

Investigation of Early-Diverging Land Plant Metabolism



Naomi Cannell
Department of Plant Sciences
University of Oxford

A thesis submitted for the degree of
Doctor of Philosophy

Hilary Term 2021

Acknowledgements

I would first like to thank Lee, without whom this thesis would not have materialised. I am so grateful for your patient guidance and encouragement, the collaborative and nurturing lab environment that you have fostered, the limitless metabolic knowledge with which you are so generous and of course for the provision of drinks at lab meetings. To have such a supportive and enthusiastic supervisor has been a joy and a privilege.

The Plant Sciences department is filled with many further generous and inspiring people without whom the successful completion of this thesis would have been difficult. Thank you to Liam, my second supervisor, for your energising support and your expertise on all things bryophyte. To Sanu, thank you for your endless modelling advice and for your patience in helping me investigating coding problems. Thank you also Nick for your advice throughout and for working so hard to keep all graduate students progressing over the past year. Finally, huge thanks to the other members of the Sweetlove Group past and present for keeping me sane in the lab and office (and more recently, from our respective homes) and for our regular celebrations of the small wins. I'll miss working with such a dynamic group of friends.

To my amazing friends and siblings, thank you for reminding me that not everything in life revolves around plant metabolism (although arguably you are wrong). To my wonderful parents, thank you for nurturing my passion for biology, and for allowing four years' worth of phone calls to always revolve around plant metabolism! And to Michael, thank you always for putting up with me, for supporting me relentlessly and always offering your eternally optimistic perspective, and for making sure I had a long commute during which to mull over my project.

Much appreciation goes to the Systems Biology Doctoral Training Programme at the University of Oxford, and the EPSRC for providing the funding for this project.

Abstract

The terrestrialisation of embryophytes was preceded by, and precipitated, the evolution of diverse metabolic capabilities to aid growth and survival on land. A broad understanding of metabolism across all land plants is important for the discovery of novel metabolic enzymes and pathways which may be used commercially and to design metabolic engineering strategies for crop improvement. However, metabolic research has mainly focused on a small number of angiosperm crops and model plants. The research described in this thesis was carried out with the goal of correcting this knowledge bias, using bioinformatic and computational techniques to characterise metabolism, and its associated evolutionary processes, in early-diverging land plants.

The work is presented as three manuscripts. In the first I used metabolic modelling to investigate central metabolic fluxes in the liverwort *Marchantia polymorpha*, which exhibits extremely low carbon assimilation rates as a consequence of the low mesophyll conductance characteristic of bryophytes. I found that my constrained stoichiometric model of *Marchantia* was unable to achieve experimentally-defined growth rates without increasing carbon conversion efficiency by requiring either a very low cell maintenance cost, or some form of nocturnal carbon recycling. In the second manuscript I carried out a comparative analysis of secondary metabolism across the Chloroplastida, using homologous relationships between metabolic genes in 72 land plants and algae. The findings, supported by gene absences/presences in a further 305 plant and algal transcriptomes, elucidate details of the evolution of several phytohormone and biopolymer biosynthetic pathways. Finally, I used phylogenetic profiling and co-expression analysis to identify candidate genes associated with the previously discussed biosynthetic pathways. The results were validated by the identification of several known gene associations, and additional genes with various potential functional relationships with the pathways in question were detected.

Contents

List of Figures	xi
List of Abbreviations	xiii
1 Introduction	1
1.1 The evolution of land plants	2
1.1.1 Overview of land plant evolution	2
1.1.2 The evolution of central plant metabolism	7
1.1.3 The evolution of secondary metabolism in plants	11
1.2 Bioinformatic and computational investigation of plant metabolism	16
1.2.1 Investigating plant metabolism is a necessity and a challenge	16
1.2.2 Bioinformatic and computational methods in plant metabolism	17
1.2.3 Flux-balance analysis of genome-scale models	22
1.3 Project rationale and objectives	24
2 Constraint-based modelling reveals a high carbon conversion efficiency in <i>Marchantia polymorpha</i> compared to that in C_3 angiosperms	35
2.1 Motivation	35
2.2 Contribution	35
2.3 Summary	36
2.4 Introduction	36
2.5 Methods	38
2.5.1 Constructing a stoichiometric model of primary metabolism in <i>Marchantia</i>	38
2.5.2 Constructing biomass reactions for <i>Arabidopsis</i> and <i>Marchantia</i>	39
2.5.3 Estimating non-growth-associated maintenance	39
2.5.4 Setting up diel models of <i>Arabidopsis</i> and <i>Marchantia</i>	40
2.5.5 Validating model predictions against experimental data	40
2.5.6 Investigating reaction fluxes	42
2.5.7 Data availability	42
2.6 Results	42
2.6.1 Model framework and characteristics	42

2.6.2	Experimental data validates our model of <i>Arabidopsis</i> metabolism, but <i>Marchantia</i> growth rates cannot be successfully met using equivalent constraints within the <i>Marchantia</i> metabolic model	47
2.6.3	Removing constraints on night-phase carbon fixation allows the <i>Marchantia</i> model to meet experimental growth rates	55
2.6.4	Comparison of flux modes between <i>Arabidopsis</i> and <i>Marchantia</i> models suggests CAM-like carbon re-fixation strategy in <i>Marchantia</i>	57
2.7	Discussion	61
3	Multiple metabolic innovations and losses are associated with major transitions in land plant evolution	71
3.1	Motivation	71
3.2	Contribution	71
3.3	Summary	72
3.4	Introduction	73
3.5	Methods	74
3.5.1	Experimental model and subject details	75
3.5.2	Production of metabolic pathway databases	75
3.5.3	Analysis of metabolic pathways	76
3.5.4	Identification of metabolic transitions	77
3.5.5	Transcriptome analysis	79
3.6	Results	79
3.6.1	Bioinformatic approach and justification	79
3.6.2	<i>De novo</i> genome annotation provides gene-to-reaction associations for previously unannotated metabolic genes	82
3.6.3	Inferred metabolic pathways capture known phylogenetic relationships	82
3.6.4	Comparative analysis identifies previously identified metabolic innovations and losses in land plant evolution	83
3.6.5	Biosynthetic capabilities for gibberellin and brassinosteroid production are encoded in the genomes of non-angiosperm plants . . .	87
3.6.6	Non-angiosperm plants lack genes encoding known enzymes for gibberellin and brassinosteroid inactivation	91
3.6.7	Bryophytes, lycophytes, ferns, and gymnosperms lack specific cytochrome P450 family enzymes involved in oxidation of structural monomers	92
3.6.8	Conserved enzymes for production of secondary monomers found in suberin and sporopollenin are limited to angiosperms and land plants, respectively	97

3.6.9	Testing metabolic gains on a broader selection of species using 1KP transcriptomes	99
3.7	Discussion	102
3.8	Limitations of the approach	105
4	Random forest analysis of phylogenetically limited gene sets facilitates identification of candidate genes associated with phytohormone and biopolymer biosynthesis	113
4.1	Motivation	113
4.2	Contribution	113
4.3	Summary	114
4.4	Introduction	115
4.5	Methods	118
4.5.1	Establishing phylogenetic and gene expression profiles for each metabolic pathway	118
4.5.2	Identifying candidate genes using Pearson's correlation coefficient .	119
4.5.3	Identifying candidate genes using random forests	119
4.6	Results	120
4.6.1	Bioinformatic approach	120
4.6.2	Gibberellin biosynthesis-associated genes	129
4.6.3	Brassinosteroid biosynthesis-associated genes	133
4.6.4	Sporopollenin biosynthesis-associated genes	136
4.6.5	Cutin and suberin biosynthesis-associated genes	139
4.7	Discussion	142
4.8	Limitations of the approach	144
5	Discussion	155
5.1	Summary of the main findings presented in this thesis	155
5.2	Central metabolism is conserved across land plants, but flux distributions are not necessarily conserved across the central metabolic network	156
5.3	Conservation of the presence of secondary metabolic pathways does not necessarily translate to functional conservation	159
5.4	Extensive experimental work is required to improve our knowledge of early-diverging land plants	160
5.5	Computational and bioinformatic metabolic investigations suffer from an inherent bias towards angiosperms	162
5.6	Further limitations are imposed by bioinformatic methods in metabolic research	164
5.7	Additional computational and experimental work can elucidate further details of early-diverging land plant metabolism	166

Bibliography	171
Appendices	
A Appendix A: Supplementary figures and tables for Chapter 2	197
B Appendix B: Supplementary figures and tables for Chapter 3	203
C Appendix C: Publication of Chapter 3	215

List of Figures

1.1	A schematic depicting the evolution of the Chloroplastida and land plant-specific evolutionary innovations and losses.	3
1.2	Illustrating the differences between C ₃ , C ₄ and CAM photosynthesis. . . .	12
1.3	Flux-balance analysis.	23
2.1	Basic schematic of <i>Arabidopsis</i> and <i>Marchantia</i> stoichiometric models. . . .	43
2.2	Comparing predicted growth rates as a proportion of experimental growth rates between the <i>Arabidopsis</i> and <i>Marchantia</i> models.	50
2.3	Parallel coordinates plot indicating the results of a sensitivity analysis on five constraints on our <i>Marchantia</i> model at a PPFD of 95 μ mol/m ² /s. . . .	54
2.4	Comparison of growth rate predictions from the <i>Marchantia</i> model under both sets of experimental conditions.	57
2.5	Major flux routes through <i>Arabidopsis</i> and <i>Marchantia</i> models with solutions able to meet experimental growth rate predictions.	59
3.1	Overview of metabolic gains and losses identified by the orthogroup analysis.	84
3.2	Presence of gene homologues encoding glucosinolate biosynthesis enzymes across 72 plant and algal species.	86
3.3	Presence of gene homologues encoding gibberellin biosynthesis and inactivation enzymes across 72 plant and algal species.	88
3.4	Presence of gene homologues encoding brassinosteroid biosynthesis and inactivation enzymes across 72 plant and algal species.	89
3.5	Presence of gene homologues encoding cutin and suberin biosynthesis enzymes across 72 plant and algal species.	94
3.6	Presence of gene homologues encoding sporopollenin biosynthesis enzymes across 72 plant and algal species.	95
4.1	Flowchart representing the bioinformatic pipeline developed to identify candidate genes functionally associated with individual metabolic pathways.	122
4.2	Simplified metabolic pathway diagrams representing the biosynthetic pathways for gibberellins, brassinosteroids, cutin, suberin and sporopollenin . .	123
4.3	Phylogenetic profiles for each metabolic pathway.	124

4.4	Transcript abundances of genes with Pearson's correlation coefficients higher than that for 99% of randomly selected genes for each metabolic pathway analysis.	127
4.5	Illustrating the results of random forest analyses for each metabolic pathway.	128
A.1	Results of fitting curves of the form $A = \frac{x \times PPFD}{y + PPFD}$ to carbon assimilation data for <i>Arabidopsis</i> and <i>Marchantia</i>	198
A.2	Illustrating differences in central carbon metabolism between the <i>Marchantia</i> model and the <i>Arabidopsis</i> model.	199
A.3	Flux differences between the <i>Marchantia</i> model and the <i>Arabidopsis</i> model stemming from different requirements for pectin as a component of biomass.	200
A.4	Flux differences between the <i>Marchantia</i> model and the <i>Arabidopsis</i> model stemming from differing NADH requirements in the chloroplast at night. .	201
B.1	Overview of methodology for the identification of metabolic innovations and losses.	204
B.2	Phylogenetic tree showing evolutionary relationships between all 72 algal and plant species involved in the orthogroup analysis.	205
B.3	Number and type of novel EC-number annotations made using E2P2 for seven species.	206
B.4	Results of <i>k</i> -medioids clustering and t-SNE analysis of representative species.	207
B.5	Presence of gene homologues encoding diacylglycerol-N,N,N-trimethylhomoserine biosynthesis enzyme across 72 land plant and algal species.	208
B.6	Presence of gene homologues encoding selenocysteine biosynthesis enzymes across 72 land plant and algal species.	209
B.7	Simplified gene tree detailing the relationships between <i>Arabidopsis</i> genes in the orthogroup containing a known brassinosteroid hydroxylation enzyme.	210
B.8	Simplified gene tree detailing the relationships between <i>Arabidopsis</i> genes in the orthogroup containing CYP86 and CYP704 enzymes.	211

List of Abbreviations

1KP		One Thousand Plant Transcriptomes Initiative
2OG		2-oxoglutarate
2PG		2-phosphoglycolate
ABA		Absciscic acid
ACH		Acyl-CoA thioesterase
ADP		Adenosine diphosphate
ATP		Adenosine triphosphate
bHLH		Basic helix-loop-helix
BR		Brassinosteroid
BTA		Betaine lipid synthase
CAM		Crassulacean acid metabolism
CBB cycle		Calvin Benson Bassham cycle
ChIP-seq		Chromatin immunoprecipitation sequencing
CIS-ACON		<i>cis</i> -aconitate
CoA		Coenzyme A
CYP		Cytochrome P450
DET2		5- α steroid dehydrogenase
DGTS		Diacylglycerol-N,N,N-trimethylhomoserine
DNA		Deoxyribonucleic acid
DUF		Domain of unknown function
E2P2		Ensemble Enzyme Prediction Pipeline
EC Number		Enzyme Commission Number
FAR		Fatty acid reductase
FBA		Flux-balance analysis
FRU		Fructose

FVA	Flux variability analysis
G3P	3-phosphoglycerate
GA	Gibberellin
GABA	γ -aminobutyric acid
GAMT	Gibberellin methyltransferase
GAox	Gibberellin oxidase
GAP	Glyceraldehyde-3-phosphate
GGP	GGP γ -glutamyl peptidase
GGT	γ -glutamyl transpeptidase
GLC	Glucose
GLN	Glutamine
GLT	Glutamic acid
GO	Gene ontology
GPAT	Glycerol-3-phosphate acyltransferase
GS-OX	Glucosinolate S-oxygenase
HET	Heteroglycan
ID	Identifier
ISOCIT	Isocitrate
JA	Jasmonic acid
KAO	Kaurenic acid oxidase
KCS	3-ketoacyl CoA synthase
KO	Kaurene oxidase
LACS	Long-chain acyl-CoA synthetase
LEU	Leucine
MAFFT	Multiple alignment using fast fourier transform
MAL	Malate
MCL	Markov clustering
MET	Methionine
mTHF	Methyl tetrahydrofolate
NAD(P)-ME . . .	NAD(P) malic enzyme
NAD(P)H	Nicotinamide adenine dinucleotide (phosphate)

NGAM	Non growth-associated maintenance
OAA	Oxaloacetate
PEP	Phosphoenolpyruvate
PEPC	Phosphoenolpyruvate carboxylase
PEPCK	Phosphoenolpyruvate carboxykinase
pFBA	parsimonious FBA
PGDB	Pathway/Genome Database
PKS	Polyketide synthase
PPFD	Photosynthetic photon flux density
PS	Photosystem
PYR	Pyruvate
RNA	Ribonucleic acid
RNAseq	RNA sequencing
ROS	Reactive oxygen species
RT-PCR	Reverse transcription polymerase chain reaction
RuBisCO	Ribulose biphosphate carboxylase/oxygenase
RuBP	Ribulose biphosphate
SBML	Systems biology markup language
SERS	seryl-tRNA synthetase
SL	Strigolactone
SOT	Sulfotransferase
SPCS	Selenocysteine synthetase
SPS	Selenophosphate synthetase
SUC	Succinate
TCA cycle	Tri-carboxylic acid cycle
THF	Tetrahydrofolate
TKPR	Tetreketide pyrone reductase
t-SNE	t-distributed stochastic neighbour embedding
UDP	uracil diphosphate
UGT	UDP-glucosyltransferase
UV	Ultraviolet
VAL	Valine

1

Introduction

The colonisation of land and subsequent adaptive radiation of embryophytes were accompanied by the development of an array of novel metabolic capabilities required for the growth and development of land plants (Delwiche and Cooper 2015; de Vries and Archibald 2018; Rensing 2018). While central metabolism is highly conserved among land plants, metabolic profiling experiments indicate huge diversity in secondary metabolic pathways across individual species and taxa (Hartmann 1996; Wink 2008). However, the challenge of experimentally investigating metabolism means that we lack a complete picture of metabolism in any species, and what we do know is heavily skewed towards well-studied angiosperm species (Rensing 2017). The overall aim of the research described in this thesis was to computationally characterise and investigate metabolism in early-diverging land plants, using genome sequence information and metabolic model construction. This will contribute to the study of the evolution of metabolism as well as to the improvement of our understanding of metabolic processes in early-diverging land plants, which may help to direct the discovery of novel metabolic pathways, enzymatic functions and compounds.

1.1 THE EVOLUTION OF LAND PLANTS

1.1.1 Overview of land plant evolution

The earliest members of the Chloroplastida, one of three lineages derived from an Archaeplastid ancestor which engulfed the free-living cyanobacterium that would eventually form the chloroplast, evolved more than a billion years ago (bya) (Delwiche and Cooper 2015). This lineage further diverged to produce the chlorophytes and streptophytes (Yoon et al. 2004), the latter giving rise to the embryophytes (land plants) about 470 million years ago (mya) (Figure 1.1). While it is unclear whether chlorophytes originally evolved in marine or freshwater environments (Dittami et al. 2017), the subsequent evolution of land plants from a freshwater algal ancestor is undisputed (de Vries et al. 2016; Rensing 2018). The conditions that would have facilitated this progression are thought to have been sporadically hydrated environments such as the shorelines of lakes, or pools which exist seasonally as a consequence of rainfall patterns. The transient nature of these aquatic environments is thought to have generated a selection pressure for the development of traits to protect against exposure and desiccation (Delwiche and Cooper 2015) and to have allowed stepwise progression to a terrestrial lifestyle. Given the ubiquitous nature of such transitional hydrological gradients in freshwater environments, it is perhaps unsurprising that terrestrial species have evolved independently within various clades of streptophyte algae (de Vries and Archibald 2018; McCourt et al. 2004). Nonetheless, only one such lineage gave rise to the full diversity of extant terrestrial macroflora (de Vries and Archibald 2018) – and the precise set of environmental conditions and ancestral characteristics which precipitated this unique evolutionary radiation are still a source of discussion among plant evolutionary biologists (Delwiche and Cooper 2015; de Vries and Archibald 2018).

From the perspective of a freshwater alga, the terrestrial environment provides a wealth of opportunity. Atmospheric CO₂ concentration is about 50 times higher than that of CO₂ dissolved in water, and dry land offers unfiltered sunlight. This combination of factors allows photosynthesis to proceed significantly faster than is possible underwater, but presents corresponding limitations. In addition to higher visible light flux, terrestrial plants are exposed to damaging ultraviolet (UV) radiation, and faster photosynthesis increases the

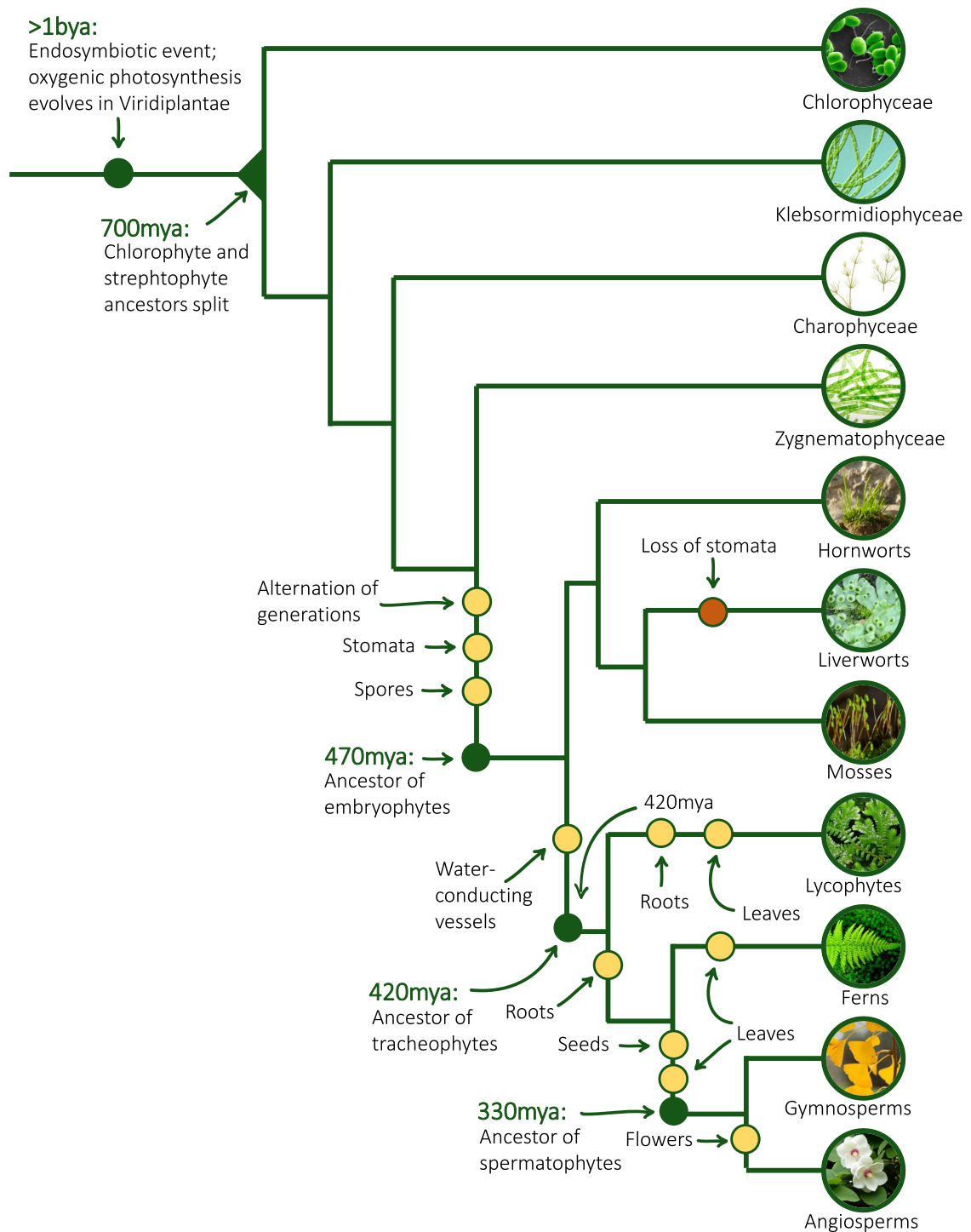


Figure 1.1: A schematic depicting the evolution of the Chloroplastida and land plant-specific evolutionary innovations and losses. Green points indicate known timings, yellow points indicate evolutionary innovations and orange points indicate evolutionary losses. mya, million years ago; bya, billion years ago.

potential for photooxidative damage as well as higher flux through photorespiration, which produces reactive oxygen species (ROS). Dry land also introduces greater daily and seasonal variations in ambient temperature which at both extremes may lead to drought stress (de Vries and Archibald 2018). The algal ancestor of embryophytes must have possessed physiological characteristics which acted as precursors for the development of terrestrial adaptations to overcome these challenges. Within the streptophyte algae, it is generally accepted that the Zygnematophyceae are the closest relatives of the land plant lineage (Ruhfel et al. 2014; Wickett et al. 2014) (Figure 1.1); as such, comparative analysis of extant zygnematophytes and land plants may provide clues to the characteristics of the ancestral embryophyte. The identification of the Zygnematophyceae as sister to all land plants was initially surprising, given the relative simplicity of zygnematophyte body plans compared to those of earlier-diverging streptophyte algae. However, this is now considered to be the result of a secondary reduction of more complex characteristics in the zygnematophyte ancestor (de Vries et al. 2016). It is possible that in itself, this reduction of complexity was involved in facilitating the transition to land – a smaller organism is likely to require smaller amounts of water to be functional, and may be able to complete its life cycle faster, which would have been adaptive in a habitat where water was only seasonally available (Delwiche and Cooper 2015). Further exaptations are likely to have involved modulation of the interaction between nucleus and plastid via the loss or transfer to the nucleus of previously plastid-encoded genes controlling gene expression and plastid division – the arrangement of several such genes are shared between zygnematophytes and embryophytes and differ from earlier-diverging algae (de Vries et al. 2016). As well as preventing mutations (as a consequence of lower irradiation reaching the nucleus) and mitigating errors in plastid-encoded gene duplication (via recombination), this would have conferred control of plastid division and production of proteins destined for the plastid to the nucleus, a characteristic thought to be associated with polyplastidy and the diversity of plastid forms present in embryophytes (de Vries et al. 2016). There are also metabolic features of streptophyte algae which are expected to have contributed to the feasibility of terrestrialisation, including the existence of protective biopolymers and phytohormones. These are discussed in detail in Section 1.1.3. While the function of many of these metabolites in algae is unclear, it has become evident that a molecular and

genetic toolbox containing many of the innovations necessary for the colonisation of land was present well before the emergence of the ancestral embryophyte.

The earliest-diverging clade of land plants are the bryophytes, comprising mosses, liverworts and hornworts. Almost every possible phylogenetic topology has been proposed for this group, but recent evidence strongly suggests that the bryophytes form a monophyletic taxon, with the liverworts and mosses as a sister-group to the earlier-diverging hornworts (Harris et al. 2020) (Figure 1.1). This is significant because it indicates that contrary to previous suggestions of liverworts representing an intermediate evolutionary stage between algae and other land plants, the common ancestor of all embryophytes was relatively complex and present-day liverworts reflect a loss of ancestral complexity (Harris et al. 2020; Puttick et al. 2018). One example of this loss of complexity relates to stomata, which are found on the sporophyte of all land plants except liverworts. The stomata of other bryophytes are less complex than those of angiosperms and do not respond to changes in atmospheric CO₂ concentrations. However, the genes regulating stomatal development are homologous across land plants (Merced and Renzaglia 2017). It is therefore thought that the ancestral embryophyte had already evolved relatively complex stomata over which it would have had active control, and that the stomata present in extant hornworts and mosses are the result of reductive evolution. Liverworts have entirely lost the ability to develop stomata and possess only simple air pores, which evolved *de novo* following this loss (Harris et al. 2020).

Bryophytes are the earliest-diverging plants in the streptophyte lineage to exhibit a haplodiplontic lifecycle (Rensing 2018), characterised by the alternation between a diploid sporophytic stage (beginning with the diploid zygote following fertilisation) and a haploid gametophytic stage. This is thought to have evolved via delayed zygotic meiosis to form a multicellular sporophyte before meiotic production of spores (Sakakibara et al. 2013). The alternation of haploid and diploid generations is conserved across land plants, however the dominant generation has transferred over the course of land plant evolution from the gametophyte (dominant in bryophytes) to the sporophyte (dominant in vascular plants) (Bowman et al. 2016). This process was facilitated by the co-option of gametophyte-specific patterns of gene expression and regulation by the sporophyte (Szovenyi et al. 2011) – for example the core signalling network responsible for the development of rhizoids in the moss

gametophyte and root hairs in the *Arabidopsis* sporophyte is conserved (Tam et al. 2015). Ferns represent an intermediate stage in this process, where both the gametophyte and sporophyte stage form independent, free-living organisms (Sigel et al. 2018).

Following the initial transition to land, complexity within the land plants arose rapidly. By 420mya, the common ancestor of modern tracheophytes had evolved (Morris et al. 2018), precipitating an explosion in land plant diversity. Presumably driven by competition for light, the evolution of water-conducting vessels allowed these plants to grow tall while maintaining water and nutrient delivery to all tissues. The development of such vascular systems was necessarily preceded by the evolution of biopolymers for structural support of the vessels themselves as well as the taller plant body plan. Lignin and suberin, responsible for reinforcement of above- and below-ground vessels in extant tracheophytes, are thought to have evolved in a step-wise manner from the pre-existing biosynthetic pathways for phenylpropanoids and cutin respectively (Niklas et al. 2017). It was the presence of vascular systems, as well as roots to provide consistent access to water even when ground surface was dry, which allowed the evolution of the first trees, lycophytes up to 50m tall, in the Carboniferous period (Hetherington et al. 2016). Interestingly, rooting systems themselves appear to have evolved at least twice independently – once in the common ancestor of lycophytes and once in the common ancestor of their sister group, the euphyllophytes – and were not present in the ancestral tracheophyte (Hetherington and Dolan 2018). Leaves – the defining feature of the euphyllophytes (the monophyletic clade containing extant ferns, gymnosperms and angiosperms) – are similarly thought to have evolved at least twice – once in the lycophytes and several times in the euphyllophytes (Harrison and Morris 2018; Tomescu 2009). The next major innovation in land plants was the evolution of seeds some time before 330mya in the common ancestor of the spermatophytes (gymnosperms and angiosperms) (Morris et al. 2018). Although analogous to the spores produced by earlier-diverging groups of plants, seeds differ in their multicellularity and contain a nutrient-containing endosperm which nourishes the developing embryo. In the earlier-diverging gymnosperms, seeds are found exposed on cones and within a gymnosperm seed the embryo is surrounded by both a maternal, nutrient-containing megagametophyte and the seed coat. Angiosperms are additionally characterised by their ability to produce an inflorescence, a trait which had

evolved by 150mya and provided angiosperms with enhanced seed protection as well as faster fertilisation via attracted pollinators (Chanderbali et al. 2016) and to which the vast diversity of this clade is often attributed.

Given the singularity of land plant evolution, a combination of genetic, morphological and metabolic exaptations were likely required in the embryophyte ancestor to facilitate the initial transition to a terrestrial environment. Once made, this transition provided huge evolutionary opportunity. The innovations we see in extant land plants and the fossil record (roots, leaves, seeds and flowers) appeared in stages and sometimes multiple times, each allowing the occupancy of additional ecological niches. The subsequent explosive adaptive radiation represented by the embryophyte lineage had enormous impact on terrestrial geology and ecosystems, and ultimately paved the way for the evolution of terrestrial forms in the animal kingdom.

1.1.2 The evolution of central plant metabolism

Oxygenic photosynthesis as carried out by cyanobacteria, algae and plants (the Cyanophyta) evolved only once (Fischer et al. 2016). This happened in the common ancestor of the Cyanophyta at some point before 2.35bya, at which point the consequential Great Oxidation Event occurred, modifying the Earth's atmosphere with the stable incorporation of gaseous oxygen and the subsequent formation of an ozone layer (Ligrone 2019). Oxygenic photosynthesis therefore had profound impacts on global geology, biochemistry and biodiversity, providing the evolutionary driver for the development of aerobic respiration, which would later support the evolution of complex life forms (Ligrone 2019).

The process of oxygenic photosynthesis comprises both light-independent reactions (the Calvin-Benson-Bassham (CBB) cycle) and light-dependent reactions (photosynthesis reactions). The light-independent CBB cycle is responsible for the assimilation of inorganic carbon (CO_2) into organic triose phosphates using the enzyme ribulose biphosphate carboxylase/oxygenase (RuBisCO), which carboxylates ribulose-1,5-bisphosphate (RuBP) to produce 3-phosphoglycerate (G3P). If O_2 is present, RuBisCO is also capable of a corresponding oxygenase reaction, combining O_2 with RuBP to produce 2-phosphoglycolate (2PG). The CBB cycle is considerably more ancient than the light-dependent reactions of oxygenic

photosynthesis, and RuBisCO is thought to have evolved in an anoxic environment where its carboxylation activity was unconstrained by atmospheric O₂ (Erb and Zarzycki 2018). However, the development of oxygenic photosynthesis in single celled organisms placed RuBisCO in the vicinity of O₂-generating enzymes, and the subsequent oxygenation of Earth's atmosphere further increased ambient O₂ concentrations. This drove evolution of increased RuBisCO carboxylase specificity at the cost of its catalytic rate (despite what is considered near-optimal positioning within this trade-off (Tcherkez et al. 2006), oxygenation of RuBP by RuBisCO is still significant) as well as a carbon salvage pathway known as photorespiration (Bauwe et al. 2012), which recycles 2PG into G3P for re-entry into the CBB cycle at considerable energetic cost.

The light-dependent photosynthetic reactions require two integral membrane reaction centres. Photosystem II (PSII) comprises a type II reaction centre in addition to a protein-chlorophyll-based light harvesting complex. When a photon is absorbed by a chlorophyll molecule, an electron is excited which, in the special pair of chlorophylls in the reaction centre is transferred to a quinone electron acceptor due to a charge separation event. The original electron lost from chlorophyll in PSII is replaced with an electron generated by the splitting of two water molecules to produce oxygen, four electrons and four protons. The electrons from the reduced quinone are passed through a cytochrome b6f complex which, in turn, transfers the electrons to plastocyanin while pumping protons across the membrane to form an electrochemical gradient which is later used to power the production of ATP via ATP synthase. Simultaneously, an excitation event occurs at photosystem I (PSI), which comprises a type I reaction centre in combination with its own light harvesting complex. This results in the reduction of a ferredoxin molecule from which the electrons are used to reduce NADP⁺ to NADPH, which provides the reducing power required for the fixation of CO₂ in the CBB cycle. The electrons removed from chlorophyll in PSI are replaced by electrons from the reduced plastocyanin resulting from PSII.

These coupled reaction centres responsible for the oxygenic photosynthesis present in cyanobacteria, algae and plants are only found individually in other organisms. The two forms of reaction centre are highly divergent, and in non-cyanobacterial organisms which possess one they are used for anoxygenic photosynthesis, relying on electron donors such as Fe²⁺

instead of water. There are various theories concerning the evolution of their coupling in the ancestor of all Cyanophyta (Soo et al. 2019). It has been suggested either that both reaction centres were gained by a non-photosynthetic ancestor from separate bacteria via horizontal gene transfer (Mathis 1990), that the possession of both reaction centres is ancestral and all bacterial lineages now containing only one were subject to the evolutionary loss of the other (Olson and Pierson 1987), or finally that both reactions centres evolved within the cyanobacteria from a single ancestral reaction centre, and the resulting specialised versions were later acquired individually by other bacterial lineages via horizontal gene transfer (Sousa et al. 2013). Although there is no definitive consensus on exactly how or when oxygenic photosynthesis evolved, the significant evolutionary distance between type I and type II reaction centres suggest that it was much later than the origin of anoxygenic photosynthesis.

Following the evolution of photosynthesis, a critical step in the evolution of photosynthetic organisms was the primary endosymbiosis which preceded the evolution of chloroplasts. This event, in which a single-celled eukaryote already containing a mitochondrion engulfed a photosynthetic cyanobacterium, is thought to have occurred between 1.5 and 1bya (Maréchal 2018). This cyanobacterial endosymbiosis was not a singularity in the evolution of eukaryotes – indeed, there is some uncertainty as to whether the chloroplasts identified across the Archaeplastida are the result of a single event or up to three independent engulfments (Rockwell et al. 2014). Additionally, multiple subsequent secondary endosymbioses (the engulfment of an already photosynthetic eukaryote by another eukaryote) are thought to be responsible for much algal diversity (Gavelis and Gile 2018). Modern chloroplasts retain cyanobacterial characteristics (for example the use of the cytochrome b6 complex (Hohmann-Marriott and Blankenship 2011)) but have a dramatically reduced genome, the consequence of gene transfer to the nuclear genome. 18% of nuclear genes in *Arabidopsis* are thought to be of cyanobacterial origin (Martin et al. 2002) and complicated signalling pathways have evolved to co-ordinate the expression of nuclear and chloroplast genes (Woodson and Chory 2008). Evidencing the complete integration of host and organelle, chloroplasts are responsible for various processes in plants in addition to carrying out photosynthesis, for example chlorophyll biosynthesis and the sensing and signalling of oxidative stress, resulting in the regulation of expression of some nuclear genes (Chan et al. 2016; Jensen and Leister 2014).

Photosynthetic processes have continued to evolve following the primary endosymbiosis that gave rise to chloroplasts. For example, the light harvesting pigments found across photosynthetic organisms are diverse but share metabolic reactions in the early stages of their biosynthetic pathways, suggesting a shared origin (Tomitani et al. 1999). Of particular interest are the varied carbon concentrating mechanisms that have evolved in algae and land plants, driven by the beneficial energetic savings available to an organism capable of decreasing flux through photorespiration. The majority of land plants and algae exhibit C_3 photosynthesis, whereby CO_2 is taken up during the day and directly assimilated into sugars via the CBB cycle, catalysed by RuBisCO and limited by RuBisCO's innate substrate promiscuity (Figure 1.2). However, distributed across the Chloroplastida there are examples of evolutionary innovations which function to increase the local concentration of CO_2 in the vicinity of RuBisCO, thereby increasing flux through the CBB cycle relative to photorespiration (Raven et al. 2008). Many chlorophyte and streptophyte algae, as well as hornworts, contain carbon concentrating mechanisms which act within single cells. In these species, specialised intracellular compartments known as pyrenoids containing extremely high levels of sequestered RuBisCO are present within chloroplasts (Villarreal and Renner 2012). Carbon transporters and the enzyme carbonic anhydrase, which catalyses the reversible conversion of CO_2 to HCO_3^- , work together to ensure a CO_2 concentration within pyrenoids up to 50 times higher than elsewhere in the cell (Goudet et al. 2020; Li et al. 2017). Later-diverging land plants have developed alternative mechanisms that require the coordination of multiple cell types. C_4 photosynthesis relies on the separation of reactions involved in carbon fixation into neighbouring cell types (Figure 1.2). Gas exchange with the environment occurs within mesophyll cells, which capture CO_2 by carboxylation of phosphoenolpyruvate to malate catalysed by the enzyme phosphoenolpyruvate carboxylase (PEPC). Malate is then transported to an inner layer of bundle sheath cells containing RuBisCO where it is decarboxylated to form a CO_2 -rich environment. Crassulacean acid metabolism (CAM) photosynthesis, which has evolved as an adaptation to minimise water loss in species inhabiting ecological niches where water is limited (Bone et al. 2015), similarly separates carboxylation reactions albeit temporally rather than spatially (Figure 1.2). Gas exchange is limited during the day by stomatal closure. When stomata are opened at night,

carbon is assimilated into malate using PEPC and stored in the vacuole. During the day, the stored malate is released and decarboxylated resulting in trapped concentrations of CO₂ up to 60 times higher than atmospheric levels (Keeley and Rundel 2003). The increase in efficiency that such mechanisms deliver is profound, and is illustrated by the multiple independent origins of these adaptations across plant lineages. Pyrenoids are considered to have evolved many times independently in algae (Meyer and Griffiths 2013) and at least five times in hornworts (Villarreal and Renner 2012), while C₄ photosynthesis is found in at least 62 independent angiosperm lineages (Sage et al. 2011). CAM photosynthesis is thought to have evolved independently at least 60 times in at least 35 tracheophyte families, and the identification of several species with intermediate C₃/CAM photosynthesis indicates that a continuum of such phenotypes exists (Bräutigam et al. 2017; Winter et al. 2015).

1.1.3 The evolution of secondary metabolism in plants

Aside from the highly-conserved metabolites synthesised by almost all land plants in the process of producing biomass and maintaining core metabolic functioning (i.e. involved in primary metabolism), there are upwards of one million additional metabolites identified across the plant kingdom (Kusano et al. 2019; Pichersky and Lewinsohn 2011). The production of these molecules, which are not absolutely essential for survival (although their loss may be significantly detrimental) falls under the umbrella of 'secondary metabolism' and include phytohormones such as gibberellins (GAs), auxins and brassinosteroids (BRs), biopolymers such as lignin, cutin, suberin and sporopollenin, and protective and defence compounds such as anthocyanins, glucosinolates and terpenes.

Many of these compounds are present in chlorophyte and streptophyte algae – and in fact are considered crucial exaptations which facilitated the transition to land. Key to desiccation resistance and stability of a multicellular form in land plants are structural biopolymers. Compounds identifiable as sporopollenin, the biopolymer responsible for reinforcement of the outer coating of seeds and spores in land plants, have been found in algae across the Chlorophyta (Delwiche et al. 1989; He et al. 2016) and molecules similar to lignins are present in some streptophytes (Delwiche et al. 1989). In land plants these biopolymers both comprise phenolic constituents which, along with similar molecules

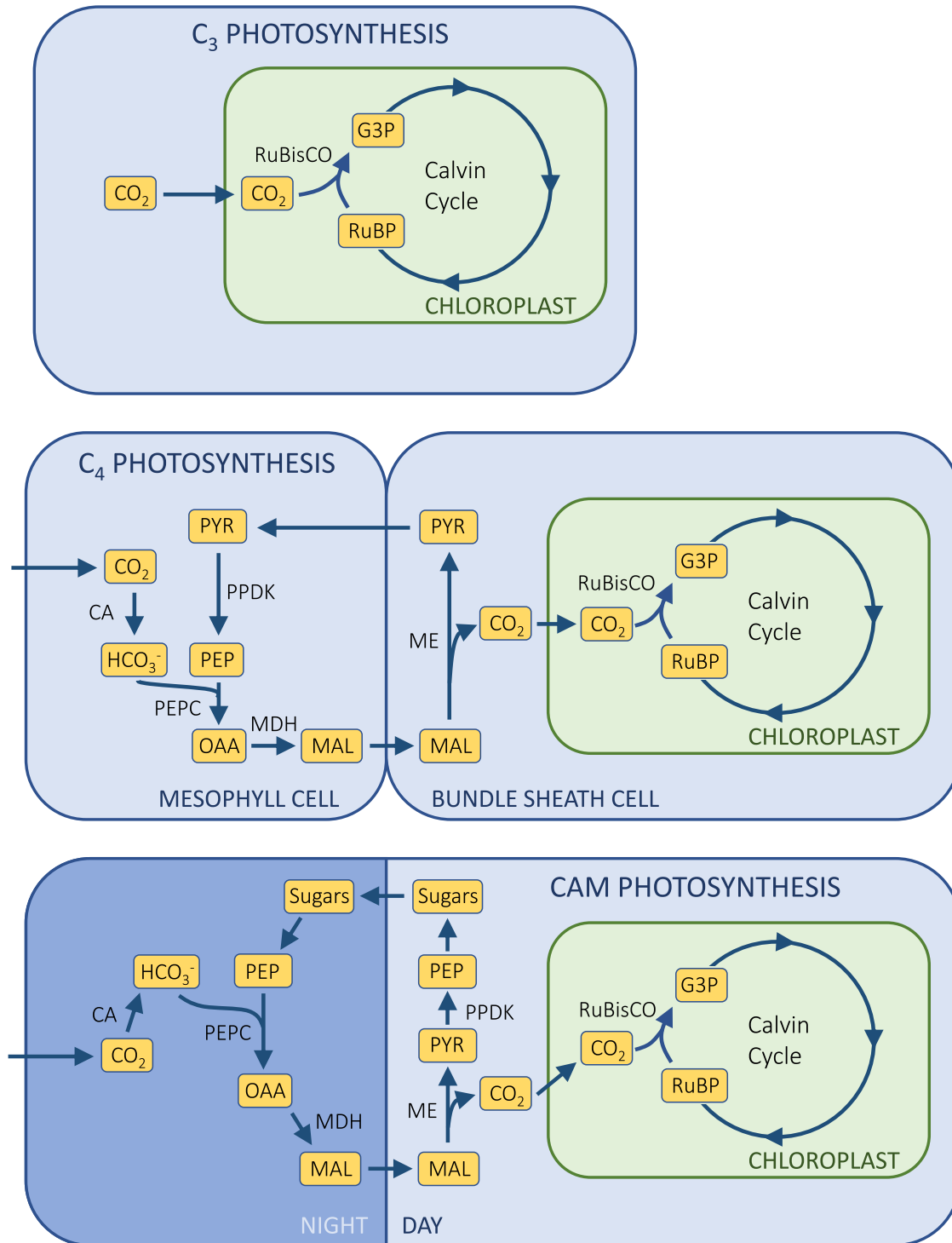


Figure 1.2: Illustrating the differences between C₃, C₄ and CAM photosynthesis. Metabolite abbreviations: CO₂, carbon dioxide; HCO₃⁻, bicarbonate; PEP, phosphoenolpyruvate; OAA, oxaloacetate; MAL, malate; PYR, pyruvate; RuBP, ribulose-1,5-bisphosphate; G3P, 3-phosphoglycolate. Enzyme abbreviations: CA, carbonic anhydrase; PEPC, phosphoenolpyruvate carboxylase; MDH, malate dehydrogenase; ME, malic enzyme; PPDK, pyruvate phosphate dikinase; RuBisCO, ribulose biphosphate carboxylase/oxygenase.

required for UV protection and immune defence, are produced via the phenylpropanoid biosynthetic pathway; correspondingly several of the genes required for phenylpropanoid biosynthesis are also present in the streptophyte algae (de Vries et al. 2017), suggesting that the basis for these metabolic pathways had already evolved well before the divergence of the ancestral embryophyte. Similarly, streptophyte algae have been found to contain various phytohormones essential for developmental processes and stress responses in embryophytes. The phytohormones auxin and abscisic acid (ABA) have been isolated from streptophyte algae (Boot et al. 2012; Hori et al. 2014) and cytokinins have been found in plants as early-diverging as the chlorophytes (Stirk et al. 2003), however land plant-like perception and signal transduction mechanisms do not appear to function in these species (Mutte et al. 2018; Ohtaka et al. 2017; Pils and Heyl 2009; Sun et al. 2020). Strigolactones are also present in streptophyte algae, where they promote rhizoid elongation (Delaux et al. 2012). This function is conserved in land plants, which additionally use strigolactones to mediate symbiotic relationships with arbuscular mycorrhizal fungi – and indeed the development of these symbiotic relationships is thought to have been important in enabling the transition to land (Delaux et al. 2015).

Beyond that metabolism already present in algal streptophytes, there has been considerable expansion of secondary metabolic diversity within land plants (Kusano et al. 2019). Structural biopolymers were clearly required in algae, presumably involved with microbial resistance and possibly UV protection and desiccation tolerance in various semi-terrestrial algae (He et al. 2016; Kondo et al. 2016). There is also evidence that the same metabolic reactions are responsible for the early stages of algal and land-plant biopolymer biosynthesis (Niklas et al. 2017). However, the monomeric composition of these compounds in embryophytes differs from those in algae (Kondo et al. 2016; Niklas et al. 2017). For example, cutin is a lipid-rich polymer which forms the major component of the land plant cuticle, protecting the aerial organs from desiccation and other abiotic stresses (Bakan and Marion 2017). Its component monomers and molecular structure is relatively conserved across land plant lineages, but is markedly different from polymers comprising the cuticle-like layer in the streptophyte alga *Klebsormidium flaccidum* (Kondo et al. 2016). This suggests that while the hydrophobic structural compounds found in *Klebsormidium* may aid in surviving

intermittent terrestrial conditions, the cutin present across land plants likely evolved in the common ancestor of all embryophytes. The similar biopolymer suberin, which shares structural and monomeric features with cutin but contains higher proportions of longer chain fatty acids and phenolics (Philippe et al. 2020), is thought to have evolved only with the common ancestor of the tracheophytes – correspondingly it is generally deposited in plant roots and stems. The biosynthetic pathway for suberin biosynthesis evolved directly from that for cutin biosynthesis. The two metabolic pathways share their early reactions and the glycerol-3-phosphate acyltransferases (GPATs) responsible for the later stages are homologous, with those involved in suberin biosynthesis arising only in the common ancestor of the tracheophytes (Yang et al. 2012). Similarly, lignin is found in all tracheophytes (Niklas et al. 2017). It is composed primarily of polymerised hydroxycinnamyl alcohols, and provides rigidity and hydrophobicity in woody tissues and the vessels of vascular systems (Weng et al. 2010). With origins in the phenylpropanoid biosynthetic pathway, reactions catalysing the early stages of lignin biosynthesis are found in streptophyte algae and also in bryophytes, where evidence from the moss *Physcomitrella patens* suggests that they are involved in the production of a cuticle enriched with phenols which could be considered ancestral to the lignins present in tracheophytes (Renault et al. 2017). Interestingly, even within the tracheophytes the biosynthesis of lignin has diversified. All tracheophytes produce lignin containing monolignols derived from *p*-coumaryl- and coniferyl-alcohols, but a third form of monolignol derived from sinapyl alcohol has been identified in several tracheophyte lineages. The biosynthetic pathway for the production of sinapyl alcohol has been shown to differ between angiosperms and the lycophyte *Selaginella moellendorffii* (Weng et al. 2008), suggesting that this monolignol may have evolved at least twice independently (Weng et al. 2010).

The increase in morphological and developmental complexity seen over the course of land plant evolution was additionally supported by further metabolic developments, such as the phytohormone-based signalling pathways that facilitate the differentiation of multiple cell types and communication between the resulting plant organs (Bowman et al. 2019). Components of some such signalling pathways are thought to have already been present in the algal ancestor of land plants. For example, the angiosperm-like metabolic pathway for the biosynthesis and signal transduction of ethylene has been found in the streptophyte

alga *Spirogyra pratensis*, producing a similar cell elongation response across algae and land plants (Ju et al. 2015). Similarly, genes encoding biosynthetic enzymes for the production of *cis*-zeatin-type cytokinins, as well as all major components of the cytokinin signalling pathway, have been identified in streptophyte algae (Rashotte 2020; Wang et al. 2020). Conversely, despite the identification of auxin, jasmonic acid (JA), abscisic acid (ABA) and strigolactone compounds in streptophyte algae, algae do not appear to contain the canonical land plant signalling pathways for these hormones. The identification of genes encoding biosynthetic enzymes for the production of auxin and JA in streptophyte algae suggest that these hormones are ancient innovations which pre-date the evolution of land plants (Han 2016; Romani 2017). However, the full angiosperm-like signalling pathways are present only in land plants. The common use of auxin and JA as molecular ‘glue’, stabilising protein-protein interactions to modulate their respective signalling pathways, suggests that these transduction pathways are related and originated via duplication of an ancestral receptor molecule in the common ancestor of all embryophytes (Wang et al. 2015). ABA is ubiquitous in all taxonomic kingdoms except the Archaea, but unlike auxin and JA its angiosperm-like biosynthesis and signalling pathways are limited to land plants. While strigolactones with apparently analogous functions have been identified in algae and land plants, and all major strigolactone biosynthetic enzymes are present across land plants (Bonhomme and Waters 2019), the angiosperm-like signalling pathway is only complete in spermatophytes, with earlier-diverging plants lacking the D14 receptor protein (Bythell-Douglas et al. 2017). The functional receptor in non-spermatophytes has not yet been identified (Waters et al. 2017). Likewise, while BRs are identified across land plants and algae (Yokota et al. 2017), the angiosperm-like mode of BR perception (via the BRI1 protein) is only utilised in the spermatophytes (Kim and Russinova 2020). GAs differ in that the hormones themselves have not been identified in algae or bryophytes. This is surprising, given that GA signalling is involved with varied aspects of growth, development and stress responses in vascular plants and exhibits extensive crosstalk with other hormones (Schwechheimer 2012). Although other components of their angiosperm-like signalling pathway are present in the bryophytes, GAs are thought to have appeared concurrently with their DELLA protein receptors in the common ancestors of the tracheophytes (Blaacutierrez et al. 2020).

1.2 BIOINFORMATIC AND COMPUTATIONAL INVESTIGATION OF PLANT METABOLISM

1.2.1 Investigating plant metabolism is a necessity and a challenge

Biochemistry underpins life on Earth, generating energy to fuel the production of biomass as well as the molecules required to construct organisms and modulate their interactions with the environment. The metabolic networks which underlie these processes are necessarily complex, particularly in plants which, subject to extreme variation in environmental conditions and unable to escape hazardous circumstances, produce an astonishing array of specialised metabolites with functions including defence, protection from biotic and abiotic stresses, and interactions with other organisms (e.g. reproduction and pollination) (Sweetlove and Ratcliffe 2011). While over one million unique metabolites are thought to be produced across plant species, the vast majority of our current scientific understanding of these compounds and their associated metabolic pathways is based upon investigation of model angiosperms and crop plants. There exists a wide knowledge gap concerning metabolism in earlier-diverging land plants and algae, which may possess as-yet-undiscovered metabolic compounds and processes. Plant metabolism is therefore an enormous source of untapped potential for beneficial exploitation. Over the past two decades, two thirds of novel drugs have been derived from plant metabolites (Rai et al. 2017). Unexplored metabolic networks in early-diverging land plants may be concealing not only novel metabolites with potentially therapeutic functions, but also novel metabolic processes for the production of previously identified metabolites. These may be of use in the development of herbicides with new modes of action as well as in the design of metabolic engineering strategies for the production of high-value biochemicals. Additionally, the lack of information available regarding early-diverging land plant metabolism severely limits our ability to draw conclusions about processes of metabolic evolution which, given the scarcity of metabolic remnants in the fossil record, is largely based on comparative analyses of extant species (Weng 2014).

Even for organisms in which metabolism is relatively well-studied, our knowledge is often insufficient for practical applications. The need for increased crop productivity as climates shift and global populations rise is a problem with an inherently metabolic solution, and much current metabolic research is geared towards genetic engineering of crop plants for

increased productivity and more efficient use of nutrients and water. Specific aims include improving photosynthetic efficiency for increased yield, reducing the need for fertiliser by developing plants capable of growing in nutrient-depleted conditions, improving the nutrient content in edible parts of crops and reducing the recalcitrance of biopolymers such as lignin for improved biofuels (Sulpice and McKeown 2015). The necessary elucidation of the enzymatic reactions responsible for the biosynthesis and catabolism of individual metabolites involved in these processes can often be challenging in itself, particularly for secondary metabolic processes which may only be identifiable in specific plant tissues under certain environmental conditions (Sweetlove and Ratcliffe 2011). Further difficulties exist because metabolic systems are under tight genetic control (often with built-in redundancies) which influence the regulation of the manufacture and breakdown of individual products, their transport between cellular compartments and across plant tissues, and their signalling crosstalk with interacting metabolites and proteins. Attempting to modulate specific areas of plant metabolism without a comprehensive understanding of their positioning within the wider plant metabolic and regulatory network may therefore lead to unexpected phenotypic consequences as a result of previously unidentified metabolic interactions (Sweetlove et al. 2017). Thorough investigation of even a single metabolic pathway therefore requires the integration of numerous techniques to identify specific compounds and investigate their biosynthesis, regulation, activity and interactions on a cellular and whole-plant scale.

1.2.2 Bioinformatic and computational methods in plant metabolism

The requirement for investigation at multiple levels of biosynthetic and regulatory coordination in the pursuit of fully characterising plant metabolic processes has led to the incorporation of various 'omics techniques into metabolic research (Sulpice and McKeown 2015). Experimental techniques such as metabolic flux analysis using isotope tracing and high-throughput mass spectrometry provide fluxomic data and broad-scale metabolomic and proteomic presence/absence information (Roscher et al. 2000; Sheth and Thaker 2014; Shih and Morgan 2020). Cell fractionation preceding mass spectrometry (or lower-throughput approaches such as fluorescent tagging of proteins combined with microscopy) can identify the subcellular location of individual portions of a metabolic pathway (Kueger et al. 2012),

and yeast two-hybrid assays can be used to identify features of an organism's protein-protein interactome (Dreze et al. 2011; Seo et al. 2011). These methods combine with genomic approaches such as whole-genome sequencing, which can identify the full complement of enzyme-encoding genes in an organism, and ChIP-seq and RT-PCR/RNAseq, which like metabolomic and fluxomic investigations can be carried out over a time series, and provide information on metabolic regulatory processes by inspection of the epigenome and transcriptome (Sheth and Thaker 2014).

Beyond the detailed experimental analysis of individual metabolic processes, the data available from high-throughput approaches such as genome and transcriptome analysis allow the study of metabolism on a much larger scale. Comparative analyses of known metabolic genes across extant species have been used to draw conclusions about the evolutionary processes leading to metabolic diversity in land plants. A recent study comparing all metabolically annotated genes in 16 land plants and algae concluded that the vast majority of metabolic innovation since the evolution of land plants has been in secondary metabolic processes, with central metabolism highly conserved between land plants and algae (Chae et al. 2014). The same study used metabolic network statistics to show that secondary metabolic processes in land plants tend to be more complex and exhibit increased interconnectivity than those in algae, while clustering analysis indicated that these increasingly complex metabolic networks evolved via a pattern of descent with modification (whereby selection for duplication of secondary metabolic genes results in expansion of enzyme families) (Chae et al. 2014). Originally, secondary metabolic pathways evolved from central metabolic branch-points (Weng 2014). For example, caffeine is metabolised from purine nucleotide precursors (Ashihara et al. 2008), and phenylpropanoid metabolism begins with the amino acid phenylalanine (Vogt 2010).

However, investigation of metabolic evolution, particularly regarding secondary metabolism, is hindered by the previously described knowledge gap affecting metabolism in early-diverging land plants. Encouragingly, large scale genomic, transcriptomic and phylogenetic analyses are increasingly being used to identify novel metabolic pathways and enzymes across a range of non-model plant species. Protein sequences are used to assign gene function based on the similarity between amino acid sequences of characterised and uncharacterised

genes (Tian and Skolnick 2003), while differential expression analyses comparing plant transcriptomes under which the production of a secondary metabolite differs can be used to identify candidate genes for the biosynthesis of the metabolite and its associated regulatory processes. For example, transcript abundances for genes involved in cytokinin biosynthesis and regulation have been used to characterise the cytokinin-mediated response to drought in soybean (Le et al. 2012) and tissue-specific roles of cytokinins in *Jatropha* (Cai et al. 2018). Moreover, machine learning algorithms capable of producing high-throughput predictions applied to genomic, transcriptomic and proteomic data have proved highly applicable to plant metabolic research (Mahood et al. 2020). Thousands of genomic features for individual genes were recently used as input for random forest and support vector machine algorithms to identify uncharacterised genes as involved with secondary metabolism in *Arabidopsis* (Moore et al. 2019), and amino acid sequences as inputs for deep neural networks have been used to predict subcellular localisation of plant proteins based on the classification of embedded targeting peptides (Armenteros et al. 2019). The latter information is crucial for the elucidation of metabolic enzyme function as well as the accuracy of computational models of metabolism (discussed below).

Additionally, contextual information for individual genes within a genome, transcriptome or phylogeny can aid in the discovery of metabolic genes and their functions. It has been shown that genes sharing similar phylogenetic profiles across species (i.e. with homologues present and absent in the same sets of species) are more likely to be functionally similar than genes with disparate phylogenetic profiles (Pellegrini et al. 1999). On this basis, candidate genes for involvement in specific metabolic functions can be identified by comparing metabolic profiles, although genomic information from a large number of species is required to maximise robustness of the results (Weißborn and Walther 2017). Similarly co-expression analysis, based on the assumption that genes with associated functions are likely to be synchronously expressed, can identify genes involved with a particular aspect of metabolism (Rhee and Mutwil 2014). This method has been used in *Arabidopsis* to identify metabolic enzymes involved in phenylalanine production (Maeda et al. 2010), as well as to characterise transcription factors regulating fatty acid biosynthesis (Han et al. 2012). Furthermore, it has recently been discovered that genes encoding various secondary

metabolic biosynthetic pathways in plants are grouped in operon-like clusters within a genome (Boycheva et al. 2014; Nützmann et al. 2016). Examples include genes for the biosynthesis of triterpenes in various species (Field and Osbourn 2008; Krokida et al. 2013), cyanogenic glucosides in the legume *Lotus japonicus* (Tako et al. 2011), alkaloids in solanaceous vegetables and the poppy *Papaver somniferum* (Itkin et al. 2013; Winzer et al. 2012) and momilactones in rice and the moss *Calohyphnum plumiforme* (Mao et al. 2020; Shimura et al. 2007). The existence of these clusters provides an opportunity for the prediction of metabolic gene functions based on a gene's proximity to known genes within the same cluster, as well as the prediction of novel metabolic pathways via the identification of clusters *de novo*. Analysis of genes co-localised with those of known function have helped to elucidate biosynthetic pathways for novel triterpenes in *Lotus japonica* and alkaloids in potato (Itkin et al. 2013; Krokida et al. 2013). Gene cluster analysis has also been combined with techniques such as co-expression analysis to identify groups of genes which are highly likely to represent as-yet-uncharacterised functional metabolic clusters (Schlöpfer et al. 2017; Wada et al. 2012).

Finally, multiple such bioinformatic strategies can be combined synergistically. Recently, a biosynthetic enzyme for anthraquinone production in the medicinal plant *Senna tora* was characterised using a combination of phylogenetic analysis to identify gene families specifically expanded in *S. tora*, differential expression analysis to compare transcriptomes produced by plant tissues which vary in anthraquinone production, and co-expression analysis to identify gene clusters for which transcript abundance profiles matched metabolomic profiles for anthraquinones (Kang et al. 2020).

Computational methods are also used for the in-depth examination of known metabolic processes. Computational modelling techniques incorporating data from the sources described above can represent the entire metabolic network of an organism *in silico* using only the chemical stoichiometry of individual enzyme-catalysed reactions, identified by the presence of their encoding genes in the genome/transcriptome of the organism in question (Gomes de Oliveira Dal'Molin and Nielsen 2013; O'Brien et al. 2015). Flux distributions can then be predicted across the entire metabolic network. The most common method of predicting network flux distributions is flux-balance analysis (FBA), a process involving the application of network constraints which mimic experimental conditions (for example

regarding light and nutrient availability) and an optimisation analysis with the aim of maximising flux through a specific reaction (for example biomass production) (Sweetlove et al. 2017). However, other methods for generating flux predictions exist. The investigation of differential fluxes between *Arabidopsis* guard cells and mesophyll cells has been facilitated by the use of a constraint-based procedure which maximises the agreement between modelled flux distributions and transcriptome datasets for each cell type (Robaina Estévez et al. 2017; Robaina Estévez and Nikoloski 2015). The inclusion of an objective function (which is not always representative of realistic metabolic demands in plants (Boyle et al. 2017)) can be avoided by the use of flux sampling, a method which generates a probability distribution for the flux through each reaction in a network and has been used to investigate the central metabolic basis for cold acclimation in *Arabidopsis* (Herrmann et al. 2019).

Genome-scale models of metabolism provide a powerful approach for the investigation of metabolism in specific organisms with minimal experimental input. The inclusion of all enzymatic reactions allows for the prediction of network-wide consequences of metabolic knockouts or exogenous knock-ins, shedding light on redundancies and alternative pathways that may cause a metabolic engineering strategy to fail (Cheung et al. 2013). Furthermore, stoichiometric modelling has the potential to address the angiosperm bias in metabolic understanding by allowing the study of metabolism in increasing numbers of early-diverging plants for which experimental data are limited but genome information is available (Banks et al. 2011; Bowman et al. 2017; Hori et al. 2014; Li et al. 2018; Neale et al. 2014; Nishiyama et al. 2018; Rensing et al. 2008).

Inferences derived from genome-scale models of plant metabolism can be improved by incorporating further information into models, for example by combining the resulting flux predictions with experimental data extracted under conditions matching those modelled by FBA. Including the subcellular compartmentalisation of individual reactions and the transport of metabolites between compartments (although this generally requires experimental investigation and for non-model species is usually derived from model plants) is essential as the subcellular localisation of enzymes and transporters affect possible flux routes (Cheung et al. 2014; Sweetlove and Fernie 2013). Hybrid models incorporating differential equations governing kinetic parameters for specific enzymes can overcome the challenge

of the lack of regulatory information in models based purely on stoichiometry (Shaw and Cheung 2018), and the integration of time-series transcriptome data to inform the upper and lower boundaries of individual fluxes has been used successfully to investigate metabolic acclimation to light and temperature in *Arabidopsis* (Töpfer et al. 2013). As the latter examples suggest, FBA carried out on genome-scale models can furthermore be expanded in scope both temporally and spatially. In the former scenario, the results of initial optimisations are fed forwards as constraints for subsequent optimisations to produce a time series of flux distributions. In the latter, multiscale optimisations are constructed by combining models for individual cell- or tissue-types into a single optimisation problem, thereby allowing the analysis of both intra- and inter-cellular metabolism. This methodology has been used to study metabolic subtypes of C_4 metabolism (Gomes de Oliveira Dal'Molin et al. 2010), which takes place across mesophyll and bundle-sheath cells, as well as resource partitioning in source and sink tissues in *Arabidopsis* (Shaw and Cheung 2018).

1.2.3 Flux-balance analysis of genome-scale models

A genome-scale stoichiometric model systematically represents all enzymatic reactions and corresponding metabolites encoded in an organisms genome and can therefore be used to accurately simulate metabolism *in silico* (Gomes de Oliveira Dal'Molin et al. 2010). FBA is the most common tool used to predict global flux distributions in such a model (Orth et al. 2010; Sweetlove et al. 2017) (Figure 1.3). Using this method, the metabolic network is represented as a matrix S of size $m \times n$, where m represents all metabolites in the network and n all reactions. The model contains the stoichiometric coefficient for metabolite m_i in reaction n_j . The flux distribution solution of the network is represented by a vector v of length n , which contains the flux through each reaction in the network. The contents of this vector must satisfy two conditions. First, that $S \cdot v = 0$. This codifies the assumption made by FBA that metabolism operates at a steady state – i.e. that the sum of fluxes consuming a metabolite matches those producing it, such that the net change in metabolite concentration is 0. This has been shown to be a reasonable assumption over short timescales (Collakova et al. 2012; Schilling et al. 2000). Second, that $l_n < v_n < u_n$, where l and u denote the lower and upper flux boundaries allowed for each reaction n . This limits the solution space

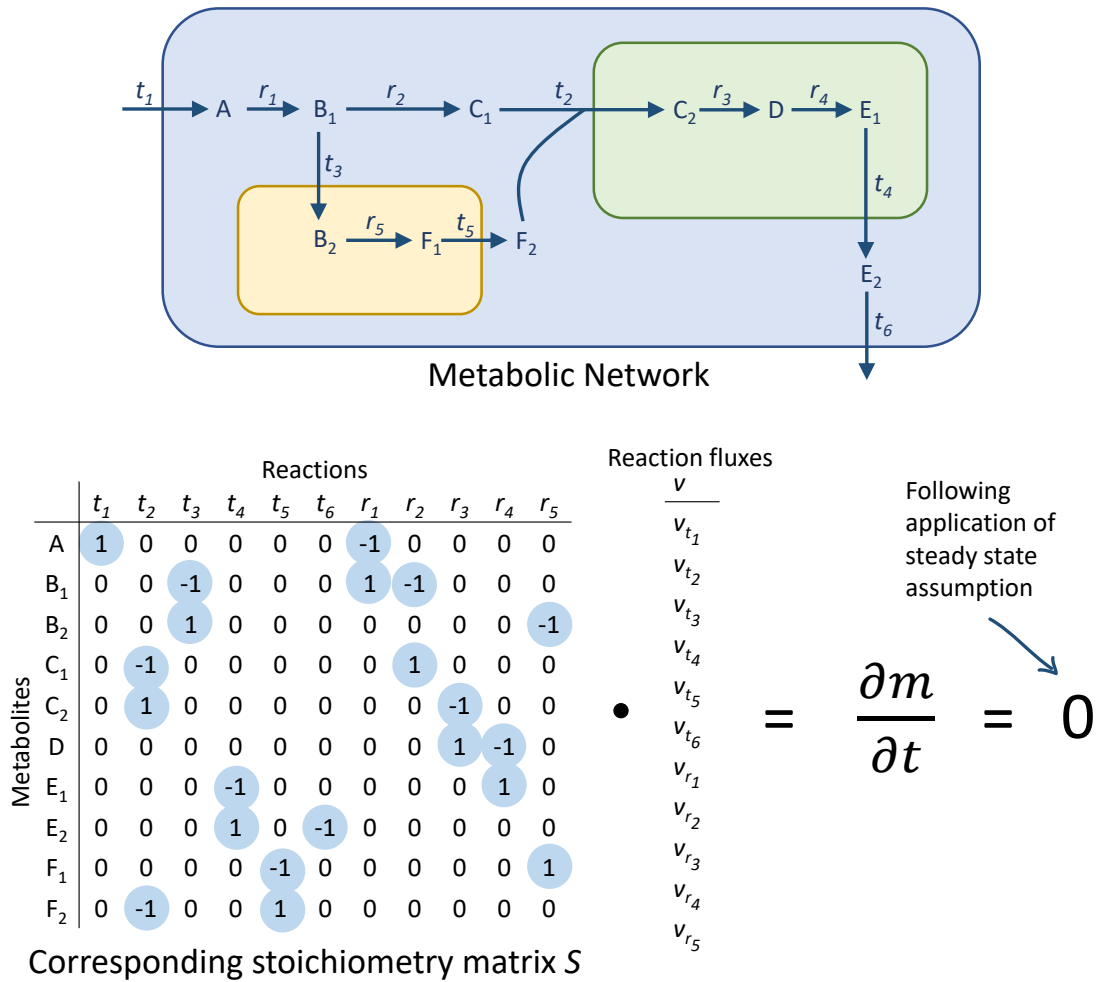


Figure 1.3: Flux-balance analysis. For an example metabolic network (top panel) a stoichiometry matrix S is produced identifying the production and consumption stoichiometries for each metabolite m in each reaction n . When combined with the known rate of metabolite flow (flux) through each equation, the rate of change in concentration of each metabolite over time can be calculated. Including a steady state assumption reduces this problem to a linear equation which can be used instead to calculate the flux (v) through each reaction. Metabolites are denoted here as capital letters, reactions are denoted as r_i (metabolic interconversions) and t_i (transport reactions), reaction fluxes are denoted as $v_{reaction}$.

for flux distributions to within the applied flux boundary constraints which mimic both experimentally defined inputs and outputs such as light intensity and carbon availability as well as known thermodynamic limitations which define the reversibility of each reaction.

Following the setup of reaction stoichiometries and flux constraints, a linear optimisation algorithm is applied to identify the flux distribution which maximises or minimises specified target outputs of the network. In plant models, the maximised output is often a biomass synthesis reaction or sugar export reaction, simulating plant growth or phloem export from a source tissue. For models intended to support the development of a metabolic engineering strategy the optimisation goal may instead be the maximisation or minimisation of flux through biosynthetic pathways for specific desirable or undesirable metabolites. Genome-scale models tend to include more reactions than metabolites ($n > m$) and as such are underdetermined (there are many possible solutions to $S \cdot v = 0$). To further limit the solution space, an additional commonly applied constraint is the minimisation of the sum of fluxes over the whole network. Known as parsimonious FBA (pFBA) (Lewis et al. 2010), this simulates efficient enzyme usage, relevant to model plants such as crops which are often grown in nitrogen-limited environments. Finally, the full range of individual reaction fluxes over which the solution space can range while still maintaining an optimised objective function can be identified using a method known as flux variability analysis (FVA) (Gudmundsson and Thiele 2010).

1.3 PROJECT RATIONALE AND OBJECTIVES

The colonisation of land by the streptophyte ancestor of land plants and the subsequent adaptive radiation of embryophytes were accompanied by the development of an array of novel metabolic capabilities required for growth and development. Examples include the biosynthesis of phytohormones such as GAs and BRs, and structural biopolymers such as cutin and lignin. Although primary metabolism is considered highly conserved among land plants, metabolic profiling experiments indicate huge unexplored diversity in secondary metabolic pathways across individual species and taxa, and the challenge of experimental metabolic investigation means that our knowledge of plant metabolic networks is heavily skewed towards well-studied angiosperm species. The goal of the research described in this

thesis was to begin to address these knowledge gaps in plant metabolism, focusing both on the metabolism of early-diverging land plants and the evolution of secondary metabolism across the embryophytes. A computational and bioinformatic approach based on scientific programming in Python and R was followed, drawing on recent advances in the development of computational methods in metabolic research (see Section 1.2.2) and in the availability of genome sequence and transcriptome information for early-diverging land plants. The benefits of such investigation are widespread. Comparative analyses of metabolism across plant lineages allow us to draw conclusions on the evolutionary processes governing the development metabolic pathways. Furthermore, an improved understanding of metabolic processes in early-diverging land plants and how they differ from those in angiosperms can direct the discovery of novel metabolic pathways, enzymes and compounds. These may have applications in the identification of useful metabolites such as drugs and herbicides, in the design of metabolic engineering strategies for crop improvement and in the use of plants for the production of high-value biochemicals.

This work is presented in three manuscripts. The first, submitted to *Plants in Silico*, focuses on central metabolism in the model bryophyte *Marchantia polymorpha*. Bryophytes exhibit high mesophyll resistance to CO₂ diffusion, and consequentially have some of the lowest carbon assimilation rates measured in land plants. I hypothesised that the resulting reduced carbon budget would favour flux modes that maximise carbon use efficiency and could lead to different flux distributions in central metabolism compared to angiosperms. To investigate this, I constructed carefully constrained stoichiometric models for *Marchantia* and *Arabidopsis* and compared flux predictions produced using FBA. I found that the prediction of reasonable growth rates in *Marchantia* requires surprisingly high carbon conversion efficiencies compared to those in *Arabidopsis*. The second manuscript, published in *Current Biology* in May 2020 (see Appendix C), presents a comparative analysis of secondary metabolism across land plants and algae. I combined genome-scale metabolic models for eight representative species spanning the major extant streptophyte lineages with homology information from a further 64 land plants and algae (derived from 62 genome sequences and two transcriptomes) to pinpoint the timings of appearances and disappearances of metabolic traits over the course of land plant evolution. The results resolved details of the evolution of biosynthetic pathways for

several phytohormones and biopolymers, and analysis of an additional 305 non-angiosperm transcriptomes supported the findings. Finally, in a third manuscript (unsubmitted) I used machine learning-based gene co-expression analysis and phylogenetic profiling methods (based on existing transcriptome data for three *Arabidopsis* accessions combined with the same homology data produced in the second manuscript) to extend our knowledge of secondary metabolism in plants by identifying and investigating candidate genes associated with biosynthetic pathways for phytohormones and biopolymers. I found several types of associated proteins, including enzymes involved in related metabolic pathways, proteins involved in the delivery of precursors and transcription factors with regulatory roles.

REFERENCES

- Armenteros, J. J. A., Salvatore, M., Emanuelsson, O., ... Nielsen, H. (2019). Detecting sequence signals in targeting peptides using deep learning. *Life Science Alliance*, 2, e201900429.
- Ashihara, H., Sano, H., & Crozier, A. (2008). Caffeine and related purine alkaloids: Biosynthesis, catabolism, function and genetic engineering. *Phytochemistry*, 69, 841–856.
- Bakan, B., & Marion, D. (2017). Assembly of the cutin polyester: From cells to extracellular cell walls. *Plants*, 6, 57.
- Banks, J. A., Nishiyama, T., Hasebe, M., ... Grigoriev, I. V. (2011). The *Selaginella* genome identifies genetic changes associated with the evolution of vascular plants. *Science*, 332, 960–963.
- Bauwe, H., Hagemann, M., Kern, R., & Timm, S. (2012). Photorespiration has a dual origin and manifold links to central metabolism. *Current Opinion in Plant Biology*, 15, 269–275.
- Blaacutierrez, M. A., Nelson, D. C., & Weijers, D. (2020). Evolution of plant hormone response pathways. *Annual Review of Plant Biology*, 71, 327–353.
- Bone, R. E., Smith, J. A. C., Arrigo, N., & Buerki, S. (2015). A macro-ecological perspective on crassulacean acid metabolism (CAM) photosynthesis evolution in afro-madagascan drylands: Eulophiinae orchids as a case study. *New Phytologist*, 208, 469–481.
- Bonhomme, S., & Waters, M. (2019). Evolution of strigolactone biosynthesis and signalling. *Strigolactones - biology and applications* (pp. 143–161). Springer International Publishing.
- Boot, K. J. M., Libbenga, K. R., Hille, S. C., ... Van Duijn, B. (2012). Polar auxin transport: An early invention. *Journal of Experimental Botany*, 63, 4213–4218.
- Bowman, J. L., Araki, T., & Kohchi, T. (2016). Marchantia: Past, present and future. *Plant and Cell Physiology*, 57, 205–209.
- Bowman, J. L., Briginshaw, L. N., Fisher, T. J., & Flores-Sandoval, E. (2019). Something ancient and something neofunctionalized - evolution of land plant hormone signaling pathways. *Current Opinion in Plant Biology*, 47, 64–72.
- Bowman, J. L., Kohchi, T., Yamato, K. T., ... Schmutz, J. (2017). Insights into land plant evolution garnered from the *Marchantia polymorpha* genome. *Cell*, 171, 287–304.e15.
- Boycheva, S., Daviet, L., Wolfender, J. L., & Fitzpatrick, T. B. (2014). The rise of operon-like gene clusters in plants. *Trends in Plant Science*, 19, 447–459.
- Boyle, N. R., Sengupta, N., & Morgan, J. A. (2017). Metabolic flux analysis of heterotrophic growth in *Chlamydomonas reinhardtii*. *PLOS ONE*, 12, e0177292.

- Bräutigam, A., Schlüter, U., Eisenhut, M., & Gowik, U. (2017). On the evolutionary origin of CAM photosynthesis. *Plant Physiology*, 174, 473–477.
- Bythell-Douglas, R., Rothfels, C. J., Stevenson, D. W. D., ... Bennett, T. (2017). Evolution of strigolactone receptors by gradual neo-functionalization of KAI2 paralogues. *BMC Biology*, 15.
- Cai, L., Zhang, L., Fu, Q., & Xu, Z. F. (2018). Identification and expression analysis of cytokinin metabolic genes IPTs, CYP735A and CKXs in the biofuel plant *Jatropha curcas*. *PeerJ*, 16(6), e4812.
- Chae, L., Kim, T., Nilo-Poyanco, R., & Rhee, S. Y. (2014). Genomic signatures of specialized metabolism in plants. *Science*, 344, 510–513.
- Chan, K. X., Mabbitt, P. D., Phua, S. Y., ... Pogson, B. J. (2016). Sensing and signaling of oxidative stress in chloroplasts by inactivation of the SAL1 phosphoadenosine phosphatase. *Proceedings of the National Academy of Sciences of the United States of America*, 113, E4567–E4576.
- Chanderbali, A. S., Berger, B. A., Howarth, D. G., ... Soltis, D. E. (2016). Evolving ideas on the origin and evolution of flowers: New perspectives in the genomic era. *Genetics*, 202, 1255–1265.
- Cheung, C. Y. M., Poolman, M. G., Fell, D. A., ... Sweetlove, L. J. (2014). A diel flux balance model captures interactions between light and dark metabolism during day-night cycles in C_3 and crassulacean acid metabolism leaves. *Plant Physiology*, 165, 917–929.
- Cheung, C. Y. M., Williams, T. C. R., Poolman, M. G., ... Sweetlove, L. J. (2013). A method for accounting for maintenance costs in flux balance analysis improves the prediction of plant cell metabolic phenotypes under stress conditions. *Plant Journal*, 75, 1050–1061.
- Collakova, E., Yen, J. Y., & Senger, R. S. (2012). Are we ready for genome-scale modeling in plants? *Plant Science*, 191–192, 53–70.
- Delaux, P. M., Radhakrishnan, G. V., Jayaraman, D., ... Ane, J. M. (2015). Algal ancestor of land plants was preadapted for symbiosis. *Proceedings of the National Academy of Sciences of the United States of America*, 112, 13390–13395.
- Delaux, P. M., Xie, X., Timme, R. E., ... Séjalon-Delmas, N. (2012). Origin of strigolactones in the green lineage. *New Phytologist*, 195, 857–871.
- Delwiche, C. F., & Cooper, E. D. (2015). The evolutionary origin of a terrestrial flora. *Current Biology*, 25, R899–R910.
- Delwiche, C. F., Graham, L. E., & Thomson, N. (1989). Lignin-like compounds and sporopollenin in *Coleochaete*, an algal model for land plant ancestry. *Science*, 245, 399–401.
- de Vries, J., & Archibald, J. M. (2018). Plant evolution: Landmarks on the path to terrestrial life. *New Phytologist*, 217, 1428–1434.
- de Vries, J., de Vries, S., Slamovits, C. H., ... Archibald, J. M. (2017). How embryophytic is the biosynthesis of phenylpropanoids and their derivatives in streptophyte algae? *Plant and Cell Physiology*, 58, 934–945.
- de Vries, J., Stanton, A., Archibald, J. M., & Gould, S. B. (2016). Streptophyte terrestrialization in light of plastid evolution. *Trends in Plant Science*, 21, 467–476.
- Dittami, S. M., Heesch, S., Olsen, J. L., & Collén, J. (2017). Transitions between marine and freshwater environments provide new clues about the origins of multicellular plants and algae. *Journal of Phycology*, 53, 731–745.
- Dreze, M., Carvunis, A. R., Charlotiaux, B., ... Yazaki, J. (2011). Evidence for network evolution in an *Arabidopsis* interactome map. *Science*, 333, 601–607.
- Erb, T. J., & Zarzycki, J. (2018). A short history of RuBisCO: The rise and fall (?) of Nature's predominant CO_2 fixing enzyme. *Current Opinion in Biotechnology*, 49, 100–107.
- Field, B., & Osbourn, A. E. (2008). Metabolic diversification - independent assembly of operon-like gene clusters in different plants. *Science*, 320, 543–547.

- Fischer, W. W., Hemp, J., & Johnson, J. E. (2016). Evolution of oxygenic photosynthesis. *Annual Review of Earth and Planetary Sciences*, 44, 647–683.
- Gavelis, G. S., & Gile, G. H. (2018). How did cyanobacteria first embark on the path to becoming plastids?: Lessons from protist symbioses. *FEMS Microbiology Letters*, 365, 209.
- Gomes de Oliveira Dal’Molin, C., & Nielsen, L. K. (2013). Plant genome-scale metabolic reconstruction and modelling. *Current Opinion in Biotechnology*, 24, 271–277.
- Gomes de Oliveira Dal’Molin, C., Quek, L. E., Palfreyman, R. W., ... Nielsen, L. K. (2010). C4GEM, a genome-scale metabolic model to study C₄ plant metabolism. *Plant Physiology*, 154, 1871–1885.
- Goudet, M. M. M., Orr, D. J., Melkonian, M., & Carmo-Silva, H., E. and Griffiths. (2020). RuBisCO and carbon-concentrating mechanism co-evolution across chlorophyte and streptophyte green algae. *New Phytologist*, 227, 810–823.
- Gudmundsson, S., & Thiele, I. (2010). Computationally efficient flux variability analysis. *BMC Bioinformatics*, 11, 489.
- Han, G. (2016). Evolution of jasmonate biosynthesis and signaling mechanisms. *Journal of Experimental Botany*, 68, 1323–1331.
- Han, X., Yin, L., & Xue, H. (2012). Co-expression analysis identifies CRC and AP1 the regulator of *Arabidopsis* fatty acid biosynthesis. *Journal of Integrative Plant Biology*, 54, 486–499.
- Harris, B. J., Harrison, C. J., Hetherington, A. M., & Williams, T. A. (2020). Phylogenomic evidence for the monophyly of bryophytes and the reductive evolution of stomata. *Current Biology*, 30, 2001–2012.e2.
- Harrison, C. J., & Morris, J. L. (2018). The origin and early evolution of vascular plant shoots and leaves. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 373.
- Hartmann, T. (1996). Diversity and variability of plant secondary metabolism: A mechanistic view. *Entomologia Experimentalis et Applicata*, 80, 177–188.
- He, X., Dai, J., & Wu, Q. (2016). Identification of sporopollenin as the outer layer of cell wall in microalga *Chlorella protothecoides*. *Frontiers in Microbiology*, 7, 1047.
- Herrmann, H. A., Dyson, B. C., Vass, L., ... Schwartz, J. M. (2019). Flux sampling is a powerful tool to study metabolism under changing environmental conditions. *npj Systems Biology and Applications*, 5, 1–8.
- Hetherington, A. J., Berry, C. M., & Dolan, L. (2016). Networks of highly branched stigmarian rootlets developed on the first giant trees. *Proceedings of the National Academy of Sciences of the United States of America*, 113, 6695–6700.
- Hetherington, A. J., & Dolan, L. (2018). Stepwise and independent origins of roots among land plants. *Nature*, 561, 235–238.
- Hohmann-Marriott, M. F., & Blankenship, R. E. (2011). Evolution of photosynthesis. *Annual Review of Plant Biology*, 62, 515–548.
- Hori, K., Maruyama, F., Fujisawa, T., ... Ohta, H. (2014). *Klebsormidium flaccidum* genome reveals primary factors for plant terrestrial adaptation. *Nature Communications*, 5.
- Itkin, M., Heinig, U., Tzfadia, O., ... Aharoni, A. (2013). Biosynthesis of antinutritional alkaloids in solanaceous crops is mediated by clustered genes. *Science*, 341, 175–179.
- Jensen, P. E., & Leister, D. (2014). Chloroplast evolution, structure and functions. *F1000 Prime Reports*, 6.
- Ju, C., Van de Poel, B., Cooper, E. D., ... Chang, C. (2015). Conservation of ethylene as a plant hormone over 450 million years of evolution. *Nature Plants*, 1, 14004.
- Kang, S., Pandey, R. P., Lee, C., ... Sohng, J. K. (2020). Genome-enabled discovery of anthraquinone biosynthesis in *Senna tora*. *bioRxiv*, 2020.04.27.063495.

- Keeley, J. E., & Rundel, P. W. (2003). Evolution of CAM and C₄ carbon-concentrating mechanisms. *International Journal of Plant Sciences*, 164, 55–77.
- Kim, E., & Russinova, E. (2020). Brassinosteroid signalling. *Current Biology*, 30, R294–R298.
- Kondo, S., Hori, K., Sasaki-Sekimoto, Y., ... Ohta, H. (2016). Primitive extracellular lipid components on the surface of the charophytic alga *Klebsormidium flaccidum* and their possible biosynthetic pathways as deduced from the genome sequence. *Frontiers in Plant Science*, 7, 952.
- Krokida, A., Delis, C., Geisler, K., ... Papadopoulou, K. K. (2013). A metabolic gene cluster in *Lotus japonicus* discloses novel enzyme functions and products in triterpene biosynthesis. *New Phytologist*, 200, 675–690.
- Kueger, S., Steinhauser, D., Willmitzer, L., & Giavalisco, P. (2012). High-resolution plant metabolomics: From mass spectral features to metabolites and from whole-cell analysis to subcellular metabolite distributions. *Plant Journal*, 70, 39–50.
- Kusano, H., Li, H., Minami, H., ... Yazaki, K. (2019). Evolutionary developments in plant specialized metabolism, exemplified by two transferase families. *Frontiers in Plant Science*, 10, 794.
- Le, D. T., Nishiyama, R., Watanabe, Y., ... Tran, L. P. (2012). Identification and expression analysis of cytokinin metabolic genes in soybean under normal and drought conditions in relation to cytokinin levels. *PLoS ONE*, 7, e42411.
- Lewis, N. E., Hixson, K. K., Conrad, T. M., ... Palsson, B. (2010). Omic data from evolved *E. coli* are consistent with computed optimal growth from genome-scale models. *Molecular Systems Biology*, 6.
- Li, F. W., Brouwer, P., Carretero-Paulet, L., ... Pryer, K. M. (2018). Fern genomes elucidate land plant evolution and cyanobacterial symbioses. *Nature Plants*, 4, 460–472.
- Li, F. W., Villarreal, J. C., & Szövényi, P. (2017). Hornworts: An overlooked window into carbon-concentrating mechanisms. *Trends in Plant Science*, 22, 275–277.
- Ligrone, R. (2019). The great oxygenation event. *Biological innovations that built the world* (pp. 129–154). Springer International Publishing.
- Maeda, H., Yoo, H., & Dudareva, N. (2010). Prephenate aminotransferase directs plant phenylalanine biosynthesis via arogenate. *Nature Chemical Biology*, 7, 19–21.
- Mahood, E. H., Kruse, L. H., & Moghe, G. D. (2020). Machine learning: A powerful tool for gene function prediction in plants. *Applications in Plant Sciences*, 8, e11376.
- Mao, L., Kawaide, H., Higuchi, T., ... Okada, K. (2020). Genomic evidence for convergent evolution of gene clusters for momilactone biosynthesis in land plants. *Proceedings of the National Academy of Sciences of the United States of America*, 117, 12472–12480.
- Maréchal, E. (2018). Primary endosymbiosis: Emergence of the primary chloroplast and the chromatophore, two independent events. *Methods in Molecular Biology*, 1829, 3–16.
- Martin, W., Rujan, T., Richly, E., ... Penny, D. (2002). Evolutionary analysis of *Arabidopsis*, cyanobacterial, and chloroplast genomes reveals plastid phylogeny and thousands of cyanobacterial genes in the nucleus. *Proceedings of the National Academy of Sciences of the United States of America*, 99, 12246–12251.
- Mathis, P. (1990). Compared structure of plant and bacterial photosynthetic reaction centers. Evolutionary implications. *Biochimica et Biophysica Acta (BBA) - Bioenergetics*, 1018, 163–167.
- McCourt, R. M., Delwiche, C. F., & Karol, K. G. (2004). Charophyte algae and land plant origins. *Trends in Ecology and Evolution*, 19, 661–666.
- Merced, A., & Renzaglia, K. S. (2017). Structure, function and evolution of stomata from a bryological perspective. *Bryophyte Diversity and Evolution*, 39, 7.
- Meyer, M., & Griffiths, H. (2013). Origins and diversity of eukaryotic CO₂-concentrating mechanisms: Lessons for the future. *Journal of Experimental Botany*, 64, 769–786.

