Tracing the evolution of the arbuscular mycorrhizal symbiosis in the plant lineage

Guru Vighnesh Radhakrishnan

Thesis submitted for the degree of
Doctor of Philosophy
to the
University of East Anglia

Research conducted at the John Innes Centre

September 2017

Abstract

The Arbuscular Mycorrhizal (AM) symbiosis is formed by ~80% of land plants with a specific group of soil fungi, the AM fungi. Through this symbiosis, plants obtain nutrients that they otherwise would not be able to access. Based on data from fossils and extant plants, it has been predicted that the AM symbiosis evolved in early land plants. Research in the past two decades utilising angiosperm model plant species has identified several plant genes that regulate the AM symbiosis. These studies have also revealed that these symbiosis genes are highly conserved in the angiosperms but whether this conservation extends to the non-flowering plants has not been explored. In the present study, a comprehensive phylogenetic analysis of the symbiosis genes was conducted, using genomic and transcriptomic data from the non-flowering plant lineages, to gain insights into the evolution of these genes in plants. The results of this analysis indicate that these genes evolved in a stepwise fashion. While some genes appeared in the algal ancestors of the charophytes, others appeared in the early land plant ancestors of liverworts. To further study the AM symbiosis in the non-flowering plants, key methods and genomic resources were established for the liverwort *Marchantia paleacea* to enable its use as a model plant for the study of the evolution of the AM symbiosis. Using the resources established, phylogenomic comparisons were conducted between *M. paleacea* and a related liverwort species, *M. polymorpha*, that is predicted to have lost the ability to engage in the AM symbiosis. These analyses revealed that the homologs of angiosperm symbiosis genes are required for functional symbioses in the liverworts. The findings presented here provide insights into the processes that contributed to the evolution and maintenance of this ancestral symbiosis in the plant lineage.

List of contents

List of tables

List of figures

List of Accompanying Materials

Supplementary File S1. Unannotated complete versions of the trees presented in Chapter 2 (PDF file).

Supplementary File S2. Annotated and unannotated complete versions of the trees presented in Chapter 4 (ZIP file containing PDF of annotated trees and NEWICK unannotated trees for each orthogroup).

Supplementary File S3. Results of the differential expression analysis conducted on RNA-seq data from mycorrhized and control samples of *Lunularia cruciata* (Excel file).

Supplementary File S4. List of *Medicago truncatula* genes used as queries for the sequence searches conducted for building the phylogenetic trees presented in Chapters 2 and 4 (Excel file).

Supplementary File S5. Sequences used for the construction of the phylogenetic trees presented in Chapters 2 and 4 (ZIP file containing FASTA files used for each tree).

List of Abbreviations

ABC	ATP-binding cassette
AM	Arbuscular Mycorrhizal
BiFC	Bimolecular Fluorescence Complementation
BLAST	Basic Local Alignment Search Tool
CCaMK	Calcium/Calmodulin-dependent Protein Kinase
CDPK	Calcium-Dependent Protein Kinase
FDR	False Discovery Rate
FP	Fahreus Plant
GFP	Green Fluorescent Protein
HR	Homologous recombination
LB	Lysogeny Broth
LRR	Leucine-Rich Repeat
MLD	Malectin-Like Domain
MM	Mucoromycotina
PCR	Polymerase Chain Reaction
PBS	Phosphate-Buffered Saline
PT	Phosphate Transporters
RN	Root-nodule
SRV	Strullu-Romand Variant
SYMRK	SYMBIOSIS RECEPTOR KINASE
TALEN	Transcription activator-like effector nucleases
WGA	Wheat Germ Agglutinin

Publications

During the course of the work presented in this thesis, the following manuscripts have been published, or have been planned for submission:

Delaux P-M, Radhakrishnan G, Jayaraman D, Cheema J, Malbreil M, Volkening J, Sekimoto H, et al. 2015. Algal ancestor of land plants was preadapted for symbiosis. *Proc National Acad Sci* 112(43):13390–13395.

Delaux P-M, Radhakrishnan G, and Oldroyd G. 2015. Tracing the evolutionary path to nitrogen-fixing crops. *Curr Opin Plant Biol* 26:95–99.

Charpentier M, Sun J, Martins T, Radhakrishnan G, Findlay K, Soumpourou E, Thouin J, et al. 2016. Nuclear-localised cyclic nucleotide–gated channels mediate symbiotic calcium oscillations. *Science* 352(6289):1102–1105

Luginbuehl L, Menard G, Kurup S, Erp H, Radhakrishnan G, Breakspear A, Oldroyd G, and Eastmond P. 2017. Fatty acids in arbuscular mycorrhizal fungi are synthesised by the host plant. *Science* 356(6343):1175–1178.

Radhakrishnan G, Cooke A, Cheema J, Vigneron N, Delaux P-M and Oldroyd G. (in preparation). Phylogenomics uncovers the 450-million-year-long commitment of plant genes to symbiosis.

Acknowledgements

First and foremost, I would like to thank my supervisor Giles Oldroyd, for all his support and encouragement during the course of my PhD. I am extremely grateful for the mentorship and guidance he provided and for helping me find my passion for evolutionary research. The amount of independence he provided me with to follow my interests was astonishing and I could not have asked for a better PhD project. I would also like to thank my secondary supervisor Cristobal Uauy for all his input on my project and for helping me get my priorities straight both in terms of the project as well as in terms of my professional goals.

I have been extremely fortunate to have had the opportunity to work with some amazing people who have supported me immensely during my PhD. I would like to thank Pierre-Marc Delaux for all his help, support, encouragement and discussions over the course of my PhD. My interest in evolution was inspired by his amazing evolutionary thinking. Thanks are also due to Myriam Charpentier, Jeremy Murray and Allan Downie for offering their time and expertise in symbiosis and research in general. Thanks to Christian Rogers, Eleni Soumpourou, Andy Breakspear and Andrey Korolev for keeping the lab running smoothly and for their help with DNA synthesis and GoldenGate cloning. I would also like to thank Julie Ellwood for all her help over the last 4 years and for always being the most efficient and effective person around. I would like to thank Ben Miller for his patient help and guidance when I first joined the lab with GoldenGate cloning, Leonie Luginbuehl for her help with all things mycorrhiza, Feng Feng for sharing his expertise on plant pathogens and for showing me how to get tobacco infiltrations right, Nuno Leitao for his amazing Arabidopsis knowledge, Jitender Cheema for all the bioinformatics help, Jongho Sun for help with calcium imaging and JIC Bioimaging, Grant, Eva and Kim for all their help with the microscopy. Thanks to the JIC media kitchen (particularly to Mary-Anne), horticultural services (Damian Alger) and stores (Kevin, Arnie, Roger, Kieran) staff for all their help.

I would also like to thank all my friends who have made the last 4 years in Norwich so amazing: KK, Leonie, Nuno, Aisling, Ramesha, Jodi, Kath, Matt, Jian, Feng, Giannis, Ponraj, Chengwu, Jan and Tom. You're the best, folks.

I could not have done any of the things that I've done over the last 4 years without the help of my family. Thank you Sreya for your unlimited patience, love and for accompanying me on this amazing journey. Abi, thank you for being the best sister in the whole world. Thank you, mom and dad for supporting me throughout my education.

I would like to thank the Gates Foundation for supporting Giles Oldroyd and for providing for my studentship through the ENSA grant to Giles.

1

General Introduction

Chapter 1: General Introduction

Plants have adapted to a wide array of environmental challenges throughout their evolutionary history (Beerling, 2017). These adaptations have given rise to the tremendous diversity of plant life that we find on earth today. To overcome some of the challenges they face, some species of plants have evolved the ability to form mutually beneficial partnerships, with other organisms, commonly referred to as "symbioses" (Shtark et al., 2010). The work presented here deals with the symbiosis formed by plants with a specific group of soil fungi called AM fungi. This symbiosis, termed the AM symbiosis, is formed by 80-90% of land plants and provides numerous benefits to both plant and fungal partners (Smith and Read, 2008; Ruhfel et al., 2014). In the present study, the evolution of plant genes regulating this symbiosis is explored in the context of plant evolution.

1.1. Evolution of the plant lineage

Green plants arose as a result of an endosymbiotic event caused by the uptake and integration of an autotrophic cyanobacterium into a heterotrophic eukaryote that gave rise to a novel cellular organelle called the plastid in the host eukaryote (Keeling, 2010). This ancestral eukaryotic lineage, wherein plastids evolved, gave rise to the members of the archaeplastid lineage - glaucophytes, red algae and green algae. Members of the green algal or green plant lineage are distinguished from other archaeplastidans by the storage of starch in the plastids as well as the possession of both chlorophyll a and b forms (Lewis and McCourt, 2004). Extant plant life exists both on land and in water, in a variety of habitats, possessing diverse morphological characteristics. Plants are predicted to be at least 750 million years old based on fossil data whereas molecular data-driven divergence time estimates suggest that they may be over a billion years old (Butterfield et al., 1994; Halverson et al., 2007; Butterfield, 2009; Herron et al., 2009; Parfrey et al., 2011) The currently accepted classifications (Ruhfel et al., 2014; Wickett et al., 2014) divide green plants, the *Viridiplantae*, into the *Chlorophyta* and the *Streptophyta*. The *Chlorophyta* are comprised of marine and freshwater species that are traditionally referred to as "green algae" and are classified into the prasinophytes and core chlorophytes (Leliaert et al., 2011).

The *Streptophyta* includes a few freshwater algae called the charophytes and the land plants (Karol et al., 2001). While the chlorophytes are thought to have originally evolved in a marine setting, the original habitat of charophytes is predicted to have

been freshwater environments (Delwiche and Cooper, 2015) and the divergence between these two algal lineages is predicted to have occurred approximately 700 million years ago (Falkowski and Knoll, 2011). Based on their order of divergence, the charophytes are divided into basal and advanced lineages. The paraphyletic basal charophyte lineage comprises the Chlorokybales and Mesostigmatales, both of which are monotypic lineages each containing a single species, *Chlorokybus atmophyticus* and *Mestostigma viride* respectively, as well as the Klebsormidiales which includes *Klebsormidium flaccidum*, the first charophyte whose draft genome has been published (Brodie and Lewis, 2007; Hori et al., 2014). The monophyletic advanced or higher charophyte lineage are relatively more species-rich and include the Charales, Coleochaetales and Zygnematales, with the Zygnematales being the most species rich with nearly 10,000 named species (Hall et al., 2008). Well-known species among these clades include *Spirogyra* (Zygnematales), *Coleochaetae* (Coleochaetales) and *Nitella* (Charales). A great diversity in body plans is found in the charophytes with both unicellular and multicellular forms but this diversity is restricted to the haploid (gametophyte) generation (Haig, 2010; Delwiche and Cooper, 2015). The diploid (sporophyte) generation on the other hand is the single-celled zygote which is quite similar among the different charophyte clades.

Land plants are predicted to have evolved from ancestral advanced charophytes and transitioned into the terrestrial environment about 470 million years ago (Gensel and Edwards, 2001; Kenrick et al., 2012). The land plants differ from the charophytes in their possession of diplobiontic life cycles with alternating haploid and diploid multicellular life phases (Graham, 1985). While the first cell division of the zygote in the charophytes is meiotic and gives rise to haploid spores, in the land plants, the zygote divides numerous times mitotically to produce a diploid embryo that eventually becomes the multicellular sporophyte (Graham and Wilcox, 2000). Based on this unique possession of an embryo within the green lineage, land plants are referred to as the embryophytes. Among the embryophytes, the liverworts, hornworts, and mosses together referred to as the bryophytes have haploid-dominant (gametophyte-dominant) life cycles while the lycophytes, monilophytes, gymnosperms and angiosperms possess life cycles which are diploid-dominant (sporophyte-dominant) (Niklas and Kutschera, 2010). While both generations are free-living in the case of the lycophytes and monilophytes, only the sporophyte is free-living in the case of the angiosperms and gymnosperms (Kenrick and Crane, 1997).

In the bryophytes, it is the gametophyte that is free-living. While free-living gametophytes of extant land plants are anchored to the soil by means of filamentous cells called rhizoids, free-living extant land plant sporophytes possess axial anchorage organs called roots that are predicted to have evolved at least twice in the land plants (Friedman et al., 2004; Jones and Dolan, 2012) and to have occurred in a piecemeal fashion (Kenrick and Strullu-Derrien, 2014). The bryophytes differ from the other embryophytes by their lack of vascular conducting tissue which evolved in the free-living sporophytes of the lycophytes, monilophytes, angiosperms and gymnosperms - together called the tracheophytes based on their possession of vasculature (Cantino et al., 2007). Seeds evolved in the ancestors of the gymnosperms and the angiosperms and together they form the seed plants, the spermatophyta (Linkies et al., 2010).

While the relationships between some of the above plant lineages remain unresolved, recent phylogenetic analyses (Ruhfel et al., 2014; Wickett et al., 2014) have defined the majority of the relationships between these lineages (Figure 1.1). The Zygnematales have been identified as being the closest extant algal relatives to the land plants and while the exact branching order remains contested, members of the paraphyletic bryophyte grade are predicted to be the earliest diverging among extant land plant lineages with the lycophytes, monilophytes, gymnosperms and angiosperms diverging at later points (Wickett et al., 2014).

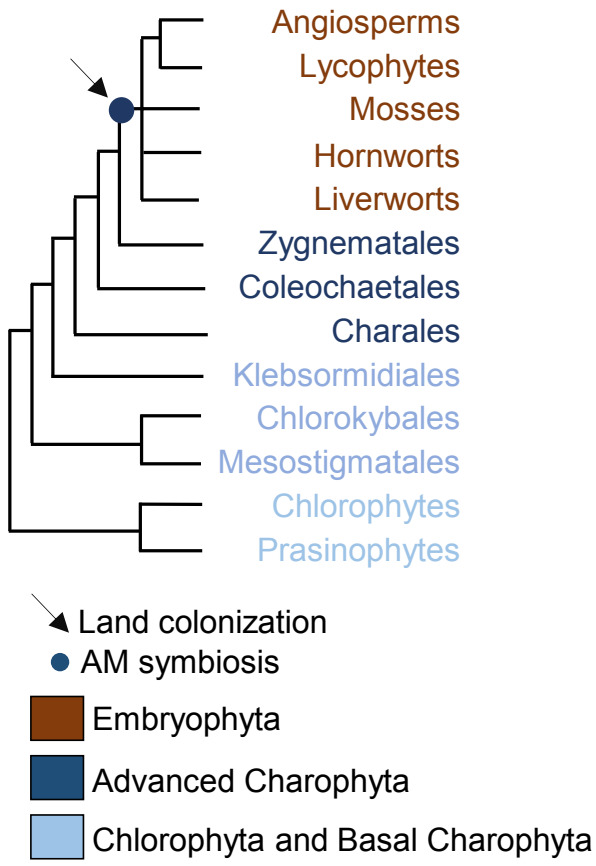

Figure 1.1. A phylogenetic tree representing the relationships among clades in the green lineage.

Phylogenetic analyses (Ruhfel et al., 2014; Wickett et al., 2014) have revealed the major relationships between the different clades in the plant lineage. The predicted evolution of AM symbiosis and land colonisation are indicated on the tree.

1.2. The Arbuscular mycorrhizal symbiosis

AM symbioses are formed by the majority (80-90%) of land plants with soil fungi belonging to the phylum Glomeromycota (Smith and Read, 2008). By forming AM symbioses, plants obtain access to growth-limiting nutrients (such as phosphate and nitrogen) and water (Smith and Smith, 2011), while AM fungi, which are obligate biotrophs, receive fixed carbon in the form of sugars(Shachar-Hill et al., 1995; Harrison, 1996; Solaimanand and Saito, 1997; Pfeffer et al., 1999; Bago et al., 2000) as well as lipids (Bravo et al., 2017; Jiang et al., 2017; Keymer et al., 2017; Luginbuehl et al., 2017) in return. By associating with AM fungi, plants obtain enhanced access to nutrients through the increased soil surface area for nutrient absorption provided by

the far-reaching AM fungal hyphal network (Finlay, 2008). The AM symbiosis also provides the plant hosts with other benefits in the form of increased resistance to disease (Liu et al., 2007; Pozo and Azcon-Aguilar, 2007)and improved tolerance to drought and salinity (Augé et al., 2015). The vast majority of discoveries relating to the AM symbiosis were made using angiosperm model plants such as *Medicago truncatula* (Huguet et al., 1995), *Lotus japonicus* (Handberg and Stougaard, 1992) and *Oryza sativa* (Nakagawa and Imaizumi-Anraku, 2015). Only recently has the occurrence of the AM symbiosis in other land plant lineages been explored.

1.2.1. The AM symbiosis in angiosperms

Studies into AM symbioses in the angiosperms have revealed that the progression of this association is under the tight regulation of the host plant (Luginbuehl and Oldroyd, 2017; MacLean et al., 2017). Imaging of plant roots inoculated with AM fungi has revealed that the process through which the plant allows fungal colonisation (or mycorrhization) of the root can be divided into distinct steps (Figure 1.2).

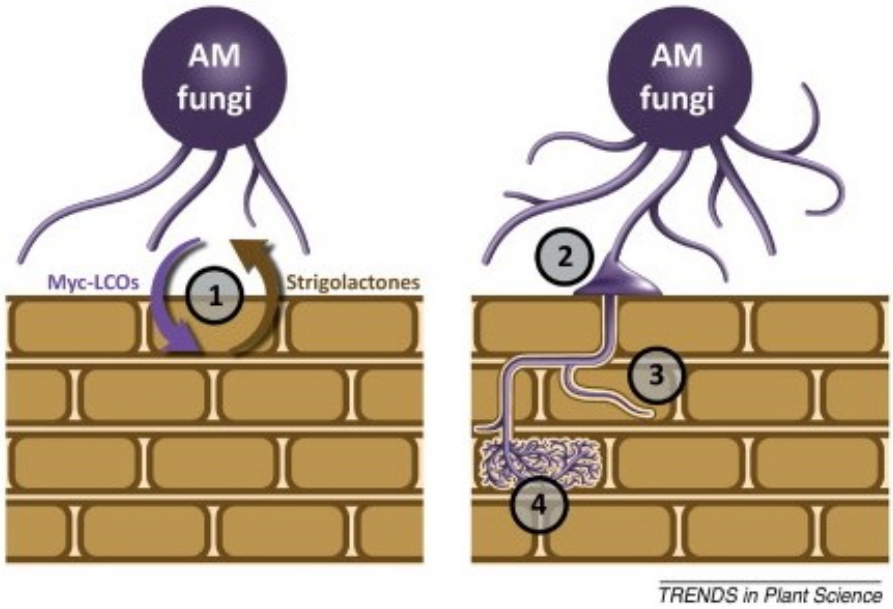

Figure 1.2. Steps in the progression of fungal colonisation during the AM symbiosis in the angiosperms.

Studies in angiosperm model plants have revealed the distinct steps through which AM fungi colonise plant roots during the AM symbiosis. These include: (1) pre-symbiotic communication between the plant and the fungus through signalling molecules; (2) attachment of fungal hyphae

to the plant epidermis and formation of hyphopodia; (3) progression of fungal colonisation through the epidermis to the root cortex; (4) formation of the symbiotic structures, arbuscules, in root cortical cells. Figure adapted and reprinted from Trends in Plant Sciences, 18/6, (Delaux et al., 2013b), Evolution of the plant-microbe symbiotic 'toolkit', 298-304, Copyright (2013), with permission from Elsevier.

The first step in the process involves communication between the plant and the fungus through the exchange of diffusible chemical signalling molecules. On the plant side, known signalling molecules include strigolactones (Besserer et al., 2006) and flavonoids (Bécard et al., 1992). Both classes of molecules have been shown to induce the germination of AM fungal spores and hyphal elongation (Bécard et al., 1992; Akiyama et al., 2005; Besserer et al., 2006), suggesting that the perception of these signals by AM fungi results in their attraction to plant roots. The fungus, on the other hand, produces chitin-based signalling molecules, called "Myc (Mycorrhization) factors", which are recognised by the plant to initiate a signalling cascade that activates a fungal accommodation programme inside the plant root (Maillet et al., 2011; Genre et al., 2013). Following this initial signal exchange, the fungal hyphae attach to the epidermis of the plant root and form a special type of appressorium called a hyphopodium, which develops from mature hyphae (Genre et al., 2005).

Following hyphopodia formation, radical rearrangement of the cytoskeletal structures and organelles of the host cell occurs (Genre et al., 2005). The host cell nucleus moves towards the site of hyphopodium attachment and travels across the cell to the opposite side. Movement of the nucleus is followed by the formation of a cellular tunnel-like structure called the pre-penetration apparatus which dictates the path followed by the fungal hyphae through the cell (Genre et al., 2005). Following this entry of the fungal hyphae through the epidermal cells into the inner cell layers, fungal colonisation may proceed in one of two ways based on the specific species of plant host and AM fungus involved - *Paris*-type or *Arum*-type (Dickson, 2004). In *Paris*- type colonisation, the hyphae proceed through the cortical cells by means of intracellular passage where they differentiate to form the sites of nutrient exchange. These can either be hyphal coils or branched tree-like structures called "arbuscules". On the other hand, *Arum*- type colonisation proceeds intercellularly between cortical cells and culminates in the penetration of an inner cortical cell to form an arbuscule. Intermediate types of

colonisation between Paris- and Arum-type have also been found to occur in some plant species (Dickson, 2004). Throughout the colonisation process, the fungal cell is kept separate from the plant cell by means of a plant-derived perifungal membrane (Parniske, 2008) (Genre and Bonfante, 2010). The interface between the arbuscules and the cortical cells is the main site of nutrient exchange between the plant host and the fungal symbiont (Luginbuehl and Oldroyd, 2017). Several plant genes that are responsible for the regulation of these distinct steps in the progression of the AM symbiosis have been identified and these are introduced in Section 1.3.

1.2.2. Fossil evidence for the AM symbiosis

The ability to form the AM symbiosis has been proposed to have evolved at least 460 million years ago based on fossil evidence from early land plants. The earliest fossils showing plants in association with fungi resembling the Glomeromycota come from exceptionally well-preserved macrofossils (Remy et al., 1994; Taylor et al., 1995), which formed through the trapping and burial of whole biotas in sediments (Seilacher et al., 1985). The macrofossils most pertinent to the evolution of the AM symbiosis come from the 407 million-year-old Rhynie chert in Scotland where fossils of early vascular plants have been discovered (Edwards, 1986). Among these are fossils attributed to *Aglaophyton major* that contain arbuscule-like structures in the inner cortical cell layers of the plant axes in the predicted sporophyte (Figure 1.3a, Figure 1.3b) (Remy et al., 1994). Later, fossils now thought to be the gametophyte of *A. major*, but at the time referred to as *Lyonophyton rhyniensis* were also found to possess similar arbuscule-like structures (Figure 1.3c, Figure 1.3d) (Taylor et al., 2005). No macrofossils exist of the earliest land plants which are predicted to have been non-vascular similar to extant bryophytes and as a result of this, it has not been possible to determine whether they associated with AM fungi (Kenrick and Crane, 1997; Renzaglia et al., 2000). The absence of macrofossil evidence of these plants has been attributed to the poor preservation of non-vascular plants in fossils owing to the softer nature of their body tissue compared to vascular plants. The earliest evidence for land plants (Figure 1.3e) and fungi some resembling the Glomeromycota come from microfossils dating to ~460-470 MYA (Redecker et al., 2000; Rubinstein et al., 2010). While there is still debate on the exact age of these Glomeromycota-like fossils owing to the possibility that fossil extraction methods used in the above study might have allowed for younger samples to contaminate older samples (Taylor et al., 2015), the proposed age of these fossils is in line with molecular data-based estimates on

when the ancestors of Glomeromycota are predicted to have evolved (Taylor and Berbee, 2006). On the plant side, the microfossil evidence comes in the form of fossilized plant spores called cryptospores. These cryptospores have been mostly found in non-marine rock sources and when discovered from marine rock sources, the abundance of cryptospores has been found to decrease with an increase in the distance from the shore (Wellman et al., 2003). This observation along with the occurence of sporopollenin, a trait conserved across the land plant lineage, in these cryptospores and their possession of anatomical similarities to spores from extant liverworts support a terrestrial origin for these cryptospores (Wellman et al., 2003). Taken together, the above observations imply co-emergence of the earliest land plants with AM fungi approximately 460 million years ago and this has been taken to suggest that the AM association may have facilitated the colonisation of land by plants (Pirozynski and Malloch, 1975; Remy et al., 1994; Brundrett, 2002).

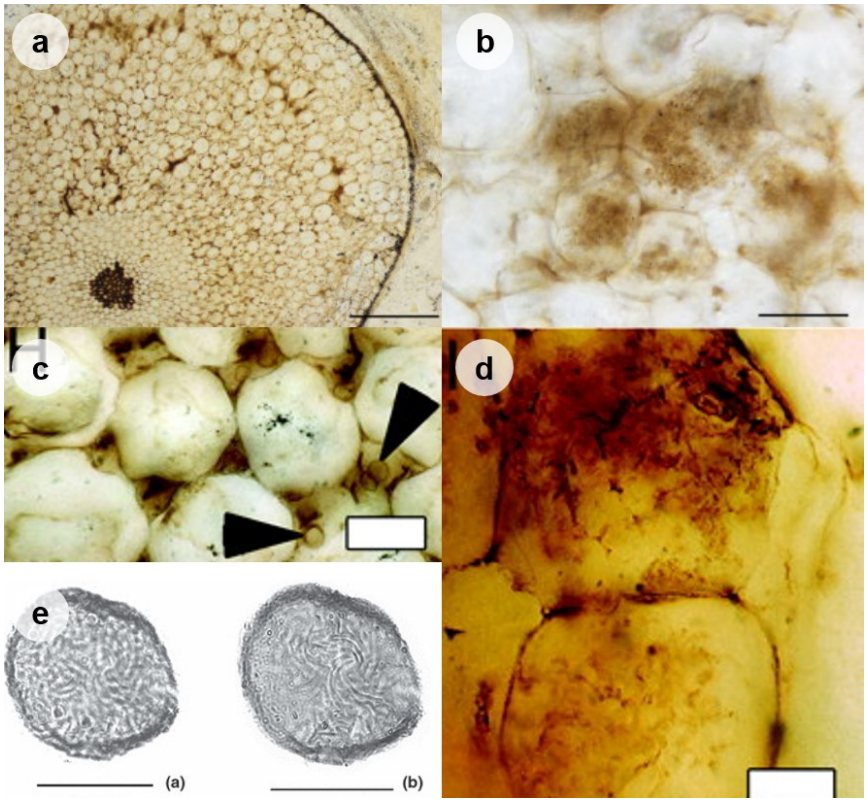

Figure 1.3. Fossil evidence for the AM symbiosis in early land plants

(a) Sections of the aerial axes of the *Aglaophyton major* sporophyte showing fungal colonisation in the cortical cells. Scale bar = 30 mm. (b) Magnified image of the arbuscule-like structures in the *A. major* sporophytes. Scale bar = 0.55 mm. (c) Sections of *Lyonophyton rhyniensis*

(gametophyte of *A. major*) exhibiting intercellular fungal colonisation. Scale bar = 50 μm. (d) Magnified image of the arbuscule-like structures found in *L. rhyniensis* (the gametophyte of *A. major*). Scale bar = 12 μm. (e) Earliest uncovered evidence for plant life on land - microfossil cryptospores from *Chomotriletes* cf. sp. dated to be at least 460 million years old. Scale bar = 20 μm. Figures (a) and (b) are reprinted with permission from (Strullu-Derrien et al., 2016) copyright 2016, John Wiley & Sons. Figures (c) and (d) are reprinted with permission from (Taylor et al., 2005) Copyright (2005) National Academy of Sciences, U.S.A. Figure (e) is reprinted with permission from (Rubinstein et al., 2010) copyright 2010, John Wiley & Sons.

1.2.3. The AM symbiosis in extant lineages of non-flowering plants

Insights into the associations between non-flowering plants and AM fungi came in the form of three major discoveries. First among these was the discovery of arbuscule-like structures in these plants demonstrated through the microscopic examination of liverworts (Ligrone et al., 2007), hornworts (Desirò et al., 2013), lycophytes (Lehnert et al., 2017), monilophytes (Pressel et al., 2016) and gymnosperms (Imhof, 1999). In the case of mosses, AM associations have been reported for the early diverging *Takakiopsida* lineage, while other later diverging lineages were predicted to have lost the ability to associate with AM fungi as no evidence for AM associations could be found for these lineages (Wang and Qiu, 2006; Field et al., 2015a). While AM associations are restricted to the sporophytes in the angiosperms and gymnosperms, they occur in both the sporophytes and gametophytes of the lycophytes and monilophytes (Brundrett, 2002). In the case of the bryophytes, the association is restricted to the gametophyte generation (Brundrett, 2002; Ligrone et al., 2007; Desirò et al., 2013). In plant species where fungal colonisation was found to occur in the sporophyte, this was found to be restricted to the roots and root colonisation by AM fungi was found to initiate at the epidermis and progress into the cortex where arbuscules or hyphal coils were formed (Brundrett, 2002; Dickson, 2004). In the bryophytes, where gametophyte colonisation by AM fungi was found to occur, this is in the main plant body with the colonisation initiating at the lower epidermis and

proceeding to the upper parenchymatous tissue where the arbuscules were formed (Ligrone et al., 2007; Desirò et al., 2013).

Following these studies into the structural features of the AM symbiosis in non-flowering plants, the advent of molecular techniques made it possible to test the identities of the fungi present in these plants using comparisons of DNA sequences that allowed the discernment of fungal species and strain identities (Redecker et al., 2003). The results of these analyses showed that the fungi colonising and forming arbuscule-like structures in non-flowering plants belonged to the same clade of Glomeromycotean fungi that associate with the angiosperms during the AM symbiosis (Russell and Bulman, 2005; Desirò et al., 2013).

The third major discovery in this area came from the analysis of the complex thalloid liverwort *Marchantia paleacea.* Humphreys and colleagues(Humphreys et al., 2010) showed that this liverwort not only interacted with AM fungi but also obtained benefits from the interaction, providing evidence for the mutualistic symbiotic nature of the associations between non-flowering plants and AM fungi. It was revealed that associating with AM fungi resulted in the host liverwort exhibiting enhanced gametophyte growth and increased asexual reproduction. These growth enhancements were found to be accompanied by increased photosynthetic carbon gain and increased uptake of soil nitrogen and phosphorus in *M. paleacea* plants harbouring AM fungi compared to those that were not in association with AM fungi. On the fungal side, evidence for the receipt of photosynthate from the plant was provided by measuring the lengths of fungal hyphae extending from the plants into the soil. Individual thalli were found to support upto 300 m of fungal mycelium. As AM fungi are obligate biotrophs incapable of obtaining carbon from the soil, it was suggested that such an extension of the mycelial network could only be possible if the fungus was obtaining carbon from the plant. This study (Humphreys et al., 2010) proved that the association formed by the liverwort *M. paleacea* with AM fungi was a mutualistic symbiosis similar to those formed by angiosperms. Following this study, similar results have also been observed using other liverworts (Field et al., 2012) and in the monilophyte *Ophioglossum vulgatum* (Field et al., 2015b)**.** These observations combined with the symbiotic status of AM fungi in the angiosperms have led to the prediction that the AM symbiosis appeared in the common ancestors of the extant land plants.

Recent studies have also identified an alternate class of fungi, belonging to the Mucoromycotina (MM), as associating with members of the early diverging lineages of liverworts, the *Haplomitriopsida* (Bidartondo et al., 2011; Desirò et al., 2013). Functional studies on the *Haplomitriopsida*-MM fungal associations, similar to those performed on the *M. paleacea*-AM fungal associations, showed that these associations were also mutualistic by demonstrating the reciprocal exchange of carbon-for-nutrients between the plant and the MM fungus (Field et al., 2015c, 2016). The occurrence of symbiotic fungi in the basal lineages of liverworts belonging to a class other than the Glomeromycota has challenged the widely-held view that the ancestral plant-fungal partnership was formed with the Glomeromycota (Field et al., 2015a). Recent phylogenetic analyses of fungi have placed the Glomeromycota and Mucoromycotina as sister lineages (Spatafora et al., 2016; Tang et al., 2016; Spatafora et al., 2017). The identity of the ancestral fungi that formed symbiosis with early land plants remains contested with the possibility that these ancestral fungi could have been Glomeromycota or Mucoromycotina or fungi from a group ancestral to these lineages (Field et al., 2015a). Also unclear is whether the level of control that host plants have over MM symbioses is similar to that observed in plants engaging in the AM symbiosis.

1.2.4. The AM symbiosis in the context of land plant evolution

Colonisation of land was a key event in the evolution of plants. Plant transition from an aqueous to a terrestrial environment, had far-reaching implications for the earth's climate and provided the foundation for the evolution of extant terrestrial ecosystems (Lenton et al., 2012; Beerling, 2017). The move to land from water presented a unique set of challenges for the first land plants: access to dissolved nutrients was much more limited, exposure to harmful UV-radiation and extreme changes in temperature, normally buffered in water (Kenrick and Crane, 1997). The terrestrial ecosystems that these early land plants faced harboured microbes such as bacteria and fungi and a beneficial interaction with fungi, such as that formed during the AM symbiosis, has been proposed to have helped early land plants obtain access to nutrients and successfully colonise land (Pirozynski and Malloch, 1975). As these early land plants lacked the extensive root systems possessed by the vascular plants, it has been proposed that the functions of modern rooting systems were shared between the root-

hair like rhizoids and the mycorrhizal associations that these plants formed (Brundrett, 2002; Jones and Dolan, 2012; Kenrick and Strullu-Derrien, 2014).

1.3. The genetic pathway controlling the AM symbiosis

Research over the last two decades into the AM symbiosis in the angiosperms has identified numerous plant genes that are involved in regulating the different processes occurring in the plant root leading to a functional symbiosis. These discoveries were largely facilitated by resources developed for the study of the AM symbiosis in angiosperm models. Some of the important resources were: (i) the ability to reliably establish the AM symbiosis *in vitro* (Fortin et al., 2005); (ii) visualisation methods to observe the progression of the AM fungus in the plant root (Genre et al., 2005); (iii) methods for the generation of plant mutants through forward and reverse genetics (Marsh et al., 2001; Watts-Williams and Cavagnaro, 2015) and (iv) the availability of genetic and genomic resources for the identification and functional characterisation of genes from these plants (Jackson et al., 2009). Advances in molecular biology in concert with the development of methods to introduce transgenes transiently and quickly into the roots of the legumes *M. truncatula* and *L. japonicus* using *Agrobacterium rhizogenes*, made these plants the models of choice for the characterisation of genes with functions in the AM symbiosis (Stiller et al., 1997; Boisson-Dernier et al., 2001).

Establishment of the AM symbiosis is governed by genes that can be broadly categorised into those possessing functions in the early signalling process and those that function downstream of this symbiotic signalling process (Figure 1.4). Several of the signalling genes were originally identified due to their roles in another symbiotic association called root-nodule (RN) symbiosis formed by legumes with a specific group of soil bacteria called rhizobia (Oldroyd and Downie, 2004). The RN symbiosis is estimated to have evolved ~60 million years ago (Sprent, 2007), much later than the AM symbiosis (~470 million years ago) (Kenrick et al., 2012; Edwards et al., 2014). As a result of the relative evolutionary novelty of the RN symbiosis, it has been proposed that genes with shared functions in the RN and AM symbioses, had functions only in the AM symbiosis in ancestral angiosperm species and were recruited into the RN symbiosis in the legumes (Parniske, 2008).

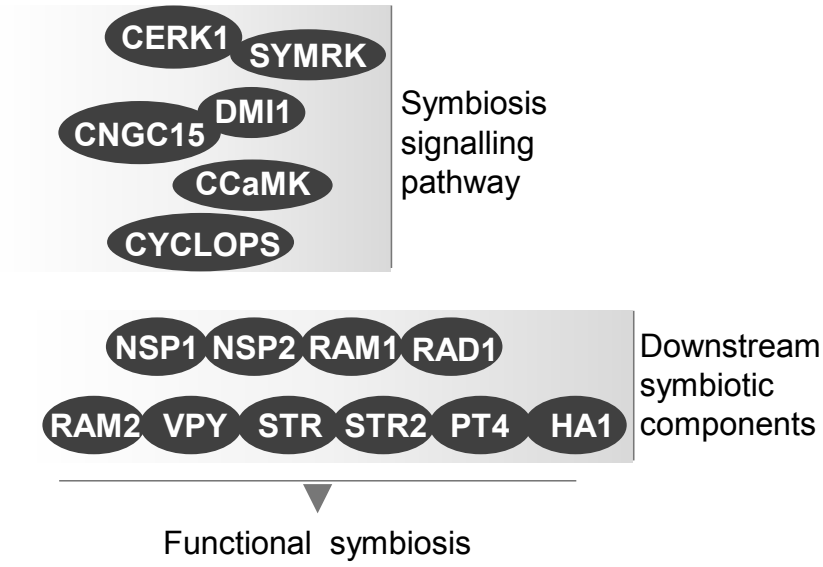

Figure 1.4. Genes identified as having functions in the regulation of the AM symbiosis in angiosperms.

Symbiosis genes can be broadly categorised into those functioning during the signalling processes and those that function in processes downstream of signalling (Oldroyd, 2013). The downstream symbiotic components have been found to have roles in regulating infection, arbuscule development and nutrient transfer that are required for a functional the AM symbiosis (Luginbuehl and Oldroyd, 2017; MacLean et al., 2017).

1.3.1. The symbiosis signalling pathway

The process of communication between plants and AM fungi is mediated by a signalling cascade consisting of plant proteins that recognise fungal signals and initiate the host plant response to the fungus in the form of gene expression changes(Oldroyd, 2013; Luginbuehl and Oldroyd, 2017; MacLean et al., 2017) (Figure 1.5). The first step in this process is the recognition of the chitin-based signalling molecules called "Myc factors" by a proposed pattern-recognition receptor-complex consisting of at least CERK1 (CHITIN ELICITIOR RECEPTOR KINASE 1) and DMI2 (DOES NOT MAKE INFECTIONS 2)/SYMRK (SYMBIOSIS RECEPTOR KINASE) (Section 1.3.1.1 & 1.3.1.2) that are present in the plant plasma membrane. This receptor-complex transduces the signal through the cytoplasm by means of an as yet-unknown secondary messenger to the outer nuclear membrane where ion-channel proteins DMI1 (DOES NOT MAKE INFECTIONS 1) and CNGC15 (CYCLIC NUCLEOTIDE GATED CHANNEL 15) subunits are present.

Upon activation, these ion-channel proteins (Section 1.3.1.3) encode a specific oscillatory calcium signature, called calcium spiking, that is decoded by a nuclear calcium-sensing protein DMI3 (DOES NOT MAKE INFECTIONS 3)/CCaMK (CALCIUM/CALMODULIN-DEPENDENT PROTEIN KINASE) (Section 1.3.1.4). CCaMK interacts with a transcription factor IPD3 (INTERACTING PROTEIN OF DMI3)/CYCLOPS (Section 1.3.1.5) that induces the expression of downstream symbiosis genes (1.3.2). It has been found that treatment with spore exudates from AM fungi alone is enough to activate this pathway, trigger calcium spiking and cause many of the gene expression changes observed during AM fungal colonisation (Maillet et al., 2011; Genre et al., 2013; Sun et al., 2015).

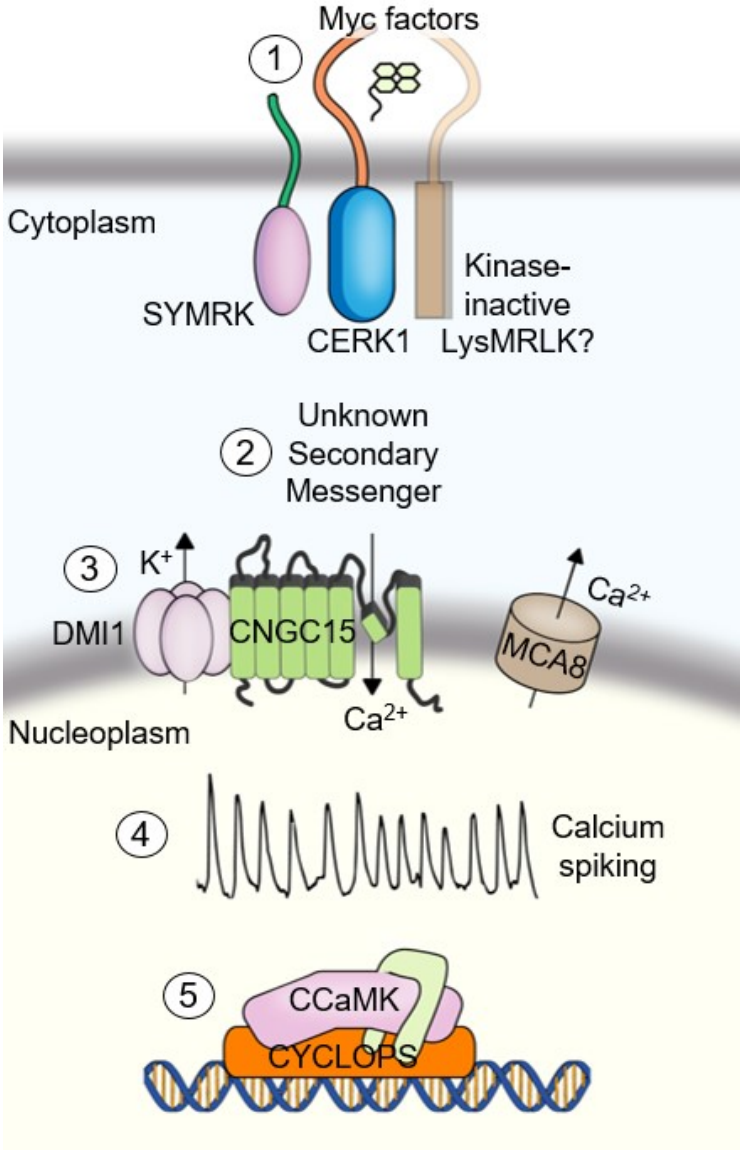

Figure 1.5. The symbiosis signalling pathway.

Initiation of plant response to AM fungi is mediated by signalling by components of the symbiosis signalling pathway (Oldroyd, 2013). The first step in this signalling process is mediated by a pattern-recognition complex at the plant plasma membrane. The kinase-active Receptor-Like Kinases (RLKs) CERK1 and SYMRK have been identified as being involved in this process. A kinase-inactive RLK has been predicted to be a part of this complex based on the composition of similar receptor-complexes required for the perception of pathogenic fungi and symbiotic bacteria. Following perception of fungal signalling molecules at the plasma membrane, the signal is transduced to the nuclear envelope by an as yet unidentified secondary messenger. Here, through the action of the ion-channel proteins DMI1 and CNGC15a, b, c, and the calcium ATPase MCA8, calcium oscillations are produced. These oscillations are then decoded by the calcium-binding protein CCaMK and through the transcription factor activity of its interacting partner CYCLOPS, activation of downstream symbiosis genes is achieved.

1.3.1.1. Symbiotic signal perception at the plant plasma membrane

Originally identified as a receptor for the recognition of pathogenic fungi (Miya et al., 2007), CERK1, a plasma membrane localised LysM-type Receptor-Like Kinase (LysMRLK) protein, has recently been shown to have a role in the recognition of AM fungal signals in *O. sativa* (Miyata et al., 2014; Zhang et al., 2015). This discovery was informed by the phenotype of knockout and RNAi lines of the *CERK1* gene which were found to be inhibited in AM fungal colonisation. Previous analyses have shown that chitin tetramers are among the symbiotic signalling molecules produced by AM fungi that are recognised by the plant (Genre et al., 2013). A role for *CERK1* in perceiving these symbiotic signals was shown through comparative analysis of the response of wildtype and *cerk1* mutant rice plants to AM fungal spore exudates as well as chitin tetramers (Carotenuto et al., 2017). In these experiments, calcium spiking was used as an indicator for the successful perception of the signals and was measured using a fluorescent calcium sensor. Wildtype rice plants exhibited calcium spiking signatures reminiscent of those observed during fungal colonisation, while

cerk1 mutant plants did not show any calcium spiking response. Analysis of mutants in *L. japonicus* and *M. truncatula* homologs of the *CERK1* gene, *NFR1* and *LYK3* revealed similar reductions in AM fungal colonisation and calcium spiking in these mutant plants compared to their respective wildtype counterparts (Zhang et al., 2015). While some LysMRLKs have been shown to possess kinases that have no phosphorylation ability (inactive kinases), CERK1 possesses an active kinase version with the ability to phosphorylate other proteins (Miya et al., 2007). Another Receptor-Like Kinase protein, SYMRK, also referred to as DMI2, has been found to be required for the perception of AM fungi. Cloned simultaneously in the legumes *M. truncatula* (Endre et al., 2002), *M. sativa* (Endre et al., 2002), *Pisum sativum* (Schneider et al., 1999) and *L. japonicus* (Stracke et al., 2002), the *SYMRK* gene, encodes a Receptor-Like Kinase protein composed of Leucine-Rich Repeat (LRR) regions and a Malectin-Like Domain (MLD). Mutations in this gene affect the ability of plants to form both AM and RN symbiosis, while overexpression of SYMRK leads to the induction of genes related to both AM and RN symbioses (Ried et al., 2014). As calcium spiking responses to rhizobial and mycorrhizal signalling molecules were abolished in the *dmi2/symrk* mutants, this gene was placed upstream of calcium spiking (Miwa et al., 2006; Kosuta et al., 2008). Although the exact function of this protein during the AM symbiosis remains elusive, based on its plasma membrane localisation and possession of both intracellular and extracellular domains, SYMRK has been proposed to be a part of a plasma membrane-localised receptor-complex required for the recognition of Myc factors, a process in which CERK1 was recently shown to be involved (Carotenuto et al., 2017). Although it remains to be tested whether CERK1 and SYMRK are indeed part of a receptor-complex, the occurrence of such a complex has been predicted based on the existence of a similar complex formed by SYMRK with other LysMRLK proteins for the recognition of signalling molecules produced by rhizobia during the RN symbiosis in *L. japonicus* (Madsen et al., 2011; Antolin-Llovera et al., 2014). A kinase-active and kinase-inactive LysMRLK protein pair along with SYMRK have been shown to be part of the symbiotic signal perception-complex. As *CERK1* is kinase-active, it has been suggested that a kinase-inactive LysMRLK protein may also be a part of the Myc-factor recognition complex but such a protein remains to be identified (Oldroyd, 2013).

1.3.1.2. Generation of oscillatory nuclear calcium signatures

Following the perception of AM fungi at the plant plasma membrane, signal transduction through the cytoplasm occurs via an as yet unidentified mechanism to the plant nucleus where calcium oscillations in the nuclear envelope and the nucleoplasm occur (Capoen et al., 2011). In *Medicago truncatula*, the potassium-permeable ion-channel protein DMI1 (Ané et al., 2004), the calcium-dependent adenosine triphosphatase MCA8 (Capoen et al., 2011) and three calcium-permeable channel proteins CNGC15a, 15b and 15c (Charpentier et al., 2016) have been found to be required for the generation of these calcium signals and are localised to the nuclear envelope. A model for the generation of symbiotic calcium oscillations by these proteins has been proposed wherein the interplay between these proteins is required for sustained calcium oscillations to occur (Granqvist et al., 2012; Charpentier et al., 2013). The model proposes that while the CNGC15s are responsible for the transport of Ca^{2+} into the nucleoplasm, simultaneous K^+ transport by DMI1 in the opposite direction counterbalances the increase in positive charges resulting from the transport of Ca^{2+}. For this simultaneous activation of DMI1 and the CNGC15s to occur, the model predicted that these proteins would need to physically interact. Replenishment of the Ca^{2+} into the nucleoplasm was proposed to be carried out by the calcium pump activity of MCA8. Evidence of physical interactions between DMI1 and the CNGC15s was found through yeast-two hybrid assays and Bimolecular Fluorescence Complementation (BiFC) studies in *Nicotiana benthamiana* leaves and *M. truncatula* hairy roots (Charpentier et al., 2016). While the *M. truncatula* mutant in *DMI1* was isolated from a forward genetic screen for its initial role in RN symbiosis and later found to have functions in the AM symbiosis, MCA8 and CNGC15s were found through targeted bioinformatic searches for calcium pumps and channels with nuclear localisation signals (Capoen et al., 2011; Charpentier et al., 2016). Following their identification through these searches, functions for these genes were confirmed using gene silencing. While mutants in MCA8 have not yet been isolated, mutants in the three *M. truncatula* CNGC15 genes showed reduced fungal colonisation when compared to wildtype plants. In *L. japonicus*, two ion channels homologous to *DMI1*, *CASTOR* and *POLLUX* have been found to be K^+-permeable and it has been found that both of these are required to perform the equivalent function of *DMI1* during the mycorrhizal symbiosis (Imaizumi-Anraku et al., 2005; Charpentier et al., 2008).

1.3.1.3. Calcium signature decoding and activation of downstream symbiosis genes

The oscillatory nuclear calcium signatures generated by the perception of Myc factors and AM fungi have been proposed to be perceived by a nuclear-localised protein called CCaMK, also referred to as DMI3 in *M. truncatula* (Shimoda et al., 2012; Miller et al., 2013). This gene was identified in *M. truncatula* from a forward screen for mutants blocked in their ability to form the RN symbiosis (Mitra et al., 2004) and its role in the AM symbiosis was elucidated later (Lévy et al., 2004). As calcium spiking responses to rhizobial and mycorrhizal signalling molecules in *ccamk* mutant plants were similar to wildtype plants, a role for *CCaMK* in signalling downstream of calcium spiking was proposed. Analysis of the domain structure of CCaMK revealed the presence of a protein kinase at the amino-terminus followed by a calmodulin-binding domain and a visinin-like domain made of four EF-hands(Patil et al., 1995). This structure implied that CCaMK could bind calcium directly through the EF-hands and indirectly through associations with calmodulin. Recent studies have shown that CCaMK is indeed dually regulated by calcium through the calmodulin-binding and visinin-like domains as predicted by the analysis of its domain structure (Shimoda et al., 2012; Miller et al., 2013). Gain-of-function mutations in the kinase domain of CCaMK have been shown to induce gene expression and physiological changes reminiscent of fungal colonisation during the AM symbiosis even in the absence of fungi or fungal signals (Takeda et al., 2012). As a result of these observations, CCaMK was proposed to be the protein translating the oscillatory calcium signatures into downstream signalling responses by means of its calcium-binding and kinase domains respectively. The activation of downstream symbiotic responses in plants possessing auto-active versions of CCaMK was found to be independent of the genes upstream of calcium spiking and therefore led to the proposal that the main objective of generating calcium spiking is the activation of CCaMK (Hayashi et al., 2010; Madsen et al., 2010). Following the activation of CCaMK, the induction of downstream symbiosis genes is facilitated by CYCLOPS, an interacting protein of CCaMK. CYCLOPS, also known as IPD3 in *M. truncatula*, was identified through screening for interacting proteins of *M. truncatula* CCaMK (Messinese et al., 2007) and through forward genetics in *L. japonicus* (Yano et al., 2008). The interaction between *M. truncatula* CCaMK and CYCLOPS was confirmed in *N. benthamiana* leaves using BiFC analysis (Messinese et al., 2007). The same was later confirmed

for the *L. japonicus* orthologs of these proteins using yeast-two hybrid assays and BiFC analysis in *N. benthamiana* leaves (Yano et al., 2008). In this study using *L. japonicus*, it was also found that CYCLOPS was phosphorylated *in vitro* by CCaMK. Mutant analyses in *L. japonicus* and *M. truncatula* revealed that CYCLOPS is essential for the progression of both AM and RN symbioses, but was not required for the generation of symbiotic calcium signatures, similar to CCaMK (Horvath et al., 2011). Characterisation of mutants in *CYCLOPS* using both *L. japonicus* and *M. truncatula* revealed the impairment of fungal colonisation in the mutants compared to wildtype (Yano et al., 2008; Horvath et al., 2011). Although sequence analysis of CYCLOPS proteins revealed the presence of a nuclear localisation signal and the presence of a coiled-coil domain (Messinese et al., 2007), for a long time after the identification of CYCLOPS, its role in symbiosis remained enigmatic. This was mainly due to the lack of functionally characterised proteins with sequence similarity to CYCLOPS. However, a recent study by Singh and colleagues (Singh et al., 2014) has shown that CYCLOPS acts as a transcription factor and induces the expression of *NIN (NODULE INCEPTION),* a gene known for its role in in RN symbiosis (Vernié et al., 2015), by binding to a region in its promoter. The study also located two serine residues in CYCLOPS that were phosphorylated by CCaMK and showed that these residues were important for the symbiotic function of CYCLOPS (Singh et al., 2014). Following this discovery, Pimprikar and colleagues (Pimprikar et al., 2016) showed that CYCLOPS also induces the expression of *RAM1* (*REQUIRED FOR ARBUSCULAR MYCORRHIZATION 1*), a gene shown to have roles in the AM symbiosis, cementing the role of CYCLOPS as a transcription factor that acts as the bridge between calcium-oscillation perception by CCaMK and downstream symbiotic gene induction responses.

1.3.2. Downstream symbiosis genes

Downstream of the symbiotic signalling cascade, several genes have been identified as being associated with the processes required for the formation of functional the AM symbiosis (Luginbuehl and Oldroyd, 2017; MacLean et al., 2017) (Figure 1.6). These include genes encoding for the GRAS-domain proteins RAM1, RAD1, NSP1 and NSP2, the glycerol-3-phosphate acyl transferase RAM2, the half-size ATP-binding cassette (ABC) transporters STR and STR2, the phosphate transporter PT4, the proton ATPase HA1 and a plant-specific protein VPY (VAPYRIN). While RAM2, STR, STR2, PT4 and HA1 have been implicated as having roles in the transfer of nutrients

between the symbiotic partners, RAD1 and VPY have been shown to function in regulating the progression of fungal colonisation in the plant root(Luginbuehl and Oldroyd, 2016; MacLean et al., 2017). On the other hand, the transcription factor RAM1 regulates both these processes through genes that are predicted to be under its transcriptional control. NSP1 and NSP2 were originally identified for their roles in activating transcriptional responses downstream of calcium spiking during RN symbiosis and were thought to be performing similar roles during the AM symbiosis (Oldroyd, 2013). Recent analyses have suggested that these two genes have roles in regulating the biosynthesis of strigolactones, which are recognised by AM fungi (Liu et al., 2011b).

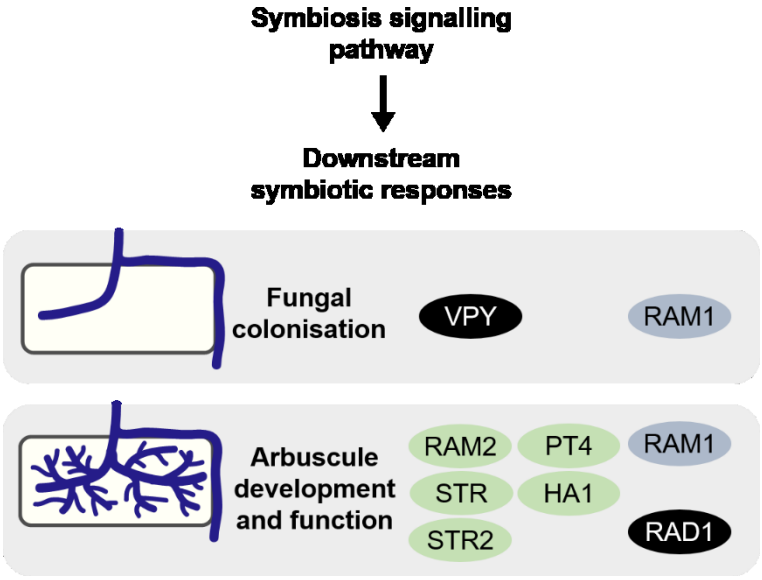

Figure 1.6. Genes controlling processes downstream of the symbiosis signalling pathway.

Following the activation of the symbiosis signalling pathway, downstream symbiotic processes required for successful fungal colonisation, arbuscule development and nutrient exchange are activated (Luginbuehl and Oldroyd, 2017; MacLean et al., 2017). The GRAS-domain transcription factor RAM1, a target of CYCLOPS, has been identified as regulating all three processes. RAD1, also a GRAS-domain protein, interacts with RAM1 and has been shown to be important for arbuscule development. The target genes of RAM1 - RAM2, STR and STR2 were shown to be required for the transfer of lipids from the plant to the fungus, while PT4 and HA1 function in the uptake of phosphate by the plant from the fungus. VPY, a plant-specific protein with domains

important for the mediation of protein-protein interactions, has been shown to have roles in the fungal infection process and is independent of RAM1 activity.

The only known target of CYCLOPS with functions in the AM symbiosis is RAM1 (REQUIRED FOR ARBUSCULAR MYCORRHIZATION 1) a GRAS-domain-containing transcription factor protein (Pimprikar et al., 2016). A role for RAM1 in the AM symbiosis was elucidated through the analysis of a mutant isolated from a forward genetic screen for reduced mycorrhizal colonisation in *M. truncatula* (Gobbato et al., 2012). Mutants in *RAM1* were found to possess defective arbuscules and imaging of these mutant plants suggested that arbuscule formation was halted prematurely. Characterisation of *ram1* mutant plants in *L. japonicus* by Pimprikar and colleagues led to a similar conclusion (Pimprikar et al., 2016). In this study, it was also observed that in *cyclops* mutants, overexpression of *RAM1* was sufficient to restore fully developed arbuscules in these plants. Based on this observation, it has been suggested that the key function of CYCLOPS may be to activate *RAM1* expression thereby leading to the initiation of downstream processes vital for arbuscule development. Adding support to this hypothesis, several genes that have been shown to be involved in regulating the establishment of the periarbuscular membrane were also found to be regulated by RAM1 (Pimprikar et al., 2016; Luginbuehl et al., 2017). Among these genes is another GRAS-domain transcription factor RAD1 (REQUIRED FOR ARBUSCULE DEVELOPMENT 1) identified recently as having a role in the regulation of arbuscule development. *L. japonicus rad1* mutant plants were found to exhibit reduced mycorrhizal colonisation and premature degeneration arbuscules compared to wildtype plants (Xue et al., 2015).

Genes required for successful nutrient exchange between the symbiotic partners to occur have also been found to be under the control of RAM1 (Pimprikar et al., 2016; Luginbuehl et al., 2017). These include *RAM2 (REQUIRED FOR ARBUSCULAR MYCORRHIZATION 1)*, *STR (STUNTED ARBUSCULE)* and *STR2 (STUNTED ARBUSCULE 2)*. *RAM2* was originally identified in the same mutant screen from which *RAM1* was identified (Wang et al., 2012). Similar to *ram1*, *M. truncatula ram2* mutant plants had reduced mycorrhizal colonisation compared to wildtype plants. RAM2 belongs to a family of glycerol-3-phosphate acyl transferases which are involved in the production of the lipid polymers cutin and suberin (Li et al., 2007;

Wang et al., 2012). Initial studies into the role of RAM2 in the AM symbiosis suggested that RAM2 might be involved in the production of lipid signalling molecules that were perceived by the AM fungus (Wang et al., 2012), but it has recently been found that RAM2 is part of a lipid biosynthetic pathway that produces lipids that are transferred to AM fungi (Bravo et al., 2017; Jiang et al., 2017; Keymer et al., 2017; Luginbuehl et al., 2017). Whether the lipids transferred to AM fungi solely fulfil a nutritional role or in addition also play a role as signal molecules is unclear. The half-size ABC type G (ABCG) transporters STR/STR2 have been implicated in the lipid transfer process as they have been suggested to be part of the machinery that transports the lipids synthesised in the plant cell to the AM fungus (Zhang et al., 2010; Bravo et al., 2017; Jiang et al., 2017).

On the flip side of this transaction, the main nutrient that plants receive from AM fungi during symbiosis is phosphate. Mutant analysis of the *PT4 (PHOSPHATE TRANSPORTER 4)* gene in *M. truncatula* (Harrison et al., 2002) revealed that it is required and responsible for the uptake of phosphate by *M. truncatula* plants from AM fungi during symbiosis. In addition, it was also observed that the arbuscules in these mutant plants degenerated prematurely and that the symbiosis was terminated in the absence of phosphate transported to the plant. Based on these observations, it was suggested that the plant may have mechanisms in place for the measurement of phosphate transported by the fungus and that in the absence of phosphate transfer by the fungus, the plant terminates the symbiosis. A role in nutrient transfer during the AM symbiosis was also revealed recently for the proton ATPase gene *HA1 (H^+ ATPASE 1)*, originally identified based on its expression specifically in arbuscule-containing cells (Wang et al., 2014). It was found that *M. truncatula* mutants in this gene had defective arbuscules as well as impaired phosphate uptake. During symbiosis, transportation of nutrients across the periarbuscular membrane is required. Based on the mycorrhizal phenotype of *ha1*, energy for the transport of nutrients was proposed to be provided through the generation of an electrochemical gradient by the proton transport action of HA1. The ability of HA1 to generate this gradient was shown through comparative measurements of the plasma membrane potential in epidermal cells from *M. truncatula* wildtype roots and transgenic roots overexpressing this protein.

A role for VPY has been proposed in the regulation of infection as well as arbuscule development processes. Analysis of *vpy M. truncatula* mutant plants have shown that

VPY is essential for the entry of AM fungi through the epidermis (Pumplin et al., 2010). Analysis of *VPY* mutant and knockdown plants revealed that in these plants, AM fungi only rarely entered the root and that fungal entry was stopped at the level of hyphopodia formation in the majority of these cases. Where fungi were seen to enter the root and infection observed to proceed to the cortical cells, no arbuscules could be found. Domain structure analysis of VPY found that this protein contains two domains that are known to be involved in mediating protein-protein interactions- an N-terminal major sperm protein domain and C-terminal Ankyrin repeats. The exact mechanism through which VPY regulates the infection and arbuscule development processes and the identity of any proteins that it may interact with are yet to be studied.

1.4. Evolution of symbiosis genes

Extant lineages of plants are the principal resources through which an understanding of the evolution of the AM symbiosis can be gained. Therefore, comparative phylogenetic analyses of the genes shown to be involved in regulating the AM symbiosis in extant land plants are required. Some studies exploring the evolution of these genes have already been conducted and these have used both experimental as well as phylogenetic approaches to study the conservation of symbiosis genes in the angiosperms (Delaux et al., 2013b). On the experimental side, studies have explored the conservation of these genes by studying the function of orthologous genes from other angiosperms. *O. sativa* mutants in the respective orthologs of *CASTOR, POLLUX, CCaMK, CYCLOPS, STR* and *STR2* have been found to be impaired in their ability to associate with AM fungi (Chen et al., 2007; Banba et al., 2008; Chen et al., 2008; Gutjahr et al., 2008; Gutjahr et al., 2012). *Petunia hybrida* mutants in *VPY* and *RAM1* have also been found to have defects in the AM symbiosis (Feddermann et al., 2010; Rich et al., 2015). The ability of these orthologs to rescue symbiosis defects in the respective mutants of their legume counterparts has also been studied. It was found that the *O. sativa SYMRK*, *CASTOR*, *POLLUX*, *CCaMK* and *CYCLOPS* genes were able to successfully rescue the mycorrhizal defects in the respective legume mutants (Godfroy et al., 2006; Chen et al., 2008; Markmann et al., 2008; Yano et al., 2008; Chen et al., 2009). Based on these results, it was concluded that the symbiotic function of these genes predated the divergence of monocots and dicots and was likely conserved across the angiosperm lineage (Parniske, 2008). Phylogenetic analyses of these genes in the angiosperms further supported this hypothesis as it was found that

these genes are highly conserved in most angiosperms (Delaux et al., 2014; Bravo et al., 2016). A summary of the findings from these studies is presented in Figure 1.7.

Orthologs of three symbiosis genes, *DMI1*, *CCaMK* and *CYCLOPS*, have also been found in members of the non-flowering plant lineages (Wang et al., 2010). The search for these orthologs was conducted using degenerate primers designed against the angiosperm versions of these genes. Using these primers, PCR and RT-PCR was conducted on DNA and RNA extracted from liverwort, hornwort, moss, lycophyte, monilophytes and gymnosperm species as genomes and transcriptomes for these species were not available at the time. From these searches, orthologs for all three genes were found in the liverworts, hornworts, mosses, lycophytes and gymnosperms, while orthologs of *CCaMK* were found in the monilophytes (Figure 1.7).

