bp (*SPL3'UTR*) and 4365 bp (35S-pART7) were excised and ligated. (B) Diagnostic digests to confirm ligation of 35S::*SPL3'UTR*-pART7, minis 3 and 4 are correct. Expected bands = *NotI/XbaI*: 2864 bp, 1615 bp, 1401 bp; *BglII*: 4600 bp, 746 bp, 534 bp. (C-F) Insertion of *mCHERRY* to generate 35S::*mCHERRY::SPL3'UTR*-pART7 construct. (C) Digestion of 35S::*SPL3'UTR*-pART7 with *XbaI* and *KpnI*. Target fragment = 5845 bp. (D) Digestion of *mCHERRY*-pJET with *XbaI* and *KpnI*. Target fragment = 756 bp. (E-F) Diagnostic digests to confirm insertion of *mCHERRY*. (E) Digest with *NotI* and *HindIII*. Expected band pattern = 3737 bp, 2664 bp. (F) Digest with *BglII*. Expected band pattern = 4600 bp, 1255 bp, 746 bp. (G-I) Substitution of the 35S promoter for *pSPL* in both 35S::*mCHERRY::SPL3'UTR*-pART7 and 35S::*SPL3'UTR*-pART7 constructs. (G) Partial digest of 35S::*SPL3'UTR* and 35S::*mCHERRY::SPL3'UTR* with *NotI* and *KpnI*. Desired fragment of 4506 bp and 5227 bp were excised. Digest of *pSPL*-pJET with *NotI* and *KpnI* - 3926 bp fragment was excised. (H) Diagnostic digest of *pSPL::mCHERRY::SPL3'UTR* to confirm substitution of *pSPL* with *XbaI* (Expected fragments = 5000 bp, 3472 bp, 677 bp) and *EcoRI* (Expected fragments = 6876, 2273). (I) Diagnostic digest of *pSPL::SPL3'UTR* with *EcoRI* (Expected fragments = 6155 bp, 2273 bp) and *XbaI* (Expected fragments = 5000 bp, 2751 bp, 677 bp).

C = control donor vector. * = miniprep that was confirmed correct and taken forward.

To verify the function of the *pSPL::SPL3'UTR* and *pSPL::mCHERRY::SPL3'UTR* constructs, the native *AtSPL* CDS was ligated into the constructs with a view to confirm rescue of function in the *spl-4* mutant (Figure 4.11). *pSPL::SPL3'UTR*, *pSPL::mCHERRY::SPL3'UTR*, *AtSPL* and *AtSPL Δ STOP* were digested with *KpnI* and the target bands excised (Figure 4.11A). Successful ligation of *AtSPL* to form the construct *pSPL::AtSPL::SPL3'UTR* was confirmed by diagnostic digest with *XhoI/EcoRI* (Figure 4.11B). Insertion of *AtSPL* in the forward orientation was further confirmed by PCR amplification (Figure 4.11D). Miniprep 2 was correct and in the forward orientation. Meanwhile, diagnostic digestion of *pSPL::AtSPL Δ STOP-mCHERRY::SPL3'UTR* with *XhoI* and *EcoRI* found no correct minipreps (Figure 4.11C). Further digests with *HindIII* and *SacI* confirmed that ligation had not been successful (data not shown). To further confirm the lack of the correct construct, PCR with internal primers was performed (Figure 4.11E). Amplification of the expected fragments was not achieved. Following several more attempts to insert *AtSPL Δ STOP*, the decision was taken to abandon construction of the tagged fusion protein and continue with the untagged version only.

To deliver the untagged complementation construct into *Arabidopsis*, insertion of the expression cassette into pART27 was required. Countless attempts were made; the amount of DNA digested with *NotI* was adjusted several times; the ligation ratio of vector:insert was adjusted; the *NotI* enzyme and ligation enzymes were replaced; three different commercial *E.coli* strains were used for transformation. However, following months of attempts, ligation into pART27 was not successful and this arm of the project was abandoned.

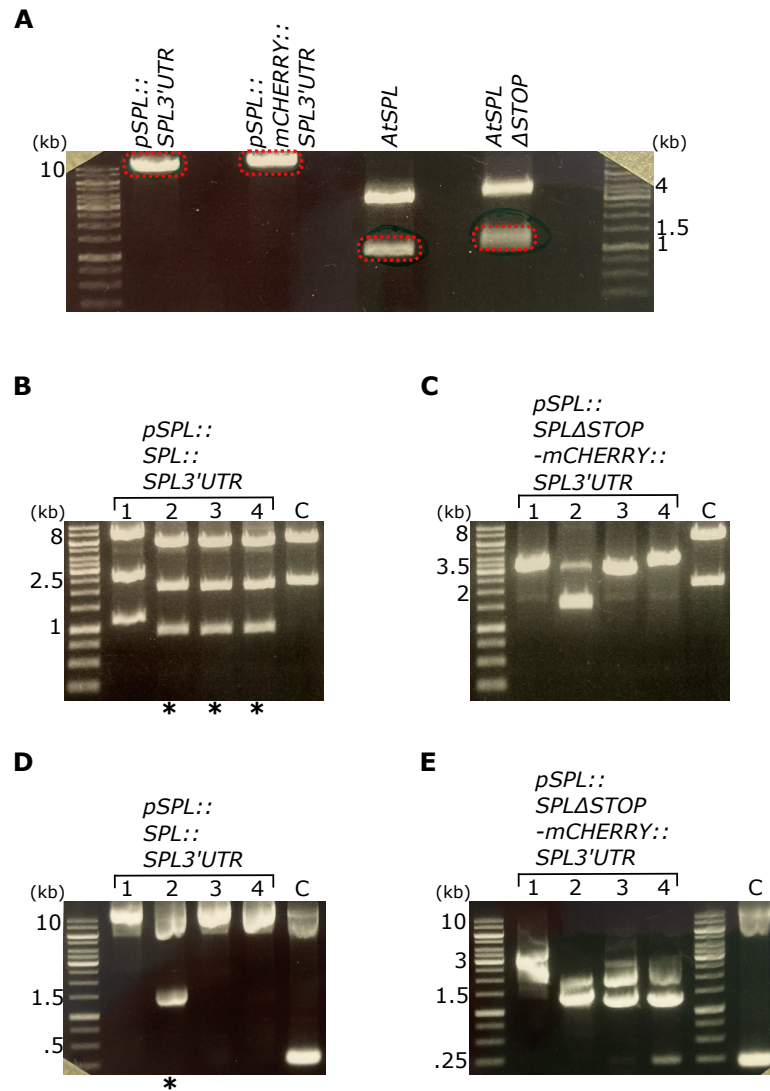


Figure 4.11: Insertion of *AtSPL* and *AtSPL*ΔSTOP into *pSPL::SPL3'UTR-part7* and *pSPL::mCHERRY::SPL3'UTR-pART7*. (A) Digestion with *KpnI* of *pSPL::SPL3'UTR* (8399 bp), *pSPL::mCHERRY::SPL3'UTR* (9149 bp), *AtSPL* (1154 bp) and *AtSPL*ΔSTOP (1151 bp). Target bands (highlighted by red dotted boxes) were excised. (B and C) Diagnostic digests to confirm correct insertion. (B) Diagnostic digest of *pSPL::AtSPL::SPL3'UTR-pART7* with *XhoI* and *EcoRI* to confirm insertion. Expected band pattern = 6199 bp, 2273 bp, 1075 bp. Minis 2 to 4 were correct. (C) Diagnostic digest of *pSPL::SPLΔSTOP-mCHERRY::SPL3'UTR-pART7* with *XhoI* and *EcoRI*. Expected band pattern = 6649 bp, 2273 bp, 1075 bp. None were correct. (D and E) PCR amplification with *AtSPL* forward primer and *SPL3'UTR* reverse primer to confirm correct orientation of insert. (D) *pSPL::SPL::SPL3'UTR*, desired amplicon = 1657 bp. Mini 2 is in correct orientation. (E) *pSPL::SPLΔSTOP-mCHERRY::SPL3'UTR-pART7*, desired amplicon = 2407 bp.

4.3 Discussion

This chapter aimed to investigate the functions of *CrSPLA*, *B* and *C* in *C. richardii* and furthermore determine whether these genes are capable of complementing *spl* knockout in *Arabidopsis*. Due to time, space and experimental limitations, these goals were not fully achieved. However, this work provides a solid ground on which to build. RNAi knockdowns were generated for all three *CrSPLs*, transformed into *Ceratopteris* and generated up until at least T1 spores. *CrSPLB-RNAi* and *CrpANDAE* were taken through to T2 spores. Preliminary analysis of *CrSPLB-RNAi* T1 mature plants did not garner a consistent mutant phenotype, however provided information about plants that should be further investigated in the T2 generation. Meanwhile, tagged and untagged complementation constructs were successfully designed and built, though unfortunately delivery into *Arabidopsis* was not achieved.

4.3.1 *CrSPL-RNAi* transformation efficiency was low

Knockdown of *C. richardii* native gene expression has previously been achieved *via* microparticle bombardment with the pANDA vector expressing double stranded RNA (dsRNA) against a target gene (Plackett et al., 2015, 2014). Three pANDA vectors containing RNAi target fragments to target each of the three *CrSPLs* were designed, constructed and delivered into callus *via* microparticle bombardment. Sporophytes that regenerated on selection were transplanted to soil and grown to maturity. From bombardment through to transplantation, many factors can influence survival: callus is not transformed and therefore dies on selection; callus is transformed but unable to regenerate sporophytes; sporophytes are unable to survive on soil. Therefore, the final number of transformants that survived to production of T1 spores was low and varied amongst lines e.g. a total of only 22 plants from 6 independent bombardments produced T1 spores from the initial ~220 bombarded calli in the line *CrSPLB-RNAi* - approximately 10 % efficiency. This is a much lower efficiency than the reported

71.58±2.56 % mean transformation efficiency (Plackett et al., 2014).

As discussed in the introduction to this chapter, there are two competing methods for transforming *C. richardii*: *Agrobacterium*-mediated transformation (Bui et al., 2015) and microparticle bombardment (Plackett et al., 2014), each with distinct advantages and limitations. Since the initial publications, several independent labs have adopted and optimised *Agrobacterium*-mediated transformation of *C. richardii* gametophytes (Bui et al., 2017; Jiang et al., 2024; Withers et al., 2023), whilst several others continue to opt for microparticle bombardment (Geng et al., 2022; Plackett et al., 2018; Xiang & Li, 2024). This is perhaps owing to the low reported efficiency of *Agrobacterium* transformation in the original publication (6.4 - 10.4 % transformation efficiency when taking account of the 1:4 ratio of hermaphrodite to males). However this is a similar efficiency to that achieved in this thesis, without the lengthy process of callus induction, and with a much reduced chance of multi-copy insertions (Bui et al., 2015). Given the very low efficiency achieved in this thesis, the reduced complexity and higher precision of *Agrobacterium* transformation make it a preferable method for future applications.

4.3.2 *CrSPLB-RNAi* T1 survival and transgenic efficiency was low

Due to limitations of time and space, *CrSPLA-RNAi* and *CrSPLC-RNAi* investigation terminated with the collection of T1 spores. *CrSPLB-RNAi* was prioritised due to findings in Chapter 3 that indicated *CrSPLB* was the most likely *SPL* candidate, namely its position within the *SPL* clade, and its specific expression in reproductive tissues in RT-PCR analysis. Work continued with *CrSPLB-RNAi* and the *CrpANDAE* empty vector control.

Transformation *via* microparticle bombardment has been shown to produce chimeric T0 transgenics, meaning that the T1 progeny will be a mix of transformed and untransformed individuals (Plackett et al., 2014). Therefore, transgene inheritance of T1 spores harvested from 25 *CrSPLB-RNAi* T0 individuals and 12 *CrpANDAE* T0

individuals was assessed by antibiotic selection. 11 of the 25 *CrSPLB-RNAi* individuals were confirmed transgenic in the gametophyte generation, whilst 8 of the 12 *CrpANDAE* individuals grew on selection. Of the 14 non-transgenic *CrSPLB-RNAi* individuals, 7 did not germinate even on control media, whilst all tested *CrpANDAE* spore stocks were viable. This could imply that these 7 *CrSPLB-RNAi* lines were transgenic, but that expression of the transgene rendered the spores inviable. However, the transformants are unlikely to be stable at this stage and it was not possible to test this hypothesis. Work continued with the viable, resistant gametophytes.

Early studies suggested that fern genes silenced by RNAi often reactivate after fertilisation of the gametophyte (Rutherford et al., 2004). Therefore, after fertilisation of resistant T1 gametophytes, T1 sporophytes were re-selected. 8 of the 11 *CrSPLB-RNAi* lines and 5 of the 8 *CrpANDAE* lines that were resistant in the gametophyte generation were also resistant in the sporophyte generation. Again, it is difficult to determine whether this was as a consequence of the transformation and selection process, or the transgenes themselves preventing fertilisation.

4.3.3 *CrSPLB-RNAi* were not phenotypically different to WT

Phenotyping and expression analysis of *CrSPLB-RNAi* was carried out to assess whether native *CrSPLB* was successfully downregulated and if this conferred a mutant phenotype. Expression analysis *via* qPCR indicated that *CrSPLB* expression in the *CrSPLB-RNAi* lines was not significantly different to *CrWT*. This could be due to several interrelated factors. Firstly, in this qPCR, we assumed that individuals arising from the same T0 spore stock represent replicates, however in the T1 generation this may not be the case; sister plants may have arisen from gametophytes with independent insertions (Plackett et al., 2014). Therefore, within-line variation likely influenced qPCR analysis. Furthermore, whilst we assume that individuals are homozygous at this stage, we cannot be certain of this until the T2 generation, and therefore variation may

occur due to heterozygosity for the transgene (Plackett et al., 2015).

Secondly, in the T1 generation we have not necessarily achieved stable homozygous integration of the transgene; microparticle bombardment can cause severe damage to chromosomes that may require several generations to repair (Liu et al., 2019). Third, particle bombardment commonly leads to multiple insertions of the transgene, which can additionally be fragmented upon entry, leading to the insertion of multiple, partial fragments (Liu et al., 2019). Therefore lines that survived selection may only contain a functional selection marker, but not a working transgene. To mitigate these limitations, genotyping and southern blot analysis would be necessary to verify that the full cassette has been inserted and determine copy number (Plackett et al., 2015). Furthermore, expression analysis in the T2 generation would be a more reliable indicator of true down-regulation of expression.

Nevertheless, whilst *CrSPLB* expression in *CrSPLB-RNAi* was not significantly different to *CrWT*, two lines did potentially have slightly downregulated expression: 3.3 and 5.2 (Figure 4.6C). Interestingly, both of these displayed a similar phenotype with regard to number of sporangia per pinnae. Line 5.2 had significantly more sporangia per pinnae than WT, whilst line 3.3 had more sporangia but this was not significant. This could imply that *CrSPLB* is responsible for the negative regulation of sporangia development. Therefore, lines 3.3 and 5.2 are good candidates for further analysis in the T2 generation. To determine whether *CrSPLB-RNAi* confers a reproductive phenotype, a more in-depth analysis of meiosis would be necessary. Determination of spores per sporangia, spore viability, and gametophyte development, such as those conducted by Withers et al. (2023), would provide a more in-depth view into whether *CrSPLB* has a conserved function in meiosis.

Studies in cucumber demonstrated that knockdown, rather than knockout, of *CsSPL* expression did not completely abolish reproductive function (Liu et al., 2018). Recently, the use of CRISPR/Cas9 for complete gene knockout has been successfully applied in *C. richardii* (Xiang & Li, 2024). Employing this method to entirely eliminate *CrSPLB*

expression could provide a more definitive assessment of whether *CrSPLB* plays an equally critical role. In contrast, RNAi knockdown may leave residual gene expression, as suggested by qPCR analysis in Figure 4.6, potentially allowing partial function and obscuring the full impact of gene loss.

4.3.4 Insertion of complementation cassettes into the plant transformation vector pART27 was not successful

In this thesis, tagged and untagged complementation cassettes were designed and built. Unfortunately, insertion of the cassettes into the plant expression vector pART27 was not achieved. Several attempts were made, and adjustments to the protocol were made at every step; amount of DNA digested, ligation ratio, transformation ratio, reagents. It was hypothesised that the ligated vector may be too large for propagation in the *E. coli* strain to hand in the lab, and therefore a commercial strain for larger vectors was tried, however this did not yield any colonies either. Following months of unsuccessful attempts, this arm of the project was dropped. In the future, I would suggest, trying a different plant expression vector. Should this not be successful, the vector that these constructs were designed from could be requested from Mendes et al., 2020, and the fragments of interest inserted via the Gateway BP/LR reactions.

Chapter 5

**Analysis of transcriptional changes
during development of the *C. richardii*
sporophyte by RNA sequencing**

5.1 Introduction

5.1.1 Genomic and transcriptomic resources in *Ceratopteris richardii*

Ceratopteris richardii has emerged as an important model for investigating the evolution of plant development due to its unique alternation of generations and phylogenetic position within the fern lineage (Conway & Di Stilio, 2020; Kinoshita & Wolf, 2022). However, the development of genomic resources for ferns has lagged far behind that of model angiosperms such as *Arabidopsis thaliana* and *Oryza sativa*. One major obstacle to fern genomics is their large genome size, which can be several times larger than that of most flowering plants (Barker & Wolf, 2010; Henry et al., 2014). The haploid genome of *C. richardii* is an estimated 9.6 Gb - more than twice the size of the human genome - and is characterised by high repeat content and large introns (Marchant et al., 2022).

The first draft genome of *C. richardii*, published in 2019, was highly fragmented and incomplete, limiting its utility as a reference for transcriptomic studies (Marchant et al., 2019). However, this was significantly improved in 2022 with the release of the v2.1 genome assembly (Marchant et al., 2022). This updated version combined long-read sequencing data with transcript evidence from RNA sequencing (RNAseq) and Isoform sequencing (Iso-seq) to produce a more complete and contiguous assembly. The scaffold N50 - a key metric indicating assembly contiguity - reached 182 Mb, with 91% of the 36,857 identified protein-coding genes anchored to the 39 chromosomes of *C. richardii*. This represents a substantial improvement over the fragmented first draft (Marchant et al., 2022; Marchant et al., 2019).

Despite this progress, the genome remains largely untested as a reference for independent transcriptomic studies. The accompanying transcriptome published by Marchant et al. (2022) sampled a wide range of tissues and developmental stages, but focused primarily on the presence or absence of genes, without presenting

quantitative gene expression data. Additionally, no other published studies have explicitly mapped RNAseq data to the v2.1 genome, leaving its utility for large-scale transcriptomic analyses unexplored. A recent study by Woudenberg et al., [2024](#) investigated auxin responses in *C. richardii* using RNAseq, but the authors did not specify which reference genome was used, making it unclear whether they utilized the v2.1 assembly.

Prior to the availability of the genome, all transcriptomic studies in *C. richardii* relied on *de novo* assembly approaches. These early studies primarily used differential gene expression (DEG) analysis to explore specific processes, such as gravity responses in germinating spores (Bushart et al., [2013](#)), sex determination in gametophytes (Atallah et al., [2018](#)), and root development in the young sporophyte (Aragón-Raygoza et al., [2022](#)). While these studies provided insights into important biological processes, they were limited by the lack of a complete genome reference and did not explore co-expression relationships or regulatory networks. Geng et al. ([2021](#)) analysed a broader range of tissues, including two gametophyte (males and hermaphrodites) and five sporophyte tissues (shoot apices, fertile fronds, roots, bulbils, and calli). However, this study focused primarily on evolutionary orthogroup analysis rather than investigating gene expression patterns across tissues. Similarly, studies on callus differentiation (Xiao & Li, [2024](#)) and stomatal development (Plackett et al., [2021](#)) relied on DEG approaches without examining broader transcriptomic relationships. As such, a comprehensive, quantitative transcriptomic analysis across developmental stages in *C. richardii* has not yet been conducted. The lack of such studies, particularly those leveraging the improved v2.1 genome, represents a key gap in the literature.

5.1.2 Read mapping and quantification approaches: HISAT2 vs Salmon

RNAseq data analysis relies on accurate methods to map sequencing reads and quantify transcript abundance (Conesa et al., 2016). The choice of mapping strategy is particularly important when working with species such as *C. richardii*, where genome annotations remain incomplete. Where a reference is available, reads can be mapped to either the genome or the transcriptome, with tools such as HISAT2 (Kim et al., 2019), or Salmon (Patro et al., 2017), respectively. The choice of tool depends on the availability and completeness of the reference, the study objectives, and computational considerations (Conesa et al., 2016).

HISAT2 (Hierarchical Indexing for Spliced Alignment of Transcripts) is a genome-based spliced aligner that maps RNAseq reads to a reference genome (Kim et al., 2019). A key feature of HISAT2 is its ability to account for splice junctions, which allows reads that span exon-intron boundaries to be correctly mapped. This is crucial for capturing alternative splicing events and detecting novel transcripts that may not be included in the current annotation. HISAT2 can map reads across both known and previously unknown splice sites, making it especially useful for species with incomplete genome annotations. However, HISAT2 is a computationally intensive method and relies on a high-quality reference genome to achieve optimal results.

Meanwhile, Salmon is a quasi-mapping tool that quantifies transcript abundance by mapping RNAseq reads to a precomputed transcriptome index (Patro et al., 2017). Unlike traditional aligners, Salmon uses k-mer matching to quickly determine the most likely origin of a read within the transcriptome, bypassing the need for full-length alignment. A key innovation of Salmon is its decoy-aware mapping feature. Non-transcribed genomic regions are included as decoy sequences in the index to absorb spurious alignments, thereby reducing false mappings. This method increases accuracy by ensuring reads belonging to repetitive or non-transcribed regions of the

genome are not incorrectly assigned to sequence-similar transcribed regions. Salmon is favoured for its fast and computationally efficient mapping, and its ability to reduce spurious mappings. However, it is heavily reliant on the quality of the transcriptome annotation - if the transcriptome index is incomplete or inaccurate, many reads will remain unmapped. Furthermore, Salmon cannot map reads outside of the annotated transcriptome, meaning that it is less capable of detecting novel transcripts and alternative splicing events.

The choice between genome-guided methods like HISAT2 (Kim et al., 2019) and transcriptome-based tools like Salmon (Patro et al., 2017) often depends on the quality of the available reference data. Genome-guided methods are generally better-suited to detecting novel transcripts and capturing expression from regions that may not be represented in a transcriptome index (Conesa et al., 2016). In contrast, transcriptome-based tools offer faster and more computationally-efficient quantification, particularly when working with well-annotated transcriptomes. In species with complex or large genomes - such as *C. richardii*, both approaches present unique challenges, and performance can vary depending on the completeness and accuracy of the references used.

5.1.3 From mapping to co-expression analysis

The successful quantification of RNAseq data, whether achieved through genome-guided or transcriptome-based approaches, results in a vast, high resolution dataset of expression values across thousands of genes and multiple tissues or experimental conditions (Finotello & Di Camillo, 2015; Oszolak & Milos, 2011; Wang, Gerstein, & Snyder, 2009). Traditionally, the primary focus of RNAseq analysis has been to identify differentially expressed genes (DEGs) between predefined groups or treatments (Rosati et al., 2024). DEGs are defined as genes that have significantly different expression across two or more conditions (Oshlack et al., 2010; Pertea et al., 2016). Whilst DEG analysis has been instrumental in identifying individual genes with

significant changes in expression (e.g. Tello-Ruiz et al., 2016), it provides a limited view of broader transcriptional programs and does not account for the relationships between genes.

In contrast, weighted gene co-expression network analysis (WGCNA) offers a more holistic approach by identifying clusters, or “modules”, of genes with highly correlated expression patterns across samples (Zhang & Horvath, 2005). These modules can be associated with specific biological processes or traits, providing insights into underlying regulatory networks that are often missed by DEG analyses. WGCNA is therefore particularly valuable in complex, multi-tissue studies.

5.1.4 Key concepts of WGCNA

WGCNA is built on the assumption that biological networks, including gene co-expression networks, exhibit a scale-free topology (Zhang & Horvath, 2005). In a scale-free network, most nodes (genes) have few connections, while a small number of nodes (hub genes) have many connections (Figure 5.1). This pattern reflects real biological systems, where a few genes often play central regulatory roles, while the majority of genes have more limited co-expression relationships (Barabási & Bonabeau, 2003). Constructing a co-expression network that mirrors this scale-free topology ensures that the resulting modules represent biologically meaningful relationships (Zhang & Horvath, 2005).

The first step in WGCNA is to calculate pairwise correlations between all genes across all samples (Zhang & Horvath, 2005). These correlation values form the basis of the co-expression network, where genes that have similar expression patterns across samples will have higher correlation values. However these raw correlations need to be transformed to reflect a scale-free network structure (Figure 5.1). In WGCNA, this is achieved by soft-thresholding, where each pairwise correlation is raised to a chosen power to emphasise strong correlations and reduce the influence of weaker correlations. This transformation shapes the network to better reflect a

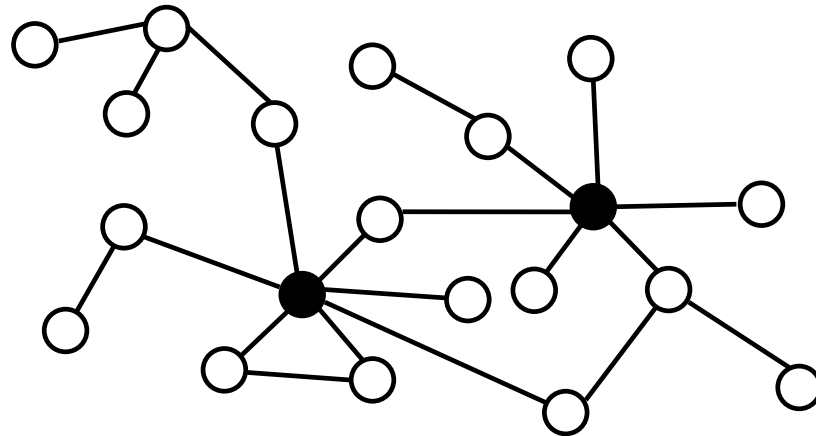


Figure 5.1: Scale-free network structure. In a scale-free network, the majority of nodes have few connections, while a few have many connections (black circles). In a biological network, highly connected nodes are termed hub genes.

biological system. The appropriate soft-threshold power is selected by testing a range of values and identifying the smallest power at which the network approximates a scale-free topology (Langfelder & Horvath, [2008](#)).

Following transformation of pairwise correlations, a topological overlap matrix (TOM) is calculated (Zhang & Horvath, [2005](#)). The TOM quantifies the similarity between two genes by considering not only their direct correlation but also their shared connections with other genes. This measure provides a more robust indicator of co-expression than correlation alone, capturing both direct and indirect interactions. The TOM reduces the influence of random noise and emphasises potentially functionally related gene clusters. A key choice at this stage is whether to use a signed or an unsigned network (Langfelder & Horvath, [2008](#); Zhang & Horvath, [2005](#)). In an unsigned network, both positive and negative correlations are treated equally, meaning genes with inverse expression patterns are considered connected. In contrast, a signed network preserves the direction of the correlation, meaning only positively correlated genes will be placed in the same module.

Once the TOM is calculated, genes are grouped into modules using hierarchical

clustering and a dynamic tree-cutting algorithm (Langfelder & Horvath, 2008; Zhang & Horvath, 2005). Each module represents a cluster of genes with highly correlated expression profiles, which may reflect shared biological functions or regulatory mechanisms. The overall expression profile of each module is summarised by its module eigengene (ME), which is the first principal component of the module's gene expression matrix. The ME acts as a representative value for the module and can be used to explore how the module's expression correlates with experimental traits or conditions. Within each module, hub genes are those with the highest connectivity to other genes in the module. These genes are identified by module membership (kME), which measures how strongly a gene's expression pattern correlates with the ME. Hub genes are thought to play key regulatory roles in the biological processes associated with the module, making them valuable targets for further investigation.

In summary, WGCNA provides a powerful framework for identifying biologically meaningful patterns in large transcriptomic datasets by grouping genes into co-expression modules based on shared expression profiles. This approach moves beyond traditional DEG analysis by offering insights into gene relationships and regulatory networks, making it particularly valuable for exploring species with relatively unexplored genomes, such as *C. richardii*.

5.1.5 Aims and Objectives

The recent publication of a high-quality genome and transcriptome for *C. richardii* represents a significant advancement for fern research. However, these resources have yet to be extensively validated through independent transcriptomic studies. To address this, this chapter begins by evaluating two widely used read-mapping strategies: HISAT2, a genome-guided method, and Salmon, a transcriptome-based method.

Following this, WGCNA is applied to explore gene co-expression patterns across various tissues and developmental stages of *C. richardii*, with particular focus on

reproductive development. The use of WGCNA provides an opportunity to explore previously uncharacterised gene networks, including modules that may contain candidate regulators involved in key developmental processes.

The primary aims of this chapter are to:

1. Evaluate and compare genome-guided and transcriptome-based read-mapping methods to determine their effectiveness in *C. richardii*.
2. Construct a weighted gene co-expression network (WGCNA) to identify modules of co-expressed genes during sporophyte development.
3. Identify modules specifically correlated with reproductive tissues, with a view to identify key hub genes during reproductive development.

5.2 Results

5.2.1 RNA preparation and sequencing

RNA sequencing of the developing *C. richardii* sporophyte was conducted to investigate transcriptional changes associated with reproduction. A total of 12 distinct tissues representing three developmental stages of the *C. richardii* sporophyte - 25 days (25d), 60 days (60d), and 130 days (130d) - were sequenced (Figure 5.2). These stages represent the young sporophyte (25d, Figure 5.2A), the vegetative sporophyte (60d, Figure 5.2C), and the reproductive sporophyte (130d, Figure 5.2H), enabling comparisons between the young and mature vegetative stages and the reproductive stage.

The shoot apex and frond primordia were sampled across all three stages, representing the developmentally youngest tissues at each age (Figure 5.2B, D, I). In addition, tissues representing frond development—including the fiddlehead (Figure 5.2E, J), expanding frond (Figure 5.2F, K) and mature frond (Figure 5.2G, L)—were tested in the 60d (Figure 5.2E-G) and 130d (Figure 5.2J-L) sporophytes. RNA was extracted from tissue samples by Dr Andrew Plackett, and three biological replicates were prepared for each tissue type. mRNA library preparation and sequencing were conducted by Novogene using the Illumina PE150 platform. FASTQ files were provided following sequencing for downstream processing and analysis. As part of this study, sequencing reads were mapped to the recently published *C. richardii* genome and transcriptome assemblies to assess mapping efficiency and prepare the data for downstream transcriptomic analysis (Marchant et al., 2022).

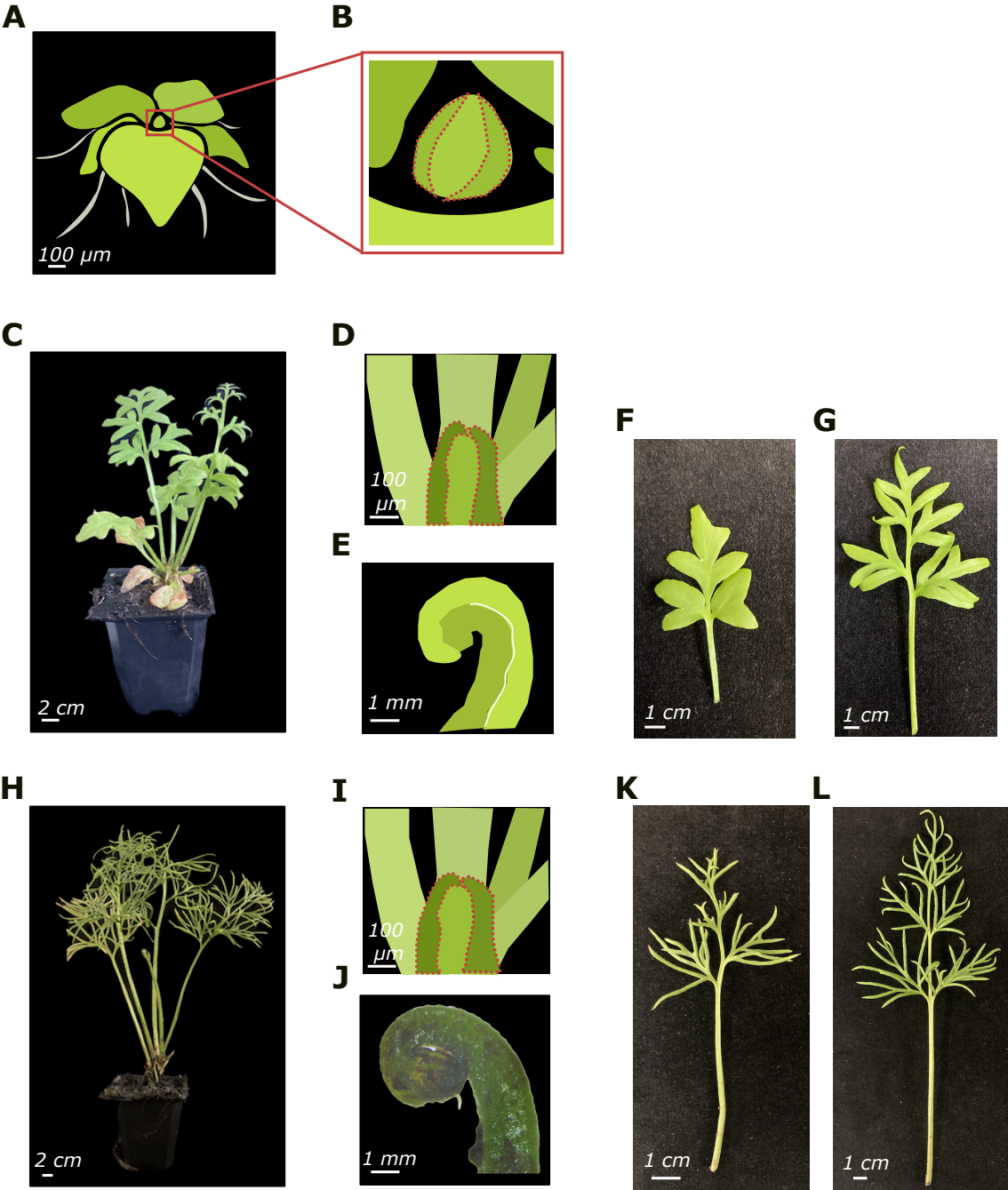


Figure 5.2: Caption overleaf.

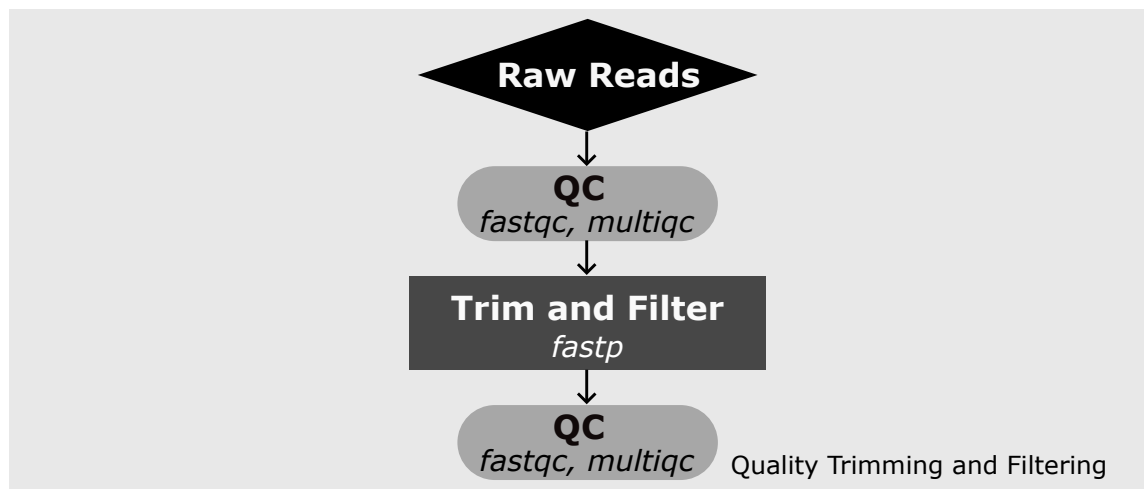
Figure 5.2 Developmental stages of *C. richardii* analysed by RNA sequencing.