- Moore, B. M., Wang, P. P., Fan, P., ... Shiu, S. H. (2019). Robust predictions of specialized metabolism genes through machine learning. *Proceedings of the National Academy of Sciences of the United States of America*, 116, 2344–2353.
- Morris, J. L., Puttick, M. N., Clark, J. W., ... Donoghue, P. C. J. (2018). The timescale of early land plant evolution. *Proceedings of the National Academy of Sciences of the United States of America*, 115, E2274–E2283.
- Mutte, S. K., Kato, H., Rothfels, C., ... Weijers, D. (2018). Origin and evolution of the nuclear auxin response system. *eLife*, 7, e33399.
- Neale, D. B., Wegrzyn, J. L., Stevens, K. A., ... Langley, C. H. (2014). Decoding the massive genome of loblolly pine using haploid DNA and novel assembly strategies. *Genome Biology*, 15(R59).
- Niklas, K. J., Cobb, E. D., & Matas, A. J. (2017). The evolution of hydrophobic cell wall biopolymers: From algae to angiosperms. *Journal of Experimental Botany*, 68, 5261–5269.
- Nishiyama, T., Sakayama, H., de Vries, J., ... Rensing, S. A. (2018). The *Chara* genome: Secondary complexity and implications for plant terrestrialization. *Cell*, 174, 448–464.e24.
- Nützmann, H. W., Huang, A., & Osbourn, A. (2016). Plant metabolic clusters – from genetics to genomics. *New Phytologist*, 211, 771–789.
- O'Brien, E. J., Monk, J. M., & Palsson, B. O. (2015). Using genome-scale models to predict biological capabilities. *Cell*, 161, 971–987.
- Ohtaka, K., Hori, K., Kanno, Y., ... Ohta, H. (2017). Primitive auxin response without TIR1 and Aux/IAA in the charophyte alga *Klebsormidium nitens*. *Plant Physiology*, 174, 1621–1632.
- Olson, J. M., & Pierson, B. K. (1987). Origin and evolution of photosynthetic reaction centers. *Origins of life and evolution of the biosphere*, 17, 419–430.
- Orth, J. D., Thiele, I., & Palsson, B. O. (2010). What is flux balance analysis? *Nature Biotechnology*, 28, 245–248.
- Pellegrini, M., Marcotte, E. M., Thompson, M. J., ... Yeates, T. O. (1999). Assigning protein functions by comparative genome analysis: Protein phylogenetic profiles. *Biochemistry*, 96, 4285–4288.
- Philippe, G., Sørensen, I., Jiao, C., ... Rose, J. K. C. (2020). Cutin and suberin: Assembly and origins of specialized lipidic cell wall scaffolds. *Current Opinion in Plant Biology*, 55, 11–20.
- Pichersky, E., & Lewinsohn, E. (2011). Convergent evolution in plant specialized metabolism. *Annual Review of Plant Biology*, 62, 549–566.
- Pils, B., & Heyl, A. (2009). Unraveling the evolution of cytokinin signaling. *Plant Physiology*, 151, 782–791.
- Puttick, M. N., Morris, J. L., Williams, T. A., ... Donoghue, P. C. J. (2018). The interrelationships of land plants and the nature of the ancestral embryophyte. *Current Biology*, 28, 733–745.e2.
- Rai, A., Saito, K., & Yamazaki, M. (2017). Integrated omics analysis of specialized metabolism in medicinal plants. *Plant Journal*, 90, 764–787.
- Rashotte, A. M. (2020). The evolution of cytokinin signaling and its role in development before angiosperms. *Seminars in Cell and Developmental Biology*, 109, 31–38.
- Raven, J. A., Cockell, C. S., De, C. L., & Rocha, L. (2008). The evolution of inorganic carbon concentrating mechanisms in photosynthesis. *Philosophical Transactions of the Royal Society B*, 363, 2641–2650.
- Renault, H., Alber, A., Horst, N. A., ... Werck-Reichhart, D. (2017). A phenol-enriched cuticle is ancestral to lignin evolution in land plants. *Nature Communications*, 8, 14713.
- Rensing, S. A. (2017). Why we need more non-seed plant models. *New Phytologist*, 216, 355–360.
- Rensing, S. A. (2018). Great moments in evolution: The conquest of land by plants. *Current Opinion in Plant Biology*, 42, 49–54.
- Rensing, S. A., Lang, D., Zimmer, A. D., ... Boore, J. L. (2008). The *Physcomitrella* genome reveals evolutionary insights into the conquest of land by plants. *Science*, 319, 64–69.

- Rhee, S. Y., & Mutwil, M. (2014). Towards revealing the functions of all genes in plants. *Trends in Plant Science*, 19, 212–221.
- Robaina Estévez, S., Daloso, D. M., Zhang, Y., & Fernie, A. R. (2017). Resolving the central metabolism of *Arabidopsis* guard cells. *Scientific Reports*, 7, 8307.
- Robaina Estévez, S., & Nikoloski, Z. (2015). Context-specific metabolic model extraction based on regularized least squares optimization. *PLOS ONE*, 10, e0131875.
- Rockwell, N. C., Lagarias, J. C., & Bhattacharya, D. (2014). Primary endosymbiosis and the evolution of light and oxygen sensing in photosynthetic eukaryotes. *Frontiers in Ecology and Evolution*, 2, 66.
- Romani, F. (2017). Origin of TAA genes in charophytes: New insights into the controversy over the origin of auxin biosynthesis. *Frontiers in Plant Science*, 8, 1616.
- Roscher, A., Kruger, N. J., & Ratcliffe, R. G. (2000). Strategies for metabolic flux analysis in plants using isotope labelling. *Journal of Biotechnology*, 77, 81–102.
- Ruhfel, B. R., Gitzendanner, M. A., Soltis, P. S., ... Burleigh, J. G. (2014). From algae to angiosperms-inferring the phylogeny of green plants (Viridiplantae) from 360 plastid genomes. *BMC Evolutionary Biology*, 14, 1–27.
- Sage, R. F., Christin, P., & Edwards, E. J. (2011). The C₄ plant lineages of planet earth. *Journal of Experimental Botany*, 62, 3155–69.
- Sakakibara, K., Ando, S., Yip, H. K., ... Bowman, J. L. (2013). KNOX2 genes regulate the haploid-to-diploid morphological transition in land plants. *Science*, 339, 1067–1070.
- Schilling, C. H., Edwards, J. S., Letscher, D., & Palsson, B. (2000). Combining pathway analysis with flux balance analysis for the comprehensive study of metabolic systems. *Biotechnology and Bioengineering*, 71, 286–306.
- Schläpfer, P., Zhang, P., Wang, C., ... Rhee, S. Y. (2017). Genome-wide prediction of metabolic enzymes, pathways, and gene clusters in plants. *Plant Physiology*, 173, 2041–2059.
- Schwechheimer, C. (2012). Gibberellin signaling in plants - the extended version. *Frontiers in Plant Science*, 2, 107.
- Seo, Y. S., Chern, M., Bartley, L. E., ... Ronald, P. C. (2011). Towards establishment of a rice stress response interactome. *PLoS Genetics*, 7, e1002020.
- Shaw, R., & Cheung, C. Y. M. (2018). A dynamic multi-tissue flux balance model captures carbon and nitrogen metabolism and optimal resource partitioning during *Arabidopsis* growth. *Frontiers in Plant Science*, 9, 884.
- Sheth, B. P., & Thaker, V. S. (2014). Plant systems biology: Insights, advances and challenges. *Planta*, 240, 33–54.
- Shih, M. L., & Morgan, J. A. (2020). Metabolic flux analysis of secondary metabolism in plants. *Metabolic Engineering Communications*, 10, e00123.
- Shimura, K., Okada, A., Okada, K., ... Yamane, H. (2007). Identification of a biosynthetic gene cluster in rice for momilactones. *Journal of Biological Chemistry*, 282, 34013–34018.
- Sigel, E. M., Schuettpelz, E., Pryer, K. M., & Der, J. P. (2018). Overlapping patterns of gene expression between gametophyte and sporophyte phases in the fern *Polypodium amorphum* (Polypodiales). *Frontiers in Plant Science*, 9, 1450.
- Soo, R. M., Hemp, J., & Hugenholtz, P. (2019). Evolution of photosynthesis and aerobic respiration in the cyanobacteria. *Free Radical Biology and Medicine*, 140, 200–205.
- Sousa, F. L., Shavit-Grievink, L., & Allen, J. F. (2013). Chlorophyll biosynthesis gene evolution indicates photosystem gene duplication, not photosystem merger, at the origin of oxygenic photosynthesis. *Genome Biology*, 5, 200–216.
- Stirk, W. A., Novák, O., Strnad, M., & Van Staden, J. (2003). Cytokinins in macroalgae. *Plant Growth Regulation*, 41, 13–24.

- Sulpice, R., & McKeown, P. C. (2015). Moving toward a comprehensive map of central plant metabolism. *Annual Review of Plant Biology*, 66, 187–210.
- Sun, Y., Pri-Tal, O., Michaeli, D., & Mosquna, A. (2020). Evolution of abscisic acid signaling module and its perception. *Frontiers in Plant Science*, 11, 934.
- Sweetlove, L. J., & Fernie, A. R. (2013). The spatial organization of metabolism within the plant cell. *Annual Review of Plant Biology*, 64, 723–746.
- Sweetlove, L. J., Nielsen, J., & Fernie, A. R. (2017). Engineering central metabolism - a grand challenge for plant biologists. *The Plant Journal*, 90, 749–763.
- Sweetlove, L. J., & Ratcliffe, R. G. (2011). Flux-balance modeling of plant metabolism. *Frontiers in Plant Science*, 2, 38.
- Szovenyi, P., Rensing, S. A., Lang, D., ... Shaw, A. J. (2011). Generation-biased gene expression in a bryophyte model system. *Molecular Biology and Evolution*, 28, 803–812.
- Takos, A. M., Knudsen, C., Lai, D., ... Rook, F. (2011). Genomic clustering of cyanogenic glucoside biosynthetic genes aids their identification in *Lotus japonicus* and suggests the repeated evolution of this chemical defence pathway. *Plant Journal*, 68, 273–286.
- Tam, T. H. Y., Catarino, B., & Dolan, L. (2015). Conserved regulatory mechanism controls the development of cells with rooting functions in land plants. *Proceedings of the National Academy of Sciences of the United States of America*, 112, E3959–E3968.
- Tcherkez, G. G. B., Farquhar, G. D., & Andrews, T. J. (2006). Despite slow catalysis and confused substrate specificity, all ribulose biphosphate carboxylases may be nearly perfectly optimized. *Proceedings of the National Academy of Sciences of the United States of America*, 103, 7246–7251.
- Tian, W., & Skolnick, J. (2003). How well is enzyme function conserved as a function of pairwise sequence identity? *Journal of Molecular Biology*, 333, 863–882.
- Tomescu, A. M. F. (2009). Megaphylls, microphylls and the evolution of leaf development. *Trends in Plant Science*, 14, 5–12.
- Tomitani, A., Okada, K., Miyashita, H., ... Tanaka, A. (1999). Chlorophyll b and phycobilins in the common ancestor of cyanobacteria and chloroplasts. *Nature*, 400, 159–162.
- Töpfer, N., Caldana, C., Grimbs, S., ... Nikoloski, Z. (2013). Integration of genome-scale modeling and transcript profiling reveals metabolic pathways underlying light and temperature acclimation in *Arabidopsis*. *The Plant Cell*, 25, 1197–1211.
- Villarreal, J. C., & Renner, S. S. (2012). Hornwort pyrenoids, carbon-concentrating structures, evolved and were lost at least five times during the last 100 million years. *Proceedings of the National Academy of Sciences of the United States of America*, 109, 18873–18878.
- Vogt, T. (2010). Phenylpropanoid biosynthesis. *Molecular Plant*, 3, 2–20.
- Wada, M., Takahashi, H., Altaf-Ul-Amin, M., ... Kanaya, S. (2012). Prediction of operon-like gene clusters in the *Arabidopsis thaliana* genome based on co-expression analysis of neighboring genes. *Gene*, 503, 56–64.
- Wang, C., Liu, Y., Li, S. S., & Han, G. Z. (2015). Insights into the origin and evolution of the plant hormone signaling machinery. *Plant Physiology*, 167, 872–886.
- Wang, X., Ding, J., Lin, S., ... Wu, H. (2020). Evolution and roles of cytokinin genes in angiosperms 2: Do ancient CKXs play housekeeping roles while non-ancient CKXs play regulatory roles? *Nature Horticultural Research*, 7.
- Waters, M. T., Gutjahr, C., Bennett, T., & Nelson, D. C. (2017). Strigolactone signaling and evolution. *Annual Review of Plant Biology*, 68, 291–322.
- Weissenborn, S., & Walther, D. (2017). Metabolic pathway assignment of plant genes based on phylogenetic profiling – a feasibility study. *Frontiers in Plant Science*, 8, 1831.

- Weng, J. (2014). The evolutionary paths towards complexity: A metabolic perspective. *The New Phytologist*, 201, 1141–9.
- Weng, J., Akiyama, T., Bonawitz, N. D., ... Chapple, C. (2010). Convergent evolution of syringyl lignin biosynthesis via distinct pathways in the lycophyte *Selaginella* and flowering plants. *The Plant Cell*, 22, 1033–45.
- Weng, J. K., Li, X., Stout, J., & Chapple, C. (2008). Independent origins of syringyl lignin in vascular plants. *Proceedings of the National Academy of Sciences of the United States of America*, 105, 7887–7892.
- Wickett, N. J., Mirarab, S., Nguyen, N., ... Leebens-Mack, J. (2014). Phylotranscriptomic analysis of the origin and early diversification of land plants. *Proceedings of the National Academy of Sciences of the United States of America*, 111, E4859–E4868.
- Wink, M. (2008). Plant secondary metabolism: Diversity, function and its evolution. *Natural Product Communications*, 3.
- Winter, K., Holtum, J. A. M., & Smith, J. A. C. (2015). Crassulacean acid metabolism: A continuous or discrete trait? *New Phytologist*, 208, 73–78.
- Winzer, T., Gazda, V., He, Z., ... Graham, I. A. (2012). A *Papaver somniferum* 10-gene cluster for synthesis of the anticancer alkaloid noscapine. *Science*, 336, 1704–1708.
- Woodson, J. D., & Chory, J. (2008). Coordination of gene expression between organellar and nuclear genomes. *Nature Reviews Genetics*, 9, 383–395.
- Yang, W., Simpson, J. P., Li-Beisson, Y., ... Ohlrogge, J. B. (2012). A land-plant-specific glycerol-3-phosphate acyltransferase family in *Arabidopsis*: Substrate specificity, *sn*-2 preference, and evolution. *Plant Physiology*, 160, 638–652.
- Yokota, T., Ohnishi, T., Shibata, K., ... Kohchi, T. (2017). Occurrence of brassinosteroids in non-flowering land plants, liverwort, moss, lycophyte and fern. *Phytochemistry*, 136, 46–55.
- Yoon, H. S., Hackett, J. D., Ciniglia, C., ... Bhattacharya, D. (2004). A molecular timeline for the origin of photosynthetic eukaryotes. *Molecular Biology and Evolution*, 21, 809–818.

2

Constraint-based modelling reveals a high carbon conversion efficiency in *Marchantia polymorpha* compared to that in C₃ angiosperms

2.1 MOTIVATION

Bryophytes exhibit significantly lower carbon assimilation rates than spermatophytes, as a consequence of their extremely low non-stomatal conductance which is thought to have evolved as an adaptation to reduce water loss in an early terrestrial environment. In this paper, we hypothesise that there may be adaptations to this difference in resource availability between bryophytes and spermatophytes which result in alternative core metabolic fluxes in these species. We explore this by constructing stoichiometric models of core metabolism for the bryophyte *Marchantia polymorpha* and the angiosperm *Arabidopsis thaliana* and using flux-balance analysis to compare carbon metabolism in each species. This paper has been submitted for publication in *Plants in Silico*.

2.2 CONTRIBUTION

A signed statement of authorship can be found at the end of this chapter. Naomi Cannell and Lee J. Sweetlove conceived the project and developed the methodology; N.C wrote code

and carried out the investigation and formal analysis; L.J.S and Sanu Shameer provided supervision; N.C. wrote the manuscript and prepared visualisations; L.J.S reviewed and edited the manuscript.

2.3 SUMMARY

Bryophytes were the first clade of land plants to evolve from an aquatic ancestor. To limit and tolerate the stress of desiccation they developed thick cell walls, which consequently reduce non-stomatal diffusion of CO₂ into the chloroplast. This low CO₂ conductance is thought to be the cause of the very low photosynthetic CO₂ assimilation rates observed in bryophytes compared to spermatophytes. Here, we use genomic and metabolic data to construct an experimentally-constrained stoichiometric model of central metabolism in the liverwort *Marchantia polymorpha*. We compared flux predictions from this model to those obtained from an experimentally-constrained stoichiometric model of the angiosperm *Arabidopsis thaliana*. We hypothesised that the difference in carbon availability between bryophytes and angiosperms would have consequences for central metabolic fluxes. From the model predictions we identified a number of metabolic pathways exhibiting alternative flux modes between the two species. By comparing predicted rates of biomass synthesis in our models against experimental growth rates, we demonstrate that *Marchantia* must exhibit higher carbon conversion efficiency (CCE) than C₃ angiosperms. Based on the models, we propose two mechanisms for generating high CCE in *Marchantia*: either *Marchantia* has an unusually low cell maintenance cost or *Marchantia* fixes respiratory CO₂ during the night, or a combination of the two. We suggest that bryophytes may be an as-yet-uninterrogated source of carbon-use-efficient metabolism that could provide novel strategies for exploitation in the metabolic engineering of crops for increased yield.

2.4 INTRODUCTION

Bryophytes are the most ancient clade of land plants, diverging more than 450 million years ago from the common ancestor of the embryophytes (Harris et al. 2020). Compared to the ancestral aquatic environment, terrestrial habitats allowed improved access to CO₂ for photosynthesis but limited water availability and increased opportunity for water loss

(Hanson et al. 2014). The later-diverging tracheophytes developed roots and vascular tissues, providing consistent access to water beneath the soil surface, and stomata, which are under fine regulatory control to reduce water loss (Delwiche and Cooper 2015; de Vries and Archibald 2018). Bryophytes lack these innovations in their vegetative gametophore, and under the evolutionary pressures of an early terrestrial environment developed, and have maintained, thick cell walls to increase resistance to water loss and to structurally support the plant under the stress of desiccation (Carriqui et al. 2019; Waite and Sack 2010).

A consequence of this adaptation to reduce water loss in bryophytes is high resistance to CO₂ diffusing from the atmosphere into the chloroplast stroma (Carriqui et al. 2019). Although some bryophytes have evolved strategies to improve this water loss/CO₂ assimilation trade-off (e.g. complex thalloid liverworts improve surface area for CO₂ absorption while reducing water loss by enclosing mesophyll cell regions within air chambers gated by hydrophobic air pores) (Green and Snelgar 1982), low non-stomatal conductance is thought to be the major factor limiting photosynthetic carbon assimilation in these species, which exhibit significantly lower net photosynthetic rates than spermatophytes (Flexas and Carriqui 2020; Gago et al. 2019). We considered that this reduced carbon budget may result in carbon-limited growth in bryophytes (contrary to spermatophytes, which are typically limited by mineral nutrient availability (Reich et al. 2006; Rubio-Asensio and Bloom 2017)). Indeed, CO₂ fertilisation has been shown to improve growth rates in liverworts (Humphreys et al. 2010), and the carbon concentrating mechanism that has evolved in some hornworts further supports this suggestion (Li et al. 2017; Villarreal and Renner 2012). We hypothesised that there may be adaptations to these fundamental differences in resource availability resulting in alternative core metabolic fluxes in bryophytes and higher plants.

To address this, we investigated potential flux distributions in central metabolism using a comparative computational approach based on flux-balance analysis (FBA) of experimentally-constrained stoichiometric models of the central metabolic networks of the bryophyte *Marchantia polymorpha* and the angiosperm *Arabidopsis thaliana*. FBA has been used extensively to analyse fluxes in large-scale metabolic networks across a range of plant organs and broad temporal scales (Gomes de Oliveira Dal'Molin and Nielsen 2018; Moreira et al. 2019; Sweetlove and Ratcliffe 2011) in various species from bacteria to angiosperms

(Arnold and Nikoloski 2014; Park et al. 2011; Saha et al. 2011), but to our knowledge has not yet been applied to bryophyte metabolism. Several bryophyte whole genome sequences have become available in recent years (Bowman et al. 2017; Rensing et al. 2008; Zhang et al. 2020), facilitating the construction of stoichiometric models of metabolism for these species. We chose *Marchantia polymorpha* as the subject of our investigation due to the additional availability of metabolic and cellular compositional information in this species (Dembitsky et al. 1993; Happ and Classen 2019; Shinmen et al. 1991).

2.5 METHODS

2.5.1 Constructing a stoichiometric model of primary metabolism in *Marchantia*

The fully compartmentalised, mass-balanced and charge-balanced model of conserved reactions of plant primary metabolism (PlantCoreMetabolism_v1_2) described by Shameer et al. (2018) was used as the starting point for metabolic modelling of *Arabidopsis* and *Marchantia*. The core model comprises all reactions required to produce the main components of biomass, either autotrophically or heterotrophically and hence represents central metabolism. These reactions are highly conserved and it was previously demonstrated that the reactions in PlantCoreMetabolism_v1_2 are present in all species in the PlantCyc metabolic database (Schl pfer et al. 2017; Shameer et al. 2018). However, since PlantCoreMetabolism_v1_2 was developed from *Arabidopsis*, we took steps to ensure the representation of relevant metabolic reactions for *Marchantia*. This was done by extracting all *Arabidopsis* genes associated with the reactions in PlantCoreMetabolism_v1_2 and identifying whether homologues exist in the *Marchantia* genome using OrthoFinder software (Cannell et al. 2020; Emms and Kelly 2019). Genes for five reactions were found to be absent in *Marchantia* and these reactions (with MetaCyc identifiers RXN-1121, RXN-10773, MYO-INOSITOL-OXYGENASE-RXN, CONIFERIN-BETA-GLUCOSIDASE-RXN and 2.4.1.111-RXN) were removed from the *Marchantia* model. Additionally, as cell membranes in *Marchantia* contain a number of long-chain fatty acids which are absent from *Arabidopsis* under normal circumstances, a further 57 reactions and 56 metabolites were added to the starting model to allow the generation of these fatty acids (the full set of additional reactions and metabolites are listed in Data 2.1).

2.5.2 Constructing biomass reactions for *Arabidopsis* and *Marchantia*

Predicting rates of biomass accumulation in a stoichiometric model requires a biomass reaction which combines individual biomass components (e.g. amino acids, cell wall polymers and fatty acids) as ‘reactants’ in the correct relative stoichiometries that match experimental determinations of cellular biomass composition for the species in question. As our aim was to investigate metabolism during vegetative growth in *Arabidopsis* and *Marchantia*, we focused on developing biomass reactions specific to *Arabidopsis* leaf tissue and *Marchantia* thallus tissue. The overall relative quantities of major biomass components for the *Arabidopsis* leaf model were taken from (Poolman et al. 2009), while more detailed fractional composition for *Arabidopsis* (e.g. ratios of individual amino acids) and all biomass composition information for *Marchantia* thallus were identified using an in-depth search of existing literature (Arnold and Nikoloski 2014; Bowman et al. 2017; Carella et al. 2019; Dembitsky et al. 1993; Happ and Classen 2019; Koide et al. 2020; Miquel and Browse 1992; Shinmen et al. 1991; Thomas et al. 1984; Zablackis et al. 1995). Where information was not available for *Marchantia* (for total protein content and soluble metabolite content and composition), biomass component proportions were taken from *Arabidopsis* to use in the *Marchantia* model. For simplicity, in both models hemicellulose and pectin polymers were represented as the monomers xylan and homogalacturonan respectively. The overall biomass composition for each species was used to calculate the molar mass of biomass in each species, using a weighted average of the molar masses of each biomass component and their relative contribution to biomass. The biomass composition for each species and their calculated stoichiometries can be found in Data 2.2 and 2.3.

2.5.3 Estimating non-growth-associated maintenance

The energetic cost of non-growth-associated maintenance (NGAM) for both species was modelled by way of constraints that lead to flux through ATP hydrolysis and NADPH oxidase reactions. The proportions of ATP hydrolysis and NADPH oxidase were constrained in a 3:1 ratio, shown to produce the most realistic flux distributions in a comparison of *Arabidopsis* FBA and *in vivo* fluxes (Cheung et al. 2014). Töpfer et al. (2020) developed a linear equation showing how NGAM costs vary with light intensity in angiosperms (ATPase

flux ($\mu\text{mol}/\text{m}^2/\text{s}$) = $(0.0049 \times \text{light intensity } (\mu\text{mol photons}/\text{m}^2/\text{s})) + 2.7851$). This equation was used to set NGAM for the *Arabidopsis* model but is not applicable to bryophytes, which have lower carbon assimilation rates than most higher plants (Fletcher et al. 2006; Gimeno et al. 2017). For *Marchantia*, a fixed value for NGAM was used for modelling across light intensities. This was estimated using data from Katoh (1983), where respiration rates were measured for a mature (non-growing) photomixotrophic culture of *Marchantia* cells in the dark. As the cells were not growing, the respiration rate reflects the energy demand for maintenance at night. A single-phase *Marchantia* model without light input was set up and the flux through the maintenance ATPase reaction was incrementally increased until the modelled respiration rate equalled the average respiration rate of the experimental culture. This ATPase flux was then used as the NGAM for the *Marchantia* model, with the 3:1 ratio (assumed to hold true in *Marchantia*) defining flux through NADPH oxidase. NGAM costs were assumed to be equal in the day and night phases for each model.

2.5.4 Setting up diel models of *Arabidopsis* and *Marchantia*

To produce diel models capable of optimising flux over both day and night phases in a single optimisation, the previously established diel FBA framework was used (Cheung et al. 2014). All metabolites and reactions in the model were duplicated and their identifiers modified to indicate their assignment to the day or night phase. Reactions were added to both models to allow diel storage of starch, amino acids and soluble metabolites, and pseudometabolites were included to constrain certain reactions during the day and night phases to a specific flux ratio. In the absence of bryophyte-specific data, all of these flux-ratio constraints were applied to the *Marchantia* model as well as the *Arabidopsis* model. These include constraining day: night nitrate uptake ratio to 3:2, constraining RuBisCO carboxylation: oxygenation ratio to 3:1 and constraining day: night biomass output to 3:1.

2.5.5 Validating model predictions against experimental data

We identified two photosynthetic CO_2 assimilation datasets each for *Arabidopsis* (Sun et al. 1999; Tanaka et al. 2013) and *Marchantia* (Fletcher et al. 2006; Gimeno et al. 2017). Data were extracted from the graphs using WebPlotDigitizer (WebPlotDigitizer 2020), and using the drc package in R a curve of the form $A = \frac{(x \times PPF D)}{(y + PPF D)}$ was fitted to each dataset, where x and y

are constants, *PPFD* refers to the light intensity in $\mu\text{mol}/\text{m}^2/\text{s}$ and *A* is net photosynthetic CO_2 assimilation in $\mu\text{mol}/\text{m}_2/\text{s}$. The fitted curves can be found in Figure A.1.

We identified four studies in *Arabidopsis* (Christophe et al. 2008; Cookson et al. 2006, 2005; Rasse and Tocquin 2006) and two in *Marchantia* (Carter and Romine 1969; Graham and Dixon 2012) which provide data on photoautotrophic growth rates of leaves and thallus tissue respectively. The growth rates between each pair of consecutive timepoints in the datasets was used to calculate the mean relative growth rate. This value (in units of leaf/thallus area or grams dry mass) was converted to units of $\mu\text{mol}/\text{m}^2/\text{s}$ using the molecular mass of biomass as calculated previously for *Arabidopsis* and *Marchantia*. If required, values of leaf/thallus mass per area were taken from Weraduwege et al. (2015) for *Arabidopsis* and Voth (1943) for *Marchantia*. Additionally, if the experiments were carried out in a photoperiod which deviated from 12 hours, the average growth rate was adjusted to represent a 12-hour photoperiod, assuming that overall biomass output during the day and night occurs in a 3:1 ratio (Gibon et al. 2004). The photosynthetic photon flux density (PPFD) for each set of experimental conditions was passed into the CO_2 assimilation curves fitted above and the CO_2 assimilation rate for each PPFD extracted. The average CO_2 assimilation rate was taken from the two curves for each species to produce an estimated CO_2 assimilation rate for each set of conditions. Table A.1 shows the sources of experimental data in *Arabidopsis* and *Marchantia* and the relevant experimental conditions.

To compare the experimental growth rates for each species to those predicted by the models, the diel models for *Arabidopsis* and *Marchantia* were set up to run with additional constraints: limiting available photon flux to the experimental PPFD, and matching CO_2 assimilation to the estimated CO_2 assimilation for each experimental condition. With these constraints applied, the models were run using parsimonious FBA (pFBA) (Lewis et al. 2012), with the maximisation of biomass production as the objective function, while also minimising the overall sum of fluxes. Python functions to set up the model and run pFBA (as in Shameer et al. (2019)) can be found in Data 2.4. The diel biomass output of each model (in units of μmol biomass/ m^2/s) was then compared to *in vivo* growth rates from the corresponding experiment.

2.5.6 Investigating reaction fluxes

Flux solutions from pFBA optimisation are not necessarily unique with respect to individual reaction fluxes (Sweetlove and Ratcliffe 2011), so to investigate flux differences through specific reactions in the *Arabidopsis* and *Marchantia* models we carried out a flux variability analysis (FVA) (Gudmundsson and Thiele 2010) to identify the upper and lower flux bounds through each reaction. Reactions were considered to carry differing fluxes between the two models if the flux ranges did not overlap.

2.5.7 Data availability

Stoichiometric models of *Arabidopsis* and *Marchantia* in SBML format can be found in Data 2.5 and 2.6. Python functions to run the models can be found in Data 2.4 and the Sweetlove Lab github at <https://github.com/ljs1002>, and a list of constraints applied can be found in Table 2.1.

2.6 RESULTS

2.6.1 Model framework and characteristics

Given recent evidence that bryophytes experience physiological barriers to carbon uptake (Flexas and Carriqui 2020; Gago et al. 2019), we set up two diel models of primary metabolism for a growing *Arabidopsis* leaf and *Marchantia* thallus with the aim of investigating the metabolic consequences of this reduced carbon budget (Figure 2.1). The *Arabidopsis* model was composed of 1985 reactions and 1881 metabolites and the *Marchantia* model contained 1975 reactions and 1877 metabolites, both comprising day- and night-phase duplicates of all those required to produce vegetative growth in an *Arabidopsis* leaf and a *Marchantia* thallus over a diurnal cycle, as well as linker reactions allowing the accumulation of metabolites in one phase for use in the other. In both models, constraints were applied to the quantities and ratios of flux through specific reactions. These were used to impose known flux limitations within which the models must operate to ensure realistic flux distributions. Table 2.1 lists these constraints.

The main distinguishing feature of the *Arabidopsis* and *Marchantia* models is the biomass reaction, which defines the proportions of metabolites output by each model as the known

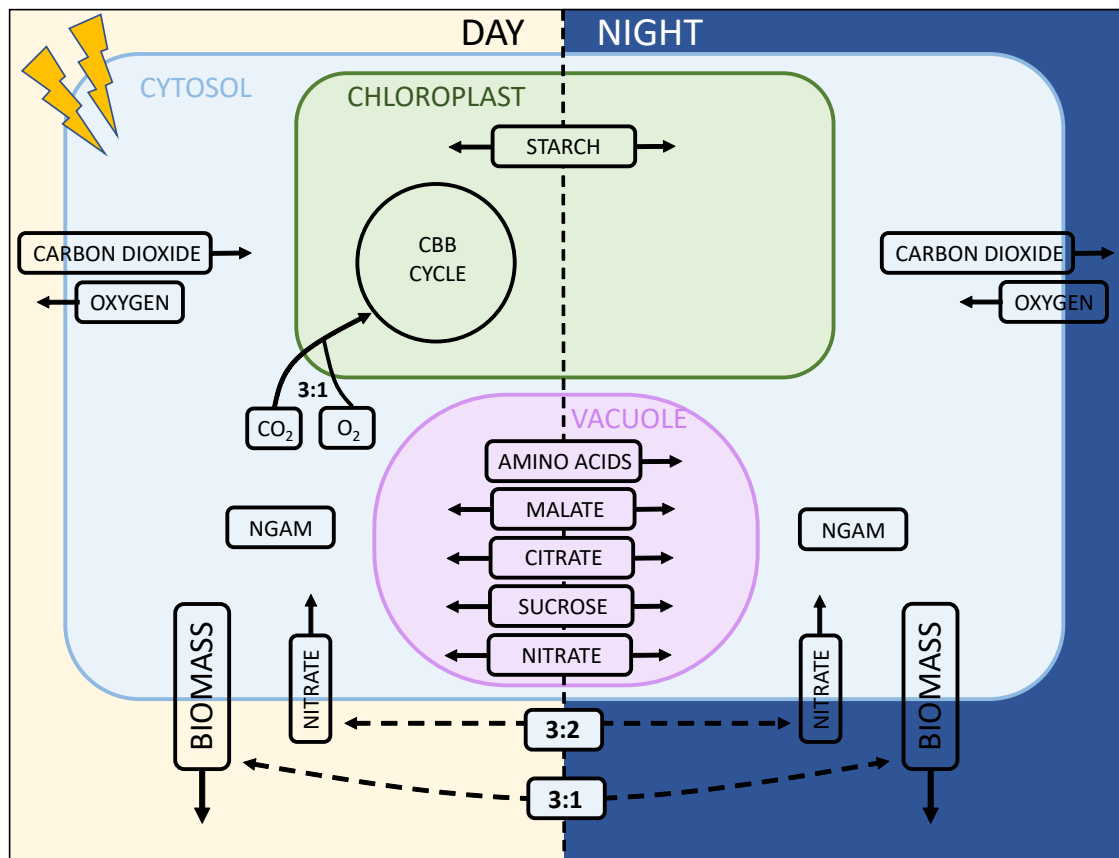


Figure 2.1: Basic schematic of *Arabidopsis* and *Marchantia* stoichiometric models, showing inputs, outputs and constraints applied. NGAM, non-growth-associated maintenance.

relative contributions of each metabolite to cellular biomass in each specific plant tissue. The biomass reactions included here account for the relative quantities and molecular composition of cell walls, lipids, soluble metabolites, proteins and starch in each species, and were constructed based on an extensive survey of the existing literature. The relative experimentally-defined quantities of each biomass component included in each model are shown in Table 2.2. Cell wall composition in *Arabidopsis* was identified from Zabackis et al. (1995) while total cell wall content in *Arabidopsis* was taken from Poolman et al. (2009). Cell wall composition of *Marchantia* was calculated from Thomas et al. (1984) and Happ and Classen (2019) while total cell wall content was taken from Thomas et al. (1984). Starch content in *Arabidopsis* was taken from Poolman et al. (2009) while starch content in *Marchantia* was taken from Koide et al. (2020). Overall protein content in *Arabidopsis* was taken from Poolman et al. (2009) and assumed to be the same in *Marchantia*, while amino acid contents were taken from Arnold and Nikoloski (2014) for *Arabidopsis* and calculated

Constraint	Python Command	Reason for application of constraint
Light input during the night set to 0	<code>model.reactions.Photon_tx2. upper_bound = 0</code>	To allow light and dark phases to be modelled concurrently with a 12hr/12hr photoperiod.
Flux through Calvin-Benson-Bassham cycle at night set to 0	As above, set upper and lower bounds to 0 for the following reactions: RIBULOSE_BISPHOSPHATE_CARBOXYLASE_RXN_p2, PHOSGLYPHOS_RXN_p2, 1_PERIOD_2_PERIOD_1_PERIOD_13_RXN_p2, TRIOSEPISOMERIZATION_RXN_p2, F16ALDOLASE_RXN_p2, F16BDEPHOS_RXN_p2, 2TRANSKETO_RXN_p2, RIBULP3EPIM_RXN_p2, PHOSPHORIBULOKINASE_RXN_p2, RIB5PISOM_RXN_p2, 1TRANSKETO_RXN_p2, SEDOHEPTULOSE_BISPHOSPHATASE_RXN_p2, SEDOBISALDOL_RXN_p2	There is evidence of downregulation of these enzymes during the night in the plastid of higher plants (Michelet et al. 2013).
Nitrate uptake constrained to 3:2 day: night ratio	See function <code>SetupDielModel</code> (Data 2.4) lines 140-142	Based on measurements in several angiosperms (Delhon et al. 1995; Macduff and Bakken 2003).
Biomass output constrained to 3:1 day: night ratio	See function <code>SetupDielModel</code> (Data 2.4) lines 33-43	Based on measurements in <i>Arabidopsis</i> (Gibon et al. 2004).
RuBisCO carboxylase/oxygenase activity constrained in a 3:1 ratio	See function <code>SetupDielModel</code> (Data 2.4) lines 146-148	Based on experimentally measured fluxes (Ma et al. 2014).
Accumulation of amino acids is allowed only during the day	See function <code>SetupDielModel</code> (Data 2.4) lines 46-60	Based on measurements in tobacco leaves (Scheible et al. 2000).
CO ₂ fixation during the night not permitted	See function <code>estimateOutputFromNetCO2</code> (Data 2.4)	The FBA procedure will lead to a maximally optimised model – if not constrained this results in the model re-fixing all CO ₂ at night which is unrealistic for C ₃ organisms.

Constraint	Python Command	Reason for application of constraint
Flux through NGAM reactions (ATP hydrolase and NADPH oxidase) constrained to a specific value for each model. Flux through ATP hydrolase and NADPH oxidase reactions constrained in a 3:1 ratio.	Executed once for each phase: <pre> model.reactions.ATPase_tx.lower_ bound = NGAM model.reactions.NADPHoxc_tx.lower_ bound = NGAM/9 model.reactions.NADPHoxm_tx.lower_ bound = NGAM/9 model.reactions.NADPHoxp_tx.lower_ bound = NGAM/9 </pre>	A portion of energy produced from respiration must be used to maintain existing cellular structures in real plant tissue. Without constraining flux through these reactions they will not run in an optimised model. 3:1 ratio taken from (Cheung et al. 2013).

Table 2.1: Flux constraints applied to *Marchantia* and *Arabidopsis* models.

from proteome data taken from Carella et al. (2019) and protein sequences from Bowman et al. (2017). Total lipid content was taken from Poolman et al. (2009) for *Arabidopsis* and from Dembitsky et al. (1993) for *Marchantia*, while individual fatty acid contents were taken from Miquel and Browse (1992) for *Arabidopsis* and Shinmen et al. (1991) for *Marchantia*. Soluble metabolite content was taken from Williams et al. (2010) for *Arabidopsis* and were assumed to be the same in *Marchantia* ¹.

The biomass compositions of *Arabidopsis* and *Marchantia* are broadly similar, but there were some notable differences. For example, total cell wall polysaccharide content is higher in *Arabidopsis*, and cell walls in *Arabidopsis* are composed of more hemicellulose and pectin and less cellulose than cell walls in *Marchantia*. Likewise, fatty acids comprise a greater proportion of biomass in *Arabidopsis*, but while *Arabidopsis* fatty acids never contain more than 18 carbons, *Marchantia* contains several 20C fatty acids. Taking an average of the molecular masses of all biomass components for each species, weighted by their percentage contribution to cellular biomass, produces a ‘molecular mass’ of biomass for each species – this was calculated at 174.8g/mol for *Arabidopsis* and 162.4g/mol for *Marchantia*.

¹Note that the biomass content of fumarate is likely underestimated here due to measurements taken from cell suspension cultures rather than *Arabidopsis* leaves. This is unlikely to affect the capacity for biomass production in either model due to the interchangeability of malate and fumarate as carbon skeletons in central metabolism.

Biomass component	Proportion of biomass in <i>Arabidopsis</i> dry mass)	Proportion of biomass in <i>Marchantia</i> dry mass)	Biomass component	Proportion of biomass in <i>Arabidopsis</i> dry mass)	Proportion of biomass in <i>Marchantia</i> dry mass)
Cellulose	7.525	15.3	Palmitic acid (16:0)	1.65	2.436
Hemicellulose (xylan)	22.575	15	Palmitoleic acid (16:1)	0.418	0.199
Pectin (homogalacturonan)	12.9	6	Hexadecadienoic acid (16:2)	0.121	0
Total cell wall	43	36.3	Hexadecatrienoic acid (16:3)	1.518	0.027
Non-diel-associated starch	5	0.5	Stearic acid (18:0)	0.11	0.027
Alanine	1.685	1.791	Oleic acid (18:1)	0.385	1.07
Arginine	0.967	1.089	Linoleic acid (18:2)	1.727	0.321
Asparagine	0.729	0.716	α -linolenic acid (18:3)	5.06	0.369
Aspartic acid	1.095	1.092	Dihomo- γ -linolenic acid (20:3)	0	0.077
Cysteine	0.309	0.293	Arachidonic acid (20:4)	0	0.812
Glutamine	0.597	0.754	Eicosapentaenoic acid (20:5)	0	0.959
Glutamic acid	1.374	1.335	Total lipids	11	6.3
Glycine	1.679	1.545	Soluble glutamic acid	1.45	1.45
Histidine	0.423	0.411	Soluble glutamine	1.6	1.6
Isoleucine	0.965	0.918	Soluble alanine	1.29	1.29
Leucine	1.707	1.801	Calcium	1.1	1.1
Lysine	1.239	1.164	Magnesium	0.44	0.44
Methionine	0.408	0.47	Potassium	2.2	2.2
Phenylalanine	0.893	0.757	Citrate	1.6	1.6
Proline	1.015	1.007	GABA	0.28	0.28
Serine	1.264	1.563	Succinate	0.22	0.22
Threonine	1.272	1.045	Soluble aspartate	0.38	0.38
Tryptophan	0.277	0.24	Sucrose	1.51	1.51
Tyrosine	0.695	0.534	Malate	1.82	1.82
Valine	1.404	1.475	Glycerol	0.09	0.09
Total protein	20	20	Fumarate	0.02	0.02
			Total soluble metabolites	14	14

Table 2.2: Experimentally-derived biomass composition of *Arabidopsis* and *Marchantia* used to specify the stoichiometry of the biomass reaction in each species' model. Note that as a bryophyte, *Marchantia* does not contain lignin. Lignin content in *Arabidopsis* leaves was considered negligible relative to other cell wall components.

NGAM refers to the respiration cost of essential processes not related to the production of biomass, for example protein turnover and the maintenance of ion gradients across membranes (Amthor 1984, 2000), and is represented in the models as flux through ATP hydrolysis and NADPH oxidase reactions. Different methods of calculating NGAM fluxes were used for each model. For *Arabidopsis*, a linear equation calculating NGAM for higher plants (Töpfer et al. 2020) was used to produce constraints for NGAM dependent on the light available to the plant. Using this equation to calculate NGAM for *Marchantia* resulted in values of NGAM incompatible with the much lower growth rates characteristic of *Marchantia*, and infeasible models when constrained to accurate carbon assimilation rates for *Marchantia*. Instead, since there is no existing experimental data regarding NGAM for *Marchantia*, NGAM for *Marchantia* was estimated from experimental night-respiration rates in non-growing cells (Katoh 1983), producing a flux constraint of $1.895 \mu\text{mol}/\text{m}^2/\text{s}$ through ATP hydrolase. Although this level of NGAM is proportionally higher, relative to carbon-use in *Marchantia*, than NGAM in *Arabidopsis*, when considered per unit mass it is more than 50% lower than the *Arabidopsis* NGAM under any of the experimental conditions considered here. This difference stems from the fact that carbon assimilation in *Marchantia* per unit mass is lower than that in *Arabidopsis*.

2.6.2 Experimental data validates our model of *Arabidopsis* metabolism, but *Marchantia* growth rates cannot be successfully met using equivalent constraints within the *Marchantia* metabolic model

To validate flux predictions from each model, the *Arabidopsis* and *Marchantia* models were run using CO_2 assimilation and PPFD constraints taken from experimental data and the experimentally-determined biomass compositions. The predicted growth rates (i.e. rates of biomass production) were compared to those measured experimentally under the corresponding growth conditions. In each case, the CO_2 assimilation rate for each species under each set of experimental conditions was calculated from the light response curves in Figure A.1 using the light intensity stated for each study. Four studies were identified for *Arabidopsis* (Christophe et al. 2008; Cookson et al. 2006, 2005; Rasse and Tocquin 2006) and two for *Marchantia* (Carter and Romine 1969; Graham and Dixon 2012) which present growth rates for each species under various experimental conditions.

The average experimental growth rate for each study was calculated and converted into units compatible with the models, i.e. $\mu\text{mol biomass/m}^2/\text{s}$. The final value was adjusted to account for differences in day length, for comparison to the modelled 12hr/12hr photoperiod. Table 2.3 shows the PPFD, calculated CO_2 assimilation level and experimental growth rate for each study compared with the light- and CO_2 -use of each model and the predicted growth rate. Figure 2.2 illustrates these modelled growth predictions for *Arabidopsis* and *Marchantia* as a proportion of the experimental growth rate achieved under each set of experimental conditions.

Species	Study	Light available ($\mu\text{mol Photons/m}^2/\text{s}$)	Calculated carbon assimilation ($\mu\text{mol CO}_2/\text{m}^2/\text{s}$)	Experimental growth rate ($\mu\text{mol biomass/m}^2/\text{s}$)	Predicted light use ($\mu\text{mol photons/m}^2/\text{s}$)	Predicted carbon assimilation ($\mu\text{mol CO}_2/\text{m}^2/\text{s}$)	Predicted growth rate ($\mu\text{mol biomass/m}^2/\text{s}$)
<i>Arabidopsis</i>	Rasse & Tocquin, 2006	120	4.08	0.393	67.3	4.08	0.517
<i>Arabidopsis</i>	Christophe et al, 2008	92.6	3.37	0.409	56.4	3.37	0.399
<i>Arabidopsis</i>	Cookson et al, 2005	108.8	3.8	0.596	63	3.8	0.471
<i>Arabidopsis</i>	Cookson et al, 2006	149.3	4.74	0.587	77.6	4.74	0.626
<i>Marchantia</i>	Carter & Romine, 1969	174	1.16	0.238	20.7	1.16	0.105
<i>Marchantia</i>	Graham & Dixon, 2012	95	0.794	0.0998	13.8	0.794	0.0124

Table 2.3: Comparison of experimental and predicted light availability, carbon use and growth rates for the *Arabidopsis* and *Marchantia* models under different experimental conditions.

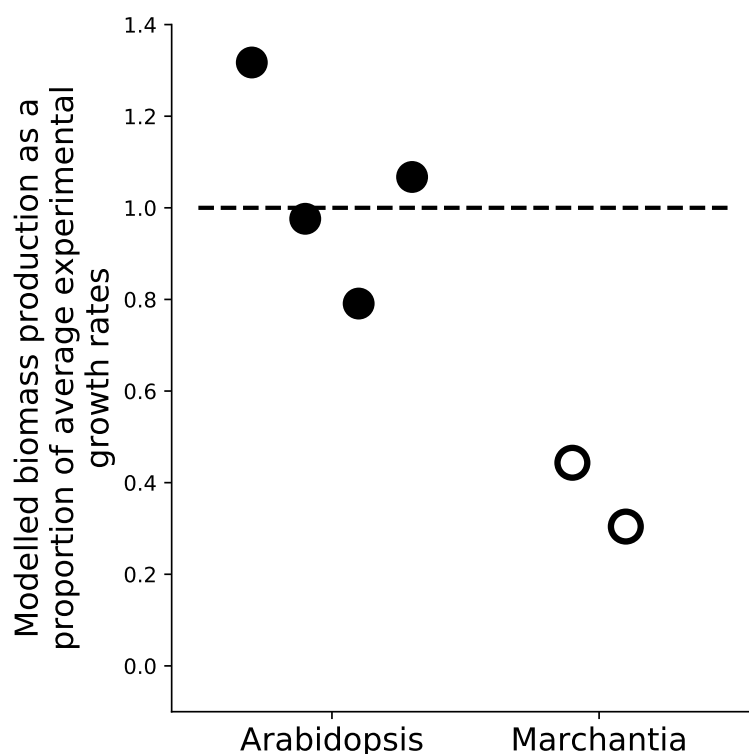


Figure 2.2: Comparing predicted growth rates as a proportion of experimental growth rates between the *Arabidopsis* and *Marchantia* models. Each datapoint represents the predicted growth rate as a proportion of the experimentally measured growth rate for each species under a specific set of experimental conditions. Filled circles show *Arabidopsis* growth rates and open circles show *Marchantia* growth rates. The dashed horizontal line indicates the point at which the predicted growth rate would be 100% of the experimentally measured growth rate.

As indicated by Table 2.3, neither of the models are light-limited – there is more light available to each model than is required for each set of experimental CO₂ assimilation rates. On the other hand, the models are both limited by CO₂ assimilation – both models maximise use of the carbon available (as expected for a model with the objective function of maximising biomass production). Figure 2.2 and Table 2.3 show that the *Arabidopsis* model is able to consistently match experimental growth rates – the maximum deviation from experimental growth rates produced by the *Arabidopsis* model is 31.7% (at 120 $\mu\text{mol photons/m}^2/\text{s}$) while in the best case the modelled growth rates are within 2% of the experimental growth rates (at 92.6 $\mu\text{mol photons/m}^2/\text{s}$). This indicates that although there are small differences between experimental and predicted growth rates under some experimental conditions (likely due to slight inaccuracies in e.g. leaf mass per area, CO₂ assimilation rates and biomass composition),

our *Arabidopsis* model appears to accurately represent core metabolism and growth in a real *Arabidopsis* leaf.

Conversely, Figure 2.2 shows large deviations between experimental and predicted growth rates for *Marchantia*, where the best match represents a 55% difference between these values. The same calculations are used for growth rate predictions in the *Marchantia* model as in the *Arabidopsis* model, and it is unlikely that these would have a greater detrimental effect on *Marchantia* growth predictions. However, the breadth of literature available to establish appropriate constraints for the *Marchantia* model (e.g. biomass composition, RuBisCO carboxylation: oxygenation ratio and mineral uptake ratios) is limited, and for several constraints it was assumed that *Marchantia* should be constrained in the same way as *Arabidopsis*. It is therefore possible that any of these applied constraints could be affecting the ability of the *Marchantia* model to reach experimental growth rates.

To identify whether any of the assumed constraints could result in the extent of discrepancy seen between experimental and predicted growth rates in *Marchantia*, we carried out a sensitivity analysis to investigate the impact of each constraint on biomass predictions. Five constraints were modified simultaneously in order to evaluate the robustness of the model to perturbations around their original values. The day: night ratios of nitrate uptake and biomass output, as well as the RuBisCO carboxylation: oxygenation ratio were modified to increase and decrease the relative proportions of flux for each pair of reactions from the original values. NGAM was calculated from experimental data for *Marchantia* (Kato 1983) and this was adjusted to between one quarter of and double its original value. As some of the biomass components of *Marchantia* were assumed due to lack of experimental information, the sensitivity of the model to the biomass formulation was tested by replacing the biomass composition shown in Table 2.2 with a number of separate biomass compositions, each of which maximise the relative content of a specific biomass component. For each major biomass component (protein, lipids, cell wall and starch), the overall biomass composition of the *Marchantia* model was modified to scale the content of the component in question to 50% total biomass, while maintaining both the relative proportions of all other biomass components and the total biomass accounted for in the model. Table 2.4 shows the five constraints that were modified and the values to which they were perturbed. The range

Constraint	Original value	Modified values
Rubisco carboxylase: oxygenase ratio	3:1	Ratio was varied from 0.1:1 to 10:1.
Nitrate uptake day: night ratio	3:2	Ratio was varied from 0.1:1 to 10:1.
Non-growth associated maintenance	Dependent on species and experimental condition	Flux through NGAM reactions was varied from one quarter of the original value to twice the original value.
Biomass output day: night ratio	3:1	Ratio was varied from 0.1:1 to 10:1.
Biomass composition	As calculated from literature search for each species	Each major biomass component (proteins, cell walls, starch and lipids) was increased individually to 50% of total biomass while keeping the ratio of other biomass components constant within the remaining 50%.

Table 2.4: Constraints modified during sensitivity analysis and the extent of each perturbation.

within which each constraint was varied was chosen to be broader than the ranges we considered biologically realistic to account for any possibility that these values in bryophytes may differ dramatically from those in angiosperms.

To vary each constraint simultaneously, a Latin hypercube sampling method (McKay et al. 1979) was used to produce randomised sets of five values, each representing a point in the sampling space for each varied constraint. 2000 such sets of values were produced and each set was used to constrain the *Marchantia* model before running FBA to identify the predicted biomass output under each set of constraints. The results of our sensitivity analysis for the *Marchantia* model at both investigated light intensities (95 and $174\mu\text{mol}/\text{m}^2/\text{s}$) can be found in Data 2.7 and 2.8. The results at $95\mu\text{mol}/\text{m}^2/\text{s}$ are additionally represented in Figure 2.3, which shows a parallel coordinates plot indicating the combinations of different constraint values which allow our *Marchantia* model to predict growth rates which meet or exceed experimentally defined rates at a light intensity of $95\mu\text{mol}/\text{m}^2/\text{s}$.

Out of 2000 sets of constraints analysed, the model constrained by a light intensity of $95\mu\text{mol}/\text{m}^2/\text{s}$ was only able to meet experimental growth rates 76 times (Figure 2.3).

Conversely, the model constrained with a light intensity of $174\mu\text{mol}/\text{m}^2/\text{s}$ within the same ranges of varying constraints remained unable to meet experimental growth rates across all tested sets of constraint values. For every set of constraints at which the model at the lower light intensity was able to reach experimental growth rates, the NGAM value applied to the model was lower than 0.849, less than 47% of our estimated NGAM. This is reflected in Figure 2.3 and suggests that the chosen value of NGAM has the greatest impact on modelled growth rate predictions compared to other applied constraints. The rate of NGAM directly impacts the amount of ATP and reducing power available to produce biomass, so increasing the cost of NGAM reduces biomass output and vice versa. It is clear that biomass composition also affects growth rate predictions – when biomass is composed primarily of lipids, it appears that even with the lowest NGAM values the model is unable to meet experimental growth rates. Lipids are the most energetically costly portion of biomass, requiring the use of more ATP and reducing power per unit of biomass produced, reducing the overall biomass production possible for a given CO_2 assimilation rate. However, the other three biomass compositions tested appear to show little difference from our normally-applied biomass (Table 2.2).

Changing the day: night nitrate uptake ratio had no major impact on biomass output (Figure 2.3). This can be explained by the fact that both the *Arabidopsis* and *Marchantia* models were allowed to take up unlimited nitrate during the day and night as long as the flux ratio between the two meets the 3:2 constraint. The most stoichiometrically efficient method of assimilating this nitrogen is to store all nitrate absorbed at night in the vacuole for use during the day (Cheung et al. 2014) – as a result of this, modifying the day: night uptake ratio simply prompts the model to adjust the relative proportions of nitrate assimilated during the day and night while maintaining the same overall rate of nitrate uptake for use during the day. The models behaved similarly with regards to the day: night biomass output ratio, the modification of which also had little effect on predicted growth rates. The RuBisCO carboxylase/oxygenase ratio of the models had little impact on biomass output (Figure 2.3) as long as oxygenation did not significantly exceed carboxylation (which to our knowledge does not occur *in vivo* in land plants). In the optimised stoichiometric models, carbon released via photorespiration during the day is re-assimilated – therefore there was only a small impact of photorespiration on biomass output while the energy available to the

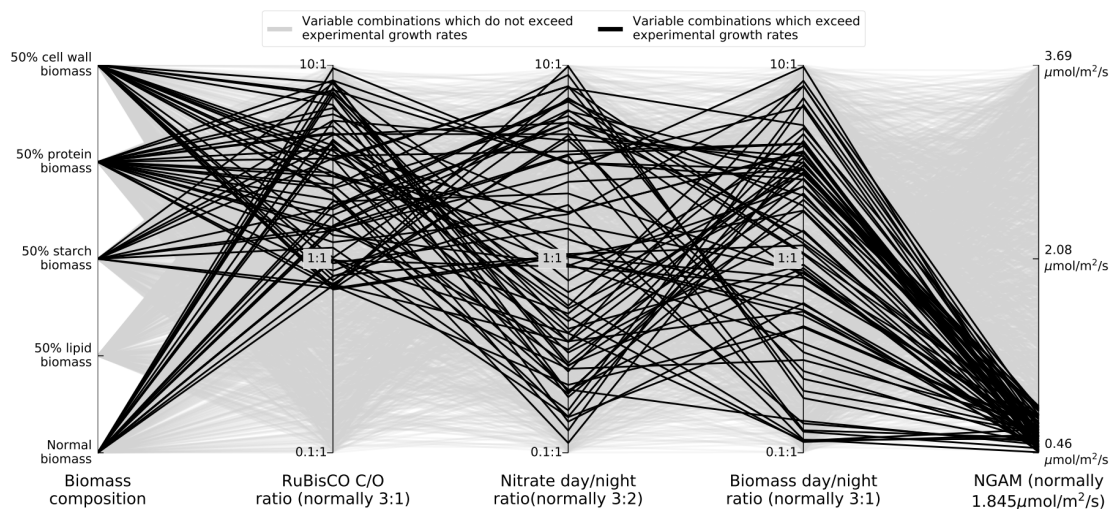


Figure 2.3: Parallel coordinates plot indicating the results of a sensitivity analysis on five constraints on our *Marchantia* model at a PPFD of $95 \mu\text{mol}/\text{m}^2/\text{s}$. The constraints varied were the biomass composition of the model (biomass compositions comprising 50% of each major biomass component were compared to normal biomass), the carboxylation: oxygenation ratio of RuBisCO, the day: night ratios of nitrate uptake and biomass production, and the energetic cost of NGAM. The x -axis shows the constraints varied during the analysis while the y -axes show the range of values within which each constraint was varied. Each plotted line connects the constraint values used for one FBA analysis on the model (2000 sets of constraints were compared). Plotted lines are shown in black when the resulting model was able to predict growth rates equal to or higher than that defined experimentally at a PPFD of $95 \mu\text{mol}/\text{m}^2/\text{s}$, while grey lines indicate combinations of constraints which do not allow the model to meet experimental growth rates. NGAM, non growth-associated maintenance.

model was high enough that the ATP- and NADH-costs of running photorespiration and re-assimilating carbon (Fernie and Bauwe 2020) were negligible. While C_3 plants *in vivo* are known to re-assimilate photorespiratory CO_2 , the extent to which this occurs appears to differ between species (Busch et al. 2013). Measurements range from about 16% in sunflower to about 42% in wheat (Pärnik and Keerberg 2007) to more than 90% in spinach (Loreto et al. 1999), while tomatoes have been modelled to re-assimilate 56-76% of photorespiratory CO_2 (Berghuijs et al. 2017). Both models described here re-assimilate 100% of CO_2 released during photorespiration. This means that our models overestimate biomass production by the extent of the disparity between the actual loss of photorespiratory CO_2 *in vivo* and the complete recapture of it in our models. This therefore cannot account for the failure of our *Marchantia* model to reach the experimentally-determined rates of biomass accumulation since the model is predicting a maximum for C_3 photosynthesis and this still is not sufficient to reach the observed growth rate.

Overall, Figure 2.3 indicates that the model at lower light intensities is sensitive to changes in biomass composition and NGAM costs, but modifying day: night biomass output ratios and nitrate uptake ratios, and RuBisCO carboxylation: oxygenation ratios have very little effect on predicted growth rates. At higher light intensities, none of our tested constraint variations were sufficient to allow the model to meet experimental growth rates. This difference likely reflects the relative importance of NGAM in both models. Since we have used a fixed NGAM value across all light intensities, the relative cost of NGAM when more energy is available to the model (i.e. at higher light intensities) is lower than when less energy is available. Reducing the value of NGAM will correspondingly have less of an impact on growth rate predictions when there is more energy available to the model overall.

2.6.3 Removing constraints on night-phase carbon fixation allows the *Marchantia* model to meet experimental growth rates

Our sensitivity analysis shows that the growth rate predictions of the *Marchantia* model are robust to many of the variations in parameters for which assumptions were made about constraint equality between angiosperms and bryophytes (Figure 2.3). The only constraint which had a major impact on growth rate predictions is the cost of NGAM, which was estimated directly from experimental data for *Marchantia*. Although it is sensible to allow for variance in this estimation, our model at lower light intensities was only able to meet experimental growth rates when NGAM was below 47% of our estimated value, which represents an extreme divergence from our estimate. Furthermore, our model at higher light intensities was unable to meet experimental growth rates even with such a divergence. This is particularly surprising because FBA optimises the objective function (in this case production of biomass) within the constraints applied (Orth et al. 2010; Sweetlove et al. 2017) – so an overestimation of metabolic flux is more likely for an accurately constrained model than an underestimation as seen here. Effectively, the observed growth rates of *Marchantia* appear to be stoichiometrically impossible for our model to produce within reasonable constraint limits, given the observed rates of photosynthetic CO₂ assimilation.

We therefore considered whether we had applied other limitations to the model, as a function of how we ran the analysis as opposed to direct constraints on reaction fluxes. Aside from the use of pFBA, which minimises the overall sum of fluxes within an optimised model

(based on the reasonable assumption that minimising the quantity of enzymes required to fulfil metabolic requirements is likely to have been selected for as it reduces the energy required for enzyme turnover (Lewis et al. 2010; Nikoloski et al. 2015)), the only other limitation imposed on the model is the requirement that CO₂ fixation only occurs during the day. However, plants exhibit a wide continuum of carbon fixation strategies – from C₃ and C₄ to obligate or facultative crassulacean acid metabolism (CAM), with a huge range of intermediate phenotypes (Silvera et al. 2010; Winter 2019; Winter et al. 2019) – and both C₄ and CAM are known to have evolved independently over 60 times (Bräutigam et al. 2017; Yang et al. 2015). Indeed, CAM plants are found within the angiosperms, gymnosperms, ferns and lycophytes (Smith and Winter 1996; Winter et al. 2019), while several groups of eukaryotic algae and hornworts have an alternative carbon concentrating mechanism based on the use of pyrenoids containing high concentrations of RuBisCO (Li et al. 2017; Villarreal and Renner 2012). It is therefore reasonable to consider that other bryophytes, which are known to have low mesophyll conductances (Carriqui et al. 2019; Gago et al. 2019) and thrive in wet environments where diffusion of CO₂ may be limited, may also have evolved mechanisms to increase their CCE.

We therefore relaxed the requirement that the *Marchantia* model restricts carbon fixation to the light phase and re-ran the optimisation under each set of experimental conditions. Figure 2.4 illustrates the comparison between the resulting predicted growth rate, the predicted growth rate before the restriction was relaxed and the experimental growth rates. The model does not allow for net carbon fixation at night (i.e. CO₂ is not allowed to enter the cell during the night phase), however we found that allowing the model to re-fix respiratory CO₂ during the night phase resulted in predicted growth rates for *Marchantia* that were close to or exceeded experimental growth rates. While the CCE of the original *Marchantia* model was 18% and 42% at lower and higher experimental PPFDs respectively, CCEs of 59% and 96% for the respective PPFDs are required to meet experimental growth rates. This is compared to between 62% and 70% for the *Arabidopsis* model across the set of modelled experimental conditions, which is able to reach higher carbons conversion efficiencies without invoking carbon recycling due to the difference in NGAM between the two models. Allowing the *Marchantia* model to re-fix respiratory CO₂ during the night results in a CCE of almost

100%. This suggests that *Marchantia in vivo* could achieve the high growth rates shown in experimental conditions despite low CO₂ assimilation rates (relative to angiosperms) by exploiting strategies that result in an extremely high CCE.

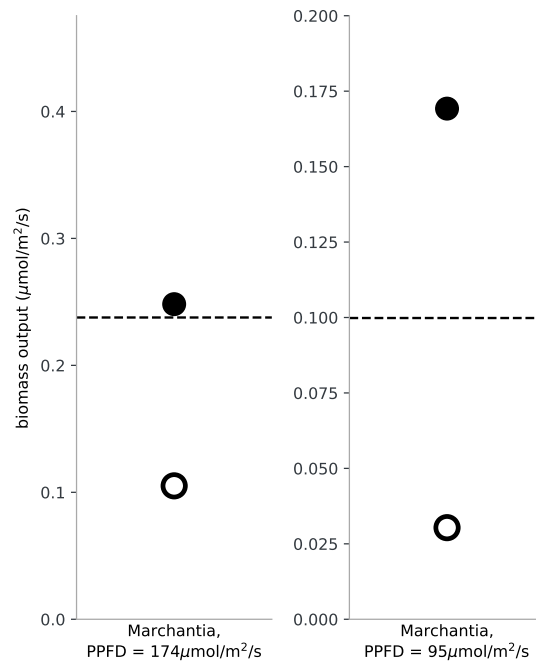


Figure 2.4: Comparison of growth rate predictions from the *Marchantia* model under both sets of experimental conditions. Open circles indicate the growth rate predictions without carbon fixation at night while closed circles indicate growth rate predictions when night-time carbon fixation is allowed. Dashed horizontal lines show experimental growth rates. PPFD, photosynthetic photon flux density.

2.6.4 Comparison of flux modes between *Arabidopsis* and *Marchantia* models suggests CAM-like carbon re-fixation strategy in *Marchantia*

To examine the metabolic strategies of the two models which can accurately predict experimentally-defined growth rates (i.e with relaxed carbon fixation constraints in the *Marchantia* model), we conducted a thorough comparison of fluxes across both central metabolic networks (using the experimental constraints defined by a PPFD of 120 μmol photons/m²/s in *Arabidopsis* and 174 μmol photons/m²/s in *Marchantia*). Reaction fluxes in the optimal pFBA solution for a model are not necessarily uniquely defined (Sweetlove and Ratcliffe 2011), so we carried out a FVA analysis on both models to determine the possible range of all reaction fluxes at the optimum. We considered a specific reaction flux to differ between models if the range of

fluxes in each model, adjusted to account for the differences in total biomass output between the two models, did not overlap between models. Using this criterion, we found that 813 reactions carried non-overlapping fluxes between the *Arabidopsis* and *Marchantia* models. The full set of reactions carrying non-overlapping fluxes between the two models can be found in Data 2.9. To identify the set of flux differences likely to have the most impact on the metabolic network as a whole, we looked specifically at reactions for which the closest fluxes between the two non-overlapping ranges differed by a factor of two or more, and at reactions which only carried flux in one model or for which fluxes ran in different directions in each model. In total, these reactions accounted for 389 of the 813 non-overlapping fluxes.

Figure 2.5 is a representation of the major flux routes through the *Arabidopsis* and *Marchantia* models. The *Arabidopsis* model predicts fluxes through the reactions expected for a plant using C₃ photosynthesis. CO₂ is taken up during the day and fixed in the chloroplast via the Calvin-Benson-Bassham (CBB) cycle. Some starch and malate are stored in the chloroplast and vacuole respectively during the day and broken down at night to provide energy for night-phase biomass production, through which CO₂ is lost from the system at night via respiration (Figure 2.5).

Conversely, for the *Marchantia* model to achieve realistic growth rates the model predicted fluxes through reactions similar to CAM photosynthesis. Carbon is fixed during the day and a substantial amount of starch is stored in the chloroplast for breakdown during the night to provide phosphoenolpyruvate (PEP) for carbon re-fixation using phosphoenolpyruvate carboxylase (PEPC). The resulting oxaloacetate is converted into malate, which is stored in the vacuole during the night phase. During the day, malate is exported from the vacuole and converted into pyruvate and other compounds that can be re-assimilated back into the CBB Cycle (Figure 2.5). Note that this is representative of CAM cycling (Tay et al. 2019) and is not full CAM because there is no net CO₂ assimilation at night and CO₂ uptake still occurs during the day, consistent with the lack of stomata in *Marchantia* to regulate gas exchange during the day and night. This demonstrates that significant metabolic carbon recovery is required in *Marchantia* in order to meet the growth rates observed in experimental systems, and suggests that some form of night-time carbon fixation may be

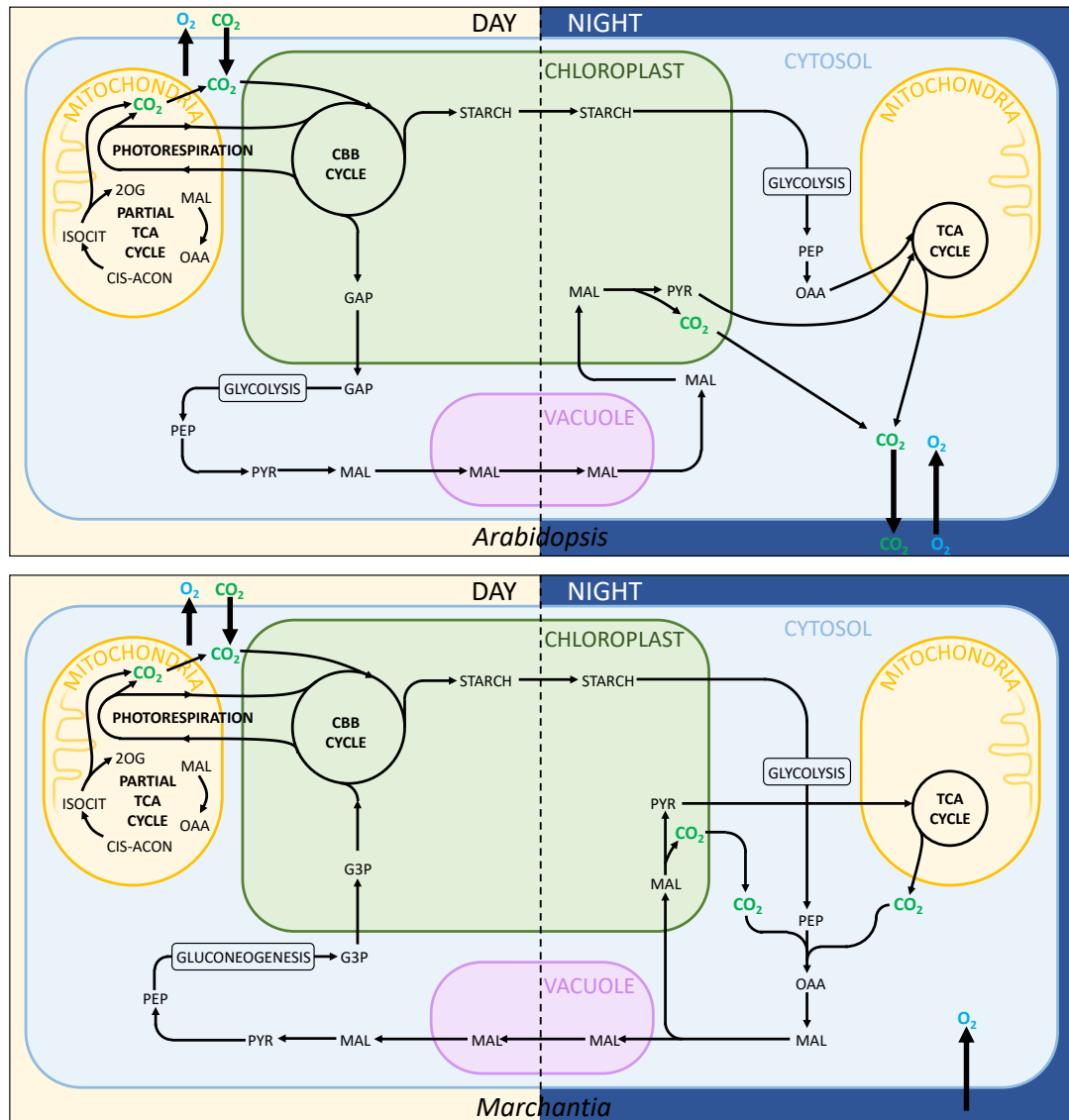


Figure 2.5: Major flux routes through *Arabidopsis* and *Marchantia* models with solutions able to meet experimental growth rate predictions. PEP, phosphoenolpyruvate; OAA, oxaloacetate; MAL, malate; PYR, pyruvate; GAP, D-glyceraldehyde 3-phosphate; G3P, 3-phospho-D-glycerate; CIS-ACON, cis-aconitate; ISOCIT, threo-isocitrate; 2OG, 2-oxoglutarate.

the strategy that allows *Marchantia* plants *in vivo* to exhibit such high growth rates despite very low CO_2 assimilation rates (Figure 2.5).

In addition to the major flux differences shown in Figure 2.5, the analysis highlighted a number of reaction groups with varying flux modes between the *Arabidopsis* and *Marchantia* models, related to the differing requirements for ATP and reducing power imposed by each species' specific biomass composition requirements (Figures A.2, A.3, A.4). For example,

in addition to starch storage during the day, the *Marchantia* model predicts the vacuolar storage of significantly more sucrose than required for direct incorporation into biomass at night (Figure A.2). During the night phase, this sucrose is converted to UDP-glucose and used for the production of cell wall polysaccharides; by contrast cell wall polysaccharides in the *Arabidopsis* model are produced using UDP-glucose derived from starch. Although in system-level optimisations such as these it is difficult to be certain of the explanation behind individual flux predictions, this is likely a mechanism for the conservation of ATP in the *Marchantia* model. ATP is required to pump protons into the vacuole to facilitate the passage of malate into the vacuole at night. While UDP-glucose production from starch (as used in the *Arabidopsis* model) requires UTP, which is produced using ATP, UDP-glucose production from sucrose incorporates UDP directly and no ATP is required.

Another example of such differences can be seen in Figure A.3, which illustrates the alternative uses of serine at night in the *Arabidopsis* and *Marchantia* models. Pectin is a higher proportion of biomass in *Arabidopsis* than in *Marchantia*, and as such Figure A.3 shows the requirement in *Arabidopsis* for more serine to be used in the regeneration of methylene-tetrahydrofolate for pectin production in the cytosol. By contrast, in *Marchantia* serine is used to produce glycine in the chloroplast, which results in a formate by-product that is broken down in the mitochondria to produce reducing power and carbon for re-assimilation via PEPC.

Finally, Figure A.4 shows the differing routes for leucine production between the *Arabidopsis* and *Marchantia* models. In the *Marchantia* model, leucine is predicted to be stored in the vacuole during the day and released for direct incorporation into protein for biomass production at night. In the *Arabidopsis* model, leucine is instead produced from additional valine stored in the vacuole during the day, via a biosynthetic pathway in the chloroplast which releases one molecule of NADH for every molecule of valine transformed. This NADH is presumably the benefit of synthesising leucine at night for *Arabidopsis* – the model predicts an increased requirement for NADH in *Arabidopsis* chloroplasts due to the higher proportion of lipids in *Arabidopsis* biomass compared to *Marchantia*. There is also a small flux through the chloroplast/cytosol glutamine-glutamate shuttle in *Arabidopsis* (Figure A.4) which is not predicted to run in the *Marchantia* model and also requires NADH – this may compensate

for the slightly reduced availability of 2-oxoglutarate in the plastid in *Arabidopsis* (used in glutamate production from aspartate) as a result of this leucine biosynthesis pathway.

2.7 DISCUSSION

We conducted a comparative flux balance analysis of central metabolism in *Arabidopsis* and *Marchantia*, the latter of which represents the first fully compartmentalised, charge-balanced stoichiometric model of metabolism in a bryophyte. We found that unlike *Arabidopsis*, in order to meet experimental growth rates the *Marchantia* model requires carbon to be re-fixed during the night phase or for the cost of cell maintenance to be dramatically reduced, or a combination of both. The modelling demonstrates that *Marchantia* requires a much higher CCE than *Arabidopsis* in order to achieve the experimentally measured growth rates of thallus tissue.

It is worth noting that as our calculation of NGAM was based on respiration rates measured for *Marchantia* cells in the dark, if *Marchantia* does indeed invoke some form of carbon recycling at night then these measured rates would represent an underestimation of the actual night-time respiration rate in *Marchantia*. However, as this would have led to an underestimation of NGAM, it would not explain the discrepancy between experimental and modelled growth rates described here (as the potentially underestimated NGAM value was still too high to allow observed growth rates to be met without additional carbon-saving measures). Maintenance costs in plants stem from cellular functions which are necessary even in the absence of growth, for example turnover and re-synthesis of proteins and other macromolecules and maintenance of ion gradients across membranes. If *Marchantia* achieves high CCE via reduced maintenance costs compared to angiosperms, it may be as a result of lower requirements for regeneration of proteins or reduced membrane permeability. While protein turnover rates have not been measured in bryophytes, bryophyte cell membranes are known to contain higher proportions of very-long-chain polyunsaturated fatty acids than angiosperm cell membranes (Dembitsky et al. 1993; Resemann et al. 2019). There is evidence that fatty acid desaturation increases membrane stability and reduces membrane permeability under cold stress in angiosperms (Gao et al. 2018; Spsychalla and Desborough 1990). However, knockout mutants of the moss *Physcomitrella patens* which are unable to

produce arachidonic acid (a polyunsaturated fatty acid which represents a major membrane component in *Physcomitrella*) do not exhibit any obvious growth defects (Resemann et al. 2019), and any functions of the increased proportions of long chain polyunsaturated fatty acids in bryophytes are not yet established.

With our experimentally-constrained maintenance cost, our stoichiometric model of *Marchantia* predicts that nocturnal respiratory CO₂ is re-fixed via PEPC and stored during the night phase as malate for use during the day (Figure 2.5). However, the high CCE we have identified in *Marchantia* could also be achieved via alternative metabolic or physiological mechanisms. For example, recent work modelling CAM photosynthesis under drought conditions has identified the reverse reaction of mitochondrial isocitrate dehydrogenase as a secondary mode of night-time carbon fixation (Töpfer et al. 2020). CCE may also be increased using strategic intracellular localisation or concentration of specific enzymes. The pyrenoid-based carbon concentrating mechanisms found in hornworts and some eukaryotic algae are dependent on a carbonic anhydrase localised to the thylakoid lumen, and a high relative concentration of carbonic anhydrase in mesophyll cells has been shown to be required for photosynthetic efficiency in some C₄ plants (DiMario et al. 2017; Hanson et al. 2014). Additionally, while our sensitivity analysis suggests this is unlikely to provide improved growth rates in *Marchantia* (Figure 2.3), the presence of the glycine cleavage system only in bundle sheath cells in plants exhibiting intermediate C₃-C₄ photosynthesis results in an increased concentration of CO₂ from photorespiration, increasing the carboxylation efficiency of RuBisCO (Ferne and Bauwe 2020; Sage et al. 2018). Finally, the subcellular location of specific organelles may aid in carbon re-fixation - it has been suggested that some C₃ plants achieve carbon re-fixation during the day by positioning chloroplasts around the periphery of photosynthetic cells with mitochondria encapsulated in the centre, thereby forcing the recapture of photorespiratory CO₂ as it diffuses out of the mitochondria (Busch et al. 2013).

Given the variation in carbon concentration and re-fixation strategies seen across land plants, and the number of times they are thought to have evolved independently, it is reasonable to consider that a similar mechanism may have developed in the liverworts, or a subset of these plants. Bryophytes do not have stomata in their gametophyte phase (Merced

and Renzaglia 2017; Renzaglia et al. 2020; Shimamura 2016), and limit water loss with thick cell walls, resulting in a low rate of gas exchange which limits their access to carbon (Carriqui et al. 2019; Gago et al. 2019). This limitation may have acted as the evolutionary driver behind the development of compensation mechanisms, which could involve any (or a combination) of the metabolic and physiological possibilities described above. Tracheophytes can regulate gas exchange and water loss and in the wild are generally more limited by access to mineral nutrients (Reich et al. 2006; Rubio-Asensio and Bloom 2017). However, there is accumulating evidence that higher plants under agricultural conditions (i.e. extensively fertilised, irrigated and treated with pesticides) are by contrast limited by access to sufficient carbon (Amthor et al. 2019; Ort et al. 2015), and recent studies engineering either increased carbon assimilation or reduced carbon efflux in crop plants have shown significantly improved yield under field conditions (López-Calcano et al. 2020; South et al. 2019). With this in mind, our finding that *Marchantia*, similarly carbon-limited, exhibits a higher CCE than *C₃* angiosperms places the bryophytes as a promising opportunity for further investigation of existing strategies for carbon recycling which could be exploited for the increased productivity of crop plants. *Marchantia* is an obvious choice for such investigation, as an emerging model organism with practical characteristics for experimental investigation and increasing genomic and metabolic resources available (Bowman et al. 2016, 2017; Cannell et al. 2020). However, further research, both experimental and computational in nature, is required to confirm the existence of any metabolic or physiological carbon re-fixation strategy in *Marchantia* and to improve the standing of bryophytes in general as model organisms. Indeed, for this work we constructed an additional stoichiometric model of *Physcomitrella patens*, a model moss species for which carbon assimilation rates are even lower than in *Marchantia* (Takemura et al. 2017), but we were unable to extract meaningful results for *Physcomitrella* due to the lack of metabolic data available to appropriately constrain the model. It seems possible that future metabolic investigation of the bryophytes may uncover a currently unrealised source of novel metabolic processes to inform innovative engineering strategies in higher plants.

REFERENCES

- Amthor, J. S. (1984). The role of maintenance respiration in plant growth. *Plant, Cell and Environment*, 7, 561–569.

- Amthor, J. S. (2000). The McCree-de Wit-Penning de Vries-Thornley respiration paradigms: 30 years later. *Annals of Botany*, 86, 1–20.
- Amthor, J. S., Bar-Even, A., Hanson, A. D., ... Tyerman, S. D. (2019). Engineering strategies to boost crop productivity by cutting respiratory carbon loss. *The Plant Cell*, 31, 297–314.
- Arnold, A., & Nikoloski, Z. (2014). Bottom-up metabolic reconstruction of *Arabidopsis* and its application to determining the metabolic costs of enzyme production. *Plant Physiology*, 165, 1380–1391.
- Berghuijs, H. N. C., Yin, X., Ho, Q. T., ... Struik, P. C. (2017). Localization of (photo)respiration and CO₂ re-assimilation in tomato leaves investigated with a reaction-diffusion model. *PLOS ONE*, 12, e0183746.
- Bowman, J. L., Araki, T., & Kohchi, T. (2016). Marchantia: Past, present and future. *Plant and Cell Physiology*, 57, 205–209.
- Bowman, J. L., Kohchi, T., Yamato, K. T., ... Schmutz, J. (2017). Insights into land plant evolution garnered from the *Marchantia polymorpha* genome. *Cell*, 171, 287–304.e15.
- Bräutigam, A., Schlüter, U., Eisenhut, M., & Gowik, U. (2017). On the evolutionary origin of CAM photosynthesis. *Plant Physiology*, 174, 473–477.
- Busch, F. A., Sage, T. L., Cousins, A. B., & Sage, R. F. (2013). C₃ plants enhance rates of photosynthesis by reassimilating photorespired and respired CO₂. *Plant, Cell and Environment*, 36, 200–212.
- Cannell, N., Emms, D. M., Hetherington, A. J., ... Sweetlove, L. J. (2020). Multiple metabolic innovations and losses are associated with major transitions in land plant evolution. *Current Biology*, 30, 1783–1800.e11.
- Carella, P., Gogleva, A., Hoey, D. J., ... Schornack, S. (2019). Conserved biochemical defenses underpin host responses to oomycete infection in an early-divergent land plant lineage. *Current Biology*, 29, 2282–2294.e5.
- Carriqui, M., Roig-Oliver, M., Brodribb, T. J., ... Flexas, J. (2019). Anatomical constraints to nonstomatal diffusion conductance and photosynthesis in lycophytes and bryophytes. *New Phytologist*, 222, 1256–1270.
- Carter, J. L., & Romine, K. G. (1969). The effects of long and short photoperiods on the rate of growth and gemmae cup production in *Marchantia polymorpha* L. *Transactions of the Kansas Academy of Science (1903-)*, 72, 98.
- Cheung, C. Y. M., Poolman, M. G., Fell, D. A., ... Sweetlove, L. J. (2014). A diel flux balance model captures interactions between light and dark metabolism during day-night cycles in C₃ and crassulacean acid metabolism leaves. *Plant Physiology*, 165, 917–929.
- Cheung, C. Y. M., Williams, T. C. R., Poolman, M. G., ... Sweetlove, L. J. (2013). A method for accounting for maintenance costs in flux balance analysis improves the prediction of plant cell metabolic phenotypes under stress conditions. *Plant Journal*, 75, 1050–1061.
- Christophe, A., Letort, V., Hummel, I., ... Lecoecur, J. (2008). A model-based analysis of the dynamics of carbon balance at the whole-plant level in *Arabidopsis thaliana*. *Functional Plant Biology*, 35, 1147.
- Cookson, S. J., Radziejowski, A., & Granier, C. (2006). Cell and leaf size plasticity in *Arabidopsis*: What is the role of endoreduplication? *Plant, Cell and Environment*, 29, 1273–1283.
- Cookson, S. J., Van Lijsebettens, M., & Granier, C. (2005). Correlation between leaf growth variables suggest intrinsic and early controls of leaf size in *Arabidopsis thaliana*. *Plant, Cell and Environment*, 28, 1355–1366.
- Delhon, P., Gojon, A., Tillard, P., & Passama, L. (1995). Diurnal regulation of NO₃⁻ uptake in soybean plants I. Changes in NO₃⁻ influx, efflux, and N utilization in the plant during the day/night cycle. *Journal of Experimental Botany*, 46, 1585–1594.

- Delwiche, C. F., & Cooper, E. D. (2015). The evolutionary origin of a terrestrial flora. *Current Biology*, 25, R899–R910.
- Dembitsky, V. M., Rezanka, T., Bychek, I. A., & Afonina, O. M. (1993). Polar lipid and fatty acid composition of some bryophytes. *Phytochemistry*, 33, 1009–1014.
- de Vries, J., & Archibald, J. M. (2018). Plant evolution: Landmarks on the path to terrestrial life. *New Phytologist*, 217, 1428–1434.
- DiMario, R. J., Clayton, H., Mukherjee, A., ... Moroney, J. V. (2017). Plant carbonic anhydrases: Structures, locations, evolution, and physiological roles. *Molecular Plant*, 10, 30–46.
- Emms, D. M., & Kelly, S. (2019). Orthofinder: Phylogenetic orthology inference for comparative genomics. *Genome Biology*, 20, 238.
- Fernie, A. R., & Bauwe, H. (2020). Wasteful, essential, evolutionary stepping stone? The multiple personalities of the photorespiratory pathway. *The Plant Journal*, 102, 666–677.
- Fletcher, B. J., Brentnall, S. J., Quick, W. P., & Beerling, D. J. (2006). BRYOCARB: A process-based model of thallose liverwort carbon isotope fractionation in response to CO₂, O₂, light and temperature. *Geochimica et Cosmochimica Acta*, 70, 5676–5691.
- Flexas, J., & Carriquí, M. (2020). Photosynthesis and photosynthetic efficiencies along the terrestrial plant's phylogeny: Lessons for improving crop photosynthesis. *The Plant Journal*, 101, 964–978.
- Gago, J., Carriquí, M., Nadal, M., ... Flexas, J. (2019). Photosynthesis optimized across land plant phylogeny. *Trends in Plant Science*, 24, 947–958.
- Gao, H., Lu, Z. M., Yang, Y., ... Cao, W. (2018). Melatonin treatment reduces chilling injury in peach fruit through its regulation of membrane fatty acid contents and phenolic metabolism. *Food Chemistry*, 245, 659–666.
- Gibon, Y., Bläsing, O. E., Palacios-Rojas, N., ... Stitt, M. (2004). Adjustment of diurnal starch turnover to short days: Depletion of sugar during the night leads to a temporary inhibition of carbohydrate utilization, accumulation of sugars and post-translational activation of ADP-glucose pyrophosphorylase in the following light period. *Plant Journal*, 39, 847–862.
- Gimeno, T. E., Ogée, J., Royles, J., ... Wingate, L. (2017). Bryophyte gas-exchange dynamics along varying hydration status reveal a significant carbonyl sulphide (COS) sink in the dark and COS source in the light. *New Phytologist*, 215, 965–976.
- Gomes de Oliveira Dal'Molin, C., & Nielsen, L. K. (2018). Plant genome-scale reconstruction: From single cell to multi-tissue modelling and omics analyses. *Current Opinion in Biotechnology*, 49, 42–48.
- Graham, T., & Dixon, M. A. (2012). Liverwort control: An ancillary role for ozone-based irrigation water treatment systems? *Horticultural Science*, 47, 361–367.
- Green, T. G. A., & Snelgar, W. P. (1982). A comparison of photosynthesis in two thalloid liverworts. *Oecologia*, 54, 275–280.
- Gudmundsson, S., & Thiele, I. (2010). Computationally efficient flux variability analysis. *BMC Bioinformatics*, 11, 489.
- Hanson, D. T., Renzaglia, K., & Villarreal, J. C. (2014). Diffusion limitation and CO₂ concentrating mechanisms in bryophytes. *Photosynthesis in bryophytes and early land plants* (pp. 95–111). Springer, Dordrecht.
- Happ, K., & Classen, B. (2019). Arabinogalactan-proteins from the liverwort *Marchantia polymorpha* l., a member of a basal land plant lineage, are structurally different to those of angiosperms. *Plants*, 8, 460.
- Harris, B. J., Harrison, C. J., Hetherington, A. M., & Williams, T. A. (2020). Phylogenomic evidence for the monophyly of bryophytes and the reductive evolution of stomata. *Current Biology*, 30, 2001–2012.e2.

- Humphreys, C. P., Franks, P. J., Rees, M., ... Beerling, D. J. (2010). Mutualistic mycorrhiza-like symbiosis in the most ancient group of land plants. *Nature Communications*, 1, 1–7.
- Katoh, K. (1983). Photosynthesis and photoassimilation of glucose during photomixotrophic growth of *Marchantia polymorpha* cells in suspension culture. *Physiologia Plantarum*, 57, 67–74.
- Koide, E., Suetsugu, N., Iwano, M., ... Nishihama, R. (2020). Regulation of photosynthetic carbohydrate metabolism by a raf-like kinase in the liverwort *Marchantia polymorpha*. *Plant and Cell Physiology*, 61, 631–643.
- Lewis, N. E., Hixson, K. K., Conrad, T. M., ... Palsson, B. (2010). Omic data from evolved *E. coli* are consistent with computed optimal growth from genome-scale models. *Molecular Systems Biology*, 6.
- Lewis, N. E., Nagarajan, H., & Palsson, B. O. (2012). Constraining the metabolic genotype-phenotype relationship using a phylogeny of *in silico* methods. *Nature Reviews Microbiology*, 10, 291–305.
- Li, F. W., Villarreal, J. C., & Szövényi, P. (2017). Hornworts: An overlooked window into carbon-concentrating mechanisms. *Trends in Plant Science*, 22, 275–277.
- López-Calcano, P. E., Brown, K. L., Simkin, A. J., ... Raines, C. A. (2020). Stimulating photosynthetic processes increases productivity and water-use efficiency in the field. *Nature Plants*, 6, 1054–1063.
- Loreto, F., Delfine, S., & Di Marco, G. (1999). Estimation of photorespiratory carbon dioxide recycling during photosynthesis. *Australian Journal of Plant Physiology*, 26, 733–736.
- Ma, F., Jazmin, L. J., Young, J. D., & Allen, D. K. (2014). Isotopically nonstationary ¹³C flux analysis of changes in *Arabidopsis thaliana* leaf metabolism due to high light acclimation. *Proceedings of the National Academy of Sciences of the United States of America*, 111, 16967–16972.
- Macduff, J. H., & Bakken, A. K. (2003). Diurnal variation in uptake and xylem contents of inorganic and assimilated N under continuous and interrupted N supply to *Phleum pratense* and *Festuca pratensis*. *Journal of Experimental Botany*, 54, 431–444.
- McKay, M. D., Beckman, R. J., & Conover, W. J. (1979). A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. *Technometrics*, 21, 239.
- Merced, A., & Renzaglia, K. S. (2017). Structure, function and evolution of stomata from a bryological perspective. *Bryophyte Diversity and Evolution*, 39, 7.
- Michelet, L., Zaffagnini, M., Morisse, S., ... Lemaire, S. D. (2013). Redox regulation of the Calvin-Benson Cycle: Something old, something new. *Frontiers in Plant Science*, 4, 470.
- Miquel, M., & Browse, J. (1992). *Arabidopsis* mutants deficient in polyunsaturated fatty acid synthesis. *The Journal of Biological Chemistry*, 267, 1502–1509.
- Moreira, T. B., Lima, J. M., Coca, G. C., & Williams, T. C. R. (2019). Insights into the spatial and temporal organisation of plant metabolism from network flux analysis. *Theoretical and Experimental Plant Physiology*, 31, 215–226.
- Nikoloski, Z., Perez-Storey, R., & Sweetlove, L. J. (2015). Inference and prediction of metabolic network fluxes. *Plant Physiology*, 169, 1443–1455.
- Ort, D. R., Merchant, S. S., Alric, J., ... Zhu, X. G. (2015). Redesigning photosynthesis to sustainably meet global food and bioenergy demand. *Proceedings of the National Academy of Sciences of the United States of America*, 112, 8529–8536.
- Orth, J. D., Thiele, I., & Palsson, B. O. (2010). What is flux balance analysis? *Nature Biotechnology*, 28, 245–248.
- Park, J. M., Kim, T. Y., & Lee, S. Y. (2011). Genome-scale reconstruction and *in silico* analysis of the *Ralstonia eutropha* h16 for polyhydroxyalkanoate synthesis, lithoautotrophic growth, and 2-methyl citric acid production. *BMC Systems Biology*, 5, 101.

