The evolution of the plant organellar proteome



Rona Costello

Linacre College

Department of Plant Sciences

University of Oxford

Thesis submitted for the degree of Doctor of Philosophy

Hilary 2021

Abstract

A hallmark of eukaryotic cells is the compartmentalisation of intracellular processes into specialised, membrane-bound compartments known as organelles. The function of organelles is dependent on a suite of proteins localised within them (the organellar proteome), the majority of which are encoded in the nucleus and targeted to the correct organelle from the cytosol. Divergent evolution of organelles in different eukaryotic lineages is thus expected to arise, in part, due to differences in the set of nuclear-encoded proteins targeted to them. However, neither the rate at which differences in protein targeting accumulate, nor the evolutionary consequences of these changes, are known. The research presented in this thesis aims to address this gap in our knowledge through the development of a phylogenomic approach to identify the changes in organellar protein targeting that have occurred during the evolution of a diverse set of land plant species. It is revealed that there has been considerable, and continual, modulation of organellar proteomes during plant evolution. The rate of change in protein targeting is assessed, and a previously hidden link between gene duplication and the rate of organellar proteome evolution is uncovered. Next, the focus is narrowed, and the changes in organellar protein targeting that occurred during the evolution of a particular plant trait – C_4 photosynthesis – in two independent C_4 grass lineages are investigated. Of particular interest here is the identification of a novel beta-glucosidase enzyme that was relocated to the chloroplast during C_4 evolution, and that may function to activate chloroplast development in the bundle sheath cells; a key feature of C_4 leaves. This analysis also highlights a gap in our understanding of the C_4 biochemical pathway in the crop plant maize. Specifically, it is unknown how the C_4 metabolite aspartate is processed in the bundle sheath cells and thus an attempt is made to identify and experimentally characterise a novel enzyme that might catalyse this missing reaction. Together, the chapters of this thesis explore the mechanisms by which molecular sequence evolution leads to changes in protein function through changes in protein localisation and the consequences thereof at a macro- (e.g., organellar evolution) and micro- (e.g., subcellular pathways in a single species) level.

“When dealing with events that happened in the remote past, perhaps billions of years ago, it is rarely possible to find definitive answers. Nonetheless, it is possible to use what we know, or think we know, to narrow down the list of possibilities. There are clues scattered throughout life, sometimes in the most unexpected places (…). These clues allow us to eliminate some possibilities, and focus on others, after the method of Sherlock Holmes. As Holmes put it: ‘When you have eliminated the impossible, whatever remains, however improbable, must be the truth.’. While it is dangerous to brandish terms like impossible at evolution, there is sleuthful satisfaction in reconstructing the most likely paths that life might have taken.”

Extract from Nick Lane’s introduction to the 2nd edition of
Power, Sex, Suicide: Mitochondria and the meaning of life

Acknowledgements

I would firstly like to thank my supervisor, Steve Kelly, for all his support and guidance throughout the PhD. I have learnt an incredible amount under your supervision and have been continually inspired by your approach to biological research. Thank you for being so generous with your time and energy, and for all your words of encouragement during the final stages of writing up. I would also like to extend my thanks to David Emms for taking the time to support the endeavours of my project and for all the helpful discussions we've had. In this regard, thanks also to Ursula Flores-Perez for showing me the ropes when I first started in the lab, and Dana Vlad for continually providing expert guidance on, well, most lab things. Thank you to Nick Crang for all his advice (shoulder to cry on) regarding protein purification and Barney Geddes for the indispensable crash course in molecular biology at the start of my PhD.

To everyone else in the Kelly Lab, past and present, I extend my warmest thanks for your company over these past few years. You've been a lovely group of people to work with and I am grateful to you all for making the Kelly Lab such a supportive, calm, and fruitful place to work. In particular to my fellow PhD students – Emily, Ross, Michael, Peatrie, Pippa, Ollie, Lisa, and Jacques – thank you for all the academic and emotional support, and for the steady supply of PhD-related memes…I couldn't have done it without them. To members of the Langdale Lab also, thank you for your equally calm and supportive presence and, most importantly, for the abundance of cake I've eaten throughout this PhD. Inordinate thanks to goes to our lab manager Julie Bull without whom nothing would be possible, and to our lab technicians Lizzie Jamison and Rob Pinches.