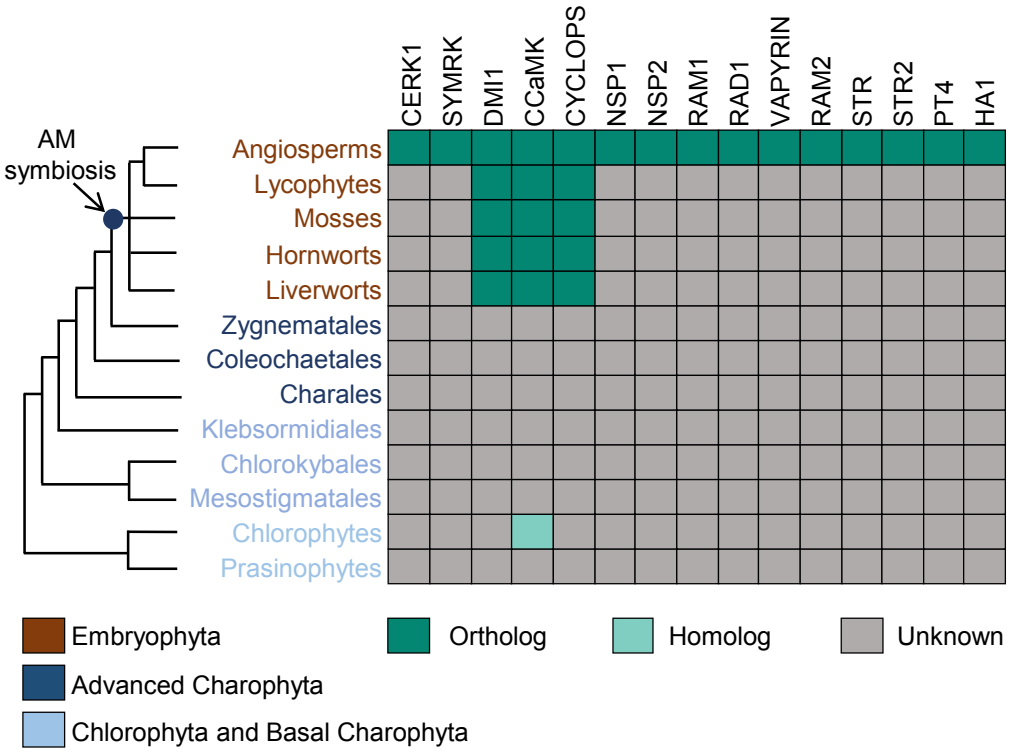

Figure 1.7. Summary of results from previous studies into the evolution of symbiosis genes in different clades of plants.

The symbiotic toolkit is highly conserved within the angiosperms (Delaux et al. 2013b). Orthologs of a few genes have been found in other land plants (Wang et al., 2010). Phylogenetic analysis of much of the symbiotic toolkit has not been carried out to-date. Homologs of *CCaMK*, belonging to a related class of proteins called CDPKs (Calcium-Dependent Protein

Kinases) have previously been found in the chlorophytes (Wang et al., 2010).

Conservation of symbiotic function in the *CCaMK* orthologs was explored by testing whether these genes could rescue the symbiotic defects in the *M. truncatula* ccamk mutant by transgenically introducing these genes. From these experiments, it was found that both liverwort and hornwort *CCaMKs* could functionally rescue the *M. truncatula* mutant, suggesting that the symbiotic function of *CCaMK* predated the evolution of angiosperms and was perhaps conserved throughout the land plant lineage. As the occurrence of symbiosis gene orthologs in the charophyte and chlorophyte green algae could not be ascertained or ruled out, it was not possible at the time to ascertain whether symbiosis genes evolved specifically in land plants or whether their evolution predated that of the land plants.

1.5. The aims of the current study

The AM symbiosis is conserved across the land plant lineage but studies into the genetic components regulating the AM symbiosis have mostly focussed on angiosperms. From these studies, several genes that regulate the ability of plants to associate with AM fungi have been identified.

In Chapter 2 of the present study, I have reconstructed the evolutionary history of these symbiosis genes to understand when they first appeared in plants and the processes that contributed to their evolution. From these phylogenetic analyses, I found that the evolution of symbiosis genes in the plant lineage likely occurred in stages. The first of these stages is predicted to have occurred in the algal ancestors of the charophytes where homologs of several canonical symbiosis genes, part of the symbiosis signalling pathway, evolved. The second stage in the evolution of the symbiosis genes is predicted to have occurred in the earliest land plant ancestors of the bryophytes, where the homologs of the remaining characterised symbiosis genes evolved.

A phenomenal amount of insight into the AM symbiosis in angiosperms has been gained over the last two decades using angiosperm model plant species. Such model systems for the study of the AM symbiosis in the non-flowering plants are yet to be established. To address this, I established a non-flowering plant model system to facilitate further study of the AM symbiosis outside the angiosperm lineage. In chapter 3, I describe the successful establishment of this model system, *Marchantia paleacea*, through the development of protocols for routine culture, mycorrhizal colonization assays, and *Agrobacterium*-mediated transformation of this organism *in vitro* as well as the generation of the required genomic resources. I also describe my attempts to test a CRISPR-Cas9 mediated mutagenesis system to generate targeted mutations in *M. paleacea*.

Although the presence of symbiosis gene orthologs in non-flowering plants has been reported, it is not yet known whether these genes have functions in symbiosis in the plants from which they were found. I aimed to test whether there is any evidence of symbiotic roles for these genes in non-flowering plants in Chapter 4. For this, I employed comparative phylogenomic approaches, previously used to compare AM host and non-host species from the angiopserms, to the liverworts. Based on the results

of these analyses presented in chapter 4, I argue that symbiosis gene homologs likely function in regulating the AM symbiosis in the liverworts.

2

Comprehensive Phylogenetic Analyses Provide Insights into the Stepwise Acquisition of the Symbiotic Toolkit in Plants

The work described in this chapter has previously been published in the following peer-reviewed research article where I am co-author:

Delaux PM, Radhakrishnan GV, Jayaraman D, Cheema J, Malbreil M, Volkening JD, Sekimoto H, Nishiyama T, Melkonian M, Pokorny L, et al. 2015. Algal ancestor of land plants was preadapted for symbiosis. Proc Natl Acad Sci U S A 112:13390-13395.

The figures used in this chapter have been adapted from the aforementioned publication.

Chapter 2: Comprehensive Phylogenetic Analyses Provide Insights into the Stepwise Acquisition of the Symbiotic Toolkit in Plants

2.1. Introduction

Studies into the fungal associations formed by the bryophytes, whose members are predicted to the earliest diverging extant land plants, have shown that these plants also form mutualistic symbioses with the AM fungi (Humphreys et al., 2010; Field et al., 2012). Furthermore, orthologs of three genes of the angiosperm symbiotic toolkit have been identified in the bryophytes (Wang et al., 2010), leading to the hypothesis that these orthologous genes of the symbiotic toolkit function in the AM symbiosis in the bryophytes. Due to the relative dearth of sequence data from the other lineages of green plants, the majority of previous studies into the evolution of the symbiotic toolkit have been restricted to the angiosperms (Delaux et al., 2013b). In cases where genome sequences are available for other lineages, these are from plants that have lost the ability to engage in the AM symbiosis (Wang and Qiu, 2006) – such as the moss *Physcomitrella patens* (Rensing et al., 2008) and the conifer *Picea albies* (Nystedt et al., 2013). The only non-flowering plant that is capable of engaging in the AM symbiosis and for which reliable genome sequence data is available is the lycophyte *Selaginella moellendorffii* (Banks et al., 2011). On the other hand, recent efforts to generate sequence data for the non-flowering plant lineages have culminated in the availability of good-quality transcriptomic data (Matasci et al., 2014) that have been used successfully to trace the evolution of genes involved in the regulation of various plant processes (Sayou et al., 2014; Seki et al., 2015; Zhong and Kellogg, 2015). In the present chapter, I used these recently generated sequence data to explore the evolution of symbiosis genes to understand when these genes evolved in the evolutionary history of plants and to study the processes that have contributed to their evolution.

2.2. Aim and experimental design

In order to bridge the gap in knowledge on the evolution of symbiosis genes in the non-flowering plant lineages, I performed phylogenetic analyses on transcriptomic data generated for over 1000 plant species as part of the 1000 plants (1KP) initiative (Matasci et al., 2014). The aim of these analyses was to understand when the

symbiotic toolkit evolved and how this toolkit has changed over the evolutionary history of plants. Green algal and bryophyte transcriptomes from the 1KP dataset were used in addition to publicly available genomes of plants belonging to various lineages and genomic/transcriptomic data from other sources (Appendix A1).

A comprehensive phylogenetic workflow was employed to obtain homologs of the symbiotic toolkit from this dataset (Figure 2.1) (methods described in Section 6.1.2). For the subset of these genes that possess multi-domain architectures or characteristic domains (*CERK1, SYMRK, CCaMK, NSP1, NSP2, RAM1, RAD1, RAM2, VPY, STR, STR2*), the corresponding Pfam domains were used for the search. For the other genes (*DMI1, CYCLOPS, PT4, HA1*) sequence similarity searches were conducted to obtain their homologs (Table 2.1) (methods described in Section 6.1.2.2).

Gene	Family	Name	Multi-gene?	Domains in Pfam?
CERK1	LysMRLK	Lysin-Motif Receptor-like kinase	Yes	Yes
SYMRK	MLD-RLK	Malectin-like domain receptor-like kinase	Yes	Yes
DMI1	Ion channel	-	No	No
CCaMK	CCaMK	Calcium- and Calmodulin-dependent protein kinase	No	Yes
CYCLOPS	-	-	No	No
NSP1	GRAS	GRAS domain protein	Yes	Yes
NSP2	GRAS	GRAS domain protein	Yes	Yes
RAM1	GRAS	GRAS domain protein	Yes	Yes
RAD1	GRAS	GRAS domain protein	Yes	Yes
RAM2	GPAT	Glycerol-3-phosphate acyltransferase	Yes	Yes
VPY	-	-	No	Yes
STR	ABCG	ABCG transporter	Yes	Yes
STR2	ABCG	ABCG transporter	Yes	Yes
PT4	PT	Phosphate transporter	Yes	No
HA1	HA	Proton ATPase	Yes	No

Table 2.1. Description of the genes of the symbiotic toolkit, the families they belong to, and the occurrence of Pfam domains in the proteins they encode.

The *M. truncatula* gene identifiers for these genes are provided in Supplementary File S4.

The resultant hits from each of these searches were aligned (methods described in Section 6.1.2.3) and phylogenetic trees were constructed (methods described in Section 6.1.2.4). For the analysis, two specific comparisons were selected based on the predicted timeline of evolution of the AM symbiosis. First, the AM symbiosis is predicted to have evolved in the first land plants sometime after their divergence from ancestral alga. The closest living lineages to the first land plants and their algal ancestors are the bryophytes and advanced charophytes, respectively (Delwiche and Cooper, 2015). Thus, a comparative analysis between these lineages was undertaken to understand what occurred between the divergence of charophytes and bryophytes that led to the evolution of the AM symbiosis in the land plant ancestors of the bryophytes. Secondly, in order to follow the evolution of the symbiotic toolkit from its appearance in the first land plants through to its current state in the extant plant lineages, a comparative analysis between the bryophytes and angiosperms was undertaken. Within the bryophytes, the analysis focussed on the liverwort lineage, as the AM symbiosis has been reliably described within this group of bryophytes.

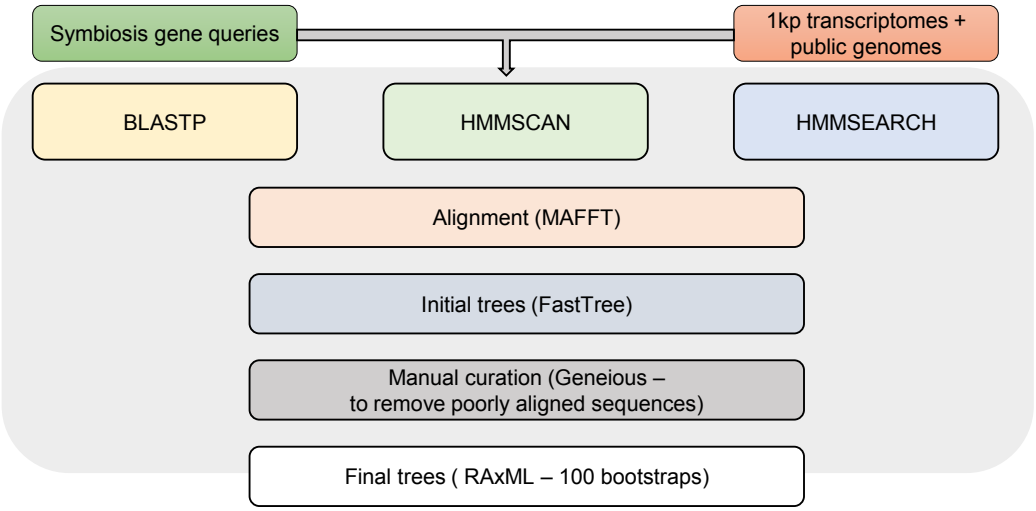

Figure 2.1. The phylogenetic workflow used for the analysis of symbiosis genes in the plant lineage.

Using angiosperm symbiosis genes as queries, transcriptomes and genomes were mined and the resultant hits were analysed using standard phylogenetic methods (sequence alignment and tree construction). Full-length proteins were used for the phylogenies as the overarching aim of the study was to test when the orthologs of the symbiosis genes first appeared. The limitation of this approach is that if different

domains of the proteins have recombined and evolved differently, this would not be detected.

2.3. Results

2.3.1. Components of the symbiotic toolkit predated the evolution of land plants

The phylogenetic reconstructions of the evolutionary history of the 15 symbiosis genes from angiosperms revealed that for 12 of these (*CERK1*, *SYMRK*, *DMI1*, *CCaMK, CYCLOPS, NSP1, NSP2, RAM1, RAD1, VPY, STR, STR2)*, true orthologs were present in the bryophytes while for the remaining genes (*RAM2, PT4, HA1*), homologs could be found (Figure 2.2). These findings suggest that although some components of the symbiotic toolkit may have diverged in the bryophyte and angiosperm lineages, the majority of the components seem to be conserved (Figure 2.2). Together, these results point to a scenario where an ancestral symbiotic toolkit similar to the one found in angiosperms may have evolved in the first land plants.

Surprisingly, a similar analysis of the symbiotic toolkit in the algae revealed that true orthologs could be found in the advanced charophytes for several genes of the symbiosis signalling pathway, including *CERK1*, *CCaMK, DMI1* and *CYCLOPS* (Figure 2.2). Furthermore, homologs were found for all the other genes except *VPY* and *RAM2* (Figure 2.2). In the basal charophytes and chlorophytes, orthologs were found for *CCaMK, DMI1* and *CYCLOPS*, and homologs were found for *STR, STR2, PT4* and *HA1*. For the other genes, no homologs or orthologs were found in these plants (Figure 2.2). The phylogenetic trees constructed for these analyses are provided in Supplementary File S1.

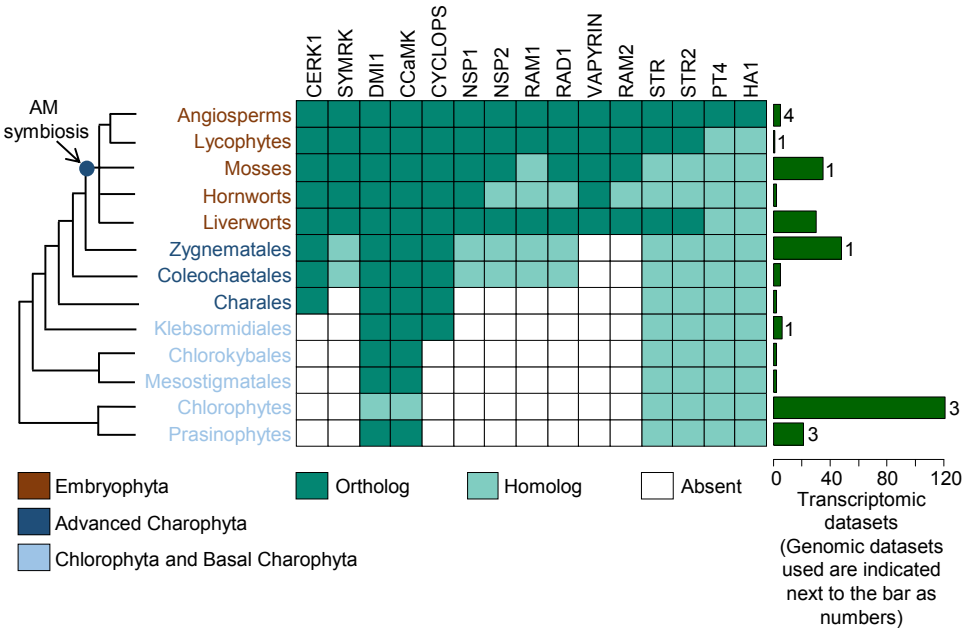

Figure 2.2. Summary of results from the phylogenetic analysis of symbiosis genes in plants.

The analysis was carried out on transcriptome and genome datasets. The number of datasets used for each clade are provided next to the orthology results. The number of genomes studied are mentioned next to the green bars representing the number of transcriptomes studied.

The discovery of components of the symbiotic toolkit outside the land plant lineage points to the evolution of some components of the symbiotic toolkit predating the evolution of the AM symbiosis. These components may have been later recruited into the AM symbiosis during land plant evolution but originally evolved for functions other than the AM symbiosis. These results show that the genes involved in symbiotic signalling had already evolved in the algal ancestors of land plants while the downstream symbiosis genes evolved in the first land plants. Further analysis into the evolution of each of these genes is presented below.

2.3.2. The evolutionary mechanisms that shaped the origin of the symbiotic toolkit

Using the phylogenies of the symbiotic toolkit (Figures 2.3-2.8), the evolutionary processes that led to the evolution of the symbiotic toolkit in plants were explored. These analyses revealed that previously proposed drivers of gene evolution such as gene duplication, gene fusion and *de novo* gene birth have influenced the evolution of

the symbiotic toolkit (Long et al., 2003; Magadum et al., 2013; Schlötterer, 2015; Albalat and Canestro, 2016).

2.3.2.1. Several genes in the symbiotic toolkit arose as a result of gene duplications

While some of the genes in the symbiotic toolkit belong to large multigene families, others belong to single gene families (Table 2.1). Multigene families are known to evolve through gene duplication occurring either in specific genomic locations or through whole-genome duplications (Friedman and Hughes, 2001). Using the phylogenies made for the symbiotic toolkit, the role of gene duplications in the evolution of the symbiotic toolkit was explored.

CERK1, a LysMRLK, has recently been shown to be involved in the perception of AM fungi to allow fungal colonisation in rice (Miya et al., 2007). From the phylogenetic reconstruction of the LysMRLK family, it was found that the clade containing CERK1 contained both land plants and advanced charophyte sequences (Figure 2.3a). CERK1 in specific species such as *S. moellendorffii* and *M. truncatula* seems to have been duplicated and these seem to be species or lineage-specific duplications as has been previously reported for genes in the LysMRLK family (Zhang et al., 2007; Zhang et al., 2009). True orthologs of CERK1 were found in the liverworts and the charophytes.

For SYMRK (an MLD-RLK Malectin-like-domain Receptor-Like Kinase), true orthologs were found in the bryophyte lineage, whereas only homologs could be found in the advanced charophyte lineages tested (Figure 2.3b). Numerous clades of bryophyte MLD-RLKs were found to be co-orthologous to the single charophyte clade suggesting that the ancestral MLD-RLK found in the common ancestor of charophytes and liverworts diversified into the many clades of MLD-RLKs found in extant liverworts.

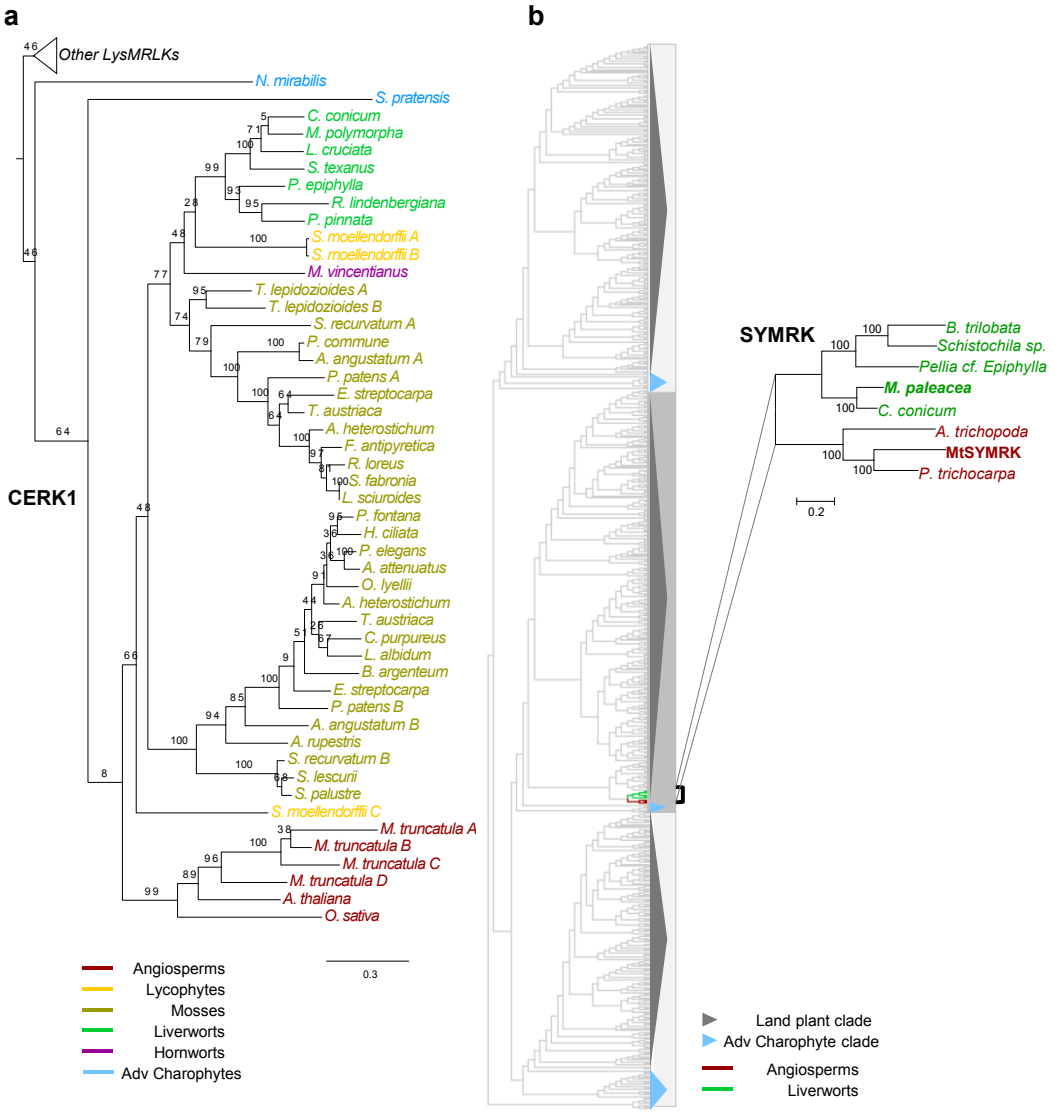

Figure 2.3. Evolutionary history of CERK1 and SYMRK

(a) A maximum-likelihood phylogenetic tree representing the evolution of CERK1 in the plant lineage. CERK1 orthologs were found in land plants and advanced charophytes. (b) Maximum-likelihood phylogenetic tree representing the evolution of MLD-RLKs. The SYMRK clade is highlighted showing that SYMRK is restricted to land plants. Land plant SYMRKs evolved as a result of an expansion of the MLD-RLKs in the land plant lineage from ancestral MLD-RLKs that have orthologs in the advanced charophytes. Adv Charophyte– Advanced Charophyte.

The ion-channel DMI1 has been shown to have functions in the AM symbiosis in the angiosperms and has previously been shown to have orthologs in the liverworts (Wang et al., 2010). Phylogenetic reconstruction of the evolutionary history of this

ion-channel family revealed that DMI1 predates the evolution of land plants and is present in both chlorophytes and charophytes. Two copies of DMI1 (named CASTOR and POLLUX after the *L. japonicus* proteins) were found in the angiosperms and this seems to have been a result of a gene duplication event that occurred after the divergence of angiosperms from its common ancestor with the liverworts. Outside the angiosperm lineage, only a single-copy of DMI1 is found (Figure 2.4a).

Numerous GRAS-domain proteins have been shown to have important roles in the establishment of the AM symbiosis (Xue et al., 2015). Using a HMM-based profile to search for the GRAS-domain that is present in these proteins, GRAS proteins were found in the liverworts as well as advanced charophytes (Figure 2.4b). No GRAS proteins were detected in the basal charophyte or chlorophyte lineages. Phylogenetic reconstruction of the entire GRAS family showed that for each of the symbiotic GRAS proteins (NSP1, NSP2, RAM1 and RAD1) orthologous genes could be identified in the liverworts. In the charophytes, true orthologs were not found for any of these genes but ancestral relationships could be ascertained where a charophyte clade was found to be ancestral to several clades containing land plant GRAS proteins. This suggests that the symbiotic GRAS proteins arose through the massive expansion of the GRAS protein lineage through gene duplications in the land plants. Intriguingly, several GRAS protein clades unrelated to land plant GRAS proteins were found in the charophytes. From the phylogeny, it is clear that the numerous GRAS proteins found in the charophytes arose through independent duplications in this lineage similar to the duplications found in the land plant lineage that gave rise to the symbiotic GRAS proteins. This finding highlights the fact that although the charophytes diverged from much earlier than the extant land plants, these algae have continued to evolve independently since their divergence and should not be thought of as primitive plants containing fewer gene family members compared to the land plants.

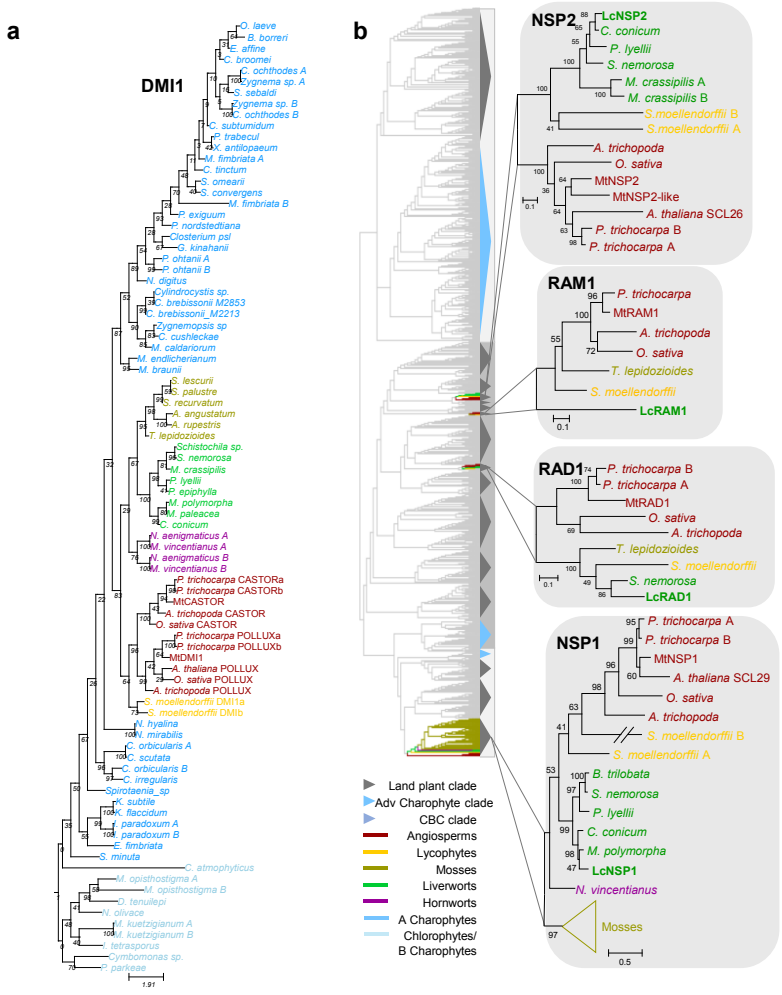

Figure 2.4. Evolutionary history of DMI1 and the GRAS proteins NSP1, NSP2, RAM1, RAD1

(a) Maximum-likelihood phylogenetic tree representing the evolution of DMI1 in the plant lineage. DMI1 orthologs were found in land plants, basal charophytes and advanced charophytes. (b) Maximum-likelihood phylogenetic tree representing the evolution of the GRAS proteins with the NSP1, NSP2, RAM1 and RAD1 clades being highlighted. The orthologs of all four genes are restricted to the land plants. Other GRAS protein homologs were found in the charophytes, but no orthologs were identified. Adv Charophyte clade – Advanced Charophyte clade, CBC – Chlorophyte & Basal Charophytes, B Charophytes – Basal Charophytes

The *RAM2* gene encodes for a glycerol-3-phosphate acyl transferase (GPAT) that has been shown to be required for the biosynthesis of lipids provided by the plant hosts to

their fungal partners during the AM symbiosis (Bravo et al., 2017; Jiang et al., 2017; Keymer et al., 2017; Luginbuehl et al., 2017). The GPATs are a multigene family, with RAM2 being one of the representative genes in the family (Table 2.1). From the phylogenetic reconstruction, GPATs were found to be specific to the land plants and no GPATs were found in the algal lineages. Focussing on the RAM2 clade, the angiosperm RAM2 genes were found to form a single monophyletic group with three bryophyte ancestral clades (Figure 2.5a). Thus, in the case of RAM2, there was a single RAM2 ortholog in the ancestral land plants, which was maintained as single-copy in the angiosperms. In the liverworts, RAM2 diversified into three copies. GPATs have previously been shown to have a role in cutin biosynthesis and RAM2 behaves biochemically similar to these GPATs. Thus, differences in the GPAT family, such as this diversification of RAM2 in the liverworts, may be due to differences in cutin biosynthesis between these lineages reflected by the absence of a cuticle in liverworts.

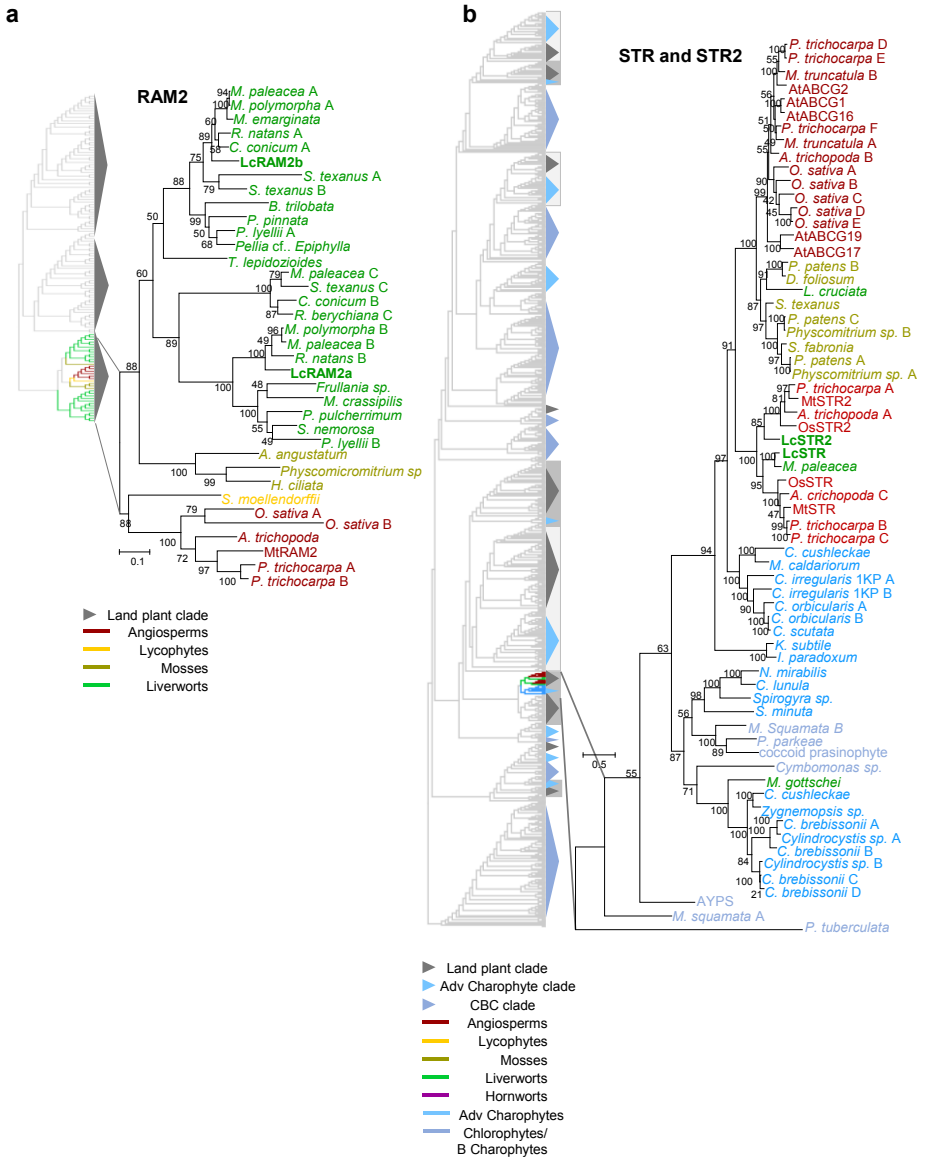

Figure 2.5. Evolutionary history of RAM2 and STR/STR2

(a) Maximum-likelihood phylogenetic tree representing the evolution of the GPATs in the plant lineage with the RAM2 clade highlighted. The RAM2 lineage has undergone diversification in the liverworts with three copies from liverworts being found to be orthologous to the angiosperm RAM2 clade (b) Maximum-likelihood phylogenetic tree representing the evolution of the half-size ABCG transporter family. The clade containing STR and STR2 from the land plants and the closest algal homologs is highlighted. The orthologs of STR and STR2 are restricted to the land plants. The presence of homologs that are ancestral to the STR and STR2 clades in the algae suggest that STR and STR2 arose through gene duplications from ancestral ABCGs. Adv Charophyte – Advanced

Charophyte, CBC – Chlorophyte & Basal Charophytes, B Charophytes – Basal Charophytes.

In the case of *STR* and *STR2,* which encode for ABC transporters, the genes are conserved as orthologs in the land plant lineage but are not found in the algal lineage (Figure 2.5b). By contrast, for the phosphate transporter PT4 and the proton ATPase HA1, clades containing numerous angiosperm genes were found to be orthologous to clades containing numerous bryophyte genes (Figure 2.6). Clear orthology relationships between the members of these clades could not be ascertained due to massive gene duplications resulting in independent diversification of these genes in the liverworts and the angiosperms. For all of the above genes, although orthologs could not be found in the algae, a clade consisting of charophyte and chlorophyte genes was found to be ancestral to the land plant clades pointing to numerous gene duplications occurring during the evolution of these genes.

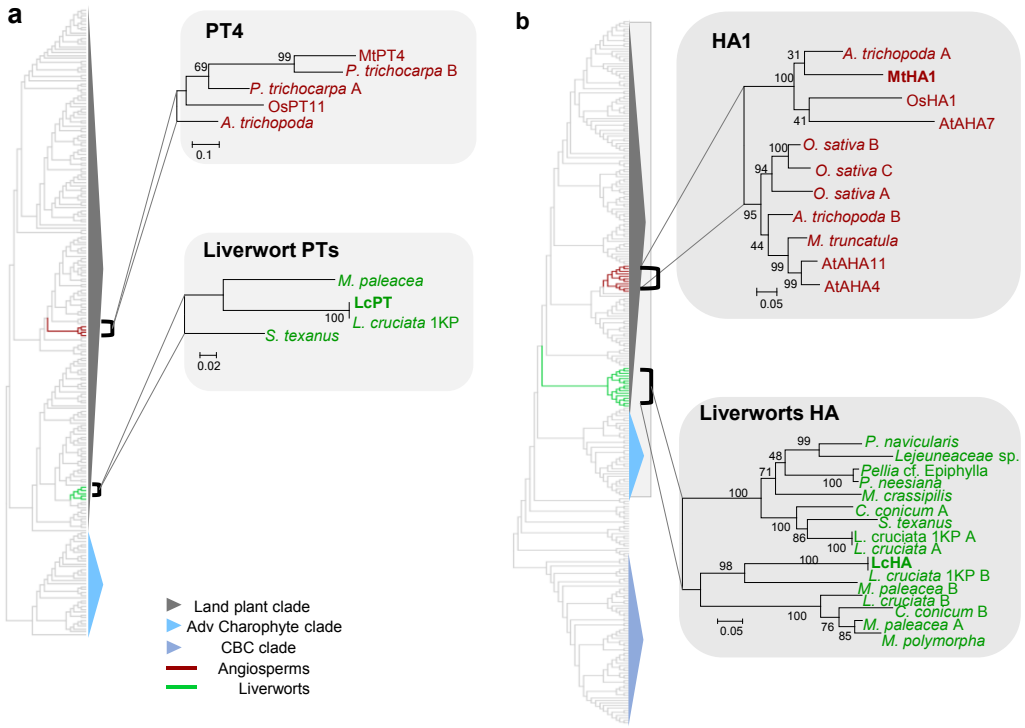

Figure 2.6. Evolutionary history of PT4 and HA1 in the plant lineage

(a) Maximum-likelihood phylogenetic tree representing the evolution of phosphate transporters in plants. The clades containing angiosperm PT4 orthologs and the closest liverwort phosphate transporters (PTs) are highlighted. True orthologs of angiosperm PT4 were not found in the liverworts due to extensive gene diversification of PTs in both angiosperms and liverworts. (b) Maximum-likelihood phylogenetic tree representing the evolution of proton ATPases in plants. The clades containing angiosperm HA1 and the closest liverwort H+-ATPases (HAs) are highlighted. Similar to PT4, true orthologs of angiosperm HA1 were not found in the liverworts due to extensive gene diversification of HAs in both angiosperms and liverworts. Adv Charophyte – Advanced Charophyte, CBC – Chlorophyte & Basal Charophytes.

The remaining genes of the symbiotic toolkit, *CCaMK, CYCLOPS* and *VPY*, have been maintained as single-copy genes in most species (Figure 2.7). Numerous single-copy gene families are known to be present in the genomes of plants and the main factor behind the maintenance of these genes as single-copy seems to be dosage sensitivity (Edger and Pires, 2009). Gene duplication has an adverse effect on the

function of these genes due to stoichiometric constraints. Therefore, any duplications arising in these genes are quickly eliminated through selection. For all three genes, orthologs were found in the liverworts. For *CCaMK* and *CYCLOPS*, orthologs were also found in the basal and advanced charophyte lineages. For *VAPYRIN*, orthologs could only be found in the liverworts and not in any of the algae.

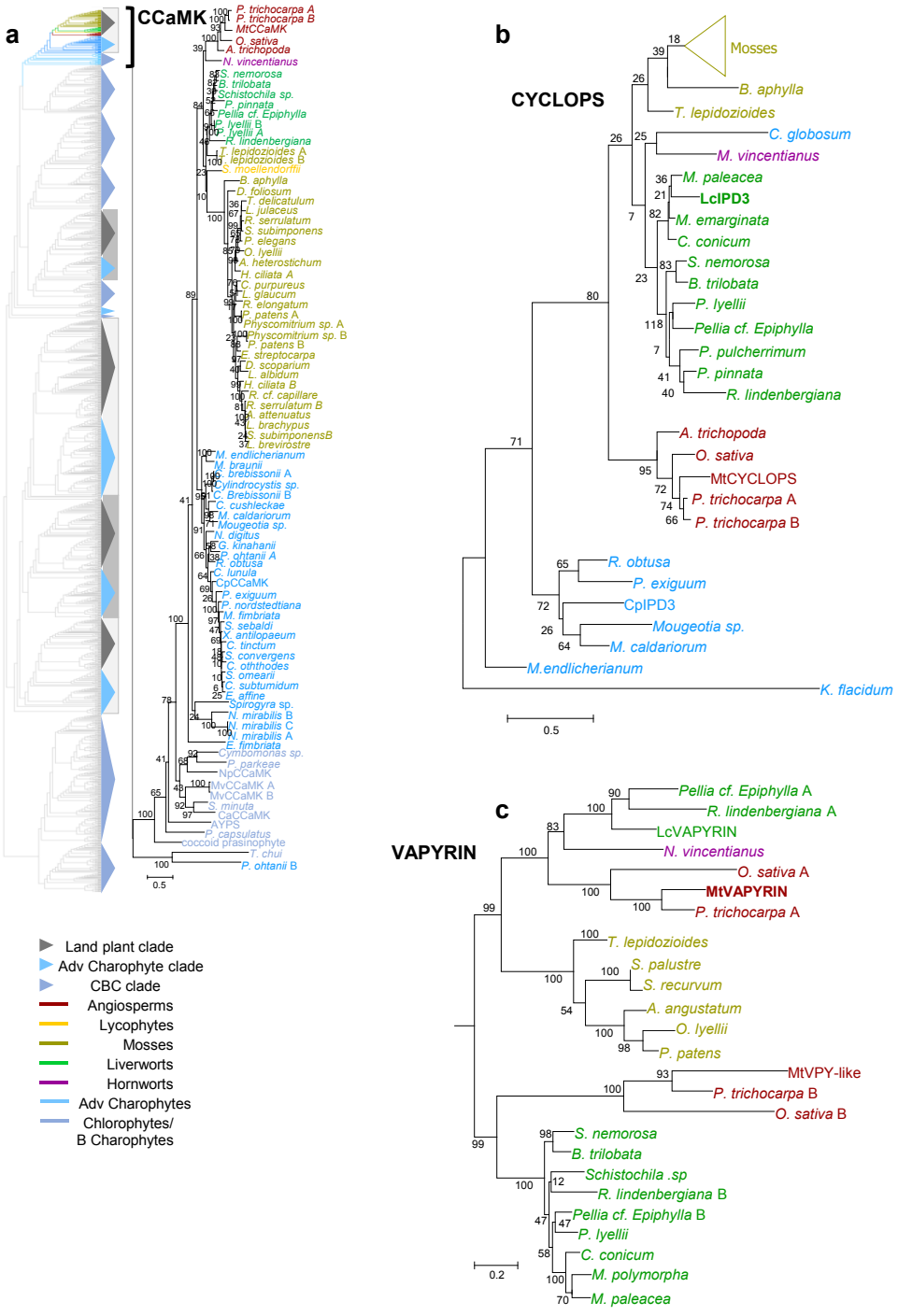

Figure 2.7. Evolutionary history of CCaMK, CYCLOPS and VAPYRIN in the plant lineage

(a) Maximum-likelihood phylogenetic tree representing the evolution of CDPKs and CCaMKs in the plant lineage. The clade containing CCaMKs is highlighted. CCaMKs were found in land plants and in both basal and advanced charophytes. (b) Maximum-likelihood phylogenetic tree representing the evolution of CYCLOPS. The orthologs of CYCLOPS were found in land plants and advanced charophytes. No other homologs were found. (c) Maximum-likelihood phylogenetic tree representing the evolution of VAPYRIN and its homolog VPY-like. Both VAPYRIN and VPY-like were restricted to the land plants. Adv Charophyte – Advanced Charophyte, CBC – Chlorophyte & Basal Charophytes, B Charophytes – Basal Charophytes.

2.3.2.2. Duplication-independent evolutionary mechanisms also contributed to the evolution of the symbiotic toolkit

Several of the genes in the symbiotic toolkit encode multi-domain proteins (Table 2.1). It was found that some of these genes (*VPY* and *RAM2*) did not have any homologs resembling their multi-domain structures in the charophyte and chlorophyte species tested (Figure 2.2). As the next step in the analysis, these datasets were searched for the presence of the constituent domains that make up these multi-domain proteins. If the constituent domains are present, it would suggest that these multi-domain proteins may have evolved in the land plants through gene fusions or domain shuffling between proteins that were present in their algal ancestors (Pasek et al., 2006). Although VPY and RAM2 could not be found in the charophytes, their constituent domains were detected as being present in the charophytes in other proteins (Table 2.2) supporting the possibility that these genes may have evolved through domain shuffling as has been reported for other cases of gene evolution.

Proteins containing the MLD domain (found in SYMRK) and the GRAS-domain (found in NSP1, NSP2, RAM1, and RAD1) were found in the advanced charophytes but not in the basal charophytes or chlorophytes (Table 2.2). The appearance of these novel domains in the charophytes may have been through *de novo* gene evolution, an as yet understudied and underappreciated form of evolution of genes (Tautz, 2014). Similarly, orthologs of *CYCLOPS* were found in the charophytes and liverworts but no genes with homology to *CYCLOPS* were found in the chlorophytes. Thus, it is

possible that *CYCLOPS* also evolved through this mechanism in the ancestors of charophytes and liverworts.

Protein	Domains	PFAM ID	Adv Charophytes		Chlorophytes & Basal charophytes	
			Fused?	Present?	Fused?	Present?
LysMRLK	LysM	PF01476	+	+	-	+
	Kinase	PF00069 PF07714		+		+
MLD-RLK	MLD	PF12819	+	+	-	-
	Kinase	PF00069 PF07714		+		+
VAPYRIN	Motile sperm	PF00635	-	+	-	+
	Ankyrin	CL0465		+		+
RAM2	HAD	PF12710	-	+	-	+
	Acyl-transferase	PF01553		+		+
GRAS	GRAS	PF03514		+		-

Table 2.2. The occurrence of proteins from the symbiotic toolkit with well-characterised domains and their constituent domains in the charophytes and chlorophytes.

2.3.3. Conservation of structural and topological features of genes in the symbiotic toolkit in the plant lineage

With the results of the phylogenetic analyses in place, conservation of specific structural features of the proteins comprising the symbiotic toolkit was explored. Previous functional studies have shown that specific motifs in two symbiotic proteins CCaMK and DMI1 are important for the establishment of a functional symbiosis (Venkateshwaran et al., 2012; Miller et al., 2013). In CCaMK, this is the calmodulin-binding domain, which is highly conserved within the angiosperms and is important for the symbiotic function of CCaMK. In DMI1, this is a stretch of residues referred to as the pore domain, which is also highly conserved in the angiosperms. Analysis of these two domains in CCaMK and DMI1 orthologs was performed to gain insight into when these structural features important for symbiotic function appeared. For both of these domains, the conservation previously found in the angiosperms was found to extend across the land plants (Figure 2.8). Surprisingly, the conservation also

extended into the advanced charophyte orthologs of these proteins. In contrast, the basal charophyte and chlorophyte orthologs had highly divergent sequences in these domains. Thus, the forms of these proteins with structural features important for symbiotic function seem to have evolved in the advanced charophytes from an ancestral form that did not contain these features as evidenced by their absence in the basal charophyte and chlorophyte sequences.

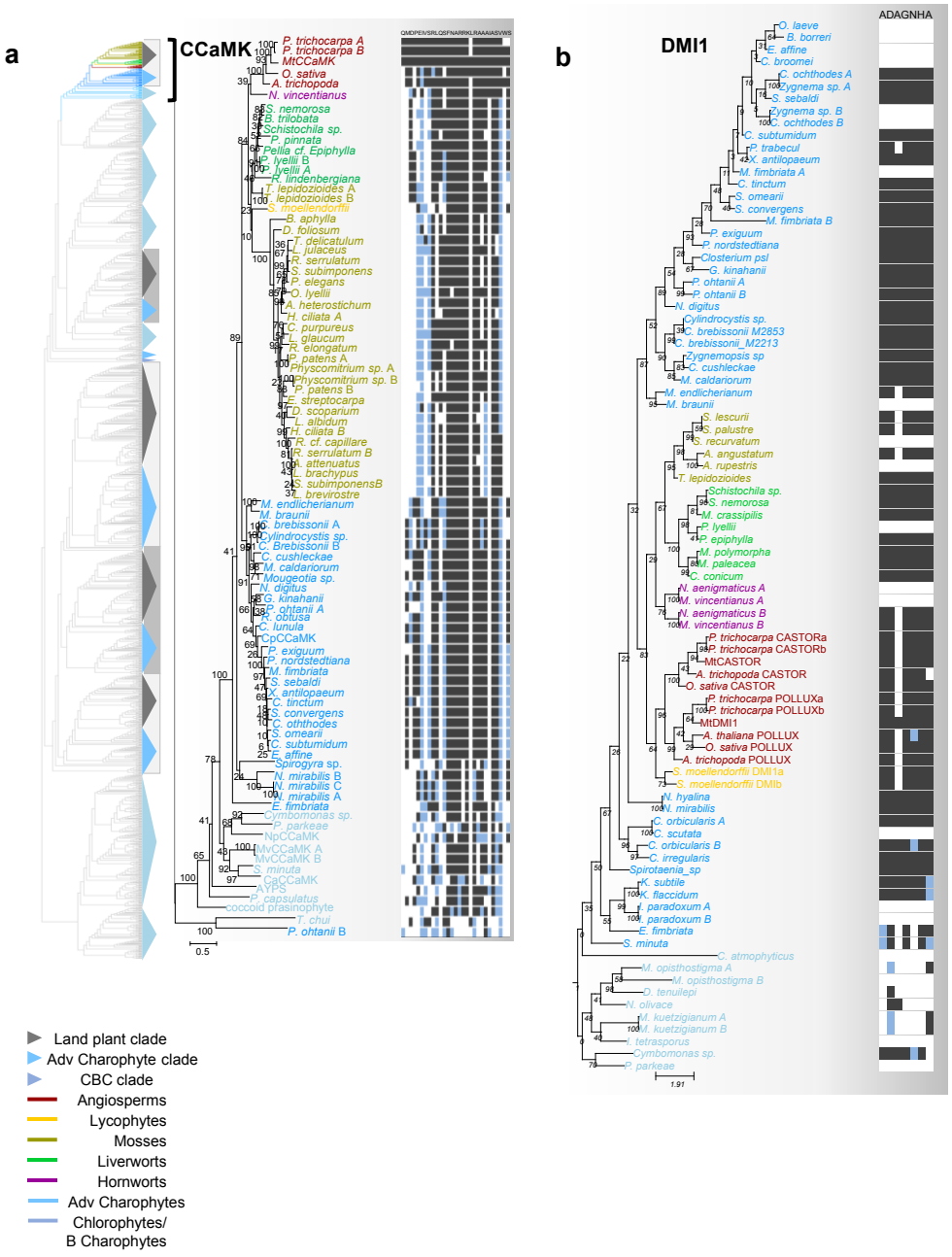

Figure 2.8. Conservation of the calmodulin-binding domain and the pore domain in CCaMK and DMI1 homologs in plants

(a) Maximum-likelihood phylogenetic tree representing the evolution of CDPKs and CCaMKs in the plant lineage. The clade containing CCaMKs is highlighted along with the respective sequence of the calmodulin-binding domain required for the symbiotic function. The sequences were compared against Medicago CCaMK where the calmodulin-binding domain is QMDPEIVSRLQSFNARRKLRAAAIASVWS. Residues in a given taxon that are identical to the Medicago sequence are coloured in black, similar residues are coloured in blue, and other residues are coloured in white. The CaM-binding domain is highly conserved within the advanced charophytes and land plants. In the basal charophytes, the corresponding region is highly divergent. (b) Maximum-likelihood phylogenetic tree representing the evolution of DMI1 in the plant lineage. Alongside the tree a multiple sequence alignment with the pore domain from each taxon is provided. The sequences were compared against Medicago DMI1 where the pore domain sequence is ADAGNHA. The residues were coloured as defined above. The pore domain was found to be highly conserved within the advanced charophytes and land plants. In the basal charophytes, the corresponding region is highly divergent. Adv Charophyte – Advanced Charophyte, CBC – Chlorophyte & Basal Charophytes, B Charophytes – Basal Charophytes.

2.3.4. Revisiting previously proposed phylogenetic hypotheses on the evolution of the symbiotic toolkit using new data

Taxon sampling has been shown to be a significant factor in determining the power of studies using phylogenetics to evaluate evolutionary hypotheses (Hillis, 1998; Zwickl and Hillis, 2002; Hillis et al., 2003; Nabhan and Sarkar, 2012). Previous studies on the evolution of genes in the symbiotic toolkit have formulated evolutionary hypotheses based on the sparse genomic data available at the time of the study (Hrabak et al., 2003; Markmann et al., 2008; Wang et al., 2015). As more sequence information from previously unavailable taxa has now become available, I revisited two of these hypotheses about the evolution of CCaMK and SYMRK.

2.3.4.1. CCaMKs and CDPKs are monophyletic groups of calcium kinases predating green plant evolution

Previous analyses on the evolution of CCaMKs have been restricted to only a few species with sequenced genomes (Hrabak et al., 2003; Wang et al., 2015). These studies have identified structural similarities between CCaMKs and a related group of kinases called CDPKs. CCaMKs and CDPKs share a similar domain architecture with an N-terminal protein kinase domain and several EF-hand domains at the C-terminus. Phylogenetic analyses of CCaMKs and CDPKs have previously concluded that CCaMKs likely evolved from CDPK ancestors (Wang et al., 2015). This was based on the observation that CDPKs were found in plants as well as protists, while CCaMKs were only found in plants (Harper and Harmon, 2005). These results suggest that CDPKs were present in the common ancestor of protists and plants while CCaMK was a novel acquisition in the plant lineage. In order to explore if these conclusions hold true with denser and wider taxon sampling, I used transcriptome and genome datasets covering the entire archaeplastid lineage (Appendix A2). Using the domain structure shared by CCaMKs and CDPKs as the criteria for the mining of the sequence data, I conducted HMM-based profile searches to search for CCaMK/CDPK homologs. The resultant hits were then aligned and a phylogenetic tree was constructed, in which the CCaMK and CDPK clades were identified based on known CCaMK and CDPK sequences from plants (Figure 2.9).

Surprisingly, the phylogeny revealed that CCaMKs were not only present in plant lineages, but also in protist species previously thought to contain only CDPKs (Ward et al., 2004; Harper and Harmon, 2005; Billker et al., 2009; Wernimont et al., 2010). Furthermore, sequences with homology to CCaMKs were found in the rhodophytes and glaucophytes, but as these sequences were solely from transcriptomes and were truncated, did not have a stop codon, and did not cover the regions representing all the canonical domains present in CCaMKs, they were designated as CCaMK-like. Multiple sequence alignments of the protist CCaMKs revealed that these sequences were indeed *bona fide* CCaMKs based on the conservation of all the domains that are thought to be present in canonical CCaMKs - the kinase, the calmodulin-binding domain and the EF-hands. CDPKs on the other hand do not possess the calmodulin-binding domain. Furthermore, the CCaMKs and CDPKs were each found to form monophyletic clades containing sequences from protists, green algae and plants, suggesting that the divergence of these two families is much more ancient than

previously thought. Future studies including more representatives of the calcium-regulated kinase family, such as those from animals, are required to further elucidate the evolution of these kinases and to understand what the calcium-decoding toolkit of the common ancestor of the protists and the green lineage looked like.

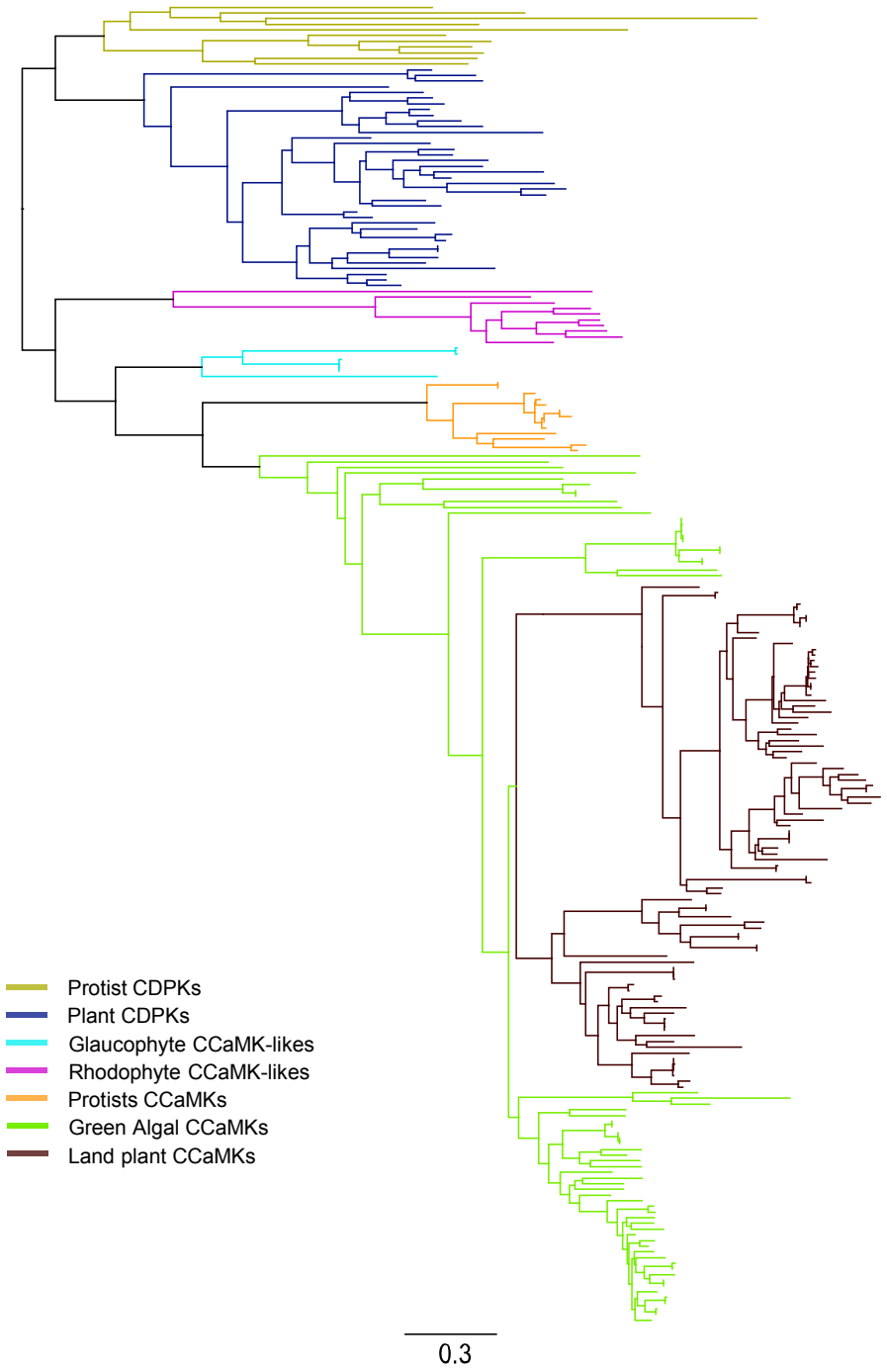

Figure 2.9. Evolutionary history of CDPKs and CCaMKs

Maximum-likelihood phylogenetic tree representing the evolution of CDPKs and CCaMKs from all plastid-containing organisms

(Archaeplastida). CCaMKs and CDPKs each form monophyletic clades. CDPKs were found in plants and protists. CCaMKs were found in plants, protists, glaucophytes and rhodophytes. The sequences found in the rhodophytes and glaucophytes are referred to as CCaMK-like as the sequences were truncated, did not have a stop codon, and did not cover the regions representing all the canonical domains present in a CCaMK. Genome sequences from these lineages will provide further clarification about these sequences.

2.3.4.2. Domain loss explains the evolution of SYMRK architecture and rejects the predisposition of SYMRK for nodulation

The protein kinase SYMRK belongs to the MLD-RLK family, comprising proteins with a MLD, numerous LRR domains and a protein kinase domain. The legume genes for *SYMRK* encode proteins that contain a MLD in addition to 3 LRR domains and a kinase domain (Figure 2.10a). Previous analysis of SYMRK proteins from angiosperms found that the SYMRK in angiosperms was present in three different domain architectures (Markmann et al., 2008). The first type of architecture found consisted of a SYMRK with 2 LRRs and a kinase domain but no MLD and this version was found to be present in all the tested monocots. The second version with an MLD, 2 LRRs and a kinase domain was found in all dicots outside the rosid clade. The third version with an MLD, 3 LRRs and a kinase domain is present in the members of the rosid clade, which contains all known angiosperm species capable of engaging in bacterial endosymbiosis. (Figure 2.10b)

SYMRK in legumes has been shown to function in both fungal and bacterial endosymbioses. Based on the finding that a specific structural version of SYMRK was only found in the clade containing species able to form bacterial endosymbioses, a role for structural variations in SYMRK architecture in determining the ability of plants to form bacterial endosymbiosis has previously been explored. For this, a complementation strategy was applied where *symrk* mutants of the legume *Lotus japonicus* defective in both bacterial and fungal symbioses were used. These mutants were transformed with *SYMRK* genes from other species to test whether these genes could complement either the bacterial, fungal or both symbioses. From these experiments, it was revealed that only the SYMRK version from the rosids (MLD, 3

LRRs and kinase) could complement both symbioses. The other two versions could complement the fungal symbiosis but not the bacterial one (Figure 2.10c). Based on these findings, it was proposed that the evolution of the extra LRR may have contributed to the eventual evolution of bacterial endosymbiosis in the rosid clade (Figure 2.10d). An alternate hypothesis was also proposed wherein, the ancestral form SYMRK may have been one with 3 LRRs and the SYMRK found in the monocots and non-rosid dicots may have been a result of the loss of an LRR domain (Figure 2.10e).

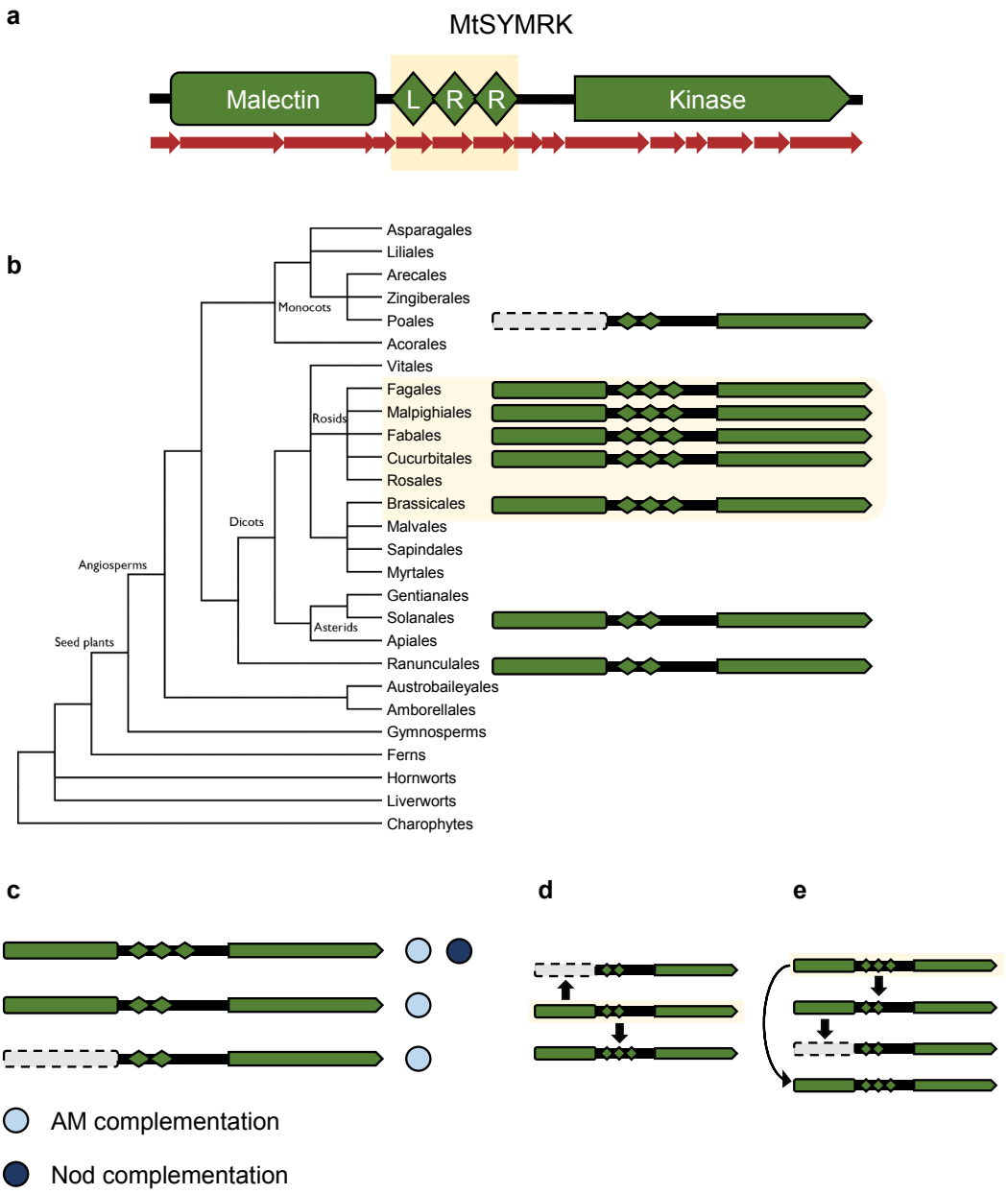

Figure 2.10. Insights gained into the evolution of SYMRK by Markmann and colleagues (Markmann et al., 2008)

(a) Domain structure of Medicago SYMRK with the corresponding exons highlighted in red. Medicago SYMRK is an MLD-RLK composed of a MLD, 3 LRRs and a kinase domain. (b) Domain structure evolution of SYMRK in the plant lineage as evidenced by the sequences used. Only Rosid SYMRKs were found to possess an MLD, 3 LRRs and a kinase. Monocot SYMRKs had no MLD and had only 2 LRRs and a kinase. Other Dicot SYMRKs had an MLD, 2 LRRs and a kinase. (c) Results from trans-complementation experiments suggested that all three forms complemented the AM- phenotype of *Lotus japonicus symrk* mutant while only the full-length Rosid SYMRK could complement the nod- phenotype. (d) and (e) The hypotheses proposed by the study about the evolution of SYMRK. In the first hypotheses (d), the ancestral SYMRK in the angiosperms had an MLD, 2 LRRs and a kinase. In the monocots, the MLD was lost and in the rosids, an LRR was gained. In the second hypotheses, (e) the ancestral SYMRK in the angiosperms had an MLD, 3 LRRs and a kinase. In the monocots, an MLD and an LRR were lost. In some dicots, an LRR was lost while the ancestral domain structure was preserved in the rosids.