- Pärnik, T., & Keerberg, O. (2007). Advanced radiogasometric method for the determination of the rates of photorespiratory and respiratory decarboxylations of primary and stored photosynthates under steady-state photosynthesis. *Physiologia Plantarum*, 129, 34–44.
- Poolman, M. G., Miguët, L., Sweetlove, L. J., & Fell, D. A. (2009). A genome-scale metabolic model of *Arabidopsis* and some of its properties. *Plant Physiology*, 151, 1570–1581.
- Rasse, D. P., & Tocquin, P. (2006). Leaf carbohydrate controls over *Arabidopsis* growth and response to elevated CO₂: An experimentally based model. *New Phytologist*, 172, 500–513.
- Reich, P. B., Hobbie, S. E., Lee, T., ... Trost, J. (2006). Nitrogen limitation constrains sustainability of ecosystem response to CO₂. *Nature*, 440, 922–925.
- Rensing, S. A., Lang, D., Zimmer, A. D., ... Boore, J. L. (2008). The *Physcomitrella* genome reveals evolutionary insights into the conquest of land by plants. *Science*, 319, 64–69.
- Renzaglia, K. S., Browning, W. B., & Merced, A. (2020). With over 60 independent losses, stomata are expendable in mosses. *Frontiers in Plant Science*, 11, 567.
- Resemann, H. C., Lewandowska, M., Gömann, J., & Feussner, I. (2019). Membrane lipids, waxes and oxylipins in the moss model organism *Physcomitrella patens*. *Plant and Cell Physiology*, 60, 1166–1175.
- Rubio-Asensio, J. S., & Bloom, A. J. (2017). Inorganic nitrogen form: A major player in wheat and *Arabidopsis* responses to elevated CO₂. *Journal of Experimental Botany*, 68, 2611–2625.
- Sage, R. F., Monson, R. K., Ehleringer, J. R., ... Percy, R. W. (2018). Some like it hot: The physiological ecology of C₄ plant evolution. *Oecologia*, 187, 941–966.
- Saha, R., Suthers, P. F., & Maranas, C. D. (2011). *Zea mays* iRS1563: A comprehensive genome-scale metabolic reconstruction of maize metabolism. *PLoS ONE*, 6, e21784.
- Scheible, W. R., Krapp, A., & Stitt, M. (2000). Reciprocal diurnal changes of phosphoenolpyruvate carboxylase expression and cytosolic pyruvate kinase, citrate synthase and NADP-isocitrate dehydrogenase expression regulate organic acid metabolism during nitrate assimilation in tobacco leaves. *Plant, Cell and Environment*, 23, 1155–1167.
- Schläpfer, P., Zhang, P., Wang, C., ... Rhee, S. Y. (2017). Genome-wide prediction of metabolic enzymes, pathways, and gene clusters in plants. *Plant Physiology*, 173, 2041–2059.
- Shameer, S., Baghalian, K., Cheung, C. Y. M., ... Sweetlove, L. J. (2018). Computational analysis of the productivity potential of CAM. *Nature Plants*, 4, 165–171.
- Shameer, S., Ratcliffe, R. G., & Sweetlove, L. J. (2019). Leaf energy balance requires mitochondrial respiration and export of chloroplast NADPH in the light. *Plant Physiology*, 180, 1947–1961.
- Shimamura, M. (2016). *Marchantia polymorpha*: Taxonomy, phylogeny and morphology of a model system. *Plant and Cell Physiology*, 57, 230–256.
- Shinmen, Y., Katoh, K., Shimizu, S., ... Yamada, H. (1991). Production of arachidonic acid and eicosapentaenoic acids by *Marchantia polymorpha* in cell culture. *Phytochemistry*, 30, 3255–3260.
- Silvera, K., Neubig, K. M., Whitten, W. M., ... Cushman, J. C. (2010). Evolution along the crassulacean acid metabolism continuum. *Functional Plant Biology*, 37, 995–1010.
- Smith, J. A. C., & Winter, K. (1996). Taxonomic distribution of crassulacean acid metabolism. *Crassulacean acid metabolism* (pp. 427–436). Springer, Berlin, Heidelberg.
- South, P. F., Cavanagh, A. P., Liu, H. W., & Ort, D. R. (2019). Synthetic glycolate metabolism pathways stimulate crop growth and productivity in the field. *Science*, 363, eaat9077.
- Spychalla, J. P., & Desborough, S. L. (1990). Fatty acids, membrane permeability, and sugars of stored potato tubers. *Plant Physiology*, 94, 1207–1213.
- Sun, J., Okita, T. W., & Edwards, G. E. (1999). Modification of carbon partitioning, photosynthetic capacity, and O₂ sensitivity in *Arabidopsis* plants with low ADP-glucose pyrophosphorylase activity. *Plant Physiology*, 119, 267–276.

- Sweetlove, L. J., Nielsen, J., & Fernie, A. R. (2017). Engineering central metabolism - a grand challenge for plant biologists. *The Plant Journal*, 90, 749–763.
- Sweetlove, L. J., & Ratcliffe, R. G. (2011). Flux-balance modeling of plant metabolism. *Frontiers in Plant Science*, 2, 38.
- Takemura, K., Kamachi, H., Kume, A., ... Hanba, Y. T. (2017). A hypergravity environment increases chloroplast size, photosynthesis, and plant growth in the moss *Physcomitrella patens*. *Journal of Plant Research*, 130, 181–192.
- Tanaka, Y., Sugano, S. S., Shimada, T., & Hara-Nishimura, I. (2013). Enhancement of leaf photosynthetic capacity through increased stomatal density in *Arabidopsis*. *New Phytologist*, 198, 757–764.
- Tay, I., Odang, K. B., & Cheung, C. Y. M. (2019). Metabolic modelling of CAM cycling, CAM idling and the C₃ CAM continuum. *BioRxiv*.
- Thomas, R. J., Behringer, F. J., Lombard, C. S., & Sparkowski, J. J. (1984). Effects of auxin on wall polysaccharide composition and enzyme activity during extension-growth of *Pellia* (Bryophyta). *Physiologia Plantarum*, 60, 502–506.
- Töpfer, N., Braam, T., Shameer, S., & Ratcliffe, R. G. (2020). CAM emerges in a leaf metabolic model under water-saving constraints in different environments. *bioRxiv*.
- Villarreal, J. C., & Renner, S. S. (2012). Hornwort pyrenoids, carbon-concentrating structures, evolved and were lost at least five times during the last 100 million years. *Proceedings of the National Academy of Sciences of the United States of America*, 109, 18873–18878.
- Voth, P. D. (1943). Effects of nutrient-solution concentration on the growth of *Marchantia polymorpha*. *Botanical Gazette*, 104, 591–601.
- Waite, M., & Sack, L. (2010). How does moss photosynthesis relate to leaf and canopy structure? Trait relationships for 10 Hawaiian species of contrasting light habitats. *New Phytologist*, 185, 156–172.
- Webplotdigitizer. (2020). Retrieved June 1, 2020, from <https://automeris.io/WebPlotDigitizer/citation.html>
- Weraduwage, S. M., Chen, J., Anozie, F. C., ... Sharkey, T. D. (2015). The relationship between leaf area growth and biomass accumulation in *Arabidopsis thaliana*. *Frontiers in Plant Science*, 6, 167.
- Williams, T. C. R., Poolman, M. G., Howden, A. J. M., ... Sweetlove, L. J. (2010). A genome-scale metabolic model accurately predicts fluxes in central carbon metabolism under stress conditions. *Plant Physiology*, 154, 311–323.
- Winter, K. (2019). Ecophysiology of constitutive and facultative CAM photosynthesis. *Journal of Experimental Botany*, 70, 6495–6508.
- Winter, K., Sage, R. F., Edwards, E. J., ... Holtum, J. A. M. (2019). Facultative crassulacean acid metabolism in a C₃-C₄ intermediate. *Journal of Experimental Botany*, 70, 6571–6579.
- Yang, X., Cushman, J. C., Borland, A. M., ... Holtum, J. A. M. (2015). A roadmap for research on crassulacean acid metabolism (CAM) to enhance sustainable food and bioenergy production in a hotter, drier world. *New Phytologist*, 207, 491–504.
- Zabackis, E., Huang, J., Müller, B., ... Albersheim, P. (1995). Characterization of the cell-wall polysaccharides of *Arabidopsis thaliana* leaves. *Plant Physiology*, 107, 1129–1138.
- Zhang, H., Zhang, L., Wu, S., ... Chen, L. (2020). AtWRKY75 positively regulates age-triggered leaf senescence through gibberellin pathway. *Plant Diversity*.

Statement of Authorship for joint/multi-authored papers for PGR thesis

To appear at the end of each thesis chapter submitted as an article/paper

The statement shall describe the candidate's and co-authors' independent research contributions in the thesis publications. For each publication there should exist a complete statement that is to be filled out and signed by the candidate and supervisor (**only required where there isn't already a statement of contribution within the paper itself**).

Title of Paper	Constraint-based modelling reveals a high carbon conversion efficiency in <i>Marchantia polymorpha</i> compared to that in C ₃ angiosperms
Publication Status	☐ Published ☐ Accepted for Publication ☒ Submitted for Publication ☐ Unpublished and unsubmitted work written in a manuscript style
Publication Details	Cannell et. al. (unpublished). Constraint-based modelling reveals a high carbon conversion efficiency in <i>Marchantia polymorpha</i> compared to that in C ₃ angiosperms

Student Confirmation

Student Name:	Naomi Cannell		
Contribution to the Paper	Naomi Cannell and Lee J. Sweetlove conceived the project and developed the methodology; N.C wrote code and carried out the investigation and formal analysis; L.J.S and Sanu Shameer provided supervision; N.C. wrote the manuscript and prepared visualisations; L.J.S reviewed and edited the manuscript.		
Signature Naomi Cannell	Date	15/01/2021	

Supervisor Confirmation

By signing the Statement of Authorship, you are certifying that the candidate made a substantial contribution to the publication, and that the description described above is accurate.

Supervisor name and title: Professor Lee J. Sweetlove		
Supervisor comments		
Signature 	Date	18/01/2020

This completed form should be included in the thesis, at the end of the relevant chapter.

3

Multiple metabolic innovations and losses are associated with major transitions in land plant evolution

3.1 MOTIVATION

This manuscript addresses a major bias in our knowledge of the metabolic capabilities of plants. Namely, that our understanding of plant metabolic systems is dominated by studies of angiosperms and in particular a handful of angiosperm model species and crops. Knowledge of metabolism of other plant lineages is scarce and there are likely to be undiscovered metabolic capabilities and mechanisms in plants and algae. We lack a systematic picture of the evolution of metabolic capabilities and how these relate to major evolutionary transitions such as the emergence of land plants. Here, we tackle this knowledge gap using genome-sequence data to reconstruct metabolic pathway databases of eight species representative of relevant plant and algal lineages. This paper was accepted for publication by *Current Biology* and was published in the 18th May 2020 issue (Cannell et al. 2020) (Appendix C).

3.2 CONTRIBUTION

A signed statement of authorship can be found at the end of this chapter. Naomi Cannell and Lee J. Sweetlove conceived the project and developed the methodology; N.C. wrote code and

carried out the investigation and formal analysis; L.J.S. provided supervision; Steven Kelly and David M. Emms provided software and computing resources; John MacKay, Liam Dolan, and Alexander J. Hetherington provided transcriptome resources; N.C. wrote the manuscript and prepared the figures; L.J.S., L.D., J.M., A.J.H., and D.M.E. reviewed and edited the manuscript.

3.3 SUMMARY

Investigating the evolution of plant biochemistry is challenging because few metabolites are preserved in fossils and because metabolic networks are difficult to experimentally characterise in diverse extant organisms. We report a comparative computational approach based on whole-genome metabolic pathway databases of eight species representative of major plant lineages, combined with homologous relationships among genes of 72 species from streptophyte algae to angiosperms. We use this genomic approach to identify metabolic gains and losses during land plant evolution. We extended our findings with additional analysis of 305 non-angiosperm plant transcriptomes. Our results revealed that genes encoding the complete biosynthetic pathway for brassinosteroid phytohormones and enzymes for brassinosteroid inactivation are present only in spermatophytes. Genes encoding only part of the biosynthesis pathway are present in ferns and lycophytes, indicating a stepwise evolutionary acquisition of this pathway. Nevertheless, brassinosteroids are ubiquitous in land plants, suggesting that brassinosteroid biosynthetic pathways differ between earlier- and later-diverging lineages. Conversely, genes for gibberellin biosynthesis and inactivation using methyltransferases are found in all land plant lineages. This suggests that bioactive gibberellins might be present in bryophytes, although they have yet to be detected experimentally. We also found that cytochrome P450 oxidases involved in cutin and suberin production are absent in genomes of non-angiosperm plants that nevertheless do contain these biopolymers. Overall, we identified significant differences in crucial metabolic processes between angiosperms and earlier-diverging land plants and resolve details of the evolutionary history of several phytohormone and structural polymer biosynthetic pathways in land plants.

3.4 INTRODUCTION

Some time before 450 million years ago, a single lineage of streptophyte algae colonised land, and the subsequent evolutionary radiation gave rise to the embryophytes, which are the dominant photosynthetic organisms in the terrestrial flora (de Vries et al. 2016). Extant embryophytes display a diverse range of morphologies, physiologies, and biochemistries. The evolution of this diversity had enormous impact on the terrestrial biotic and abiotic environment, affecting nutrient cycles and hydrology and modifying earth sediments and atmosphere (Morris et al. 2018). Many of the adaptations that accompanied the colonisation of land involved evolution of new metabolic capabilities (Rensing 2018). Examples include metabolic pathways for the biosynthesis of phytohormones coordinating plant growth, of specialised metabolites that defend against pathogens and provide tolerance of abiotic stresses, and of structural polymers such as lignin and suberin that provide support to stems and roots (Berens et al. 2017; Davies et al. 2018; Niklas et al. 2017). This diversity of metabolic enzymes is a consequence of substantial adaptive gene family expansion (Nelson and Werck-Reichhart 2011).

Investigating the full complement of metabolic pathways in an organism is challenging. Much of our current knowledge of plant metabolism is the result of experimental characterisation of the kinetic properties of purified enzymes or *in vivo* metabolic flux analysis using isotope tracers (Basler et al. 2018). The laborious nature of these approaches means that we lack a complete picture of metabolism for any given species, even for well-studied angiosperms. Moreover, the focus on a few model species and crops means that current knowledge of plant metabolism is heavily skewed toward angiosperms (Rensing 2017), which means that we have only a partial picture of the evolution of plant metabolism over the course of land plant history.

Recently, *in silico* systems approaches have been applied to plant metabolism (Gomes de Oliveira Dal'Molin and Nielsen 2018; Moreira et al. 2019), facilitating the development of computational representations of entire metabolic networks from genome sequences. Furthermore, the availability of transcriptomes and whole-genome sequences of algae and earlier-diverging land plants has provided researchers with valuable resources for metabolic

modeling (Bowman et al. 2017; Hori et al. 2014; Rensing et al. 2008). These data are a source of metabolic information that remains largely uninterrogated. A single study has analysed the evolution of metabolism from genome sequence data, revealing that the main metabolic innovations after the appearance of vascular plants relate to specialised metabolism (Chae et al. 2014). However, the study analysed just 16 species providing a sparse sample of plant evolution and only provided a top-level overview of metabolic pathway types, not specific functions.

Here, we exploit a wider range of genome sequence and transcriptome data to investigate in detail the evolution of metabolism in the Chloroplastida. A comparative approach was used to identify metabolic innovations and losses occurring during the evolution of streptophyte algae and land plants. Our starting point was the generation of genome-scale metabolic pathway databases for seven species spanning from algae to land plants (two charophytes, two bryophytes, a lycophyte, a fern, and a gymnosperm). From these databases, and an additional pre-existing database for *Arabidopsis* (Mueller et al. 2003), the presence and absence of the full range of known metabolic pathways (and therefore the appearance and disappearance of various metabolic traits over the course of land plant evolution in these species) was inferred. Individual metabolic gene annotations from these databases were then used in combination with information on homologous genes in other organisms to confirm inferred metabolic innovations or losses and extend their supporting evidence across 64 further plant and algal genomes (bringing the total number of species analysed to 72). To provide further support for metabolic innovations, an additional 305 non-angiosperm transcriptomes were analysed. The identification of several known evolutionary innovations in plant metabolism validates the method. The analysis revealed new information on the occurrence of biosynthetic enzymes for phytohormones and structural compounds across land plants and algae.

3.5 METHODS

Datasets and code generated during this study are available in Data 3.1, 3.2, 3.3, 3.4, and 3.5.

3.5.1 Experimental model and subject details

This study used a total of 377 genome and transcriptome datasets. All angiosperm genome sequences used in the orthogroup analysis were downloaded from the Phytozome database (<https://phytozome.jgi.doe.gov/>). Non-angiosperm genome and transcriptome sequences used in the orthogroup analysis were either downloaded from Phytozome, GenBank (Benson et al. 2017), Fernbase (Li et al. 2018), TreeGenes (Wegrzyn et al. 2019), or ConGenie (Sundell et al. 2015), or from individual publications; specific sources for each species can be found in the key resources table (included in Appendix C). All 305 transcriptomes used in the transcriptome analysis were taken from the One Thousand Plant Transcriptomes Initiative (Leebens-Mack et al. 2019), and unique identifiers for each species are documented in the key resources table.

3.5.2 Production of metabolic pathway databases

Metabolic pathway/genome databases (PGDBs) for seven streptophyte species (*Klebsormidium nitens*, *Chara braunii*, *Marchantia polymorpha*, *Physcomitrella patens*, *Selaginella moellendorffii*, *Salvinia cucullata* and *Picea glauca*) were generated (Data 3.1) using the Pathway Tools software (Karp et al. 2016), which uses genome annotations to infer the presence of metabolic reactions in an organism and a rule-based approach to import entire metabolic pathways once certain thresholds of reaction evidence are reached (Figure B.1). The software was set up to utilise both MetaCyc and PlantCyc as reference metabolic pathway databases. Pathway Tools computes a likelihood score, incorporating information about the proportion of metabolic reactions in the pathway with corresponding annotations in an organism's genome, the uniqueness of these reactions across metabolic pathways and whether any are considered 'key' reactions to the pathway. A pathway is imported into the PGDB if the likelihood score is greater than a specified threshold value. This was set relatively low (0.15) to minimise the number of false negative pathway inferences. To be included in a PGDB, a pathway must also meet several further criteria: it must be a natural metabolic pathway (not genetically engineered), must not be missing any specified key reactions, and must have an expected taxonomic range matching the organism in question. The latter criterion is problematic for studies such as this which include earlier-diverging land plants and algae on which relatively little metabolic research has been carried out, increasing the likelihood of false negatives and

limiting our ability to extend current knowledge beyond well-studied higher plants. Despite this taxonomy restriction, the Pathway Tools software also intrinsically reduces the number of false negative inferences caused by missing genome annotations – given a sufficiently high pathway likelihood score, complete metabolic pathways are incorporated into a PGDB, including any reactions for which an associated genome annotation has not been found. The low pathway likelihood score inclusion threshold maximises this effect.

Efforts were made to further mitigate the possibility of false negative inferences by improving the genome annotations on which the PGDBs were based (Figure B.1). Prior to PGDB construction, the raw protein sequences of all chosen species were passed through the E2P2 software (Ensemble-Enzyme Prediction Pipeline (Schlöpfer et al. 2017)), a machine learning-based algorithm developed for the extraction of metabolism-specific information from genome sequences. The algorithm is trained on a set of more than 142,000 protein sequences, compiled from enzyme and non-enzyme protein sequences from any organism taken from SwissProt (Bateman 2019), BRENDA (Jeske et al. 2019), MetaCyc or PlantCyc, with the stipulation that the entry must either have been manually curated or has experimental support (Schlöpfer et al. 2017). E2P2 automatically annotates input sequences for each organism with EC numbers and MetaCyc-specific reaction identifiers (E2P2 annotations for each of the seven species are available in Data 3.1). Following genome annotation with E2P2, Pathway Tools was used to produce two PGDBs, individually based on the previously published genome annotation for each organism and the E2P2 annotation produced here. The union of the metabolic pathways inferred in each case became the final PGDB for each organism. It is important to note that since this method, like the majority of gene annotation efforts, is based on sequence homology and protein domain identification, it has little power to identify novel metabolic enzymes (and therefore reactions and pathways). Rather, this approach attempts to produce the most complete set of metabolic pathways possible for each organism, given the totality of existing knowledge of metabolic reactions.

3.5.3 Analysis of metabolic pathways

The PGDBs for each of the seven streptophyte species detailed above and the well-curated *Arabidopsis* PGDB (Mueller et al. 2003) were consolidated into a binary presence/absence

table for all metabolic pathways present in any species PGDB. To compare the patterns of metabolic pathway presence/absence across species, k -medioids clustering followed by an unsupervised clustering technique known as t-distributed stochastic neighbor embedding (t-SNE (Van Der Maaten and Hinton 2008)) was carried out. For the clustering of species, t-SNE was run 500,000 times and the solution with the minimal final cost function value was chosen.

3.5.4 Identification of metabolic transitions

For a broader multi-species comparison, 64 plant and algal species in addition to the eight species described above were compared, taking the gene-reaction associations identified in the eight original species as the starting point. The additional species include seven chlorophyte algae, two further charophytes (*Spirogloea muscicola* and *Mesotaenium endlicherianum*), one further bryophyte (*Sphagnum fallax*), two further lycophytes (*Selaginella tamariscina* and *Isoetes echinospora*), one further fern (*Azolla filiculoides*), six further gymnosperms and 45 further angiosperms. All angiosperm genome sequence data were taken from the Phytozome online repository (Goodstein et al. 2012); sources for additional non-angiosperm species are identified in the key resources table (Appendix C). Genome sequence data was used for all additional species except the lycophyte *I. echinospora*, for which the transcriptome was used (available on NCBI's GenBank (Benson et al. 2017), GenBank: GGKY000000000.1). Genes for all 72 species were sorted into orthogroups using OrthoFinder software (Emms and Kelly 2019), which infers orthogroups of genes across species from sequence data (where an orthogroup is a collection of orthologous and paralogous genes descending from one gene in the common ancestor of all species included in the analysis). Orthogroups are produced using an all-versus-all DIAMOND (Buchfink et al. 2014) search followed by the normalisation of the resulting scores for gene length and phylogenetic distance. A gene network is then constructed, weighted with the normalised scores and this network is clustered into orthogroups using Markov clustering (MCL). Orthogroups for all 72 species and corresponding orthogroup gene trees are available as Data 3.2 and 3.3. A script to query this orthogroup data based on *Arabidopsis* accession IDs is provided in Data 3.4.

Both the gene-reaction association data and the orthogroup data were processed using a Python script (available as Data 3.5) to identify pathways representing possible metabolic

innovations or losses in land plant evolution. For every metabolic pathway across all representative species PGDBs, support for each metabolic reaction in the pathway in each of the genomes of the eight representative species was tabulated from the gene-reaction association data used by Pathway Tools. Equivalent support for each metabolic reaction in the 72 species on which orthogroups had been calculated was identified mainly on the basis of *Arabidopsis* annotations since *Arabidopsis* is the best studied species included in the analysis and likely has the highest quality genome annotation. If an *Arabidopsis* gene is annotated with a reaction, genes in other organisms found in the orthogroup corresponding to this *Arabidopsis* gene are considered functionally homologous and associated with the same reaction. If no *Arabidopsis* gene is annotated with a reaction, but at least two of the seven further representative species' genome annotations contain annotations for this reaction, then as long as an orthogroup exists which contains at least one such annotated gene from all the representative species, this orthogroup is considered associated with the reaction and corresponding genes in other organisms in this orthogroup are again considered functionally homologous.

Evidence for each metabolic pathway from both the gene-reaction associations and OrthoFinder were consolidated into a set of tables containing this information across the eight representative species (Figure B.1). Reaction presence was evaluated sequentially from earlier-diverging to later-diverging species. If this progression along the phylogeny was associated with a single change in the proportion of reactions evidenced (e.g. at a single point the proportion of reactions for which there is genetic evidence increases and remains high for the rest of the progression, or decreases and remains low), the metabolic pathway was flagged as involved in a possible metabolic innovation or loss in land plant evolution and sent to a results file containing all such pathways. The orthogroups associated with each reaction in each flagged metabolic pathway were then inspected to confirm that the evidence pattern held across the 64 further species included in the orthogroups. Evidence for each flagged metabolic pathway in each of the eight species annotated was scrutinised to remove false positive hits (e.g. where existing literature showed the identified pattern to be incorrect) and those for which limited reaction information was available, and to ensure that assumptions were not being made based on unreasonably large orthogroups.

Some orthogroups are too large to assume that the spectrum of included genes may have similar function; in these situations gene trees for each orthogroup were investigated and genes were only considered candidates for functional homology if closer relationships were observed within subtrees. For the purposes of removing false positives, the gene trees were investigated when genes required for pathways fell in one of the largest 200 orthogroups (containing more than 860 genes across the 72 species). For the genes presented in the results section of this work, all gene trees of any size were inspected.

3.5.5 Transcriptome analysis

Seven blast databases were created from all available transcripts for charophytes, liverworts, mosses, hornworts, lycophytes, ferns and gymnosperms individually using DIAMOND. For each gene, we identified whether reverse blast hits of the *Arabidopsis* gene in question are present in the relevant species' transcriptomes. Genes from two species are considered to be reverse blast hits when each is the top-scoring DIAMOND result in a sequence comparison against the other species' genome/transcriptome. For the initial blast, the total number of DIAMOND results was set to 100, and the single top scoring genes from the top 15 scoring species were extracted for reverse-comparison in a second DIAMOND search back against the *Arabidopsis* genome. Transcripts from each set of species which pass this reciprocal test were added to the multiple sequence alignment of all the genes in the corresponding orthogroup using MAFFT (Katoh and Standley 2013) and the resulting alignments trimmed using trimAl (Capella-Gutiérrez et al. 2009) with the 'gappyout' option. We then built a phylogenetic tree from the new multiple sequence alignment using IQ-TREE (Nguyen et al. 2015) with the model 'LG+I+G4' and inspected each phylogeny manually to ensure that the transcripts identified are indeed most closely related to the original *Arabidopsis* gene in question (as opposed to forming an outgroup, or being homologues of alternative *Arabidopsis* genes).

3.6 RESULTS

3.6.1 Bioinformatic approach and justification

The metabolic capabilities of an organism can be inferred from its genome sequence. However, even when the genome is well annotated, this is a time-consuming process, involving complex

algorithms to infer the presence or absence of whole metabolic pathways. Therefore, rather than attempting to make a detailed comparison of metabolic pathways present in all plant species for which genome sequences are available, the initial comparison was simplified to a smaller subset of species representing each of the major lineages in plant evolution, i.e. charophyte algae, bryophytes, lycophytes, ferns, gymnosperms, and angiosperms. The choice of these organisms was made to maximise the breadth of plant phylogeny covered while maintaining a feasible number of datasets for analysis. Since well-curated information on *Arabidopsis* metabolism is already available (Mueller et al. 2003), efforts here were focused on earlier-diverging land plants and algae and directed mainly based on the availability of sequenced genomes, which are sparse for such organisms. In fact, sequence data for four of the included species have only become available in the last two years. Two streptophyte algae genomes have been sequenced. Both — *Klebsormidium nitens* (Hori et al. 2014) and *Chara braunii* (Nishiyama et al. 2018) — were included in this analysis. Two bryophytes were chosen: the moss *Physcomitrella patens* (Rensing et al. 2008), a model organism and the bryophyte with the most well-developed genome annotation, and *Marchantia polymorpha* (Bowman et al. 2017), an emerging model organism. *Selaginella moellendorffii* (Banks et al. 2011) was selected as a representative lycophyte and the fern *Salvinia cucullata* (Li et al. 2018) was selected as a representative fern. Finally, *Picea glauca* was selected as a representative gymnosperm (Van Ghelder et al. 2019). Note that due to concerns about the reliability of assembly of the conifer megagenome, we used transcriptome data for this species.

Metabolic pathways from the MetaCyc (Caspi et al. 2018) and PlantCyc (Schl pfer et al. 2017) databases were identified from the genome sequences of these seven species using the Pathway Tools software (Karp et al. 2016). To increase the robustness of the analysis, two parallel Pathway Tools analyses were carried out, one based on annotation files compiled from each species' published genome annotation, and one using EC number annotations generated from E2P2, a machine learning-based algorithm for metabolic annotation (Schl pfer et al. 2017). The union of the predicted sets of pathways from these two approaches formed the final pathway/genome database (PGDB) for each species. Possible metabolic innovations and losses were inferred when the number of genes encoding enzymes associated with metabolic

pathways either increased or decreased at any point, suggesting the gain or loss of the metabolic pathway at the branchpoint in question.

Because the eight species are only a representative sample of a large phylogenetic space, each identified metabolic gain or loss across the eight species was mapped onto genetic data for 72 phylogenetically ordered species from the Chloroplastida (a full, labeled phylogeny of all included species is provided as Figure B.2). The species investigated consist of 46 angiosperms, seven gymnosperms (*Ginkgo biloba* (Guan et al. 2016), *Gnetum montanum* (Wan et al. 2018), *Pseudotsuga menziesii* (Neale et al. 2017), *Pinus lambertiana* (Stevens et al. 2016), *Pinus taeda* (Neale et al. 2014), *Picea glauca* (Van Ghelder et al. 2019), *Picea abies* (Nystedt et al. 2013)), two ferns (*Salvinia cucullata* (Li et al. 2018), *Azolla filiculoides* (Li et al. 2018)), three lycophytes (*Selaginella moellendorffii* (Banks et al. 2011), *Selaginella tamariscina* (Xu et al. 2018), *Isoetes echinospora* (unpublished data; GenBank: GGKY000000000.1), two mosses (*Physcomitrella patens* (Rensing et al. 2008), *Sphagnum fallax* (Goodstein et al. 2012)), one liverwort (*Marchantia polymorpha* (Bowman et al. 2017), four charophytes (*Klebsormidium nitens* (Hori et al. 2014), *Chara braunii* (Nishiyama et al. 2018), *Spirogloea muscicola* (Cheng et al. 2019), *Mesotaenium endlicherianum* (Cheng et al. 2019)), and seven chlorophytes (*Ostreococcus lucimarinus*, *Micromonas* sp. RCC299, *Micromonas pusilla* CCMP1545, *Coccomyxa subellipsoidea* C-169, *Chromochloris zofingiensis*, *Volvox carteri*, *Chlamydomonas reinhardtii* (Goodstein et al. 2012)). For 70 of these species, genome sequences were used; for *P. glauca* and *I. echinospora*, transcriptome data were used. All genes from the 72 species were placed into orthogroups using OrthoFinder software (Emms and Kelly 2019). Each of the metabolic gains/losses identified from the eight-species comparison was then tracked across the 72-species orthogroups using gene trees. In the following sections, only metabolic innovations and losses that held true beyond the eight representatives of the major evolutionary groups are presented. To provide additional support for results that relate to non-angiosperms (which are under-represented in terms of genome availability), 305 non-angiosperm land plant and algal transcriptomes taken from the One Thousand Plant Transcriptome Initiative (Leebens-Mack et al. 2019) were analysed to identify homologues of the genes required for the presented metabolic pathways in these species.

3.6.2 *De novo* genome annotation provides gene-to-reaction associations for previously unannotated metabolic genes

EC number annotations identified using E2P2 were compared to published genome annotations for each of the seven non-angiosperm land plants and algae for which PGDBs were generated. Novel metabolic gene annotations from E2P2 were identified in all seven species, both in the form of completely new annotations for genes previously described only as open reading frames and as additional annotations for genes already associated with another function or metabolic reaction. An overview of the number and classification (i.e. completely new annotations or additional annotations) of novel metabolic gene annotations for each species is shown in Figure B.3, and tables containing EC number and MetaCyc reaction ID annotations produced by E2P2 for each species are provided in Data 3.1, which also contains full PGDBs for each species based on the union of both annotations, provided as lists of MetaCyc/PlantCyc metabolic pathway IDs. Additionally, orthogroups identified using OrthoFinder across the complete set of 72 species is provided in Data 3.2, and the gene trees for each orthogroup are provided in Data 3.3. A Python script allowing this information to be queried based on associations between MetaCyc/PlantCyc reaction identifiers and *Arabidopsis* genes is supplied as Data 3.4.

3.6.3 Inferred metabolic pathways capture known phylogenetic relationships

To confirm that the PGDBs of representative species contain information applicable to the study of plant evolution, *k*-medioids clustering followed by a dimensionality reduction technique known as t-SNE (Van Der Maaten and Hinton 2008) was carried out to generate a cluster map of metabolic pathway similarity among the species to examine how this relates to phylogeny. The analysis grouped species based on the presence and absence of the metabolic pathways in each species' PGDB. As can be seen in Figure B.4, the green algae form a two-member cluster distinct from the land plants. Terrestrial plants were split into two clusters, one containing all land plants up to and including *Picea* and the other containing only *Arabidopsis* (Figure B.4). This isolation of *Arabidopsis* is potentially a consequence of the presence of angiosperm-specific metabolic pathways in its PGDB. However, it may also be that the difference is influenced by the level of manual curation involved in the production

of the *Arabidopsis* PGDB - the other PGDBs developed in this analysis were not manually curated. Nevertheless, the nearest neighbors of *Arabidopsis* as placed by t-SNE on this graph are the two next most recently diverging land plants, the euphyllophytes *Picea* and *Salvinia*. The two bryophytes, *Physcomitrella* and *Marchantia*, clustered together and were grouped with the earliest diverging member of the vascular plants, the lycophyte *Selaginella*.

Overall, a clear distinction between earlier-diverging and later-diverging lineages from streptophyte algae to angiosperms can be seen on the graph. Based only on the presence and absence of metabolic pathways in each species' PGDB, it was thus possible to capture the known phylogenetic relationships between the plant species analysed (Figure B.4). This indicates that the metabolic information gathered and included in the PGDBs reflects true differences between the species and supports the use of such information in the analysis of the evolution of metabolism.

3.6.4 Comparative analysis identifies previously identified metabolic innovations and losses in land plant evolution

Orthogroup analysis identified 50 metabolic pathways that showed gains or losses over the course of land plant evolution (Figure 3.1). To prioritise metabolic pathways for further investigation, we considered the novelty of the observation and the number of reactions in the pathway. The latter criterion was used because some of the metabolic “pathways” as categorized by Pathway Tools are extremely short, containing only one or two reactions and these were not considered further. The 17 pathways highlighted by dark blue circles in Figure 3.1 were the ones that were further investigated. Table B.1 contains orthogroup and gene IDs for each of the pathways investigated in the following sections.

The reliability of the approach was demonstrated by the fact that three known metabolic changes during land plant evolution were correctly identified. The first example is the metabolic pathway for diacylglycerol-N,N,N-trimethylhomoserine (DGTS) biosynthesis, which is thought to have been lost in spermatophytes where DGTS has been replaced by phosphatidylcholine (Riekhof et al. 2014; Rozentsvet 2004; Rozentsvet et al. 2000). Our analysis picked out this metabolic pathway because of the presence of genes encoding the required enzyme only in the genomes of algae, bryophytes, lycophytes, and ferns (Figures 3.1 and B.5). Homologous genes were not identified in any gymnosperm or angiosperm

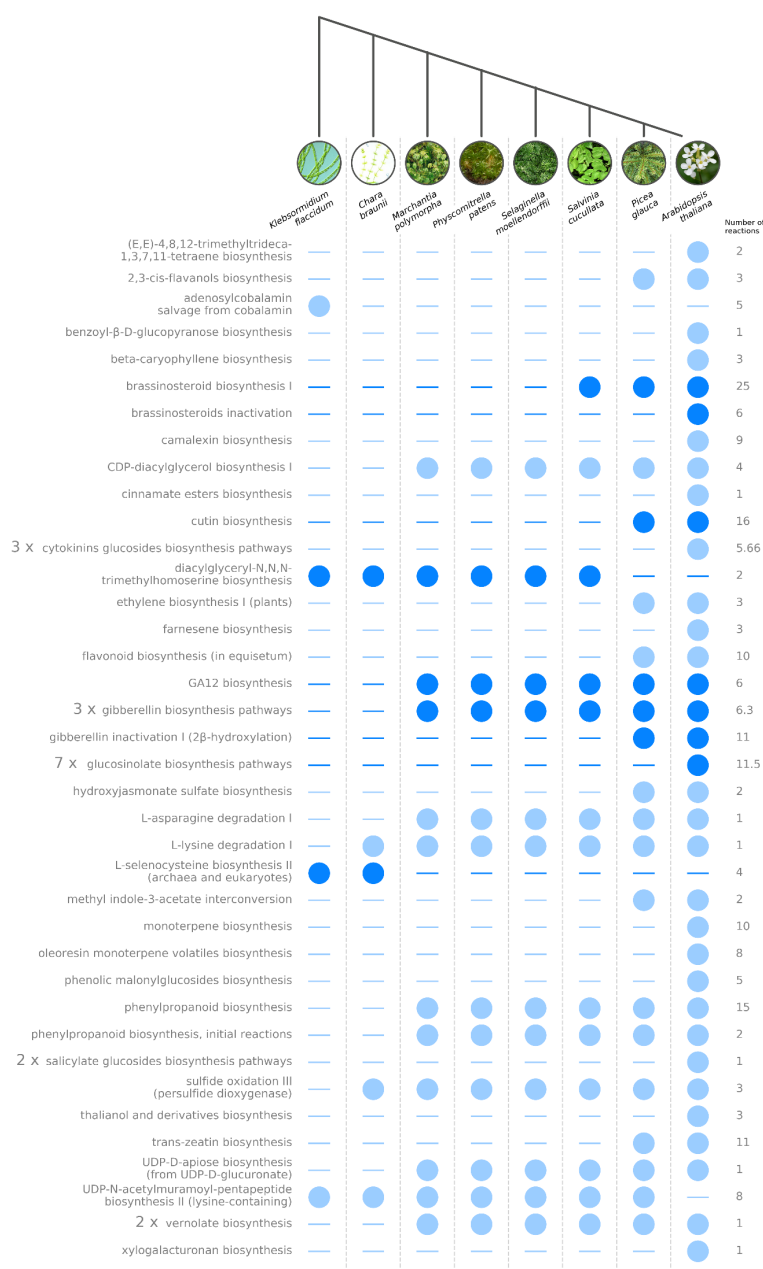


Figure 3.1: Overview of metabolic gains and losses identified by the orthogroup analysis. For each representative species, the presence or absence of the metabolic pathway in question is denoted by a filled circle (presence) or a horizontal line (absence). The far-right column shows the number of reactions in each metabolic pathway; where multiple similar reactions are grouped, e.g. for glucosinolate biosynthesis, this number denotes the average number of reactions across the grouped pathways. Dark blue circles indicate the metabolic pathways discussed in further detail. All metabolic pathway names are taken from the MetaCyc or PlantCyc metabolic pathway databases; Table B.2 contains a list of the MetaCyc/PlantCyc IDs corresponding to each of the metabolic pathway names listed. Figure B.2 shows a full phylogeny for the 72 species across which these metabolic gains/losses were identified, while Figure B.3 contains an overview of the additional EC number annotations for each representative species for which PGDBs were constructed. Figure B.4 shows the results of clustering these PGDBs based on metabolic pathway content.

species. This pattern is consistent with the previous identification of DGTS in ferns but not seed plants (Rozentsvet 2004; Rozentsvet et al. 2000).

The second example is the capacity for biosynthesis of selenocysteine, which is considered an ancestral character, found in bacteria, mammals, and green algae but lost in land plants (Novoselov et al. 2002). The pathway of selenocysteine biosynthesis was identified by our approach as a metabolic loss during land plant evolution (Figure 3.1). The gene encoding the first enzyme in the pathway is found in all plant taxa analysed, while the genes encoding the full combination of enzymes required for this pathway are found only in the algae (Figure B.6).

Finally, our analysis identified several metabolic pathways involved in glucosinolate biosynthesis from various amino acids as a metabolic capability that has been gained during land plant evolution (Figure 3.1). We found only one or two genes encoding the involved enzymes in algae and non-angiosperm plants and never the complete pathway (Figure 3.2B). The genes were more commonly found in angiosperms, but the complete pathway was only encoded by the genomes of the Brassicaceae in our analysis (Figure 3.2). This is in agreement with the hypothesis that glucosinolates are a synapomorphy of the Capparales (Clausen et al. 2015; Fang et al. 2019; Rodman et al. 1998). The analysis shown in Figure 3.2 also reveals a more detailed picture of the evolution of glucosinolate biosynthesis. The presence of genes for several different glucosinolate-related enzymes across the angiosperms is unsurprising, given that several of these encoding genes are members of large gene families (e.g. glucosyltransferases and monooxygenases) and the evolution of glucosinolate-specific enzymes likely evolved from similar pre-existing pathways such as those for cyanogenic glucosides production (Halkier and Gershenzon 2006). However, the presence of genes encoding the enzyme glucosinolate γ -glutamyl peptidase (GGP) in gymnosperms and flavin-containing monooxygenases (GS-OX) in species of every clade from chlorophyte algae to angiosperms suggests a conserved ancient function for these genes outside of the biosynthesis of glucosinolates.

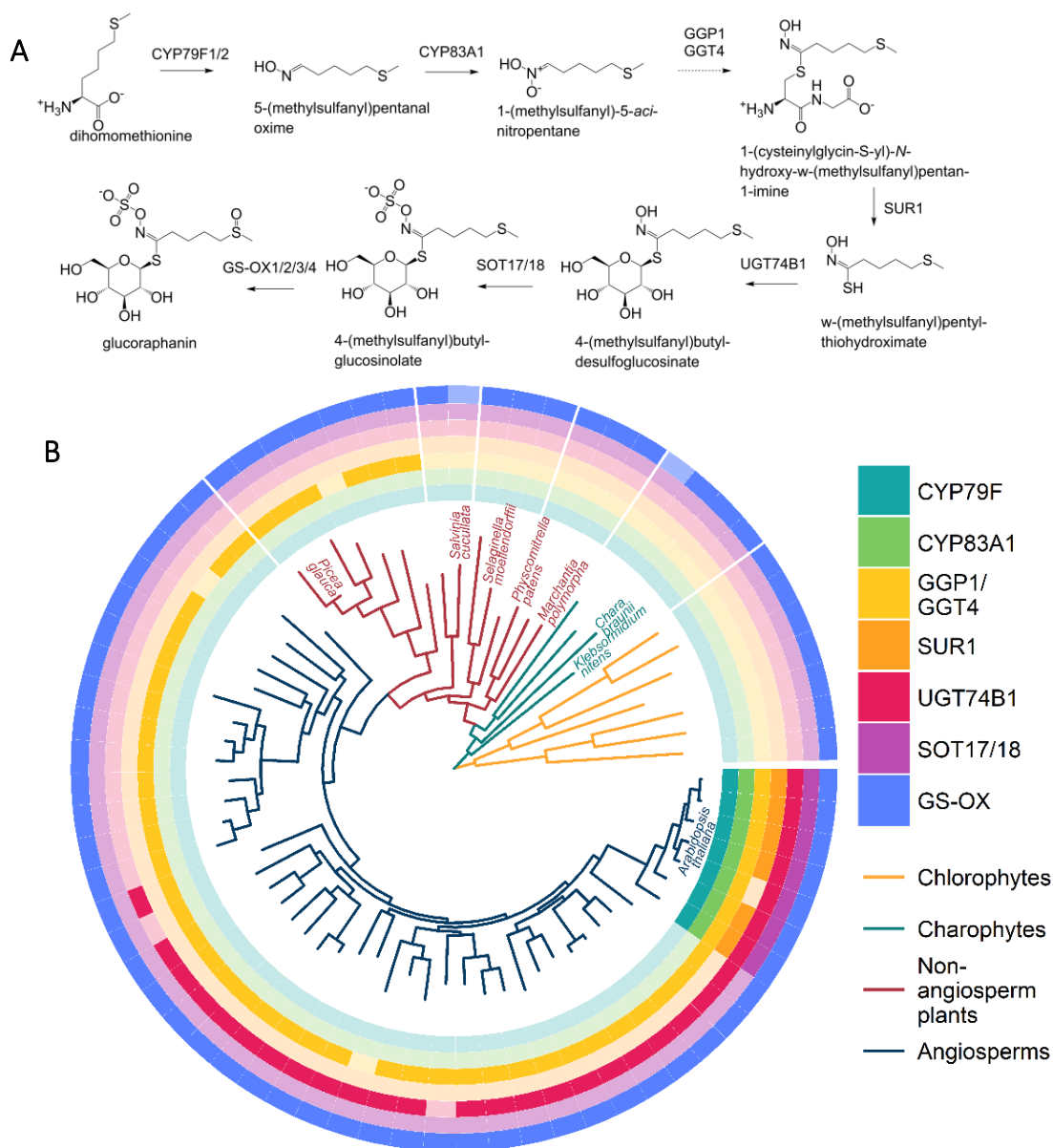


Figure 3.2: Presence of gene homologues encoding glucosinolate biosynthesis enzymes across 72 plant and algal species. **A:** Biosynthetic pathway for glucosinolates. Although several variants of this pathway were identified as possible evolutionary transitions (all utilising the same enzymes), only glucosinolate biosynthesis from a dihomomethionine precursor is shown. **B:** Homologue presence (coloured rings) for genes involved in glucosinolate biosynthesis across 72 studied species. Dark colours indicate the presence of a homologue while light colours indicate absence. Orthogroup information for the genes involved is provided in Table B.1. CYP, cytochrome P450; GGP, γ -glutamyl peptidase; GGT, γ -glutamyl transpeptidase; SUR, alkyl-thiohydroximate C-S lyase; UGT, UDP-glucosyltransferase; GS-OX, glucosinolate S-oxygenase.

3.6.5 Biosynthetic capabilities for gibberellin and brassinosteroid production are encoded in the genomes of non-angiosperm plants

The appearance of enzymatic capabilities for synthesis of certain plant hormones has been linked to evolutionary adaptations to terrestrial environments (Berens et al. 2017; Wang et al. 2015). Consistent with this, genes encoding enzymes involved in both gibberellin (GA) and brassinosteroid (BR) biosynthesis were associated with the transition to land in the 72-species comparison - i.e. were absent from all algal species analysed (Figures 3.3 and 3.4). The enzymes comprising the GA biosynthesis pathway from *ent*-kaurene to GA₁₂ and its conversion to bioactive forms (Figure 3.3A) were not found in the charophyte or chlorophyte algal genomes (Figure 3.3B). Similarly, genes encoding enzymes for the majority of the metabolic pathway of BR biosynthesis were not found in these two algal groups (Figure 3.4B). The algae do, however, contain homologues of DET2 encoding $\alpha 5$ steroid dehydrogenase, the first enzyme in the pathway that our analysis suggests is present throughout the land plant lineage (Figure 3.4B).

Previously, it has been suggested that GAs are confined to vascular plants (Hayashi et al. 2010). However, our analysis demonstrates that the enzymes required for their biosynthesis are encoded in bryophyte genomes. The two enzymes responsible for biosynthesis of the inactive GA precursor GA₁₂ (*ent*-kaurene oxidase (KO) and kaurenoic acid oxidase (KAO) (Figure 3.3A)) are present in all land plants except the mosses *Physcomitrella patens* and *Sphagnum fallax*, which are missing KAO (Figure 3.3B). This supports previous reports that *P. patens* is missing KAO (Miyazaki et al. 2018; Nelson and Werck-Reichhart 2011; Zi et al. 2014) and extends this finding to an additional moss species. The presence of KO in all land plants is consistent with observations that its metabolite product, the GA₁₂ intermediate KA, has been widely found in plants (Kumar et al. 2016; Noguchi et al. 2018). However, there has been no metabolomic evidence for the occurrence of recognised active forms of GA in bryophytes (Hirano et al. 2007; Miyazaki et al. 2018). Yet, our analysis provides evidence for the presence of genes encoding GA20ox and GA3ox enzymes in all land plant clades analysed, including the liverworts. These enzymes are 2-oxoglutarate-dependent dioxygenases responsible for the oxidation of GA₁₂ into bioactive GAs (Figure 3.3A). GA oxidase homologues have been previously identified in *P. patens* (Huang et al. 2015), and

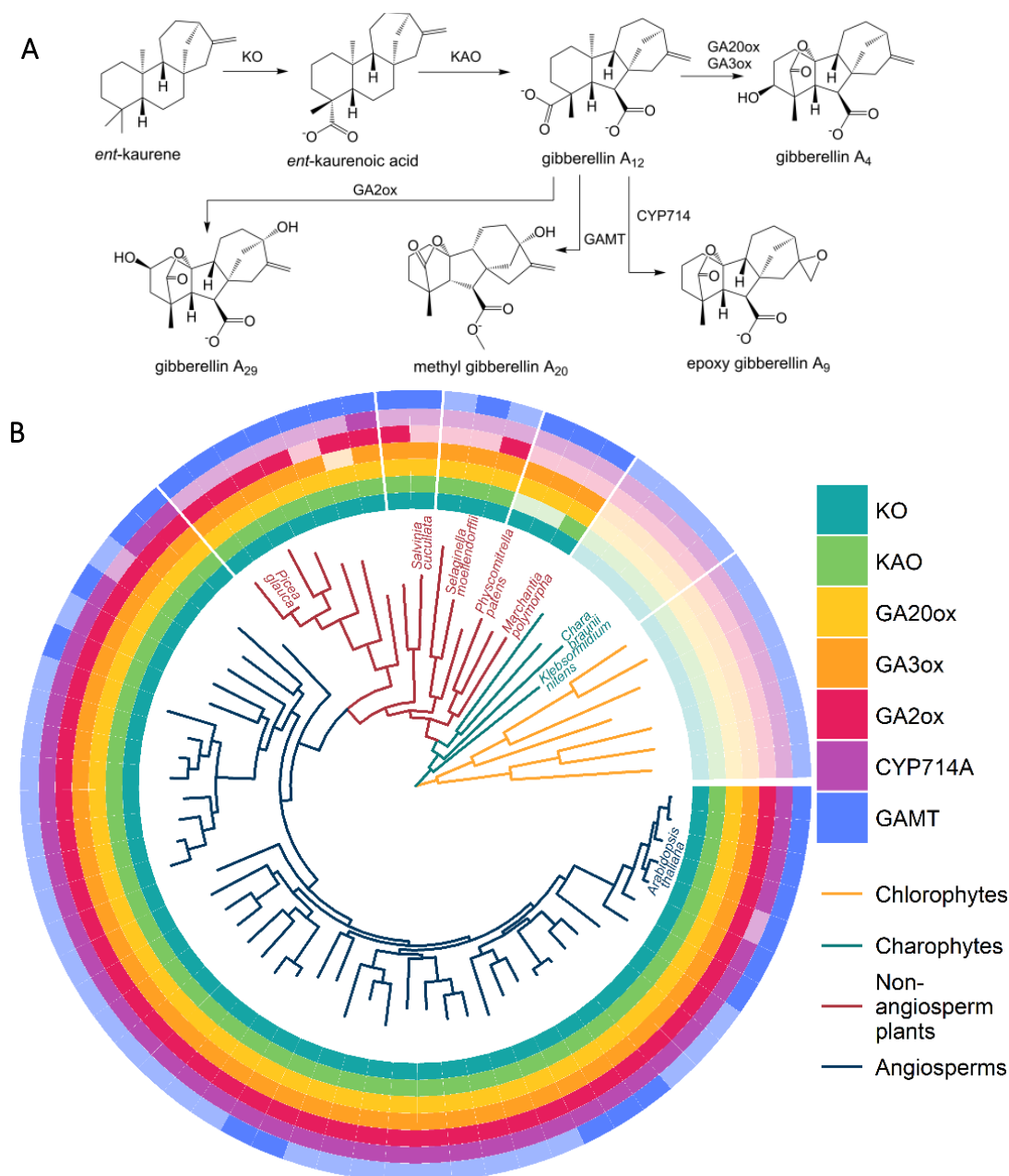


Figure 3.3: Presence of gene homologues encoding gibberellin biosynthesis and inactivation enzymes across 72 plant and algal species. **A:** The gibberellin biosynthetic/inactivation pathway. **B:** Phylogeny showing the presence and absence of homologues for genes (coloured rings) involved in gibberellin biosynthesis across 72 analysed species. Darker ring colours indicate gene presence while lighter colours indicate absence. Internal phylogeny colours show broad plant groupings. Orthogroup information for the genes involved is provided in Table B.1. KO, *ent*-kaurene oxidase; KAO, *ent*-kaurenoic acid oxidase; GA20ox, gibberellin 20 oxidase; GA3ox, gibberellin 3 oxidase; GA2ox, gibberellin 2 oxidase; CYP, cytochrome P450; GAMT, gibberellin methyltransferase.

the identification here of homologues in moss *Sphagnum fallax* and the liverwort *Marchantia polymorpha* supports the evolution of GA oxidases in bryophytes.

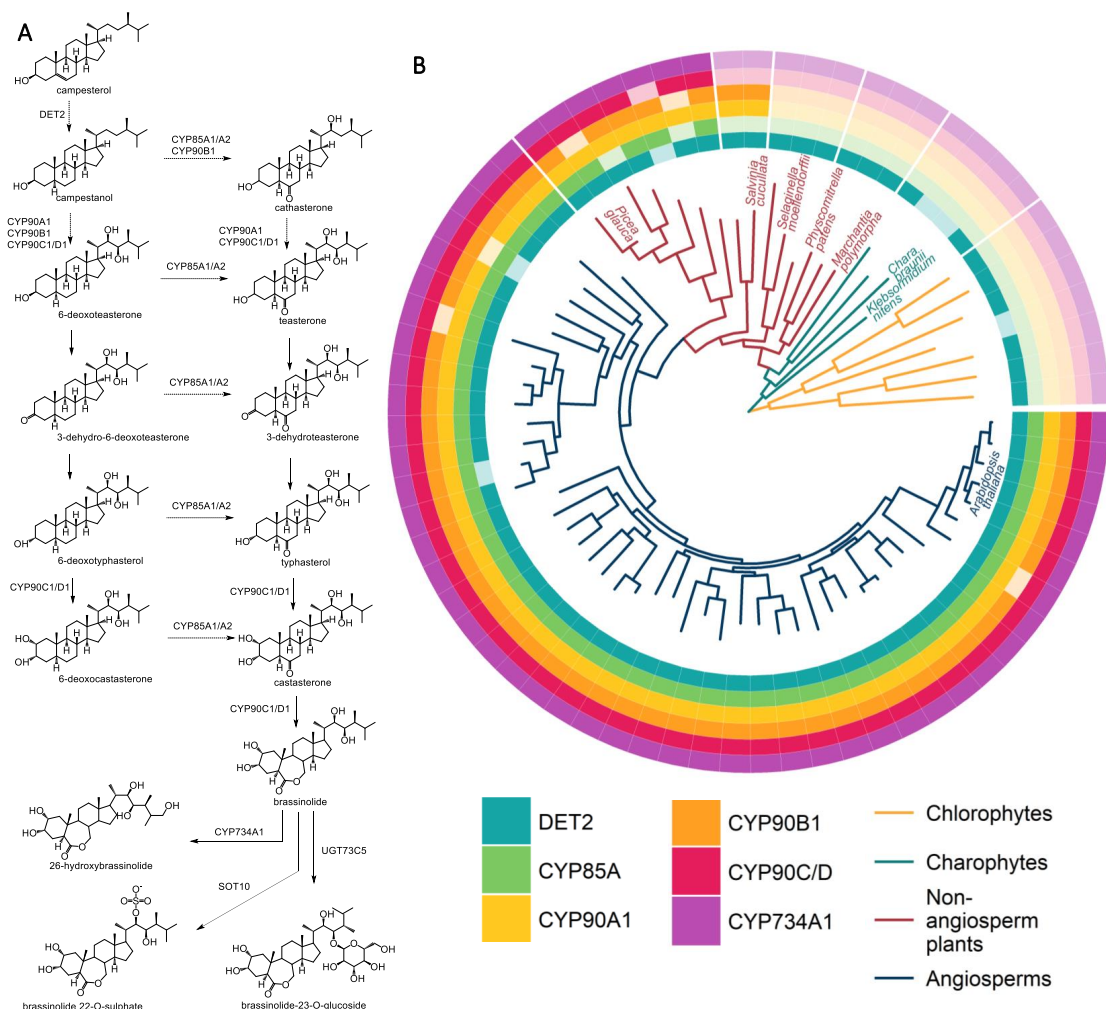


Figure 3.4: Presence of gene homologues encoding brassinosteroid biosynthesis and inactivation enzymes across 72 plant and algal species. **A:** The brassinosteroid biosynthetic/inactivation pathway. **B:** Evidence (coloured rings) for homologues of brassinosteroid biosynthetic enzymes across 72 plant species. Darker colours indicate presence of gene homologues, while lighter colours indicate absence. Internal phylogeny colours show broad plant groupings. Degradation steps are shown only for brassinolide; however, the same enzymes also act to degrade castasterone. Orthogroup information for the genes involved is provided in Table B.1. A simplified gene tree showing relationships between homologous genes across species in the orthogroup containing CYP734A1 is provided as Figure B.7. DET2, 5 α steroid dehydrogenase; CYP, cytochrome P450; UGT, UDP glucose:cytokinin glucosyltransferase; SOT, brassinosteroid sulfotransferase.

Conversely, beyond the initial reaction catalysed by enzymes encoding DET2, our analysis did not identify genes encoding BR biosynthesis enzymes in bryophytes or algae. Following the synthesis of campestanol by the ubiquitous DET2 enzyme, BR biosynthesis follows a series of oxidative modifications in two overlapping pathways (Figure 3.4A) catalysed by cytochrome P450 (CYP) enzymes. All *Arabidopsis* enzymes involved are members of the

CYP85 clan - either the CYP85A subfamily (involved in C6 oxidation and ring extension) or the CYP90A/B/C/D subfamilies (which catalyse C22 and C23 hydroxylation). The BR castasterone has been identified in bryophytes, albeit at concentrations that are orders of magnitude lower than in angiosperms (Yokota et al. 2017). The lack of CYP85A and CYP90 homologues in bryophytes and algae suggests either that the biosynthetic route for castasterone diverges from that shown in Figure 3.4A in these species or that other cytochrome P450s are involved in its biosynthesis. Green algae are known to contain CYP85 clan enzymes with divergent sequences from those in angiosperms (Nelson and Werck-Reichhart 2011) — it is possible that these genes and their relatives in the bryophytes are responsible for BR biosynthesis in these species but were replaced by CYP85A and CYP90 genes in spermatophytes. The latter scenario may also explain our finding of the absence of homologues of genes encoding BR biosynthesis enzymes in the lycophytes (Figure 3.4B). BRs have been found in lycophytes (Cheon et al. 2013; Yokota et al. 2017), but we did not find specifically identifiable CYP85A and CYP90A/B/C/D homologues in either the *S. moellendorffii* and *S. tamariscina* genomes or in the *I. echinospora* transcriptome (Figure 3.4B). However, we did identify genes in *S. moellendorffii* and *S. tamariscina* forming a sister group to the clade containing specific CYP90C and CYP90D genes in other species. While these genes are not specifically identifiable as CYP90 class C or D, their close relationship suggests they may be candidate genes for BR biosynthesis in lycophytes. Six gymnosperm genomes were found to encode specific CYP90D1 genes. CYP90C1 is the closest relative of CYP90D1 in angiosperms; presumably, the former arose from a gene duplication after the divergence of the gymnosperms. However, these genes are functionally redundant (Ohnishi et al. 2006), and we therefore consider the angiosperm-like BR biosynthesis pathway to be complete in the gymnosperms. Three of the gymnosperm genomes analysed contain homologues of all angiosperm BR biosynthetic genes, while the genomes of the ferns *S. cucullata* and *A. filiculoides* contain three biosynthetic genes (and are missing CYP90C/D and CYP85A homologues; Figure 3.4B), suggesting that BR biosynthesis in ferns represents an intermediate stage of evolution before the complete angiosperm-like pathway was established in the gymnosperms.

3.6.6 Non-angiosperm plants lack genes encoding known enzymes for gibberellin and brassinosteroid inactivation

A key feature of the use of specialised metabolites as hormones in angiosperms is rapid regulation of the concentration of bioactive forms of the molecules by enzymatic conversion to inactive forms. Our analysis suggests that many of these mechanisms are absent from non-angiosperm plants (Figures 3.3 and 3.4). GAs are inactivated in angiosperms using either GA 2-oxidases, GA 13-oxidases, GA 16,17-oxidases, or GA methyltransferases (GAMTs) (He et al. 2019; Rieu et al. 2008). Two orthogroups were identified that contained genes encoding GA 2-oxidases, one containing genes encoding enzymes that act on C19 GAs and one containing those that act on C20 GAs. With the exception of single homologues in *Isoetes echinospora*, *Azolla filiculoides*, and *Ginkgo biloba*, GA 2-oxidases acting on C20 GAs were not found in non-angiosperm plants. C19-acting homologues were not found in non-seed plants but were found in six gymnosperm species, while every angiosperm species contained both C19 and C20-acting homologues (Figure 3.3B).

Cytochrome P450 enzymes encoding GA 13-oxidases and GA 16,17-oxidases have been identified in rice (CYP714B1, CYP714B2, CYP714D1) and *Arabidopsis* (CYP714A1, CYP714A2) (Magome et al. 2013; Zhang et al. 2011; Zhu et al. 2006). These genes are all members of a single orthogroup containing genes exclusively from angiosperm species (Figure 3.3B) and *G. biloba*. One explanation for the presence of a CYP714A homologue in *G. biloba* is that GA inactivation via CYP714A enzymes may have evolved before the divergence of the gymnosperms but was lost in gymnosperms other than *G. biloba*. An additional GA 13-oxidase (CYP72A9) has recently been identified in *Arabidopsis* (He et al. 2019). CYP72A9 is the only CYP72A gene that encodes an enzyme with GA13ox activity in *Arabidopsis*. Our orthogroup analysis shows that similar CYP72A genes are present in other dicots - but the close grouping of CYP72A9 with other CYP genes in *Arabidopsis* makes it difficult to suggest candidate genes with GA 13-oxidase activity in other species.

Arabidopsis GAMTs are members of the SABATH family of methyltransferases, which act on various plant hormones and signaling molecules (Varbanova et al. 2007). As shown in Figure 3.3B, homologues of *Arabidopsis* GAMT1 and GAMT2 are inconsistently distributed across land plants and are not found in algae. The pattern of GAMT homology could mean

that genes in several branches of earlier-diverging land plants have convergently evolved methyltransferases with similar sequences to *Arabidopsis* GAMTs. Or, more likely, that GAMT-like methyltransferases are ancestral, and there has been sequence divergence in most angiosperms and some lycophytes. Regardless, it appears that bryophytes, lycophytes, ferns, and gymnosperms have the capacity to inactivate GAs, despite GAs having not yet been found in bryophytes.

There are several mechanisms for BR inactivation in angiosperms including hydroxylation, glucosylation, and sulfonation (Marsolais et al. 2007; Poppenberger et al. 2005; Turk et al. 2003). Each of these mechanisms is catalysed by a single gene product in *Arabidopsis*. The known *Arabidopsis* BR hydroxylation enzyme is CYP734A1, also known as CYP72B1. This is a member of a large orthogroup containing CYP72 clan genes including the CYP714 family responsible for GA inactivation. Although this orthogroup includes algal homologues, gene trees place the algal genes as an outgroup, suggesting that they are homologous to the ancestor of all angiosperm CYP genes in this orthogroup. Bryophyte, lycophyte and fern genes in this orthogroup cluster with alternative CYP genes, and there are no direct homologues of CYP734A1 in these species (Figure B.7).

Glucosylation of BRs in *Arabidopsis* is carried out by a UDP glucosyltransferase enzyme, UGT73C5, while sulfonation is carried out by BR sulfotransferase enzyme (SOT10). Orthogroup analysis places both these genes as members of large orthogroups containing other glucosyltransferases and sulfotransferases, respectively. However, both *Arabidopsis* genes in question were grouped closely with additional *Arabidopsis* genes known to have functions outside of BR inactivation — we were therefore unable to distinguish homologues in other species.

3.6.7 Bryophytes, lycophytes, ferns, and gymnosperms lack specific cytochrome P450 family enzymes involved in oxidation of structural monomers

The transition of plants to terrestrial environments was accompanied by the evolution of mechanisms that protect against desiccation, UV radiation, and exposure to a new complement of pathogens (Delwiche and Cooper 2015; Delwiche et al. 1989). The ability of land plants to withstand these stresses results in part from the evolution of several amino acid/lipid-based insoluble biopolymers such as cutin, suberin, and sporopollenin. Our

analysis identified the cutin biosynthetic pathway as a metabolic innovation during land plant evolution. As all three of these structural polymers are produced via biosynthetic pathways that use very similar (or overlapping) biochemical reactions, and several orthogroups are shared between pathways, all three pathways were investigated. To date, suberin has been found only in vascular plants, whereas cutin has been found across the land plants. Sporopollenin is found across land plants and has been isolated from the cell walls of several species of chlorophyte and charophyte algae (Delwiche et al. 1989; He et al. 2016). All three polymers are formed primarily of fatty-acid-derived monomers of varying chain lengths (sporopollenin C12–C18, cutin C16–C18, and suberin C18–C24 fatty acids), and suberin also contains significant amounts of phenolic compounds (Niklas et al. 2017; Vishwanath et al. 2015). The metabolic pathways for all three polymers rely on the activation of free fatty acids with coenzyme A (CoA) and use cytochrome P450 enzymes for the oxidation of fatty acids into polymer-specific monomers, several of which are shared by cutin and suberin (Figure 3.5A).

We identified homologues of the LACS genes required for fatty acid activation in cutin and suberin biosynthesis in the genomes of sampled land plants from all clades as well as in charophyte and chlorophyte algal genomes (Figure 3.5B). For sporopollenin biosynthesis, homologues of the ACOS5 gene responsible for the same process were not found in algae but were present in all land plants, while the ACH enzymes that remove CoA were found in charophyte algae and land plants (Figure 3.6B).

This suggests that the capability for biosynthesis of precursor molecules required for the production of cutin/suberin monomers were present prior to the evolution of land plants. However, the majority of homologues of the *Arabidopsis* genes encoding enzymes responsible for subsequent reactions in both metabolic pathways were identified only in land plants (Figures 3.5B and 3.6B) - this is consistent with the lack of cutin and suberin in algal species (Niklas et al. 2017). For sporopollenin, the lack of ACOS5 homologues in algae that do nevertheless contain sporopollenin suggests the use of a different, as-yet-unknown, metabolic pathway for sporopollenin biosynthesis in chlorophytes and charophytes.

Further investigation of the gene tree of the orthogroup containing *Arabidopsis* CYP86A and CYP86B subfamily genes, responsible for the ω -hydroxylation of fatty acids in cutin and

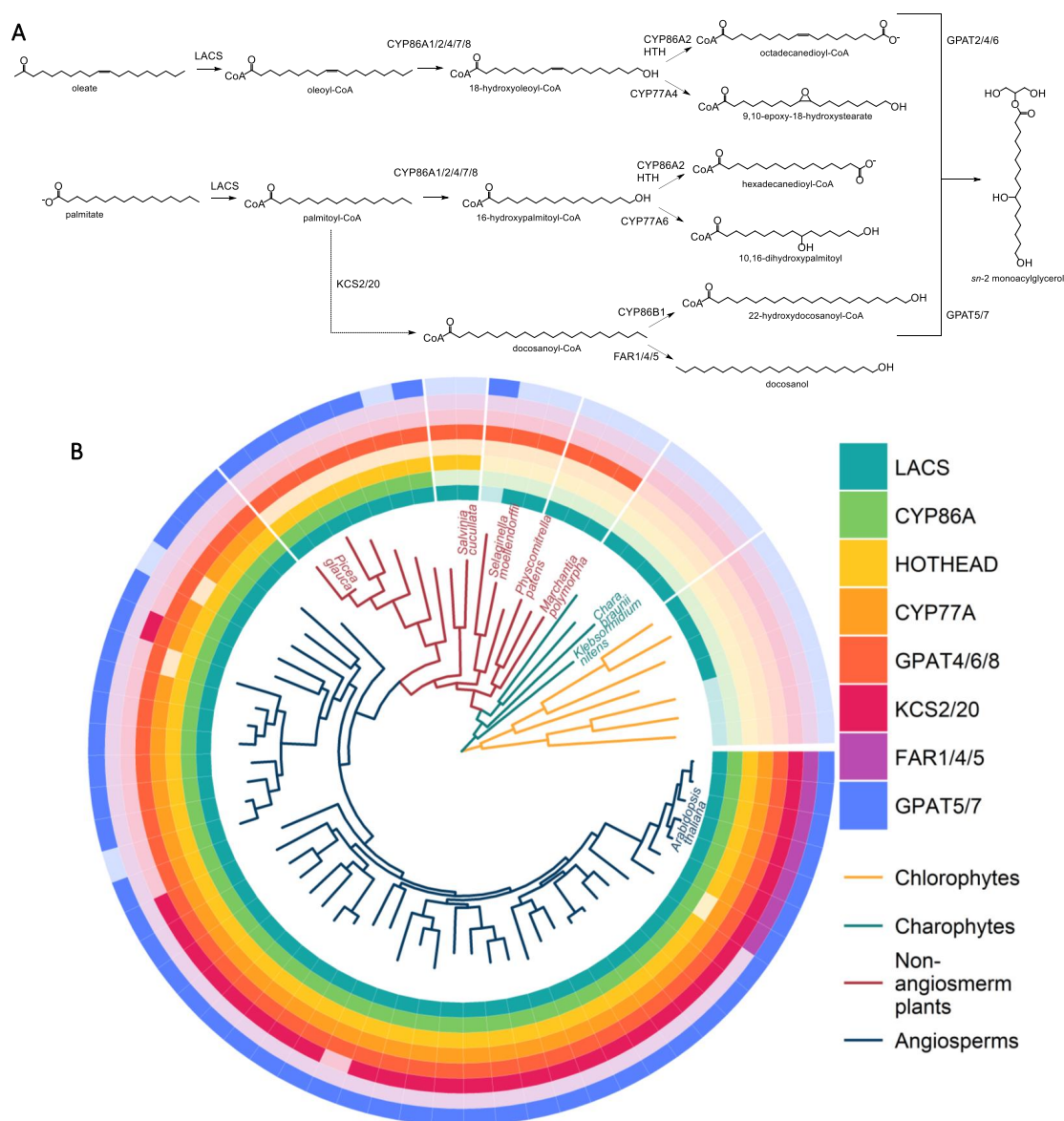


Figure 3.5: Presence of gene homologues encoding cutin and suberin biosynthesis enzymes across 72 plant and algal species. **A:** The cutin/suberin biosynthetic pathway. **B:** Evidence (coloured rings) for homologues of required enzymes across 72 plant species. Darker colours indicate presence of gene homologues while lighter colours indicate absence. Internal phylogeny colours show broad plant groupings. Orthogroup information for the genes involved is provided in Table B.1. A simplified gene tree showing relationships between homologous genes across species in the orthogroup containing CYP86 genes is provided as Figure B.8. LACS, long-chain acyl-CoA synthetase; CYP, cytochrome P450; KCS, 3-ketoacyl CoA synthase; FAR, fatty acid reductase; GPAT, glycerol-3-phosphate acyltransferase.

suberin biosynthesis, shows that, while gymnosperm and angiosperm species contain specific homologues of these *Arabidopsis* genes (Figure 3.5B), earlier-diverging land plants and algae

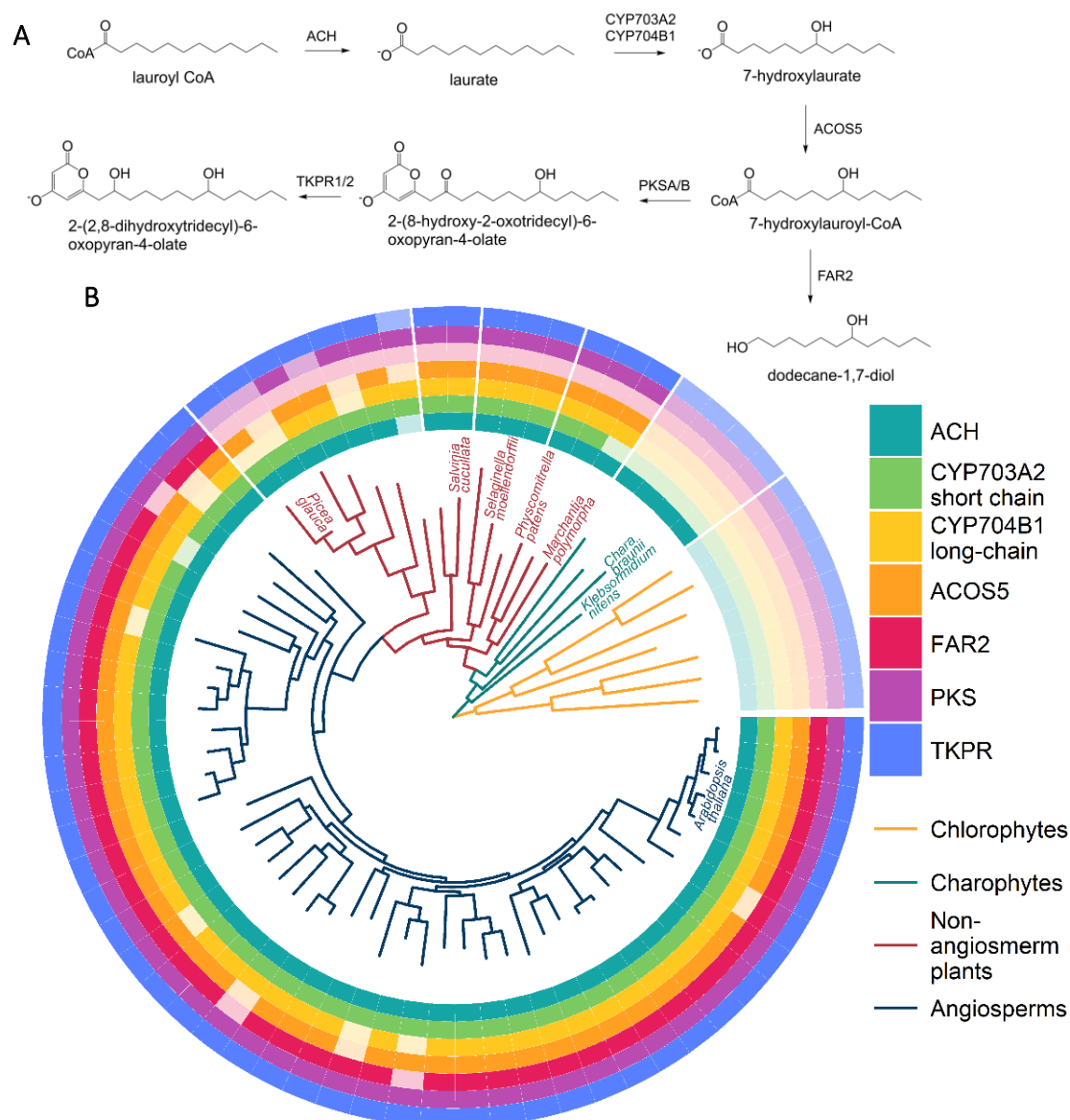


Figure 3.6: Presence of gene homologues encoding sporopollenin biosynthesis enzymes across 72 plant and algal species. **A:** The sporopollenin biosynthesis pathway. Although sporopollenin is produced via analogous reactions from several fatty acid precursors, the depicted pathway shows only those with a lauric acid precursor. **B:** Phylogeny of 72 species showing the presence and absence of homologues of sporopollenin biosynthetic genes (coloured rings). Darker colours indicate the presence of homologues while lighter colours indicate absence. Internal phylogeny colours show broad plant groupings. Orthogroup information for the genes involved is provided in Table B.1. ACH, acyl-CoA thioesterase; CYP, cytochrome P450; ACOS5, fatty-acyl CoA synthetase; FAR, fatty acid reductase; PKS, polyketide synthase; TKPR, tetraketide pyrone reductase.

contain homologues that can only be identified more broadly as related to the CYP86 and CYP96 subfamilies (Figure B.8). Direct homologues of the oxidoreductase gene HOTHEAD are also identified in the spermatophytes as well as in both fern species investigated but

not in the lycophytes, bryophytes, or algae (Figure 3.5B). These gene products in earlier-diverging land plants may also catalyse the ω -hydroxylation of fatty acids but could differ in characteristics such as chain-length preferences. This would be consistent with previously described shorter chain -lengths in the cutin biopolymers of earlier-diverging land plants (Fich et al. 2016). However, the lack of direct subfamily homologues mean that no specific function can be inferred for these species.

Similarly, we found CYP77A homologues, involved in in-chain hydroxylation and/or epoxidation, across angiosperm species. However, related genes identified in other land plants could only be classified more generally as CYP77 genes (Figure 3.5B). We found no CYP77 homologues in algae (Figure 3.5B), which is as expected given the lack of evidence for the occurrence of cutin or suberin in algae (Kondo et al. 2016). The presence of genes encoding the LACS enzyme across land plants, combined with the lack of specific CYP86A and CYP77A genes in non-seed plants and the lack of HOTHEAD in lycophytes and bryophytes, suggests that the early stages of cutin and suberin biosynthesis are conserved. However, the fatty acid monomers of cutin and suberin must be biosynthesised via alternative gene products (likely from CYP86/77 families) in bryophytes and lycophytes and later-diverging land plants, which may conceivably produce monomers of different chain lengths and with different molecular modifications. The evolution of HOTHEAD before the divergence of ferns, CYP86A/B in the ancestor of spermatophytes and CYP77A only in the angiosperms suggests that the angiosperm-like pathway for cutin and suberin biosynthesis may have evolved in stages, replacing an alternative set of enzymes in the bryophytes and lycophytes.

Regarding sporopollenin CYP enzymes, we found specific homologues of CYP704B1, which catalyses the hydroxylation of long-chain fatty acids in all land plant clades. Homologues of this gene were not found in the sampled algae. We found homologues of CYP703A2 responsible for hydroxylation of mid-length-chain fatty acids only in the vascular plants and the mosses *Physcomitrella patens* and *Sphagnum fallax* but not in the liverwort *Marchantia polymorpha* or any algae (Figure 3.6B). This suggests that the use of lauric acid derivatives in sporopollenin biosynthesis originated in the bryophytes.

Despite the lack of specific CYP704B1 and CYP703A2 homologues, sporopollenin is found in algae. One possible explanation for this is that the oxidation reactions for

sporopollenin biosynthesis are carried out by multifunctional CYP86 clan enzymes in algae, which when duplicated after the colonisation of land became specialised for the biosynthesis of sporopollenin and other biopolymers.

3.6.8 Conserved enzymes for production of secondary monomers found in suberin and sporopollenin are limited to angiosperms and land plants, respectively

The final stages of sporopollenin biosynthesis require polyketide synthases (PKSA/B) and tetraketide pyrone reductases (TKPR1/2) to produce tetraketide pyrones. No homologues of PKSA/B or TKPR1/2 were identified in the sampled algae (Figure 3.6B); our analysis suggests that tetraketide pyrones may be terrestrial plant-specific sporopollenin monomers.

Suberin and sporopollenin also contain various alcohols, the production of which requires fatty acid reductase (FAR) enzymes. In all but three of the angiosperm species (*Spirodela polyrhiza*, *Solanum tuberosum*, and *Malus domestica*), we found homologues of *Arabidopsis* FAR2, specific to sporopollenin biosynthesis. Homologues of *Arabidopsis* FARs specialised for suberin production (FAR1/4/5) were only found in the Brassicaceae (Figure 3.5). However, FAR family genes are found in all land plant taxa; those species missing FAR1/2/4/5 contain other FARs, which are more similar to *Arabidopsis* FARs of unknown or alternative function. This could mean that there has been significant sequence divergence from an ancestral specialised FAR. Or, that earlier-diverging land plants make use of multifunctional FARs for the production of these two polymers. Another possibility is that earlier-diverging land plant suberin and sporopollenin do not contain alcohols.

The use of very-long-chain fatty acids as suberin monomers means that a fatty acid elongation process is necessary. In angiosperms the enzymes responsible are 3-ketoacyl-CoA synthases (KCS2/20). As with FARs, we found homologues of *Arabidopsis* KCS genes in algae and land plants. However, in all species except dicotyledonous angiosperms and one monocot these genes are more related to alternative *Arabidopsis* KCS genes than KCS2/20 (Figure 3.5B).

After very-long-chain fatty acid production and oxidation, hydroxy fatty acids destined for cutin and suberin polymers are combined with glycerol-3-phosphate at the *sn*-2 position in a reaction catalysed by glycerol 3-phosphate acyltransferase (GPAT) enzymes. GPAT4/6/8 gene products catalyse the transfer of long-chain fatty acids, while GPAT5/7 catalyse the transfer of suberin-specific, very-long-chain fatty acids. Despite the lack of specific genes

encoding enzymes of the preceding oxidation and reduction steps, we found at least one homologue of *Arabidopsis* GPAT4/6/8 genes in all land plants (but none in algae) (Figure 3.5B). Consistent with evidence of reduced chain-length biopolymer monomers in earlier-diverging land plants (Fich et al. 2016) and presence of suberin only in vascular plants, specific homologues of GPAT5/7 were found only in the angiosperms and gymnosperms, and a single lycophyte (Figure 3.5B). It may be that the remaining vascular plants contain homologues of GPAT genes not yet characterised as suberin related.

The pattern of homology that we have described for structural molecule biosynthetic pathways supports the hypothesis that cutin appeared for the first time in the earliest land plants (Domínguez et al. 2015). LACS genes already present in algae for the initial activation of free fatty acids were likely involved in cutin synthesis in these species. Indeed, homologous genes encoding LACS are present in almost all species in this study. Similarly, the final enzymes in the cutin biosynthesis pathway (GPAT4/6/8) are encoded by all land plant genomes analysed (Figure 3.5B). However, non-seed plants, although containing genes belonging to the CYP86, CYP77, KCS, and FAR gene families, do not contain genes homologous to specific CYP86A/B and CYP77A subfamilies, or to KCS2/20 and FAR1/4/5. This suggests that oxidation reactions in cutin and suberin biosynthesis across taxa have been significantly modified since their origin in bryophytes and vascular plants, respectively, and the specific pattern of homologue presence of angiosperm biosynthetic genes in non-angiosperm plants suggest that the intermediate steps of the angiosperm-like pathway may have evolved in stages. The presence of chlorophyte algal genes in orthogroups containing *Arabidopsis* KCS and FAR genes suggest that, when suberin biosynthesis did develop, it was achieved by the co-option of pre-existing cutin biosynthetic CYP genes in combination with ancient elongase and reductase enzymes, the latter of which may have been multifunctional and also involved in sporopollenin biosynthesis. However, the lack of *Arabidopsis*-like FAR1/4/5 genes in non-Brassicaceae shows that significant modifications to suberin biosynthesis have occurred even within the angiosperms. With the exception of specific FAR2 enzymes, all land plant clades contain homologues of the enzymes required for sporopollenin biosynthesis. The missing homologues in algae suggest that algal sporopollenin is produced using alternative enzymes and may differ in its composition.

3.6.9 Testing metabolic gains on a broader selection of species using 1KP transcriptomes

To test our results from the orthogroup genome analysis, and to compensate for the relative lack of genome sequences available for non-angiosperm land plants and algae, we carried out an additional analysis of transcriptomes released by the One Thousand Plant Transcriptomes Initiative (Leebens-Mack et al. 2019). This also allowed the analysis to be extended to hornworts, for which there are no available genome sequences. While transcriptome analysis cannot be used to draw conclusions about gene absences, it can be used to confirm the presence of gene homologues. A total of 305 transcriptomes were analysed, comprising 81 gymnosperms, 78 ferns, 20 lycophytes, 37 mosses, 29 liverworts, 14 hornworts, and 46 charophytes. For each of the metabolic pathways that we analysed in detail in this paper, the presence of homologues of genes associated with metabolic reactions were tracked across the transcriptome dataset (Table 3.1). Comparison of the transcriptome results with the genome analysis (Figures 3.3, 3.4, 3.5, and 3.6) reveals broad corroboration for the conclusions drawn regarding the evolution of metabolic pathways for GA and BR biosynthesis/inactivation and the biosynthesis of structural polymers, thereby strengthening the conclusions. The transcriptome data also revealed additional details and new information, described in the subsequent sections.

Metabolic Pathway	Gene ID	<i>Arabidopsis</i> accession	Charophytes	Liverworts	Mosses	Hornworts	Lycophytes	Ferns	Gymnosperms
Gibberellin Biosynthesis	KO	AT5G25900	0/15	14/15	15/15	10/13	10/15	15/15	15/15
	KAO	AT1G05160; AT2G32440	0/19	9/16	0/21	9/12	9/18	16/16	16/17
	GA20ox	AT5G07200; AT5G51810; AT1G60980	0/16	8/21	19/22	1/12	7/16	16/17	16/16
	GA3ox	AT1G15550; AT1G80340	0/16	0/20	7/17	5/12	6/18	11/17	6/23
	GA20ox19	AT2G34555; AT1G30040	0/16	0/19	0/19	0/12	0/16	0/21	10/17
	GA20ox20	AT4G21200; AT1G50960	0/15	5/19	4/18	4/12	2/17	1/21	18/18
	CYP714	AT5G24910; AT5G24900	0/18	0/17	0/16	0/12	0/15	0/17	0/16
	GAMT	AT4G26420; AT5G56300	8/8	14/15	18/18	12/12	16/16	19/19	15/15
Brassinosteroid biosynthesis	DET2	AT2G38050	0/15	0/15	0/15	0/14	0/15	0/15	0/15
	CYP85A	AT5G38970; AT3G30180	0/22	0/16	0/16	0/12	0/16	0/20	0/19
	CYP90A1	AT5G05690	0/15	0/15	0/15	0/12	11/15	15/15	13/15
	CYP90B1	AT3G50660	0/15	0/15	0/15	0/12	0/15	15/15	15/15
	CYP90C/D	AT4G36380; AT3G13730	0/20	0/18	0/22	0/12	0/15	0/22	24/24
	CYP734A1	AT2G26710	0/15	0/15	0/14	0/12	0/15	0/15	0/15
Cutin/Suberin Biosynthesis	LACS	AT2G47240; AT1G64400; AT1G49430	22/22	21/21	24/24	12/13	18/18	30/30	32/33
	CYP86A	AT5G58860; AT2G45970	0/17	5/18	0/17	3/12	1/17	5/17	2/26
	HOTHEAD	AT1G72970	0/15	0/14	0/15	0/10	0/12	0/15	15/15
	CYP77A	AT3G10570; AT5G04660	0/16	0/17	0/17	0/12	0/16	0/17	0/19
	GPAT4/6/8	AT1G01610; AT2G38110; AT4G00400	0/0	1/16	0/19	0/10	18/18	21/23	29/30
	KCS2/20	AT1G04220; AT5G43760	0/17	0/16	0/19	0/10	0/15	0/20	0/19
	FAR1/4/5	AT5G22500; AT3G44540; AT3G44550	0/16	0/10	0/18	0/10	0/15	0/20	0/18
	GPAT5/7	AT3G11430; AT5G06090	0/2	0/16	0/17	0/10	0/18	0/17	0/20
Sporopollenin biosynthesis	ACH	AT1G01710	15/15	15/15	15/15	12/12	15/15	15/15	15/15
	CYP703A2	AT1G01280	0/15	1/15	14/15	2/12	12/15	15/15	15/15
	CYP704B1	AT1G69500	0/15	14/15	15/15	9/12	11/15	15/15	2/15
	ACOS5	AT1G62940	0/15	3/15	3/15	2/15	5/15	15/15	2/15
	FAR2	AT3G11980	0/15	0/10	0/15	0/10	0/15	0/15	0/15
	TKPR	AT4G35420; AT1G68540	0/17	7/20	9/18	4/11	15/17	23/23	24/24
	PKS	AT1G02050; AT4G34850	0/18	0/19	0/16	0/11	0/17	0/15	0/18

Table 3.1: Results of search for homologues of investigated genes in transcriptomes of 305 non-angiosperm plants and algae from the One Thousand Plant Transcriptomes Initiative. The numerator indicates the number of species in which transcripts encoding the enzyme (or enzyme group) were found. The denominator indicates the number of transcriptomes of each taxon analysed. A maximum of 15 unique transcriptomes per taxon were analysed, but note that the number of species hits will exceed 15 when more than one gene is responsible for a specified reaction, for example, in the case of LACS. Items are marked in bold when homologues of the gene/gene group in question were found in the corresponding transcriptomes.

For GA biosynthesis, the transcriptome data provide full support for the results drawn from the genome data with one minor exception: we did not find GA 3-oxidase transcripts in liverworts (Table 3.1), whereas this gene was identified in the genome of *Marchantia polymorpha* (Figure 3.3). The transcriptome data also reveal that hornworts are more similar to liverworts than to mosses with respect to GA biosynthesis: hornwort transcriptomes contain homologues of KAO, which are missing from moss genomes and transcriptomes (Table 3.1; Figure 3.3) (Miyazaki et al. 2018; Zi et al. 2014).