Tissues from the *C. richardii* sporophyte at three developmental ages were analysed by RNA sequencing: the young 25 day old (25d) sporophyte (A), the mature 60 day old (60d) vegetative sporophyte (C) and the mature 130 day old (130d) reproductive sporophyte (H). (A-B) Tissues extracted from the young 25d sporophyte. (A) Diagrammatic representation of the 25d sporophyte. (B) Close-up of the 25d shoot apex. Two tissue types were analysed in the 25d sporophyte: the shoot apex and the surrounding developing frond primordia (highlighted by red dotted line) were separated and RNA extracted. (C-G) Tissues extracted from the 60d vegetative sporophyte. (C) A representative 60d vegetative sporophyte. (D) Shoot apex emerging from the base at the centre of fronds. The shoot apex and surrounding frond primordia (highlighted by red dotted line) were analysed separately. (E) Diagrammatic representation of the vegetative fiddlehead which will develop into a frond. (F) Expanding vegetative frond. (G) Mature vegetative frond. (H-L) Tissues extracted from the 130d reproductive sporophyte. (H) A representative 130d reproductive sporophyte. (I) Shoot apex emerging from the base of the reproductive sporophyte. The shoot apex and surrounding frond primordia (highlighted by red dotted line) were separated before extraction. (J) Reproductive fiddlehead which will develop into a reproductive frond. (K) Expanding reproductive frond. (L) Mature reproductive frond.

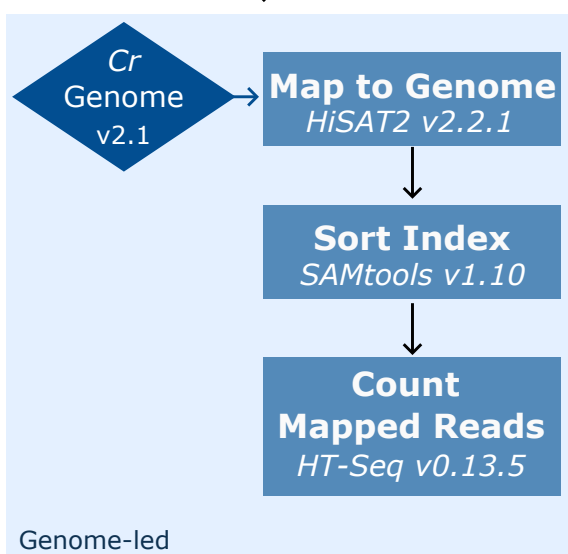
5.2.2 Comparison of genome-led and transcriptome-led mapping of RNAseq reads

Given that the *Ceratopteris richardii* genome (v2.1) and transcriptome (v2.0) (Marchant et al., 2022) were newly published at the time of this study, the accuracy and completeness of the reference genome were uncertain. To account for potential limitations in the genome assembly and annotation, two different read processing methods were compared: an alignment-based approach using HISAT2 and an alignment-free method using Salmon. HISAT2 maps reads directly to the reference genome, whereas Salmon quasi-aligns reads to an index built using the reference transcriptome plus non-transcribed regions from the genome that act as “decoys” (Kim et al., 2019; Patro et al., 2017). The workflow for the two processing methods is shown in Figure 5.3. Raw reads were first quality checked, trimmed and filtered (Figure 5.3A) before either aligning to the genome with HISAT2 (Figure 5.3B) or quasi-aligning to the transcriptome with Salmon (Figure 5.3C).

A



B



C

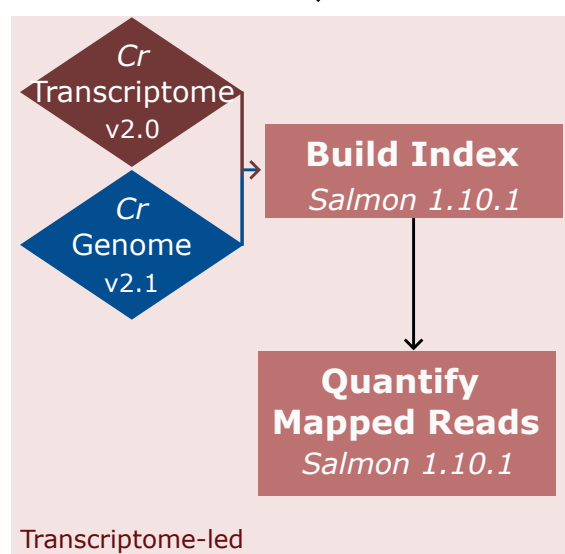


Figure 5.3: Overview of the RNAseq processing pipeline. (A) Raw RNAseq reads underwent quality control (QC) with FastQC and MultiQC to assess sequencing quality. Reads were then trimmed and filtered with fastp to remove adapters and low-quality bases, and subjected to another round of QC to ensure data quality. The trimmed reads were then processed through two parallel analysis approaches: genome-led (B) and transcriptome-led (C) workflows. (B) Trimmed reads were mapped to the *Ceratopteris richardii* genome (v2.1) using HISAT2 v2.2.1. The resulting SAM files were converted to BAM format, sorted, and indexed using SAMtools v1.10. Mapped reads were quantified at the gene level using HTSeq v0.13.5. (C) The *C. richardii* transcriptome (v2.0) was indexed for decoy-aware mapping using Salmon v1.10.1, incorporating the genome (v2.1) as decoy sequences to reduce spurious alignments. Trimmed reads were then quasi-mapped and quantified using Salmon.

Mapping efficiency was compared between HISAT2 and Salmon to evaluate the influence of reference quality on read alignment. Mapping efficiency was found to differ substantially between the two methods (Figure 5.4). HISAT2 successfully mapped approximately 90 % of reads in all samples to the *C. richardii* genome assembly (Figure 5.4A). Mapping was uniform across all samples, with a slightly reduced mapping rate in all mature frond samples, possibly due to tissue-specific differences in RNA quality or transcript abundance. The remaining 10 % comprised approximately 4 % unmapped reads (read that could not be aligned to any position in the reference genome) and approximately 6 % multi-mapped reads (reads that aligned equally well to multiple positions in the reference).

Meanwhile, approximately 50 % of reads were mapped to the reference transcriptome by Salmon (Figure 5.4B). Approximately 25 % were unmapped and a further roughly 20 % were classified as decoys. Decoys refer to non-transcribed regions of the *C. richardii* genome that were included in the Salmon index to capture non-specific or spurious alignments, typically representing unannotated or repetitive sequences that are not part of the validated transcriptome (Patro et al., 2017). The reads classified as decoys or unmapped by Salmon were further investigated through *de novo* assembly followed by mapping to the genome with HISAT2 (Appendix C). The vast majority of reads were found to represent genes that are annotated in the v2.1 genome but not included in the v2.0 reference transcriptome, likely due to limited experimental validation. This discrepancy may be explained by the tissue sampling strategy used in Marchant et al., 2022, which included RNAseq from only ten tissue types. It is possible that key tissues or developmental windows were missed, resulting in an incomplete representation of the fern's transcriptome in the current reference database. This suggests that the lower mapping rate achieved by Salmon was not due to poor quality sequencing, but rather to an incomplete reference transcriptome.

Given the more comprehensive mapping achieved by HISAT2, the genome-mapped reads were selected for subsequent expression analysis. Gene-level counts were

aggregated with HT-seq (v0.13.5), resulting in expression quantification of 36,857 annotated genes.

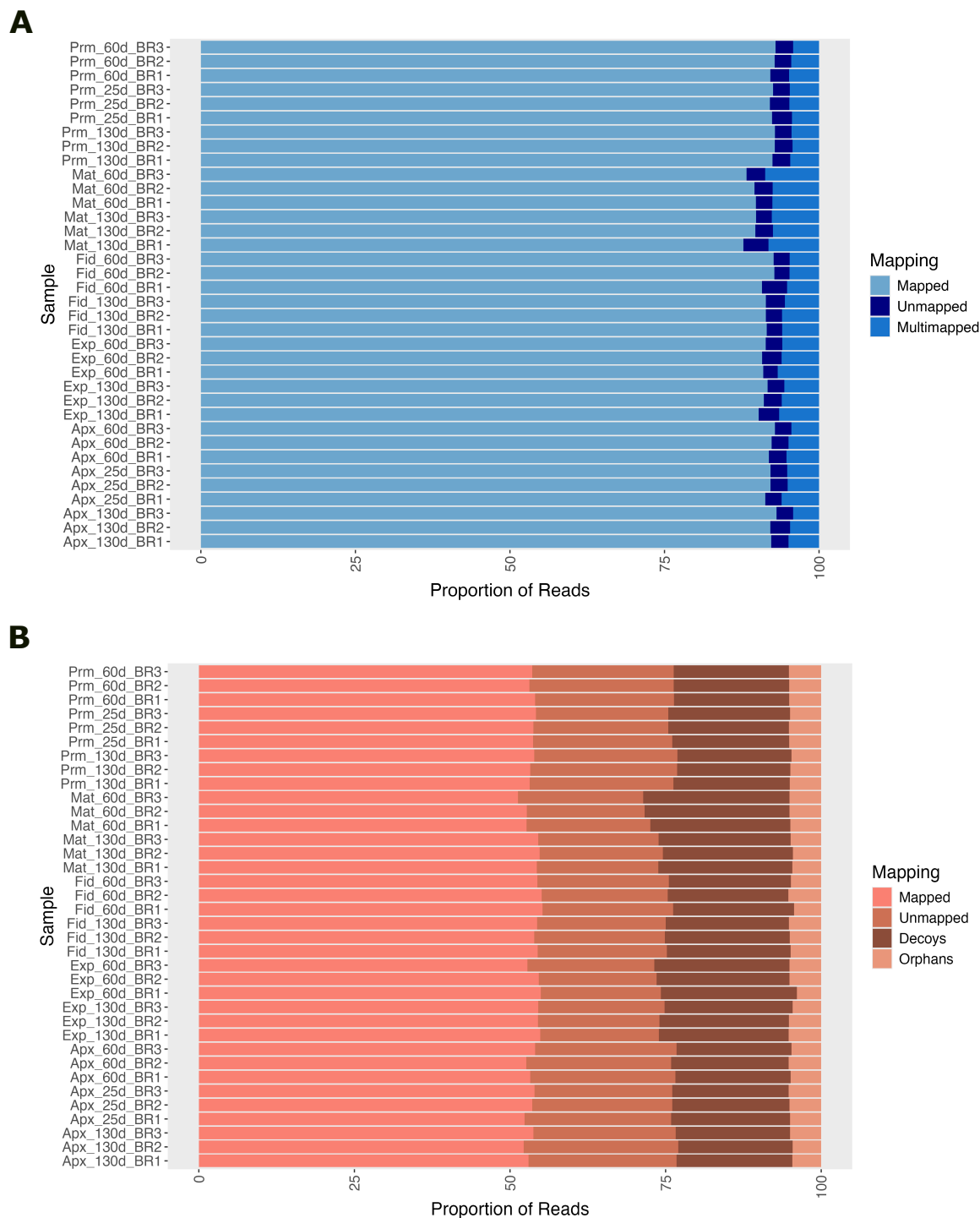


Figure 5.4: Mapping proportions resulting from HISAT2 and Salmon alignment methods. (A) HISAT2. Proportion of RNAseq reads that confidently aligned to the genome (mapped), did not align to the genome (unmapped) or were aligned to more than one place in the genome (multimapped). (B) Salmon. Proportion of reads that confidently aligned to the transcriptome (mapped), did not align to the transcriptome or the genome (unmapped), aligned more confidently to the genome than the transcriptome (decoys) or only one read of a pair could be successfully aligned (orphans).

5.2.3 Data Normalisation

Raw read counts cannot be directly compared across samples due to variations in transcript length, sequencing depth, and potential biases introduced during the sequencing process (Conesa et al., 2016). Therefore, within-sample normalisation was performed using fragments per kilobase of transcript per million (FPKM) values to standardize expression levels. Genes were then filtered to retain only genes with $FPKM \geq 0.5$ in at least three samples. 23,099 genes with robust expression across multiple samples were retained. Subsequently, the data were log-transformed using the equation $\log_2(FPKM + 1)$ to stabilize variance and reduce the influence of extreme values. The log-transformed normalized counts had a uniform distribution of gene expression across samples, with slight variations in the upper range of developmentally older tissues, potentially reflecting biological or technical variability (Figure 5.5).

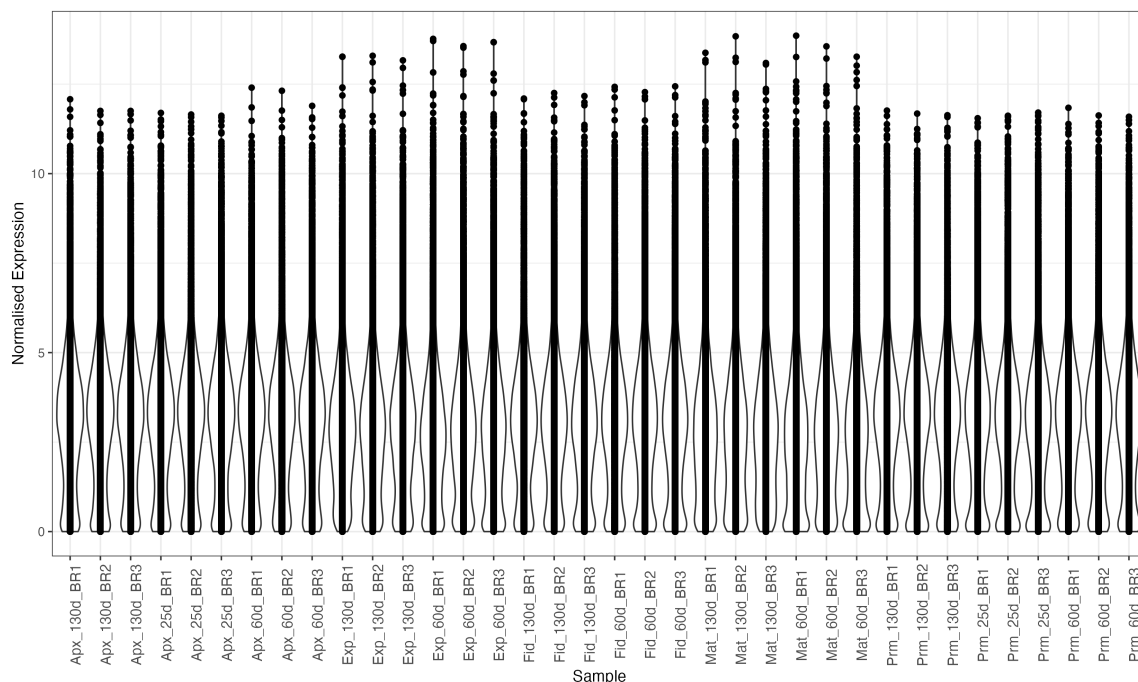


Figure 5.5: Violin plot showing the distribution of FPKM-normalised expression values across all samples. Each violin represents the distribution of normalized expression values for a single sample, with the width indicating the density of values at a given expression level.

5.2.4 Sample clustering and dimensionality reduction

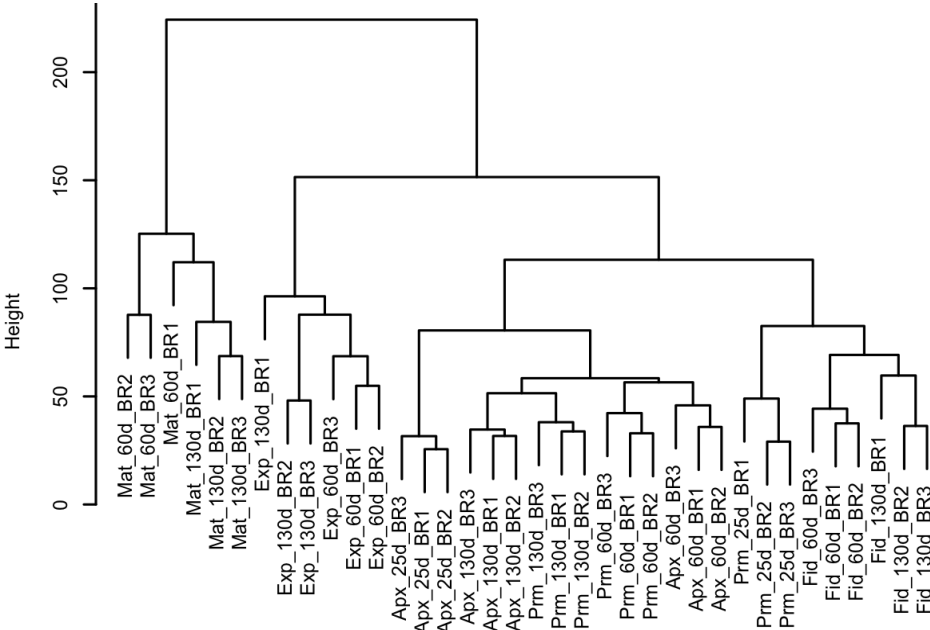
To assess the relationships between samples and ensure that biological replicates clustered as expected, hierarchical clustering (Figure 5.6) and Principal Component Analysis (PCA) (Figure 5.7) were performed. Hierarchical clustering revealed clear groupings of samples based on their overall expression profiles, with biological replicates clustering together (Figure 5.6A). The sample distance matrix provided a quantitative overview of pairwise Euclidean distances between samples (Figure 5.6B). Consistent with the clustering analysis, samples from the same tissue type and developmental stage exhibited smaller distances.

The clustering patterns identified by hierarchical analysis were reflected in the PCA results (Figure 5.7). PC1 and PC2 together accounted for 65.1% of the total variance,

with PC1 explaining the majority (50%). The samples clustered primarily by tissue type (represented by colours) and developmental age (represented by shapes), indicating that transcriptomic profiles are distinct across these variables. The separation along PC1 reflects differences between tissue types, with samples arranged from the developmentally youngest tissue (Shoot Apex) on the left through to the oldest (Mature Frond) on the right, indicating that PC1 captures overall developmental progression.

Along PC2, further separation was observed between specific tissue types by plant age, highlighting finer distinctions between samples beyond overall developmental progression (Figure 5.7). Interestingly, shoot apex and frond primordia samples clustered closely together regardless of plant age, suggesting minimal transcriptomic differences between these tissues across developmental stages. However, an exception was observed for the 25d frond primordia samples, which clustered with the fiddlehead samples from the 60d and 130d sporophytes. This clustering pattern suggests that the transcriptomic profile of the 25d frond primordia may be more similar to the fiddleheads than to the frond primordia at later developmental stages.

A



B

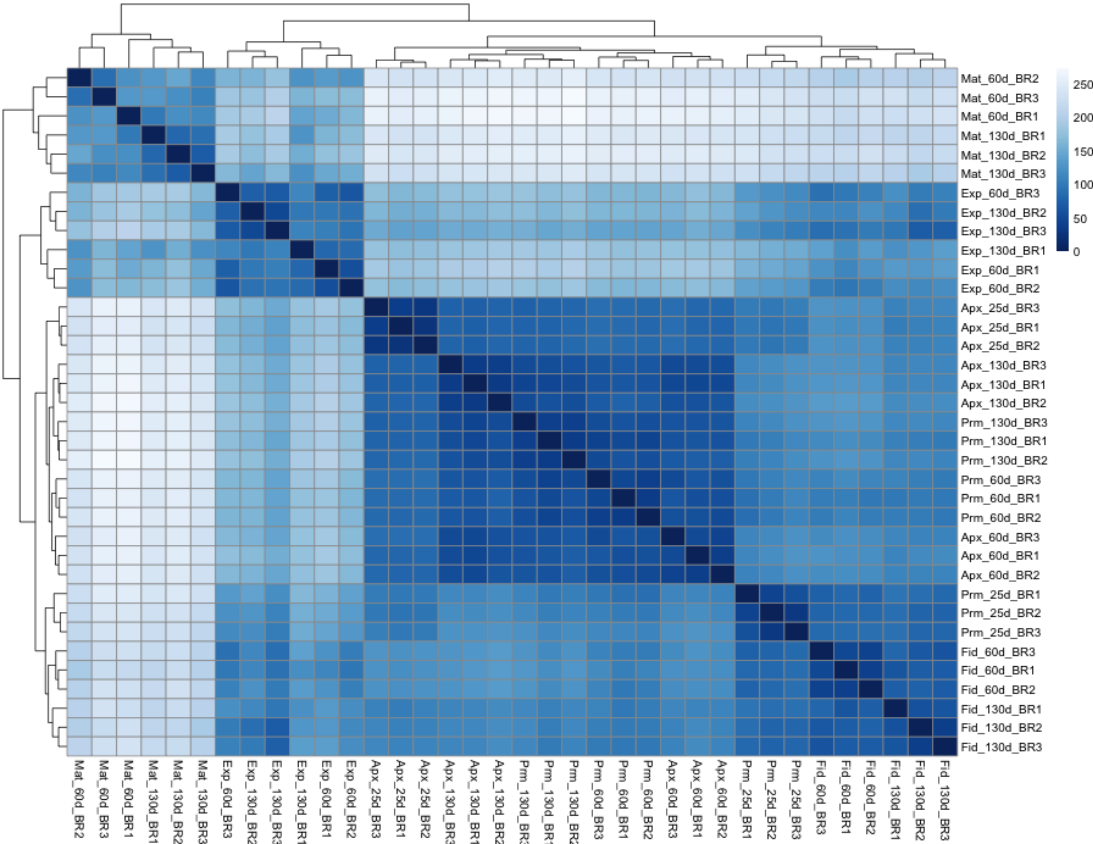


Figure 5.6: Caption overleaf.

Figure 5.6 Hierarchical clustering dendrogram (A) and sample distance matrix (B) of RNA sequencing data from *Ceratopteris richardii* sporophyte samples. (A) Hierarchical clustering dendrogram generated using average linkage clustering (hclust) based on Euclidean sample distances calculated from log-transformed FPKM values. Samples cluster primarily according to a combination of tissue type and developmental stage, indicating consistent biological relationships among samples. (B) Heatmap of the sample distance matrix, with colours representing Euclidean distances between samples. Darker shades indicate greater similarity. The clustering patterns observed in both panels reflect the expected grouping of biological replicates and highlight the robustness of the dataset.

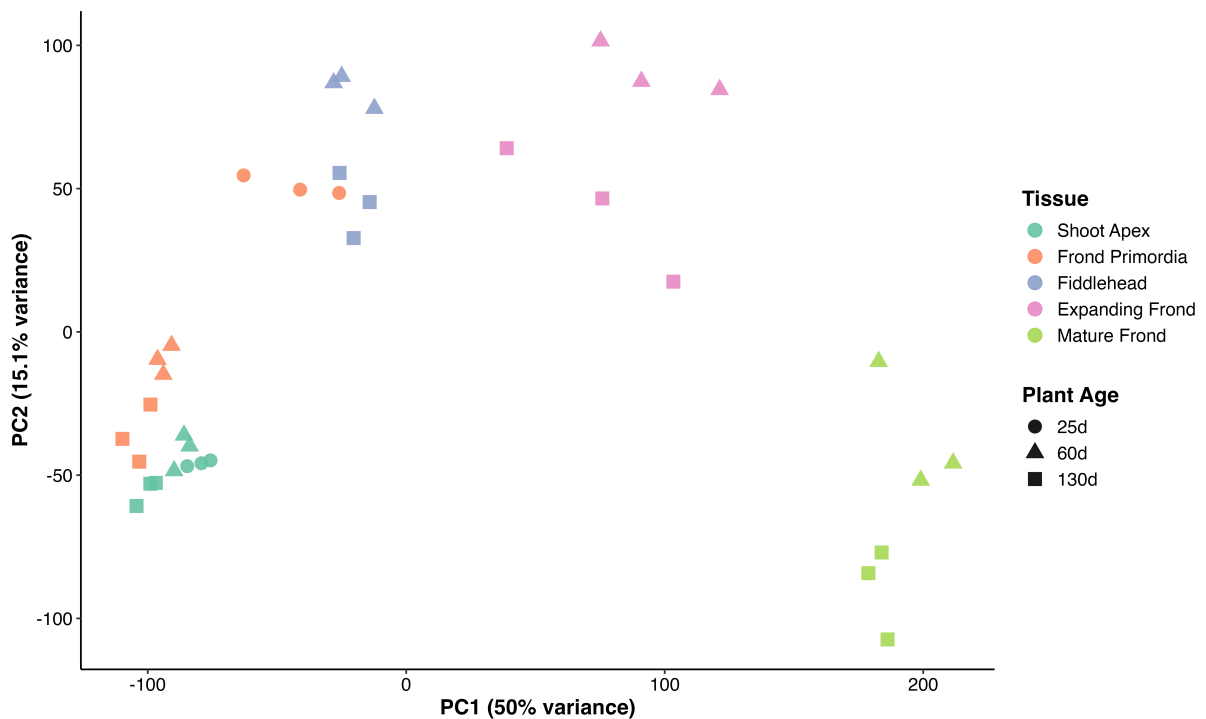


Figure 5.7: Principle Component Analysis (PCA) of normalised gene expression data. PCA was performed to visualise sample variation across tissue types and plant ages of *C. richardii* sporophytes. PC1 and PC2 together account for 65.1% of the total variance. Samples cluster primarily by tissue type (represented by different colours) and plant age (represented by different shapes), indicating that transcriptomic profiles are distinct across these variables. The separation along PC1 highlights differences between tissue types from the developmentally youngest tissue (Shoot Apex, left) through to the oldest (Mature frond, right). while PC2 further separates specific tissue types by plant age.

5.2.5 Weighted Gene Co-expression Network Analysis (WGCNA)

To identify groups of co-expressed genes across the sampled tissues and developmental stages of *Ceratopteris richardii*, a Weighted Gene Correlation Network Analysis (WGCNA) was conducted using the normalized and filtered FPKM expression values (Langfelder & Horvath, 2008; Zhang & Horvath, 2005). The analysis involved key steps including the determination of a soft-thresholding power to achieve a scale-free network, construction of a topological overlap matrix (TOM) to assess gene connectivity, dynamic tree cutting to detect initial modules (clusters of genes with shared expression patterns), and merging of closely related modules to refine the network structure.

5.2.5.1 Soft-threshold selection for scale-free network construction

A critical step in WGCNA is selecting an appropriate soft-threshold power to construct a scale-free network, where a few nodes (genes) have many connections, while most nodes have fewer connections (Langfelder & Horvath, 2008; Zhang & Horvath, 2005). A range of soft-threshold powers was tested to determine the power at which the network topology best approximated a scale-free structure (Figure 5.8). The scale-free topology fit index (signed R^2) (Figure 5.8A) and mean connectivity (Figure 5.8B) were plotted for each tested power. The scale-free topology fit index quantifies how well the network adheres to a scale-free structure, with values closer to 1 indicating that the network exhibits the expected structure. Meanwhile, the mean connectivity refers to the average sum of connection strengths between each gene. Retaining a moderate mean connectivity ensures that the network captures meaningful gene-gene interactions without being overly sparse (which risks missing co-expressed gene clusters) or overly dense (which can introduce noise and obscure relationships).

Initial analysis suggested a soft-threshold power of 5, however, this resulted in a

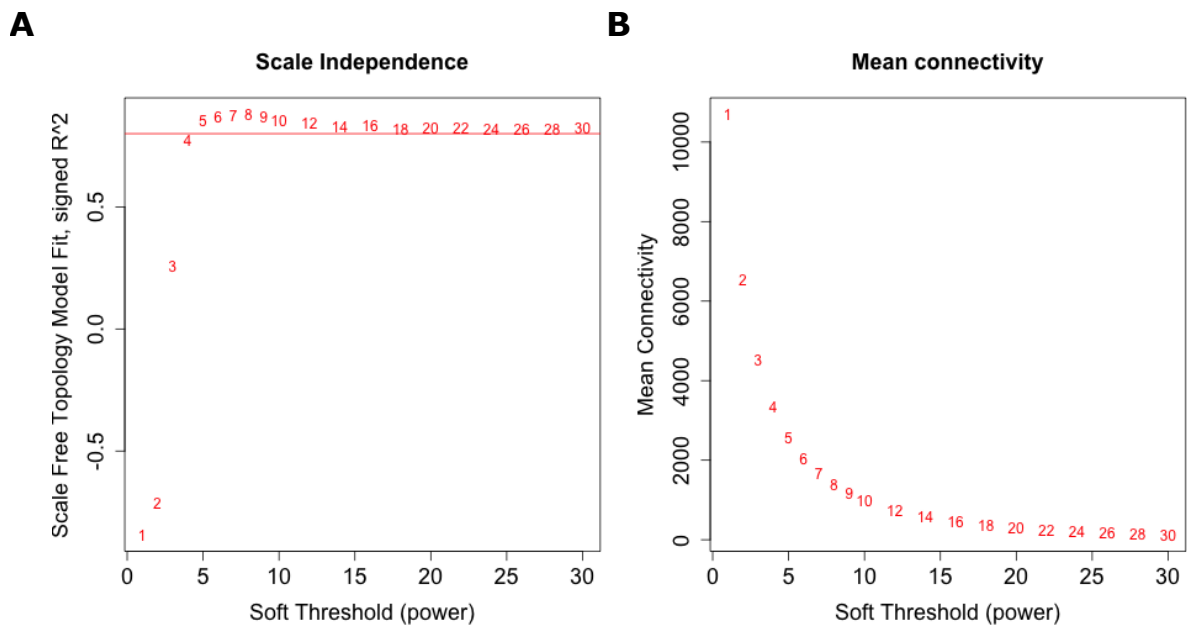


Figure 5.8: Soft-threshold power selection for WGCNA. (A) The scale-free topology fit index plotted against different soft-threshold powers. A threshold of 14 was selected, corresponding to an R^2 of at least 0.8 (red line). (B) The mean connectivity as a function of soft-threshold power, showing a decline in connectivity as the power increases.

shallow slope of -0.455 in the log-log plot of connectivity vs frequency, indicating that the resulting network was not adequately filtering out weak correlations and was deviating from the expected scale-free structure. A higher power of 14 was selected, which produced a slope closer to the theoretical ideal of -1 (-0.920) and retained a robust scale-free topology fit ($R^2 = 0.831$) while maintaining reasonable mean connectivity (574). An unsigned network approach was chosen to capture both positive and negative gene correlations to account for the possibility that genes may be co-expressed in opposing directions across different tissue types.

5.2.5.2 Gene co-expression network construction

Using the selected soft-threshold power of 14, a topological overlap matrix (TOM) was constructed to quantify the similarity between genes based on their shared network connectivity (Langfelder & Horvath, 2008). The TOM-based dissimilarity ($1 -$

TOM) was used to cluster genes hierarchically, producing a dendrogram in which genes with similar expression patterns were grouped together (Figure 5.9A). The resulting dendrogram provides a visual representation of the gene clusters, with leaves representing individual genes and branches representing groups of genes with highly similar expression patterns.

Dynamic tree cutting was employed to identify modules, which are groups of genes with highly correlated expression patterns across tissue types, from the gene dendrogram (Langfelder & Horvath, 2008; Langfelder et al., 2008). These modules represent distinct co-expression clusters that may reflect shared biological functions or regulatory pathways. A total of 46 modules were initially detected using the dynamic tree cut method with a minimum cluster size of 30 genes. Each module was assigned a unique colour to aid visualization (Figure 5.9B). The size of the modules varied, with larger modules representing core sets of co-expressed genes that are likely involved in broad, fundamental biological processes. Smaller modules may represent more specialized or context-specific gene sets. The largest module contained 5124 genes and the smallest 41 genes.

To reduce redundancy and improve biological interpretability, closely related modules were merged based on the similarity of their module eigengenes (MEs). MEs represent the overall expression profile of each module across all samples and can be used to summarize module behaviour. Hierarchical clustering of MEs revealed groups of modules with highly similar expression profiles (Figure 5.9D). Modules were merged if their dissimilarity measure ($1 - \text{correlation}$) was below 0.25, meaning that modules with eigengene correlations above 0.75 were combined into a single module, resulting in the final set of 21 merged modules (Figure 5.9C-D, Table 5.1). This threshold has been commonly applied in practice and is recommended in the official WGCNA tutorial (Langfelder & Horvath, 2014). Each module is assigned a unique, arbitrary colour label to aid visualisation and distinguish the groups of co-expressed (or inversely expressed) genes. After merging, module membership varied: the largest

module (red) contained 13,165 genes, while the smallest (darkorange2) contained 41 genes (Table [5.1](#)).

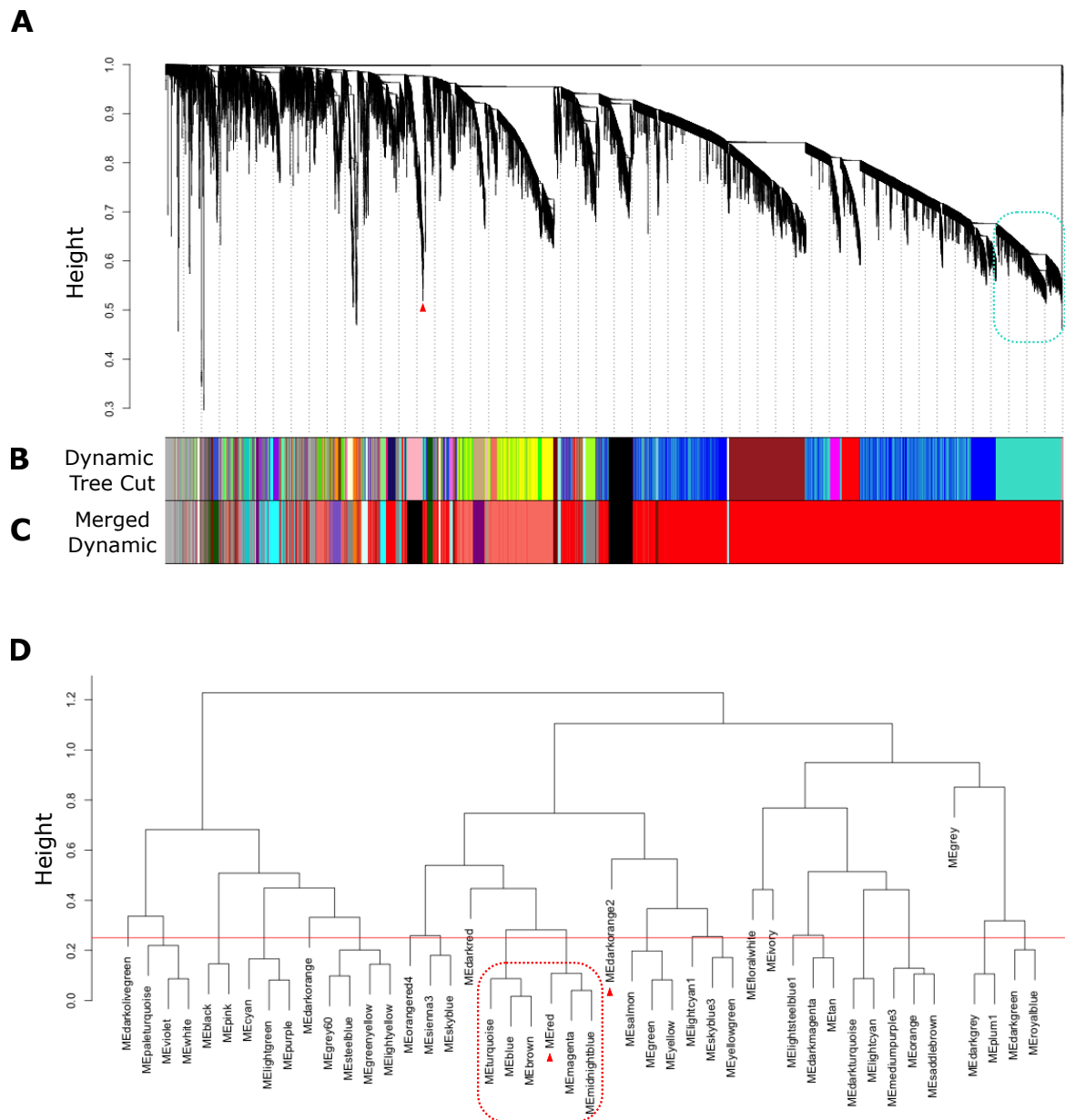


Figure 5.9: Gene module detection and merging. (A) Gene clustering dendrogram based on Topological Overlap Matrix (TOM) dissimilarity. Leaves represent individual genes (example indicated by red arrowhead); branches represent co-expressed gene groups, with the turquoise module highlighted by a turquoise dotted square. (B) Initial modules detected via dynamic tree cut. (C) Final modules after merging. (D) Hierarchical clustering of module eigengenes; leaves represent modules coloured by co-expression patterns. Modules below the dissimilarity threshold (red line at 0.25) were merged. The smallest (MEdarkorange2) and largest (MERed) modules are marked with red arrowheads; modules that were merged with red are enclosed in a red dotted square.