I would also like to thank the organisations, and staff within them, that have funded and supported this PhD. They are BBSRC, the Oxford DTP, the Department of Plant Sciences, and the Gatsby Plant Sciences Network. On this latter one, many thanks go to Roxanna Clayton for her hard work organising several excellent PhD training and network weekends, and of course for driving me to most of them.

I am immensely grateful to the friends that have supported me throughout this PhD. The list is certainly too long to include here, but my heartfelt thanks go to Amy Howard, Jenny Barrett, Fernanda Carneiro, and Lorna James for bringing much joy to my post-graduate experience. To my friends at OUUEG, thank you all for the many excellent pub nights, diving trips, and general silliness over the years.

Bundles of gratitude go to my family, on all sides. Your unwavering support, love, and – at times – shelter have kept me going throughout. Finally, an immeasurable thanks goes to Jack Leland. This has been a crazy journey and I couldn't have wished for a better person to share it with.

Table of Contents

Abstract ..... ***i***

Acknowledgements ..... ***iii***

Abbreviations ..... ***xiii***

Chapter 1: General Introduction ***1***

1.1 Introduction **1**

1.2 The majority of organellar proteins are encoded in the nucleus and synthesised in the cytosol **2**

1.3 Target signals direct protein sorting in the cell **6**

1.4 Organellar proteomes are highly dynamic in time and space **10**

1.5 Changes in protein targeting as a mechanism of organellar proteome evolution **12**

1.6 Computational prediction of target signals **16**

1.6.1 A brief history of target signal prediction 16

1.7 Molecular phylogenies can be used to reconstruct ancestral protein function **20**

1.8 C_4 photosynthesis is a plant trait split between several subcellular compartments **21**

1.9 Thesis aims and overview **24**

Chapter 2: Changes in organellar protein targeting have occurred at a continuous rate throughout plant evolution and the pace of change is accelerated by gene duplication ***25***

2.1 Introduction 25

2.1.1 Background 25

2.1.2 Author Contributions 27

2.2 Results 28

2.2.1 Predicted organellar proteome size correlates with total nuclear proteome size 28

2.2.2 Widespread gain and loss of organellar target signals have occurred throughout the evolution of plants 30

2.2.3 Changes to the organellar proteome have occurred continuously throughout plant evolution 34

2.2.4 Changes in organellar protein targeting are associated with an increased rate of molecular sequence evolution 36

2.2.5 Changes in organellar protein targeting occur more frequently following gene duplication 37

2.2.6 There is no difference in the frequency of changes in organellar protein targeting following single-gene or whole-genome duplications 40

2.2.7 Changes in the regulatory machinery are the main changes occurring to the chloroplast and mitochondrial proteomes 42

2.2.8 Using an alternative gene-tree inference method replicates the main findings presented in this chapter 46

2.2.9 Using alternative organellar target signal prediction tools replicates the main findings presented in this chapter 49

2.3 Discussion 52

2.4 Materials and methods 58

2.4.1 Problem definition and approach 58

2.4.2 Availability of data and material 58

2.4.3 Construction of orthogroups and inference of species-tree-reconciled gene trees 58
2.4.4 Prediction of organellar-targeted proteins 60
2.4.5 Ancestral state estimation of organellar protein targeting 61
2.4.6 Identification of changes in organellar protein targeting during evolution .. 62
2.4.7 Comparing the rate of molecular sequence evolution for gene-tree branches associated with and without a change in protein targeting 63
2.4.8 Incorporation of divergence time data 64
2.4.9 Identification of changes in organellar protein targeting following gene duplication and non-duplication (speciation) events 65
2.4.10 Functional term enrichment analysis 66
2.4.11 Monte Carlo resampling 67

Chapter 3: Identification of changes in protein targeting associated with the evolution of C_4 photosynthesis 68

3.1 Introduction 68
3.1.1 Background 69

3.2 Results 77
3.2.1 Identification of the changes in protein targeting that occurred concurrent with the evolution of C_4 photosynthesis in two monocot lineages 77
3.2.2 Species-tree branches encompassing an origin of C_4 are associated with an elevated rate of change in protein targeting to the chloroplast and secretory pathway 79
3.2.3 No ancestral changes in protein targeting were identified among known C_4-cycle proteins coincident with the evolution of C_4 photosynthesis 81
3.2.4 Incorporation of transcriptomic data sets to identify changes in protein targeting relevant to the evolution of C_4 photosynthesis 83