The transcriptome data also provide additional information about the evolution of GA inactivation enzymes: homologues of GAMT genes were identified in the transcriptomes of charophyte algae as well as land plants (Table 3.1). Additionally, although homologues of GA 2-oxidase genes acting on C20 GAs were not found in the genomes of *M. polymorpha*, *S. fallax*, or *P. patens* (Figure 3.3), they were identified in the transcriptomes of five other liverwort species, four other moss species, and four hornworts (Table 3.1). The presence of an additional inactivation mechanism in bryophytes provides further support for the hypothesis that GAs are likely present in the bryophytes even though this has yet to be demonstrated experimentally. The presence of GAMT homologues in charophyte algae transcriptomes indicates that these genes are more ancient in origin than previously discussed.

For the BR biosynthesis pathway, the transcriptome data were in broad agreement with the genome analysis (Figure 3.4; Table 3.1) with one substantive exception: we were unable to identify transcripts of DET2 in the transcriptome data despite the clear presence of DET2 gene homologues in all plant and algal genomes investigated (Figure 3.4). We also did not find transcripts of CYP85A or CYP734A1 genes in any species, although these are present in the genomes of gymnosperms (Figure 3.4). The transcriptome data support the absence of genes for BR biosynthesis in the bryophytes, including hornworts. Additionally, the transcriptome data extend our understanding of the evolution of BR biosynthetic genes in the CYP90 family. Genome analysis showed the presence of CYP90A1 and CYP90B1 in ferns and genes for the complete pathway are identified first in gymnosperms (Figure 3.4). Transcriptome analysis corroborated these findings, as well as that of lycophyte genes forming a sister group to the clade of genes containing gymnosperm and angiosperm CYP90C1 and CYP90D1 (Table 3.1). Additionally, the fern transcripts were similar to the lycophytes with

respect to CYP90C1 and CYP90D1. It seems likely that these genes in ferns and lycophytes represent a more generic version (perhaps with differing substrate specificity) of specific CYP90C/D genes in gymnosperms and angiosperms. Finally, lycophyte transcriptomes were shown to contain homologues of CYP90A1, which were not identified in the genome analysis. Combining these results further supports the stepwise evolution hypothesised for the BR biosynthesis pathway found in angiosperms, beginning with the evolution of CYP90A1 and genes similar to CYP90C/D in lycophytes, both of which are maintained in ferns with the addition of CYP90B1, before the completion of the biosynthetic pathway with the evolution of CYP90D1 in gymnosperms.

Finally, the transcriptome data are in full agreement with the genome analysis with respect to sporopollenin biosynthesis (Figure 3.6; Table 3.1) apart from the lack of PKS transcripts identified in any land plant taxon. The presence of transcripts involved in cutin and suberin biosynthesis were more scattered - we were unable to identify transcripts of several genes - specifically GPAT5/7 in vascular plants, HOTHEAD in ferns, and GPAT4/6/8 in bryophytes - which were identified in the genome analysis (Figure 3.5). Transcriptome analysis of cutin and suberin biosynthetic genes does, however, identify homologues of CYP86A genes in the liverworts, hornworts, lycophytes, and ferns, which suggests that the early stages of biosynthesis for these structural compounds are more similar across land plants than indicated by the genome analysis (Figure 3.5). Additionally, a single liverwort was found to contain a homologue of CYP703A2, which was not identified in the genome of *M. polymorpha* (Figure 3.5).

3.7 DISCUSSION

We have developed a scalable computational pipeline to systematically analyse and compare metabolic capabilities encoded in genomes and transcriptomes and have used it to advance our knowledge of the evolution of metabolic pathways in the Chloroplastida. The approach was validated by its ability to identify known metabolic innovations associated with plant evolution. The analysis allowed us to resolve several additional metabolic innovations in relation to land plant evolution.

Our analysis of enzymes involved in GA and BR biosynthesis strongly supports the view that the biosynthetic capability for these two hormones is correlated with the transition of plants to land but that non-spermatophyte plants do not contain the full, or conserved, pathways for synthesis and inactivation of bioactive forms of the hormones as recognised in angiosperms (Figure 3.4; Table 3.1). Moreover, our analysis suggests that, while the complete angiosperm-like GA biosynthetic pathway is found in all vascular plants, the BR biosynthetic pathway shows a stepwise increase in complexity with the evolution of CYP90A1 in lycophytes (Table 3.1) and CYP90B1 in ferns (Figure 3.4; Table 3.1) before the completion of the pathway in the gymnosperms (Figure 3.4; Table 3.1). The latter is a surprising observation because BRs have been found across plant taxa - in angiosperms, gymnosperms, ferns, lycophytes, the moss *P. patens*, the liverwort *M. polymorpha*, and the chlorophyte *C. vulgaris* (Bajguz 2009; Yokota et al. 2017). This may indicate that BR production in algae, bryophytes, lycophytes, and ferns is carried out by enzymes that differ from those used in spermatophytes.

A surprising finding was the presence of genes in bryophytes encoding GA biosynthetic and inactivation enzymes (Figure 3.3; Table 3.1) despite the fact that recognised bioactive forms of GA have not been detected in the few bryophytes that have been analysed (Hirano et al. 2007; Miyazaki et al. 2018). A plausible interpretation of this is that bioactive GAs are present at very low concentrations in bryophytes (as is the case with BRs (Yokota et al. 2017)), and this has hindered previous attempts at their identification. Mass spectrometry instrumentation has increased substantially in sensitivity in recent years so this would be worth revisiting. A key feature for the regulatory function of phytohormones is the capacity to rapidly inactivate the bioactive forms of the molecules using specific enzymes. Overall, the angiosperms were found to have a greater capacity for various kinds of GA and BR inactivation than non-angiosperms, and homologues of genes encoding the angiosperm BR inactivation mechanisms were not found in non-seed plants (Figures 3.3 and 3.4; Table 3.1). Combined with the lack of some GA inactivation enzymes identified in non-seed plants, this indicates that the majority of known angiosperm inactivation mechanisms for both hormones are relatively recent innovations in land plant evolution. Non-seed plants must either utilise alternative inactivation processes or a different set of genes to carry out BR inactivation reactions. As indicated by the presence of a GAMT gene in *Marchantia*

polymorpha (Figure 3.3), and GA 2-oxidases acting on C20 GAs in bryophytes (Table 3.1), bryophytes may be limited to GA inactivation via these mechanisms. The lack of conserved inactivation mechanisms between taxa for both phytohormones indicates that different plant groups make use of alternative or divergent enzymes for homeostasis of GAs and BRs. The presence of GAMT homologues in charophyte algal transcriptomes indicates that these genes are more ancient in origin than previously discussed and suggests that GAMTs in land plants evolved from similar methyltransferases in algae, likely with alternative substrates, given the lack of homologues of GA biosynthetic genes in algae.

The analysis also revealed details about the likely evolution of several structural polymers. The capability for biosynthesis of precursor molecules required for the production of cutin/suberin monomers were present prior to the evolution of land plants, with the LACS gene encoding the enzyme long-chain acyl-CoA synthetase required for fatty acid activation present in the genomes of sampled land plants from all clades as well as in charophyte and chlorophyte algal genomes (Figure 3.5B). However, even though cutin is thought to be present in all vascular plants, the CYP enzymes involved in angiosperm cutin and suberin biosynthesis appear to have evolved in stages, with only CYP86A present in all land plant taxa (Table 3.1), followed by HOTHEAD in the ferns and gymnosperms and CYP77A in the angiosperms (Figure 3.5; Table 3.1). For sporopollenin biosynthesis, homologues of the ACOS5 gene responsible for fatty acid activation were not found in algae but were present in all land plants, while the ACH enzymes, which remove CoA, were found in charophyte algae and land plants (Figure 3.6B). CYP enzymes involved in sporopollenin biosynthesis are conserved only in land plants (Figure 3.6; Table 3.1), but sporopollenin has been identified in chlorophyte algae - suggesting that alternative CYP enzymes (likely members of the CYP86 clan) are responsible for sporopollenin biosynthesis in these species.

Taken together, the ubiquitous presence in land plants and algae of genes encoding enzymes acting early in the biosynthetic pathways for suberin, cutin, and sporopollenin suggests that none of these biopolymer biosynthetic pathways developed *de novo* in the land plants. It may be that certain enzymes involved in cutin and suberin biosynthesis evolved directly from those involved in sporopollenin biosynthesis; a common evolutionary origin of all biopolymers has previously been hypothesised (Fich et al. 2016; Niklas et al. 2017).

However, homologues of the angiosperm enzymes responsible for intermediate and end-point reactions are often absent from earlier-diverging land plant genomes (Figures 3.5B and 3.6B), even when these plants are known to contain such biopolymers. There are several possible explanations for this. It may be that different mechanisms, or at least alternative enzymes (for example differing in substrate specificity or activity), are utilised for biopolymer production in earlier-diverging land plants. The enzyme families in question tend to be expanded in the angiosperm species in this analysis; this diversity may be linked to subfunctionalisation conferring increases in substrate specificity or catalytic efficiency in later-diverging land plants. It is plausible that these differences result in alternative monomer composition of biopolymers in divergent plant taxa, a phenomenon that has already been observed between some species (Caldicott and Eglinton 1976; Fernández et al. 2016). Another possibility is that the same reactions, leading to the same end products, are catalysed by enzymes with common ancestry but that exhibit sequence divergence between earlier- and later-diverging land plants, and as such are not identified as homologous. Finally, the presence of similar biopolymers but absence of homologous genes across divergent plant taxa could be the result of convergent evolution, which is a common theme in land plant evolution (Huang et al. 2016; Sage et al. 2011) and has already been identified in the biosynthesis of the biopolymer lignin, where lycophytes and spermatophytes have independently developed the ability to produce monomers derived from sinapyl alcohol (Weng et al. 2010).

3.8 LIMITATIONS OF THE APPROACH

The data presented here demonstrate the power of a systematic comparative analysis of large genome and transcriptome datasets to reveal new understanding about the evolution of biological pathways and processes. Nevertheless, there are some limitation to the approach. These stem mainly from the reliance on sequence homology as a tool for annotating genes with metabolic functions. First, in the case that two organisms do not have homologous genes for specific enzymes in common it is difficult to speculate as to how this reflects their biochemistry. For example, although cutin is biosynthesised in all land plants, angiosperm and non-angiosperm species do not share the specific cytochrome P450 enzymes used by higher plants (Figure 3.5). Without experimental analysis, it is impossible to determine

whether this is indicative of a difference in gene function between these species (and non-angiosperm plants produce alternative cutin monomers) or whether the lack of homology is caused by superficial divergence at the sequence level (and enzyme function may in fact be conserved across land plants). Conversely, while sequence-based homology can identify candidates for functionally analogous genes across species, experimental analysis is also required to confirm these predictions. This may be particularly true for the cytochrome P450 genes, which exist in large, highly interrelated gene families that have been subject to rapid expansion and subfunctionalisation over the course of plant evolution (Chae et al. 2014; Nelson and Werck-Reichhart 2011).

Second, plants exhibit a huge range of specialised metabolism, and it is reasonable to expect that non-angiosperm plants, which are relatively understudied, have the capacity for metabolic functions that are thus far undiscovered. The use of a *de novo* metabolic annotation tool trained across plant and non-plant organisms allows for the detection of metabolic reactions in non-angiosperms where homologues may be present in microbes or fungi rather than angiosperms; however, this method does not have the power to detect previously unstudied metabolic pathways. Increasing knowledge of non-angiosperm plant metabolism will require extensive experimental and computational investigation. Metabolic profiling over different conditions and growth stages can be used in combination with flux analysis to elucidate the compounds produced by individual species and their role in the wider metabolic network. Building up such a knowledge base of metabolism in non-angiosperm plants is the only way to improve the accuracy of metabolic pathway inference and would have additional far-reaching benefits, for example, in improving the prospects for using non-angiosperm plants as experimentally tractable test beds for metabolic engineering strategies.

Finally, gene homology inferences are dependent on a set of high-quality genomes as their foundation. Plant genomes tend to be larger and more complex than animal genomes, and the high incidence of repetitive elements affects the quality of genome assembly and annotation (Claros et al. 2012). Incorrect construction of gene models may lead to the inference of gaps in gene presence in species where there are none. This is particularly a problem for gymnosperm genomes, which can be up to 30Gb and are rich in repetitive DNA, mostly transposable elements (De La Torre et al. 2014). It may be that these issues with gymnosperm

genome assembly are the cause of some missing homologues in individual genomes in this study - for example, in Figure 3.6 several gymnosperms are missing genes that are identified in all other species analysed. We have attempted to mitigate this problem in this study by maximising our sample size, making use of the majority of genome sequences available for non-angiosperm species and establishing further gene presences using transcriptome analysis of an additional 305 species. Furthermore, our findings are limited to identifying evolutionary trends across whole plant taxa, as opposed to basing conclusions on the presence or absence of genes in individual species.

REFERENCES

- Bajguz, A. (2009). Isolation and characterization of brassinosteroids from algal cultures of *Chlorella vulgaris* Beijerinck (Trebouxiophyceae). *Journal of Plant Physiology*, 166, 1946–1949.
- Banks, J. A., Nishiyama, T., Hasebe, M., ... Grigoriev, I. V. (2011). The *Selaginella* genome identifies genetic changes associated with the evolution of vascular plants. *Science*, 332, 960–963.
- Basler, G., Fernie, A. R., & Nikoloski, Z. (2018). Advances in metabolic flux analysis toward genome-scale profiling of higher organisms. *Bioscience Reports*, 38(6), BSR20170224.
- Bateman, A. (2019). Uniprot: A worldwide hub of protein knowledge. *Nucleic Acids Research*, 47, D506–D515.
- Benson, D. A., Cavanaugh, M. A., Clark, K., ... Sayers, E. W. (2017). GenBank. *Nucleic Acids Research*, 45, D37–D42.
- Berens, M. L., Berry, H. M., Mine, A., ... Tsuda, K. (2017). Evolution of hormone signaling networks in plant defense. *Annual review of phytopathology*, 55, 401–425.
- Bowman, J. L., Kohchi, T., Yamato, K. T., ... Schmutz, J. (2017). Insights into land plant evolution garnered from the *Marchantia polymorpha* genome. *Cell*, 171, 287–304.e15.
- Buchfink, B., Xie, C., & Huson, D. H. (2014). Fast and sensitive protein alignment using DIAMOND. *Nature Methods*, 12, 59–60.
- Caldicott, A. B., & Eglinton, G. (1976). Cutin acids from bryophytes: An ω -1 hydroxy alkanolic acid in two liverwort species. *Phytochemistry*, 15, 1139–1143.
- Cannell, N., Emms, D. M., Hetherington, A. J., ... Sweetlove, L. J. (2020). Multiple metabolic innovations and losses are associated with major transitions in land plant evolution. *Current Biology*, 30, 1783–1800.e11.
- Capella-Gutiérrez, S., Silla-Martínez, J. M., & Gabaldón, T. (2009). TrimAl: A tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics*, 25, 1972–1973.
- Caspi, R., R. and Billington, Fulcher, C. A., Keseler, I. M., ... Karp, P. D. (2018). The MetaCyc database of metabolic pathways and enzymes. *Nucleic acids research*, 46, D633–D639.
- Chae, L., Kim, T., Nilo-Poyanco, R., & Rhee, S. Y. (2014). Genomic signatures of specialized metabolism in plants. *Science*, 344, 510–513.
- Cheng, S., Xian, W., Fu, Y., ... Melkonian, G. K. (2019). Genomes of subaerial Zygnematophyceae provide insights into land plant evolution. *Cell*, 179, 1057–1067.e14.
- Cheon, J., Fujioka, S., Dilkes, B. P., & Choe, S. (2013). Brassinosteroids regulate plant growth through distinct signaling pathways in *Selaginella* and *Arabidopsis*. *PLoS ONE*, 8.
- Claros, M. G., Bautista, R., Guerrero-Fernández, D., ... Fernández-Pozo, N. (2012). Why assembling plant genome sequences is so challenging. *Biology*, 1, 439–459.

- Clausen, M., Kannangara, R. M., Olsen, C. E., ... Møller, B. L. (2015). The bifurcation of the cyanogenic glucoside and glucosinolate biosynthetic pathways. *The Plant journal*, 84, 558–73.
- Davies, K. M., Albert, N. W., Zhou, Y., & Schwinn, K. E. (2018). Functions of flavonoid and betalain pigments in abiotic stress tolerance in plants. *Annual plant reviews* (pp. 1–41).
- De La Torre, A. R., Birol, I., Bousquet, J., ... Bohlmann, J. (2014). Insights into conifer giga-genomes. *Plant Physiology*, 166, 1724–1732.
- Delwiche, C. F., & Cooper, E. D. (2015). The evolutionary origin of a terrestrial flora. *Current Biology*, 25, R899–R910.
- Delwiche, C. F., Graham, L. E., & Thomson, N. (1989). Lignin-like compounds and sporopollenin in *Coleochaete*, an algal model for land plant ancestry. *Science*, 245, 399–401.
- de Vries, J., Stanton, A., Archibald, J. M., & Gould, S. B. (2016). Streptophyte terrestrialization in light of plastid evolution. *Trends in Plant Science*, 21, 467–476.
- Domínguez, E., Heredia-Guerrero, J. A., & Heredia, A. (2015). Plant cutin genesis: Unanswered questions. *Trends in Plant Science*, 20, 551–8.
- Emms, D. M., & Kelly, S. (2019). Orthofinder: Phylogenetic orthology inference for comparative genomics. *Genome Biology*, 20, 238.
- Fang, C., Fernie, A. R., & Luo, J. (2019). Exploring the diversity of plant metabolism. *Trends in Plant Science*, 24, 83–98.
- Fernández, V., Guzmán-Delgado, P., Graça, J., ... Gil, L. (2016). Cuticle structure in relation to chemical composition: Re-assessing the prevailing model. *Frontiers in Plant Science*, 7.
- Fich, E. A., Segerson, N. A., & Rose, J. K. C. (2016). The plant polyester cutin: Biosynthesis, structure, and biological roles. *Annual Review of Plant Biology*, 67, 207–233.
- Gomes de Oliveira Dal'Molin, C., & Nielsen, L. K. (2018). Plant genome-scale reconstruction: From single cell to multi-tissue modelling and omics analyses. *Current Opinion in Biotechnology*, 49, 42–48.
- Goodstein, D. M., Shu, S., Howson, R., ... Rokhsar, D. S. (2012). Phytozome: A comparative platform for green plant genomics. *Nucleic acids research*, 40, D1178–86.
- Guan, R., Zhao, Y., Zhang, H., ... Chen, W. (2016). Draft genome of the living fossil *Ginkgo biloba*. *GigaScience*, 5, 49.
- Halkier, B. A., & Gershenzon, J. (2006). Biology and biochemistry of glucosinolates. *Annual Review of Plant Biology*, 57, 303–333.
- Hayashi, K. I., Horie, K., Hiwatashi, Y., ... Nozaki, H. (2010). Endogenous diterpenes derived from *ent*-kaurene, a common gibberellin precursor, regulate protonema differentiation of the moss *Physcomitrella patens*. *Plant Physiology*, 153, 1085–1097.
- He, J., Chen, Q., Xin, P., ... Wang, G. (2019). CYP72A enzymes catalyse 13-hydrolyzation of gibberellins. *Nature Plants*, 5, 1057–1065.
- He, X., Dai, J., & Wu, Q. (2016). Identification of sporopollenin as the outer layer of cell wall in microalga *Chlorella protothecoides*. *Frontiers in Microbiology*, 7, 1047.
- Hirano, K., Nakajima, M., Asano, K., ... Matsuoka, M. (2007). The GID1-mediated gibberellin perception mechanism is conserved in the lycophyte *Selaginella moellendorffii* but not in the bryophyte *Physcomitrella patens*. *The Plant Cell*, 19, 3058–3079.
- Hori, K., Maruyama, F., Fujisawa, T., ... Ohta, H. (2014). *Klebsormidium flaccidum* genome reveals primary factors for plant terrestrial adaptation. *Nature Communications*, 5.
- Huang, R., O'Donnell, A. J., Barboline, J. J., & Barkman, T. J. (2016). Convergent evolution of caffeine in plants by co-option of exapted ancestral enzymes. *Proceedings of the National Academy of Sciences of the United States of America*, 113, 10613–10618.
- Huang, Y., Wang, X., Ge, S., & Rao, G. (2015). Divergence and adaptive evolution of the gibberellin oxidase genes in plants. *BMC Evolutionary Biology*, 15, 207.

- Jeske, L., Placzek, S., Schomburg, I., ... Schomburg, D. (2019). BRENDA in 2019: A European ELIXIR core data resource. *Nucleic Acids Research*, 47, D542–D549.
- Karp, P. D., Latendresse, M., Paley, S. M., ... Caspi, R. (2016). Pathway Tools version 19.0 update: Software for pathway/genome informatics and systems biology. *Briefings in bioinformatics*, 17, 877–90.
- Katoh, K., & Standley, D. M. (2013). Mafft Multiple Sequence Alignment Software Version 7: Improvements in performance and usability. *Molecular Biology and Evolution*, 30, 772–780.
- Kondo, S., Hori, K., Sasaki-Sekimoto, Y., ... Ohta, H. (2016). Primitive extracellular lipid components on the surface of the charophytic alga *Klebsormidium flaccidum* and their possible biosynthetic pathways as deduced from the genome sequence. *Frontiers in Plant Science*, 7, 952.
- Kumar, S., Kempinski, C., Zhuang, X., ... Chappell, J. (2016). Molecular diversity of terpene synthases in the liverwort *Marchantia polymorpha*. *The Plant Cell*, 28, 2632–2650.
- Leebens-Mack, J. H., Barker, M. S., Carpenter, E. J., ... Wong, G. K. S. (2019). One thousand plant transcriptomes and the phylogenomics of green plants. *Nature*, 574, 679–685.
- Li, F. W., Brouwer, P., Carretero-Paulet, L., ... Pryer, K. M. (2018). Fern genomes elucidate land plant evolution and cyanobacterial symbioses. *Nature Plants*, 4, 460–472.
- Magome, H., Nomura, T., Hanada, A., ... Yamaguchi, S. (2013). CYP714B1 and CYP714B2 encode gibberellin 13-oxidases that reduce gibberellin activity in rice. *Proceedings of the National Academy of Sciences of the United States of America*, 110, 1947–52.
- Marsolais, F., Boyd, J., Paredes, Y., ... Varin, L. (2007). Molecular and biochemical characterization of two brassinosteroid sulfotransferases from *Arabidopsis*, AtST4a (At2g14920) and AtST1 (At2g03760). *Planta*, 225, 1233–44.
- Miyazaki, S., Hara, M., Ito, S., ... Nakajima, M. (2018). An ancestral gibberellin in a moss *Physcomitrella patens*. *Molecular Plant*, 11, 1097–1100.
- Moreira, T. B., Lima, J. M., Coca, G. C., & Williams, T. C. R. (2019). Insights into the spatial and temporal organisation of plant metabolism from network flux analysis. *Theoretical and Experimental Plant Physiology*, 31, 215–226.
- Morris, J. L., Puttick, M. N., Clark, J. W., ... Donoghue, P. C. J. (2018). The timescale of early land plant evolution. *Proceedings of the National Academy of Sciences of the United States of America*, 115, E2274–E2283.
- Mueller, L. A., Zhang, P., & Rhee, S. Y. (2003). Aracyc: A biochemical pathway database for *Arabidopsis*. *Plant Physiology*, 132, 453–460.
- Neale, D. B., McGuire, P. E., Wheeler, N. C., ... Wegrzyn, J. L. (2017). The Douglas-Fir genome sequence reveals specialization of the photosynthetic apparatus in Pinaceae. *G3: Genes, Genomes, Genetics*, 7, 3157–3167.
- Neale, D. B., Wegrzyn, J. L., Stevens, K. A., ... Langley, C. H. (2014). Decoding the massive genome of loblolly pine using haploid DNA and novel assembly strategies. *Genome Biology*, 15(R59).
- Nelson, D., & Werck-Reichhart, D. (2011). A P450-centric view of plant evolution. *The Plant Journal*, 66, 194–211.
- Nguyen, L., Schmidt, H. A., von Haeseler, A., & Minh, B. Q. (2015). IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Molecular Biology and Evolution*, 32, 268–274.
- Niklas, K. J., Cobb, E. D., & Matas, A. J. (2017). The evolution of hydrophobic cell wall biopolymers: From algae to angiosperms. *Journal of Experimental Botany*, 68, 5261–5269.
- Nishiyama, T., Sakayama, H., de Vries, J., ... Rensing, S. A. (2018). The *Chara* genome: Secondary complexity and implications for plant terrestrialization. *Cell*, 174, 448–464.e24.

- Noguchi, S., C. and Miyazaki, Kawaide, H., Gotoh, O., ... Aoyama, Y. (2018). Characterization of moss *ent*-kaurene oxidase (CYP701B1) using a highly purified preparation. *Journal of biochemistry*, 163, 69–76.
- Novoselov, S. V., Rao, M., Onoshko, N. V., ... Gladyshev, V. N. (2002). Selenoproteins and selenocysteine insertion system in the model plant cell system, *Chlamydomonas reinhardtii*. *The EMBO journal*, 21, 3681–93.
- Nystedt, B., Street, N. R., Wetterbom, A., ... Jansson, S. (2013). The Norway spruce genome sequence and conifer genome evolution. *Nature*, 497, 579–584.
- Ohnishi, T., Szatmari, A. M., Watanabe, B., ... Mizutani, M. (2006). C-23 hydroxylation by *Arabidopsis* CYP90C1 and CYP90D1 reveals a novel shortcut in brassinosteroid biosynthesis. *The Plant Cell*, 18, 3275–3288.
- Poppenberger, B., Fujioka, S., Soeno, K., ... Bowles, D. (2005). The UGT73C5 of *Arabidopsis thaliana* glucosylates brassinosteroids. *Proceedings of the National Academy of Sciences of the United States of America*, 102, 15253–8.
- Rensing, S. A. (2017). Why we need more non-seed plant models. *New Phytologist*, 216, 355–360.
- Rensing, S. A. (2018). Great moments in evolution: The conquest of land by plants. *Current Opinion in Plant Biology*, 42, 49–54.
- Rensing, S. A., Lang, D., Zimmer, A. D., ... Boore, J. L. (2008). The *Physcomitrella* genome reveals evolutionary insights into the conquest of land by plants. *Science*, 319, 64–69.
- Riekhof, W. R., Naik, S., Bertrand, H., ... Voelker, D. R. (2014). Phosphate starvation in fungi induces the replacement of phosphatidylcholine with the phosphorus-free betaine lipid diacylglycerol-N,N,N-trimethylhomoserine. *Eukaryotic cell*, 13, 749–57.
- Rieu, I., Eriksson, S., Powers, S. J., ... Phillips, A. L. (2008). Genetic analysis reveals that C19-GA 2-oxidation is a major gibberellin inactivation pathway in *Arabidopsis*. *The Plant Cell*, 20, 2420–36.
- Rodman, J., Soltis, P., Soltis, D., ... Karol, K. (1998). Parallel evolution of glucosinolate biosynthesis inferred from congruent nuclear and plastid gene phylogenies. *American journal of botany*, 85, 997.
- Rozentsvet, O. A. (2004). Comparative examination of distribution of phospholipids and a betaine lipid DGTS in tropical fern species. *Biochemical Systematics and Ecology*, 32, 303–311.
- Rozentsvet, O. A., Dembitsky, V. M., & Saksonov, S. V. (2000). Occurrence of diacylglyceroltrimethylhomoserines and major phospholipids in some plants. *Phytochemistry*, 54, 401–407.
- Sage, R. F., Christin, P., & Edwards, E. J. (2011). The C₄ plant lineages of planet earth. *Journal of Experimental Botany*, 62, 3155–69.
- Schläpfer, P., Zhang, P., Wang, C., ... Rhee, S. Y. (2017). Genome-wide prediction of metabolic enzymes, pathways, and gene clusters in plants. *Plant Physiology*, 173, 2041–2059.
- Stevens, K. A., Wegrzyn, J. L., Zimin, A., ... Langley, C. H. (2016). Sequence of the sugar pine megagenome. *Genetics*, 204, 1613–1626.
- Sundell, D., Mannapperuma, C., Netotea, S., ... Street, N. R. (2015). The Plant Genome Integrative Explorer Resource: PlantGenIE.org. *New Phytologist*, 208, 1149–1156.
- Turk, E. M., Fujioka, S., Seto, H., ... Neff, M. M. (2003). CYP72B1 inactivates brassinosteroid hormones: An intersection between photomorphogenesis and plant steroid signal transduction. *Plant physiology*, 133, 1643–53.
- Van Der Maaten, L., & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9, 2579–2605.
- Van Ghelder, C., Parent, G. J., Rigault, P., ... MacKay, J. (2019). The large repertoire of conifer nlr resistance genes includes drought responsive and highly diversified nrls. *Scientific Reports*, 9.

- Varbanova, M., Yamaguchi, S., Yang, Y., ... Picharsky, E. (2007). Methylation of gibberellins by *Arabidopsis* GAMT1 and GAMT2. *The Plant Cell*, 19, 32–45.
- Vishwanath, S. J., Delude, C., Domergue, F., & Rowland, O. (2015). Suberin: Biosynthesis, regulation, and polymer assembly of a protective extracellular barrier. *Plant Cell Reports*, 34, 573–586.
- Wan, T., Liu, Z. M., Li, L. F., ... Wang, X. M. (2018). A genome for gnetophytes and early evolution of seed plants. *Nature Plants*, 4, 82–89.
- Wang, C., Liu, Y., Li, S. S., & Han, G. Z. (2015). Insights into the origin and evolution of the plant hormone signaling machinery. *Plant Physiology*, 167, 872–886.
- Wegrzyn, J. L., Staton, M. A., Street, N. R., ... Ficklin, S. (2019). Cyberinfrastructure to improve forest health and productivity: The role of tree databases in connecting genomes, phenomes, and the environment. *Frontiers in Plant Science*, 10.
- Weng, J., Akiyama, T., Bonawitz, N. D., ... Chapple, C. (2010). Convergent evolution of syringyl lignin biosynthesis via distinct pathways in the lycophyte *Selaginella* and flowering plants. *The Plant Cell*, 22, 1033–45.
- Xu, Z., Xin, T., Bartels, D., ... Chen, S. (2018). Genome analysis of the ancient tracheophyte *Selaginella tamariscina* reveals evolutionary features relevant to the acquisition of desiccation tolerance. *Molecular Plant*, 11, 983–994.
- Yokota, T., Ohnishi, T., Shibata, K., ... Kohchi, T. (2017). Occurrence of brassinosteroids in non-flowering land plants, liverwort, moss, lycophyte and fern. *Phytochemistry*, 136, 46–55.
- Zhang, Y., Zhang, B., Yan, D., ... He, Z. (2011). Two *Arabidopsis* cytochrome P450 monooxygenases, CYP714A1 and CYP714A2, function redundantly in plant development through gibberellin deactivation. *The Plant Journal*, 67, 342–53.
- Zhu, Y., Nomura, T., Xu, Y., ... He, Z. (2006). ELONGATED UPPERMOST INTERNODE encodes a cytochrome P450 monooxygenase that epoxidizes gibberellins in a novel deactivation reaction in rice. *The Plant Cell*, 18, 442–456.
- Zi, J., Mafu, S., & Peters, R. J. (2014). To gibberellins and beyond! Surveying the evolution of (di)terpenoid metabolism. *Annual Review of Plant Biology*, 65, 259–286.

Statement of Authorship for joint/multi-authored papers for PGR thesis

To appear at the end of each thesis chapter submitted as an article/paper

The statement shall describe the candidate's and co-authors' independent research contributions in the thesis publications. For each publication there should exist a complete statement that is to be filled out and signed by the candidate and supervisor (**only required where there isn't already a statement of contribution within the paper itself**).

Title of Paper	Multiple Metabolic Innovations and Losses Are Associated with Major Transitions in Land Plant Evolution
Publication Status	☒ Published ☐ Accepted for Publication ☐ Submitted for Publication ☐ Unpublished and unsubmitted work written in a manuscript style
Publication Details	Cannell et. al., 2020. Multiple Metabolic Innovations and Losses Are Associated with Major Transitions in Land Plant Evolution. <i>Current Biology</i> 30 (10):1783-1800.e11

Student Confirmation

Student Name:	Naomi Cannell		
Contribution to the Paper	Naomi Cannell and Lee J. Sweetlove conceived the project and developed the methodology; N.C. wrote code and carried out the investigation and formal analysis; L.J.S. provided supervision; Steven Kelly and David M. Emms provided software and computing resources; John MacKay, Liam Dolan, and Alexander J. Hetherington provided transcriptome resources; N.C. wrote the manuscript and prepared the figures; L.J.S., L.D., J.M., A.J.H., and D.M.E. reviewed and edited the manuscript.		
Signature Naomi Cannell	Date	15/01/2021	

Supervisor Confirmation

By signing the Statement of Authorship, you are certifying that the candidate made a substantial contribution to the publication, and that the description described above is accurate.

Supervisor name and title: Professor Lee J. Sweetlove		
Supervisor comments This is completely Naomi's work. The co-authors provided either access to data or computer resources but all of the analysis and the development of the narrative was Naomi's. Naomi also wrote the manuscript with the other co-authors commenting on drafts.		
Signature 	Date	18/01/2020

This completed form should be included in the thesis, at the end of the relevant chapter.

4

Random forest analysis of phylogenetically limited gene sets facilitates identification of candidate genes associated with phytohormone and biopolymer biosynthesis

4.1 MOTIVATION

Secondary metabolic networks in plants are complex and highly interconnected, both stoichiometrically (by the overlapping use of intermediate metabolites between metabolic pathways), and in terms of the signalling cascades that may simultaneously affect several metabolic systems by regulating the expression of multiple metabolic genes. However, even for well-studied metabolic pathways, the full extent of these interactions with the wider plant metabolic network are not well understood. In this manuscript, we aim to address the gaps in our understanding of specific secondary metabolic pathways, using bioinformatic and computational analyses to identify candidate genes associated with the phytohormone and biopolymer biosynthetic pathways discussed in Chapter 3.

4.2 CONTRIBUTION

A signed statement of authorship can be found at the end of this chapter. Naomi Cannell, Steven Kelly and Lee J. Sweetlove conceived the project and developed the methodology;

N.C. wrote code and carried out the investigation and formal analysis; L.J.S. provided supervision; S.K. and David M. Emms provided software and computing resources; N.C. wrote the manuscript and prepared the figures; L.J.S. reviewed and edited the manuscript.

4.3 SUMMARY

Despite progress in our understanding of plant secondary metabolism, gaps exist in our knowledge of individual metabolic pathways, their involvement in plant growth and development and the extent and mechanisms of their interaction with the plant metabolic network. To address these, we developed a computational pipeline combining gene association methods based on phylogenetic profiling and gene co-expression analysis to discover candidate genes associated with specific metabolic pathways. *Arabidopsis* transcriptome data were used to produce gene expression profiles for metabolic pathways of interest, and the presence/absence of genes encoding the relevant enzymes across the Chloroplastida established a phylogenetic profile. *Arabidopsis* genes with phylogenetic profiles matching that of each metabolic pathway were extracted and used as the basis for candidate gene identification by *Arabidopsis* transcriptome data mining. First, by calculating Pearson's correlation coefficient for each transcript compared to the metabolic pathway expression profile. Second, using random forests for regression analysis followed by the extraction of important genes from the training set. This methodology was applied to biosynthetic pathways for gibberellin and brassinosteroid phytohormones and the biopolymers cutin, suberin and sporopollenin. The identified candidate genes encoded proteins linked to each pathway in various ways. These included proteins involved in metabolite precursor delivery such as PVA12, involved in the intracellular trafficking of brassinosteroid precursor sterols; enzymes participating in connected metabolic pathways such as C4H, which catalyses an early reaction in the biosynthesis of phenylpropanoids, important sporopollenin constituents; and transcription factors with regulatory roles such as WRKY43, likely involved in abscisic acid/gibberellin crosstalk during germination.

4.4 INTRODUCTION

The colonisation of land more than 470 million years ago by the common ancestor of all embryophytes was enabled by a set of metabolic adaptations which conferred protection from the novel pressures of a terrestrial environment (Morris et al. 2018; Rensing 2018). These included the biosynthesis of specialised biochemicals to counteract abiotic stressors such as ultraviolet (UV) radiation (Brunetti et al. 2019); phytohormones involved in complex signalling networks for the coordination of plant growth, development and immunity (Berens et al. 2017); and structural polymers to provide mechanical support and physical protection (Niklas et al. 2017). In extant land plants, these metabolic innovations have diversified into a complex network of metabolic pathways and signalling processes which affect almost every aspect of plant form and function (Fang et al. 2019).

Recent work by Cannell et al. (2020) has established details of the evolution of several metabolic pathways involved with phytohormone and biopolymer biosynthesis associated with the transition to land. Specifically, it was discovered that genes encoding the full gibberellin (GA) biosynthetic pathway are present in all land plants despite a lack of bioactive GA hormones identified in bryophytes. Mosses appear to be an exception, with a single enzyme in the GA biosynthesis pathway not encoded by moss genomes (Miyazaki et al. 2018; Zi et al. 2014)). Conversely it was found that the full angiosperm-like brassinosteroid (BR) biosynthetic pathway is encoded only in the genomes of spermatophytes, and shows a step-wise evolutionary process originating with a pair of involved CYP90 genes identified in ferns (Cannell et al. 2020). This is despite the fact that BR metabolites have been identified across the land plant lineage (Yokota et al. 2017). Similarly, the evolution of the angiosperm-like metabolic pathway for biosynthesis of the structural polymers cutin and suberin was shown to have evolved in stages, starting with the presence of CYP86A in the bryophytes and with enzymes for the whole metabolic pathway only encoded in angiosperms genomes (Cannell et al. 2020). This was surprising as cutin is considered to be present in earlier-diverging land plants (Chang et al. 2020; Niklas et al. 2017). Furthermore, while the vast majority of enzymes for another structural polymer, sporopollenin, were found only in land plant genomes, sporopollenin-like biopolymers have been identified in chlorophytic algae

(Delwiche et al. 1989; He et al. 2016), suggesting alternative biosynthetic processes for the production of this compound in earlier- and later- diverging species of the Chloroplastida.

Despite continuing progress in evolutionary and functional studies of secondary metabolism, there are still major gaps in our understanding of the occurrence and biosynthesis of such metabolites – and in particular, their function in different plant tissues and developmental stages, and in response to various environmental conditions and various stressors is not broadly established. For example, although the central mechanisms of GA and BR biosynthesis and signalling are fairly well established, we lack a comprehensive understanding of the molecular mechanisms of many individual phytohormone-mediated processes in plants, as well as of the crosstalk between these and other hormone signalling pathways (Shigenaga et al. 2017; Verma et al. 2016). Hence, although GAs are implicated in leaf senescence, the specific molecular interactions which mediate this process are unknown (Chen et al. 2017; Chen et al. 2014). GAs are also involved in cell proliferation in growing tissues, however their specific roles (and the functions of several GA-regulated genes in this context) are unresolved (Claeys et al. 2014). Indeed, novel GAs are still being discovered – the bioactive GA₁₂ hydrolysis product DHGA₁₂ was recently identified in *Arabidopsis* and is involved in modulating GA and abscisic acid (ABA) balance in developing plants (Liu et al. 2019). Similarly, BRs are implicated in a range of processes for which detail is lacking at the molecular level. For example, BRs are important mediators of plant/micro-organism interactions, both as an immune response to pathogens but also in beneficial interactions such as mycorrhizal symbioses (Wang et al. 2012; Wei and Li 2016). In the latter context BRs are thought to promote arbuscular mycorrhizal interactions by affecting nodule formation, however the process by which this occurs has not yet been established (McGuinness et al. 2019). Additionally, BR signalling regulates plant fertility by modulating the expression levels of various pollen-expressed transcription factors – however the specific mechanism of interaction is unknown (Yan et al. 2020).

By contrast, structural polymers do not hold regulatory or signalling roles but important questions about their formation and function remain unresolved. Questions remain about the molecular composition of cutin, suberin and sporopollenin and it is likely that this composition may differ between plant species. This knowledge gap limits the investigation

of the full biosynthetic and regulatory processes involved in the production of structural polymers in plants (Bhunja et al. 2018). For example, sporopollenin is considered the toughest naturally occurring polymer in existence, and although a basic biosynthetic pathway is established we are only recently beginning to achieve a detailed understanding of all its molecular constituents and chemical characteristics (Li et al. 2019; Xue et al. 2020). While the composition of cutin is more established, it is still unknown how cutin monomers are exported from their proposed site of synthesis, the endoplasmic reticulum (ER), across the plasma membrane and cell wall for polymerisation and deposition – in fact the full polymerisation process itself is still a mystery (Fich et al. 2016). Although recent progress has been made in understanding the regulation of suberin biosynthesis and deposition (Cohen et al. 2020; Ursache et al. 2020), there are similarly still outstanding questions about the biosynthesis, polymerisation and degradation processes for this polymer.

Experimental elucidation of metabolic systems is challenging, involving time-consuming methods such as reverse genetics, metabolic flux analysis and the purification of enzymes for analysis of kinetic and regulatory properties (Basler et al. 2018; Thorneycroft 2001). However, the increased availability of genomic and transcriptomic data across land plants facilitates *in silico* identification of functional and regulatory relationships between genes as a method of identifying candidates for involvement in outstanding metabolic processes described above. One commonly used method is co-expression analysis, which is based on the assumption that genes involved in similar processes exhibit correlated expression profiles and has successfully been used to establish details of various metabolic processes in plants (Gachon et al. 2005; Han et al. 2012; Kim and Buell 2015; Persson et al. 2005). An alternative is phylogenetic profiling, which is based on evidence that related genes are likely to have co-evolved and therefore share a phylogenetic profile when orthologues are compared across species (Ranea et al. 2007; Škunca and Dessimoz 2015).

These methods, based on the comparison of gene expression levels and presence/absence of homologous genes across species, have independently been used to predict the functions of previously uncharacterised genes in various species (Ruprecht et al. 2011; Ternes et al. 2006). Here, we integrate these two guilt-by-association approaches, using transcriptome data sampled from the root, floral bud and aerial seedling for 17 *Arabidopsis* accessions (Lei

et al. 2017) with gene homology information across 72 land plants and algae (Cannell et al. 2020; Emms and Kelly 2019). The integration reduces the search-space for relationships between genes and therefore increases confidence in identified gene candidates. We apply this strategy to identify genes, either enzymatic or regulatory, that are associated with biosynthetic pathways for the GA and BR phytohormones as well as the structural polymers cutin, suberin and sporopollenin.

4.5 METHODS

4.5.1 Establishing phylogenetic and gene expression profiles for each metabolic pathway

For each metabolic pathway of interest, the earliest stage of land plant evolution at which the known pathway was present in full was identified from Cannell et al. (2020). This identification was made based on gene homology information produced using OrthoFinder software (Emms and Kelly 2019), which infers orthogroups containing genes identified as homologous between species. OrthoFinder was run using 70 genome sequences and two transcriptomes for a total of 72 species (comprising 11 chlorophyte and charophyte algae, three bryophytes, three lycophytes, two ferns, seven gymnosperms and 46 angiosperms) (Cannell et al. 2020). From the resulting orthogroups it was possible to identify presence/absence patterns across the 72 species for all genes encoding a metabolic pathway, and these patterns defined the phylogenetic profile for that metabolic pathway (i.e. if a metabolic pathway is only identified in full in all angiosperms, the phylogenetic profile is defined by presence in angiosperms and absence in other plants and algae). Based on the phylogenetic profile of the metabolic pathway in question, the same orthogroups were used to identify a list of all other *Arabidopsis* genes which show the same presence/absence pattern as the full metabolic pathway (i.e. they had matching phylogenetic profiles). This subset of *Arabidopsis* genes was used as the starting point for the identification of candidate genes related to each metabolic pathway.

A gene expression profile was produced for each metabolic pathway using *Arabidopsis* transcriptome dataset taken from Lei et al. (2017) (this dataset was also used as the basis for all subsequent analyses). The transcriptome dataset covers 17 *Arabidopsis* accessions sampled from the floral bud, root and aerial seedling for a total of 51 transcript abundance measurements for each *Arabidopsis* gene. The seedlings were grown under long-day conditions

at 20°C with the exception of a six-week vernalisation period during which they were subject to short days at 4°C (Lei et al. 2017). We calculated a 51-datapoint gene expression profile for each full metabolic pathway by taking the mean logged and normalised transcript abundance of genes encoding all of the biosynthetic enzymes involved in a metabolic pathway of interest for each of the 51 *Arabidopsis* transcriptome datasets.

4.5.2 Identifying candidate genes using Pearson's correlation coefficient

The first method of candidate gene identification relied on identifying genes whose transcript abundances across the 51 *Arabidopsis* samples were correlated with each metabolic pathway's gene expression profile to the greatest degree. To identify a threshold above which correlations were to be considered significant, a random sample of 1000 genes was taken from the full *Arabidopsis* transcriptome dataset and Pearson's correlation coefficient was calculated for each randomly-selected gene against each pathway's gene expression profile. The distribution of the resulting 1000 correlation coefficients for each metabolic pathway was used to find the value of correlation coefficient for the metabolic pathway in question which would be higher than that for 99% of all randomly selected genes.

Following identification of the 99th percentile correlation coefficient threshold, the full *Arabidopsis* transcriptome dataset was subsetting to include only the set of genes matching the phylogenetic profile of each metabolic pathway. Pearson's correlation coefficients were then calculated against the mean pathway expression profile for each gene contained in this set and those with positive Pearson's correlation coefficients higher than 99% of randomly chosen genes were selected for further investigation (up to a maximum of two genes, prioritising those with the highest correlation coefficients).

4.5.3 Identifying candidate genes using random forests

Our second method of identifying candidate genes used random forests trained on *Arabidopsis* transcriptome data to identify the genes whose transcript abundances have the most power to predict each metabolic pathway's expression profile. The random forests algorithm (Breiman 2001) is a tree-based, non-parametric ensemble learning technique which has increasingly been applied to genetic data (Exposito-Alonso et al. 2019; Mao et al. 2014; Stephan et al. 2015). We chose to use this method due to its ability to capture complex,

non-linear relationships within datasets, while allowing easy identification of the genes in the training set with maximum predictive power.

Again, for each metabolic pathway the transcript abundances of a subset of *Arabidopsis* genes matching the pathway's phylogenetic profile was identified. A random forest regressor class from the scikit-learn module in Python (Pedregosa et al. 2011) was instantiated and the transcript abundances of the phylogenetically subset genes were used as training data to grow a forest of 1000 trees, using one third of the training genes to split each tree at each node. The gene expression profile of the metabolic pathway in question formed the response variable which the random forest was trained to predict. After training the random forest, two measures of variable importance were calculated for each gene (an out-of-bag permutation measure and a drop-column measure, both calculated using Python module `rfpimp` (*parrr/random-forest-importances* 2020)). The genes were ranked by each individual measure of importance before taking an average rank for each gene, calculated using the mean importance ranking across both importance measures. Genes with the highest average rank across both measures were considered to have the most power to predict the response variable. The genes with the highest four average ranks in each metabolic pathway analysis were selected for further investigation.

4.6 RESULTS

4.6.1 Bioinformatic approach

To identify candidate genes associated with the metabolic pathways for biosynthesis of GAs, BRs, cutin, suberin and sporopollenin we developed the bioinformatic pipeline described in Figure 4.1. For each metabolic pathway, a phylogenetic profile was defined by the set of species in which the full metabolic pathway – i.e. genes encoding at least one enzyme for each reaction in the metabolic pathway – was present (diagrams representing each metabolic pathway are presented in Figure 4.2). The phylogenetic profile for each metabolic pathway was identified using gene homology data across 72 species of land plants and algae taken from (Cannell et al. 2020) and is summarised in Figure 4.3. Additionally, using a 51-sample transcriptome dataset comprising 17 *Arabidopsis* accessions sampled from the root, floral bud and aerial seedling (Lei et al. 2017), a 51-datapoint gene expression profile for each metabolic

pathway was produced, with each datapoint consisting of the mean logged, normalised abundance of the transcripts encoding the metabolic pathway enzymes.

This same transcriptome data formed the basis for identification of candidate associated genes. For each metabolic pathway, transcript abundances of *Arabidopsis* genes for which homologous genes across land plants and algae showed a matching phylogenetic profile were extracted from the transcriptome dataset. These subset transcript abundances were then used for two methods of candidate gene identification. First, Pearson's correlation coefficients were calculated for each subset gene's transcript abundances against the target gene expression profile identified for the metabolic pathway in question. The genes were ranked in order of their correlation coefficients, with the highest positive correlations considered to indicate the genes most likely to be related to the metabolic pathway in question (as long as the correlation coefficient was also higher than that for 99% of randomly selected genes). Second, the genes were used to train a random forest regressor with the metabolic pathway's target gene expression profile as the response variable. Two measures of variable importance (an out-of-bag permutation measure and a drop-column measure) were calculated for each gene in the training set. The average rank of each gene's importance from each measure was calculated and genes with the highest-ranked average importance were considered the most likely to be related to the metabolic pathway in question.

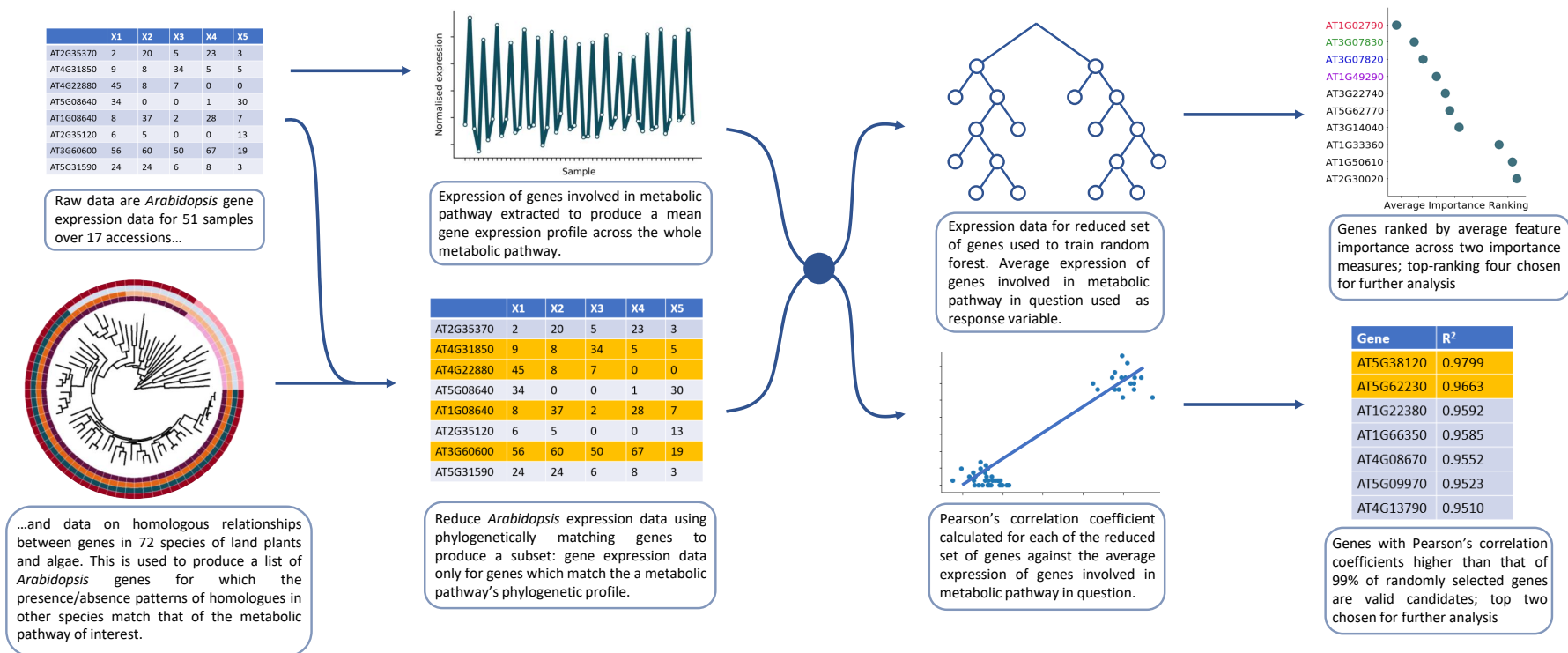


Figure 4.1: Flowchart representing the bioinformatic pipeline developed to identify candidate genes functionally associated with individual metabolic pathways. A phylogenetic profile was established for each metabolic pathway based on gene presence/absence information for each involved gene. Separately, a gene expression profile was calculated for each metabolic pathway by averaging the individual expression levels of involved genes. Using homology data across 72 land plants and algae, sets of genes whose phylogenetic distribution matched that of the metabolic pathway in question were extracted from a broader gene expression dataset. This subset was used to train a random forest regressor class as well as for the calculation of Pearson's correlation coefficients between each gene and the metabolic pathway's expression profile. The top four most predictive genes taken from ranking random forest gene importance measures and the genes with the top two Pearson's correlation coefficients were extracted for further analysis.

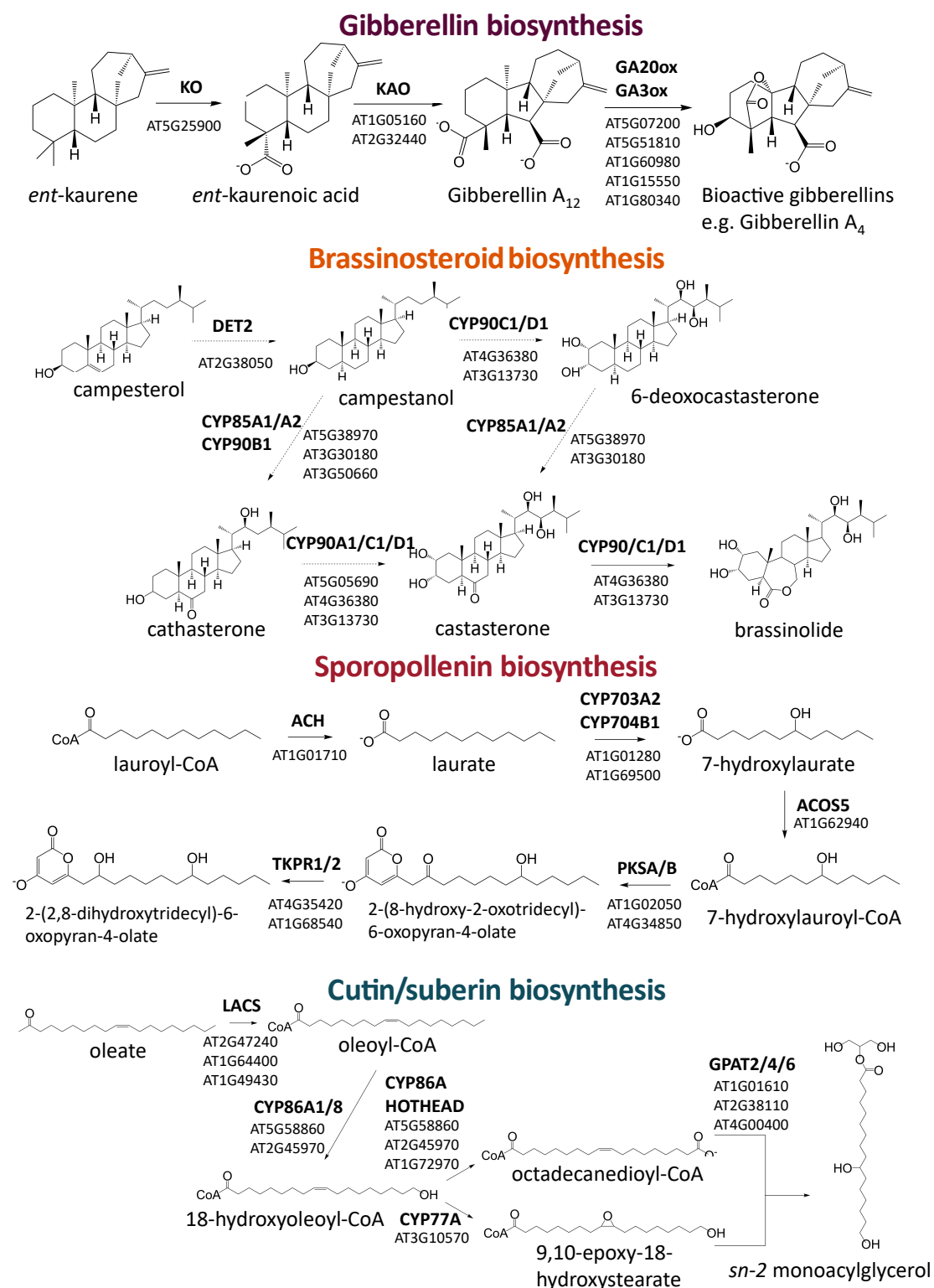


Figure 4.2: Simplified metabolic pathway diagrams representing the biosynthetic pathways for gibberellins, brassinosteroids, cutin, suberin and sporopollenin. For each metabolic pathway, metabolite names are shown below the corresponding structures. For each metabolic reaction, the gene names of the encoding enzymes are stated in bold with the relevant *Arabidopsis* gene IDs listed below. KO, ent-kaurene oxidase; KAO, ent-kaurenoic acid oxidase; GA20ox, gibberellin 20-oxidase; GA3ox, gibberellin 3-oxidase; DET2, 5 α steroid dehydrogenase; CYP, cytochrome P450; ACH, acyl-CoA thioesterase; ACOS5, fatty-acyl CoA synthetase; TKPR, tetraketide pyrone reductase; PKS, polyketide synthase; LACS, long-chain acyl-CoA synthetase; GPAT, glycerol-3-phosphate acyltransferase

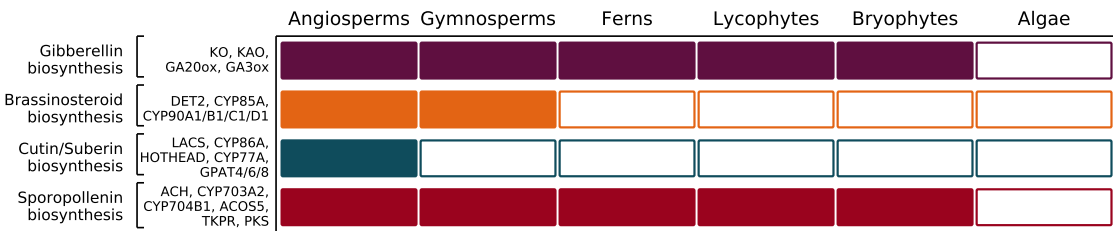


Figure 4.3: Phylogenetic profiles for each metabolic pathway. Solid boxes indicate species where the metabolic pathway is considered present in full and empty boxes indicate absence. Gene names to the left of each row indicate the genes used to create the gene expression profile for each metabolic pathway. Abbreviations as for Figure 4.2.

Table 4.1 shows a summary of the number of *Arabidopsis* genes in each subset (i.e. the number with phylogenetic profiles matching that of the metabolic pathway in question), as well as the candidate genes produced for each metabolic pathway which had either Pearson’s correlation coefficients higher than those of 99% of randomly selected genes, or ranked in the top ten average importance values from our random forest analysis. These genes were considered the most likely to be associated with each metabolic pathway. For brevity and clarity, the number of candidate genes chosen for further investigation from each Pearson’s analysis was limited to the top two genes within the significant set. The top-ranked four genes were chosen from each random forest analysis for further investigation.

Metabolic pathway	Orthogroups matching phylogenetic profile	<i>Arabidopsis</i> genes in matching orthogroups	Candidate genes – Pearson's correlation analysis		Candidate genes – random forest analysis	
			Gene ID	Gene Name	Gene ID	Gene Name
Gibberellin biosynthesis	83	584	AT2G42660	unk homeodomain	AT2G46130	WRKY43
			AT2G35350	PLL1	AT5G17800	MYB56
			AT1G30270	CIPK23	AT4G24130	DUF538
			AT1G51420	SPP1	AT5G17430	PLT4
					AT3G03820	SAUR29
					AT4G34790	SAUR3
					AT1G14390	LRRK family
					AT2G46920	POL
					AT1G46264	HSFB4
					AT4G38850	SAUR15
Brassinosteroid biosynthesis	32	122	AT3G60600	VAP27-1	AT5G41590	LURP1-like protein
			AT2G45140	PVA12	AT4G22880	ANS
			AT2G23300	LRRK family	AT1G10585	AIB1
					AT4G31850	PGR3
					AT4G38400	EXPL2
					AT1G73530	ORRM6
					AT1G23420	INO
					AT5G63590	FLS3
					AT1G67710	ARR11
					AT3G46620	DUF1117

Metabolic pathway	Orthogroups matching phylogenetic profile	Arabidopsis genes in matching orthogroups	Candidate genes – Pearson's correlation analysis		Candidate genes – random forest analysis	
			Gene ID	Gene Name	Gene ID	Gene Name
Cutin/suberin biosynthesis	43	145	AT3G14040	Pectin lyase	AT1G02790	PGA4
			AT1G02790	PGA4	AT3G07830	PGA3
			AT3G07850	Pectin lyase	AT3G07820	Pectin lyase
			AT3G07820	Pectin lyase	AT1G49290	PEG2
			AT3G07830	PGA3	AT3G22740	HMT3
			AT5G48140	Pectin lyase	AT5G62770	DUF1645
			AT4G32510	HCO3 transporter	AT3G14040	Pectin lyase
			AT3G07840	Pectin lyase	AT1G33360	CLPX3
			AT3G51410	DUF241	AT1G50610	PRK5
					AT2G30020	AP2C1
Sporopollenin biosynthesis	83	574	AT1G06520	GPAT1	AT2G39750	methyltransferase
			AT3G50780	BTB/POZ	AT4G24260	KOR3
			AT5G45810	CIPK19	AT3G01311	Actin cross-linking protein
			AT2G40790	CXXS2	AT2G30490	C4H
			AT2G38110	GPAT6	AT2G34355	Major facilitator family
			AT3G11430	GPAT5	AT1G02110	DUF630/632
			AT2G47160	BOR1	AT3G03820	SAUR29
					AT2G13820	LPTG11
					AT3G24140	FMA
					AT1G47480	α/β hydrolases family

Table 4.1: The top candidate genes identified as associated with each metabolic pathway, using Pearson's correlation coefficient and random forests. For genes identified via Pearson's correlation analysis, all genes with correlation coefficients higher than that of 99% of randomly selected genes are listed; the top two are later discussed in detail (highlighted in bold). For genes identified using random forests, the top ten most important genes are listed; the top four of these are later discussed in detail (highlighted in bold).

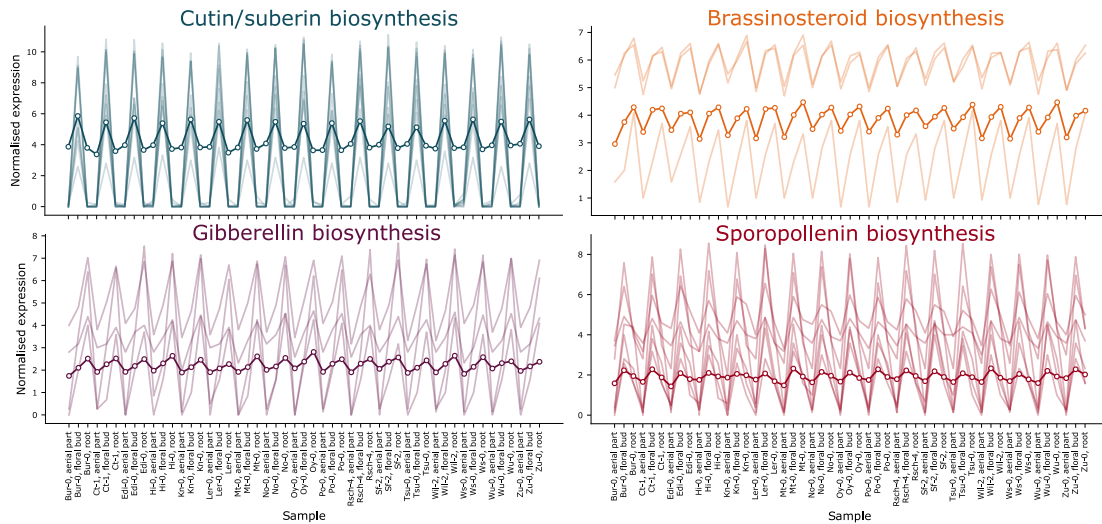


Figure 4.4: Transcript abundances of genes with Pearson’s correlation coefficients higher than that for 99% of randomly selected genes for each metabolic pathway analysis. For each metabolic pathway, the average expression profile of the metabolic pathway is shown in bold. Finer lines indicate the transcript abundances of individual highly correlated genes. The minimum Pearson’s correlation coefficient required for a gene to be included is 0.954 for cutin/suberin biosynthesis, 0.918 for brassinosteroid biosynthesis, 0.902 for gibberellin biosynthesis and 0.862 for sporopollenin biosynthesis.

Figure 4.4 presents a comparison between the average gene expression profile of each metabolic pathway and the transcript abundances of each *Arabidopsis* gene identified as correlating more strongly with the gene expression profile than 99% of randomly selected genes. Strong correlations for all metabolic pathways are evident in Figure 4.4, however the 99th percentile correlation cutoff differs for each pathway. The average expression profiles for cutin and suberin biosynthesis and BR biosynthesis correlate more strongly with individual transcript abundances (the lowest correlation coefficients presented are 0.954 and 0.918 respectively) than those for GA and sporopollenin biosynthesis (for which the lowest correlation coefficients presented are 0.902 and 0.862 respectively).

Figure 4.5 shows the accuracy of each metabolic pathway’s random forest’s gene expression profile predictions following training with the genes matching the pathway’s phylogenetic profile. Error predictions were made using out-of-bag samples for each random forest, and mirroring our Pearson’s correlation results, show that the random forests constructed for cutin and suberin biosynthesis and BR biosynthesis have greater accuracy in predicting the response variable (Figures 4.5A, B) than the random forests constructed for GA and sporopollenin biosynthesis (Figures 4.5C, D). This is reflected in the importance rankings of

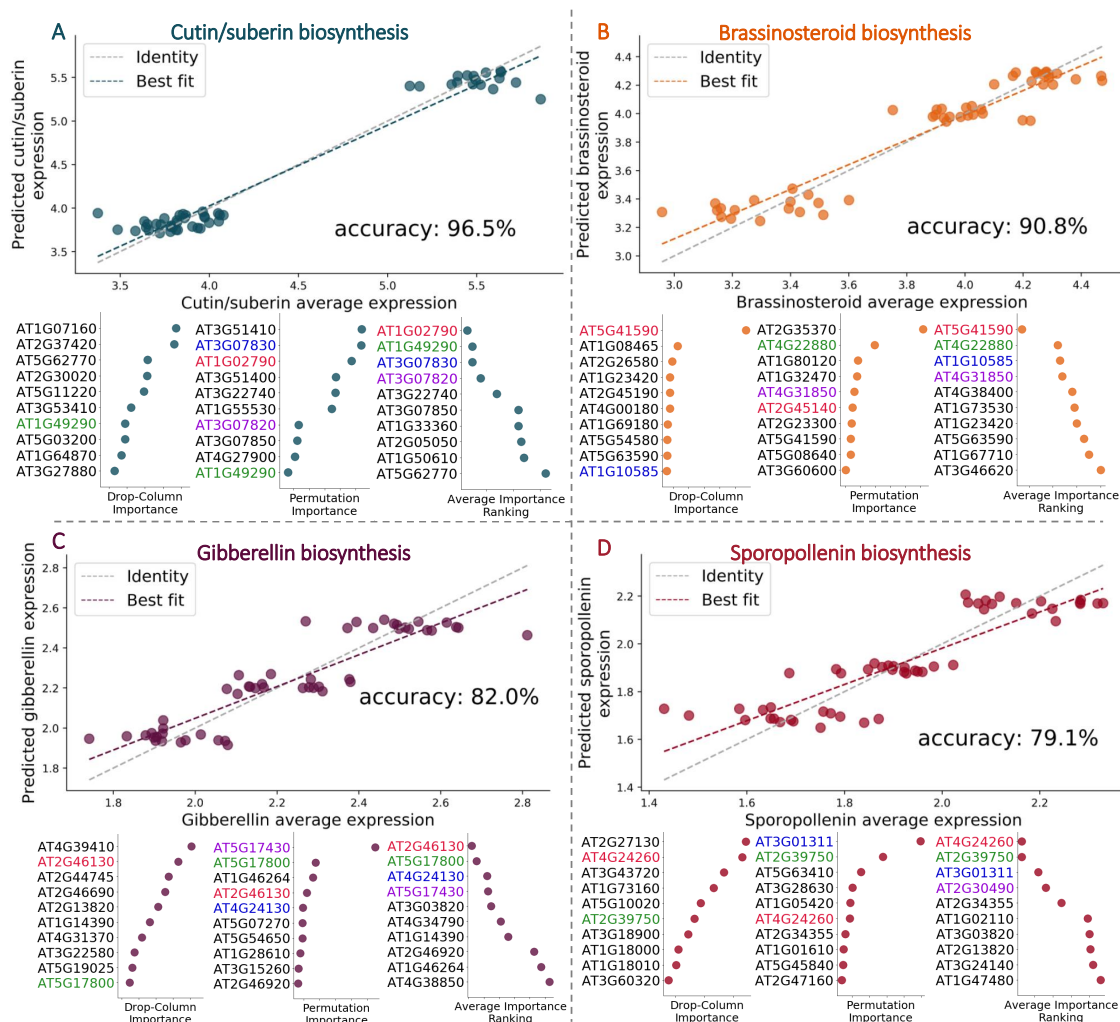


Figure 4.5: Illustrating the results of random forest analyses for each metabolic pathway. For each metabolic pathway, the upper panel shows the accuracy of the trained random forest in predicting the response variable (calculated using out-of-bag samples in each case). The scatter plot represents the out-of-bag prediction of the response variable and the coloured line indicates the line of best fit for the predicted values against the known response variable. The grey line indicates a perfect correlation for comparison. The out-of-bag prediction accuracy is noted in the bottom left corner of each graph. Expression values are logged and scaled. For each metabolic pathway, the left lower panel shows the top 10 most important genes as predicted using a drop-column measure of importance, in order of importance from top to bottom. The centre lower panel shows the top 10 most important genes as predicted using a permutation measure of importance, in order of importance from top to bottom. The right lower panel shows the top 10 genes ranked by their average importance, calculated from the drop-column and permutation importance. The four top-ranked genes are chosen for further analysis in each case – these are coloured on the y axis of the lower right panel and the matching genes in the lower left and centre panels can be identified by the corresponding colour.

individual genes in these analyses. The most important genes related to cutin and suberin biosynthesis and BR biosynthesis have high average rankings (indicating that the genes with the highest average ranking were ranked highly by both independent measures of gene importance) and the genes with highest average rankings can be seen in the top ten rankings of at least one measure of importance. On the other hand, the genes with the highest average ranking for sporopollenin biosynthesis were not necessarily present in the top ten ranking of either importance measure, indicating that the two separate measures of importance produced more incongruous results for the random forest used to analyse sporopollenin biosynthesis than for the other metabolic pathways. The Pearson's correlation coefficient results for sporopollenin biosynthesis correspondingly indicate that linear relationships between the genes matching the phylogenetic profile for sporopollenin biosynthesis and the sporopollenin biosynthesis gene expression profile are weaker than those for other metabolic pathways, with the correlation coefficient at its highest for AT1G06520 at 0.885, while the highest correlation coefficients for other metabolic pathways exceed 0.9.

4.6.2 Gibberellin biosynthesis-associated genes

GAs are biosynthesised in all areas of active plant growth and are involved in various developmental processes. The six genes that we have identified as candidates related to GA biosynthesis are correspondingly involved in divergent aspects of plant biology.

Two of the identified genes are associated with growth and stem cell maintenance in the root, shoot and floral meristems. The protein phosphatase PLL1 (POLTERGEIST-LIKE 1/AT2G35350) acts redundantly alongside POL (POLTERGEIST/AT2G46920) as part of a broader signalling cascade to positively regulate the expression of transcription factors responsible for specifying stem cell identity (Song et al. 2006). This functionality is required both in the embryo, where *pol pll1* double mutants present with a loss of asymmetric cell division and failure to appropriately specify cell types (Song et al. 2008), as well as in stem cell populations present at the shoot/flower and root meristems, where POL/PLL1 are required for the expression of WUS (WUSCHEL/AT2G17950) and WOX5 (WUSCHEL RELATED HOMEODOMAIN 5/AT3G11260) respectively, which maintain stem cell specificity in these regions (Song et al. 2008). Contrary to PLL1, which acts to maintain stem cells, GA

signalling is associated with cell differentiation and plant growth (Fleet and Sun 2005), and the homeodomain-containing transcription factor KNOX acts to maintain shoot meristems by directly repressing GA biosynthesis genes as well as initiating the degradation of GAs via cytokinin-induced activation of GA 2-oxidases (Fleet and Sun 2005; Jasinski et al. 2005). GAs are also thought to be required for appropriate cell differentiation the embryo (Fleet and Sun 2005) and have been shown to regulate homeotic genes necessary for normal flower development in *Arabidopsis* (Yu et al. 2004). While there does not appear to be any evidence of a direct interaction between GA biosynthesis and signalling and PLL1 expression, the two systems are known to be antagonistically involved in the same processes across plant structures and developmental stages and it therefore makes sense that PLL1 has been identified by our analysis as associated with GA biosynthesis.

The second meristem-related GA biosynthesis-associated gene is PLT4 (PLETHORA4/AT5G17430), one of several related transcription factors which control cell differentiation in the root meristem. Similar to PLL1, PLT4 functions redundantly with related PLT1/2/3 to promote cell division and prevent differentiation (Santuari et al. 2016). This occurs via a gradient of PLT expression highest at the centre of the root meristem which decreases through the division zone. GAs are known to regulate root growth in *Arabidopsis* via modulation of cell elongation (Ubeda-Tomás et al. 2008), however GAs have also been shown to positively regulate the size of the root meristem by promoting mitosis in proximal meristem cells (Ubeda-Tomás et al. 2009). So although it appears unlikely that GAs interact directly with PLT4, the shared involvement of GAs and PLT4 in maintenance of the root meristem and the identification of PLT4 by our random forest analysis acts as a good validator of the method's ability to identify GA-related genes.

The third GA biosynthesis-associated candidate gene we identified was MYB56 (AT5G17800), a transcription factor involved in seed development. MYB56 is a member of the twenty first R2R3 MYB transcription factor family, which includes several other MYB proteins involved in the biosynthesis of structural compounds used to thicken secondary cell walls (Zhang et al. 2013). *Arabidopsis* plants with reduced expression of MYB56 show reduced seed size while overexpression of MYB56 results in larger seeds than the wild type – this occurs due to the modified development of the seed endothelium and integument, likely as a result of direct

regulation various cell wall-related genes (Zhang et al. 2013). Separately, MYB56 has been shown to act as a negative regulator of the flowering inducer protein FT, antagonised by E3 ubiquitin ligases which target MYB56 for degradation (Chen et al. 2015). While MYB56 specifically has not yet been linked to GA biosynthesis or signalling, GAs are known to play a role in regulating flowering time and seed development (Mutasa-Göttgens and Hedden 2009; Singh et al. 2002) and MYB transcription factors are known targets of DELLA proteins in response to GA signalling in various other contexts (Gocal et al. 2001; Liu et al. 2017; Mutasa-Göttgens and Hedden 2009) – given the association with GA biosynthesis shown here it seems plausible that MYB56 fits into this signalling framework.

An additional GA biosynthesis associated candidate is WRKY43 (AT2G46130), a transcription factor which is thought to act following seed development as part of the regulatory processes involved with controlling germination. WRKY43 is a member of the large WRKY family of transcription factors which function in stress response pathways as well as in plant development (Phukan et al. 2016). WRKY43 is expressed in the developing seed, and has been shown to act as an ABA-dependent regulator of seed germination under abiotic stress conditions, accomplished via the direct and indirect regulation of genes affecting seed storage protein content and membrane fatty acid profile, as well as the salt tolerance gene ZAT10 (Geilen et al. 2017). *wrky43* mutants have higher concentrations of polyunsaturated fatty acids than wild type seeds, and have been shown to have greater germination success under high salt and low temperature conditions (Geilen et al. 2017). During germination, ABA signalling works antagonistically with GA signalling – the former promotes dormancy while the latter induces germination (Liu and Hou 2018). Interestingly, WRKY genes in rice have been shown to be associated with both ABA and GA. ABA-inducible OsWRKY51 and OsWRKY71 have been shown to mediate ABA/GA signalling crosstalk in rice seeds, repressing GA signalling in response to ABA (Xie et al. 2006). Additionally, three further rice WRKYs have been shown to additively repress signalling of both ABA and GA in seeds (Zhang et al. 2015). Given the evidence for GA as well as ABA associations in rice, in combination with our identification of WRKY43 as associated with GA biosynthesis, it seems possible that *Arabidopsis* WRKY43 may be similarly involved in ABA/GA crosstalk.

Further analysis of possible WRKY43-mediated effects on GA signalling in *Arabidopsis* seeds may provide further insight into ABA/GA crosstalk in this species.

Finally, two genes of unknown function were also identified by our analysis of GA biosynthesis-associated genes. The first, AT2G42660, is a homeodomain-containing transcription factor, thought to be a member of the MYB family. Although the function of AT2G42660 is uncharacterised, other MYB transcription factors are known to be involved in GA signalling (Gocal et al. 2001; Qi et al. 2014). Combining information from several broad-scale studies of gene expression shows that AT2G42660 is expressed in the root in *Arabidopsis* (Czechowski et al. 2004), and is down regulated in *Arabidopsis* embryos defective in auxin signalling (Möller et al. 2017). It was also found to be upregulated in high-growth allotriploid *Populus* seedlings along with various other genes related to GA, auxin and jasmonic acid (JA) metabolism (Cheng et al. 2015). The association of AT2G42660 with GA biosynthesis in our analysis, alongside its expression in the root and embryo (where GA signalling is known to play an important developmental role) and previous association with auxin metabolism, suggests that AT2G42660 may be an interesting candidate as a regulator of auxin/GA crosstalk during plant development in *Arabidopsis*.

The second uncharacterised gene identified by our analysis as GA biosynthesis-associated, AT4G24130, contains the domain of unknown function DUF538. While the specific functions of DUF538-containing proteins have yet to be identified, they do appear to be related to stress responses in plants – DUF538-containing proteins have recently been suggested to act as pectin methylesterases as part of cell wall-related defence responses to biotic stressors (Gholizadeh 2020), and have previously been associated with the activation of redox enzymes as well as chlorophyll binding, both under stress conditions (Gholizadeh 2011, 2018). AT4G24130 has been identified in several studies as upregulated in response to ABA (Böhmer and Schroeder 2011; Leonhardt et al. 2004), a major signalling molecule in plant responses to abiotic stress (Tuteja 2007). Although there is little information available regarding the function of AT4G24130, its association with ABA in previous studies combined with the suggestion in this work that it may be related to GA biosynthesis indicate that it may be involved in mediating the interaction between these two hormone signalling

pathways, perhaps as part of the stress response where the two hormones have been shown to play antagonistic roles (Shu et al. 2018).

4.6.3 Brassinosteroid biosynthesis-associated genes

BRs are steroid hormones biosynthesised from sterol precursors via a series of parallel reactions catalysed by cytochrome P450 enzymes from the CYP85 and CYP90 subfamilies (Figure 4.2). Like GAs, BRs are involved in broad ranging aspects of plant growth and development, including reproduction, differentiation and stress responses. As intercellular signals, BRs are recognised at the cell surface by BRI1 (BRASSINOSTEROID INSENSITIVE 1/AT4G39400) and its homologues. This triggers an intracellular cascade comprising the BZR1 (BRASSINAZOLE-RESISTANT 1/AT1G75080) and BES1 (BRASSINAZOLE-RESISTANT 2/AT1G19350) transcription factors, affecting downstream expression of BR-regulated genes (Nolan et al. 2020).

One of the genes we identified as BR biosynthesis-associated is anthocyanidin synthase/leucoanthocyanidin dioxygenase (ANS/LDOX/AT4G22880), a 2-oxoglutarate-dependent oxygenase involved in the late stages of anthocyanin production where it is used to convert leucoanthocyanidins to anthocyanidins (Wilmouth et al. 2002). There is extensive evidence that BRs are involved in the regulation of anthocyanin biosynthesis (Peng et al. 2011; Vergara et al. 2018; Yuan et al. 2015). LDOX is specifically implicated in this process – the expression of LDOX has been shown to be significantly reduced in *Arabidopsis* mutants deficient in BR production (Yuan et al. 2015). The identification of LDOX as an important gene via our random forest analysis therefore validates the ability of the method to produce true positive genes which are associated with the metabolic pathway in question.

The two genes with the highest Pearson's correlation coefficient against the mean expression of BR biosynthesis genes both encode vesicle-associated membrane protein-associated proteins (VAPs) – VAP27-1 (AT3G60600) and PVA12 (AT2G45140), highly conserved proteins associated with various cellular membranes in plants. PVA12 is localised to the ER, where it interacts with the sterol binding protein ORP3a (AT5G02100) (Saravanan et al. 2009). Here, it is thought to play a role in sterol trafficking between the ER and Golgi apparatus (Saravanan et al. 2009), possibly representing one stage in the transport of sterols

from the ER, where they are biosynthesised, to the plasma membrane where they are known to accumulate (Marie-Andrée et al. 1978; Moreau et al. 1998). While sterols are an important component of plant membranes, they are also the precursors of BR hormones (Bajguz et al. 2020) – it is therefore unsurprising that the transcript abundances of a gene whose function impacts BR precursor location and availability is correlated with the expression of BR biosynthetic genes themselves. However, the link between BR biosynthesis and VAP27-1 is less obvious. VAP27-1 is also associated with the ER, where it is thought to interact with the cytoskeleton and plasma membrane-associated protein NET3C (AT2G47920) to establish contact sites between the ER and the plasma membrane (Wang et al. 2014, 2016). VAP27-1 at these contact sites has been associated with both autophagy and endocytosis (Wang et al. 2019). Although it is unclear how this may interact with BR biosynthesis, BR signalling is a complex multicellular process which relies on the recycling of the extracellular receptor BRI1 between the plasma membrane and internal endocytic compartments via endocytosis (Di Rubbo et al. 2013; Russinova et al. 2004), and under starvation conditions the BR signalling master regulator BES1 is specifically targeted for degradation via autophagy (Nolan et al. 2017) – so it is possible there is a link in one of these contexts.

Another BR biosynthesis-associated gene candidate we identified is PGR3 (AT4G31850), a nuclear-encoded chloroplast-localised protein that plays a role in regulating photosynthetic electron transport (Yamazaki et al. 2004). PGR3 acts to stabilise RNA of specific chloroplastic genes including the *petL* (photosynthetic electron transport L) operon and promote translation at the ribosome, as well as promoting the formation of the ribosome itself (Cai et al. 2011; Fujii et al. 2013; Rojas et al. 2018). BRs are involved in various aspects of chloroplast development and photosynthetic regulation (Siddiqui et al. 2018), so it is possible that the predictive power of the transcript abundances of PGR3 derives simply from the fact that it is involved in a process related to those regulated by BR signalling. However, it is also possible that BRs are involved with the regulation of this gene. It has recently been identified that NOT4A, a cytosolic E3 ubiquitin ligase, is required for PGR3 expression in *Arabidopsis* – *not4a* mutants are not able to transcribe PGR3 and it is thought that NOT4A suppresses negative regulators of PGR3 (Bailey et al. 2020). BR signalling mediates varied

biological processes by regulating protein degradation via E3 ubiquitin ligases (Mazzucotelli et al. 2006) – it is conceivable that a similar form of BR-mediated regulation is at work here.

Random forest analysis of BR biosynthesis also produced two genes of unknown specific function - AT5G41590, annotated as LURP1-like, and AT1G10585, a basic helix-loop-helix (bHLH) DNA-binding superfamily protein recently designated AIB1 (Lumba et al. 2014). AT5G41590 is a member of the LOR family of genes in *Arabidopsis* which are defined by a conserved domain of unknown function (DUF567) (Knoth and Eulgem 2008). The only characterised gene in this family, LURP1, is a pathogen response gene required for *Arabidopsis* to elicit an immune response to the oomycete *Hyaloperonospora parasitica* (Knoth and Eulgem 2008) – however the other LOR genes do not elicit an obvious defence-related pattern of expression, suggesting that they may have alternative functions. While there is no existing information implicating AT5G41590 in BR signalling, another gene annotated as LURP1-like has been shown to be upregulated in response to BRs (Gupta et al. 2015), suggesting that there is precedent for BR interaction with these genes.

There are at least 133 bHLH genes in *Arabidopsis* (Heim et al. 2003) – these genes encode transcription factors and are involved in the regulation of a wide range of processes including growth, development and response to environmental factors (Toledo-Ortiz et al. 2003). There are several known examples of BRs modulating the expression of bHLH transcription factors to effect downstream gene expression (Friedrichsen et al. 2002; Zhang et al. 2009) and AIB1 specifically has been suggested to play a role in negatively regulating salt homeostasis in *Arabidopsis* as in the context of ABA signalling (Lumba et al. 2014), and has additionally been implicated in JA signalling in *Arabidopsis* (Sasaki-Sekimoto et al. 2013). Furthermore, AIB1 (along with several known BR-related transcriptions factors) was found to be significantly differentially expressed in *Arabidopsis* mutants presenting with modified methylation statuses of the protein phosphatase 2A catalytic subunit, a factor known to be important in BR signalling in *Arabidopsis* (Creighton et al. 2017). It seems that this gene plays a central role in hormone signalling in *Arabidopsis* and may be a candidate regulator of significant crosstalk among hormone systems. The evidence from our analysis that AIB1 is related to BR biosynthesis supports this information and suggests that further experimental investigation of this gene may extend our knowledge of hormone signalling regulation in *Arabidopsis*.

4.6.4 Sporopollenin biosynthesis-associated genes

Known as one of the most robust and inert natural compounds in existence, sporopollenin is the major component of the outer pollen wall, and forms an extremely tough coating which provides pollen grains with protection from environmental damage. The recalcitrance of sporopollenin to chemical degradation has rendered it extremely difficult to investigate in detail on a molecular level. Nonetheless, there is evidence that sporopollenin is composed mainly of polyketide-derived monomers in addition to various fatty acids and phenylpropanoid derivatives (Li et al. 2019; Quilichini et al. 2015). Although in spermatophytes sporopollenin is limited to the pollen wall, sporopollenin-like polymers are found across the Chloroplastida and with the exception of specific fatty acid reductase enzymes the full angiosperm-like biosynthetic pathway is found in all land plant taxa (Figure 4.3).

Cinnamate-4-hydroxylase (C4H/AT2G30490), identified as related to sporopollenin biosynthesis, catalyses one of the initial reactions in the phenylpropanoid biosynthesis pathway, the downstream reactions of which produce a huge range of flavonoids which are crucial in protecting plants from UV damage and as components of structural polymers, as well as monolignols for the production of lignin (Schilmiller et al. 2009). It is well-established that sporopollenin contains various phenolic compounds (Quilichini et al. 2015; Xue et al. 2020), and mutations in C4H are unable to develop pollen (Schilmiller et al. 2009). Furthermore, interruption of the MYB4 (AT4G38620) transcription factor responsible for negative regulation of C4H resulted in abnormal sporopollenin development and warped pollen grains, presumably stemming from a resulting imbalance in the concentrations of different metabolites required for sporopollenin production (Preston et al. 2004). Given the established connection between the production of phenylpropanoid precursors by C4H and sporopollenin production, the identification by our random forest analysis of C4H as an important predictor of sporopollenin biosynthetic gene transcript abundance acts as a validator of the capacity of this method to identify true functional relationships between genes.

Sporopollenin is produced in the anther, where pollen grains develop in a centralised area called the locule. Immediately surrounding the locule is the tapetum, which is known to supply the locule with nutrients and metabolites including sporopollenin precursors, and is thought to be the location of sporopollenin monomer biosynthesis before transport

into the locule for polymerisation and deposition (Quilichini et al. 2015). One of the genes we identified as associated with sporopollenin biosynthesis, GPAT1 (AT1G06520), is a mitochondrial *sn*-2-specific glycerol-3-phosphate acyltransferase (GPAT) expressed in flowers and known to be essential for both anther and tapetum development (Zheng et al. 2003). GPATs form a family of nine genes involved with glycerolipid biosynthesis in various contexts in plants. GPAT4-8 are well characterised and involved in the biosynthesis of lipids for incorporation into cutin and suberin polymers (see Chapter 3). The precise function of GPAT1 is less well established (Yang et al. 2012), but it is known to act as an acyltransferase on medium and long chain acyl-CoAs (16C-22C) and lacks the additional phosphatase activity of the cutin-associated GPATs (Yang et al. 2012). *Arabidopsis* mutants deficient in GPAT1 showed changes in the lipid profile of cell membranes in the tapetum and pollen grains, as well as modified ER morphology (Zheng et al. 2003) resulting in much-reduced male fertility, suggesting that GPAT1 is required for the production of glycerolipids destined for these membranes. GPAT1 mutants were also defective in the deposition of the outer pollen wall, leading to irregular exine deposition and collapsed pollen walls (Li et al. 2012; Zheng et al. 2003). Although sporopollenin, the main component of the pollen exine, also contains several lipid-derived monomers (Quilichini et al. 2015), there has been no previous indication that GPAT1 is involved in the production of sporopollenin precursors. The more likely process by which a GPAT1 defect could affect the pollen wall is related to inhibited export of sporopollenin precursors from the tapetum – many of these compounds are secreted into the locule via membrane-bound vesicles which are presumably affected by GPAT1-derived alteration of the lipid components of the plasma membrane, and *gpat1* mutants have been shown to exhibit impaired tapetal secretion (Zheng et al. 2003). It is therefore likely that, rather than stemming from a direct biochemical or regulatory relationship, the correlation between GPAT1 transcript abundance and expression of sporopollenin biosynthetic genes is the result of shared involvement in the process of sporopollenin deposition.