Table 5.1: Module membership and the corresponding number of genes in merged modules.

Module	Number of Genes
black	1129
cyan	897
darkgreen	354
darkgrey	222
darkmagenta	454
darkolivegreen	91
darkorange	126
darkorange2	41
darkred	178
darkturquoise	348
floralwhite	51
grey	724
grey60	872
ivory	53
lightcyan1	213
lightsteelblue1	60
mediumpurple3	308
orangered4	281
paleturquoise	321
red	13165
salmon	3211

5.2.6 WGCNA identified gene modules associated with tissue type and plant age

The relationship between the merged gene modules and the developmental stages of *C. richardii* sporophytes was explored through module-trait correlation analysis. Pearson correlation analysis between module eigengenes (MEs) and binary traits identified significant associations between specific modules and tissue types. A heatmap displays the module-trait relationships, with correlation values and corresponding p-values labelled within the cells (Figure 5.10). Positive correlations are indicated by red shading, while negative correlations are shaded blue.

The strength and direction of correlations provide insights into the biological relevance of each module across the sampled tissues. Several modules showed

associations primarily with plant age, regardless of tissue type (Figure 5.10). For example, modules such as MEdarkred and MEdarkgrey were significantly correlated with the 60d sporophyte, suggesting involvement in developmental processes specific to the vegetative sporophyte. Meanwhile, modules such as MEgrey were more strongly correlated with the younger 25d sporophyte, indicating a role in early sporophyte development.

In contrast, other modules appeared to be associated with specific tissue types, independent of plant age (Figure 5.10). For instance, MEorangered4, MEdarkturquoise, and MEred all displayed strong negative correlations with mature fronds in both the 60d and 130d samples, suggesting that these modules may be associated with gene networks that are less active as fronds mature, and potentially involved in early developmental processes that diminish in later stages. Conversely, MEgrey60 and MEblack were positively correlated with mature fronds, indicating associations with processes that are maintained or activated during later stages of frond maturation, regardless of whether the frond is vegetative or reproductive.

A select few modules exhibited more specific associations with particular tissues at defined developmental stages (Figure 5.10). For example, MElightcyan1 was uniquely correlated with the 25-day shoot apex, implying a highly specialized regulatory role in the youngest shoot apex tissue. Of particular interest were MEivory and MEdarkorange, which were strongly correlated with tissues from the 130-day sporophyte – the developmental stage associated with sporangia formation and meiosis. MEivory was most strongly associated with the fiddlehead and expanding frond tissues, suggesting a potential role in early reproductive tissue development. Meanwhile, MEdarkorange was uniquely correlated with both expanding and mature fronds, implying a role in later-stage processes. These findings suggest that distinct gene co-expression networks are associated with the progression from vegetative growth to reproductive development in *C. richardii* sporophytes, with MEivory and MEdarkorange modules potentially contributing to this developmental transition.

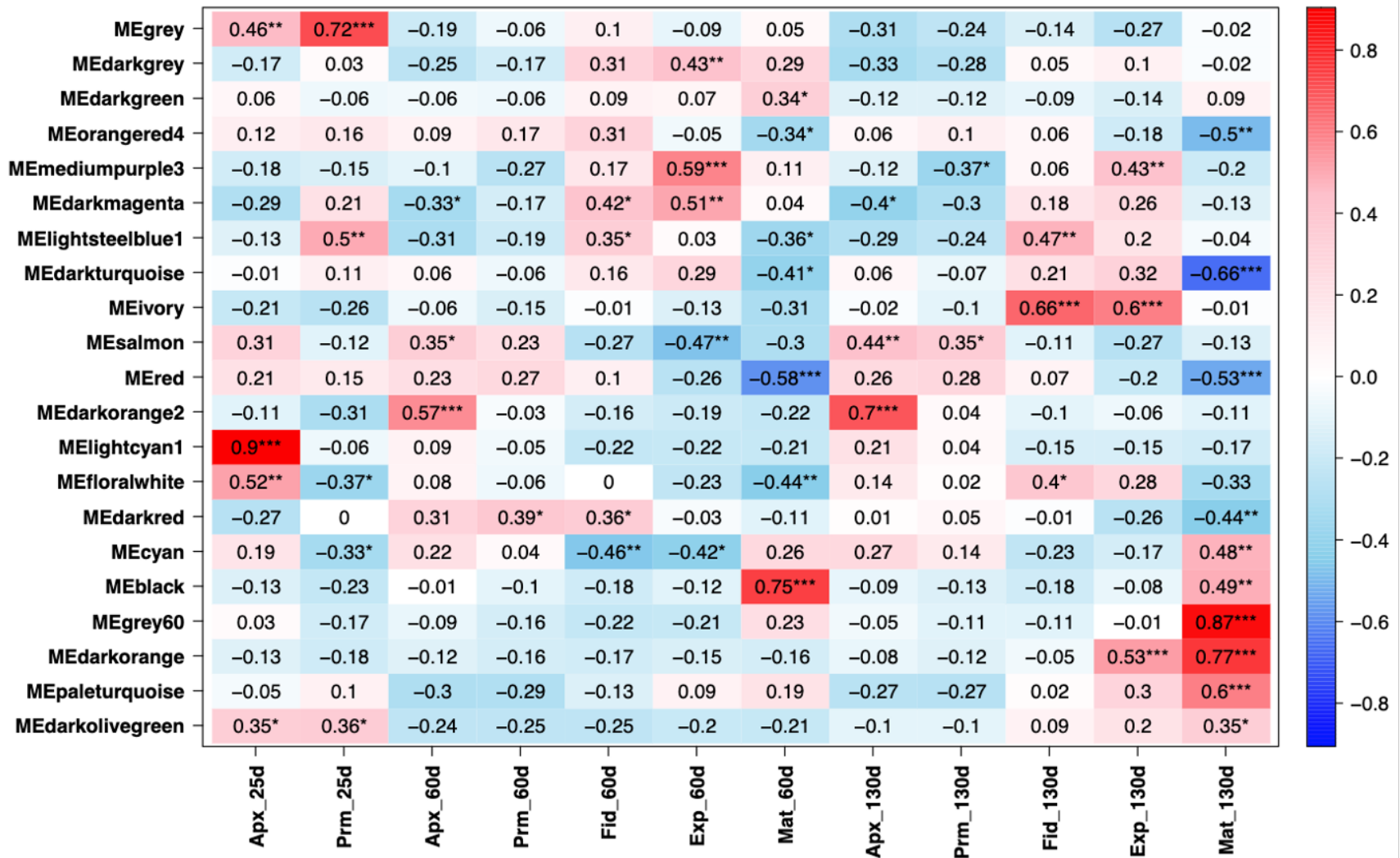


Figure 5.10: Caption overleaf.

Figure 5.10 Gene modules exhibit distinct correlations with developmental stages and plant age in *Ceratopteris richardii*. The heatmap displays Pearson correlations between module eigengenes (rows) and tissue types (columns) across different developmental stages of *C. richardii* sporophytes. Positive correlations are represented by red shading, while negative correlations are indicated by blue shading. The intensity of the color reflects the strength of the correlation. Correlation values are shown within each cell, with asterisks indicating statistical significance ($p < 0.05$ (*), $p < 0.01$ (**), $p < 0.001$ (***)).

5.2.6.1 Validation of module gene expression patterns

To confirm that the expression profiles of individual genes within each module were consistent with their corresponding ME, expression plots were generated for all modules (Figures 5.11 and 5.12). The MEgrey module was excluded from this analysis, as it represents a collection of genes that could not be assigned to a specific module and is therefore uninformative. Figure 5.11 shows the FPKM-normalized expression of genes that exhibited a positive correlation with their ME across the sampled tissues, while Figure 5.12 displays the expression profiles of genes negatively correlated with their ME.

Overall, the expression patterns within modules were consistent with their assigned MEs, confirming that the WGCNA approach effectively grouped co-expressed genes (Figures 5.11 and 5.12). This was particularly evident in smaller modules such as darkorange, darkorange2, ivory, and lightsteelblue1, where the expression profiles of individual genes closely followed the trends observed in the module eigengene. In larger modules, including red, black, and salmon, the expression plots were more densely packed due to the higher number of genes. While this density made it more challenging to visually assess whether every gene precisely followed the ME, the overall patterns still suggested correlation with the ME. Furthermore, in all modules, positively and negatively correlated genes exhibited mirrored expression profiles across the sampled tissues, further supporting the robustness of the network construction and the biological relevance of the identified modules.

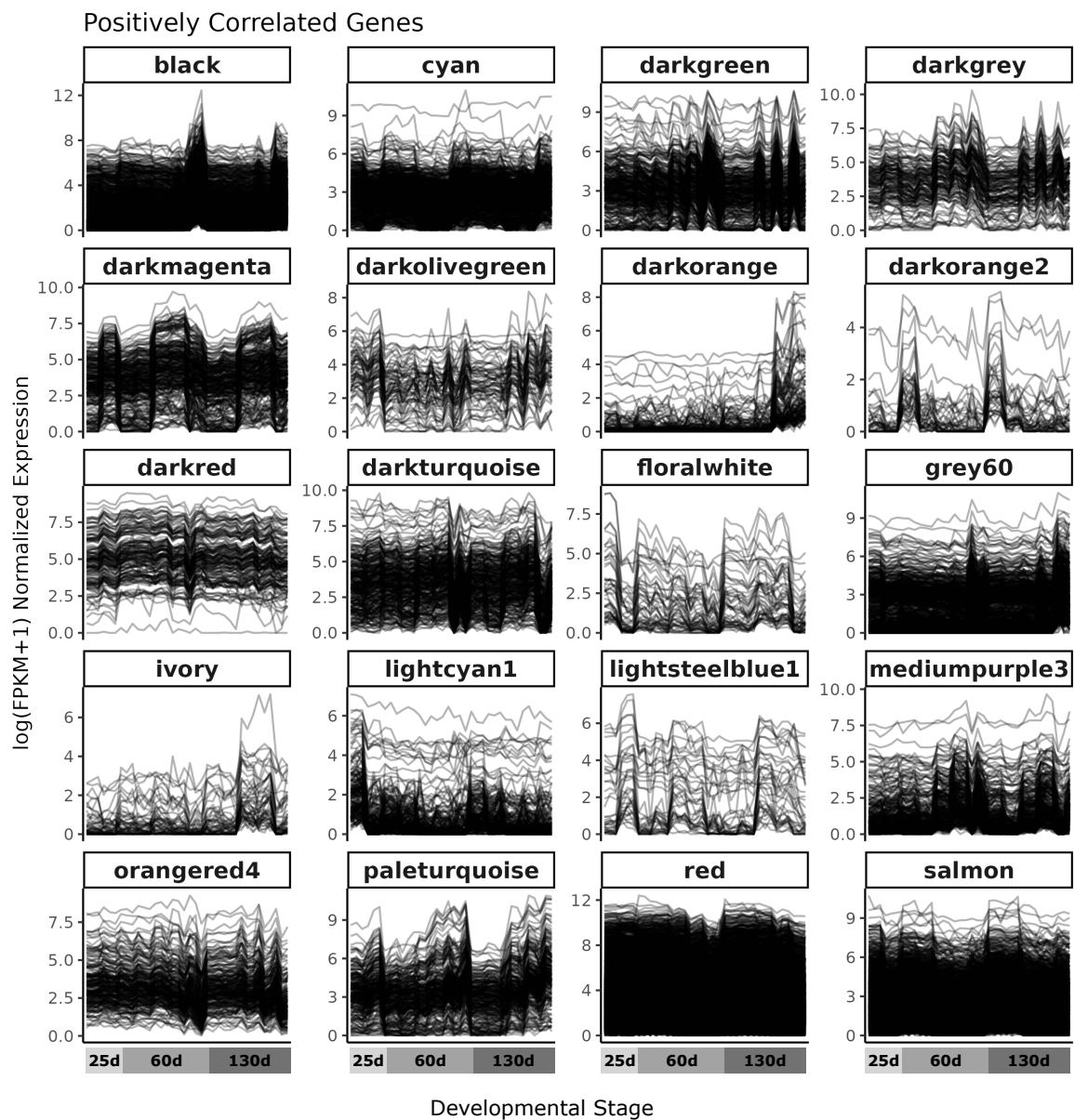


Figure 5.11: Expression profiles of positively correlated genes within each module. The normalized expression profiles ($\log_2(\text{FPKM}+1)$) of genes that positively correlate with their respective module eigengenes are shown for all modules except grey. Each line represents the expression of a single gene across the sampled tissues. Developmental stages are ordered from youngest (25d) to oldest (130d) along the x-axis.

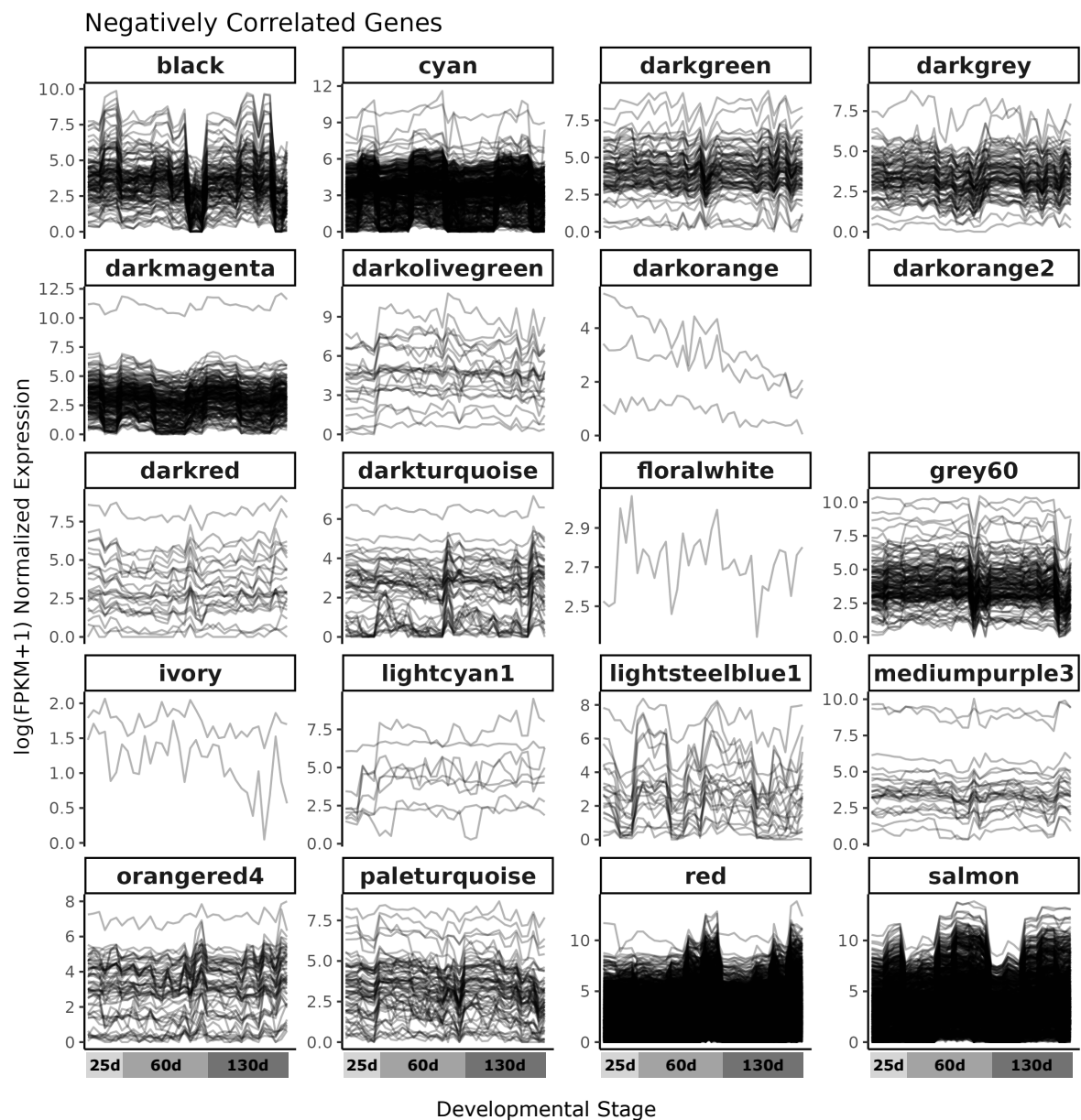


Figure 5.12: Expression profiles of negatively correlated genes within each module. The normalized expression profiles ($\log_2(\text{FPKM}+1)$) of genes that negatively correlate with their respective module eigengenes are shown for all modules except grey. Each line represents the expression of a single gene across the sampled tissues. Developmental stages are ordered from youngest (25d) to oldest (130d) along the x-axis.

5.2.7 Exploring reproductive gene modules in the developing sporophyte

Gene co-expression modules identified through WGCNA often reflect distinct biological functions. Two modules, ivory and darkorange, stood out for their strong association with reproductive tissues during the development of *C. richardii* sporophytes (Figure 5.10). Both modules exhibited strong correlations with tissues from the 130d stage, particularly the fiddlehead, expanding frond and mature frond. Furthermore, these modules were not significantly correlated with other tissues or developmental stages, including the 130d shoot apex and frond primordia and all other vegetative tissues.

The specificity of ivory and darkorange to reproductive tissues suggests that genes within these modules may regulate key processes such as sporangium formation and meiosis. Both modules were selected for further analysis to identify key genes and biological pathways potentially involved in reproductive development. The following sections delve into the ivory and darkorange modules individually, analysing their gene expression patterns, functional enrichment, and co-expression networks to uncover their biological roles.

5.2.7.1 Ivory module: gene-trait correlation analysis

The ivory module comprises 53 genes that exhibit a distinct expression pattern during the fiddlehead and expanding frond stages of the 130-day old sporophyte. To confirm this association was consistent across individual genes, their expression patterns were correlated with tissue type. The resulting gene-trait correlation heatmap (Figure 5.13) highlights that genes in this module are strongly correlated with fiddlehead and expanding fronds from the 130-day-old sporophytes. The majority of the genes (51 out of 53) exhibited a strong positive correlation with these tissues, although weak positive correlations with other tissues were observed in some genes.

Conversely, 2 genes - Ceric.11G019000 and Ceric.11G047500 - exhibited a negative correlation with the 130d fiddlehead and expanding fronds. These genes may play a contrasting regulatory role, potentially acting to suppress specific processes during reproductive frond development.

To further investigate the structure of this module, hub genes were identified by calculating intramodular connectivity (kWithin) and module membership (kME). Hub genes are considered central regulators within the module, as they exhibit strong co-expression relationships with other genes and are likely to play key roles within the biological processes associated with the module. In total, six hub genes were identified in the ivory module, each with a kME value of ≥ 0.9 (Table 5.2; Figure 5.13, closed circles). Notably, three of these hub genes - Ceric.08G001000, Ceric.15G075600, and Ceric.24G049200 - cluster tightly together in the gene-trait correlation heatmap, suggesting that they may act in concert to regulate key reproductive processes during these stages of development. The remaining three hub genes showed a more dispersed pattern in the heatmap but still exhibited a high degree of connectivity with other genes in the module. These differences in clustering patterns could reflect functional specialization within the module, with some hub genes forming a core regulatory network, while others may act as independent regulators or interact with other modules during reproductive development.

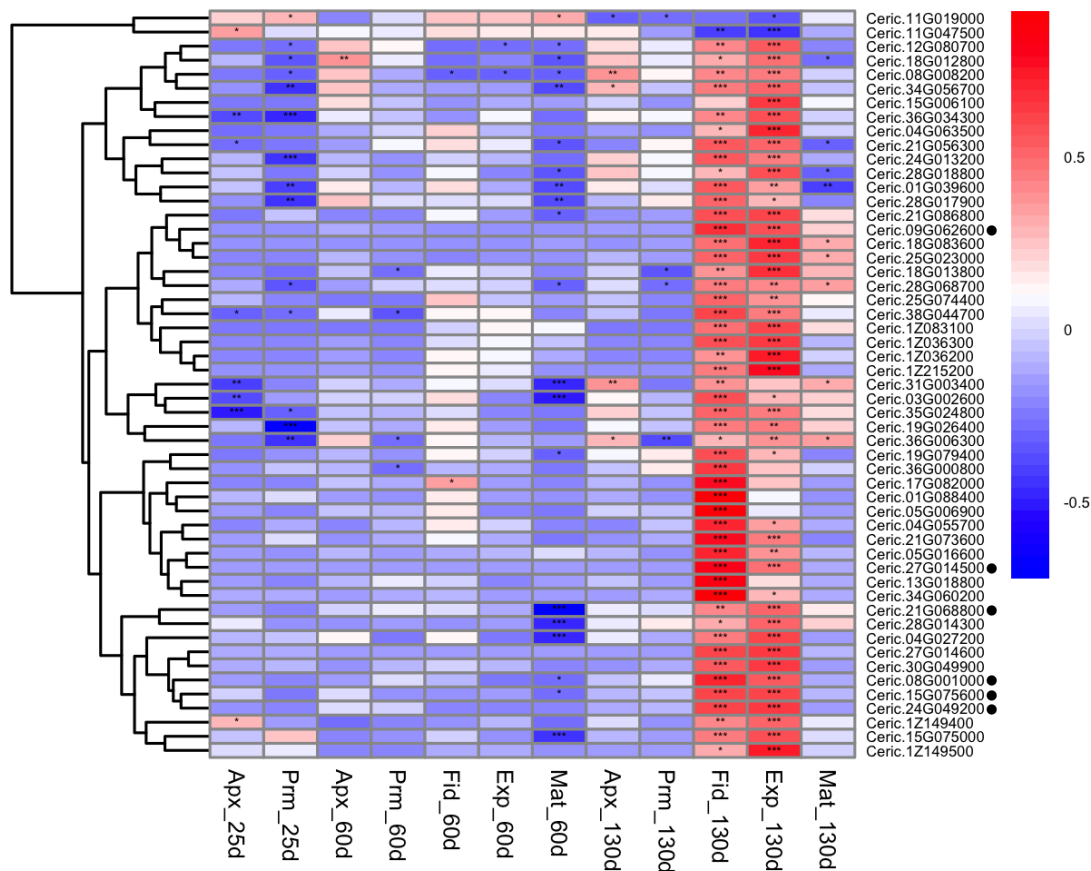


Figure 5.13: Gene-trait correlation heatmap for the ivory module. Correlations for all 53 genes belonging to this module are shown. Genes (rows) are clustered by similar correlation, whilst tissues (columns) are not clustered to enable visualisation of developmental progression. Blue indicates negative correlation, white no correlation and red positive correlation. Hub genes are indicated by filled circle.

Table 5.2: Annotation and putative function of hub genes in the ivory module.

Gene ID	Annotation	Protein Domains	Putative Function
Ceric.08G001000	HOMEODOMAIN-LEUCINE ZIPPER PROTEIN ANTHOCYANINLESS 2-RELATED	Homeobox, START	Transcription factor, anthocyanin regulation
Ceric.27G014500	No hit	–	–
Ceric.09G062600	BES1/BZR1	BES1_N	Transcription factor, brassinosteroid signalling
Ceric.24G049200	No Hit	–	–
Ceric.15G075600	No Hit	–	–
Ceric.21G086800	Flavonoid 3'- monooxygenase / Flavonoid 3'- hydroxylase	P450	Flavonoid biosynthesis

5.2.7.2 Ivory module: Gene Ontology enrichment analysis

To further explore the potential functions of genes within the ivory module, Gene Ontology (GO) enrichment analysis was performed to identify significantly enriched Biological Process (BP) and Molecular Function (MF) terms (Figure 5.14). The GO annotations were obtained from the reference genome assembly, where terms were assigned automatically using InterPro2GO, based on comparisons with most genomes (Marchant et al., 2022). It is important to note that these annotations were not empirically derived and may therefore be incomplete or imprecise. These annotations were used as the basis for an enrichment analysis against the *Arabidopsis* GO database to identify overrepresented terms.

Of the 53 genes in the ivory module, 20 were annotated with GO terms. GO enrichment analysis identified a total of 9 BP (Figure 5.14A) and 8 MF (Figure 5.14B) terms, although only one term remained statistically significant after multiple testing correction. This likely reflects the limited number of annotated genes in this module. The significantly enriched term, enzyme inhibitor activity (MF), suggests a potential regulatory function for some genes in this module, possibly involved in modulating

enzymatic activity during developmental transitions.

Although the remaining GO terms were not statistically significant after correction, they still provide insight into the potential biological roles of genes within the ivory module. In the biological process category (BP), several terms were related to transport processes (e.g., xenobiotic membrane transport, transmembrane transport) and metabolic activities (e.g., lipid metabolic process, lignin catabolic process, metabolic process) (Figure 5.14A). These terms are indicative of the active metabolism occurring in the fiddlehead and expanding fronds, where both cell expansion and tissue differentiation take place.

In the molecular function (MF) category, several notable terms were identified, including NAD⁺ binding, hexosyltransferase activity, DNA-binding transcription factor activity, and structural constituent of cell wall (Figure 5.14B). The presence of transcription factor activity supports the idea that key regulatory genes may be driving developmental changes in these tissues. Additional manual investigation identified 6 transcription factors (TFs) (Table 5.3), all of which were positively correlated with the module eigengene and two of which (Ceric.08G001000 and Ceric.09G062600) were also identified as hub genes (Table 5.2). Terms related to lipid binding and ATPase-coupled transmembrane transporter activity suggest the presence of genes involved in membrane-associated processes, potentially facilitating transport across membranes during rapid tissue growth and differentiation.

Upon manual investigation, a functional dichotomy was observed between the positive and negative genes within the module. Genes with positive correlations to the module eigengene were primarily involved in transcriptional activation and gene expression upregulation (e.g. Ceric.28G018800 was identified as a histone methyltransferase), whilst a notable negatively correlated gene, Ceric.11G019000, was identified as a homolog of histone deacetylase, suggesting coincidental downregulation of transcriptional repression.

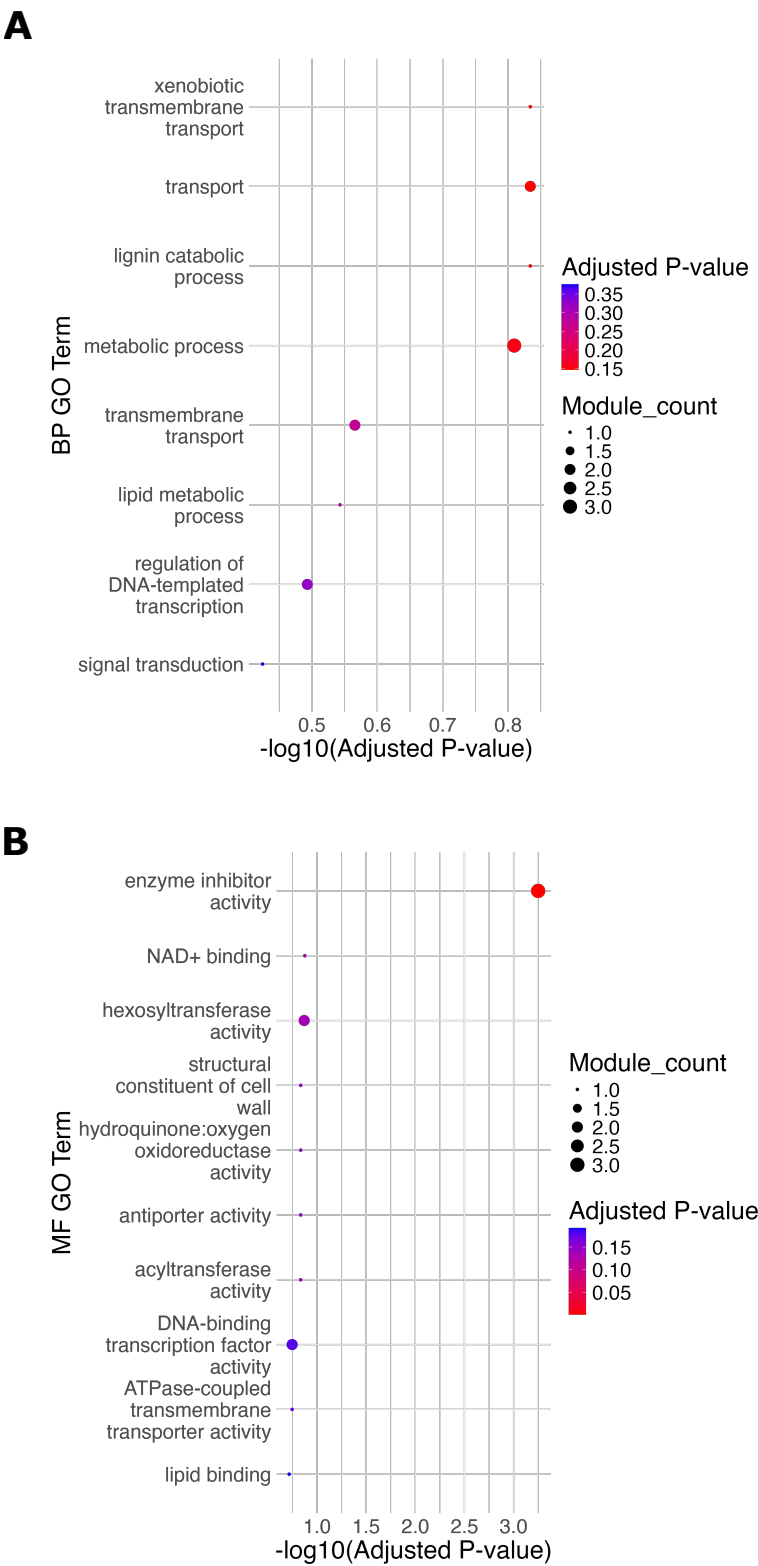


Figure 5.14: Gene Ontology (GO) enrichment analysis of the ivory module. Enriched GO terms were identified in the categories: (A) Biological Process (BP) and (B) Molecular Function (MF). Dot size indicates the number of genes associated with the term (Module_count) and colour indicates the adjusted p-value ranging from blue (less significant) to red (more significant).

Table 5.3: Transcription factors identified within the ivory module.

Gene ID	TF Family	Annotation	Putative Function
Ceric.08G001000	Homeobox	HOMEBOX-LEUCINE ZIPPER PROTEIN ANTHOCYANINLESS 2-RELATED	Regulation of developmental processes, anthocyanin biosynthesis
Ceric.09G062600	BES1/BZR1	BES1/BZR1 plant transcription factor, N-terminal	Brassinosteroid signalling pathway
Ceric.38G044700	Myb	MYB family TF	Regulate growth and development
Ceric.05G016600	HOX/HEX	TF HEX, contains HOX and HALZ domains	Regulates developmental processes and stress responses
Ceric.01G088400	ERF/AP2	Integrase-type DNA-binding superfamily protein	Regulates developmental processes and stress responses
Ceric.28G068700	Myb	Myb TF	Regulate growth and development
Ceric.28G017900	WRKY	WRKY DNA-binding protein 50	Regulate dormancy, germination, gametophyte and seed development

5.2.7.3 Ivory module: Network visualisation

To further explore the relationships between genes within the ivory module, a network map was constructed to visualise gene-gene correlations (Figure 5.15). The network includes only strong correlations between genes, with edges representing correlations of $|r| \geq 0.7$. The edge widths reflect the strength of the correlation, where thicker edges indicate stronger associations. Positive correlations are represented by grey edges, and negative correlations by red edges. Nodes represent individual genes, and their size is proportional to their module membership (kME), with larger nodes indicating genes that are more central to the module.

The nodes were further styled to distinguish between different functional categories (Figure 5.15). Hub genes are highlighted as diamonds with a red border,

signifying their central importance in the network. Additionally, TFs are shaded green to emphasize their likely regulatory roles. Notably, two genes are both hub genes and TFs (Tables 5.2 and 5.3, indicating that these genes may serve as critical regulators of the developmental processes associated with the module. These dual-function genes are represented as green diamonds with a red border, underscoring their potentially pivotal roles in orchestrating reproductive tissue development.

The ivory module network exhibits a densely interconnected structure, with hub genes occupying central positions and forming the core of the network. The remaining four transcription factors are distributed more evenly across the network. In contrast, several genes are less well connected and positioned at the periphery of the network, suggesting more limited involvement in the core processes represented by this module. Interestingly, both genes that exhibited a negative correlation with the module eigengene are located at the network's outskirts. This positioning suggests that these negatively correlated genes may be functionally distinct from the core network and could represent specialized roles or context-specific expression patterns. Their reduced connectivity implies that they may be regulated by different mechanisms than the core, positively correlated genes, potentially reflecting unique functions within the broader developmental context.

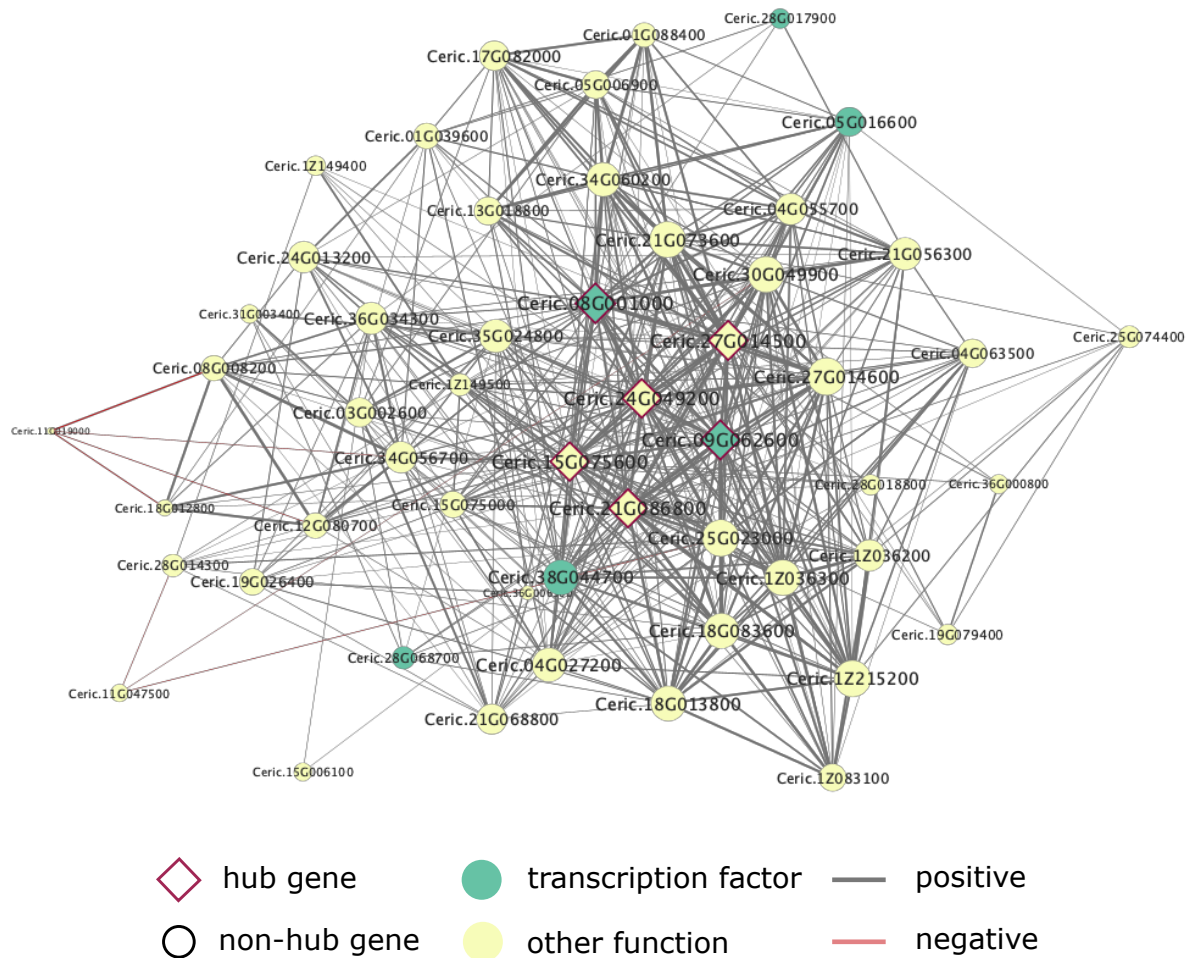


Figure 5.15: Network visualisation of the ivory module. Gene co-expression network highlighting the relationships between genes based on Pearson correlation values. Nodes represent individual genes, while edges (connections between nodes) represent significant correlations between gene expression profiles across tissue samples. Node size is proportional to intramodular connectivity (kME), edge width is proportional to correlation value. Diamond with red border = hub gene, green node = transcription factor. Grey edge = positive correlation, red edge = negative correlation.