As this analysis was performed on a dataset containing 11 angiosperm species and covering only monocot and dicot lineages, I repeated this phylogenetic analysis by including all land plant species from the 1kp data and included plant genomes sequenced since then (Appendix A3). As the original study was restricted to dicots and monocots and importantly did not include any outgroups (such as basal angiosperm species or any non-flowering plant species), the aim was to provide context to the evolution of this protein by providing appropriate outgroups and increasing the taxon sampling to test the hypotheses proposed by the original study. For the identification of SYMRK orthologs, the previously constructed SYMRK tree (Figure 2.3b) containing charophyte, bryophyte and angiosperm sequences was used and new hits from the 1kp and genome sequenced angiosperm datasets were added to this tree. On this tree, the MLD, LRR and kinase domains were then annotated (Figure 2.11).

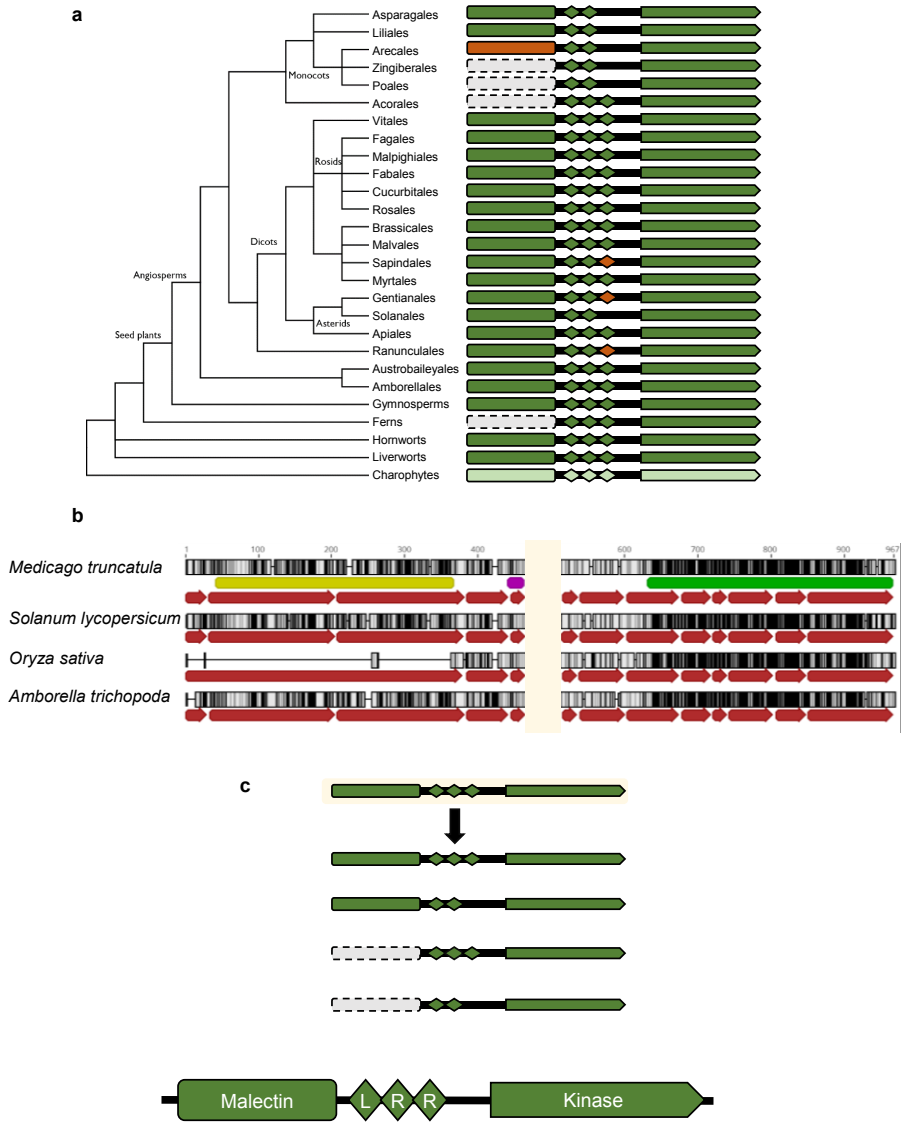

Figure 2.11. Evolution of SYMRK domain architecture in the plant lineage

(a) Domain structures of SYMRKs from across the plant lineage are superimposed onto the species tree representing the clades of the plants species studied here. Orthologs of SYMRK are represented with domains coloured dark green while the closest homolog of SYMRK from charophytes (derived from the phylogeny described in Fig 2.3) is represented with domains coloured light green. From the phylogeny, it is apparent that the ancestors of land plants, angiosperms, dicots and monocots each possessed a SYMRK with an MLD, 3 LRRs and a kinase. Where a clade contains taxa that possess the domain as well as those that have lost the domain, these are coloured in orange. Regions corresponding to the 4th imperfect LRR, previously reported to be present in SYMRKs

by Markmann et al., 2008, were found in all tested SYMRKs in the current study. (b) Multiple sequence alignment of SYMRKs from the rosid dicot *M. truncatula*, the non-rosid dicot *S. lycopersicum*, the monocot *O. sativa*, and the basal angiosperm *A. trichopoda*. The regions corresponding to the MLD is coloured yellow, the LRRs purple, and the kinase domain green. The exons in each sequence are highlighted by red arrows. Monocot SYMRKs have lost the MLD through changes in the first three exons. Only 2 LRRs are maintained in some clades of monocots and dicots and these seem to have been caused by changes in the sixth and seventh exons of the gene. (c) The hypothesis for the evolution of SYMRK supported by the comprehensive analysis of SYMRKs from across the plant lineage. The ancestral SYMRK in land plants possessed an MLD, 3 LRRs and a kinase. This ancestral domain structure is maintained in most lineages, while some lineages have specifically lost the MLD or an LRR or both through loss of exons.

From the analysis, I found that the SYMRK ortholog in liverworts possessed an MLD, 3 LRRs and a kinase, suggesting that in the common ancestor of liverworts and angiosperms, SYMRK had the "full-length" domain architecture found in the rosids (Figure 2.11a). Analysis of the domain architecture of SYMRK across the tree revealed that this "full-length" architecture has been retained throughout the angiosperm lineage. The other architectures found in the monocots and some dicots were due to independent losses of either the MLD or an LRR domain (Figure 2.11b,c). Taken together, these results reject the hypothesis that there was a gain of a LRR domain in the rosids and that this then resulted in a predisposition in SYMRK in the rosids leading to the evolution of SYMRK function in bacterial endosymbiosis. Whether the differences found in the ability of different plant SYMRKs to differentially complement fungal or bacterial endosymbioses are due to the differences in the domain architecture will now need to be revisited using SYMRKs from all representative clades of land plants. The phylogenetic analysis done here would provide a basis for selection of representative SYMRK forms with different domain architectures for these experiments.

2.4. Discussion

2.4.1. Stepwise evolution of the symbiotic toolkit in the plant lineage

Studies in experimental evolution have shown that the evolution of genetic pathways associated with specific traits progresses in a stepwise fashion with three major steps (Blount et al., 2012). Starting from the ancestral state where very few parts of the pathway exist, a potentiation event results in the acquisition of new genetic components paving the way for the evolution of the genetic pathway (Blount et al., 2008). Following this potentiation/predisposition stage, actualisation occurs whereby a preliminary form of the trait and the genetic pathway determining it evolve (Blount et al., 2012). Finally, further refinement of this primitive form of the trait continues to increase the efficiency of the trait and the benefits provided by the trait to fitness (Blount et al., 2008; Blount et al., 2012; Blount, 2016).

Using phylogenetic analyses, I found that the symbiotic toolkit followed a similar stepwise evolutionary pattern in plants (Figure 2.12). By studying extant plant lineages that diverged at different stages, it was possible to infer the ancestral states present during different points of plant evolution. It was not until the ancestor of the streptophytes (advanced charophytes and embryophytes) that the orthologs of the majority of the symbiotic toolkit appeared. The chlorophytes and basal charophytes possess orthologs of only two genes (*CCaMK* and *DMI1)* of the symbiotic toolkit and do not possess the rest of the symbiotic toolkit. By the time advanced charophytes diverged, orthologs (*CERK1*, *CYCLOPS*) and homologs (*SYMRK, NSP1, NSP2, RAM1, RAD1*) of more symbiosis genes had appeared. Furthermore, specific structural features that are required for the proper functioning (as shown for CCaMK and DMI1) of the symbiotic toolkit had also evolved in the advanced charophytes. These evolutionary events in the algal ancestor of land plants possibly paved the way for the evolution of mycorrhization in the first land plants. This is supported by the presence of orthologs/homologs of all genes in the symbiotic toolkit in the liverworts, which diverged soon after land plant colonisation.

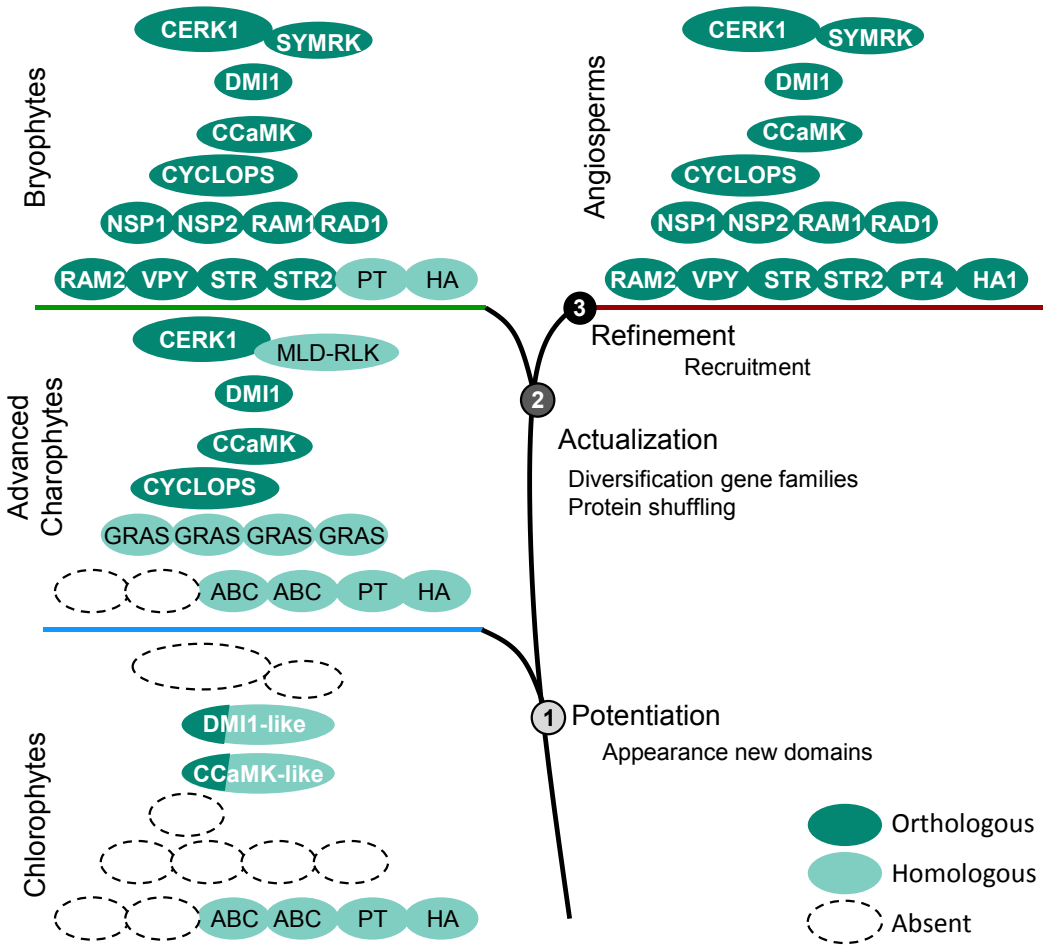

Figure 2.12. Acquisition of the symbiotic toolkit in a stepwise fashion in plants

New domains appeared in the advanced charophytes leading to the advent of novel protein families and to the evolution of the signalling module through the neo-functionalisation of existing components (DMI1 and CCaMK). The majority of the downstream symbiosis genes appeared in the first land plants through gene duplication in gene families (SYMRK, STR, STR2) or through combining existing domains to form novel proteins (RAM2, VAPYRIN). Finally, lineage-specific expansions led to the independent recruitment of some proteins (PT and HA).

The presence of orthologs and co-orthologs of all the genes of the symbiotic toolkit in the liverworts supports the long-held view based on fossil evidence that the AM symbiosis evolved once in the first land plants. It remains to be tested whether these orthologs/co-orthologs of the symbiotic toolkit function in symbiosis in the liverworts. With the advent of molecular methods allowing for the characterisation of genes in

the liverworts species, such studies might be possible in the near future (Ishizaki et al., 2016). Furthermore, recent studies have shown that the basal lineages of liverworts do not associate with AM fungi but do so with a related group of fungi belonging to the Mucoromycotina (Field et al., 2015c). It is not yet clear whether the ancestral association that the first land plants had with fungi were with the Glomeromycotean AM fungi, with the Mucoromycotina, or with the common ancestor of these lineages. Studies in the basal liverwort lineages to determine whether the homologs of the symbiotic toolkit function in the basal liverwort-Mucoromycotina symbiosis would shed further light on the early evolution of plant-fungal symbioses.

The discovery of orthologs of the symbiotic toolkit in the charophytes opens up questions about the function of these genes in these algae, which are not known to form any symbioses with fungi. The long-standing suggestion that aquatic algae may form some primitive associations with fungal-like organisms is a possibility and warrants further investigation (Pirozynski and Malloch, 1975). It could also be that these genes are functioning in processes that are not directly related to symbiosis. For example, CCaMK and DMI1 could be playing a role in general calcium signalling that is known to be important for a variety of physiological and biochemical functions. CERK1 and the SYMRK homolog, both being receptor kinases, could be involved in microbial perception either related to defence or other biotic interactions. The GRAS protein homologs could be involved in any number of processes related to development as has been shown for their angiosperm counterparts (Bolle, 2004; Hirsch and Oldroyd, 2009). With the upcoming development of genetic tools for plants in these algal lineages, the exact roles of these genes in these organisms can be tested and will provide further insight into the processes leading to the evolution of the symbiotic toolkit in plants (Delwiche and Cooper, 2015; Domozych et al., 2017).

2.4.2. Phylotranscriptomics provides novel insights into the evolution of symbiosis

The cost and effort required to obtain high-quality genomes have made analyses spanning large taxonomic distances very difficult (Denton et al., 2014). Transcriptomes, on the other hand, are generally less complex than whole genomes and are much easier to obtain. The major limitation of transcriptomic studies has been the inherent uncertainty associated with transcriptomic datasets – unless genes are expressed in the condition that the sequenced samples were collected in, they will not

be captured using transcriptomics. Recent studies across the tree of life have shown that transcriptomics can be effectively used to quickly obtain insights about the evolution of previously unexplored clades (Wickett et al., 2014; Janouškovec et al., 2017). However, these studies have employed analyses on housekeeping genes that are reliably expressed across a range of conditions in many organisms. Thus, the value of transcriptomics for evolutionary studies (phylotranscriptomics) of genetic pathways with spatial or temporal specificity has been unclear.

Here, using transcriptomic datasets from across the green lineage, I have studied the evolution of the symbiotic toolkit. Interestingly, even though some of these genes are transcriptionally regulated with spatial and temporal specificity to symbiotic organs and processes, these genes were picked up through transcriptomics of non-symbiotic tissues of plants. Indeed, previous studies have suggested that increasing sequencing depth can offset some of the limitations of using transcriptome sequencing to obtain insights into organisms for which genomic resources do not exist (Sims et al., 2014). The major limitation of using data from transcriptomic sources is that it is impossible to prove absence of genes solely based on these data and it is imperative that observations of absence are tested as and when more genomic data for the clades where these observations are made become available. For example, in the current study, the observation that homologs of certain symbiosis genes are absent in the chlorophytes and charophytes was made using both currently available genomic and transcriptomic data and this will warrant revisiting as and when new genomic data becomes available for these clades.

2.4.3. Improved phylogenetic sampling challenges previously favoured hypotheses about the evolution of symbiosis genes

The vast majority of previous phylogenetic analyses of genes in the symbiotic toolkit have been conducted exclusively on the angiosperm lineage (Delaux et al., 2013b). Furthermore, some of these studies were conducted at a time when very few plant genomes were publicly available. Based on the data available at the time, these studies have proposed hypotheses about the evolution of the symbiotic toolkit. As more sequence data such as those from the 1000 plants project are becoming available, the hypotheses proposed by earlier studies warrant revisiting. Here, I revisited specific hypotheses about the evolution of CCaMK and SYMRK. In both these cases, the

addition of sequence data from novel taxonomic clades led to novel insights into the evolution of these proteins. In the case of CCaMK, it was found that the previously proposed hypothesis about CCaMK being a plant-specific protein can be rejected and that CCaMKs possibly evolved in the common ancestor of all plastid-containing organisms (the archaeplastida). With SYMRK, it was found that the widely accepted hypothesis that a novel domain architecture evolved in the rosids and paved the way for SYMRK to be recruited into bacterial endosymbiosis from its fungal symbiotic function does not hold true upon inclusion of SYMRK orthologs from across the land plant lineage (Markmann et al., 2008). It was found that this domain architecture is not a novelty but represents the ancestral state in the land plant lineage. The variation in domain architecture previously observed is due to loss of specific domains in various clades of the angiosperms and this was not previously inferred due to missing data. Together, these results point to the value of revisiting previous phylogenetic analyses to investigate whether the conclusions inferred from these studies are still justified upon reanalysis using larger and denser taxonomic data.

3

Establishment of a Liverwort Model System for the Study of Plant-microbe Interactions

Chapter 3: Establishment of a Liverwort Model System for the Study of Plant-microbe Interactions

3.1. Introduction

Our knowledge of the AM symbiosis in plants is derived mostly from studies into the AM symbiosis in angiosperms (Delaux et al., 2013b). The fossil evidence (Remy et al., 1994; Taylor et al., 2005; Strullu-Derrien et al., 2016) and the occurrence of symbiotic associations in the major lineages of land plants (Wang and Qiu, 2006) suggest that the AM symbiosis likely evolved in ancestral land plants. Although the occurrence of AM associations in non-flowering plants has been reported, the question of whether these represent *bona fide* symbiotic associations remained unexplored until recently. Using the complex thalloid liverwort *Marchantia paleacea*, it has been shown that associating with AM fungi increases the plant dry mass and enhances phosphorus and nitrogen uptake from the soil proving that the association formed by this liverwort with AM fungi is a *bona fide* mutualistic symbiotic association (Humphreys et al., 2010). Following this discovery, similar studies into AM symbiotic associations formed by other non-flowering plants have found that these relationships are also symbiotic in nature (Field et al., 2012; Field et al., 2015b). Studies into the genetic pathways regulating the AM symbiosis in angiosperms were largely made possible by the existence of genomic resources (Sato et al., 2008; Young et al., 2011) and the development of tractable molecular methods allowing for the discovery and functional analysis of angiosperm genes involved in the AM symbiosis (Handberg and Stougaard, 1992; Huguet et al., 1995). Such studies have not been possible outside the angiosperm lineage as these genomic resources and molecular tools were yet to be developed for a non-flowering plant (Rensing, 2017) (Figure 3.1). Indeed, no reliable non-flowering plant model system exists for the study of the AM symbiosis. Currently, the only non-flowering plant model with established tractable transformation methods and a publicly available high-quality genome sequence is the moss *Physcomitrella patens* (Rensing et al., 2008). As mosses do not associate with AM fungi (Wang and Qiu, 2006), *P. patens* cannot be used as a model to study the evolution of the AM symbiosis.

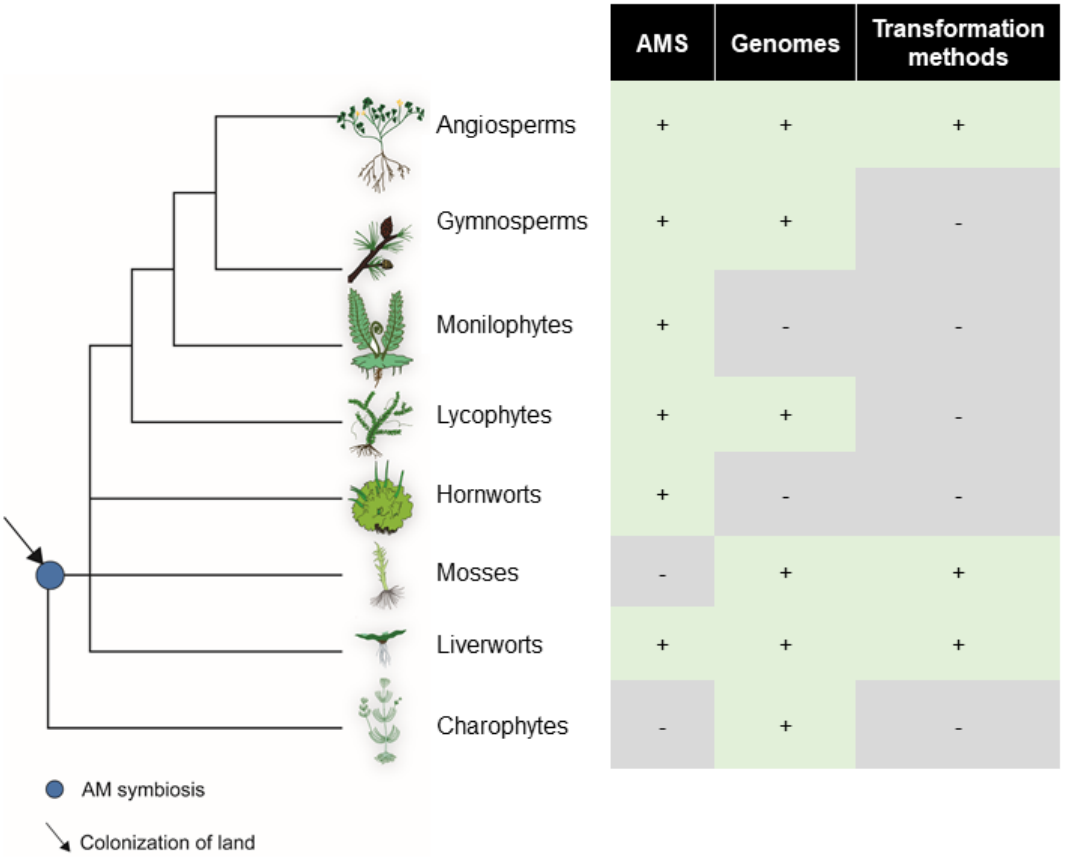

Figure 3.1. Phylogenetic representation of the streptophyte lineage along with the resources available for species in the given clades.

The AM ability and the availability of sequenced genomes and transformation methods for species in each clade are presented alongside the tree.

The aim of the present study was to establish a suitable non-vascular plant model system to study AM symbioses outside the angiosperm lineage. For this purpose, I chose to focus on the liverwort lineage due to the availability of recently developed genomic resources and molecular methods for the liverwort *Marchantia polymorpha*. The aim was to establish, in a liverwort, the key resources and techniques that have enabled successful studies of the AM symbiosis in the angiosperms. These include:

a) Reliable methods for AM colonisation in the laboratory
b) *In vitro* culturability
c) Stable integration of foreign DNA into the plant genome
d) Genomic and transcriptomic resources that allow gene discovery

e) Methods to disrupt genes and generate mutants
f) An established fungal pathogen for comparative analysis between symbiosis and pathogenesis

3.2. Results

3.2.1. Establishment of an *in vitro* AM symbiosis in a *Marchantia* species

To study the AM symbiosis in any plant species, the ability to reliably reproduce the plant- AM fungal association *in vitro* is key. The major limitation to this is the obligate biotrophic lifestyle of AM fungi – the AM fungi need the association with their plant hosts to grow and cannot be propagated in pure culture (Wewer et al., 2014). To overcome this, numerous techniques for the propagation of AM fungi have been developed using different plant hosts and in different growth conditions and media (Fortin et al., 2005). For my experiments, I made use of two of these commonly used systems - carrot hairy roots harbouring AM fungi, sterile spores of AM fungi obtained commercially. In addition to the *in vitro* tests, I also made use of a soil-based system containing chive roots that were previously inoculated with AM fungi. In all the above cases, the AM fungus used was *Rhizophagus irregularis* DAOM197198, for which the genome sequence is publicly available (Tisserant et al., 2013). Using these three approaches, I tested whether gametophytes of various *Marchantia* species can be colonised by AM fungi.

M. polymorpha is now a well-established model system for the study of liverworts. Considering the genetic and genomic resources available in this species, we felt it optimal to work with a close relative of *M. polymorpha*. This species has been taxonomically divided into three subspecies – *ruderalis*, *polymorpha* and *montivagans* (Bischler-Causse and Boisselier-Dubayle, 1991). The subspecies for which the genome sequence is available is *ruderalis*. Studies into the occurrence of AM fungi in liverworts have previously found that among these subspecies, only *montivagans* is colonised by AM fungi (Ligrone et al., 2007). To test whether any of the *M. polymorpha* subspecies could support the AM symbiosis in the laboratory and to select a suitable model system for the study of the AM symbiosis in liverworts, I tested the AM symbiotic ability of *M. polymorpha ssp. ruderalis, M. polymorpha ssp. montivagans* and *M. paleacea. M. paleacea* was included in this experiment as the

AM association in this species has been previously described in detail (Humphreys et al., 2010). The colonisation of the thalli was observed using ink staining methods that were originally developed for quantifying the AM symbiosis in angiosperm roots but have been adapted for other plant lineages. For the *M. polymorpha* subspecies, no AM fungal structures were observed in the thalli in any of the conditions used. On the other hand, *M. paleacea* was reliably colonised in all three experimental systems. (Table 3.1).

Species	Using carrot hairy roots	Using fungal spores	Using chive roots
M. polymorpha ssp. ruderalis TAK1	0/5	0/15	0/9
M. polymorpha ssp. ruderalis TAK2BC3	0/4	0/21	0/13
M. polymorpha ssp. ruderalis CAM1	0/5	0/12	0/8
M. polymorpha ssp. ruderalis CAM2	0/5	0/13	0/12
M. polymorpha ssp. montivagans	0/4	0/10	0/7
M. paleacea	5/5	19/19	22/22

Table 3.1. *In vitro* mycorrhization experiments in Marchantia species using the AM fungus *Rhizophagus irregularis*.

Strains of *Marchantia polymorpha ssp. ruderalis*, *Marchantia polymorpha ssp. montivagans* and *Marchantia paleacea* were used for the experiments. The experiments were conducted either on sterile tissue culture medium or soil depending on the type of inoculum. Carrot hairy-root cultures harbouring AM fungi or fungal spores were inoculated on plates. For the experiments on soil, chive roots harbouring AM fungi were used as the inoculum. Colonisation of AM fungi was observed by sectioning and ink staining of the thalli at 6 wpi. Only *M. paleacea* was colonised by AM fungi in these experiments. The numbers in the table represent the number of independent experiments conducted. For each experiment, 3 plants were tested.

The progression of AM colonisation was studied further in the *M. paleacea* thalli. As previously reported, colonised thalli accumulated a characteristic dark pigment on the underside of the thallus around the midrib (Humphreys et al., 2010). This dark pigment was specifically found in the colonised thalli but not in the non-colonised thalli (Figure 3.2a,b). Sectioning and ink staining of the colonised thalli revealed that the red pigment accumulated around colonised cells (Figure 3.2c-f). The fungal colonisation initiated at the lower epidermis and proceed through the parenchymatous tissue. On the upper region of the parenchymatous tissue, arbuscules (Figure 3.2g,h) were formed just below the chlorophyllous tissue.

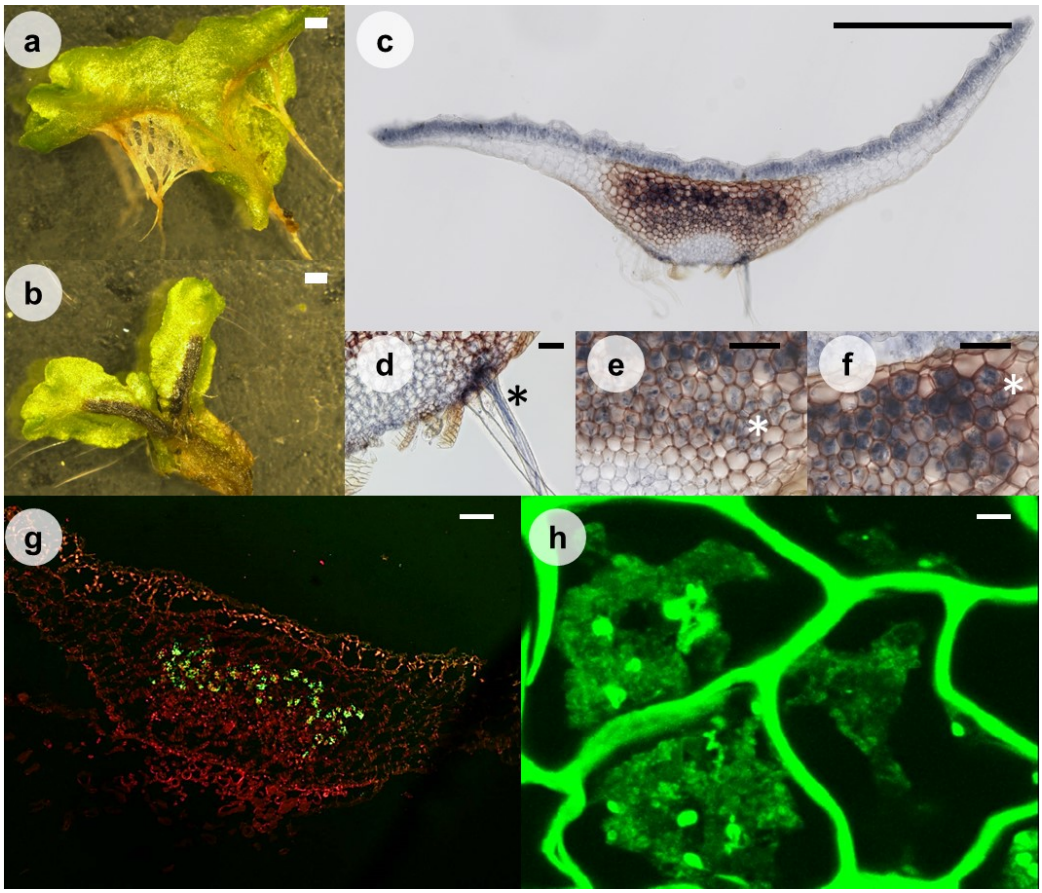

Figure 3.2. Imaging mycorrhizal colonisation in the liverwort *M. paleacea* 8 weeks post-inoculation with the AM fungus *R. irregularis*.

Underside of thalli from (a) non-inoculated and (b) inoculated *M. paleacea* plants showing specific accumulation of brownish pigment near the midrib of inoculated plants. (c)-(f) Ink staining of transverse sections of *M. paleacea* thalli to stain the fungus blue. The fungus was found to (d) infect and enter the thallus through the lower epidermis (e) and colonise

the parenchymatous tissue below the upper epidermis (f) forming arbuscules in these cells. The structures described are marked with asterisks. (g -h) Paraffin sectioning and WGA staining of colonised thalli was carried out following fixation of the tissue to image arbuscules. (g) A transverse section of a colonised thallus stained with propidium iodide and WGA. (h) Magnification of a transverse section of thallus colonised by the fungus showing the structure of an arbuscule. Scale bars: (a)-(c) 1 mm; (d) 50 µm; (e)-(g) 100 µm; (h) 5 µm. While I developed the ink staining protocol in collaboration with Pierre-Marc Delaux, the images of ink-stained thalli (c)-(f) were provided by Nicolas Vigneron and Pierre-Marc Delaux.

3.2.2. Developing methods to establish *M. paleacea* as a laboratory model to study the AM symbiosis

As none of the *M. polymorpha* subspecies were colonised by AM fungi under the conditions tested here, this species could not be used as a model to study the AM symbiosis in liverworts. By contrast, *M. paleacea* was found to be reliably colonised in all the experimental conditions tested. Therefore, the methods for routine culture and genetic transformation previously established for the manipulation of *M. polymorpha* ssp. ruderalis (Ishizaki et al., 2016) were adapted for *M. paleacea*.

3.2.2.1. *In vitro* culture of *M. paleacea*

Using the conditions for routine culture previously developed for *M. polymorpha*, it was found that *M. paleacea* cultures could also be maintained using these conditions. Both gemmae and thallus cuttings could successfully be used to initiate cultures as in the case of *M. polymorpha* but side-by-side growth comparisons indicated that the *M. paleacea* thalli did not grow as well as their *M. polymorpha* counterparts in these conditions (Figure 3.3).

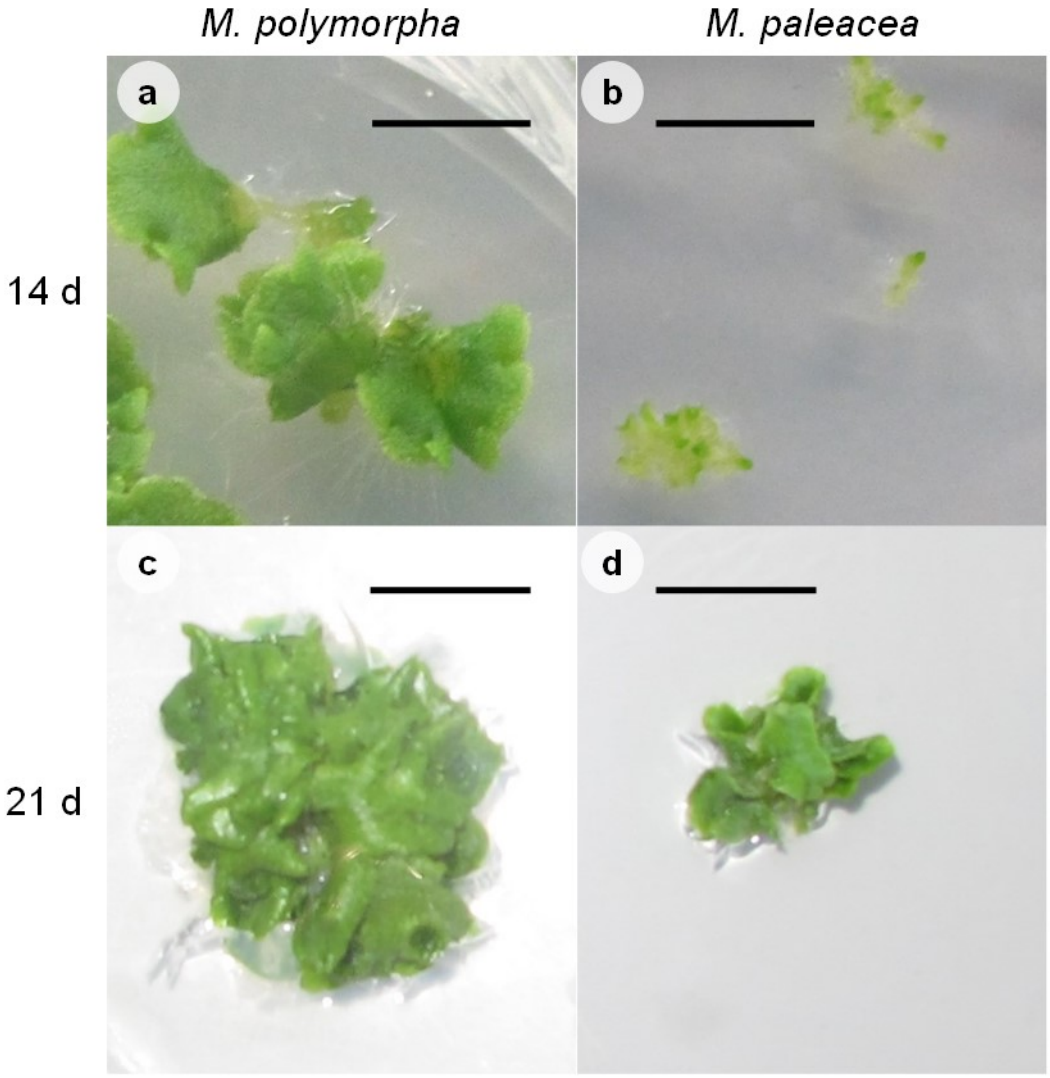

Figure 3.3. Comparing growth of *M. polymorpha* and *M. paleacea* gametophytes on sterile tissue culture media.

M. polymorpha and *M. paleacea* plants were grown on half-strength Gamborg's B5 medium from gemmae for 14 and 21 days. Scale bars: 1 cm.

The observed difference in growth could be due to inherent differences in these two species and indeed *M. polymorpha* has been reported to reach larger sizes (thallus width: 8-15 mm) (Shimamura, 2016) than *M. paleacea* (thallus width: 5-8 mm) (Borovichev and Bakalin, 2014). On the other hand, the *in vitro* grown *M. paleacea* thalli were markedly different in appearance (Figure 3.3) from those grown on soil (Figure 3.2a,b). Therefore, it was hypothesized that the growth medium used may not support optimal development of *M. paleacea*. The growth medium used, Gamborg's B5 medium was originally developed for the maintenance of calli and cell suspension

cultures of angiosperm species such as *Glycine max* and the composition was optimised specifically for angiosperm tissue culture (Gamborg et al., 1968). Previous studies have pointed at liverworts having nutritional requirements that are different from those of the angiosperms (Fonseca et al., 2009). To test if optimal growth of *M. paleacea* could be achieved on growth media other than Gamborg's B5, several well-known growth media used for plant tissue culture such as Minimal (M), FP (Fahreus Plant) and Strullu-Romand Variant (SRV) media. From these tests, it was observed that there were noticeable differences in the growth of *M. paleacea* thalli in these different media. To quantify the growth of the thalli, surface area was measured from microscopy images using the Easy Leaf software (Easlon and Bloom, 2014). From these measurements it was revealed that the plants grew best (as measured by thallus surface area) in the SRV medium (Figure 3.4) that has previously been used to successfully establish co-cultures of the liverwort *Lunularia cruciata* with AM fungi (Fonseca et al., 2009).

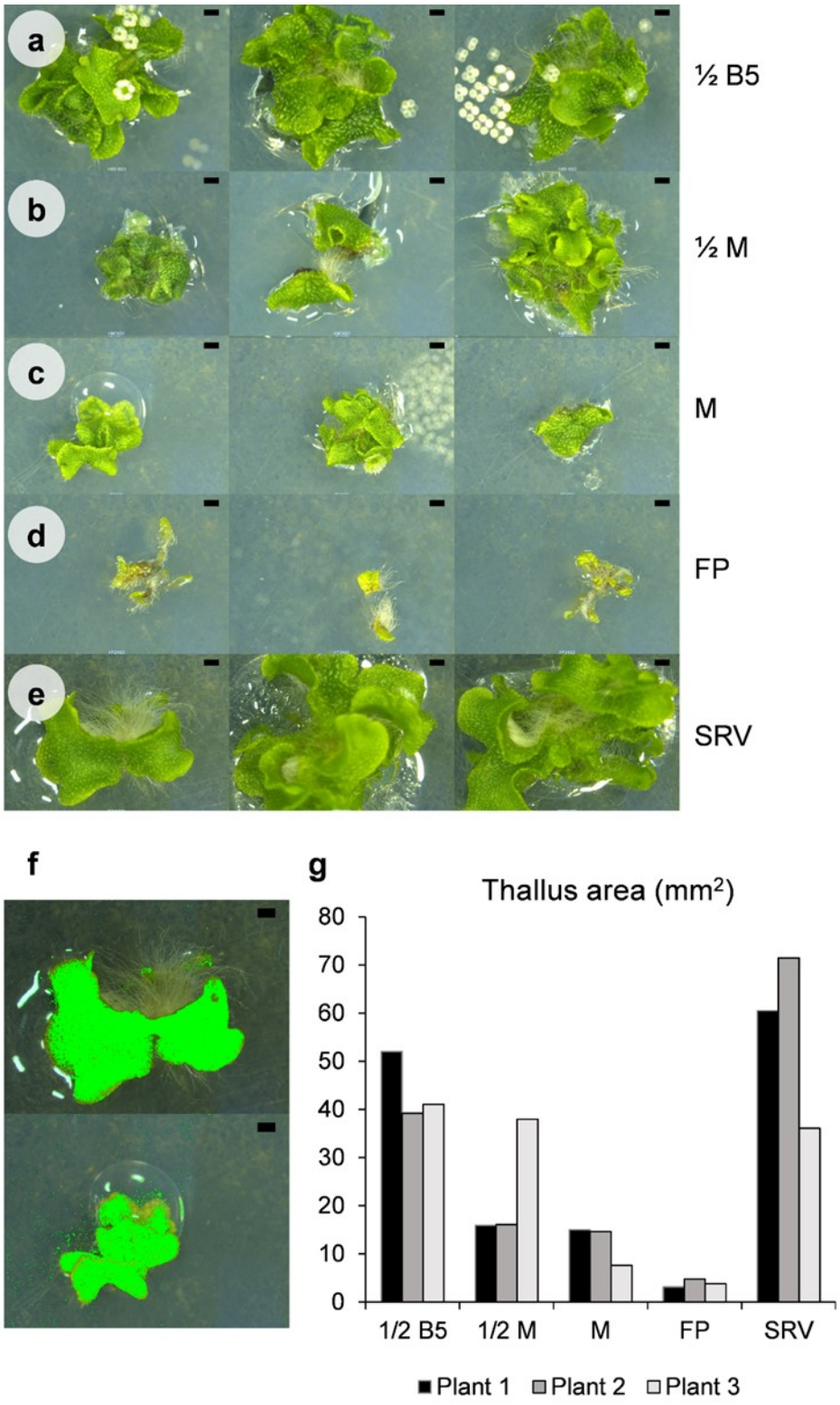

Figure 3.4. Comparison of growth media for culturing *M. paleacea*.

Different tissue culture media were tested for growing *M. paleacea*. Gemmae were plated and grown for 2 weeks on (a) half-strength Gamborg's B5 (b) half-strength M (c) M (d) FP and (e) SRV media. (f)

The surface area of the thalli was measured using the image analysis tool EasyLeaf. (g) Thallus area measurements (mm^2) of thalli grown on different media using EasyLeaf (Easlon and Bloom, 2014). Measurements were made on 3 independent plants grown on individual plates for each treatment. As observations were made only for 3 plants, statistical tests were not performed as the data do not provide appropriate statistical power and the raw data are presented. These results will need to confirmed in the future using more replicates.

3.2.2.2. Induction of the *M. paleacea* sexual cycle

With the protocols for the culture and maintenance of the gametophyte generation of *M. paleacea* in place, methods to induce the transition into the sporophyte generation were explored. For *M. polymorpha*, supplementation of far-red light to the normal growth conditions has been reported to be sufficient to induce this transition to the sexual cycle (Ishizaki et al., 2008). Therefore, similar experiments were conducted to test if this is also the case for *M. paleacea*. Plants were grown in conditions previously reported to successfully initiate the sporophyte formation in *M. polymorpha* on soil (Ishizaki et al., 2008) as well as on plates (Althoff et al., 2014). As with the experiments on mycorrhizal colonisation, both *M. polymorpha* and *M. paleacea* plants were used for these experiments. Both on soil and on plates, far-red supplementation was sufficient for the induction of the sporophyte for the *M. polymorpha* plants used. By contrast, sporophyte induction was not observed on any of the *M. paleacea* plants tested (Table 3.2).

Species	On soil	On plates
M. polymorpha ssp. ruderalis TAK1	+	+
M. polymorpha ssp. ruderalis TAK2BC3	+	+
M. polymorpha ssp. ruderalis CAM1	+	+
M. polymorpha ssp. ruderalis CAM2	+	+
M. polymorpha ssp. montivagans	+	+
M. paleacea	-	-

Table 3.2. Sexual cycle induction of Marchantia species *in vitro*.

Far-red supplementation was used for the induction of sexual structures on the gametophytes of different *Marchantia* species. Thalli were grown on soil and on tissue culture media as previously described (Ishizaki et al., 2008; Althoff et al., 2014). All *M. polymorpha* thalli formed sexual structures while none of the *M. paleacea* thalli did. For each *Marchantia* accession tested, 3 independent pots (containing ~100 thalli) were used for the soil tests and 3 independent plates (containing 3-5 thalli) were used for the tests performed on tissue culture media.

3.2.2.3. *Agrobacterium tumefaciens* mediated genetic transformation of *M. paleacea*

Agrobacterium-mediated transformation protocols developed for *M. polymorpha* have enabled studies in this species that have provided insights into the evolution of various processes such as plant growth (Kato et al., 2015), development (Honkanen et al., 2016; Proust et al., 2016) and photosynthesis (Shimakawa et al., 2017). These studies were made possible by the development of a spore-based transformation method reported by Ishizaki and colleagues (Ishizaki et al., 2008). This method was found to yield a large number of *M. polymorpha* transformants (500-600) from a

single experiment and its success was attributed to the fact that a large number of spores could be used as the starting material for the transformation. This was enabled by the organisation of *M. polymorpha* spores into self-contained structures called sporangia, each of which contains ~10^5 spores. In the method used by Ishizaki and colleagues, for each transformation experiment, one sporangia was suspended into the liquid transformation medium and co-cultured with *A. tumefaciens*. The sporelings generated from the ~10^5 spores were then transferred on to selective media and 500 to 600 positive transfomants are obtained from selection.

As *M. paleacea* could not be induced to produce sporophytes, a similar transformation protocol to the one described above could not be used. Other studies have reported that vegetative *M. polymorpha* tissue could also be transformed using *A. tumefaciens* albeit at a much smaller scale than the spore-based transformation method(Kubota et al., 2013; Tsuboyama-Tanaka and Kodama, 2015). These methods employed regenerating thalli from cuttings or gemmalings as the starting material for the transformation. I tested these methods on *M. paleacea* to see if transformants could be obtained.

The thallus-cutting method developed for *M. polymorpha* (Kubota et al., 2013) requires the manual cutting of individual thalli grown from gemmae to expose the meristematic cells and is labour-intensive. Therefore, this method cannot be used to generate large numbers of transformants like the spore-based transformation method. To circumvent this limitation of the thallus-cutting method, a modification of this method was devised for the transformation of *M. paleacea*. Instead of cutting individual thalli, thalli were aggregated and blended in a laboratory blender containing the transformation medium. The resultant chopped suspension was used for co-culturing with *A. tumefaciens* and following this, the same steps as those previously reported for the *M. polymorpha* thallus-cutting transformation were followed. 4-6 weeks after the plating of the transformation suspension onto selective medium, putative transformants were obtained (Figure 3.5). Plants that passed the initial antibiotic selection screening were also screened for fluorescence to ensure successful integration of the transfomation vector and as fluorescence could be observed for all of the plants that passed the antibiotic selection, it was assumed that no false positives had escaped the screening process. Successful integration of T-DNA into the *M. paleacea* genome in these putative transformants will need to be tested using molecular methods in the future. To check whether the Agrobacterium strains used

had any effects on the transformation efficiency, comparisons were made between the Agrobacterium strains AGL1 and GV3101 and these revealed that the GV3101 strain yielded higher transformation efficiency (Figure 3.5). Initial tests for the development of this transformation protocol were conducted in 125 ml flasks using 25 ml of liquid media, the same as that used for the transformation of *M. polymorpha*. Once this protocol was established, I tested whether the method would be amenable to scale-up and found that doubling the transformation volume yielded similar transformation efficiencies. Furthermore, to test whether scaling-down would also be possible, I adapted this protocol to 2 ml 96-well plates and found that positive transformants could be obtained using co-culture volumes as small as 400 µl. While similar transformation efficiencies (20-30%) were observed for both large and small volumes, the total number of transformants obtained from a single co-culture was much more for the larger volumes as more starting thallus material could be used for these compared to the smaller volumes (Figure 3.5). It was also observed that the quality of starting material influenced the transformation efficiency. The use of healthy-looking (4-6 week old) thalli yielded reliable efficiencies but use of old thalli (>6 weeks) as the starting material for the transformation drastically reduced transformation efficiences (data not shown).

Although the blend-based transformation method yielded large numbers of transformants and could be adapted to either large or small volumes depending on the number of transformants required, this protocol required that the thalli be pre-cultured for 4-6 weeks from gemmae before they could be used for transformation. To test whether this lead time to transformation could be circumvented, I tested a method previously developed for *M. polymorpha* where 3 to 5-day old gemmalings were transformed successfully (Tsuboyama-Tanaka and Kodama, 2015). This method was termed “agartrap” transformation due to the different transformation steps performed on gemmalings trapped in solidified media containing agar. Trials of this method revealed that the only modification necessary to adapt this protocol to *M. paleacea* was an increase in the initial growth period for the gemmae before transformation. Following this modification, the same protocol used for the transformation of *M. polymorpha* yielded positive transformants of *M. paleacea*. Quantification of successful transformation events revealed that the agartrap transformation method had a higher transformation efficiency than the thallus-blending transformation (Figure

3.5) but was comparatively more labor-intensive as it involved manually placing each gemma on to plates before transformation could be conducted.

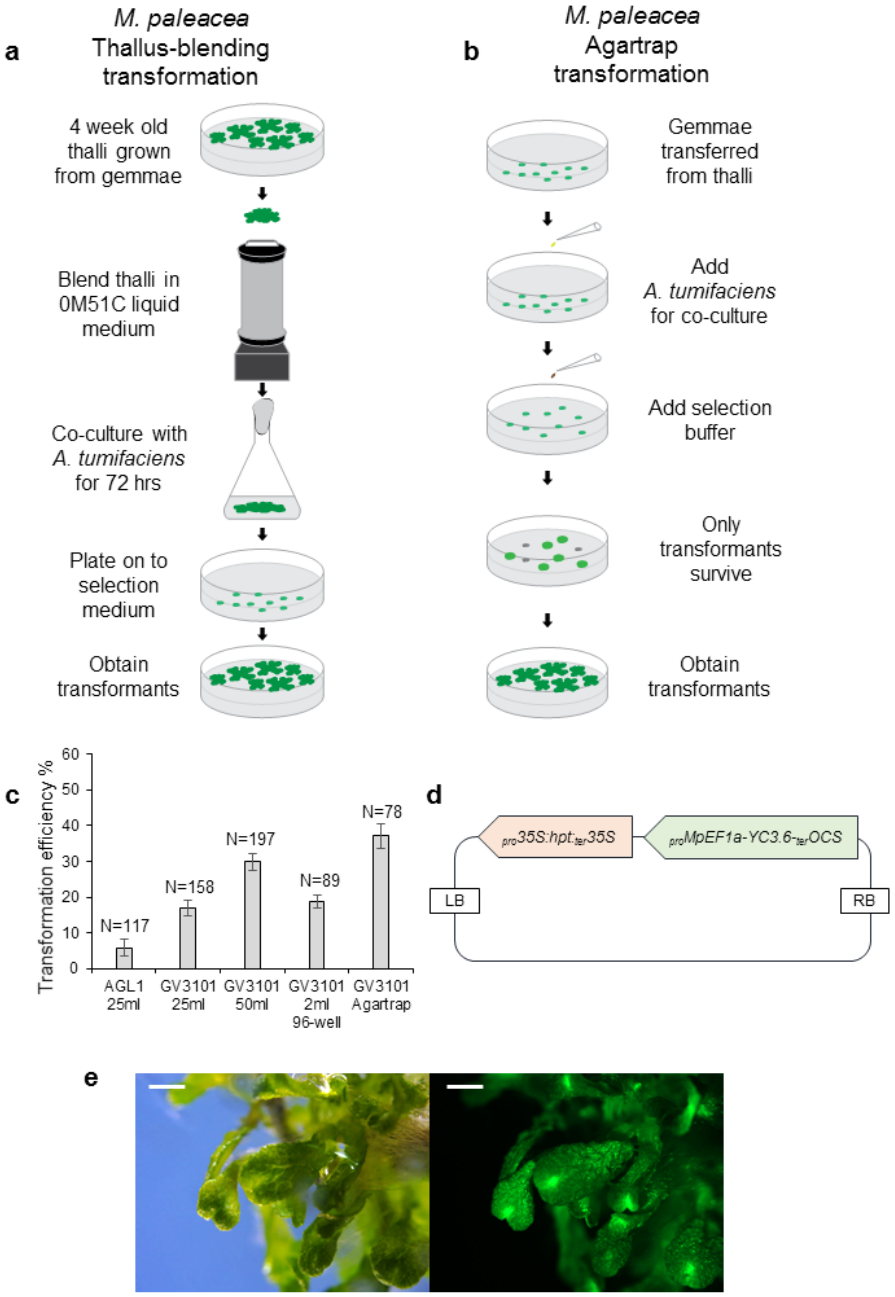

Figure 3.5. Agrobacterium-mediated transformation of *M. paleacea*.

(a) Workflow describing the thallus-blending transformation protocol for *M. paleacea*. (b) Workflow describing the agartrap transformation protocol for *M. paleacea*. (c) Efficiency of the thallus-blending and agartrap protocols for transformation of *M. paleacea* were measured. All transformation experiments were conducted in batches of three to calculate transformation efficiency. The total thalli pieces/gemmae (N) that were transformed across the three batches using each setup are indicated. Transformation efficiency (%) was calculated by measuring the

number of live plants exhibiting fluorescence in the Green Fluorescent Protein (GFP) wavelength after 4 weeks of growth on selective medium following transformation. The *A. tumefaciens* strains AGL1 and GV3101 were tested using 25 ml of transformation medium in 125 ml flasks in 4 flasks each. GV3101 was also tested using 50 ml of transformation medium in 250 ml flasks as well as 400 µl in 2 ml 96-well plates. Agartrap transformation was conducted using the GV3101 strain alone. (d) Vector map of the construct used for the transformation experiments. The vector was constructed using pICH50505 (Icon Genetics) and contains two GoldenGate level-1 modules: a hygromycin resistance module driven using the 35S promoter and a fluorescent marker driven using the *M. polymorpha* EF1a promoter. (e) Bright-field and epifluorescence images of *M. paleacea* transformants showing expression of fluorescent proteins. Scale bars: 1 mm.

3.2.2.4. Fungal pathogen of *M. paleacea*

Recent molecular genetic studies have revealed that angiosperms use similar mechanisms and employ members of the same gene families to detect both symbiotic and pathogenic fungi (Zipfel and Oldroyd, 2017). The evolution of these mechanisms remains relatively unexplored and the question of whether basal land plant lineages also possess the ability to sense and distinguish between symbiotic and pathogenic fungi remains unanswered. To address this, pathogenic fungi that infect basal land plants such as the liverworts need to be identified. Although reports on fungal pathogens infecting mosses exist, no liverwort fungal pathogens have been reported to-date (Davey and Currah, 2006).

Using samples of *M. polymorpha* collected from the wild, our collaborators at Duke University were able to isolate endophytic fungi that had positive or negative growth effects on *M. polymorpha* (Jessica Nelson, Duke University, personal communication). From these experiments, they identified an Ascomycete fungus, *Trichoderma virens* was identified that infected and killed *M. polymorpha*. I obtained a sample of this fungus *T. virens* from a fungal culture collection to test whether this fungus could also infect *M. paleacea.* Inoculation of *M. paleacea* thalli with the fungal cultures resulted in the death of the thalli. The progression of the infection was accompanied by yellowing followed by the death of the thalli suggesting that infection

by the fungus led to the necrosis of the plants (Figure 3.6a-d). SEM imaging of the plant surface infected by the fungus suggested that the fungus grew on both the thallus and the rhizoids (Figure 3.6e and f). Further imaging and experiments such as cell-death staining are required to understand the mode of infection of the fungus and the response of *M. paleacea* to infection by *T. virens*.

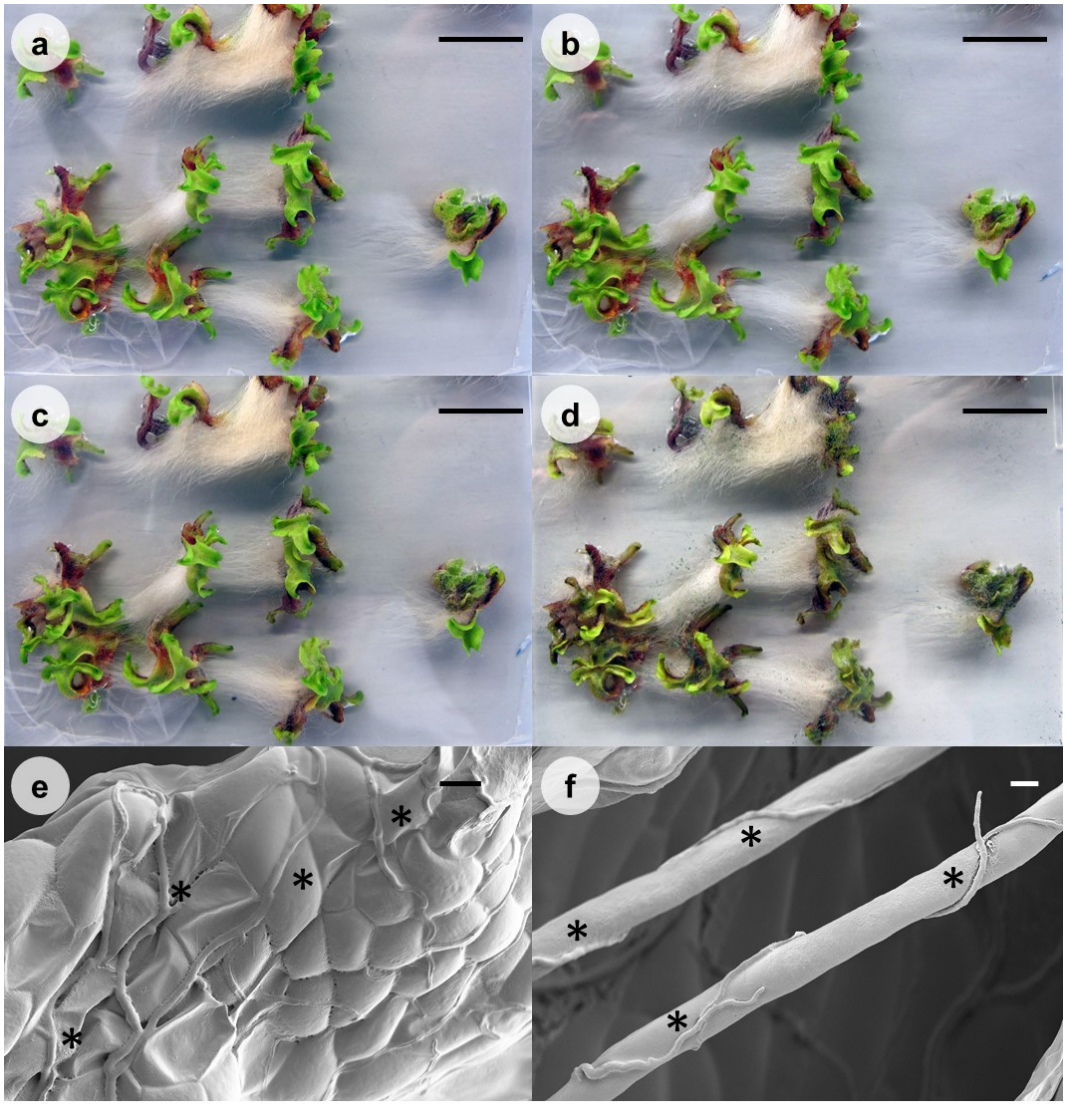

Figure 3.6. Infection of *M. paleacea* by a fungal pathogen.

2-week old *M. paleacea* thalli were inoculated with *Trichoderma virens*. Pictures of the inoculated thalli were taken at (a) 2, (b) 4, (c) 6 and (d) 8 days post-inoculation. At 8 days post-inoculation the fungi had grown over the plants and the plants showed signs of necrosis such as yellow colouration. SEM images of the infected plants showed the fungus, indicated using asterisks, colonised both (e) thallus and (f) rhizoid

surfaces. Scale bars: (a) – (d) 1 cm; (e) 20 µm; (f) 10 µm. SEM imaging was performed by Elaine Barclay (John Innes Centre, UK)

3.2.3. Developing genomic resources for *M. paleacea*

To set up genomic resources for *M. paleacea*, short-read Illumina sequencing data was generated using genomic DNA and RNA extracted from *M. paleacea* gametophyte tissue. For the genome sequencing, 218 million 100bp paired-end reads with an average insert size of 336 bp and 74 million 100bp mate-pair reads with an average insert size of 4311 bp were generated. For the transcriptome sequencing, 10 million 300bp paired-end reads were generated from a normalised library subjected to size selection for the removal of fragments smaller than 350bp and those larger than 600bp. Using these reads, I constructed numerous genome and transcriptome assemblies following the workflow detailed below (Figure 3.7).

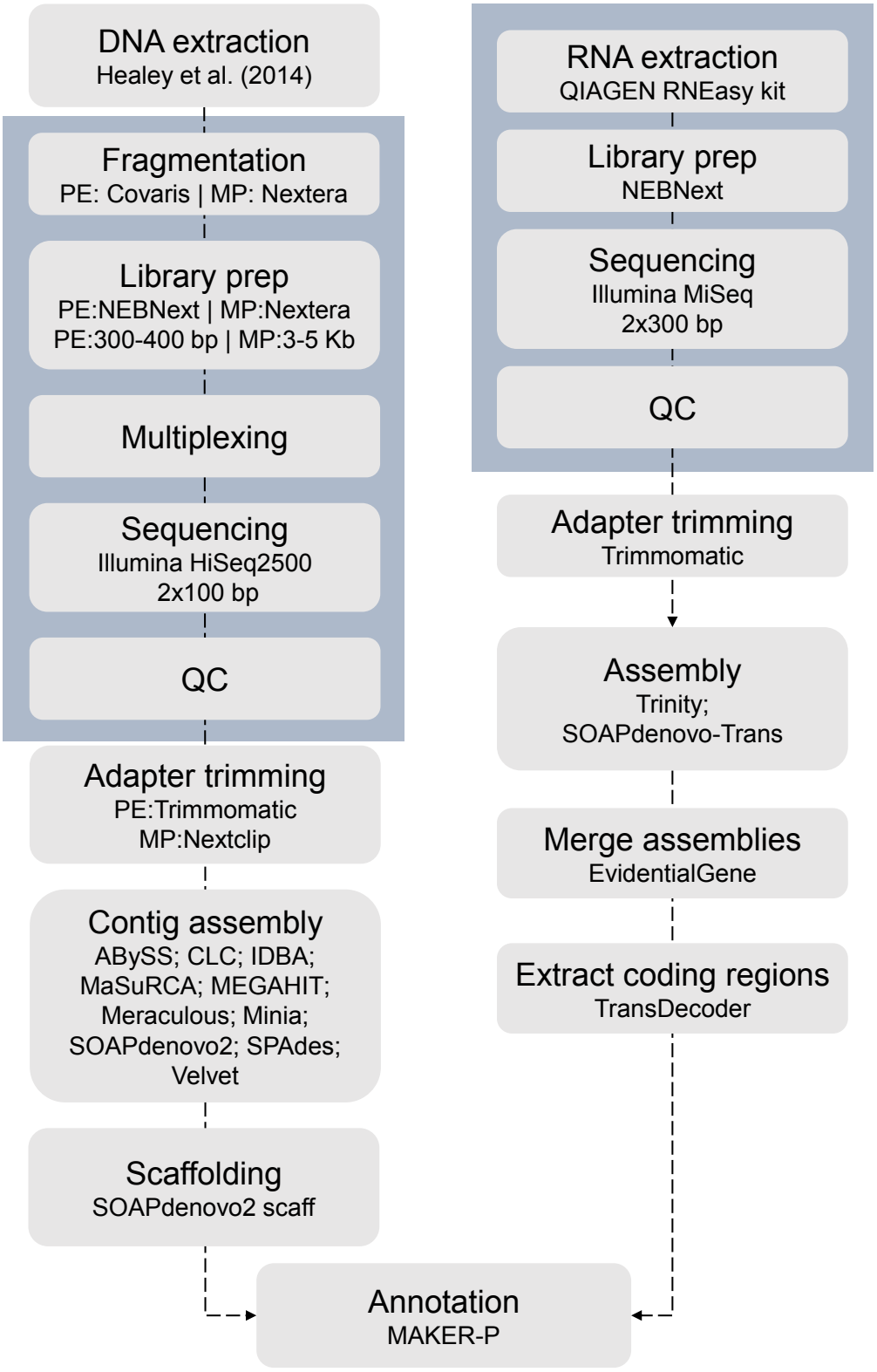

Figure 3.7. Workflow used for the assembly and annotation of the *M. paleacea* genome and transcriptome.