The third gene we identified as associated with sporopollenin biosynthesis is KOR3 (AT4G24260), one of a small family of membrane-anchored endo-1,4- β -D-glucanases. Its originally identified homologue, KOR, is required for the production of normal primary cell walls and xylem vessels in *Arabidopsis*, where it is thought to stabilise cellulose/hemicellulose

interactions from the interface between the cell wall and the plasma membrane (Maloney et al. 2012). The expression of KOR3 has been shown to be more restricted in its expression than that of KOR (which is present throughout *Arabidopsis* tissues) and KOR3 is suggested to be expressed specifically in locations where cell walls develop and require reinforcement (Mølhøj et al. 2001). In support of this, KOR3 has been implicated in the development of root syncytia (which involves various cellular modifications including the thickening of cell walls) following infection by cyst nematodes in *Arabidopsis* (Wieczorek et al. 2007), and *kor3 Arabidopsis* mutants challenged with infection by *Pseudomonas syringae* and *Botrytis cinerea* showed increased callose deposition following infection, suggesting that KOR3 may act as a negative regulator of callose formation. While there is no existing work investigating a possible link between KOR3 and pollen development, the various contexts of plant development and responses in which KOR3 is involved with reinforcing cell wall structures and particularly callose development, combined with its identification here as related to the biosynthesis of sporopollenin, suggests that it may be worth investigating whether it plays a role in the fortification of the pollen wall – either relating specifically to the deposition of sporopollenin, or more likely, the formation of the callose covering which develops as an outer layer to separate pollen grains during development (Nishikawa et al. 2005).

Finally, three unnamed genes were also identified as associated with sporopollenin biosynthesis. AT3G01311 is annotated as a putative actin cross-linking protein containing a domain of unknown function (DUF569). Very little information is available pertaining to this specific gene in *Arabidopsis*, however active reorganisation of the actin skeleton is required throughout the pollen lifecycle in plants – from the meiotic development of the gametes to pollen germination and tube growth (Genuardo et al. 1998; Jia et al. 2013; Van Lammeren et al. 1989). As there is no evidence to suggest that actin cross-binding proteins are directly associated with the process of sporopollenin biosynthesis or deposition, it is likely this relationship between the expression of AT3G01311 and sporopollenin stems from shared involvement in pollen development, and it would be interesting to investigate whether a pollen-specific function of AT3G01311 can be identified. The second unnamed gene is AT2G39750, annotated as an S-adenosyl-L-methionine (SAM)-dependent methyltransferase. SAM-dependent methyltransferases are found across plant

tissues and are involved in a broad spectrum of plant metabolic processes (Joshi and Chiang 1998) – as such it is difficult to speculate on the function of this gene without further information. However, SAM-dependent methyltransferases are known to participate in the process of pollen development, either by methylation of metabolic compounds themselves or by acting as DNA methyltransferases which impact gene expression (Fellenberg et al. 2008; Hu et al. 2014). Given its importance in predicting sporopollenin gene transcript abundance in our analysis (the top ranked gene, present in the top ten genes of both importance measures), it seems plausible that this methyltransferase is similarly involved in pollen development, either directly associated with sporopollenin biosynthesis (e.g. methylation of constituent sporopollenin compounds) or more generally as part of an associated process. Lastly, AT3G50780 encodes a BTB-domain-containing protein of the H-subfamily, a grouping of five genes known to respond to sucrose signalling in *Arabidopsis* (Hwang et al. 2018). AT3G50780 itself appears to undergo a slow increase in expression in response to sucrose application, peaking at 36 hours after application (Hwang et al. 2018). While the specific function of AT3G50780 has not been investigated, sucrose signalling is known to play diverse roles in plant growth and development including pollen development (Ruan 2014; Zhang et al. 2001) – repression of a sucrose invertase in tobacco anthers blocked pollen development to the point of male sterility by affecting carbohydrate supply in the tapetum (Goetz et al. 2001). Without further information about the expression patterns and molecular interactions of AT3G50780, it is impossible to suggest a specific function for this gene as it relates to sporopollenin biosynthesis, however it may be an interesting candidate for involvement in the sucrose signalling network in developing flowers.

4.6.5 Cutin and suberin biosynthesis-associated genes

Like sporopollenin, cutin and suberin are hydrophobic biopolymers which provide structural support as well as protection from desiccation and pathogens, albeit across a much broader range of plant tissues. Both are lipid-based polymers and share a significant portion of their biosynthetic pathways, which are catalysed predominantly by cytochrome P450 enzymes from the CYP86 and CYP77 subfamilies. Cutin is the main component of the plant cuticle and is found covering all aerial plant organs from leaves to anthers (Fich et al. 2016), while suberin

is more limited in its deposition and is found at the root epidermis as well as abscission zones and is often deposited in response to physical tissue damage (Philippe et al. 2020).

Both Pearson's correlation and random forest analyses of genes related to cutin and suberin biosynthesis identified one of the same genes – PGA4 (AT1G02790), a polygalacturonase expressed in the tapetum of the anther in *Arabidopsis* (Imoto et al. 2005). Interestingly, all four of the other genes identified as related to cutin and suberin biosynthesis by our analysis also encode proteins associated with the breakdown of pectin polymers, which are an important component of the primary cell wall, as well as the middle lamella, pollen intine and pollen tube wall in plants (Yang et al. 2018). These are PGA3 (AT3G07830), a polygalacturonase expressed in the dehiscence zones of anthers and siliques, two unidentified pectin lyase genes homologous to PGA3 and PGA4 (AT3G14040 and AT3G07820, the former of which appears to be expressed during flower development and the latter under hypoxic stress conditions) and PEG2 (AT1G49290), a paternally expressed imprinted gene expressed in the endosperm of developing seeds.

Unlike the other four identified genes, PEG2 does not appear to be directly involved in pectin breakdown. Although the function of the encoded protein is unknown (and it does not contain any known functional domains) PEG2 is known to be responsible for hybrid incompatibility, and *Arabidopsis peg2* mutants are able to develop viable triploid seeds (Wolff et al. 2015). Analysis of the gene expression effects in *peg2* mutants showed that down-regulated downstream genes are significantly enriched in polygalacturonases, suggesting that PEG2 acts as a promoter of pectin hydrolysis which represents a barrier to endosperm cellularisation during seed development (Wolff et al. 2015). While there is no evidence to suggest that PEG2 is directly involved with the process of cutin biosynthesis, cutin is the basis of a layer of cuticle that forms during seed development in tight association with the endosperm, and which may be of endospermic origin (De Giorgi et al. 2015), suggesting correlated involvement in a specific stage of seed development may be the source of a relationship between the expression of PEG2 and cutin biosynthetic genes.

Polygalacturonases are thought to be involved in the various stages of a plant's lifecycle where disconnection between neighbouring cells via cell wall and lamella degradation is

required – for example abscission, pollen development and anther dehiscence (González-Carranza et al. 2007). However, they are also important in the production of cell walls – following the synthesis of pectin polymers, hydrolysis is often required for the specific modification of polymers before incorporation into the cell wall matrix. Cell wall polysaccharides and cutin monomers are likely to be produced concurrently, as the plant cuticle contains components of both (Dominguez et al. 2011). Additionally, on the surface of plant organs where cutin is deposited, it is attached to the cell wall of the outermost cells by an intermediate, pectin-rich layer (Mazurek et al. 2017; Shi et al. 2011), suggesting the simultaneous production of both cutin and pectin is necessary for successful cuticle deposition. Moreover, the SHN family of transcription factors in *Arabidopsis* have been shown to have downstream effects on both cutin and suberin biosynthetic genes as well as polygalacturonases and pectin lyases (Shi et al. 2011), providing further evidence that the processes mediated by these genes may be related. It has even been suggested that interaction with cell wall polymers such as pectin may be necessary in the production of cutin monomers (Guzman-Puyol et al. 2015). Furthermore, under organ growth conditions, polygalacturonases are required to degrade cell wall pectins in order to promote the elasticity of the cell wall (Yang et al. 2018) while the cuticle must be continually produced and deposited (Fich et al. 2016) – providing another circumstance where the expression of the two sets of genes may well be coordinated. There are additional situations during a plant's life history where both pectin and the plant cuticle are degraded in synchrony, for example during fruit ripening (Fasoli et al. 2016), which could provide an additional condition under which the expression of one set of genes may be used to predict that of the other.

There are therefore various possible sources of a relationship between the expression of polygalacturonase and pectin lyase genes and cutin and suberin biosynthetic genes. However, the strong relationship between the two sets of genes, particularly PGA4, warrants further investigation into the specific function of PGA4 in the anther and the possibility of shared regulatory mechanisms between the two sets of genes. It also may provide a starting point for investigating the functions of the pectin lyases AT3G14040 and AT3G07820, which have not previously been linked to the development of the plant cuticle.

4.7 DISCUSSION

We have developed a bioinformatic pipeline which combines phylogenetic and gene expression-based methods of gene association analysis and used it to identify genes which are associated with the biosynthesis of phytohormones and biopolymers and which represent promising candidates for further analysis. Both random forest regression analyses and Pearson's correlation coefficients were used to identify associated genes, as this allowed the identification of both linear and non-linear relationships between gene transcript abundances. Random forest regressors were most accurate for BR biosynthesis and cutin and suberin biosynthesis, for which random forests could predict the response variable with more than 90% accuracy, while the random forests used for GA and sporopollenin biosynthesis were able to predict the response variables with 82% and 79.1% accuracy respectively (Figure 4.5).

There are several possible reasons for this disparity between random forest accuracy across different metabolic pathways. For example, if consecutive segments of a metabolic pathway are carried out in a spatially or temporally distinct manner, this may produce a less coherent averaged gene expression profile for the whole pathway that is difficult to match to genes which are related to each step. Another possibility is that the assumption that genes whose phylogenetic profile matches that of the metabolic pathway are more likely to be related does not hold true for all metabolic pathways. For example, GA biosynthetic genes are present in all land plants, however GA metabolites have not yet been identified in bryophytes (Cannell et al. 2020; Hayashi et al. 2010). The lack of experimental evidence for GAs in these species suggests that if they are present, it is likely to be at very low concentrations and GAs may have an alternative use in bryophytes compared to their wide-reaching functions in later-diverging land plants. It may therefore be the case that there are a significant number of GA-associated genes which are not homologous between early- and later-diverging land plants – if this is true then they would not be identified in a phylogenetic profile focused on genes present in all land plants. Similarly, genes involved in sporopollenin biosynthesis are present in all land plants, and while the use of sporopollenin across land plants is analogous (it is used to encase seeds in spermatophytes and covers spores in earlier-diverging land plants), it is possible that in concert with the genetic changes that

led to the evolution of seed production, the genes involved with sporopollenin transport, deposition and regulation may also differ between these sets of species. Further evidence that phylogenetic profiling may be less relevant for GA biosynthesis comes from the correlation coefficients identified for this pathway. For all other metabolic pathways investigated, Pearson's correlation coefficients calculated for all genes with matching phylogenetic profiles resulted in a higher than expected number of genes with correlation coefficients greater than that of 99% of randomly selected genes – indicating that the phylogenetic profile is successfully concentrating genes likely to have correlated expression levels with the metabolic pathway in question. However, this is not the case for GA biosynthesis, for which the number of genes within the phylogenetic profile with correlation coefficients higher than that for 99% of randomly selected genes is slightly under 1%.

Despite the possibly reduced relevance of the phylogenetic profile to GA and sporopollenin biosynthesis, the homology of biosynthetic genes themselves for both compounds across the land plants does suggest that at least some associated genes are likely to be present in the corresponding set of species, and the random forest accuracy of 79.1% was considered high enough to be a source of potentially valuable information on gene functional relationships – therefore the genes resulting from the both analyses were explored in detail as with those identified as related to the other metabolic pathways. In characterising the potential relationships between the genes identified as associated with each metabolic pathway and the metabolic pathway in question, we found that several of our identified genes have known associations with the metabolic pathway question and therefore act as validators of our method's ability to produce true positive gene associations. Examples of these include LDOX, which was identified as associated with BR biosynthesis, and C4H, which was identified as associated with sporopollenin biosynthesis. In addition to these genes, we were able to identify associations between metabolic pathways and genes for which some function is already known. In this context, the association of the gene in question with a specific metabolic pathway may provide a basis on which to expand our knowledge of the gene in question's function, and in some cases a possible mechanistic link between signalling systems, for example in the case of WRKY43, which was identified as associated with GA biosynthesis but has previously been linked to ABA. Finally, for each metabolic pathway at least one gene

which is thus far uncharacterised was found associated. In many cases these genes are thought to be transcription factors, and their association here with a specific metabolic pathway may act as an important clue for directing further research into their individual functions.

4.8 LIMITATIONS OF THE APPROACH

The availability of extensive gene expression and homology data as a result of continuing advances in genomic sequencing technology provides the opportunity for broad-ranging exploratory analysis of metabolic and gene regulatory processes. Here, we have exploited such data to produce a set of genes as candidates for association with several metabolic pathways, either in a functional or regulatory context. The method has been validated by our pipeline's identification of several genes which are known to be associated with the metabolic pathways in question, but there are nonetheless several limitations to the approach. Firstly, co-expression analyses cannot provide information about causality – so although we are able to identify that the activity of two genes are synchronised spatially and temporally and this indicates involvement in the same biological process, the suggestion of a specific regulatory context by necessity requires experimental investigation to confirm.

Additionally, the presence of genes involved in a metabolic pathway across a specific set of plant species may not necessarily imply that these genes are involved in the same processes in all included species. For example, homologues of *Arabidopsis* DELLA proteins (which are important mediators of GA-regulated gene expression) in the bryophyte *Physcomitrella patens* are not responsive to GA (Briones-Moreno et al. 2017), suggesting they are not involved in GA-regulated processes in these species. It is therefore likely that there will be genes related to GA biosynthetic genes in later diverging land plants which are not present in the bryophytes and vice versa. The phylogenetic profiling component of our analysis may also have been limited by the stringency with which we defined genes as being a member of the subset which matched our pathway's phylogenetic profile. A gene was only included in the phylogenetically matched subset if homologues were present in all species in which we defined the pathway to be present and none of the species in which we defined the pathway to be absent. While this gave us the highest likelihood of identifying true-positive genes as candidates for association with our metabolic pathways, it also may have limited our ability

to identify some associated genes – for example genes which are present in a wider range of species but were recruited for involvement with the metabolic pathway in question when it evolved, or more simply genes for which there may be discrepancies in our homology data where more flexibility in our profile definition (e.g. allowing a single-species deviation from the required profile) would have allowed inclusion.

Finally, in the situation where metabolic pathways have evolved in a step-wise manner (for example BR and cutin/suberin biosynthesis), limiting our analysis to species in which only the whole metabolic pathway is present may have reduced our capacity to identify associated genes, since presumably genes involved in regulatory processes are likely to evolve in tandem with the metabolic pathways they control. Identifying genes associated with only the segments of the metabolic pathways present in earlier-diverging land plants (as identified in Chapter 3) may also provide clues as to the functions of these partial pathways in earlier-diverging species. However, if the metabolic function of the full pathway in later-diverging plants is indeed different to that of the partial pathway in earlier-diverging plants, our use of *Arabidopsis* transcriptome data would produce a mismatch between the sets of homologous genes whose phylogenetic profile matches that of the partial pathway and those genes actually associated with the metabolic pathway in *Arabidopsis*. It would therefore be more informative for such an investigation to identify associated genes using a transcriptome dataset for an earlier-diverging species.

REFERENCES

- Bailey, M., Ivanauskaite, A., Grimmer, J., ... Gibbs, D. (2020). The *Arabidopsis* NOT4A E3 ligase coordinates PGR3 expression to regulate chloroplast protein translation. *bioRxiv*.
- Bajguz, A., Chmur, M., & Gruszka, D. (2020). Comprehensive overview of the brassinosteroid biosynthesis pathways: Substrates, products, inhibitors, and connections. *Frontiers in Plant Science*, 11, 1034.
- Basler, G., Fernie, A. R., & Nikoloski, Z. (2018). Advances in metabolic flux analysis toward genome-scale profiling of higher organisms. *Bioscience Reports*, 38(6), BSR20170224.
- Berens, M. L., Berry, H. M., Mine, A., ... Tsuda, K. (2017). Evolution of hormone signaling networks in plant defense. *Annual review of phytopathology*, 55, 401–425.
- Bhunia, R. K., Showman, L. J., Jose, A., & Nikolau, B. J. (2018). Combined use of cutinase and high-resolution mass-spectrometry to query the molecular architecture of cutin. *Plant Methods*, 14, 1–17.
- Böhmer, M., & Schroeder, J. I. (2011). Quantitative transcriptomic analysis of abscisic acid-induced and reactive oxygen species-dependent expression changes and proteomic profiling in *Arabidopsis* suspension cells. *Plant Journal*, 67, 105–118.

- Breiman, L. (2001). Random forests. *Machine Learning*, 45, 5–32.
- Briones-Moreno, A., Hernández-García, J., Vargas-Chávez, C., ... Blázquez, M. A. (2017). Evolutionary analysis of DELLA-associated transcriptional networks. *Frontiers in Plant Science*, 8, 626.
- Brunetti, C., Sebastiani, F., & Tattini, M. (2019). Review: ABA, flavonols, and the evolvability of land plants. *Plant Science*, 280, 448–454.
- Cai, W., Okuda, K., Peng, L., & Shikanai, T. (2011). PROTON GRADIENT REGULATION 3 recognizes multiple targets with limited similarity and mediates translation and rna stabilization in plastids. *The Plant Journal*, 67, 318–327.
- Cannell, N., Emms, D. M., Hetherington, A. J., ... Sweetlove, L. J. (2020). Multiple metabolic innovations and losses are associated with major transitions in land plant evolution. *Current Biology*, 30, 1783–1800.e11.
- Chang, C., Kong, L., Liu, Y., ... Gong, Z. (2020). Origins and evolution of cuticle biosynthetic machinery in land plants. *Plant Physiology*, 184, 1998–2010.
- Chen, L., Bernhardt, A., Lee, J., & Hellmann, H. (2015). Identification of *Arabidopsis* MYB56 as a novel substrate for CRL3BPM E3 ligases. *Molecular Plant*, 8, 242–250.
- Chen, L., Xiang, S., Chen, Y., ... Yu, D. (2017). *Arabidopsis* wrky45 interacts with the DELLA protein RGL1 to positively regulate age-triggered leaf senescence. *Molecular Plant*, 10, 1174–1189.
- Chen, M., Maodzeka, A., Zhou, L., ... Jiang, L. (2014). Removal of DELLA repression promotes leaf senescence in *Arabidopsis*. *Plant Science*, 219–220, 26–34.
- Cheng, S., Zhu, X., Liao, T., ... Kang, X. (2015). Gene expression differences between high-growth *Populus* allotriploids and their diploid parents. *Forests*, 6, 839–857.
- Claeys, H., De Bodt, S., & Inzé, D. (2014). Gibberellins and DELLAs: Central nodes in growth regulatory networks. *Trends in Plant Science*, 19, 231–239.
- Cohen, H., Fedyuk, V., Wang, C., ... Aharoni, A. (2020). SUBERMAN regulates developmental suberization of the *Arabidopsis* root endodermis. *The Plant Journal*, 102, 431–447.
- Creighton, M. T., Kolton, A., Kataya, A. R. A., ... Lillo, C. (2017). Methylation of protein phosphatase 2A-influence of regulators and environmental stress factors. *Plant, Cell and Environment*, 40, 2347–2358.
- Czechowski, T., Bari, R. P., Stitt, M., ... Udvardi, M. K. (2004). Real-time RT-PCR profiling of over 1400 *Arabidopsis* transcription factors: Unprecedented sensitivity reveals novel root-and shoot-specific genes. *Plant Journal*, 38, 366–379.
- De Giorgi, J., Piskurewicz, U., Loubery, S., ... Lopez-Molina, L. (2015). An endosperm-associated cuticle is required for *Arabidopsis* seed viability, dormancy and early control of germination. *PLoS Genetics*, 11.
- Delwiche, C. F., Graham, L. E., & Thomson, N. (1989). Lignin-like compounds and sporopollenin in *Coleochaete*, an algal model for land plant ancestry. *Science*, 245, 399–401.
- Di Rubbo, S., Irani, N. G., Kim, S. Y., ... Russinova, E. (2013). The clathrin adaptor complex AP-2 mediates endocytosis of BRASSINOSTEROID INSENSITIVE1 in *Arabidopsis* W. *The Plant Cell*, 25, 2986–2997.
- Dominguez, E., Heredia-Guerrero, J. A., & Heredia, A. (2011). The biophysical design of plant cuticles: An overview. *New Phytologist*, 189, 938–949.
- Emms, D. M., & Kelly, S. (2019). Orthofinder: Phylogenetic orthology inference for comparative genomics. *Genome Biology*, 20, 238.
- Exposito-Alonso, M., Burbano, H. A., & Bossdorf, O. (2019). Natural selection on the *Arabidopsis thaliana* genome in present and future climates. *Nature*, 573, 126–129.
- Fang, C., Fernie, A. R., & Luo, J. (2019). Exploring the diversity of plant metabolism. *Trends in Plant Science*, 24, 83–98.

- Fasoli, M., Dell'Anna, R., Dal Santo, S., ... Zenoni, S. (2016). Pectins, hemicelluloses and celluloses show specific dynamics in the internal and external surfaces of grape berry skin during ripening. *Plant and Cell Physiology*, 57, 1332–1349.
- Fellenberg, C., Milkowski, C., Hause, B., ... Vogt, T. (2008). Tapetum-specific location of a cation-dependent O-methyltransferase in *Arabidopsis thaliana*. *The Plant Journal*, 56, 132–145.
- Fich, E. A., Segerson, N. A., & Rose, J. K. C. (2016). The plant polyester cutin: Biosynthesis, structure, and biological roles. *Annual Review of Plant Biology*, 67, 207–233.
- Fleet, C. M., & Sun, T. P. (2005). A DELLAcate balance: The role of gibberellin in plant morphogenesis. *Current Opinion in Plant Biology*, 8, 77–85.
- Friedrichsen, D. M., Nemhauser, J., Muramitsu, T., ... Chory, J. (2002). Three redundant brassinosteroid early response genes encode putative bHLH transcription factors required for normal growth. *Genetics*, 162, 1445–1456.
- Fujii, S., Sato, N., & Shikanai, T. (2013). Mutagenesis of individual pentatricopeptide repeat motifs affects RNA binding activity and reveals functional partitioning of *Arabidopsis* PROTON GRADIENT REGULATION3. *The Plant Cell*, 25, 3079–3088.
- Gachon, C. M. M., Langlois-Meurinne, M., Henry, Y., & Saindrenan, P. (2005). Transcriptional co-regulation of secondary metabolism enzymes in *Arabidopsis*: Functional and evolutionary implications. *Plant Molecular Biology*, 58, 229–245.
- Geilen, K., Heilmann, M., Hillmer, S., & Böhmer. (2017). WRKY43 regulates polyunsaturated fatty acid content and seed germination under unfavourable growth conditions. *Nature Scientific Reports*, 7, 14235.
- Genuardo, G., Errico, A., Tiezzi, A., & Conicella, C. (1998). Tubulin and F-actin distribution during microsporogenesis in a 2n pollen producer of *Solanum*. *Genome*, 41(5).
- Gholizadeh, A. (2011). Heterologous expression of stress-responsive DUF538 domain containing protein and its morpho-biochemical consequences. *Protein Journal*, 30, 351–358.
- Gholizadeh, A. (2018). Chlorophyll binding ability of non-chloroplastic DUF538 protein superfamily in plants. *Proceedings of the National Academy of Sciences India Section B - Biological Sciences*, 88, 967–976.
- Gholizadeh, A. (2020). Pectin methylesterase activity of plant DUF538 protein superfamily. *Physiology and Molecular Biology of Plants*, 26(4), 829–839.
- Gocal, G. F. W., Sheldon, C. C., Gubler, F., ... King, R. W. (2001). GAMYB-like genes, flowering, and gibberellin signaling in *Arabidopsis*. 127, 1682–1693.
- Goetz, M., Godt, D. E., Guivarc, A., ... Roitsch, T. (2001). Induction of male sterility in plants by metabolic engineering of the carbohydrate supply. *Proceedings of the National Academy of Sciences of the United States of America*, 98(11), 6522–6527.
- González-Carranza, Z. H., Elliott, K. A., & Roberts, J. A. (2007). Expression of polygalacturonases and evidence to support their role during cell separation processes in *Arabidopsis thaliana*. *Journal of Experimental Botany*, 58, 3719–3730.
- Gupta, A., Singh, M., & Laxmi, A. (2015). Multiple interactions between glucose and brassinosteroid signal transduction pathways in *Arabidopsis* are uncovered by whole-genome transcriptional profiling. *Plant Physiology*, 168, 1091–1105.
- Guzman-Puyol, S., Benítez, J. J., Domínguez, E., ... Heredia-Guerrero, J. A. (2015). Pectin-lipid self-assembly: Influence on the formation of polyhydroxy fatty acids nanoparticles. *PLOS ONE*, 10, e0124639.
- Han, X., Yin, L., & Xue, H. (2012). Co-expression analysis identifies CRC and AP1 the regulator of *Arabidopsis* fatty acid biosynthesis. *Journal of Integrative Plant Biology*, 54, 486–499.

- Hayashi, K. I., Horie, K., Hiwatashi, Y., ... Nozaki, H. (2010). Endogenous diterpenes derived from *ent*-kaurene, a common gibberellin precursor, regulate protonema differentiation of the moss *Physcomitrella patens*. *Plant Physiology*, 153, 1085–1097.
- He, X., Dai, J., & Wu, Q. (2016). Identification of sporopollenin as the outer layer of cell wall in microalga *Chlorella protothecoides*. *Frontiers in Microbiology*, 7, 1047.
- Heim, M. A., Jakoby, M., Werber, M., ... Bailey, P. C. (2003). The basic helix-loop-helix transcription factor family in plants: A genome-wide study of protein structure and functional diversity. *Molecular Biology and Evolution*, 20, 735–747.
- Hu, J., Wang, Z., Zhang, L., & Sun, M. (2014). The *Arabidopsis* Exine Formation Defect (EFD) gene is required for primexine patterning and is critical for pollen fertility. *New Phytologist*, 203, 140–154.
- Hwang, J. E., Kim, Y. J., Shin, M. H., & Park, J. (2018). AtBTH1, a BTB domain-containing protein, influences sugar-responsive germination and seedling growth in *Arabidopsis*. *Genetics and Molecular Research*, 17(4).
- Imoto, K., Yokoyama, R., & Nishitani, K. (2005). Comprehensive approach to genes involved in cell wall modifications in *Arabidopsis thaliana*. *Plant Molecular Biology*, 58, 177–192.
- Jasinski, S., Piazza, P., Craft, J., ... Tsiantis, M. (2005). KNOX action in *Arabidopsis* is mediated by coordinate regulation of cytokinin and gibberellin activities. *Current Biology*, 15, 1560–1565.
- Jia, H., Li, J., Zhu, J., ... An, L. (2013). *Arabidopsis* CROLIN1, a novel plant actin-binding protein, functions in cross-linking and stabilizing actin filaments. *Journal of Biological Chemistry*, 288, 32277–32288.
- Joshi, C. P., & Chiang, V. L. (1998). Conserved sequence motifs in plant S-adenosyl-L-methionine-dependent methyltransferases. *Plant Molecular Biology*, 37, 663–674.
- Kim, J., & Buell, C. R. (2015). A revolution in plant metabolism: Genome-enabled pathway discovery. *Physiology*, 169, 1532–1539.
- Knoth, C., & Eulgem, T. (2008). The oomycete response gene LURP1 is required for defense against *Hyaloperonospora parasitica* in *Arabidopsis thaliana*. *The Plant Journal*, 55, 53–64.
- Lei, L., Steffen, J. G., Osborne, E. J., & Toomajian, C. (2017). Plant organ evolution revealed by phylotranscriptomics in *Arabidopsis thaliana*. *Scientific Reports*, 7.
- Leonhardt, N., Kwak, J. M., Robert, N., ... Schroeder, J. I. (2004). Microarray expression analyses of *Arabidopsis* guard cells and isolation of a recessive abscisic acid hypersensitive protein phosphatase 2C mutant. *The Plant Cell*, 16, 596–615.
- Li, F. S., Phyto, P., Jacobowitz, J., ... Weng, J. K. (2019). The molecular structure of plant sporopollenin. *Nature Plants*, 5, 41–46.
- Li, G., Köllner, T. G., Yin, Y., ... Chen, F. (2012). Nonseed plant *Selaginella moellendorffii* has both seed plant and microbial types of terpene synthases. *Proceedings of the National Academy of Sciences of the United States of America*, 109, 14711–14715.
- Liu, B., De Storme, N., & Geelen, D. (2017). Gibberellin induces diploid pollen formation by interfering with meiotic cytokinesis. *Plant Physiology*, 173, 338–353.
- Liu, H., Guo, S., Lu, M., ... Wang, W. (2019). Biosynthesis of DHGA₁₂ and its roles in *Arabidopsis* seedling establishment. *Nature Communications*, 10, 1768.
- Liu, X., & Hou, X. (2018). Antagonistic regulation of ABA and GA in metabolism and signaling pathways. *Frontiers in Plant Science*, 9, 251.
- Lumba, S., Toh, S., Handfield, L. F., ... McCourt, P. (2014). A mesoscale abscisic acid hormone interactome reveals a dynamic signaling landscape in *Arabidopsis*. *Developmental Cell*, 29, 360–372.

- Maloney, V. J., Samuels, A. L., & Mansfield, S. D. (2012). The endo-1,4- β -glucanase Korrigan exhibits functional conservation between gymnosperms and angiosperms and is required for proper cell wall formation in gymnosperms. *New Phytologist*, 193, 1076–1087.
- Mao, R., Kumar, P. K. R., Guo, C., ... Liang, C. (2014). Comparative analyses between retained introns and constitutively spliced introns in *Arabidopsis thaliana* using random forest and support vector machine. *PLoS ONE*, 9, e104049.
- Marie-Andrée, Hartmann-Bouillon, & Benveniste, P. (1978). Sterol biosynthetic capability of purified membrane fractions from maize coleoptiles. *Phytochemistry*, 17, 1037–1042.
- Mazurek, S., Garroum, I., Daraspe, J., ... Nawrath, C. (2017). Connecting the molecular structure of cutin to ultrastructure and physical properties of the cuticle in petals of *Arabidopsis*. *Plant Physiology*, 173, 1146–1163.
- Mazzucotelli, E., Belloni, S., Marone, D., ... Mastrangelo, A. (2006). The E3 ubiquitin ligase gene family in plants: Regulation by degradation. *Current Genomics*, 7, 509–522.
- McGuinness, P. N., Reid, J. B., & Foo, E. (2019). The role of gibberellins and brassinosteroids in nodulation and arbuscular mycorrhizal associations. *Frontiers in Plant Science*, 10.
- Miyazaki, S., Hara, M., Ito, S., ... Nakajima, M. (2018). An ancestral gibberellin in a moss *Physcomitrella patens*. *Molecular Plant*, 11, 1097–1100.
- Mølhøj, M., Jørgensen, B., Ulvskov, P., & Borkhardt, B. (2001). Two *Arabidopsis thaliana* genes, KOR2 and KOR3, which encode membrane-anchored endo-1,4- β -D-glucanases, are differentially expressed in developing leaf trichomes and their support cells. *Plant Molecular Biology*, 46, 263–275.
- Möller, B. K., Ten Hove, C. A., Xiang, D., ... Weijers, D. (2017). Auxin response cell-autonomously controls ground tissue initiation in the early *Arabidopsis* embryo. *Proceedings of the National Academy of Sciences of the United States of America*, 114, E2533–E2539.
- Moreau, P., Hartmann, M., Perret, A., ... Cassagne, C. (1998). Transport of sterols to the plasma membrane of leek seedlings. *Plant Physiology*, 117, 931–937.
- Morris, J. L., Puttick, M. N., Clark, J. W., ... Donoghue, P. C. J. (2018). The timescale of early land plant evolution. *Proceedings of the National Academy of Sciences of the United States of America*, 115, E2274–E2283.
- Mutasa-Göttgens, E., & Hedden, P. (2009). Gibberellin as a factor in floral regulatory networks. *Journal of Experimental Botany*, 60, 1979–1989.
- Niklas, K. J., Cobb, E. D., & Matas, A. J. (2017). The evolution of hydrophobic cell wall biopolymers: From algae to angiosperms. *Journal of Experimental Botany*, 68, 5261–5269.
- Nishikawa, S. I., Zinkl, G. M., Swanson, R. J., ... Preuss, D. (2005). Callose (β -1,3 glucan) is essential for *Arabidopsis* pollen wall patterning, but not tube growth. *BMC Plant Biology*, 5, 22.
- Nolan, T. M., Brennan, B., Yang, M., ... Li, Z. (2017). Selective autophagy of BES1 mediated by DSK2 balances plant growth and survival. *Developmental Cell*, 41, 33–46.
- Nolan, T. M., Vukasinović, N., Liu, D., ... Yin, Y. (2020). Brassinosteroids: Multidimensional regulators of plant growth, development, and stress responses. *The Plant Cell*, 32, 298–318.
- Parrr/random-forest-importances. (2020). Retrieved October 1, 2020, from <https://github.com/parrr/random-forest-importances>
- Pedregosa, F., Michel, V., Grisel, O., & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830.
- Peng, Z., Han, C., Yuan, L., ... Ren, C. (2011). Brassinosteroid enhances jasmonate-induced anthocyanin accumulation in *Arabidopsis* seedlings. *Journal of Integrative Plant Biology*, 53, 632–640.

- Persson, S., Wei, H., Milne, J., ... Somerville, C. R. (2005). Identification of genes required for cellulose synthesis by regression analysis of public microarray data sets. *Proceedings of the National Academy of Sciences of the United States of America*, 102, 8633–8638.
- Philippe, G., Sørensen, I., Jiao, C., ... Rose, J. K. C. (2020). Cutin and suberin: Assembly and origins of specialized lipidic cell wall scaffolds. *Current Opinion in Plant Biology*, 55, 11–20.
- Phukan, U. J., Jeena, G. S., & Shukla, R. K. (2016). WRKY transcription factors: Molecular regulation and stress responses in plants. *Frontiers in Plant Science*, 7, 760.
- Preston, J., Wheeler, J., Heazlewood, J., ... Parish, R. W. (2004). AtMYB32 is required for normal pollen development in *Arabidopsis thaliana*. *Plant Journal*, 40, 979–995.
- Qi, T., Huang, H., Wu, D., ... Xie, D. (2014). *Arabidopsis* DELLA and JAZ proteins bind the WD-repeat/ bHLH/MYB complex to modulate gibberellin and jasmonate signaling synergy. *The Plant Cell*, 26, 1118–1133.
- Quilichini, T. D., Grienberger, E., & Douglas, C. J. (2015). The biosynthesis, composition and assembly of the outer pollen wall: A tough case to crack. *Phytochemistry*, 113, 170–182.
- Ranea, J. A. G., Yeats, C., Grant, A., & Orengo, C. A. (2007). Predicting protein function with hierarchical phylogenetic profiles: The Gene3D Phylo-Tuner Method applied to eukaryotic genomes. *PLoS Computational Biology*, 3, e237.
- Rensing, S. A. (2018). Great moments in evolution: The conquest of land by plants. *Current Opinion in Plant Biology*, 42, 49–54.
- Rojas, M., Ruwe, H., Miranda, R. G., ... Barkan, A. (2018). Unexpected functional versatility of the pentatricopeptide repeat proteins PGR3, PPR5 and PPR10. *Nucleic Acids Research*, 46, 10448–10459.
- Ruan, Y. L. (2014). Sucrose metabolism: Gateway to diverse carbon use and sugar signaling. *Annual Review of Plant Biology*, 65, 33–67.
- Ruprecht, C., Mutwil, M., Saxe, F., ... Persson, S. (2011). Large-scale co-expression approach to dissect secondary cell wall formation across plant species. *Frontiers in Plant Science*, 2.
- Russinova, E., Borst, J., Kwaaitaal, M., ... De Vries, S. C. (2004). Heterodimerization and endocytosis of *Arabidopsis* brassinosteroid receptors BRI1 and AtSERK3 (BAK1). 16(12), 3216–3229.
- Santuari, L., Sanchez-Perez, G. F., Luijten, M., ... Heidstra, R. (2016). The PLETHORA gene regulatory network guides growth and cell differentiation in *Arabidopsis* roots. 28, 2937–2951.
- Saravanan, R. S., Slabaugh, E., Singh, V. R., ... Brandizzi, F. (2009). The targeting of the oxysterol-binding protein ORP3a to the endoplasmic reticulum relies on the plant VAP33 homolog PVA12. *The Plant Journal*, 58, 817–830.
- Sasaki-Sekimoto, Y., Jikumaru, Y., Obayashi, T., ... Shirasu, K. (2013). Basic helix-loop-helix transcription factors JASMONATE-ASSOCIATED MYC2-LIKE1 (JAM1), JAM2, and JAM3 are negative regulators of jasmonate responses in *Arabidopsis*. 163, 291–304.
- Schilmiller, A. L., Stout, J., Weng, J. K., ... Chapple, C. (2009). Mutations in the cinnamate 4-hydroxylase gene impact metabolism, growth and development in *Arabidopsis*. *Plant Journal*, 60, 771–782.
- Shi, J., Malitsky, S., De Oliveira, S., ... Aharoni, A. (2011). SHINE transcription factors act redundantly to pattern the archetypal surface of *Arabidopsis* flower organs. *PLoS Genetics*, 7, e1001388.
- Shigenaga, A. M., Berens, M. L., Tsuda, K., & Argueso, C. T. (2017). Towards engineering of hormonal crosstalk in plant immunity. *Current Opinion in Plant Biology*, 38, 164–172.
- Shu, K., Zhou, W., Chen, F., ... Yang, W. (2018). Absciscic acid and gibberellins antagonistically mediate plant development and abiotic stress responses. *Frontiers in Plant Science*, 9, 416.
- Siddiqui, H., Hayat, S., & Bajguz, A. (2018). Regulation of photosynthesis by brassinosteroids in plants. *Acta Physiologiae Plantarum*, 40, 59.

- Singh, D. P., Jermakow, A. M., & Swain, S. M. (2002). Gibberellins are required for seed development and pollen tube growth in *Arabidopsis*. *The Plant Cell*, 14, 3133–3147.
- Škunca, N., & Dessimoz, C. (2015). Phylogenetic profiling: How much input data is enough? *PLOS ONE*, 10, e0114701.
- Song, S. K., Hofhuis, H., Lee, M. M., & Clark, S. E. (2008). Key divisions in the early arabidopsis embryo require POL and PLL1 phosphatases to establish the root stem cell organizer and vascular axis. *Developmental Cell*, 15, 98–109.
- Song, S. K., Lee, M. M., & Clark, S. E. (2006). POL and PLL1 phosphatases are CLAVATA1 signaling intermediates required for *Arabidopsis* shoot and floral stem cells. *Development*, 133, 4691–4698.
- Stephan, J., Stegle, O., & Beyer, A. (2015). A random forest approach to capture genetic effects in the presence of population structure. *Nature Communications*, 6, 1–10.
- Ternes, P., Sperling, P., Albrecht, S., ... Heinz, E. (2006). Identification of fungal sphingolipid C9-methyltransferases by phylogenetic profiling. *Journal of Biological Chemistry*, 281, 5582–5592.
- Thornycroft, D. (2001). Using gene knockouts to investigate plant metabolism. *Journal of Experimental Botany*, 52, 1593–1601.
- Toledo-Ortiz, G., Huq, E., & Quail, P. H. (2003). The *Arabidopsis* basic/helix-loop-helix transcription factor family. *The Plant Cell*, 15, 1749–1770.
- Tuteja, N. (2007). Absciscic acid and abiotic stress signaling. *Plant Signaling and Behavior*, 2, 135–138.
- Ubeda-Tomás, S., Federici, F., Casimiro, I., ... Bennett, M. J. (2009). Gibberellin signaling in the endodermis controls *Arabidopsis* root meristem size. *Current Biology*, 19, 1194–1199.
- Ubeda-Tomás, S., Swarup, R., Coates, J., ... Bennett, M. J. (2008). Root growth in *Arabidopsis* requires gibberellin/DELLA signalling in the endodermis. *Nature Cell Biology*, 10, 625–628.
- Ursache, R., De Jesus Vieira-Teixeira, C., Tendon, V. D., ... Vermeer, J. (2020). GDSL-domain containing proteins mediate suberin biosynthesis and degradation, enabling developmental plasticity of the endodermis during lateral root emergence. *bioRxiv*.
- Van Lammeren, A. A. M., Bednara, J., & Willemsse, M. T. M. (1989). Organization of the actin cytoskeleton during pollen development in *Gasteria verrucosa* (Mill.) H. Duval visualized with rhodamine-phalloidin. *Planta*, 178, 531–539.
- Vergara, A. E., Díaz, K., Carvajal, R., ... Pérez-Donoso, A. G. (2018). Exogenous applications of brassinosteroids improve color of red table grape (*Vitis vinifera* L. Cv. “Redglobe”) berries. *Frontiers in Plant Science*, 9, 363.
- Verma, V., Ravindran, P., & Kumar, P. P. (2016). Plant hormone-mediated regulation of stress responses. *BMC Plant Biology*, 16, 1–10.
- Wang, P., Hawkins, T. J., Richardson, C., ... Hussey, P. J. (2014). The plant cytoskeleton, NET3C, and VAP27 mediate the link between the plasma membrane and endoplasmic reticulum. *Current Biology*, 24, 1397–1405.
- Wang, P., Pleskot, R., Zang, J., ... Wang, J. (2019). Plant AtEH/Pan1 proteins drive autophagosome formation at ER-PM contact sites with actin and endocytic machinery. *Nature Communications*, 10, 5132.
- Wang, P., Richardson, C., Hawkins, T. J., ... Hussey, P. J. (2016). Plant VAP27 proteins: Domain characterization, intracellular localization and role in plant development. *New Phytologist*, 210, 1311–1326.
- Wang, Z. Y., Bai, M. Y., Oh, E., & Zhu, J. Y. (2012). Brassinosteroid signaling network and regulation of photomorphogenesis. *Annual Review of Genetics*, 46, 701–724.
- Wei, Z., & Li, J. (2016). Brassinosteroids regulate root growth, development, and symbiosis. *Molecular Plant*, 9, 86–100.

- Wieczorek, K., Hofmann, J., Blöchl, A., ... Grundler, F. M. W. (2007). *Arabidopsis* endo-1,4- β -glucanases are involved in the formation of root syncytia induced by *Heterodera schachtii*. *The Plant Journal*, 53, 336–351.
- Wilmouth, R. C., Turnbull, J. J., Welford, R. W. D., ... Schofield, C. J. (2002). Structure and mechanism of anthocyanidin synthase from *Arabidopsis thaliana*. *Structure*, 10, 93–103.
- Wolff, P., Jiang, H., Wang, G., ... Köhler, C. (2015). Paternally expressed imprinted genes establish postzygotic hybridization barriers in *Arabidopsis thaliana*. *eLife*, 4, e10074.
- Xie, Z., Zhang, Z. L., Zou, X., ... Shen, Q. J. (2006). Interactions of two abscisic-acid induced WRKY genes in repressing gibberellin signaling in aleurone cells. *Plant Journal*, 46, 231–242.
- Xue, J. S., Zhang, B., Zhan, H. D., ... Yang, Z. N. (2020). Phenylpropanoid derivatives are essential components of sporopollenin in vascular plants. *Molecular Plant*, 13, 1644–1653.
- Yamazaki, H., Tasaka, M., & Shikanai, T. (2004). PPR motifs of the nucleus-encoded factor, PGR3, function in the selective and distinct steps of chloroplast gene expression in *Arabidopsis*. *The Plant Journal*, 38, 152–163.
- Yan, M., Xie, D., Cao, J., ... Yu, J. (2020). Brassinosteroid-mediated reactive oxygen species are essential for tapetum degradation and pollen fertility in tomato. *The Plant Journal*, 102, 931–947.
- Yang, W., Simpson, J. P., Li-Beisson, Y., ... Ohlrogge, J. B. (2012). A land-plant-specific glycerol-3-phosphate acyltransferase family in *Arabidopsis*: Substrate specificity, *sn*-2 preference, and evolution. *Plant Physiology*, 160, 638–652.
- Yang, Y., Yu, Y., Liang, Y., ... Cao, J. (2018). A profusion of molecular scissors for pectins: Classification, expression, and functions of plant polygalacturonases. *Frontiers in Plant Science*, 9, 1208.
- Yokota, T., Ohnishi, T., Shibata, K., ... Kohchi, T. (2017). Occurrence of brassinosteroids in non-flowering land plants, liverwort, moss, lycophyte and fern. *Phytochemistry*, 136, 46–55.
- Yu, J. Q., Huang, L. F., Hu, W. H., ... Nogués, S. (2004). A role for brassinosteroids in the regulation of photosynthesis in *Cucumis sativus*. *Journal of Experimental Botany*, 55, 1135–1143.
- Yuan, L. B., Peng, Z. H., Zhi, T. T., ... Ren, C. M. (2015). Brassinosteroid enhances cytokinin-induced anthocyanin biosynthesis in *Arabidopsis* seedlings. *Biologia Plantarum*, 59, 99–105.
- Zhang, L., Gu, L., Ringler, P., ... Shen, Q. J. (2015). Three WRKY transcription factors additively repress abscisic acid and gibberellin signaling in aleurone cells. *Plant Science*, 236, 214–222.
- Zhang, L. Y., Bai, M. Y., Wu, J., ... Wang, Z. Y. (2009). Antagonistic HLH/bHLH transcription factors mediate brassinosteroid regulation of cell elongation and plant development in rice and *Arabidopsis*. *The Plant Cell*, 21, 3767–3780.
- Zhang, Y., Liang, W., Shi, J., ... Zhang, D. (2013). MYB56 encoding a R2R3 MYB transcription factor regulates seed size in *Arabidopsis thaliana*. *Journal of Integrative Plant Biology*, 55, 1166–1178.
- Zhang, Y., Shewry, P. R., Jones, H., ... Halford, N. G. (2001). Expression of antisense SnRK1 protein kinase sequence causes abnormal pollen development and male sterility in transgenic barley. *Plant Journal*, 28, 431–441.
- Zheng, Z., Xia, Q., Dauk, M., ... Zou, J. (2003). *Arabidopsis* AtGPAT1, a member of the membrane-bound glycerol-3-phosphate acyltransferase gene family, is essential for tapetum differentiation and male fertility. *The Plant Cell*, 15, 1872–1887.
- Zi, J., Mafu, S., & Peters, R. J. (2014). To gibberellins and beyond! Surveying the evolution of (di)terpenoid metabolism. *Annual Review of Plant Biology*, 65, 259–286.

Statement of Authorship for joint/multi-authored papers for PGR thesis

To appear at the end of each thesis chapter submitted as an article/paper

The statement shall describe the candidate's and co-authors' independent research contributions in the thesis publications. For each publication there should exist a complete statement that is to be filled out and signed by the candidate and supervisor (**only required where there isn't already a statement of contribution within the paper itself**).

Title of Paper	Random forest analysis of phylogenetically limited gene sets facilitates identification of candidate genes associated with phytohormone and biopolymer biosynthesis
Publication Status	☐ Published ☐ Accepted for Publication ☐ Submitted for Publication ☒ Unpublished and unsubmitted work written in a manuscript style
Publication Details	Cannell et. al. (unpublished). Random forest analysis of phylogenetically limited gene sets facilitates identification of candidate genes associated with phytohormone and biopolymer biosynthesis.

Student Confirmation

Student Name:	Naomi Cannell		
Contribution to the Paper	Naomi Cannell and Lee J. Sweetlove conceived the project and developed the methodology; N.C. wrote code and carried out the investigation and formal analysis; L.J.S. provided supervision; Steven Kelly and David M. Emms provided software and computing resources; N.C. wrote the manuscript and prepared the figures; L.J.S reviewed and edited the manuscript.		
Signature Naomi Cannell	Date	15/01/2021	

Supervisor Confirmation

By signing the Statement of Authorship, you are certifying that the candidate made a substantial contribution to the publication, and that the description described above is accurate.

Supervisor name and title: Professor Lee J. Sweetlove		
Supervisor comments		
Signature 	Date	18/01/2020

This completed form should be included in the thesis, at the end of the relevant chapter.

5

Discussion

5.1 SUMMARY OF THE MAIN FINDINGS PRESENTED IN THIS THESIS

The goal of this thesis was to use computational and bioinformatic tools to characterise metabolism in early-diverging land plants, aiming to help address the current angiosperm-centric view of plant metabolism. The work was presented as three manuscripts.

In the first (Chapter 2), I focused on central metabolism in the liverwort *Marchantia polymorpha*. I used metabolic modelling combined with flux-balance analysis (FBA) to investigate the core flux consequences of the reduced carbon budget available to *Marchantia* as a result of its low mesophyll conductance (Flexas and Carriqui 2020; Gago et al. 2019). My finding that a carefully constructed computational model of *Marchantia* metabolism is unable to predict experimentally defined growth rates without invoking some form of metabolic carbon recycling or a much-reduced cell maintenance cost suggest that *Marchantia in vivo* exhibits a much higher carbon conversion efficiency (CCE) than that of C_3 angiosperms. While my metabolic model predicts crassulacean acid metabolism (CAM)-like carbon recycling at night via phosphoenolpyruvate carboxylase to accommodate this requirement for a higher CCE, there are various other ways a plant could increase CCE (see Section 2.7). These include metabolic mechanisms such as some form of CAM idling or cycling, or perhaps decreased maintenance costs as a result of reductions in protein turnover or membrane leakiness. Processes to improve CCE are found across land plants and algae (Sage et al. 2011;

Scheurwater et al. 1998; Silvera et al. 2010; Villarreal and Renner 2012) and experimental work is required to identify whether such an adaptation is present in *Marchantia*.

In the second (Chapter 3), I investigated the evolution of metabolism across land plants and algae. Using eight reference species representing each major clade of land plants, I used a comparative method to identify possible metabolic transitions over the course of land plant evolution based on the presence or absence of genes encoding enzymes. Identified metabolic transitions (i.e. evolutionary gains or losses of metabolic capabilities) were confirmed by comparison of gene presences/absences in a further 64 land plants and algae (based on 63 genomes and one transcriptome) and supported by analysis of transcriptomes of 305 early-diverging land plants and algae. Several of the metabolic innovations identified relate to the biosynthesis of phytohormones and biopolymers. My results resolve details of the evolutionary processes governing the appearance of these biosynthetic pathways.

Finally, my third manuscript (Chapter 4) focused on the identification of genes associated with specific metabolic pathways of interest. I used the guilt-by-association approaches of phylogenetic profiling (based on the same gene homology data produced in Chapter 3) and co-expression analysis (using statistical correlations and machine learning). This methodology was applied to the biosynthetic pathways for phytohormones and biopolymers discussed in Chapter 3, and was validated by the identification of genes known to be associated with these metabolic pathways. A number of new candidate genes associated with these metabolic pathways were identified. A variety of potential functional relationships to the metabolic pathways were uncovered which may help to direct research into the broader functional context of these metabolic processes.

5.2 CENTRAL METABOLISM IS CONSERVED ACROSS LAND PLANTS, BUT FLUX DISTRIBUTIONS ARE NOT NECESSARILY CONSERVED ACROSS THE CENTRAL METABOLIC NETWORK

The lack of any central metabolic pathways (i.e. those involved in carbon metabolism absolutely essential for plant growth and survival) identified as metabolic transitions in Chapter 3 indicates that central metabolism is highly conserved across land plants. Work by Chae et al. (2014) in which metabolic genes were compared across 16 land plants and

algae supports this finding, showing that most metabolic innovation since the emergence of land plants concerns secondary metabolism. Central metabolic processes in land plants are generally considered to be subject to purifying selection (Weng et al. 2012). Functioning to power normal growth and development, central metabolic enzymes have evolved towards higher catalytic efficiency and specificity than secondary metabolic enzymes (Bar-Even et al. 2011), which are conversely characterised by higher substrate promiscuity lending itself to greater evolvability and diversification following gene duplication (Weng 2014). Central metabolic enzymes show preferential loss of duplicate genes following local gene duplications compared to secondary metabolic enzymes (Moghe and Last 2015), which are more likely to retain additional copies (Chae et al. 2014). This is thought to be related to the requirement for precise gene product dosage in central metabolism (Freeling and Thomas 2006). Consequently, gene families involved in central metabolic pathways show a lower extent of gene number expansion than those involved in secondary metabolic pathways, which show high levels of diversity at the species level (Pott et al. 2019), rendering comparative analyses of these processes a productive tool for gaining evolutionary insights.

However, although the work presented in Chapter 3 supports existing evidence that the metabolic capabilities of central metabolism (i.e. the enzymes present) are conserved, my comparison of central metabolic models of *Arabidopsis* and *Marchantia* (Chapter 2) indicates that flux distributions across the central metabolic network are not necessarily maintained across land plants. The varied morphologies and physiologies across land plants dictate the accessibility of growth substrates as well as the nature of the metabolites required for biomass accumulation. For example, the thicker cell walls found in *Marchantia* compared to *Arabidopsis* result in lower rates of carbon assimilation in *Marchantia*, while the lack of a root system in *Marchantia* likely affects the absorption of nutrients. Furthermore, I identified various differences in biomass composition between *Marchantia* and *Arabidopsis* (Section 2.6.1), including lower lipid content with longer carbon chains and lower non-transitory starch content in *Marchantia*. These species differences necessitate varying flux through the central metabolic pathways which produce these compounds, although the central metabolic enzymes required remain the same. An example of this can be seen by comparing central metabolic fluxes in developing embryos of sunflower, soybean and rapeseed oilseeds, for

which metabolic flux analyses show widely varying flux distributions between species despite conserved central metabolic enzymes (Alonso et al. 2007; Goffman et al. 2005; Sriram et al. 2004). Sunflower seeds are heterotrophic and produce ATP to power the biosynthesis of fatty acids from triose phosphate, supported by high flux through the TCA cycle (Alonso et al. 2007). Similarly, soybean uses triose phosphate as its main oilseed carbon source. However the green soybean seeds and pods are photosynthetic and the majority of the energy required for fatty acid production is thought to be generated via photosynthesis light reactions and glycolysis with low flux through TCA cycle reactions (Alonso et al. 2007; Sriram et al. 2004). Finally, TCA cycle reactions in rapeseed embryos have been demonstrated to carry no net flux, with all ATP required generated via photosynthesis, and unusually utilise RuBisCO without the Calvin-Benson-Bassham cycle to recapture metabolic CO₂ to generate 3-phosphoglycerate as a substrate for fatty acid production (Goffman et al. 2005; Schwender et al. 2004). This results in an extremely high CCE of more than 80% in rapeseed embryos compared to 50% and 55% in sunflower and soybean embryos respectively (Alonso et al. 2007; Schwender et al. 2004).

I found that the model of *Marchantia* metabolism requires higher levels of CCE than that of *Arabidopsis* to meet experimental growth rates. The presented *Marchantia* model demonstrated that this requirement could be met using a CAM-like flux distribution. However, there are various other metabolic adaptations which could be responsible for improving carbon capture in this species. The majority of these adaptations, as found across the Chloroplastida, are similarly derived from the same core set of central metabolic enzymes, varying their regulation and localisation to beneficially influence flux distributions. For example, it has been shown that the organic acid metabolism required for both the day and night portions of the CAM cycle is already active in C₃ plants, albeit carrying much lower flux (Bräutigam et al. 2017). The evolution of CAM in a C₃ lineage therefore requires only an amplification of existing expression patterns in existing metabolic genes to modify flux through specific pathways, explaining the existence of species which exhibit varying levels of CAM along a continuum from C₃ to obligate CAM plants (Bräutigam et al. 2017; Silvera et al. 2010). Similarly, although requiring more dramatic regulatory and morphological modifications, both C₄ photosynthesis and the pyrenoid-based carbon

concentrating mechanism identified in hornworts are based on the same central metabolic enzymes present in C_3 species. The evolution of the former requires changes in the sites of expression of decarboxylating enzymes (NAD-ME, NADH-ME or PEPCK depending on the C_4 subtype) as well as RuBisCO (Heckmann et al. 2013), while the latter requires high a level of carbonic anhydrase expression (Hanson et al. 2014), each resulting in the release of CO_2 at high concentrations in the vicinity of RuBisCO to increase its rate of carboxylation.

5.3 CONSERVATION OF THE PRESENCE OF SECONDARY METABOLIC PATHWAYS DOES NOT NECESSARILY TRANSLATE TO FUNCTIONAL CONSERVATION

The presence of the same secondary metabolic pathway in two plants does not necessarily equate to homologous pathway function in both. For example, phytohormones such as abscisic acid and cytokinins, which are universal across land plants, have inevitably diversified in function as plants have evolved from algae to angiosperms, alongside the development of increasingly complex morphologies and physiologies (Rashotte 2020; Takezawa et al. 2011). My work in Chapter 3 may indicate a further example of this. I identified genes encoding the full gibberellin (GA) biosynthetic pathway in the genome of *Marchantia*, as well as in hornwort transcriptome data. This indicates that bryophytes possess the metabolic capacity to produce GAs. However, the lack of bioactive GA metabolites identified in bryophytes suggests that if they are present (presumably at low concentrations or only under specific conditions), they are unlikely to be as ubiquitously important in bryophytes as they are in higher plants (in angiosperms, GAs are involved in multiple functions including germination, response to abiotic and biotic stress, flowering and fruit set (Hernández-García et al. 2020)). This possible functional difference is supported by a lack of genes encoding the angiosperm-like transduction and signalling pathways for GAs in bryophytes (Blaacutierrez et al. 2020). While DELLA transcription factors (which control the downstream response to GAs in angiosperms) are present in bryophytes, homologues of the angiosperm GID1 receptor protein have not been identified in these species (Yoshida et al. 2018), and DELLA proteins in *Marchantia* and the moss *Physcomitrella patens* are considered unable to interact with angiosperm-like GID1 (Hernández-García et al. 2019). It is thought that DELLA proteins are ancestral in land plants, but were recruited for involvement in GA signalling

only in the common ancestor of all vascular plants (Hernández-García et al. 2019). Similar inconsistencies are present between bryophytes and angiosperms in the perception and signalling processes relating to additional phytohormones – bryophytes do not contain homologues of the angiosperm strigolactone receptor D14, and the canonical BRI1-dependent brassinosteroid (BR) perception module is also thought to have evolved after the divergence of bryophytes (Berens et al. 2017; Blaacutierrez et al. 2020). To further understand the roles of such metabolites and their mechanisms of use in bryophytes, it will be important to investigate whether and under which circumstances the relevant biosynthetic genes are expressed, and to identify the presence/absence of the full complement of known angiosperm-like signalling, crosstalk and regulatory interactions for these phytohormones in bryophytes. For GAs, such investigation may aid in the identification of a function for GA biosynthetic genes other than the biosynthesis of known bioactive GAs, and additionally may direct efforts to isolate bioactive GAs in bryophytes by identifying the conditions under which they are most likely to be produced. It is important to note that this established lack of conserved signalling, transduction and regulatory mechanisms between earlier- and later-diverging land plants suggest that the identification of candidate genes associated with phytohormone biosynthetic pathways in *Arabidopsis* in Chapter 4 cannot be assumed relevant to earlier-diverging plant lineages.

5.4 EXTENSIVE EXPERIMENTAL WORK IS REQUIRED TO IMPROVE OUR KNOWLEDGE OF EARLY-DIVERGING LAND PLANTS

As I report in Chapter 3, there are considerable differences in secondary metabolism between earlier- and later-diverging land plants. I identified differences in the metabolic pathways for BR and cutin biosynthesis between bryophytes and angiosperms, although all land plants are known to contain these compounds. Furthermore, my results from Chapter 2 indicate that although central metabolic pathways are conserved across land plants, their modes of function may differ between plant species. As described above, these differences in metabolic capacity are often mirrored in other important aspects of plant growth and functioning, for example in regulation, perception and transduction of metabolic signals and response to environmental stimuli (Berens et al. 2017; Blaacutierrez et al. 2020). Thorough investigation of metabolic

networks in early-diverging land plants is therefore required to bring our understanding of the involved processes up to the standard of angiosperm-based metabolic knowledge.

Such investigation is interesting for its own sake, but it may also prove exceedingly useful as a tool to facilitate metabolic research in angiosperms and crop species. It has long been suggested that *Marchantia* and other experimentally tractable bryophytes may have potential as testbeds for metabolic engineering strategies ultimately aimed for crop plants (Sweetlove et al. 2017). Research involving *Marchantia* benefits from its short generation time, easy growth under laboratory conditions and dominant haploid gametophyte (Ishizaki et al. 2016; Sauret-Güeto et al. 2020). Additionally, many gene families and regulatory modules which contain multiple redundant components in angiosperms are represented by conserved but more compact systems in *Marchantia*, enabling easier investigation of the results of genetic perturbations (Ishizaki et al. 2016; Sauret-Güeto et al. 2020). Metabolic engineering strategies could therefore be investigated and optimised in this more predictable model system before transferral to the target angiosperm. However, the results described in Chapter 2 suggest that we may not have a complete picture of the operation of central metabolic processes in *Marchantia*, and the findings of a metabolic engineering approach cannot be extrapolated between species if their wider metabolic context is not well understood.

Metabolic investigation of early-diverging land plants may also have commercial applications. For example, the lack of angiosperm-like genes encoding cutin and suberin biosynthesis in bryophytes and lycophytes identified in Chapter 3 indicates that alternative monomers may be required as precursors for these polymers in earlier-diverging land plants. Biopolymer composition is of interest for researchers developing biofuels, as recalcitrance to degradation is a major barrier to cost-effectiveness in these products (Himmel et al. 2007). An understanding of alternative compositions in early-diverging land plants may provide opportunities for the genetic engineering of biofuel crops containing biopolymers with alternative monomeric compositions, rendering them more amenable to breakdown. Furthermore, given the extent of secondary metabolic diversity identified across angiosperms, it seems likely that early-diverging land plants may constitute a valuable source of previously unknown metabolic processes and enzymes. These could provide novel

therapeutic metabolites as well as enzymes for the production of high-value metabolites or with commercially-relevant catalytic properties.

5.5 COMPUTATIONAL AND BIOINFORMATIC METABOLIC INVESTIGATIONS SUFFER FROM AN INHERENT BIAS TOWARDS ANGIOSPERMS

As a consequence of the discussed lack of understanding of early-diverging land plant metabolism, bioinformatic methods in plant science are intrinsically biased towards the investigation of angiosperms. Although advanced computational techniques are available for the prediction of gene function based on genome sequences and expression data (including machine learning based-approaches such as *k*-means clustering and random forests, and statistical approaches such as correlation and network analyses) (Mahood et al. 2020), these are limited by the scope of available training data, and traditional genome annotation methods assign homology based on comparison to known gene sequences. As the vast majority of plant gene data relates to angiosperms (particularly *Arabidopsis* and a select few crop species), there are limitations on the extent to which it is possible to annotate gene functions which are specific to earlier-diverging land plants and have not been identified in well-studied species (Rhee and Mutwil 2014). This means that bioinformatic identification of novel metabolic enzymes and pathways (which are likely to be abundant in the relatively understudied early-diverging land plants) is challenging. It is therefore important to prioritise the development of *de novo* bioinformatic approaches for the identification of metabolic genes (for example genome context analysis (Rhee and Mutwil 2014)) as well as experimental metabolic profiling-based methods in order to gain a better understanding of the metabolic capabilities of early-diverging land plants. Ultimately, the identification of novel gene function requires experimental support, as homology-based annotation inevitably leads to more of the same!

This angiosperm bias in genome annotation created a particular difficulty in Chapter 3, where despite the use of a machine-learning based annotation algorithm based on gene sequences from across plants, animals, bacteria and fungi (Schlöpfer et al. 2017), the identified metabolic transitions were biased toward metabolic innovations as opposed to losses. This is presumably because metabolic genes in early-diverging land plants for which homologues

were lost in later-diverging plants are unlikely to be annotated. Additionally, in the case of metabolic pathways which must be present in early-diverging land plants (because the end product is known to be present, e.g. the cutin biosynthetic pathway) I was only able to draw conclusions on the evolutionary processes leading to the 'angiosperm-like' metabolic pathway, as only genes homologous to those in angiosperms have been identified as relating to each biosynthetic process in earlier-diverging plants.

Similarly, our understanding of the physiological function of specific metabolites is often limited to those described in angiosperms. Gene sequence comparison predicting functional homology is often extrapolated to include predictions of subcellular compartment (Rhee and Mutwil 2014), although it is not unreasonable to consider that the localisation of functionally homologous genes may differ between plant taxa. The lack of specific information on compartmentation of metabolites in bryophytes meant that in Chapter 2 I was forced to assume that the subcellular localisation of metabolic processes in *Marchantia* was equivalent to that in *Arabidopsis*. Given the highly conserved nature of central metabolism across land plants, the majority of these assumptions are likely true. However, there are known differences in central metabolic compartmentation between plants. For example, it has been demonstrated that in addition to the universal plastidic AGPase (the enzyme responsible for the production of the starch precursor ADP-glucose) found across land plants, grasses possess a cytosolic isomer of AGPase as well as an ADP-glucose transporter which transports the ADP-glucose produced in the cytosol to the plastid for use in starch biosynthesis (Comparot-Moss and Denyer 2009). The cytosolic AGPase carries more flux in grasses than its plastidic counterpart, and is hypothesised to provide an energetic advantage resulting from a reduced need for ATP transport into the plastid (Comparot-Moss and Denyer 2009). The existence of this species-specific difference in compartmentation within angiosperms suggests that there is a possibility of further compartmentation differences between taxa, particularly those as divergent as angiosperms and bryophytes. Unfortunately, to my knowledge there is no comprehensive study on metabolic compartmentation in bryophytes. Beyond subcellular compartmentation, research concerning the molecular composition of biomass in bryophytes is sparse. Indeed, while for *Marchantia* enough information is available that I was able to produce a reasonably constrained model of central metabolism (although this still required

some *Arabidopsis*-based assumptions to be made regarding the comprised proteins and soluble metabolites), I attempted to formulate an additional model of central metabolism in the model moss *Physcomitrella patens* but was unable to find enough biomass and growth data to impose accurate constraints.

5.6 FURTHER LIMITATIONS ARE IMPOSED BY BIOINFORMATIC METHODS IN METABOLIC RESEARCH

The methods used in this thesis are computational, based on the manipulation of genome and transcriptome data to draw conclusions on evolutionary and metabolic processes. While such sequence data represents a crucial resource for plant research, there are limitations on the extent to which we can use it to draw inferences. First, as mentioned in Chapter 3, plant genomes are complex and are particularly difficult to assemble correctly due to the high incidence of repetitive sequence features (Claros et al. 2012). Although this is unlikely to affect inferences based on well-studied species such as *Arabidopsis*, it may cause problems with gene identification in other species – particularly gymnosperms, which have extremely large genomes rich in transposable elements (De La Torre et al. 2014). Genome assembly inaccuracies may therefore cause species to be identified as missing specific gene homologues which are in reality present in the genome. In Chapter 3, the metabolic pathways for BR and sporopollenin biosynthesis show a particularly scattered pattern of encoding gene presences/absences across gymnosperm species (Figures 3.4 and 3.6) – it is possible these inconsistencies stem from difficulties in gymnosperm genome assembly as opposed to truly missing genes. Second, sequenced transcriptomes are specific to the life stage, tissue type and environmental conditions under which the sequenced material was extracted. It is therefore likely that (particularly for secondary metabolic processes) genes encoding enzymes of interest may not be identified in a transcriptome analysis if the plant is not exposed to conditions under which the genes are expressed, leading to the recording of false absences in comparative studies. I used transcriptome data in Chapters 3 and 4 of this thesis and in both cases I have attempted to mitigate this problem by expanding the scope of the transcriptome data analysed to account for variations in gene expression between individual samples. For Chapter 3, I used transcriptome data from 305 species taken from the One

Thousand Plant Transcriptomes Initiative (Leebens-Mack et al. 2019) to support my results, ensuring that as many representatives of each plant taxon as possible were included (the minimum number of per-taxon representatives was 14, which was the total number of hornwort transcriptomes available at the time). For Chapter 4, I ensured that the *Arabidopsis* transcriptome data used for phylogenetic and co-expression analysis spanned several tissue types and *Arabidopsis* accessions (Lei et al. 2017).

The gene annotation data upon which the inclusion of metabolic pathways into my model of central metabolism in *Marchantia* in Chapter 2 is based is robust, given the extent of conservation of central metabolism across land plants. However, there are limitations to metabolic modelling using FBA. FBA is capable of quickly generating flux predictions across a large-scale metabolic network because it represents the entire network as a system of computationally tractable linear equations (Orth et al. 2010). This means FBA is unable to account for the non-linear regulatory processes that govern the expression of genes encoding metabolic enzymes. Nonetheless, FBA has been demonstrated to accurately predict network-wide flux distributions (Cheung et al. 2013; Williams et al. 2010), and it has been considered that since the role of regulation at the level of individual enzymes is to maintain a metabolic steady state (the presence of which is a key assumption of FBA) this regulatory function is in fact intrinsic to the method (Shameer et al. 2019). A second limitation is in the assumption of FBA that the modelled organism has evolved towards a singular goal, for example the optimal accumulation of biomass or maximal nitrogen use efficiency. These are represented in my model of *Marchantia* metabolism by the biomass objective function and the use of parsimonious FBA. However, plants in reality are subject to varying environmental conditions and have likely evolved to optimise a number of competing growth and metabolic efficiency goals at once depending on their surroundings, suggesting that maximising the accumulation of biomass is not always a reasonable approach for FBA models (Boyle et al. 2017). This is supported by the existence of metabolic phenomena such as futile cycles, which consume energy without promoting growth (Nguyen-Quoc and Foyer 2001). Additionally, different plant organs may be in different stages of growth at a single timepoint – for an actively growing euphyllophyte, individual leaves will reach their final size before the plant is fully grown, and thereafter function to provide metabolites to support growth in the rest

of the plant (Shameer et al. 2020). Regardless, the models in Chapter 2 aim to represent a growing *Marchantia* thallus and *Arabidopsis* leaf existing under laboratory conditions with minimal environmental stress. In this context, the optimisation of biomass production is a reasonable cellular objective.

5.7 ADDITIONAL COMPUTATIONAL AND EXPERIMENTAL WORK CAN ELUCIDATE FURTHER DETAILS OF EARLY-DIVERGING LAND PLANT METABOLISM

The accurate modelling of a species' metabolism using FBA benefits from the identification of species-specific metabolic constraints. Although my model of *Marchantia* incorporated *Marchantia*-specific constraints where possible, the lack of available information on biomass components and enzymatic parameters for *Marchantia* and other early-diverging land plants required angiosperm-based assumptions to be made regarding these constraints, and extensive analysis was undertaken to ensure that these assumptions did not significantly impact results (Section 2.6.2). Despite the minimal effect that most of these assumptions had on the flux predictions of the *Marchantia* model, the production of a model supported by data taken exclusively from *Marchantia* is ultimately necessary to draw robust conclusions, and to improve our understanding of *Marchantia* metabolism to the point that this plant is a realistic candidate as a testbed species. This will require metabolic profiling to investigate and quantify the individual components of biomass in *Marchantia* thallus tissue, and further experimental work to accurately measure NGAM in *Marchantia* as well as enzyme parameters such as the carboxylation: oxygenation ratio of *Marchantia* RuBisCO. Extension of metabolic profiling to the rhizoids and sporophyte of *Marchantia* would allow the relatively straightforward development of a multi-tissue model, which could provide further insights into whole-plant metabolism in bryophytes, for example regarding the nutritional dependency of the sporophyte on the gametophyte. Additionally, most liverworts including subspecies of *Marchantia polymorpha* form mycorrhizal symbioses with fungi (Shimamura 2016), which enhance nutrient uptake by the plant (Humphreys et al. 2010). This interaction is particularly interesting from an evolutionary perspective, as fungal symbioses have been hypothesised to represent a key adaptation allowing the colonisation of land by the ancestral embryophyte (de Vries and Archibald 2018). A comprehensive multi-tissue model of *Marchantia* metabolism

could be combined with a fungal model to study liverwort-fungal symbioses, identifying any differences compared to angiosperm-like mycorrhizal symbioses and possibly shedding light on the evolution of this trait.

Beyond the investigation of *Marchantia* as a model organism, the extent of available genome and transcriptome data for early-diverging land plants is continually and rapidly increasing. Indeed, since the 2020 acceptance of Chapter 3 for publication, three individual hornwort genomes have been sequenced (Li et al. 2020; Yang et al. 2020) (only transcriptomes were available for this taxon when the work for Chapter 3 was carried out). These provide additional material for the discovery of novel metabolic processes in early-diverging land plants using *de novo* identification methods, and can be further included in analyses such as those described in this thesis. The production of computational models of central metabolism across various species of early-diverging land plants would allow a detailed investigation of taxon-specific metabolic adaptations, such as the use of a pyrenoid-based carbon concentration mechanism in hornworts. Metabolic annotation of additional genomes would also allow for extended versions of the analysis carried out in Chapter 3, which with the inclusion of further species may be able to provide higher resolution results for a wider range of metabolic pathways. Furthermore, the method is not exclusive to the analysis of genes encoding metabolic enzymes, and could be extended to investigate signalling, transduction, transport and regulatory processes across land plants.

REFERENCES

- Alonso, A. P., Goffman, F. D., Ohlrogge, J. B., & Shachar-Hill, Y. (2007). Carbon conversion efficiency and central metabolic fluxes in developing sunflower (*Helianthus annuus* L.) embryos. *Plant Journal*, 52, 296–308.
- Bar-Even, A., Noor, E., Savir, Y., ... Milo, R. (2011). The moderately efficient enzyme: Evolutionary and physicochemical trends shaping enzyme parameters. *Biochemistry*, 50, 4402–4410.
- Berens, M. L., Berry, H. M., Mine, A., ... Tsuda, K. (2017). Evolution of hormone signaling networks in plant defense. *Annual review of phytopathology*, 55, 401–425.
- Blaacutierrez, M. A., Nelson, D. C., & Weijers, D. (2020). Evolution of plant hormone response pathways. *Annual Review of Plant Biology*, 71, 327–353.
- Boyle, N. R., Sengupta, N., & Morgan, J. A. (2017). Metabolic flux analysis of heterotrophic growth in *Chlamydomonas reinhardtii*. *PLOS ONE*, 12, e0177292.
- Bräutigam, A., Schlüter, U., Eisenhut, M., & Gowik, U. (2017). On the evolutionary origin of CAM photosynthesis. *Plant Physiology*, 174, 473–477.
- Chae, L., Kim, T., Nilo-Poyanco, R., & Rhee, S. Y. (2014). Genomic signatures of specialized metabolism in plants. *Science*, 344, 510–513.

- Cheung, C. Y. M., Williams, T. C. R., Poolman, M. G., ... Sweetlove, L. J. (2013). A method for accounting for maintenance costs in flux balance analysis improves the prediction of plant cell metabolic phenotypes under stress conditions. *Plant Journal*, 75, 1050–1061.
- Claros, M. G., Bautista, R., Guerrero-Fernández, D., ... Fernández-Pozo, N. (2012). Why assembling plant genome sequences is so challenging. *Biology*, 1, 439–459.
- Comparot-Moss, S., & Denyer, K. (2009). The evolution of the starch biosynthetic pathway in cereals and other grasses. *Journal of Experimental Botany*, 60, 2481–2492.
- De La Torre, A. R., Birol, I., Bousquet, J., ... Bohlmann, J. (2014). Insights into conifer giga-genomes. *Plant Physiology*, 166, 1724–1732.
- de Vries, J., & Archibald, J. M. (2018). Plant evolution: Landmarks on the path to terrestrial life. *New Phytologist*, 217, 1428–1434.
- Flexas, J., & Carriqui, M. (2020). Photosynthesis and photosynthetic efficiencies along the terrestrial plant's phylogeny: Lessons for improving crop photosynthesis. *The Plant Journal*, 101, 964–978.
- Freeling, M., & Thomas, B. C. (2006). Gene-balanced duplications, like tetraploidy, provide predictable drive to increase morphological complexity. *Genome Research*, 16, 805–814.
- Gago, J., Carriqui, M., Nadal, M., ... Flexas, J. (2019). Photosynthesis optimized across land plant phylogeny. *Trends in Plant Science*, 24, 947–958.
- Goffman, F. D., Alonso, A. P., Schwender, J., ... Ohlrogge, J. B. (2005). Light enables a very high efficiency of carbon storage in developing embryos of rapeseed. *Plant Physiology*, 138, 2269–2279.
- Hanson, D. T., Renzaglia, K., & Villarreal, J. C. (2014). Diffusion limitation and CO₂ concentrating mechanisms in bryophytes. *Photosynthesis in bryophytes and early land plants* (pp. 95–111). Springer, Dordrecht.
- Heckmann, D., Schulze, S., Denton, A., ... Lercher, M. J. (2013). Predicting C₄ photosynthesis evolution: Modular, individually adaptive steps on a Mount Fuji fitness landscape. *Cell*, 153, 1579.
- Hernández-García, J., Briones-Moreno, A., & Blázquez, M. A. (2020). Origin and evolution of gibberellin signaling and metabolism in plants. *Seminars in Cell and Developmental Biology*, 109, 46–54.
- Hernández-García, J., Briones-Moreno, A., Dumas, R., & Blázquez, M. A. (2019). Origin of gibberellin-dependent transcriptional regulation by molecular exploitation of a transactivation domain in DELLA proteins. *Molecular Biology and Evolution*, 36, 908–918.
- Himmel, M. E., Ding, S. Y., Johnson, D. K., ... Foust, T. D. (2007). Biomass recalcitrance: Engineering plants and enzymes for biofuels production. *Science*, 315, 804–807.
- Humphreys, C. P., Franks, P. J., Rees, M., ... Beerling, D. J. (2010). Mutualistic mycorrhiza-like symbiosis in the most ancient group of land plants. *Nature Communications*, 1, 1–7.
- Ishizaki, K., Nishihama, R., Yamato, K. T., & Kohchi, T. (2016). Molecular genetic tools and techniques for *Marchantia polymorpha* research. *Plant and Cell Physiology*, 57, 262–270.
- Leebens-Mack, J. H., Barker, M. S., Carpenter, E. J., ... Wong, G. K. S. (2019). One thousand plant transcriptomes and the phylogenomics of green plants. *Nature*, 574, 679–685.
- Lei, L., Steffen, J. G., Osborne, E. J., & Toomajian, C. (2017). Plant organ evolution revealed by phylotranscriptomics in *Arabidopsis thaliana*. *Scientific Reports*, 7.
- Li, F. W., Nishiyama, T., Waller, M., ... Szövényi, P. (2020). *Anthoceros* genomes illuminate the origin of land plants and the unique biology of hornworts. *Nature Plants*, 6, 259–272.
- Mahood, E. H., Kruse, L. H., & Moghe, G. D. (2020). Machine learning: A powerful tool for gene function prediction in plants. *Applications in Plant Sciences*, 8, e11376.

- Moghe, G. D., & Last, R. L. (2015). Something old, something new: Conserved enzymes and the evolution of novelty in plant specialized metabolism. *Plant Physiology*, 169, 1512–1523.
- Nguyen-Quoc, B., & Foyer, C. H. (2001). A role for ‘futile cycles’ involving invertase and sucrose synthase in sucrose metabolism of tomato fruit. *Journal of Experimental Botany*, 52, 881–889.
- Orth, J. D., Thiele, I., & Palsson, B. O. (2010). What is flux balance analysis? *Nature Biotechnology*, 28, 245–248.
- Pott, D. M., Osorio, S., & Vallarino, J. G. (2019). From central to specialized metabolism: An overview of some secondary compounds derived from the primary metabolism for their role in conferring nutritional and organoleptic characteristics to fruit. *Frontiers in Plant Science*, 10, 835.
- Rashotte, A. M. (2020). The evolution of cytokinin signaling and its role in development before angiosperms. *Seminars in Cell and Developmental Biology*, 109, 31–38.
- Rhee, S. Y., & Mutwil, M. (2014). Towards revealing the functions of all genes in plants. *Trends in Plant Science*, 19, 212–221.
- Sage, R. F., Christin, P., & Edwards, E. J. (2011). The C₄ plant lineages of planet earth. *Journal of Experimental Botany*, 62, 3155–69.
- Sauret-Güeto, S., Frangedakis, E., Silvestri, L., ... Haseloff, J. (2020). Systematic tools for reprogramming plant gene expression in a simple model, *Marchantia polymorpha*. *BioRxiv*.
- Scheurwater, I., Cornelissen, C., Dictus, F., ... Lambers, H. (1998). Why do fast- and slow-growing grass species differ so little in their rate of root respiration, considering the large differences in rate of growth and ion uptake? *Plant, Cell and Environment*, 21, 995–1005.
- Schläpfer, P., Zhang, P., Wang, C., ... Rhee, S. Y. (2017). Genome-wide prediction of metabolic enzymes, pathways, and gene clusters in plants. *Plant Physiology*, 173, 2041–2059.
- Schwender, J., Goffman, F., Ohlrogge, J. B., & Shachar-Hill, Y. (2004). RuBisCO without the Calvin Cycle improves the carbon efficiency of developing green seeds. *Nature*, 432, 779–782.
- Shameer, S., Ratcliffe, R. G., & Sweetlove, L. J. (2019). Leaf energy balance requires mitochondrial respiration and export of chloroplast NADPH in the light. *Plant Physiology*, 180, 1947–1961.
- Shameer, S., Vallarino, J. G., Fernie, A. R., ... Sweetlove, L. J. (2020). Flux balance analysis of metabolism during growth by osmotic cell expansion and its application to tomato fruits. *The Plant Journal*, 103, 68–82.
- Shimamura, M. (2016). *Marchantia polymorpha*: Taxonomy, phylogeny and morphology of a model system. *Plant and Cell Physiology*, 57, 230–256.
- Silvera, K., Neubig, K. M., Whitten, W. M., ... Cushman, J. C. (2010). Evolution along the crassulacean acid metabolism continuum. *Functional Plant Biology*, 37, 995–1010.
- Sriram, G., Fulton, D. B., Iyer, V. V., ... Shanks, J. V. (2004). Quantification of compartmented metabolic fluxes in developing soybean embryos by employing biosynthetically directed fractional ¹³C labeling, two-dimensional [¹³C, ¹H] nuclear magnetic resonance, and comprehensive isotopomer balancing. *Plant Physiology*, 136, 3043–3057.
- Sweetlove, L. J., Nielsen, J., & Fernie, A. R. (2017). Engineering central metabolism - a grand challenge for plant biologists. *The Plant Journal*, 90, 749–763.
- Takezawa, D., Komatsu, K., & Sakata, Y. (2011). ABA in bryophytes: How a universal growth regulator in life became a plant hormone? *Journal of Plant Research*, 124, 437–453.
- Villarreal, J. C., & Renner, S. S. (2012). Hornwort pyrenoids, carbon-concentrating structures, evolved and were lost at least five times during the last 100 million years. *Proceedings of the National Academy of Sciences of the United States of America*, 109, 18873–18878.
- Weng, J. (2014). The evolutionary paths towards complexity: A metabolic perspective. *The New Phytologist*, 201, 1141–9.

- Weng, J. K., Philippe, R. N., & Noel, J. P. (2012). The rise of chemodiversity in plants. *Science*, 336, 1667–1670.
- Williams, T. C. R., Poolman, M. G., Howden, A. J. M., ... Sweetlove, L. J. (2010). A genome-scale metabolic model accurately predicts fluxes in central carbon metabolism under stress conditions. *Plant Physiology*, 154, 311–323.
- Yang, Y., Sun, P., Lv, L., ... Liu, J. (2020). Prickly waterlily and rigid hornwort genomes shed light on early angiosperm evolution. *Nature Plants*, 6, 215–222.
- Yoshida, H., Tanimoto, E., Hirai, T., ... Ueguchi-Tanaka, M. (2018). Evolution and diversification of the plant gibberellin receptor GID1. *Proceedings of the National Academy of Sciences of the United States of America*, 115(33), E7844–E7853.

Bibliography

- Alonso, A. P., Goffman, F. D., Ohlrogge, J. B., & Shachar-Hill, Y. (2007). Carbon conversion efficiency and central metabolic fluxes in developing sunflower (*Helianthus annuus* L.) embryos. *Plant Journal*, 52, 296–308.
- Amthor, J. S. (1984). The role of maintenance respiration in plant growth. *Plant, Cell and Environment*, 7, 561–569.
- Amthor, J. S. (2000). The McCree-de Wit-Penning de Vries-Thornley respiration paradigms: 30 years later. *Annals of Botany*, 86, 1–20.
- Amthor, J. S., Bar-Even, A., Hanson, A. D., ... Tyerman, S. D. (2019). Engineering strategies to boost crop productivity by cutting respiratory carbon loss. *The Plant Cell*, 31, 297–314.
- Armenteros, J. J. A., Salvatore, M., Emanuelsson, O., ... Nielsen, H. (2019). Detecting sequence signals in targeting peptides using deep learning. *Life Science Alliance*, 2, e201900429.
- Arnold, A., & Nikoloski, Z. (2014). Bottom-up metabolic reconstruction of *Arabidopsis* and its application to determining the metabolic costs of enzyme production. *Plant Physiology*, 165, 1380–1391.
- Ashihara, H., Sano, H., & Crozier, A. (2008). Caffeine and related purine alkaloids: Biosynthesis, catabolism, function and genetic engineering. *Phytochemistry*, 69, 841–856.
- Bailey, M., Ivanauskaite, A., Grimmer, J., ... Gibbs, D. (2020). The *Arabidopsis* NOT4A E3 ligase coordinates PGR3 expression to regulate chloroplast protein translation. *bioRxiv*.
- Bajguz, A. (2009). Isolation and characterization of brassinosteroids from algal cultures of *Chlorella vulgaris* Beijerinck (Trebouxioophyceae). *Journal of Plant Physiology*, 166, 1946–1949.
- Bajguz, A., Chmur, M., & Gruszka, D. (2020). Comprehensive overview of the brassinosteroid biosynthesis pathways: Substrates, products, inhibitors, and connections. *Frontiers in Plant Science*, 11, 1034.
- Bakan, B., & Marion, D. (2017). Assembly of the cutin polyester: From cells to extracellular cell walls. *Plants*, 6, 57.
- Banks, J. A., Nishiyama, T., Hasebe, M., ... Grigoriev, I. V. (2011). The *Selaginella* genome identifies genetic changes associated with the evolution of vascular plants. *Science*, 332, 960–963.
- Bar-Even, A., Noor, E., Savir, Y., ... Milo, R. (2011). The moderately efficient enzyme: Evolutionary and physicochemical trends shaping enzyme parameters. *Biochemistry*, 50, 4402–4410.
- Basler, G., Fernie, A. R., & Nikoloski, Z. (2018). Advances in metabolic flux analysis toward genome-scale profiling of higher organisms. *Bioscience Reports*, 38(6), BSR20170224.
- Bateman, A. (2019). Uniprot: A worldwide hub of protein knowledge. *Nucleic Acids Research*, 47, D506–D515.
- Bauwe, H., Hagemann, M., Kern, R., & Timm, S. (2012). Photorespiration has a dual origin and manifold links to central metabolism. *Current Opinion in Plant Biology*, 15, 269–275.
- Benson, D. A., Cavanaugh, M. A., Clark, K., ... Sayers, E. W. (2017). GenBank. *Nucleic Acids Research*, 45, D37–D42.
- Berens, M. L., Berry, H. M., Mine, A., ... Tsuda, K. (2017). Evolution of hormone signaling networks in plant defense. *Annual review of phytopathology*, 55, 401–425.

- Berghuijs, H. N. C., Yin, X., Ho, Q. T., ... Struik, P. C. (2017). Localization of (photo)respiration and CO₂ re-assimilation in tomato leaves investigated with a reaction-diffusion model. *PLOS ONE*, 12, e0183746.
- Bhunia, R. K., Showman, L. J., Jose, A., & Nikolau, B. J. (2018). Combined use of cutinase and high-resolution mass-spectrometry to query the molecular architecture of cutin. *Plant Methods*, 14, 1–17.
- Blaacutierrez, M. A., Nelson, D. C., & Weijers, D. (2020). Evolution of plant hormone response pathways. *Annual Review of Plant Biology*, 71, 327–353.
- Böhmer, M., & Schroeder, J. I. (2011). Quantitative transcriptomic analysis of abscisic acid-induced and reactive oxygen species-dependent expression changes and proteomic profiling in *Arabidopsis* suspension cells. *Plant Journal*, 67, 105–118.
- Bone, R. E., Smith, J. A. C., Arrigo, N., & Buerki, S. (2015). A macro-ecological perspective on crassulacean acid metabolism (CAM) photosynthesis evolution in afro-madagascan drylands: Eulophiinae orchids as a case study. *New Phytologist*, 208, 469–481.
- Bonhomme, S., & Waters, M. (2019). Evolution of strigolactone biosynthesis and signalling. *Strigolactones - biology and applications* (pp. 143–161). Springer International Publishing.
- Boot, K. J. M., Libbenga, K. R., Hille, S. C., ... Van Duijn, B. (2012). Polar auxin transport: An early invention. *Journal of Experimental Botany*, 63, 4213–4218.
- Bowman, J. L., Araki, T., & Kohchi, T. (2016). Marchantia: Past, present and future. *Plant and Cell Physiology*, 57, 205–209.
- Bowman, J. L., Briginshaw, L. N., Fisher, T. J., & Flores-Sandoval, E. (2019). Something ancient and something neofunctionalized - evolution of land plant hormone signaling pathways. *Current Opinion in Plant Biology*, 47, 64–72.
- Bowman, J. L., Kohchi, T., Yamato, K. T., ... Schmutz, J. (2017). Insights into land plant evolution garnered from the *Marchantia polymorpha* genome. *Cell*, 171, 287–304.e15.
- Boycheva, S., Daviet, L., Wolfender, J. L., & Fitzpatrick, T. B. (2014). The rise of operon-like gene clusters in plants. *Trends in Plant Science*, 19, 447–459.
- Boyle, N. R., Sengupta, N., & Morgan, J. A. (2017). Metabolic flux analysis of heterotrophic growth in *Chlamydomonas reinhardtii*. *PLOS ONE*, 12, e0177292.
- Bräutigam, A., Schlüter, U., Eisenhut, M., & Gowik, U. (2017). On the evolutionary origin of CAM photosynthesis. *Plant Physiology*, 174, 473–477.
- Breiman, L. (2001). Random forests. *Machine Learning*, 45, 5–32.
- Briones-Moreno, A., Hernández-García, J., Vargas-Chávez, C., ... Blázquez, M. A. (2017). Evolutionary analysis of DELLA-associated transcriptional networks. *Frontiers in Plant Science*, 8, 626.
- Brunetti, C., Sebastiani, F., & Tattini, M. (2019). Review: ABA, flavonols, and the evolvability of land plants. *Plant Science*, 280, 448–454.
- Buchfink, B., Xie, C., & Huson, D. H. (2014). Fast and sensitive protein alignment using DIAMOND. *Nature Methods*, 12, 59–60.
- Busch, F. A., Sage, T. L., Cousins, A. B., & Sage, R. F. (2013). C₃ plants enhance rates of photosynthesis by reassimilating photorespired and respired CO₂. *Plant, Cell and Environment*, 36, 200–212.
- Bythell-Douglas, R., Rothfels, C. J., Stevenson, D. W. D., ... Bennett, T. (2017). Evolution of strigolactone receptors by gradual neo-functionalization of KAI2 paralogues. *BMC Biology*, 15.
- Cai, L., Zhang, L., Fu, Q., & Xu, Z. F. (2018). Identification and expression analysis of cytokinin metabolic genes IPTs, CYP735A and CKXs in the biofuel plant *Jatropha curcas*. *PeerJ*, 16(6), e4812.