3.2.5 Investigation of ancestral changes in protein targeting among extant genes with high leaf-expression 85
3.2.6 A putative C_4-related change in protein targeting was identified for a beta-glucosidase involved in cytokinin signalling in maize 92
3.2.7 Analysis of a change in targeting for an aspartate aminotransferase highlights a gap in our understanding of C_4 photosynthesis in maize 101
3.3 Discussion 105
3.4 Materials and methods 110
3.4.1 Orthogroup and phylogenetic tree inference 110
3.4.2 Prediction of protein targeting 111
3.4.3 Ancestral state estimation 111
3.4.4 Analysis of gene expression data 112
3.4.5 Functional term enrichment analysis and prediction of transmembrane helices 113
Chapter 4: Bioinformatic and molecular characterisation of a putative aspartate aminotransferase localised in the bundle sheath cells of maize 114
4.1 Introduction 114
4.1.1 Background 115
4.1.2 Aims and outline 126
4.2 Results 127
4.2.1 A bioinformatic search for candidate aminotransferases operating in maize bundle sheath cells identifies the gene *AILP1* 127
4.2.2 *ZmAILP1* belongs to a highly conserved plant specific gene family containing many paralogous genes 131

4.2.3 Investigation of *AILP1* gene expression across species shows upregulation in the photosynthetically active tissue of C_4 species 134

4.2.4 No amino-lyase enzyme activity could be detected for purified recombinant *Zm*AILP1 138

4.2.5 *Zm*AILP1 and *At*AILP1 localise to the cytosol in *Arabidopsis* protoplasts 144

4.2.6 The generation of loss-of-function *ailp1* mutants in Arabidopsis using CRISPR-Cas9 147

4.2.7 *ailp1* mutants display a growth rate and flowering time phenotype indistinguishable from wild type plants 151

4.2.8 Despite *AtAILP1* being highly expressed in the leaf, no perturbation in photosynthesis was found for *ailp1* plants 154

4.2.9 Leaves from *ailp1* plants have a higher nitrogen content than wild type plants 158

4.3 Discussion 160

4.4 Materials and methods 165

4.4.1 Phylogenetic analyses 165

4.4.2 Analysis of gene expression data 165

4.4.3 Strains, growth conditions and media 166

4.4.4 Plant material and growth conditions 166

4.4.5 Modification of protein sequences for cloning and expression 166

4.4.6 Generation of constructs for protein expression of *ZmAILP1* in *E. coli* 167

4.4.7 Expression of recombinant *ZmAILP1* in *E. coli* 169

4.4.8 Protein purification of recombinant *Zm*AILP1 170

4.4.9 Enzyme assay to assess the ability of AILP1 to liberate ammonia from amino acids 171

4.4.10 Generation of constructs for protoplast expression 172

4.4.11 *Arabidopsis thaliana* protoplast generation and transformation 173

4.4.12 Confocal analysis of transformed protoplasts 174
4.4.13 CRISPR-Cas9 vector construction 175
4.4.14 *Agrobacterium*-mediated *Arabidopsis* transformation and identification of *ailp1* mutants 177
4.4.15 Phenotypic characterisation of *Arabidopsis* CRISPR mutants 178

Chapter 5: General discussion 180

5.1 A first attempt to identify the changes in organellar protein targeting that occurred during the evolution of a diverse set of eukaryotic species 180

5.2 Two recent studies investigating chloroplast transit peptide gains corroborate several findings in this thesis and provide insight into the origin of novel transit peptides 181

5.3 Changes in protein targeting following gene duplication may arise through mechanisms of gene neofunctionalisation or subfunctionalisation 185

5.4 The ancestral state estimation method used is expected to underestimate the total number of changes in protein targeting that occurred during plant evolution 189

5.5 What is the functional significance of identified changes in protein targeting? 191

5.5.1 Investigation of the changes in protein targeting that occurred during the evolution of C_4 photosynthesis 192

5.6 More work is needed to identify the function of AILP1 194

5.7 It is still unclear what happens to aspartate in the bundle sheath cells of maize 196

Appendices 200

Appendix 2.1: Worked example of, and justification for, the 75% retention score used in the identification of changes in protein targeting 200

Appendix 2.2: Comparison of gene-tree–species-tree mapping approaches 205

Appendix 2.3: Functional terms (MapMan) enriched among proteins that underwent a change in organellar targeting during plant evolution..... 209