5.2.7.4 Darkorange module: gene-trait correlation analysis

The darkorange module comprises 126 genes and exhibits a strong positive correlation with the expanding and mature reproductive fronds of the 130-day

sporophyte (Figure 5.10). The expression profiles of individual genes within this module were assessed across different tissue types. To enable visualisation, the top 50 most correlated genes to the expanding and mature 130-day fronds are shown in Figure 5.16.

The heatmap reveals a consistent pattern of negative correlation across most tissue types, except in the expanding and mature reproductive fronds, where correlations are strongly positive. The mature frond shows the highest correlation for most genes, while the expanding frond exhibits a slightly lower but still distinct correlation compared to other tissues. Out of the 10 hub genes identified within the darkorange module (Table 5.4), six are ranked among the top 50 most correlated genes and are shown in Figure 5.10. Notably, five of these six hub genes (indicated by filled circles) cluster tightly together, suggesting co-regulation and shared functional roles in reproductive frond development. The remaining hub gene falls into a more distant cluster, indicating it may have a partially distinct regulatory role or respond to different developmental cues.

Two genes—Ceric.02G004600 (a MYB transcription factor) and Ceric.25G047600 (a cytochrome P450 homolog)—show stronger correlation with the expanding 130-day frond compared to the mature frond. The higher expression of these two genes in the expanding frond suggests they are critical for early sporangial development. In contrast, the mature frond likely supports processes related to sporangium maturation, meiosis, and spore formation, which would require a different set of regulatory genes. This developmental progression was further explored through GO enrichment analysis.

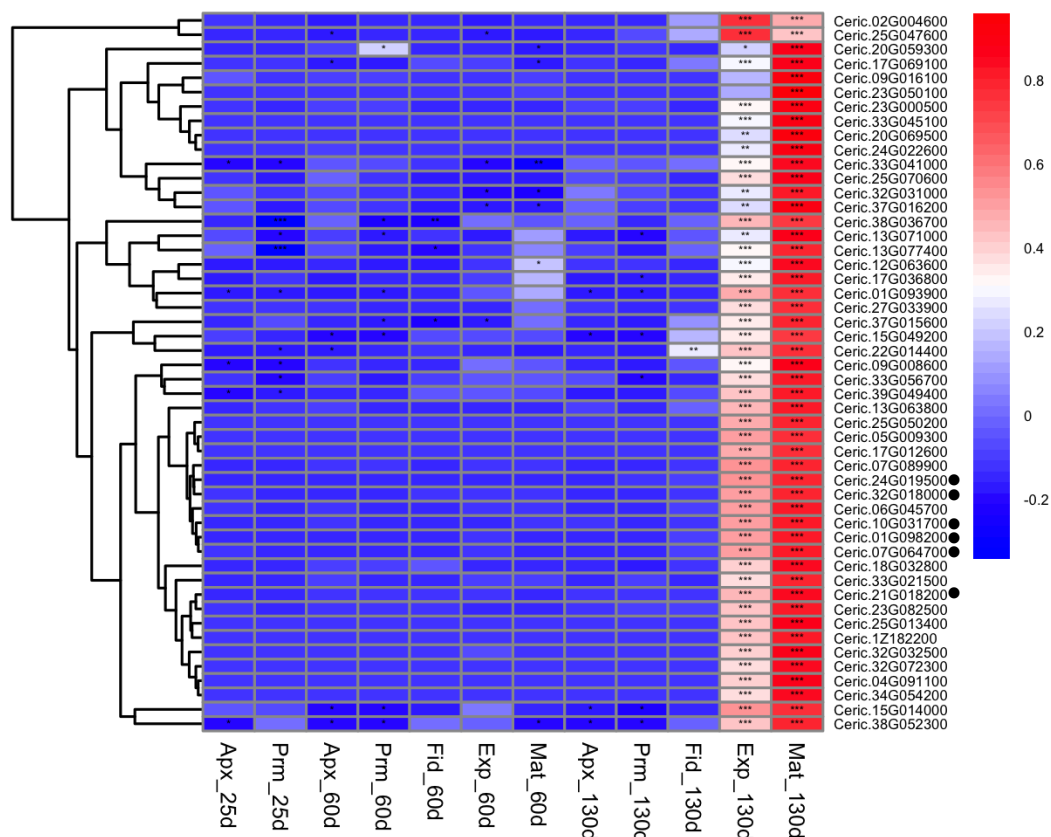


Figure 5.16: Gene-trait correlation heatmap for the darkorange module. Gene-trait correlation heatmap for the darkorange module. Expression correlation of the top 50 genes from the darkorange module across different tissue types of *Ceratopterus richardii*. Genes (rows) are clustered by similar correlation patterns, while tissues (columns) are ordered to reflect developmental progression. Blue indicates negative correlation, white indicates no correlation, and red indicates positive correlation. Hub genes are indicated by filled circles.

5.2.7.5 Darkorange module: Gene Ontology enrichment analysis

To further investigate the potential functions of genes within the darkorange module, GO enrichment analysis was conducted, identifying significantly overrepresented BP, MF, and Cellular Component (CC) terms (Figure 5.17). Among the enriched terms, copper ion transmembrane transport (BP; Figure 5.17A), heme binding (MF; Figure 5.17B), and copper ion transmembrane transporter activity (MF) remained

Table 5.4: Annotation and putative functions of hub genes in the darkorange module.

Gene ID	Annotation	Protein Domains	Putative Function
Ceric.24G019500	KOG1603 heavy metal-associated	Copper chaperone	Metal ion transport
Ceric.10G031700	HOMEODOMAIN PROTEIN KNOTTED-1-LIKE 3-RELATED	Homeobox, KNOX1	Regulation of DNA-templated transcription
Ceric.18G025500	Acetyl-CoA carboxylase	-	Fatty acid synthesis
Ceric.21G018200	No hit	-	-
Ceric.01G098200	Bifunctional inhibitor/lipid-transfer protein/seed storage 2S albumin superfamily protein	-	Lipid transfer and membrane stabilisation
Ceric.03G048700	Bifunctional inhibitor/lipid-transfer protein/seed storage 2S albumin superfamily protein	-	Lipid transfer and membrane stabilisation
Ceric.32G018000	Cupredoxin superfamily protein	-	Copper ion binding
Ceric.07G064700	Bifunctional inhibitor/lipid-transfer protein/seed storage 2S albumin superfamily protein	-	Lipid transfer and membrane stabilisation
Ceric.38G061600	RNA dependent DNA polymerase	-	DNA methylation
Ceric.24G042000	No hit	-	-

significant after correction for multiple testing, highlighting metal ion transport and redox-related processes as key molecular functions associated with reproductive frond development.

Additional enriched BP terms included reciprocal meiotic recombination, DNA metabolic processes, and DNA repair (Figure 5.17A). In the MF category, the enrichment of terms such as heme binding and oxidoreductase activity (Figure 5.17B) further suggests roles in redox regulation. In the CC category, enrichment of nuclear chromosome, nucleus, and intracellular anatomical structure (Figure 5.17C) reflects the localization of darkorange module genes to the nucleus. The presence of

nuclear-associated terms is consistent with processes involving DNA replication, repair, and chromosome segregation, further supporting the potential involvement of this module in meiotic events.

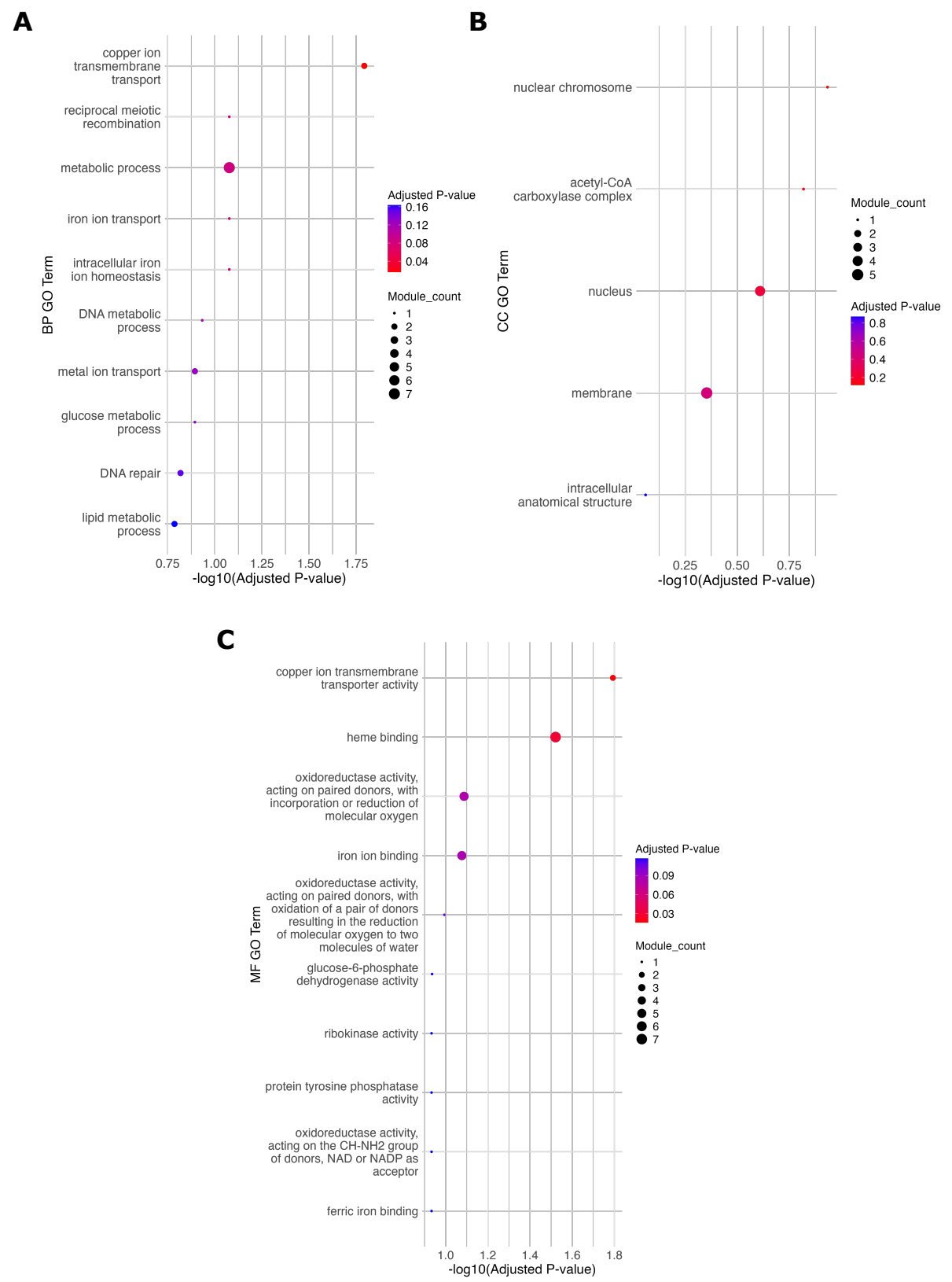


Figure 5.17: Caption overleaf.

Figure 5.17 Gene Ontology (GO) enrichment analysis of the darkorange module. Enriched terms were identified across three GO categories: (A) Biological Process (BP), (B) Molecular Function (MF) and (C) Cellular Compartment (CC). Dot size indicates the number of genes associated with the term (Module_count) and colour indicates the adjusted p-value ranging from blue (less significant) to red (more significant).

5.2.7.6 Darkorange module: network visualisation

To further explore gene relationships within the darkorange module, a network map was constructed, filtered to include only strong correlations ($|r| \geq 0.8$) (Figure 5.18). The resulting structure is very dense, with many genes exhibiting a module membership (kME) ≥ 0.9 , indicating strong co-expression relationships across the module. Six of the ten hub genes occupied central positions.

The most central hub gene was identified as a homeobox/*KNOTTED-1-LIKE 3-RELATED* TF, suggesting a potential regulatory role within this module. Other central hub genes included a mix of regulatory proteins and enzymes involved in metal ion transport, epigenetic regulation, lipid transfer and cell membrane stabilisation, highlighting a potential integration of regulatory and metabolic processes (Table 5.4).

The network contains three additional highly connected TFs: two MYBs and a *basic helix-loop-helix (bHLH)* (Table 5.5). These TFs are evenly dispersed across three edges of the network rather than clustered at the centre, indicating that they may regulate different aspects of the module's activity (Figure 5.18). Two less connected transcription factors – *PHYTOCHROME RAPIDLY REGULATED 1 (PAR1)* and another *bHLH* TF (Table 5.5) – form a closely grouped cluster on the periphery of the network (Figure 5.18).

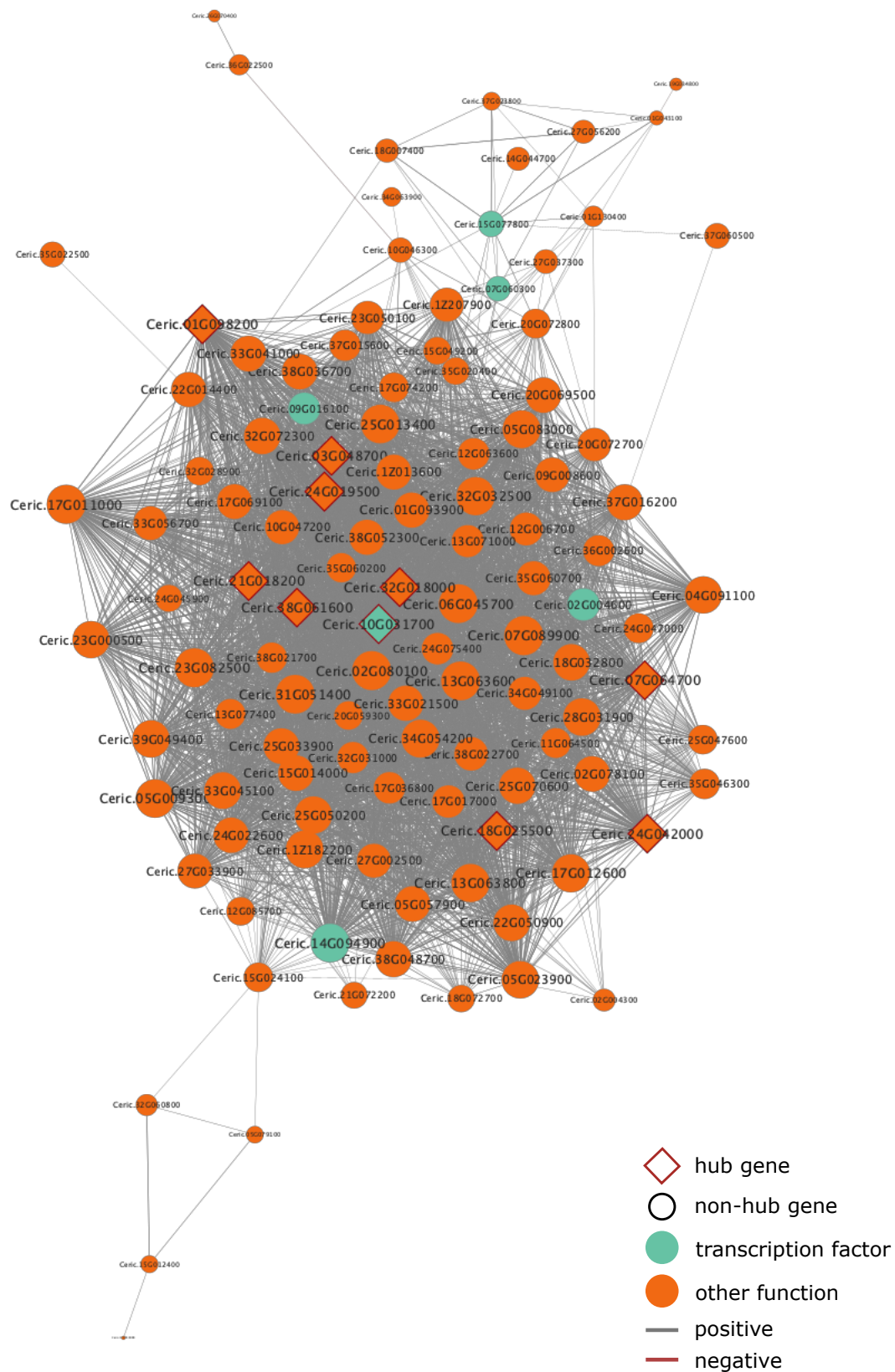


Figure 5.18: *Caption overleaf.*

Figure 5.18 Network visualisation of the darkorange module. Gene co-expression network highlighting the relationships between genes based on Pearson correlation values. Nodes represent individual genes, while edges (connections between nodes) represent significant correlations between gene expression profiles across tissue samples. Node size is proportional to intramodular connectivity (kME), edge width is proportional to correlation value. Diamond with red border = hub gene; green node = transcription factor. Grey edge = positive correlation; red edge = negative correlation.

Table 5.5: Transcription factors identified in the darkorange module along with their corresponding annotations and putative functions.

Gene ID	TF Family	Annotation	Putative Function
Ceric.10G031700	Homeobox	HOMEBOX PROTEIN 3-KNOTTED-1-LIKE RELATED	Regulation of DNA-templated transcription
Ceric.14G094900	bHLH	Basic helix-loop-helix (bHLH) DNA-binding superfamily protein	Regulation of developmental processes and stress responses
Ceric.02G004600	MYB	MYB domain protein 101	Anther/pollen development
Ceric.09G016100	MYB	MYB domain protein 80 (formerly 103)	Anther development and/or functionality
Ceric.15G077800	bHLH	Basic helix-loop-helix (bHLH) DNA-binding superfamily protein	Regulation of developmental processes and stress responses
Ceric.07G060300	Atypical bHLH	PHYTOCHROME RAPIDLY REGULATED 1 (PAR1)	Regulation of development and shade avoidance response
Ceric.17G049600	MYB	MYB family transcription factor	Regulation of development, metabolism and/or stress responses

5.2.8 Co-expression network assignment of *SPL* homologs

The three *SPOROCTELESS* (*SPL*) homologs identified and discussed in Chapters 3 and 4 were assigned to two distinct co-expression modules within the weighted gene co-expression network. *SPLA* and *SPLC* were both assigned to the red module, which contained 13,165 genes. The red module exhibited a moderate positive correlation with shoot apex and frond primordia tissues across all ages, followed by a progressive decrease in correlation as fronds matured, becoming strongly negative in the mature fronds of both the 60-day and 130-day sporophytes. Given the size and heterogeneity of the red module, a detailed investigation of its expression profile and hub genes was not feasible. The lack of a distinct, tissue-specific expression pattern within this module suggests that *SPLA* and *SPLC* are unlikely to play a specialized regulatory role in reproductive tissue development.

SPLB was assigned to the cyan module, consisting of 897 genes, and showed a positive correlation with the ME. The cyan module exhibited a consistent expression profile across tissues, regardless of plant age, with higher correlation in the shoot apex and mature fronds and lower correlation in the frond primordia, fiddlehead and expanding fronds at all developmental stages sampled. The module contained genes involved in a wide range of biological processes, including transcriptional regulation, metabolism, nutrient transport, and stress responses.

Among the positively correlated genes in the cyan module were several TFs in addition to *SPLB*, including *WRKY40*, *bHLH* and a nuclear transcription factor Y subunit, all of which may play roles in regulating tissue-specific gene expression. Genes involved in carbohydrate metabolism were also prominent, such as Phosphoenolpyruvate carboxykinase and pyruvate kinase, suggesting active metabolic processes in tissues showing high correlation with MEcyan. Additionally, transport proteins - including potassium channels, choline transporters and major facilitator family proteins - were identified, indicating potential roles in nutrient transport and ion homeostasis. Several stress-related genes, such as peroxidases and universal stress protein domain-containing proteins, were also present.

In contrast, genes that were negatively correlated with the cyan module showed the opposite expression profile, with higher expression in the frond primordia, fiddlehead and expanding fronds. Many of these genes were associated with chloroplast function and photosynthesis, including magnesium-chelatase subunit CHLI, ferredoxin-NADP reductase and chloroplast channel forming outer membrane proteins, all of which are involved in chloroplast biogenesis and photosynthetic activity. A number of TFs were also identified among the negatively correlated genes, including GATA zinc finger proteins, C2H2 zinc finger proteins, and bZIPs, which may regulate genes associated with photosynthetically active tissues. Other notable genes in the negatively correlated set include histidine kinases, heat shock proteins, and superoxide dismutases, indicating potential roles in stress responses and oxidative

stress regulation.

5.3 Discussion

This chapter set out to address two key gaps in *C. richardii* transcriptomics: the limited testing of recently published genomic resources, and the lack of comprehensive co-expression network analyses to investigate regulatory processes during fern development. Despite the publication of a high-quality genome and accompanying transcriptome in 2022 (Marchant et al., 2022), these resources remain unvalidated in independent studies. The comparison of HISAT2 and Salmon revealed that genome-guided mapping was more effective, likely due to incomplete representation of transcripts in the transcriptome (Kim et al., 2019; Patro et al., 2017).

A weighted gene co-expression network analysis (WGCNA) was conducted to explore gene expression patterns associated with reproductive development in *C. richardii* (Langfelder & Horvath, 2008; Zhang & Horvath, 2005). WGCNA enabled the identification of co-expressed gene modules, each representing distinct sets of genes with correlated expression profiles across different tissues and developmental stages. By focusing on modules significantly correlated with reproductive fronds, this study uncovered regulatory networks that may play key roles in fern reproductive development. Notably, several transcription factors (TFs) were identified as hub genes in these modules, suggesting potential roles in coordinating developmental processes such as tissue differentiation, reproductive structure formation, and stress responses.

5.3.1 Mapping Performance Reveals a Well-Resolved *C. richardii* Genome but a Limited Transcriptome

The recent publication of the *C. richardii* genome (v2.1) and accompanying reference transcriptome (v2.0) marked a significant advancement in genomic resources for this model fern species (Marchant et al., 2022). However, the comparison of genome-guided and transcriptome-based RNA-seq mapping

approaches revealed important limitations in the current transcriptome. HISAT2, a genome-guided method, consistently outperformed Salmon, a transcriptome-based method, in alignment rates across all tissue types (Kim et al., 2019; Patro et al., 2017). This suggests that the v2.0 reference transcriptome remains incomplete and does not capture the full transcript diversity of *C. richardii* (Marchant et al., 2022).

The investigation of unmapped and decoy reads provided further evidence of these gaps. The majority of unmapped reads aligned to annotated genes in the v2.1 genome, indicating that many transcripts were missed in the original transcriptome assembly (Appendix C) (Marchant et al., 2022). The published transcriptome was assembled using reads from ten tissues and developmental stages, including two gametophyte tissues (“immature” and “mature”) as well as whole young sporophyte, roots, stem, fiddlehead and expanding frond (from unspecified plant ages), mature vegetative and reproductive fronds, and sporangia. While these tissues capture a broad range of developmental stages, the limited sampling may have resulted in the exclusion of transcripts expressed in more specific or transient conditions.

In contrast, the tissues sampled in this study included shoot apices and frond primordia across distinct developmental time points (25-day, 60-day, and 130-day plants) and captured the progression of frond development in both vegetative and reproductive sporophytes. These additional tissues likely contain unique transcriptional programs that were not represented in the original assembly. Future efforts to improve *C. richardii* transcriptome resources should prioritize the integration of existing RNAseq datasets, including those generated in this chapter. Multiple independent studies have already produced high-quality transcriptome data from diverse tissue types and experimental conditions, including additional root stages (Aragón-Raygoza et al., 2022), callus tissue (Xiao & Li, 2024), and a comprehensive library of gametophyte and sporophyte tissues (Geng et al., 2021). Transcriptome data from studies investigating environmental responses, such as gravity sensing (Bushart et al., 2013) and hormonal treatments (Atallah et al., 2018; Woudenberg et al., 2024),

could further enrich the reference transcriptome by broadening the range of represented transcripts. Incorporating these datasets into the reference transcriptome would significantly enhance its completeness, by capturing transcripts expressed in specialized tissues or under specific developmental or environmental conditions.

Many of the above studies have focused primarily on identifying differentially expressed genes (DEGs), leaving their raw sequencing data as an underutilized resource. Leveraging these datasets to update the *C. richardii* transcriptome could improve annotation accuracy and better reflect the species' full transcriptional diversity. While many of these studies already conducted *de novo* assembly (Aragón-Raygoza et al., 2022; Atallah et al., 2018; Bushart et al., 2013; Geng et al., 2021), further integration could involve aligning these assembled transcripts to the high-quality v2.1 genome to identify and correct misassemblies, refine gene models, and capture missing transcripts (Lischer & Shimizu, 2017). Additionally, raw RNAseq reads from these studies could be directly mapped to the genome using HISAT2 or similar genome-guided approaches to identify novel transcripts and refine existing annotations (Kim et al., 2019). Combining these transcriptomes would likely improve the accuracy and completeness of transcriptome-based mapping approaches, such as Salmon, and enhance the utility of *C. richardii* as a model system for transcriptomic research (Patro et al., 2017).

5.3.2 Normalisation and filtering of mapped reads

To ensure accurate co-expression network construction, mapped read counts must be normalised and filtered (Conesa et al., 2016). In this study, counts were normalized using FPKM (Fragments Per Kilobase of transcript per Million mapped reads) rather than TPM (Transcripts Per Million) or raw count-based normalization methods such as DESeq2 (Love et al., 2014) and edgeR (Robinson et al., 2010). FPKM and TPM are both widely used normalization methods for RNAseq data, with the primary difference being how they scale gene expression values within each sample

(Conesa et al., 2016). TPM scales transcript abundances so that the total expression across all transcripts in a sample sums to one million, making it useful for comparing relative transcript abundances within a sample. FPKM, on the other hand, normalizes by gene length and sequencing depth across samples, which makes it better suited for between-sample comparisons. For co-expression network analysis, where the goal is to identify consistent correlations between genes across multiple samples, FPKM is a more suitable choice (Langfelder & Horvath, 2008).

In contrast, the normalization methods employed in DESeq2 and edgeR are designed for differential expression analysis rather than co-expression analysis (Love et al., 2014; Robinson et al., 2010). These methods use sophisticated algorithms to account for library size differences and variability in read counts, making them highly effective for identifying genes that are differentially expressed between groups. However, WGCNA is sensitive to the range and distribution of expression values, and methods like DESeq2's variance-stabilizing transformation (VST) or edgeR's TMM normalization reduce variance across samples, potentially reducing the sensitivity of correlation-based analyses (Langfelder & Horvath, 2008; Love et al., 2014; Robinson et al., 2010; Zhang & Horvath, 2005).

Following normalization, genes with FPKM < 0.5 in fewer than three samples were removed from the dataset to reduce noise from low-expressed genes. While traditional differential expression analyses often retain low-abundance genes to maximize statistical power, WGCNA is more sensitive to spurious correlations, making the removal of low-count genes essential for constructing biologically meaningful networks (Langfelder & Horvath, 2008; Zhang & Horvath, 2005). Filtering lowly expressed genes ensures that the co-expression relationships identified are robust and less influenced by random noise.

5.3.3 Co-expression modules revealed distinct gene networks across developmental stages

Through the application of WGCNA, this study identified 21 distinct co-expression modules in *C. richardii*, each with unique associations to tissue types, developmental stages or plant age. These modules represent distinct gene networks that may drive key biological processes during the development of the fern sporophyte, providing a valuable framework for future research.

While the initial focus was on the three *CrSPL* homologs identified in earlier chapters, these genes were assigned to the red (*CrSPLA* and *CrSPLC*) and cyan (*CrSPLB*) modules, which did not display specific associations with reproductive tissues. The red module was primarily associated with the shoot apex and frond primordia tissues, reflecting a broader role in early vegetative growth, while the cyan module exhibited correlations that primarily reflected plant age rather than specific developmental processes. Genes correlated with the cyan module suggested a network with a more general role in supporting frond development and tissue maintenance, rather than being specifically involved in reproductive development. Consequently, the focus shifted to modules with clearer biological relevance, such as MEivory and MEdarkorange, to investigate their potential roles in tissue differentiation and reproductive development.

The correlation of some modules with particular plant ages (e.g., MEdarkred and MEgrey) highlights the potential to investigate age-dependent gene expression changes during the transition from early vegetative growth to reproductive development. In contrast, modules that correlate with specific tissue types, such as MEivory and MEdarkorange, suggest the presence of genes that may play specialized roles in tissue differentiation or reproductive transitions. This specificity makes the identified modules a valuable resource for researchers interested in exploring distinct biological questions, from developmental regulation to responses to environmental

stimuli. For example, modules like MEgrey60, which correlate with mature fronds, could be of interest to researchers studying frond maintenance, nutrient uptake, or senescence processes, while modules such as MEivory and MEdarkorange may provide insight into the genetic regulation of reproductive tissue development.

5.3.4 Regulatory networks specific to early reproductive frond development: insights from the ivory module

The ivory module offers valuable insights into the transcriptional networks underpinning critical developmental transitions in *C. richardii*. Its strong association with the fiddlehead and expanding fronds of 130-day-old sporophytes suggests that it governs processes essential for tissue differentiation and expansion during the reproductive transition. The presence of hub genes and transcription factors within this module highlights its regulatory complexity and potential centrality to developmental programming.

The ivory module contains seven transcription factors (TFs), representing diverse regulatory families. These TFs include members of the MYB, WRKY, AP2/ERF, and homeobox families (Table 5.3), which are known to play critical roles in plant development (Bakshi & Oelmüller, 2014; Dubos et al., 2010; Elhiti & Stasolla, 2009; Ma et al., 2024). MYB transcription factors, for example, are widely recognized for their involvement in growth regulation and tissue differentiation, particularly during reproductive development and pollen formation (Dubos et al., 2010). Similarly, WRKY transcription factors are versatile regulators involved in processes ranging from stress responses to developmental regulation (Bakshi & Oelmüller, 2014). Their co-expression in the ivory module highlights the potential for coordinated activity in reproductive fronds where developmental reprogramming to specify sporangia is crucial (Conway & Di Stilio, 2020).

Notably, two TFs were also identified as hub genes, underscoring their central regulatory roles. The HOMEBOX-LEUCINE ZIPPER PROTEIN ANTHOCYANINLESS

2-RELATED TF (Ceric.08G001000) has been implicated in anthocyanin biosynthesis, a pathway associated with pigmentation changes (Kubo et al., 1999). This is particularly intriguing in the context of the ivory module, as sporangia turn red as they mature, suggesting a potential link between this gene and flavonoid biosynthesis in pigmentation processes. Another hub gene, flavonoid 3'-monooxygenase (Ceric.21G086800), also plays a key role in flavonoid biosynthesis, specifically in modifying flavonoids that contribute to pigmentation, UV protection, and structural integrity (Morris et al., 2016). The co-expression of these genes suggests a coordinated regulatory mechanism controlling flavonoid biosynthesis during reproductive tissue development. Together, these genes may drive the production and deposition of pigments in maturing sporangial tissues. This highlights the ivory module's likely involvement in regulating flavonoid-related processes during reproductive development in *C. richardii*.

GO enrichment analysis highlighted processes associated with transport and metabolism, particularly xenobiotic membrane transport, lipid metabolism, and structural components of the cell wall. The physiological demands of fiddlehead and expanding frond tissues, such as active cell wall remodelling, nutrient transport, and lipid synthesis are well documented in plant biology. For instance, studies have shown that cell wall remodelling facilitates cell differentiation during embryogenic transitions in ferns (Grzyb et al., 2024), while lipid-modified proteins are integral to maintaining cell wall integrity and signalling during expansion (Quinn et al., 2023). Additionally, xenobiotic transport processes may reflect the metabolic requirements of actively growing tissues, which often experience elevated oxidative stress due to heightened respiration and metabolic activity. This aligns with evidence suggesting that plants employ xenobiotic transport and detoxification pathways to mitigate oxidative damage and maintain cellular function under stress (Viciomini et al., 2024). Together, these findings suggest that the ivory module encompasses genes that are integral to the metabolic and structural demands of tissue differentiation and growth

in *C. richardii*.

The functional dichotomy within the ivory module adds an intriguing layer of complexity to its regulatory potential. Genes positively correlated with the module eigengene predominantly encode transcriptional activators, including a histone methyltransferase (e.g., Ceric.28G018800), which likely promotes gene expression during frond development through chromatin remodelling (He et al., 2021). In contrast, negatively correlated genes, such as Ceric.11G019000, a histone deacetylase homolog, suggest a concurrent mechanism of transcriptional repression (Kumar et al., 2021). This juxtaposition of transcriptional activation and the repression of transcriptional silencing suggests a finely tuned balance that promotes gene induction while simultaneously reducing inhibitory pathways.

The ivory module provides valuable insights into the molecular mechanisms underlying reproductive fiddlehead and expanding frond development in *C. richardii*. The identification of transcription factors as both hub genes and key regulators highlights the importance of coordinated gene networks in facilitating developmental transitions. The enrichment of GO terms related to transport and metabolic processes underscores the physiological demands of rapidly expanding tissues, while the functional dichotomy between transcriptional activation and the suppression of repression reveals the sophisticated regulatory balance required to fine-tune gene expression during reproductive frond maturation. Together, these findings offer a foundation for future investigations into the genetic and regulatory pathways that govern fern development, providing potential targets for studying broader aspects of plant growth and adaptation.

5.3.5 Regulatory networks driving later reproductive frond development: insights from the darkorange module

The darkorange module sheds light on the transcriptional networks underpinning later stages of reproductive frond development in *C. richardii*. Its strong positive

correlation with the expanding and mature fronds of 130d sporophytes suggests a central role in coordinating gene expression during the progression from sporangium initiation to maturation and spore development. The presence of gene homologs with roles in reproductive processes in *Arabidopsis* and the enrichment of terms related to meiosis, suggest that this module features genetic networks responsible for the specific regulation of reproductive processes.

Seven TFs were placed in the darkorange module, spanning diverse families such as homeobox, MYB, and basic helix-loop-helix (bHLH) (Table 5.5). These TFs are well-known for their roles in developmental regulation, including anther and pollen development, stress responses, and cell differentiation (Chan et al., 1998; Dubos et al., 2010; Hao et al., 2021). The central hub gene, *HOMEBOX PROTEIN KNOTTED-1-LIKE 3-RELATED* (*KNAT3-related*, Ceric.10G031700), underscores the importance of transcriptional regulation within this module. Homeobox genes, particularly those in the KNOX family, are extensively documented for their roles in regulating meristem activity, leaf development, and organ patterning (Hake et al., 2004; Sakakibara et al., 2008). Notably, *KNAT3* (the *Arabidopsis* homolog of Ceric.10G031700) plays a critical role in integument and ovule development alongside *KNAT4*, with loss-of-function mutants exhibiting infertility due to defective ovule formation (Chen et al., 2023). The centrality of Ceric.10G031700 within the darkorange network suggests it may play a similarly essential role in coordinating reproductive transitions during frond maturation in *C. richardii*, which may have been co-opted for seed-specific developmental processes in the seed plants.

Two MYB TFs (Ceric.02G004600 and Ceric.09G016100) are linked to reproductive tissue specification, with evidence suggesting roles in anther and pollen development in other plant systems (Dubos et al., 2010). The spatial distribution of these TFs within the network suggests functional specialization, with MYBs localized to peripheral regions, potentially influencing discrete processes within the broader developmental program (Figure 5.18). Interestingly, MYB80 (Ceric.09G016100) is involved in tapetal

layer development and pollen viability in *Arabidopsis* (Dubos et al., 2010). In *C. richardii*, tapetum layers enclose the sporogenous cells within the sporangia, hinting at a potential conserved regulatory pathway for tapetum layer formation during sporangial development (Withers et al., 2023).

GO enrichment analysis highlighted processes associated with metal ion transport (copper, iron and ferric ion transport), redox regulation (heme binding, oxidoreductase activity) and meiosis (reciprocal meiotic recombination, DNA metabolic processes, and DNA repair) (Figure 5.17). The enrichment of terms related to meiosis suggests that darkorange genes may be responsible for the specific regulation of meiosis within the sporangia of reproductive fronds. Reciprocal meiotic recombination is essential for the formation of crossovers between homologous chromosomes during prophase I of meiosis, whilst enrichment of DNA repair indicates the active resolution of double-strand breaks, a critical step in meiotic recombination (Hunter, 2015).