- Cai, W., Okuda, K., Peng, L., & Shikanai, T. (2011). PROTON GRADIENT REGULATION 3 recognizes multiple targets with limited similarity and mediates translation and rna stabilization in plastids. *The Plant Journal*, 67, 318–327.
- Caldicott, A. B., & Eglinton, G. (1976). Cutin acids from bryophytes: An ω -1 hydroxy alkanolic acid in two liverwort species. *Phytochemistry*, 15, 1139–1143.
- Cannell, N., Emms, D. M., Hetherington, A. J., ... Sweetlove, L. J. (2020). Multiple metabolic innovations and losses are associated with major transitions in land plant evolution. *Current Biology*, 30, 1783–1800.e11.
- Capella-Gutiérrez, S., Silla-Martínez, J. M., & Gabaldón, T. (2009). TrimAl: A tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics*, 25, 1972–1973.
- Carella, P., Gogleva, A., Hoey, D. J., ... Schornack, S. (2019). Conserved biochemical defenses underpin host responses to oomycete infection in an early-divergent land plant lineage. *Current Biology*, 29, 2282–2294.e5.
- Carriquí, M., Roig-Oliver, M., Brodribb, T. J., ... Flexas, J. (2019). Anatomical constraints to nonstomatal diffusion conductance and photosynthesis in lycophytes and bryophytes. *New Phytologist*, 222, 1256–1270.
- Carter, J. L., & Romine, K. G. (1969). The effects of long and short photoperiods on the rate of growth and gemmae cup production in *Marchantia polymorpha* L. *Transactions of the Kansas Academy of Science (1903-)*, 72, 98.
- Caspi, R., R. and Billington, Fulcher, C. A., Keseler, I. M., ... Karp, P. D. (2018). The MetaCyc database of metabolic pathways and enzymes. *Nucleic acids research*, 46, D633–D639.
- Chae, L., Kim, T., Nilo-Poyanco, R., & Rhee, S. Y. (2014). Genomic signatures of specialized metabolism in plants. *Science*, 344, 510–513.
- Chan, K. X., Mabbitt, P. D., Phua, S. Y., ... Pogson, B. J. (2016). Sensing and signaling of oxidative stress in chloroplasts by inactivation of the SAL1 phosphoadenosine phosphatase. *Proceedings of the National Academy of Sciences of the United States of America*, 113, E4567–E4576.
- Chanderbali, A. S., Berger, B. A., Howarth, D. G., ... Soltis, D. E. (2016). Evolving ideas on the origin and evolution of flowers: New perspectives in the genomic era. *Genetics*, 202, 1255–1265.
- Chang, C., Kong, L., Liu, Y., ... Gong, Z. (2020). Origins and evolution of cuticle biosynthetic machinery in land plants. *Plant Physiology*, 184, 1998–2010.
- Chen, L., Bernhardt, A., Lee, J., & Hellmann, H. (2015). Identification of *Arabidopsis* MYB56 as a novel substrate for CRL3BPM E3 ligases. *Molecular Plant*, 8, 242–250.
- Chen, L., Xiang, S., Chen, Y., ... Yu, D. (2017). *Arabidopsis* wrky45 interacts with the DELLA protein RGL1 to positively regulate age-triggered leaf senescence. *Molecular Plant*, 10, 1174–1189.
- Chen, M., Maodzeka, A., Zhou, L., ... Jiang, L. (2014). Removal of DELLA repression promotes leaf senescence in *Arabidopsis*. *Plant Science*, 219–220, 26–34.
- Cheng, S., Xian, W., Fu, Y., ... Melkonian, G. K. (2019). Genomes of subaerial Zygnematophyceae provide insights into land plant evolution. *Cell*, 179, 1057–1067.e14.
- Cheng, S., Zhu, X., Liao, T., ... Kang, X. (2015). Gene expression differences between high-growth *Populus* allotriploids and their diploid parents. *Forests*, 6, 839–857.
- Cheon, J., Fujioka, S., Dilkes, B. P., & Choe, S. (2013). Brassinosteroids regulate plant growth through distinct signaling pathways in *Selaginella* and *Arabidopsis*. *PLoS ONE*, 8.
- Cheung, C. Y. M., Poolman, M. G., Fell, D. A., ... Sweetlove, L. J. (2014). A diel flux balance model captures interactions between light and dark metabolism during day-night cycles in C_3 and crassulacean acid metabolism leaves. *Plant Physiology*, 165, 917–929.
- Cheung, C. Y. M., Williams, T. C. R., Poolman, M. G., ... Sweetlove, L. J. (2013). A method for accounting for maintenance costs in flux balance analysis improves the prediction of plant cell metabolic phenotypes under stress conditions. *Plant Journal*, 75, 1050–1061.

- Christophe, A., Letort, V., Hummel, I., ... Lecoecur, J. (2008). A model-based analysis of the dynamics of carbon balance at the whole-plant level in *Arabidopsis thaliana*. *Functional Plant Biology*, 35, 1147.
- Claeys, H., De Bodt, S., & Inzé, D. (2014). Gibberellins and DELLAs: Central nodes in growth regulatory networks. *Trends in Plant Science*, 19, 231–239.
- Claros, M. G., Bautista, R., Guerrero-Fernández, D., ... Fernández-Pozo, N. (2012). Why assembling plant genome sequences is so challenging. *Biology*, 1, 439–459.
- Clausen, M., Kannangara, R. M., Olsen, C. E., ... Møller, B. L. (2015). The bifurcation of the cyanogenic glucoside and glucosinolate biosynthetic pathways. *The Plant journal*, 84, 558–73.
- Cohen, H., Fedruk, V., Wang, C., ... Aharoni, A. (2020). SUBERMAN regulates developmental suberization of the *Arabidopsis* root endodermis. *The Plant Journal*, 102, 431–447.
- Collakova, E., Yen, J. Y., & Senger, R. S. (2012). Are we ready for genome-scale modeling in plants? *Plant Science*, 191–192, 53–70.
- Comparot-Moss, S., & Denyer, K. (2009). The evolution of the starch biosynthetic pathway in cereals and other grasses. *Journal of Experimental Botany*, 60, 2481–2492.
- Cookson, S. J., Radziejewski, A., & Granier, C. (2006). Cell and leaf size plasticity in *Arabidopsis*: What is the role of endoreduplication? *Plant, Cell and Environment*, 29, 1273–1283.
- Cookson, S. J., Van Lijsebettens, M., & Granier, C. (2005). Correlation between leaf growth variables suggest intrinsic and early controls of leaf size in *Arabidopsis thaliana*. *Plant, Cell and Environment*, 28, 1355–1366.
- Creighton, M. T., Kolton, A., Kataya, A. R. A., ... Lillo, C. (2017). Methylation of protein phosphatase 2A-influence of regulators and environmental stress factors. *Plant, Cell and Environment*, 40, 2347–2358.
- Czechowski, T., Bari, R. P., Stitt, M., ... Udvardi, M. K. (2004). Real-time RT-PCR profiling of over 1400 *Arabidopsis* transcription factors: Unprecedented sensitivity reveals novel root-and shoot-specific genes. *Plant Journal*, 38, 366–379.
- Davies, K. M., Albert, N. W., Zhou, Y., & Schwinn, K. E. (2018). Functions of flavonoid and betalain pigments in abiotic stress tolerance in plants. *Annual plant reviews* (pp. 1–41).
- De Giorgi, J., Piskurewicz, U., Loubéry, S., ... Lopez-Molina, L. (2015). An endosperm-associated cuticle is required for *Arabidopsis* seed viability, dormancy and early control of germination. *PLoS Genetics*, 11.
- De La Torre, A. R., Birol, I., Bousquet, J., ... Bohlmann, J. (2014). Insights into conifer giga-genomes. *Plant Physiology*, 166, 1724–1732.
- Delaux, P. M., Radhakrishnan, G. V., Jayaraman, D., ... Ane, J. M. (2015). Algal ancestor of land plants was preadapted for symbiosis. *Proceedings of the National Academy of Sciences of the United States of America*, 112, 13390–13395.
- Delaux, P. M., Xie, X., Timme, R. E., ... Séjalon-Delmas, N. (2012). Origin of strigolactones in the green lineage. *New Phytologist*, 195, 857–871.
- Delhon, P., Gojon, A., Tillard, P., & Passama, L. (1995). Diurnal regulation of NO₃⁻ uptake in soybean plants I. Changes in NO₃⁻ influx, efflux, and N utilization in the plant during the day/night cycle. *Journal of Experimental Botany*, 46, 1585–1594.
- Delwiche, C. F., & Cooper, E. D. (2015). The evolutionary origin of a terrestrial flora. *Current Biology*, 25, R899–R910.
- Delwiche, C. F., Graham, L. E., & Thomson, N. (1989). Lignin-like compounds and sporopollenin in *Coleochaete*, an algal model for land plant ancestry. *Science*, 245, 399–401.
- Dembitsky, V. M., Rezanka, T., Bychek, I. A., & Afonina, O. M. (1993). Polar lipid and fatty acid composition of some bryophytes. *Phytochemistry*, 33, 1009–1014.

- de Vries, J., & Archibald, J. M. (2018). Plant evolution: Landmarks on the path to terrestrial life. *New Phytologist*, 217, 1428–1434.
- de Vries, J., de Vries, S., Slamovits, C. H., ... Archibald, J. M. (2017). How embryophytic is the biosynthesis of phenylpropanoids and their derivatives in streptophyte algae? *Plant and Cell Physiology*, 58, 934–945.
- de Vries, J., Stanton, A., Archibald, J. M., & Gould, S. B. (2016). Streptophyte terrestrialization in light of plastid evolution. *Trends in Plant Science*, 21, 467–476.
- Di Rubbo, S., Irani, N. G., Kim, S. Y., ... Russinova, E. (2013). The clathrin adaptor complex AP-2 mediates endocytosis of BRASSINOSTEROID INSENSITIVE1 in *Arabidopsis* W. *The Plant Cell*, 25, 2986–2997.
- DiMario, R. J., Clayton, H., Mukherjee, A., ... Moroney, J. V. (2017). Plant carbonic anhydrases: Structures, locations, evolution, and physiological roles. *Molecular Plant*, 10, 30–46.
- Dittami, S. M., Heesch, S., Olsen, J. L., & Collén, J. (2017). Transitions between marine and freshwater environments provide new clues about the origins of multicellular plants and algae. *Journal of Phycology*, 53, 731–745.
- Dominguez, E., Heredia-Guerrero, J. A., & Heredia, A. (2011). The biophysical design of plant cuticles: An overview. *New Phytologist*, 189, 938–949.
- Domínguez, E., Heredia-Guerrero, J. A., & Heredia, A. (2015). Plant cutin genesis: Unanswered questions. *Trends in Plant Science*, 20, 551–8.
- Dreze, M., Carvunis, A. R., Charlotiaux, B., ... Yazaki, J. (2011). Evidence for network evolution in an *Arabidopsis* interactome map. *Science*, 333, 601–607.
- Emms, D. M., & Kelly, S. (2019). Orthofinder: Phylogenetic orthology inference for comparative genomics. *Genome Biology*, 20, 238.
- Erb, T. J., & Zarzycki, J. (2018). A short history of RuBisCO: The rise and fall (?) of Nature's predominant CO₂ fixing enzyme. *Current Opinion in Biotechnology*, 49, 100–107.
- Exposito-Alonso, M., Burbano, H. A., & Bossdorf, O. (2019). Natural selection on the *Arabidopsis thaliana* genome in present and future climates. *Nature*, 573, 126–129.
- Fang, C., Fernie, A. R., & Luo, J. (2019). Exploring the diversity of plant metabolism. *Trends in Plant Science*, 24, 83–98.
- Fasoli, M., Dell'Anna, R., Dal Santo, S., ... Zenoni, S. (2016). Pectins, hemicelluloses and celluloses show specific dynamics in the internal and external surfaces of grape berry skin during ripening. *Plant and Cell Physiology*, 57, 1332–1349.
- Fellenberg, C., Milkowski, C., Hause, B., ... Vogt, T. (2008). Tapetum-specific location of a cation-dependent O-methyltransferase in *Arabidopsis thaliana*. *The Plant Journal*, 56, 132–145.
- Fernández, V., Guzmán-Delgado, P., Graça, J., ... Gil, L. (2016). Cuticle structure in relation to chemical composition: Re-assessing the prevailing model. *Frontiers in Plant Science*, 7.
- Fernie, A. R., & Bauwe, H. (2020). Wasteful, essential, evolutionary stepping stone? The multiple personalities of the photorespiratory pathway. *The Plant Journal*, 102, 666–677.
- Fich, E. A., Segerson, N. A., & Rose, J. K. C. (2016). The plant polyester cutin: Biosynthesis, structure, and biological roles. *Annual Review of Plant Biology*, 67, 207–233.
- Field, B., & Osbourn, A. E. (2008). Metabolic diversification - independent assembly of operon-like gene clusters in different plants. *Science*, 320, 543–547.
- Fischer, W. W., Hemp, J., & Johnson, J. E. (2016). Evolution of oxygenic photosynthesis. *Annual Review of Earth and Planetary Sciences*, 44, 647–683.
- Fleet, C. M., & Sun, T. P. (2005). A DELLAcate balance: The role of gibberellin in plant morphogenesis. *Current Opinion in Plant Biology*, 8, 77–85.

- Fletcher, B. J., Brentnall, S. J., Quick, W. P., & Beerling, D. J. (2006). BRYOCARB: A process-based model of thallose liverwort carbon isotope fractionation in response to CO₂, O₂, light and temperature. *Geochimica et Cosmochimica Acta*, 70, 5676–5691.
- Flexas, J., & Carriqui, M. (2020). Photosynthesis and photosynthetic efficiencies along the terrestrial plant's phylogeny: Lessons for improving crop photosynthesis. *The Plant Journal*, 101, 964–978.
- Freeling, M., & Thomas, B. C. (2006). Gene-balanced duplications, like tetraploidy, provide predictable drive to increase morphological complexity. *Genome Research*, 16, 805–814.
- Friedrichsen, D. M., Nemhauser, J., Muramitsu, T., ... Chory, J. (2002). Three redundant brassinosteroid early response genes encode putative bHLH transcription factors required for normal growth. *Genetics*, 162, 1445–1456.
- Fujii, S., Sato, N., & Shikanai, T. (2013). Mutagenesis of individual pentatricopeptide repeat motifs affects RNA binding activity and reveals functional partitioning of *Arabidopsis* PROTON GRADIENT REGULATION3. *The Plant Cell*, 25, 3079–3088.
- Gachon, C. M. M., Langlois-Meurinne, M., Henry, Y., & Saindrenan, P. (2005). Transcriptional co-regulation of secondary metabolism enzymes in *Arabidopsis*: Functional and evolutionary implications. *Plant Molecular Biology*, 58, 229–245.
- Gago, J., Carriqui, M., Nadal, M., ... Flexas, J. (2019). Photosynthesis optimized across land plant phylogeny. *Trends in Plant Science*, 24, 947–958.
- Gao, H., Lu, Z. M., Yang, Y., ... Cao, W. (2018). Melatonin treatment reduces chilling injury in peach fruit through its regulation of membrane fatty acid contents and phenolic metabolism. *Food Chemistry*, 245, 659–666.
- Gavelis, G. S., & Gile, G. H. (2018). How did cyanobacteria first embark on the path to becoming plastids?: Lessons from protist symbioses. *FEMS Microbiology Letters*, 365, 209.
- Geilen, K., Heilmann, M., Hillmer, S., & Böhmer. (2017). WRKY43 regulates polyunsaturated fatty acid content and seed germination under unfavourable growth conditions. *Nature Scientific Reports*, 7, 14235.
- Genuardo, G., Errico, A., Tiezzi, A., & Conicella, C. (1998). Tubulin and F-actin distribution during microsporogenesis in a 2n pollen producer of *Solanum*. *Genome*, 41(5).
- Gholizadeh, A. (2011). Heterologous expression of stress-responsive DUF538 domain containing protein and its morpho-biochemical consequences. *Protein Journal*, 30, 351–358.
- Gholizadeh, A. (2018). Chlorophyll binding ability of non-chloroplastic DUF538 protein superfamily in plants. *Proceedings of the National Academy of Sciences India Section B - Biological Sciences*, 88, 967–976.
- Gholizadeh, A. (2020). Pectin methylesterase activity of plant DUF538 protein superfamily. *Physiology and Molecular Biology of Plants*, 26(4), 829–839.
- Gibon, Y., Bläsing, O. E., Palacios-Rojas, N., ... Stitt, M. (2004). Adjustment of diurnal starch turnover to short days: Depletion of sugar during the night leads to a temporary inhibition of carbohydrate utilization, accumulation of sugars and post-translational activation of ADP-glucose pyrophosphorylase in the following light period. *Plant Journal*, 39, 847–862.
- Gimeno, T. E., Ogée, J., Royles, J., ... Wingate, L. (2017). Bryophyte gas-exchange dynamics along varying hydration status reveal a significant carbonyl sulphide (COS) sink in the dark and COS source in the light. *New Phytologist*, 215, 965–976.
- Gocal, G. F. W., Sheldon, C. C., Gubler, F., ... King, R. W. (2001). GAMYB-like genes, flowering, and gibberellin signaling in *Arabidopsis*. 127, 1682–1693.
- Goetz, M., Godt, D. E., Guivarc, A., ... Roitsch, T. (2001). Induction of male sterility in plants by metabolic engineering of the carbohydrate supply. *Proceedings of the National Academy of Sciences of the United States of America*, 98(11), 6522–6527.

- Goffman, F. D., Alonso, A. P., Schwender, J., ... Ohlrogge, J. B. (2005). Light enables a very high efficiency of carbon storage in developing embryos of rapeseed. *Plant Physiology*, 138, 2269–2279.
- Gomes de Oliveira Dal'Molin, C., & Nielsen, L. K. (2013). Plant genome-scale metabolic reconstruction and modelling. *Current Opinion in Biotechnology*, 24, 271–277.
- Gomes de Oliveira Dal'Molin, C., & Nielsen, L. K. (2018). Plant genome-scale reconstruction: From single cell to multi-tissue modelling and omics analyses. *Current Opinion in Biotechnology*, 49, 42–48.
- Gomes de Oliveira Dal'Molin, C., Quek, L. E., Palfreyman, R. W., ... Nielsen, L. K. (2010). C4GEM, a genome-scale metabolic model to study C₄ plant metabolism. *Plant Physiology*, 154, 1871–1885.
- González-Carranza, Z. H., Elliott, K. A., & Roberts, J. A. (2007). Expression of polygalacturonases and evidence to support their role during cell separation processes in *Arabidopsis thaliana*. *Journal of Experimental Botany*, 58, 3719–3730.
- Goodstein, D. M., Shu, S., Howson, R., ... Rokhsar, D. S. (2012). Phytozome: A comparative platform for green plant genomics. *Nucleic acids research*, 40, D1178–86.
- Goudet, M. M. M., Orr, D. J., Melkonian, M., & Carmo-Silva, H., E. and Griffiths. (2020). RuBisCO and carbon-concentrating mechanism co-evolution across chlorophyte and streptophyte green algae. *New Phytologist*, 227, 810–823.
- Graham, T., & Dixon, M. A. (2012). Liverwort control: An ancillary role for ozone-based irrigation water treatment systems? *Horticultural Science*, 47, 361–367.
- Green, T. G. A., & Snelgar, W. P. (1982). A comparison of photosynthesis in two thalloid liverworts. *Oecologia*, 54, 275–280.
- Guan, R., Zhao, Y., Zhang, H., ... Chen, W. (2016). Draft genome of the living fossil *Ginkgo biloba*. *GigaScience*, 5, 49.
- Gudmundsson, S., & Thiele, I. (2010). Computationally efficient flux variability analysis. *BMC Bioinformatics*, 11, 489.
- Gupta, A., Singh, M., & Laxmi, A. (2015). Multiple interactions between glucose and brassinosteroid signal transduction pathways in *Arabidopsis* are uncovered by whole-genome transcriptional profiling. *Plant Physiology*, 168, 1091–1105.
- Guzman-Puyol, S., Benítez, J. J., Domínguez, E., ... Heredia-Guerrero, J. A. (2015). Pectin-lipid self-assembly: Influence on the formation of polyhydroxy fatty acids nanoparticles. *PLOS ONE*, 10, e0124639.
- Halkier, B. A., & Gershenzon, J. (2006). Biology and biochemistry of glucosinolates. *Annual Review of Plant Biology*, 57, 303–333.
- Han, G. (2016). Evolution of jasmonate biosynthesis and signaling mechanisms. *Journal of Experimental Botany*, 68, 1323–1331.
- Han, X., Yin, L., & Xue, H. (2012). Co-expression analysis identifies CRC and AP1 the regulator of *Arabidopsis* fatty acid biosynthesis. *Journal of Integrative Plant Biology*, 54, 486–499.
- Hanson, D. T., Renzaglia, K., & Villarreal, J. C. (2014). Diffusion limitation and CO₂ concentrating mechanisms in bryophytes. *Photosynthesis in bryophytes and early land plants* (pp. 95–111). Springer, Dordrecht.
- Happ, K., & Classen, B. (2019). Arabinogalactan-proteins from the liverwort *Marchantia polymorpha* l., a member of a basal land plant lineage, are structurally different to those of angiosperms. *Plants*, 8, 460.
- Harris, B. J., Harrison, C. J., Hetherington, A. M., & Williams, T. A. (2020). Phylogenomic evidence for the monophyly of bryophytes and the reductive evolution of stomata. *Current Biology*, 30, 2001–2012.e2.

- Harrison, C. J., & Morris, J. L. (2018). The origin and early evolution of vascular plant shoots and leaves. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 373.
- Hartmann, T. (1996). Diversity and variability of plant secondary metabolism: A mechanistic view. *Entomologia Experimentalis et Applicata*, 80, 177–188.
- Hayashi, K. I., Horie, K., Hiwatashi, Y., ... Nozaki, H. (2010). Endogenous diterpenes derived from *ent*-kaurene, a common gibberellin precursor, regulate protonema differentiation of the moss *Physcomitrella patens*. *Plant Physiology*, 153, 1085–1097.
- He, J., Chen, Q., Xin, P., ... Wang, G. (2019). CYP72A enzymes catalyse 13-hydrolyzation of gibberellins. *Nature Plants*, 5, 1057–1065.
- He, X., Dai, J., & Wu, Q. (2016). Identification of sporopollenin as the outer layer of cell wall in microalga *Chlorella protothecoides*. *Frontiers in Microbiology*, 7, 1047.
- Heckmann, D., Schulze, S., Denton, A., ... Lercher, M. J. (2013). Predicting C₄ photosynthesis evolution: Modular, individually adaptive steps on a Mount Fuji fitness landscape. *Cell*, 153, 1579.
- Heim, M. A., Jakoby, M., Werber, M., ... Bailey, P. C. (2003). The basic helix-loop-helix transcription factor family in plants: A genome-wide study of protein structure and functional diversity. *Molecular Biology and Evolution*, 20, 735–747.
- Hernández-García, J., Briones-Moreno, A., & Blázquez, M. A. (2020). Origin and evolution of gibberellin signaling and metabolism in plants. *Seminars in Cell and Developmental Biology*, 109, 46–54.
- Hernández-García, J., Briones-Moreno, A., Dumas, R., & Blázquez, M. A. (2019). Origin of gibberellin-dependent transcriptional regulation by molecular exploitation of a transactivation domain in DELLA proteins. *Molecular Biology and Evolution*, 36, 908–918.
- Herrmann, H. A., Dyson, B. C., Vass, L., ... Schwartz, J. M. (2019). Flux sampling is a powerful tool to study metabolism under changing environmental conditions. *npj Systems Biology and Applications*, 5, 1–8.
- Hetherington, A. J., Berry, C. M., & Dolan, L. (2016). Networks of highly branched stigmarian rootlets developed on the first giant trees. *Proceedings of the National Academy of Sciences of the United States of America*, 113, 6695–6700.
- Hetherington, A. J., & Dolan, L. (2018). Stepwise and independent origins of roots among land plants. *Nature*, 561, 235–238.
- Himmel, M. E., Ding, S. Y., Johnson, D. K., ... Foust, T. D. (2007). Biomass recalcitrance: Engineering plants and enzymes for biofuels production. *Science*, 315, 804–807.
- Hirano, K., Nakajima, M., Asano, K., ... Matsuoka, M. (2007). The GID1-mediated gibberellin perception mechanism is conserved in the lycophyte *Selaginella moellendorffii* but not in the bryophyte *Physcomitrella patens*. *The Plant Cell*, 19, 3058–3079.
- Hohmann-Marriott, M. F., & Blankenship, R. E. (2011). Evolution of photosynthesis. *Annual Review of Plant Biology*, 62, 515–548.
- Hori, K., Maruyama, F., Fujisawa, T., ... Ohta, H. (2014). *Klebsormidium flaccidum* genome reveals primary factors for plant terrestrial adaptation. *Nature Communications*, 5.
- Hu, J., Wang, Z., Zhang, L., & Sun, M. (2014). The *Arabidopsis* Exine Formation Defect (EFD) gene is required for primexine patterning and is critical for pollen fertility. *New Phytologist*, 203, 140–154.
- Huang, R., O'Donnell, A. J., Barboline, J. J., & Barkman, T. J. (2016). Convergent evolution of caffeine in plants by co-option of exapted ancestral enzymes. *Proceedings of the National Academy of Sciences of the United States of America*, 113, 10613–10618.
- Huang, Y., Wang, X., Ge, S., & Rao, G. (2015). Divergence and adaptive evolution of the gibberellin oxidase genes in plants. *BMC Evolutionary Biology*, 15, 207.

- Humphreys, C. P., Franks, P. J., Rees, M., ... Beerling, D. J. (2010). Mutualistic mycorrhiza-like symbiosis in the most ancient group of land plants. *Nature Communications*, 1, 1–7.
- Hwang, J. E., Kim, Y. J., Shin, M. H., & Park, J. (2018). AtBTH1, a BTB domain-containing protein, influences sugar-responsive germination and seedling growth in *Arabidopsis*. *Genetics and Molecular Research*, 17(4).
- Imoto, K., Yokoyama, R., & Nishitani, K. (2005). Comprehensive approach to genes involved in cell wall modifications in *Arabidopsis thaliana*. *Plant Molecular Biology*, 58, 177–192.
- Ishizaki, K., Nishihama, R., Yamato, K. T., & Kohchi, T. (2016). Molecular genetic tools and techniques for *Marchantia polymorpha* research. *Plant and Cell Physiology*, 57, 262–270.
- Itkin, M., Heinig, U., Tzfadia, O., ... Aharoni, A. (2013). Biosynthesis of antinutritional alkaloids in solanaceous crops is mediated by clustered genes. *Science*, 341, 175–179.
- Jasinski, S., Piazza, P., Craft, J., ... Tsiantis, M. (2005). KNOX action in *Arabidopsis* is mediated by coordinate regulation of cytokinin and gibberellin activities. *Current Biology*, 15, 1560–1565.
- Jensen, P. E., & Leister, D. (2014). Chloroplast evolution, structure and functions. *F1000 Prime Reports*, 6.
- Jeske, L., Placzek, S., Schomburg, I., ... Schomburg, D. (2019). BRENDA in 2019: A European ELIXIR core data resource. *Nucleic Acids Research*, 47, D542–D549.
- Jia, H., Li, J., Zhu, J., ... An, L. (2013). *Arabidopsis* CROLIN1, a novel plant actin-binding protein, functions in cross-linking and stabilizing actin filaments. *Journal of Biological Chemistry*, 288, 32277–32288.
- Joshi, C. P., & Chiang, V. L. (1998). Conserved sequence motifs in plant S-adenosyl-L-methionine-dependent methyltransferases. *Plant Molecular Biology*, 37, 663–674.
- Ju, C., Van de Poel, B., Cooper, E. D., ... Chang, C. (2015). Conservation of ethylene as a plant hormone over 450 million years of evolution. *Nature Plants*, 1, 14004.
- Kang, S., Pandey, R. P., Lee, C., ... Sohng, J. K. (2020). Genome-enabled discovery of anthraquinone biosynthesis in *Senna tora*. *bioRxiv*, 2020.04.27.063495.
- Karp, P. D., Latendresse, M., Paley, S. M., ... Caspi, R. (2016). Pathway Tools version 19.0 update: Software for pathway/genome informatics and systems biology. *Briefings in bioinformatics*, 17, 877–90.
- Katoh, K. (1983). Photosynthesis and photoassimilation of glucose during photomixotrophic growth of *Marchantia polymorpha* cells in suspension culture. *Physiologia Plantarum*, 57, 67–74.
- Katoh, K., & Standley, D. M. (2013). MAFFT Multiple Sequence Alignment Software Version 7: Improvements in performance and usability. *Molecular Biology and Evolution*, 30, 772–780.
- Keeley, J. E., & Rundel, P. W. (2003). Evolution of CAM and C₄ carbon-concentrating mechanisms. *International Journal of Plant Sciences*, 164, 55–77.
- Kim, E., & Russinova, E. (2020). Brassinosteroid signalling. *Current Biology*, 30, R294–R298.
- Kim, J., & Buell, C. R. (2015). A revolution in plant metabolism: Genome-enabled pathway discovery. *Physiology*, 169, 1532–1539.
- Knoth, C., & Eulgem, T. (2008). The oomycete response gene LURP1 is required for defense against *Hyaloperonospora parasitica* in *Arabidopsis thaliana*. *The Plant Journal*, 55, 53–64.
- Koide, E., Suetsugu, N., Iwano, M., ... Nishihama, R. (2020). Regulation of photosynthetic carbohydrate metabolism by a raf-like kinase in the liverwort *Marchantia polymorpha*. *Plant and Cell Physiology*, 61, 631–643.
- Kondo, S., Hori, K., Sasaki-Sekimoto, Y., ... Ohta, H. (2016). Primitive extracellular lipid components on the surface of the charophytic alga *Klebsormidium flaccidum* and their possible biosynthetic pathways as deduced from the genome sequence. *Frontiers in Plant Science*, 7, 952.

- Krokida, A., Delis, C., Geisler, K., ... Papadopoulou, K. K. (2013). A metabolic gene cluster in *Lotus japonicus* discloses novel enzyme functions and products in triterpene biosynthesis. *New Phytologist*, 200, 675–690.
- Kueger, S., Steinhäuser, D., Willmitzer, L., & Giavalisco, P. (2012). High-resolution plant metabolomics: From mass spectral features to metabolites and from whole-cell analysis to subcellular metabolite distributions. *Plant Journal*, 70, 39–50.
- Kumar, S., Kempinski, C., Zhuang, X., ... Chappell, J. (2016). Molecular diversity of terpene synthases in the liverwort *Marchantia polymorpha*. *The Plant Cell*, 28, 2632–2650.
- Kusano, H., Li, H., Minami, H., ... Yazaki, K. (2019). Evolutionary developments in plant specialized metabolism, exemplified by two transferase families. *Frontiers in Plant Science*, 10, 794.
- Le, D. T., Nishiyama, R., Watanabe, Y., ... Tran, L. P. (2012). Identification and expression analysis of cytokinin metabolic genes in soybean under normal and drought conditions in relation to cytokinin levels. *PLoS ONE*, 7, e42411.
- Leebens-Mack, J. H., Barker, M. S., Carpenter, E. J., ... Wong, G. K. S. (2019). One thousand plant transcriptomes and the phylogenomics of green plants. *Nature*, 574, 679–685.
- Lei, L., Steffen, J. G., Osborne, E. J., & Toomajian, C. (2017). Plant organ evolution revealed by phylotranscriptomics in *Arabidopsis thaliana*. *Scientific Reports*, 7.
- Leonhardt, N., Kwak, J. M., Robert, N., ... Schroeder, J. I. (2004). Microarray expression analyses of *Arabidopsis* guard cells and isolation of a recessive abscisic acid hypersensitive protein phosphatase 2C mutant. *The Plant Cell*, 16, 596–615.
- Lewis, N. E., Hixson, K. K., Conrad, T. M., ... Palsson, B. (2010). Omic data from evolved *E. coli* are consistent with computed optimal growth from genome-scale models. *Molecular Systems Biology*, 6.
- Lewis, N. E., Nagarajan, H., & Palsson, B. O. (2012). Constraining the metabolic genotype-phenotype relationship using a phylogeny of *in silico* methods. *Nature Reviews Microbiology*, 10, 291–305.
- Li, F. S., Phyo, P., Jacobowitz, J., ... Weng, J. K. (2019). The molecular structure of plant sporopollenin. *Nature Plants*, 5, 41–46.
- Li, F. W., Brouwer, P., Carretero-Paulet, L., ... Pryer, K. M. (2018). Fern genomes elucidate land plant evolution and cyanobacterial symbioses. *Nature Plants*, 4, 460–472.
- Li, F. W., Nishiyama, T., Waller, M., ... Szövényi, P. (2020). *Anthoceros* genomes illuminate the origin of land plants and the unique biology of hornworts. *Nature Plants*, 6, 259–272.
- Li, F. W., Villarreal, J. C., & Szövényi, P. (2017). Hornworts: An overlooked window into carbon-concentrating mechanisms. *Trends in Plant Science*, 22, 275–277.
- Li, G., Köllner, T. G., Yin, Y., ... Chen, F. (2012). Nonseed plant *Selaginella moellendorffii* has both seed plant and microbial types of terpene synthases. *Proceedings of the National Academy of Sciences of the United States of America*, 109, 14711–14715.
- Ligrone, R. (2019). The great oxygenation event. *Biological innovations that built the world* (pp. 129–154). Springer International Publishing.
- Liu, B., De Storme, N., & Geelen, D. (2017). Gibberellin induces diploid pollen formation by interfering with meiotic cytokinesis. *Plant Physiology*, 173, 338–353.
- Liu, H., Guo, S., Lu, M., ... Wang, W. (2019). Biosynthesis of DHGA₁₂ and its roles in *Arabidopsis* seedling establishment. *Nature Communications*, 10, 1768.
- Liu, X., & Hou, X. (2018). Antagonistic regulation of ABA and GA in metabolism and signaling pathways. *Frontiers in Plant Science*, 9, 251.
- López-Calcano, P. E., Brown, K. L., Simkin, A. J., ... Raines, C. A. (2020). Stimulating photosynthetic processes increases productivity and water-use efficiency in the field. *Nature Plants*, 6, 1054–1063.

- Loreto, F., Delfine, S., & Di Marco, G. (1999). Estimation of photorespiratory carbon dioxide recycling during photosynthesis. *Australian Journal of Plant Physiology*, 26, 733–736.
- Lumba, S., Toh, S., Handfield, L. F., ... McCourt, P. (2014). A mesoscale abscisic acid hormone interactome reveals a dynamic signaling landscape in *Arabidopsis*. *Developmental Cell*, 29, 360–372.
- Ma, F., Jazmin, L. J., Young, J. D., & Allen, D. K. (2014). Isotopically nonstationary ^{13}C flux analysis of changes in *Arabidopsis thaliana* leaf metabolism due to high light acclimation. *Proceedings of the National Academy of Sciences of the United States of America*, 111, 16967–16972.
- Macduff, J. H., & Bakken, A. K. (2003). Diurnal variation in uptake and xylem contents of inorganic and assimilated N under continuous and interrupted N supply to *Phleum pratense* and *Festuca pratensis*. *Journal of Experimental Botany*, 54, 431–444.
- Maeda, H., Yoo, H., & Dudareva, N. (2010). Prephenate aminotransferase directs plant phenylalanine biosynthesis via arogonate. *Nature Chemical Biology*, 7, 19–21.
- Magome, H., Nomura, T., Hanada, A., ... Yamaguchi, S. (2013). CYP714B1 and CYP714B2 encode gibberellin 13-oxidases that reduce gibberellin activity in rice. *Proceedings of the National Academy of Sciences of the United States of America*, 110, 1947–52.
- Mahood, E. H., Kruse, L. H., & Moghe, G. D. (2020). Machine learning: A powerful tool for gene function prediction in plants. *Applications in Plant Sciences*, 8, e11376.
- Maloney, V. J., Samuels, A. L., & Mansfield, S. D. (2012). The endo-1,4- β -glucanase Korrigan exhibits functional conservation between gymnosperms and angiosperms and is required for proper cell wall formation in gymnosperms. *New Phytologist*, 193, 1076–1087.
- Mao, L., Kawaide, H., Higuchi, T., ... Okada, K. (2020). Genomic evidence for convergent evolution of gene clusters for momilactone biosynthesis in land plants. *Proceedings of the National Academy of Sciences of the United States of America*, 117, 12472–12480.
- Mao, R., Kumar, P. K. R., Guo, C., ... Liang, C. (2014). Comparative analyses between retained introns and constitutively spliced introns in *Arabidopsis thaliana* using random forest and support vector machine. *PLoS ONE*, 9, e104049.
- Maréchal, E. (2018). Primary endosymbiosis: Emergence of the primary chloroplast and the chromatophore, two independent events. *Methods in Molecular Biology*, 1829, 3–16.
- Marie-Andrée, Hartmann-Bouillon, & Benveniste, P. (1978). Sterol biosynthetic capability of purified membrane fractions from maize coleoptiles. *Phytochemistry*, 17, 1037–1042.
- Marsolais, F., Boyd, J., Paredes, Y., ... Varin, L. (2007). Molecular and biochemical characterization of two brassinosteroid sulfotransferases from *Arabidopsis*, AtST4a (At2g14920) and AtST1 (At2g03760). *Planta*, 225, 1233–44.
- Martin, W., Rujan, T., Richly, E., ... Penny, D. (2002). Evolutionary analysis of *Arabidopsis*, cyanobacterial, and chloroplast genomes reveals plastid phylogeny and thousands of cyanobacterial genes in the nucleus. *Proceedings of the National Academy of Sciences of the United States of America*, 99, 12246–12251.
- Mathis, P. (1990). Compared structure of plant and bacterial photosynthetic reaction centers. Evolutionary implications. *Biochimica et Biophysica Acta (BBA) - Bioenergetics*, 1018, 163–167.
- Mazurek, S., Garroum, I., Daraspe, J., ... Nawrath, C. (2017). Connecting the molecular structure of cutin to ultrastructure and physical properties of the cuticle in petals of *Arabidopsis*. *Plant Physiology*, 173, 1146–1163.
- Mazzucotelli, E., Belloni, S., Marone, D., ... Mastrangelo, A. (2006). The E3 ubiquitin ligase gene family in plants: Regulation by degradation. *Current Genomics*, 7, 509–522.
- McCourt, R. M., Delwiche, C. F., & Karol, K. G. (2004). Charophyte algae and land plant origins. *Trends in Ecology and Evolution*, 19, 661–666.

- McGuinness, P. N., Reid, J. B., & Foo, E. (2019). The role of gibberellins and brassinosteroids in nodulation and arbuscular mycorrhizal associations. *Frontiers in Plant Science*, 10.
- McKay, M. D., Beckman, R. J., & Conover, W. J. (1979). A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. *Technometrics*, 21, 239.
- Merced, A., & Renzaglia, K. S. (2017). Structure, function and evolution of stomata from a bryological perspective. *Bryophyte Diversity and Evolution*, 39, 7.
- Meyer, M., & Griffiths, H. (2013). Origins and diversity of eukaryotic CO₂-concentrating mechanisms: Lessons for the future. *Journal of Experimental Botany*, 64, 769–786.
- Michelet, L., Zaffagnini, M., Morisse, S., ... Lemaire, S. D. (2013). Redox regulation of the Calvin-Benson Cycle: Something old, something new. *Frontiers in Plant Science*, 4, 470.
- Miquel, M., & Browse, J. (1992). *Arabidopsis* mutants deficient in polyunsaturated fatty acid synthesis. *The Journal of Biological Chemistry*, 267, 1502–1509.
- Miyazaki, S., Hara, M., Ito, S., ... Nakajima, M. (2018). An ancestral gibberellin in a moss *Physcomitrella patens*. *Molecular Plant*, 11, 1097–1100.
- Moghe, G. D., & Last, R. L. (2015). Something old, something new: Conserved enzymes and the evolution of novelty in plant specialized metabolism. *Plant Physiology*, 169, 1512–1523.
- Mølhøj, M., Jørgensen, B., Ulvskov, P., & Borkhardt, B. (2001). Two *Arabidopsis thaliana* genes, KOR2 and KOR3, which encode membrane-anchored endo-1,4-β-D-glucanases, are differentially expressed in developing leaf trichomes and their support cells. *Plant Molecular Biology*, 46, 263–275.
- Möller, B. K., Ten Hove, C. A., Xiang, D., ... Weijers, D. (2017). Auxin response cell-autonomously controls ground tissue initiation in the early *Arabidopsis* embryo. *Proceedings of the National Academy of Sciences of the United States of America*, 114, E2533–E2539.
- Moore, B. M., Wang, P. P., Fan, P., ... Shiu, S. H. (2019). Robust predictions of specialized metabolism genes through machine learning. *Proceedings of the National Academy of Sciences of the United States of America*, 116, 2344–2353.
- Moreau, P., Hartmann, M., Perret, A., ... Cassagne, C. (1998). Transport of sterols to the plasma membrane of leek seedlings. *Plant Physiology*, 117, 931–937.
- Moreira, T. B., Lima, J. M., Coca, G. C., & Williams, T. C. R. (2019). Insights into the spatial and temporal organisation of plant metabolism from network flux analysis. *Theoretical and Experimental Plant Physiology*, 31, 215–226.
- Morris, J. L., Puttick, M. N., Clark, J. W., ... Donoghue, P. C. J. (2018). The timescale of early land plant evolution. *Proceedings of the National Academy of Sciences of the United States of America*, 115, E2274–E2283.
- Mueller, L. A., Zhang, P., & Rhee, S. Y. (2003). Aracyc: A biochemical pathway database for *Arabidopsis*. *Plant Physiology*, 132, 453–460.
- Mutasa-Göttgens, E., & Hedden, P. (2009). Gibberellin as a factor in floral regulatory networks. *Journal of Experimental Botany*, 60, 1979–1989.
- Mutte, S. K., Kato, H., Rothfels, C., ... Weijers, D. (2018). Origin and evolution of the nuclear auxin response system. *eLife*, 7, e33399.
- Neale, D. B., McGuire, P. E., Wheeler, N. C., ... Wegrzyn, J. L. (2017). The Douglas-Fir genome sequence reveals specialization of the photosynthetic apparatus in Pinaceae. *G3: Genes, Genomes, Genetics*, 7, 3157–3167.
- Neale, D. B., Wegrzyn, J. L., Stevens, K. A., ... Langley, C. H. (2014). Decoding the massive genome of loblolly pine using haploid DNA and novel assembly strategies. *Genome Biology*, 15(R59).
- Nelson, D., & Werck-Reichhart, D. (2011). A P450-centric view of plant evolution. *The Plant Journal*, 66, 194–211.

- Nguyen, L., Schmidt, H. A., von Haeseler, A., & Minh, B. Q. (2015). IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Molecular Biology and Evolution*, 32, 268–274.
- Nguyen-Quoc, B., & Foyer, C. H. (2001). A role for ‘futile cycles’ involving invertase and sucrose synthase in sucrose metabolism of tomato fruit. *Journal of Experimental Botany*, 52, 881–889.
- Niklas, K. J., Cobb, E. D., & Matas, A. J. (2017). The evolution of hydrophobic cell wall biopolymers: From algae to angiosperms. *Journal of Experimental Botany*, 68, 5261–5269.
- Nikoloski, Z., Perez-Storey, R., & Sweetlove, L. J. (2015). Inference and prediction of metabolic network fluxes. *Plant Physiology*, 169, 1443–1455.
- Nishikawa, S. I., Zinkl, G. M., Swanson, R. J., ... Preuss, D. (2005). Callose (β -1,3 glucan) is essential for *Arabidopsis* pollen wall patterning, but not tube growth. *BMC Plant Biology*, 5, 22.
- Nishiyama, T., Sakayama, H., de Vries, J., ... Rensing, S. A. (2018). The *Chara* genome: Secondary complexity and implications for plant terrestrialization. *Cell*, 174, 448–464.e24.
- Noguchi, S., C. and Miyazaki, Kawaide, H., Gotoh, O., ... Aoyama, Y. (2018). Characterization of moss *ent*-kaurene oxidase (CYP701B1) using a highly purified preparation. *Journal of biochemistry*, 163, 69–76.
- Nolan, T. M., Brennan, B., Yang, M., ... Li, Z. (2017). Selective autophagy of BES1 mediated by DSK2 balances plant growth and survival. *Developmental Cell*, 41, 33–46.
- Nolan, T. M., Vukasinović, N., Liu, D., ... Yin, Y. (2020). Brassinosteroids: Multidimensional regulators of plant growth, development, and stress responses. *The Plant Cell*, 32, 298–318.
- Novoselov, S. V., Rao, M., Onoshko, N. V., ... Gladyshev, V. N. (2002). Selenoproteins and selenocysteine insertion system in the model plant cell system, *Chlamydomonas reinhardtii*. *The EMBO journal*, 21, 3681–93.
- Nützmann, H. W., Huang, A., & Osbourn, A. (2016). Plant metabolic clusters – from genetics to genomics. *New Phytologist*, 211, 771–789.
- Nystedt, B., Street, N. R., Wetterbom, A., ... Jansson, S. (2013). The Norway spruce genome sequence and conifer genome evolution. *Nature*, 497, 579–584.
- O’Brien, E. J., Monk, J. M., & Palsson, B. O. (2015). Using genome-scale models to predict biological capabilities. *Cell*, 161, 971–987.
- Ohnishi, T., Szatmari, A. M., Watanabe, B., ... Mizutani, M. (2006). C-23 hydroxylation by *Arabidopsis* CYP90C1 and CYP90D1 reveals a novel shortcut in brassinosteroid biosynthesis. *The Plant Cell*, 18, 3275–3288.
- Ohtaka, K., Hori, K., Kanno, Y., ... Ohta, H. (2017). Primitive auxin response without TIR1 and Aux/IAA in the charophyte alga *Klebsormidium nitens*. *Plant Physiology*, 174, 1621–1632.
- Olson, J. M., & Pierson, B. K. (1987). Origin and evolution of photosynthetic reaction centers. *Origins of life and evolution of the biosphere*, 17, 419–430.
- Ort, D. R., Merchant, S. S., Alric, J., ... Zhu, X. G. (2015). Redesigning photosynthesis to sustainably meet global food and bioenergy demand. *Proceedings of the National Academy of Sciences of the United States of America*, 112, 8529–8536.
- Orth, J. D., Thiele, I., & Palsson, B. O. (2010). What is flux balance analysis? *Nature Biotechnology*, 28, 245–248.
- Park, J. M., Kim, T. Y., & Lee, S. Y. (2011). Genome-scale reconstruction and in silico analysis of the *Ralstonia eutropha* h16 for polyhydroxyalkanoate synthesis, lithoautotrophic growth, and 2-methyl citric acid production. *BMC Systems Biology*, 5, 101.
- Pärnik, T., & Keerberg, O. (2007). Advanced radiogasometric method for the determination of the rates of photorespiratory and respiratory decarboxylations of primary and stored photosynthates under steady-state photosynthesis. *Physiologia Plantarum*, 129, 34–44.

- Parrr/random-forest-importances. (2020). Retrieved October 1, 2020, from <https://github.com/parrr/random-forest-importances>
- Pedregosa, F., Michel, V., Grisel, O., & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830.
- Pellegrini, M., Marcotte, E. M., Thompson, M. J., ... Yeates, T. O. (1999). Assigning protein functions by comparative genome analysis: Protein phylogenetic profiles. *Biochemistry*, 96, 4285–4288.
- Peng, Z., Han, C., Yuan, L., ... Ren, C. (2011). Brassinosteroid enhances jasmonate-induced anthocyanin accumulation in *Arabidopsis* seedlings. *Journal of Integrative Plant Biology*, 53, 632–640.
- Persson, S., Wei, H., Milne, J., ... Somerville, C. R. (2005). Identification of genes required for cellulose synthesis by regression analysis of public microarray data sets. *Proceedings of the National Academy of Sciences of the United States of America*, 102, 8633–8638.
- Philippe, G., Sørensen, I., Jiao, C., ... Rose, J. K. C. (2020). Cutin and suberin: Assembly and origins of specialized lipidic cell wall scaffolds. *Current Opinion in Plant Biology*, 55, 11–20.
- Phukan, U. J., Jeena, G. S., & Shukla, R. K. (2016). WRKY transcription factors: Molecular regulation and stress responses in plants. *Frontiers in Plant Science*, 7, 760.
- Pichersky, E., & Lewinsohn, E. (2011). Convergent evolution in plant specialized metabolism. *Annual Review of Plant Biology*, 62, 549–566.
- Pils, B., & Heyl, A. (2009). Unraveling the evolution of cytokinin signaling. *Plant Physiology*, 151, 782–791.
- Poolman, M. G., Miguët, L., Sweetlove, L. J., & Fell, D. A. (2009). A genome-scale metabolic model of *Arabidopsis* and some of its properties. *Plant Physiology*, 151, 1570–1581.
- Poppenberger, B., Fujioka, S., Soeno, K., ... Bowles, D. (2005). The UGT73C5 of *Arabidopsis thaliana* glucosylates brassinosteroids. *Proceedings of the National Academy of Sciences of the United States of America*, 102, 15253–8.
- Pott, D. M., Osorio, S., & Vallarino, J. G. (2019). From central to specialized metabolism: An overview of some secondary compounds derived from the primary metabolism for their role in conferring nutritional and organoleptic characteristics to fruit. *Frontiers in Plant Science*, 10, 835.
- Preston, J., Wheeler, J., Heazlewood, J., ... Parish, R. W. (2004). AtMYB32 is required for normal pollen development in *Arabidopsis thaliana*. *Plant Journal*, 40, 979–995.
- Puttick, M. N., Morris, J. L., Williams, T. A., ... Donoghue, P. C. J. (2018). The interrelationships of land plants and the nature of the ancestral embryophyte. *Current Biology*, 28, 733–745.e2.
- Qi, T., Huang, H., Wu, D., ... Xie, D. (2014). *Arabidopsis* DELLA and JAZ proteins bind the WD-repeat/ bHLH/MYB complex to modulate gibberellin and jasmonate signaling synergy. *The Plant Cell*, 26, 1118–1133.
- Quilichini, T. D., Grienemberger, E., & Douglas, C. J. (2015). The biosynthesis, composition and assembly of the outer pollen wall: A tough case to crack. *Phytochemistry*, 113, 170–182.
- Rai, A., Saito, K., & Yamazaki, M. (2017). Integrated omics analysis of specialized metabolism in medicinal plants. *Plant Journal*, 90, 764–787.
- Ranea, J. A. G., Yeats, C., Grant, A., & Orengo, C. A. (2007). Predicting protein function with hierarchical phylogenetic profiles: The Gene3D Phylo-Tuner Method applied to eukaryotic genomes. *PLoS Computational Biology*, 3, e237.
- Rashotte, A. M. (2020). The evolution of cytokinin signaling and its role in development before angiosperms. *Seminars in Cell and Developmental Biology*, 109, 31–38.
- Rasse, D. P., & Tocquin, P. (2006). Leaf carbohydrate controls over *Arabidopsis* growth and response to elevated CO₂: An experimentally based model. *New Phytologist*, 172, 500–513.

- Raven, J. A., Cockell, C. S., De, C. L., & Rocha, L. (2008). The evolution of inorganic carbon concentrating mechanisms in photosynthesis. *Philosophical Transactions of the Royal Society B*, 363, 2641–2650.
- Reich, P. B., Hobbie, S. E., Lee, T., ... Trost, J. (2006). Nitrogen limitation constrains sustainability of ecosystem response to CO₂. *Nature*, 440, 922–925.
- Renault, H., Alber, A., Horst, N. A., ... Werck-Reichhart, D. (2017). A phenol-enriched cuticle is ancestral to lignin evolution in land plants. *Nature Communications*, 8, 14713.
- Rensing, S. A. (2017). Why we need more non-seed plant models. *New Phytologist*, 216, 355–360.
- Rensing, S. A. (2018). Great moments in evolution: The conquest of land by plants. *Current Opinion in Plant Biology*, 42, 49–54.
- Rensing, S. A., Lang, D., Zimmer, A. D., ... Boore, J. L. (2008). The *Physcomitrella* genome reveals evolutionary insights into the conquest of land by plants. *Science*, 319, 64–69.
- Renzaglia, K. S., Browning, W. B., & Merced, A. (2020). With over 60 independent losses, stomata are expendable in mosses. *Frontiers in Plant Science*, 11, 567.
- Resemann, H. C., Lewandowska, M., Gömann, J., & Feussner, I. (2019). Membrane lipids, waxes and oxylipins in the moss model organism *Physcomitrella patens*. *Plant and Cell Physiology*, 60, 1166–1175.
- Rhee, S. Y., & Mutwil, M. (2014). Towards revealing the functions of all genes in plants. *Trends in Plant Science*, 19, 212–221.
- Riekhof, W. R., Naik, S., Bertrand, H., ... Voelker, D. R. (2014). Phosphate starvation in fungi induces the replacement of phosphatidylcholine with the phosphorus-free betaine lipid diacylglyceryl-N,N,N-trimethylhomoserine. *Eukaryotic cell*, 13, 749–57.
- Rieu, I., Eriksson, S., Powers, S. J., ... Phillips, A. L. (2008). Genetic analysis reveals that C19-GA 2-oxidation is a major gibberellin inactivation pathway in *Arabidopsis*. *The Plant Cell*, 20, 2420–36.
- Robaina Estévez, S., Daloso, D. M., Zhang, Y., & Fernie, A. R. (2017). Resolving the central metabolism of *Arabidopsis* guard cells. *Scientific Reports*, 7, 8307.
- Robaina Estévez, S., & Nikoloski, Z. (2015). Context-specific metabolic model extraction based on regularized least squares optimization. *PLOS ONE*, 10, e0131875.
- Rockwell, N. C., Lagarias, J. C., & Bhattacharya, D. (2014). Primary endosymbiosis and the evolution of light and oxygen sensing in photosynthetic eukaryotes. *Frontiers in Ecology and Evolution*, 2, 66.
- Rodman, J., Soltis, P., Soltis, D., ... Karol, K. (1998). Parallel evolution of glucosinolate biosynthesis inferred from congruent nuclear and plastid gene phylogenies. *American journal of botany*, 85, 997.
- Rojas, M., Ruwe, H., Miranda, R. G., ... Barkan, A. (2018). Unexpected functional versatility of the pentatricopeptide repeat proteins PGR3, PPR5 and PPR10. *Nucleic Acids Research*, 46, 10448–10459.
- Romani, F. (2017). Origin of TAA genes in charophytes: New insights into the controversy over the origin of auxin biosynthesis. *Frontiers in Plant Science*, 8, 1616.
- Roscher, A., Kruger, N. J., & Ratcliffe, R. G. (2000). Strategies for metabolic flux analysis in plants using isotope labelling. *Journal of Biotechnology*, 77, 81–102.
- Rozentsvet, O. A. (2004). Comparative examination of distribution of phospholipids and a betaine lipid DGTS in tropical fern species. *Biochemical Systematics and Ecology*, 32, 303–311.
- Rozentsvet, O. A., Dembitsky, V. M., & Saksonov, S. V. (2000). Occurrence of diacylglyceryltrimethylhomoserines and major phospholipids in some plants. *Phytochemistry*, 54, 401–407.

- Ruan, Y. L. (2014). Sucrose metabolism: Gateway to diverse carbon use and sugar signaling. *Annual Review of Plant Biology*, 65, 33–67.
- Rubio-Asensio, J. S., & Bloom, A. J. (2017). Inorganic nitrogen form: A major player in wheat and *Arabidopsis* responses to elevated CO₂. *Journal of Experimental Botany*, 68, 2611–2625.
- Ruhfel, B. R., Gitzendanner, M. A., Soltis, P. S., ... Burleigh, J. G. (2014). From algae to angiosperms-inferring the phylogeny of green plants (Viridiplantae) from 360 plastid genomes. *BMC Evolutionary Biology*, 14, 1–27.
- Ruprecht, C., Mutwil, M., Saxe, E., ... Persson, S. (2011). Large-scale co-expression approach to dissect secondary cell wall formation across plant species. *Frontiers in Plant Science*, 2.
- Russinova, E., Borst, J., Kwaaitaal, M., ... De Vries, S. C. (2004). Heterodimerization and endocytosis of *Arabidopsis* brassinosteroid receptors BRI1 and AtSERK3 (BAK1). 16(12), 3216–3229.
- Sage, R. F., Christin, P., & Edwards, E. J. (2011). The C₄ plant lineages of planet earth. *Journal of Experimental Botany*, 62, 3155–69.
- Sage, R. F., Monson, R. K., Ehleringer, J. R., ... Pearcy, R. W. (2018). Some like it hot: The physiological ecology of C₄ plant evolution. *Oecologia*, 187, 941–966.
- Saha, R., Suthers, P. F., & Maranas, C. D. (2011). *Zea mays* iRS1563: A comprehensive genome-scale metabolic reconstruction of maize metabolism. *PLoS ONE*, 6, e21784.
- Sakakibara, K., Ando, S., Yip, H. K., ... Bowman, J. L. (2013). KNOX2 genes regulate the haploid-to-diploid morphological transition in land plants. *Science*, 339, 1067–1070.
- Santuari, L., Sanchez-Perez, G. F., Luijten, M., ... Heidstra, R. (2016). The PLETHORA gene regulatory network guides growth and cell differentiation in *Arabidopsis* roots. 28, 2937–2951.
- Saravanan, R. S., Slabaugh, E., Singh, V. R., ... Brandizzi, F. (2009). The targeting of the oxysterol-binding protein ORP3a to the endoplasmic reticulum relies on the plant VAP33 homolog PVA12. *The Plant Journal*, 58, 817–830.
- Sasaki-Sekimoto, Y., Jikumaru, Y., Obayashi, T., ... Shirasu, K. (2013). Basic helix-loop-helix transcription factors JASMONATE-ASSOCIATED MYC2-LIKE1 (JAM1), JAM2, and JAM3 are negative regulators of jasmonate responses in *Arabidopsis*. 163, 291–304.
- Sauret-Güeto, S., Frangedakis, E., Silvestri, L., ... Haseloff, J. (2020). Systematic tools for reprogramming plant gene expression in a simple model, *Marchantia polymorpha*. *BioRxiv*.
- Scheible, W. R., Krapp, A., & Stitt, M. (2000). Reciprocal diurnal changes of phosphoenolpyruvate carboxylase expression and cytosolic pyruvate kinase, citrate synthase and NADP-isocitrate dehydrogenase expression regulate organic acid metabolism during nitrate assimilation in tobacco leaves. *Plant, Cell and Environment*, 23, 1155–1167.
- Scheurwater, I., Cornelissen, C., Dictus, F., ... Lambers, H. (1998). Why do fast- and slow-growing grass species differ so little in their rate of root respiration, considering the large differences in rate of growth and ion uptake? *Plant, Cell and Environment*, 21, 995–1005.
- Schilling, C. H., Edwards, J. S., Letscher, D., & Palsson, B. (2000). Combining pathway analysis with flux balance analysis for the comprehensive study of metabolic systems. *Biotechnology and Bioengineering*, 71, 286–306.
- Schilmiller, A. L., Stout, J., Weng, J. K., ... Chapple, C. (2009). Mutations in the cinnamate 4-hydroxylase gene impact metabolism, growth and development in *Arabidopsis*. *Plant Journal*, 60, 771–782.
- Schläpfer, P., Zhang, P., Wang, C., ... Rhee, S. Y. (2017). Genome-wide prediction of metabolic enzymes, pathways, and gene clusters in plants. *Plant Physiology*, 173, 2041–2059.
- Schwechheimer, C. (2012). Gibberellin signaling in plants - the extended version. *Frontiers in Plant Science*, 2, 107.
- Schwender, J., Goffman, F., Ohlrogge, J. B., & Shachar-Hill, Y. (2004). RuBisCO without the Calvin Cycle improves the carbon efficiency of developing green seeds. *Nature*, 432, 779–782.

- Schwender, J., & Hay, J. O. (2012). Predictive modeling of biomass component tradeoffs in *Brassica napus* developing oilseeds based on *in silico* manipulation of storage metabolism. *Plant Physiology*, 160, 1218–1236.
- Seo, Y. S., Chern, M., Bartley, L. E., ... Ronald, P. C. (2011). Towards establishment of a rice stress response interactome. *PLoS Genetics*, 7, e1002020.
- Shameer, S., Baghalian, K., Cheung, C. Y. M., ... Sweetlove, L. J. (2018). Computational analysis of the productivity potential of CAM. *Nature Plants*, 4, 165–171.
- Shameer, S., Ratcliffe, R. G., & Sweetlove, L. J. (2019). Leaf energy balance requires mitochondrial respiration and export of chloroplast NADPH in the light. *Plant Physiology*, 180, 1947–1961.
- Shameer, S., Vallarino, J. G., Fernie, A. R., ... Sweetlove, L. J. (2020). Flux balance analysis of metabolism during growth by osmotic cell expansion and its application to tomato fruits. *The Plant Journal*, 103, 68–82.
- Shaw, R., & Cheung, C. Y. M. (2018). A dynamic multi-tissue flux balance model captures carbon and nitrogen metabolism and optimal resource partitioning during *Arabidopsis* growth. *Frontiers in Plant Science*, 9, 884.
- Sheth, B. P., & Thaker, V. S. (2014). Plant systems biology: Insights, advances and challenges. *Planta*, 240, 33–54.
- Shi, J., Malitsky, S., De Oliveira, S., ... Aharoni, A. (2011). SHINE transcription factors act redundantly to pattern the archetypal surface of *Arabidopsis* flower organs. *PLoS Genetics*, 7, e1001388.
- Shigenaga, A. M., Berens, M. L., Tsuda, K., & Argueso, C. T. (2017). Towards engineering of hormonal crosstalk in plant immunity. *Current Opinion in Plant Biology*, 38, 164–172.
- Shih, M. L., & Morgan, J. A. (2020). Metabolic flux analysis of secondary metabolism in plants. *Metabolic Engineering Communications*, 10, e00123.
- Shimamura, M. (2016). *Marchantia polymorpha*: Taxonomy, phylogeny and morphology of a model system. *Plant and Cell Physiology*, 57, 230–256.
- Shimura, K., Okada, A., Okada, K., ... Yamane, H. (2007). Identification of a biosynthetic gene cluster in rice for momilactones. *Journal of Biological Chemistry*, 282, 34013–34018.
- Shinmen, Y., Katoh, K., Shimizu, S., ... Yamada, H. (1991). Production of arachidonic acid and eicosapentaenoic acids by *Marchantia polymorpha* in cell culture. *Phytochemistry*, 30, 3255–3260.
- Shu, K., Zhou, W., Chen, F., ... Yang, W. (2018). Absciscic acid and gibberellins antagonistically mediate plant development and abiotic stress responses. *Frontiers in Plant Science*, 9, 416.
- Siddiqui, H., Hayat, S., & Bajguz, A. (2018). Regulation of photosynthesis by brassinosteroids in plants. *Acta Physiologiae Plantarum*, 40, 59.
- Sigel, E. M., Schuettelpelz, E., Pryer, K. M., & Der, J. P. (2018). Overlapping patterns of gene expression between gametophyte and sporophyte phases in the fern *Polypodium amorphum* (Polypodiales). *Frontiers in Plant Science*, 9, 1450.
- Silvera, K., Neubig, K. M., Whitten, W. M., ... Cushman, J. C. (2010). Evolution along the crassulacean acid metabolism continuum. *Functional Plant Biology*, 37, 995–1010.
- Singh, D. P., Jermakow, A. M., & Swain, S. M. (2002). Gibberellins are required for seed development and pollen tube growth in *Arabidopsis*. *The Plant Cell*, 14, 3133–3147.
- Škunca, N., & Dessimoz, C. (2015). Phylogenetic profiling: How much input data is enough? *PLOS ONE*, 10, e0114701.
- Smith, J. A. C., & Winter, K. (1996). Taxonomic distribution of crassulacean acid metabolism. *Crassulacean acid metabolism* (pp. 427–436). Springer, Berlin, Heidelberg.

- Song, S. K., Hofhuis, H., Lee, M. M., & Clark, S. E. (2008). Key divisions in the early arabidopsis embryo require POL and PLL1 phosphatases to establish the root stem cell organizer and vascular axis. *Developmental Cell*, 15, 98–109.
- Song, S. K., Lee, M. M., & Clark, S. E. (2006). POL and PLL1 phosphatases are CLAVATA1 signaling intermediates required for *Arabidopsis* shoot and floral stem cells. *Development*, 133, 4691–4698.
- Soo, R. M., Hemp, J., & Hugenholtz, P. (2019). Evolution of photosynthesis and aerobic respiration in the cyanobacteria. *Free Radical Biology and Medicine*, 140, 200–205.
- Sousa, F. L., Shavit-Grievink, L., & Allen, J. F. (2013). Chlorophyll biosynthesis gene evolution indicates photosystem gene duplication, not photosystem merger, at the origin of oxygenic photosynthesis. *Genome Biology*, 5, 200–216.
- South, P. F., Cavanagh, A. P., Liu, H. W., & Ort, D. R. (2019). Synthetic glycolate metabolism pathways stimulate crop growth and productivity in the field. *Science*, 363, eaat9077.
- Spychalla, J. P., & Desborough, S. L. (1990). Fatty acids, membrane permeability, and sugars of stored potato tubers. *Plant Physiology*, 94, 1207–1213.
- Sriram, G., Fulton, D. B., Iyer, V. V., ... Shanks, J. V. (2004). Quantification of compartmented metabolic fluxes in developing soybean embryos by employing biosynthetically directed fractional ¹³C labeling, two-dimensional [¹³C, ¹H] nuclear magnetic resonance, and comprehensive isotopomer balancing. *Plant Physiology*, 136, 3043–3057.
- Stephan, J., Stegle, O., & Beyer, A. (2015). A random forest approach to capture genetic effects in the presence of population structure. *Nature Communications*, 6, 1–10.
- Stevens, K. A., Wegrzyn, J. L., Zimin, A., ... Langley, C. H. (2016). Sequence of the sugar pine megagenome. *Genetics*, 204, 1613–1626.
- Stirk, W. A., Novák, O., Strnad, M., & Van Staden, J. (2003). Cytokinins in macroalgae. *Plant Growth Regulation*, 41, 13–24.
- Sulpice, R., & McKeown, P. C. (2015). Moving toward a comprehensive map of central plant metabolism. *Annual Review of Plant Biology*, 66, 187–210.
- Sun, J., Okita, T. W., & Edwards, G. E. (1999). Modification of carbon partitioning, photosynthetic capacity, and O₂ sensitivity in *Arabidopsis* plants with low ADP-glucose pyrophosphorylase activity. *Plant Physiology*, 119, 267–276.
- Sun, Y., Pri-Tal, O., Michaeli, D., & Mosquna, A. (2020). Evolution of abscisic acid signaling module and its perception. *Frontiers in Plant Science*, 11, 934.
- Sundell, D., Mannapperuma, C., Netotea, S., ... Street, N. R. (2015). The Plant Genome Integrative Explorer Resource: PlantGenIE.org. *New Phytologist*, 208, 1149–1156.
- Sweetlove, L. J., & Fernie, A. R. (2013). The spatial organization of metabolism within the plant cell. *Annual Review of Plant Biology*, 64, 723–746.
- Sweetlove, L. J., Nielsen, J., & Fernie, A. R. (2017). Engineering central metabolism - a grand challenge for plant biologists. *The Plant Journal*, 90, 749–763.
- Sweetlove, L. J., & Ratcliffe, R. G. (2011). Flux-balance modeling of plant metabolism. *Frontiers in Plant Science*, 2, 38.
- Szovenyi, P., Rensing, S. A., Lang, D., ... Shaw, A. J. (2011). Generation-biased gene expression in a bryophyte model system. *Molecular Biology and Evolution*, 28, 803–812.
- Takemura, K., Kamachi, H., Kume, A., ... Hanba, Y. T. (2017). A hypergravity environment increases chloroplast size, photosynthesis, and plant growth in the moss *Physcomitrella patens*. *Journal of Plant Research*, 130, 181–192.
- Takezawa, D., Komatsu, K., & Sakata, Y. (2011). ABA in bryophytes: How a universal growth regulator in life became a plant hormone? *Journal of Plant Research*, 124, 437–453.

- Takos, A. M., Knudsen, C., Lai, D., ... Rook, F. (2011). Genomic clustering of cyanogenic glucoside biosynthetic genes aids their identification in *Lotus japonicus* and suggests the repeated evolution of this chemical defence pathway. *Plant Journal*, 68, 273–286.
- Tam, T. H. Y., Catarino, B., & Dolan, L. (2015). Conserved regulatory mechanism controls the development of cells with rooting functions in land plants. *Proceedings of the National Academy of Sciences of the United States of America*, 112, E3959–E3968.
- Tanaka, Y., Sugano, S. S., Shimada, T., & Hara-Nishimura, I. (2013). Enhancement of leaf photosynthetic capacity through increased stomatal density in *Arabidopsis*. *New Phytologist*, 198, 757–764.
- Tay, I., Odang, K. B., & Cheung, C. Y. M. (2019). Metabolic modelling of CAM cycling, CAM idling and the C₃ CAM continuum. *BioRxiv*.
- Tcherkez, G. G. B., Farquhar, G. D., & Andrews, T. J. (2006). Despite slow catalysis and confused substrate specificity, all ribulose biphosphate carboxylases may be nearly perfectly optimized. *Proceedings of the National Academy of Sciences of the United States of America*, 103, 7246–7251.
- Ternes, P., Sperling, P., Albrecht, S., ... Heinz, E. (2006). Identification of fungal sphingolipid C9-methyltransferases by phylogenetic profiling. *Journal of Biological Chemistry*, 281, 5582–5592.
- Thomas, R. J., Behringer, F. J., Lombard, C. S., & Sparkowski, J. J. (1984). Effects of auxin on wall polysaccharide composition and enzyme activity during extension-growth of *Pellia* (Bryophyta). *Physiologia Plantarum*, 60, 502–506.
- Thorncroft, D. (2001). Using gene knockouts to investigate plant metabolism. *Journal of Experimental Botany*, 52, 1593–1601.
- Tian, W., & Skolnick, J. (2003). How well is enzyme function conserved as a function of pairwise sequence identity? *Journal of Molecular Biology*, 333, 863–882.
- Toledo-Ortiz, G., Huq, E., & Quail, P. H. (2003). The *Arabidopsis* basic/helix-loop-helix transcription factor family. *The Plant Cell*, 15, 1749–1770.
- Tomescu, A. M. F. (2009). Megaphylls, microphylls and the evolution of leaf development. *Trends in Plant Science*, 14, 5–12.
- Tomitani, A., Okada, K., Miyashita, H., ... Tanaka, A. (1999). Chlorophyll b and phycobilins in the common ancestor of cyanobacteria and chloroplasts. *Nature*, 400, 159–162.
- Töpfer, N., Braam, T., Shameer, S., & Ratcliffe, R. G. (2020). CAM emerges in a leaf metabolic model under water-saving constraints in different environments. *bioRxiv*.
- Töpfer, N., Caldana, C., Grimbs, S., ... Nikoloski, Z. (2013). Integration of genome-scale modeling and transcript profiling reveals metabolic pathways underlying light and temperature acclimation in *Arabidopsis*. *The Plant Cell*, 25, 1197–1211.
- Turk, E. M., Fujioka, S., Seto, H., ... Neff, M. M. (2003). CYP72B1 inactivates brassinosteroid hormones: An intersection between photomorphogenesis and plant steroid signal transduction. *Plant physiology*, 133, 1643–53.
- Tuteja, N. (2007). Absciscic acid and abiotic stress signaling. *Plant Signaling and Behavior*, 2, 135–138.
- Ubeda-Tomás, S., Federici, F., Casimiro, I., ... Bennett, M. J. (2009). Gibberellin signaling in the endodermis controls *Arabidopsis* root meristem size. *Current Biology*, 19, 1194–1199.
- Ubeda-Tomás, S., Swarup, R., Coates, J., ... Bennett, M. J. (2008). Root growth in *Arabidopsis* requires gibberellin/DELLA signalling in the endodermis. *Nature Cell Biology*, 10, 625–628.
- Ursache, R., De Jesus Vieira-Teixeira, C., Tendon, V. D., ... Vermeer, J. (2020). GDSL-domain containing proteins mediate suberin biosynthesis and degradation, enabling developmental plasticity of the endodermis during lateral root emergence. *bioRxiv*.

- Van Der Maaten, L., & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9, 2579–2605.
- Van Ghelder, C., Parent, G. J., Rigault, P., ... MacKay, J. (2019). The large repertoire of conifer nlr resistance genes includes drought responsive and highly diversified nrls. *Scientific Reports*, 9.
- Van Lammeren, A. A. M., Bednara, J., & Willemse, M. T. M. (1989). Organization of the actin cytoskeleton during pollen development in *Gasteria verrucosa* (Mill.) H. Duval visualized with rhodamine-phalloidin. *Planta*, 178, 531–539.
- Varbanova, M., Yamaguchi, S., Yang, Y., ... Picharsky, E. (2007). Methylation of gibberellins by *Arabidopsis* GAMT1 and GAMT2. *The Plant Cell*, 19, 32–45.
- Vergara, A. E., Díaz, K., Carvajal, R., ... Pérez-Donoso, A. G. (2018). Exogenous applications of brassinosteroids improve color of red table grape (*Vitis vinifera* L. Cv. “Redglobe”) berries. *Frontiers in Plant Science*, 9, 363.
- Verma, V., Ravindran, P., & Kumar, P. P. (2016). Plant hormone-mediated regulation of stress responses. *BMC Plant Biology*, 16, 1–10.
- Villarreal, J. C., & Renner, S. S. (2012). Hornwort pyrenoids, carbon-concentrating structures, evolved and were lost at least five times during the last 100 million years. *Proceedings of the National Academy of Sciences of the United States of America*, 109, 18873–18878.
- Vishwanath, S. J., Delude, C., Domergue, F., & Rowland, O. (2015). Suberin: Biosynthesis, regulation, and polymer assembly of a protective extracellular barrier. *Plant Cell Reports*, 34, 573–586.
- Vogt, T. (2010). Phenylpropanoid biosynthesis. *Molecular Plant*, 3, 2–20.
- Voth, P. D. (1943). Effects of nutrient-solution concentration on the growth of *Marchantia polymorpha*. *Botanical Gazette*, 104, 591–601.
- Wada, M., Takahashi, H., Altaf-Ul-Amin, M., ... Kanaya, S. (2012). Prediction of operon-like gene clusters in the *Arabidopsis thaliana* genome based on co-expression analysis of neighboring genes. *Gene*, 503, 56–64.
- Waite, M., & Sack, L. (2010). How does moss photosynthesis relate to leaf and canopy structure? Trait relationships for 10 Hawaiian species of contrasting light habitats. *New Phytologist*, 185, 156–172.
- Wan, T., Liu, Z. M., Li, L. F., ... Wang, X. M. (2018). A genome for gnetophytes and early evolution of seed plants. *Nature Plants*, 4, 82–89.
- Wang, C., Liu, Y., Li, S. S., & Han, G. Z. (2015). Insights into the origin and evolution of the plant hormone signaling machinery. *Plant Physiology*, 167, 872–886.
- Wang, P., Hawkins, T. J., Richardson, C., ... Hussey, P. J. (2014). The plant cytoskeleton, NET3C, and VAP27 mediate the link between the plasma membrane and endoplasmic reticulum. *Current Biology*, 24, 1397–1405.
- Wang, P., Pleskot, R., Zang, J., ... Wang, J. (2019). Plant AtEH/Pan1 proteins drive autophagosome formation at ER-PM contact sites with actin and endocytic machinery. *Nature Communications*, 10, 5132.
- Wang, P., Richardson, C., Hawkins, T. J., ... Hussey, P. J. (2016). Plant VAP27 proteins: Domain characterization, intracellular localization and role in plant development. *New Phytologist*, 210, 1311–1326.
- Wang, X., Ding, J., Lin, S., ... Wu, H. (2020). Evolution and roles of cytokinin genes in angiosperms 2: Do ancient CKXs play housekeeping roles while non-ancient CKXs play regulatory roles? *Nature Horticultural Research*, 7.
- Wang, Z. Y., Bai, M. Y., Oh, E., & Zhu, J. Y. (2012). Brassinosteroid signaling network and regulation of photomorphogenesis. *Annual Review of Genetics*, 46, 701–724.
- Waters, M. T., Gutjahr, C., Bennett, T., & Nelson, D. C. (2017). Strigolactone signaling and evolution. *Annual Review of Plant Biology*, 68, 291–322.

- Webplotdigitizer. (2020). Retrieved June 1, 2020, from <https://automeris.io/WebPlotDigitizer/citation.html>
- Wegrzyn, J. L., Staton, M. A., Street, N. R., ... Ficklin, S. (2019). Cyberinfrastructure to improve forest health and productivity: The role of tree databases in connecting genomes, phenomes, and the environment. *Frontiers in Plant Science*, 10.
- Wei, Z., & Li, J. (2016). Brassinosteroids regulate root growth, development, and symbiosis. *Molecular Plant*, 9, 86–100.
- Weissenborn, S., & Walther, D. (2017). Metabolic pathway assignment of plant genes based on phylogenetic profiling – a feasibility study. *Frontiers in Plant Science*, 8, 1831.
- Weng, J. (2014). The evolutionary paths towards complexity: A metabolic perspective. *The New Phytologist*, 201, 1141–9.
- Weng, J., Akiyama, T., Bonawitz, N. D., ... Chapple, C. (2010). Convergent evolution of syringyl lignin biosynthesis via distinct pathways in the lycophyte *Selaginella* and flowering plants. *The Plant Cell*, 22, 1033–45.
- Weng, J. K., Li, X., Stout, J., & Chapple, C. (2008). Independent origins of syringyl lignin in vascular plants. *Proceedings of the National Academy of Sciences of the United States of America*, 105, 7887–7892.
- Weng, J. K., Philippe, R. N., & Noel, J. P. (2012). The rise of chemodiversity in plants. *Science*, 336, 1667–1670.
- Weraduwage, S. M., Chen, J., Anozie, F. C., ... Sharkey, T. D. (2015). The relationship between leaf area growth and biomass accumulation in *Arabidopsis thaliana*. *Frontiers in Plant Science*, 6, 167.
- Wickett, N. J., Mirarab, S., Nguyen, N., ... Leebens-Mack, J. (2014). Phylotranscriptomic analysis of the origin and early diversification of land plants. *Proceedings of the National Academy of Sciences of the United States of America*, 111, E4859–E4868.
- Wieczorek, K., Hofmann, J., Blöchl, A., ... Grundler, F. M. W. (2007). *Arabidopsis* endo-1,4- β -glucanases are involved in the formation of root syncytia induced by *Heterodera schachtii*. *The Plant Journal*, 53, 336–351.
- Williams, T. C. R., Poolman, M. G., Howden, A. J. M., ... Sweetlove, L. J. (2010). A genome-scale metabolic model accurately predicts fluxes in central carbon metabolism under stress conditions. *Plant Physiology*, 154, 311–323.
- Wilmouth, R. C., Turnbull, J. J., Welford, R. W. D., ... Schofield, C. J. (2002). Structure and mechanism of anthocyanidin synthase from *Arabidopsis thaliana*. *Structure*, 10, 93–103.
- Wink, M. (2008). Plant secondary metabolism: Diversity, function and its evolution. *Natural Product Communications*, 3.
- Winter, K. (2019). Ecophysiology of constitutive and facultative CAM photosynthesis. *Journal of Experimental Botany*, 70, 6495–6508.
- Winter, K., Holtum, J. A. M., & Smith, J. A. C. (2015). Crassulacean acid metabolism: A continuous or discrete trait? *New Phytologist*, 208, 73–78.
- Winter, K., Sage, R. F., Edwards, E. J., ... Holtum, J. A. M. (2019). Facultative crassulacean acid metabolism in a C₃-C₄ intermediate. *Journal of Experimental Botany*, 70, 6571–6579.
- Winzer, T., Gazda, V., He, Z., ... Graham, I. A. (2012). A *Papaver somniferum* 10-gene cluster for synthesis of the anticancer alkaloid noscapine. *Science*, 336, 1704–1708.
- Wolff, P., Jiang, H., Wang, G., ... Köhler, C. (2015). Paternally expressed imprinted genes establish postzygotic hybridization barriers in *Arabidopsis thaliana*. *eLife*, 4, e10074.
- Woodson, J. D., & Chory, J. (2008). Coordination of gene expression between organellar and nuclear genomes. *Nature Reviews Genetics*, 9, 383–395.
- Xie, Z., Zhang, Z. L., Zou, X., ... Shen, Q. J. (2006). Interactions of two abscisic-acid induced WRKY genes in repressing gibberellin signaling in aleurone cells. *Plant Journal*, 46, 231–242.

- Xu, Z., Xin, T., Bartels, D., ... Chen, S. (2018). Genome analysis of the ancient tracheophyte *Selaginella tamariscina* reveals evolutionary features relevant to the acquisition of desiccation tolerance. *Molecular Plant*, 11, 983–994.
- Xue, J. S., Zhang, B., Zhan, H. D., ... Yang, Z. N. (2020). Phenylpropanoid derivatives are essential components of sporopollenin in vascular plants. *Molecular Plant*, 13, 1644–1653.
- Yamazaki, H., Tasaka, M., & Shikanai, T. (2004). PPR motifs of the nucleus-encoded factor, PGR3, function in the selective and distinct steps of chloroplast gene expression in *Arabidopsis*. *The Plant Journal*, 38, 152–163.
- Yan, M., Xie, D., Cao, J., ... Yu, J. (2020). Brassinosteroid-mediated reactive oxygen species are essential for tapetum degradation and pollen fertility in tomato. *The Plant Journal*, 102, 931–947.
- Yang, W., Simpson, J. P., Li-Beisson, Y., ... Ohlrogge, J. B. (2012). A land-plant-specific glycerol-3-phosphate acyltransferase family in *Arabidopsis*: Substrate specificity, *sn-2* preference, and evolution. *Plant Physiology*, 160, 638–652.
- Yang, X., Cushman, J. C., Borland, A. M., ... Holtum, J. A. M. (2015). A roadmap for research on crassulacean acid metabolism (CAM) to enhance sustainable food and bioenergy production in a hotter, drier world. *New Phytologist*, 207, 491–504.
- Yang, Y., Sun, P., Lv, L., ... Liu, J. (2020). Prickly waterlily and rigid hornwort genomes shed light on early angiosperm evolution. *Nature Plants*, 6, 215–222.
- Yang, Y., Yu, Y., Liang, Y., ... Cao, J. (2018). A profusion of molecular scissors for pectins: Classification, expression, and functions of plant polygalacturonases. *Frontiers in Plant Science*, 9, 1208.
- Yokota, T., Ohnishi, T., Shibata, K., ... Kohchi, T. (2017). Occurrence of brassinosteroids in non-flowering land plants, liverwort, moss, lycophyte and fern. *Phytochemistry*, 136, 46–55.
- Yoon, H. S., Hackett, J. D., Ciniglia, C., ... Bhattacharya, D. (2004). A molecular timeline for the origin of photosynthetic eukaryotes. *Molecular Biology and Evolution*, 21, 809–818.
- Yoshida, H., Tanimoto, E., Hirai, T., ... Ueguchi-Tanaka, M. (2018). Evolution and diversification of the plant gibberellin receptor GID1. *Proceedings of the National Academy of Sciences of the United States of America*, 115(33), E7844–E7853.
- Yu, J. Q., Huang, L. F., Hu, W. H., ... Nogués, S. (2004). A role for brassinosteroids in the regulation of photosynthesis in *Cucumis sativus*. *Journal of Experimental Botany*, 55, 1135–1143.
- Yuan, L. B., Peng, Z. H., Zhi, T. T., ... Ren, C. M. (2015). Brassinosteroid enhances cytokinin-induced anthocyanin biosynthesis in *Arabidopsis* seedlings. *Biologia Plantarum*, 59, 99–105.
- Zabackis, E., Huang, J., Müller, B., ... Albersheim, P. (1995). Characterization of the cell-wall polysaccharides of *Arabidopsis thaliana* leaves. *Plant Physiology*, 107, 1129–1138.
- Zhang, H., Zhang, L., Wu, S., ... Chen, L. (2020). AtWRKY75 positively regulates age-triggered leaf senescence through gibberellin pathway. *Plant Diversity*.
- Zhang, L., Gu, L., Ringler, P., ... Shen, Q. J. (2015). Three WRKY transcription factors additively repress abscisic acid and gibberellin signaling in aleurone cells. *Plant Science*, 236, 214–222.
- Zhang, L. Y., Bai, M. Y., Wu, J., ... Wang, Z. Y. (2009). Antagonistic HLH/bHLH transcription factors mediate brassinosteroid regulation of cell elongation and plant development in rice and *Arabidopsis*. *The Plant Cell*, 21, 3767–3780.
- Zhang, Y., Liang, W., Shi, J., ... Zhang, D. (2013). MYB56 encoding a R2R3 MYB transcription factor regulates seed size in *Arabidopsis thaliana*. *Journal of Integrative Plant Biology*, 55, 1166–1178.
- Zhang, Y., Shewry, P. R., Jones, H., ... Halford, N. G. (2001). Expression of antisense SnRK1 protein kinase sequence causes abnormal pollen development and male sterility in transgenic barley. *Plant Journal*, 28, 431–441.

- Zhang, Y., Zhang, B., Yan, D., ... He, Z. (2011). Two *Arabidopsis* cytochrome P450 monooxygenases, CYP714A1 and CYP714A2, function redundantly in plant development through gibberellin deactivation. *The Plant Journal*, 67, 342–53.
- Zheng, Z., Xia, Q., Dauk, M., ... Zou, J. (2003). *Arabidopsis* AtGPAT1, a member of the membrane-bound glycerol-3-phosphate acyltransferase gene family, is essential for tapetum differentiation and male fertility. *The Plant Cell*, 15, 1872–1887.
- Zhu, Y., Nomura, T., Xu, Y., ... He, Z. (2006). ELONGATED UPPERMOST INTERNODE encodes a cytochrome P450 monooxygenase that epoxidizes gibberellins in a novel deactivation reaction in rice. *The Plant Cell*, 18, 442–456.
- Zi, J., Mafu, S., & Peters, R. J. (2014). To gibberellins and beyond! Surveying the evolution of (di)terpenoid metabolism. *Annual Review of Plant Biology*, 65, 259–286.

Appendices



Appendix A: Supplementary figures and tables for Chapter 2

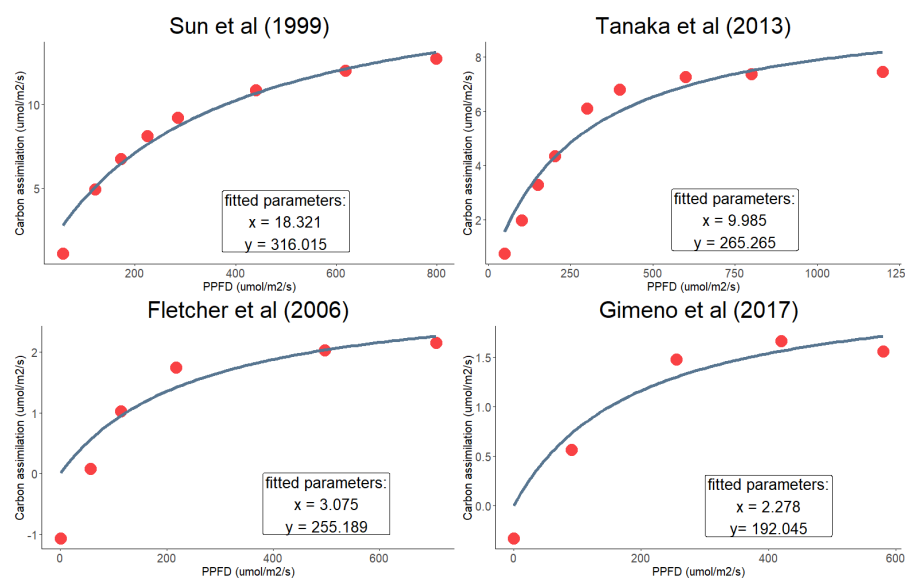


Figure A.1: Results of fitting curves of the form $A = \frac{(x \times PPFD)}{(y + PPFD)}$ to carbon assimilation data for *Arabidopsis* (top row) and *Marchantia* (bottom row). Top left: Sun et al. 1999, top right: Tanaka et al. 2013, bottom left: Fletcher et al. 2006, bottom right: Gimeno et al. 2017. A, photosynthetic carbon assimilation rate; PPFD, photosynthetic photon flux density.