For the genome assembly, the aim was to find the optimal assembly that was most contiguous and complete. To measure contiguity, traditional measures such as the

contig N50 statistic were used. For measuring the completeness, I used a combination of approaches. To maximise the coverage of the coding regions of the genome in the assembly, the successful mapping rate of the reads from the transcriptome sequencing was used as the measure. Another measure that was used was the BUSCO score which is calculated based on the presence of homologs of genes that are conserved as single-copy genes in most sequenced plant genomes (Simão et al., 2015). Finally, the K-mer Analysis Toolkit (Mapleson et al., 2017) was used to maximise the carryover of information present in the read data to the assemblies. 28 genome assemblies were made using publicly available genome assembly tools. The tools used employ algorithms based on the de Bruijn graph data structure. Previous studies have shown that one of the main determinants of the final quality of the assembly generated is the value of k-mer size used for the de Bruijn graph construction (Chikhi and Medvedev, 2014). Therefore, numerous k-mer values were used to generate multiple assemblies with each of the assembly tools (Table 3.3a). The best assembly from these was selected based on the contiguity and completeness criteria and the contigs from this assembly were used to generate the scaffolds for version 1.0 of the *M. paleacea* draft genome assembly. The final size of this assembly was 238.61 Mb which was close to the genome size value of 241.17 Mb estimated from the read data using k-mergenie (Chikhi and Medvedev, 2014).

For the transcriptome assembly, the de Bruijn graph assemblers Trinity (Grabherr et al., 2011) and SOAPdenovo-Trans (Xie et al., 2014) were used to each produce an assembly. The two assemblies were then merged using the EvidentialGene pipeline as it has previously been shown that different transcriptome assemblers tend to have complementary results with each assembler assembling its own unique set of real transcripts that other assemblers do not assemble (Nakasugi et al., 2014). The coding sequences from the transcriptome assembly were extracted using TransDecoder (Haas et al., 2013) and used to perform the annotation of the *M. paleacea* genome. Transcriptome assemblies of two other liverworts, *Lunularia cruciata* and *Pellia endiviifolia* were also generated using publicly available read data from NCBI with the same transcriptome assembly workflow as that used for *M. paleacea* (Table 3.3b).

a

Assembler	k-mer size	% BUSCO score	% of RNAseq reads mapped	No. of contigs	N50 (Kb)	Total size (Mb)
SPAdes	61-91 (10)	59.93	78.1	150,455	26.30	235.3
MEGAHIT	91-111 (4)	59.79	78.1	61,719	18.37	243.2
MEGAHIT	51-121 (10)	60.07	78.0	68,919	17.42	240.7
IDBA	60-120 (5)	59.38	77.9	63,371	18.73	238.9
IDBA	70-100 (3)	59.38	77.8	83,169	15.82	233.6
IDBA	50-100 (10)	59.10	77.8	78,072	16.41	233.1
CLC	12-64*	59.58	77.7	27,184	19.34	203.3
ABySS	93	59.51	77.6	103,283	17.03	243.9
ABySS	91	60.14	77.5	146,116	16.76	239.8
ABySS	81	59.93	77.5	201,305	16.95	237.4
SOAPdenovo2	91	57.99	77.3	513,804	10.25	223
SOAPdenovo2	81	58.47	77.2	813,926	11.75	217.5
MaSuRCA	61	59.86	76.6	23,770	31.89	246.8
MaSuRCA	71*	60.00	76.5	21,748	33.11	248.9
SOAPdenovo2	71	58.06	76.4	1,289,170	10.85	209.2
Meraculous	81	60.14	76.2	37,007	18.13	233.8
MEGAHIT	51-93 (6)	59.65	76.2	112,091	14.24	227.3
Velvet	61	60.14	76.0	60,972	72.71	230
SOAPdenovo2	61	57.29	76.0	1,883,058	9.56	199.4
Meraculous	71	60.21	75.9	40,285	17.95	230.1
Meraculous	91	58.26	75.8	52,500	11.83	230.7
Meraculous	61	60.14	75.7	46,901	16.84	225
Velvet	81	63.82	75.6	81,774	28.46	361.3
SOAPdenovo2	51	55.97	75.5	2,570,341	7.98	188.2
Velvet	71	59.93	74.8	123,391	13.57	342.7
minia	71	57.22	74.7	196,460	10.57	205.4
minia	61	56.53	74.3	239,703	9.72	195.9
ABySS	31	56.60	72.8	1,902,268	8.31	161.8

b

	Marchantia paleacea	*Marchantia polymorpha*	*Lunularia cruciata*	*Pellia endivifolia*
Genome				
Assembly size (Mb)	238.61	205.72	-	-
Scaffolds	22669	4137	-	-
N50 length (Kb)	77.78	372.13	-	-
BUSCO score	862/1440	856/1440	-	-
Average depth	112x	64x	-	-
G+C (%)	40.3	41.1	-	-
Transcriptome				
Transcripts	35388	29453	37093	34469
Assembly size (Mb)	38.7	39.7	47.1	43.7
G+C (%)	48.3	48.9	43.8	46.4

Table 3.3. Assembly statistics for *M. paleacea* genome and transcriptome.

(a) Statistics for the various genome assemblies made. (b) Statistics for the draft genome and transcriptome assemblies of *M. paleacea* compared

to those previously published for *M. polymorpha*. Statistics are also provided for the transcriptome assemblies of *L. cruciata* and *P. endivifolia* constructed in the current study.

Basic Local Alignment Search Tool (BLAST) searches for the liverwort genes previously identified as being orthologous/homologous (Chapter 2) revealed that these genes were present in the *M. paleacea* genome. In the *M. paleacea* transcriptome, the liverwort counterparts of most genes except for *RAM1* and *VAPYRIN* were found (Figure 3.8a). As both genes are specifically transcriptionally induced during the AM symbiosis in the angiosperms, it could be that they behave similarly in the liverworts and would only be detected in transcriptomic data generated from AM colonised thalli. The material that I used for my experiments was from non-colonised *M. paleacea* thalli grown on sterile media. The assembly of the *M. paleacea* genome and transcriptome allowed for the first study of how intron-exon structures of symbiosis gene orthologs have changed during the evolution of plants. Previous studies have reported that for orthologous genes, the intron-exon structures remain relatively unchanged and are conserved over long evolutionary times (Betts et al., 2001). To test if this is the case for symbiosis gene orthologs as well, I chose *CCaMK* and *CYCLOPS* for the analysis as these are multi-exonic genes that are present as single-copy orthologs in the land plant lineage (Chapter 2). The analysis revealed that both genes have a conserved gene structure in both the liverworts and in the angiosperms with 7 exons in the case of *CCaMK* and 11 in the case of *CYCLOPS* (Figure 3.8b). Although the length of the entire genes has changed, this was mostly due to changes in the intron lengths and not in the exon lengths. The total exon lengths of *CCaMK* and *CYCLOPS* in *M. paleacea*, *M. truncatula* and *O. sativa* were 1698, 1572, 1551 bp and 1839, 1542, 1527 bp respectively.

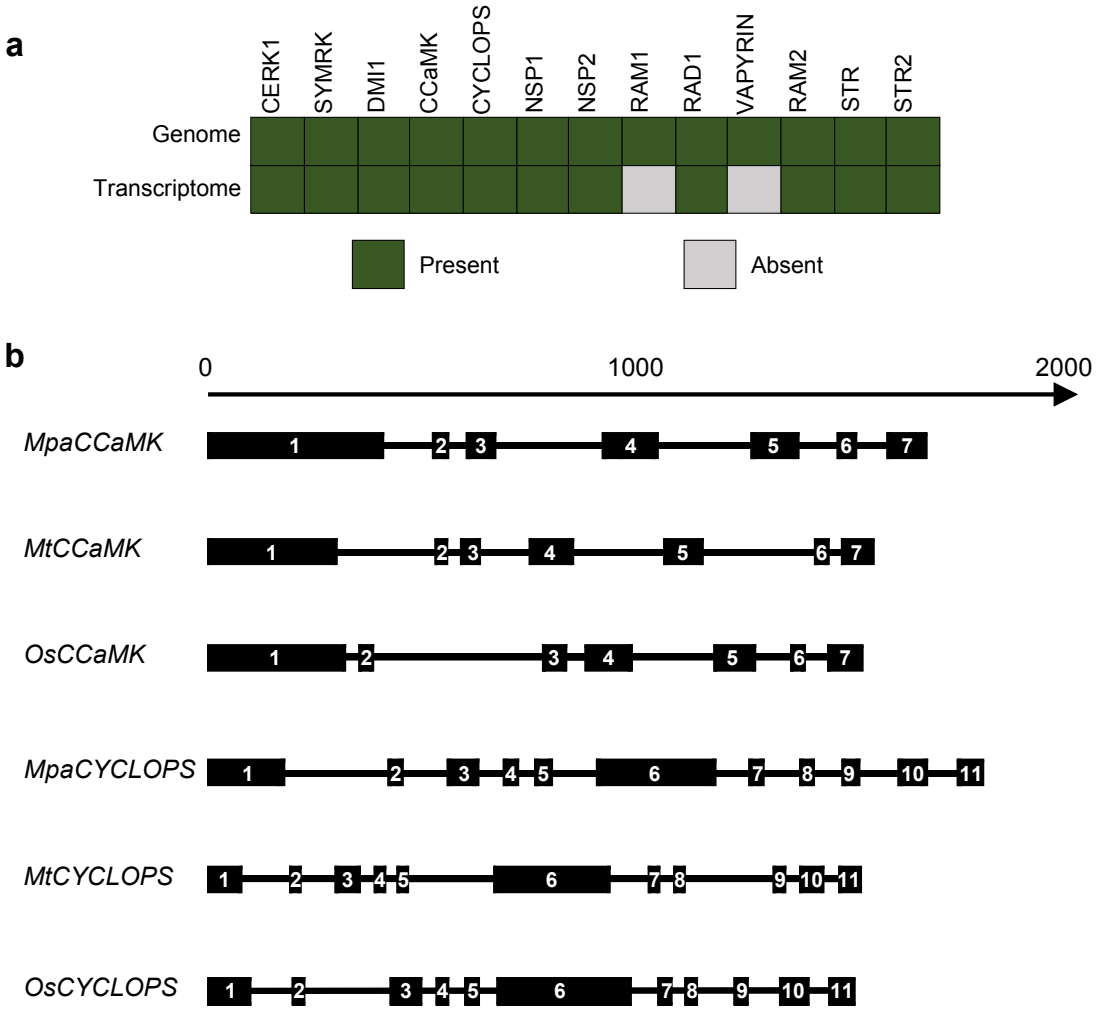

Figure 3.8. Homologs of symbiosis genes in the *M. paleacea* genome

(a) Summary of homologs of liverwort symbiosis genes discovered in the genome and transcriptome of *M. paleacea*. (b) The genome allowed for the comparison of intron-exon structures of *CCaMK* and *CYCLOPS* between the liverwort *M. paleacea*, the monocot *O. sativa* and the dicot *M. truncatula*. Conserved structures were observed for both genes in these three species. The *M. paleacea* genes are prefixed Mpa, the *M. truncatula* genes Mt and the *O. sativa* genes Os.

3.2.4. CRISPR/Cas9-based mutagenesis for reverse genetics in *M. paleacea*

Having established the resources required for transformation as well as genomic analyses in *M. paleacea*, I tested whether molecular methods used to disrupt genes

could be adapted for use in *M. paleacea*. Such methods have previously been established for *M. polymorpha* where homologous recombination (HR)-mediated gene targeting (Ishizaki et al., 2013a) and CRISPR/Cas9-mediated mutagenesis (Sugano et al., 2014) were recently shown to be effective for knocking out genes to study their functions. As the targeting efficiency of both HR and CRISPR/Cas9-mediated mutagenesis were found to be relatively low in *M. polymorpha*, I tested the utility of CRISPR/Cas9-mediated mutagenesis for generation of mutants in *M. paleacea* due to the ease with which different constructs can be designed to target different genes with a change in short stretch of sequence (~20bp) in the constructs (Belhaj et al., 2015). For this purpose, I adapted the previously published CRISPR/Cas9 system used in *M. polymorpha* to produce a GoldenGate compatible vector containing a hygromycin selection cassette at the position closest to the left border sequence of the T-DNA and a fluorescent marker (mCherry) at the position closest to the right border sequence (Figure 3.9c). The reasoning behind placing markers at both ends of the T-DNA was to allow screening for the integration of the entire T-DNA. Between these two marker cassettes, the Cas9 and U6-guide RNA modules previously used for mutagenesis in *M. polymorpha* (Sugano et al., 2014) were included with the only change being in the guide RNA target sequence. It was apparent that a gene which would produce a visually recognisable phenotype when mutated in *M. paleacea* would be ideal for testing this system. Therefore, the *NOP1* gene which has been shown to be required for air pore formation in *M. polymorpha* (Ishizaki et al., 2013b) was selected for this purpose where mutating this gene resulted in the complete absence of air pores on the thallus surface (Figure 3.9a,b). Therefore, I obtained the *NOP1* ortholog from *M. paleacea* and designed a CRISPR/Cas9 vector targeting 21bp of the *M. paleacea NOP1* gene (GATAGTCTTTGTGAGAGGAT**AGG**).

This vector was then transformed into *M. paleacea* in several independent batches to obtain a total of 1733 putative transformants following selection on hygromycin. None of the transformants exhibited the expected air pore phenotype or any observable differences in the air pore structure. While some differences in the level and pattern of fluorescence could be observed in different transformants, mCherry expression could be observed in 1687 of the transformants using fluorescence microscopy. 96 of these transformants were randomly selected and genotyping was carried out using PCR and Sanger sequencing. The results of the genotyping

confirmed that none of these transformants had mutations in the *NOP1* gene, suggesting that the system used was not successful at inducing mutations in this gene.

The observed lack of mutations could be caused by any of the following: i) the T-DNA that was integrated into the genome of *M. paleacea* might not be whole and could lack some parts of the CRISPR/Cas9 system, ii) the guide RNA target used for the gene might not be optimal, iii) the promoter driving the Cas9 or the guide RNA might not be optimal. As integration of the entire T-DNA carrying the CRISPR/Cas9 components could be confirmed for at least 1687 out of the 1733 putative transformants, the first possibility could be ruled out. To check whether the target sequence was suboptimal and was the reason behind the observed lack of mutations, two other guide RNAs designed to target the NOP1 gene were used (GACCACCGAGGTGAAGCAGT**TGG;** GGAACATTTTTGAAAACGTT**GGG**) to construct CRISPR/Cas9 vectors and transformed into *M. paleacea*. Screening of 417 and 338 putative transformants containing these constructs mirrored the results obtained using the first *NOP1* guide RNA – no mutants were observed. For the second and third targets, 48 plants each were genotyped using PCR and Sanger sequencing. These observations suggested that the choice of guide RNA was probably not the cause of failure of the system to generate mutants.

With regards to the promoters, I used the *MpoEF1a* promoter to drive a fluorescent marker for the initial transformation optimisation (described in Section 3.2.2.2) and it was found that expression of the fluorescent marker was ubiquitous with increased concentration in the meristematic regions (Figure 3.5) as previously observed in *M. polymorpha* (Althoff et al., 2014). As the expression of genes driven by this promoter are similar between *M. polymorpha* and *M. paleacea* and as this promoter has previously been used to generate knockouts in *M. polymorpha,* it is unlikely to be the cause of the failure to generate mutations.

To test whether the cause of the failure to generate mutations could be the *U6* promoter used to drive guide RNA expression, I obtained from *M. paleacea* the closest homolog of the *U6-1* gene of *M. polymorpha* (whose promoter was used in the constructs) and compared the promoter sequences of the two genes (Figure 3.9d). These comparisons revealed that the promoter regions from *M. paleacea* and *M. polymorpha* were quite divergent (41.5% identity). By contrast, the promoters of the EF1a genes from these two species were 76.6% identical (Figure 3.9e). As it has

previously been suggested that U6 genes that are evolutionary divergent cannot be used interchangeably, it is possible that the U6-1 promoter from *M. polymorpha* is not optimal for the expression of guide RNA sequences in *M. paleacea* and is the cause of the failure of the CRISPR/Cas9 system used here to induce mutations. Alternatively, it could be that the GoldenGate vector system developed here as a modification from the previously published vector (Ishizaki et al., 2013) is faulty and is the cause of the failure to generate mutations in *M. paleacea*. This will need to be tested in the future by employing this GoldenGate vector system to generate the same mutations as those generated by Ishizaki et al., 2013 in *M. polymorpha* by using the exact same guide RNA target sequence.

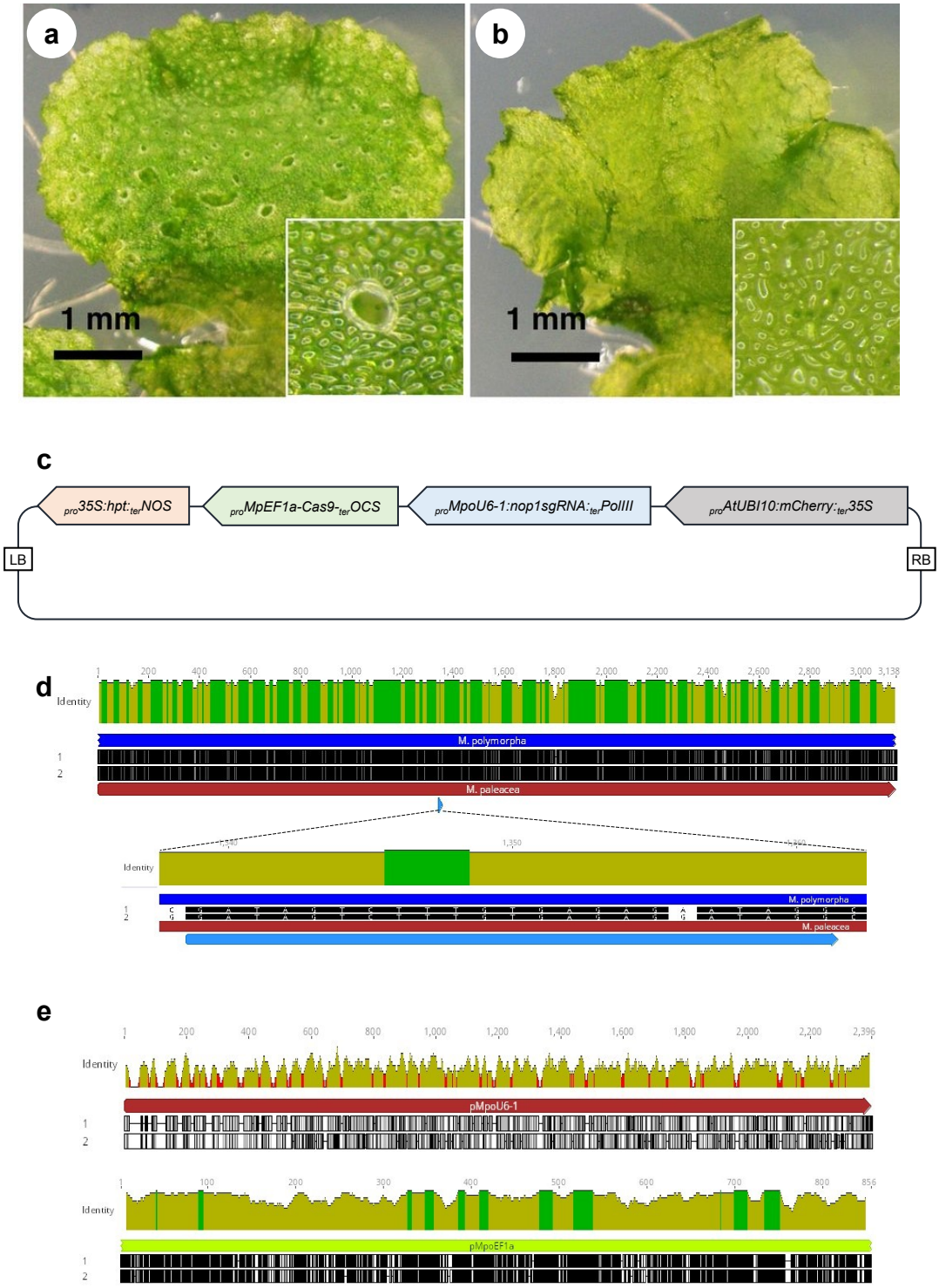

Figure 3.9. CRISPR/Cas9-mediated mutagenesis of *M. paleacea*.

Mutagenesis of *M. paleacea* was attempted using a CRISPR/Cas9 vector designed to knockout the *NOP1* gene in *M. paleacea*. This gene has previously been shown to be important for the formation of air pores in

M. polymorpha and knockouts in this gene resulted in the complete abolition of air pore formation in *M. polymorpha* (Ishizaki et al., 2013). (a) Wildtype *M. polymorpha* thalli showing air pores and (b) *nop1 M. polymorpha* thalli exhibiting the absence of air pores reported previously (Ishizaki et al. 2013). These images are reprinted as is from (Ishizaki et al. 2013) and are copyright by the American Society of Plant Biologists. (c) To knockout the *NOP1* gene in *M. paleacea*, a GoldenGate binary vector was constructed with a hygromycin resistance marker close to the left border of the T-DNA and a mCherry fluorescent marker at the right border to allow for the screening of plants with successfully integrated T-DNA. The CRISPR/Cas9 system previously used for *M. polymorpha* (Sugano et al., 2014) with the *MpoEF1a* promoter driving a human codon-optimised *Cas9* and the *MpoU6-1* promoter driving the guide RNA were adapted to *M. paleacea* and used. (d) The *NOP1* genes of *M. polymorpha* and *M. paleacea* were aligned for comparison. The 21bp sequence for which the CRISPR guide was designed for *M. paleacea* is indicated. (e) Alignments of the promoter regions of the *U6-1* small RNA and *EF1a* gene are provided to show that the *U6-1* promoters of *M. polymorpha* and *M. paleacea* show relatively less similarity.

3.3. Discussion

Although much has been learnt about the processes involved in the AM symbiosis and its regulation, our understanding of AM symbioses is largely based on studies in angiosperms (Delaux et al., 2013b). Given the antiquity of the AM symbiosis in the plant lineage, it is important to understand whether the insights gained into the AM symbiosis in angiosperms are true for other land plants is important. A major limitation to this goal has been the absence of a reliable non-flowering model system with a sequenced genome and established molecular methods to study the AM symbiosis. Here, I showed that the liverwort *Marchantia paleacea* can be used as a non-flowering plant model to study the AM symbiosis. For this, methods for AM colonisation *in vitro*, routine culturing and for the delivery of foreign DNA using Agrobacterium-mediated transformation were developed. The culturing and transformation methods that were successfully used for *M. paleacea* were adapted from those developed for the related liverwort species *M. polymorpha*. It was found

that the spore-based transformation method that has been successfully used to characterise genes in *M. polymorpha* could not be applied to *M. paleacea* as sporophyte production in this species could not be induced in the lab. The far-red supplementation method used for these induction experiments has only been shown to work for *M. polymorpha*, therefore it is possible that other liverworts such as *M. paleacea* require either additional or entirely different stimuli to induce sporophyte formation. Identification of these stimuli would allow for the induction of sporophytes in *M. paleacea* in the future, at which time the spore-based transformation method may be applicable to this species.

Although gametophyte-based transformation methods have been established for *M. polymorpha*, these are labour-intensive and only yield tens of transformants from a single batch of transformation compared to the hundreds of transformants yielded by the spore-based transformation method from a single experiment. While obtaining tens of transformants is enough for certain targeted applications such as studies of protein subcellular localisation and similar applications for the characterisation of individual proteins, transformation numbers in the order of magnitude of the spore-based method are required for large-scale applications such as forward genetic screens and for mutagenesis methods that work at low efficiencies.

To enable the application of both small-scale and large-scale approaches in *M. paleacea*, two gametophyte-based transformation methods used previously in *M. polymorpha* were adapted. A modification of the published thallus-cutting method, wherein the majority of steps are conducted in liquid culture, was developed, replacing the manual cutting steps with a laboratory blender to make this method amenable to scaling up. This method, termed the blend-based transformation method, was surprisingly similarly effective at both small volumes (2 ml) as well as large volumes (50 ml) suggesting that this method could be used for both large-scale and small-scale applications with appropriate scaling up or down. As applications such as forward genetic screens require transformants in the hundreds of thousands (Honkanen et al., 2016), further studies will need to be conducted to test the upper limit of the transformation volumes for this method and to check whether increased volume of transformation affects the transformation efficiency. An alternate transformation method called “agartrap” transformation developed in *M. polymorpha*, conducted on solid culture media, was also found to be applicable to *M. paleacea* with a few modifications. This method has the advantage of a reduced lead time to getting

transformants compared to the blend-based method but is not as optimal for large-scale studies as the blend-based method. This is primarily due to the nature of the starting material in these methods. As increasing the input to the transformation in the agartrap method involves manually increasing the number of gemmae transformed as opposed to a simple increase in the amount of thalli blended in the blend-based method, the agartrap method is not as scalable.

Mycorrhization experiments revealed that while AM associations could be reliably established using different types of inocula in *M. paleacea*, the subspecies ruderalis and montivagans of *M. polymorpha* could not accommodate the AM fungus *Rhizophagus irregularis*. While this confirms previous results observed in *M. polymorpha ssp. ruderalis* from the wild, it contests the observations of mycorrhizal fungi in samples of the *M. polymorpha ssp. montivagans* from the wild. As only one AM fungal species was used in my experiments, it is possible that the species used was not compatible with *montivagans* but this scenario is unlikely given the broad host range of AM fungi (Lin et al., 2014) and by the observation that *M. paleacea* was colonised by the AM fungal species used. As reliable molecular methods for the discrimination between *M. polymorpha* subspecies are yet to be established, the principal method used for the discrimination between these subspecies is visual. Therefore, it is possible that the montivagans sample used in our experiments was misidentified during collection and that it is either a different species or subspecies of liverwort, such as *polymorpha* or *ruderalis*, both of which have been proposed to be non-mycorrhizal. In addition to the mycorrhization assays, a fungal pathogen that could infect both *M. polymorpha* and *M. paleacea* was also identified. Future studies using such pathogenic and symbiotic fungal partners of liverworts will provide valuable insight into how liverworts are able distinguish between symbiotic and pathogenic fungi and if the genetic mechanisms that liverworts use for this purpose are similar to those used by their angiosperm counterparts.

To facilitate the use of *M. paleacea* as a model liverwort to study plant-fungal interactions, draft genome and transcriptome assemblies for *M. paleacea* were constructed through short-read sequencing. These assemblies were optimised to ensure maximum contiguity and completeness. Comparisons of the *M. paleacea* assembly to the recently published draft genome of *M. polymorpha* revealed that the BUSCO scores of both assemblies were similar (Mpa-862 vs Mpo-856) suggesting a comparable level of completeness. The *M. polymorpha* assembly was more

contiguous (N50=372Kb; number of scaffolds=4137) than the *M. paleacea* assembly (N50=78Kb; number of scaffolds=22669), and this can be attributed to the differences in the data used for the construction of these assemblies. While the previously constructed *M. polymorpha* assembly made use of mate-pair libraries of several sizes, the *M. paleacea* assembly described here was constructed using a single mate-pair library. As reliable estimates of the genome sizes of these species are yet to be obtained from methods such as flow cytometry, it is not possible to know the level of completeness of these assemblies. Both these assemblies fall short when compared to the genome assembly of the moss, *P. patens*, for which a chromosome-level assembly is available and the development of such high-quality assemblies for these liverwort species will prove invaluable to the research community. However, such a high-quality genome may not be necessary for applications such as the identification and characterisation of genes as recent studies have been successful at these applications in *M. polymorpha* with the currently available draft genome assembly.

With the transformation methods and genomic resources in place, the possibility of mutating genes in *M. paleacea* was explored using a CRISPR/Cas9-mediated mutagenesis system successfully used in *M. polymorpha*. For these experiments, the *M. paleacea* ortholog of a gene required for air pore formation in *M. polymorpha* was chosen as a visual marker as mutating this gene results in the complete absence of air pores in *M. polymorpha* and a similar phenotype in *M. paleacea* was expected. Visual and genotypic screening of transformed plants revealed that no mutations were induced by the CRISPR/Cas9 system. Explorations into why the system was not behaving as expected revealed that the likely source of failure to induce mutations was the promoter used for driving the guide RNA. *U6* small RNA promoters have been shown to be effective for driving the guide RNA for CRISPR/Cas9 applications in numerous plant species and the promoter used in the present study was the small RNA *U6* promoter, *pMpoU6-1*, which was previously used as part of a CRISPR/Cas9 system to successfully induce mutations in *M. polymorpha*. Comparison of this promoter with that of the closest homolog from *M. paleacea* of the gene it encodes revealed that these promoters were dissimilar (41.5% identical). Previous studies have suggested that dissimilar *U6* promoters from evolutionarily distant species cannot be used interchangeably (Wang et al., 2008) and the low similarity between the *U6-1* promoters of *M. polymorpha* and *M. paleacea* might be the reason for the failure of the system to induce mutations. Replacing the *MpoU6-1* promoter used in the current

study with the equivalent promoter from *M. paleacea* identified here would allow for testing this hypothesis. In addition, it is possible that the GoldenGate vector system developed here is faulty and is the cause for the failure to generate mutations in *M. paleacea*. This could be tested in the future by using the exact same guide RNA target sequence used by Ishizaki et al., 2013 to generate mutations in *M. polymorpha* using the GoldenGate vector system developed here to see if this is indeed the case.

Alternatively, a method of mutagenesis other than the CRISPR/Cas9 system could be used. Recently, it was shown that TALEN (transcription activator-like effector nucleases)-mediated gene targeting could induce mutations in *M. polymorpha* at a high efficiency (20%) (Kopischke et al., 2017). Therefore, this could provide an alternate approach to obtaining mutants in *M. paleacea*.

The resources and methods described in this chapter provide a starting point for the use of *M. paleacea* as a model system to study the evolution of plant-fungal interactions. Further development of molecular methods and genetic resources for this species will enable studies of how liverworts perceive and interact with their fungal partners and perhaps provide clues into how the AM symbiosis evolved in the ancestral land plants.

4

Phylogenomics Uncovers a 450-million-year-long Commitment of Plant Genes to Symbiosis

Chapter 4: Deciphering the Evolutionary Trajectories of Symbiosis Genes in Land Plants

4.1. Introduction

The AM symbiosis is formed by members of all major lineages of land plants – the bryophytes, lycophytes, monilophytes, gymnosperms and angiosperms (Wang and Qiu, 2006). Studies focussing on specific clades have revealed that independent losses of the AM trait have occurred numerous times in the land plant lineage (Ligrone et al., 2007; Brundrett, 2009). Comparative phylogenomic studies into trait loss have shown that the loss of a given trait is usually accompanied by a concerted loss of genes associated with the trait (Albalat and Canestro, 2016; Martí-Solans et al., 2016). This gene co-elimination occurs when these genes do not possess pleiotropic functions and their fates are completely intertwined with that of the trait they are associated with i.e., the genes are "committed" to the trait.

Phylogenomic approaches have recently been employed to compare angiosperm species that can form AM symbioses with those that have lost the ability to form this association (Figure 4.1) (Delaux et al., 2014; Favre et al., 2014; Bravo et al., 2016). These analyses revealed that several well-characterised symbiosis genes (such as *DMI2, CCaMK, IPD3, RAM1, RAD1, VPY, STR* and *STR2*) are committed to the AM symbiosis in angiosperms and this is reflected by consistent gene loss in all angiosperm species that have lost the AM symbiosis, but retention in those species where the symbiosis occurs. By looking for additional genes that exhibit this specific presence-absence pattern between AM host and non-host angiosperm species, these studies could predict that previously unknown genes that may also be committed to the AM association.

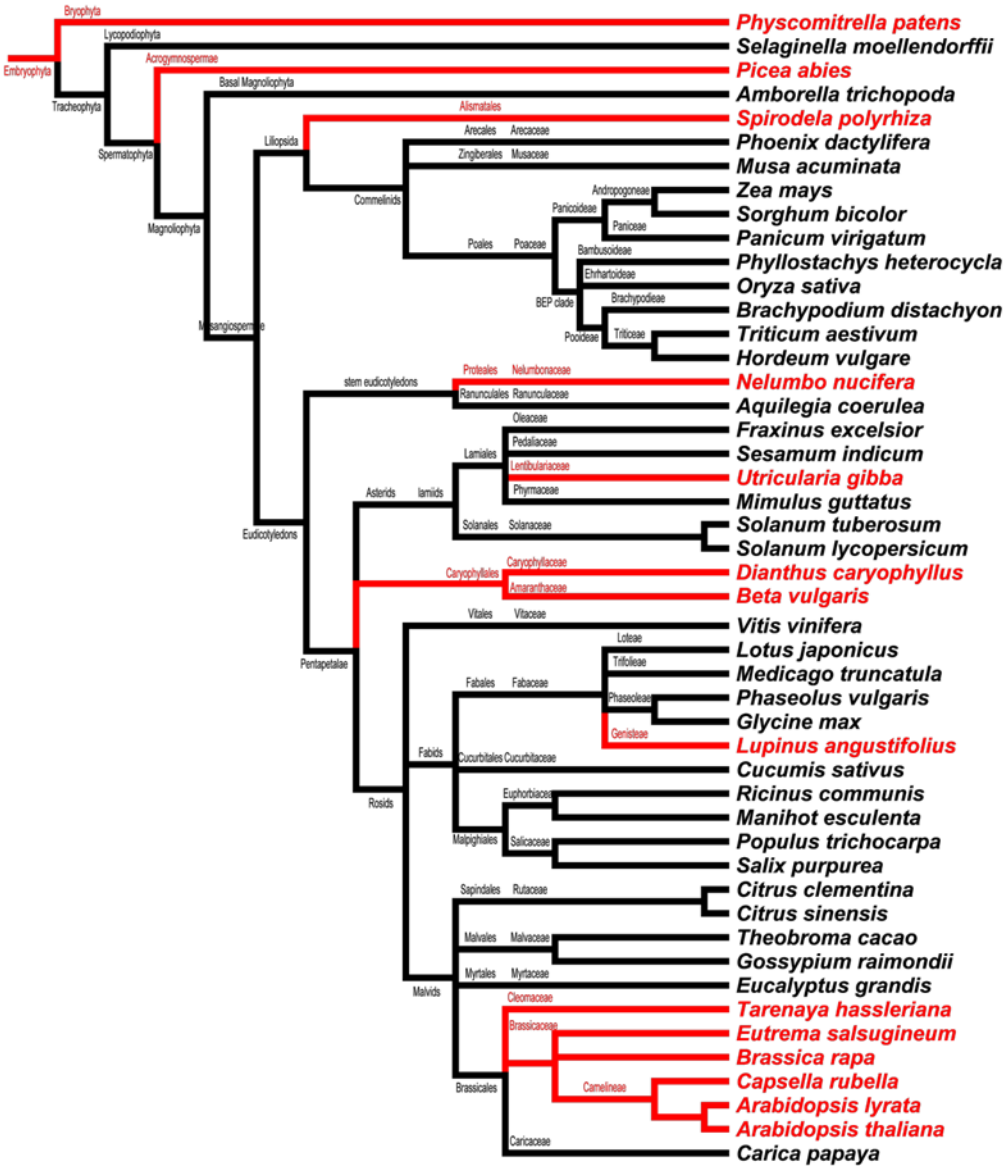

Figure 4.1. Phylogenetic representation of the species used for comparative phylogenomics of AM host and non-host species.

The genomes of the species listed were used to discover a correlated set of genes that were specifically lost in angiosperm species that have lost the ability to form the AM symbiosis. Species and clades that are able to form AM symbioses (AM hosts) are portrayed in black while those that are unable to form the AM symbiosis (AM non-hosts) are portrayed in red. Reprinted by permission from Macmillan Publishers Ltd: Nature Plants (Bravo et al., 2016), copyright (2016).

It was shown that a large percentage of these genes (88% of the genes with probe-sets available) exhibit a particular expression pattern - they are transcriptionally induced upon the AM symbiosis (Bravo et al., 2016). Furthermore, when mutants in some of these predicted symbiosis genes were analysed, they were found to be impaired in the AM symbiosis, confirming a symbiotic function for these genes. Based on these findings, it has been suggested that these patterns of elimination and transcriptional induction could be used as predictors for the involvement of genes in symbiosis. The comparative phylogenomic analysis in angiosperms (Bravo et al., 2016) found genes (138 in the legume *Medicago truncatula*) belonging to a total of 72 orthogroups (groups of genes that descended from a single gene in the last common ancestor of a group of species) committed to the AM symbiosis. Similar phylogenomic studies into the loss of other traits have usually been limited to individual biosynthetic pathways (Martí-Solans et al., 2016) and developmental traits (Zhang et al., 2013) controlled by a small number of genes and have therefore found only a few genes committed to a given trait. Further research is required to understand why such a relatively large set of genes became committed to the AM trait and to learn more about the processes that led to the recruitment and commitment of these genes.

Homologs of symbiosis genes have been found outside the angiosperms but the symbiotic function of these genes in the plants to which they belong is yet to be investigated (Wang et al., 2010). The main hurdles to testing the function of symbiosis gene homologs in these plants have been the lack of genomic resources and the inability to generate targeted mutations in non-flowering plant species able to form the AM symbiosis (Rensing, 2017). The discovery that specific co-elimination and transcriptional induction patterns can be used as predictors for the involvement of genes in symbiosis could help circumvent these limitations by providing an alternative approach to test if the symbiosis gene homologs from non-flowering plants have roles in symbiosis. If homologs of angiosperm symbiosis genes in non-flowering plants are transcriptionally induced upon symbiosis and are specifically lost in non-flowering plant species that have lost the ability to form the AM symbiosis, then this would support a role for these homologs during symbiosis in these plants.

The aim of the current study was to examine whether homologs of angiosperm symbiosis genes function during the AM symbiosis in non-flowering plants. As my efforts to utilise targeted mutagenesis systems to generate mutants in symbiosis gene homologs in *M. paleacea* were not successful, this hypothesis could not be tested

using reverse genetic approaches. Therefore, alternate approaches utilising comparative phylogenomics were employed to test whether homologs of symbiosis genes exhibited the indicators of symbiotic function as do their angiosperm counterparts – co-elimination upon loss of the AM symbiosis and specific transcriptional induction during the AM symbiosis. The availability of an *M. paleacea* genome, described in Chapter 3, made this work possible.

To explore whether symbiosis gene homologs were co-eliminated upon the loss of symbiosis in non-flowering plant species, a comparative genomic analysis was conducted using the liverworts *M. paleacea*, an AM host, and *M. polymorpha*, an AM non-host estimated to have diverged ~42 million years ago (Villarreal A et al., 2016) (Figure 4.2). Transcriptomes of the liverwort AM hosts *Lunularia cruciata* and *Pellia endiviifolia* were assembled from publicly available data and used as outgroups for the comparisons between *M. paleacea* and *M. polymorpha* (Alaba et al., 2015; Delaux et al., 2015). To test whether angiosperm symbiosis gene homologs from non-flowering plants are transcriptionally induced during the AM symbiosis, differential expression analysis was performed on sequence data generated from mycorrhized and control plants of the species *Lunularia cruciata*.

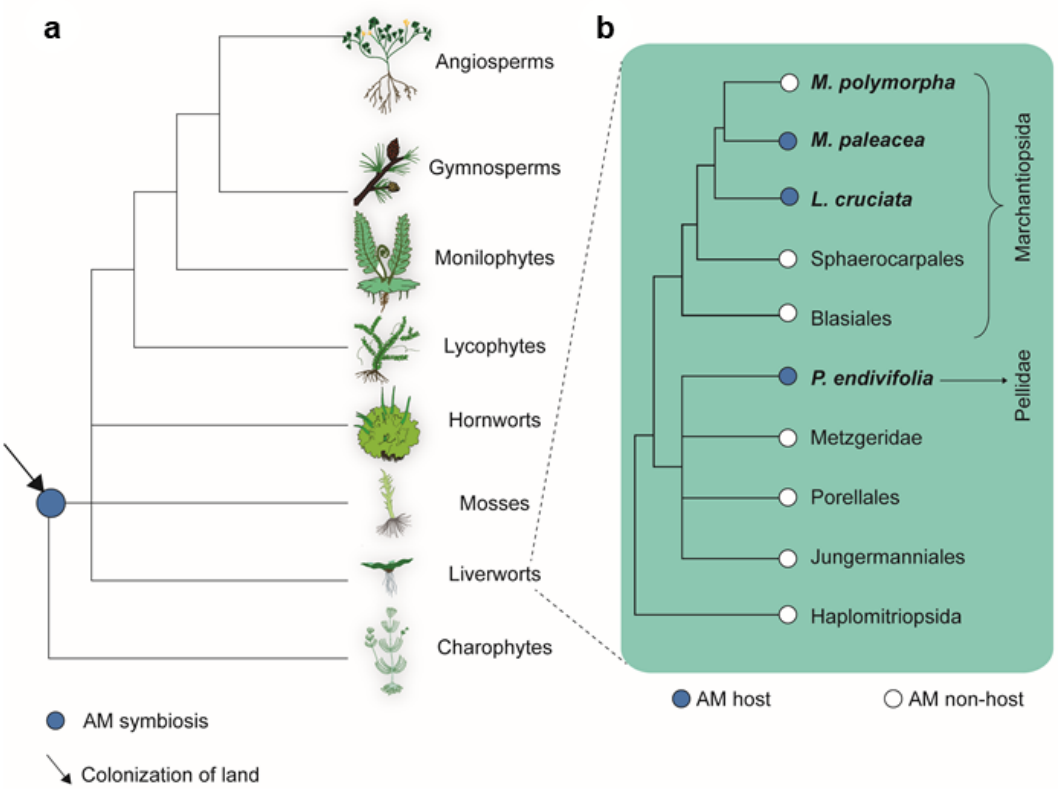

Figure 4.2. The AM symbiosis in the land plant lineage and the occurrence of the AM symbiosis in different clades of liverworts.

(a) A phylogenetic tree representing the relationships between clades within the streptophyte lineage. The predicted points at which the AM symbiosis evolved and land colonisation occurred are indicated on the tree. (b) A simplified representation of the different clades of liverworts with the species used in the current study highlighted. If any member of a given clade or if the highlighted species has been reported to form the AM symbiosis, they are represented by filled circles (AM hosts), if not they are represented using empty circles (AM non-hosts). The AM host status of the liverworts was based on previously published results (Ligrone et al., 2007).

4.2. Results

4.2.1. Reconstruction of the evolutionary history of genes committed to symbiosis in the angiosperms

To obtain homologs of angiosperm symbiosis genes in the liverworts, phylogenetic comparisons of angiosperms and liverworts were required for the 72 symbiosis-committed orthogroups. For such analyses, two types of methods are generally used: those that infer orthology based on sequence similarity measures (using BLAST) and those based on phylogenetic tree reconstruction (Koonin, 2005). While phylogenetic tree reconstructions take longer and require more resources than sequence similarity-based orthology inference methods, they have been shown to be relatively more accurate (Smith and Pease, 2017). Before embarking on the time-consuming phylogenetic tree construction, orthology inference was performed using the OrthoFinder algorithm (Emms and Kelly, 2015) to check whether sequence similarity-based orthology inference can be used instead of phylogenetic tree reconstruction to reliably obtain liverwort homologs of symbiosis genes. To do this, a test was conducted using a phylogenetic tree of LysMRLK genes to check whether the results of the OrthoFinder algorithm could faithfully mirror phylogenetic tree reconstruction. In parallel, the proteins predicted from the genomes of these species were clustered into orthogroups using OrthoFinder. The OrthoFinder classification divided the LysMRLK family into 9 orthogroups (Figure 4.3a), whereas, species reconciliation of the LysMRLK phylogeny suggested that this family only had 4 orthogroups (Figure 4.3b). Each of these orthogroups contained a clade with singleton genes from liverwort species basal to a clade containing either single or multiple representative(s) from each angiosperm species. These results suggest that the OrthoFinder classification divided the LysMRLK family into fragmented orthogroups. It has previously been reported that sequence similarity-based orthology inference methods tend to produce fragmented orthogroups (Emms and Kelly, 2015; Bravo et al., 2016). Hence, phylogenetic tree reconstructions were performed for the 72 angiosperm symbiosis-committed orthogroups as this approach proved more reliable.

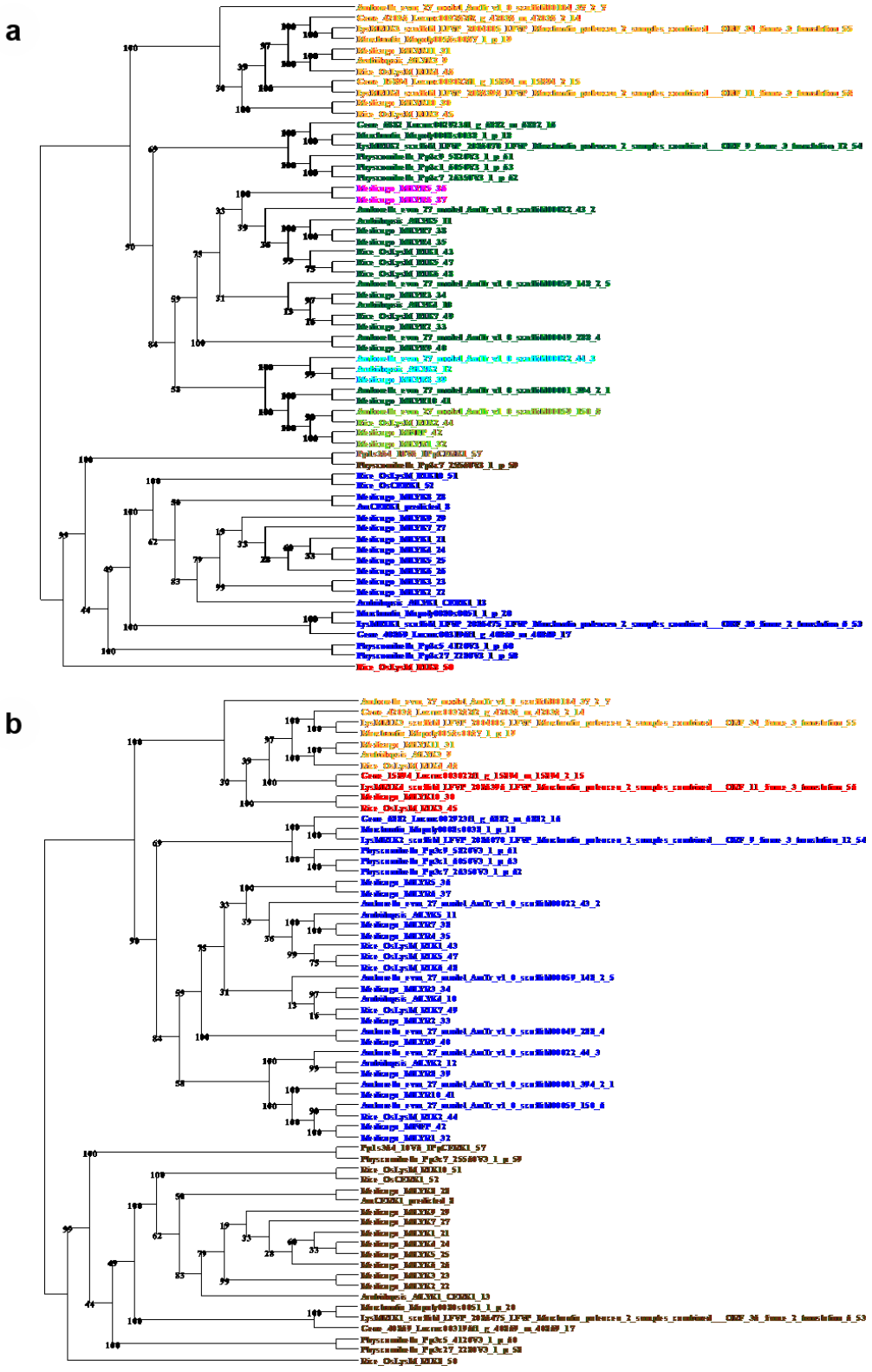

Figure 4.3. Comparison of orthology inference results from a sequence similarity-based method to those from a phylogenetic tree-based method using LysMRLK proteins from 7 species.

The following species were included in the tree: *Amborella trichopoda*, *Arabidopsis thaliana*, *Medicago truncatula*, *Oryza sativa*, *Lunularia cruciata*, *Marchantia paleacea* and *Marchantia polymorpha*. (a) OrthoFinder results were used to classify the LysMRLK tree per the orthogroup each protein was assigned to. The proteins were classified into 9 different orthogroups each represented by a different colour. (b) 4 orthogroups in the LysMRLK family as determined through manual reconciliation using the species tree. Each orthogroup consisted of a single liverwort representative and either single or multiple angiosperm representatives.

4.2.2. Homologs of symbiosis-committed angiosperm genes are present in the liverworts

Phylogenetic reconstructions of the evolutionary history of the 72 orthogroups committed to symbiosis in the angiosperms revealed the presence of homologs for 66 of these orthogroups in the liverworts (Figure 4.4a). The liverwort genes were classified as either being conserved as orthologs (if there was no evidence for lineage-wide duplications in either the liverwort or angiosperm lineages in the orthogroup of interest) or co-orthologs (if there was evidence for duplications in either the liverwort or angiosperm or both lineages in the orthogroup of interest) (Figure 4.4b). True orthologs of the angiosperm genes were found for 19 orthogroups, while homologs (co-orthologs) separated by rounds of duplication in either the angiosperm or liverwort or both lineages were found for the remaining 47 orthogroups. The presence of homologs for a majority (92%; 66 of 72) of the symbiosis-committed angiosperm genes in the liverworts mirrors the results obtained from the analysis of a subset of symbiosis genes, where homologs of all canonical symbiosis genes were found in the liverworts (Chapter 2). The phylogenetic trees used for this analysis are provided in Supplementary file S2.

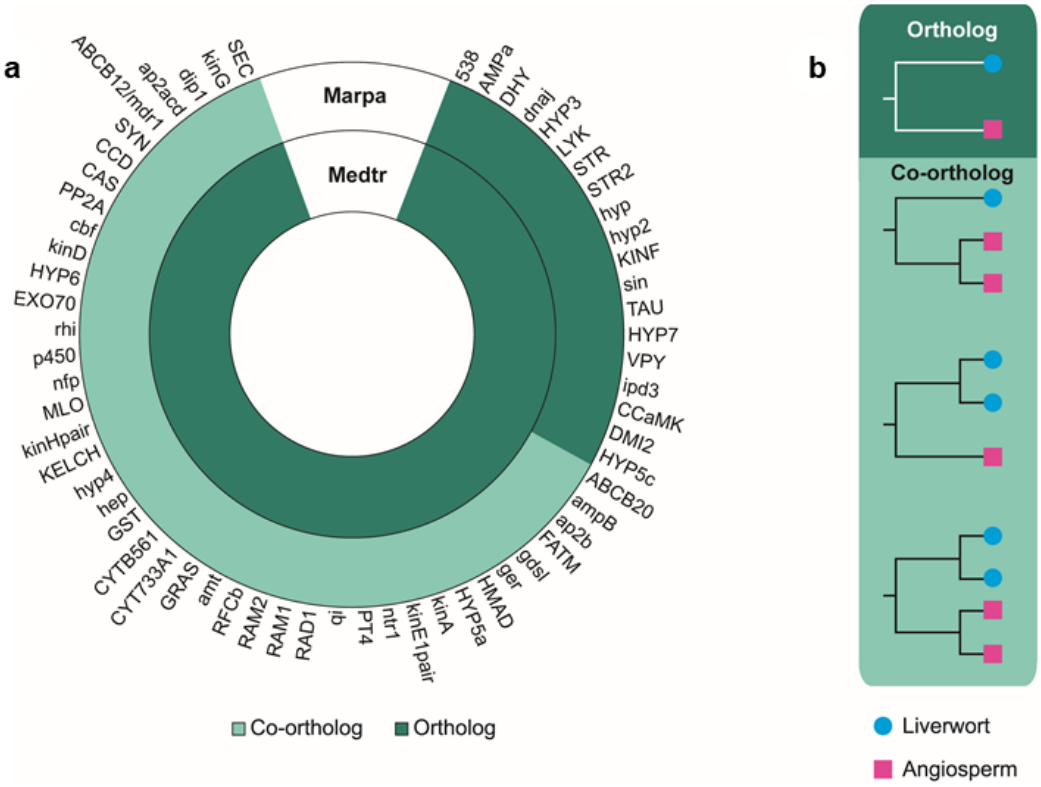

Figure 4.4. Inference of orthology of angiosperm symbiosis-committed genes in the liverworts.

(a) Summary of the phylogenetic analysis performed to find liverwort homologs of angiosperm genes committed to symbiosis. The relationships observed between liverworts and angiosperms for each orthogroup are represented using *Medicago truncatula* (Medtr) and *Marchantia paleacea* (Marpa) as references. Homologs/orthologs were found for 66 of the 72 angiosperm orthogroups in the liverworts. (b) The different types of orthology relationships observed between liverwort and angiosperm genes were categorised into orthologous (one-to-one) or co-orthologous (one-to-many, many-to-one and many-to-many).

4.2.3. Homologs of genes committed to symbiosis in the angiosperms are transcriptionally induced during symbiosis in the liverworts

To check whether the liverwort homologs of the symbiosis-committed angiosperm genes are transcriptionally induced during symbiosis, differential expression analysis was performed using transcriptomic data from the liverwort *L. cruciata* (provided by

Pierre-Marc Delaux and Christophe Roux, LRSV Toulouse). *L. cruciata* thalli were grown from gemmae and inoculated with spores of *R. irregularis* (AM samples) or water (mock) and grown for 8 weeks. Genes that were upregulated at least 1.5 times in the AM samples compared to the mock samples were regarded as transcriptionally induced. From the analysis, I found that genes homologous to 30 orthogroups were transcriptionally induced specifically in the mycorrhized samples (Figure 4.5a). These genes were then compared to their homologous/orthologous counterparts in Rice and Medicago using published RNA-seq data (Fiorilli et al., 2015; Luginbuehl et al., 2017) to check whether the transcriptional induction during symbiosis is conserved among these species. It was observed that for 24 orthogroups, at least one gene from *L. cruciata*, Rice and Medicago were transcriptionally induced in all species. This set included both orthologs and co-orthologs. For a subset of these orthogroups (*STR, PT4, RAM2*), qPCR was used to measure the expression of genes in *M. paleacea* in order to check whether the results observed in *L. cruciata* could be regarded as representative of the scenario in all liverworts. Gene expression analysis showed a similar transcriptional induction of these genes was observed in *M. paleacea* to that in *L. cruciata* (Figure 4.5b). (qPCR data were provided by Aisling Cooke, John Innes Centre, UK). The results of the differential expression analysis using RNA-seq data for *L. cruciata* are provided in Supplementary file S3.

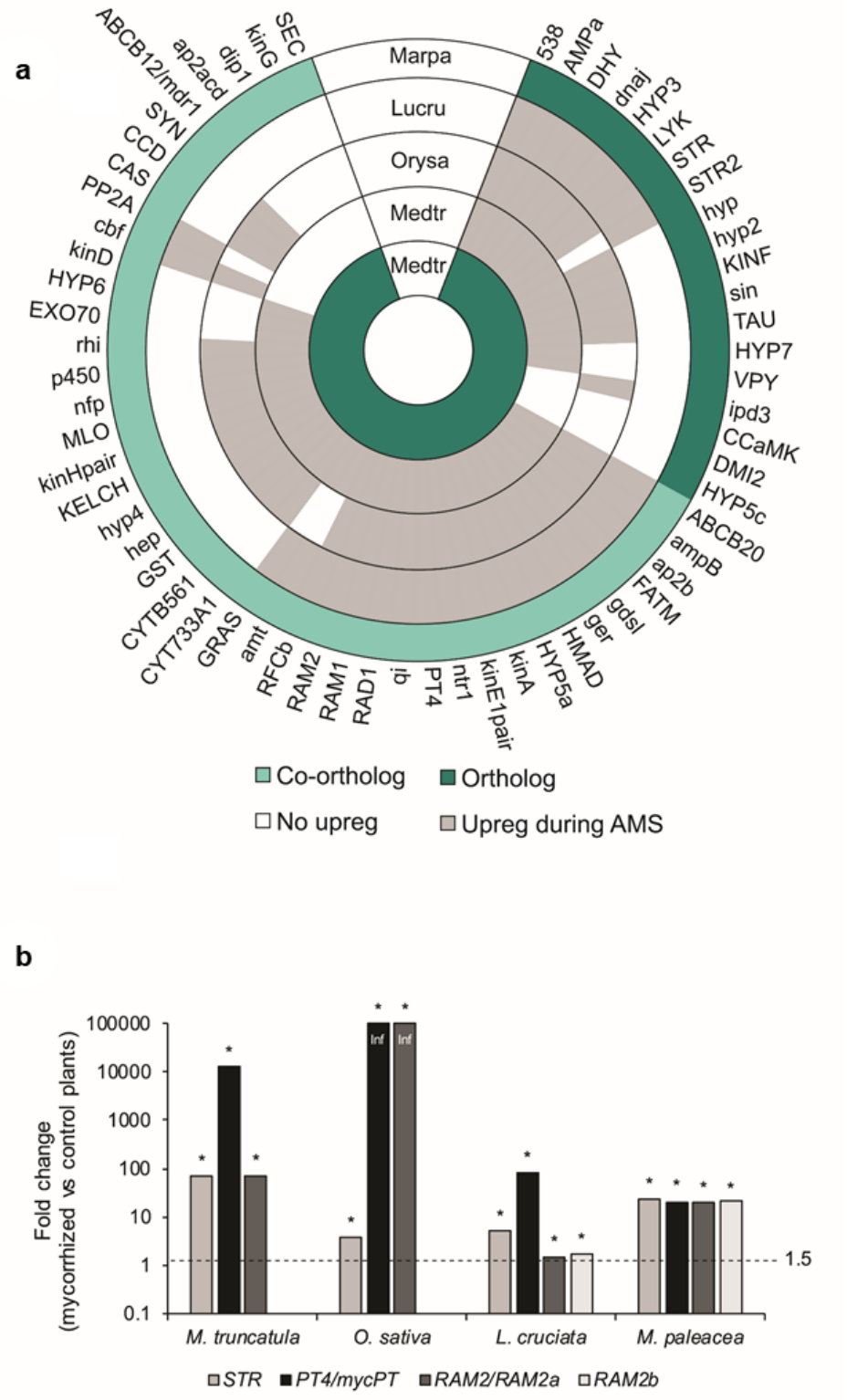

Figure 4.5. Expression analysis revealed that homologs of angiosperm symbiosis-committed genes are transcriptionally induced during the AM symbiosis in the liverworts.

(a) Expression analysis of homologs of angiosperm symbiosis-committed genes in the liverwort *L. cruciata* revealed that they are transcriptionally

induced during symbiosis much like their homologs in the angiosperms *O. sativa* and *M. truncatula*. Differential expression analysis on AM and mock thalli of *L. cruciata* was carried out using edgeR (fold change>2; False Discovery Rate (FDR)-corrected $p < 0.05$). The samples were generated and sequenced by Pierre-Marc Delaux and Christophe Roux (LRSV Toulouse). *M. truncatula* and *O. sativa* RNA-seq data were obtained from published studies (Fiorilli et al., 2015; Luginbuehl et al., 2017). (b) Expression analysis of *M. paleacea* genes belonging to the *STR*, *PT4* and *RAM2* orthogroups using qPCR. Fold change values comparing mycorrhized to control samples are provided for the *M. paleacea* qPCR experiments in addition to the RNA-seq results from *L. cruciata*, *O. sativa* and *M. truncatula* for comparison. *STR* is conserved as orthologs in all the above species and these orthologs are transcriptionally induced during symbiosis. *RAM2* is conserved as a singleton gene in the angiosperms but as three copies in the liverworts. The fold change values of two of these *RAM2* homologs (*RAM2a, RAM2b*) in liverworts are indicated. The phosphate transporter family to which *PT4* belongs was found to have independently diversified in liverwort and angiosperm lineages. One of the liverwort homologs of *PT4*, designated *mycPT*, was found to be transcriptionally induced during symbiosis. qPCR data were provided by Aisling Cooke (John Innes Centre, UK). Asterisks indicate significant ($p < 0.05$) fold changes greater than 1.5. Key: Marpa: *M. paleacea*; Lucru: *L. cruciata*; Orysa: *O. sativa*; Medtr: *M. truncatula.*

4.2.4. Signatures of co-elimination reveal symbiosis-committed genes in the liverworts

Although transcriptional induction provides evidence for these genes having a role in symbiosis in the liverworts, it is not as conclusive as demonstrations of loss-of-function phenotypes obtained through genetics. As discussed earlier, nature has itself provided numerous loss-of-function events of symbiosis in the angiosperms. It is from these cases that the phenomenon of commitment of genes to symbiosis was discovered. If co-elimination of angiosperm symbiosis gene homologs occurs upon loss of symbiosis in the liverworts, then that would imply commitment of these genes to symbiosis in the liverworts and thereby provide support for these genes having roles

in symbiosis in the liverworts. To test if this is the case, I used recently sequenced genomes of two strains of the AM non-host liverwort *Marchantia polymorpha* (Honkanen et al., 2016; Delmans et al., 2017). BLAST searches were conducted on the genomes of the two *M. polymorpha* strains for the symbiosis gene homologs found in *M. paleacea*. For the first set of searches, I used the genome of the *M. polymorpha* TAK strain that was collected in Japan. Searching this genome for the genes found in *M. paleacea* revealed that while most of these genes were conserved in *M. polymorpha*, 17 were absent (Figure 4.6). I then carried out a second set of searches on an independent genome assembly of the *M. polymorpha* CAM strain collected in the UK and confirmed my observations from the TAK genome. Additionally, I found that 17 orthogroups were retained in *M. polymorpha* despite exhibiting transcriptional induction upon symbiosis suggesting that although these genes might have already been recruited into symbiosis, they are not committed to symbiosis in the liverworts.

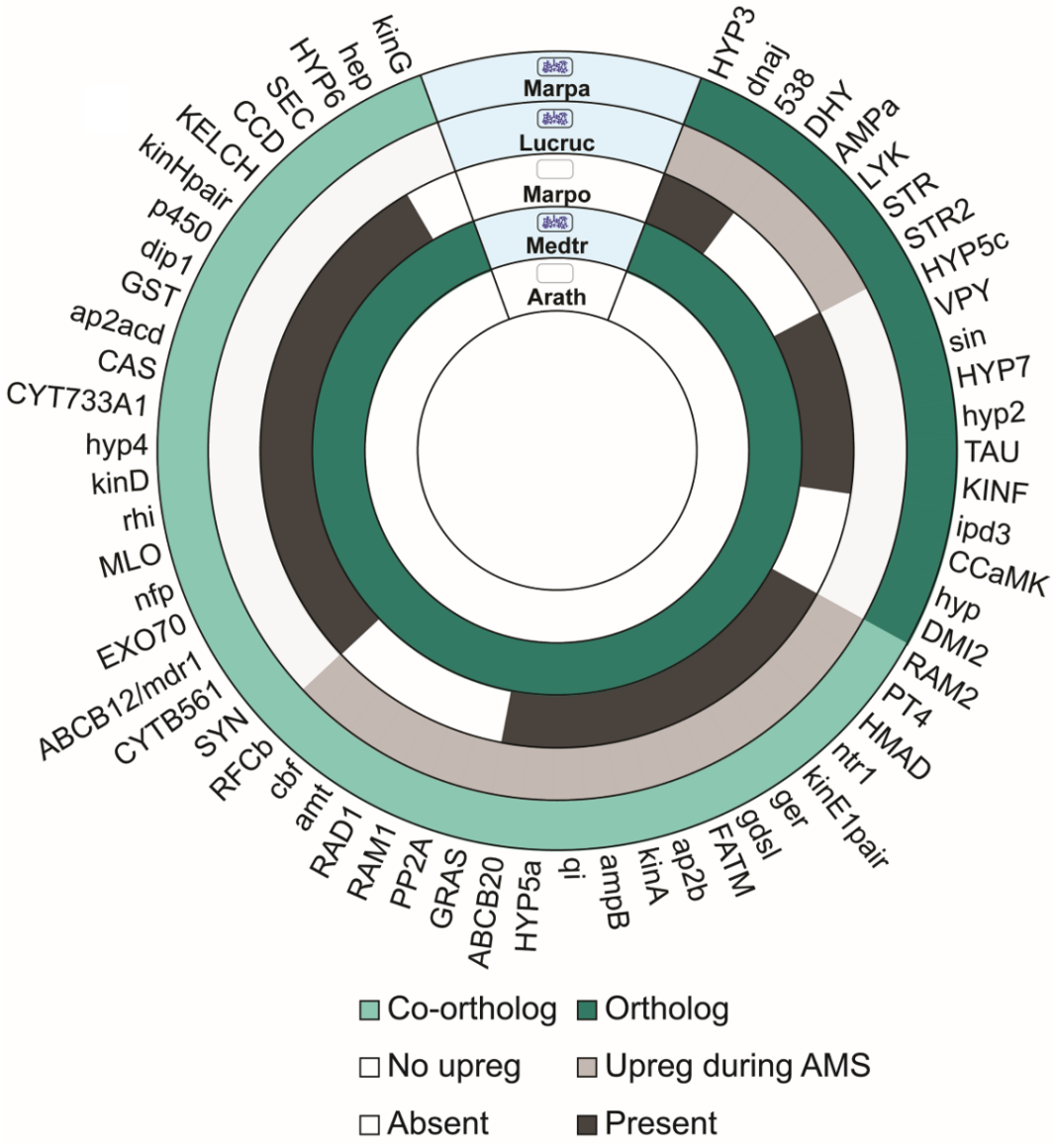

Figure 4.6. Comparative genomics of the liverwort AM non-host *M. polymorpha* and the liverwort host *M. paleacea* revealed that homologs of angiosperm symbiosis-committed genes are lost upon loss of symbiosis in the liverworts.