The enrichment of terms related to redox state regulation further support a role in meiotic regulation. Indeed, the entire process of plant sexual reproduction is redox regulated (Zinta et al., 2016). The process of homologous recombination and double-strand break repair generates oxidative stress, and the activity of heme-binding proteins and oxidoreductases may be necessary to prevent DNA damage and maintain cellular homeostasis (Li et al., 2024). Copper and iron ion homeostasis are also important in the maintenance of redox states, reinforcing the hypothesis that the darkorange network is responsible for regulating the complex processes governing meiosis (Ravet & Pilon, 2013). The enrichment of terms to the nucleus and nuclear chromosome further support this. Together, these enriched terms suggest that the darkorange module features a network of highly connected genes responsible for the tight regulation of meiosis within the sporangia.

5.3.6 Identifying future targets in reproductive development

The identification of the ivory and darkorange modules provides a robust foundation for understanding the transcriptional regulation of reproductive frond development in *C. richardii*. While initial analyses of the *CrSPL* homologs yielded limited insights due to their placement in non-specific modules, the exploration of these alternative modules has revealed new candidate genes with compelling roles in reproductive transitions. The ivory module, associated with the early stages of reproductive frond differentiation, highlights processes such as transcriptional activation, lipid metabolism, and structural remodelling essential for tissue expansion. In contrast, the darkorange module is linked to later stages, encompassing meiosis, sporangium maturation, and spore formation.

Genes such as Ceric.08G001000 (a homeobox transcription factor in ivory) and Ceric.10G031700 (*KNAT3-related* in darkorange) emerge as promising candidates for future functional studies due to their likely roles in regulating reproductive transitions. Likewise, MYB80 (Ceric.09G016100) in the darkorange module and flavonoid biosynthesis genes in the ivory module present intriguing opportunities to explore potential conservation of regulatory pathways between ferns and seed plants. These findings underscore the power of network-based approaches like WGCNA to uncover previously uncharacterised regulators of plant development. Future work should aim to experimentally validate the roles of these hub genes and transcription factors, leveraging tools like CRISPR-Cas9 or RNA interference to dissect their contributions to fern reproductive biology. Together, these findings provide a valuable framework for investigating the molecular mechanisms underpinning reproductive development in ferns and their evolutionary links to seed plant systems.

Chapter 6

General Discussion

6.1 Introduction: Uncovering the genetic basis of reproductive development

The evolution of seeds represents one of the most transformative innovations in land plant history, facilitating the diversification and dominance of seed plants in terrestrial ecosystems (Harris & Davies, 2016; Linkies et al., 2010; Willis & McElwain, 2014). Since plants first appeared on land approximately 500 million years ago, a series of key evolutionary innovations set the stage for the emergence of the seed habit (Morris et al., 2018; Nie et al., 2019; Strother & Foster, 2021). Early bryophytes, characterised by their gametophyte-dominant life cycle and reliance on free water for fertilisation, were constrained by short-lived, dependent sporophytes (Chang & Graham, 2011; Niklas & Kutschera, 2010; Qiu et al., 2006; Shaw et al., 2011; Wickett et al., 2014). The rise of vascular plants (tracheophytes) marked a major shift to sporophyte dominance, supported by vascular tissues that enabled larger and more complex plant forms (Kenrick & Crane, 1997; Kenrick & Strullu-Derrien, 2014; Niklas, 1997). Despite this shift, early vascular plants, such as lycophytes and ferns, retained free-living gametophytes and continued to rely on water for fertilisation (Bateman & DiMichele, 1994).

In gymnosperms and angiosperms, the evolution of heterospory and highly reduced gametophytes that are entirely dependent on the sporophyte allowed reproduction without the need for water (Bateman & DiMichele, 1994; Linkies et al., 2010; Raven, 2002). These reproductive innovations were underpinned by genetic networks that regulated key developmental processes (Pabón-Mora et al., 2019). In angiosperms, for instance, genes like *SPOROCTELESS* (*SPL*) are pivotal for ovule development, the precursor to the seed (Schiefthaler et al., 1999; Yang et al., 1999).

This thesis aimed to address key gaps in our understanding of how genetic networks underpinning reproductive development in seed plants have evolved from

their non-seed plant ancestors. While genes such as *SPL* are well-studied in angiosperms, their evolutionary conservation and roles in earlier-diverging lineages remain largely unexplored. To address this, the thesis focused on the model fern *Ceratopteris richardii*, which occupies a key evolutionary position as a member of the monilophytes, bridging bryophytes and seed plants (Hickok et al., [1995](#); Pryer et al., [2001](#)). The work aimed to uncover the roles of *SPL* homologs in ferns, explore their evolutionary relationships with seed plant *SPL* genes, and broaden the investigation to identify other genetic networks involved in reproductive development.

6.2 Chapter 3: Evolutionary insights into *C. richardii* SPL

In Chapter 3 the evolutionary and functional significance of *SPL* and *SPEAR* homologs in *C. richardii* was explored, revealing insights into their roles in land plant evolution. Three *SPEAR* homologs, *CrSPLA*, *CrSPLB*, and *CrSPLC*, were identified, all exhibiting high conservation in the *SPL* and *EAR* domains, consistent with findings from other embryophytes (Chen et al., 2014; Liu et al., 2018; Ren et al., 2018; Rojas-Gracia et al., 2017). Phylogenetic analyses placed *CrSPLB* and *CrSPLC* within the *SPL* clade, alongside angiosperm and gymnosperm sequences, suggesting that *SPL* function predates the evolution of seeds. This finding raises the possibility that *SPL* genes played ancestral roles in embryophyte reproduction before being co-opted for seed-specific functions in flowering plants.

6.2.1 Phylogenetic insights into *SPL* evolution

The phylogenetic analyses in this study represent a significant advancement over previous work by incorporating a broader range of sequences from diverse plant lineages, including non-flowering plants such as gymnosperms, ferns, lycophytes, and bryophytes (Chen et al., 2014). This comprehensive sampling enabled the placement of fern *CrSPLB* and *CrSPLC* within the *SPL* clade, highlighting their potential conservation as key regulators of reproduction. Importantly, this expanded dataset revealed that, contrary to previous thought, *SPL*-like sequences likely evolved early in land plants, with gene duplications occurring in vascular plants coinciding with the transition to sporophyte dominance (Chen et al., 2014; Kenrick & Strullu-Derrien, 2014). However, the phylogeny also highlights ongoing challenges in resolving deep evolutionary relationships (Lemmon et al., 2009). Limited availability of genomic and transcriptomic data from basal lineages, particularly non-flowering embryophytes, constrained the ability to fully resolve relationships among early-diverging *SPL* sequences (Xie et al., 2024). The placement of lycophyte sequences, for example,

varied between maximum likelihood and Bayesian analyses, reflecting the need for more comprehensive taxon sampling to improve the resolution of deep branches. These findings underscore the importance of expanding phylogenetic datasets to include newly sequenced taxa as they become available. Complementary approaches, such as structural-based alignments, could also aid in resolving deep nodes, particularly in cases where conserved domains are interspersed with poorly aligned regions (Carpentier & Chomilier, 2019; Malik et al., 2020; Rajapaksa et al., 2023).

6.2.2 Expression patterns of *CrSPLs* in reproductive development

Expression analyses of *CrSPLA*, *CrSPLB*, and *CrSPLC* provided further insights into their potential functions. While *CrSPLA* was broadly expressed in gametophyte and sporophyte tissues, suggesting a role in general growth, *CrSPLB* exhibited strong expression in reproductive tissues only, indicating a specialized role in reproduction. Conversely, *CrSPLC* was not strongly expressed in gametophyte tissues, but expression was detected in reproductive sporophyte tissues, indicating a potential sporophyte-specific role in reproduction. The expression patterns of *CrSPLB* and *CrSPLC* are consistent with observations in flowering plants, where *SPL* genes are strongly expressed in the anthers and ovules (Liu et al., 2018; Ren et al., 2018; Rojas-Gracia et al., 2017; Schiefthaler et al., 1999; Yang et al., 1999). These findings reinforced *CrSPLB* and *CrSPLC* as good candidates for a conserved regulatory role in reproduction.

6.2.3 Conservation and divergence of *CrSPL* function in land plants

Together, these findings establish *CrSPLB* and *CrSPLC* as strong candidates for *SPL*-like reproductive functions in *C. richardii*. Their placement within the *SPL* clade and association with reproductive tissues support the hypothesis that *SPL* functions evolved early in land plants and were subsequently modified in vascular plants. By integrating phylogenetic and expression data, this chapter lays the groundwork for

understanding how ancient genetic networks were adapted to drive key reproductive innovations in land plant evolution. Despite these advances, discrepancies between expression methods (RNAseq, RT-PCR, and qPCR) highlight the importance of using complementary approaches to validate findings. Future studies should employ spatial expression analyses, such as *in situ* hybridization or promoter-reporter constructs, to confirm the tissue-specific localization of *CrSPL* genes (Jackson, 1991; Jefferson et al., 1987). These analyses will provide critical insights into the regulatory networks underpinning fern reproduction.

6.3 Chapter 4: Investigating the functional roles of *CrSPLs*

Building on the findings of Chapter 3, work in Chapter 4 sought to investigate the functional roles of *CrSPLA*, *CrSPLB*, and *CrSPLC* in *C. richardii* and assess whether these genes could complement the *spl* knockout in *Arabidopsis* (Schiefthaler et al., 1999; Yang et al., 1999). While these goals were not fully realized due to technical challenges, this work establishes a foundation for future studies. RNAi knockdowns were successfully generated for all three *CrSPL* genes, and plants progressed to T1 spores. However, due to time and space limitations, only one line, *CrSPLB-RNAi*, was carried through to the T2 generation, based on findings from Chapter 3 that suggested *CrSPLB* as the most likely candidate for an SPL-like function in reproduction. Preliminary analysis of T1 *CrSPLB-RNAi* plants did not yield obvious phenotypes, but potential candidate lines for further investigation in the T2 generation were identified. Complementation constructs were designed but could not be delivered into *Arabidopsis*, leaving this objective for future work.

6.3.1 Challenges in microparticle bombardment and transformation efficiency

Transformation efficiency using microparticle bombardment was significantly lower than previously reported, with only 10% efficiency compared to the 70% efficiency reported in earlier studies (Plackett et al., 2014). This low efficiency, combined with the labour-intensive nature of microparticle bombardment, suggests that future studies should explore alternative transformation methods. *Agrobacterium*-mediated transformation, which has shown promise in recent studies (Jiang et al., 2024; Withers et al., 2023), offers a more precise and less laborious approach and may improve reproducibility and reduce challenges associated with

multi-copy insertions.

6.3.2 Preliminary insights from *CrSPLB-RNAi* lines

Despite the lack of a strong phenotypic outcome, two *CrSPLB-RNAi* lines (3.3 and 5.2) showed slightly reduced *CrSPLB* expression and increases in sporangia number, with line 5.2 displaying a statistically significant increase. These preliminary results suggest that *CrSPLB* may play a role in regulating sporangia development, possibly as a negative regulator. This aligns with the broader role of SPLs in reproductive regulation observed in flowering plants (Schiefthaler et al., 1999; Yang et al., 1999). However, the incomplete knockdown typically achieved with RNAi may have left residual *CrSPLB* expression, masking more severe phenotypic effects. Employing CRISPR/Cas9 to achieve complete knockout could provide a more definitive test of *CrSPLB*'s function (Xiang & Li, 2024). Future work should also include detailed phenotyping of spore viability, meiosis, and gametophyte development to explore the potential role of *CrSPLB* in reproductive processes, coupled with spatial and temporal expression patterns of *CrSPLB* as discussed above.

6.3.3 Future directions for functional analysis of *CrSPL* genes

The work in chapter 4 provides a valuable starting point for functional studies of *CrSPL* genes in *C. richardii*. The preliminary findings from *CrSPLB-RNAi* suggest its potential role in sporangia development, while the challenges encountered highlight the need for methodological improvements. While the design and construction of complementation vectors represented a significant achievement, their inability to propagate in the pART27 vector prevented their use in *Arabidopsis* transformation. Future efforts should explore alternative vectors or cloning approaches, such as Gateway cloning, to successfully deliver these constructs. By refining transformation techniques, employing CRISPR-based knockouts, and advancing complementation assays, future research can build upon this work to elucidate the role of *CrSPL* genes

and their evolutionary conservation across land plants.

6.4 Chapter 5: Exploring transcriptional networks in reproductive development

While the RNAi experiments in Chapter 4 did not yield definitive phenotypes for the *CrSPL* homologs, they underscored the challenges of uncovering functional roles for these genes in *C. richardii*. This prompted a broader exploration of transcriptional networks underpinning reproductive development to identify new candidate regulators. By leveraging co-expression network analysis, Chapter 5 shifts focus from specific *CrSPL* homologs to a genome-wide approach, aiming to uncover novel genetic pathways associated with key developmental processes in *C. richardii*.

This chapter addressed two key gaps in the study of *C. richardii* transcriptomics: the limited validation of newly available genomic resources and the lack of comprehensive co-expression network analyses to investigate regulatory processes during fern development (Marchant et al., 2022). By comparing genome-guided and transcriptome-based RNA-seq mapping approaches, this study demonstrated the advantages of HISAT2 over Salmon, highlighting the incompleteness of the reference transcriptome (Kim et al., 2019; Marchant et al., 2022; Patro et al., 2017). Furthermore, the application of WGCNA revealed gene networks associated with key developmental processes, particularly those linked to reproductive transitions, providing insights into the transcriptional regulation of fern development (Langfelder & Horvath, 2008; Zhang & Horvath, 2005).

6.4.1 Evaluating the current transcriptomic resources

The recent publication of the high-quality *C. richardii* genome (v2.1) and accompanying transcriptome (v2.0) marked a significant advancement for fern genomics (Marchant et al., 2022). However, this study demonstrated that the transcriptome remains incomplete, as evidenced by the superior performance of

genome-guided mapping with HISAT2 compared to transcriptome-based mapping with Salmon (Kim et al., 2019; Patro et al., 2017). The higher alignment rates achieved by HISAT2 suggest that many transcripts are either missing or poorly represented in the reference transcriptome. This discrepancy likely arises from limited tissue sampling during the initial transcriptome assembly, which excluded transient or tissue-specific transcriptional programs (Marchant et al., 2022).

By incorporating additional tissues and developmental stages, including shoot apices and frond primordia across distinct time points, this study captured transcriptional profiles not represented in the original dataset. Future efforts to expand and refine the reference transcriptome should integrate existing datasets from diverse experimental conditions and tissue types (Aragón-Raygoza et al., 2022; Atallah et al., 2018; Bushart et al., 2013; Geng et al., 2021; Woudenberg et al., 2024; Xiao & Li, 2024). These additions would improve annotation accuracy, better reflect transcriptional diversity, and enhance the utility of transcriptome-based analyses for *C. richardii*.

6.4.2 Advancing Co-expression network analysis in ferns

The construction of a weighted gene co-expression network revealed 21 distinct modules, each representing a set of genes with correlated expression patterns across developmental stages and tissue types. While initial efforts focused on the placement of *CrSPL* homologs, these genes were assigned to modules with no clear correlation to reproductive tissues. For instance, *CrSPLA* and *CrSPLC* were placed in the red module, associated with early vegetative development, while *CrSPLB* was assigned to the cyan module, linked to plant age rather than reproductive development. Consequently, the focus shifted to modules such as ivory and darkorange, which were strongly correlated with reproductive tissues.

The ivory module provided insights into the transcriptional networks underpinning early reproductive transitions, highlighting processes such as cell wall remodeling,

lipid metabolism, and transcriptional regulation. In contrast, the darkorange module was associated with later stages of reproductive frond development, encompassing meiosis, sporangium maturation, and spore formation. These findings underscore the power of co-expression analysis to uncover novel gene networks and candidate regulators, offering a broader perspective than single-gene analyses.

6.4.3 Insights into reproductive development from the ivory and darkorange modules

The ivory module was strongly associated with fiddlehead and expanding fronds, suggesting a role in tissue differentiation and expansion during early reproductive transitions. The identification of transcription factors such as MYB, WRKY, and homeobox proteins as hub genes highlighted the regulatory complexity of this module (Bakshi & Oelmüller, 2014; Dubos et al., 2010; Elhiti & Stasolla, 2009; Ma et al., 2024). For instance, Ceric.08G001000, a homeobox TF linked to flavonoid biosynthesis, may regulate pigmentation processes associated with sporangium maturation (Kubo et al., 1999). GO enrichment analysis further revealed processes related to structural remodelling, transport, and lipid metabolism, reflecting the physiological demands of rapidly growing tissues (Grzyb et al., 2024; Quinn et al., 2023).

In contrast, the darkorange module was associated with expanding and mature reproductive fronds, emphasizing its role in later stages of reproductive development. Key processes identified included meiosis, DNA repair, and redox regulation. Hub genes such as *KNAT3-related* (Ceric.10G031700) and MYB80 (Ceric.09G016100) underscore the potential conservation of regulatory pathways between ferns and seed plants (Chen et al., 2023; Dubos et al., 2010). Notably, MYB80 is involved in tapetum layer formation in angiosperms, suggesting a conserved role in sporangium development (Dubos et al., 2010). Similarly, *KNAT3* essential for integument and ovule development in *Arabidopsis*, could play a parallel role in coordinating reproductive processes in *C. richardii* (Chen et al., 2023). The enrichment of GO terms related to

metal ion transport and redox homeostasis further supports the involvement of darkorange genes in regulating meiosis and maintaining cellular integrity during the energetically demanding process of spore formation. These findings underscore the importance of the darkorange module in orchestrating the complex molecular and cellular processes required for reproductive success in ferns.

6.4.4 A framework for future investigations

The findings in this chapter provide a valuable framework for understanding the transcriptional regulation of reproductive development in *C. richardii*. While the *CrSPL* homologs did not emerge as key players in these processes, the identification of modules like ivory and darkorange has uncovered novel candidate genes for future study. Hub genes such as Ceric.08G001000 (*HD-ZIP ANTHOCYANINLESS 2-RELATED*), Ceric.10G031700 (*KNAT3-related*), and Ceric.09G016100 (*MYB80*) represent promising targets for functional validation using tools like CRISPR-Cas9 or RNAi (Plackett et al., 2018; Xiang & Li, 2024). These genes offer insights into the conserved and divergent mechanisms underlying reproductive development in ferns and seed plants, bridging the gap between non-seed and seed plant evolution. By integrating transcriptomic data with experimental approaches, future research can build on the insights gained from this chapter to uncover the genetic networks driving reproductive transitions in ferns, further enhancing our understanding of land plant evolution.

6.5 Conclusion

This thesis has advanced our understanding of the evolutionary and functional roles of reproductive regulators in ferns, providing valuable insights into the conservation and divergence of genetic networks across land plants. By integrating phylogenetic, expression, and network analyses, this work has identified key candidates, such as *CrSPLB*, *MYB80* and *KNAT3-related*, that may bridge the evolutionary gap between non-seed and seed plants. While challenges were encountered in functional validation and transformation techniques, these limitations underscore the importance of refining methodologies and expanding genomic resources. The findings presented here lay the groundwork for future studies to explore the molecular mechanisms underlying reproductive development in ferns and their evolutionary links to seed plant systems. By leveraging emerging technologies such as CRISPR/Cas9 and improved transcriptomic resources, the field is well-positioned to uncover the ancient regulatory pathways that shaped the evolution of plant reproduction.

References

- Aagesen, L. (2004). The information content of an ambiguously alignable region, a case study of the trnI intron from the rhamnaceae. *Organisms Diversity & Evolution*, 4(1), 35–49. <https://doi.org/https://doi.org/10.1016/j.ode.2003.11.003>
- Agrawal, P. K., Kohli, A., Twyman, R. M., & Christou, P. (2005). Transformation of plants with multiple cassettes generates simple transgene integration patterns and high expression levels. *Molecular Breeding*, 16, 247–260.
- Aguiar, V. R. C., Castelli, E. C., Single, R. M., Bashirova, A., Ramsuran, V., Kulkarni, S., Augusto, D. G., Martin, M. P., Gutierrez-Arcelus, M., Carrington, M., & Meyer, D. (2023). Comparison between qpcr and rna-seq reveals challenges of quantifying hla expression [Epub 2023 Jan 28]. *Immunogenetics*, 75(3), 249–262. <https://doi.org/10.1007/s00251-023-01296-7>
- Anders, S., Pyl, P. T., & Huber, W. (2014). Htseq—a python framework to work with high-throughput sequencing data. *Bioinformatics*, 31(2), 166–169. <https://doi.org/10.1093/bioinformatics/btu638>
- Andrews, S. (2010). Fastqc: A quality control tool for high throughput sequence data.
- Aragón-Raygoza, A., Herrera-Estrella, L., & Cruz-Ramírez, A. (2022). Transcriptional analysis of *Ceratopteris richardii* young sporophyte reveals conservation of stem cell factors in the root apical meristem. *Frontiers in Plant Science*, 13. <https://doi.org/10.3389/fpls.2022.924660>
- Atallah, N. M., Vitek, O., Gaiti, F., Tanurdzic, M., & Banks, J. A. (2018). Sex determination in *Ceratopteris richardii* is accompanied by transcriptome changes that drive

- epigenetic reprogramming of the young gametophyte. *G3 Genes/Genomes/Genetics*, 8(7), 2205–2214. <https://doi.org/10.1534/g3.118.200292>
- Aya, K., Kobayashi, M., Tanaka, J., Ohyanagi, H., Suzuki, T., Yano, K., Takano, T., Yano, K., & Matsuoka, M. (2014). De Novo Transcriptome Assembly of a Fern, *Lygodium japonicum*, and a Web Resource Database, Ljtrans DB. *Plant and Cell Physiology*, 56(1), e5–e5. <https://doi.org/10.1093/pcp/pcu184>
- Bakshi, M., & Oelmüller, R. (2014). Wrky transcription factors: Jack of many trades in plants. *Plant Signal Behav*, 9(2), e27700. <https://doi.org/10.4161/psb.27700>
- Banks, J. A. (1999). Gametophyte development in ferns. *Annual Review of Plant Physiology and Plant Molecular Biology*, 50, 163–186. <https://doi.org/10.1146/annurev.arplant.50.1.163>
- Barabási, A. L., & Bonabeau, E. (2003). Scale-free networks. *Sci Am*, 288(5), 60–9. <https://doi.org/10.1038/scientificamerican0503-60>
- Barker, M. S., & Wolf, P. G. (2010). Unfurling fern biology in the genomics age. *BioScience*, 60(3), 177–185. <https://doi.org/10.1525/bio.2010.60.3.4>
- Baroux, C., Spillane, C., & Grossniklaus, U. (2002). Evolutionary origins of the endosperm in flowering plants. *Genome Biology*, 3(9), reviews1026.1. <https://doi.org/10.1186/gb-2002-3-9-reviews1026>
- Bateman, R. M., & DiMichele, W. A. (1994). Heterospory: The most iterative key innovation in the evolutionary history of the plant kingdom. *Biological Reviews*, 69(3), 345–417. <https://doi.org/https://doi.org/10.1111/j.1469-185X.1994.tb01276.x>
- Battaglia, R., Colombo, M., & Kater, M. M. (2018). The ins and outs of ovule development. In *Annual plant reviews online* (pp. 70–106). John Wiley & Sons, Ltd. <https://doi.org/https://doi.org/10.1002/9781119312994.apr0410>

- Beerling, D. J., & Franks, P. J. (2009). Evolution of stomatal function in plants: Key adaptations to the terrestrial environment. *New Phytologist*, 183(3), 631–647. <https://doi.org/10.1111/j.1469-8137.2009.02899.x>
- Bell, G. (1982). *The masterpiece of nature: The evolution and genetics of sexuality*. Routledge.
- Bennici, A. (2008). Origin and early evolution of land plants: Problems and considerations. *Communicative & Integrative Biology*, 1(2), 212–218. <https://doi.org/10.4161/cib.1.2.6987>
- Berger, F., Hamamura, Y., Ingouff, M., & Higashiyama, T. (2008). Double fertilization—caught in the act. *Trends in plant science*, 13(8), 437–443.
- Bergfeld, R., & Schopfer, P. (1986). Differentiation of a functional aleurone layer within the seed coat of *sinapis alba* L. *Annals of botany*, 57(1), 25–33.
- Bernstein, E., Caudy, A. A., Hammond, S. M., & Hannon, G. J. (2001). Role for a bidentate ribonuclease in the initiation step of rna interference. *Nature*, 409(6818), 363–366. <https://doi.org/10.1038/35053110>
- Berr, A., McCallum, E. J., Menard, R., Meyer, D., Fuchs, J., Dong, A. W., & Shen, W. H. (2010). Arabidopsis set domain group2 is required for h3k4 trimethylation and is crucial for both sporophyte and gametophyte development. *Plant Cell*, 22(10), 3232–3248. <https://doi.org/10.1105/tpc.110.079962>
- Bezodis, W., Prescott, H., & Dickinson, H. (2022). Misexpression of the homologues of bryophyte gametophyte-to-sporophyte control genes in arabidopsis results in germline reprogramming and phenotypes that mirror apomictic development. *bioRxiv*. <https://doi.org/10.1101/2022.09.15.508103>
- Blighe, K. (2018). Corlevelplot: Visualise correlation results and test significancies of these [Accessed: 2025-01-05].
- Borg, M., Rutley, N., Kagale, S., Hamamura, Y., Gherghinoiu, M., Kumar, S., Sari, U., Esparza-Franco, M. A., Sakamoto, W., Rozwadowski, K., et al. (2014). An

- ear-dependent regulatory module promotes male germ cell division and sperm fertility in arabidopsis. *The Plant Cell*, 26(5), 2098–2113.
- Borrill, P. (2020). Blurring the boundaries between cereal crops and model plants. *New Phytologist*, 228(6), 1721–1727. <https://doi.org/https://doi.org/10.1111/nph.16229>
- Bower, F. O. (1908). *The origin of a land flora*. Macmillan; Company, limited.
- Bowman, J. L. (2022). The origin of a land flora. *Nature Plants*, 8, 1352–1369. <https://doi.org/10.1038/s41477-022-01283-y>
- Bowman, J. L., Kohchi, T., Yamato, K. T., Jenkins, J., Shu, S., Ishizaki, K., Yamaoka, S., Nishihama, R., Nakamura, Y., Berger, F., et al. (2017). Insights into land plant evolution garnered from the marchantia polymorpha genome. *Cell*, 171(2), 287–304.
- Bowman, J. L., Smyth, D. R., & Meyerowitz, E. M. (1991). Genetic interactions among floral homeotic genes of arabidopsis. *Development*, 112(1), 1–20.
- Brambilla, V., Battaglia, R., Colombo, M., Masiero, S., Bencivenga, S., Kater, M. M., & Colombo, L. (2007). Genetic and molecular interactions between bell1 and mads box factors support ovule development in arabidopsis. *Plant Cell*, 19(8), 2544–2556. <https://doi.org/10.1105/tpc.107.051797>
- Brown, R. C., & Lemmon, B. E. (2011). Spores before sporophytes: Hypothesizing the origin of sporogenesis at the algal–plant transition. *New Phytologist*, 190(4), 875–881. <https://doi.org/https://doi.org/10.1111/j.1469-8137.2011.03709.x>
- Browning, B., Jordan, G., Dalton, P., Grove, S., Wardlaw, T., & Turner, P. (2010). Succession of mosses, liverworts and ferns on coarse woody debris, in relation to forest age and log decay in tasmanian wet eucalypt forest. *Forest Ecology and Management*, 260(10), 1896–1905. <https://doi.org/https://doi.org/10.1016/j.foreco.2010.08.038>

- Bui, L. T., Cordle, A. R., Irish, E. E., et al. (2015). Transient and stable transformation of *Ceratopteris richardii* gametophytes. *BMC Research Notes*, 8, 214. <https://doi.org/10.1186/s13104-015-1193-x>
- Bui, L. T., Pandzic, D., Youngstrom, C. E., Wallace, S., Irish, E. E., Szövényi, P., & Cheng, C.-L. (2017). A fern gene mirrors in promoting apogamy in *ceratopteris richardii*. *The Plant Journal*, 90(1), 122–132. <https://doi.org/https://doi.org/10.1111/tpj.13479>
- Bushart, T. J., Cannon, A. E., ul Haque, A., San Miguel, P., Mostajeran, K., Clark, G. B., Porterfield, D. M., & Roux, S. J. (2013). Rna-seq analysis identifies potential modulators of gravity response in spores of *Ceratopteris* (parkeriaceae): Evidence for modulation by calcium pumps and apyrase activity. *American Journal of Botany*, 100, 161–174. <https://doi.org/10.3732/ajb.1200292>
- Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., & Madden, T. L. (2009). BLAST+: Architecture and applications. *BMC Bioinformatics*, 10, 421. <https://doi.org/10.1186/1471-2105-10-421>
- Cantino, P. D., Doyle, J. A., Graham, S. W., Judd, W. S., Olmstead, R. G., Soltis, D. E., Soltis, P. S., & Donoghue, M. J. (2007). Towards a phylogenetic nomenclature of tracheophyta. *Taxon*, 56(3), E1–E44.
- Carpentier, M., & Chomilier, J. (2019). Protein multiple alignments: Sequence-based versus structure-based programs. *Bioinformatics*, 35(20), 3970–3980. <https://doi.org/10.1093/bioinformatics/btz236>
- Chan, R. L., Gago, G. M., Palena, C. M., & Gonzalez, D. H. (1998). Homeoboxes in plant development. *Biochimica et Biophysica Acta (BBA) - Gene Structure and Expression*, 1442(1), 1–19. [https://doi.org/https://doi.org/10.1016/S0167-4781\(98\)00119-5](https://doi.org/https://doi.org/10.1016/S0167-4781(98)00119-5)
- Chang, Y., & Graham, S. W. (2011). Inferring the higher-order phylogeny of mosses (bryophyta) and relatives using a large, multigene plastid data set. *American Journal of Botany*, 98(5), 839–849. <https://doi.org/https://doi.org/10.3732/ajb.0900384>

- Chen, G. H., Sun, J. Y., Liu, M., Liu, J., & Yang, W. C. (2014). Sporocyteless is a novel embryophyte-specific transcription repressor that interacts with tpl and tcp proteins in arabidopsis. *Journal of Genetics and Genomics*, 41(12), 617–625. <https://doi.org/10.1016/j.jgg.2014.08.009>
- Chen, J. J., Wang, W., Qin, W. Q., Men, S. Z., Li, H. L., Mitsuda, N., Ohme-Takagi, M., & Wu, A. M. (2023). Transcription factors knat3 and knat4 are essential for integument and ovule formation in arabidopsis. *Plant Physiol*, 191(1), 463–478. <https://doi.org/10.1093/plphys/kiac513>
- Chen, S., Zhou, Y., Chen, Y., & Gu, J. (2018). Fastp: An ultra-fast all-in-one fastq preprocessor. *Bioinformatics*, 34(17), i884–i890. <https://doi.org/10.1093/bioinformatics/bty560>
- Chitwood, D. H., & Sinha, N. R. (2016). Evolutionary and environmental forces sculpting leaf development. *Current Biology*, 26(7), R297–R306. <https://doi.org/10.1016/j.cub.2016.02.033>
- Christenhusz, M. J., & Byng, J. W. (2016). The number of known plants species in the world and its annual increase. *Phytotaxa*, 261(3), 201–217.
- Christensen, C. A., King, E. J., Jordan, J. R., & Drews, G. N. (1997). Megagametogenesis in arabidopsis wild type and the gf mutant. *Sexual Plant Reproduction*, 10(1), 49–64. <https://doi.org/10.1007/s004970050067>
- Chuang, C.-F., & Meyerowitz, E. M. (2000). Specific and heritable genetic interference by double-stranded rna in *Arabidopsis thaliana*. *Proceedings of the National Academy of Sciences USA*, 97, 4985–4990. <https://doi.org/10.1073/pnas.97.9.4985>
- Clarke, J. T., Warnock, R. C., & Donoghue, P. C. (2011). Establishing a time-scale for plant evolution. *New Phytologist*, 192(1), 266–301.
- Coen, E. S., & Meyerowitz, E. M. (1991). The war of the whorls: Genetic interactions controlling flower development. *Nature*, 353(6339), 31–37.