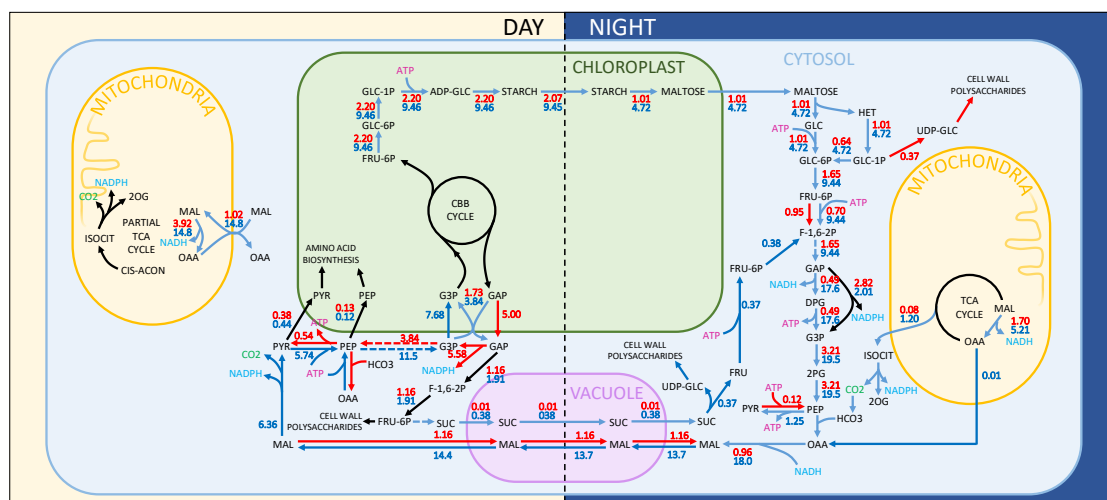
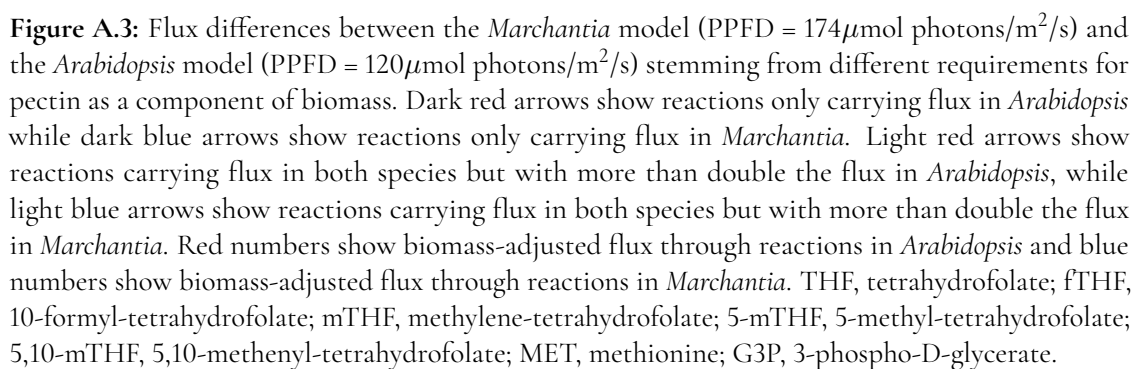
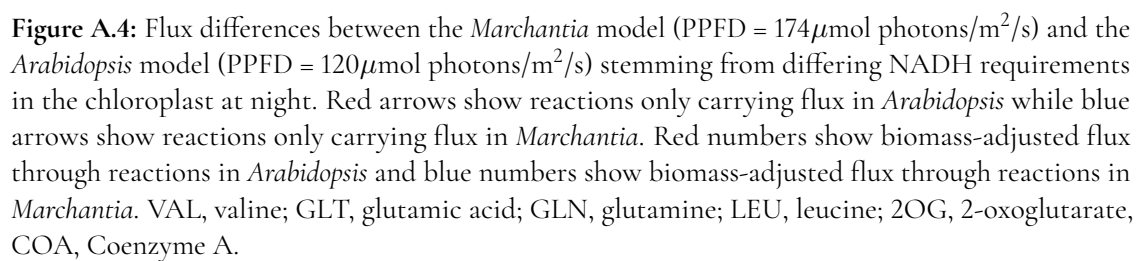


Figure A.2: Illustrating differences in central carbon metabolism between the *Marchantia* model at PPFD = $174 \mu\text{mol photons/m}^2/\text{s}$ and the *Arabidopsis* model at PPFD = $120 \mu\text{mol photons/m}^2/\text{s}$. Red arrows show reactions only carrying flux in *Arabidopsis* while dark blue arrows show reactions only carrying flux in *Marchantia*. Light blue arrows show reactions carrying flux in both species but with more than double the flux in *Marchantia*. Red numbers show biomass-adjusted flux through reactions in *Arabidopsis* and blue numbers show biomass-adjusted flux through reactions in *Marchantia*. FRU-6P, fructose-6-phosphate; GLC-6P, glucose-6-phosphate; GLC-1P, glucose-1-phosphate; ADP-GLC, ADP-glucose; UDP-GLC, UDP-glucose; GLC, glucose; HET, heteroglycans; F-1,6-2P, fructose-1,6-bisphosphate; GAP, glyceraldehyde 3-phosphate; DGP, 3-phospho-D-glyceroyl phosphate; G3P, 3-phospho-D-glycerate; 2PG, 2-phospho-D-glycerate; PEP, phosphoenolpyruvate; MAL, malate; OAA, oxaloacetate; ISOCIT, D-threo-isocitrate; ZOG, 2-oxoglutarate, HCO_3 , hydrogencarbonate; PYR, pyruvate; SUC, sucrose; FRU, fructose; CIS-ACON, cis-aconitate.





Species	Study	PPFD ($\mu\text{mol}/\text{m}^2/\text{s}$)	Photoperiod
<i>Arabidopsis</i>	Rasse & Tocquin, 2006	120	Short day (8 hours)
<i>Arabidopsis</i>	Christophe et al, 2008	92.6	12 hours
<i>Arabidopsis</i>	Cookson et al, 2005	108.8	Short day (10 hours)
<i>Arabidopsis</i>	Cookson et al, 2006	149.3	Long day (14 hours)
<i>Marchantia</i>	Carter & Romine, 1969	174	Short day (8 hours)
<i>Marchantia</i>	Graham & Dixon, 2012	95	Long day (16 hours)
<i>Marchantia</i>			Long day (14 hours)

Table A.1: Sources of experimental data and corresponding experimental conditions for *Arabidopsis* and *Marchantia* validation datasets.

B

Appendix B: Supplementary figures and tables for Chapter 3

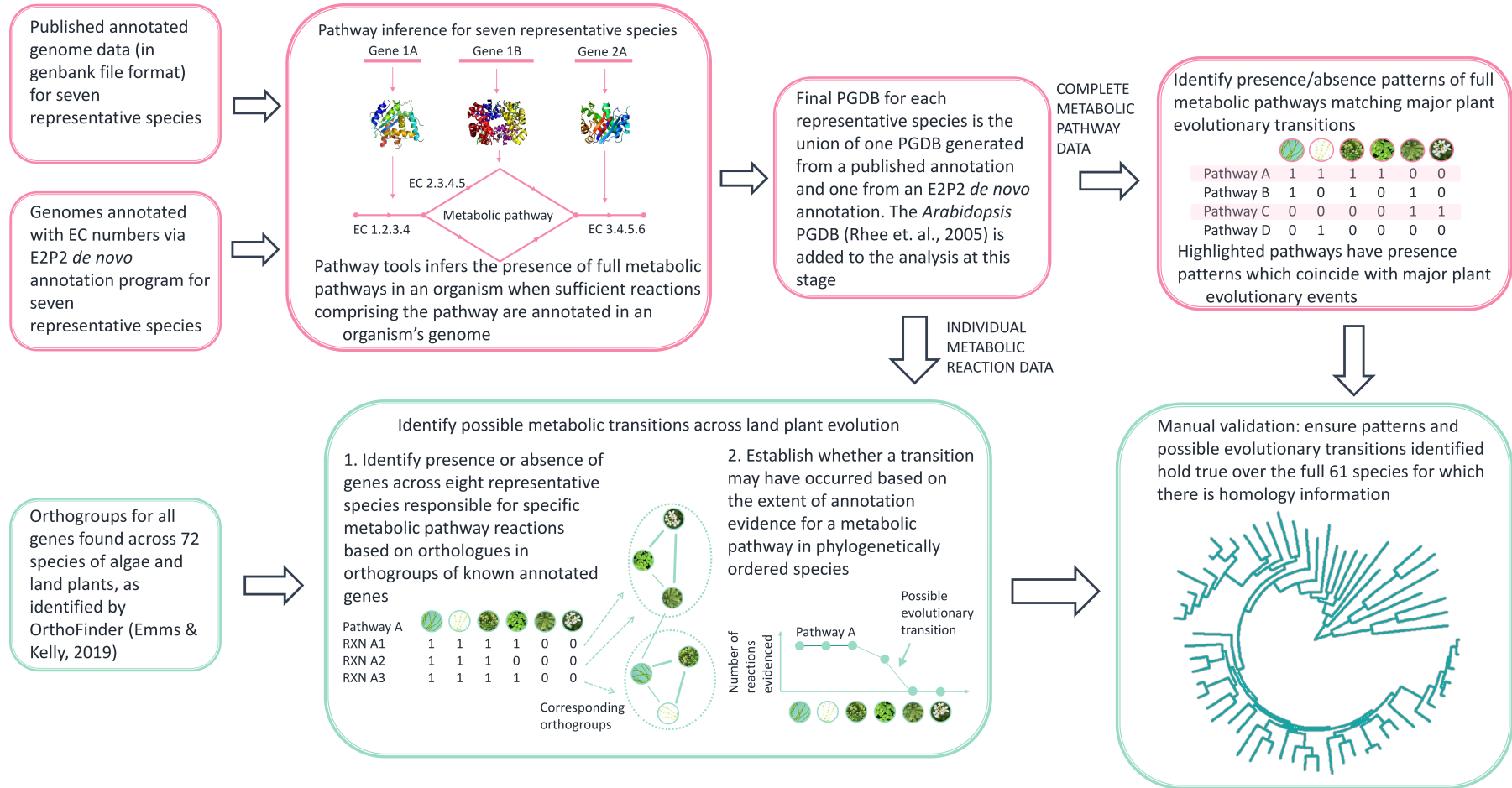


Figure B.1: Overview of methodology for the identification of metabolic innovations and losses. Red: metabolic pathway inference approach. Metabolic pathways for each of seven representative species were inferred from published genome data and genomes annotated using E2P2. Final PGDBs for each species are the union of the metabolic pathways inferred using both input methods. Full inferred metabolic pathways were analysed to identify presence/absence patterns matching major transitions in land plant evolution. Green: enzyme orthogroup approach. OrthoFinder was used to identify orthogroups containing homologous genes across 72 plant and algal species. Reaction-level evidence for each pathway identified in the metabolic pathway inference approach was evaluated to identify possible gains or losses of enzymes across the seven representative species. The potential metabolic developments identified by both approaches were then validated by evaluating reaction evidence across all 72 species.



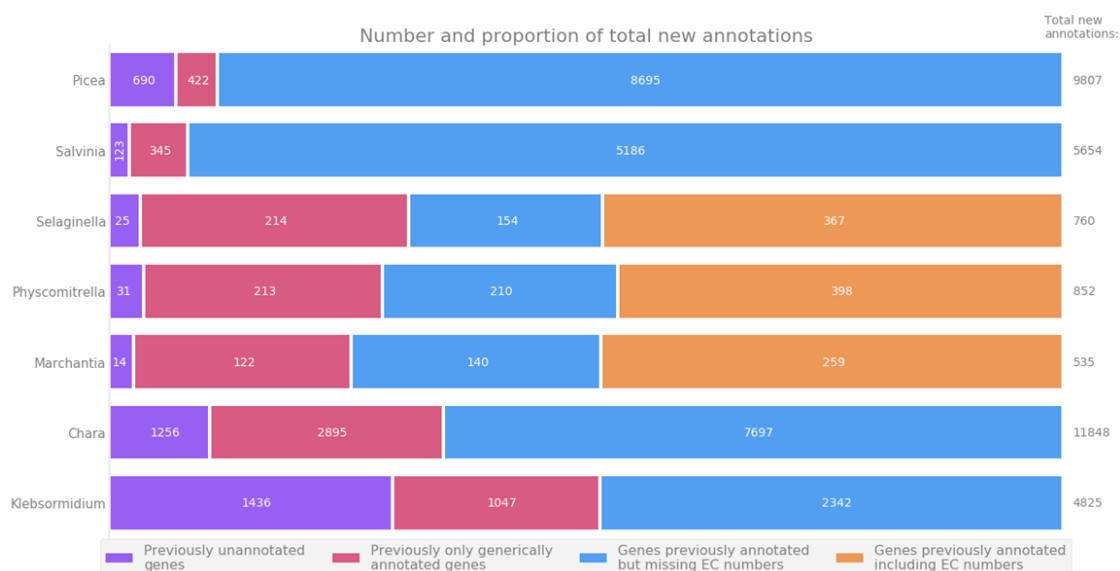


Figure B.3: Number and type of novel EC-number annotations made using E2P2 for seven species. All genes included have new EC number associations. Unannotated genes (purple) are previously identified only as open reading frames. Generically annotated genes (pink) are previously annotated non-specifically, e.g. 'putative', 'superfamily', 'homologue of'. Annotated genes missing EC numbers (blue) are previously annotated with specific functional IDs such as Pfam, InterPro or GO but lacking reaction associations with EC numbers. Annotated genes including EC numbers (orange) are those previously annotated with EC numbers but for which new additional EC numbers have been identified. EC number annotations for each species are available in Data 3.1.

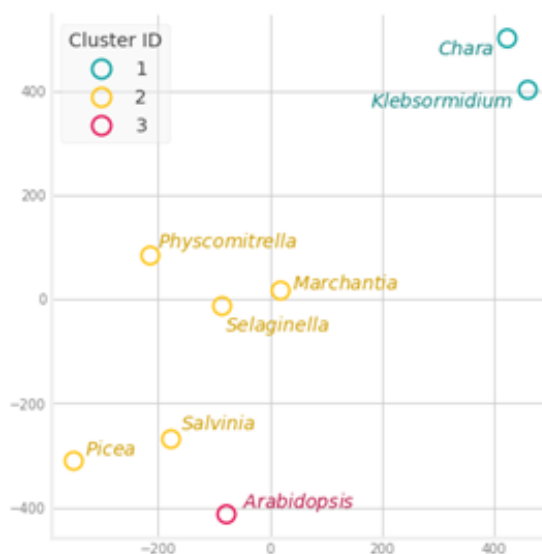


Figure B.4: Results of k -medoids clustering and t-SNE analysis of representative species, based on the presence and absence of all metabolic pathways identified in any PGDB. The colours of the species data points indicate their cluster as assigned by k -medoids clustering, while the relative locations are the result of t-SNE. Axis scales are arbitrary numbers indicating the relative difference between different species PGDBs. PGDBs for each species are available in Data 3.1.

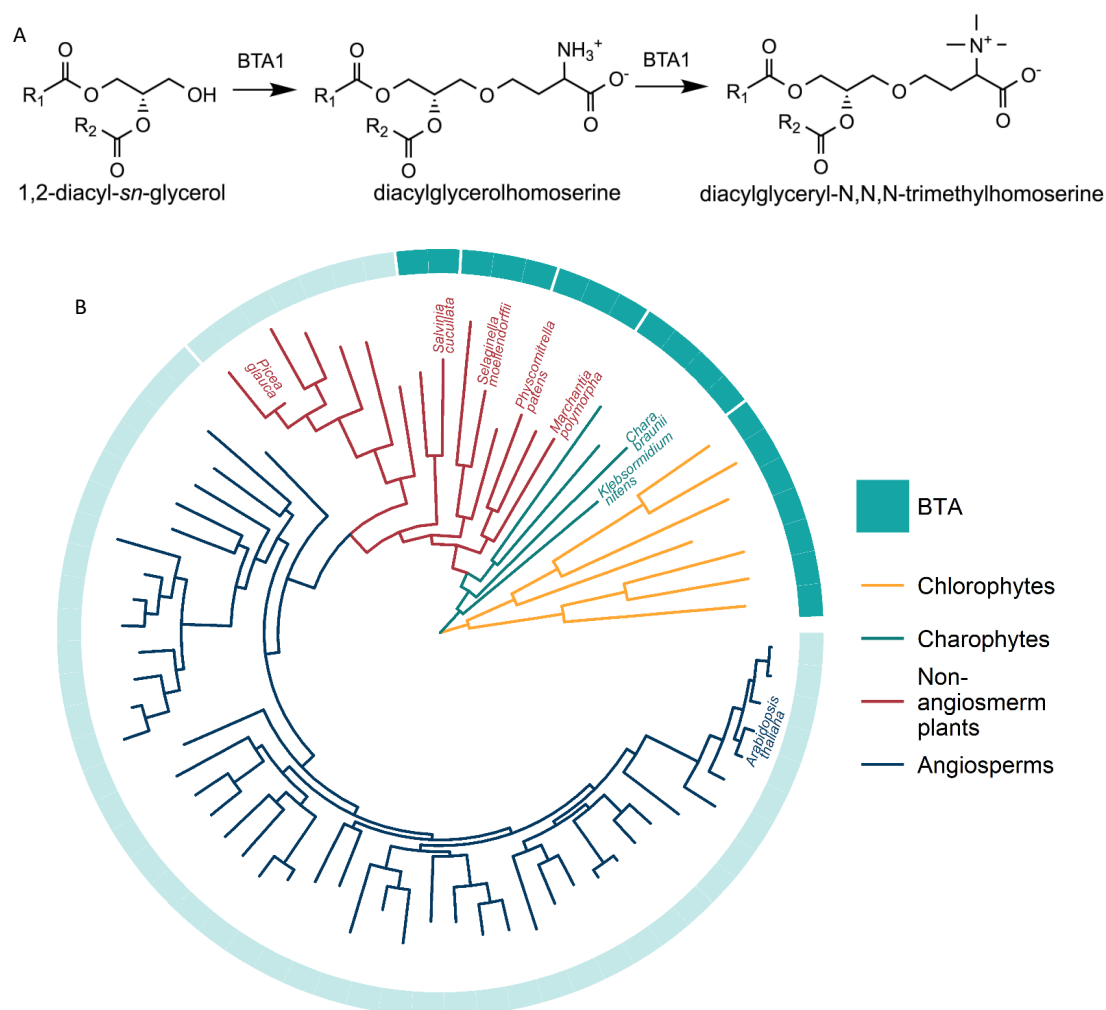


Figure B.5: Presence of gene homologues encoding diacylglyceryl-N,N,N-trimethylhomoserine biosynthesis enzyme across 72 land plant and algal species. **A:** The known biosynthetic pathway for DGTS production. **B:** Phylogeny showing the presence of homologues of genes annotated as DGTS biosynthetic enzymes across the 72 species analysed. For each gene, darker colouring indicates homologue presence while lighter colouring indicates homologue absence. BTA, betaine lipid synthase.

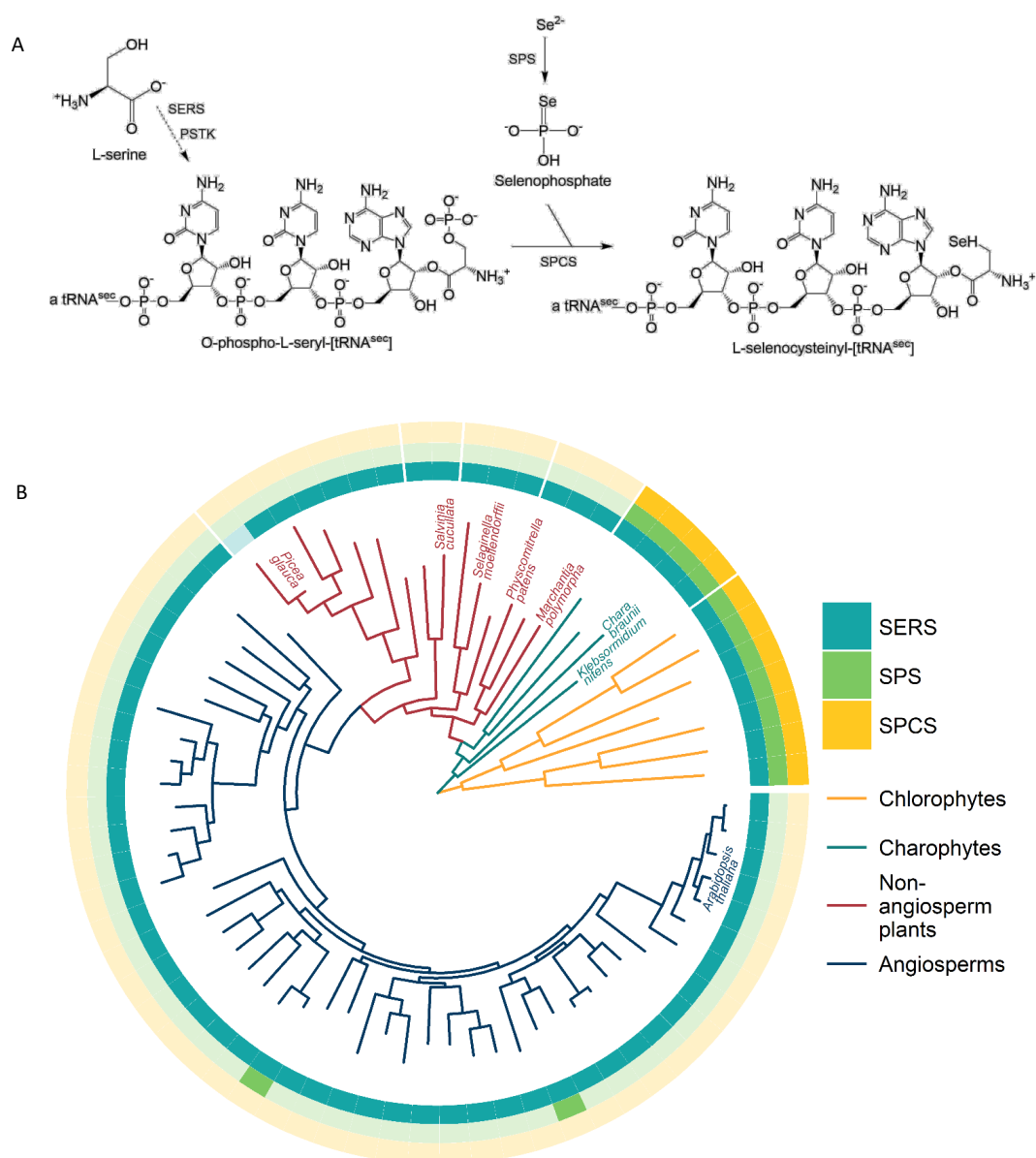


Figure B.6: Presence of gene homologues encoding selenocysteine biosynthesis enzymes across 72 land plant and algal species. *A*: The selenocysteine biosynthesis pathway. *B*: Evidence for homologues of genes involved in selenocysteine biosynthesis across 72 species. Dark colours indicate the presence of a homologue while light colours indicate absence. SERS, seryl-tRNA synthetase; SPS, selenophosphate synthetase; SPCS, selenocysteine synthetase.

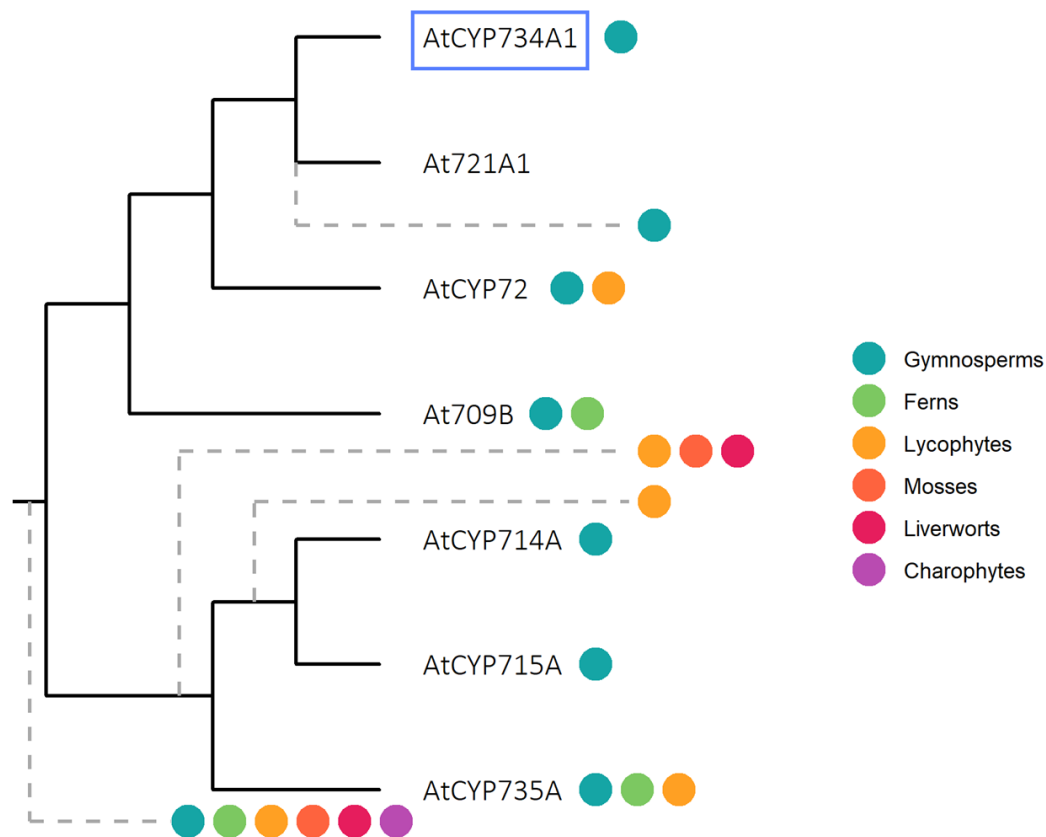


Figure B.7: Simplified gene tree detailing the relationships between *Arabidopsis* genes in the orthogroup containing a known brassinosteroid hydroxylation enzyme. The *Arabidopsis* enzyme in question is highlighted with a blue box, while the positions of homologous genes in other species are shown by coloured spots (connected by a dotted line when they do not fit clearly with a single set of *Arabidopsis* genes). This representation is based on the full gene tree for orthogroup OG0000029 (Data 3.3).

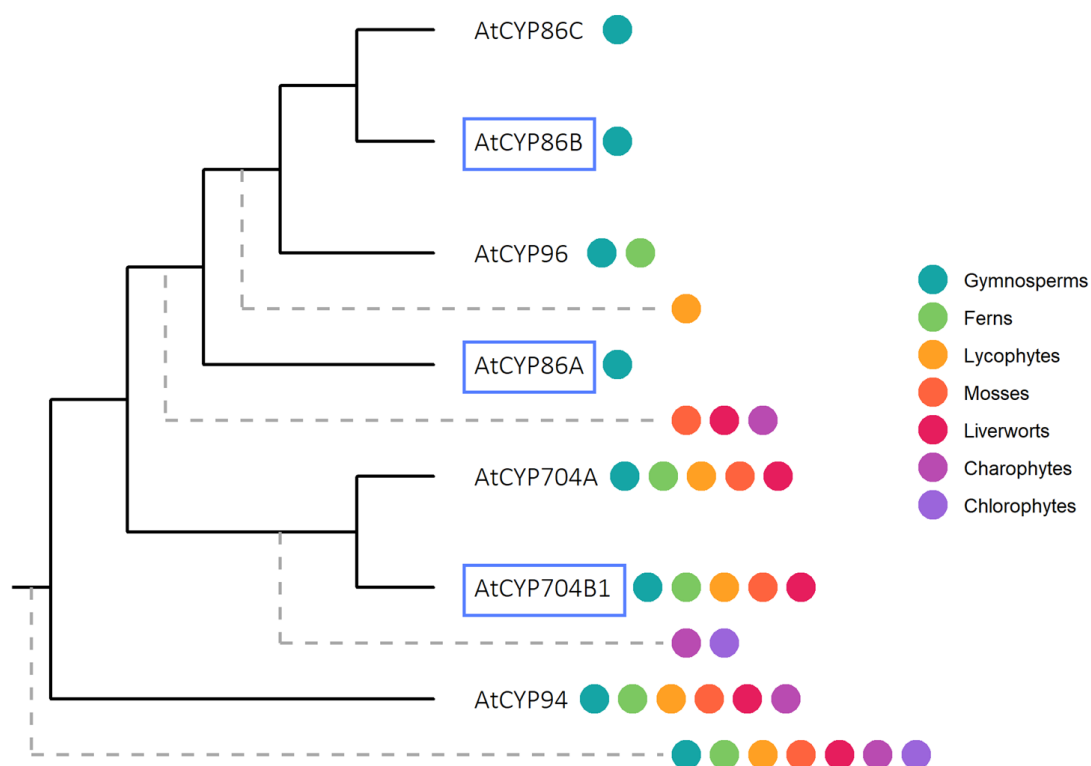


Figure B.8: Simplified gene tree detailing the relationships between *Arabidopsis* genes in the orthogroup containing CYP86 and CYP704 enzymes. The *Arabidopsis* enzymes in question are highlighted with a blue box while the positions of homologous genes in other species are shown by coloured spots (connected by a dotted line when they do not fit clearly with a single set of *Arabidopsis* genes). This representation is based on the full gene tree for orthogroup OG0000031 (Data 3.3).

Metabolic Pathway	Gene ID	Arabidopsis/Chlamydomonas accession	Orthogroup
Gibberellin biosynthesis	KO	AT5G25900	OG0003245
	KAO	AT1G05160; AT2G32440	OG0000021
	GA20ox	AT5G07200; AT5G51810; AT1G60980	OG0000903
	GA3ox	AT1G15550; AT1G80340	OG0001512
	GA2ox19	AT2G34555; AT1G30040	OG0000874
	GA2ox20	AT4G21200; AT1G50960	OG0002247
	CYP714	AT5G24910; AT5G24900	OG0000031
	GAMT	AT4G26420; AT5G56300	OG0000092
Brassinosteroid biosynthesis	DET2	AT2G38050	OG0006889
	CYP85A	AT5G38970; AT3G30180	OG0000021
	CYP90A1	AT5G05690	OG0000021
	CYP90B1	AT3G50660	OG0000021
	CYP90C/D	AT4G36380; AT3G13730	OG0000021
	CYP734A1	AT2G26710	OG0000031
Cutin/Suberin biosynthesis	LACS	AT2G47240; AT1G64400; AT1G49430	OG0001084
	CYP86A	AT5G58860; AT2G45970	OG0000029
	HOTHEAD	AT1G72970	OG0000268
	CYP77A	AT3G10570; AT5G04660	OG0000363
	GPAT4/6/8	AT1G01610; AT2G38110; AT4G00400	OG0000338
	KCS2/20	AT1G04220; AT5G43760	OG0000123
	FAR1/4/5	AT5G22500; AT3G44540; AT3G44550	OG0000321
	GPAT5/7	AT3G11430; AT5G06090	OG0000338
Sporopollenin biosynthesis	ACH	AT1G01710	OG0003066
	CYP703A2	AT1G01280	OG0000008
	CYP704B1	AT1G69500	OG0000029
	ACOS5	AT1G62940	OG0008202
	FAR2	AT3G11980	OG0000321
	TKPR	AT4G35420; AT1G68540	OG0000048
	PKS	AT1G02050; AT4G34850	OG0000170
Glucosinolate biosynthesis	CYP79F	AT1G16410; AT1G16400	OG0015239
	CYP83A1	AT4G13770	OG0000008
	GGP1	AT4G30530	OG0001274
	GGT4	AT4G29210	OG0001569
	SUR1	AT2G20610	OG0000871
	UGT74B1	AT1G24100	OG0000113
	SOT	AT1G18590; AT1G74090	OG0000174
	GS-OX	AT1G65860; AT1G12140	OG0000686
DGTS biosynthesis	BTA	Cre07.g324200.t1.2	OG0009410
Selenocysteine biosynthesis	SERS	Cre03.g189400.t1.2	OG0002977
	SPS	Cre13.g562501.t1.1	OG0012387
	SPCS	Cre12.g547750.t1.2	OG0014814

Table B.1: Orthogroup membership and *Arabidopsis* or *Chlamydomonas* accession numbers for all metabolic pathways discussed in Figures 3.2-3.6 and B.5-B.6.

Metabolic pathway name	MetaCyc/PlantCyc ID
(E,E)-4,8,12-trimethyltrideca-1,3,7,11-tetraene biosynthesis	PWY-6668
2,3-cis-flavanols biosynthesis	PWY-6035
adenosylcobalamin salvage from cobalamin	PWY-6268
benzoyl- β -D-glucopyranose biosynthesis	PWY-7468
β -caryophyllene biosynthesis	PWY-6275
brassinosteroid biosynthesis I	PWY-699
brassinosteroids inactivation	PWY-6546
camalexin biosynthesis	CAMALEXIN-SYN
CDP-diacylglycerol biosynthesis I	PWY-5667
cinnamate esters biosynthesis	PWY-5968
cutin biosynthesis	PWY-321
cytokinins 7-N-glucoside biosynthesis	PWY-2881
cytokinins 9-N-glucoside biosynthesis	PWY-2901
cytokinin-O-glucosides biosynthesis	PWY-2902
diacylglycerol-N,N,N-trimethylhomoserine biosynthesis	PWY-6795
ethylene biosynthesis I (plants)	ETHYL-PWY
farnesene biosynthesis	PWY-5725
flavonoid biosynthesis (in equisetum)	PWY-6787
GA ₁₂ biosynthesis	PWY-5034
gibberellin biosynthesis I (non C-3, non C-13 hydroxylation)	PWY-5070
gibberellin biosynthesis II (early C-3 hydroxylation)	PWY-5036
gibberellin biosynthesis III (early C-13 hydroxylation)	PWY-5035
gibberellin inactivation I (2 β -hydroxylation)	PWY-102
glucosinolate biosynthesis from homomethionine	PWY-1187
glucosinolate biosynthesis from phenylalanine	PWY-2821
glucosinolate biosynthesis from tryptophan	PWY-601
glucosinolate biosynthesis from trihomomethionine	PWYQT-4472
glucosinolate biosynthesis from tetrahomomethionine	PWYQT-4473
glucosinolate biosynthesis from pentahomomethionine	PWYQT-4474
glucosinolate biosynthesis from hexahomomethionine	PWYQT-4475
hydroxyjasmonate sulfate biosynthesis	PWY-6235
L-asparagine degradation I	ASPARAGINE-DEG1-PWY
L-lysine degradation I	PWY0-461
L-selenocysteine biosynthesis II (archaea and eukaryotes)	PWY-6281
methyl indole-3-acetate interconversion	PWY-6303
monoterpene biosynthesis	PWY-3041
oleoresin monoterpene volatiles biosynthesis	PWY-5423
phenolic malonylglucosides biosynthesis	PWY-6930
phenylpropanoid biosynthesis	PWY-361
phenylpropanoid biosynthesis, initial reactions	PWY1F-467
salicylate glucosides biosynthesis III	PWY-6624
salicylate glucosides biosynthesis IV	PWY-6762

Metabolic pathway name	MetaCyc/PlantCyc ID
sulfide oxidation III (persulfide dioxygenase)	PWY-5285
thalianol and derivatives biosynthesis	PWY-5992
trans-zeatin biosynthesis	PWY-2681
UDP-D-apiose biosynthesis (from UDP-D-glucuronate)	PWY-5113
UDP-N-acetylmuramoyl-pentapeptide biosynthesis II (lysine-containing)	PWY-6386
vernolate biosynthesis I	PWY-6014
vernolate biosynthesis II	PWY-7634
xylogalacturonan biosynthesis	PWY-5980

Table B.2: MetaCyc/PlantCyc IDs for every metabolic pathway identified as a metabolic innovation or loss over the course of land plant evolution.

C

Appendix C: Publication of Chapter 3

Current Biology

Multiple Metabolic Innovations and Losses Are Associated with Major Transitions in Land Plant Evolution

Highlights

- Angiosperm-like brassinosteroid biosynthetic pathway evolved in a stepwise manner
- Gibberellin biosynthesis and inactivation genes are present in bryophytes
- This suggests bryophytes may contain as-yet-undiscovered bioactive gibberellins
- Known cutin/suberin biosynthetic genes are absent from non-angiosperm plants

Authors

Naomi Cannell, David M. Emms, Alexander J. Hetherington, John MacKay, Steven Kelly, Liam Dolan, Lee J. Sweetlove

Correspondence

lee.sweetlove@plants.ox.ac.uk

In Brief

In a comparative analysis of genes in 72 plants and algae, Cannell et al. identify innovations and losses in metabolic capabilities for biosynthesis/inactivation of phytohormones and biosynthesis of structural polymers over the course of land plant evolution. Analysis of transcriptomes from 305 non-angiosperm plants supports the findings.



Multiple Metabolic Innovations and Losses Are Associated with Major Transitions in Land Plant Evolution

Naomi Cannell,¹ David M. Emms,¹ Alexander J. Hetherington,¹ John MacKay,¹ Steven Kelly,¹ Liam Dolan,¹ and Lee J. Sweetlove^{1,2,*}

¹Department of Plant Sciences, University of Oxford, South Parks Road, Oxford OX1 3RB, UK

²Lead Contact

*Correspondence: lee.sweetlove@plants.ox.ac.uk

<https://doi.org/10.1016/j.cub.2020.02.086>

SUMMARY

Investigating the evolution of plant biochemistry is challenging because few metabolites are preserved in fossils and because metabolic networks are difficult to experimentally characterize in diverse extant organisms. We report a comparative computational approach based on whole-genome metabolic pathway databases of eight species representative of major plant lineages, combined with homologous relationships among genes of 72 species from streptophyte algae to angiosperms. We use this genomic approach to identify metabolic gains and losses during land plant evolution. We extended our findings with additional analysis of 305 non-angiosperm plant transcriptomes. Our results revealed that genes encoding the complete biosynthetic pathway for brassinosteroid phytohormones and enzymes for brassinosteroid inactivation are present only in spermatophytes. Genes encoding only part of the biosynthesis pathway are present in ferns and lycophytes, indicating a stepwise evolutionary acquisition of this pathway. Nevertheless, brassinosteroids are ubiquitous in land plants, suggesting that brassinosteroid biosynthetic pathways differ between earlier- and later-diverging lineages. Conversely, genes for gibberellin biosynthesis and inactivation using methyltransferases are found in all land plant lineages. This suggests that bioactive gibberellins might be present in bryophytes, although they have yet to be detected experimentally. We also found that cytochrome P450 oxidases involved in cutin and suberin production are absent in genomes of non-angiosperm plants that nevertheless do contain these biopolymers. Overall, we identified significant differences in crucial metabolic processes between angiosperms and earlier-diverging land plants and resolve details of the evolutionary history of several phytohormone and structural polymer biosynthetic pathways in land plants.

INTRODUCTION

Some time before 450 million years ago, a single lineage of streptophyte algae colonized land, and the subsequent evolutionary radiation gave rise to the embryophytes, which are the dominant photosynthetic organisms in the terrestrial flora [1]. Extant embryophytes display a diverse range of morphologies, physiologies, and biochemistries. The evolution of this diversity had enormous impact on the terrestrial biotic and abiotic environment, affecting nutrient cycles and hydrology and modifying earth sediments and atmosphere [2]. Many of the adaptations that accompanied the colonization of land involved evolution of new metabolic capabilities [3]. Examples include metabolic pathways for the biosynthesis of phytohormones coordinating plant growth, of specialized metabolites that defend against pathogens and provide tolerance of abiotic stresses, and of structural polymers such as lignin and suberin that provide support to stems and roots [4–6]. This diversity of metabolic enzymes is a consequence of substantial adaptive gene family expansion [7].

Investigating the full complement of metabolic pathways in an organism is challenging. Much of our current knowledge of plant metabolism is the result of experimental characterization of the kinetic properties of purified enzymes or *in vivo* metabolic flux analysis using isotope tracers [8]. The laborious nature of these approaches means that we lack a complete picture of metabolism for any given species, even for well-studied angiosperms. Moreover, the focus on a few model species and crops means that current knowledge of plant metabolism is heavily skewed toward angiosperms [9], which means that we have only a partial picture of the evolution of plant metabolism over the course of land plant history.

Recently, *in silico* systems approaches have been applied to plant metabolism [10, 11], facilitating the development of computational representations of entire metabolic networks from genome sequences. Furthermore, the availability of transcriptomes and whole-genome sequences of algae and earlier-diverging land plants has provided researchers with valuable resources for metabolic modeling [12–14]. These data are a source of metabolic information that remains largely uninterrogated. A single study has analyzed the evolution of metabolism from genome sequence data, revealing that the main metabolic innovations after the appearance of vascular plants relate to



specialized metabolism [15]. However, the study analyzed just 16 species providing a sparse sample of plant evolution and only provided a top-level overview of metabolic pathway types, not specific functions.

Here, we exploit a wider range of genome sequence and transcriptome data to investigate in detail the evolution of metabolism in the Chloroplastida. A comparative approach was used to identify metabolic innovations and losses occurring during the evolution of streptophyte algae and land plants. Our starting point was the generation of genome-scale metabolic pathway databases for seven species spanning from algae to land plants (two charophytes, two bryophytes, a lycophyte, a monilophyte, and a gymnosperm). From these databases, and an additional pre-existing database for *Arabidopsis* [16], the presence and absence of the full range of known metabolic pathways (and therefore the appearance and disappearance of various metabolic traits over the course of land plant evolution in these species) was inferred. Individual metabolic gene annotations from these databases were then used in combination with information on homologous genes in other organisms to confirm inferred metabolic innovations or losses and extend their supporting evidence across 64 further plant and algal genomes (bringing the total number of species analyzed to 72). To provide further support for metabolic innovations, an additional 305 non-angiosperm transcriptomes were analyzed. The identification of several known evolutionary innovations in plant metabolism validates the method. The analysis revealed new information on the occurrence of biosynthetic enzymes for phytohormones and structural compounds across land plants and algae.

RESULTS

Bioinformatic Approach and Justification

The metabolic capabilities of an organism can be inferred from its genome sequence. However, even when the genome is well annotated, this is a time-consuming process, involving complex algorithms to infer the presence or absence of whole metabolic pathways. Therefore, rather than attempting to make a detailed comparison of metabolic pathways present in all plant species for which genome sequences are available, the initial comparison was simplified to a smaller subset of species representing each of the major lineages in plant evolution, i.e., charophyte algae, bryophytes, lycophytes, ferns, gymnosperms, and angiosperms. The choice of these organisms was made to maximize the breadth of plant phylogeny covered while maintaining a feasible number of datasets for analysis. Since well-curated information on *Arabidopsis* metabolism is already available [16], efforts here were focused on earlier-diverging land plants and algae and directed mainly based on the availability of sequenced genomes, which are sparse for such organisms. In fact, sequence data for four of the included species have only become available in the last 2 years. Two streptophyte algae genomes have been sequenced. Both—*Klebsormidium nitens* [13] and *Chara braunii* [17]—were included in this analysis. Two bryophytes were chosen: the moss *Physcomitrella patens* [12], a model organism and the bryophyte with the most well-developed genome annotation, and *Marchantia polymorpha* [14], an emerging model organism. *Selaginella moellendorffii* [18] was selected as a representative lycophyte and the fern *Salvinia cucullata* [19] was selected as a representative monilophyte.

Finally, *Picea glauca* was selected as a representative gymnosperm [20]. Note that due to concerns about the reliability of assembly of the conifer megagenome, we used transcriptome data for this species.

Metabolic pathways from the MetaCyc [21] and PlantCyc [22] databases were identified from the genome sequences of these seven species using the Pathway Tools software [23]. To increase the robustness of the analysis, two parallel Pathway Tools analyses were carried out, one based on annotation files compiled from each species' published genome annotation, and one using EC number annotations generated from E2P2, a machine learning-based algorithm for metabolic annotation [22]. The union of the predicted sets of pathways from these two approaches formed the final pathway/genome database (PGDB) for each species. Possible metabolic innovations and losses were inferred when the number of genes encoding enzymes associated with metabolic pathways either increased or decreased at any point, suggesting the gain or loss of the metabolic pathway at the branchpoint in question.

Because the eight species are only a representative sample of a large phylogenetic space, each identified metabolic gain or loss across the eight species was mapped onto genetic data for 72 phylogenetically ordered species from the Chloroplastida (a full, labeled phylogeny of all included species is provided as Figure S1). The species investigated consist of 46 angiosperms, seven gymnosperms (*Ginkgo biloba* [24], *Gnetum montanum* [25], *Pseudotsuga menziesii* [26], *Pinus lambertiana* [27], *Pinus taeda* [28], *Picea glauca* [20], *Picea abies* [29]), two ferns (*Salvinia cucullata* [19], *Azolla filiculoides* [19]), three lycophytes (*Selaginella moellendorffii* [18], *Selaginella tamariscina* [30], *Isoetes echinospora* (unpublished data; GenBank: GGKY000000000.1), two mosses (*Physcomitrella patens* [12], *Sphagnum fallax* [31]), one liverwort (*Marchantia polymorpha* [14]), four charophytes (*Klebsormidium nitens* [13], *Chara braunii* [17], *Spiroglaea muscicola* [32], *Mesotaenium endlicherianum* [32]), and seven chlorophytes (*Ostreococcus lucimarinus*, *Micromonas* sp. RCC299, *Micromonas pusilla* CCMP1545, *Coccomyxa subellipsoidea* C-169, *Chromochloris zofingiensis*, *Volvox carteri*, *Chlamydomonas reinhardtii* [31]). For 70 of these species, genome sequences were used; for *Picea glauca* and *Isoetes echinospora*, transcriptome data were used. All genes from the 72 species were placed into orthogroups using OrthoFinder software [33]. Each of the metabolic gains/losses identified from the eight-species comparison was then tracked across the 72-species orthogroups using gene trees. In the following sections, only metabolic innovations and losses that held true beyond the eight representatives of the major evolutionary groups are presented. To provide additional support for results that relate to non-angiosperms (which are under-represented in terms of genome availability), 305 non-angiosperm land plant and algal transcriptomes taken from the One Thousand Plant Transcriptome Initiative [34] were analyzed to identify homologs of the genes required for the presented metabolic pathways in these species.

De Novo Genome Annotation Provides Gene-to-Reaction Associations for Previously Unannotated Metabolic Genes

EC number annotations identified using E2P2 were compared to published genome annotations for each of the seven

non-angiosperm land plants and algae for which PGDBs were generated. Novel metabolic gene annotations from E2P2 were identified in all seven species, both in the form of completely new annotations for genes previously described only as open reading frames and as additional annotations for genes already associated with another function or metabolic reaction. An overview of the number and classification (i.e., completely new annotations or additional annotations) of novel metabolic gene annotations for each species is shown in Figure S2, and tables containing EC number and MetaCyc reaction ID annotations produced by E2P2 for each species are provided in Data S1, which also contains full PGDBs for each species based on the union of both annotations, provided as lists of MetaCyc/PlantCyc metabolic pathway IDs. Additionally, orthogroups identified using OrthoFinder across the complete set of 72 species is provided in Data S2, and the gene trees for each orthogroup are provided in Data S3. A Python script allowing this information to be queried based on associations between MetaCyc/PlantCyc reaction identifiers and *Arabidopsis* genes is supplied as Data S4.

Inferred Metabolic Pathways Capture Known Phylogenetic Relationships

To confirm that the PGDBs of representative species contain information applicable to the study of plant evolution, k-medoids clustering followed by a dimensionality reduction technique known as t-SNE [35] was carried out to generate a cluster map of metabolic pathway similarity among the species to examine how this relates to phylogeny. The analysis grouped species based on the presence and absence of the metabolic pathways in each species' PGDB. As can be seen in Figure S3, the green algae form a two-member cluster distinct from the land plants. Terrestrial plants were split into two clusters, one containing all land plants up to and including *Picea* and the other containing only *Arabidopsis* (Figure S3). This isolation of *Arabidopsis* is potentially a consequence of the presence of angiosperm-specific metabolic pathways in its PGDB. However, it may also be that the difference is influenced by the level of manual curation involved in the production of the *Arabidopsis* PGDB—the other PGDBs developed in this analysis were not manually curated. Nevertheless, the nearest neighbors of *Arabidopsis* as placed by t-SNE on this graph are the two next most recently diverging land plants, the euphyllophytes *Picea* and *Salvinia*. The two bryophytes, *Physcomitrella* and *Marchantia*, clustered together and were grouped with the earliest diverging member of the vascular plants, the lycophyte *Selaginella*.

Overall, a clear distinction between earlier-diverging and later-diverging lineages from streptophyte algae to angiosperms can be seen on the graph. Based only on the presence and absence of metabolic pathways in each species' PGDB, it was thus possible to capture the known phylogenetic relationships between the plant species analyzed (Figure S3). This indicates that the metabolic information gathered and included in the PGDBs reflects true differences between the species and supports the use of such information in the analysis of the evolution of metabolism.

Comparative Analysis Identifies Previously Identified Metabolic Innovations and Losses in Land Plant Evolution

Orthogroup analysis identified 50 metabolic pathways that showed gains or losses over the course of land plant evolution

(Figure 1). To prioritize metabolic pathways for further investigation, we considered the novelty of the observation and the number of reactions in the pathway. The latter criterion was used because some of the metabolic “pathways” as categorized by Pathway Tools are extremely short, containing only one or two reactions and these were not considered further. The 17 pathways highlighted by dark blue circles in Figure 1 were the ones that were further investigated. Table S1 contains orthogroup and gene IDs for each of the pathways investigated in the following sections.

The reliability of the approach was demonstrated by the fact that three known metabolic changes during land plant evolution were correctly identified. The first example is the metabolic pathway for diacylglycerol-N,N,N-trimethylhomoserine (DGTS) biosynthesis, which is thought to have been lost in spermatophytes where DGTS has been replaced by phosphatidylcholine [36–38]. Our analysis picked out this metabolic pathway because of the presence of genes encoding the required enzyme only in the genomes of algae, bryophytes, lycophytes, and ferns (Figures 1 and S4). Homologous genes were not identified in any gymnosperm or angiosperm species. This pattern is consistent with the previous identification of DGTS in ferns but not seed plants [36, 37].

The second example is the capacity for biosynthesis of selenocysteine, which is considered an ancestral character, found in bacteria, mammals, and green algae but lost in land plants [39]. The pathway of selenocysteine biosynthesis was identified by our approach as a metabolic loss during land plant evolution (Figure 1). The gene encoding the first enzyme in the pathway is found in all plant taxa analyzed, while the genes encoding the full combination of enzymes required for this pathway are found only in the algae (Figure S5).

Finally, our analysis identified several metabolic pathways involved in glucosinolate biosynthesis from various amino acids as a metabolic capability that has been gained during land plant evolution (Figure 1). We found only one or two genes encoding the involved enzymes in algae and non-angiosperm plants and never the complete pathway (Figure 2B). The genes were more commonly found in angiosperms, but the complete pathway was only encoded by the genomes of the Brassicaceae in our analysis (Figure 2). This is in agreement with the hypothesis that glucosinolates are a synapomorphy of the Capparales [40–42]. The analysis shown in Figure 2 also reveals a more detailed picture of the evolution of glucosinolate biosynthesis. The presence of genes for several different glucosinolate-related enzymes across the angiosperms is unsurprising, given that several of these encoding genes are members of large gene families (e.g., glucosyltransferases and monooxygenases) and the evolution of glucosinolate-specific enzymes likely evolved from similar pre-existing pathways such as cyanogenic glucosides production [43]. However, the presence of genes encoding the enzyme glucosinolate γ -glutamyl peptidase (GGP) in gymnosperms and flavin-containing monooxygenases (GS-OX) in species of every clade from chlorophyte algae to angiosperms suggests a conserved ancient function for these genes outside of the biosynthesis of glucosinolates.

Biosynthetic Capabilities for Gibberellin and Brassinosteroid Production Are Encoded in the Genomes of Non-angiosperm Plants

The appearance of enzymatic capabilities for synthesis of certain plant hormones has been linked to evolutionary adaptations to

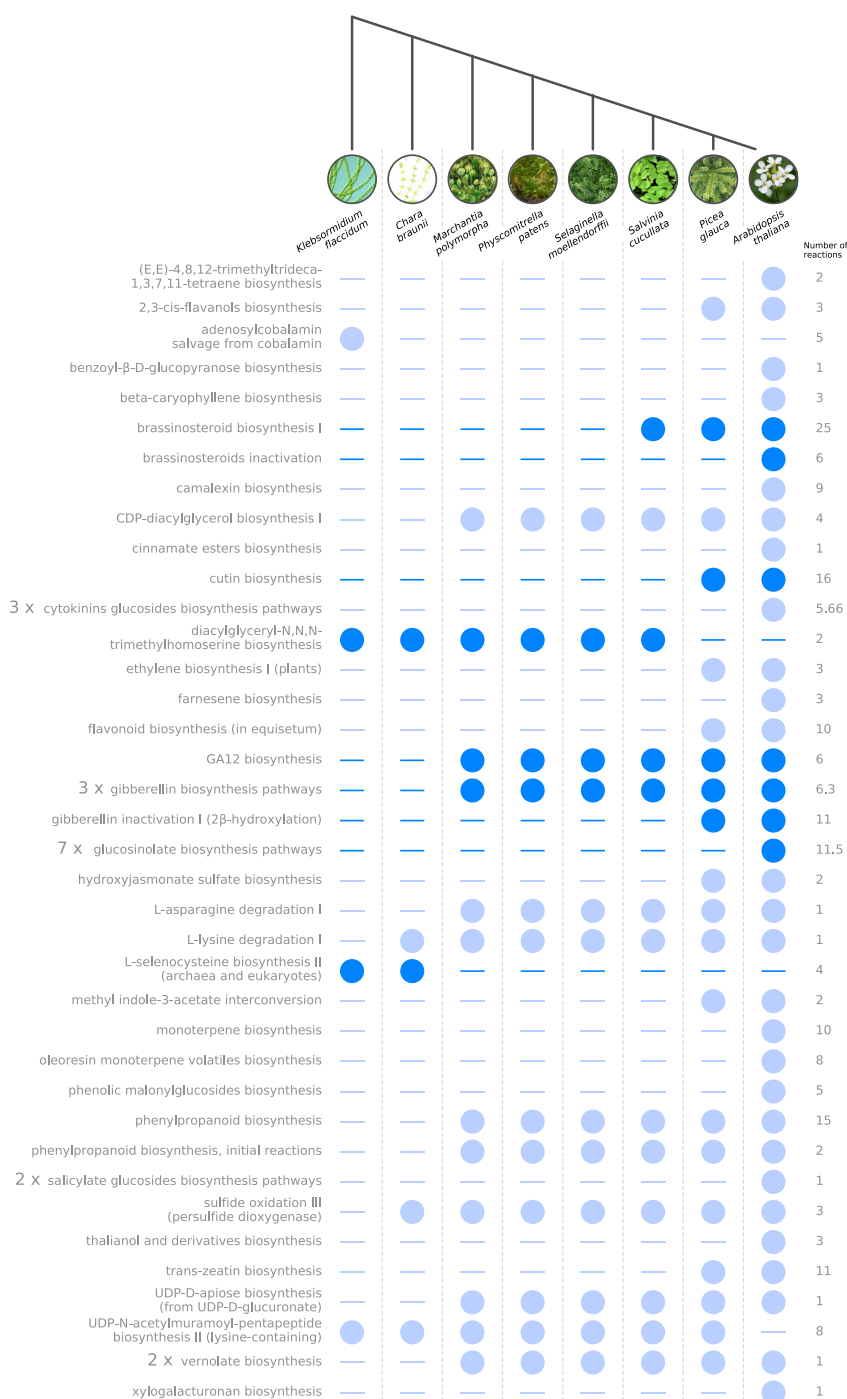


Figure 1. Overview of Metabolic Gains and Losses Identified by the Orthogroup Analysis

For each representative species, the presence or absence of the metabolic pathway in question is denoted by a filled circle (presence) or a horizontal line (absence). The far-right column shows the number of reactions in each metabolic pathway; where multiple similar reactions are grouped, e.g., for glucosinolate biosynthesis, this number denotes the average number of reactions across the grouped pathways. Dark blue circles indicate the metabolic pathways discussed in further detail. All metabolic pathway names are taken from the MetaCyc or PlantCyc metabolic pathway databases; Table S2 contains a list of the MetaCyc/PlantCyc IDs corresponding to each of the metabolic pathway names listed. Figure S1 shows a full phylogeny for the 72 species across which these metabolic gains/losses were identified, while Figure S2 contains an overview of the additional EC number annotations for each representative species for which PGDBs were constructed. Figure S3 shows the results of clustering these PGDBs based on metabolic pathway content.

the metabolic pathway of brassinosteroid biosynthesis were not found in these two algal groups (Figure 4B). The algae do, however, contain homologs of DET2 encoding $\alpha 5$ steroid dehydrogenase, the first enzyme in the pathway that our analysis suggests is present throughout the land plant lineage (Figure 4B).

Previously, it has been suggested that GAs are confined to vascular plants [45]. However, our analysis demonstrates that the enzymes required for their biosynthesis are encoded in bryophyte genomes. The two enzymes responsible for biosynthesis of the inactive GA precursor GA₁₂ (*ent*-kaurene oxidase [KO] and kaurenoic acid oxidase [KAO] [Figure 3A]) are present in all land plants except the mosses *Physcomitrella patens* and *Sphagnum fallax*, which are missing KAO (Figure 3B). This supports previous reports that *P. patens* is missing KAO [7, 46, 47] and extends this finding to an additional moss species. The presence of KO in all land plants is consistent with observations that its metabolite product, the GA₁₂ intermediate

terrestrial environments [4, 44]. Consistent with this, genes encoding enzymes involved in both gibberellin (GA) and brassinosteroid biosynthesis were associated with the transition to land in the 72-species comparison—i.e., were absent from all algal species analyzed (Figures 3 and 4). The enzymes comprising the GA biosynthesis pathway from *ent*-kaurene to GA₁₂ and its conversion to bioactive forms (Figure 3A) were not found in the charophyte or chlorophyte algal genomes (Figure 3B). Similarly, genes encoding enzymes for the majority of

KA, has been widely found in plants [48, 49]. However, there has been no metabolomic evidence for the occurrence of recognized active forms of GA in bryophytes [47, 50]. Yet, our analysis provides evidence for the presence of genes encoding GA20ox and GA3ox enzymes in all land plant clades analyzed, including the liverworts. These enzymes are 2-oxoglutarate-dependent dioxygenases responsible for the oxidation of GA₁₂ into bioactive GAs (Figure 3A). GA oxidase homologs have been previously identified in *P. patens* [51], and the identification here of homologs in

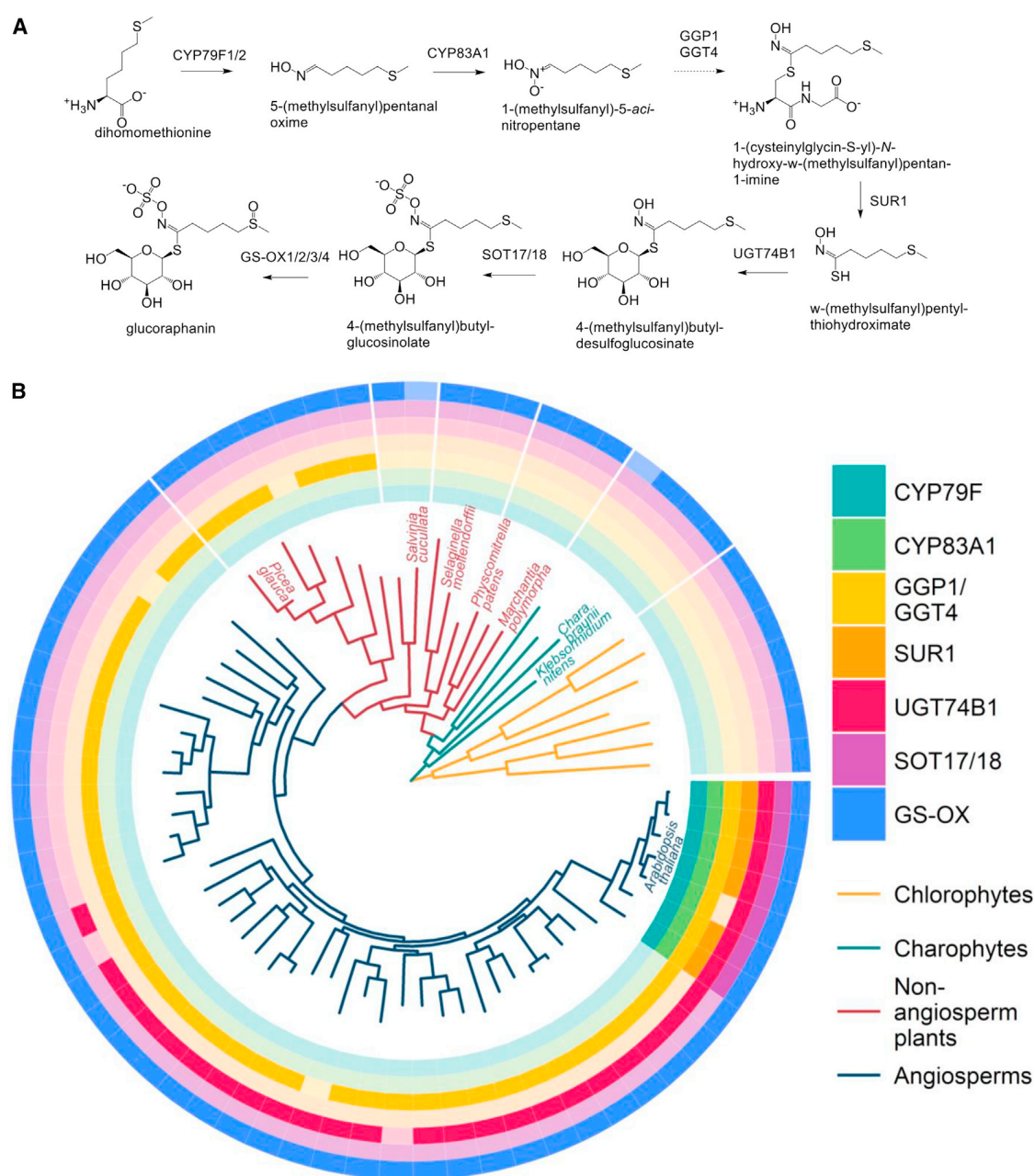


Figure 2. Presence of Gene Homologs Encoding Glucosinolate Biosynthesis Enzymes across 72 Plant and Algal Species

(A) Biosynthetic pathway for glucosinolates. Although several variations on this pathway were identified as possible evolutionary transitions (all utilizing the same enzymes), only glucosinolate biosynthesis from a dihomomethionine precursor is shown.

(B) Homolog presence (colored rings) for genes involved in glucosinolate biosynthesis across 72 studied species. Dark colors indicate the presence of a homolog while light colors indicate absence. Orthogroup information for the genes involved is provided in Table S1. Figures S4 and S5 show the presence of gene homologs for the additional validation pathways - DGTS and selenocysteine biosynthesis. CYP, cytochrome P450; GGP, γ -glutamyl peptidase; GGT, γ -glutamyl transpeptidase; SUR, alkyl-thiohydroxamate C-S lyase; UGT, UDP-glucosyltransferase; GS-OX, glucosinolate S-oxygenase.

moss *Sphagnum fallax* and the liverwort *Marchantia polymorpha* supports the evolution of GA oxidases in bryophytes.

Conversely, beyond the initial reaction catalyzed by enzymes encoding DET2, our analysis did not identify genes encoding brassinosteroid biosynthesis enzymes in bryophytes or algae. Following the synthesis of campestanol by the ubiquitous DET2

enzyme, brassinosteroid biosynthesis follows a series of oxidative modifications in two overlapping pathways (Figure 4A) catalyzed by cytochrome P450 (CYP) enzymes. All *Arabidopsis* enzymes involved are members of the CYP85 clan—either the CYP85A subfamily (involved in C6 oxidation and ring extension) or the CYP90A/B/C/D subfamilies (which catalyze C22 and C23

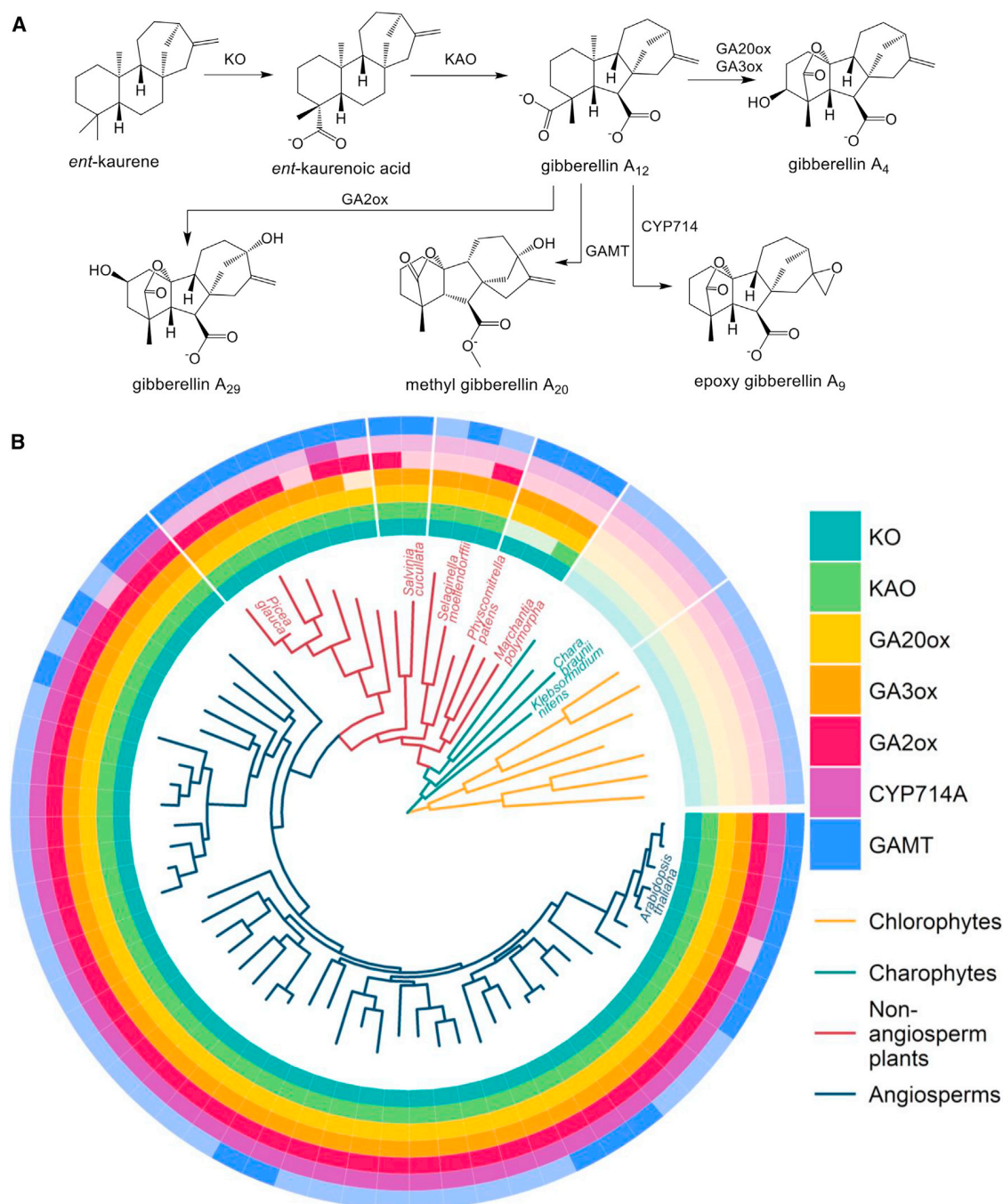


Figure 3. Presence of Gene Homologs Encoding Gibberellin Biosynthesis and Inactivation Enzymes across 72 Plant and Algal Species

(A) The gibberellin biosynthetic/inactivation pathway.

(B) Phylogeny showing the presence and absence of homologs for genes (colored rings) involved in gibberellin biosynthesis across 72 analyzed species. Darker ring colors indicate gene presence while lighter colors indicate absence. Internal phylogeny colors show broad plant groupings. Orthogroup information for the genes involved is provided in Table S1. KO, *ent*-kaurene oxidase; KAO, *ent*-kaurenoic acid oxidase; GA20ox, gibberellin 20 oxidase; GA3ox, gibberellin 3 oxidase; GA2ox, gibberellin 2 oxidase; CYP, cytochrome P450; GAMT, gibberellin methyltransferase.

hydroxylation). The brassinosteroid castasterone has been identified in bryophytes, albeit at concentrations that are orders of magnitude lower than in angiosperms [52]. The lack of CYP85A and CYP90 homologs in bryophytes and algae suggests either

that the biosynthetic route for castasterone diverges from that shown in Figure 4A in these species or that other cytochrome P450s are involved in its biosynthesis. Green algae are known to contain CYP85 clan enzymes with divergent sequences from

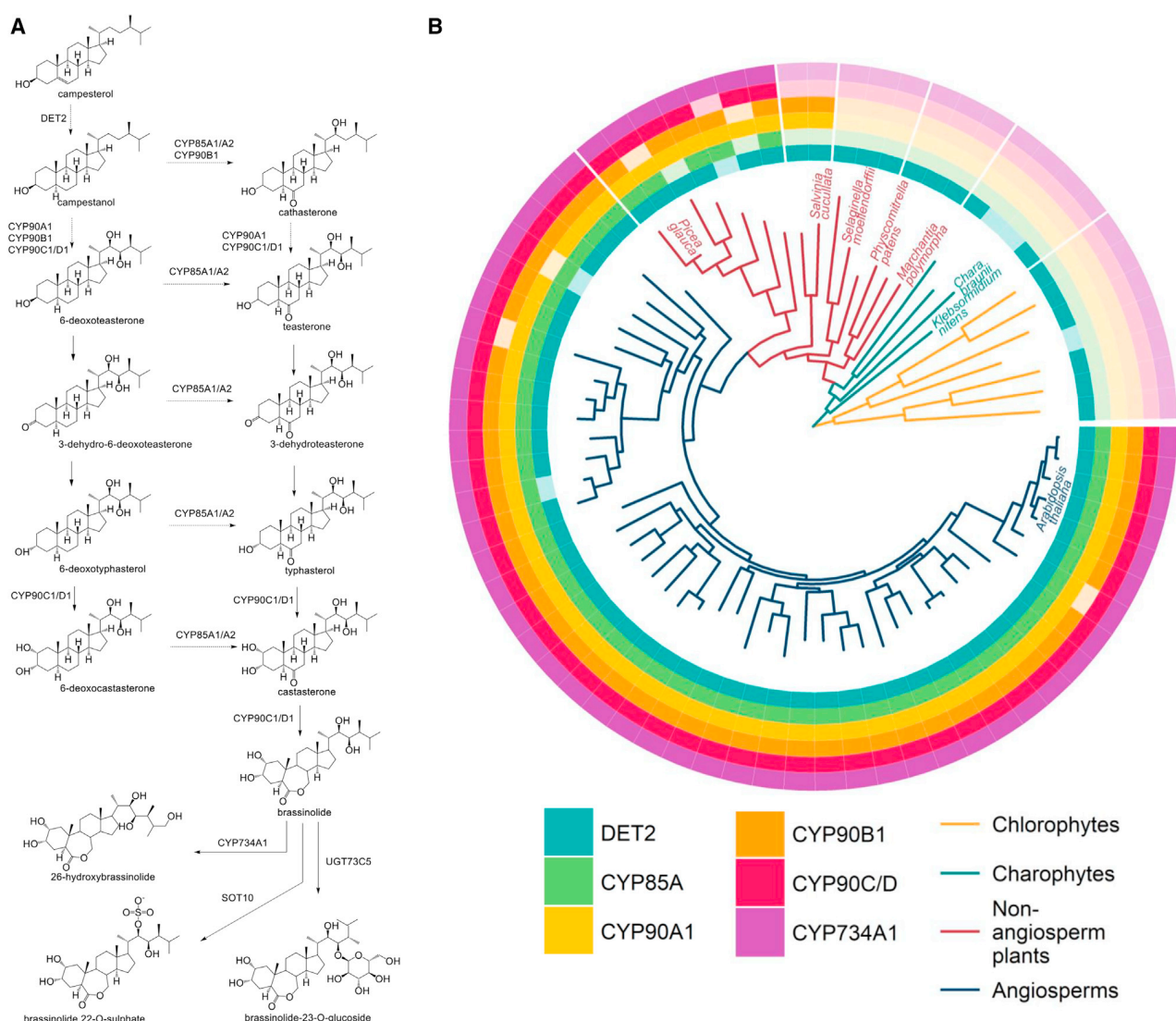


Figure 4. Presence of Gene Homologs Encoding Brassinosteroid Biosynthesis and Inactivation Enzymes across 72 Plant and Algal Species

(A) The brassinosteroid biosynthetic/inactivation pathway.

(B) Evidence (colored rings) for homologs of brassinosteroid biosynthetic enzymes across 72 plant species. Darker colors indicate presence of gene homologs, while lighter colors indicate absence. Internal phylogeny colors show broad plant groupings. Degradation steps are shown only for brassinolide; however, the same enzymes also act to degrade castasterone. Orthogroup information for the genes involved is provided in Table S1. A simplified tree showing relationships between homologous genes across species in the orthogroup containing CYP734A1 is provided as Figure S6A. DET2, 5 α steroid dehydrogenase; CYP, cytochrome P450; UGT, UDP glucose:cytokinin glucosyltransferase; SOT, brassinosteroid sulfotransferase.

those in angiosperms [7]—it is possible that these genes and their relatives in the bryophytes are responsible for brassinosteroid biosynthesis in these species but were replaced by CYP85A and CYP90 genes in spermatophytes. The latter scenario may also explain our finding of the absence of homologs of genes encoding brassinosteroid biosynthesis enzymes in the lycophytes (Figure 4B). Brassinosteroids have been found in lycophytes [52, 53], but we did not find specifically identifiable CYP85A and CYP90A/B/C/D homologs in either the *S. moellendorffii* and *S. tamariscina* genomes or in the *I. echinospora* transcriptome (Figure 4B). However, we did identify genes in *S. moellendorffii* and *S. tamariscina* forming a sister group to the clade containing

specific CYP90C and CYP90D genes in other species. While these genes are not specifically identifiable as CYP90 class C or D, their close relationship suggests they may be candidate genes for brassinosteroid biosynthesis in lycophytes. Six gymnosperm genomes were found to encode specific CYP90D1 genes. CYP90C1 is the closest relative of CYP90D1 in angiosperms; presumably, the former arose from a gene duplication after the divergence of the gymnosperms. However, these genes are functionally redundant [54], and we therefore consider the angiosperm-like brassinosteroid biosynthesis pathway to be complete in the gymnosperms. Three of the gymnosperm genomes analyzed contain homologs of all angiosperm brassinosteroid

biosynthetic genes, while the genomes of the ferns *S. cucullata* and *A. filiculoides* contain three biosynthetic genes (and are missing CYP90C/D and CYP85A homologs; Figure 4B), suggesting that brassinosteroid biosynthesis in ferns represents an intermediate stage of evolution before the complete angiosperm-like pathway was established in the gymnosperms.

Non-angiosperm Plants Lack Genes Encoding Known Enzymes for GA and Brassinosteroid Inactivation

A key feature of the use of specialized metabolites as hormones in angiosperms is rapid regulation of the concentration of bioactive forms of the molecules by enzymatic conversion to inactive forms. Our analysis suggests that many of these mechanisms are absent from non-angiosperm plants (Figures 3 and 4). GAs are inactivated in angiosperms using either GA 2-oxidases, GA 13-oxidases, GA 16,17-oxidases, or GA methyltransferases (GAMTs) [55, 56]. Two orthogroups were identified that contained genes encoding GA 2-oxidases, one containing genes encoding enzymes that act on C19 GAs and one containing those that act on C20 GAs. With the exception of single homologs in *Isoetes echinospora*, *Azolla filiculoides*, and *Ginkgo biloba*, GA 2-oxidases acting on C20 GAs were not found in non-angiosperm plants. C19-acting homologs were not found in non-seed plants but were found in six gymnosperm species, while every angiosperm species contained both C19 and C20-acting homologs (Figure 3B).

Cytochrome P450 enzymes encoding GA 13-oxidases and GA 16,17-oxidases have been identified in rice (CYP714B1, CYP714B2, CYP714D1) and *Arabidopsis* (CYP714A1, CYP714A2) [57–59]. These genes are all members of a single orthogroup containing genes exclusively from angiosperm species (Figure 3B) and *G. biloba*. One explanation for the presence of a CYP714A homolog in *G. biloba* is that GA inactivation via CYP714A enzymes may have evolved before the divergence of the gymnosperms but was lost in gymnosperms other than *G. biloba*. An additional GA 13-oxidase (CYP72A9) has recently been identified in *Arabidopsis* [56]. CYP72A9 is the only CYP72A gene that encodes an enzyme with GA13ox activity in *Arabidopsis*. Our orthogroup analysis shows that similar CYP72A genes are present in other dicots—but the close grouping of CYP72A9 with other CYP genes in *Arabidopsis* makes it difficult to suggest candidate genes with GA 13-oxidase activity in other species.

Arabidopsis GAMTs are members of the SABATH family of methyltransferases, which act on various plant hormones and signaling molecules [60]. As shown in Figure 3B, homologs of *Arabidopsis* GAMT1 and GAMT2 are inconsistently distributed across land plants and are not found in algae. The pattern of GAMT homology could mean that genes in several branches of earlier-diverging land plants have convergently evolved methyltransferases with similar sequences to *Arabidopsis* GAMTs. Or, more likely, that GAMT-like methyltransferases are ancestral, and there has been sequence divergence in most angiosperms and some lycophytes. Regardless, it appears that bryophytes, lycophytes, ferns, and gymnosperms have the capacity to inactivate GAs, despite GAs having not yet been found in bryophytes.

There are several mechanisms for brassinosteroid inactivation in angiosperms including hydroxylation, glucosylation, and sulfonation [61–63]. Each of these mechanisms is catalyzed by a single gene product in *Arabidopsis*. The known *Arabidopsis*

brassinosteroid hydroxylation enzyme is CYP734A1, also known as CYP72B1. This is a member of a large orthogroup containing CYP72 clan genes including the CYP714 family responsible for GA inactivation. Although this orthogroup includes algal homologs, gene trees place the algal genes as an outgroup, suggesting that they are homologous to the ancestor of all angiosperm CYP genes in this orthogroup. Bryophyte, lycophyte and fern genes in this orthogroup cluster with alternative CYP genes, and there are no direct homologs of CYP734A1 in these species (Figure S6A).

Glucosylation of brassinosteroids in *Arabidopsis* is carried out by a UDP glucosyltransferase enzyme, UGT73C5, while sulfonation is carried out by brassinosteroid sulfotransferase enzyme (SOT10). Orthogroup analysis places both these genes as members of large orthogroups containing other glucosyltransferases and sulfotransferases, respectively. However, both *Arabidopsis* genes in question were grouped closely with additional *Arabidopsis* genes known to have functions outside of brassinosteroid inactivation—we were therefore unable to distinguish homologs in other species.

Bryophytes, Lycophytes, Ferns, and Gymnosperms Lack Specific Cytochrome P450 Family Enzymes Involved in Oxidation of Structural Monomers

The transition of plants to terrestrial environments was accompanied by the evolution of mechanisms that protect against desiccation, UV radiation, and exposure to a new complement of pathogens [64, 65]. The ability of land plants to withstand these stresses results in part from the evolution of several amino acid/lipid-based insoluble biopolymers such as cutin, suberin, and sporopollenin. Our analysis identified the cutin biosynthetic pathway as a metabolic innovation during land plant evolution. As all three of these structural polymers are produced via biosynthetic pathways that use very similar (or overlapping) biochemical reactions, and several orthogroups are shared between pathways, all three pathways were investigated. To date, suberin has been found only in vascular plants, whereas cutin has been found across the land plants. Sporopollenin is found across land plants and has been isolated from the cell walls of several species of chlorophyte and charophyte algae [64, 66]. All three polymers are formed primarily of fatty-acid-derived monomers of varying chain lengths (sporopollenin C12–C18, cutin C16–C18, and suberin C18–C24 fatty acids), and suberin also contains significant amounts of phenolic compounds [67, 68]. The metabolic pathways for all three polymers rely on the activation of free fatty acids with coenzyme A (CoA) and use cytochrome P450 enzymes for the oxidation of fatty acids into polymer-specific monomers, several of which are shared by cutin and suberin (Figure 5A).

We identified homologs of the LACS genes required for fatty acid activation in cutin and suberin biosynthesis in the genomes of sampled land plants from all clades as well as in charophyte and chlorophyte algal genomes (Figure 5B). For sporopollenin biosynthesis, homologs of the ACOS5 gene responsible for the same process were not found in algae but were present in all land plants, while the ACH enzymes that remove CoA were found in charophyte algae and land plants (Figure 6B).

This suggests that the capability for biosynthesis of precursor molecules required for the production of cutin/suberin monomers were present prior to the evolution of land plants. However,

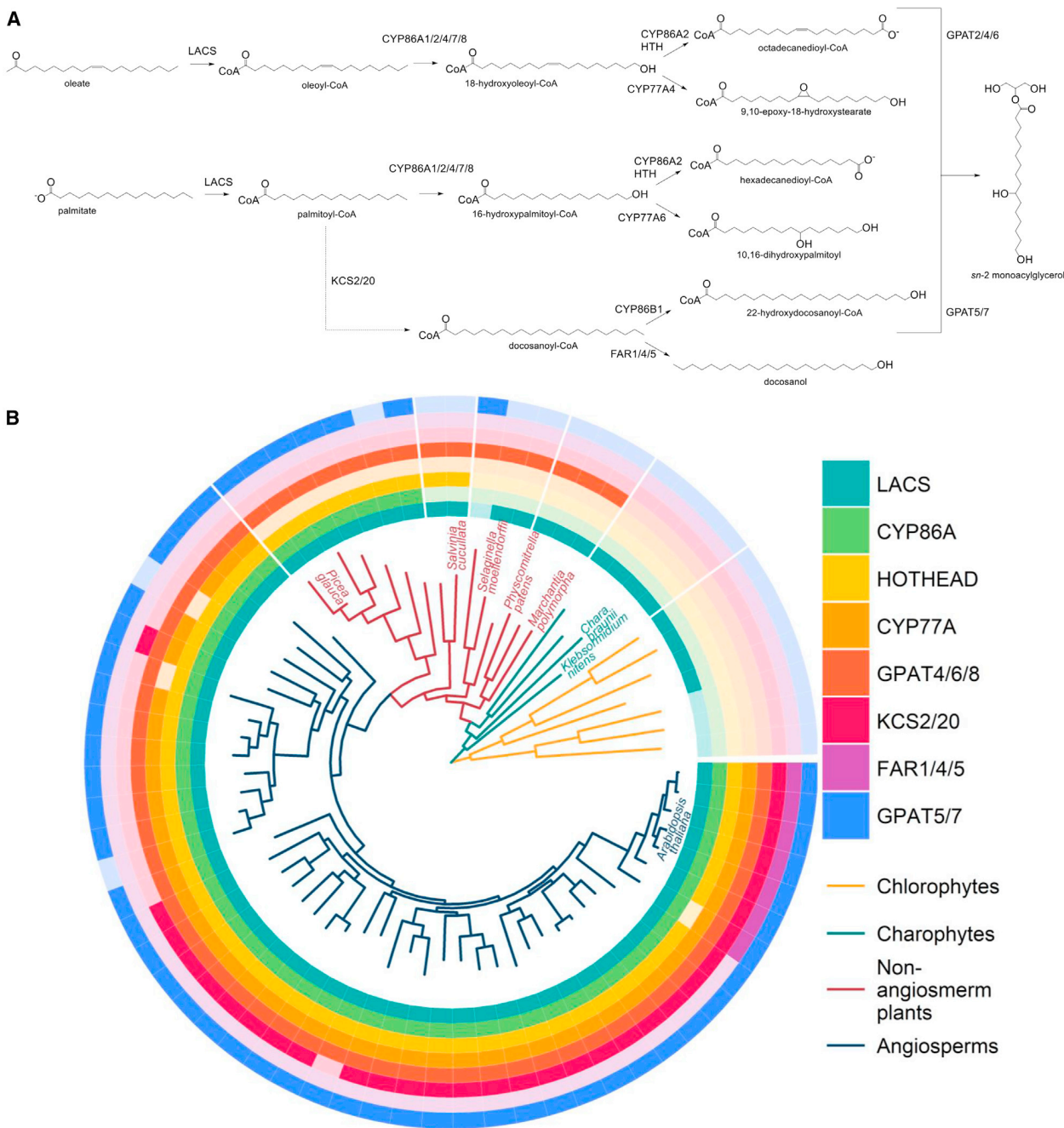


Figure 5. Presence of Gene Homologs Encoding Cutin and Suberin Biosynthesis Enzymes across 72 Plant and Algal Species

(A) The cutin/suberin biosynthetic pathway. (B) Evidence (colored rings) for homologs of required enzymes across 72 plant species. Darker colors indicate presence of gene homologs while lighter colors indicate absence. Internal phylogeny colors show broad plant groupings. Orthogroup information for the genes involved is provided in Table S1. A simplified gene tree showing relationships between homologous genes across species in the orthogroup containing CYP86 genes is provided as Figure S6B. LACS, long-chain acyl-CoA synthetase; CYP, cytochrome P450; KCS, 3-ketoacyl CoA synthase; FAR, fatty acid reductase; GPAT, glycerol-3-phosphate acyltransferase.

the majority of homologs of the *Arabidopsis* genes encoding enzymes responsible for subsequent reactions in both metabolic pathways were identified only in land plants (Figures 5B and 6B)—this is consistent with the lack of cutin and suberin in algal species [68]. For sporopollenin, the lack of ACOS5 homologs in

algae that do nevertheless contain sporopollenin suggests the use of a different, as-yet-unknown, metabolic pathway for sporopollenin biosynthesis in chlorophytes and charophytes. Further investigation of the gene tree of the orthogroup containing *Arabidopsis* CYP86A and CYP86B subfamily genes,

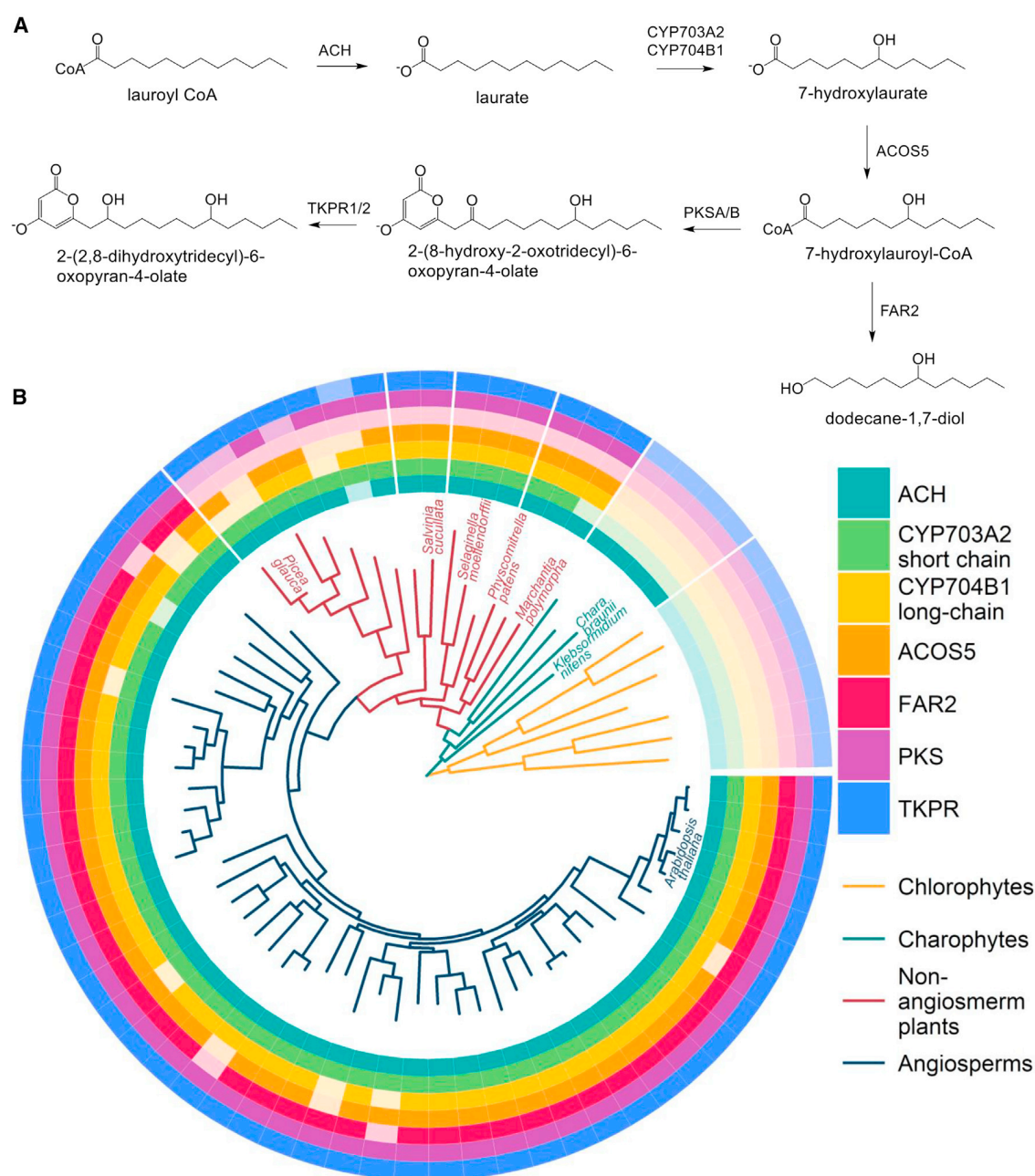


Figure 6. Presence of Gene Homologs Encoding Sporopollenin Biosynthesis Enzymes across 72 Plant and Algal Species

(A) The sporopollenin biosynthesis pathway. Although sporopollenin is produced via analogous reactions from several fatty acid precursors, the depicted pathway shows only those with a lauric acid precursor.

(B) Phylogeny of 72 species showing the presence and absence of homologs of sporopollenin biosynthetic genes (colored rings). Darker colors indicate the presence of homologs while lighter colors indicate absence. Internal phylogeny colors show broad plant groupings. Orthogroup information for the genes involved is provided in Table S1. ACH, acyl-CoA thioesterase; CYP, cytochrome P450; ACOS5, fatty-acyl CoA synthetase; FAR, fatty acid reductase; PKS, polyketide synthase; TKPR, tetraketide pyrone reductase.

responsible for the ω -hydroxylation of fatty acids in cutin and suberin biosynthesis, shows that, while gymnosperm and angiosperm species contain specific homologs of these *Arabidopsis* genes (Figure 5B), earlier-diverging land plants and algae contain homologs that can only be identified more broadly as related to the CYP86 and CYP96 subfamilies (Figure S6B). Direct

homologs of the oxidoreductase gene HOTHEAD are also identified in the spermatophytes as well as in both fern species investigated but not in the lycophytes, bryophytes, or algae (Figure 5B). These gene products in earlier-diverging land plants may also catalyze the ω -hydroxylation of fatty acids but could differ in characteristics such as chain-length preferences. This

would be consistent with previously described shorter chain-lengths in the cutin biopolymers of earlier-diverging land plants [69]. However, the lack of direct subfamily homologs mean that no specific function can be inferred for these species.