Genome comparisons between *M. paleacea* and the liverwort AM non-host *M. polymorpha* revealed the absence of several but not all angiosperm symbiosis-committed gene homologs in the *M. polymorpha* genome including well-characterised symbiosis genes such as *CCaMK*, *IPD3*, *DMI2*, *RAM1*, *RAD1*, *STR* and *STR2*. This finding is in contrast with the absence of all of these genes in angiosperm AM non-hosts such as *Arabidopsis thaliana*. The orthology relationships for each of the orthogroups between liverworts and angiosperms are presented between *M. paleacea* and *M. truncatula*. The transcriptional induction status of the

liverwort orthologs/co-orthologs is presented for *L. cruciata*. Key: Marpa: *M. paleacea*, Lucruc: *L. cruciata*, Marpo: *M. polymorpha*, Medtr: *M. truncatula*, Arath: *A. thaliana*. In the figure, the AM host status of each species is indicated with a plant cell containing a blue arbuscule (AM host) or an empty cell (AM non-host) above the species name.

To check whether the absence of genes in the *M. polymorpha* genome could be due to the incompleteness of the *M. polymorpha* assembly, a synteny analysis was conducted using BLAST between *M. paleacea* and *M. polymorpha* for the genes absent in the *M. polymorpha* genome. Conserved genomic blocks between *M. paleacea* and *M. polymorpha* were identified for 11 of the 17 missing genes in the *M. polymorpha* genome. In order to confirm the absence of these 11 genes in *M. polymorpha*, homologous sequences to the missing genes were screened in the identified genomic blocks. For 6 of the *M. polymorpha* genomic blocks, no homologous sequence was identified, confirming gene loss. For 5 of the *M. polymorpha* genomic blocks, sequences with similarity to the *M. paleacea* genes were detected. Comparison of these homologous regions with *M. paleacea* revealed that the *M. polymorpha* regions had accumulated disruptive mutations, including frameshifts and deletions, that likely caused the pseudogenisation of symbiosis gene homologs in *M. polymorpha* sometime after its divergence from its shared AM host ancestor with *M. paleacea* (Figure 4.7). The discovery of these pseudogenes in syntenic locations in *M. polymorpha* confirms that these genes were truly lost.

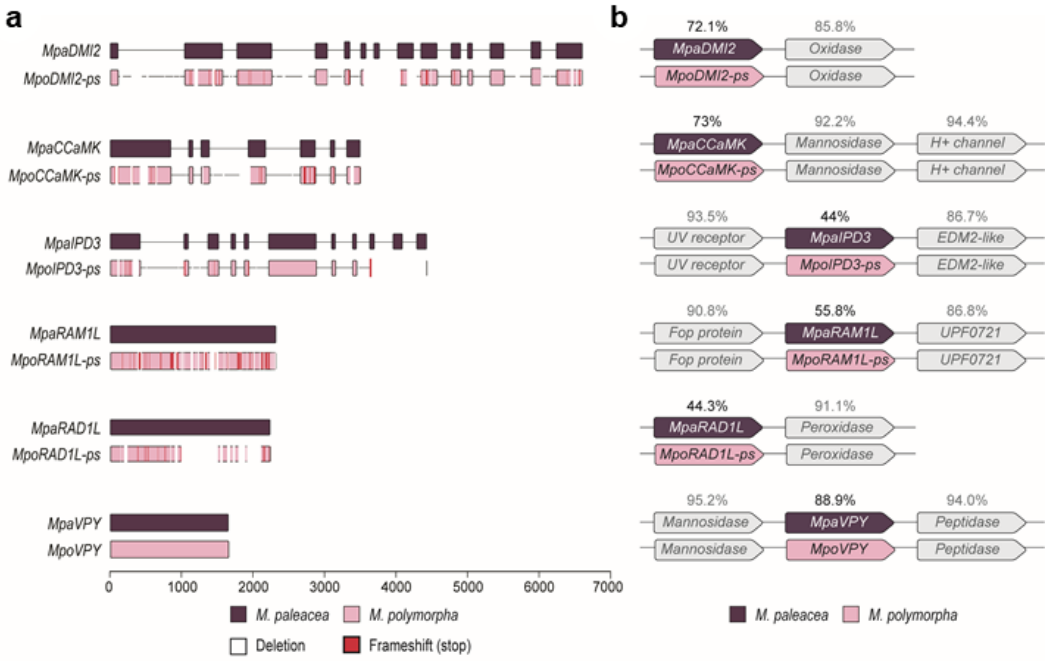

Figure 4.7. Homologs of angiosperm symbiosis-committed genes are pseudogenised in the AM non-host liverwort *M. polymorpha*.

(a) Synteny analysis between *M. paleacea* and *M. polymorpha* revealed the pseudogenisation of *DMI2*, *CCaMK* and *IPD3* orthologs as well as *RAM1* and *RAD1* co-orthologs in *M. polymorpha*. The pseudogenisation is evidenced by numerous deletions and disruptive mutations in the genomic regions corresponding to the genes in *M. polymorpha*. The alignment for the *VPY* ortholog which has not accumulated such mutations is provided for comparison. The *M. paleacea* and *M. polymorpha* genes are prefixed with "Mpa" and "Mpo" respectively. The pseudogenes are indicated with a "ps" suffix. (b) Sequence identity was calculated for the coding sequences of genes from surrounding genomic regions for the aforementioned genes between *M. polymorpha* and *M. paleacea*. This analysis revealed that changes were isolated to the symbiosis gene homologs and did not affect the surrounding genes in *M. polymorpha* as evidenced by the reduction in sequence identity specifically in the pseudogenes.

4.2.5. Loss of symbiosis was accompanied by a change in the selective constraints on homologs of symbiosis genes in the liverworts

Gene disruption is usually a result of changes in the selective constraints on these genes following a change in the fitness benefit provided by these genes or the trait controlling these genes (Chapple and Guigó, 2008). To test whether changes in the selective constraint was the reason for the pseudogenisation events observed in *M. polymorpha*, the ratio of synonymous-to-non-synonymous substitution ratio (ω), a measure of selection, was calculated for homologs of angiosperm symbiosis-committed genes from the liverworts *M. paleacea*, *M. polymorpha* and *L. cruciata*. The analysis was restricted to genes present as single-copy orthologs in each of the liverworts. This analysis was conducted using three evolutionary models provided by the PAML framework (Yang, 1997) each allowing for different scenarios of evolution (Figure 4.8). The results from these models were compared to ascertain which mode best explained the evolution of the given sequences (Table 4.1). PAML was used as an initial test to check whether any changes in the selection regimen (including relaxation of purifying selection and positive selection) had specifically occurred in the *M. polymorpha* branch as has previously been demonstrated for the *NSP1* gene in the Brassicaceae (Delaux et al., 2013a). Following this initial test, the RELAX framework (Wertheim et al., 2015), which was specifically designed to test whether relaxation of selection (across the whole gene in all sites including those under purifying, positive and neutral selection) has taken place in a specific lineage was used.

Pseudogenes are usually removed from large-scale analyses employing PAML as the input files for PAML require codon alignments and as pseudogenes usually contain numerous stop codons, these cannot be directly used as input for the codon alignments required by PAML. A software package called PAL2NAL (Suyama et al., 2006) was specifically developed to address this issue whereby stop codons and the corresponding non-stop codons in homologous non-pseudogenic sequences are removed following codon alignment. This “stop-codon filtered” file can be used as input to PAML to study pseudogenes and indeed many previous studies have used the PAML framework to study the selective constraints that have governed the evolution

of pseudogenes (Mundy and Cook, 2003; Zhang et al., 2003; Khachane and Harrison, 2009; Cai and Patel, 2010). The above approach was also applied in the present study.

For the PAML analysis, first, the one-ratio model M0, where a single ω ratio is assumed for all the branches of the tree, was used. Based on the ω value observed using this model, it is possible to check whether the gene/genes on the tree are under purifying selection (ω<1), neutral evolution (ω=1) or positive selection (ω>1). All the tested genes had ω < 1, suggesting that these genes were under purifying selection in the liverwort lineage (Table 4.1). Next, the free-ratios model, which allows for a flexible number of ω values and thereby changes in the ω over time along the branches, was tested to determine whether it explained the evolution of the genes better than the one-ratio model. For all the genes tested, the free-ratio model explained their evolution significantly better than a single ω ratio, suggesting that there had been changes in the ω ratio over time (Table 4.1). To determine whether the changes in ω predicted by the free-ratios model for these genes could be explained by a change in ω specifically in the *M. polymorpha* sequences, the two-ratios model was used to allow for changes in the background ω value specifically in the branch leading to *M. polymorpha*. For 13 of the 28 tested genes, the model which assumes *M. polymorpha* to have a significantly different ω ratio best explains their evolution. In each of these cases, the branch leading to *M. polymorpha* had a significantly higher ω compared to the background (Table 4.1). This included all identified pseudogenes in addition to 8 other retained genes suggesting that these genes experienced a change in the purifying selection following the loss of the AM symbiosis in this lineage. With the PAML branch-model tests alone it is not possible to ascertain whether this change in the purifying selection on these genes was due to positive selection acting on a few sites, or relaxation of purifying selection across the entire gene.

Recently, the RELAX framework (Wertheim et al., 2015) was developed which models the ω value across every site in the gene along each branch of the phylogeny to test for relaxation of purifying selection. This framework tests for relaxation of selection in the branch of interest (foreground branch) by checking how the ω value of different sites have changed in the foreground branch compared to the background branch (the rest of the phylogeny). Sites that have a ω value < 1 in the background branch, upon relaxation of selective constraints, tend to increase to a ω value of 1 in the foreground branch. Sites that have a ω value >1 in the background branch, upon relaxation of selective constraints, tend to decrease to a value closer to 1. Thus, by

measuring the proportion of sites with ω values in the three categories (ω<1, ω=1, ω>1) in the foreground and background branches, this framework can test whether purifying selection is intensified or relaxed in any given branch of a phylogeny relative to the other branches. I used this framework to test whether the *M. polymorpha* genes/pseudogenes experienced a relaxation of selection (Figure 4.9). The results of this analysis suggested that there was a significant relaxation of selection in *M. polymorpha* for all pseudogenes and for 3 of the retained genes (Table 4.2). While the results of this analysis provide support for the relaxation of purifying selection in these genes/pseudogenes, it does not rule out the possibility that positive selection on specific sites in these genes/pseudogenes may have also occurred during their evolution. To fully understand how selective constraints have changed in these genes, specific sites in the gene need to be considered and for this, more sophisticated models such as the branch-site models implemented in PAML (Yang, 1997) will need to be used.

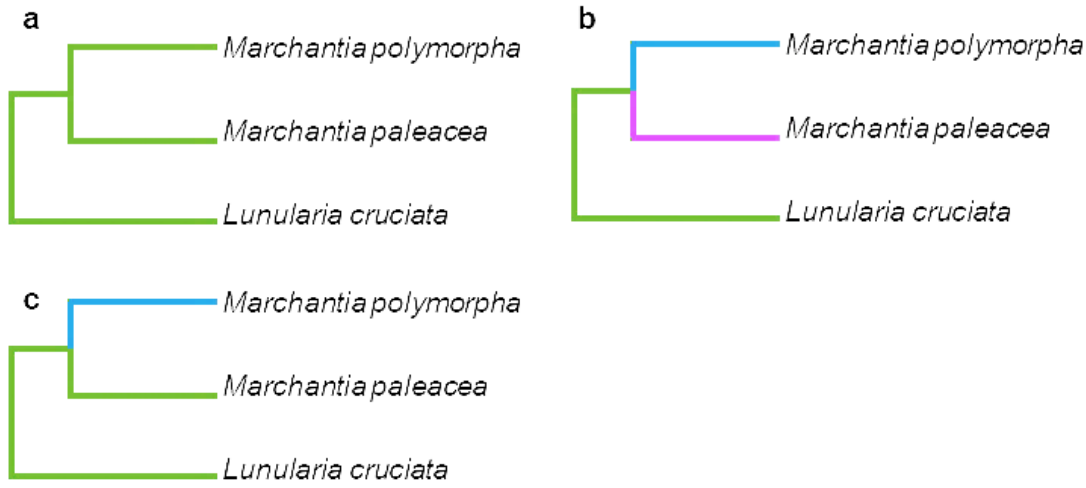

Figure 4.8. Selection constraints on the homologs of angiosperm symbiosis-committed genes were measured using the PAML framework with the aim to determine whether selection was relaxed specifically in the *M. polymorpha* lineage.

(a) The first model used was the one-ratio model, which allows for a single ω value for the entire tree. (b) The second model, the free-ratio model, allows for a separate ω value for each branch in the tree. (c) The third model used, the two-ratio model, allows for the selection of background and foreground sets of branches on the tree and for the

calculation of a separate ω value for each set. This model was used to check whether selection was relaxed specifically in the branch leading to the *M. polymorpha* genes. Different colours for the branches on the tree indicate different ω values as allowed by the models described above. The one-ratio model is referred to as the M0 model, the other two models (two-ratio and free-ratio) do not have a M0-style model number associated with them but are specified solely by setting the model=n option: free-ratio (model=1) and two-ratio (model=2).

Family	One-ratio	Free-ratio				Two-ratio				
	ω	Lc ω	Mpa ω	Mpo ω	p-value	Lc/Mpa ω	Mpo ω	Fold Change	p-value	Significant increase?
VPY	0.13	0.09	0.27	0.16	5.58E-11	0.06	2.45	44.32	2.71E-07	YES
RAD1L*	0.30	0.22	0.17	0.76	5.31E-12	0.14	0.85	6.24	3.30E-11	YES
CCaMK*	0.30	0.02	0.48	0.32	1.17E-13	0.15	0.67	4.43	8.51E-14	YES
exo70	0.04	0.14	0.03	0.23	8.53E-12	0.02	0.08	4.23	6.82E-06	YES
RAM1L*	0.13	0.05	0.23	241.00	2.10E-02	0.13	0.53	4.13	9.64E-03	YES
KINF	0.22	0.12	0.14	0.52	5.64E-06	0.18	0.66	3.74	9.09E-04	YES
CAS	0.19	0.15	0.15	0.67	8.97E-16	0.12	0.46	3.73	1.75E-15	YES
DMI2*	0.32	11.81	1.61	0.02	8.22E-13	0.21	0.74	3.49	8.50E-14	YES
IPD3*	0.48	0.04	999.00	0.04	3.93E-04	0.32	0.85	2.64	3.29E-04	YES
CYT733A1	0.19	0.31	0.39	0.83	2.65E-02	0.16	0.40	2.57	1.64E-02	YES
nfp	0.17	0.12	4.69	0.41	3.56E-03	0.13	0.30	2.26	1.36E-03	YES
MLO	0.20	0.14	0.10	0.34	1.79E-03	0.16	0.33	2.10	2.39E-03	YES
TAU	0.12	0.15	999.00	0.30	2.80E-02	0.10	0.20	2.08	5.69E-02	YES
p450	0.29	0.12	0.12	0.19	8.72E-05	0.18	nd	nd	nd	NO
dnaj	0.15	0.12	0.82	0.66	1.43E-02	0.13	nd	nd	nd	NO
CYTB561	0.14	0.16	0.15	0.41	2.20E-01	0.13	2.31	17.28	5.30E-01	NO
ap2b	0.04	0.10	0.13	0.11	1.15E-01	0.03	0.07	1.98	1.15E-01	NO
ap2acd	0.18	0.16	999.00	0.35	7.47E-02	0.16	0.29	1.80	4.57E-02	NO
HYP6	0.19	0.00	182.71	0.05	3.84E-02	0.18	0.28	1.57	3.77E-02	NO
hyp4	0.13	0.16	0.15	0.29	1.57E-01	0.12	0.19	1.55	1.33E-01	NO
HYP3	0.24	0.08	0.03	0.08	8.26E-02	0.22	0.32	1.46	1.39E-01	NO
538	0.09	0.04	0.00	0.09	6.04E-01	0.09	0.11	1.17	8.92E-01	NO
kinA	0.13	0.18	0.21	0.27	9.45E-01	0.13	0.14	1.10	7.28E-01	NO
HYP5c	0.19	0.13	0.22	0.15	4.51E-01	0.18	0.20	1.08	7.28E-01	NO
SEC	0.10	0.09	2.20	0.05	8.15E-01	0.10	0.11	1.08	8.68E-01	NO
dip1	0.04	0.25	0.18	0.22	9.40E-01	0.04	0.04	0.89	7.12E-01	NO
kinD	0.08	0.20	0.15	0.83	2.03E-01	0.08	0.07	0.89	8.22E-01	NO
hyp2	0.23	0.21	0.32	0.26	4.16E-01	0.25	0.21	0.85	4.62E-01	NO

Table 4.1. Results of the PAML analysis to determine whether purifying selection on homologs of angiosperm symbiosis-committed genes was relaxed following the loss of symbiosis in *M. polymorpha*.

ω values were obtained for the homologs of angiosperm symbiosis-committed genes conserved as single-copy orthologs within the liverworts. *L. cruciata, M. polymorpha* and *M. paleacea* sequences were used for the analysis. The results obtained using the one-ratio, free-ratio and two-ratio models are presented. For the two-ratio model, the *M.*

polymorpha branch was set as the foreground branch. The free-ratio and two-ratio models were tested against the one-ratio model to check whether they could explain the evolution of the genes significantly ($p<0.05$) better using likelihood-ratio tests. The results of these tests are presented. For 13 orthogroups, a significant increase in the ω value was observed for the *M. polymorpha* sequence suggesting a relaxation of purifying selection (indicated in green). This set included all the pseudogenes described previously (marked using asterisks on the table).

Figure 4.9. Selection constraints on the homologs of angiosperm symbiosis-committed genes were measured using the RELAX framework with the aim to determine whether selection was relaxed specifically in the *M. polymorpha* lineage.

Using the RELAX framework, a test was carried out to test whether the *M. polymorpha* homologs of the angiosperm symbiosis-committed genes had undergone a relaxation of selection following the loss of symbiosis by comparing them against their counterparts from *M. paleacea* and *L. cruciata*.

Orthogroup	k	P-value	Log-likelihood ratio
IPD3*	0.45	0.007794683	7.08
CCaMK*	0.39	1.30E-10	32.34
DMI2*	0.29	1.60E-10	40.91
RAD1L*	0.35	1.62E-09	36.38
RAM1L*	0.02	1.15865E-05	19.23
CAS	0.54	9.27245E-05	15.28
HYP6	0	0.02332	5.14
VPY	0.3	5.50E-15	7.71
duf538	50	0.01754204	5.64
hyp2	1.89	0.019442292	5.46
ap2acd	0.73	0.299481831	1.08
ap2b	0.83	0.358525666	0.84
CYTB561	0.46	0.715213333	0.13
dnaj	0.87	1	0
hyp4	0.49	0.380712899	0.77
kinA	0.7	0.614999	0.25
MLO	0.89	0.86541	0.03
nfp	0.9	1	-0.07
p450	0.3	0.2246648	1.47
SEC	0.74	0.69965	0.15
TAU	0.83	0.576099	0.31
CYT733A1	2.14	0.360928566	0.83
dip1	1.33	0.60176689	0.27
HYP3	13.31	0.122967	2.38
HYP5c	1.37	0.1967654	1.67
kinD	1.23	0.49098	0.47
KINF	1.28	0.4424089	0.59

Table 4.2. Results of the RELAX test to determine whether purifying selection on homologs of angiosperm symbiosis-committed genes was relaxed following the loss of symbiosis in *M. polymorpha*.

For 8 of the orthogroups, a significant relaxation of selection ($k<1$; $p < 0.05$) was detected for the *M. polymorpha* branch in the trees containing liverwort homologs of the angiosperm symbiosis-committed genes (indicated in green). This set included all the detected pseudogenes (marked with asterisks in the table). 7 of these were also detected as experiencing relaxed selection using the PAML framework (Table 4.1).

4.2.6. A conserved set of genes is committed to symbiosis in the land plants

Although 24 of the 72 orthogroups had genes transcriptionally induced in both angiosperms and liverworts during the AM symbiosis, not all orthogroups were lost or pseudogenised upon loss of symbiosis in the liverworts. In contrast, genes belonging to all these orthogroups are lost in the numerous cases of symbiosis loss in the angiosperms. Thus, despite the transcriptional induction supporting a role for of these genes in symbiosis in liverworts, not all of them were found to be committed to symbiosis in the liverworts as they are in the angiosperms. To test whether the number of orthogroups committed to symbiosis differed from the angiosperms in other non-flowering plant lineages as well, I looked at AM non-hosts from two other plant lineages, the mosses and the gymnosperms using the genomes of the moss *P. patens* (Rensing et al., 2008) and the gymnosperm *Picea abies* (Nystedt et al., 2013). BLAST searches and phylogenetic analyses were conducted to obtain homologs of angiosperm symbiosis-committed genes from these plants. The phylogenetic trees constructed from this analysis are provided in Supplementary File S2. From this analysis, I found that the number of genes co-eliminated in these lineages differed from that in the angiosperm lineage (Figure 4.10) confirming that the number of genes committed to symbiosis was different for each of the lineages tested. Despite the observed difference in the number of genes committed to symbiosis in each lineage, genes belonging to 13 orthogroups were absent in all these AM non-host species tested. These 13 orthogroups were found to be retained in host species of both angiosperm and liverwort lineages. Together, these observations support the idea that these 13 orthogroups are conserved across the land plant lineage and became committed to symbiosis during the early evolution of land plants. This conserved set of 13 symbiosis-committed genes included several well-characterised symbiosis-committed genes such as *DMI2, RAM1, RAD1, STR* and *STR2*.

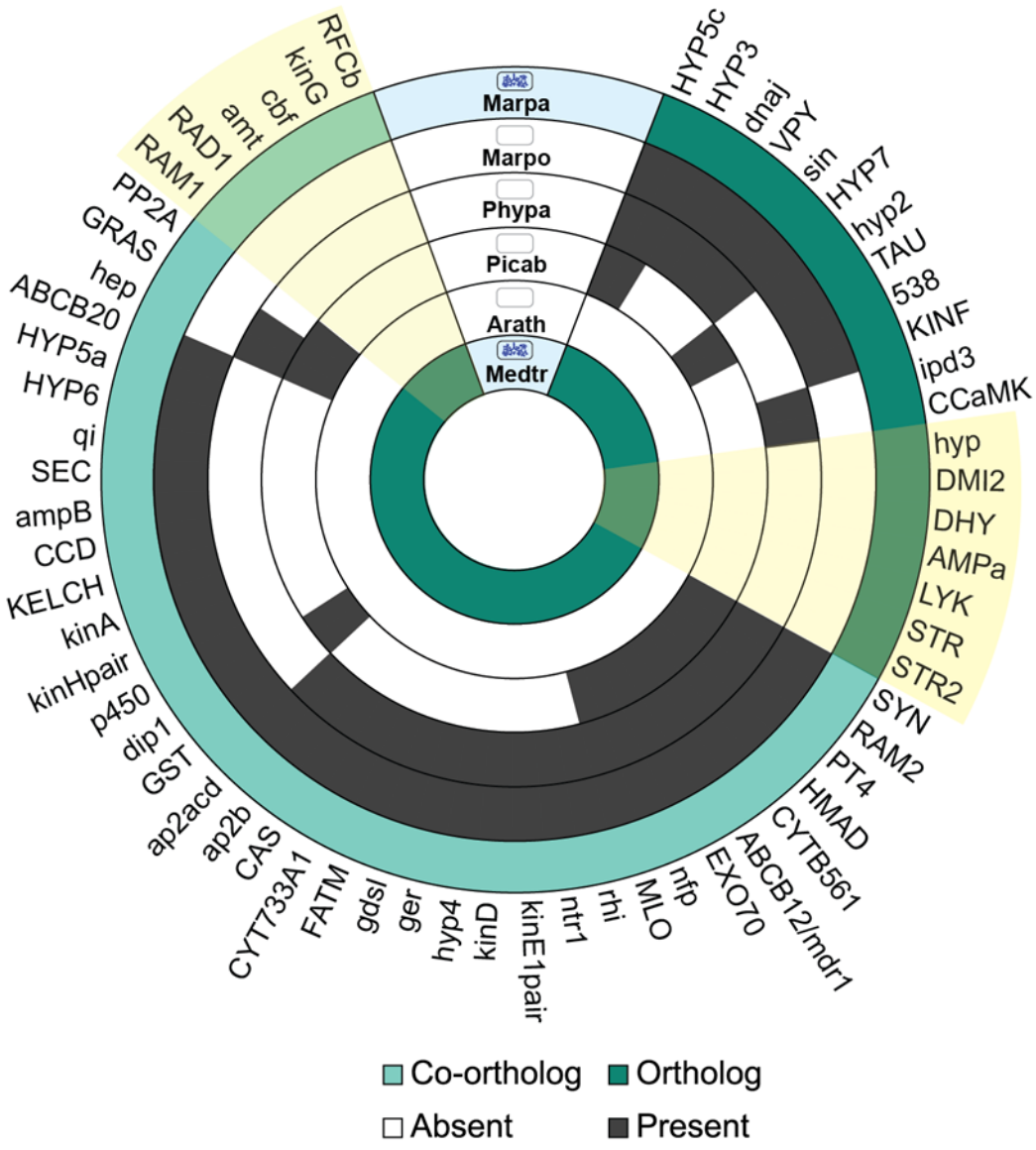

Figure 4.10. Representation of the loss of symbiosis gene homologs across the land plant lineage.

Phylogenetic analyses were conducted to study the presence-absence variation in homologs of angiosperm symbiosis-committed genes in AM non-host species from other lineages of land plants. The AM non-hosts used for this analysis were the liverwort *M. polymorpha*, the moss *P. patens*, the gymnosperm *P. abies* and the angiosperm *A. thaliana*. *M. paleacea* and *M. truncatula* were used as the AM host references for the liverworts and angiosperms respectively to elucidate the orthology relationships between these two clades. Genes belonging to 13 orthogroups were found to be lost in each of the AM non-host species tested (highlighted in yellow). Key: Marpa: *M. paleacea*, Marpo: *M. polymorpha*, Phypa: *P. patens*, Picab: *P. abies*, Arath: *A. thaliana*, Medtr:

M. truncatula. In the figure, the AM host status of each species is indicated using a plant cell containing a blue arbuscule (AM host) or an empty cell (AM non-host) above the species name.

4.3. Discussion

The discovery of symbiosis gene homologs in non-flowering plants led to the hypothesis that these homologs have functions in symbiosis across all land plants (Wang et al., 2010). Testing this hypothesis has hitherto not been possible due to the lack of methods for the generation of mutants in AM host species from these lineages (Rensing, 2017). The work presented in this chapter has begun to address this through the comparative analysis of AM host and non-host species from liverworts and angiosperms. The results of this analysis support a role in the AM symbiosis for the liverwort homologs of angiosperm symbiosis genes. These liverwort genes exhibited two predictors of symbiotic function, transcriptional induction upon symbiosis and specific co-elimination upon the loss of symbiosis.

Recent phylogenomic studies (Delaux et al., 2014; Favre et al., 2014; Bravo et al., 2016) in the angiosperms have shown that a large set of genes (belonging to 72 orthogroups) with functions in the AM symbiosis are committed to the AM symbiosis in the angiosperms i.e. genes belonging to these 72 orthogroups are consistently lost in angiosperm species that have lost the AM symbiosis and the fate of these genes is tightly linked to the fate of the symbiotic trait. Phylogenetic analysis of liverworts and angiosperms showed that homologs of 66 of these orthogroups were present in the liverworts. Differential expression analysis of mycorrhized and control samples of the liverwort *L. cruciata* revealed that homologs of 30 of these orthogroups are transcriptionally induced during symbiosis in this liverwort, a characteristic exhibited by several of these orthogroups in the angiosperms. Using a comparative genomic analysis of related AM host and non-host liverworts, *M. paleacea* and *M. polymorpha* respectively, it was revealed that 17 of these orthogroups were lost upon the loss of symbiosis in the AM non-host *M. polymorpha*.

Comparisons of the transcriptional induction status and commitment status of the orthogroups indicate that despite exhibiting transcriptional induction during symbiosis, a different 17 orthogroups were not committed to symbiosis in the liverworts, as they are retained in *M. polymorpha*. As these genes exhibited transcriptional induction upon symbiosis, it is likely that they were already recruited

to perform functions during symbiosis in the liverworts. In contrast, all 17 orthogroups are both transcriptionally induced upon symbiosis as well as lost upon the loss of symbiosis in the angiosperms. The observed uncoupling of the transcriptional induction during symbiosis and the elimination upon loss of symbiosis for genes in the liverworts suggests that the processes of recruitment and commitment of genes to symbiosis occurred in distinct stages during the evolution of these genes in the land plant lineage.

Although it is not yet possible to predict what leads to the loss of the AM symbiosis in specific clades after a long history of maintenance of this trait in the land plant lineage, our understanding of what happens following the loss of the AM symbiosis has been expanded as a result of this study. Detailed sequence analysis revealed that the accumulation of deleterious mutations causing pseudogenisation of homologs of canonical symbiosis genes accompanied the loss of symbiosis in *M. polymorpha*. Estimation of selective constraints on other symbiosis gene homologs revealed that selection was relaxed in several of these genes. Long-term relaxation of selective constraints and the absence of purifying selection have previously been reported to be the major drivers of gene loss (Chapple and Guigó, 2008) and is likely that this is the major driving factor behind the loss of genes following the loss of the AM symbiosis in plants.

It is unclear as to what caused the loss of the AM symbiosis and resulted in the change in the selective constraints on the symbiosis genes in these AM non-host plant species after nearly 400 million years of maintenance of these genes in the ancestors of these lineages. It has previously been suggested that the specific changes in the environment of the plant that render the maintenance of the AM symbiosis obsolete or even harmful could drive the loss of this trait (Delaux et al., 2014). Based on currently available data, it is not possible to ascertain whether the process of loss of symbiosis was a result of adaptive or regressive evolution in the numerous independent cases of loss of this trait. A case for adaptive evolution being the driver of loss of the AM symbiosis could be made based on observations that some symbiosis genes such as *RAM2* have been found to be required by filamentous plant pathogens for colonisation of the plant root (Wang et al., 2012). In the past, similar hijacking of symbiosis genes by plant pathogens to facilitate pathogenic infection may have caused plants to be faced with a scenario where the costs of retaining the ability to associating with AM fungi, viz. allowing for pathogenic infection, may have outweighed the benefits of the AM

association and led to the adaptive deactivation of key symbiosis genes. Alternatively, loss of the AM association in certain lineages may have been the result of adaptation to specific niches where the benefits provided by the AM association are rendered useless (such as the evolution of specialized nutrient acquisition strategies such as carnivory) or access to AM fungal partners is limited (such as in plants that have secondarily moved to aquatic environments) (Brundrett, 2009). In these cases, the loss of the AM association may have been a case of regressive evolution where the loss was driven by the lack of fitness benefits associated with retaining the AM trait and the genes associated with the trait. Further studies comparing closely related AM host and non-host species representing the entire spectrum of land plant lineages are required to learn more about the factors that lead to the loss of this highly conserved trait in the land plants. With the ever-increasing amount of genomic data from clades of plants that were previously unexplored, such studies will soon be possible and provide even more insights into the processes driving the maintenance and loss of this ancient symbiosis.

5

General Discussion

Chapter 5: General Discussion

Research in angiosperm model plant species such as *Medicago truncatula* (Huguet et al., 1995), *Lotus japonicus* (Handberg and Stougaard, 1992) and *Oryza sativa* (Nakagawa and Imaizumi-Anraku, 2015) has uncovered the variety of processes that occur in the plant root to successfully accommodate AM fungi during the formation of the AM symbiosis (Luginbuehl and Oldroyd, 2017; MacLean et al., 2017). Through forward and reverse genetic studies in these species, the genes regulating these processes have been identified and extensively characterised (Oldroyd, 2013). Several of the initially identified genes with functions in the AM symbiosis were ones known to possess functions in regulating the RN symbiosis (Catoira et al., 2000; Kistner et al., 2005). Due to their functions in both AM and RN symbioses, these genes are referred to as the common symbiosis (SYM) genes (Oldroyd et al., 2009). This set of genes with functions identified in both AM and RN symbioses includes *DMI2/SYMRK* (Endre et al., 2002; Stracke et al., 2002), *DMI1/POLLUX/CASTOR* (Ané et al., 2004; Imaizumi-Anraku et al., 2005; Charpentier et al., 2008), CNGC15a,b,c (Charpentier et al., 2016), MCA8 (Capoen et al., 2011), NENA (Groth et al., 2010), NUP85 (Saito et al., 2007), NUP133 (Kanamori et al., 2006), *DMI3/CCaMK* (Lévy et al., 2004; Mitra et al., 2004; Tirichine et al., 2006), *IPD3/CYCLOPS* (Messinese et al., 2007; Yano et al., 2008), *NSP1* (Smit et al., 2005; Delaux et al., 2013a; Nagae et al., 2014), *NSP2* (Kalo et al., 2005; Maillet et al., 2011) and *VAPYRIN* (Feddermann et al., 2010; Pumplin et al., 2010; Murray et al., 2011). Studies into the evolution of these genes have found that these genes are highly conserved in the angiosperms (Zhu et al., 2006; Pumplin et al., 2010; Delaux et al., 2013b). Functional tests of conservation have also been conducted using trans-complementation assays of *L. japonicus ccamk*, *cyclops*, *castor*, *pollux* and *symrk* mutants. These tests revealed that the respective orthologs of these genes from the monocot *O. sativa* could rescue the symbiosis defects in these mutants and restore functional AM symbioses (Chen et al., 2007; Banba et al., 2008; Chen et al., 2008; Markmann et al., 2008; Yano et al., 2008). In addition to the identification of genes with functions in both AM and RN symbioses, several genes which function specifically during the AM symbiosis have also been identified. These include genes such as *RAM1* (Gobbato et al., 2012; Pimprikar et al., 2016; Keymer et al., 2017), *RAM2* (Wang et al., 2012), *RAD1* (Park et al., 2015; Xue et al., 2015), *STR* (Zhang et al., 2010; Gutjahr et al., 2012), *STR2* (Zhang et al., 2010; Gutjahr et al., 2012) and *PT4* (Harrison et al., 2002). These genes were also found to be conserved

within the angiosperms (Gutjahr et al., 2012; Delaux et al., 2013b; Delaux et al., 2014; Xue et al., 2015; Bravo et al., 2016).

Despite the occurrence of the AM symbiosis in all major lineages of land plants, studies into the genetic mechanisms governing the AM symbiosis in non-flowering plants have been lacking. It has been shown that the relationships formed by members of non-flowering plant lineages with AM fungi are *bona fide* symbiotic associations with both partners receiving benefits from the association (Humphreys et al., 2010). Additionally, genes governing the AM symbiosis have been found outside the angiosperms. Orthologs of three symbiosis genes (*DMI1, CCaMK and CYCLOPS*) were identified in bryophytes, whose members are considered to be the earliest diverging land plants (Wang et al., 2010). Furthermore, trans-complementation assays using *M. truncatula ccamk* mutants revealed that transgenic introduction of the liverwort and hornwort *CCaMK* orthologs could restore a functional symbiosis (Wang et al., 2010). Based on these observations, it was suggested that the orthologs of symbiosis genes in the bryophytes may function during symbiosis. Furthermore, as orthologs of these genes were found to be conserved in several lineages of land plants, it was suggested that these genes might perform similar functions in all land plants (Wang et al., 2010). Testing the functions of these symbiosis gene orthologs in bryophytes was not possible at the time due to the unavailability of suitable model systems with the molecular tools required for the study of the AM symbiosis in non-flowering plants.

The aim of the current study was to learn more about the AM symbiosis in non-flowering plants:

(i) Using comprehensive phylogenetic analyses to understand the stages and processes involved in the evolution of symbiosis genes in plants
(ii) Through the establishment and study of a suitable non-flowering plant model system
(iii) By testing whether symbiosis gene homologs in non-flowering plants have symbiotic functions

Since the previous study examining symbiosis gene homologs outside the angiosperm lineage was published, numerous discoveries about novel genes with functions in symbiosis have been made (Luginbuehl and Oldroyd, 2017; MacLean et al., 2017). Furthermore, in the past few years, genomes and transcriptomes of plants belonging

to several clades relevant to the evolution of the AM symbiosis such as the charophytes and bryophytes have become available (Delaux et al., 2013b; Matasci et al., 2014; Rensing, 2017). In Chapter 2, using genomic and transcriptomic data from these clades, I conducted a comprehensive phylogenetic analysis of genes shown to have roles in the AM symbiosis in the angiosperms. Previous studies into the evolution of symbiosis genes have focussed mainly on the angiosperm lineage and found that these genes are highly conserved within angiosperm species with a few exceptions (e.g., *A. thaliana*) (Delaux et al., 2014; Favre et al., 2014; Bravo et al., 2016). From the phylogenetic analysis I conducted, it was evident that the evolution of symbiosis genes occurred in distinct steps during the evolution of plants (Figure 2.3-Figure 2.8).

The first step occurred in the ancestors of charophytes sometime after their divergence from the ancestors of chlorophytes. Analysis of advanced charophytes indicate that orthologs of several symbiotic signalling genes (*CERK1*, *DMI1*, *CCaMK*, *CYCLOPS*) appeared during this stage. While the orthologs of DMI1 and CCaMK could also be found in the basal charophytes, comparisons of structural features, known to be important for the respective calcium-encoding and -decoding functions of these proteins, showed that these features were not present in the basal charophytes. By contrast, these features were found to be conserved in the respective advanced charophyte orthologs (Figure 2.8). This finding suggests that in addition to a stepwise acquisition of genes, acquisition of novel structural features in pre-existing components may have also played a role in the evolution of symbiotic proteins. The phenomenon of such changes in structural features contributing to the acquisition of novel functions is well-known (Harlin-Cognato et al., 2006) and has also been previously reported to have occurred during the evolution of the Gibberellin-DELLA signalling module in land plants (Yasumura et al., 2007; Sun, 2011). The second step in the evolution of symbiosis genes in plants occurred in the ancestors of the bryophytes sometime after their divergence from the ancestors of charophytes. During this stage, orthologs of several of the remaining symbiosis genes (*DMI2/SYMRK*, *NSP1, NSP2, RAM1, RAM2, VPY, STR, STR2*) evolved. Thus, in the common ancestors of liverworts and angiosperms, orthologs of most canonical symbiosis genes had evolved. Despite this conservation in most genes, it was found that for a few genes (*PT4, HA1*) only homologs could be found in the liverworts and that independent

diversification of these genes had occurred in both liverwort and angiosperm lineages as a result of independent gene duplications in these lineages.

Stepwise acquisition of genes over long periods of evolution have previously been reported for other processes such as vertebrate blood coagulation (Doolittle, 2009) and eukaryotic centriole-assembly (Carvalho-Santos et al., 2010). Interestingly, in these processes, the pathways studied were found to function in the respective processes throughout their evolution. The progressive steps involving acquisition of additional genes were found to result in increased complexity in form or regulation of the relatively primitive form of the process that had appeared during the initial stages of gene acquisition. It is unclear whether the same is true for plant-fungal symbioses. As no forms of symbiosis between charophytes and mycorrhizal or related fungi have been reported, it is not possible to ascertain whether the orthologs of the symbiosis genes in the charophytes function in a charophyte-fungal symbiosis. At this stage, it is equally likely that these charophyte genes could be performing roles in processes other than symbiosis. The absence in the charophytes of several key symbiosis genes (*DMI2*/*SYMRK, RAM1, RAD1* and *VPY)* required for AM fungal colonisation supports this possibility. Furthermore, based on the identities of the symbiosis gene orthologs found in the charophytes it is tempting to predict roles for these genes outside of symbiosis. For instance, CERK1 is well-known for its role in pathogenic fungal perception in a wide range of plants and it was recently shown that this function of CERK1 is conserved in the moss *P. patens* suggesting CERK1 functions in pathogen perception maybe conserved across all land plants. Thus, it is possible that this function of CERK1 is even more ancient than previously thought and that the charophyte CERK1 ortholog could be playing a role in charophyte pathogen perception. With regards to DMI1, CCaMK and CYCLOPS, as calcium signalling has been shown to be required for an astonishing variety of cellular processes across the tree of life, it is possible that the charophyte orthologs of these proteins play similar calcium-encoding and decoding roles as their counterparts in land plants but in a non-symbiotic context. In this scenario, while the orthologs of these genes would have originally evolved in a non-symbiotic context, the appearance of the other symbiosis gene orthologs in the ancestral land plants may have contributed to the co-option of these genes into symbiosis signalling. Thus, the evolution of these genes in the charophytes may have made the eventual evolution of symbiosis in the land plants possible and could be regarded as a pre-adaptation to the evolution of symbiosis.

Similar co-option of ancestral algal genes to perform similar roles but in novel pathways in the land plants has been found to have occurred in the case of the *KNOX* and *BELL* genes where genes belonging to these families regulate mating in the chlorophyte *Chlamydomonas* (Lee et al., 2008) while their homologs have been recruited into the regulation of sporophyte development in the land plants (Singer and Ashton, 2007; Sakakibara et al., 2008; Furumizu et al., 2015).

Detailed analysis of the evolution of these symbiosis genes revealed that several well-known processes contributed to their evolution: (i) several of the canonical symbiosis genes evolved through gene duplications from homologous ancestral genes at various points during the evolution of plants; (ii) Gene fusions were also found to have played an important role in the evolution of several symbiosis genes through the fusion of existing domains that gave rise to novel domain combinations resulting in the evolution of these genes; (iii) *De novo* evolution from previously non-coding sequences was found to have led to the appearance of at least one symbiosis gene in plants. While the domain architectures of the multi-domain symbiotic proteins such as CERK1, DMI2/SYMRK, CCaMK, VAPYRIN and RAM2 are conserved in the different plant species where they are found, it remains to be tested whether these different domains have undergone any recombination with equivalent domains present in other respective homologous multi-domain proteins. Indeed, several of these genes belong to large multigene families that could provide the source material for extensive recombination to take place. To test whether this is the case, the phylogenetic workflow used in the present study would need to be modified through the inclusion of individual phylogenies of the constituent protein domains and comparisons of these phylogenies to those of the full-length protein trees as well as the species tree. Alternatively, phylogenetic network construction methods allowing for detection of reticulate evolution would need to be employed. The value of the application of such methods to study the evolution of multigene families has been especially apparent for genes of the Major Histocompatibility Complex in animals (Jakobsen et al., 1998; Wittzell et al., 1999; Bos and Waldman, 2006) where extensive recombination has been shown to have taken place.

In addition to the analysis of symbiosis genes across the plant lineage, previously proposed evolutionary hypotheses for the symbiosis genes *CCaMK* (Harper and Harmon, 2005; Wang et al., 2015) and *SYMRK* (Markmann et al., 2008) were also revisited in Chapter 2 using a dataset representing more diverse taxa with higher

species density than the original studies. The results of the *CCaMK* phylogenetic analysis (Figure 2.9) revealed that this gene is not plant-specific as previously thought (Wang et al., 2015). This discovery calls into question some of the hypotheses previously proposed regarding the evolution of three related protein families regulated by calcium – the CAMKs, CCaMKs and CDPKs. Based on the occurrence of CCaMKs only in plants as opposed to the occurrence of CDPKs in both plants and protists, previous studies have proposed that CCaMKs are likely to have evolved from ancestral CDPKs (Chen et al., 2017). The analysis of recently generated protist transcriptomes (Keeling et al., 2014) in the current study revealed that CCaMKs are also found in the protist lineage. The phylogeny suggests that CCaMKs and CDPKs are monophyletic protein families that likely arose independently in the ancestors of plants and protists. Following their appearance in the plant-protist ancestors, the CDPKs diversified in both plant and protist lineages, while the CCaMKs have been maintained as singleton genes in most protist and plant genomes. These protein kinases are thought to be differentiated by their ability to bind calcium (Harmon et al., 2000). CCaMKs have been shown to bind calcium both directly through their EF-hand domains and indirectly through their calmodulin-binding domain. On the other hand, the CDPKs only possess EF-hand domains and the CAMKs only possess a calmodulin-binding domain and therefore these proteins are capable of only direct and indirect binding of calcium respectively. It has recently been discovered that CDPKs also likely possess calmodulin-binding abilities (Bender et al., 2017). Together, these results suggest that previously published classifications of the calcium-regulated kinase families are likely incorrect and warrant revisiting.

Based on a previous phylogenetic analysis of SYMRK proteins from angiosperms (Markmann et al., 2008), it was proposed that changes in its domain architecture accommodated the recruitment of SYMRK to a novel function in bacterial endosymbiosis from its ancient function in fungal symbiosis. This change in domain architecture was predicted to have occurred specifically in the lineage leading to the rosids - the clade containing all known species capable of forming symbioses with bacteria. Phylogenetic analysis (Figure 2.11) using a taxonomically denser and more diverse set of genomes and transcriptomes compared to the original study revealed that the novel domain architecture predicted to have evolved in the rosids was, in fact, the ancestral form in the land plants. This conclusion was reached based on the occurrence of this domain architecture in basal angiosperm and non-flowering plant

species. Analysis of changes in the domain architecture showed that SYMRKs from specific non-rosid lineages of angiosperms have lost certain domains. These results indicated that previously observed correlations between changes in domain architecture of SYMRK and the occurrence of bacterial endosymbioses was a result of undersampling. Together, the data in Chapter 2 speak to the utility of phylotranscriptiomic analyses and to the importance of taxon sampling in formulating phylogenetic hypotheses.

Studies into the molecular network governing the AM symbiosis in angiosperms have largely been made possible by the availability of numerous model plant species with established molecular methods (Handberg and Stougaard, 1992; Huguet et al., 1995; Nakagawa and Imaizumi-Anraku, 2015). Such studies have not been possible in the non-flowering plants as all the plant species established as models to-date in these lineages are incapable of associating with AM fungi. To address this, in Chapter 3, I tested candidate liverwort species for their ability to associate with AM fungi using methods established for *in vitro* colonisation of angiosperms by AM fungi. In these experiments, I found that the liverwort model plant species *M. polymorpha* was not colonised by the AM fungus *R. irregularis* and therefore could not be used as a model for studying the AM symbiosis. On the other hand, the liverwort *M. paleacea* was reliably colonised by *R. irregularis* in the experimental conditions tested (Figure 3.2). *M. paleacea* thalli inoculated with *R. irregularis* were sectioned and imaged to document the progression of fungal colonisation. The imaging showed that fungal infection of the thallus initiated at the ventral epidermis through the rhizoids and culminated in the formation of arbuscules in the upper parenchymatous cells of the thallus. *M. paleacea* was selected as the species for which other tools were developed to establish it as a model liverwort for the study of symbiosis. Interestingly, it was observed that upon mycorrhizal colonisation, *M. paleacea* thalli accumulated a brown pigment and that this accumulation was specific to colonised thalli. This has also been previously reported (Humphreys et al., 2010). One of the main limitations to identifying genes with functions in the AM symbiosis in plants has been the inability to visualise the progression of AM fungal colonisation without the use of staining and microscopy techniques. The only published macroscopic large-scale forward genetic screen for genes with functions in the AM symbiosis was made possible by a similar visual pigmentation marker for mycorrhizal colonisation in maize (Paszkowski et al., 2006). Thus, the pigmentation observed upon mycorrhization in *M. paleacea* might

allow for similar forward genetic screens for genes with functions in the AM symbiosis in this liverwort.

Methods for *in vitro* propagation and transformation using *Agrobacterium tumefaciens* were modified from those previously developed for *M. polymorpha* (Ishizaki et al., 2016) and successfully applied to *M. paleacea* (Figure 3.3). As induction of the sporophyte generation of *M. paleacea* proved unsuccessful, transformation methods solely making use of the gametophyte generation were used (Figure 3.5). Through a modification of the thallus-cutting-based transformation method previously established for *M. polymorpha* (Kubota et al., 2013), a high-throughput method for the transformation of *M. paleacea* gametophytes was established. An alternate method developed for the transformation of *M. polymorpha* gametophytes (Tsuboyama-Tanaka and Kodama, 2015) was also tested on *M. paleacea*. With minor modifications, this "agartrap" method could be applied to *M. paleacea* to generate transformants at higher efficiency than the thallus-cutting-based method. It was found that the blend-based method could be used for both small-scale and large-scale studies as this method was amenable to scaling with a simple adjustment of the size of input material and the transformation volume. Thus, it may be possible to adapt this method for applications such as forward genetic screens in the future. On the other hand, the agartrap method is more suited to the study of individual genes and similar targeted approaches.

Recent studies in angiosperms have shown that these plants use similar mechanisms for the detection of pathogenic and symbiotic fungi (Miyata et al., 2014; Zhang et al., 2015). These observations were enabled by the comparative analysis of plant perception of and response to pathogenic and symbiotic fungi. To enable such studies in the liverworts, pathogenic fungi that infect liverworts are required. To address this, a pathogenic fungus, *Trichoderma virens*, previously identified as capable of infecting *M. polymorpha* (Jessica Nelson, personal communication) was tested on *M. paleacea* thalli (Figure 3.6). These tests revealed that the pathogen was also capable of infecting *M. paleacea*. SEM images of infected plants indicated the growth of the fungus on both thallus and rhizoid surfaces.

The genome and transcriptome sequencing of *M. paleacea* conducted to establish sequence resources (Figure 3.7) are also described in Chapter 3. I made 28 genome assemblies using 10 assembly tools from Illumina data and selected the most

contiguous and complete version to designate as the draft genome assembly of *M. paleacea* (Table 3.3). *De novo* transcriptome assembly was conducted in parallel to aid the annotation of genic regions of the *M. paleacea* genome. Having established these methods and resources for *M. paleacea*, the possibility of inducing targeted mutations in genes in this plant was explored using a CRISPR/Cas9-mediated mutagenesis system previously used for *M. polymorpha* (Sugano et al., 2014). Preliminary experiments revealed that the system used for *M. polymorpha* cannot be used as is for inducing mutations in *M. paleacea*. Changes in the promoter used for driving guide RNA expression are likely required to adapt this system for use in *M. paleacea*. Alternatively, other mutagenesis approaches such as TALEN-mediated gene targeting could be used as it has recently been shown that a TALEN system could be used to induce mutations in *M. polymorpha* at a high efficiency (20%) (Kopischke et al., 2017).

As it was not possible to develop reverse genetic tools to knockout symbiosis gene homologs in *M. paleacea*, in Chapter 4, I attempted to test whether these genes have functions in the AM symbiosis in liverworts using an alternate approach. Phylogenomic studies (Delaux et al., 2014; Favre et al., 2014; Bravo et al., 2016) have revealed that a set of angiosperm genes belonging to 72 orthogroups are "committed" to symbiosis – they are specifically lost when a plant species loses the ability to form the AM symbiosis. Several of these genes (88% of genes with microarray probe-sets available) were also found to exhibit a specific pattern of transcriptional induction during the progression of the AM symbiosis in angiosperms (Bravo et al., 2016). Using these co-elimination and transcriptional induction patterns as indicators of symbiotic function, I tested whether non-flowering plant homologs of these genes are likely to perform symbiotic functions in non-flowering plants. By reconstructing the phylogenies of these symbiosis-committed genes, I found that homologs for 66 of these genes could be found in the liverworts (Figure 4.4). RNA-seq analysis revealed that 30 of these homologs are transcriptionally induced in the liverworts during symbiosis like their angiosperm counterparts (Figure 4.5). Comparative analysis of the liverwort AM host *M. paleacea* and non-host *M. polymorpha* species revealed that 17 of the 66 orthogroups conserved in liverworts are also specifically lost upon the loss of symbiosis in the liverworts (Figure 4.6). This list of 17 genes included several canonical symbiosis genes such as *SYMRK, CCaMK, CYCLOPS, RAM1, RAD1, STR* and *STR2*, suggesting that these genes are indeed required for the AM symbiosis in

liverworts. To gain insights into how these committed genes are lost upon the loss of symbiosis, syntenic analysis of genomic regions harbouring symbiosis gene homologs was conducted between AM host and non-host liverworts. The results of this analysis indicated that homologs of 5 canonical symbiosis genes were pseudogenised in the AM non-host liverwort, *M. polymorpha*, upon the loss of symbiosis (Figure 4.7). By modelling the selective constraints operating on these pseudogenes and other genes retained in *M. polymorpha* (Table 4.1,Table 4.2), it was revealed that the relaxation of purifying selection caused the pseudogenisation events and likely resulted in the complete deletion of other symbiosis genes.

The AM symbiosis occurs in completely different tissues (thalli versus roots) and generations (gametophyte vs sporophyte) in the liverworts and angiosperms. The finding that symbiosis gene homologs are required for symbiosis in the liverworts suggests that, despite these differences, the same genetic pathway regulates the AM symbioses in liverworts and angiosperms. As the AM symbiosis is predicted to have evolved in the dominant gametophyte generation of the earliest land plants (Brundrett, 2002), the exact details on how this symbiosis was transferred to the sporophyte generation remain unclear. The sporophytes in these early land plants are predicted to have been entirely dependent on the gametophytes for their survival and were not free-living similar to extant bryophytes (Kenrick and Crane, 1997). Thus, as in the case of extant bryophytes, the AM symbiosis is likely to have been restricted to the gametophyte generation in these early land plants. The transfer of the AM symbiosis to the sporophyte generation may have occurred in the land plants either accompanying or post the evolution of the first free-living sporophytes. As no macrofossils of these early land plants have been discovered to-date (Brundrett, 2002), it is not yet possible to test whether this was indeed the case but some clues about the evolution of the AM symbiosis in the sporophytes are available from fossils of early vascular plants whose ancestors had already evolved a free-living sporophyte generation (Kenrick and Crane, 1997). Fossils of these plants show that unlike extant land plants, where the sporophyte and gametophyte generations are morphologically different, these plants had gametophyte and sporophyte generations that were quite similar. These early tracheophytes, referred to as protracheophytes, did not possess roots and surprisingly it was found that arbuscule-like structures characteristic of the AM symbiosis occurred in the aerial axes of these plants (Remy et al., 1994). As these aerial axes could be considered as being analogous to the parenchymatous cells where

arbuscules are formed in extant bryophytes, these structures may represent intermediate stages in the evolution of the AM symbiosis. Thus, as AM symbiosis moved to the sporophyte generation it did so in tissues equivalent to those in the gametophyte that harbour the AM symbiotic structures. From these primitive states, with the evolution of more complex structures such as roots, the AM symbiosis could then have moved underground to give rise to forms of the AM symbiosis similar to those found in the sporophytes of extant vascular plants.

If this was indeed how the AM symbiosis evolved in the sporophyte generation, then this would have necessitated the sporophyte expression of symbiosis genes that were originally expressed solely in the gametophyte generation. Such changes may have been facilitated by changes in the expression patterns of a few key symbiotic transcriptional regulators such as *CYCLOPS* and *RAM1*. Alternatively, the AM symbioses formed in the gametophytes and those formed in the sporophytes could be analogous processes that evolved independently. This scenario would have required that each of these symbioses recruited a conserved set of genes to form analogous structures, a process usually referred to as a "deep homology" (Scotland, 2010). Such a deep homology has been observed in the gene network controlling the formation of rhizoids in the bryophytes and root hairs in the angiosperms where a conserved set of genes was independently recruited into the processes regulating the formation of these analogous structures (Jang et al., 2011; Pires et al., 2013; Tam et al., 2015; Breuninger et al., 2016; Proust et al., 2016). Alternatively, as the exact nature of the life cycle of the earliest land plants are not known, owing to the lack of macrofossils of these plants, it is also plausible that the evolution of the AM symbiosis occurred simultaneously in the gametophytic and sporophytic generations of the earliest land plants, the extinct ancestral bryophytes. In this scenario, the earliest land plants may have possessed isomorphic gametophytic and sporophytic generations as do the fossilized plants from the Rhynie Chert (Kenrick and Crane, 1997) unlike the dimorphic generations found in extant bryophytes and vascular plants.

Based on the phylogenomic analyses conducted in Chapter 2 and 4, a model for the evolution of genes regulating the AM symbiosis in plants is presented here (Figure 5.1). Orthologs of canonical symbiosis genes appeared at various points during the evolution of plants either *de novo* or from pre-existing genes through gene duplication and fusion The recruitment of these genes for their functions in the AM symbiosis occurred in early land plants. The genes recruited for symbiosis in early land plants

included those that had pleiotropic functions in other processes as well as those that solely possessed specialised functions in symbiosis. This second set of genes that had functions only in symbiosis did not provide fitness benefits to plants outside of their symbiotic functions and thus their fates became completely determined by the fate of the symbiosis trait. These genes were therefore "committed" to the AM symbiosis. This initial set of symbiosis-committed genes included several well-characterised symbiosis genes such as *DMI2/SYMRK, RAM1, RAD1, STR* and *STR2* that have remained committed to symbiosis throughout land plant evolution. On the other hand, other symbiosis genes such as *RAM2* and *VPY*, which originally possessed pleiotropic functions in early land plants, lost these functions, became specialised for their symbiotic functions and thereby became committed to symbiosis at later points in evolution. To understand how and why these genes lost their pleiotropic functions, identification of these functions would be necessary and this could be achieved by mutating these genes in extant bryophytes such as *P. patens* and *M. polymorpha* that have retained these genes but have lost the ability to form the AM symbiosis.

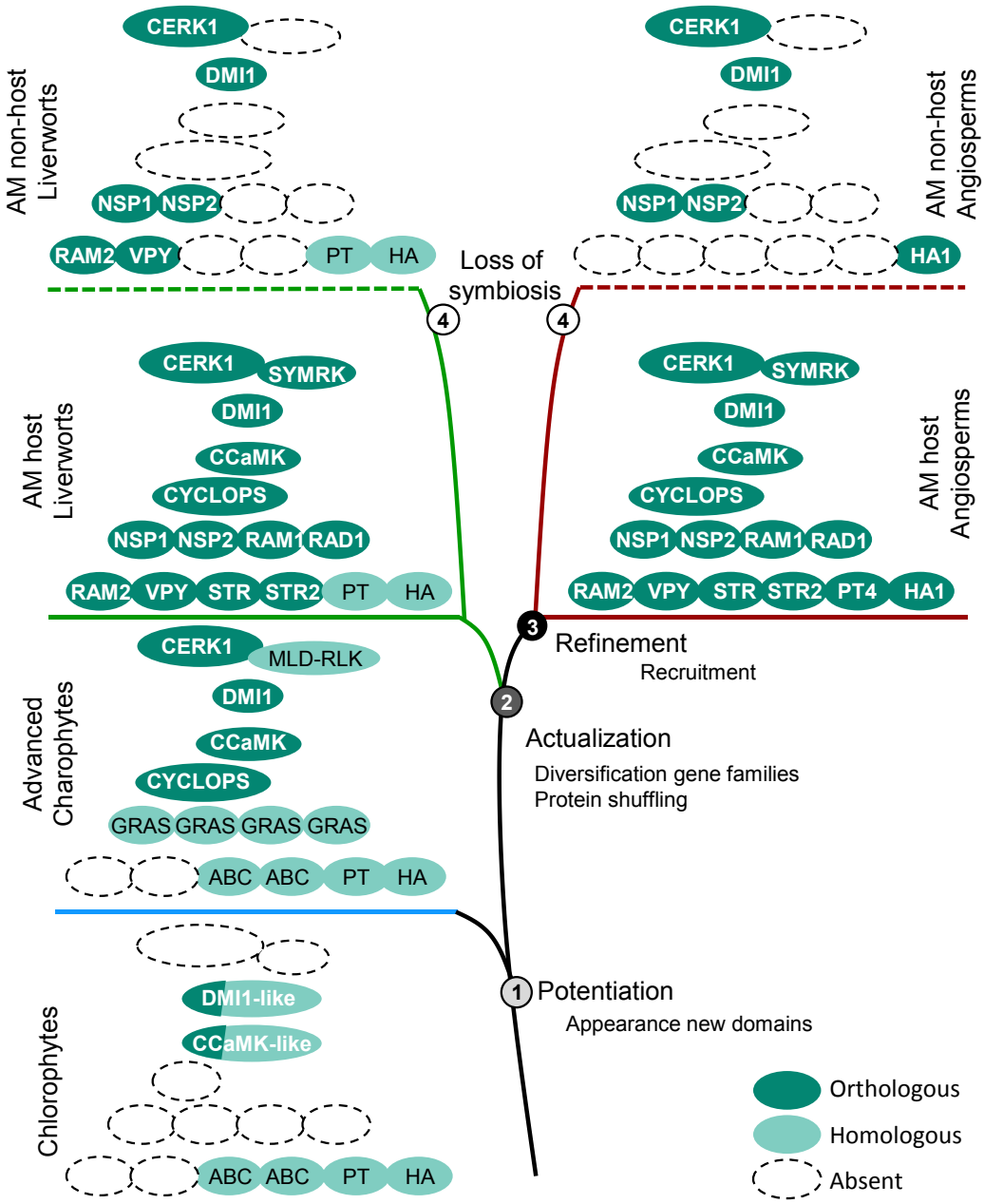

Figure 5.1. A model for the evolution of symbiosis genes in plants.

The model presented here summarises the findings made in chapters 2 and 4 using the resources developed in Chapter 3. Symbiosis genes evolved in a stepwise manner in plants. Several of these genes had specialised roles solely in symbiosis in early land plants i.e. they were "committed" to symbiosis. Some of the remaining genes became committed to symbiosis at later points during plant evolution while others continue to have non-symbiotic roles in extant plant lineages. Genes that were lost in each

lineage following the loss of symbiosis were inferred as being committed to symbiosis in that lineage.

5.1. Conclusions and outlook

The AM symbiosis is an ancient trait that evolved in early land plants and is conserved in the major lineages of land plants. The findings presented here show that despite the numerous changes that have occurred in the plant lineage over time, several genetic components have been conserved across the plant lineage. Transcriptomic and genomic data from previously unexplored clades of plants were used to obtain novel insights into the evolution of the AM symbiosis. By sequencing the *M. paleacea* genome and establishing resources to enable the use of this liverwort as a non-vascular plant model for the study of the AM symbiosis, an avenue for addressing long-standing questions on the evolution of symbiosis was provided. These resources enabled comparative phylogenomic analyses that showed that symbiosis gene homologs likely have roles in symbiosis across all land plants. Furthermore, it was found that plants that lose the ability to form the AM symbiosis lose symbiosis genes from their genome through the relaxation of selective constraints and pseudogenisation.

The work described has made inroads into understanding the evolution of a trait that arose at least 450 million years ago and likely contributed to plant colonisation of land, an event critical to the development of life on earth. Despite the insights gained through the present study, several important questions regarding the evolution of symbiosis remain to be answered. Recent studies have shown that basal lineages of liverworts form symbioses with a different class of fungi (Mucoromycotina) to the class that most other plants associate with (Glomeromycota) (Field et al., 2015c). Previous analyses have uncovered orthologs of canonical symbiosis genes in these basal liverworts (Wang et al., 2010). Whether these symbiosis gene orthologs are also involved in the basal liverwort association with mucuromycotinous fungi will need to be examined. Moreover, reverse genetic resources have recently been developed for AM non-host liverworts (Ishizaki et al., 2013a; Sugano et al., 2014; Kopischke et al., 2017), adapting these methods for use in AM host liverworts will allow for the definitive analysis of whether symbiosis gene homologs function during symbiosis in these plants. Additional studies aimed at identifying the causative events that lead to the loss of symbiosis after such a long history of maintenance in the plant lineage are

also required. Analysis of related AM host and non-host species from across the plant lineage will provide more insights in this area. The discovery of symbiosis gene homologs in algae that are not capable of forming the AM symbiosis presents the possibility that symbiosis gene homologs were not always involved in the AM symbiosis, studies in these algal species will be necessary to understand the function of these genes in these plants. Such studies will provide further insights into the evolution of this ancient interaction between plants and fungi.

6

Materials and Methods

Chapter 6: Materials and Methods

6.1. Bioinformatics

6.1.1. A brief introduction to the methods used for the study of gene evolution

To explore the results of previous evolutionary analyses and of the current study, a background in the major methods used for the study of gene evolution is required and a brief introduction to these is presented below. Examining the evolution of genes (and the proteins they encode) much like the evolutionary study of the traits that these genes determine is, by necessity, quite forensic in its approach (Baum and Smith, 2012). The principal aim of such studies is to model hypothetical ancestral states (genes/traits) and the trajectory taken (process of evolution) to reach observable present states (extant genes/traits). In the case of trait evolution, fossils provide complementary testing mechanisms that can be used to calibrate and test models based on data observed in extant lineages (Forest, 2009). Until recently, such complementary data did not exist for gene evolution. With the advent of ancient DNA sequencing methods (Hagelberg et al., 2015), this is now changing but such data is often only available from samples that are a few hundred years old compared to the millions of years that plants have evolved on land. Thus, sophisticated models and methods are required to overcome this limitation for the study of gene evolution. The field of phylogenetics was developed to address this need (Felsenstein, 2003).

6.1.1.1. Phylogenetic trees

One of the cornerstones of evolutionary analysis is a simple method for the organisation of biological data that can be traced to Charles Darwin's "The Origin of Species" - the phylogenetic tree (Darwin, 1869). This data structure is perfectly suited for representing evolutionary relationships inferred between species and genes (Baum and Smith, 2012). While extant and observable species or genes are placed at the tips of the tree, the points at which the different branches of the tree are connected, called nodes, are used to represent biological entities (species or genes) that existed in the past - the ancestral states (Figure 6.1). Phylogenetic trees are ideal for depicting the evolutionary relationships among genes as the fundamental processes through which

genes evolve (such as gene duplication, horizontal gene transfer, gene loss, speciation) can be represented by the simple addition or removal of branches.