- Conant, G. C., & Wolfe, K. H. (2008). Turning a hobby into a job: How duplicated genes find new functions [Epub 2008 Oct 28]. *Nature Reviews Genetics*, 9, 938–950. <https://doi.org/10.1038/nrg2482>
- Conesa, A., Madrigal, P., Tarazona, S., et al. (2016). A survey of best practices for rna-seq data analysis. *Genome Biology*, 17, 13. <https://doi.org/10.1186/s13059-016-0881-8>
- Cong, M., Zhu, T., Li, Y., Yang, W., & Wei, Y. (2023). Ancient ecological disaster site is now a refuge: Bryophyte diversity in volcanic lava caves of jingpo lake world geopark. *Diversity*, 15(7). <https://doi.org/10.3390/d15070842>
- Conway, S. J., & Di Stilio, V. S. (2020). An ontogenetic framework for functional studies in the model fern *ceratopteris richardii*. *Developmental Biology*, 457(1), 20–29. <https://doi.org/https://doi.org/10.1016/j.ydbio.2019.08.017>
- Cox, C. J., Li, B., Foster, P. G., Embley, T. M., & Civián, P. (2014). Conflicting Phylogenies for Early Land Plants are Caused by Composition Biases among Synonymous Substitutions. *Systematic Biology*, 63(2), 272–279. <https://doi.org/10.1093/sysbio/syt109>
- Crane, P. R., Friis, E. M., & Pedersen, K. R. (1995). The origin and early diversification of angiosperms. *Nature*, 374(6517), 27–33.
- Crepet, W., & Friis, E. (1987). The evolution of insect pollination in angiosperms. In E. Friis, W. Chaloner, & P. Crane (Eds.), *The origin of angiosperms and their biological consequences* (pp. 181–201). Cambridge University Press.
- Danecek, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., Whitwham, A., Keane, T., McCarthy, S. A., Davies, R. M., & Li, H. (2021). Twelve years of SAMtools and BCFtools [giab008]. *GigaScience*, 10(2). <https://doi.org/10.1093/gigascience/giab008>
- Darriba, D., Posada, D., Kozlov, A. M., Stamatakis, A., Morel, B., & Flouri, T. (2020). Modeltest-ng: A new and scalable tool for the selection of dna and protein evolutionary models. *Molecular Biology and Evolution*, 37(1), 291–294. <https://doi.org/10.1093/molbev/msz189>

- Darwin, C. (1879). *Letter to j.d. hooker, july 22, 1879* (F. Burkhardt & S. Smith, Eds.; Vol. 27). Cambridge University Press.
- de Sousa, F., Foster, P. G., Donoghue, P. C. J., Schneider, H., & Cox, C. J. (2019). Nuclear protein phylogenies support the monophyly of the three bryophyte groups (bryophyta schimp.) *New Phytologist*, 222(1), 565–575. <https://doi.org/https://doi.org/10.1111/nph.15587>
- De La Torre, A. R., Piot, A., Liu, B., Wilhite, B., Weiss, M., & Porth, I. (2019). Functional and morphological evolution in gymnosperms: A portrait of implicated gene families. *Evolutionary Applications*, 13(1), 210–227. <https://doi.org/10.1111/eva.12839>
- De Vries, J., & Archibald, J. M. (2018). Plant evolution: Landmarks on the path to terrestrial life. *New Phytologist*, 217(4), 1428–1434.
- Debeaujon, I., Léon-Kloosterziel, K. M., & Koornneef, M. (2000). Influence of the testa on seed dormancy, germination, and longevity in arabidopsis [Erratum in: Plant Physiol. 2001 Feb;125(2):1139-41]. *Plant Physiology*, 122(2), 403–414. <https://doi.org/10.1104/pp.122.2.403>
- Delaux, P.-M., Hetherington, A., Coudert, Y., Delwiche, C., Dunand, C., et al. (2019). Reconstructing trait evolution in plant evo–devo studies [hal-02355085]. *Current Biology*, 29(21), R1110–R1118. <https://doi.org/10.1016/j.cub.2019.09.044>
- Delwiche, C. F., & Cooper, E. D. (2015). The evolutionary origin of a terrestrial flora. *Current Biology*, 25(19), R899–R910.
- Dessimoz, C., & Gil, M. (2010). Phylogenetic assessment of alignments reveals neglected tree signal in gaps. *Genome Biology*, 11, R37. <https://doi.org/10.1186/gb-2010-11-4-r37>
- Dingwall, C., & Laskey, R. A. (1991). Nuclear targeting sequences – a consensus? *Trends in Biochemical Sciences*, 16, 478–481. [https://doi.org/https://doi.org/10.1016/0968-0004\(91\)90184-W](https://doi.org/https://doi.org/10.1016/0968-0004(91)90184-W)
- Dipp-Álvarez, M., & Cruz-Ramírez, A. (2019). A phylogenetic study of the ant family points to a preant gene as the ancestor of basal and euant transcription factors

- in land plants. *Frontiers in Plant Science*, 10.
<https://doi.org/10.3389/fpls.2019.00017>
- Donoghue, P., Harrison, C. J., Paps, J., & Schneider, H. (2021). The evolutionary emergence of land plants. *Current Biology*, 31(19), R1281–R1298.
<https://doi.org/https://doi.org/10.1016/j.cub.2021.07.038>
- Donoghue, P., & Paps, J. (2020). Plant evolution: Assembling land plants. *Current Biology*, 30(2), R81–R83.
- Doyle, J. A. (1977). Patterns of evolution in early angiosperms. In *Developments in palaeontology and stratigraphy* (pp. 501–546, Vol. 5). Elsevier.
- Doyle, J. A. (1998). Phylogeny of vascular plants. *Annual Review of Ecology, Evolution, and Systematics*, 29(Volume 29, 1998), 567–599. <https://doi.org/https://doi.org/10.1146/annurev.ecolsys.29.1.567>
- Doyle, J. A. (2006). Seed ferns and the origin of angiosperms. *The Journal of the Torrey Botanical Society*, 133(1), 169–209.
- Dubos, C., Stracke, R., Grotewold, E., Weisshaar, B., Martin, C., & Lepiniec, L. (2010). Myb transcription factors in *arabidopsis*. *Trends in Plant Science*, 15(10), 573–581. <https://doi.org/10.1016/j.tplants.2010.06.005>
- Edwards, D., & Salter, R. J. (1982). Sporopollenin: Its structure and evolution. *Annals of Botany*, 49(3), 393–400. <https://doi.org/10.1093/oxfordjournals.aob.a087885>
- Eisen, J. A., & Wu, M. (2002). Phylogenetic analysis and gene functional predictions: Phylogenomics in action. *Theoretical Population Biology*, 61(4), 481–487. <https://doi.org/https://doi.org/10.1006/tpbi.2002.1594>
- Elbashir, S. M., Lendeckel, W., & Tuschl, T. (2001). Rna interference is mediated by 21- and 22-nucleotide rnas. *Genes & Development*, 15(2), 188–200. <https://doi.org/10.1101/gad.862301>
- Elhiti, M., & Stasolla, C. (2009). Structure and function of homodomain-leucine zipper (hd-zip) proteins. *Plant Signal Behav*, 4(2), 86–88. <https://doi.org/10.4161/psb.4.2.7692>

- Emms, D. M., & Kelly, S. (2019). Orthofinder: Phylogenetic orthology inference for comparative genomics. *Genome Biology*, 20, 238. <https://doi.org/10.1186/s13059-019-1832-y>
- Ewels, P., Magnusson, M., Lundin, S., & Käller, M. (2016). Multiqc: Summarize analysis results for multiple tools and samples in a single report. *Bioinformatics*, 32(19), 3047–3048. <https://doi.org/10.1093/bioinformatics/btw354>
- Ewing, E. P. (1977). Selection at the haploid and diploid phases: Cyclical variation. *Genetics*, 87(1), 195–207.
- Ezquer, I., Mizzotti, C., Nguema-Ona, E., Gotte, M., Beauzamy, L., Viana, V. E., Dubrulle, N., de Oliveira, A. C., Caporali, E., Koroney, A. S., Boudaoud, A., Driouich, A., & Colombo, L. (2016). The developmental regulator seedstick controls structural and mechanical properties of the arabidopsis seed coat. *Plant Cell*, 28(10), 2478–2492. <https://doi.org/10.1105/tpc.16.00454>
- Favaro, R., Pinyopich, A., Battaglia, R., Kooiker, M., Borghi, L., Ditta, G., Yanofsky, M. F., Kater, M. M., & Colombo, L. (2003). Mads-box protein complexes control carpel and ovule development in arabidopsis. *Plant Cell*, 15(11), 2603–2611. <https://doi.org/10.1105/tpc.015123>
- Fedoroff, N. V., Battisti, D. S., Beachy, R. N., Cooper, P. J. M., Fischhoff, D. A., Hodges, C. N., Knauf, V. C., Lobell, D., Mazur, B. J., Molden, D., Reynolds, M. P., Ronald, P. C., Rosegrant, M. W., Sanchez, P. A., Vonshak, A., & Zhu, J.-K. (2010). Radically rethinking agriculture for the 21st century. *Science*, 327(5967), 833–834. <https://doi.org/10.1126/science.1186834>
- Ferrari, C., Shivhare, D., Hansen, B. O., Pasha, A., Esteban, E., Provart, N. J., Kragler, F., Fernie, A., Tohge, T., & Mutwil, M. (2020). Expression atlas of *Selaginella moellendorffii* provides insights into the evolution of vasculature, secondary metabolism, and roots. *The Plant Cell*, 32(4), 853–870. <https://doi.org/10.1105/tpc.19.00780>

- Finch-Savage, W. E., & Leubner-Metzger, G. (2006). Seed dormancy and the control of germination. *New phytologist*, 171(3), 501–523.
- Finotello, F., & Di Camillo, B. (2015). Measuring differential gene expression with rna-seq: Challenges and strategies for data analysis. *Briefings in functional genomics*, 14(2), 130–142.
- Flouri, T., Izquierdo-Carrasco, F., Darriba, D., Aberer, A. J., Nguyen, L.-T., Minh, B. Q., von Haeseler, A., & Stamatakis, A. (2014). The phylogenetic likelihood library. *Systematic Biology*, 64(2), 356–362. <https://doi.org/10.1093/sysbio/syu084>
- Floyd, S. K., & Friedman, W. E. (2000). Evolution of endosperm developmental patterns among basal flowering plants. *International Journal of Plant Sciences*, 161(S6), S57–S81.
- Food and Agriculture Organization of the United Nations. (2023). Faostat: Food and agriculture data [Accessed: 2024-08-19]. <http://www.fao.org/faostat/en/>
- Fouracre, J. P., & Harrison, C. J. (2022). How was apical growth regulated in the ancestral land plant? insights from the development of non-seed plants. *Plant Physiology*, 190(1), 100–112. <https://doi.org/10.1093/plphys/kiac313>
- Friedman, W. E., & Ryerson, K. C. (2009). Reconstructing the ancestral female gametophyte of angiosperms: Insights from amborella and other ancient lineages of flowering plants. *American Journal of Botany*, 96(1), 129–143.
- Friedman, W. E., & Williams, J. H. (2004). Developmental evolution of the sexual process in ancient flowering plant lineages. *The Plant Cell*, 16(suppl_1), S119–S132.
- Friis, E. M., Pedersen, K. R., & Crane, P. R. (2006). Cretaceous angiosperm flowers: Innovation and evolution in plant reproduction. *Palaeogeography, palaeoclimatology, palaeoecology*, 232(2-4), 251–293.
- Frohlich, M. W., & Chase, M. W. (2007). After a dozen years of progress the origin of angiosperms is still a great mystery. *Nature*, 450(7173), 1184–1189.
- Fusaro, A. F., Matthew, L., Smith, N. A., Curtin, S. J., Dedic-Hagan, J., Ellacott, G. A., Watson, J. M., Wang, M.-B., Brosnan, C., Carroll, B. J., & Waterhouse, P. M.

- (2006). Rna interference-inducing hairpin rnas in plants act through the viral defence pathway. *EMBO Reports*, 7(11), 1168–1175. <https://doi.org/10.1038/sj.embor.7400837>
- Gao, C., & Nielsen, K. K. (2013). Comparison between agrobacterium-mediated and direct gene transfer using the gene gun. In S. Sudowe & A. B. Reske-Kunz (Eds.), *Biolistic dna delivery: Methods and protocols* (pp. 3–16). Humana Press. https://doi.org/10.1007/978-1-62703-110-3_1
- Gao, L., SU, Y.-J., & Wang, T. (2010). Plastid genome sequencing, comparative genomics, and phylogenomics: Current status and prospects. *Journal of Systematics and Evolution*, 48(2), 77–93.
- Garber, M., Grabherr, M. G., Guttman, M., & Trapnell, C. (2011). Computational methods for transcriptome annotation and quantification using rna-seq. *Nature Methods*, 8, 469–477. <https://doi.org/10.1038/nmeth.1613>
- Geng, Y., Cai, C., McAdam, S. A., Banks, J. A., Wisecaver, J. H., & Zhou, Y. (2021). A de novo transcriptome assembly of *Ceratopteris richardii* provides insights into the evolutionary dynamics of complex gene families in land plants. *Genome Biology and Evolution*, 13(3), evab042. <https://doi.org/10.1093/gbe/evab042>
- Geng, Y., Yan, A., & Zhou, Y. (2022). Positional cues and cell division dynamics drive meristem development and archegonium formation in *Ceratopteris* gametophytes. *Communications Biology*, 5, 650. <https://doi.org/10.1038/s42003-022-03627-y>
- Gerrienne, P., Dilcher, D. L., Bergamaschi, S., Milagres, I., Pereira, E., & Rodrigues, M. A. C. (2006). An exceptional specimen of the early land plant *Cooksonia paranensis*, and a hypothesis on the life cycle of the earliest eutracheophytes. *Review of Palaeobotany and Palynology*, 142(3-4), 123–130.
- Godfray, H. C. J., Beddington, J. R., Crute, I. R., Haddad, L., Lawrence, D., Muir, J. F., Pretty, J., Robinson, S., Thomas, S. M., & Toulmin, C. (2010). Food security: The

- challenge of feeding 9 billion people. *Science*, 327(5967), 812–818. <https://doi.org/10.1126/science.1185383>
- Goodstein, D. M., Shu, S., Howson, R., Neupane, R., Hayes, R. D., Fazo, J., Mitros, T., Dirks, W., Hellsten, U., Putnam, N., & Rokhsar, D. S. (2012). Phytozome: A comparative platform for green plant genomics. *Nucleic Acids Research*, 40(D1), D1178–D1186. <https://doi.org/10.1093/nar/gkr944>
- Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., Chen, Z., Mauceli, E., Hacohen, N., Gnirke, A., Rhind, N., di Palma, F., Birren, B. W., Nusbaum, C., Lindblad-Toh, K., ... Regev, A. (2011). Full-length transcriptome assembly from rna-seq data without a reference genome. *Nature Biotechnology*, 29(7), 644–652. <https://doi.org/10.1038/nbt.1883>
- Graham, L. E. (1993). *Origin of land plants*. Wiley.
- Graham, L. K., & Wilcox, L. W. (2000). The origin of alternation of generations in land plants: A focus on matrotrophy and hexose transport. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 355(1398), 757–767.
- Graves, S., Piepho, H.-P., & with help from Sundar Dorai-Raj, L. S. (2024). *Multcompview: Visualizations of paired comparisons* [R package version 0.1-10]. <https://github.com/ltselzer/multcompview>
- Grzyb, M., Sujkowska-Rybkowska, M., & Mikuła, A. (2024). Cell wall remodeling and callose deposition during the embryogenic transition in the tree fern *Cyathea delgadii* Sternb. *Plant Cell, Tissue and Organ Culture*, 156, 30. <https://doi.org/10.1007/s11240-023-02654-9>
- Guan, R., Zhao, Y., Zhang, H., Fan, G., Liu, X., Zhou, W., Shi, C., Wang, J., Liu, W., Liang, X., Fu, Y., Ma, K., Zhao, L., Zhang, F., Lu, Z., Lee, S. M.-Y., Xu, X., Wang, J., Yang, H., ... Chen, W. (2016). Draft genome of the living fossil *Ginkgo biloba*. *GigaScience*, 5(1), s13742-016-0154–1. <https://doi.org/10.1186/s13742-016-0154-1>

- Haas, B. J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P. D., Bowden, J., Couger, M. B., Eccles, D., Li, B., Lieber, M., Macmanes, M. D., Ott, M., Orvis, J., Pochet, N., Strozzi, F., Weeks, N., Westerman, R., William, T., Dewey, C. N., ... Regev, A. (2013). De novo transcript sequence reconstruction from rna-seq using the trinity platform for reference generation and analysis. *Nature Protocols*, 8(8), 1494–1512. <https://doi.org/10.1038/nprot.2013.084>
- Hackenberg, D., & Twell, D. (2019). Chapter eleven - the evolution and patterning of male gametophyte development. In U. Grossniklaus (Ed.), *Plant development and evolution* (pp. 257–298, Vol. 131). Academic Press. <https://doi.org/https://doi.org/10.1016/bs.ctdb.2018.10.008>
- Hake, S., Smith, H. M., Holtan, H., Magnani, E., Mele, G., & Ramirez, J. (2004). The role of knox genes in plant development. *Annu Rev Cell Dev Biol*, 20, 125–51. <https://doi.org/10.1146/annurev.cellbio.20.031803.093824>
- Hannon, G. J. (2002). Rna interference. *Nature*, 418, 244–251. <https://doi.org/10.1038/418244a>
- Hao, S., Ariizumi, T., & Ezura, H. (2017). Sexual sterility is essential for both male and female gametogenesis in tomato. *Plant and Cell Physiology*, 58(1), 22–34. <https://doi.org/10.1093/pcp/pcw214>
- Hao, Y., Zong, X., Ren, P., Qian, Y., & Fu, A. (2021). Basic helix-loop-helix (bhlh) transcription factors regulate a wide range of functions in arabidopsis. *International Journal of Molecular Sciences*, 22(13). <https://doi.org/10.3390/ijms22137152>
- Harris, B., Harrison, C., Hetherington, A., & Williams, T. (2020). Phylogenomic evidence for the monophyly of bryophytes and the reductive evolution of stomata. *Current Biology*, 30(11), 2001–2012.e2. <https://doi.org/10.1016/j.cub.2020.03.048>
- Harris, L. W., & Davies, J. T. (2016). A complete fossil-calibrated phylogeny of seed plant families as a tool for comparative analyses: Testing the ‘time for

- speciation' hypothesis. *PLOS ONE*, 11(10), 1–16.
<https://doi.org/10.1371/journal.pone.0162907>
- Harrison, C. J. (2017). Development and genetics in the evolution of land plant body plans. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 372(1713), 20150490. <https://doi.org/10.1098/rstb.2015.0490>
- Harrison, C. J., & Morris, J. L. (2018). The origin and early evolution of vascular plant shoots and leaves. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 373(1739), 20160496. <https://doi.org/10.1098/rstb.2016.0496>
- Haughn, G. W., & Somerville, C. R. (1988). Genetic control of morphogenesis in arabidopsis. *Developmental Genetics*, 9(2), 73–89.
- He, K., Cao, X., & Deng, X. (2021). Histone methylation in epigenetic regulation and temperature responses [Epigenetics]. *Current Opinion in Plant Biology*, 61, 102001. <https://doi.org/https://doi.org/10.1016/j.pbi.2021.102001>
- Heath, T. A., Hedtke, S. M., & Hillis, D. M. (2008). Taxon sampling and the accuracy of phylogenetic analyses. *Journal of systematics and evolution*, 46(3), 239.
- Hedtke, S. M., Townsend, T. M., & Hillis, D. M. (2006). Resolution of phylogenetic conflict in large data sets by increased taxon sampling. *Systematic Biology*, 55(3), 522–529.
- Hemsley, A. R. (1994). The origin of the land plant sporophyte: An interpolational scenario. *Biological Reviews*, 69(3), 263–273.
- Henry, T. A., Bainard, J. D., & Newmaster, S. G. (2014). Genome size evolution in ontario ferns (polypodiidae): Evolutionary correlations with cell size, spore size, and habitat type and an absence of genome downsizing. *Genome*, 57(10), 555–566.
<https://doi.org/10.1139/gen-2014-0090>
- Hetherington, A. J. (2021). New views on old seeds: A new description of genomosperma sheds light on early seed evolution. *New Phytologist*, 229(3), 1189–1191. <https://doi.org/https://doi.org/10.1111/nph.16875>

- Hickok, L. G., Warne, T. R., & Fribourg, R. S. (1995). The biology of the fern *Ceratopteris* and its use as a model system. *International Journal of Plant Sciences*, 156(3), 332–345.
- Hickok, L. G., Warne, T. R., & Slocum, M. K. (1987). *Ceratopteris richardii*: Applications for experimental plant biology. *American Journal of Botany*, 74(8), 1304–1316.
<https://doi.org/https://doi.org/10.1002/j.1537-2197.1987.tb08743.x>
- Hillis, D. M., Pollock, D. D., McGuire, J. A., & Zwickl, D. J. (2003). Is sparse taxon sampling a problem for phylogenetic inference? *Systematic biology*, 52(1), 124.
- Hiss, M., Meyberg, R., Westermann, J., Haas, F. B., Schneider, L., Schallenberg-Rüdinger, M., Ullrich, K. K., & 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), 606–620.
- Hofmeister, W. (1851). *Vergleichende untersuchungen der keimung, entfaltung und fruchtbildung höherer kryptogamen:(moose, farnn, equisetaceen, rhizocarpeen und lycopodiaceen) und der samenbildung der coniferen*. F. Hofmeister.
- Hofmeister, W. (1862). *On the germination, development, and fructification of the higher cryptogamia, and on the fructification of the coniferae*. Ray Society.
- Hong, M. (2005). Molecular genetic analyses of microsporogenesis and megasporogenesis in flowering plants. *Annual Review of Plant Biology*, 56, 393–434. <https://doi.org/10.1146/annurev.arplant.55.031903.141717>
- Honma, T., & Goto, K. (2001). Complexes of MADS-box proteins are sufficient to convert leaves into floral organs. *Nature*, 409(6819), 525–529.
- Huang, H., Mizukami, Y., Hu, Y., & Ma, H. (1993). Isolation and characterization of the binding sequences for the product of the Arabidopsis floral homeotic gene *AGAMOUS*. *Nucleic Acids Research*, 21(20), 4769–4776.
<https://doi.org/10.1093/nar/21.20.4769>

- Hunter, N. (2015). Meiotic recombination: The essence of heredity. *Cold Spring Harb Perspect Biol*, 7(12). <https://doi.org/10.1101/cshperspect.a016618>
- Ismagul, A., Yang, N., Maltseva, E., et al. (2018). A biolistic method for high-throughput production of transgenic wheat plants with single gene insertions. *BMC Plant Biology*, 18, 135. <https://doi.org/10.1186/s12870-018-1326-1>
- Ito, T., Wellmer, F., Yu, H., Das, P., Ito, N., Alves-Ferreira, M., Riechmann, J. L., & Meyerowitz, E. M. (2004). The homeotic protein *agamous* controls microsporogenesis by regulation of sporocyteless. *Nature*, 430(6997), 356–360. <https://doi.org/10.1038/nature02733>
- Jackson, D. (1991). In situ hybridization in plants [Chapter contribution]. In D. J. Bowles, S. J. Gurr, & M. McPherson (Eds.), *Molecular plant pathology: A practical approach*. Oxford University Press.
- Jalali, M., Zaborowska, J., & Jalali, M. (2017). Chapter 1 - the polymerase chain reaction: Pcr, qpcr, and rt-pcr. In M. Jalali, F. Y. Saldanha, & M. Jalali (Eds.), *Basic science methods for clinical researchers* (pp. 1–18). Academic Press. <https://doi.org/10.1016/B978-0-12-803077-6.00001-1>
- Jefferson, R. A., Kavanagh, T. A., & Bevan, M. W. (1987). Gus fusions: Beta-glucuronidase as a sensitive and versatile gene fusion marker in higher plants [Published 1987 Dec 20]. *The EMBO Journal*, 6(13), 3901–3907. <https://doi.org/10.1002/j.1460-2075.1987.tb02730.x>
- Jiang, W., Deng, F., Babla, M., Chen, C., Yang, D., Tong, T., Qin, Y., Chen, G., Marchant, B., Soltis, P., Soltis, D. E., Zeng, F., & Chen, Z.-H. (2024). Efficient gene editing of a model fern species through gametophyte-based transformation. *Plant Physiology*, 196(4), 2346–2361. <https://doi.org/10.1093/plphys/kiad473>
- Johansen, L. K., & Carrington, J. C. (2001). Silencing on the spot: Induction and suppression of rna silencing in the *Agrobacterium*-mediated transient expression system. *Plant Physiology*, 126, 930–938. <https://doi.org/10.1104/pp.126.3.930>

- Johnson, G. P., & Renzaglia, K. S. (2008). Embryology of ceratopteris richardii (pteridaceae, tribe ceratopterideae), with emphasis on placental development. *Journal of Plant Research*, 121, 581–592.
- Kagale, S., Links, M. G., & Rozwadowski, K. (2010). Genome-wide analysis of ethylene-responsive element binding factor-associated amphiphilic repression motif-containing transcriptional regulators in arabidopsis. *Plant Physiol*, 152(3), 1109–34. <https://doi.org/10.1104/pp.109.151704>
- Kagale, S., & Rozwadowski, K. (2011). Ear motif-mediated transcriptional repression in plants: An underlying mechanism for epigenetic regulation of gene expression. *Epigenetics*, 6(2), 141–6. <https://doi.org/10.4161/epi.6.2.13627>
- Karol, K. G., McCourt, R. M., Cimino, M. T., & Delwiche, C. F. (2001). The closest living relatives of land plants. *Science*, 294(5550), 2351–2353.
- Karol, K. G., Arumuganathan, K., Boore, J. L., Duffy, A. M., Everett, K. D., Hall, J. D., Hansen, S. K., Kuehl, J. V., Mandoli, D. F., Mishler, B. D., Olmstead, R. G., Renzaglia, K. S., & Wolf, P. G. (2010). Complete plastome sequences of equisetum arvense and isoetes flaccida: Implications for phylogeny and plastid genome evolution of early land plant lineages. *BMC Evolutionary Biology*, 10, 321. <https://doi.org/10.1186/1471-2148-10-321>
- Kato, M., & Akiyama, H. (2005). Interpolation hypothesis for origin of the vegetative sporophyte of land plants. *TAXON*, 54(2), 443–450. <https://doi.org/https://doi.org/10.2307/25065371>
- Katoh, K., Kuma, K.-i., Toh, H., & Miyata, T. (2005). MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucleic Acids Research*, 33(2), 511–518. <https://doi.org/10.1093/nar/gki198>
- Kemena, C., & Notredame, C. (2009). Upcoming challenges for multiple sequence alignment methods in the high-throughput era. *Bioinformatics*, 25(19), 2455–2465. <https://doi.org/10.1093/bioinformatics/btp452>

- Kenrick, P. (1994). Alternation of generations in land plants: New phylogenetic and palaeobotanical evidence. *Biological Reviews*, 69(3), 293–330. <https://doi.org/https://doi.org/10.1111/j.1469-185X.1994.tb01273.x>
- Kenrick, P., & Crane, P. R. (1997). *The origin and early diversification of land plants: A cladistic study*. Smithsonian Institution Press.
- Kenrick, P., & Strullu-Derrien, C. (2014). The origin and early evolution of roots. *Plant Physiology*, 166(2), 570–580. <https://doi.org/10.1104/pp.114.244517>
- Kenrick, P., Wellman, C. H., Schneider, H., & Edgecombe, G. D. (2012). A timeline for terrestrialization: Consequences for the carbon cycle in the palaeozoic. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 367(1588), 519–536. <https://doi.org/10.1098/rstb.2011.0271>
- Kim, D., Paggi, J. M., Park, C., Bennett, C., & Salzberg, S. L. (2019). Graph-based genome alignment and genotyping with hisat2 and hisat-genotype. *Nature Biotechnology*, 37, 907–915. <https://doi.org/10.1038/s41587-019-0201-4>
- Kim, J., Harter, K., & Theologis, A. (1997). Protein-protein interactions among the aux/iaa proteins. *Proceedings of the National Academy of Sciences of the United States of America*, 94(22), 11786–11791. <https://doi.org/10.1073/pnas.94.22.11786>
- Kinosian, S. P., & Wolf, P. G. (2022). The biology of *Ceratopteris richardii* as a tool to understand plant evolution [PMID: 35311640; PMCID: PMC8979586]. *eLife*, 11, e75019. <https://doi.org/10.7554/eLife.75019>
- Kohchi, T., Yamato, K. T., Ishizaki, K., Yamaoka, S., & Nishihama, R. (2021). Development and molecular genetics of marchantia polymorpha. *Annual review of plant biology*, 72(1), 677–702.
- Kohli, A., Gahakwa, D., Vain, P., et al. (1999). Transgene expression in rice engineered through particle bombardment: Molecular factors controlling stable expression and transgene silencing. *Planta*, 208, 88–97. <https://doi.org/10.1007/s004250050538>

- Kolde, R. (2018). *Pheatmap: Pretty heatmaps* [R package version 1.0.12]. <https://github.com/raivokolde/pheatmap>
- Kozlov, A. M., Darriba, D., Flouri, T., Morel, B., & Stamatakis, A. (2019). RAXML-NG: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference. *Bioinformatics*, 35(21), 4453–4455. <https://doi.org/10.1093/bioinformatics/btz305>
- Kubo, H., Peeters, A. J. M., Aarts, M. G. M., Pereira, A., & Koornneef, M. (1999). Anthocyaninless2, a homeobox gene affecting anthocyanin distribution and root development in arabidopsis. *The Plant Cell*, 11(7), 1217–1226. <https://doi.org/10.1105/tpc.11.7.1217>
- Kumar, V., Thakur, J. K., & Prasad, M. (2021). Histone acetylation dynamics regulating plant development and stress responses. *Cellular and Molecular Life Sciences*, 78(10), 4467–4486. <https://doi.org/10.1007/s00018-021-03794-x>
- Landberg, K., Pederson, E. R., Viaene, T., Bozorg, B., Friml, J., Jönsson, H., Thelander, M., & 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), 1406–1419.
- Lang, D., Zimmer, A. D., Rensing, S. A., & Reski, R. (2010). Exploring plant biodiversity: The *Physcomitrella* genome and beyond. *Trends in Plant Science*, 15(9), 543–549. <https://doi.org/10.1016/j.tplants.2010.06.002>
- Lang, W. H. (1937). *On the plant-remains from the Downtonian of England and Wales*. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, 227(544), 245–291.
- Langfelder, P., & Horvath, S. (2008). WGCNA: An R package for weighted correlation network analysis. *BMC Bioinformatics*, 9, 559. <https://doi.org/10.1186/1471-2105-9-559>

- Langfelder, P., & Horvath, S. (2014). *Tutorial for the wgcna package for r* [Accessed January 2025]. https://edo98811.github.io/WGCNA_official_documentation/FemaleLiver-02-networkConstr-blockwise.pdf
- Langfelder, P., Zhang, B., & Horvath, S. (2008). Defining clusters from a hierarchical cluster tree: The dynamic tree cut package for r. *Bioinformatics*, 24(5), 719–720. <https://doi.org/10.1093/bioinformatics/btm563>
- Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with bowtie 2. *Nature Methods*, 9(4), 357–359. <https://doi.org/10.1038/nmeth.1923>
- Langmead, B., Wilks, C., Antonescu, V., & Charles, R. (2018). Scaling read aligners to hundreds of threads on general-purpose processors. *Bioinformatics*. <https://doi.org/10.1093/bioinformatics/bty648>
- Lemieux, C., Otis, C., & Turmel, M. (2016). Comparative chloroplast genome analyses of streptophyte green algae uncover major structural alterations in the klebsormidiophyceae, coleochaetophyceae and zygnematophyceae. *Frontiers in plant science*, 7, 697.
- Lemmon, A. R., Brown, J. M., Stanger-Hall, K., & Moriarty Lemmon, E. (2009). The impact of missing data on phylogenetic estimates obtained by maximum likelihood and bayesian inference. *Systematic Biology*, 58(1), 130–145.
- Lersten, N. R. (2004). *Flowering plant embryology: With emphasis on economic species*. Blackwell Publishing.
- Letunic, I., & Bork, P. (2021). Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. *Nucleic Acids Research*, 49(W1), W293–W296. <https://doi.org/10.1093/nar/gkab301>
- Lewis, L. A., & McCourt, R. M. (2004). Green algae and the origin of land plants. *American Journal of Botany*, 91(10), 1535–1556. <https://doi.org/https://doi.org/10.3732/ajb.91.10.1535>

- Li, C. (2023). *Huperzia.asiatica.pep.fa.gz* [Online resource]. <https://doi.org/10.6084/m9.figshare.23299682.v1>
- Li, F.-W., Brouwer, P., Carretero-Paulet, L., Cheng, S., De Vries, J., Delaux, P.-M., Eily, A., Koppers, N., Kuo, L.-Y., Li, Z., Simenc, M., Small, I., Wafula, E. K., Wheeler, G., Zhang, Y., Pryer, K. M., Rothfels, C. J., Sigel, E. M., dePamphilis, C. W., ... Marchant, D. B. (2018). Fern genomes elucidate land plant evolution and cyanobacterial symbioses. *Nature Plants*, 4, 460–472. <https://doi.org/10.1038/s41477-018-0188-8>
- Li, F.-W., Nishiyama, T., Waller, M., Frangedakis, E., Keller, J., Li, Z., Fernandez-Pozo, N., Barker, M. S., Bennett, T., Blázquez, M. A., et al. (2020). Anthoceros genomes illuminate the origin of land plants and the unique biology of hornworts. *Nature plants*, 6(3), 259–272.
- Li, L.-C., Qin, G.-J., Tsuge, T., Hou, X.-H., Ding, M.-Y., Aoyama, T., Oka, A., Chen, Z., Gu, H., Zhao, Y., & Qu, L.-J. (2008). Sporocyteless modulates yucca expression to regulate the development of lateral organs in arabidopsis. *New Phytologist*, 179(3), 751–764. <https://doi.org/https://doi.org/10.1111/j.1469-8137.2008.02514.x>
- Li, W., & Godzik, A. (2006). Cd-hit: A fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, 22(13), 1658–1659. <https://doi.org/10.1093/bioinformatics/btl158>
- Li, X., Yang, C., Wu, H., Chen, H., Gao, X., Zhou, S., Zhang, T.-C., & Ma, W. (2024). Dsb-induced oxidative stress: Uncovering crosstalk between dna damage response and cellular metabolism. *DNA Repair*, 141, 103730. <https://doi.org/https://doi.org/10.1016/j.dnarep.2024.103730>
- Ligrone, R., Duckett, J. G., & Renzaglia, K. S. (2012). Major transitions in the evolution of early land plants: a bryological perspective. *Annals of Botany*, 109(5), 851–871. <https://doi.org/10.1093/aob/mcs017>

- Linkies, A., Graeber, K., Knight, C., & Leubner-Metzger, G. (2010). The evolution of seeds. *New Phytologist*, 186(4), 817–831. <https://doi.org/https://doi.org/10.1111/j.1469-8137.2010.03249.x>
- Lischer, H. E. L., & Shimizu, K. K. (2017). Reference-guided de novo assembly approach improves genome reconstruction for related species. *BMC Bioinformatics*, 18, 474. <https://doi.org/10.1186/s12859-017-1911-6>
- Liu, J., Nannas, N. J., Fu, F., Shi, J., Aspinwall, B., Parrott, W. A., & Dawe, R. K. (2019). Genome-scale sequence disruption following biolistic transformation in rice and maize. *The Plant Cell*, 31(2), 368–383. <https://doi.org/10.1105/tpc.18.00613>
- Liu, Q., Rand, T. A., Kalidas, S., Du, F., Kim, H.-E., Smith, D. P., & Wang, X. (2003). R2d2, a bridge between the initiation and effector steps of the *drosophila* rai pathway. *Science*, 301(5641), 1921–1925. <https://doi.org/10.1126/science.1088710>
- Liu, X., Ning, K., Che, G., Yan, S., Han, L., Gu, R., Li, Z.-M., Weng, Y., & Zhang, X. (2018). CsSPL functions as an adaptor between HD-ZIP III and CsWUS transcription factors regulating anther and ovule development in *Cucumis sativus* (cucumber). *Plant Journal*, 94(3), 535–547. <https://doi.org/10.1111/tpj.13877>
- Lloyd, R. M. (1974). Systematics of the genus *ceratopteris* brongn. (parkeriaceae) ii. taxonomy. *Brittonia*, 26, 139. <https://doi.org/10.2307/2805883>
- Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for rna-seq data with deseq2. *Genome Biology*, 15, 550. <https://doi.org/10.1186/s13059-014-0550-8>
- Lucas, W. J., Groover, A., Lichtenberger, R., Furuta, K., Yadav, S.-R., Helariutta, Y., He, X.-Q., Fukuda, H., Kang, J., Brady, S. M., Patrick, J. W., Sperry, J., Yoshida, A., López-Millán, A.-F., Grusak, M. A., & Kachroo, P. (2013). The plant vascular system: Evolution, development and functions. *Journal of Integrative Plant Biology*, 55(4), 294–388. <https://doi.org/https://doi.org/10.1111/jipb.12041>

- Ma, H. (1994). The unfolding drama of flower development: Recent results from genetic and molecular analyses. *Genes Dev*, 8(7), 745–56.
- Ma, Z., Hu, L., & Jiang, W. (2024). Understanding ap2/erf transcription factor responses and tolerance to various abiotic stresses in plants: A comprehensive review. *International Journal of Molecular Sciences*, 25(2), 893. <https://doi.org/10.3390/ijms25020893>
- Mable, B. K., & Otto, S. P. (1998). The evolution of life cycles with haploid and diploid phases. *BioEssays*, 20(6), 453–462.
- Madeira, F., Madhusoodanan, N., Lee, J., Eusebi, A., Niewielska, A., Tivey, A. R. N., Lopez, R., & Butcher, S. (2024). The embl-ebi job dispatcher sequence analysis tools framework in 2024. *Nucleic acids research*, 52(W1), W521–W525. <https://doi.org/10.1093/nar/gkae241>
- Malik, A. J., Poole, A. M., & Allison, J. R. (2020). Structural phylogenetics with confidence. *Molecular Biology and Evolution*, 37(9), 2711–2726. <https://doi.org/10.1093/molbev/msaa100>
- Marchant, D. B., Chen, G., Cai, S., et al. (2022). Dynamic genome evolution in a model fern. *Nature Plants*, 8, 1038–1051. <https://doi.org/10.1038/s41477-022-01226-7>
- Marchant, D. B., Sessa, E. B., Wolf, P. G., et al. (2019). The c-fern (*Ceratopteris richardii*) genome: Insights into plant genome evolution with the first partial homosporous fern genome assembly. *Scientific Reports*, 9, 18181. <https://doi.org/10.1038/s41598-019-53968-8>
- Matzke, M. A., & Mosher, R. A. (2014). Rna-directed dna methylation: An epigenetic pathway of increasing complexity (vol 15, 394, 2014). *Nature Reviews Genetics*, 15(8). <https://doi.org/10.1038/nrg3391>
- McDaniel, S. F. (2021). Bryophytes are not early diverging land plants. *New Phytologist*, 230(4), 1300–1304. <https://doi.org/10.1111/nph.17241>
- Meade, L. E., Plackett, A. R. G., & Hilton, J. (2021). Reconstructing development of the earliest seed integuments raises a new hypothesis for the evolution of ancestral

- seed-bearing structures. *New Phytologist*, 229(3), 1782–1794. <https://doi.org/https://doi.org/10.1111/nph.16792>
- Mendes, M. A., Petrella, R., Cucinotta, M., Vignati, E., Gatti, S., Pinto, S. C., Bird, D. C., Gregis, V., Dickinson, H., Tucker, M. R., & Colombo, L. (2020). The rna-dependent dna methylation pathway is required to restrict sporocyteless/nozzle expression to specify a single female germ cell precursor in arabidopsis. *Development*, 147(23). <https://doi.org/10.1242/dev.194274>
- Meyerowitz, E. M., Bowman, J. L., Brockman, L. L., Drews, G. N., Jack, T., Sieburth, L. E., & Weigel, D. (1991). A genetic and molecular model for flower development in arabidopsis thaliana. *Development*, 113(Supplement_1), 157–167.
- Miki, D., & Shimamoto, K. (2004). Simple rnai vectors for stable and transient suppression of gene function in rice. *Plant Cell Physiology*, 45(4), 490–495. <https://doi.org/10.1093/pcp/pch048>
- Mishler, B. D., & Churchill, S. P. (1984). A cladistic approach to the phylogeny of the “bryophytes”. *Brittonia*, 36, 406–424. <https://doi.org/10.2307/2806602>
- Mishler, B. D., & Churchill, S. P. (1985). Transition to a land flora: Phylogenetic relationships of the green algae and bryophytes. *Cladistics*, 1, 305–328. <https://doi.org/10.1111/j.1096-0031.1985.tb00431.x>
- Mizukami, Y., & Ma, H. (1992). Ectopic expression of the floral homeotic gene agamous in transgenic arabidopsis plants alters floral organ identity. *Cell*, 71(1), 119–131.
- Mizzotti, C., Ezquer, I., Paolo, D., Rueda-Romero, P., Guerra, R. F., Battaglia, R., Rogachev, I., Aharoni, A., Kater, M. M., & Colombo, L. (2020). Seedstick is a master regulator of development and metabolism in the arabidopsis seed coat. *Planta*, 251(2). <https://doi.org/10.1007/s00425-019-03414-x>
- Modrusan, Z., Reiser, L., Feldmann, K. A., Fischer, R. L., & Haughn, G. W. (1994). Homeotic transformation of ovules into carpel-like structures in arabidopsis. *Plant Cell*, 6(3), 333–349. <https://doi.org/10.1105/tpc.6.3.333>

- Moody, L. A. (2019). The 2d to 3d growth transition in the moss *Physcomitrella patens* [Growth and development]. *Current Opinion in Plant Biology*, 47, 88–95. <https://doi.org/10.1016/j.pbi.2018.10.001>
- Morris, C., Reem, N. T., & Walker, J. C. (2016). Flavonoids: A colorful model for the regulation of cellular processes. *Frontiers in Plant Science*, 7, 410. <https://doi.org/10.3389/fpls.2016.00410>
- Morris, J. L., Puttick, M. N., Clark, J. W., Edwards, D., Kenrick, P., Pressel, S., Wellman, C. H., Yang, Z., Schneider, H., & Donoghue, P. C. J. (2018). The timescale of early land plant evolution. *Proceedings of the National Academy of Sciences*, 115(10), E2274–E2283. <https://doi.org/10.1073/pnas.1719588115>
- Nickrent, D. L., Parkinson, C. L., Palmer, J. D., & Duff, R. J. (2000). Multigene phylogeny of land plants with special reference to bryophytes and the earliest land plants. *Molecular Biology and Evolution*, 17(12), 1885–1895. <https://doi.org/10.1093/oxfordjournals.molbev.a026290>
- Nie, Y., Foster, C. S. P., Zhu, T., Yao, R., Duchêne, D. A., Ho, S. Y. W., & Zhong, B. (2019). Accounting for Uncertainty in the Evolutionary Timescale of Green Plants Through Clock-Partitioning and Fossil Calibration Strategies. *Systematic Biology*, 69(1), 1–16. <https://doi.org/10.1093/sysbio/syz032>
- Niklas, K. J. (1997). *The evolutionary biology of plants*. University of Chicago Press.
- Niklas, K. J. (2000). The evolution of plant body plans—a biomechanical perspective. *Annals of Botany*, 85(4), 411–438. <https://doi.org/10.1006/anbo.1999.1060>
- Niklas, K. J., & Kutschera, U. (2010). The evolution of the land plant life cycle. *New Phytologist*, 185(1), 27–41.
- Nishiyama, T., & Kato, M. (1999). Molecular phylogenetic analysis among bryophytes and tracheophytes based on combined data of plastid coded genes and the 18s rRNA gene. *Molecular Biology and Evolution*, 16(8), 1027–1036.
- Olmedo-Monfil, V., Duran-Figueroa, N., Hamamura, Y., Gillmor, C. S., Conner, J. A., Ana, P. M., Kinoshita, T., Harada, J. J., Holder, A., Gehring, M., & Sundaresan, V.