Appendix 2.4: Localisation terms (from the Plant Proteome Database) enriched among proteins that underwent a change in organellar targeting during plant evolution 213

Appendix 2.5: Functional terms (MapMan) enriched among proteins that did not undergo a change in protein targeting to A) the chloroplast or B) the mitochondrion during plant evolution. 215

Appendix 3.1: Breakdown of the number of gains and losses in protein targeting identified to each organelle along branches of the species tree associated with the evolution of C_4 photosynthesis 218

Appendix 3.2: Details about the transcriptomic data sets used in Chapter 3 and Chapter 4 of this thesis 219

Appendix 3.3: Gene trees and expression data supplementary to the main text in Chapter 3..... 220

Appendix 4.1: The set of species (and their phylogenetic relationship) used in the construction of the AILP1 orthogroup and gene tree 224

Appendix 4.2: Binding and cleavage sites of the sgRNAs used in the editing of the *Arabidopsis AILP1* gene 225

Appendix 4.3: Gel electrophoresis from screening T2 CRISPR edited plants ... 226

Appendix 4.4: Expression of different malic dehydrogenase proteins in maize 227

Appendix 4.5: Vector maps of the protein expression constructs used to express recombinant *Zm*AlLP1 in *E. coli* **228**

Appendix 4.6: Golden gate parts used for plant transformation work **230**

References ***235***

Abbreviations

2-PG	2-phosphoglycolate
3-PGA	3-phosphoglycerate
ADP	adenosine diphosphate
AILP1	aluminium-induced-like protein homolog 1
ATP	adenosine triphosphate
BGLU1	beta-glucosidase 1
BLAST	basic local alignment search tool
BS cells	bundle sheath cells
CA	carbonic anhydrase
Cas9	CRISPR associated protein 9
CRISPR	clustered regularly interspaced short palindromic repeats
cTP	chloroplast transit peptide
DNA	deoxyribonucleic acid
EC numbers	enzyme commission numbers
ER	endoplasmic reticulum
GFP	green fluorescent protein
GO	gene ontology
GWD	genome-wide duplication
HIS-tag	polyhistidine-tag
kDa	kilodalton
M cell	mesophyll cell
MBP	maltose binding protein
MDH	malate dehydrogenase
mRNA	messenger RNA
mTP	mitochondrial target peptide
mya	million years ago
NAD	Nicotinamide adenine dinucleotide
NAD-ME	NAD-dependent malic enzyme
NADP	nicotinamide adenine dinucleotide phosphate
NADP-ME	NADP-dependent malic enzyme

NCBI	National Center for Biotechnology Information
NLS	nuclear localisation signal
PCR	polymerase chain reaction
PEP	phosphoenolpyruvate
PEP-CK	PEP-carboxykinase
PEPC	PEP carboxylase
PPDK	pyruvate phosphate dikinase
PPR	pentatricopeptide repeat
PS II	photosystem II
PTS1	peroxisomal target signal 1
PTS2	peroxisomal target signal 2
RNA	ribonucleic acid
RNA-seq	RNA sequencing
rpm	revolution per minute
RuBisCO	Ribulose-1,5-Bisphosphate Carboxylase and Oxygenase
RuBP	Ribulose-1,5-bisphosphate
SDS-PAGE	sodium dodecyl sulphate polyacrylamide gel electrophoresis
sgRNA	short guide RNA
SP	signal peptide
TPM	transcripts per million
U	enzyme unit
WT	wild type

Chapter 1: General Introduction

1.1 Introduction

The cellular structure of eukaryotic cells is distinguished from that of bacteria and archaea by the presence of several membrane-bound subcellular compartments known as organelles. Indeed, the word 'eukaryote' comes from the Greek *eu* for "true" and *karyon* for "nut" or "kernel", the reference here being to cells that possess one particular organelle – the nucleus – which is the "true kernel". Bacteria and archaea on the other hand are *pro*karyotes, "before kernel". As well as the nucleus, eukaryotic cells contain several other distinct organelles that sequester different biochemical reactions of the cell. The efficiency of these reactions is facilitated by this compartmentalisation and also by being housed within the specialised structures exhibited by different organelles. Although some prokaryotes have been shown to possess some level of cellular compartmentalisation, for example the aggregation of enzymes into protein-bound metabolosomes (Yeates et al. 2011; Kerfeld and Erbilgin 2015) or the sequestration of DNA and ribosomes into nucleus-like structures (Lindsay et al. 2001; Chaikeeratisak et al. 2017), none have been shown to possess the rich tapestry of membrane-bound compartments present in eukaryotic cells.