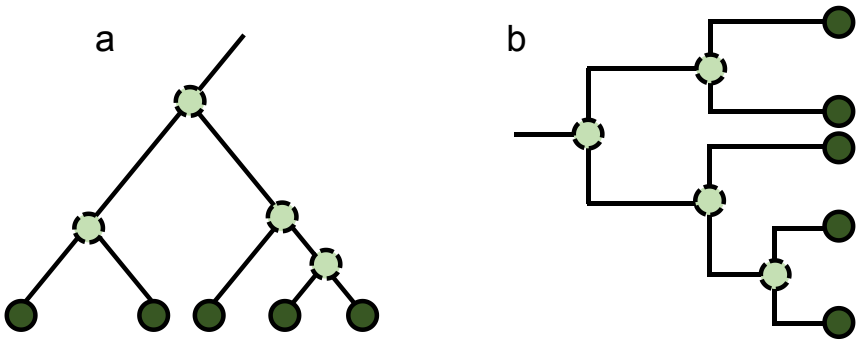

Figure 6.1. Phylogenetic trees are ideal for the representation of gene evolution.

Phylogenetic trees are usually represented either in the (a) phylogram (triangular) or (b) cladogram (rectangular) formats. In the phylogram and cladogram presented here, the evolutionary history of five hypothetical genes (dark green circles) is provided to illustrate the utility of phylogenetic trees. These genes diverged from shared common ancestors (light green circles with dotted outlines) at various points in their evolutionary history and this shared ancestry is reflected through the shared branching points from which the lines leading to the genes emerge.

6.1.1.2. The concept of homology

With advent of genomic data, it became possible to reconstruct the evolutionary history of genes through the comparison of a given set of genomes from different organisms (Tatusov et al., 1997). To signify how genes from different organisms are related, the concept of homology is generally used (Koonin, 2005). Although this concept was originally used for the comparative study of organs between different animals (Owen, 1848), it is now used for the evolutionary analysis of gene relationships. Homology is broadly defined as the property of shared ancestry and genes that arose from shared ancestors are referred to as homologs. Homologs can be further classified into orthologs and paralogs. If two genes (each from a single species) arose from a shared ancestral gene in the last common ancestor of the two species being compared, then they are referred to as orthologs. On the other hand, genes (either in the same species or different species) that are related through a gene

duplication event are referred to as paralogs or co-orthologs. A representation of the homology relationships discussed above is provided using a phylogenetic tree in Figure 6.2.

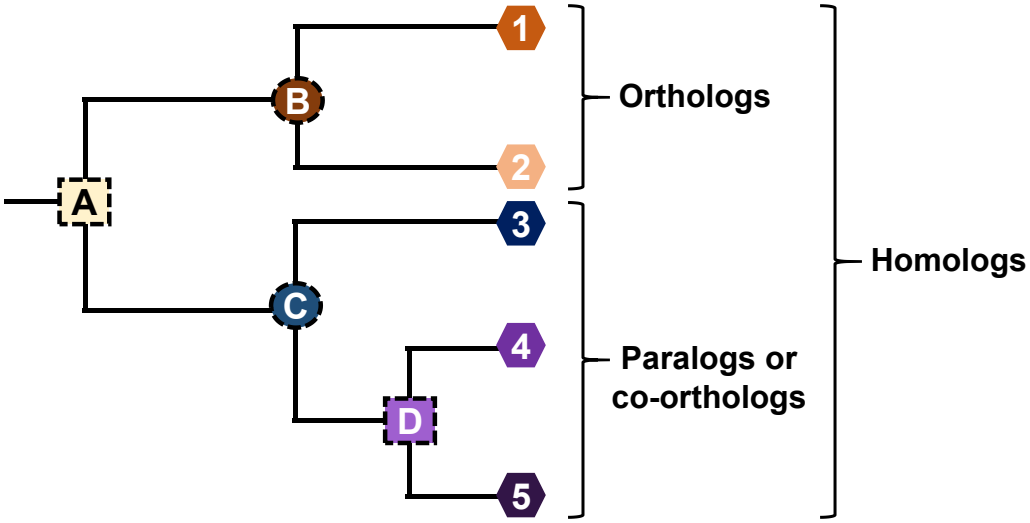

Figure 6.2. Homologs, orthologs and paralogs explained using a phylogenetic tree.

Through the inference of phylogenetic trees, it is possible to ascertain the evolutionary relationships between genes in extant species. In the tree above; the relationships between five hypothetical genes (polygons) is presented. If the branching event that gave rise to these genes was due to speciation (rectangles), then the resultant genes are referred to as orthologs. If the branching was a result of gene duplication, then the resultant genes are referred to as paralogs. These paralogs share a co-orthologous relationship with genes present in other species that diverged from a common ancestor that only had the single ancestral copy of the gene from which the paralogs evolved.

6.1.1.3. Phylogenetic tree construction

Sophisticated methods for inferring the evolutionary relationships between any given set of genes have been developed over the past few decades (Felsenstein, 2003; Holder and Lewis, 2003; Yang and Rannala, 2012). While the actual tools used for the different steps in this process may vary, most phylogenetic analyses follow a workflow similar to that described in Figure 6.3, where the process starts with the collection of the genes/proteins for which evolutionary relationships are to be inferred.

Their sequences are then aligned, and following the selection of a suitable model for the processes assumed to have governed the evolution of these genes, phylogenetic tree estimations are made. While several methods for phylogenetic tree estimation have been developed, the most widely used among these are Maximum-Likelihood based and Bayesian methods (Holder and Lewis, 2003). As both these methods have been shown to have their own advantages and shortcomings, where possible, they are often used together in most evolutionary studies. For larger datasets, ML-based methods are often used as currently available Bayesian methods are computationally intensive and are not as scalable as ML-based methods (Liu et al., 2011a).

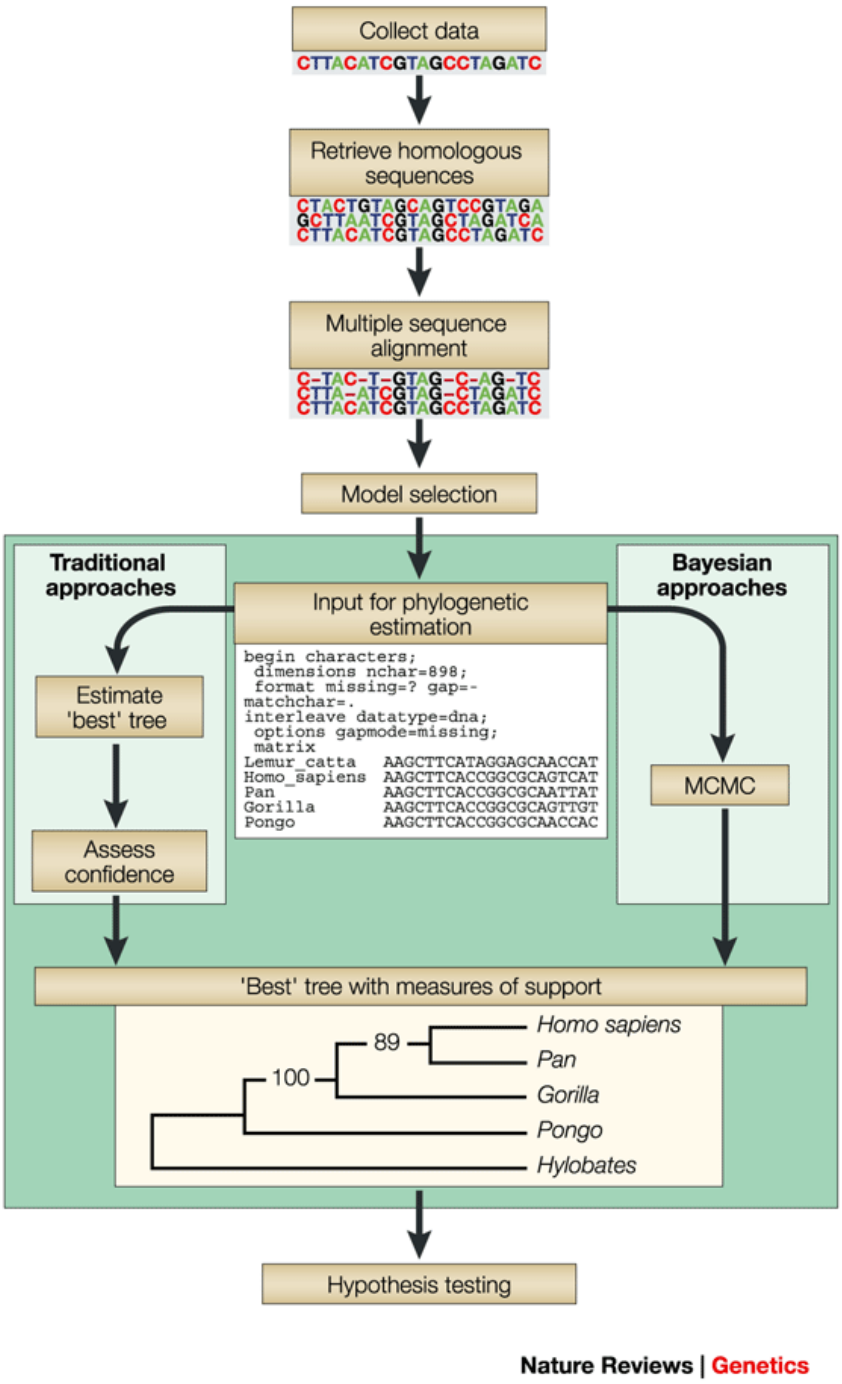

Figure 6.3. General scheme used for the inference of phylogenetic trees.

Figure reprinted by permission from Macmillan Publishers Ltd: Nature Reviews Genetics (Holder and Lewis, 2003), copyright (2003). The initial step in phylogenetic tree reconstruction is the collection of data (gene/protein sequences) using which the tree is to be constructed. This data collection is usually performed in the form of sequence searches to retrieve sequences similar to a query. These sequences are then aligned and the relationships between them are inferred using a suitable

evolutionary model and phylogenetic estimation method. The results of these analyses are presented usually as a phylogenetic tree along with the confidence measures on these results calculated for each branch of the tree. With this tree in hand, it is then possible to test various hypothesis about the evolution of these genes.

6.1.2. Workflow for phylogenetic analysis

The phylogenetic analyses presented in the current study (chapters 2 and 4) were conducted according to the following scheme:

a) Sequence search
b) Multiple Sequence Alignment
c) Construction of initial trees
d) Manual curation of trees
e) Construction of final trees

The details on how each of these steps was conducted are provided below. The UNIX scripts used for this purpose are presented in Appendix A2.

6.1.2.1. Datasets used

The genomic and transcriptomic data used in the current study and the sections of the study in which they were used are presented in Appendix A1.

6.1.2.2. Sequence search

Sequence searches were conducted using BLAST (v2.2.30) (Altschul et al., 1990) and HMMER (v3.1b2) (Finn et al., 2015) to obtain homologs of the symbiosis genes using the respective *M. truncatula* proteins as queries. The sequences used as queries are presented in Supplementary File S4. HMMSCAN was used for genes encoding for proteins with characteristic domains that are part of the Pfam (v29.0) database (Bateman et al., 2002). BLAST was used for the remaining genes. For those genes for which BLAST yielded less than 200 hits, HMMSEARCH was used to obtain further hits. Details of the respective tools used for the different sequence searches conducted in the current study are presented in Table 6.1.

Protein/ orthogroup	Sequence search type	Chapter	Figure
CERK1	BLAST	2	2.3
SYMRK	HMMSCAN	2	2.3
DMI1	BLAST	2	2.4, 2.8
NSP1	HMMSCAN	2	2.4
NSP2	HMMSCAN	2	2.4
RAM1	HMMSCAN	2	2.4
RAD1	HMMSCAN	2	2.4
RAM2	HMMSCAN	2	2.5
STR	BLAST	2	2.5
STR2	BLAST	2	2.5
PT4	BLAST	2	2.6
HA1	BLAST	2	2.6
CCaMK	HMMSCAN	2	2.7, 2.8
CYCLOPS	BLAST	2	2.7
VAPYRIN	HMMSCAN	2	2.7
CDPK/CCaMK	HMMSCAN	2	2.9
SYMRK	BLAST	2	2.10
538	BLAST	4	4.4, 4.5, 4.6, 4.10
AMPa	BLAST	4	4.4, 4.5, 4.6, 4.10
DHY	BLAST	4	4.4, 4.5, 4.6, 4.10
dnaj	BLAST	4	4.4, 4.5, 4.6, 4.10
HYP3	BLAST	4	4.4, 4.5, 4.6, 4.10
LYK	BLAST	4	4.4, 4.5, 4.6, 4.10
STR	BLAST	4	4.4, 4.5, 4.6, 4.10
STR2	BLAST	4	4.4, 4.5, 4.6, 4.10
HYP3	BLAST	4	4.4, 4.5, 4.6, 4.10
hyp2	HMMSEARCH	4	4.4, 4.5, 4.6, 4.10
KINF	BLAST	4	4.4, 4.5, 4.6, 4.10
sin	HMMSEARCH	4	4.4, 4.5, 4.6, 4.10
TAU	HMMSEARCH	4	4.4, 4.5, 4.6, 4.10
HYP7	HMMSEARCH	4	4.4, 4.5, 4.6, 4.10
VPY	BLAST	4	4.4, 4.5, 4.6, 4.10
ipd3	BLAST	4	4.4, 4.5, 4.6, 4.10
CCaMK	BLAST	4	4.4, 4.5, 4.6, 4.10
DMI2	BLAST	4	4.4, 4.5, 4.6, 4.10
HYP5c	HMMSEARCH	4	4.4, 4.5, 4.6, 4.10
ABCB20	BLAST	4	4.4, 4.5, 4.6, 4.10
ampB	BLAST	4	4.4, 4.5, 4.6, 4.10
ap2b	BLAST	4	4.4, 4.5, 4.6, 4.10

continued overleaf

Table 6.1. Details of the sequence searches used to collect sequences for phylogenetic analyses.

Protein/ orthogroup	Sequence search type	Chapter	Figure
FATM	HMMSEARCH	4	4.4, 4.5, 4.6, 4.10
gdsl	HMMSEARCH	4	4.4, 4.5, 4.6, 4.10
ger	BLAST	4	4.4, 4.5, 4.6, 4.10
HMAD	HMMSEARCH	4	4.4, 4.5, 4.6, 4.10
HYP5a	HMMSEARCH	4	4.4, 4.5, 4.6, 4.10
kinA	BLAST	4	4.4, 4.5, 4.6, 4.10
kinE1pair	BLAST	4	4.4, 4.5, 4.6, 4.10
ntr1	HMMSEARCH	4	4.4, 4.5, 4.6, 4.10
PT4	BLAST	4	4.4, 4.5, 4.6, 4.10
qi	HMMSEARCH	4	4.4, 4.5, 4.6, 4.10
RAD1	BLAST	4	4.4, 4.5, 4.6, 4.10
RAM1	BLAST	4	4.4, 4.5, 4.6, 4.10
RAM2	BLAST	4	4.4, 4.5, 4.6, 4.10
RFC	BLAST	4	4.4, 4.5, 4.6, 4.10
amt	BLAST	4	4.4, 4.5, 4.6, 4.10
GRAS	BLAST	4	4.4, 4.5, 4.6, 4.10
CYT733A1	BLAST	4	4.4, 4.5, 4.6, 4.10
CYTB561	HMMSEARCH	4	4.4, 4.5, 4.6, 4.10
GST	HMMSEARCH	4	4.4, 4.5, 4.6, 4.10
hep	HMMSEARCH	4	4.4, 4.5, 4.6, 4.10
hyp4	HMMSEARCH	4	4.4, 4.5, 4.6, 4.10
KELCH	BLAST	4	4.4, 4.5, 4.6, 4.10
kinHpair	BLAST	4	4.4, 4.5, 4.6, 4.10
MLO	BLAST	4	4.4, 4.5, 4.6, 4.10
nfp	BLAST	4	4.4, 4.5, 4.6, 4.10
p450	BLAST	4	4.4, 4.5, 4.6, 4.10
rhi	HMMSEARCH	4	4.4, 4.5, 4.6, 4.10
EXO70	BLAST	4	4.4, 4.5, 4.6, 4.10
HYP6	HMMSEARCH	4	4.4, 4.5, 4.6, 4.10
kinD	HMMSEARCH	4	4.4, 4.5, 4.6, 4.10
cbf	HMMSEARCH	4	4.4, 4.5, 4.6, 4.10
PP2A	HMMSEARCH	4	4.4, 4.5, 4.6, 4.10
CAS	BLAST	4	4.4, 4.5, 4.6, 4.10
CCD	BLAST	4	4.4, 4.5, 4.6, 4.10
SYN	BLAST	4	4.4, 4.5, 4.6, 4.10
ABCB12/mdr1	BLAST	4	4.4, 4.5, 4.6, 4.10
ap2acd	BLAST	4	4.4, 4.5, 4.6, 4.10
dip1	BLAST	4	4.4, 4.5, 4.6, 4.10
kinG	BLAST	4	4.4, 4.5, 4.6, 4.10
SEC	BLAST	4	4.4, 4.5, 4.6, 4.10

Table 6.1. continued.

6.1.2.3. Sequence alignment

Multiple sequence alignment of the hits obtained from the sequence searches were conducted was MAFFT (v7.305) on `auto` mode for the construction of the initial trees (Katoh et al., 2002). For the construction of the curated final trees, MAFFT was run on L-INS-i mode using the BLOSUM62 matrix for alignment scoring and 1000 refinement iterations were conducted. Manual curation of alignments was conducted using Geneious (v10.1) (Biomatters Ltd., NZ) to remove spuriously aligned sequences (Kearse et al., 2012). The sequences used for making the alignments and trees presented in this study are provided in Supplementary File S5.

6.1.2.4. Tree construction

For the first round of tree constructions, FastTree (v2.1.4) was used to construct Maximum-Likelihood-based trees using the alignments described in Section 6.1.1.3 (Arkin et al., 2009). The tree constructions were performed using extra settings recommended in the manual for increased accuracy (-pseudo -spr 4 -mlacc -slownni 2; http://www.microbesonline.org/fasttree/#FAQ). Manual curation of the initial trees to extract the clades of interest for further analysis was performed using Geneious by converting FastTree output into NEXUS format using NCLconverter (v2.1; http://ncl.sourceforge.net/). The final trees presented in the figures in the present study were constructed using RAxML (v8.1.2) after determining the evolutionary model that fit each alignment the best (using the "PROTGAMMAAUTO" feature of RAxML) (Stamatakis, 2006). 20 initial trees were constructed using the estimated model and the best tree from this set was used to compute 1000 bootstrap iterations. The details of the models used for the construction of the trees presented in this study are presented in Table 6.2. For the subtrees presented in Chapter 2, clades of interest were extracted from the above trees, realigned using MAFFT and Maximum-Likelihood trees were built using MEGA6 with the appropriate models previously determined using RAxML. 100 bootstrap computations were performed for these trees.

Protein/orthogroup	Chapter	Figure where tree is used	Model used for tree construction
CERK1	2	2.3	WAG
SYMRK	2	2.3	WAG
DMI1	2	2.4, 2.8	LG
NSP1	2	2.4	JTT
NSP2	2	2.4	JTT
RAM1	2	2.4	JTT
RAD1	2	2.4	JTT
RAM2	2	2.5	LG
STR	2	2.5	LG
STR2	2	2.5	LG
PT4	2	2.6	LG
HA1	2	2.6	LG
CCaMK	2	2.7, 2.8	LG
CYCLOPS	2	2.7	JTT
VAPYRIN	2	2.7	LG
CDPK/CCaMK	2	2.9	LG
SYMRK	2	2.10	WAG
538	4	4.4, 4.5, 4.6, 4.10	LG
AMPa	4	4.4, 4.5, 4.6, 4.10	JTT
DHY	4	4.4, 4.5, 4.6, 4.10	LG
dnaj	4	4.4, 4.5, 4.6, 4.10	JTT
HYP3	4	4.4, 4.5, 4.6, 4.10	JTT
LYK	4	4.4, 4.5, 4.6, 4.10	JTT
STR	4	4.4, 4.5, 4.6, 4.10	JTTDCMUT
STR2	4	4.4, 4.5, 4.6, 4.10	LG
HYP3	4	4.4, 4.5, 4.6, 4.10	JTT
hyp2	4	4.4, 4.5, 4.6, 4.10	VT
KINF	4	4.4, 4.5, 4.6, 4.10	JTT
sin	4	4.4, 4.5, 4.6, 4.10	VT
TAU	4	4.4, 4.5, 4.6, 4.10	LG
HYP7	4	4.4, 4.5, 4.6, 4.10	DCMUT
VPY	4	4.4, 4.5, 4.6, 4.10	JTT
ipd3	4	4.4, 4.5, 4.6, 4.10	JTT
CCaMK	4	4.4, 4.5, 4.6, 4.10	LG
DMI2	4	4.4, 4.5, 4.6, 4.10	JTTDCMUT
HYP5c	4	4.4, 4.5, 4.6, 4.10	JTT
ABCB20	4	4.4, 4.5, 4.6, 4.10	LG
ampB	4	4.4, 4.5, 4.6, 4.10	LG
ap2b	4	4.4, 4.5, 4.6, 4.10	JTT

continued overleaf

Table 6.2. Models used for the construction of the phylogenetic trees in the present study.

Protein/orthogroup	Chapter	Figure where tree is used	Model used for tree construction
FATM	4	4.4, 4.5, 4.6, 4.10	JTT
gdsl	4	4.4, 4.5, 4.6, 4.10	LG
ger	4	4.4, 4.5, 4.6, 4.10	LG
HMAD	4	4.4, 4.5, 4.6, 4.10	WAG
HYP5a	4	4.4, 4.5, 4.6, 4.10	VT
kinA	4	4.4, 4.5, 4.6, 4.10	LG
kinE1pair	4	4.4, 4.5, 4.6, 4.10	JTTDCMUT
ntr1	4	4.4, 4.5, 4.6, 4.10	LG
PT4	4	4.4, 4.5, 4.6, 4.10	LG
qi	4	4.4, 4.5, 4.6, 4.10	WAG
RAD1	4	4.4, 4.5, 4.6, 4.10	JTT
RAM1	4	4.4, 4.5, 4.6, 4.10	LG
RAM2	4	4.4, 4.5, 4.6, 4.10	LG
RFC	4	4.4, 4.5, 4.6, 4.10	JTT
amt	4	4.4, 4.5, 4.6, 4.10	LG
GRAS	4	4.4, 4.5, 4.6, 4.10	JTT
CYT733A1	4	4.4, 4.5, 4.6, 4.10	LG
CYTB561	4	4.4, 4.5, 4.6, 4.10	LG
GST	4	4.4, 4.5, 4.6, 4.10	JTT
hep	4	4.4, 4.5, 4.6, 4.10	LG
hyp4	4	4.4, 4.5, 4.6, 4.10	JTT
KELCH	4	4.4, 4.5, 4.6, 4.10	JTT
kinHpair	4	4.4, 4.5, 4.6, 4.10	LG
MLO	4	4.4, 4.5, 4.6, 4.10	JTT
nfp	4	4.4, 4.5, 4.6, 4.10	JTT
p450	4	4.4, 4.5, 4.6, 4.10	JTT
rhi	4	4.4, 4.5, 4.6, 4.10	LG
EXO70	4	4.4, 4.5, 4.6, 4.10	JTT
HYP6	4	4.4, 4.5, 4.6, 4.10	JTT
kinD	4	4.4, 4.5, 4.6, 4.10	JTT
cbf	4	4.4, 4.5, 4.6, 4.10	JTT
PP2A	4	4.4, 4.5, 4.6, 4.10	JTT
CAS	4	4.4, 4.5, 4.6, 4.10	LG
CCD	4	4.4, 4.5, 4.6, 4.10	JTT
SYN	4	4.4, 4.5, 4.6, 4.10	JTT
ABCB12/mdr1	4	4.4, 4.5, 4.6, 4.10	JTT
ap2acd	4	4.4, 4.5, 4.6, 4.10	VT
dip1	4	4.4, 4.5, 4.6, 4.10	JTT
kinG	4	4.4, 4.5, 4.6, 4.10	JTT
SEC	4	4.4, 4.5, 4.6, 4.10	JTT

Table 6.2. continued.

6.1.2.5. Tree viewing and editing

The phylogenetic trees constructed using FastTree, RAxML and MEGA were analysed and edited manually using Geneious (v10.1) and Dendroscope (v3.5.9). Tree annotation and editing for production of the final figures was performed using Illustrator CS5 (Adobe Inc., USA).

6.1.3. Generation of sequence resources

The genome and transcriptome of *M. paleacea* were assembled using data generated in this study (deposited to NCBI sequence read archive under the accession SRR5196885). Details of the sequencing and assembly are provided in the subsections below. The transcriptomes of *L. cruciata* and *P. endiviifolia* were assembled from publicly available data (Alaba et al., 2015; Delaux et al., 2015) using the procedure detailed in Section 6.1.2.3. Prediction of the amino acid sequences encoded by coding sequences from the genome assemblies of *Closterium*, and *Spirogyra* sp. and from the recently published transcriptome assemblies of *Nitella mirabilis*, *Mesostigma viride*, *Spirogyra* sp. and *Coleochaete orbicularis* was done using AUGUSTUS (v3.2) with the parameter set defined for *Arabidopsis thaliana* (Stanke et al., 2004).

6.1.3.1. Nucleic acid extraction from *M. paleacea*

Genomic DNA from 8-week old thalli was extracted as described previously for obtaining genomic DNA from the hornwort *Anthoceros* (Szovenyi et al., 2015). RNA extraction was also carried out from 8-week old thalli using the RNeasy mini plant kit (Qiagen, USA) following the manufacturer's protocols. Fragmentation of RNA and cDNA synthesis were performed using kits from New England Biolabs (Ipswich, MA) per the manufacturer's protocols with minor modifications. Briefly, 1µg of total RNA was used to purify mRNA on Oligo dT coupled to paramagnetic beads (NEBNext Poly(A) mRNA Magnetic Isolation Module). Purified mRNA was fragmented and eluted from the beads in one step by incubation in 2x first strand buffer at 94ºC for 7 min, followed by first strand cDNA synthesis using random primed reverse transcription (NEBNext RNA First Strand Synthesis Module), followed by random primed second strand synthesis using an enzyme mixture of DNA PolI, RnaseH and *E. coli* DNA ligase (NEBNext Second Strand Synthesis Module). The RNA extraction and cDNA synthesis were performed by LGC Genomics GmbH (Berlin, Germany).

6.1.3.2. Genome sequencing and assembly

For the genome sequencing, short-insert paired-end and long insert mate-pair libraries were produced. For the paired-end library, the DNA fragmentation was done using the Covaris (Covaris Inc., Woburn, MA). The NEBNext Ultra DNA Library Prep Kit (New England Biolabs, Ipswich, MA) was used for the library preparation, and the bead size selection for 300-400 bp fragments based on the manufacturer's protocol. The library had an average insert size of 336 bp. For the mate-pair library, preparation was done using the Nextera Mate-Pair DNA library prep kit (Illumina, San Diego, CA) following the manufacturer's protocol. After enzymatic tagmentation, gel size selection was used to obtain 3-5 kb fragments. The average size that was recovered from the gel was 4311 bp. Sequencing was carried out on an Illumina HiSeq2500 on 2x100bp Rapid Run mode. The library preparation and sequencing were carried out by GENEWIZ (South Plainfield, NJ).

FASTQ files of the reads from the sequenced paired-end and mate-pair libraries were provided by GENEWIZ. Trimming of adapter and low-quality sequences was performed on the paired-end library using Trimmomatic v0.33 with the following parameters (ILLUMINACLIP:TruSeq2-PE.fa:2:30:10 LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:12). The trimmed paired-end reads were used for the assembly of the contigs using multiple assemblers. The following assemblers were used: ABySS (Simpson et al., 2009), SOAPdenovo2 (Luo et al., 2012), CLC Genomics Workbench (https://www.qiagenbioinformatics.com/products/clc-genomics-workbench/), IDBA-UD (Peng et al., 2012), MEGAHIT (Li et al., 2015; Watts-Williams and Cavagnaro, 2015), SPAdes (Bankevich et al., 2012), Platanus (Kajitani et al., 2014), MaSuRCa (Zimin et al., 2013), Meraculous (Chapman et al., 2011), Velvet (Zerbino and Birney, 2008) and Minia (Chikhi and Rizk, 2012). For each of these assemblers, different k-mer sizes/ranges were used as detailed in Table 3.3. Scaffolding of the contigs was performed using the scaffolder of SOAPdenovo2 with the mate-pair library after processing the reads through the NextClip pipeline to only retain predicted genuine long insert mate-pairs. Assembly completeness was measured using BUSCO (Simão et al., 2015) by benchmarking against the plants dataset. Assembly statistics were calculated using the ABySS-fac tool from ABySS (Simpson et al., 2009).

6.1.3.3. Transcriptome sequencing and assembly

M. paleacea cDNA, obtained as described in Section 6.1.2.1, was purified and concentrated on MinElute Columns (Qiagen, USA) and used to construct an Illumina library using the Ovation Rapid DR Multiplex System 1-96 (NuGEN, USA). The library was amplified using MyTaq (Bioline, USA) and standard Illumina TruSeq amplification primers. PCR primer and small fragments were removed by Agencourt XP bead purification. MyTaq remnants were removed through an additional purification step on Qiagen MinElute Columns. Normalisation was performed using Trimmer Kit (Evrogen, USA). The normalised library was re-amplified using MyTaq (Bioline, USA) and standard Illumina TruSeq amplification primers. The normalised library was size selected on a LMP-Agarose gel, removing fragments smaller than 350bp and those larger than 600bp. Sequencing was done on an Illumina MiSeq on 2x300bp mode. The RNA extractions, cDNA synthesis library preparation and sequencing were carried out by LGC Genomics GmbH (Berlin, Germany).

De novo transcriptomes were constructed from the trimmed RNA-seq reads (Trimmomatic v0.33) using the Trinity and SOAPdenovo-Trans assembly packages using default parameters (Grabherr et al., 2011; Xie et al., 2014). The two assemblies were then merged using the EvidentialGene pipeline (Nakasugi et al., 2014). Primary isoforms were extracted from the merged transcriptome assembly and used for further analysis. The same process was repeated on publicly available read data to construct the *L. cruciata* and *P. endiviifolia* transcriptomes.

6.1.3.4. Annotation of the *M. paleacea* genome

Protein and coding sequence predictions from the transcriptome were made using TransDecoder (v2.1.0) using the default parameters (Haas et al., 2013). Annotation of the draft genome assembly of *M. paleacea* was performed using MAKER-P (v2.31.9) (Campbell et al., 2014). To aid the prediction of gene models, gene predictions from *P. patens* and *M. polymorpha* and coding sequences predicted from the *M. paleacea* transcriptome were provided as input to the MAKER pipeline.

6.1.4. Intron-exon structure analysis

To study the evolution of intron-exon structures of *CCaMK* and *CYCLOPS* genes in the plant lineage, *CCaMK* and *CYCLOPS* coding sequences were extracted from the

transcriptomes of *M. paleacea*, *M. truncatula* and *O. sativa* using their respective protein sequences as queries for BLAST searches. The coding sequences were extracted using the "Find Open Reading Frames" tool in Geneious (v10.1). Using these coding sequences as queries the respective genomes of these plants were searched using BLAST to get the genomic regions containing these genes. Once the *CCaMK* and *CYCLOPS* genes from these species were obtained, the coding sequences and genic regions were aligned using MAFFT (default settings) to deduce the intron-exon structures for these genes. The lengths of the introns and exons were calculated using Geneious (v10.1) and these were plotted using Illustrator CS5 (Adobe, USA).

6.1.5. Orthogroup construction

OrthoFinder (Emms and Kelly, 2015) was used to classify the predicted protein sequences from the genomes of *Amborella trichopoda*, *Arabidopsis thaliana*, *Medicago truncatula*, *Oryza sativa*, *Lunularia cruciata*, *Marchantia paleacea* and *Marchantia polymorpha*. All vs all BLAST searches were conducted among these seven species. The results of this BLAST analysis was provided as input to OrthoFinder. OrthoFinder output classifying the proteins encoded by these genomes was used to annotate a tree containing LysMRLK genes previously identified from these genomes (Chapter 2) per the orthogroup each protein was classified into. This was performed manually using the Geneious (v10.1) tree viewer.

6.1.6. Differential expression analysis of *L. cruciata*

RNA-seq data generated from *L. cruciata* samples inoculated with *R. irregularis* spores (AM) and water(mock)-treated samples were provided by Pierre-Marc Delaux and Christophe Roux (LRSV, Toulouse). The details of how these data were generated have been described previously (Delaux et al., 2015). Expression analysis of AM and mock *L. cruciata* samples was performed using the following workflow: trimming of reads using Trimmomatic v0.33, read mapping on to the draft transcriptome using STAR (Dobin et al., 2013), extraction of read counts (aggregated by genes) using the featureCounts command from Rsubread (Shi et al., 2013) and differential expression analysis using edgeR (Robinson et al., 2010). Genes induced greater than 1.5 fold (FDR-corrected $p < 0.05$) in the AM samples compared to the mock were regarded as being significantly transcriptionally induced. FDR correction was performed using the Benjamini-Hochberg method (Benjamini and Hochberg, 1995). The results of the differential expression analysis are provided in Supplementary File S3.

6.1.7. Pseudogene identification and analysis

For the genes whose homologs were found in *L. cruciata* and *M. paleacea* but not in the genome and transcriptome of *M. polymorpha*, detailed searches were conducted to ensure that the absence of these genes was not due to assembly incompleteness. For this, the coding sequences of the respective *M. paleacea* genes were used as queries to BLAST search the genome assemblies of the *M. polymorpha* TAK-1 and CAM1 strains. The genomic regions of *M. polymorpha* obtained as hits were then aligned with the corresponding *M. paleacea* genomic regions for comparisons using Mauve v2.3.1 (Darling et al., 2004) and visualised using Geneious. Synteny analysis was performed to find the corresponding genes in these regions by mapping the transcripts onto the *M. polymorpha* and *M. paleacea* genomic regions upstream and downstream of each gene using Geneious. The identities of the preceding and subsequent genes were confirmed by aligning the *M. paleacea* and *M. polymorpha* sequences. Their functional identity was discovered by BLAST searches against the SWISSPROT database (Boeckmann et al., 2003) and using Pfam.

6.1.8. Estimation of selective constraints using PAML and RELAX

The ratios of synonymous to non-synonymous substitutions (ω) were calculated using PAML following the user manual (Yang, 1997). Briefly, ω ratios for the different models were calculated by aligning the appropriate codon sequences after gap removal and stop codon removal using PAL2NAL (Suyama et al., 2006), followed by ω calculations using PAML using default parameters. A user-defined tree constructed with the orthologous genomic regions from *M. polymorpha*, *M. paleacea* and *L. cruciata* was used to specify the different model parameters. A test for the relaxation of purifying selection on the *M. polymorpha* branch of the tree for each of the genes was conducted using the RELAX framework (Wertheim et al., 2015) using the codon alignments constructed for the PAML analysis.

6.2. Experimental methods

6.2.1. Plant growth and media

M. paleacea thalli used previously by Humphreys and colleagues (Humphreys et al., 2010) were kindly provided by Katie Field and David Beerling (University of

Sheffield). *M. polymorpha* thalli were kindly provided by Jim Hasseloff, University of Cambridge. Plants were grown and maintained on soil (10% peat mixed with sand) in controlled environment rooms under a 16/8-hour day-night cycle at 22ºC and fluorescent illumination with a light intensity of 100 μmol/m^2s. Gemma from these thalli were collected and sterilised based on previously described methods (Vujičić et al., 2010). Briefly, gemma cups were collected using a micropipette tip and placed into a micro-centrifuge tube. 5% sodium hypochlorite solution was used to sterilise the gemmae for about 30 seconds followed by rinsing with sterile water 5 times to remove residual sodium hypochlorite solution. The sterilised gemmae were grown on Gamborg's half-strength B5 medium (composition provided in Table 6.3) on sterile tissue culture plates under a 16/8-hour day-night cycle at 22ºC under fluorescent illumination with a light intensity of 100 μmol/m^2s. Gemma cup production was induced by transferring 4 week old thalli onto Gamborg's half-strength B5 medium supplemented with 1% sucrose.

6.2.2. Media and antibiotics used

The composition of the media used for plant and bacterial growth in the current study are detailed in Table 6.3. Antibiotics (Table 6.4) were added to the media where necessary for the selection of transgenic plants and bacteria.

Medium	Composition for 1 litre
Gamborg's B5	KNO3 2500 mg, CaCl2(2H2O) 150 mg, MgSO4(7H2O) 250 mg, (NH4)2SO4 134 mg, NaH2PO4(H_2O) 150 mg, KI 0.75 mg, H3BO3 3.0 mg, MnSO4(H_2O) 10 mg, ZnSO4(7H2O) 2.0 mg, Na2MoO4(2H2O) 0.25 mg, CuSO4(5H2O) 0.025 mg, CoCl2(6H2O) 0.025 mg, Ferric-EDTA 43 mg, sucrose 2%, pH 5.5, inositol 100 mg, nicotinic acid 1.0 mg, pyridoxine.HCl 1.0 mg, thiamine. HCl 10 mg, kinetin 0.1 mg, 2,4-D 1.0 mg, 10 g agar
M (Minimal)	MgSO4(7H2O) 731 mg, KNO3 80 mg, KCl 65 mg, KH2PO4 4.8 mg, Ca(NO3)2 4H2O 288 mg, NaFe EDTA 8 mg, KI 0.75 mg, MnCl2(4H2O) 6 mg, ZnSO4(7H2O) 2.65 mg, H3BO3 1.5 mg, CuSO4(5H2O) 0.13 mg, Na2MoO4(2H2O) 0.0024 mg, glycine 3 mg, thiamin Cl 0.1 mg, pyridoxin HCl 0.1 mg, nicotinic acid 0.5 mg, myo inositol 50 mg, sucrose 10 g, pH 5.5 For solid medium add Phytagel 5 g
Mod FP (Modified FP)	CaCl2(2H2O) 0.1 g, MgSO4 0.12 g, KHPO4 0.01 g, Na2HPO4(12H2O) 0.150 g, ferric citrate 5 mg, H3BO3 2.86 g, MnSO4 2.03 g, ZnSO4(7H2O) 0.22 g, CuSO4(5H2O) 0.08 g, H2MoO4(4H2O) 0.08 g, NH4NO3 0.5 mM, Formedium agar 8 g, pH 6.0
SRV	3.0 mM MgSO4·7H2O, 0.75 mM KNO3, 0.87 mM KCl, 1.52 mM Ca(NO3)2·4H2O, 0.03 mM KH2PO4; 11 µM MnSO4·H_2O, 1 µM ZnSO4·7H2O, 30 µM H3BO3, 0.96 µM CuSO4·5H2O, 0.02 µM NaFe·EDTA, 0.03 µM (NH4)6Mo7O24·4H2O, 0.01 µM Na2MoO4·2H2O, 0.49 µm V2O5, 0.85 µM CoSO4·7H2O; 5 g l−1 Phytagel (Sigma); pH 5.5 before
SOC (Super Optimal broth with Catabolite repression)	Tryptone 20 g, yeast extract 5 g, NaCl 0.58 g, KCl 0.19 g, MgCl2 2.03 g, MgSO4(7H2O) 2.46 g, Glucose 3.6 g
LB (Lysogeny Broth)	Tryptone 10 g, yeast extract 5 g, NaCl 5 g. To obtain solid medium, 10 g Lab M No.1 agar was added
OM51C	Made by combining 10× 0M51C stock solution 100 ml, sucrose 20 g, L-glutamate 0.3 g, Casamino acid 1.0 g. Adjust to pH 5.5 with 1N KOH. 10 × 0M51C stock solution (4 L) KNO3 80 g, NH4NO3 16 g, MgSO4·7H2O 14.8 g, CaCl2·H_2O 12 g, KH 11 g, EDTA-NaFe (III) 1.6 g, B5 micro components 40 ml, B5 vitamin 40 ml, 0.75% KI 4 ml B5 micro components (100 ml) NaMoO4·2H2O 25 mg, CuSO4·5H2O 2.5 mg, CoCl2·6H2O 2.5 mg, ZnSOO 200 mg, MnSO4·7H2O 1 g, HBO 300 mg B5 vitamin (100 ml) Inositol 10 g, nicotinic acid 100 mg, pyridoxine hydrochloride 100 mg, thiamine hydrochloride 1 g

Table 6.3. Composition of media used for the growth of plants and bacteria in the present study

Antibiotic	Dissolved in	Concentration (µg/ml)
Hygromycin	Water	10
Spectinomycin	Water	50
Carbenicillin	Water	100
Kanamycin	Water	25
Rifampicin	methanol	50
Gentamycin	Water	40
Cefotaxime	Water	100

Table 6.4. Antibiotics used for the selection of transformed plants and bacteria

6.2.3. *In vitro* mycorrhization assays

M. paleacea and *M. polymorpha* thalli grown on plates as described in Section 6.2.1 for 2 weeks were transferred to a 1:10 mixture of peat and sand. To the above soil mixture, chive root inocula harbouring *R. irregularis* prepared as described previously (Rasmussen et al., 2016) were added to obtain a final concentration of 1:5 soil:inocula to facilitate colonisation of *M. paleacea* thalli by AM fungi. The plants were grown for 8 weeks under a 16/8-hour day-night cycle at 22ºC under fluorescent illumination with a light intensity of 100 µmol/m^2s before imaging of mycorrhizal colonisation was performed. For the mycorrhization assays on tissue culture plates, thalli grown from gemmae for 2 weeks on M medium (Table 6.3) were inoculated with 500 *R. irregularis* spores (sterile spores obtained as a solution containing 2500 spores/ml from Agronutrition Ltd, France). The plants were grown for 4 weeks before imaging of mycorrhizal colonisation was performed. Carrot hairy-root cultures harbouring the AM fungus *R. irregularis*, prepared and maintained as previously described (Kosuta et al., 2008), were provided by Jongho Sun (John Innes Centre, UK). To test mycorrhizal colonisation using the carrot hairy-root cultures as inocula, thalli grown from gemmae for 2 weeks were placed on tissue culture plates containing carrot hairy-root cultures grown on M medium (Table 6.3). Fungal colonisation was observed after 4 weeks.

6.2.4. Sectioning and imaging of *M. paleacea* thalli

6.2.4.1. Ink staining

Thalli were collected in 2 ml microfuge tubes and soil/agar attached to the rhizoids was cleaned using tap water. Transverse sections of thalli were obtained by hand using

a scalpel and tweezers under a dissecting scope. The sections were placed on 50mm round petri dishes to which 10% KOH solution was added until all the sections were covered by the solution. After a 3-hour incubation, the sections were rinsed using sterile water and stained with ink (Waterman Ltd, Paris) for 1hr. This was followed by rinsing using sterile water and an incubation in 30% H_2O_2 for 2hrs. Sections were rinsed again using sterile water and placed in 70% ethanol overnight. The above steps were performed at room temperature. The sections were then placed on glass slides and imaged using a Leica M80 stereo microscope.

6.2.4.2. Wheat Germ Agglutinin (WGA) staining

Thalli were collected and cleaned as described above. Fixation, embedding, sectioning, dewaxing and mounting were performed as described previously (Wegel, 2017). Briefly, segments of thalli were fixed in 4% formaldehyde, freshly prepared from paraformaldehyde, using vacuum infiltration. Following overnight fixation, the samples were washed twice with 1xPBS and dehydrated in an ethanol series (30%, 50%, 70%) for 30mineach. The samples were then placed in tissue cassettes, submerged in 70% ethanol and transferred to a tissue processor for wax embedding. The cassettes were then placed in the hot wax chamber of a wax embedding station and 3-4 thallus segments were placed on a thin layer of wax in a metal mould. They were then covered with wax and the wax was allowed to set. 8-12 μm thin sections were obtained using a rotary microtome and placed on Poly-L-lysine slides. Dewaxing of sections was performed using Histo-Clear (National Diagnostics, USA). The sections were then stained using the WGA-Alexa Fluor 488 conjugate (ThermoFisher Scientific, USA) by incubating in a 1x Phosphate-Buffered Saline (PBS) solution containing 0.2 μg/ml WGA-Alexa Fluor 488 in the dark for 30minfollowed by washing using 1x PBS solution. The sections were mounted using Histomount and visualised on a TCS SP5 confocal microscope (Leica Microsystems, UK).

6.2.5. Comparison of *M. paleacea* growth on different media

M. paleacea gemmae were obtained as described in Section 6.2.1. These gemmae were placed on half-strength Gamborg's B5, half-strength M, M, FP and SRV media. The compositions of these media are listed in Table 6.3. The plants were grown for 2 weeks in the conditions described in Section 6.2.1 and the growth of the plants was quantified using a MZ16 stereo microscope (Leica Microsystems, UK). Growth

surface area measurements were made using the EasyLeafArea tool (Easlon and Bloom, 2014) from the microscopy images.

6.2.6. Sexual cycle induction of *M. paleacea*

Sexual cycle induction experiments of *M. paleacea* and *M. polymorpha* were conducted in controlled environment rooms plates under a 16/8-hour day-night cycle at 22º C, 50% relative humidity, with a light intensity of 100 μmol/m^2s. Illumination was provided using LED lights with a light spectrum of 400-760 nm (GrowSun 12W Spot with 6 bands; E-shine systems, China) to provide far-red illumination in addition to visible light. A similar setup was previously shown to induce the sexual cycle in *M. polymorpha* (Ishizaki et al., 2008). Plants were either grown on soil as described in Section 6.2.1 or on tissue culture plates containing half-strength Gamborg's B5 medium supplemented with 1% glucose as described previously (Althoff et al., 2014).

6.2.7. Inoculation of *M. paleacea* with *Trichoderma virens*

M. paleacea thalli grown for 2 weeks from gemmae on half-strength Gamborg's B5 medium were inoculated with a solution of *Trichoderma virens* (ATCC® 9645™) cells re-suspended in sterile water from lyophilised powder. Following inoculation, the plants were grown as described in Section 6.2.1. Pictures of the inoculated thalli were taken at 2, 4, 6 and 8 days post-inoculation using a PowerShot G11 camera (Canon, UK).

6.2.8. Transformation of *M. paleacea* thalli

6.2.8.1. Golden Gate cloning

Golden Gate cloning following the published common syntax for plant synthetic biology (Patron et al., 2015) was used to produce a binary vector (based on backbone vectors from Icon Genetics GmbH, Germany) for the transformation of *M. paleacea*. The level 0(L0), 1(L1) and 2(L2) parts were used to construct the L2 binary expression vector. The L0 parts were synthesised by Life Technologies (ThermoFisher Scientific, UK) after the sequences were domesticated *in silico* to remove BsaI, BpiI and DraIII Type-III restriction sites using Geneious (v10.1). Golden Gate digestion-ligation reactions were performed using BsaI and BpiI for the digestion to obtain L1 and L2 plasmids respectively using the T4 DNA ligase for the ligation as previously described (Engler and Marillonnet, 2013). For the L2 binary expression vector used for the

optimisation of transformation protocols described in Section 3.2.2.2, the L1 modules pL1M-R1-p35S-hptII-tNOS and pL1M-R2-pMpoEF1a-tOCS were used in positions 1 and 2 respectively. For the L2 binary expression vector used for testing CRISPR/Cas9 mutagenesis in *M. paleacea*, the L1 modules pL1M-R1-p35S-hptII-tNOS, pL1M-R2-MpoEF1a-HsCas9-tOCS, pL1M-R3-pMpoU61-nop1target-tPolIII, pL1M-R4-pAtUBI10-mCherry-t35S were used in positions 1, 2, 3 and 4 respectively. The L1 modules were constructed from the appropriate promoter (PU), coding sequence (SC) and terminator (T) L0 modules based on the standardised names of the L1 modules as recommended by Patron and colleagues (Patron et al., 2015).

6.2.8.2. Plasmid amplification using *Escherichia coli*

To amplify the L1 and L2 vectors, chemically competent versions of the *E. coli* strains DH5α and DH10 (Invitrogen, UK) were used respectively. For the bacterial transformation, 1ul of the Golden Gate cloning reaction mixture was added to 20µl of the cells and incubated for 30minon ice. A heat-shock of 42ºC was provided to the cells for 30s, followed by an incubation on ice for 5min. 900 µl of SOC medium (Table 6.3) was added to the cells and they were incubated at 37ºC for 1-1.5h at 220 rpm. 150 µl of the culture was then spread on to plates containing LB medium with the appropriate antibiotics and incubated overnight at 37ºC. Colony PCR was used to check whether the plasmid of interest was successfully transformed. For this purpose, a PCR was set up using 2 µl of forward primer (10uM) (CCCGCCAATATATCCTGTC), 2 µl of reverse primer (10uM) (GCGGACGTTTTTAATGTACTG) and 1 µl of water. Using a pipette tip, bacterial cells belonging to single colonies from the plates incubated overnight were transferred into the PCR mixture. The PCR cycling conditions used for DNA amplification were: 95ºC for 10 min; 30 cycles of 95ºC for 10 s, 52ºC for 20 s and 72ºC for 2mineach; followed by a final extension step at 72ºC for 5min. TRIS/Acetic acid/EDTA gel electrophoresis was performed on a 1%(w/v) agarose gel to visualise the colony PCR results. The gel was stained in a 1 µg/ml ethidium bromide solution for 1min. For plasmid amplification, single positive colonies were grown in 10 ml of liquid LB medium (Table 6.3) overnight at 37ºC at 220 rpm. Plasmid extraction was performed using QIAprep Spin Miniprep Kits (Qiagen, USA) following the manufacturer's instructions. Plasmids were sequenced (Eurofins Genomics, Germany) to ensure the accurate assembly of L0 and L1 components into L1 and L2 plasmids respectively.

6.2.8.3. Transformation of *Agrobacterium tumefaciens* to facilitate plant transformation

Chemically competent *Agrobacterium tumefaciens* AGL1 and GV3101 strains, prepared as previously detailed (Talhinhas et al., 2008), were transformed with L2 plasmids by mixing 1µg of the plasmid with 100µl of the competent cells in a 1.5 ml microfuge tube. The tubes were then frozen in liquid nitrogen for 1minand incubated at 37ºC for 5min. 500µl of LB liquid medium was added to the bacteria and they were incubated at 28ºC for 2-3 hrs at 220 rpm. The cells were spread on to LB plates with the appropriate antibiotics and were incubated for 2-3 days at 28ºC. Colony PCR to confirm the presence of the plasmid of interest in positive clones was performed as described in Section 6.2.8.2.

6.2.8.4. Transformation of *M. paleacea* through blending of thalli

M. paleacea thalli grown for 6 weeks on Gamborg's half-strength B5 medium were used for the transformation. Approximately 2g of thalli was added to 50 ml of 0M51C liquid medium (Table 6.3) and placed in a laboratory blender (Waring, USA). The thalli were blended for 30 seconds at 6000 rpm. The resultant mixture was transferred to a sterile 250 ml conical flask and incubated in a growth chamber with continuous white light (60 µmol/m^2s) at 22º C for 3 days. *A. tumefaciens* GV3101 or AGL1 containing the plasmid of interest were streaked out onto LB plates with appropriate selection and incubated for 2 days at 28ºC. A single positive colony was inoculated into 5 ml of liquid LB medium containing the appropriate antibiotics and cultured at 28ºC for 2 days at 220 rpm. 1 ml of the resultant culture was centrifuged for 15minat 2000g. The supernatant was discarded and the pellet was re-suspended in 0M51C liquid medium (Table 6.3) supplemented with 100µM acetosyringone. This was then incubated at 28ºC for 6 hours at 220 rpm.

1 ml of the above suspension was added to the *M. paleacea* blended thalli in the 250 ml conical flask after the 3-day pre-culture period. Acetosyringone was added to this mixture to a final concentration of 100µM. The flasks were then incubated under continuous white light (60 µmol/m^2s) at 22º C at 130 rpm for 2 days. The liquid from the flasks containing the thalli-agrobacteria co-culture was transferred into 50 ml plastic tubes and the liquid was discarded taking care not to lose any of the thalli

pieces. 30 ml of sterile water was added to the thalli and vortexed for 30 seconds. After allowing the thalli to settle, the supernatant was discarded taking care not to lose any of the thalli. This washing step was repeated 5 times. The thalli were then spread onto plates containing Gamborg's half-strength B5 medium with the appropriate antibiotics in addition to 100 µg/ml Cefotaxime to limit overgrowth of agrobacteria. The plants were grown under continuous light (60 µmol/m^2s) at 22ºC for 3-4 weeks at which point putative transformants could be observed. Transformation efficiency was measured using a DMR/MZFLIII microscope (Leica Microsystems, UK) to visualise the transformation marker GFP included in the L2 vectors transformed (Section 6.2.8.1).

6.2.8.5. Agartrap transformation of *M. paleacea* gemmalings

Gemmae produced from thalli as previously described (Section 6.2.1) were transferred to 60mm circular plates containing 10 ml of Gamborg's half-strength solid medium supplemented with 1% sucrose (Table 6.3). Approximately 50 gemmae were spread evenly onto each plate and incubated for 5 days under continuous light (60 µmol/m^2s) at 22ºC. *A. tumefaciens* GV3101 containing the plasmid of interest was streaked out onto LB plates with appropriate selection and incubated for 2 days at 28ºC. A single positive colony was inoculated into 5 ml of liquid LB medium containing the appropriate antibiotics and cultured at 28ºC for 2 days at 220 rpm. 1 ml of this culture was spread onto plate containing LB medium and incubated at 28ºC for 1-2 days to obtain a lawn culture. Using a toothpick, bacterial cells were then transferred into 1 ml of the transformation buffer (10mM $MgCl_2$, 10mM MES-NaOH pH 5.7, 150µM acetosyringone) to obtain a final Optical density of 0.5. 1 ml of this bacterial solution was added to the plates with pre-cultured *M. paleacea* gemmalings and left for 1min. The solution was then removed using a micropipette and the plates were incubated under continuous light (60 µmol/m^2s) at 22ºC for 3 days. Following this incubation, the surface of the media in the plates was washed with sterile water to remove any excess agrobacteria. 1 ml of selection buffer (1µl of 100 mg/ml Cefotaxime and the appropriate antibiotics) was added to each plate. The plants were incubated for 3 weeks under continuous light (60 µmol/m^2s) at 22ºC before the transformation efficiency was measured as described in Section 6.2.8.4.

6.2.9. CRISPR/Cas9-mediated mutagenesis of *M. paleacea*

The CRISPR/Cas9 system previously used for generating mutants in *M. polymorpha* was adapted for GoldenGate cloning through the in silico domestication of parts as previously described (Engler et al., 2008). These parts were then synthesised by Life Technologies, UK and cloned to make level 2 binary vectors as described in Section 6.2.8.1 which was then transformed into *M. paleacea* using the blend-based transformation methods as described in Section 6.2.8.4. Visual screening of transformants for the *nop1* air pore phenotype was performed using a MZ16 microscope (Leica Microsystems, UK). Fluorescent screening was performed using a DMR/MZFLIII microscope (Leica Microsystems, UK) to visualise the transformation marker mCherry included in the L2 vectors transformed. Screening for mutations by PCR and sequencing was performed as described in Section 6.2.8.2 using a primer pair flanking the sequences (F: GCAAGAAACCGACGAGAGGA, R: GTGCTGACAACATTGGCCAG) in the *NOP1* gene used to design the guide RNA.

7

References

References

Akiyama, K., Matsuzaki, K.-i., and Hayashi, H. (2005). Plant sesquiterpenes induce hyphal branching in arbuscular mycorrhizal fungi. Nature **435,** 824.

Alaba, S., Piszczalka, P., Pietrykowska, H., Pacak, A.M., Sierocka, I., Nuc, P.W., Singh, K., Plewka, P., Sulkowska, A., Jarmolowski, A., Karlowski, W.M., and Szweykowska-Kulinska, Z. (2015). The liverwort Pellia endiviifolia shares microtranscriptomic traits that are common to green algae and land plants. The New Phytologist **206,** 352-367.

Albalat, R., and Canestro, C. (2016). Evolution by gene loss. Nat Rev Genet **17,** 379-391.

Althoff, F., Kopischke, S., Zobell, O., Ide, K., Ishizaki, K., Kohchi, T., and Zachgo, S. (2014). Comparison of the MpEF1α and CaMV35 promoters for application in Marchantia polymorpha overexpression studies. Transgenic Research **23,** 235-244.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990). Basic Local Alignment Search Tool. J Mol Biol **215,** 403-410.

Ané, J.-M., Kiss, G.B., Riely, B.K., Penmetsa, R.V., Oldroyd, G.E.D., Ayax, C., Lévy, J., Debellé, F., Baek, J.-M., Kalo, P., Rosenberg, C., Roe, B.A., Long, S.R., Dénarié, J., and Cook, D.R. (2004). Medicago truncatula DMI1 Required for Bacterial and Fungal Symbioses in Legumes. Science **303,** 1364-1367.

Antolin-Llovera, M., Ried, M.K., and Parniske, M. (2014). Cleavage of the SYMBIOSIS RECEPTOR-LIKE KINASE ectodomain promotes complex formation with Nod factor receptor 5. Current biology : CB **24,** 422-427.

Arkin, M.N.P., Paramvir, S.D., and Adam, P. (2009). FastTree: Computing Large Minimum Evolution Trees with Profiles instead of a Distance Matrix.

Augé, R.M., Toler, H.D., and Saxton, A.M. (2015). Arbuscular mycorrhizal symbiosis alters stomatal conductance of host plants more under drought than under amply watered conditions: a meta-analysis. Mycorrhiza **25,** 13-24.

Bago, B., Pfeffer, P.E., and Shachar-Hill, Y. (2000). Carbon Metabolism and Transport in Arbuscular Mycorrhizas. Plant Physiology **124,** 949-958.

Banba, M., Gutjahr, C., Miyao, A., Hirochika, H., Paszkowski, U., Kouchi, H., and Imaizumi-Anraku, H. (2008). Divergence of Evolutionary Ways Among Common sym Genes: CASTOR and CCaMK Show Functional

Conservation Between Two Symbiosis Systems and Constitute the Root of a Common Signaling Pathway. Plant and Cell Physiology **49,** 1659-1671.

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A.A., Dvorkin, M., Kulikov, A.S., Lesin, V.M., Nikolenko, S.I., Pham, S., and Prjibelski, A.D. (2012). SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. Journal of computational biology **19,** 455-477.

Banks, J.A., Nishiyama, T., Hasebe, M., Bowman, J.L., Gribskov, M., dePamphilis, C., Albert, V.A., Aono, N., Aoyama, T., Ambrose, B.A., Ashton, N.W., Axtell, M.J., Barker, E., Barker, M.S., Bennetzen, J.L., Bonawitz, N.D., Chapple, C., Cheng, C., Correa, L.G.G., Dacre, M., DeBarry, J., Dreyer, I., Elias, M., Engstrom, E.M., Estelle, M., Feng, L., Finet, C., Floyd, S.K., Frommer, W.B., Fujita, T., Gramzow, L., Gutensohn, M., Harholt, J., Hattori, M., Heyl, A., Hirai, T., Hiwatashi, Y., Ishikawa, M., Iwata, M., Karol, K.G., Koehler, B., Kolukisaoglu, U., Kubo, M., Kurata, T., Lalonde, S., Li, K., Li, Y., Litt, A., Lyons, E., Manning, G., Maruyama, T., Michael, T.P., Mikami, K., Miyazaki, S., Morinaga, S.i., Murata, T., Mueller- Roeber, B., Nelson, D.R., Obara, M., Oguri, Y., Olmstead, R.G., Onodera, N., Petersen, B.L., Pils, B., Prigge, M., Rensing, S.A., Riaño-Pachón, D.M., Roberts, A.W., Sato, Y., Scheller, H.V., Schulz, B., Schulz, C., Shakirov, E.V., Shibagaki, N., Shinohara, N., Shippen, D.E., Sørensen, I., Sotooka, R., Sugimoto, N., Sugita, M., Sumikawa, N., Tanurdzic, M., Theißen, G., Ulvskov, P., Wakazuki, S., Weng, J.K., Willats, W.W.G.T., Wipf, D., Wolf, P.G., Yang, L., Zimmer, A.D., Zhu, Q., Mitros, T., Hellsten, U., Loqué, D., Otillar, R., Salamov, A., Schmutz, J., Shapiro, H., Lindquist, E., Lucas, S., Rokhsar, D., and Grigoriev, I.V. (2011). The Selaginella Genome Identifies Genetic Changes Associated with the Evolution of Vascular Plants. Science **332,** 960-963.

Bateman, A., Birney, E., Cerruti, L., Durbin, R., Etwiller, L., Eddy, S.R., Griffiths-Jones, S., Howe, K.L., Marshall, M., and Sonnhammer, E.L. (2002). The Pfam protein families database. Nucleic Acids Res **30**.

Baum, D.A., and Smith, S.D. (2012). Tree Thinking: An Introduction to Phylogenetic Biology. (Macmillan Learning).

Bécard, G., Douds, D., and Pfeffer, P. (1992). Extensive in vitro hyphal growth of vesicular-arbuscular mycorrhizal fungi in the presence of CO2 and flavonols. Applied and Environmental Microbiology **58,** 821-825.

Beerling, D. (2017). The emerald planet: how plants changed Earth's history. (Oxford University Press).

Belhaj, K., Chaparro-Garcia, A., Kamoun, S., Patron, N.J., and Nekrasov, V. (2015). Editing plant genomes with CRISPR/Cas9. Current opinion in biotechnology **32,** 76-84.

Bender, K.W., Blackburn, R.K., Monaghan, J., Derbyshire, P., Menke, F.L.H., Zipfel, C., Goshe, M.B., Zielinski, R.E., and Huber, S.C. (2017). Autophosphorylation-based Calcium (Ca2+) Sensitivity Priming and Ca2+/Calmodulin Inhibition of Arabidopsis thaliana Ca2+-dependent Protein Kinase 28 (CPK28). Journal of Biological Chemistry **292,** 3988-4002.

Benjamini, Y., and Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society. Series B (Methodological) **57,** 289-300.

Besserer, A., Puech-Pagès, V., Kiefer, P., Gomez-Roldan, V., Jauneau, A., Roy, S., Portais, J.-C., Roux, C., Bécard, G., and Séjalon-Delmas, N. (2006). Strigolactones stimulate arbuscular mycorrhizal fungi by activating mitochondria. PLoS biology **4,** e226.

Betts, M.J., Guigó, R., Agarwal, P., and Russell, R.B. (2001). Exon structure conservation despite low sequence similarity: a relic of dramatic events in evolution? The EMBO Journal **20,** 5354-5360.

Bidartondo, M.I., Read, D.J., Trappe, J.M., Merckx, V., Ligrone, R., and Duckett, J.G. (2011). The dawn of symbiosis between plants and fungi. Biology Letters **7,** 574-577.

Billker, O., Lourido, S., and Sibley, L.D. (2009). Calcium-dependent signaling and kinases in apicomplexan parasites. Cell Host Microbe **5,** 612-622.

Bischler-Causse, H., and Boisselier-Dubayle, M.C. (1991). Lectotypification of Marchantia polymorpha L. Journal of Bryology **16,** 361-365.

Blount, Z.D., Barrick, J.E., Davidson, C.J., and Lenski, R.E. (2012). Genomic Analysis of a Key Innovation in an Experimental E. coli Population. Nature **489,** 513-518.

Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, M.-C., Estreicher, A., Gasteiger, E., Martin, M.J., Michoud, K., O'donovan, C., and Phan, I. (2003). The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. Nucleic acids research **31,** 365-370.

Boisson-Dernier, A., Chabaud, M., Garcia, F., Bécard, G., Rosenberg, C., and Barker, D.G. (2001). Agrobacterium rhizogenes-Transformed Roots of Medicago truncatula for the Study of Nitrogen-Fixing and Endomycorrhizal Symbiotic Associations. Molecular Plant-Microbe Interactions **14,** 695-700.

Bolle, C. (2004). The role of GRAS proteins in plant signal transduction and development. Planta **218,** 683-692.

Borovichev, E.A., and Bakalin, V.A. (2014). A survey of Marchantiales from the Russian Far East II. Note on Marchantia paleacea Bertol. Arctoa **23,** 25-28.

Bravo, A., York, T., Pumplin, N., Mueller, L.A., and Harrison, M.J. (2016). Genes conserved for arbuscular mycorrhizal symbiosis identified through phylogenomics **2,** 15208.

Bravo, A., Brands, M., Wewer, V., Dörmann, P., and Harrison, M.J. (2017). Arbuscular mycorrhiza-specific enzymes FatM and RAM2 fine-tune lipid biosynthesis to promote development of arbuscular mycorrhiza. New Phytologist **214,** 1631-1645.

Breuninger, H., Sakayama, H., Nishiyama, T., and Dolan, L. (2016). Diversification of a bHLH transcription factor family in streptophytes led to the evolution of antagonistically acting genes controlling root hair growth. Curr. Biol **26,** 1622-1628.

Brodie, J., and Lewis, J. (2007). Unravelling the algae: the past, present, and future of algal systematics. (CRC Press).

Brundrett, M.C. (2002). Coevolution of roots and mycorrhizas of land plants. New Phytol **154**.