- (2010). Control of female gamete formation by a small rna pathway in arabidopsis. *Nature*, 464(7288), 628–U125. <https://doi.org/10.1038/nature08828>
- One Thousand Plant Transcriptomes Initiative. (2019). One thousand plant transcriptomes and the phylogenomics of green plants. *Nature*, 574(7780), 679–685. <https://doi.org/10.1038/s41586-019-1693-2>
- Oshlack, A., Robinson, M. D., & Young, M. D. (2010). From rna-seq reads to differential expression results [Epub 2010 Dec 22. PMID: 21176179; PMCID: PMC3046478]. *Genome Biology*, 11(12), 220. <https://doi.org/10.1186/gb-2010-11-12-220>
- Ozsolak, F., & Milos, P. M. (2011). Rna sequencing: Advances, challenges and opportunities. *Nature reviews genetics*, 12(2), 87–98.
- Pabón-Mora, N., Di Stilio, V. S., & Becker, A. (2019). Editorial: Genetic regulatory mechanisms underlying developmental shifts in plant evolution. *Front Plant Sci*, 10, 710. <https://doi.org/10.3389/fpls.2019.00710>
- Patro, R., Duggal, G., Love, M. I., Irizarry, R. A., & Kingsford, C. (2017). Salmon provides fast and bias-aware quantification of transcript expression [Epub 2017 Mar 6]. *Nature Methods*, 14(4), 417–419. <https://doi.org/10.1038/nmeth.4197>
- Pennisi, E. (2009). On the origin of flowering plants.
- Pertea, M., Kim, D., Pertea, G. M., Leek, J. T., & Salzberg, S. L. (2016). Transcript-level expression analysis of rna-seq experiments with hisat, stringtie and ballgown. *Nature protocols*, 11(9), 1650–1667.
- Petersen, K. B., & Burd, M. (2017). Why did heterospory evolve? *Biological Reviews*, 92(3), 1739–1754.
- Petersen, K. B., & Burd, M. (2018). The adaptive value of heterospory: Evidence from selaginella. *Evolution*, 72(5), 1080–1091. <https://doi.org/10.1111/evo.13484>
- Pinto, S. C., Mendes, M. A., Coimbra, S., & Tucker, M. R. (2019). Revisiting the female germline and its expanding toolbox. *Trends in Plant Science*, 24(5), 455–467. <https://doi.org/10.1016/j.tplants.2019.02.003>

- Pinyopich, A., Ditta, G. S., Savidge, B., Liljegren, S. J., Baumann, E., Wisman, E., & Yanofsky, M. F. (2003). Assessing the redundancy of mads-box genes during carpel and ovule development. *Nature*, 424(6944), 85–88. <https://doi.org/10.1038/nature01741>
- Pires, N. D., & Dolan, L. (2012). Morphological evolution in land plants: New designs with old genes. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 367(1588), 508–518. <https://doi.org/10.1098/rstb.2011.0252>
- Plackett, A. R. G., Emms, D. M., Kelly, S., Hetherington, A. M., & Langdale, J. A. (2021). Conditional stomatal closure in a fern shares molecular features with flowering plant active stomatal responses [Epub 2021 Aug 26. PMID: 34450089]. *Current Biology*, 31(20), 4560–4570.e5. <https://doi.org/10.1016/j.cub.2021.08.008>
- Plackett, A. R. G., Rabbinowitsch, E. H., & Langdale, J. A. (2015). Protocol: Genetic transformation of the fern *Ceratopteris richardii* through microparticle bombardment. *Plant Methods*, 11, 37. <https://doi.org/10.1186/s13007-015-0080-8>
- Plackett, A. R., Conway, S. J., Hewett Hazelton, K. D., Rabbinowitsch, E. H., Langdale, J. A., & Di Stilio, V. S. (2018). *LEAFY* maintains apical stem cell activity during shoot development in the fern *Ceratopteris richardii* (C. S. Hardtke & S. McCormick, Eds.). *eLife*, 7, e39625. <https://doi.org/10.7554/eLife.39625>
- Plackett, A. R., Huang, L., Sanders, H. L., & Langdale, J. A. (2014). High-efficiency stable transformation of the model fern species *ceratopteris richardii* via microparticle bombardment. *Plant Physiology*, 165(1), 3–14. <https://doi.org/10.1104/pp.113.231357>
- Plomion, C., Aury, J.-M., Amselem, J., et al. (2018). Oak genome reveals facets of long lifespan. *Nature Plants*, 4, 440–452. <https://doi.org/10.1038/s41477-018-0172-3>
- Prestianni, C., & Gerrienne, P. (2010). Early seed plant radiation: An ecological hypothesis. In M. Vecoli, G. Clement, & B. MeyerBerthaud (Eds.), *The terrestrialization process: Modelling complex interactions at the*

- biosphere-geosphere interface* (pp. 71–80, Vol. 339). Geological Society Publishing House. <https://doi.org/10.1144/SP339.7>
- Prigge, M. J., & Bezanilla, M. (2010). Evolutionary crossroads in developmental biology: *Physcomitrella patens*. *Development*, 137(21), 3535–3543.
- Proctor, M. (2014). The diversification of bryophytes and vascular plants in evolving terrestrial environments. In D. Hanson & S. Rice (Eds.), *Photosynthesis in bryophytes and early land plants* (pp. 59–78). Springer.
- Pryer, K. M., Schneider, H., Smith, A. R., Cranfill, R., Wolf, P. G., Hunt, J. S., & Sipes, S. D. (2001). Horsetails and ferns are a monophyletic group and the closest living relatives to seed plants. *Nature*, 409(6820), 618–622. <https://doi.org/10.1038/35054555>
- Puttick, M. N., Morris, J. L., Williams, T. A., Cox, C. J., Edwards, D., Kenrick, P., Pressel, S., Wellman, C. H., Schneider, H., Pisani, D., & Donoghue, P. C. (2018). The interrelationships of land plants and the nature of the ancestral embryophyte. *Current Biology*, 28(5), 733–745.e2. <https://doi.org/https://doi.org/10.1016/j.cub.2018.01.063>
- Qiu, Y., Li, Z., & Köhler, C. (2024). Ancestral duplication of mads-box genes in land plants empowered the functional divergence between sporophytes and gametophytes. *New Phytologist*, 244(2), 358–363. <https://doi.org/https://doi.org/10.1111/nph.20065>
- Qiu, Y.-L., Li, L., Wang, B., Chen, Z., Knoop, V., Groth-Malonek, M., Dombrowska, O., Lee, J., Kent, L., Rest, J., Estabrook, G. F., Hendry, T. A., Taylor, D. W., Testa, C. M., Ambros, M., Crandall-Stotler, B., Duff, R. J., Stech, M., Frey, W., ... Davis, C. C. (2006). The deepest divergences in land plants inferred from phylogenomic evidence. *Proceedings of the National Academy of Sciences*, 103(42), 15511–15516. <https://doi.org/10.1073/pnas.0603335103>

- Qiu, Y.-L., Taylor, A. B., & McManus, H. A. (2012). Evolution of the life cycle in land plants. *Journal of Systematics and Evolution*, 50(3), 171–194. <https://doi.org/10.1111/j.1759-6831.2012.00188.x>
- Qiu, Y.-L., & Yu, J. (2003). *Azolla*—a model organism for plant genomic studies. *Genomics, Proteomics & Bioinformatics*, 1(1), 15–25. [https://doi.org/10.1016/s1672-0229\(03\)01004-0](https://doi.org/10.1016/s1672-0229(03)01004-0)
- Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., Peplies, J., & Glöckner, F. O. (2013). The silva ribosomal rna gene database project: Improved data processing and web-based tools. *Nucleic Acids Research*. <https://doi.org/10.1093/nar/gks1219>
- Quinlan, A. R., & Hall, I. M. (2010). Bedtools: A flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26, 841–842. <https://api.semanticscholar.org/CorpusID:268129191>
- Quinn, O., Kumar, M., & Turner, S. (2023). The role of lipid-modified proteins in cell wall synthesis and signaling. *Plant Physiology*, 194(1), 51–66. <https://doi.org/10.1093/plphys/kiad491>
- R Core Team. (2024). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria. <https://www.R-project.org/>
- Rajapaksa, S., Konagurthu, A. S., & Lesk, A. M. (2023). Sequence and structure alignments in post-alfafold era. *Current Opinion in Structural Biology*, 79, 102539. <https://doi.org/https://doi.org/10.1016/j.sbi.2023.102539>
- Raubeson, L. A., & Jansen, R. K. (1992). Chloroplast dna evidence on the ancient evolutionary split in vascular land plants. *Science*, 255(5052), 1697–1699. <https://doi.org/10.1126/science.255.5052.1697>
- Raven, J. A. (2002). Selection pressures on stomatal evolution. *New Phytologist*, 153(3), 371–386. <https://doi.org/10.1046/j.0028-646X.2001.00334.x>

- Raven, P. H., Evert, R. F., & Eichhorn, S. E. (2012). *Biology of plants* (8th). W.H. Freeman; Company.
- Ravet, K., & Pilon, M. (2013). Copper and iron homeostasis in plants: The challenges of oxidative stress. *Antioxid Redox Signal*, 19(9), 919–32. <https://doi.org/10.1089/ars.2012.5084>
- Reiser, L., Modrusan, Z., Margossian, L., Samach, A., Ohad, N., Haughn, G. W., & Fischer, R. L. (1995). The bell1 gene encodes a homeodomain protein involved in pattern formation in the arabidopsis ovule primordium. *Cell*, 83(5), 735–742.
- Ren, L., Tang, D., Zhao, T., Zhang, F., Liu, C., Xue, Z., Shi, W., Du, G., Shen, Y., Li, Y., & Cheng, Z. (2018). Os regulates meiotic fate acquisition in rice. *New Phytologist*, 218(2), 789–803. <https://doi.org/https://doi.org/10.1111/nph.15017>
- Rensing, S. A., Lang, D., Zimmer, A. D., Terry, A., Salamov, A., Shapiro, H., Nishiyama, T., Perroud, P.-F., Lindquist, E. A., Kamisugi, Y., et al. (2008). The physcomitrella genome reveals evolutionary insights into the conquest of land by plants. *Science*, 319(5859), 64–69.
- Robinson, M. D., McCarthy, D. J., & Smyth, G. K. (2010). Edger: A bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1), 139–140. <https://doi.org/10.1093/bioinformatics/btp616>
- Robinsonbeers, K., Pruitt, R. E., & Gasser, C. S. (1992). Ovule development in wild-type arabidopsis and 2 female-sterile mutants. *Plant Cell*, 4(10), 1237–1249. <https://doi.org/10.1105/tpc.4.10.1237>
- Rodkiewicz, B. (1970). Callose in cell walls during megasporogenesis in angiosperms. *Planta*, 93, 39–47.
- Roe, E., Keith. (1975). Origin of the alternation of generations in plants: Reconsideration of the traditional theories. *Biologist*.
- Rojas-Gracia, P., Roque, E., Medina, M., Rochina, M., Hamza, R., Angarita-Díaz, M. P., Moreno, V., Pérez-Martín, F., Lozano, R., Cañas, L., Beltrán, J. P., & Gómez-Mena, C. (2017). The parthenocarpic hydra mutant reveals a new

- function for a sporocyteless-like gene in the control of fruit set in tomato. *New Phytologist*, 214(3), 1198–1212.
<https://doi.org/https://doi.org/10.1111/nph.14433>
- Ronquist, F., Teslenko, M., van der Mark, P., Ayres, D. L., Darling, A., Höhna, S., Larget, B., Liu, L., Suchard, M. A., & Huelsenbeck, J. P. (2012). MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. *Systematic Biology*, 61(3), 539–542. <https://doi.org/10.1093/sysbio/sys029>
- Rosati, D., Palmieri, M., Brunelli, G., Morrione, A., Iannelli, F., Frullanti, E., & Giordano, A. (2024). Differential gene expression analysis pipelines and bioinformatic tools for the identification of specific biomarkers: A review [PMID: 38510977; PMCID: PMC10951429]. *Computational and Structural Biotechnology Journal*, 23, 1154–1168. <https://doi.org/10.1016/j.csbj.2024.02.018>
- Rothwell, G. W. (1986). Classifying the earliest gymnosperms. In R. A. Spicer & B. A. Thomas (Eds.), *Systematic and taxonomic approaches in palaeobotany* (pp. 137–161, Vol. 31). Systematics Association.
- Rubinstein, C. V., Gerrienne, P., de la Puente, G. S., Astini, R. A., & Steemans, P. (2010). Early middle ordovician evidence for land plants in argentina (eastern gondwana). *New Phytologist*, 188(2), 365–369.
<https://doi.org/https://doi.org/10.1111/j.1469-8137.2010.03433.x>
- Rutherford, G., Tanurdzic, M., Hasebe, M., & Banks, J. A. (2004). A systemic gene silencing method suitable for high throughput, reverse genetic analyses of gene function in fern gametophytes. *BMC Plant Biology*, 4, 6.
<https://doi.org/10.1186/1471-2229-4-6>
- Sakakibara, K., Nishiyama, T., Deguchi, H., & Hasebe, M. (2008). Class 1 knox genes are not involved in shoot development in the moss *Physcomitrella patens* but do function in sporophyte development. *Evolution & Development*, 10(5), 555–566.
<https://doi.org/https://doi.org/10.1111/j.1525-142X.2008.00271.x>

- Sanchez-Vera, V., Kenchappa, C. S., Landberg, K., Bressendorff, S., Schwarzbach, S., Martin, T., Mundy, J., Petersen, M., Thelander, M., & Sundberg, E. (2017). Autophagy is required for gamete differentiation in the moss *Physcomitrella patens*. *Autophagy*, 13(11), 1939–1951.
- Sauquet, H., von Balthazar, M., Magallón, S., Doyle, J. A., Endress, P. K., Bailes, E. J., Barroso de Morais, E., Bull-Hereñu, K., Carrive, L., Chartier, M., Chomicki, G., Coiro, M., Cornette, R., El Ottra, J. H., Epicoco, C., Foster, C. S., Jabbour, F., Haevermans, A., Haevermans, T., ... Schönenberger, J. (2017). The ancestral flower of angiosperms and its early diversification. *Nature Communications*, 8, 16047. <https://doi.org/10.1038/ncomms16047>
- Schedlbauer, M. D., & Klekowski, J., Edward J. (2008). Antheridogen activity in the fern *Ceratopteris thalictroides* (L.) Brongn. *Botanical Journal of the Linnean Society*, 65(4), 399–413. <https://doi.org/10.1111/j.1095-8339.1972.tb02280.x>
- Schiefthaler, U., Balasubramanian, S., Sieber, P., Chevalier, D., Wisman, E., & Schneitz, K. (1999). Molecular analysis of nozzle, a gene involved in pattern formation and early sporogenesis during sex organ development in *Arabidopsis thaliana*. *Proceedings of the National Academy of Sciences of the United States of America*, 96(20), 11664–11669. <https://doi.org/10.1073/pnas.96.20.11664>
- Schneitz, K., Hulskamp, M., & Pruitt, R. E. (1995). Wild-type ovule development in *Arabidopsis thaliana*: A light microscope study of cleared whole-mount tissue. *Plant Journal*, 7(5), 731–749. <https://doi.org/10.1046/j.1365-313X.1995.7050731.x>
- Schuettpelz, E., & Pryer, K. M. (2008). Fern phylogeny. In T. A. Ranker & C. H. Haufler (Eds.), *Biology and evolution of ferns and lycophytes* (pp. 395–416). Cambridge University Press. <https://doi.org/10.1017/CBO9780511541827.019>
- Schuettpelz, E., & Pryer, K. M. (2009). Evidence for a cenozoic radiation of ferns in an angiosperm-dominated canopy. *Proceedings of the National Academy of Sciences*, 106(27), 11200–11205.

- Schuettpelez, E., Schneider, H., Huie, L., Windham, M. D., & Pryer, K. M. (2007). A molecular phylogeny of the fern family pteridaceae: Assessing overall relationships and the affinities of previously unsampled genera. *Molecular Phylogenetics and Evolution*, 44(3), 1172–1185. <https://doi.org/https://doi.org/10.1016/j.ympev.2007.04.011>
- Schuettpelez, E., Schneider, H., Smith, A. R., Hovenkamp, P. H., Prado, J., Rouhan, G., Salino, A., Sundue, M. A., Almeida, T. E., Parris, B., Sessa, E. B., Field, A. R., de Gasper, A. L., Rothfels, C. J., Windham, M. D., Lehnert, M., Dauphin, B., Ebihara, A., Lehtonen, S., ... Zhou, X.-M. (2016). A community-derived classification for extant lycophytes and ferns. *Journal of Systematics and Evolution*, 54. <https://api.semanticscholar.org/CorpusID:39980610>
- Schweizer, P., Pokorny, J., Schulze-Lefert, P., & Dudler, R. (2000). Double-stranded rna interferes with gene function at the single-cell level in cereals. *The Plant Journal*, 24, 895–903. <https://doi.org/10.1046/j.1365-313x.2000.00940.x>
- Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., Amin, N., Schwikowski, B., & Ideker, T. (2003). Cytoscape: A software environment for integrated models of biomolecular interaction networks. *Genome Research*, 13(11), 2498–2504. <https://doi.org/10.1101/gr.1239303>
- Shaw, A. J., Szövényi, P., & Shaw, B. (2011). Bryophyte diversity and evolution: Windows into the early evolution of land plants. *American Journal of Botany*, 98(3), 352–369. <https://doi.org/https://doi.org/10.3732/ajb.1000316>
- Shaw, J., & Renzaglia, K. (2004). Phylogeny and diversification of bryophytes. *American Journal of Botany*, 91(10), 1557–1581. <https://doi.org/https://doi.org/10.3732/ajb.91.10.1557>
- Shiraishi, H., Okada, K., & Shimura, Y. (1993). Nucleotide sequences recognized by the agamous mads domain of arabidopsis thaliana in vitro. *The Plant Journal*, 4(2), 385–398. <https://doi.org/https://doi.org/10.1046/j.1365-313X.1993.04020385.x>

- Shore, P., & Sharrocks, A. D. (1995). The mads-box family of transcription factors. *European Journal of Biochemistry*, 229(1), 1–13. <https://doi.org/10.1111/j.1432-1033.1995.tb20430.x>
- Sieber, P., Gheyselinck, J., Gross-Hardt, R., Laux, T., Grossniklaus, U., & Schneitz, K. (2004). Pattern formation during early ovule development in arabidopsis thaliana. *Developmental Biology*, 273(2), 321–334. <https://doi.org/https://doi.org/10.1016/j.ydbio.2004.05.037>
- Simmons, M. P., Richardson, D., & Reddy, A. S. N. (2008). Incorporation of gap characters and lineage-specific regions into phylogenetic analyses of gene families from divergent clades: An example from the kinesin superfamily across eukaryotes. *Cladistics*, 24(3), 372–384. <https://doi.org/https://doi.org/10.1111/j.1096-0031.2007.00183.x>
- Smith, A. B., & McGowan, A. J. (2007). The shape of the phanerozoic marine palaeodiversity curve: How much can be predicted from the sedimentary rock record of western europe? *Palaeontology*, 50(4), 765–774. <https://doi.org/https://doi.org/10.1111/j.1475-4983.2007.00693.x>
- Smyth, D. R., Bowman, J. L., & Meyerowitz, E. M. (1990). Early flower development in arabidopsis. *Plant Cell*, 2(8), 755–767. <https://doi.org/10.1105/tpc.2.8.755>
- Soltis, D. E., Albert, V. A., Leebens-Mack, J., Palmer, J. D., Wing, R. A., dePamphilis, C. W., Ma, H., Carlson, J. E., Altman, N., Kim, S., et al. (2008). The amborella genome: An evolutionary reference for plant biology. *Genome biology*, 9, 1–6.
- Soltis, P. S., Marchant, D. B., Van de Peer, Y., & Soltis, D. E. (2015). Polyploidy and genome evolution in plants [Genomes and evolution]. *Current Opinion in Genetics & Development*, 35, 119–125. <https://doi.org/https://doi.org/10.1016/j.gde.2015.11.003>
- Song, L., & Florea, L. (2015). Rcorrector: Efficient and accurate error correction for illumina rna-seq reads. *GigaScience*, 4, 48. <https://doi.org/10.1186/s13742-015-0089-y>

- Sparks, C., & Jones, H. (2004). Transformation of wheat by biolistics. In *Transgenic crops of the world: Essential protocols* (pp. 19–34). Springer.
- Srivastava, V., Vasil, V., & Vasil, I. (1996). Molecular characterization of the fate of transgenes in transformed wheat (*triticum aestivum* L.) *Theoretical and Applied Genetics*, 92, 1031–1037.
- Stebbins, G. L. (1974). *Flowering plants: Evolution above the species level*. Harvard University Press.
- Stein, D. B. (1971). Gibberellin-induced fertility in the fern *ceratopteris thalictroides* (L.) brongn. 1. *Plant Physiology*, 48(4), 416–418. <https://doi.org/10.1104/pp.48.4.416>
- Stein, W. (1993). Modeling the evolution of stelar architecture in vascular plants. *International Journal of Plant Sciences*, 154(2), 229–263. <https://doi.org/10.1086/297106>
- Stoger, E., Williams, S., Keen, D., & Christou, P. (1998). Molecular characteristics of transgenic wheat and the effect on transgene expression. *Transgenic research*, 7, 463–471.
- Stothard, P. (2000). The sequence manipulation suite: Javascript programs for analyzing and formatting protein and dna sequences. *Biotechniques*, 28, 1102–1104.
- Stout, S. C., Clark, G. B., Archer-Evans, S., & Roux, S. J. (2003). Rapid and efficient suppression of gene expression in a single-cell model system, *ceratopteris richardii*. *Plant Physiology*, 131(3), 1165–1168.
- Strasburger, E. (1894). The periodic reduction of the number of the chromosomes in the life-history of living organisms. *Annals of Botany*, 8(31), 281–316.
- Strother, P. K., & Foster, C. (2021). A fossil record of land plant origins from charophyte algae. *Science*, 373(6556), 792–796. <https://doi.org/10.1126/science.abj2927>
- Sun, X. (2022). Genome assembly and annotation of *torreya grandis* [Dataset]. <https://figshare.com/>

- Szemenyei, H., Hannon, M., & Long, J. A. (2008). Topless mediates auxin-dependent transcriptional repression during arabidopsis embryogenesis. *Science*, 319(5868), 1384–6. <https://doi.org/10.1126/science.1151461>
- Szövényi, P., Frangedakis, E., Ricca, M., Quandt, D., Wicke, S., & Langdale, J. A. (2015). Establishment of *anthoceros agrestis* as a model species for studying the biology of hornworts. *BMC plant biology*, 15, 1–7.
- Takashima-Oleivera, T. T., Medeiros, P. S., & Tavares-Martins, A. C. (2020). Bryophyte communities across the ecological succession process in the caxiuanã national forest, pará, brazil. *Anais da Academia Brasileira de Ciências*, 92, e20180626. <https://doi.org/10.1590/0001-3765202020180626>
- Talavera, G., & Castresana, J. (2007). Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. *Systematic Biology*, 56(4), 564–577. <https://doi.org/10.1080/10635150701472164>
- Taylor, E. L., Taylor, T. N., & Krings, M. (2009). *Paleobotany: The biology and evolution of fossil plants*. Academic Press.
- Taylor, S. C., Nadeau, K., Abbasi, M., Lachance, C., Nguyen, M., & Fenrich, J. (2019). The ultimate qpcr experiment: Producing publication quality, reproducible data the first time [Epub 2019 Jan 14]. *Trends in Biotechnology*, 37(7), 761–774. <https://doi.org/10.1016/j.tibtech.2018.12.002>
- Tello-Ruiz, M. K., Stein, J., Wei, S., Preece, J., Olson, A., Naithani, S., Amarasinghe, V., Dharmawardhana, P., Jiao, Y., Mulvaney, J., Kumari, S., Chougule, K., Elser, J., Wang, B., Thomason, J., Bolser, D. M., Kerhornou, A., Walts, B., Fonseca, N. A., ... Ware, D. (2016). Gramene 2016: Comparative plant genomics and pathway resources. *Nucleic Acids Research*, 44(D1), D1133–D1140. <https://doi.org/10.1093/nar/gkv1179>
- Thermo Fisher Scientific. (2003). Gateway® technology [[Accessed January 2022]].
- Thompson, J. D., Higgins, D. G., & Gibson, T. J. (1994). CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence

- weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Research*, 22(22), 4673–4680. <https://doi.org/10.1093/nar/22.22.4673>
- Tomescu, A. M. (2009). Megaphylls, microphylls and the evolution of leaf development. *Trends in Plant Science*, 14(1), 5–12. <https://doi.org/https://doi.org/10.1016/j.tplants.2008.10.008>
- Tucker, M. R., Araujo, A. C. G., Paech, N. A., Hecht, V., Schmidt, E. D. L., Rossell, J. B., de Vries, S. C., & Koltunow, A. M. G. (2003). Sexual and apomictic reproduction in dandelion (*taraxacum officinale*) are closely interrelated developmental pathways. *Plant Cell*, 15(6), 1372–1386. <https://doi.org/10.1105/tpc.010323>
- Turmel, M., Otis, C., & Lemieux, C. (2013). Tracing the evolution of streptophyte algae and their mitochondrial genome. *Genome Biology and Evolution*, 5(10), 1817–1835. <https://doi.org/10.1093/gbe/evt135>
- Turmel, M., Otis, C., & Lemieux, C. (2006). The chloroplast genome sequence of chara vulgaris sheds new light into the closest green algal relatives of land plants. *Molecular biology and evolution*, 23(6), 1324–1338.
- Unamba, C. I. N., Nag, A., & Sharma, R. K. (2015). Next generation sequencing technologies: The doorway to the unexplored genomics of non-model plants. *Frontiers in Plant Science*, 6. <https://www.frontiersin.org/journals/plant-science/articles/10.3389/fpls.2015.01074>
- Van de Peer, Y., Maere, S., & Meyer, A. (2009). The evolutionary significance of ancient genome duplications [Epub 2009 Aug 4]. *Nature Reviews Genetics*, 10(10), 725–732. <https://doi.org/10.1038/nrg2600>
- Vicidomini, C., Palumbo, R., Moccia, M., & Roviello, G. N. (2024). Oxidative processes and xenobiotic metabolism in plants: Mechanisms of defense and potential therapeutic implications. *Journal of Xenobiotics*, 14(4), 1541–1569. <https://doi.org/10.3390/jox14040084>
- Wagih, O. (2017). Ggseqlogo: A versatile r package for drawing sequence logos. *Bioinformatics*. <https://doi.org/10.1093/bioinformatics/btx469>

- Wan, T., Liu, Z.-M., Li, L.-F., et al. (2018). Data from: A genome for gnetophytes and early evolution of seed plants [[Dataset]]. <https://doi.org/10.5061/dryad.0vm37>
- Wang, L.-S., Leebens-Mack, J., Wall, P. K., Beckmann, K., DePamphilis, C. W., & Warnow, T. (2009). The impact of multiple protein sequence alignment on phylogenetic estimation. *IEEE/ACM transactions on computational biology and bioinformatics*, 8(4), 1108–1119.
- Wang, Z., Gerstein, M., & Snyder, M. (2009). Rna-seq: A revolutionary tool for transcriptomics. *Nature reviews genetics*, 10(1), 57–63.
- Waterhouse, A. M., Procter, J. B., Martin, D. M. A., Clamp, M., & Barton, G. J. (2009). Jalview version 2 - a multiple sequence alignment editor and analysis workbench. *Bioinformatics*. <https://doi.org/10.1093/bioinformatics/btp033>
- Watkins Jr, J., & Cardelús, C. L. (2012). Ferns in an angiosperm world: Cretaceous radiation into the epiphytic niche and diversification on the forest floor. *International Journal of Plant Sciences*, 173(6), 695–710.
- Wei, B. Y., Zhang, J. Z., Pang, C. X., Yu, H., Guo, D. S., Jiang, H., Ding, M. X., Chen, Z. Y., Tao, Q., Gu, H. Y., Qu, L. J., & Qin, G. J. (2015). The molecular mechanism of sporocyteless/nozzle in controlling arabidopsis ovule development. *Cell Research*, 25(1), 121–134. <https://doi.org/10.1038/cr.2014.145>
- Weigel, D., & Meyerowitz, E. M. (1994). The abcs of floral homeotic genes. *Cell*, 78(2), 203–209.
- Wickett, N. J., Mirarab, S., Nguyen, N., Warnow, T., Carpenter, E., Matasci, N., Ayyampalayam, S., Barker, M. S., Burleigh, J. G., Gitzendanner, M. A., Ruhfel, B. R., Wafula, E., Der, J. P., Graham, S. W., Mathews, S., Melkonian, M., Soltis, D. E., Soltis, P. S., Miles, N. W., ... Leebens-Mack, J. (2014). Phylotranscriptomic analysis of the origin and early diversification of land plants. *Proceedings of the National Academy of Sciences*, 111(45), E4859–E4868. <https://doi.org/10.1073/pnas.1323926111>

- Wickham, H. (2016). *Ggplot2: Elegant graphics for data analysis*. Springer-Verlag New York. <https://ggplot2.tidyverse.org>
- Wickham, H., François, R., Henry, L., Müller, K., & Vaughan, D. (2023). *Dplyr: A grammar of data manipulation* [R package version 1.1.4, <https://github.com/tidyverse/dplyr>]. <https://dplyr.tidyverse.org>
- Willis, C. G., Baskin, C. C., Baskin, J. M., Auld, J. R., Venable, D. L., Cavender-Bares, J., Donohue, K., Rubio de Casas, R., & Group, T. N. G. W. (2014). The evolution of seed dormancy: Environmental cues, evolutionary hubs, and diversification of the seed plants. *New Phytologist*, 203(1), 300–309. <https://doi.org/https://doi.org/10.1111/nph.12782>
- Willis, K. J., & McElwain, J. C. (2014). *The evolution of plants*. Oxford University Press, USA.
- Windsor, J. B., Symonds, V. V., Mendenhall, J., & Lloyd, A. M. (2000). Arabidopsis seed coat development: Morphological differentiation of the outer integument. *The Plant Journal*, 22(6), 483–493. <https://doi.org/10.1046/j.1365-313x.2000.00756.x>
- Wing, S., & Boucher, L. (1998). Ecological aspects of the cretaceous flowering plant radiation. *Annual Review of Earth and Planetary Sciences*, 26, 379–421. <https://doi.org/10.1146/annurev.earth.26.1.379>
- Withers, K. A., Falls, K., Youngstrom, C. E., Nguyen, T., DeWald, A., Yarvis, R. M., Simons, G. P., Flanagan, R., Bui, L. T., Irish, E. E., & Cheng, C.-L. (2023). A ceratopteris excess microsporocytes1 suppresses reproductive transition in the fern vegetative leaves. *Plant Science*, 335, 111812. <https://doi.org/https://doi.org/10.1016/j.plantsci.2023.111812>
- Woudenberg, S., Dipp Alvarez, M., Rienstra, J., Levitsky, V., Mironova, V., Scarpella, E., Kuhn, A., & Weijers, D. (2024). Analysis of auxin responses in the fern *Ceratopteris richardii* identifies the developmental phase as a major determinant for response properties. *Development*, 151(20), dev203026. <https://doi.org/10.1242/dev.203026>

- Xiang, D., & Li, G. (2024). Control of leaf development in the water fern *Ceratopteris richardii* by the auxin efflux transporter crpinma in the crispr/cas9 analysis. *BMC Plant Biology*, 24, 322. <https://doi.org/10.1186/s12870-024-05009-4>
- Xiao, Y.-L., & Li, G.-S. (2024). Differential expression and co-localization of transcriptional factors during callus transition to differentiation for shoot organogenesis in the water fern *Ceratopteris richardii*. *Annals of Botany*, 133(3), 495–507. <https://doi.org/10.1093/aob/mcae006>
- Xie, L., Gong, X., Yang, K., Huang, Y., Zhang, S., Shen, L., Sun, Y., Wu, D., Ye, C., Zhu, Q.-H., & Fan, L. (2024). Technology-enabled great leap in deciphering plant genomes. *Nature Plants*, 10(4), 551–566. <https://doi.org/10.1038/s41477-024-01655-6>
- Yang, J.-H., Zhang, M.-F., & Yu, J.-Q. (2008). Relationship between cytoplasmic male sterility and spl-like gene expression in stem mustard. *Physiologia Plantarum*, 133(2), 426–434. <https://doi.org/10.1111/j.1399-3054.2008.01064.x>
- Yang, W. C., Shi, D. Q., & Chen, Y. H. (2010). Female gametophyte development in flowering plants. In S. Merchant, W. R. Briggs, & D. Ort (Eds.), *Annual review of plant biology*, vol 61 (pp. 89–108, Vol. 61). Annual Reviews. <https://doi.org/10.1146/annurev-arplant-042809-112203>
- Yang, W. C., Ye, D., Xu, J., & Sundaresan, V. (1999). The sporocyteless gene of arabidopsis is required for initiation of sporogenesis and encodes a novel nuclear protein. *Genes & Development*, 13(17), 2108–2117. <https://doi.org/10.1101/gad.13.17.2108>
- Youngstrom, C. E., Withers, K. A., Irish, E. E., & Cheng, C. L. (2022). Vascular function of the t3/modern clade wuschel-related homeobox transcription factor genes predate apical meristem-maintenance function. *BMC Plant Biology*, 22(1), 210. <https://doi.org/10.1186/s12870-022-03590-0>
- Yu, J. (2022). *Lycopodium clavatum* genome annotation [Dataset]. <https://doi.org/10.6084/m9.figshare.20493417.v3>

- Zhang, B., & Horvath, S. (2005). A general framework for weighted gene co-expression network analysis. *Statistical Applications in Genetics and Molecular Biology*, 4(1). <https://doi.org/10.2202/1544-6115.1128>
- Zhao, S., Ye, Z., & Stanton, R. (2020). Misuse of rpkm or tpm normalization when comparing across samples and sequencing protocols [Epub 2020 Apr 13]. *RNA*, 26(8), 903–909. <https://doi.org/10.1261/rna.074922.120>
- Zinta, G., Khan, A., AbdElgawad, H., Verma, V., & Srivastava, A. K. (2016). Unveiling the redox control of plant reproductive development during abiotic stress. *Frontiers in Plant Science*, 7. <https://doi.org/10.3389/fpls.2016.00700>
- Zwickl, D. J., & Hillis, D. M. (2002). Increased taxon sampling greatly reduces phylogenetic error. *Systematic biology*, 51(4), 588–598.