All extant eukaryotes are thought to form a monophyletic group (Koonin 2010; Archibald 2015a; Koonin 2015; Roger et al. 2017). As such, the last eukaryotic common ancestor was most likely a cell that contained an already complex endomembrane system and the majority of eukaryotic-associated organelles. This includes the mitochondrion, an organelle formed from the endosymbiosis of an α–proteobacterium. Plants also form a monophyletic group and arose following another endosymbiosis event, this time of a cyanobacterium which became the chloroplast (Archibald 2015b). Each type of organelle in the eukaryotic cell contains a unique suite of proteins (the organellar proteome) which is necessary for organellar function. Although many of the functions of organelles are conserved between species, their independent evolution in different eukaryotic lineages

has resulted in the adaptation of organellar function in response to the changing life history of the host organism. A good example of this is the unique biochemical functions taken on by plant mitochondria in response to the acquisition of a second energy-producing organelle and the evolution of the plant cell (Mackenzie and McIntosh 1999). For example, in plants, mitochondria house several reactions of the photorespiratory pathway; a metabolic cycle that only occurs concomitantly with oxygenic photosynthesis (Bauwe et al. 2010). Such changes in mitochondrial biochemistry can only have arisen through changes in the protein content of plant mitochondria. Beyond such drastic evolutionary transitions however, it is less clear to what extent organellar proteomes have been modified during the evolution of different eukaryotic groups.

The work in this thesis aims to bridge this knowledge gap by estimating when changes in the proteomes of organelles have occurred during the evolution of land plants. Particular attention will be paid to the major energy organelles (the chloroplast, mitochondrion and peroxisome) and the evolution of one particular plant trait: C_4 photosynthesis.

1.2 The majority of organellar proteins are encoded in the nucleus and synthesised in the cytosol

In eukaryotic cells, almost all organellar proteins are encoded in the nucleus, synthesised in the cytosol and delivered to the correct organelle during or after cytosolic translation. The chloroplast and mitochondrion are the exception to this rule as both contain a genome derived from their free-living bacterial ancestors (Gray and Doolittle 1982; Gray and Lang 1998). However, these chloroplast and mitochondrion genomes are significantly reduced in size and complexity compared to those found in free-living cyanobacteria and α–proteobacteria today (Martin and Herrmann 1998; Gray et al. 1999; Martin et al. 2002). This reduction in chloroplast and mitochondrion gene content occurred during endosymbiosis and subsequent organellogenesis and was the result of gene loss (through

mutation and drift), and the transfer of a significant proportion of endosymbiont DNA to the nucleus of the host cell (Martin and Herrmann 1998; Timmis et al. 2004).

The products of some of the genes that were transferred to the nuclear genome provided their function to a new localisation in the cell (Martin and Herrmann 1998; Martin et al. 2002; Timmis et al. 2004). However, the products of many of these transferred genes were still required in the organelle from whence they came and consequently acquired the ability to be transported from the cytosol back into the organelle (as illustrated in Figure 1.1 for the endosymbiotic event that gave rise to the chloroplast). The continued loss and transfer of genetic material from the chloroplast and mitochondrion during evolution means that today these organelles typically encode fewer than 200 and 100 genes, respectively (Barbrook et al. 2010; Smith and Keeling 2015; Daniell et al. 2016). To illustrate this, the distribution of chloroplast genome size across different major groups of photosynthetic eukaryotes is shown in Figure 1.2 A. Given that the chloroplast and mitochondrial proteome in *Arabidopsis* is predicted to be in excess of 2000 proteins (Millar et al. 2001; Leister and Kleine 2008; Cui et al. 2011; van Wijk and Baginsky 2011), more than 90% of these proteins are encoded in the nucleus. Thus, chloroplast and mitochondrial proteomes retain much of the complexity of their bacterial ancestors despite a severe reduction in genome size (illustrated for the chloroplast in Figure 1.2 B). For the other organelles of the cell, all of their proteins are encoded in the nucleus. In plant cells this includes the nucleus, peroxisome, vacuole, and organelles of the secretory pathway such as the endoplasmic reticulum, Golgi, lysosome, and endosome. It should be noted that in this thesis the term ‘chloroplast’ is used to denote all types of plastid that derived from the primary endosymbiosis of a cyanobacterium.