Similarly, we found CYP77A homologs, involved in in-chain hydroxylation and/or epoxidation, across angiosperm species. However, related genes identified in other land plants could only be classified more generally as CYP77 genes (Figure 5B). We found no CYP77 homologs in algae (Figure 5B), which is as expected given the lack of evidence for the occurrence of cutin or suberin in algae [70]. The presence of genes encoding the LACS enzyme across land plants, combined with the lack of specific CYP86A and CYP77A genes in non-seed plants and the lack of HOTHED in lycophytes and bryophytes, suggests that the early stages of cutin and suberin biosynthesis are conserved. However, the fatty acid monomers of cutin and suberin must be biosynthesized via alternative gene products (likely from CYP86/77 families) in bryophytes and lycophytes and later-diverging land plants, which may conceivably produce monomers of different chain lengths and with different molecular modifications. The evolution of HOTHED before the divergence of ferns, CYP86A/B in the ancestor of spermatophytes and CYP77A only in the angiosperms suggests that the angiosperm-like pathway for cutin and suberin biosynthesis may have evolved in stages, replacing an alternative set of enzymes in the bryophytes and lycophytes.

Regarding sporopollenin CYP enzymes, we found specific homologs of CYP704B1, which catalyzes the hydroxylation of long-chain fatty acids in all land plant clades. Homologs of this gene were not found in the sampled algae. We found homologs of CYP703A2 responsible for hydroxylation of mid-length-chain fatty acids only in the vascular plants and the mosses *Physcomitrella patens* and *Sphagnum fallax* but not in the liverwort *Marchantia polymorpha* or any algae (Figure 6B). This suggests that the use of lauric acid derivatives in sporopollenin biosynthesis originated in the bryophytes.

Despite the lack of specific CYP704B1 and CYP703A2 homologs, sporopollenin is found in algae. One possible explanation for this is that the oxidation reactions for sporopollenin biosynthesis are carried out by multifunctional CYP86 clan enzymes in algae, which when duplicated after the colonization of land became specialized for the biosynthesis of sporopollenin and other biopolymers.

Conserved Enzymes for Production of Secondary Monomers Found in Suberin and Sporopollenin Are Limited to Angiosperms and Land Plants, Respectively

The final stages of sporopollenin biosynthesis require polyketide synthases (PKSA/B) and tetraketide pyrone reductases (TKPR1/2) to produce tetraketide pyrones. No homologs of PKSA/B or TKPR1/2 were identified in the sampled algae (Figure 6B); our analysis suggests that tetraketide pyrones may be terrestrial plant-specific sporopollenin monomers.

Suberin and sporopollenin also contain various alcohols, the production of which requires fatty acid reductase (FAR) enzymes. In all but three of the angiosperm species (*Spirodela polyrrhiza*, *Solanum tuberosum*, and *Malus domestica*), we found homologs of *Arabidopsis* FAR2, specific to sporopollenin biosynthesis. Homologs of *Arabidopsis* FARs specialized for suberin production (FAR1/4/5) were only found in the Brassicaceae

(Figure 5). However, FAR family genes are found in all land plant taxa; those species missing FAR1/2/4/5 contain other FARs, which are more similar to *Arabidopsis* FARs of unknown or alternative function. This could mean that there has been significant sequence divergence from an ancestral specialized FAR. Or, that earlier-diverging land plants make use of multifunctional FARs for the production of these two polymers. Another possibility is that earlier-diverging land plant suberin and sporopollenin do not contain alcohols.

The use of very-long-chain fatty acids as suberin monomers means that a fatty acid elongation process is necessary. In angiosperms the enzymes responsible are 3-ketoacyl-CoA synthases (KCS2/20). As with FARs, we found homologs of *Arabidopsis* KCS genes in algae and land plants. However, in all species except dicotyledonous angiosperms and one monocot these genes are more related to alternative *Arabidopsis* KCS genes than KCS2/20 (Figure 5B).

After very-long-chain fatty acid production and oxidation, hydroxy fatty acids destined for cutin and suberin polymers are combined with glycerol-3-phosphate at the *sn*-2 position in a reaction catalyzed by glycerol 3-phosphate acyltransferase (GPAT) enzymes. GPAT4/6/8 gene products catalyze the transfer of long-chain fatty acids, while GPAT5/7 catalyze the transfer of suberin-specific, very-long-chain fatty acids. Despite the lack of specific genes encoding enzymes of the preceding oxidation and reduction steps, we found at least one homolog of *Arabidopsis* GPAT4/6/8 genes in all land plants (but none in algae) (Figure 5B). Consistent with evidence of reduced chain-length biopolymer monomers in earlier-diverging land plants [69] and presence of suberin only in vascular plants, specific homologs of GPAT5/7 were found only in the angiosperms and gymnosperms, and a single lycophyte (Figure 5B). It may be that the remaining vascular plants contain homologs of GPAT genes not yet characterized as suberin related.

The pattern of homology that we have described for structural molecule biosynthetic pathways supports the hypothesis that cutin appeared for the first time in the earliest land plants [71]. LACS genes already present in algae for the initial activation of free fatty acids were likely involved in cutin synthesis in these species. Indeed, homologous genes encoding LACS are present in almost all species in this study. Similarly, the final enzymes in the cutin biosynthesis pathway (GPAT4/6/8) are encoded by all land plant genomes analyzed (Figure 5B). However, non-seed plants, although containing genes belonging to the CYP86, CYP77, KCS, and FAR gene families, do not contain genes homologous to specific CYP86A/B and CYP77A subfamilies, or to KCS2/20 and FAR1/4/5. This suggests that oxidation reactions in cutin and suberin biosynthesis across taxa have been significantly modified since their origin in bryophytes and vascular plants, respectively, and the specific pattern of homolog presence of angiosperm biosynthetic genes in non-angiosperm plants suggest that the intermediate steps of the angiosperm-like pathway may have evolved in stages. The presence of chlorophyte algal genes in orthogroups containing *Arabidopsis* KCS and FAR genes suggest that, when suberin biosynthesis did develop, it was achieved by the co-option of pre-existing cutin biosynthetic CYP genes in combination with ancient elongase and reductase enzymes, the latter of which may have been multifunctional and also involved in sporopollenin biosynthesis. However, the lack of

Table 1. Results of Search for Homologs of Investigated Genes in Transcriptomes of 305 Non-angiosperm Plants and Algae from the One Thousand Plant Transcriptomes Initiative

Metabolic Pathway	Gene ID	<i>Arabidopsis</i> Accession	Charophytes	Liverworts	Mosses	Hornworts	Lycophytes	Ferns	Gymnosperms
Gibberellin biosynthesis	KO	AT5G25900	0/15	14/15 ^a	15/15 ^a	10/13 ^a	10/15 ^a	15/15 ^a	15/15 ^a
	KA0	AT1G05160; AT2G32440	0/19	9/16 ^a	0/21	9/12 ^a	9/18 ^a	16/16 ^a	16/17 ^a
	GA20ox	AT5G07200; AT5G51810; AT1G60980	0/16	8/21 ^a	19/22 ^a	1/12 ^a	7/16 ^a	16/17 ^a	16/16 ^a
	GA3ox	AT1G15550; AT1G80340	0/16	0/20	7/17 ^a	5/12 ^a	6/18 ^a	11/17 ^a	6/23 ^a
	GA2ox19	AT2G34555; AT1G30040	0/16	0/19	0/19	0/12	0/16	0/21	10/17 ^a
	GA2ox20	AT4G21200; AT1G50960	0/15	5/19 ^a	4/18 ^a	4/12 ^a	2/17 ^a	1/21 ^a	18/18 ^a
	CYP714	AT5G24910; AT5G24900	0/18	0/17	0/16	0/12	0/15	0/17	0/16
	GAMT	AT4G26420; AT5G56300	8/8 ^a	14/15 ^a	18/18 ^a	12/12 ^a	16/16 ^a	19/19 ^a	15/15 ^a
Brassinosteroid biosynthesis	DET2	AT2G38050	0/15	0/15	0/15	0/14	0/15	0/15	0/15
	CYP85A	AT5G38970; AT3g30180	0/22	0/16	0/16	0/12	0/16	0/20	0/19
	CYP90A1	AT5G05690	0/15	0/15	0/15	0/12	11/15 ^a	15/15 ^a	13/15 ^a
	CYP90B1	AT3G50660	0/15	0/15	0/15	0/12	0/15	15/15 ^a	15/15 ^a
	CYP90C/D	AT4G36380; AT3G13730	0/20	0/18	0/22	0/12	0/15	0/22	24/24 ^a
	CYP734A1	AT2G26710	0/15	0/15	0/14	0/12	0/15	0/15	0/15
Cutin/suberin biosynthesis	LACS	AT2G47240; AT1G64400; AT1G49430	22/22 ^a	21/21 ^a	24/24 ^a	12/13 ^a	18/18 ^a	30/30 ^a	32/33 ^a
	CYP86A	AT5G58860; AT2G45970	0/17	5/16 ^a	0/17	3/12 ^a	1/17 ^a	5/17 ^a	2/26 ^a
	HOTHEAD	AT1G72970	0/15	0/14	0/15	0/10	0/12	0/15	15/15 ^a
	CYP77A	AT3G10570; AT5G04660	0/16	0/17	0/17	0/12	0/16	0/17	0/19
	GPAT4/6/8	AT1G01610; AT2G38110; AT4G00400	0/0	1/16 ^a	0/19	0/10	18/18 ^a	21/23 ^a	29/30 ^a
	KCS2/20	AT1G04220; AT5G43760	0/17	0/16	0/19	0/10	0/15	0/20	0/19
	FAR1/4/5	AT5G22500; AT3G44540; AT3G44550	0/16	0/10	0/18	0/10	0/15	0/20	0/18
	GPAT5/7	AT3G11430; AT5G06090	0/2	0/16	0/17	0/10	0/18	0/17	0/20
Sporopollenin	ACH	AT1G01710	15/15 ^a	15/15 ^a	15/15 ^a	12/12 ^a	15/15 ^a	15/15 ^a	15/15 ^a
	CYP703A2	AT1G01280	0/15	1/15 ^a	14/15 ^a	2/12 ^a	12/15 ^a	15/15 ^a	15/15 ^a
	CYP704B1	AT1G69500	0/15	14/15 ^a	15/15 ^a	9/12 ^a	11/15 ^a	15/15 ^a	2/15 ^a
	ACOS5	AT1G62940	0/15	3/15 ^a	3/15 ^a	2/15 ^a	5/15 ^a	15/15 ^a	2/15 ^a
	FAR2	AT3G11980	0/15	0/10	0/15	0/10	0/15	0/15	0/15
	TKPR	AT4G35420; AT1G68540	0/17	7/20 ^a	9/18 ^a	4/11 ^a	15/17 ^a	23/23 ^a	24/24 ^a
	PKS	AT1G02050; AT4G34850	0/18	0/19	0/16	0/11	0/17	0/15	0/18

The numerator indicates the number of species in which transcripts encoding the enzyme (or enzyme group) were found. The denominator indicates the number of transcriptomes of each taxon analyzed. A maximum of 15 unique transcriptomes per taxon were analyzed, but note that the number of species hits will exceed 15 when more than one gene is responsible for a specified reaction, for example, in the case of LACS.

^aHomologs of the gene/gene group in question were found in the corresponding transcriptomes.

Arabidopsis-like FAR1/4/5 genes in non-Brassicaceae shows that significant modifications to suberin biosynthesis have occurred even within the angiosperms. With the exception of specific FAR2 enzymes, all land plant clades contain homologs of the enzymes required for sporopollenin biosynthesis. The missing homologs in algae suggest that algal sporopollenin is produced using alternative enzymes and may differ in its composition.

Testing Metabolic Gains on a Broader Selection of Species Using 1KP Transcriptomes

To test our results from the orthogroup genome analysis, and to compensate for the relative lack of genome sequences available for non-angiosperm land plants and algae, we carried out an

additional analysis of transcriptomes released by the One Thousand Plant Transcriptomes Initiative [34]. This also allowed the analysis to be extended to hornworts, for which there are no available genome sequences. While transcriptome analysis cannot be used to draw conclusions about gene absences, it can be used to confirm the presence of gene homologs. A total of 305 transcriptomes were analyzed, comprising 81 gymnosperms, 78 ferns, 20 lycophytes, 37 mosses, 29 liverworts, 14 hornworts, and 46 charophytes. For each of the metabolic pathways that we analyzed in detail in this paper, the presence of homologs of genes associated with metabolic reactions were tracked across the transcriptome dataset (Table 1). Comparison of the transcriptome results with the genome analysis (Figures 3,

4, 5, and 6) reveals broad corroboration for the conclusions drawn regarding the evolution of metabolic pathways for GA and brassinosteroid biosynthesis/inactivation and the biosynthesis of structural polymers, thereby strengthening the conclusions. The transcriptome data also revealed additional detail and new information, described in the subsequent sections.

For GA biosynthesis, the transcriptome data provide full support for the results drawn from the genome data with one minor exception: we did not find GA3ox transcripts in liverworts (Table 1), whereas this gene was identified in the genome of *Marchantia polymorpha* (Figure 3). The transcriptome data also reveal that hornworts are more similar to liverworts than to mosses with respect to GA biosynthesis: hornwort transcriptomes contain homologs of KAO, which are missing from moss genomes and transcriptomes (Table 1; Figure 3) [46, 47].

The transcriptome data also provide additional information about the evolution of GA inactivation enzymes: homologs of GAMT genes were identified in the transcriptomes of charophyte algae as well as land plants (Table 1). Additionally, although homologs of GA2-oxidase genes acting on C20 GAs were not found in the genomes of *M. polymorpha*, *S. fallax*, or *P. patens* (Figure 3), they were identified in the transcriptomes of five other liverwort species, four other moss species, and four hornworts (Table 1). The presence of an additional inactivation mechanism in bryophytes provides further support for the hypothesis that GAs are likely present in the bryophytes even though this has yet to be demonstrated experimentally. The presence of GAMT homologs in charophyte algae transcriptomes indicates that these genes are more ancient in origin than previously discussed.

For the brassinosteroid biosynthesis pathway, the transcriptome data were in broad agreement with the genome analysis (Figure 4; Table 1) with one substantive exception: we were unable to identify transcripts of DET2 in the transcriptome data despite the clear presence of DET2 gene homologs in all plant and algal genomes investigated (Figure 4). We also did not find transcripts of CYP85A or CYP734A1 genes in any species, although these are present in the genomes of gymnosperms (Figure 4). The transcriptome data support the absence of genes for brassinosteroid biosynthesis in the bryophytes, including hornworts. Additionally, the transcriptome data extend our understanding of the evolution of brassinosteroid biosynthetic genes in the CYP90 family. Genome analysis showed the presence of CYP90A1 and CYP90B1 in ferns and genes for the complete pathway are identified first in gymnosperms (Figure 4). Transcriptome analysis corroborated these findings, as well as that of lycophyte genes forming a sister group to the clade of genes containing gymnosperm and angiosperm CYP90C1 and CYP90D1 (Table 1). Additionally, the fern transcripts were similar to the lycophytes with respect to CYP90C1 and CYP90D1. It seems likely that these genes in ferns and lycophytes represent a more generic version (perhaps with differing substrate specificity) of specific CYP90C/D genes in gymnosperms and angiosperms. Finally, lycophyte transcriptomes were shown to contain homologs of CYP90A1, which were not identified in the genome analysis. Combining these results further supports the stepwise evolution hypothesized for the brassinosteroid biosynthesis pathway found in angiosperms, beginning with the evolution of CYP90A1 and genes similar to CYP90C/D in lycophytes, both of which are maintained in ferns

with the addition of CYP90B1, before the completion of the biosynthetic pathway with the evolution of CYP90D1 in gymnosperms.

Finally, the transcriptome data are in full agreement with the genome analysis with respect to sporopollenin biosynthesis (Figure 6; Table 1) apart from the lack of PKS transcripts identified in any land plant taxon. The presence of transcripts involved in cutin and suberin biosynthesis were more scattered—we were unable to identify transcripts of several genes—specifically GPAT5/7 in vascular plants, HOTHEAD in ferns, and GPAT4/6/8 in bryophytes—which were identified in the genome analysis (Figure 5). Transcriptome analysis of cutin and suberin biosynthetic genes does, however, identify homologs of CYP86A genes in the liverworts, hornworts, lycophytes, and ferns, which suggests that the early stages of biosynthesis for these structural compounds are more similar across land plants than indicated by the genome analysis (Figure 5). Additionally, a single liverwort was found to contain a homolog of CYP703A2, which was not identified in the genome of *M. polymorpha* (Figure 5).

DISCUSSION

We have developed a scalable computational pipeline to systematically analyze and compare metabolic capabilities encoded in genomes and transcriptomes and have used it to advance our knowledge of the evolution of metabolic pathways in the Chloroplastida. The approach was validated by its ability to identify known metabolic innovations associated with plant evolution. The analysis allowed us to resolve several additional metabolic innovations in relation to land plant evolution.

Our analysis of enzymes involved in GA and brassinosteroid biosynthesis strongly supports the view that the biosynthetic capability for these two hormones is correlated with the transition of plants to land but that non-spermatophyte plants do not contain the full, or conserved, pathways for synthesis and inactivation of bioactive forms of the hormones as recognized in angiosperms (Figure 4; Table 1). Moreover, our analysis suggests that, while the complete angiosperm-like GA biosynthetic pathway is found in all vascular plants, the brassinosteroid biosynthetic pathway shows a stepwise increase in complexity with the evolution of CYP90A1 in lycophytes (Table 1) and CYP90B1 in ferns (Figure 4; Table 1) before the completion of the pathway in the gymnosperms (Figure 4; Table 1). The latter is a surprising observation because brassinosteroids have been found across plant taxa—in angiosperms, gymnosperms, ferns, lycophytes, the moss *P. patens*, the liverwort *M. polymorpha*, and the chlorophyte *C. vulgaris* [52, 72]. This may indicate that brassinosteroid production in algae, bryophytes, lycophytes, and ferns is carried out by enzymes that differ from those used in spermatophytes.

A surprising finding was the presence of genes in bryophytes encoding GA biosynthetic and inactivation enzymes (Figure 3; Table 1) despite the fact that recognized bioactive forms of GA have not been detected in the few bryophytes that have been analyzed [47, 50]. A plausible interpretation of this is that bioactive GAs are present at very low concentrations in bryophytes (as is the case with brassinosteroids [52]), and this has hindered previous attempts at their identification. Mass spectrometry instrumentation has increased substantially in sensitivity in recent years so this would be worth

revisiting. A key feature for the regulatory function of phytohormones is the capacity to rapidly inactivate the bioactive forms of the molecules using specific enzymes. Overall, the angiosperms were found to have a greater capacity for various kinds of GA and brassinosteroid inactivation than non-angiosperms, and homologs of genes encoding the angiosperm brassinosteroid inactivation mechanisms were not found in non-seed plants (Figures 3 and 4; Table 1). Combined with the lack of some GA inactivation enzymes identified in non-seed plants, this indicates that the majority of known angiosperm inactivation mechanisms for both hormones are relatively recent innovations in land plant evolution. Non-seed plants must either utilize alternative inactivation processes or a different set of genes to carry out brassinosteroid inactivation reactions. As indicated by the presence of a GAMT gene in *Marchantia polymorpha* (Figure 3), and GA 2-oxidases acting on C20 GAs in bryophytes (Table 1), bryophytes may be limited to GA inactivation via these mechanisms. The lack of conserved inactivation mechanisms between taxa for both phytohormones indicates that different plant groups make use of alternative or divergent enzymes for homeostasis of GAs and brassinosteroids. The presence of GAMT homologs in charophyte algae transcriptomes indicates that these genes are more ancient in origin than previously discussed and suggests that GAMTs in land plants evolved from similar methyltransferases in algae, likely with alternative substrates, given the lack of homologs of GA biosynthetic genes in algae.

The analysis also revealed details about the likely evolution of several structural polymers. The capability for biosynthesis of precursor molecules required for the production of cutin/suberin monomers were present prior to the evolution of land plants, with the LACS gene encoding the enzyme long-chain acyl-CoA synthetase required for fatty acid activation present in the genomes of sampled land plants from all clades as well as in charophyte and chlorophyte algal genomes (Figure 5B). However, even though cutin is thought to be present in all vascular plants, the CYP enzymes involved in angiosperm cutin and suberin biosynthesis appear to have evolved in stages, with only CYP86A present in all land plant taxa (Table 1), followed by HOTHEAD in the ferns and gymnosperms and CYP77A in the angiosperms (Figure 5; Table 1). For sporopollenin biosynthesis, homologs of the ACOS5 gene responsible for fatty acid activation were not found in algae but were present in all land plants, while the ACH enzymes, which remove CoA, were found in charophyte algae and land plants (Figure 6B). CYP enzymes involved in sporopollenin biosynthesis are conserved only in land plants (Figure 6; Table 1), but sporopollenin has been identified in chlorophyte algae—suggesting that alternative CYP enzymes (likely members of the CYP86 clan) are responsible for sporopollenin biosynthesis in these species.

Taken together, the ubiquitous presence in land plants and algae of genes encoding enzymes acting early in the biosynthetic pathways for suberin, cutin, and sporopollenin suggests that none of these biopolymer biosynthetic pathways developed *de novo* in the land plants. It may be that certain enzymes involved in cutin and suberin biosynthesis evolved directly from those involved in sporopollenin biosynthesis; a common evolutionary origin of all biopolymers has previously been hypothesized

[68, 69]. However, homologs of the angiosperm enzymes responsible for intermediate and end-point reactions are often absent from earlier-diverging land plant genomes (Figures 5B and 6B), even when these plants are known to contain such biopolymers. There are several possible explanations for this. It may be that different mechanisms, or at least alternative enzymes (for example differing in substrate specificity or activity), are utilized for biopolymer production in earlier-diverging land plants. The enzyme families in question tend to be expanded in the angiosperm species in this analysis; this diversity may be linked to subfunctionalization conferring increases in substrate specificity or catalytic efficiency in later-diverging land plants. It is plausible that these differences result in alternative monomer composition of biopolymers in divergent plant taxa, a phenomenon that has already been observed between some species [73, 74]. Another possibility is that the same reactions, leading to the same end products, are catalyzed by enzymes with common ancestry but that exhibit sequence divergence between earlier- and later-diverging land plants, and as such are not identified as homologous. Finally, the presence of similar biopolymers but absence of homologous genes across divergent plant taxa could be the result of convergent evolution, which is a common theme in land plant evolution [75, 76] and has already been identified in the biosynthesis of the biopolymer lignin, where lycophytes and spermatophytes have independently developed the ability to produce monomers derived from sinapyl alcohol [77].

Limitations of the Approach

The data presented here demonstrate the power of a systematic comparative analysis of large genome and transcriptome datasets to reveal new understanding about the evolution of biological pathways and processes. Nevertheless, there are some limitations to the approach. These stem mainly from the reliance on sequence homology as a tool for annotating genes with metabolic functions. First, in the case that two organisms do not have homologous genes for specific enzymes in common it is difficult to speculate as to how this reflects their biochemistry. For example, although cutin is biosynthesized in all land plants, angiosperm and non-angiosperm species do not share the specific cytochrome P450 enzymes used by higher plants (Figure 5). Without experimental analysis, it is impossible to determine whether this is indicative of a difference in gene function between these species (and non-angiosperm plants produce alternative cutin monomers) or whether the lack of homology is caused by superficial divergence at the sequence level (and enzyme function may in fact be conserved across land plants). Conversely, while sequence-based homology can identify candidates for functionally analogous genes across species, experimental analysis is also required to confirm these predictions. This may be particularly true for the cytochrome P450 genes, which exist in large, highly interrelated gene families that have been subject to rapid expansion and subfunctionalization over the course of plant evolution [7, 15].

Second, plants exhibit a huge range of specialized metabolism, and it is reasonable to expect that non-angiosperm plants, which are relatively understudied, have the capacity for metabolic functions that are thus far undiscovered. The use of a *de novo* metabolic annotation tool trained across plant and

non-plant organisms allows for the detection of metabolic reactions in non-angiosperms where homologs may be present in microbes or fungi rather than angiosperms; however, this method does not have the power to detect previously unstudied metabolic pathways. Increasing knowledge of non-angiosperm plant metabolism will require extensive experimental and computational investigation. Metabolic profiling over different conditions and growth stages can be used in combination with flux analysis to elucidate the compounds produced by individual species and their role in the wider metabolic network. Building up such a knowledge base of metabolism in non-angiosperm plants is the only way to improve the accuracy of metabolic pathway inference and would have additional far-reaching benefits, for example, in improving the prospects for using non-angiosperm plants as experimentally tractable test beds for metabolic engineering strategies.

Finally, gene homology inferences are dependent on a set of high-quality genomes as their foundation. Plant genomes tend to be larger and more complex than animal genomes, and the high incidence of repetitive elements affects the quality of genome assembly and annotation [78]. Incorrect construction of gene models may lead to the inference of gaps in gene presence in species where there are none. This is particularly a problem for gymnosperm genomes, which can be up to 30 Gb and are rich in repetitive DNA, mostly transposable elements [79]. It may be that these issues with gymnosperm genome assembly are the cause of some missing homologs in individual genomes in this study—for example, in Figure 5 several gymnosperms are missing genes that are identified in all other species analyzed. We have attempted to mitigate this problem in this study by maximizing our sample size, making use of the majority of genome sequences available for non-angiosperm species and establishing further gene presences using transcriptome analysis of an additional 305 species. Furthermore, our findings are limited to identifying evolutionary trends across whole plant taxa, as opposed to basing conclusions on the presence or absence of genes in individual species.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- [KEY RESOURCES TABLE](#)
- [LEAD CONTACT AND MATERIALS AVAILABILITY](#)
- [EXPERIMENTAL MODEL AND SUBJECT DETAILS](#)
- [METHOD DETAILS](#)
 - Production of metabolic pathway databases
- [QUANTIFICATION AND STATISTICAL ANALYSIS](#)
 - Analysis of metabolic pathways
 - Identification of metabolic transitions
 - Transcriptome analysis
- [DATA AND CODE AVAILABILITY](#)

SUPPLEMENTAL INFORMATION

Supplemental Information can be found online at <https://doi.org/10.1016/j.cub.2020.02.086>.

ACKNOWLEDGMENTS

N.C. is grateful for a studentship from the Oxford Systems Biology Doctoral Training Program funded by the EPSRC. S.K. and D.M.E. were supported by the European Union's Horizon 2020 research and innovation program under grant agreement number 637765. S.K. is a Royal Society University Research Fellow. A.J.H. was funded by the George Grosvenor Freeman Fellowship by Examination in Sciences, Magdalen College (Oxford).

AUTHOR CONTRIBUTIONS

N.C. and L.J.S. conceived the project and developed the methodology; N.C. wrote code and carried out the investigation and formal analysis; L.J.S. provided supervision; S.K. and D.M.E. provided software and computing resources; J.M., L.D., and A.J.H. provided transcriptome resources; N.C. wrote the manuscript and prepared the figures; L.J.S., L.D., J.M., A.J.H., and D.M.E. reviewed and edited the manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: October 17, 2019

Revised: February 4, 2020

Accepted: February 27, 2020

Published: March 26, 2020

REFERENCES

1. de Vries, J., Stanton, A., Archibald, J.M., and Gould, S.B. (2016). Streptophyte terrestrialization in light of plastid evolution. *Trends Plant Sci.* 21, 467–476.
2. Morris, J.L., Puttick, M.N., Clark, J.W., Edwards, D., Kenrick, P., Pressel, S., Wellman, C.H., Yang, Z., Schneider, H., and Donoghue, P.C.J. (2018). The timescale of early land plant evolution. *Proc. Natl. Acad. Sci. USA* 115, E2274–E2283.
3. Rensing, S.A. (2018). Great moments in evolution: the conquest of land by plants. *Curr. Opin. Plant Biol.* 42, 49–54.
4. Berens, M.L., Berry, H.M., Mine, A., Argueso, C.T., and Tsuda, K. (2017). Evolution of hormone signaling networks in plant defense. *Annu. Rev. Phytopathol.* 55, 401–425.
5. Davies, K.M., Albert, N.W., Zhou, Y., and Schwinn, K.E. (2018). Functions of flavonoid and betalain pigments in abiotic stress tolerance in plants. *Ann. Plant Rev.* Published online April 20, 2018. <https://doi.org/10.1002/9781119312994.apr0604>.
6. Niklas, K.J., Cobb, E.D., and Matas, A.J. (2017). The evolution of hydrophobic cell wall biopolymers: from algae to angiosperms. *J. Exp. Bot.* 68, 5261–5269.
7. Nelson, D., and Werck-Reichhart, D. (2011). A P450-centric view of plant evolution. *Plant J.* 66, 194–211.
8. Basler, G., Fernie, A.R., and Nikoloski, Z. (2018). Advances in metabolic flux analysis toward genome-scale profiling of higher organisms. *Biosci. Rep.* 38. Published online November 23, 2018. <https://doi.org/10.1042/BSR20170224>.
9. Rensing, S.A. (2017). Why we need more non-seed plant models. *New Phytol.* 216, 355–360.
10. Gomes de Oliveira Dal'Molin, C., and Nielsen, L.K. (2018). Plant genome-scale reconstruction: from single cell to multi-tissue modelling and omics analyses. *Curr. Opin. Biotechnol.* 49, 42–48.
11. Moreira, T.B., Lima, J.M., Coca, G.C., and Williams, T.C.R. (2019). Insights into the spatial and temporal organisation of plant metabolism from network flux analysis. *Theor. Exp. Plant Physiol.* 37, 215–226.
12. Rensing, S.A., Lang, D., Zimmer, A.D., Terry, A., Salamov, A., Shapiro, H., Nishiyama, T., Perroud, P.-F., Lindquist, E.A., Kamisugi, Y., et al. (2008).

- The *Physcomitrella* genome reveals evolutionary insights into the conquest of land by plants. *Science* 319, 64–69.
13. Hori, K., Maruyama, F., Fujisawa, T., Togashi, T., Yamamoto, N., Seo, M., Sato, S., Yamada, T., Mori, H., Tajima, N., et al. (2014). *Klebsormidium flaccidum* genome reveals primary factors for plant terrestrial adaptation. *Nat. Commun.* 5, 3978.
 14. Bowman, J.L., Kohchi, T., Yamato, K.T., Jenkins, J., Shu, S., Ishizaki, K., Yamaoka, S., Nishihama, R., Nakamura, Y., Berger, F., et al. (2017). Insights into land plant evolution garnered from the *Marchantia polymorpha* genome. *Cell* 171, 287–304.
 15. Chae, L.; Kim, T.; Nilo-Poyanco, R.; Rhee, S.Y. Genomic signatures of specialized metabolism in plants. *Science* 344, 510–513.
 16. Mueller, L.A., Zhang, P., and Rhee, S.Y. (2003). AraCyc: a biochemical pathway database for *Arabidopsis*. *Plant Physiol.* 132, 453–460.
 17. Nishiyama, T., Sakayama, H., de Vries, J., Buschmann, H., Saint-Marcoux, D., Ullrich, K.K., Haas, F.B., Vanderstraeten, L., Becker, D., Lang, D., et al. (2018). The *Chara* genome: secondary complexity and implications for plant terrestrialization. *Cell* 174, 448–464.
 18. Banks, J.A.; Nishiyama, T.; Hasebe, M.; Bowman, J.L.; Gribskov, M.; DePamphilis, C.; Albert, V.A.; Aono, N.; Aoyama, T.; Ambrose, B.A., et al. The *Selaginella* genome identifies genetic changes associated with the evolution of vascular plants. *Science* 332, 960–963.
 19. Li, F.W., Brouwer, P., Carretero-Paulet, L., Cheng, S., de Vries, J., Delaux, P.M., Eily, A., Koppers, N., Kuo, L.Y., Li, Z., et al. (2018). Fern genomes elucidate land plant evolution and cyanobacterial symbioses. *Nat. Plants* 4, 460–472.
 20. Van Ghelder, C., Parent, G.J., Rigault, P., Prunier, J., Giguère, I., Caron, S., Stival Sena, J., Deslauriers, A., Bousquet, J., Esmejaud, D., and MacKay, J. (2019). The large repertoire of conifer NLR resistance genes includes drought responsive and highly diversified RNLs. *Sci. Rep.* 9, 11614.
 21. Caspi, R., Billington, R., Fulcher, C.A., Keseler, I.M., Kothari, A., Krummenacker, M., Latendresse, M., Midford, P.E., Ong, Q., Ong, W.K., et al. (2018). The MetaCyc database of metabolic pathways and enzymes. *Nucleic Acids Res.* 46 (D1), D633–D639.
 22. Schläpfer, P., Zhang, P., Wang, C., Kim, T., Banf, M., Chae, L., Dreher, K., Chavali, A.K., Nilo-Poyanco, R., Bernard, T., et al. (2017). Genome-wide prediction of metabolic enzymes, pathways, and gene clusters in plants. *Plant Physiol.* 173, 2041–2059.
 23. Karp, P.D., Latendresse, M., Paley, S.M., Krummenacker, M., Ong, Q.D., Billington, R., Kothari, A., Weaver, D., Lee, T., Subhraveti, P., et al. (2016). Pathway Tools version 19.0 update: software for pathway/genome informatics and systems biology. *Brief. Bioinform.* 17, 877–890.
 24. Guan, R., Zhao, Y., Zhang, H., Fan, G., Liu, X., Zhou, W., Shi, C., Wang, J., Liu, W., Liang, X., et al. (2016). Draft genome of the living fossil *Ginkgo biloba*. *Gigascience* 5, 49.
 25. Wan, T., Liu, Z.M., Li, L.F., Leitch, A.R., Leitch, I.J., Lohaus, R., Liu, Z.J., Xin, H.P., Gong, Y.B., Liu, Y., et al. (2018). A genome for gnetophytes and early evolution of seed plants. *Nat. Plants* 4, 82–89.
 26. Neale, D.B., McGuire, P.E., Wheeler, N.C., Stevens, K.A., Crepeau, M.W., Cardeno, C., Zimin, A.V., Puiu, D., Perte, G.M., Sezen, U.U., et al. (2017). The Douglas-Fir genome sequence reveals specialization of the photosynthetic apparatus in Pinaceae. *G3 (Bethesda)* 7, 3157–3167.
 27. Stevens, K.A., Wegrzyn, J.L., Zimin, A., Puiu, D., Crepeau, M., Cardeno, C., Paul, R., Gonzalez-Ibeas, D., Koriabine, M., Holtz-Morris, A.E., et al. (2016). Sequence of the sugar pine megagenome. *Genetics* 204, 1613–1626.
 28. Neale, D.B., Wegrzyn, J.L., Stevens, K.A., Zimin, A.V., Puiu, D., Crepeau, M.W., Cardeno, C., Koriabine, M., Holtz-Morris, A.E., Liechty, J.D., et al. (2014). Decoding the massive genome of loblolly pine using haploid DNA and novel assembly strategies. *Genome Biol.* 15, R59.
 29. Nystedt, B., Street, N.R., Wetterbom, A., Zuccolo, A., Lin, Y.C., Scofield, D.G., Vezzi, F., Delhomme, N., Giacomello, S., Alexeyenko, A., et al. (2013). The Norway spruce genome sequence and conifer genome evolution. *Nature* 497, 579–584.
 30. Xu, Z., Xin, T., Bartels, D., Li, Y., Gu, W., Yao, H., Liu, S., Yu, H., Pu, X., Zhou, J., et al. (2018). Genome analysis of the ancient tracheophyte *Selaginella tamariscina* reveals evolutionary features relevant to the acquisition of desiccation tolerance. *Mol. Plant* 11, 983–994.
 31. Goodstein, D.M., Shu, S., Howson, R., Neupane, R., Hayes, R.D., Fazo, J., Mitros, T., Dirks, W., Hellsten, U., Putnam, N., and Rokhsar, D.S. (2012). Phytozome: a comparative platform for green plant genomics. *Nucleic Acids Res.* 40, D1178–D1186.
 32. Cheng, S., Xian, W., Fu, Y., Marin, B., Keller, J., Wu, T., Sun, W., Li, X., Xu, Y., Zhang, Y., et al. (2019). Genomes of subaerial Zygnematophyceae provide insights into land plant evolution. *Cell* 179, 1057–1067.
 33. Emms, D.M., and Kelly, S. (2019). OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol.* 20, 238.
 34. Leebens-Mack, J.H., Barker, M.S., Carpenter, E.J., Deyholos, M.K., Gitzendanner, M.A., Graham, S.W., Grosse, I., Li, Z., Melkonian, M., Mirarab, S., et al.; One Thousand Plant Transcriptomes Initiative (2019). One thousand plant transcriptomes and the phylogenomics of green plants. *Nature* 574, 679–685.
 35. Van Der Maaten, L., and Hinton, G. (2008). Visualizing data using t-SNE. *J. Mach. Learn. Res.* 9, 2579–2605.
 36. Rozentsvet, O.A. (2004). Comparative examination of distribution of phospholipids and a betaine lipid DGTS in tropical fern species. *Biochem. Syst. Ecol.* 32, 303–311.
 37. Rozentsvet, O.A., Dembitsky, V.M., and Saksonov, S.V. (2000). Occurrence of diacylglyceryltrimethylhomoserines and major phospholipids in some plants. *Phytochemistry* 54, 401–407.
 38. Riekhof, W.R., Naik, S., Bertrand, H., Benning, C., and Voelker, D.R. (2014). Phosphate starvation in fungi induces the replacement of phosphatidylcholine with the phosphorus-free betaine lipid diacylglyceryl-N,N,N-trimethylhomoserine. *Eukaryot. Cell* 13, 749–757.
 39. Novoselov, S.V., Rao, M., Onoshko, N.V., Zhi, H., Kryukov, G.V., Xiang, Y., Weeks, D.P., Hatfield, D.L., and Gladyshev, V.N. (2002). Selenoproteins and selenocysteine insertion system in the model plant cell system, *Chlamydomonas reinhardtii*. *EMBO J.* 21, 3681–3693.
 40. Rodman, J., Soltis, P., Soltis, D., Sytsma, K., and Karol, K. (1998). Parallel evolution of glucosinolate biosynthesis inferred from congruent nuclear and plastid gene phylogenies. *Am. J. Bot.* 85, 997.
 41. Clausen, M., Kannangara, R.M., Olsen, C.E., Blomstedt, C.K., Gleadow, R.M., Jørgensen, K., Bak, S., Motawie, M.S., and Møller, B.L. (2015). The bifurcation of the cyanogenic glucoside and glucosinolate biosynthetic pathways. *Plant J.* 84, 558–573.
 42. Fang, C., Fernie, A.R., and Luo, J. (2019). Exploring the diversity of plant metabolism. *Trends Plant Sci.* 24, 83–98.
 43. Halkier, B.A., and Gershenzon, J. (2006). Biology and biochemistry of glucosinolates. *Annu. Rev. Plant Biol.* 57, 303–333.
 44. Wang, C., Liu, Y., Li, S.S., and Han, G.Z. (2015). Insights into the origin and evolution of the plant hormone signaling machinery. *Plant Physiol.* 167, 872–886.
 45. Hayashi, K., Horie, K., Hiwatashi, Y., Kawaide, H., Yamaguchi, S., Hanada, A., Nakashima, T., Nakajima, M., Mander, L.N., Yamane, H., et al. (2010). Endogenous diterpenes derived from ent-kaurene, a common gibberellin precursor, regulate protonema differentiation of the moss *Physcomitrella patens*. *Plant Physiol.* 153, 1085–1097.
 46. Zi, J., Mafu, S., and Peters, R.J. (2014). To gibberellins and beyond! Surveying the evolution of (di)terpenoid metabolism. *Annu. Rev. Plant Biol.* 65, 259–286.
 47. Miyazaki, S., Hara, M., Ito, S., Tanaka, K., Asami, T., Hayashi, K.-I., Kawaide, H., and Nakajima, M. (2018). An ancestral gibberellin in a moss *Physcomitrella patens*. *Mol. Plant* 11, 1097–1100.
 48. Kumar, S., Kempinski, C., Zhuang, X., Norris, A., Mafu, S., Zi, J., Bell, S.A., Nybo, S.E., Kinison, S.E., Jiang, Z., et al. (2016). Molecular diversity of

- terpene synthases in the liverwort *Marchantia polymorpha*. *Plant Cell* 28, 2632–2650.
49. Noguchi, C., Miyazaki, S., Kawaide, H., Gotoh, O., Yoshida, Y., and Aoyama, Y. (2018). Characterization of moss ent-kaurene oxidase (CYP701B1) using a highly purified preparation. *J. Biochem.* 163, 69–76.
 50. Hirano, K., Nakajima, M., Asano, K., Nishiyama, T., Sakakibara, H., Kojima, M., Katoh, E., Xiang, H., Tanahashi, T., Hasebe, M., et al. (2007). The GID1-mediated gibberellin perception mechanism is conserved in the Lycophyte *Selaginella moellendorffii* but not in the Bryophyte *Physcomitrella patens*. *Plant Cell* 19, 3058–3079.
 51. Huang, Y., Wang, X., Ge, S., and Rao, G.Y. (2015). Divergence and adaptive evolution of the gibberellin oxidase genes in plants genome evolution and evolutionary systems biology. *BMC Evol. Biol.* 15. Published online September 2015. <https://doi.org/10.1186/s12862-015-0490-2>.
 52. Yokota, T., Ohnishi, T., Shibata, K., Asahina, M., Nomura, T., Fujita, T., Ishizaki, K., and Kohchi, T. (2017). Occurrence of brassinosteroids in non-flowering land plants, liverwort, moss, lycophyte and fern. *Phytochemistry* 136, 46–55.
 53. Cheon, J., Fujioka, S., Dilkes, B.P., and Choe, S. (2013). Brassinosteroids regulate plant growth through distinct signaling pathways in *Selaginella* and *Arabidopsis*. *PLoS ONE* 8, e81938.
 54. Ohnishi, T., Szatmari, A.M., Watanabe, B., Fujita, S., Bancos, S., Koncz, C., Lafos, M., Shibata, K., Yokota, T., Sakata, K., et al. (2006). C-23 hydroxylation by *Arabidopsis* CYP90C1 and CYP90D1 reveals a novel shortcut in brassinosteroid biosynthesis. *Plant Cell* 18, 3275–3288.
 55. Rieu, I., Eriksson, S., Powers, S.J., Gong, F., Griffiths, J., Woolley, L., Benlloch, R., Nilsson, O., Thomas, S.G., Hedden, P., and Phillips, A.L. (2008). Genetic analysis reveals that C19-GA 2-oxidation is a major gibberellin inactivation pathway in *Arabidopsis*. *Plant Cell* 20, 2420–2436.
 56. He, J., Chen, Q., Xin, P., Yuan, J., Ma, Y., Wang, X., Xu, M., Chu, J., Peters, R.J., and Wang, G. (2019). CYP72A enzymes catalyse 13-hydroxylation of gibberellins. *Nat. Plants* 5, 1057–1065.
 57. Zhu, Y., Nomura, T., Xu, Y., Zhang, Y., Peng, Y., Mao, B., Hanada, A., Zhou, H., Wang, R., Li, P., et al. (2006). ELONGATED UPPERMOST INTERNODE encodes a cytochrome P450 monooxygenase that epoxidizes gibberellins in a novel deactivation reaction in rice. *Plant Cell* 18, 442–456.
 58. Zhang, Y., Zhang, B., Yan, D., Dong, W., Yang, W., Li, Q., Zeng, L., Wang, J., Wang, L., Hicks, L.M., and He, Z. (2011). Two *Arabidopsis* cytochrome P450 monooxygenases, CYP714A1 and CYP714A2, function redundantly in plant development through gibberellin deactivation. *Plant J.* 67, 342–353.
 59. Magome, H., Nomura, T., Hanada, A., Takeda-Kamiya, N., Ohnishi, T., Shinma, Y., Katsumata, T., Kawaide, H., Kamiya, Y., and Yamaguchi, S. (2013). CYP714B1 and CYP714B2 encode gibberellin 13-oxidases that reduce gibberellin activity in rice. *Proc. Natl. Acad. Sci. USA* 110, 1947–1952.
 60. Varbanova, M., Yamaguchi, S., Yang, Y., McKelvey, K., Hanada, A., Borochoy, R., Yu, F., Jikumaru, Y., Ross, J., Cortes, D., et al. (2007). Methylation of gibberellins by *Arabidopsis* GAMT1 and GAMT2. *Plant Cell* 19, 32–45.
 61. Turk, E.M., Fujioka, S., Seto, H., Shimada, Y., Takatsuto, S., Yoshida, S., Denzel, M.A., Torres, Q.I., and Neff, M.M. (2003). CYP72B1 inactivates brassinosteroid hormones: an intersection between photomorphogenesis and plant steroid signal transduction. *Plant Physiol.* 133, 1643–1653.
 62. Poppenberger, B., Fujioka, S., Soeno, K., George, G.L., Vaistij, F.E., Hirayama, S., Seto, H., Takatsuto, S., Adam, G., Yoshida, S., and Bowles, D. (2005). The UGT73C5 of *Arabidopsis thaliana* glucosylates brassinosteroids. *Proc. Natl. Acad. Sci. USA* 102, 15253–15258.
 63. Marsolais, F., Boyd, J., Paredes, Y., Schinas, A.-M., Garcia, M., Elzein, S., and Varin, L. (2007). Molecular and biochemical characterization of two brassinosteroid sulfotransferases from *Arabidopsis*, AtST4a (At2g14920) and AtST1 (At2g03760). *Planta* 225, 1233–1244.
 64. Delwiche, C.F.; Graham, L.E.; Thomson, N. Lignin-like compounds and sporopollenin in Coleochaete, an algal model for land plant ancestry. *Science* 245, 399–401.
 65. Delwiche, C.F., and Cooper, E.D. (2015). The evolutionary origin of a terrestrial flora. *Curr. Biol.* 25, R899–R910.
 66. He, X., Dai, J., and Wu, Q. (2016). Identification of sporopollenin as the outer layer of cell wall in microalga *Chlorella protothecoides*. *Front. Microbiol.* 7, 1047.
 67. Vishwanath, S.J., Delude, C., Domergue, F., and Rowland, O. (2015). Suberin: biosynthesis, regulation, and polymer assembly of a protective extracellular barrier. *Plant Cell Rep.* 34, 573–586.
 68. Niklas, K.J., Cobb, E.D., and Matas, A.J. (2017). The evolution of hydrophobic cell wall biopolymers: from algae to angiosperms. *J. Exp. Bot.* 68, 5261–5269.
 69. Fich, E.A., Segerson, N.A., and Rose, J.K.C. (2016). The plant polyester cutin: biosynthesis, structure, and biological roles. *Annu. Rev. Plant Biol.* 67, 207–233.
 70. Kondo, S., Hori, K., Sasaki-Sekimoto, Y., Kobayashi, A., Kato, T., Yuno-Ohta, N., Nobusawa, T., Ohtaka, K., Shimojima, M., and Ohta, H. (2016). Primitive extracellular lipid components on the surface of the charophytic alga *Klebsormidium flaccidum* and their possible biosynthetic pathways as deduced from the genome sequence. *Front. Plant Sci.* 7, 952.
 71. Domínguez, E., Heredia-Guerrero, J.A., and Heredia, A. (2015). Plant cutin genesis: unanswered questions. *Trends Plant Sci.* 20, 551–558.
 72. Bajguz, A. (2009). Isolation and characterization of brassinosteroids from algal cultures of *Chlorella vulgaris* Beijerinck (Trebouxiophyceae). *J. Plant Physiol.* 166, 1946–1949.
 73. Caldicott, A.B., and Eglinton, G. (1976). Cutin acids from bryophytes: An ω -1 hydroxy alkanolic acid in two liverwort species. *Phytochemistry* 15, 1139–1143.
 74. Fernández, V., Guzmán-Delgado, P., Graça, J., Santos, S., and Gil, L. (2016). Cuticle structure in relation to chemical composition: Re-assessing the prevailing model. *Front. Plant Sci.* 7, 427.
 75. Huang, R., O'Donnell, A.J., Barboline, J.J., and Barkman, T.J. (2016). Convergent evolution of caffeine in plants by co-option of exapted ancestral enzymes. *Proc. Natl. Acad. Sci. USA* 113, 10613–10618.
 76. Sage, R.F., Christin, P.-A., and Edwards, E.J. (2011). The C(4) plant lineages of planet Earth. *J. Exp. Bot.* 62, 3155–3169.
 77. Weng, J.-K., Akiyama, T., Bonawitz, N.D., Li, X., Ralph, J., and Chapple, C. (2010). Convergent evolution of syringyl lignin biosynthesis via distinct pathways in the lycophyte *Selaginella* and flowering plants. *Plant Cell* 22, 1033–1045.
 78. Claros, M.G., Bautista, R., Guerrero-Fernández, D., Benzerki, H., Seoane, P., and Fernández-Pozo, N. (2012). Why assembling plant genome sequences is so challenging. *Biology (Basel)* 1, 439–459.
 79. De La Torre, A.R., Birol, I., Bousquet, J., Ingvarsson, P.K., Jansson, S., Jones, S.J.M., Keeling, C.I., MacKay, J., Nilsson, O., Ritland, K., et al. (2014). Insights into conifer giga-genomes. *Plant Physiol.* 166, 1724–1732.
 80. Katoh, K., and Standley, D.M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* 30, 772–780.
 81. Capella-Gutiérrez, S., Silla-Martínez, J.M., and Gabaldón, T. (2009). trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* 25, 1972–1973.
 82. Nguyen, L.-T., Schmidt, H.A., von Haeseler, A., and Minh, B.Q. (2015). IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* 32, 268–274.
 83. Benson, D.A., Cavanaugh, M., Clark, K., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J., and Sayers, E.W. (2017). GenBank. *Nucleic Acids Res.* 45 (D1), D37–D42.
 84. Wegrzyn, J.L., Staton, M.A., Street, N.R., Main, D., Grau, E., Herndon, N., Buehler, S., Falk, T., Zaman, S., Ramnath, R., et al. (2019). Cyberinfrastructure to improve forest health and productivity: The role of

- tree databases in connecting genomes, phenomes, and the environment. *Front. Plant Sci.* **10**, 813.
85. Sundell, D., Mannapperuma, C., Netotea, S., Delhomme, N., Lin, Y.C., Sjödin, A., Van de Peer, Y., Jansson, S., Hvidsten, T.R., and Street, N.R. (2015). The Plant Genome Integrative Explorer Resource: *New Phytol.* **208**, 1149–1156 PlantGenIE.org.
86. UniProt Consortium (2019). UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res.* **47** (D1), D506–D515.
87. Jeske, L., Placzek, S., Schomburg, I., Chang, A., and Schomburg, D. (2019). BRENDA in 2019: a European ELIXIR core data resource. *Nucleic Acids Res.* **47** (D1), D542–D549.
88. Buchfink, B., Xie, C., and Huson, D.H. (2015). Fast and sensitive protein alignment using DIAMOND. *Nat. Methods* **12**, 59–60.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT OR RESOURCE	SOURCE	IDENTIFIER
Deposited Data		
<i>Chlamydomonas reinhardtii</i> genome	Phytozome	N/A
<i>Volvox carteri</i> genome	Phytozome	N/A
<i>Coccomyxa subellipsoidea</i> C-169 genome	Phytozome	N/A
<i>Micromonas pusilla</i> CCMP1545 genome	Phytozome	N/A
<i>Micromonas</i> sp. RCC299 v3.0 genome	Phytozome	N/A
<i>Chromochloris zofingiensis</i> genome	Phytozome	N/A
<i>Ostreococcus lucimarinus</i> genome	Phytozome	N/A
<i>Klebsormidium nitens</i> genome	GenBank	GenBank: DF236950-DF238763
<i>Chara braunii</i> genome	[17]	https://bioinformatics.psb.ugent.be/gdb/Chara_braunii/
<i>Mesotaenium endlicherianum</i> genome	[32]	NCBI BioProject: PRJNA543678
<i>Spirogloea muscicola</i> genome	[32]	NCBI BioProject: PRJNA543679
<i>Marchantia polymorpha</i> genome	Phytozome	N/A
<i>Physcomitrella patens</i> genome	Phytozome	N/A
<i>Sphagnum fallax</i> genome	Phytozome	N/A
<i>Selaginella moellendorffii</i> genome	Phytozome	N/A
<i>Selaginella tamariscina</i> genome	[30]	NCBI Biosample: SAMN07840812
<i>Isoetes echinospora</i> transcriptome	GenBank	GenBank: GGKY000000000.1
<i>Salvinia cucullata</i> genome	Fernbase	https://www.fernbase.org/
<i>Azolla filiculoides</i> genome	Fernbase	https://www.fernbase.org/
<i>Ginkgo biloba</i> genome	[24]	https://doi.org/10.5524/100613
<i>Gnetum montanum</i> genome	[25]	https://doi.org/10.5061/dryad.0vm37
<i>Pseudotsuga menziesii</i> genome	TreeGenes	https://treegenesdb.org/
<i>Pinus lambertiana</i> genome	TreeGenes	https://treegenesdb.org/
<i>Pinus taeda</i> genome	ConGenIE	http://congenie.org/
<i>Picea abies</i> genome	ConGenIE	http://congenie.org/
<i>Picea glauca</i> transcriptome	GenBank	GenBank: GGJI000000000.1
<i>Amborella trichopoda</i> genome	Phytozome	N/A
<i>Spirodela polyrrhiza</i> genome	Phytozome	N/A
<i>Zostera marina</i> genome	Phytozome	N/A
<i>Asparagus officinalis</i> genome	Phytozome	N/A
<i>Ananas comosus</i> genome	Phytozome	N/A
<i>Musa acuminata</i> genome	Phytozome	N/A
<i>Brachypodium distachyon</i> genome	Phytozome	N/A
<i>Hordeum vulgare</i> genome	Phytozome	N/A
<i>Triticum aestivum</i> genome	Phytozome	N/A
<i>Oryza sativa</i> genome	Phytozome	N/A
<i>Oropetium thomaeum</i> genome	Phytozome	N/A
<i>Setaria italica</i> genome	Phytozome	N/A
<i>Sorghum bicolor</i> genome	Phytozome	N/A
<i>Zea mays</i> genome	Phytozome	N/A
<i>Chenopodium quinoa</i> genome	Phytozome	N/A
<i>Daucus carota</i> genome	Phytozome	N/A
<i>Lactuca sativa</i> genome	Phytozome	N/A
<i>Helianthus annuus</i> genome	Phytozome	N/A

(Continued on next page)

Continued

REAGENT OR RESOURCE	SOURCE	IDENTIFIER
<i>Solanum lycopersicum</i> genome	Phytozome	N/A
<i>Solanum tuberosum</i> genome	Phytozome	N/A
<i>Mimulus guttatus</i> genome	Phytozome	N/A
<i>Olea europaea</i> genome	Phytozome	N/A
<i>Vitis vinifera</i> genome	Phytozome	N/A
<i>Eucalyptus grandis</i> genome	Phytozome	N/A
<i>Fragaria vesca</i> genome	Phytozome	N/A
<i>Malus domestica</i> genome	Phytozome	N/A
<i>Prunus persica</i> genome	Phytozome	N/A
<i>Cucumis sativus</i> genome	Phytozome	N/A
<i>Glycine max</i> genome	Phytozome	N/A
<i>Medicago truncatula</i> genome	Phytozome	N/A
<i>Cicer arietinum</i> genome	Phytozome	N/A
<i>Linum usitatissimum</i> genome	Phytozome	N/A
<i>Ricinus communis</i> genome	Phytozome	N/A
<i>Manihot esculenta</i> genome	Phytozome	N/A
<i>Citrus clementina</i> genome	Phytozome	N/A
<i>Citrus sinensis</i> genome	Phytozome	N/A
<i>Gossypium raimondii</i> genome	Phytozome	N/A
<i>Theobroma cacao</i> genome	Phytozome	N/A
<i>Carica papaya</i> genome	Phytozome	N/A
<i>Brassica oleracea capitata</i> genome	Phytozome	N/A
<i>Eutrema salsugineum</i> genome	Phytozome	N/A
<i>Arabidopsis lyrata</i> genome	Phytozome	N/A
<i>Arabidopsis thaliana</i> genome	Phytozome	N/A
<i>Boechera stricta</i> genome	Phytozome	N/A
<i>Capsella rubella</i> genome	Phytozome	N/A
<i>Capsella grandiflora</i> genome	Phytozome	N/A
<i>Klebsormidium nitens</i> PGDB	This manuscript	N/A
<i>Chara braunii</i> PGDB	This manuscript	N/A
<i>Marchantia polymorpha</i> PGDB	This manuscript	N/A
<i>Physcomitrella patens</i> PGDB	This manuscript	N/A
<i>Selaginella moellendorffii</i> PGDB	This manuscript	N/A
<i>Salvinia cucullata</i> PGDB	This manuscript	N/A
<i>Picea glauca</i> PGDB	This manuscript	N/A
<i>Arabidopsis thaliana</i> PGDB	[16]	https://www.plantcyc.org/downloads
<i>Encephalartos barteri</i>	1KP Initiative	GNQG
<i>Stangeria eriopus</i>	1KP Initiative	KAWQ
<i>Dioon edule</i>	1KP Initiative	WLIC
<i>Cycas micholitzii</i>	1KP Initiative	XZUY
<i>Ginkgo biloba</i>	1KP Initiative	SGTW
<i>Gnetum montanum</i>	1KP Initiative	GTHK
<i>Ephedra sinica</i>	1KP Initiative	VDAO
<i>Agathis macrophylla</i>	1KP Initiative	ACWS
<i>Chamaecyparis lawsoniana</i>	1KP Initiative	AIGO
<i>Pseudolarix amabilis</i>	1KP Initiative	AQFM
<i>Nothotsuga longibracteata</i>	1KP Initiative	AREG
<i>Widdringtonia cedarbergensis</i>	1KP Initiative	AUDE
<i>Picea engelmannii</i>	1KP Initiative	AWQB

(Continued on next page)

Continued

REAGENT OR RESOURCE	SOURCE	IDENTIFIER
<i>Lepidothamnus</i> sp.	1KP Initiative	BBDD
<i>Austrotaxus spicata</i>	1KP Initiative	BTTS
<i>Platycladus orientalis</i>	1KP Initiative	BUWV
<i>Manoao colensoi</i>	1KP Initiative	CDFR
<i>Tetraclinis</i> sp.	1KP Initiative	CGDN
<i>Cryptomeria japonica</i>	1KP Initiative	DSXO
<i>Pinus radiata</i>	1KP Initiative	DZQM
<i>Torreya taxifolia</i>	1KP Initiative	EFMS
<i>Prumnopitys andina</i>	1KP Initiative	EGLZ
<i>Cunninghamia lanceolata</i>	1KP Initiative	ESYX
<i>Pilgerodendron uviferum</i>	1KP Initiative	ETCJ
<i>Taxodium distichum</i>	1KP Initiative	FHST
<i>Dacrycarpus compactus</i>	1KP Initiative	FMWZ
<i>Calocedrus decurrens</i>	1KP Initiative	FRPM
<i>Tsuga heterophylla</i>	1KP Initiative	GAMH
<i>Cedrus libani</i>	1KP Initiative	GGEA
<i>Cephalotaxus harringtonia</i>	1KP Initiative	GJTI
<i>Diselma archeri</i>	1KP Initiative	GKCZ
<i>Sequoia sempervirens</i>	1KP Initiative	HBGV
<i>Acropyle pancheri</i>	1KP Initiative	HILW
<i>Cryptomeria japonica</i>	1KP Initiative	HOUF
<i>Torreya nucifera</i>	1KP Initiative	HQOM
<i>Amentotaxus argotaenia</i>	1KP Initiative	IAJW
<i>Callitris gracilis</i>	1KP Initiative	IFLI
<i>Pinus parviflora</i>	1KP Initiative	IIOL
<i>Pseudotsuga wilsoniana</i>	1KP Initiative	IOVS
<i>Dacrydium balansae</i>	1KP Initiative	IZGN
<i>Pinus ponderosa</i>	1KP Initiative	JBND
<i>Neocallitropsis pancheri</i>	1KP Initiative	JDQB
<i>Phyllocladus hypophyllus</i>	1KP Initiative	JRNA
<i>Keteleeria evelyniana</i>	1KP Initiative	JUWL
<i>Parasitaxus usta</i>	1KP Initiative	JZVE
<i>Sundacarpus amarus</i>	1KP Initiative	KLGF
<i>Pinus jeffreyi</i>	1KP Initiative	MFTM
<i>Microcachrys tetragona</i>	1KP Initiative	MHGD
<i>Agathis robusta</i>	1KP Initiative	MIXZ
<i>Thujopsis dolabrata</i>	1KP Initiative	NKIN
<i>Cathaya argyrophylla</i>	1KP Initiative	NPRL
<i>Metasequoia glyptostroboides</i>	1KP Initiative	NRXL
<i>Cunninghamia lanceolata</i>	1KP Initiative	OUOI
<i>Papuacedrus papuana</i>	1KP Initiative	OVIJ
<i>Halocarpus bidwillii</i>	1KP Initiative	OWFC
<i>Glyptostrobus pensilis</i>	1KP Initiative	OXGJ
<i>Saxegothaea conspicua</i>	1KP Initiative	QCGM
<i>Sequoiadendron giganteum</i> <i>Glaucum</i>	1KP Initiative	QFAE
<i>Falcatifolium taxoides</i>	1KP Initiative	QHBI
<i>Cupressus dupreziana</i>	1KP Initiative	QNGJ
<i>Taiwania cryptomerioides</i>	1KP Initiative	QSNJ
<i>Callitris macleayana</i>	1KP Initiative	RMMV

(Continued on next page)

Continued

REAGENT OR RESOURCE	SOURCE	IDENTIFIER
<i>Falcatifolium taxoides</i>	1KP Initiative	ROWR
<i>Wollemia nobilis</i>	1KP Initiative	RSCE
<i>Podocarpus coriaceus</i>	1KP Initiative	SCEB
<i>Fokienia hodginsii</i>	1KP Initiative	UEVI
<i>Nageia nagi</i>	1KP Initiative	UUJS
<i>Thuja plicata</i>	1KP Initiative	VFYZ
<i>Retrophillum minus</i>	1KP Initiative	VGSX
<i>Abies lasiocarpa</i>	1KP Initiative	VSRH
<i>Larix speciosa</i>	1KP Initiative	WVWN
<i>Taxus baccata</i>	1KP Initiative	WWSS
<i>Cephalotaxus harringtonia</i>	1KP Initiative	WYAJ
<i>Athrotaxis cupressoides</i>	1KP Initiative	XIRK
<i>Podocarpus rubens</i>	1KP Initiative	XLGK
<i>Juniperus scopulorum</i>	1KP Initiative	XMGP
<i>Microbiota decussata</i>	1KP Initiative	XQSG
<i>Araucaria rulei</i>	1KP Initiative	XTZO
<i>Sciadopitys verticillata</i>	1KP Initiative	YFZK
<i>Pseudotaxus chienii</i>	1KP Initiative	YLPM
<i>Austrocedrus chilensis</i>	1KP Initiative	YYPE
<i>Tmesipteris parva</i>	1KP Initiative	ALVQ
<i>Botrypus virginianus</i>	1KP Initiative	BEGM
<i>Equisetum diffusum</i>	1KP Initiative	CAPN
<i>Danaea nodosa</i>	1KP Initiative	DFHO
<i>Sceptridium dissectum</i>	1KP Initiative	EEAQ
<i>Equisetum hyemale</i>	1KP Initiative	JVSZ
<i>Psilotum nudum</i>	1KP Initiative	QGVV
<i>Psilotum nudum</i>	1KP Initiative	QVMR
<i>Marattia attenuata</i>	1KP Initiative	UGNK
<i>Marattia sp.</i>	1KP Initiative	UXCS
<i>Ophioglossum petiolatum</i>	1KP Initiative	WTJG
<i>Osmundastrum cinnamomeum</i>	1KP Initiative	BIVQ
<i>Asplenium sp.</i>	1KP Initiative	BMIF
<i>Adiantum raddianum</i>	1KP Initiative	BMJR
<i>Polypodium glycyrrhiza</i>	1KP Initiative	CJNT
<i>Anemia tomentosa</i>	1KP Initiative	CQPW
<i>Azolla cf. caroliniana</i>	1KP Initiative	CVEG
<i>Gaga arizonica</i>	1KP Initiative	DCDT
<i>Thyrsopteris elegans</i>	1KP Initiative	EWXK
<i>Deparia lobato-crenata</i>	1KP Initiative	FCHS
<i>Pteris ensiformis</i>	1KP Initiative	FLTD
<i>Polystichum acrostichoides</i>	1KP Initiative	FQGG
<i>Cyathea (Alsophila) spinulosa</i>	1KP Initiative	GANB
<i>Myriopteris rufa</i>	1KP Initiative	GSXD
<i>Polypodium hesperium</i>	1KP Initiative	GYFU
<i>Gymnocarpium dryopteris</i>	1KP Initiative	HEGQ
<i>Cystopteris utahensis</i>	1KP Initiative	HNDZ
<i>Onoclea sensibilis</i>	1KP Initiative	HTFH
<i>Polypodium hesperium</i>	1KP Initiative	IXLH
<i>Bolbitis repanda</i>	1KP Initiative	JBLI

(Continued on next page)

Continued

REAGENT OR RESOURCE	SOURCE	IDENTIFIER
<i>Pilularia globulifera</i>	1KP Initiative	KIIX
<i>Asplenium platyneuron</i>	1KP Initiative	KJZG
<i>Cystopteris fragilis</i>	1KP Initiative	LHLE
<i>Dipteris conjugata</i>	1KP Initiative	MEKP
<i>Thelypteris acuminata</i>	1KP Initiative	MROH
<i>Dennstaedtia davallioides</i>	1KP Initiative	MTGC
<i>Vittaria appalachiana</i>	1KP Initiative	NDUV
<i>Lindsaea linearis</i>	1KP Initiative	NOKI
<i>Nephrolepis exaltata</i>	1KP Initiative	NWWI
<i>Homalosorus pycnocarpus</i>	1KP Initiative	OCZL
<i>Davallia fejeensis</i>	1KP Initiative	OQWW
<i>Phymatosorus grossus</i>	1KP Initiative	ORJE
<i>Lygodium japonicum</i>	1KP Initiative	PBUU
<i>Ceratopteris thalictroides</i>	1KP Initiative	PIVW
<i>Culcita macrocarpa</i>	1KP Initiative	PNZO
<i>Pteris vittata</i>	1KP Initiative	POPJ
<i>Asplenium nidus</i>	1KP Initiative	PSKY
<i>Hymenophyllum bivalve</i>	1KP Initiative	QIAD
<i>Osmundastrum cinnamomeum</i>	1KP Initiative	RFMZ
<i>Didymochlaena truncatula</i>	1KP Initiative	RFRB
<i>Ricciocarpos natans</i>	1KP Initiative	RICC
<i>Vittaria lineata</i>	1KP Initiative	SKYV
<i>Hymenophyllum cupressiforme</i>	1KP Initiative	TRPJ
<i>Crepidomanes venosum</i>	1KP Initiative	TWFZ
<i>Diplazium wichurae</i>	1KP Initiative	UFJN
<i>Pityrogramma trifoliata</i>	1KP Initiative	UJTT
<i>Pleopeltis polypodioides</i>	1KP Initiative	UJWU
<i>Osmunda sp.</i>	1KP Initiative	UOMY
<i>Athyrium filix-femina</i>	1KP Initiative	URCP
<i>Plagiogyria japonica</i>	1KP Initiative	UWOD
<i>Osmunda javanica</i>	1KP Initiative	VIBO
<i>Blechnum spicant</i>	1KP Initiative	VITX
<i>Lonchitis hirsuta</i>	1KP Initiative	VVRN
<i>Adiantum aleuticum</i>	1KP Initiative	WCLG
<i>Leucostegia immersa</i>	1KP Initiative	WGTU
<i>Cryptogramma acrostichoides</i>	1KP Initiative	WQML
<i>Argyrochosma nivea</i>	1KP Initiative	XDDT
<i>Sticherus lobatus</i>	1KP Initiative	XDVM
<i>Cystopteris fragilis</i>	1KP Initiative	XXHP
<i>Notholaena montieliae</i>	1KP Initiative	YCKE
<i>Lindsaea microphylla</i>	1KP Initiative	YIXP
<i>Woodsia scopulina</i>	1KP Initiative	YJJY
<i>Osmunda regalis</i>	1KP Initiative	YKSS
<i>Polypodium amorphum</i>	1KP Initiative	YLJA
<i>Cystopteris protrusa</i>	1KP Initiative	YOWV
<i>Woodsia ilvensis</i>	1KP Initiative	YQEC
<i>Phlebodium pseudoaureum</i>	1KP Initiative	ZQYU
<i>Parahemionitis cordata</i>	1KP Initiative	ZXJO
<i>Selaginella lepidophylla</i>	1KP Initiative	ABIJ

(Continued on next page)

Continued

REAGENT OR RESOURCE	SOURCE	IDENTIFIER
<i>Huperzia myrsinites</i>	1KP Initiative	CBAE
<i>Lycopodium annotinum</i>	1KP Initiative	ENQF
<i>Huperzia squarrosa</i>	1KP Initiative	GAON
<i>Huperzia lucidula</i>	1KP Initiative	GKAG
<i>Huperzia selago</i>	1KP Initiative	GTUO
<i>Selaginella wallacei</i>	1KP Initiative	JKAA
<i>Selaginella willdenowii</i>	1KP Initiative	KJYC
<i>Selaginella selaginoides</i>	1KP Initiative	KUXM
<i>Selaginella apoda</i>	1KP Initiative	LGDQ
<i>Huperzia selago</i>	1KP Initiative	NYBX
<i>Isoetes tegetiformans</i>	1KP Initiative	PKOX
<i>Lycopodium deuterodensum</i>	1KP Initiative	PQTO
<i>Isoetes</i> sp.	1KP Initiative	PYHZ
<i>Lycopodiella appressa</i>	1KP Initiative	ULKT
<i>Pseudolycopodiella caroliniana</i>	1KP Initiative	UPMJ
<i>Diphasiastrum digitatum</i>	1KP Initiative	WAFT
<i>Dendrolycopodium obscurum</i>	1KP Initiative	XNXF
<i>Selaginella kraussiana</i>	1KP Initiative	ZFGK
<i>Selaginella stauntoniana</i>	1KP Initiative	ZZOL
<i>Blasia</i> sp.	1KP Initiative	AEXY
<i>Radula lindenbergiana</i>	1KP Initiative	BNCU
<i>Frullania</i> spp.	1KP Initiative	CHJJ
<i>Treubia lacunosa</i>	1KP Initiative	FITN
<i>Sphaerocarpos texanus</i>	1KP Initiative	HERT
<i>Marchantia paleacea</i>	1KP Initiative	HMHL
<i>Ptilidium pulcherrimum</i>	1KP Initiative	HPXA
<i>Marchantia paleacea</i>	1KP Initiative	IHWQ
<i>Conocephalum conicum</i>	1KP Initiative	ILBQ
<i>Scapania nemorosa</i>	1KP Initiative	IRBN
<i>Barbilophozia barbata</i>	1KP Initiative	JHFI
<i>Marchantia polymorpha</i>	1KP Initiative	JPYU
<i>Porella navicularis</i>	1KP Initiative	KRUQ
<i>Schistochila</i> sp.	1KP Initiative	LGOW
<i>Metzgeria crassipilis</i>	1KP Initiative	NRWZ
<i>mixed species of Plagiochilaceae/Fissidentaceae</i>	1KP Initiative	NWQC
<i>Barbilophozia barbata</i>	1KP Initiative	OFTV
<i>Pellia</i> sp. (cf. <i>epiphylla</i> (L.) Corda)	1KP Initiative	PIUF
<i>Calypogeia fissa</i>	1KP Initiative	RTMU
<i>Monoclea gottschei</i>	1KP Initiative	TFDQ
<i>Marchantia emarginata</i>	1KP Initiative	TFYI
<i>Frullania</i> sp.	1KP Initiative	TGKW
<i>Lunularia cruciata</i>	1KP Initiative	TXVB
<i>Porella pinnata</i>	1KP Initiative	UUHD
<i>Ricciocarpos natans</i>	1KP Initiative	WJLO
<i>Bazzania trilobata</i>	1KP Initiative	WZYK
<i>Odontoschisma prostratum</i>	1KP Initiative	YBQN
<i>Pallavicinia lyellii</i>	1KP Initiative	YFGP
<i>Noteroclada confluens</i>	1KP Initiative	YPSN

(Continued on next page)

Continued

REAGENT OR RESOURCE	SOURCE	IDENTIFIER
<i>Racomitrium elongatum</i>	1KP Initiative	ABCD
<i>Diphyscium foliosum</i>	1KP Initiative	AWOI
<i>Scouleria aquatica</i>	1KP Initiative	BPSG
<i>Orthotrichum lyellii</i>	1KP Initiative	CMEQ
<i>Thuidium delicatulum</i>	1KP Initiative	EEMJ
<i>Ceratodon purpureus</i>	1KP Initiative	FFPD
<i>Sphagnum lescurii</i>	1KP Initiative	GOWD
<i>Syntrichia princeps</i>	1KP Initiative	GRKU
<i>Buxbaumia aphylla</i>	1KP Initiative	HRWG
<i>Tetraphis pellucida</i>	1KP Initiative	HVBQ
<i>Leucodon julaceus</i>	1KP Initiative	IGUH
<i>Rhynchostegium serrulatum</i>	1KP Initiative	JADL
<i>Encalypta streptocarpa</i>	1KP Initiative	KEFD
<i>Stereodon subimponens</i>	1KP Initiative	LNSF
<i>Climacium dendroides</i>	1KP Initiative	MIRS
<i>Dicranum scoparium</i>	1KP Initiative	NGTD
<i>Pseudotaxiphyllum elegans</i>	1KP Initiative	QKQO
<i>Anomodon attenuatus</i>	1KP Initiative	QMWB
<i>Sphagnum palustre</i>	1KP Initiative	RCBT
<i>Racomitrium varium</i>	1KP Initiative	RDOO
<i>Leucobryum glaucum</i>	1KP Initiative	RGKI
<i>Takakia lepidozoides</i>	1KP Initiative	SKQD
<i>Polytrichum commune</i>	1KP Initiative	SZYG
<i>Calliergon cordifolium</i>	1KP Initiative	TAVP
<i>Neckera douglasii</i>	1KP Initiative	TMAJ
<i>Sphagnum recurvum</i>	1KP Initiative	UHLI
<i>Claopodium rostratum</i>	1KP Initiative	VBMM
<i>Leucobryum albidum</i>	1KP Initiative	VMXJ
<i>Aulacomnium heterostichum</i>	1KP Initiative	WNGH
<i>Loeskeobryum brevirostre</i>	1KP Initiative	WSPM
<i>Rosulabryum cf. capillare</i>	1KP Initiative	XWHK
<i>Physcomitrium sp.</i>	1KP Initiative	YEPO
<i>Hedwigia ciliata</i>	1KP Initiative	YWNF
<i>Leucodon brachypus</i>	1KP Initiative	ZACW
<i>Timmia austriaca</i>	1KP Initiative	ZQRI
<i>Atrichum angustatum</i>	1KP Initiative	ZTHV
<i>Phaeomegaceros coriaceus</i>	1KP Initiative	AKXB
<i>Leiosporoceros dussii</i>	1KP Initiative	ANON
<i>Anthoceros agrestis</i>	1KP Initiative	BSNI
<i>Nothoceros aenigmaticus</i>	1KP Initiative	DXOU
<i>Paraphymatoceros hallii</i>	1KP Initiative	FAJB
<i>Leiosporoceros dussii</i>	1KP Initiative	FANS
<i>Anthoceros angustus</i>	1KP Initiative	IQJU
<i>Phaeoceros carolinianus</i>	1KP Initiative	RXRQ
<i>Nothoceros vincentianus</i>	1KP Initiative	TCBC
<i>Anthoceros agrestis</i>	1KP Initiative	TWUW
<i>Megaceros flagellaris</i>	1KP Initiative	UCRN
<i>Phaeoceros carolinianus</i>	1KP Initiative	WCZB
<i>Phaeoceros carolinianus</i>	1KP Initiative	WEEQ

(Continued on next page)

Continued

REAGENT OR RESOURCE	SOURCE	IDENTIFIER
<i>Phaeoceros carolinianus</i>	1KP Initiative	ZFRE
<i>Penium margaritaceum</i>	1KP Initiative	AEKF
<i>Entransia fimbria ta</i>	1KP Initiative	BFIK
<i>Cosmarium tinctorum</i>	1KP Initiative	BHBK
<i>Desmidium aptogonum</i>	1KP Initiative	DFDS
<i>Closterium lunula</i>	1KP Initiative	DRFX
<i>Chaetosphaeridium globosum</i>	1KP Initiative	DRGY
<i>Netrium digitus</i>	1KP Initiative	FFGR
<i>Klebsormidium subtile</i>	1KP Initiative	FQLP
<i>Xanthidium antilopaeum</i>	1KP Initiative	GBGT
<i>Onychonema laeve</i>	1KP Initiative	GGWH
<i>Euastrum affine</i>	1KP Initiative	GYRP
<i>Spirogyra</i> sp.	1KP Initiative	HAOX
<i>Cosmarium broomei</i>	1KP Initiative	HIDG
<i>Cosmarium ochthodes</i>	1KP Initiative	HJVM
<i>Staurostrum sebaldi</i>	1KP Initiative	ISHC
<i>Cylindrocystis cushleackae</i>	1KP Initiative	JOJQ
<i>Gonatozygon kinahanii</i>	1KP Initiative	KEYW
<i>Nucleotaenium eifelense</i>	1KP Initiative	KMNX
<i>Micrasterias fimbriata</i>	1KP Initiative	MCHJ
<i>Zygnemopsis</i> sp.	1KP Initiative	MFZO
<i>Cosmarium granatum</i>	1KP Initiative	MNNM
<i>Pleurotaenium trabecula</i>	1KP Initiative	MOYY
<i>Chara vulgaris</i>	1KP Initiative	MWXT
<i>Mesotaenium kramstae</i>	1KP Initiative	NBYP
<i>Spirotaenia minuta</i>	1KP Initiative	NNHQ
<i>Coleochaete irregularis</i>	1KP Initiative	QPDY
<i>Bambusina borreri</i>	1KP Initiative	QWV
<i>Cylindrocystis brebissonii</i>	1KP Initiative	RPGL
<i>Phymatodocis nordstedtiana</i>	1KP Initiative	RPQV
<i>Staurodesmus omearii</i>	1KP Initiative	RPRU
<i>Cosmocladium cf. constrictum</i>	1KP Initiative	RQFE
<i>Planotaenium ohtanii</i>	1KP Initiative	SNOX
<i>Cosmarium ochthodes</i>	1KP Initiative	STKJ
<i>Spirotaenia</i> sp.	1KP Initiative	TPHT
<i>Cylindrocystis</i> sp.	1KP Initiative	VAZE
<i>Coleochaete scutata</i>	1KP Initiative	VQBJ
<i>Staurodesmus convergens</i>	1KP Initiative	WCQU
<i>Mesotaenium endlicherianum</i>	1KP Initiative	WDCW
<i>Cosmarium subtumidum</i>	1KP Initiative	WDGV
<i>Zygnema</i> sp.	1KP Initiative	WGMD
<i>Mesotaenium braunii</i>	1KP Initiative	WSJO
<i>Roya obtusa</i>	1KP Initiative	XRTZ
<i>Cylindrocystis brebissonii</i>	1KP Initiative	YOXI
<i>Penium exiguum</i>	1KP Initiative	YSQT
<i>Mougeotia</i> sp.	1KP Initiative	ZRMT
Software and Algorithms		
E2P2	[22]	https://gitlab.com/rhee-lab/E2P2/tree/master
Pathway Tools	[23]	RRID:SCR_013786; http://bioinformatics.ai.sri.com/ptools/

(Continued on next page)

Continued

REAGENT OR RESOURCE	SOURCE	IDENTIFIER
OrthoFinder	[33]	RRID:SCR_017118; https://github.com/davideemms/OrthoFinder
MAFFT	[80]	RRID:SCR_011811; https://mafft.cbrc.jp/alignment/software/
trimAl	[81]	RRID:SCR_017334; http://trimal.cgenomics.org/
IQ-TREE	[82]	RRID:SCR_017254; http://www.iqtree.org/
Metabolic Innovation Identifier Script	This manuscript	N/A

LEAD CONTACT AND MATERIALS AVAILABILITY

This study did not generate new unique reagents.

A key resources table is included. Datasets and code generated during this study are available in [Figures S1, S2, S3, S4, S5, and S6](#); [Table 1](#); [Tables S1 and S2](#); and [Data S1, S2, S3, S4, S5, and S6](#).

Requests for further information and resources should be directed to and will be fulfilled by the Lead Contact, Lee Sweetlove (lee.sweetlove@plants.ox.ac.uk).

EXPERIMENTAL MODEL AND SUBJECT DETAILS

This study used a total of 377 genome and transcriptome datasets. All angiosperm genome sequences used in the orthogroup analysis were downloaded from the Phytozome database (<https://phytozome.jgi.doe.gov/>). Non-angiosperm genome and transcriptome sequences used in the orthogroup analysis were either downloaded from Phytozome, GenBank [83], Fernbase [19], TreeGenes [84], or ConGenie [85], or from individual publications; specific sources for each species can be found in the key resources table. All 305 transcriptomes used in the transcriptome analysis were taken from the One Thousand Plant Transcriptomes Initiative [34], and unique identifiers for each species are documented in the key resources table.

METHOD DETAILS**Production of metabolic pathway databases**

Metabolic pathway/genome databases (PGDBs) for seven streptophyte species (*Klebsormidium nitens*, *Chara braunii*, *Marchantia polymorpha*, *Physcomitrella patens*, *Selaginella moellendorffii*, *Salvinia cucullata* and *Picea glauca*) were generated ([Data S1](#)) using the Pathway Tools software [23], which uses genome annotations to infer the presence of metabolic reactions in an organism and a rule-based approach to import entire metabolic pathways once certain thresholds of reaction evidence are reached ([Data S5](#)). The software was set up to utilize both MetaCyc and PlantCyc as reference metabolic pathway databases. Pathway Tools computes a likelihood score, incorporating information about the proportion of metabolic reactions in the pathway with corresponding annotations in an organism's genome, the uniqueness of these reactions across metabolic pathways and whether any are considered 'key' reactions to the pathway. A pathway is imported into the PGDB if the likelihood score is greater than a specified threshold value. This was set relatively low (0.15) to minimize the number of false negative pathway inferences. To be included in a PGDB, a pathway must also meet several further criteria: it must be a natural metabolic pathway (not genetically engineered), must not be missing any specified key reactions, and must have an expected taxonomic range matching the organism in question. The latter criterion is problematic for studies such as this which include earlier-diverging land plants and algae on which relatively little metabolic research has been carried out, increasing the likelihood of false negatives and limiting our ability to extend current knowledge beyond well-studied higher plants. Despite this taxonomy restriction, the Pathway Tools software also intrinsically reduces the number of false negative inferences caused by missing genome annotations – given a sufficiently high pathway likelihood score, complete metabolic pathways are incorporated into a PGDB, including any reactions for which an associated genome annotation has not been found. The low pathway likelihood score inclusion threshold maximizes this effect.

Efforts were made to further mitigate the possibility of false negative inferences by improving the genome annotations on which the PGDBs were based ([Data S5](#)). Prior to PGDB construction, the raw protein sequences of all chosen species were passed through the E2P2 software (Ensemble-Enzyme Prediction Pipeline [22]), a machine learning-based algorithm developed for the extraction of metabolism-specific information from genome sequences. The algorithm is trained on a set of more than 142,000 protein sequences, compiled from enzyme and non-enzyme protein sequences from any organism taken from SwissProt [86], BRENDA [87], MetaCyc or PlantCyc, with the stipulation that the entry must either have been manually curated or has experimental support [22]. E2P2 automatically annotates input sequences for each organism with EC numbers and MetaCyc-specific reaction identifiers (E2P2 annotations for each of the seven species are available in [Data S1](#)). Following genome annotation with E2P2, Pathway Tools was used to produce two PGDBs, individually based on the previously published genome annotation for each organism and the E2P2 annotation produced here. The union of the metabolic pathways inferred in each case became the final PGDB for each organism. It is important to note that since this method, like the majority of gene annotation efforts, is based on sequence homology and protein domain identification, it has little power to identify novel metabolic enzymes (and therefore reactions and pathways). Rather, this approach attempts to

produce the most complete set of metabolic pathways possible for each organism, given the totality of existing knowledge of metabolic reactions.

QUANTIFICATION AND STATISTICAL ANALYSIS

Analysis of metabolic pathways

The PGDBs for each of the seven streptophyte species detailed above and the well-curated *Arabidopsis* PGDB [16] were consolidated into a binary presence/absence table for all metabolic pathways present in any species PGDB. To compare the patterns of metabolic pathway presence/absence across species, k-medoids clustering followed by an unsupervised clustering technique known as t-distributed stochastic neighbor embedding (t-SNE [35]) was carried out. For the clustering of species, t-SNE was run 500,000 times and the solution with the minimal final cost function value was chosen.

Identification of metabolic transitions

For a broader multi-species comparison, 64 plant and algal species in addition to the eight species described above were compared, taking the gene-reaction associations identified in the eight original species as the starting point. The additional species include seven chlorophyte algae, two further charophytes (*Spiroglaea muscicola* and *Mesotaenium endlicherianum*), one further bryophyte (*Sphagnum fallax*), two further lycophytes (*Selaginella tamariscina* and *Isoetes echinospora*), one further fern (*Azolla filiculoides*), six further gymnosperms and 45 further angiosperms. All angiosperm genome sequence data were taken from the Phytozome online repository [31]; sources for additional non-angiosperm species are identified in the key resources table. Genome sequence data was used for all additional species except the lycophyte *I. echinospora*, for which the transcriptome was used (available on NCBI's GenBank [83], GenBank: GGKY00000000.1). Genes for all 72 species were sorted into orthogroups using OrthoFinder software [33], which infers orthogroups of genes across species from sequence data (where an orthogroup is a collection of orthologous and paralogous genes descending from one gene in the common ancestor of all species included in the analysis). Orthogroups are produced using an all-versus-all DIAMOND [88] search followed by the normalization of the resulting scores for gene length and phylogenetic distance. A gene network is then constructed, weighted with the normalized scores and this network is clustered into orthogroups using Markov clustering (MCL). Orthogroups for all 72 species and corresponding orthogroup gene trees are available as [Data S2](#) and [Data S3](#). A script to query this orthogroup data based on *Arabidopsis* accession IDs is provided in [Data S4](#).

Both the gene-reaction association data and the orthogroup data were processed using a Python script (available as [Data S6](#)) to identify pathways representing possible metabolic innovations or losses in land plant evolution. For every metabolic pathway across all representative species PGDBs, support for each metabolic reaction in the pathway in each of the genomes of the eight representative species was tabulated from the gene-reaction association data used by Pathway Tools. Equivalent support for each metabolic reaction in the 72 species on which orthogroups had been calculated was identified mainly on the basis of *Arabidopsis* annotations since *Arabidopsis* is the best studied species included in the analysis and likely has the highest quality genome annotation. If an *Arabidopsis* gene is annotated with a reaction, genes in other organisms found in the orthogroup corresponding to this *Arabidopsis* gene are considered functionally homologous and associated with the same reaction. If no *Arabidopsis* gene is annotated with a reaction, but at least two of the seven further representative species' genome annotations contain annotations for this reaction, then as long as an orthogroup exists which contains at least one such annotated gene from all the representative species, this orthogroup is considered associated with the reaction and corresponding genes in other organisms in this orthogroup are again considered functionally homologous.

Evidence for each metabolic pathway from both the gene-reaction associations and OrthoFinder were consolidated into a set of tables containing this information across the eight representative species ([Data S5](#)). Reaction presence was evaluated sequentially from earlier-diverging to later-diverging species. If this progression along the phylogeny was associated with a single change in the proportion of reactions evidenced (e.g., at a single point the proportion of reactions for which there is genetic evidence increases and remains high for the rest of the progression, or decreases and remains low), the metabolic pathway was flagged as involved in a possible metabolic innovation or loss in land plant evolution and sent to a results file containing all such pathways. The orthogroups associated with each reaction in each flagged metabolic pathway were then inspected to confirm that the evidence pattern held across the 64 further species included in the orthogroups. Evidence for each flagged metabolic pathway in each of the eight species annotated was scrutinised to remove false positive hits (e.g., where existing literature showed the identified pattern to be incorrect) and those for which limited reaction information was available, and to ensure that assumptions were not being made based on unreasonably large orthogroups. Some orthogroups are too large to assume that the spectrum of included genes may have similar function; in these situations gene trees for each orthogroup were investigated and genes were only considered candidates for functional homology if closer relationships were observed within subtrees. For the purposes of removing false positives, the gene trees were investigated when genes required for pathways fell in one of the largest 200 orthogroups (containing more than 860 genes across the 72 species). For the genes presented in the results section of this work, all gene trees of any size were inspected.

Transcriptome analysis

Seven blast databases were created from all available transcripts for charophytes, liverworts, mosses, hornworts, lycophytes, ferns and gymnosperms individually using DIAMOND. For each gene, we identified whether reverse blast hits of the *Arabidopsis* gene in question are present in the relevant species' transcriptomes. Genes from two species are considered to be reverse blast hits when

each is the top-scoring DIAMOND result in a sequence comparison against the other species' genome/transcriptome. For the initial blast, the total number of DIAMOND results was set to 100, and the single top scoring genes from the top 15 scoring species were extracted for reverse-comparison in a second DIAMOND search back against the *Arabidopsis* genome. Transcripts from each set of species which pass this reciprocal test were added to the multiple sequence alignment of all the genes in the corresponding orthogroup using MAFFT [80] and the resulting alignments trimmed using trimAl [81] with the 'gappyout' option. We then built a phylogenetic tree from the new multiple sequence alignment using IQ-TREE [82] with the model 'LG+I+G4' and inspected each phylogeny manually to ensure that the transcripts identified are indeed most closely related to the original *Arabidopsis* gene in question (as opposed to forming an outgroup, or being homologs of alternative *Arabidopsis* genes).

DATA AND CODE AVAILABILITY

The published article contains all datasets, code and supporting information generated during this study.