Appendix A: Growth media and buffer recipes

Growth media

C-fern growth media

- 100 mL 10x Macronutrient Solution
- 10 mL 100x Chelated Iron Solution
- 5 mL 200x Micronutrient Solution
- Up to 1 L with dH₂O
- Adjust to pH 6.0 using KOH

Macronutrients 10x (1L)

- 1.25 g NH₄NO₃
- 5 g KH₂PO₄
- 1.2 g MgSO₄·7H₂O
- 0.26 g CaCl₂·2H₂O
- Made up to 1L with dH₂O

Micronutrients 200x (1L)

- 0.05 g MnSO₄·H₂O
- 0.074 g CuSO₄·5H₂O
- 0.104 g ZnSO₄·7H₂O
- 0.372 g H₃BO₃
- 0.0074 g (NH₄)MoO₄·4H₂O
- Made up to 1L with dH₂O

Chelated Iron 100x (1L)

- 2.78 g $\text{FeSO}_4 \cdot 7\text{H}_2\text{O}$
- 3.73 g Disodium EDTA- $2\text{H}_2\text{O}$
- Made up to 1L with dH_2O

LB media (1 L)

- **10 g** Bacto-Tryptone
- **5 g** Bacto-Yeast Extract
- **10 g** NaCl
- Made up to 1 L with dH_2O

1x MS 2% Sucrose Media (1 L)

- **0.5 g** Plant MES hydrate
- **4.16 g** Murashige-Skoog
- **20 g** Sucrose
- Made up to 1L with dH_2O
- Adjust to pH 5.8 using KOH

1/2x MS 2% Sucrose Media (1 L)

- **0.5 g** Plant MES hydrate
- **2.08 g** Murashige-Skoog
- **20 g** Sucrose
- Made up to 1L with dH_2O
- Adjust to pH 5.8 using KOH

Buffers

2% CTAB Extraction Buffer

(100mM Tris-HCl (pH 9.0); 2% CTAB; 1.4M NaCl; 20mM EDTA)

2% CTAB Extraction Buffer (400mL final volume):

- 40mL 1M Tris-HCl pH9.0 (1/10 dilution)
- 16mL 500mM EDTA (1/25 dilution)
- 8g CTAB (Hexadecyltrimethylammonium bromide)
- 32.72g NaCl
- dH₂O up to final volume of 400 mL

1M Tris-HCl pH9.0 (400mL final volume):

- 48.46g Tris-HCl
- Up to 400 mL dH₂O
- Adjust pH to 9.0 with 1M HCl

0.5M EDTA pH 8.0 (400mL final volume):

- 58.448 EDTA (MW: 292.24)
- Up to 400mL with dH₂O

TAE 50x Stock Solution (1L)

- **242 g** Tris base (MW = 121)
- **57.1 mL** Glacial acetic acid
- **100 mL** 0.5 M EDTA (pH 8.0)
- Made up to 1 L with dH₂O

Appendix B: Plasmid cloning results

Table B.1: Table summarising the results of all plasmids attempted for cloning. Each row represents an individual miniprep with details on plasmid concentration, diagnostic digest results, glycerol stock generation, and sequencing verification. Bolded rows highlight plasmids that were successfully cloned (digest and sequencing confirmed) and subsequently used in further experimental steps. Non-bolded rows indicate plasmids that were either unsuccessful or not pursued further.

Miniprep	Conc (ng/ul)	Digest Correct	Glycerol stock	Seq. verified
CrSPL_A-A4-pJET1.2 miniprep 1	239.7	No	-	-
CrSPL_A-A4-pJET1.2 miniprep 2	490.6	Yes	Yes	Yes
CrSPL_A-A7-pJET1.2 miniprep 1	600.2	Yes	Yes	Yes
CrSPL_A-A7-pJET1.2 miniprep 2	553.8	Yes	-	No
CrSPL_B-B5-pJET1.2 miniprep 1	575.9	No	-	No
CrSPL_B-B5-pJET1.2 miniprep 2	226.9	Yes	-	No
CrSPL_B-B10-pJET1.2 miniprep 1	139.7	No	-	-
CrSPL_B-B10-pJET1.2 miniprep 2	79.2	No	-	-

Miniprep	Conc (ng/ul)	Digest Correct	Glycerol stock	Seq. verified
CrSPL_B-B10-pJET1.2 miniprep 3	282.7	Yes	-	No
CrSPL_B-B10-pJET1.2 miniprep 4	341.2	No	-	-
CrSPL_B-B10-pJET1.2 miniprep 5	85.6	No	-	-
CrSPL_B-B10-pJET1.2 miniprep 6	369.4	No	-	-
CrSPL_B-B10-pJET1.2 miniprep 7	263.9	Yes	Yes	Yes
CrSPL_B-B10-pJET1.2 miniprep 8	236.7	No	-	-
AtSPL-T2-pJET1.2 miniprep 1	422.4	Yes	Yes	Yes
AtSPL-T2-pJET1.2 miniprep 2	526.6	No	-	-
AtSPL-T1-pJET1.2 miniprep 1	297.2	Yes	-	No
AtSPL-T1-pJET1.2 miniprep 2	301	No	-	-
AtSPLpro-16 miniprep 1	130.4	Yes	-	-
AtSPLpro-16 miniprep 2	759.5	Yes	-	-
AtSPLpro-16 miniprep 3	224.4	Yes	Yes	Yes
AtSPLpro-16 miniprep 4	515.9	No	-	-
AtSPLpro-17 miniprep 1	483.8	Yes	Yes	No
AtSPLpro-17 miniprep 2	760.1	Yes	-	No

Miniprep	Conc (ng/ul)	Digest Correct	Glycerol stock	Seq. verified
AtSPLpro-17 miniprep 3	534.5	Yes	-	No
AtSPLpro-17 miniprep 4	189.4	No	-	No
AtSPL3'UTR-16 miniprep 1	377.1	Yes	Yes	No
AtSPL3'UTR-16 miniprep 2	504.4	Yes	No	No
AtSPL3'UTR-16 miniprep 3	409.8	No	-	-
AtSPL3'UTR-16 miniprep 4	382.8	Yes	-	No
AtSPL3'UTR-17 miniprep 1	402.9	Yes	Yes	Yes
AtSPL3'UTR-17 miniprep 2	353	No	-	-
AtSPL3'UTR-17 miniprep 3	402.3	No	-	-
AtSPL3'UTR-17 miniprep 4	396.8	No	-	-
AtSPL3'UTR-16 miniprep 5	96.7	No	-	-
AtSPL3'UTR-16 miniprep 6	72	Yes	-	No
AtSPL3'UTR-16 miniprep 7	67.9	Yes	-	No
AtSPL3'UTR-16 miniprep 8	94.8	Yes	Yes	No
AtSPL3'UTR miniprep 5	41.6	No	-	-
AtSPL3'UTR miniprep 6	69.8	No	-	-
AtSPL3'UTR miniprep 7	39.2	No	-	-
AtSPL3'UTR miniprep 8	63.6	Yes	No	No
AtSPLpro-NotI-KpnI miniprep 1	79.6	Yes	-	No
AtSPLpro-NotI-KpnI miniprep 2	99.5	Yes	No	No

Miniprep	Conc (ng/ul)	Digest Correct	Glycerol stock	Seq. verified
AtSPLpro-NotI-KpnI miniprep 3	82.7	Yes	-	No
AtSPLpro-NotI-KpnI miniprep 4	36.7	Yes	-	No
AtSPLpro-NotI-KpnI miniprep 5	289.4	No	-	No
AtSPLpro-NotI-KpnI miniprep 6	239.4	No	-	No
AtSPLpro-NotI-KpnI miniprep 7	307.2	No	-	No
AtSPLpro-NotI-KpnI miniprep 8	395.3	No	-	No
AtSPLpro-NotI-KpnI miniprep 9	196.7	No	-	-
AtSPLpro-NotI-KpnI miniprep 10	196.8	No	-	-
AtSPLpro-NotI-KpnI miniprep 11	364	No	-	-
AtSPLpro-NotI-KpnI miniprep 12	166.1	No	-	-
AtSPLpro-NotI-KpnI miniprep 13	215.2	No	-	-
AtSPLpro-NotI-KpnI miniprep 14	317.1	No	-	-

Miniprep	Conc (ng/ul)	Digest Correct	Glycerol stock	Seq. verified
AtSPLpro-NotI-KpnI miniprep 15	200	No	-	-
AtSPLpro-NotI-KpnI miniprep 16	281	No	-	-
CrSPL_C miniprep 1	594.4	Yes	Yes	Yes
CrSPL_C miniprep 2	757	Yes	-	No
CrSPL_C miniprep 3	322.4	No	-	-
CrSPL_C miniprep 4	269.4	Yes	-	No
AtSPLpro-NotI-KpnI miniprep 17	538.9	Yes	-	No
AtSPLpro-NotI-KpnI miniprep 18	428.6	Yes	Yes	Yes
AtSPLpro-NotI-KpnI miniprep 19	511.5	Yes	-	No
AtSPLpro-NotI-KpnI miniprep 20	162.4	No	-	-
AtSPL3'UTR-NotI-XbaI miniprep 1	443.4	Yes	No	No
AtSPL3'UTR-NotI-XbaI miniprep 2	478.4	Yes	Yes	Yes
AtSPL3'UTR-NotI-XbaI miniprep 3	276.8	Yes	-	No
AtSPL3'UTR-NotI-XbaI miniprep 4	264.8	Yes	-	No

Miniprep	Conc (ng/ul)	Digest Correct	Glycerol stock	Seq. verified
mCHERRY1 miniprep 1	197.9	Yes	yes	Yes
mCHERRY1 miniprep 2	274.5	Yes	-	No
mCHERRY1 miniprep 3	194.8	Yes	-	No
mCHERRY1 miniprep 4	206.4	Yes	-	No
mCHERRY2 miniprep 1	210.8	Yes	-	No
mCHERRY2 miniprep 2	164.2	Yes	-	No
mCHERRY2 miniprep 3	322.6	Yes	-	No
mCHERRY2 miniprep 4	288.1	Yes	yes	Yes
mCHERRY-XbaI-KlpnI miniprep 1	154.4	Yes	No	No
mCHERRY-XbaI-KlpnI miniprep 2	208.9	Yes	Yes	Yes
mCHERRY-XbaI-KlpnI miniprep 3	85	Yes	-	No
mCHERRY-XbaI-KlpnI miniprep 4	69.1	Yes	-	No
CrSPL_C-RNAi miniprep 1	270.1	Yes	No	No
CrSPL_C-RNAi miniprep 2	224.3	Yes	-	No
pART7::AtSPL3'UTR miniprep 1	871.5	No	-	-
pART7::AtSPL3'UTR miniprep 2	640.6	No	-	-

Miniprep	Conc (ng/ul)	Digest Correct	Glycerol stock	Seq. verified
pART7::AtSPL3'UTR miniprep 3	826.4	Yes	NA	Yes
pART7::AtSPL3'UTR miniprep 4	682.8	Yes	NA	No
CrSPL_C-RNAi miniprep 3	210	Yes	Yes	Yes
CrSPL_C-RNAi miniprep 4	285.8	No	-	No
CrSPL_C-RNAi miniprep 5	325.7	Yes	-	No
CrSPL_C-RNAi miniprep 6	269	Yes	-	No

Appendix C: *De novo* assembly of Salmon unmapped reads

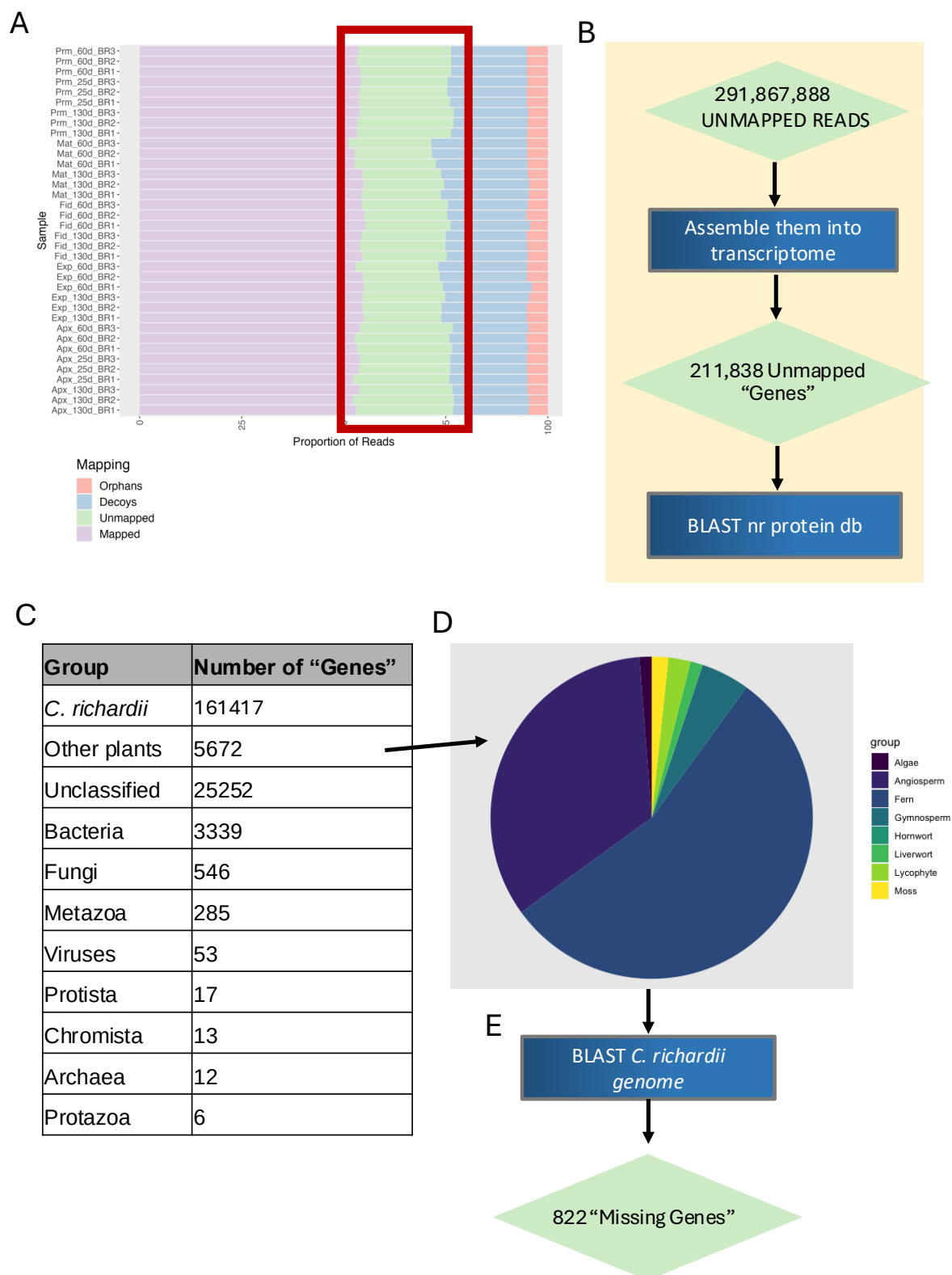


Figure C.1: De novo assembly of unmapped reads from quasi-alignment with Salmon yielded 822 "missing genes". Caption overleaf.

Figure C.1 *De novo* assembly of unmapped reads from quasi-alignment with Salmon yielded 822 “missing genes”. (A) Approximately 24 % of reads were unmapped when quasi-aligned with Salmon. These were extracted for downstream analysis. (B) 291,867,888 unmapped reads were *de novo* assembled with Trinity, resulting in 211,838 “genes”. These “genes” were then blasted against the non-redundant protein database. (C) The vast majority of reads were assigned to *C. richardii*. The 5672 “other plants” hits were explored in more detail. (D) Pie chart shows the 5672 “other plant” genes, and which lineage they were assigned to. (E) The 5672 other plant genes were BLASTed against the *C. richardii* database only. 822 did not have a hit, representing 822 new, previously unknown genes, that are not represented in the current genome assembly.

Appendix D: Module Gene Lists

Table D.1: GO annotation and best hits for genes in the ivory module.

Hub genes are in bold.

Gene ID	kME	Corr	GO Terms	Best Hit
Ceric.08G001000	0.941	+	GO:0003677 GO:0008289	HOMEBOX-LEUCINE ZIPPER PROTEIN ANTHOCYANINLESS 2-RELATED
Ceric.27G014500	0.934	+	-	-
Ceric.09G062600	0.933	+	-	BES1/BZR1 plant transcription factor, N-terminal
Ceric.24G049200	0.927	+	-	-
Ceric.15G075600	0.926	+	-	-
Ceric.21G086800	0.921	+	GO:0004497 GO:0005506 GO:0016705 GO:0020037 GO:0055114	Flavonoid 3'-monooxygenase / Flavonoid 3'-hydroxylase
Ceric.27G014600	0.897	+	-	-

Gene ID	kME	Corr	GO Terms	Best Hit
Ceric.25G023000	0.890	+	GO:0005507 GO:0016491 GO:0046274 GO:0048046 GO:0052716 GO:0055114	MULTI-COPPER OXIDASE // LACCASE-2
Ceric.1Z215200	0.889	+	GO:0004857	Plant invertase/pectin methylesterase inhibitor superfamily protein
Ceric.21G073600	0.883	+	-	AWPM-19-like family protein
Ceric.30G049900	0.883	+	GO:0008152 GO:0016758	Trans-zeatin O-beta-D-glucosyltransferase / Zeatin O-glucosyltransferase
Ceric.1Z036300	0.878	+	GO:0004857	Plant invertase/pectin methylesterase inhibitor superfamily protein

Gene ID	kME	Corr	GO Terms	Best Hit
Ceric.18G013800	0.873	+	GO:0005524 GO:0006810 GO:0016021 GO:0016887 GO:0042626 GO:0055085	ABC transporter family protein
Ceric.18G083600	0.872	+	-	LATE EMBRYOGENESIS ABUNDANT PROTEIN-LIKE PROTEIN
Ceric.38G044700	0.871	+	-	MYB family transcription factor, putative, expressed
Ceric.04G027200	0.859	+	-	-
Ceric.34G060200	0.859	+	-	-
Ceric.35G024800	0.851	+	-	plant-specific domain TIGR01615 family protein, expressed
Ceric.21G056300	0.844	+	GO:0008152 GO:0016746	Glycerol-3-phosphate 1-O- acyltransferase

Gene ID	kME	Corr	GO Terms	Best Hit
Ceric.36G034300	0.837	+	-	-
Ceric.24G013200	0.831	+	-	(1 of 28) PF14009 - Domain of unknown function (DUF4228) (DUF4228)
Ceric.1Z036200	0.831	+	GO:0004857	Plant invertase/pectin methylesterase inhibitor superfamily protein
Ceric.34G056700	0.831	+	GO:0005515 GO:0008270	Ring finger domain (zf-RING_2)
Ceric.04G055700	0.829	+	-	-
Ceric.21G068800	0.821	+	GO:0006629	CGI-141-RELATED/LIPASE CONTAINING PROTEIN
Ceric.17G082000	0.814	+	-	-
Ceric.03G002600	0.808	+	-	-
Ceric.05G016600	0.805	+	GO:0003677	Transcription factor HEX, contains HOX and HALZ domains

Gene ID	kME	Corr	GO Terms	Best Hit
Ceric.04G063500	0.799	+	-	CVNH domain (CVNH)
Ceric.13G018800	0.795	+	-	-
Ceric.05G006900	0.793	+	-	-
Ceric.1Z083100	0.786	+	-	-
Ceric.19G026400	0.777	+	-	PF13668 - Ferritin-like domain (Ferritin_2)
Ceric.15G075000	0.776	+	-	Protein of unknown function (DUF793)
Ceric.01G039600	0.770	+	GO:0005199	-
Ceric.08G008200	0.767	+	GO:0005515 GO:0007165	Toll-Interleukin-Resistance (TIR) domain family protein
Ceric.12G080700	0.764	+	-	-
Ceric.01G088400	0.752	+	GO:0003700 GO:0006355	AP2 domain containing protein, expressed
Ceric.28G068700	0.733	+	-	Myb transcription factor, putative, expressed

Gene ID	kME	Corr	GO Terms	Best Hit
Ceric.25G074400	0.727	+	-	-
Ceric.1Z149500	0.727	+	-	PF03945 - delta endotoxin, N-terminal domain (Endotoxin_N)
Ceric.19G079400	0.709	+	-	-
Ceric.28G018800	0.708	+	GO:0005515	histone methyltransferase
Ceric.1Z149400	0.702	+	-	-
Ceric.28G017900	0.702	+	GO:0003700 GO:0006355 GO:0043565	WRKY DNA-binding protein 50
Ceric.36G000800	0.692	+	GO:0005215 GO:0006810 GO:0016020	Aquaporin-like superfamily protein

Gene ID	kME	Corr	GO Terms	Best Hit
Ceric.31G003400	0.687	+	GO:0006855 GO:0015238 GO:0015297 GO:0016020 GO:0055085	MATE efflux family protein
Ceric.15G006100	0.684	+	-	RNA-directed DNA polymerase / Revertase
Ceric.18G012800	0.660	+	-	-
Ceric.36G006300	0.633	+	-	-
Ceric.11G019000	-0.559	-	GO:0070403	sirtuin 2: histone deacetylase
Ceric.11G047500	-0.677	-	-	-

Table D.2: GO annotation and best hits for genes in the darkorange module.
Hub genes are in bold.

Gene ID	kME	Corr	GO Terms	Best Hit
Ceric.24G019500	0.989896861 +		GO:0030001 GO:0046872	copper chaperone
Ceric.10G031700	0.989315116 +		GO:0003677 GO:0005634 GO:0006355	HOMEODOMAIN PROTEIN KNOTTED-1- LIKE 3-RELATED
Ceric.18G025500	0.988700766 +		GO:0003989 GO:0006633 GO:0009317 GO:0016874	acetyl Co-enzyme A carboxylase carboxyltransferase alpha subunit
Ceric.21G018200	0.987790564 +		GO:0006486 GO:0008378 GO:0016020 GO:0016757	Galactosylxylosylprotein 3-beta- galactosyltransferase

Gene ID	kME	Corr	GO Terms	Best Hit
Ceric.01G098200	0.987335325 +			Bifunctional inhibitor/lipid-transfer protein/seed storage 2S albumin superfamily protein
Ceric.03G048700	0.985140494 +			Bifunctional inhibitor/lipid-transfer protein/seed storage 2S albumin superfamily protein
Ceric.32G018000	0.98490737 +		GO:0009055	Cupredoxin superfamily protein
Ceric.07G064700	0.984735177 +			Bifunctional inhibitor/lipid-transfer protein/seed storage 2S albumin superfamily protein
Ceric.23G082500	0.983928835 +			DNA-binding HORMA family protein
Ceric.38G061600	0.983496674 +			GAG-POL-RELATED RETROTRANSPOSON.RNA dependent DNA polymerase
Ceric.02G080100	0.983460041 +			plant lipid transfer protein/seed storage helical domain

Gene ID	kME	Corr	GO Terms	Best Hit
Ceric.24G042000	0.982030726 +			
Ceric.17G011000	0.981189806 +			Pectate lyase superfamily protein (Pectate_lyase_3)
Ceric.06G045700	0.979857673 +			remorin
Ceric.13G063600	0.979711543 +		GO:0005375 GO:0016021 GO:0035434	copper transporter
Ceric.31G051400	0.979253426 +			
Ceric.25G013400	0.977340387 +		GO:0005506 GO:0016705 GO:0020037 GO:0055114	cytochrome P450, family 704, subfamily B, polypeptide 1
Ceric.14G094900	0.976004527 +		GO:0046983	

Gene ID	kME	Corr	GO Terms	Best Hit
Ceric.05G009300	0.973898751	+	GO:0004672 GO:0004713 GO:0005524 GO:0006468	basic helix-loop-helix (bHLH) DNA-binding superfamily protein
Ceric.32G032500	0.973130806	+		
Ceric.13G063800	0.972327003	+	GO:0005375 GO:0016021 GO:0035434	CTR type copper ion transporter
Ceric.17G012600	0.96959105	+	GO:0006826 GO:0006879 GO:0008199	Ferroxidase / Hephaestin
Ceric.34G054200	0.969120673	+		FASCICLIN-LIKE ARABINOGALACTAN FAMILY PROTEIN
Ceric.05G083000	0.96854659	+		
Ceric.05G083000	0.96854659	+		

Gene ID	kME	Corr	GO Terms	Best Hit
Ceric.1Z182200	0.966399556 +			
Ceric.28G031900	0.963472589 +		GO:0004747 GO:0006014	Fructokinase
Ceric.25G050200	0.960634943 +			
Ceric.05G023900	0.960241138 +		GO:0006694 GO:0008152 GO:0008168	sterol methyltransferase 1
Ceric.39G049400	0.959605884 +		GO:0006520 GO:0016491 GO:0016639 GO:0055114	Glutamate dehydrogenase
Ceric.04G091100	0.956025593 +			solute carrier family 50 (sugar transporter) (SLC50A, SWEET)
Ceric.23G000500	0.955472189 +			Fasciclin domain (Fasciclin
Ceric.33G045100	0.953090251 +			K17769 - mitochondrial import receptor subunit TOM22 (TOM22)

Gene ID	kME	Corr	GO Terms	Best Hit
Ceric.22G050900	0.950229839 +			ALDEHYDE DEHYDROGENASE FAMILY 2 MEMBER B7, MITOCHONDRIAL
Ceric.18G032800	0.949220903 +			
Ceric.32G072300	0.947990242 +			
Ceric.25G070600	0.945499979 +			ribulose-1,5-bisphosphate carboxylase/oxygenase small subunit 1
Ceric.38G048700	0.941272356 +		GO:0005515	
Ceric.37G016200	0.940236619 +			MA3 domain-containing protein
Ceric.25G033900	0.938993187 +		GO:0003677 GO:0005524 GO:0005634 GO:0006259 GO:0006281 GO:0007131 GO:0008094	DNA repair (Rad51) family protein

Gene ID	kME	Corr	GO Terms	Best Hit
Ceric.15G014000	0.93764677	+		
Ceric.33G041000	0.937275812	+	GO:0003824 GO:0008152	acyl-CoA synthetase 5
Ceric.38G036700	0.936316191	+	GO:0004190 GO:0006508 GO:0006629	aspartic proteinase A1
Ceric.02G078100	0.935878654	+	GO:0016021	Predicted small molecule transporter
Ceric.24G022600	0.934976783	+	GO:0003682	BAH domain
Ceric.05G057900	0.934582826	+		SWI/SNF-RELATED CHROMATIN BINDING PROTEIN
Ceric.20G069500	0.934174884	+		SMALL HEAT SHOCK PROTEIN
Ceric.33G021500	0.932344813	+		
Ceric.01G093900	0.931283077	+		Bifunctional inhibitor/lipid-transfer protein/seed storage 2S albumin superfamily protein

Gene ID	kME	Corr	GO Terms	Best Hit
Ceric.38G052300	0.929790236	+		
Ceric.1Z013600	0.928942162	+		
Ceric.22G014400	0.926877933	+		root hair specific 8
Ceric.17G069100	0.926359803	+		
Ceric.27G033900	0.924488806	+		ASCICLIN-LIKE ARABINOGALACTAN FAMILY PROTEIN
Ceric.09G008600	0.922571919	+	GO:0009058 GO:0016844	Calcium-dependent phosphotriesterase superfamily protein
Ceric.35G060700	0.916626539	+		cAMP-regulated phosphoprotein 19-related protein
Ceric.10G047200	0.91647066	+		
Ceric.27G002500	0.915571336	+		
Ceric.33G056700	0.91318269	+	GO:0005515 GO:0016020	

Gene ID	kME	Corr	GO Terms	Best Hit
Ceric.1Z207900	0.910569726	+	GO:0004497 GO:0005506 GO:0016705 GO:0020037 GO:0055114	CYTOCHROME P450 704B1
Ceric.38G022700	0.90823411	+		PTHR33513:SF4 - GB
Ceric.20G072700	0.903272552	+	GO:0008152 GO:0016787	
Ceric.23G050100	0.901043987	+		
Ceric.34G049100	0.900917277	+	GO:0003676 GO:0003677 GO:0005634 GO:0006260 GO:0006281 GO:0006310 GO:0008270	replication protein A 1A
Ceric.12G006700	0.900025538	+		

Gene ID	kME	Corr	GO Terms	Best Hit
Ceric.17G017000	0.893304889	+		APOLIPOPROTEIN D
Ceric.02G004600	0.891241145	+		myb domain protein 101
Ceric.13G071000	0.889576569	+		
Ceric.09G016100	0.88383867	+		myb domain protein 86
Ceric.12G063600	0.882407715	+	GO:0005506 GO:0016705 GO:0020037 GO:0055114	Cytochrome P450, CYP197 superfamily
Ceric.24G075400	0.881982591	+	GO:0003924 GO:0005525 GO:0005622 GO:0006886 GO:0006913 GO:0007165 GO:0007264 GO:0015031 GO:0016020	RAS-RELATED PROTEIN RABA4C

Gene ID	kME	Corr	GO Terms	Best Hit
Ceric.17G036800	0.880335255 +			Reverse transcriptase (RNA-dependent DNA polymerase)
Ceric.32G031000	0.877000395 +			
Ceric.36G002600	0.875033335 +		GO:0000228 GO:0005515	Rad21/Rec8-like family protein
Ceric.37G015600	0.874841855 +		GO:0008152 GO:0016787	
Ceric.38G021700	0.864872704 +		GO:0004601 GO:0006979 GO:0020037 GO:0055114	peroxidase precursor, putative, expressed
Ceric.35G046300	0.861020039 +			
Ceric.11G064500	0.860075462 +			
Ceric.13G077400	0.859925422 +		GO:0004725 GO:0016311 GO:0016791	Phosphotyrosine protein phosphatases superfamily protein

Gene ID	kME	Corr	GO Terms	Best Hit
Ceric.24G047000	0.856536091 +		GO:0004497 GO:0005506 GO:0016705 GO:0020037 GO:0055114	Cytochrome P450, CYP197 superfamily
Ceric.17G074200	0.854042619 +		GO:0030001 GO:0046872	copper chaperone
Ceric.20G072800	0.852415008 +		GO:0008152 GO:0016787	PEPTIDASE M20 FAMILY MEMBER // IAA-AMINO ACID HYDROLASE ILR1-LIKE 1-RELATED
Ceric.35G060200	0.850507811 +			
Ceric.25G047600	0.84965933 +		GO:0004497 GO:0005506 GO:0016705 GO:0020037 GO:0055114	cytochrome P450, family 707, subfamily A, polypeptide 3

Gene ID	kME	Corr	GO Terms	Best Hit
Ceric.15G024100	0.849438046 +		GO:0005515 GO:0043531	target of AVRB operation1
Ceric.20G059300	0.840895681 +			SMALL HEAT-SHOCK PROTEIN HSP20 FAMILY
Ceric.12G085700	0.836859616 +			
Ceric.32G028900	0.834917085 +			Flagellar Associated Protein 45
Ceric.15G049200	0.825668229 +		GO:0009765 GO:0016020	photosystem II light harvesting complex gene B1B2
Ceric.35G020400	0.821320379 +			PTHR31223:SF11 - LOG FAMILY PROTEIN YJL055W
Ceric.18G072700	0.819585583 +			kinesin-like protein, identical, putative, expressed
Ceric.24G045900	0.818200324 +			
Ceric.21G072200	0.810013722 +		GO:0008152 GO:0016491 GO:0055114	aldehyde dehydrogenase 2C4

Gene ID	kME	Corr	GO Terms	Best Hit
Ceric.15G077800	0.80219791	+	GO:0008152 GO:0016597 GO:0046983	basic helix-loop-helix (bHLH) DNA-binding superfamily protein
Ceric.35G022500	0.796457807	+	GO:0004190 GO:0006508	Eukaryotic aspartyl protease family protein
Ceric.37G060500	0.793916427	+		
Ceric.07G060300	0.788642159	+		PAR1 protein
Ceric.27G037300	0.7784326	+	GO:0009733	Auxin responsive protein (Auxin_inducible)
Ceric.18G007400	0.771607476	+	GO:0004345 GO:0006006 GO:0050661 GO:0055114	glucose-6-phosphate dehydrogenase 3
Ceric.14G044700	0.769656548	+		5-amino-6-(5-phosphoribosylamino)uracil reductase

Gene ID	kME	Corr	GO Terms	Best Hit
Ceric.27G056200	0.769487374	+	GO:0004601 GO:0006979 GO:0020037 GO:0055114	peroxidase
Ceric.1Z121900	0.766791982	+	GO:0003676	RIBONUCLEASE P SUBUNIT P25
Ceric.02G004300	0.75631725	+	GO:0005634	CYCLINE // CYCLIN-B2-1-RELATED
Ceric.22G035200	0.745331126	+		
Ceric.32G060800	0.74521052	+		
Ceric.01G130400	0.734550205	+	GO:0016020 GO:0016021 GO:0022891 GO:0055085	MFS transporter, SP family, solute carrier family 2 (myo-inositol transporter), member 13 (SLC2A13, ITR)
Ceric.21G084600	0.731451504	+		
Ceric.34G063900	0.711518364	+		Protein of unknown function (DUF288)
Ceric.37G023800	0.698084109	+		

Gene ID	kME	Corr	GO Terms	Best Hit
Ceric.17G049600	0.697522964 +			MYB family transcription factor, putative, expressed
Ceric.15G012400	0.689374028 +			
Ceric.05G079100	0.684235772 +		GO:0004672 GO:0004713 GO:0005524 GO:0006468 GO:0030246	Concanavalin A-like lectin protein kinase family protein
Ceric.01G043100	0.647227974 +			
Ceric.33G052700	0.643167092 +		GO:0005096	NSP (nuclear shuttle protein)- interacting GTPase
Ceric.05G063000	0.628444359 +		GO:0005515 GO:0008270	RING/U-box superfamily protein
Ceric.02G020200	0.625153492 +			
Ceric.15G061600	0.62002974 +		GO:0016021	TRANSMEMBRANE SUPERFAMILY PROTEIN

Gene ID	kME	Corr	GO Terms	Best Hit
Ceric.28G005900	0.61283335	+	GO:0006950	Adenine nucleotide alpha hydrolases-like superfamily protein
Ceric.07G052900	0.606622096	+		Sperm-tail PG-rich repeat (SHIPPO- rpt)
Ceric.16G075000	0.587646511	+	GO:0004672 GO:0004713 GO:0005524 GO:0006468	Ca ²⁺ /calmodulin-dependent protein kinase, EF-Hand protein superfamily
Ceric.17G043800	0.57166387	+	GO:0009405	delta endotoxin, N-terminal domain (Endotoxin_N)
Ceric.12G015000	0.502024864	+	GO:0003993	HAD superfamily, subfamily IIIB (Acid phosphatase)
Ceric.26G070400	- 0.616453743	-		
Ceric.19G034800	- 0.624415798	-		FAD NAD BINDING OXIDOREDUCTASES

Gene ID	kME	Corr	GO Terms	Best Hit
Ceric.36G022500	- 0.732664305	-		jasmonate-zim-domain protein 5