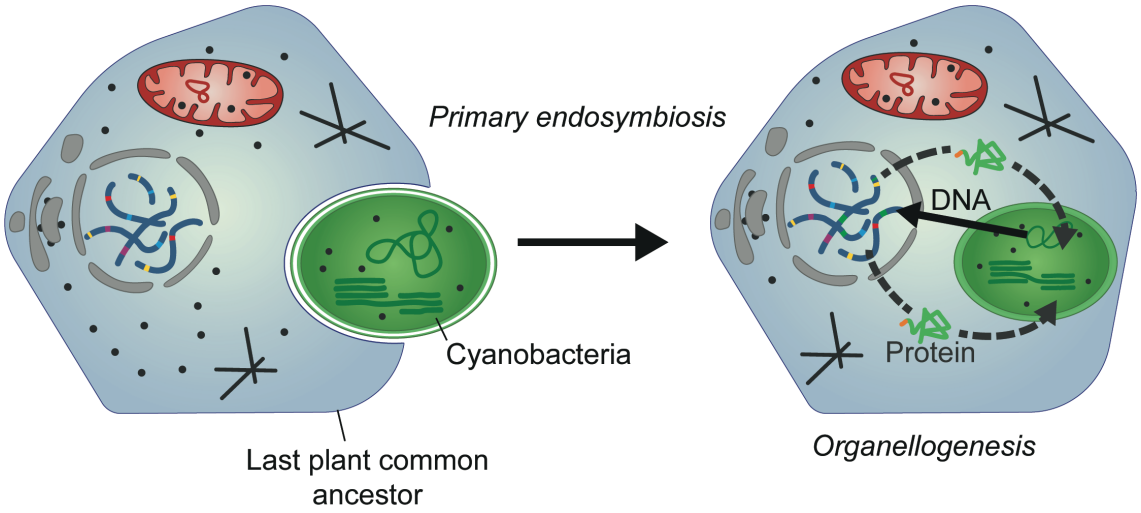


Figure 1.1. A simplified illustration of the primary endosymbiosis event involving a eukaryotic host cell and a cyanobacteria that led to the establishment of the chloroplast in the last common ancestor of all Archaeplastida (plants). During the initial endosymbiosis, a cyanobacterial cell (whose proteome was entirely encoded by its own cyanobacterial genome) was sequestered by a eukaryotic cell as illustrated by the image on the left-hand side of the arrow. Subsequently, the cyanobacterial cell was established as a new organelle in a process known as organellogenesis (illustrated on the right-hand side of the arrow) which involved the transfer of DNA from the cyanobacterial endosymbiont to the host cell nucleus and the evolution of mechanisms to allow many nuclear-encoded proteins of cyanobacterial and non-cyanobacterial origin to be targeted back to the nascent chloroplast. Image inspired by (Dacks and Field 2007).