Brundrett, M.C. (2009). Mycorrhizal associations and other means of nutrition of vascular plants: understanding the global diversity of host plants by resolving conflicting information and developing reliable means of diagnosis. Plant and Soil **320,** 37-77.

Butterfield, N., Knoll, A., and Swett, K. (1994). Paleobiology of the Neoproterozoic Svanbergfjellet formation, Spitsbergen. Fossils Strata.

Butterfield, N.J. (2009). Modes of pre-Ediacaran multicellularity. Precambrian Research **173,** 201-211.

Campbell, M.S., Law, M., Holt, C., Stein, J.C., Moghe, G.D., Hufnagel, D.E., Lei, J., Achawanantakun, R., Jiao, D., and Lawrence, C.J. (2014). MAKER-P: a tool kit for the rapid creation, management, and quality control of plant genome annotations. Plant physiology **164,** 513-524.

Cantino, P.D., Doyle, J.A., Graham, S.W., Judd, W.S., Olmstead, R.G., Soltis, D.E., Soltis, P.S., and Donoghue, M.J. (2007). Towards a phylogenetic nomenclature of Tracheophyta. Taxon **56**.

Capoen, W., Sun, J., Wysham, D., Otegui, M.S., Venkateshwaran, M., Hirsch, S., Miwa, H., Downie, J.A., Morris, R.J., and Ané, J.-M. (2011). Nuclear membranes control symbiotic calcium signaling of legumes. Proceedings of the National Academy of Sciences **108,** 14348-14353.

Carotenuto, G., Chabaud, M., Miyata, K., Capozzi, M., Takeda, N., Kaku, H., Shibuya, N., Nakagawa, T., Barker, D.G., and Genre, A. (2017). The rice LysM receptor-like kinase OsCERK1 is required for the perception of short-chain chitin oligomers in arbuscular mycorrhizal signaling. New Phytologist **214,** 1440-1446.

Carvalho-Santos, Z., Machado, P., Branco, P., Tavares-Cadete, F., Rodrigues-Martins, A., Pereira-Leal, J.B., and Bettencourt-Dias, M. (2010). Stepwise evolution of the centriole-assembly pathway. J Cell Sci **123,** 1414-1426.

Catoira, R., Galera, C., de Billy, F., Penmetsa, R.V., Journet, E.P., Maillet, F., Rosenberg, C., Cook, D., Gough, C., and D\'nari. (2000). Four genes of Medicago truncatula controlling components of a nod factor transduction pathway. The Plant cell **12,** 1647--1666.

Chapman, J.A., Ho, I., Sunkara, S., Luo, S., Schroth, G.P., and Rokhsar, D.S. (2011). Meraculous: de novo genome assembly with short paired-end reads. PloS one **6,** e23501.

Chapple, C.E., and Guigó, R. (2008). Relaxation of Selective Constraints Causes Independent Selenoprotein Extinction in Insect Genomes. PLOS ONE **3,** e2968.

Charpentier, M., Vaz Martins, T., Granqvist, E., Oldroyd, G.E.D., and Morris, R.J. (2013). The role of DMI1 in establishing Ca(2+) oscillations in legume symbioses. Plant Signaling & Behavior **8,** e22894.

Charpentier, M., Bredemeier, R., Wanner, G., Takeda, N., Schleiff, E., and Parniske, M. (2008). Lotus japonicus CASTOR and POLLUX are ion channels essential for perinuclear calcium spiking in legume root endosymbiosis. The Plant Cell **20,** 3467-3479.

Charpentier, M., Sun, J., Vaz Martins, T., Radhakrishnan, G.V., Findlay, K., Soumpourou, E., Thouin, J., Very, A.A., Sanders, D., Morris, R.J., and

Oldroyd, G.E. (2016). Nuclear-localized cyclic nucleotide-gated channels mediate symbiotic calcium oscillations. Science **352,** 1102-1105.

Chen, C., Ane, J.M., and Zhu, H. (2008). OsIPD3, an ortholog of the Medicago truncatula DMI3 interacting protein IPD3, is required for mycorrhizal symbiosis in rice. New Phytol **180,** 311-315.

Chen, C., Gao, M., Liu, J., and Zhu, H. (2007). Fungal Symbiosis in Rice Requires an Ortholog of a Legume Common Symbiosis Gene Encoding a Ca²⁺/Calmodulin-Dependent Protein Kinase. Plant Physiology **145,** 1619-1628.

Chen, C., Fan, C., Gao, M., and Zhu, H. (2009). Antiquity and function of CASTOR and POLLUX, the twin ion channel-encoding genes key to the evolution of root symbioses in plants. Plant Physiology **149,** 306-317.

Chen, F., Zhang, L., and Cheng, Z.-M. (2017). The calmodulin fused kinase novel gene family is the major system in plants converting Ca2+ signals to protein phosphorylation responses. Sci Rep-Uk **7,** 4127.

Chikhi, R., and Rizk, G. (2012). Space-efficient and exact de Bruijn graph representation based on a Bloom filter. In International Workshop on Algorithms in Bioinformatics (Springer), pp. 236-248.

Chikhi, R., and Medvedev, P. (2014). Informed and automated k-mer size selection for genome assembly. Bioinformatics **30,** 31-37.

Darling, A.C., Mau, B., Blattner, F.R., and Perna, N.T. (2004). Mauve: multiple alignment of conserved genomic sequence with rearrangements. Genome research **14,** 1394-1403.

Darwin, C. (1869). On the Origin of Species by Means of Natural Selection: Or the Preservation of Favoured Races in the Struggle for Life. (D. Appleton).

Davey, M.L., and Currah, R.S. (2006). Interactions between mosses (Bryophyta) and fungi. Canadian Journal of Botany **84,** 1509-1519.

Delaux, P.-M., Bécard, G., and Combier, J.-P. (2013a). NSP1 is a component of the Myc signaling pathway. New Phytologist **199,** 59-65.

Delaux, P.-M., Varala, K., Edger, P.P., Coruzzi, G.M., Pires, J.C., and Ané, J.-M. (2014). Comparative Phylogenomics Uncovers the Impact of Symbiotic Associations on Host Genome Evolution. PLOS Genetics **10,** e1004487.

Delaux, P.-M., Radhakrishnan, G.V., Jayaraman, D., Cheema, J., Malbreil, M., Volkening, J.D., Sekimoto, H., Nishiyama, T., Melkonian, M., Pokorny, L., Rothfels, C.J., Sederoff, H.W., Stevenson, D.W., Surek, B., Zhang, Y.,

Sussman, M.R., Dunand, C., Morris, R.J., Roux, C., Wong, G.K.-S., Oldroyd, G.E.D., and Ané, J.-M. (2015). Algal ancestor of land plants was preadapted for symbiosis. Proceedings of the National Academy of Sciences **112,** 13390-13395.

Delaux, P.M., Sejalon-Delmas, N., Bécard, G., and Ané, J.M. (2013b). Evolution of the plant-microbe symbiotic 'toolkit'. Trends Plant Sci **18**.

Delmans, M., Pollak, B., and Haseloff, J. (2017). MarpoDB: An Open Registry for Marchantia Polymorpha Genetic Parts. Plant and Cell Physiology **58,** e5-e5.

Delwiche, C., and Cooper, E. (2015). The Evolutionary Origin of a Terrestrial Flora. Current Biology **25**.

Denton, J.F., Lugo-Martinez, J., Tucker, A.E., Schrider, D.R., Warren, W.C., and Hahn, M.W. (2014). Extensive Error in the Number of Genes Inferred from Draft Genome Assemblies. PLoS Computational Biology **10,** e1003998.

Desirò, A., Duckett, J.G., Pressel, S., Villarreal, J.C., and Bidartondo, M.I. (2013). Fungal symbioses in hornworts: a chequered history. Proceedings of the Royal Society B: Biological Sciences **280**.

Dickson, S. (2004). The Arum–Paris continuum of mycorrhizal symbioses. New Phytologist **163,** 187-200.

Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R. (2013). STAR: ultrafast universal RNA-seq aligner. Bioinformatics **29,** 15-21.

Domozych, D., Sørensen, I., and Popper, Z.A. (2017). Editorial: Charophytes: Evolutionary Ancestors of Plants and Emerging Models for Plant Research. Frontiers in Plant Science **8**.

Doolittle, R.F. (2009). Step-by-step evolution of vertebrate blood coagulation. Cold Spring Harbor symposia on quantitative biology **74,** 35-40.

Easlon, H.M., and Bloom, A.J. (2014). Easy Leaf Area: Automated Digital Image Analysis for Rapid and Accurate Measurement of Leaf Area. Applications in Plant Sciences **2,** 1400033.

Edger, P.P., and Pires, J.C. (2009). Gene and genome duplications: the impact of dosage-sensitivity on the fate of nuclear genes. Chromosome Research **17,** 699.

Edwards, D., Morris, J.L., Richardson, J.B., and Kenrick, P. (2014). Cryptospores and cryptophytes reveal hidden diversity in early land floras. New Phytologist **202,** 50-78.

Edwards, D.S. (1986). Aglaophyton major, a non- vascular land- plant from the Devonian Rhynie Chert. Botanical Journal of the Linnean Society **93,** 173-204.

Emms, D.M., and Kelly, S. (2015). OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. Genome Biology **16,** 157.

Endre, G., Kereszt, A., Kevei, Z., Mihacea, S., Kalo, P., and Kiss, G.B. (2002). A receptor kinase gene regulating symbiotic nodule development. Nature **417,** 962-966.

Engler, C., and Marillonnet, S. (2013). Combinatorial DNA Assembly Using Golden Gate Cloning. In Synthetic Biology, K.M. Polizzi and C. Kontoravdi, eds (Totowa, NJ: Humana Press), pp. 141-156.

Engler, C., Kandzia, R., and Marillonnet, S. (2008). A one pot, one step, precision cloning method with high throughput capability. PloS one **3,** e3647.

Falkowski, P., and Knoll, A.H. (2011). Evolution of Primary Producers in the Sea. (Elsevier Science).

Favre, P., Bapaume, L., Bossolini, E., Delorenzi, M., Falquet, L., and Reinhardt, D. (2014). A novel bioinformatics pipeline to discover genes related to arbuscular mycorrhizal symbiosis based on their evolutionary conservation pattern among higher plants. BMC plant biology **14,** 333.

Feddermann, N., Muni, R.R., Zeier, T., Stuurman, J., Ercolin, F., Schorderet, M., and Reinhardt, D. (2010). The PAM1 gene of petunia, required for intracellular accommodation and morphogenesis of arbuscular mycorrhizal fungi, encodes a homologue of VAPYRIN. The Plant journal : for cell and molecular biology **64,** 470-481.

Felsenstein, J. (2003). Inferring Phylogenies. (Sinauer).

Field, K.J., Pressel, S., Duckett, J.G., Rimington, W.R., and Bidartondo, M.I. (2015a). Symbiotic options for the conquest of land. Trends in ecology & evolution **30,** 477-486.

Field, K.J., Cameron, D.D., Leake, J.R., Tille, S., Bidartondo, M.I., and Beerling, D.J. (2012). Contrasting arbuscular mycorrhizal responses of vascular and non-vascular plants to a simulated Palaeozoic CO2 decline. Nature Communications **3**.

Field, K.J., Leake, J.R., Tille, S., Allinson, K.E., Rimington, W.R., Bidartondo, M.I., Beerling, D.J., and Cameron, D.D. (2015b). From mycoheterotrophy

to mutualism: mycorrhizal specificity and functioning in Ophioglossum vulgatum sporophytes. New Phytologist **205,** 1492-1502.

Field, K.J., Rimington, W.R., Bidartondo, M.I., Allinson, K.E., Beerling, D.J., Cameron, D.D., Duckett, J.G., Leake, J.R., and Pressel, S. (2015c). First evidence of mutualism between ancient plant lineages (Haplomitriopsida liverworts) and Mucoromycotina fungi and its response to simulated Palaeozoic changes in atmospheric CO2. New Phytologist **205,** 743-756.

Field, K.J., Rimington, W.R., Bidartondo, M.I., Allinson, K.E., Beerling, D.J., Cameron, D.D., Duckett, J.G., Leake, J.R., and Pressel, S. (2016). Functional analysis of liverworts in dual symbiosis with Glomeromycota and Mucoromycotina fungi under a simulated Palaeozoic CO2 decline. ISME J **10,** 1514-1526.

Finlay, R.D. (2008). Ecological aspects of mycorrhizal symbiosis: with special emphasis on the functional diversity of interactions involving the extraradical mycelium. Journal of experimental botany **59,** 1115-1126.

Finn, R.D., Clements, J., Arndt, W., Miller, B.L., Wheeler, T.J., Schreiber, F., Bateman, A., and Eddy, S.R. (2015). HMMER web server: 2015 update. Nucleic Acids Res **43,** W30-38.

Fiorilli, V., Vallino, M., Biselli, C., Faccio, A., Bagnaresi, P., and Bonfante, P. (2015). Host and non-host roots in rice: cellular and molecular approaches reveal differential responses to arbuscular mycorrhizal fungi. Frontiers in Plant Science **6,** 636.

Fonseca, H.M.A.C., Ferreira, J.I.L., Berbara, R.L.L., and Zatorre, N.P. (2009). Dominance of Paris-Type Morphology on Mycothallus of Lunularia Cruciata Colonised by Glomus Proliferum. Braz J Microbiol **40,** 96-101.

Forest, F. (2009). Calibrating the Tree of Life: fossils, molecules and evolutionary timescales. Annals of Botany **104,** 789-794.

Fortin, J.A., Declerck, S., and Strullu, D.-G. (2005). In Vitro Culture of Mycorrhizas. In In Vitro Culture of Mycorrhizas, S. Declerck, J.A. Fortin, and D.-G. Strullu, eds (Berlin, Heidelberg: Springer Berlin Heidelberg), pp. 3-14.

Friedman, R., and Hughes, A.L. (2001). Pattern and timing of gene duplication in animal genomes. Genome research **11,** 1842-1847.

Friedman, W.E., Moore, R.C., and Purugganan, M.D. (2004). The evolution of plant development. American Journal of Botany **91,** 1726-1741.

Furumizu, C., Alvarez, J.P., Sakakibara, K., and Bowman, J.L. (2015). Antagonistic Roles for KNOX1 and KNOX2 Genes in Patterning the Land Plant Body Plan Following an Ancient Gene Duplication. PLOS Genetics **11,** e1004980.

Gamborg, O.L., Miller, R.A., and Ojima, K. (1968). Nutrient requirements of suspension cultures of soybean root cells. Experimental Cell Research **50,** 151-158.

Genre, A., and Bonfante, P. (2010). The Making of Symbiotic Cells in Arbuscular Mycorrhizal Roots. In Arbuscular Mycorrhizas: Physiology and Function, H. Koltai and Y. Kapulnik, eds (Dordrecht: Springer Netherlands), pp. 57-71.

Genre, A., Chabaud, M., Timmers, T., Bonfante, P., and Barker, D.G. (2005). Arbuscular mycorrhizal fungi elicit a novel intracellular apparatus in Medicago truncatula root epidermal cells before infection. The Plant Cell **17,** 3489-3499.

Genre, A., Chabaud, M., Balzergue, C., Puech- Pagès, V., Novero, M., Rey, T., Fournier, J., Rochange, S., Bécard, G., and Bonfante, P. (2013). Short-chain chitin oligomers from arbuscular mycorrhizal fungi trigger nuclear Ca2+ spiking in Medicago truncatula roots and their production is enhanced by strigolactone. New Phytologist **198,** 190-202.

Gensel, P.G., and Edwards, D. (2001). Plants invade the land: evolutionary and environmental perspectives. (Columbia University Press).

Gobbato, E., Marsh, J.F., Vernie, T., Wang, E., Maillet, F., Kim, J., Miller, J.B., Sun, J., Bano, S.A., Ratet, P., Mysore, K.S., Dénarié, J., Schultze, M., and Oldroyd, G.E. (2012). A GRAS-type transcription factor with a specific function in mycorrhizal signaling. Current biology : CB **22**.

Godfroy, O., Debelle, F., Timmers, T., and Rosenberg, C. (2006). A rice calcium- and calmodulin-dependent protein kinase restores nodulation to a legume mutant. Molecular plant-microbe interactions : MPMI **19,** 495-501.

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., Friedman, N., and Regev, A. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nature biotechnology **29,** 644-652.

Graham, L.E. (1985). The Origin of the Life Cycle of Land Plants: A simple modification in the life cycle of an extinct green alga is the likely origin of the first land plants. American Scientist **73,** 178-186.

Graham, L.K., and 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 B: Biological Sciences **355,** 757-767.

Granqvist, E., Wysham, D., Hazledine, S., Kozlowski, W., Sun, J., Charpentier, M., Martins, T.V., Haleux, P., Tsaneva-Atanasova, K., Downie, J.A., Oldroyd, G.E.D., and Morris, R.J. (2012). Buffering Capacity Explains Signal Variation in Symbiotic Calcium Oscillations. Plant Physiology **160,** 2300-2310.

Groth, M., Takeda, N., Perry, J., Uchida, H., Draexl, S., Brachmann, A., Sato, S., Tabata, S., Kawaguchi, M., Wang, T.L., and Parniske, M. (2010). NENA, a Lotus japonicus homolog of Sec13, is required for rhizodermal infection by arbuscular mycorrhiza fungi and rhizobia but dispensable for cortical endosymbiotic development. Plant Cell **22**.

Gutjahr, C., Banba, M., Croset, V., An, K., Miyao, A., An, G., Hirochika, H., Imaizumi-Anraku, H., and Paszkowski, U. (2008). Arbuscular Mycorrhiza–Specific Signaling in Rice Transcends the Common Symbiosis Signaling Pathway. The Plant Cell **20,** 2989-3005.

Gutjahr, C., Radovanovic, D., Geoffroy, J., Zhang, Q., Siegler, H., Chiapello, M., Casieri, L., An, K., An, G., Guiderdoni, E., Kumar, C.S., Sundaresan, V., Harrison, M.J., and Paszkowski, U. (2012). The half-size ABC transporters STR1 and STR2 are indispensable for mycorrhizal arbuscule formation in rice. The Plant journal : for cell and molecular biology **69,** 906-920.

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., Henschel, R., LeDuc, R.D., Friedman, N., and Regev, A. (2013). De novo transcript sequence reconstruction from RNA-Seq: reference generation and analysis with Trinity. Nature protocols **8,** 10.1038/nprot.2013.1084.

Hagelberg, E., Hofreiter, M., and Keyser, C. (2015). Ancient DNA: the first three decades. Philosophical Transactions of the Royal Society B: Biological Sciences **370,** 20130371.

Haig, D. (2010). What do we know about Charophyte (Streptophyta) life cycles? Journal of phycology **46,** 860-867.

Hall, J.D., Karol, K.G., McCourt, R.M., and Delwiche, C.F. (2008). Phylogeny of the conjugating green algae based on chloroplast and mitochondrial nucleotide sequence data. Journal of Phycology **44,** 467-477.

Halverson, G.P., Maloof, A.C., Schrag, D.P., Dudás, F.Ö., and Hurtgen, M. (2007). Stratigraphy and geochemistry of a ca 800 Ma negative carbon isotope interval in northeastern Svalbard. Chemical Geology **237,** 5-27.

Handberg, K., and Stougaard, J. (1992). Lotus Japonicus, an Autogamous, Diploid Legume Species for Classical and Molecular-Genetics. The Plant Journal **2,** 487-496.

Harlin-Cognato, A., Hoffman, E.A., and Jones, A.G. (2006). Gene cooption without duplication during the evolution of a male-pregnancy gene in pipefish. Proceedings of the National Academy of Sciences of the United States of America **103,** 19407-19412.

Harmon, A.C., Gribskov, M., and Harper, J.F. (2000). CDPKs - a kinase for every Ca2+ signal? Trends Plant Sci **5,** 154-159.

Harper, J.F., and Harmon, A. (2005). Plants, symbiosis and parasites: a calcium signalling connection. Nat Rev Mol Cell Biol **6,** 555-566.

Harrison, M.J. (1996). A sugar transporter from Medicago truncatula: altered expression pattern in roots during vesicular-arbuscular (VA) mycorrhizal associations. The Plant journal : for cell and molecular biology **9,** 491-503.

Harrison, M.J., Dewbre, G.R., and Liu, J. (2002). A Phosphate Transporter from Medicago truncatula Involved in the Acquisition of Phosphate Released by Arbuscular Mycorrhizal Fungi. The Plant Cell **14,** 2413-2429.

Hayashi, T., Banba, M., Shimoda, Y., Kouchi, H., Hayashi, M., and Imaizumi-Anraku, H. (2010). A dominant function of CCaMK in intracellular accommodation of bacterial and fungal endosymbionts. The Plant journal : for cell and molecular biology **63,** 141-154.

Herron, M.D., Hackett, J.D., Aylward, F.O., and Michod, R.E. (2009). Triassic origin and early radiation of multicellular volvocine algae. Proceedings of the National Academy of Sciences **106,** 3254-3258.

Hillis, D.M. (1998). Taxonomic sampling, phylogenetic accuracy, and investigator bias. Syst Biol **47**.

Hillis, D.M., Pollock, D.D., McGuire, J.A., and Zwickl, D.J. (2003). Is sparse taxon sampling a problem for phylogenetic inference? Syst Biol **52**.

Hirsch, S., and Oldroyd, G.E.D. (2009). GRAS-domain transcription factors that regulate plant development. Plant Signaling & Behavior **4,** 698-700.

Holder, M., and Lewis, P.O. (2003). Phylogeny estimation: traditional and Bayesian approaches. Nat Rev Genet **4,** 275-284.

Honkanen, S., Jones, Victor A., Morieri, G., Champion, C., Hetherington, Alexander J., Kelly, S., Proust, H., Saint-Marcoux, D., Prescott, H., and Dolan, L. (2016). The Mechanism Forming the Cell Surface of Tip-Growing Rooting Cells Is Conserved among Land Plants. Current Biology **26,** 3238-3244.

Hori, K., Maruyama, F., Fujisawa, T., Togashi, T., Yamamoto, N., Seo, M., Sato, S., Yamada, T., Mori, H., and Tajima, N. (2014). Klebsormidium flaccidum genome reveals primary factors for plant terrestrial adaptation. Nature communications **5**.

Horvath, B., Yeun, L.H., Domonkos, A., Halasz, G., Gobbato, E., Ayaydin, F., Miro, K., Hirsch, S., Sun, J.H., Tadege, M., Ratet, P., Mysore, K.S., Ané, J.M., Oldroyd, G.E., and Kaló, P. (2011). Medicago truncatulaIPD3 Is a member of the common symbiotic signaling pathway required for rhizobial and mycorrhizal symbioses. Mol Plant-Microbe Interact **24**.

Hrabak, E.M., Chan, C.W., Gribskov, M., Harper, J.F., Choi, J.H., Halford, N., Kudla, J., Luan, S., Nimmo, H.G., Sussman, M.R., Thomas, M., Walker-Simmons, K., Zhu, J.K., and Harmon, A.C. (2003). The Arabidopsis CDPK-SnRK superfamily of protein kinases. Plant Physiol **132,** 666-680.

Huguet, T., Duc, G., Sagan, M., Olivieri, I., and Prosperi, J.M. (1995). The legume Medicago truncatula as a model plant. Colloq Inra**,** 223-228.

Humphreys, C.P., Franks, P.J., Rees, M., Bidartondo, M.I., Leake, J.R., and Beerling, D.J. (2010). Mutualistic mycorrhiza-like symbiosis in the most ancient group of land plants. Nature Communications **1**.

Imaizumi-Anraku, H., Takeda, N., Charpentier, M., Perry, J., Miwa, H., Umehara, Y., Kouchi, H., Murakami, Y., Mulder, L., Vickers, K., Pike, J., Downie, J.A., Wang, T., Sato, S., Asamizu, E., Tabata, S., Yoshikawa, M., Murooka, Y., Wu, G.J., Kawaguchi, M., Kawasaki, S., Parniske, M.,

and Hayashi, M. (2005). Plastid proteins crucial for symbiotic fungal and bacterial entry into plant roots. Nature **433,** 527-531.

Imhof, S. (1999). Subterranean structures and mycorrhiza of the achlorophyllous Burmannia tenella (Burmanniaceae). Canadian Journal of Botany-Revue Canadienne De Botanique **77,** 637-643.

Ishizaki, K., Chiyoda, S., Yamato, K.T., and Kohchi, T. (2008). Agrobacterium-mediated transformation of the haploid liverwort Marchantia polymorpha L., an emerging model for plant biology. Plant and Cell Physiology **49,** 1084-1091.

Ishizaki, K., Nishihama, R., Yamato, K.T., and Kohchi, T. (2016). Molecular Genetic Tools and Techniques for Marchantia polymorpha Research. Plant and Cell Physiology **57,** 262-270.

Ishizaki, K., Johzuka-Hisatomi, Y., Ishida, S., Iida, S., and Kohchi, T. (2013a). Homologous recombination-mediated gene targeting in the liverwort Marchantia polymorpha L. Sci Rep-Uk **3**.

Ishizaki, K., Mizutani, M., Shimamura, M., Masuda, A., Nishihama, R., and Kohchi, T. (2013b). Essential Role of the E3 Ubiquitin Ligase NOPPERABO1 in Schizogenous Intercellular Space Formation in the Liverwort Marchantia polymorpha. The Plant Cell Online **25,** 4075-4084.

Jackson, S.B.C., Gregory, D.M., and Scott, A. (2009). Three Sequenced Legume Genomes and Many Crop Species: Rich Opportunities for Translational Genomics.

Jang, G., Yi, K., Pires, N.D., Menand, B., and Dolan, L. (2011). RSL genes are sufficient for rhizoid system development in early diverging land plants. Development **138,** 2273-2281.

Janouškovec, J., Gavelis, G.S., Burki, F., Dinh, D., Bachvaroff, T.R., Gornik, S.G., Bright, K.J., Imanian, B., Strom, S.L., Delwiche, C.F., Waller, R.F., Fensome, R.A., Leander, B.S., Rohwer, F.L., and Saldarriaga, J.F. (2017). Major transitions in dinoflagellate evolution unveiled by phylotranscriptomics. Proceedings of the National Academy of Sciences **114,** E171-E180.

Jiang, Y., Wang, W., Xie, Q., Liu, N., Liu, L., Wang, D., Zhang, X., Yang, C., Chen, X., Tang, D., and Wang, E. (2017). Plants transfer lipids to sustain colonization by mutualistic mycorrhizal and parasitic fungi. Science **356,** 1172-1175.

Jones, V.A.S., and Dolan, L. (2012). The evolution of root hairs and rhizoids. Annals of Botany **110,** 205-212.

Kajitani, R., Toshimoto, K., Noguchi, H., Toyoda, A., Ogura, Y., Okuno, M., Yabana, M., Harada, M., Nagayasu, E., and Maruyama, H. (2014). Efficient de novo assembly of highly heterozygous genomes from whole-genome shotgun short reads. Genome research **24,** 1384-1395.

Kalo, P., Gleason, C., Edwards, A., Marsh, J., Mitra, R.M., Hirsch, S., Jakab, J., Sims, S., Long, S.R., Rogers, J., Kiss, G.B., Downie, J.A., and Oldroyd, G.E. (2005). Nodulation signaling in legumes requires NSP2, a member of the GRAS family of transcriptional regulators. Science **308,** 1786-1789.

Kanamori, N., Madsen, L.H., Radutoiu, S., Frantescu, M., Quistgaard, E.M., Miwa, H., Downie, J.A., James, E.K., Felle, H.H., Haaning, L.L., Jensen, T.H., Sato, S., Nakamura, Y., Tabata, S., Sandal, N., and Stougaard, J. (2006). A nucleoporin is required for induction of Ca2+ spiking in legume nodule development and essential for rhizobial and fungal symbiosis. Proc Natl Acad Sci U S A **103,** 359-364.

Karol, K.G., McCourt, R.M., Cimino, M.T., and Delwiche, C.F. (2001). The closest living relatives of land plants. Science **294**.

Kato, H., Ishizaki, K., Kouno, M., Shirakawa, M., Bowman, J.L., Nishihama, R., and Kohchi, T. (2015). Auxin-Mediated Transcriptional System with a Minimal Set of Components Is Critical for Morphogenesis through the Life Cycle in Marchantia polymorpha. PLOS Genetics **11,** e1005084.

Katoh, K., Misawa, K., Kuma, K.Ä., and Miyata, T. (2002). MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res **30**.

Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., Buxton, S., Cooper, A., Markowitz, S., Duran, C., Thierer, T., Ashton, B., Meintjes, P., and Drummond, A. (2012). Geneious Basic: An integrated and extendable desktop software platform for the organization and analysis of sequence data. Bioinformatics **28,** 1647-1649.

Keeling, P.J. (2010). The endosymbiotic origin, diversification and fate of plastids. Philosophical Transactions of the Royal Society B: Biological Sciences **365,** 729-748.

Keeling, P.J., Burki, F., Wilcox, H.M., Allam, B., Allen, E.E., Amaral-Zettler, L.A., Armbrust, E.V., Archibald, J.M., Bharti, A.K., Bell, C.J., Beszteri,

B., Bidle, K.D., Cameron, C.T., Campbell, L., Caron, D.A., Cattolico, R.A., Collier, J.L., Coyne, K., Davy, S.K., Deschamps, P., Dyhrman, S.T., Edvardsen, B., Gates, R.D., Gobler, C.J., Greenwood, S.J., Guida, S.M., Jacobi, J.L., Jakobsen, K.S., James, E.R., Jenkins, B., John, U., Johnson, M.D., Juhl, A.R., Kamp, A., Katz, L.A., Kiene, R., Kudryavtsev, A., Leander, B.S., Lin, S., Lovejoy, C., Lynn, D., Marchetti, A., McManus, G., Nedelcu, A.M., Menden-Deuer, S., Miceli, C., Mock, T., Montresor, M., Moran, M.A., Murray, S., Nadathur, G., Nagai, S., Ngam, P.B., Palenik, B., Pawlowski, J., Petroni, G., Piganeau, G., Posewitz, M.C., Rengefors, K., Romano, G., Rumpho, M.E., Rynearson, T., Schilling, K.B., Schroeder, D.C., Simpson, A.G.B., Slamovits, C.H., Smith, D.R., Smith, G.J., Smith, S.R., Sosik, H.M., Stief, P., Theriot, E., Twary, S.N., Umale, P.E., Vaulot, D., Wawrik, B., Wheeler, G.L., Wilson, W.H., Xu, Y., Zingone, A., and Worden, A.Z. (2014). The Marine Microbial Eukaryote Transcriptome Sequencing Project (MMETSP): Illuminating the Functional Diversity of Eukaryotic Life in the Oceans through Transcriptome Sequencing. PLOS Biology **12,** e1001889.

Kenrick, P., and Crane, P.R. (1997). The origin and early evolution of plants on land. Nature **389**.

Kenrick, P., and Strullu-Derrien, C. (2014). The Origin and Early Evolution of Roots. Plant Physiology **166,** 570-580.

Kenrick, P., Wellman, C.H., Schneider, H., and Edgecombe, G.D. (2012). A timeline for terrestrialization: consequences for the carbon cycle in the Palaeozoic. Phil. Trans. R. Soc. B **367,** 519-536.

Keymer, A., Pimprikar, P., Wewer, V., Huber, C., Brands, M., Bucerius, S.L., Delaux, P.-M., Klingl, V., Röpenack-Lahaye, E.v., Wang, T.L., Eisenreich, W., Dörmann, P., Parniske, M., and Gutjahr, C. (2017). Lipid transfer from plants to arbuscular mycorrhiza fungi. eLife **6,** e29107.

Kistner, C., Winzer, T., Pitzschke, A., Mulder, L., Sato, S., Kaneko, T., Tabata, S., Sandal, N., Stougaard, J., Webb, K.J., Szczyglowski, K., and Parniske, M. (2005). Seven *Lotus japonicus* Genes Required for Transcriptional Reprogramming of the Root during Fungal and Bacterial Symbiosis. The Plant Cell **17,** 2217-2229.

Koonin, E.V. (2005). Orthologs, Paralogs, and Evolutionary Genomics. Annual Review of Genetics **39,** 309-338.

Kopischke, S., Schüßler, E., Althoff, F., and Zachgo, S. (2017). TALEN-mediated genome-editing approaches in the liverwort Marchantia polymorpha yield high efficiencies for targeted mutagenesis. Plant Methods **13,** 20.

Kosuta, S., Hazledine, S., Sun, J., Miwa, H., Morris, R.J., Downie, J.A., and Oldroyd, G.E.D. (2008). Differential and chaotic calcium signatures in the symbiosis signaling pathway of legumes. Proceedings of the National Academy of Sciences of the United States of America **105,** 9823-9828.

Kubota, A., Ishizaki, K., Hosaka, M., and Kohchi, T. (2013). Efficient Agrobacterium-Mediated Transformation of the Liverwort Marchantia polymorpha Using Regenerating Thalli. Biosci Biotech Bioch **77,** 167-172.

Lee, J.-H., Lin, H., Joo, S., and Goodenough, U. (2008). Early Sexual Origins of Homeoprotein Heterodimerization and Evolution of the Plant KNOX/BELL Family. Cell **133,** 829-840.

Lehnert, M., Krug, M., and Kessler, M. (2017). A review of symbiotic fungal endophytes in lycophytes and ferns – a global phylogenetic and ecological perspective. Symbiosis **71,** 77-89.

Leliaert, F., Verbruggen, H., and Zechman, F.W. (2011). Into the deep: New discoveries at the base of the green plant phylogeny. Bioessays **33**.

Lenton, T.M., Crouch, M., Johnson, M., Pires, N., and Dolan, L. (2012). First plants cooled the Ordovician. Nature Geosci **5,** 86-89.

Lévy, J., Bres, C., Geurts, R., Chalhoub, B., Kulikova, O., Duc, G., Journet, E.P., Ané, J.M., Lauber, E., Bisseling, T., Dénarié, J., Rosenberg, C., and Debellé, F. (2004). A putative Ca2+ and calmodulin-dependent protein kinase required for bacterial and fungal symbioses. Science **303**.

Lewis, L.A., and McCourt, R.M. (2004). Green algae and the origin of land plants. Amer J Bot **91**.

Li, D., Liu, C.-M., Luo, R., Sadakane, K., and Lam, T.-W. (2015). MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. Bioinformatics **31,** 1674-1676.

Li, Y., Beisson, F., Koo, A.J.K., Molina, I., Pollard, M., and Ohlrogge, J. (2007). Identification of acyltransferases required for cutin biosynthesis and production of cutin with suberin-like monomers. Proceedings of the National Academy of Sciences **104,** 18339-18344.

Ligrone, R., Carafa, A., Lumini, E., Bianciotto, V., Bonfante, P., and Duckett, J.G. (2007). Glomeromycotean associations in liverworts: a molecular, cellular, and taxonomic analysis. American Journal of Botany **94,** 1756-1777.

Lin, K., Limpens, E., Zhang, Z., Ivanov, S., Saunders, D.G.O., Mu, D., Pang, E., Cao, H., Cha, H., Lin, T., Zhou, Q., Shang, Y., Li, Y., Sharma, T., van Velzen, R., de Ruijter, N., Aanen, D.K., Win, J., Kamoun, S., Bisseling, T., Geurts, R., and Huang, S. (2014). Single Nucleus Genome Sequencing Reveals High Similarity among Nuclei of an Endomycorrhizal Fungus. PLOS Genetics **10,** e1004078.

Linkies, A., Graeber, K., Knight, C., and Leubner- Metzger, G. (2010). The evolution of seeds. New Phytologist **186,** 817-831.

Liu, J., Maldonado-Mendoza, I., Lopez-Meyer, M., Cheung, F., Town, C.D., and Harrison, M.J. (2007). Arbuscular mycorrhizal symbiosis is accompanied by local and systemic alterations in gene expression and an increase in disease resistance in the shoots. The Plant Journal **50,** 529-544.

Liu, K., Linder, C.R., and Warnow, T. (2011a). RAxML and FastTree: Comparing Two Methods for Large-Scale Maximum Likelihood Phylogeny Estimation. PLOS ONE **6,** e27731.

Liu, W., Kohlen, W., Lillo, A., Op den Camp, R., Ivanov, S., Hartog, M., Limpens, E., Jamil, M., Smaczniak, C., Kaufmann, K., Yang, W.C., Hooiveld, G.J., Charnikhova, T., Bouwmeester, H.J., Bisseling, T., and Geurts, R. (2011b). Strigolactone biosynthesis in Medicago truncatula and rice requires the symbiotic GRAS-type transcription factors NSP1 and NSP2. Plant Cell **23,** 3853-3865.

Long, M., Betran, E., Thornton, K., and Wang, W. (2003). The origin of new genes: glimpses from the young and old. Nat Rev Genet **4,** 865-875.

Luginbuehl, L., and Oldroyd, G.E.D. (2016). Calcium signaling and transcriptional regulation in arbuscular mycorrhizal symbiosis. In Molecular Mycorrhizal Symbiosis (John Wiley & Sons, Inc.), pp. 125-140.

Luginbuehl, L.H., and Oldroyd, G.E. (2017). Understanding the Arbuscule at the Heart of Endomycorrhizal Symbioses in Plants. Current Biology **27,** R952-R963.

Luginbuehl, L.H., Menard, G.N., Kurup, S., Van Erp, H., Radhakrishnan, G.V., Breakspear, A., Oldroyd, G.E.D., and Eastmond, P.J. (2017). Fatty acids in arbuscular mycorrhizal fungi are synthesized by the host plant. Science.

Luo, R., Liu, B., Xie, Y., Li, Z., Huang, W., Yuan, J., He, G., Chen, Y., Pan, Q., and Liu, Y. (2012). SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. Gigascience **1,** 18.

MacLean, A.M., Bravo, A., and Harrison, M.J. (2017). Plant signaling and metabolic pathways enabling arbuscular mycorrhizal symbiosis. Plant Cell.

Madsen, E.B., Antolín-Llovera, M., Grossmann, C., Ye, J., Vieweg, S., Broghammer, A., Krusell, L., Radutoiu, S., Jensen, O.N., Stougaard, J., and Parniske, M. (2011). Autophosphorylation is essential for the in vivo function of the Lotus japonicus Nod factor receptor 1 and receptor-mediated signalling in cooperation with Nod factor receptor 5. The Plant Journal **65,** 404-417.

Madsen, L.H., Tirichine, L., Jurkiewicz, A., Sullivan, J.T., Heckmann, A.B., Bek, A.S., Ronson, C.W., James, E.K., and Stougaard, J. (2010). The molecular network governing nodule organogenesis and infection in the model legume Lotus japonicus **1,** 10.

Magadum, S., Banerjee, U., Murugan, P., Gangapur, D., and Ravikesavan, R. (2013). Gene duplication as a major force in evolution. Journal of genetics **92,** 155-161.

Maillet, F., Poinsot, V., Andre, O., Puech-Pages, V., Haouy, A., Gueunier, M., Cromer, L., Giraudet, D., Formey, D., Niebel, A., Martinez, E.A., Driguez, H., Becard, G., and Denarie, J. (2011). Fungal lipochitooligosaccharide symbiotic signals in arbuscular mycorrhiza. Nature **469,** 58-63.

Mapleson, D., Garcia Accinelli, G., Kettleborough, G., Wright, J., and Clavijo, B.J. (2017). KAT: a K-mer analysis toolkit to quality control NGS datasets and genome assemblies. Bioinformatics **33,** 574-576.

Markmann, K., Giczey, G., and Parniske, M. (2008). Functional Adaptation of a Plant Receptor- Kinase Paved the Way for the Evolution of Intracellular Root Symbioses with Bacteria. PLOS Biology **6,** e68.

Marsh, J.F., The Plant Laboratory, D.o.B.U.o.Y.P.O.B.Y.Y.O.Y.W.U.K., Schultze, M., and The Plant Laboratory, D.o.B.U.o.Y.P.O.B.Y.Y.O.Y.W.U.K. (2001). Analysis of arbuscular mycorrhizas using symbiosis- defective plant mutants. New Phytologist **150,** 525-532.

Martí-Solans, J., Belyaeva, O.V., Torres-Aguila, N.P., Kedishvili, N.Y., Albalat, R., and Cañestro, C. (2016). Coelimination and Survival in Gene Network

Evolution: Dismantling the RA-Signaling in a Chordate. Molecular Biology and Evolution **33,** 2401-2416.

Matasci, N., Hung, L.-H., Yan, Z., Carpenter, E.J., Wickett, N.J., Mirarab, S., Nguyen, N., Warnow, T., Ayyampalayam, S., Barker, M., Burleigh, J.G., Gitzendanner, M.A., Wafula, E., Der, J.P., dePamphilis, C.W., Roure, B., Philippe, H., Ruhfel, B.R., Miles, N.W., Graham, S.W., Mathews, S., Surek, B., Melkonian, M., Soltis, D.E., Soltis, P.S., Rothfels, C., Pokorny, L., Shaw, J.A., DeGironimo, L., Stevenson, D.W., Villarreal, J.C., Chen, T., Kutchan, T.M., Rolf, M., Baucom, R.S., Deyholos, M.K., Samudrala, R., Tian, Z., Wu, X., Sun, X., Zhang, Y., Wang, J., Leebens-Mack, J., and Wong, G.K.-S. (2014). Data access for the 1,000 Plants (1KP) project. GigaScience **3,** 17-17.

Messinese, E., Mun, J.-H., Yeun, L.H., Jayaraman, D., Rougé, P., Barre, A., Lougnon, G., Schornack, S., Bono, J.-J., Cook, D.R., and Ané, J.-M. (2007). A Novel Nuclear Protein Interacts With the Symbiotic DMI3 Calcium- and Calmodulin-Dependent Protein Kinase of Medicago truncatula. Molecular Plant-Microbe Interactions **20,** 912-921.

Miller, J.B., Pratap, A., Miyahara, A., Zhou, L., Bornemann, S., Morris, R.J., and Oldroyd, G.E. (2013). Calcium/Calmodulin-dependent protein kinase is negatively and positively regulated by calcium, providing a mechanism for decoding calcium responses during symbiosis signaling. The Plant Cell **25,** 5053-5066.

Mitra, R.M., Gleason, C.A., Edwards, A., Hadfield, J., Downie, J.A., Oldroyd, G.E., and Long, S.R. (2004). A Ca2+/calmodulin-dependent protein kinase required for symbiotic nodule development: gene identification by transcript-based cloning. Proceedings of the National Academy of Sciences of the United States of America **101,** 4701-4705.

Miwa, H., Sun, J., Oldroyd, G.E.D., and Downie, J.A. (2006). Analysis of Nod-Factor-Induced Calcium Signaling in Root Hairs of Symbiotically Defective Mutants of Lotus japonicus. Molecular Plant-Microbe Interactions **19,** 914-923.

Miya, A., Albert, P., Shinya, T., Desaki, Y., Ichimura, K., Shirasu, K., Narusaka, Y., Kawakami, N., Kaku, H., and Shibuya, N. (2007). CERK1, a LysM receptor kinase, is essential for chitin elicitor signaling in Arabidopsis. Proc Natl Acad Sci U S A **104,** 19613-19618.

Miyata, K., Kozaki, T., Kouzai, Y., Ozawa, K., Ishii, K., Asamizu, E., Okabe, Y., Umehara, Y., Miyamoto, A., Kobae, Y., Akiyama, K., Kaku, H., Nishizawa, Y., Shibuya, N., and Nakagawa, T. (2014). The Bifunctional Plant Receptor, OsCERK1, Regulates Both Chitin-Triggered Immunity and Arbuscular Mycorrhizal Symbiosis in Rice. Plant and Cell Physiology **55,** 1864-1872.

Murray, J.D., Duvvuru Muni, R., Torres-Jerez, I., Tang, Y., Allen, S., Andriankaja, M., Li, G., Laxmi, A., Cheng, X., Wen, J., Vaughan, D., Schultze, M., Sun, J., Charpentier, M., Oldroyd, G., Tadege, M., Ratet, P., Mysore, K.S., Chen, R., and Udvardi, M.K. (2011). Vapyrin, a gene essential for intracellular progression of arbuscular mycorrhizal symbiosis, is also essential for infection by rhizobia in the nodule symbiosis of Medicago truncatula. The Plant journal : for cell and molecular biology **65**.

Nabhan, A.R., and Sarkar, I.N. (2012). The impact of taxon sampling on phylogenetic inference: a review of two decades of controversy. Briefings in Bioinformatics **13,** 122-134.

Nagae, M., Takeda, N., and Kawaguchi, M. (2014). Common symbiosis genes CERBERUS and NSP1 provide additional insight into the establishment of arbuscular mycorrhizal and root nodule symbioses in Lotus japonicus. Plant Signaling & Behavior **9,** e28544.

Nakagawa, T., and Imaizumi-Anraku, H. (2015). Rice arbuscular mycorrhiza as a tool to study the molecular mechanisms of fungal symbiosis and a potential target to increase productivity. Rice **8,** 32.

Nakasugi, K., Crowhurst, R., Bally, J., and Waterhouse, P. (2014). Combining Transcriptome Assemblies from Multiple De Novo Assemblers in the Allo-Tetraploid Plant Nicotiana benthamiana. PLOS ONE **9,** e91776.

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

Nystedt, B., Street, N.R., Wetterbom, A., Zuccolo, A., Lin, Y.-C., Scofield, D.G., Vezzi, F., Delhomme, N., Giacomello, S., Alexeyenko, A., Vicedomini, R., Sahlin, K., Sherwood, E., Elfstrand, M., Gramzow, L., Holmberg, K., Hallman, J., Keech, O., Klasson, L., Koriabine, M., Kucukoglu, M., Kaller, M., Luthman, J., Lysholm, F., Niittyla, T., Olson, A., Rilakovic, N., Ritland, C., Rossello, J.A., Sena, J., Svensson, T., Talavera-Lopez, C., Theiszen, G., Tuominen, H., Vanneste, K., Wu, Z.-Q., Zhang, B., Zerbe,

P., Arvestad, L., Bhalerao, R., Bohlmann, J., Bousquet, J., Garcia Gil, R., Hvidsten, T.R., de Jong, P., MacKay, J., Morgante, M., Ritland, K., Sundberg, B., Lee Thompson, S., Van de Peer, Y., Andersson, B., Nilsson, O., Ingvarsson, P.K., Lundeberg, J., and Jansson, S. (2013). The Norway spruce genome sequence and conifer genome evolution. Nature **497,** 579-584.

Oldroyd, G.E.D. (2013). Speak, friend, and enter: signalling systems that promote beneficial symbiotic associations in plants. Nat Rev Microbiol **11**.

Oldroyd, G.E.D., and Downie, J.A. (2004). Calcium, kinases and nodulation signalling in legumes. Nat Rev Mol Cell Biol **5,** 566-576.

Oldroyd, G.E.D., Harrison, M.J., and Paszkowski, U. (2009). Reprogramming plant cells for endosymbiosis. Science **324,** 753--754.

Owen, R. (1848). On the Archetype and Homologies of the Vertebrate Skeleton. (author).

Parfrey, L.W., Lahr, D.J., Knoll, A.H., and Katz, L.A. (2011). Estimating the timing of early eukaryotic diversification with multigene molecular clocks. Proceedings of the National Academy of Sciences **108,** 13624-13629.

Park, H.-J., Floss, D.S., Levesque-Tremblay, V., Bravo, A., and Harrison, M.J. (2015). Hyphal Branching during Arbuscule Development Requires Reduced Arbuscular Mycorrhiza1. Plant Physiology **169,** 2774-2788.

Parniske, M. (2008). Arbuscular mycorrhiza: the mother of plant root endosymbioses. Nat Rev Micro **6,** 763-775.

Pasek, S., Risler, J.-L., and Brézellec, P. (2006). Gene fusion/fission is a major contributor to evolution of multi-domain bacterial proteins. Bioinformatics **22,** 1418-1423.

Paszkowski, U., Jakovleva, L., and Boller, T. (2006). Maize mutants affected at distinct stages of the arbuscular mycorrhizal symbiosis. The Plant Journal **47,** 165-173.

Patil, S., Takezawa, D., and Poovaiah, B.W. (1995). Chimeric plant calcium/calmodulin-dependent protein kinase gene with a neural visinin-like calcium-binding domain. Proceedings of the National Academy of Sciences **92,** 4897-4901.

Patron, N.J., Orzaez, D., Marillonnet, S., Warzecha, H., Matthewman, C., Youles, M., Raitskin, O., Leveau, A., Farre, G., Rogers, C., Smith, A., Hibberd, J., Webb, A.A., Locke, J., Schornack, S., Ajioka, J., Baulcombe, D.C., Zipfel, C., Kamoun, S., Jones, J.D., Kuhn, H., Robatzek, S., Van

Esse, H.P., Sanders, D., Oldroyd, G., Martin, C., Field, R., O'Connor, S., Fox, S., Wulff, B., Miller, B., Breakspear, A., Radhakrishnan, G., Delaux, P.M., Loque, D., Granell, A., Tissier, A., Shih, P., Brutnell, T.P., Quick, W.P., Rischer, H., Fraser, P.D., Aharoni, A., Raines, C., South, P.F., Ane, J.M., Hamberger, B.R., Langdale, J., Stougaard, J., Bouwmeester, H., Udvardi, M., Murray, J.A., Ntoukakis, V., Schafer, P., Denby, K., Edwards, K.J., Osbourn, A., and Haseloff, J. (2015). Standards for plant synthetic biology: a common syntax for exchange of DNA parts. New Phytol **208,** 13-19.

Peng, Y., Leung, H.C., Yiu, S.-M., and Chin, F.Y. (2012). IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. Bioinformatics **28,** 1420-1428.

Pfeffer, P.E., Douds, D.D., Bécard, G., and Shachar-Hill, Y. (1999). Carbon Uptake and the Metabolism and Transport of Lipids in an Arbuscular Mycorrhiza. Plant Physiology **120,** 587-598.

Pimprikar, P., Carbonnel, S., Paries, M., Katzer, K., Klingl, V., Bohmer, M.J., Karl, L., Floss, D.S., Harrison, M.J., Parniske, M., and Gutjahr, C. (2016). A CCaMK-CYCLOPS-DELLA Complex Activates Transcription of RAM1 to Regulate Arbuscule Branching. Current biology : CB **26,** 987-998.

Pires, N.D., Yi, K., Breuninger, H., Catarino, B., Menand, B., and Dolan, L. (2013). Recruitment and remodeling of an ancient gene regulatory network during land plant evolution. Proceedings of the National Academy of Sciences **110,** 9571-9576.

Pirozynski, K.A., and Malloch, D.W. (1975). The origin of land plants: a matter of mycotrophism. Bio Systems **6,** 153-164.

Pozo, M.J., and Azcon-Aguilar, C. (2007). Unraveling mycorrhiza-induced resistance. Current opinion in plant biology **10,** 393-398.

Pressel, S., Bidartondo, M.I., Field, K.J., Rimington, W.R., and Duckett, J.G. (2016). Pteridophyte fungal associations: Current knowledge and future perspectives. Journal of Systematics and Evolution **54,** 666-678.

Proust, H., Honkanen, S., Jones, Victor A., Morieri, G., Prescott, H., Kelly, S., Ishizaki, K., Kohchi, T., and Dolan, L. (2016). RSL Class I Genes Controlled the Development of Epidermal Structures in the Common Ancestor of Land Plants. Current Biology **26,** 93-99.

Pumplin, N., Mondo, S.J., Topp, S., Starker, C.G., Gantt, J.S., and Harrison, M.J. (2010). Medicago truncatula Vapyrin is a novel protein required for arbuscular mycorrhizal symbiosis. The Plant journal : for cell and molecular biology **61,** 482-494.

Rasmussen, S.R., Füchtbauer, W., Novero, M., Volpe, V., Malkov, N., Genre, A., Bonfante, P., Stougaard, J., and Radutoiu, S. (2016). Intraradical colonization by arbuscular mycorrhizal fungi triggers induction of a lipochitooligosaccharide receptor **6,** 29733.

Redecker, D., Kodner, R., and Graham, L.E. (2000). Glomalean fungi from the Ordovician. Science **289,** 1920-1921.

Redecker, D., Hijri, I., and Wiemken, A. (2003). Molecular identification of arbuscular mycorrhizal fungi in roots: Perspectives and problems. Folia Geobotanica **38,** 113-124.

Remy, W., Taylor, T.N., Hass, H., and Kerp, H. (1994). Four hundred-million-year-old vesicular arbuscular mycorrhizae. Proc Natl Acad Sci U S A **91,** 11841-11843.

Rensing, S.A. (2017). Why we need more non-seed plant models. New Phytologist**,** n/a-n/a.

Rensing, S.A., Lang, D., Zimmer, A.D., Terry, A., Salamov, A., Shapiro, H., Nishiyama, T., Perroud, P.F., Lindquist, E.A., Kamisugi, Y., Tanahashi, T., Sakakibara, K., Fujita, T., Oishi, K., Shin, I.T., Kuroki, Y., Toyoda, A., Suzuki, Y., Hashimoto, S., Yamaguchi, K., Sugano, S., Kohara, Y., Fujiyama, A., Anterola, A., Aoki, S., Ashton, N., Barbazuk, W.B., Barker, E., Bennetzen, J.L., Blankenship, R., Cho, S.H., Dutcher, S.K., Estelle, M., Fawcett, J.A., Gundlach, H., Hanada, K., Heyl, A., Hicks, K.A., Hughes, J., Lohr, M., Mayer, K., Melkozernov, A., Murata, T., Nelson, D.R., Pils, B., Prigge, M., Reiss, B., Renner, T., Rombauts, S., Rushton, P.J., Sanderfoot, A., Schween, G., Shiu, S.H., Stueber, K., Theodoulou, F.L., Tu, H., Van de Peer, Y., Verrier, P.J., Waters, E., Wood, A., Yang, L., Cove, D., Cuming, A.C., Hasebe, M., Lucas, S., Mishler, B.D., Reski, R., Grigoriev, I.V., Quatrano, R.S., and Boore, J.L. (2008). The Physcomitrella genome reveals evolutionary insights into the conquest of land by plants. Science **319,** 64-69.

Renzaglia, K.S., Duff, R.J., Nickrent, D.L., and Garbary, D.J. (2000). Vegetative and reproductive innovations of early land plants: implications for a unified phylogeny. Philos Transs R Soc Lon B **355**.

Rich, M.K., Schorderet, M., Bapaume, L., Falquet, L., Morel, P., Vandenbussche, M., and Reinhardt, D. (2015). The Petunia GRAS Transcription Factor ATA/RAM1 Regulates Symbiotic Gene Expression and Fungal Morphogenesis in Arbuscular Mycorrhiza. Plant Physiol **168,** 788-797.

Ried, M.K., Antolin-Llovera, M., and Parniske, M. (2014). Spontaneous symbiotic reprogramming of plant roots triggered by receptor-like kinases. Elife **3**.

Robinson, M.D., McCarthy, D.J., and Smyth, G.K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics **26,** 139-140.

Rubinstein, C.V., Gerrienne, P., de la Puente, G.S., Astini, R.A., and Steemans, P. (2010). Early Middle Ordovician evidence for land plants in Argentina (eastern Gondwana). New Phytologist **188,** 365-369.

Ruhfel, B.R., Gitzendanner, M.A., Soltis, P.S., Soltis, D.E., and Burleigh, J.G. (2014). From algae to angiosperms–inferring the phylogeny of green plants (Viridiplantae) from 360 plastid genomes. BMC Evolutionary Biology **14,** 23.

Russell, J., and Bulman, S. (2005). The liverwort Marchantia foliacea forms a specialized symbiosis with arbuscular mycorrhizal fungi in the genus Glomus. New Phytologist **165,** 567-579.

Saito, K., Yoshikawa, M., Yano, K., Miwa, H., Uchida, H., Asamizu, E., Sato, S., Tabata, S., Imaizumi-Anraku, H., Umehara, Y., Kouchi, H., Murooka, Y., Szczyglowski, K., Downie, J.A., Parniske, M., Hayashi, M., and Kawaguchi, M. (2007). NUCLEOPORIN85 is required for calcium spiking, fungal and bacterial symbioses, and seed production in Lotus japonicus. Plant Cell **19**.

Sakakibara, K., Nishiyama, T., Deguchi, H., and 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,** 555-566.

Sato, S., Nakamura, Y., Kaneko, T., Asamizu, E., Kato, T., Nakao, M., Sasamoto, S., Watanabe, A., Ono, A., Kawashima, K., Fujishiro, T., Katoh, M., Kohara, M., Kishida, Y., Minami, C., Nakayama, S., Nakazaki, N., Shimizu, Y., Shinpo, S., Takahashi, C., Wada, T., Yamada, M., Ohmido,

N., Hayashi, M., Fukui, K., Baba, T., Nakamichi, T., Mori, H., and Tabata, S. (2008). Genome Structure of the Legume, Lotus japonicus. DNA Res **15,** 227-239.

Sayou, C., Monniaux, M., Nanao, M.H., Moyroud, E., Brockington, S.F., Thévenon, E., Chahtane, H., Warthmann, N., Melkonian, M., Zhang, Y., Wong, G.K.-S., Weigel, D., Parcy, F., and Dumas, R. (2014). A Promiscuous Intermediate Underlies the Evolution of LEAFY DNA Binding Specificity. Science.

Schlötterer, C. (2015). Genes from scratch – the evolutionary fate of de novo genes. Trends in Genetics **31,** 215-219.

Schneider, A., Walker, S.A., Poyser, S., Sagan, M., Ellis, T.H.N., and Downie, J.A. (1999). Genetic mapping and functional analysis of a nodulation-defective mutant (sym19) of pea (Pisum sativum L.). Molecular and General Genetics MGG **262,** 1-11.

Scotland, R.W. (2010). Deep homology: a view from systematics. Bioessays **32,** 438-449.

Seilacher, A., Reif, W.E., Westphal, F., Riding, R., Clarkson, E.N.K., and Whittington, H.B. (1985). Sedimentological, Ecological and Temporal Patterns of Fossil Lagerstatten [and Discussion]. Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences **311,** 5-24.

Seki, H., Tamura, K., and Muranaka, T. (2015). P450s and UGTs: Key Players in the Structural Diversity of Triterpenoid Saponins. Plant and Cell Physiology **56,** 1463-1471.

Shachar-Hill, Y., Pfeffer, P.E., Douds, D., Osman, S.F., Doner, L.W., and Ratcliffe, R.G. (1995). Partitioning of Intermediary Carbon Metabolism in Vesicular-Arbuscular Mycorrhizal Leek. Plant Physiology **108,** 7-15.

Shi, W., Shi, M.W., PGM, L.I., SequenceMatching, R., ChIPSeq, G., and GeneRegulation, G. (2013). Package 'Rsubread'.

Shimakawa, G., Ishizaki, K., Tsukamoto, S., Tanaka, M., Sejima, T., and Miyake, C. (2017). The Liverwort, *Marchantia*, Drives Alternative Electron Flow Using a Flavodiiron Protein to Protect PSI. Plant Physiology **173,** 1636-1647.

Shimamura, M. (2016). Marchantia polymorpha : Taxonomy, Phylogeny and Morphology of a Model System. Plant and Cell Physiology **57,** 230-256.

Shimoda, Y., Han, L., Yamazaki, T., Suzuki, R., Hayashi, M., and Imaizumi-Anraku, H. (2012). Rhizobial and fungal symbioses show different requirements for calmodulin binding to calcium calmodulin–dependent protein kinase in Lotus japonicus. The Plant Cell **24,** 304-321.

Shtark, O.Y., Borisov, A.Y., Zhukov, V.A., Provorov, N.A., and Tikhonovich, I.A. (2010). Intimate Associations of Beneficial Soil Microbes with Host Plants. In Soil Microbiology and Sustainable Crop Production, G.R. Dixon and E.L. Tilston, eds (Dordrecht: Springer Netherlands), pp. 119-196.

Simão, F.A., Waterhouse, R.M., Ioannidis, P., Kriventseva, E.V., and Zdobnov, E.M. (2015). BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics **31,** 3210-3212.

Simpson, J.T., Wong, K., Jackman, S.D., Schein, J.E., Jones, S.J., and Birol, I. (2009). ABySS: a parallel assembler for short read sequence data. Genome research **19,** 1117-1123.

Sims, D., Sudbery, I., Ilott, N.E., Heger, A., and Ponting, C.P. (2014). Sequencing depth and coverage: key considerations in genomic analyses. Nat Rev Genet **15,** 121-132.

Singer, S.D., and Ashton, N.W. (2007). Revelation of ancestral roles of KNOX genes by a functional analysis of Physcomitrella homologues. Plant cell reports **26,** 2039-2054.

Singh, S., Katzer, K., Lambert, J., Cerri, M., and Parniske, M. (2014). CYCLOPS, A DNA-Binding Transcriptional Activator, Orchestrates Symbiotic Root Nodule Development. Cell Host & Microbe **15,** 139-152.

Smit, P., Raedts, J., Portyanko, V., and Debell. (2005). NSP1 of the GRAS protein family is essential for rhizobial Nod factor-induced transcription. Science **308,** 1789--1791.

Smith, S.A., and Pease, J.B. (2017). Heterogeneous molecular processes among the causes of how sequence similarity scores can fail to recapitulate phylogeny. Briefings in Bioinformatics **18,** 451-457.

Smith, S.E., and Read, D.J. (2008). Mycorrhizal Symbiosis. (New York: Academic).

Smith, S.E., and Smith, F.A. (2011). Roles of Arbuscular Mycorrhizas in Plant Nutrition and Growth: New Paradigms from Cellular to Ecosystem Scales. Annual Review of Plant Biology **62,** 227-250.

Solaimanand, M.Z., and Saito, M. (1997). Use of sugars by intraradical hyphae of arbuscular mycorrhizal fungi revealed by radiorespirometry. New Phytologist **136,** 533-538.

Spatafora, J.W., Aime, M.C., Grigoriev, I.V., Martin, F., Stajich, J.E., and Blackwell, M. (2017). The Fungal Tree of Life: from Molecular Systematics to Genome-Scale Phylogenies. Microbiology Spectrum **5**.

Spatafora, J.W., Chang, Y., Benny, G.L., Lazarus, K., Smith, M.E., Berbee, M.L., Bonito, G., Corradi, N., Grigoriev, I., and Gryganskyi, A. (2016). A phylum-level phylogenetic classification of zygomycete fungi based on genome-scale data. Mycologia **108,** 1028-1046.

Sprent, J.I. (2007). Evolving ideas of legume evolution and diversity: a taxonomic perspective on the occurrence of nodulation. New Phytologist **174,** 11-25.

Stamatakis, A. (2006). RAxML-VI-HPC: Maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics **22**.

Stanke, M., Steinkamp, R., Waack, S., and Morgenstern, B. (2004). AUGUSTUS: a web server for gene finding in eukaryotes. Nucleic acids research **32,** W309-W312.

Stiller, J., Martirani, L., Tuppale, S., Chian, R.-J., Chiurazzi, M., and Gresshoff, P.M. (1997). High frequency transformation and regeneration of transgenic plants in the model legume Lotus japonicus. Journal of experimental botany **48,** 1357-1365.

Stracke, S., Kistner, C., Yoshida, S., Mulder, L., Sato, S., Kaneko, T., Tabata, S., Sandal, N., Stougaard, J., Szczyglowski, K., and Parniske, M. (2002). A plant receptor-like kinase required for both bacterial and fungal symbiosis. Nature **417,** 959-962.

Strullu-Derrien, C., Kenrick, P., and Selosse, M.-A. (2016). Origins of the mycorrhizal symbioses. In Molecular Mycorrhizal Symbiosis (John Wiley & Sons, Inc.), pp. 1-20.

Sugano, S.S., Shirakawa, M., Takagi, J., Matsuda, Y., Shimada, T., Hara-Nishimura, I., and Kohchi, T. (2014). CRISPR/Cas9-Mediated Targeted Mutagenesis in the Liverwort Marchantia polymorpha L. Plant and Cell Physiology **55,** 475-481.

Sun, J., Miller, J.B., Granqvist, E., Wiley-Kalil, A., Gobbato, E., Maillet, F., Cottaz, S., Samain, E., Venkateshwaran, M., Fort, S., Morris, R.J., Ané, J.-M., Dénarié, J., and Oldroyd, G.E.D. (2015). Activation of Symbiosis

Signaling by Arbuscular Mycorrhizal Fungi in Legumes and Rice. The Plant Cell Online.

Sun, T.-p. (2011). The Molecular Mechanism and Evolution of the GA–GID1–DELLA Signaling Module in Plants. Current Biology **21**.

Suyama, M., Torrents, D., and Bork, P. (2006). PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. Nucleic Acids Res **34**.

Szovenyi, P., Frangedakis, E., Ricca, M., Quandt, D., Wicke, S., and Langdale, J.A. (2015). Establishment of Anthoceros agrestis as a model species for studying the biology of hornworts. BMC Plant Biol **15,** 98.

Takeda, N., Maekawa, T., and Hayashi, M. (2012). Nuclear-localized and deregulated calcium-and calmodulin-dependent protein kinase activates rhizobial and mycorrhizal responses in Lotus japonicus. The Plant Cell **24,** 810-822.

Talhinhas, P., Muthumeenakshi, S., Neves-Martins, J., Oliveira, H., and Sreenivasaprasad, S. (2008). Agrobacterium-mediated transformation and insertional mutagenesis in Colletotrichum acutatum for investigating varied pathogenicity lifestyles. Molecular biotechnology **39,** 57-67.