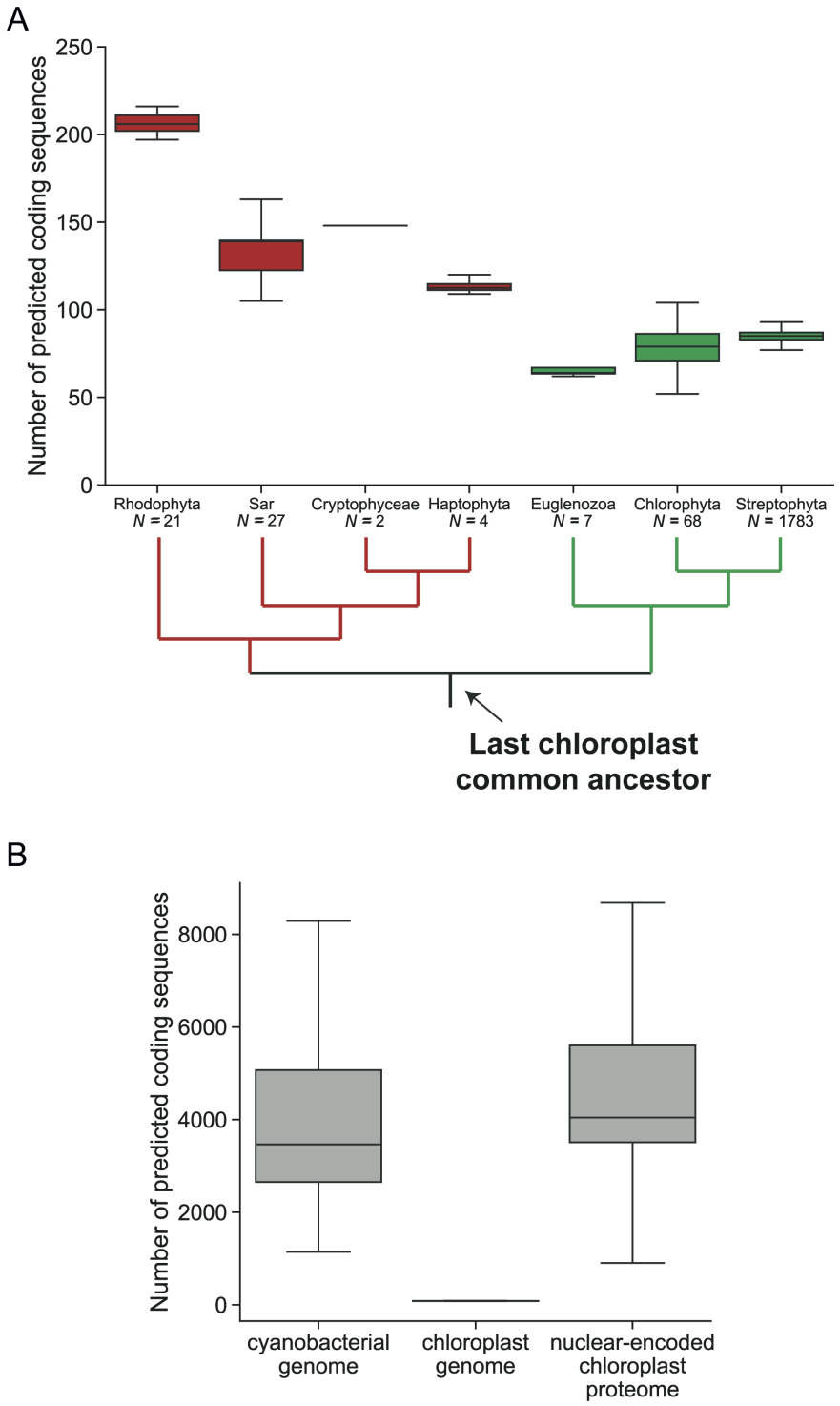


Figure 1.2. Comparison of cyanobacterial genome, chloroplast genome and chloroplast nuclear-encoded proteome size (number of predicted coding sequences). A) Distribution of chloroplast genome size across major groups of photosynthetic eukaryotes. The number of species sampled in each group is shown (*N*). The phylogenetic tree shows a proposed evolutionary relationship between the chloroplast-genomes of the groups as estimated by (Keeling 2013); note that branch

lengths are not drawn to scale and do not correspond to evolutionary distance. The red branches (left of the root bifurcation) show descendants from a red alga. Here, the chloroplasts in Sar, cryptophyceae, and haptophyta derive from an ancestral secondary endosymbiosis event involving a single-celled eukaryote and a red alga. The green branches (right of the root bifurcation) show descendants from a green alga. Here, the chloroplast in euglenozoa derives from a secondary endosymbiosis event involving a single-celled eukaryote and a green alga. **B)** Comparison of cyanobacterial genome size (N = 190) with the chloroplast genome and proteome size of the 42 plant species studied in Chapter 2 of this thesis. Chloroplast genome sizes were obtained from NCBI. Nuclear-encoded chloroplast proteome sizes were predicted by the methods in Chapter 2.

1.3 Target signals direct protein sorting in the cell

To enable cytosolic-synthesised organellar proteins to reach their intended destination they contain within their polypeptide sequences a short stretch of amino acids (a target signal) that facilitates protein translocation to a specific organelle. Broadly, there are five different types of organellar import target signals: the chloroplast transit peptide (cTP), mitochondrial target peptide (mTP), peroxisomal target signals 1 & 2 (PTS1 and PTS2), nuclear localisation signal (NLS), and signal peptide (SP). With the exception of the PTS1 and NLS, these target signals are located at the N-terminus of a nascent organellar protein (preprotein) and are cleaved upon entry into the organelle (Figure 1.3). The PTS1 and NLS on the other hand are located at the C-terminus and in the middle of organellar proteins, respectively, and are not cleaved upon organellar entry. In all cases though, these target signals function by interacting with specific sorting receptor proteins in the cytosol and, in some cases, mediating protein translocation across the organellar membrane (Schatz and Dobberstein 1996; Kunze and Berger 2015). The presence of an organellar target signal at the correct location within a protein sequence is typically sufficient to determine the intracellular localisation of that protein (Jarvis 2008).