Tam, T.H.Y., Catarino, B., and Dolan, L. (2015). Conserved regulatory mechanism controls the development of cells with rooting functions in land plants. Proceedings of the National Academy of Sciences **112,** E3959-E3968.

Tang, N., San Clemente, H., Roy, S., Bécard, G., Zhao, B., and Roux, C. (2016). A Survey of the Gene Repertoire of Gigaspora rosea Unravels Conserved Features among Glomeromycota for Obligate Biotrophy. Frontiers in Microbiology **7**.

Tatusov, R.L., Koonin, E.V., and Lipman, D.J. (1997). A genomic perspective on protein families. Science **278**.

Tautz, D. (2014). The discovery of de novo gene evolution. Perspectives in biology and medicine **57,** 149-161.

Taylor, T.N., Kerp, H., and Hass, H. (2005). Life history biology of early land plants: Deciphering the gametophyte phase. Proceedings of the National Academy of Sciences of the United States of America **102,** 5892-5897.

Taylor, T.N., Remy, W., Hass, H., and Kerp, H. (1995). Fossil arbuscular mycorrhizae from the Early Devonian. Mycologia**,** 560-573.

Tirichine, L., Imaizumi-Anraku, H., Yoshida, S., Murakami, Y., Madsen, L.H., Miwa, H., Nakagawa, T., Sandal, N., Albrektsen, A.S., Kawaguchi, M., Downie, A., Sato, S., Tabata, S., Kouchi, H., Parniske, M., Kawasaki, S., and Stougaard, J. (2006). Deregulation of a Ca2+/calmodulin-dependent kinase leads to spontaneous nodule development. Nature **441,** 1153-1156.

Tisserant, E., Malbreil, M., Kuo, A., Kohler, A., Symeonidi, A., Balestrini, R., Charron, P., Duensing, N., Frei dit Frey, N., Gianinazzi-Pearson, V., Gilbert, L.B., Handa, Y., Herr, J.R., Hijri, M., Koul, R., Kawaguchi, M., Krajinski, F., Lammers, P.J., Masclaux, F.G., Murat, C., Morin, E., Ndikumana, S., Pagni, M., Petitpierre, D., Requena, N., Rosikiewicz, P., Riley, R., Saito, K., San Clemente, H., Shapiro, H., van Tuinen, D., Bécard, G., Bonfante, P., Paszkowski, U., Shachar-Hill, Y.Y., Tuskan, G.A., Young, J.P.W., Sanders, I.R., Henrissat, B., Rensing, S.A., Grigoriev, I.V., Corradi, N., Roux, C., and Martin, F. (2013). Genome of an arbuscular mycorrhizal fungus provides insight into the oldest plant symbiosis. Proceedings of the National Academy of Sciences **110,** 20117-20122.

Tsuboyama-Tanaka, S., and Kodama, Y. (2015). AgarTrap-mediated genetic transformation using intact gemmae/gemmalings of the liverwort Marchantia polymorpha L. Journal of plant research **128,** 337-344.

Venkateshwaran, M., Cosme, A., Han, L., Banba, M., Satyshur, K.A., Schleiff, E., Parniske, M., Imaizumi-Anraku, H., and Ané, J.-M. (2012). The Recent Evolution of a Symbiotic Ion Channel in the Legume Family Altered Ion Conductance and Improved Functionality in Calcium Signaling. The Plant Cell **24,** 2528-2545.

Vernié, T., Kim, J., Frances, L., Ding, Y., Sun, J., Guan, D., Niebel, A., Gifford, M.L., de Carvalho-Niebel, F., and Oldroyd, G.E. (2015). The NIN transcription factor coordinates diverse nodulation programs in different tissues of the Medicago truncatula root. The Plant Cell **27,** 3410-3424.

Villarreal A, J.C., Crandall-Stotler, B.J., Hart, M.L., Long, D.G., and Forrest, L.L. (2016). Divergence times and the evolution of morphological complexity in an early land plant lineage (Marchantiopsida) with a slow molecular rate. New Phytologist **209,** 1734-1746.

Vujičić, M., Cvetić, T., Sabovljević, A., and Sabovljević, M. (2010). Axenically culturing the bryophytes: A case study of the liverwort Marchantia

polymorpha L. ssp. ruderalis Bischl. & Boisselier (Marchantiophyta, Marchantiaceae). Kragujevac Journal of Science**,** 73-81.

Wang, B., and Qiu, Y.L. (2006). Phylogenetic distribution and evolution of mycorrhizas in land plants. Mycorrhiza **16,** 299-363.

Wang, B., Yeun, L.H., Xue, J.-Y., Liu, Y., Ané, J.-M., and Qiu, Y.-L. (2010). Presence of three mycorrhizal genes in the common ancestor of land plants suggests a key role of mycorrhizas in the colonization of land by plants. New Phytologist **186,** 514-525.

Wang, E., Schornack, S., Marsh, J.F., Gobbato, E., Schwessinger, B., Eastmond, P., Schultze, M., Kamoun, S., and Oldroyd, G.E. (2012). A common signaling process that promotes mycorrhizal and oomycete colonization of plants. Current biology : CB **22,** 2242-2246.

Wang, E., Yu, N., Bano, S.A., Liu, C., Miller, A.J., Cousins, D., Zhang, X., Ratet, P., Tadege, M., and Mysore, K.S. (2014). A H+-ATPase that energizes nutrient uptake during mycorrhizal symbioses in rice and Medicago truncatula. The Plant cell **26,** 1818-1830.

Wang, J.-P., Munyampundu, J.-P., Xu, Y.-P., and Cai, X.-Z. (2015). Phylogeny of Plant Calcium and Calmodulin-Dependent Protein Kinases (CCaMKs) and Functional Analyses of Tomato CCaMK in Disease Resistance. Frontiers in Plant Science **6,** 1075.

Wang, M.-B., Helliwell, C.A., Wu, L.-M., Waterhouse, P.M., Peacock, W.J., and Dennis, E.S. (2008). Hairpin RNAs derived from RNA polymerase II and polymerase III promoter-directed transgenes are processed differently in plants. RNA **14,** 903-913.

Ward, P., Equinet, L., Packer, J., and Doerig, C. (2004). Protein kinases of the human malaria parasite Plasmodium falciparum: the kinome of a divergent eukaryote. BMC Genomics **5,** 79.

Watts-Williams, S.J., and Cavagnaro, T.R. (2015). Using mycorrhiza-defective mutant genotypes of non-legume plant species to study the formation and functioning of arbuscular mycorrhiza: a review. Mycorrhiza **25,** 587-597.

Wegel, E. (2017). Fluorescence In Situ Hybridization in Oat. In Oat: Methods and Protocols, S. Gasparis, ed (New York, NY: Springer New York), pp. 3-21.

Wernimont, A.K., Artz, J.D., Finerty, P., Lin, Y.-H., Amani, M., Allali-Hassani, A., Senisterra, G., Vedadi, M., Tempel, W., Mackenzie, F., Chau, I., Lourido, S., Sibley, L.D., and Hui, R. (2010). Structures of apicomplexan

calcium-dependent protein kinases reveal mechanism of activation by calcium. Nat Struct Mol Biol **17,** 596-601.

Wertheim, J.O., Murrell, B., Smith, M.D., Kosakovsky Pond, S.L., and Scheffler, K. (2015). RELAX: Detecting Relaxed Selection in a Phylogenetic Framework. Molecular Biology and Evolution **32,** 820-832.

Wewer, V., Brands, M., and Dormann, P. (2014). Fatty acid synthesis and lipid metabolism in the obligate biotrophic fungus Rhizophagus irregularis during mycorrhization of Lotus japonicus. Plant Journal **79,** 398-412.

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., Rothfels, C.J., Pokorny, L., Shaw, A.J., DeGironimo, L., Stevenson, D.W., Surek, B., Villarreal, J.C., Roure, B., Philippe, H., dePamphilis, C.W., Chen, T., Deyholos, M.K., Baucom, R.S., Kutchan, T.M., Augustin, M.M., Wang, J., Zhang, Y., Tian, Z., Yan, Z., Wu, X., Sun, X., Wong, G.K.-S., and Leebens-Mack, J. (2014). Phylotranscriptomic analysis of the origin and early diversification of land plants. Proceedings of the National Academy of Sciences **111,** E4859-E4868.

Xie, Y., Wu, G., Tang, J., Luo, R., Patterson, J., Liu, S., Huang, W., He, G., Gu, S., Li, S., Zhou, X., Lam, T.-W., Li, Y., Xu, X., Wong, G.K.-S., and Wang, J. (2014). SOAPdenovo-Trans: de novo transcriptome assembly with short RNA-Seq reads. Bioinformatics **30,** 1660-1666.

Xue, L., Cui, H., Buer, B., Vijayakumar, V., Delaux, P.M., Junkermann, S., and Bucher, M. (2015). Network of GRAS transcription factors involved in the control of arbuscule development in Lotus japonicus. Plant Physiol **167,** 854-871.

Yang, Z. (1997). PAML: a program package for phylogenetic analysis by maximum likelihood. Bioinformatics **13,** 555-556.

Yang, Z., and Rannala, B. (2012). Molecular phylogenetics: principles and practice. Nat Rev Genet **13,** 303-314.

Yano, K., Yoshida, S., Muller, J., Singh, S., Banba, M., Vickers, K., Markmann, K., White, C., Schuller, B., Sato, S., Asamizu, E., Tabata, S., Murooka, Y., Perry, J., Wang, T.L., Kawaguchi, M., Imaizumi-Anraku, H., Hayashi,

M., and Parniske, M. (2008). CYCLOPS, a mediator of symbiotic intracellular accommodation. Proc Natl Acad Sci U S A **105**.

Yasumura, Y., Crumpton-Taylor, M., Fuentes, S., and Harberd, N.P. (2007). Step-by-Step Acquisition of the Gibberellin-DELLA Growth-Regulatory Mechanism during Land-Plant Evolution. Current Biology **17,** 1225-1230.

Young, N.D., Debelle, F., Oldroyd, G.E.D., Geurts, R., Cannon, S.B., Udvardi, M.K., Benedito, V.A., Mayer, K.F.X., Gouzy, J., Schoof, H., Van de Peer, Y., Proost, S., Cook, D.R., Meyers, B.C., Spannagl, M., Cheung, F., De Mita, S., Krishnakumar, V., Gundlach, H., Zhou, S., Mudge, J., Bharti, A.K., Murray, J.D., Naoumkina, M.A., Rosen, B., Silverstein, K.A.T., Tang, H., Rombauts, S., Zhao, P.X., Zhou, P., Barbe, V., Bardou, P., Bechner, M., Bellec, A., Berger, A., Berges, H., Bidwell, S., Bisseling, T., Choisne, N., Couloux, A., Denny, R., Deshpande, S., Dai, X., Doyle, J.J., Dudez, A.-M., Farmer, A.D., Fouteau, S., Franken, C., Gibelin, C., Gish, J., Goldstein, S., Gonzalez, A.J., Green, P.J., Hallab, A., Hartog, M., Hua, A., Humphray, S.J., Jeong, D.-H., Jing, Y., Jocker, A., Kenton, S.M., Kim, D.-J., Klee, K., Lai, H., Lang, C., Lin, S., Macmil, S.L., Magdelenat, G., Matthews, L., McCorrison, J., Monaghan, E.L., Mun, J.-H., Najar, F.Z., Nicholson, C., Noirot, C., O/'Bleness, M., Paule, C.R., Poulain, J., Prion, F., Qin, B., Qu, C., Retzel, E.F., Riddle, C., Sallet, E., Samain, S., Samson, N., Sanders, I., Saurat, O., Scarpelli, C., Schiex, T., Segurens, B., Severin, A.J., Sherrier, D.J., Shi, R., Sims, S., Singer, S.R., Sinharoy, S., Sterck, L., Viollet, A., Wang, B.-B., Wang, K., Wang, M., Wang, X., Warfsmann, J., Weissenbach, J., White, D.D., White, J.D., Wiley, G.B., Wincker, P., Xing, Y., Yang, L., Yao, Z., Ying, F., Zhai, J., Zhou, L., Zuber, A., Denarie, J., Dixon, R.A., May, G.D., Schwartz, D.C., Rogers, J., Quetier, F., Town, C.D., and Roe, B.A. (2011). The Medicago genome provides insight into the evolution of rhizobial symbioses. Nature **480,** 520-524.

Zerbino, D.R., and Birney, E. (2008). Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome research **18,** 821-829.

Zhang, Q., Blaylock, L.A., and Harrison, M.J. (2010). Two Medicago truncatula half-ABC transporters are essential for arbuscule development in arbuscular mycorrhizal symbiosis. Plant Cell **22,** 1483-1497.

Zhang, R., Guo, C., Zhang, W., Wang, P., Li, L., Duan, X., Du, Q., Zhao, L., Shan, H., Hodges, S.A., Kramer, E.M., Ren, Y., and Kong, H. (2013).

Disruption of the petal identity gene APETALA3-3 is highly correlated with loss of petals within the buttercup family (Ranunculaceae). Proceedings of the National Academy of Sciences **110,** 5074-5079.

Zhang, X., Dong, W., Sun, J., Feng, F., Deng, Y., He, Z., Oldroyd, G.E., and Wang, E. (2015). The receptor kinase CERK1 has dual functions in symbiosis and immunity signalling. The Plant journal : for cell and molecular biology **81,** 258-267.

Zhang, X.C., Cannon, S.B., and Stacey, G. (2009). Evolutionary genomics of LysM genes in land plants. BMC Evol Biol **9,** 183.

Zhang, X.C., Wu, X., Findley, S., Wan, J., Libault, M., Nguyen, H.T., Cannon, S.B., and Stacey, G. (2007). Molecular evolution of lysin motif-type receptor-like kinases in plants. Plant Physiol **144,** 623-636.

Zhong, J., and Kellogg, E.A. (2015). Stepwise evolution of corolla symmetry in CYCLOIDEA2-like and RADIALIS-like gene expression patterns in Lamiales. American Journal of Botany **102,** 1260-1267.

Zhu, H., Riely, B.K., Burns, N.J., and Ané, J.-M. (2006). Tracing Nonlegume Orthologs of Legume Genes Required for Nodulation and Arbuscular Mycorrhizal Symbioses. Genetics **172,** 2491-2499.

Zimin, A.V., Marçais, G., Puiu, D., Roberts, M., Salzberg, S.L., and Yorke, J.A. (2013). The MaSuRCA genome assembler. Bioinformatics **29,** 2669-2677.

Zipfel, C., and Oldroyd, G.E.D. (2017). Plant signalling in symbiosis and immunity. Nature **543,** 328-336.

Zwickl, D.J., and Hillis, D.M. (2002). Increased taxon sampling greatly reduces phylogenetic error. Syst Biol **51**.

Appendix

Appendix

A1. List of genomes and transcriptomes used for the phylogenetic analyses conducted in the current study

Species	Dataset	Database	1kp Identifier	Phylogenetic position	Phylogenetic position - B	Phylogenetic position - C
Acrosiphonia sp	transcriptome	onekp.com/project.html	JIWJ	Green Algae	Chlorophytes	
Amborella trichopoda	genome	www.amborella.org/		Angiosperms		
Andreaea rupestris	transcriptome	onekp.com/project.html	WOGB	Mosses		
Ankistrodesmus sp	transcriptome	onekp.com/project.html	OTQG	Green Algae	Chlorophytes	
Anomodon attenuatus	transcriptome	onekp.com/project.html	QMWB	Mosses		
Aphanochaete repens	transcriptome	onekp.com/project.html	IJMT	Green Algae	Chlorophytes	
Arabidopsis thaliana	genome	TAIR		Angiosperms		
Asteromonas gracilis	transcriptome	onekp.com/project.html	ZFXU	Green Algae	Chlorophytes	
Asteromonas gracilis	transcriptome	onekp.com/project.html	MNPL	Green Algae	Chlorophytes	
Asteromonas gracilis	transcriptome	onekp.com/project.html	NTLE	Green Algae	Chlorophytes	
Atrichum angustatum	transcriptome	onekp.com/project.html	ZTHV	Mosses		
Aulacomnium heterostichum	transcriptome	onekp.com/project.html	WNGH	Mosses		
Bambusina borreri	transcriptome	onekp.com/project.html	QWFV	Green Algae	Charophytes	Zygnemophyceae
Barbilophozia barbata	transcriptome	onekp.com/project.html	OFTV	Liverworts		
Bathycoccus prasinos	transcriptome	onekp.com/project.html	MCPK	Green Algae	Chlorophytes	Prasinophyte
Bazzania trilobata	transcriptome	onekp.com/project.html	WZYK	Liverworts		
Blasia sp	transcriptome	onekp.com/project.html	AEXY	Liverworts		

Blastophysa cf.	transcriptome	onekp.com/project.html	VHIJ	Green Algae	Chlorophytes	
Bolbocoleon piliferum	transcriptome	onekp.com/project.html	LSHT	Green Algae	Chlorophytes	
Botryococcus braunii	transcriptome	onekp.com/project.html	ETGN	Green Algae	Chlorophytes	
Botryococcus sudeticus	transcriptome	onekp.com/project.html	VJDZ	Green Algae	Chlorophytes	
Botryococcus terribilis	transcriptome	onekp.com/project.html	QYXY	Green Algae	Chlorophytes	
Brachiomonas submarina	transcriptome	onekp.com/project.html	GUBD	Green Algae	Chlorophytes	
Bryopsis plumosa	transcriptome	onekp.com/project.html	JTIG	Green Algae	Chlorophytes	
Bryum argenteum	transcriptome	onekp.com/project.html	JMXW	Mosses		
Buxbaumia aphylla	transcriptome	onekp.com/project.html	HRWG	Mosses		
Calypogeia fissa	transcriptome	onekp.com/project.html	RTMU	Liverworts		
Carteria crucifera	transcriptome	onekp.com/project.html	VIAU	Green Algae	Chlorophytes	
Carteria obtusa	transcriptome	onekp.com/project.html	RUIF	Green Algae	Chlorophytes	
Cephaleuros virescens	transcriptome	onekp.com/project.html	YDCQ	Green Algae	Chlorophytes	
Ceratodon purpureus	transcriptome	onekp.com/project.html	FFPD	Mosses		
Chaetopeltis orbicularis	transcriptome	onekp.com/project.html	BAZF	Green Algae	Chlorophytes	
Chaetosphaeridium globosum	transcriptome	NCBI		Green Algae	Charophytes	Coleochaetales
Chaetosphaeridium globosum	transcriptome	onekp.com/project.html	DRGY	Green Algae	Charophytes	Coleochaetales
Chara vulgaris	transcriptome	onekp.com/project.html	MWXT	Green Algae	Charophytes	Charales
Chlamydomonas bilatus	transcriptome	onekp.com/project.html	SRGS	Green Algae	Chlorophytes	
Chlamydomonas bilatus	transcriptome	onekp.com/project.html	MULF	Green Algae	Chlorophytes	
Chlamydomonas bilatus	transcriptome	onekp.com/project.html	OVHR	Green Algae	Chlorophytes	
Chlamydomonas cribrum	transcriptome	onekp.com/project.html	BCYF	Green Algae	Chlorophytes	
Chlamydomonas moewusii	transcriptome	onekp.com/project.html	JRGZ	Green Algae	Chlorophytes	
Chlamydomonas noctigama	transcriptome	onekp.com/project.html	VALZ	Green Algae	Chlorophytes	
Chlamydomonas reinhardtii	genome	www.phytozome.net/		Green Algae	Chlorophytes	
Chlamydomonas sp.	transcriptome	onekp.com/project.html	XOZZ	Green Algae	Chlorophytes	

Chlamydomonas sp.	transcriptome	onekp.com/project.html	TSBQ	Green Algae	Chlorophytes	
Chlamydomonas sp.	transcriptome	onekp.com/project.html	AOUJ	Green Algae	Chlorophytes	
Chlorella minutissima	transcriptome	onekp.com/project.html	MWAN	Green Algae	Chlorophytes	
Chlorokybus atmophyticus	transcriptome	onekp.com/project.html	AZZW	Green Algae	Charophytes	Chlorokybales
Chlorokybus atmophyticus	transcriptome	NCBI		Green Algae	Charophytes	Chlorokybales
Chloromonas oogama	transcriptome	onekp.com/project.html	IHOI	Green Algae	Chlorophytes	
Chloromonas perforata	transcriptome	onekp.com/project.html	QRTH	Green Algae	Chlorophytes	
Chloromonas reticulata	transcriptome	onekp.com/project.html	ZLBP	Green Algae	Chlorophytes	
Chloromonas reticulata	transcriptome	onekp.com/project.html	LBRP	Green Algae	Chlorophytes	
Chloromonas reticulata	transcriptome	onekp.com/project.html	GDUD	Green Algae	Chlorophytes	
Chloromonas rosae	transcriptome	onekp.com/project.html	AJUW	Green Algae	Chlorophytes	
Chloromonas subdivisa	transcriptome	onekp.com/project.html	GFUR	Green Algae	Chlorophytes	
Chloromonas tughillensis	transcriptome	onekp.com/project.html	UTRE	Green Algae	Chlorophytes	
Chlorosarcinopsis halophila	transcriptome	onekp.com/project.html	KSFK	Green Algae	Chlorophytes	
Cladophora glomerata	transcriptome	onekp.com/project.html	VBLH	Green Algae	Chlorophytes	
Climacium dendroides	transcriptome	onekp.com/project.html	MIRS	Mosses		
Closterium lunula	transcriptome	onekp.com/project.html	DRFX	Green Algae	Charophytes	Zygnemophyceae
Closterium peracerosum-strigosum-littorale complex.	transcriptome	newly sequenced		Green Algae	Charophytes	Desmidiales
coccoid prasinophyte	transcriptome	onekp.com/project.html	XJGM	Green Algae	Chlorophytes	Prasinophytes
Coccomyxa pringsheimii	transcriptome	onekp.com/project.html	GXBM	Green Algae	Chlorophytes	
Coccomyxa subellipsoidea	genome	www.phytozome.net/		Green Algae	Chlorophytes	
Codium fragile	transcriptome	onekp.com/project.html	GYBH	Green Algae	Chlorophytes	
Coleochaete irregularis	transcriptome	onekp.com/project.html	QPDY	Green Algae	Charophytes	Coleochaetales
Coleochaete orbicularis	transcriptome	NCBI		Green Algae	Charophytes	Coleochaetales
Coleochaete scutata	transcriptome	onekp.com/project.html	VQBJ	Green Algae	Charophytes	Coleochaetales

Conocephalum conicum	transcriptome	onekp.com/project.html	ILBQ	Liverworts		
Cosmarium broomei	transcriptome	onekp.com/project.html	HIDG	Green Algae	Charophytes	Desmidiales
Cosmarium granatum	transcriptome	onekp.com/project.html	MNNM	Green Algae	Charophytes	Desmidiales
Cosmarium ochthodes	transcriptome	onekp.com/project.html	STKJ	Green Algae	Charophytes	Desmidiales
Cosmarium ochthodes	transcriptome	onekp.com/project.html	HJVM	Green Algae	Charophytes	Zygnematales
Cosmarium subtumidum	transcriptome	onekp.com/project.html	WDGV	Green Algae	Charophytes	Desmidiales
Cosmarium tinctum	transcriptome	onekp.com/project.html	BHBK	Green Algae	Charophytes	Desmidiales
Cosmocladium cf.	transcriptome	onekp.com/project.html	RQFE	Green Algae	Charophytes	Zygnemophyceae
Cylindrocapsa geminella	transcriptome	onekp.com/project.html	DZPJ	Green Algae	Chlorophytes	
Cylindrocystis brebissonii	transcriptome	onekp.com/project.html	YLBK	Green Algae	Charophytes	Zygnemophyceae
Cylindrocystis brebissonii	transcriptome	onekp.com/project.html	YOXI	Green Algae	Charophytes	Zygnemophyceae
Cylindrocystis brebissonii	transcriptome	onekp.com/project.html	RPGL	Green Algae	Charophytes	Zygnemophyceae
Cylindrocystis cushleckae	transcriptome	onekp.com/project.html	JOJQ	Green Algae	Charophytes	Zygnemophyceae
Cylindrocystis sp	transcriptome	onekp.com/project.html	VAZE	Green Algae	Charophytes	Zygnemophyceae
Cymbomonas sp	transcriptome	onekp.com/project.html	XIVI	Green Algae	Chlorophytes	Prasinophytes
Desmidium aptogonum	transcriptome	onekp.com/project.html	DFDS	Green Algae	Charophytes	Desmidiales
Dicranum scoparium	transcriptome	onekp.com/project.html	NGTD	Mosses		
Diphyscium foliosum	transcriptome	onekp.com/project.html	AWOI	Mosses		
Dolichomastix tenuilepis	transcriptome	onekp.com/project.html	XOAL	Green Algae	Chlorophytes	Prasinophytes
Dunaliella primolecta	transcriptome	onekp.com/project.html	WDWX	Green Algae	Chlorophytes	
Dunaliella salina	transcriptome	onekp.com/project.html	RHVC	Green Algae	Chlorophytes	
Dunaliella salina	transcriptome	onekp.com/project.html	NDPQ	Green Algae	Chlorophytes	
Dunaliella salina	transcriptome	onekp.com/project.html	SYJM	Green Algae	Chlorophytes	
Dunaliella salina	transcriptome	onekp.com/project.html	UKUC	Green Algae	Chlorophytes	
Dunaliella tertiolecta	transcriptome	onekp.com/project.html	ZDIZ	Green Algae	Chlorophytes	
Encalypta streptocarpa	transcriptome	onekp.com/project.html	KEFD	Mosses		

Entocladia endozoica	transcriptome	onekp.com/project.html	OQON	Green Algae	Chlorophytes	
Entransia fimbriata	transcriptome	onekp.com/project.html	BFIK	Green Algae	Charophytes	Klebsormidiales
Eremosphaera viridis	transcriptome	onekp.com/project.html	MNCB	Green Algae	Chlorophytes	
Euastrum affine	transcriptome	onekp.com/project.html	GYRP	Green Algae	Charophytes	Desmidiales
Eudorina elegans	transcriptome	onekp.com/project.html	RNAT	Green Algae	Chlorophytes	
Fontinalis antipyretica	transcriptome	onekp.com/project.html	DHWX	Mosses		
Fritschiella tuberosa	transcriptome	onekp.com/project.html	VFIV	Green Algae	Chlorophytes	
Frullania	transcriptome	onekp.com/project.html	TGKW	Liverworts		
Frullania spp.	transcriptome	onekp.com/project.html	CHJJ	Liverworts		
Geminella sp	transcriptome	onekp.com/project.html	PFUD	Green Algae	Chlorophytes	
Golenkinia longispicula	transcriptome	onekp.com/project.html	BZSH	Green Algae	Chlorophytes	
Gonatozygon kinahanii	transcriptome	onekp.com/project.html	KEYW	Green Algae	Charophytes	Desmidiales
Gonium pectorale	transcriptome	onekp.com/project.html	KUJU	Green Algae	Chlorophytes	
Haematococcus pluvialis	transcriptome	onekp.com/project.html	ODXI	Green Algae	Chlorophytes	
Haematococcus pluvialis	transcriptome	onekp.com/project.html	AGIO	Green Algae	Chlorophytes	
Haematococcus pluvialis	transcriptome	onekp.com/project.html	KFEB	Green Algae	Chlorophytes	
Hafniomonas reticulata	transcriptome	onekp.com/project.html	FXHG	Green Algae	Chlorophytes	
Halochlorococcum marinum	transcriptome	onekp.com/project.html	ALZF	Green Algae	Chlorophytes	
Hedwigia ciliata	transcriptome	onekp.com/project.html	YWNF	Mosses		
Helicodictyon planctonicum	transcriptome	onekp.com/project.html	AJAU	Green Algae	Chlorophytes	
Heterochlamydomonas inaequalis	transcriptome	onekp.com/project.html	IRYH	Green Algae	Chlorophytes	
Ignatius tetrasporus	transcriptome	onekp.com/project.html	KADG	Green Algae	Chlorophytes	
Interfilum paradoxum	transcriptome	onekp.com/project.html	FPCO	Green Algae	Charophytes	Klebsormidiales
Klebsormidium flaccidum	genome	www.plantmorphogenesis.bio.titech.ac.jp/~algae_genome_project/klebsormidium/index.html		Green Algae	Charophytes	Klebsormidiales

Klebsormidium flaccidum	transcriptome	NCBI		Green Algae	Charophytes	Klebsormidiales
Klebsormidium subtile	transcriptome	onekp.com/project.html	FQLP	Green Algae	Charophytes	Klebsormidiales
Leptosira obovata	transcriptome	onekp.com/project.html	ZNUM	Green Algae	Chlorophytes	
Leucobryum albidum	transcriptome	onekp.com/project.html	VMXJ	Mosses		
Leucobryum glaucum	transcriptome	onekp.com/project.html	RGKI	Mosses		
Leucodon brachypus	transcriptome	onekp.com/project.html	ZACW	Mosses		
Leucodon julaceus	transcriptome	onekp.com/project.html	IGUH	Mosses		
Lobochlamys segnis	transcriptome	onekp.com/project.html	OFUE	Green Algae	Chlorophytes	
Lobomonas rostrata	transcriptome	onekp.com/project.html	JKKI	Green Algae	Chlorophytes	
Loeskeobryum brevirostre	transcriptome	onekp.com/project.html	WSPM	Mosses		
Lunularia cruciata	transcriptome	onekp.com/project.html	TXVB	Liverworts		
Lunularia cruciata	transcriptome	newly sequenced		Liverworts		
Mantoniella squamata	transcriptome	onekp.com/project.html	QXSZ	Green Algae	Chlorophytes	Prasinophytes
Marchantia emarginata	transcriptome	onekp.com/project.html	TFYI	Liverworts		
Marchantia paleacea	transcriptome	onekp.com/project.html	HMHL	Liverworts		
Marchantia paleacea	transcriptome	onekp.com/project.html	IHWO	Liverworts		
Marchantia paleacea	transcriptome	onekp.com/project.html	LFVP	Liverworts		
Marchantia polymorpha	transcriptome	onekp.com/project.html	JPYU	Liverworts		
Medicago truncatula	genome	jcvi.org/medicago/		Angiosperms		
Mesostigma viride	transcriptome	NCBI		Green Algae	Charophytes	Mesostigmatales
Mesostigma viride	transcriptome	onekp.com/project.html	KYIO	Green Algae	Charophytes	Mesostigmatales
Mesotaenium braunii	transcriptome	onekp.com/project.html	WSJO	Green Algae	Charophytes	Zygnemophyceae
Mesotaenium caldariorum	transcriptome	onekp.com/project.html	HKZW	Green Algae	Charophytes	Zygnemophyceae
Mesotaenium endlicherianum	transcriptome	onekp.com/project.html	WDCW	Green Algae	Charophytes	Zygnemophyceae
Mesotaenium kramstae	transcriptome	onekp.com/project.html	NBYP	Green Algae	Charophytes	Desmidiales
Metzgeria crassipilis	transcriptome	onekp.com/project.html	NRWZ	Liverworts		

Micrasterias fimbriata	transcriptome	onekp.com/project.html	MCHJ	Green Algae	Charophytes	Zygnemophyceae
Micromonas pusilla	genome	www.phytozome.net/		Green Algae	Chlorophytes	Prasinophytes
Microspora cf.	transcriptome	onekp.com/project.html	FOYQ	Green Algae	Chlorophytes	
Microthamnion kuetzigianum	transcriptome	onekp.com/project.html	EATP	Green Algae	Chlorophytes	
Microthamnion kuetzingianum	transcriptome	onekp.com/project.html	DXNY	Green Algae	Chlorophytes	
Microthamnion kuetzingianum	transcriptome	onekp.com/project.html	ZDOF	Green Algae	Chlorophytes	
mixed species	transcriptome	onekp.com/project.html	AEKF	Liverworts		
mixed species	transcriptome	onekp.com/project.html	NWQC	Liverworts		
Monoclea gottschei	transcriptome	onekp.com/project.html	TFDQ	Liverworts		
Monomastix opisthostigma	transcriptome	onekp.com/project.html	BTFM	Green Algae	Chlorophytes	Prasinophytes
Mougeotia sp	transcriptome	onekp.com/project.html	ZRMT	Green Algae	Charophytes	Zygnemophyceae
Nannochloris atomus	transcriptome	onekp.com/project.html	MFYC	Green Algae	Chlorophytes	
Neochloris oleoabundans	transcriptome	onekp.com/project.html	EEJO	Green Algae	Chlorophytes	
Neochloris sp	transcriptome	onekp.com/project.html	GJIY	Green Algae	Chlorophytes	
Neochlorosarcina sp	transcriptome	onekp.com/project.html	USIX	Green Algae	Chlorophytes	
Nephroselmis olivacea	transcriptome	onekp.com/project.html	MMKU	Green Algae	Chlorophytes	Prasinophytes
Nephroselmis pyriformis	transcriptome	onekp.com/project.html	ISIM	Green Algae	Chlorophytes	Prasinophytes
Netrium digitus	transcriptome	onekp.com/project.html	FFGR	Green Algae	Charophytes	Zygnemophyceae
Nitella mirabilis	transcriptome	NCBI		Green Algae	Charophytes	Charales
Nothoceros aenigmaticus	transcriptome	onekp.com/project.html	DXOU	Hornworts		
Nothoceros vincentianus	transcriptome	onekp.com/project.html	TCBC	Hornworts		
Nucleotaenium eifelense	transcriptome	onekp.com/project.html	KMNX	Green Algae	Charophytes	Zygnemophyceae
Ochlochaete sp	transcriptome	onekp.com/project.html	CQQP	Green Algae	Chlorophytes	
Odontoschisma prostratum	transcriptome	onekp.com/project.html	YBQN	Liverworts		
Oedogonium cardiacum	transcriptome	onekp.com/project.html	DVYE	Green Algae	Chlorophytes	
Oedogonium foveolatum	transcriptome	onekp.com/project.html	SDPC	Green Algae	Chlorophytes	

Oltmannsiellopsis viridis	transcriptome	onekp.com/project.html	NSTT	Green Algae	Chlorophytes	
Oltmannsiellopsis viridis	transcriptome	onekp.com/project.html	PZBH	Green Algae	Chlorophytes	
Oltmannsiellopsis viridis	transcriptome	onekp.com/project.html	QJYX	Green Algae	Chlorophytes	
Onychonema laeve	transcriptome	onekp.com/project.html	GGWH	Green Algae	Charophytes	Desmidiales
Oogamochlamys gigantea	transcriptome	onekp.com/project.html	XDLL	Green Algae	Chlorophytes	
Orthotrichum lyellii	transcriptome	onekp.com/project.html	CMEQ	Mosses		
Oryza sativa	genome	www.phytozome.net		Angiosperms		
Ostreococcus lucimarinus	genome	www.phytozome.net/		Green Algae	Chlorophytes	Prasinophytes
Ostreococcus tauri	genome	www.phytozome.net/		Green Algae	Chlorophytes	Prasinophytes
Pallavicinia lyellii	transcriptome	onekp.com/project.html	YFGP	Liverworts		
Pandorina morum	transcriptome	onekp.com/project.html	RYJX	Green Algae	Chlorophytes	
Parachlorella kessleri	transcriptome	onekp.com/project.html	AKCR	Green Algae	Chlorophytes	
Pediastrum duplex	transcriptome	onekp.com/project.html	QGQH	Green Algae	Chlorophytes	
Pediastrum duplex	transcriptome	onekp.com/project.html	YOHC	Green Algae	Chlorophytes	
Pediastrum duplex	transcriptome	onekp.com/project.html	JRDV	Green Algae	Chlorophytes	
Pediastrum duplex	transcriptome	onekp.com/project.html	NLOM	Green Algae	Chlorophytes	
Pediastrum duplex	transcriptome	onekp.com/project.html	XKWQ	Green Algae	Chlorophytes	
Pediastrum duplex	transcriptome	onekp.com/project.html	XTON	Green Algae	Chlorophytes	
Pedinomonas minor	transcriptome	onekp.com/project.html	RRSV	Green Algae	Chlorophytes	
Pedinomonas tuberculata	transcriptome	onekp.com/project.html	PUAN	Green Algae	Chlorophytes	
Pellia cf. epiphylla	transcriptome	onekp.com/project.html	PIUF	Liverworts		
Pellia neesiana	transcriptome	onekp.com/project.html	JHFI	Liverworts		
Penium exiguum	transcriptome	onekp.com/project.html	YSQT	Green Algae	Charophytes	Desmidiales
Penium margaritaceum	transcriptome	NCBI		Green Algae	Charophytes	Desmidiales
Percursaria percursa	transcriptome	onekp.com/project.html	OAEZ	Green Algae	Chlorophytes	
Phacotus lenticularis	transcriptome	onekp.com/project.html	ZIVZ	Green Algae	Chlorophytes	

Philonotis fontana	transcriptome	onekp.com/project.html	ORKS	Mosses		
Phymatodocis nordstedtiana	transcriptome	onekp.com/project.html	RPQV	Green Algae	Charophytes	Desmidiales
Physcomitrella patens	genome	www.phytozome.net		Mosses		
Physcomitrium sp.	transcriptome	onekp.com/project.html	YEPO	Mosses		
Picocystis salinarum	transcriptome	onekp.com/project.html	TGNL	Green Algae	Chlorophytes	Prasinophytes
Pirula salina	transcriptome	onekp.com/project.html	NQYP	Green Algae	Chlorophytes	
Plagiomnium insigne	transcriptome	onekp.com/project.html	BGXB	Mosses		
Planophila laetevirens	transcriptome	onekp.com/project.html	CBNG	Green Algae	Chlorophytes	
Planophila terrestris	transcriptome	onekp.com/project.html	LETF	Green Algae	Chlorophytes	
Planotaenium ohtanii	transcriptome	onekp.com/project.html	SNOX	Green Algae	Charophytes	Zygnemophyceae
Pleurastrum insigne	transcriptome	onekp.com/project.html	PRIQ	Green Algae	Chlorophytes	
Pleurotaenium trabecula	transcriptome	onekp.com/project.html	MOYY	Green Algae	Charophytes	Desmidiales
Polytrichum commune	transcriptome	onekp.com/project.html	SZYG	Mosses		
Populus trichocarpa	genome	www.phytozome.net		Angiosperms		
Porella navicularis	transcriptome	onekp.com/project.html	KRUQ	Liverworts		
Porella pinnata	transcriptome	onekp.com/project.html	UUHD	Liverworts		
Prasinococcus capsulatus	transcriptome	onekp.com/project.html	XMCL	Green Algae	Chlorophytes	Prasinophytes
Prasinoderma coloniale	transcriptome	onekp.com/project.html	HYHN	Green Algae	Chlorophytes	Prasinophytes
Prasiola crispa	transcriptome	onekp.com/project.html	WCLV	Green Algae	Chlorophytes	
Prototheca wickerhamii	transcriptome	onekp.com/project.html	BILC	Green Algae	Chlorophytes	
Pseudoscourfieldia marina	transcriptome	onekp.com/project.html	JMTE	Green Algae	Chlorophytes	Prasinophytes
Pseudotaxiphyllum elegans	transcriptome	onekp.com/project.html	QKQO	Mosses		
Pteromonas angulosa	transcriptome	onekp.com/project.html	LNIL	Green Algae	Chlorophytes	
Pteromonas sp	transcriptome	onekp.com/project.html	ACRY	Green Algae	Chlorophytes	
Ptilidium pulcherrimum	transcriptome	onekp.com/project.html	HPXA	Liverworts		
Pycnococcus provasolii	transcriptome	onekp.com/project.html	MXEZ	Green Algae	Chlorophytes	Prasinophytes

Pyramimonas parkeae	transcriptome	onekp.com/project.html	TNAW	Green Algae	Chlorophytes	Prasinophytes
Racomitrium elongatum	transcriptome	onekp.com/project.html	ABCD	Mosses		
Radula lindenbergiana	transcriptome	onekp.com/project.html	BNCU	Liverworts		
Rhynchostegium serrulatum	transcriptome	onekp.com/project.html	JADL	Mosses		
Ricciocarpos natans	transcriptome	onekp.com/project.html	WJLO	Liverworts		
Rosulabryum cf. capillare	transcriptome	onekp.com/project.html	XWHK	Mosses		
Roya obtusa	transcriptome	onekp.com/project.html	XRTZ	Green Algae	Charophytes	Zygnemophyceae
Scapania nemorosa	transcriptome	onekp.com/project.html	IRBN	Liverworts		
Scenedesmus dimorphus	transcriptome	onekp.com/project.html	PZIF	Green Algae	Chlorophytes	
Scherffelia dubia	transcriptome	onekp.com/project.html	FMVB	Green Algae	Chlorophytes	
Schistochila sp	transcriptome	onekp.com/project.html	LGOW	Liverworts		
Scourfieldia sp	transcriptome	onekp.com/project.html	EGNB	Green Algae	Chlorophytes	
Selaginella moellendorffii	genome	www.phytozome.net		Lycophytes		
Spermatozopsis exsultans	transcriptome	onekp.com/project.html	MXDS	Green Algae	Chlorophytes	
Spermatozopsis similis	transcriptome	onekp.com/project.html	ENAU	Green Algae	Chlorophytes	
Sphaerocarpos texanus	transcriptome	onekp.com/project.html	HERT	Liverworts		
Sphagnum lescurii	transcriptome	onekp.com/project.html	GOWD	Mosses		
Sphagnum palustre	transcriptome	onekp.com/project.html	RCBT	Mosses		
Sphagnum recurvum	transcriptome	onekp.com/project.html	UHLI	Mosses		
Spirogyra sp	transcriptome	onekp.com/project.html	HAOX	Green Algae	Charophytes	Zygnemophyceae
Spirogyra sp.	genome			Green Algae	Charophytes	Zygnemophyceae
Spirogyra sp.	transcriptome	NCBI		Green Algae	Charophytes	Zygnemophyceae
Spirotaenia minuta	transcriptome	onekp.com/project.html	NNHQ	Green Algae	Charophytes	Zygnemophyceae
Spirotaenia sp	transcriptome	onekp.com/project.html	TPHT	Green Algae	Charophytes	Zygnemophyceae
Staurastrum sebaldi	transcriptome	onekp.com/project.html	ISHC	Green Algae	Charophytes	Desmidiales
Staurodesmus convergens	transcriptome	onekp.com/project.html	WCQU	Green Algae	Charophytes	Desmidiales

Staurodesmus omearii	transcriptome	onekp.com/project.html	RPRU	Green Algae	Charophytes	Desmidiales
Stephanosphaera pluvialis	transcriptome	onekp.com/project.html	ZLQE	Green Algae	Chlorophytes	
Stereodon subimponens	transcriptome	onekp.com/project.html	LNSF	Mosses		
Stichococcus bacillaris	transcriptome	onekp.com/project.html	WXRI	Green Algae	Chlorophytes	
Stigeoclonium helveticum	transcriptome	onekp.com/project.html	JMUI	Green Algae	Chlorophytes	
Syntrichia princeps	transcriptome	onekp.com/project.html	GRKU	Mosses		
Takakia lepidozioides	transcriptome	onekp.com/project.html	SKQD	Mosses		
Tetraselmis chui	transcriptome	onekp.com/project.html	HVNO	Green Algae	Chlorophytes	
Tetraselmis cordiformis	transcriptome	onekp.com/project.html	DUMA	Green Algae	Chlorophytes	
Tetraselmis striata	transcriptome	onekp.com/project.html	HHXJ	Green Algae	Chlorophytes	
Thuidium delicatulum	transcriptome	onekp.com/project.html	EEMJ	Mosses		
Timmia austriaca	transcriptome	onekp.com/project.html	ZQRI	Mosses		
Trebouxia arboricola	transcriptome	onekp.com/project.html	NKXU	Green Algae	Chlorophytes	
Trentepohlia annulata	transcriptome	onekp.com/project.html	NATT	Green Algae	Chlorophytes	
Treubia lacunosa	transcriptome	onekp.com/project.html	FITN	Liverworts		
Uronema belkae	transcriptome	onekp.com/project.html	RAWF	Green Algae	Chlorophytes	
Uronema sp.	transcriptome	onekp.com/project.html	ISGT	Green Algae	Chlorophytes	
Vitreochlamys sp	transcriptome	onekp.com/project.html	QWRA	Green Algae	Chlorophytes	
Volvox aureus	transcriptome	onekp.com/project.html	POIR	Green Algae	Chlorophytes	
Volvox aureus	transcriptome	onekp.com/project.html	JWGT	Green Algae	Chlorophytes	
Volvox aureus	transcriptome	onekp.com/project.html	WRSL	Green Algae	Chlorophytes	
Volvox carteri	genome	www.phytozome.net/		Green Algae	Chlorophytes	
Volvox globator	transcriptome	onekp.com/project.html	ISPU	Green Algae	Chlorophytes	
Xanthidium antilopaeum	transcriptome	onekp.com/project.html	GBGT	Green Algae	Charophytes	Desmidiales

For the analyses conducted in Section 2.3.4.1, all mmetsp (https://imicrobe.us/project/view/104), 1kp transcriptomes (onekp.com/project) and ENSEMBL protist genomes were used
For the analyses conducted in Section 2.3.4.2, all 1kp (onekp.com/project) plant transcriptomes were used
For Chapter 4, previously sequenced genomes were used from Phytozome; details of the genomes and transcriptomes assembled for use in Chapter 4, are described in Chapter 6.

A2. UNIX scripts used for phylogenetic analysis

Script for BLAST, alignment and initial tree building

```
#!/bin/bash

#Phylogenetics pipeline

#Typical usage of the pipeline:
./blast_align_tree.sh query_fasta_file_name working_folder_name
no_of_blast_hits_required

working_folder=$2
#set the default value of working_folder to phylogenetics
if [ -d $working_folder ]; then
cd $working_folder
#echo "Make sure you have kept the query in the./query/ folder and type the name and
extension of the FASTA file (eg: query.fa) containing the protein sequences that you
would like to find homologs for, followed by [ENTER]:"
#first argument to shell script is the query
# put your query file with one or more FASTA sequences in the query folder and put
the name here (the file must be saved like something.faa)
no_results=${3:-400}
no_blast=2000 #Please change this value to reflect how many blast results you would
like from each database that you want to find orthologs in
query=$1
query_folder=`basename $query.faa`

mkdir -p $working_folder/query_list/$query_folder
mkdir -p $working_folder/alignment/$query_folder
mkdir -p $working_folder/tree/$query_folder
mkdir -p
$working_folder/results_from_search/$query_folder/fasta_hits/all_databases_concaten
ated
mkdir -p
$working_folder/results_from_search/$query_folder/dnas/fasta_hits/all_databases_con
catenated
mkdir -p $working_folder/nexus/$query_folder

if [ -f $working_folder/query/$query ]; then

        echo "Your file $query was found"
        fromdos $working_folder/query/$query
        #search for all > and add fasta identifiers to new file (for reference)
        grep '>' $working_folder/query/$query >
$working_folder/query_list/$query_folder/$query.list.faa
```

```
	#add line number to the list - this line number would correspond to number of the
sequence in $query_$no
	cat -n $working_folder/query_list/$query_folder/$query.list.faa >
$working_folder/query_list/$query_folder/$query.list.faa.numbered
	mkdir -p $working_folder/temp/$query_folder
	cd $working_folder/temp/$query_folder

	#split sequences to separate numbers
	awk '/^>/{s=++d".faa"} {print > s}' $working_folder/query/$query
	#change file name to fasta header
	for x in $working_folder/temp/$query_folder/*.faa; do mv $x `awk 'sub(/^>/, "")'
$x`; done
	#add.faa extension
	for x in $working_folder/temp/$query_folder/*; do mv $x $x.faa; done
	#remove special characters from file name
	find. -print | while read file; do file_clean=${file//[ ()&\'|"\,]/_}; mv "$file"
"$file_clean"; done
	#add original query name to the sequences
	for x in $working_folder/temp/$query_folder/*.faa; do mv $x "$query"_"`basename
$x`"; done

	cd $working_folder

	for e in $working_folder/temp/$query_folder/*.faa;
	do
		for f in $working_folder/blast_dbs/*.fasta;
		do
			blastp -query $e -out
$working_folder/results_from_search/$query_folder/`basename $e.faa`.x.`basename $f
.faa`.blastout -db $working_folder/blast_dbs/`basename $f.fasta` -outfmt "6 sseqid" -
max_target_seqs $no_blast -num_threads 16;
			uniq
$working_folder/results_from_search/$query_folder/`basename $e.faa`.x.`basename $f
.faa`.blastout | head -n $no_results >
$working_folder/results_from_search/$query_folder/`basename $e.faa`.x.`basename
$f.faa`.blastout.uniq

		done
		for g in $working_folder/blast_dbs/*.fasta;
		do
			blastdbcmd -db $working_folder/blast_dbs/`basename
$g.fasta` -dbtype prot -entry_batch
$working_folder/results_from_search/$query_folder/`basename $e .faa`.x.`basename
$g.faa`.blastout.uniq -outfmt %f -out
$working_folder/results_from_search/$query_folder/fasta_hits/`basename
$e.faa`.x.`basename $g.faa`.hits.faa

		done

			#blast done!
```

```
                        #alignment time!

                cat
$working_folder/results_from_search/$query_folder/fasta_hits/`basename $e.faa`.x.*
>>
$working_folder/results_from_search/$query_folder/fasta_hits/all_databases_concaten
ated/`basename $e .faa`.all_databases_concatenated.faa

                #cat
$working_folder/results_from_search/$query_folder/dnas/fasta_hits/`basename
$e.faa`.x.* >>
$working_folder/results_from_search/$query_folder/dnas/fasta_hits/all_databases_con
catenated/`basename $e .faa`.all_databases_concatenated.faa

                sed -i "s/'>'/'>query_'/g" $e
                cat $e >>
$working_folder/results_from_search/$query_folder/fasta_hits/all_databases_concaten
ated/`basename $e.faa`.all_databases_concatenated.faa
                #mothur
"#unique.seqs(fasta=$working_folder/results_from_search/$query_folder/fasta_hits/all_
databases_concatenated/`basename $e.faa`.all_databases_concatenated.faa)"
                #translate FASTA_headers to shorten the fasta headers
                awk '/^>/{$0=$0"_"(++i)}1'
$working_folder/results_from_search/$query_folder/fasta_hits/all_databases_concaten
ated/`basename $e.faa`.all_databases_concatenated.faa >
$working_folder/results_from_search/$query_folder/fasta_hits/all_databases_concaten
ated/`basename $e.faa`.all_databases_concatenated.unique.faa
                translate_fasta_headers --
out=$working_folder/results_from_search/$query_folder/fasta_hits/all_databases_conc
atenated/`basename $e.faa`.all_databases_concatenated.shortened.faa
$working_folder/results_from_search/$query_folder/fasta_hits/all_databases_concaten
ated/`basename $e .faa`.all_databases_concatenated.unique.faa
                #remove spaces from original FASTA headers in the tab file
                tr ' ' '_' <
$working_folder/results_from_search/$query_folder/fasta_hits/all_databases_concaten
ated/`basename $e.faa`.all_databases_concatenated.shortened.faa.translation.tab >
$working_folder/results_from_search/$query_folder/fasta_hits/all_databases_concaten
ated/`basename $e
.faa`.all_databases_concatenated.shortened.faa.translation.tab.spacerem
                #remove punctuation marks in the tab file
                tr [:punct:] '_' <
$working_folder/results_from_search/$query_folder/fasta_hits/all_databases_concaten
ated/`basename
$e.faa`.all_databases_concatenated.shortened.faa.translation.tab.spacerem >
$working_folder/results_from_search/$query_folder/fasta_hits/all_databases_concaten
ated/`basename $e
.faa`.all_databases_concatenated.shortened.faa.translation.tab.final;
        done
        #alignment using aligner
```

```
	for c in
$working_folder/results_from_search/$query_folder/fasta_hits/all_databases_concaten
ated/*.shortened.faa;
	do
		mafft --auto --thread 16 $c >
$working_folder/alignment/$query_folder/`basename $c
results_from_search/fasta_hits/all_databases_concatenated/`.kalign.fas

	done


	#construct maximum-likelihood tree using FastTreeMP
	for m in $working_folder/alignment/$query_folder/*.fas;
	do



		#Make tree using FastTree from alignment
		FastTreeMP -pseudo -spr 4 -mlacc 2 -slownni $m >
$working_folder/tree/$query_folder/`basename $m`.fastree;

	done
	for a in $working_folder/alignment/$query_folder/*.fas;
	do
	#rename fasta headers in alignment according to original fasta header of blast
results
	taxnameconvert.pl
$working_folder/results_from_search/$query_folder/fasta_hits/all_databases_concaten
ated/`basename $a.kalign.fas`.translation.tab.final $a $a.final.fas
	#taxnameconvert.pl
$working_folder/results_from_search/$query_folder/fasta_hits/all_databases_concaten
ated/`basename $a.prank.fas`.translation.tab.final $a $a.final.fas;
	NCLconverter -faafasta -enexus $a.final.fas -o$a.final.fas;
	#seqret -sformat1 fasta -osformat2 nexus -sequence -outseq $a.final.fas.nexus;
	done

	for t in $working_folder/tree/$query_folder/*.fastree;
	do
		#rename taxa in tree according to original fasta header of blast results
		taxnameconvert.pl
$working_folder/results_from_search/$query_folder/fasta_hits/all_databases_concaten
ated/`basename $t.kalign.fas.fastree`.translation.tab.final $t $t.final.tree
		NCLconverter -frelaxedphyliptree -enexus $t.final.tree -o$t.final.tree;
		sed -i -e '1,6d' $t.final.tree.nex
		#taxnameconvert.pl
$working_folder/results_from_search/$query_folder/fasta_hits/all_databases_concaten
ated/`basename $t.prank.fas.fastree`.translation.tab.final $t $t.final.tree;
	done

	cp $working_folder/alignment/$query_folder/*.nex
$working_folder/nexus/$query_folder/
```

```
        cp $working_folder/tree/$query_folder/*.nex
$working_folder/nexus/$query_folder/

        for x in $working_folder/nexus/$query_folder/*fas.nex;
        do
                cat $x $working_folder/nexus/$query_folder/`basename
$x.final.fas.nex`.fastree.final.tree.nex > $x.geneious.nexus;
        done

else
        echo 'Your query file was not found, please check the file name and the folder
that you are in. Make sure you have kept the query in the./query/ folder and type the
name and extension of the FASTA file (eg: query.fa) containing the protein sequences
that you would like to find homologs for'
fi

else
        echo 'Your working_folder does not exist; please make sure you either pass a
working_folder argument to the script or run the command mkdir -p
~/phylogenetics/query/'

fi
```

Script for hmmsearch, alignment and initial tree building

```
#!/bin/bash

#Phylogenetics pipeline

#Typical usage of the pipeline:
./hmmer_align_tree.sh query_fasta_file_name working_folder_name
no_of_hits_required

working_folder=$2
#set the default value of working_folder to phylogenetics
if [ -d $working_folder ]; then
cd $working_folder
#echo "Make sure you have kept the query in the./query/ folder and type the name and
extension of the FASTA file (eg: query.fa) containing the protein sequences that you
would like to find homologs for, followed by [ENTER]:"
#first argument to shell script is the query
# put your query file with one or more FASTA sequences in the query folder and put
the name here (the file must be saved like something.faa)
no_results=${3:-400}
no_blast=2000 #Please change this value to reflect how many blast results you would
like from each database that you want to find orthologs in
query=$1
query_folder=`basename $query.faa`

mkdir -p $working_folder/query_list/$query_folder
mkdir -p $working_folder/alignment/$query_folder
mkdir -p $working_folder/tree/$query_folder
mkdir -p
$working_folder/results_from_search/$query_folder/fasta_hits/all_databases_concaten
ated
mkdir -p
$working_folder/results_from_search/$query_folder/dnas/fasta_hits/all_databases_con
catenated
mkdir -p $working_folder/nexus/$query_folder

if [ -f $working_folder/query/$query ]; then

        echo "Your file $query was found"
        fromdos $working_folder/query/$query
        #search for all > and add fasta identifiers to new file (for reference)
        grep '>' $working_folder/query/$query >
$working_folder/query_list/$query_folder/$query.list.faa
        #add line number to the list - this line number would correspond to number of the
sequence in $query_$no
        cat -n $working_folder/query_list/$query_folder/$query.list.faa >
$working_folder/query_list/$query_folder/$query.list.faa.numbered
        mkdir -p $working_folder/temp/$query_folder
        cd $working_folder/temp/$query_folder
```

```
	#split sequences to separate numbers
	awk '/^>/{s=++d".faa"} {print > s}' $working_folder/query/$query
	#change file name to fasta header
	for x in $working_folder/temp/$query_folder/*.faa; do mv $x `awk 'sub(/^>/, "")'
$x`; done
	#add.faa extension
	for x in $working_folder/temp/$query_folder/*; do mv $x $x.faa; done
	#remove special characters from file name
	find. -print | while read file; do file_clean=${file//[ ()&\'|"\,]/_}; mv "$file"
"$file_clean"; done
	#add original query name to the sequences
	for x in $working_folder/temp/$query_folder/*.faa; do mv $x "$query"_"`basename
$x`"; done

	cd $working_folder

	for e in $working_folder/temp/$query_folder/*.faa;
	do
		for f in $working_folder/blast_dbs/*.fasta;
		do
			#blastp -query $e -out
$working_folder/results_from_search/$query_folder/`basename $e.faa`.x.`basename $f
.faa`.blastout -db $working_folder/blast_dbs/`basename $f.fasta` -outfmt "6 sseqid" -
max_target_seqs $no_blast -num_threads 16;
			jackhmmer --tblout
$working_folder/results_from_search/$query_folder/`basename $e.faa`.x.`basename $f
.faa`.tblout --cpu 16 $e $f
			grep -v "^#"
$working_folder/results_from_search/$query_folder/`basename $e.faa`.x.`basename $f
.faa`.tblout | awk '{print $1}' >>
$working_folder/results_from_search/$query_folder/`basename $e.faa`.x.`basename
$f.faa`.blastout
			uniq
$working_folder/results_from_search/$query_folder/`basename $e.faa`.x.`basename $f
.faa`.blastout | head -n $no_results >
$working_folder/results_from_search/$query_folder/`basename $e.faa`.x.`basename
$f.faa`.blastout.uniq

		done
		#change the -max_target_seqs from 50 if desired :)
		for g in $working_folder/blast_dbs/*.fasta;
		do
			blastdbcmd -db $working_folder/blast_dbs/`basename
$g.fasta` -dbtype prot -entry_batch
$working_folder/results_from_search/$query_folder/`basename $e .faa`.x.`basename
$g.faa`.blastout.uniq -outfmt %f -out
$working_folder/results_from_search/$query_folder/fasta_hits/`basename
$e.faa`.x.`basename $g.faa`.hits.faa
```

```
                done

                cat
$working_folder/results_from_search/$query_folder/fasta_hits/`basename $e.faa`.x.*
>>
$working_folder/results_from_search/$query_folder/fasta_hits/all_databases_concaten
ated/`basename $e .faa`.all_databases_concatenated.faa


                sed -i "s/'>'/'>query_'/g" $e
                cat $e >>
$working_folder/results_from_search/$query_folder/fasta_hits/all_databases_concaten
ated/`basename $e.faa`.all_databases_concatenated.faa
                #translate FASTA_headers to shorten the fasta headers
                awk '/^>/{$0=$0"_"(++i)}1'
$working_folder/results_from_search/$query_folder/fasta_hits/all_databases_concaten
ated/`basename $e.faa`.all_databases_concatenated.faa >
$working_folder/results_from_search/$query_folder/fasta_hits/all_databases_concaten
ated/`basename $e.faa`.all_databases_concatenated.unique.faa
                translate_fasta_headers --
out=$working_folder/results_from_search/$query_folder/fasta_hits/all_databases_conc
atenated/`basename $e.faa`.all_databases_concatenated.shortened.faa
$working_folder/results_from_search/$query_folder/fasta_hits/all_databases_concaten
ated/`basename $e .faa`.all_databases_concatenated.unique.faa
                #remove spaces from original FASTA headers in the tab file
                tr ' ' '_' <
$working_folder/results_from_search/$query_folder/fasta_hits/all_databases_concaten
ated/`basename $e.faa`.all_databases_concatenated.shortened.faa.translation.tab >
$working_folder/results_from_search/$query_folder/fasta_hits/all_databases_concaten
ated/`basename $e
.faa`.all_databases_concatenated.shortened.faa.translation.tab.spacerem
                #remove punctuation marks in the tab file
                tr [:punct:] '_' <
$working_folder/results_from_search/$query_folder/fasta_hits/all_databases_concaten
ated/`basename
$e.faa`.all_databases_concatenated.shortened.faa.translation.tab.spacerem >
$working_folder/results_from_search/$query_folder/fasta_hits/all_databases_concaten
ated/`basename $e
.faa`.all_databases_concatenated.shortened.faa.translation.tab.final;
        done
        #alignment using aligner
        for c in
$working_folder/results_from_search/$query_folder/fasta_hits/all_databases_concaten
ated/*.shortened.faa;
        do
                mafft --auto --thread 16 $c >
$working_folder/alignment/$query_folder/`basename $c
results_from_search/fasta_hits/all_databases_concatenated/`.kalign.fas
        done
```

```
	#construct maximum-likelihood tree using FastTreeMP
	for m in $working_folder/alignment/$query_folder/*.fas;
	do
		#Make tree using FastTree from alignment
		FastTreeMP -pseudo -spr 4 -mlacc 2 -slownni $m >
$working_folder/tree/$query_folder/`basename $m`.fastree;

	done
	for a in $working_folder/alignment/$query_folder/*.fas;
	do
	#rename fasta headers in alignment according to original fasta header of blast
results
	taxnameconvert.pl
$working_folder/results_from_search/$query_folder/fasta_hits/all_databases_concaten
ated/`basename $a.kalign.fas`.translation.tab.final $a $a.final.fas
	NCLconverter -faafasta -enexus $a.final.fas -o$a.final.fas;
	done

	for t in $working_folder/tree/$query_folder/*.fastree;
	do
		#rename taxa in tree according to original fasta header of blast results
		taxnameconvert.pl
$working_folder/results_from_search/$query_folder/fasta_hits/all_databases_concaten
ated/`basename $t.kalign.fas.fastree`.translation.tab.final $t $t.final.tree
		NCLconverter -frelaxedphyliptree -enexus $t.final.tree -o$t.final.tree;
		sed -i -e '1,6d' $t.final.tree.nex
	done

	cp $working_folder/alignment/$query_folder/*.nex
$working_folder/nexus/$query_folder/
	cp $working_folder/tree/$query_folder/*.nex
$working_folder/nexus/$query_folder/

	for x in $working_folder/nexus/$query_folder/*fas.nex;
	do
		cat $x $working_folder/nexus/$query_folder/`basename
$x.final.fas.nex`.fastree.final.tree.nex > $x.geneious.nexus;
	done



else
	echo 'Your query file was not found, please check the file name and the folder
that you are in. Make sure you have kept the query in the./query/ folder and type the
name and extension of the FASTA file (eg: query.fa) containing the protein sequences
that you would like to find homologs for'
fi

else
```

```
    echo 'Your working_folder does not exist; please make sure you either pass a
working_folder argument to the script or run the command mkdir -p
~/phylogenetics/query/'

fi
```

Script for hmmscan and extracting sequences containing specific domains

```
#!/bin/bash
#Script to do hmmscan of proteins from species and extract the pfam annotation for all
the predicted proteins
#This will be followed by getting the sequences of containing specific domains from
these species
query=$2
hmmscan --cpu 16 -o output/hmmscan_results/$query.hmmscan --domtblout
output/hmmscan_results/$query.hmmscan.domtbl --cut_ga Pfam-A.hmm
search_dbs/$query
family=$1
hmmscan_out=$query.hmmscan.domtbl
hmmscan_query=$query
grep -i $family output/hmmscan_results/$hmmscan_out | awk '{print $4}' >
output/final_family_list_of_hits/$family.$hmmscan_out.list
uniq output/final_family_list_of_hits/$family.$hmmscan_out.list
output/final_family_list_of_hits/$family.$hmmscan_out.list.uniq
esl-sfetch --index search_dbs/$hmmscan_query
esl-sfetch -f search_dbs/$hmmscan_query
output/final_family_list_of_hits/$family.$hmmscan_out.list.uniq >
output/final_family_fasta/$family.`basename $hmmscan_query`
```

Script for building final trees using RAxML

```
#!/bin/bash
input=$1
mkdir -p raxml/input/
mkdir -p raxml/ml_search/
mkdir -p raxml/bootstrap/
mkdir -p raxml/bipartition/
sed -i 's/(//g' $input
sed -i 's/)//g' $input
cp $input raxml/input/
cd raxml/input/
raxmlHPC-PTHREADS -p 12345 -m PROTGAMMAAUTO -# 20 -s $input -n
ml_search.`basename $input` -T 16
raxmlHPC-PTHREADS -p 12345 -b 12345 -m PROTGAMMAAUTO -# 100 -s $input -n
bootstrap.`basename $input` -T 16
raxmlHPC-PTHREADS -m PROTGAMMAAUTO -p 12345 -f b -t
RAxML_bestTree.ml_search.`basename $input` -z
RAxML_bootstrap.bootstrap.`basename $input` -n bipartition.`basename $input` -T 16
```

```
mv *ml_search.`basename $input`../ml_search/
mv *bootstrap.`basename $input`../bootstrap/
mv *bipartition.`basename $input`../bipartition/
```