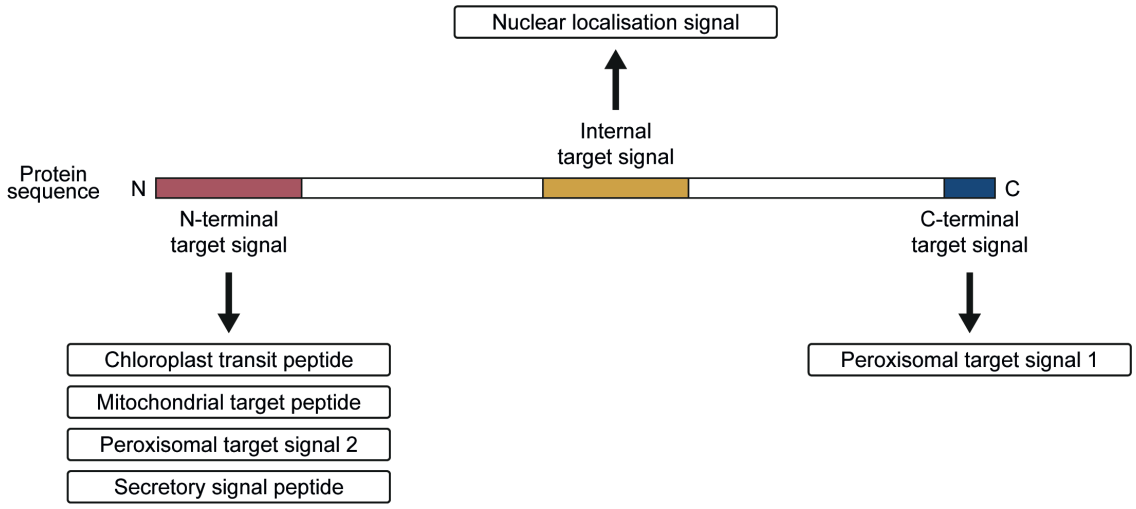


Figure 1.3. The different types of plant-cell organellar target signals and their typical location within an organellar preprotein.

The mechanisms of protein translocation are complex and unique to each type of organelle. For example, preproteins containing a chloroplast transit peptide are first recognised by two cytosolic chaperones that interact with the transit peptide and meditate preprotein transport to the chloroplast (Jarvis 2008). The identity of these cytosolic chaperones depends on the phosphorylation state of the transit peptide: phosphorylated transit peptides are recognised by Hsp70 and a 14-3-3 protein whereas unphosphorylated transit peptides are recognised by Hsp70 and Hsp90 (Li and Chiu 2010). Once at the chloroplast membrane the transit peptide binds sequentially to a series of TOC (translocon at the outer envelope membrane) and TIC (translocon at the inner envelope membrane) receptor proteins. These TOC and TIC proteins are embedded in the outer and inner chloroplast membrane, respectively, and guide the preprotein through a pore in the membrane (Li and Chiu 2010; Kunze and Berger 2015; Thomson et al. 2020). In *Arabidopsis* a minimum of three different TOC (Toc33, Toc159, Toc75) and five different TIC (Tic20, Tic21, Tic56, Tic100, Tic214) proteins make up the translocon pore complex. In the intermembrane space an additional TIC protein (Tic22) acts as a chaperone while another two TIC proteins (Tic40, Tic110) are embedded in the inner membrane with

stromal-facing appendages that facilitate preprotein import. Thus, in total a minimum of 13 different proteins are involved in the translocation of a chloroplast protein from the cytosol to the chloroplast stroma. In the stroma, additional chaperones support protein folding and hydrolyse ATP to drive protein translocation in a one-directional manner, while the chloroplast transit peptide is cleaved by a stromal processing peptidase (reviewed in Kunze and Berger 2015; Thomson et al. 2020).

For proteins targeted to the mitochondrion, a functionally analogous series of chaperones, inner membrane proteins, and outer membrane proteins to those in the chloroplast (dubbed TIM and TOM proteins in this case) are used for preprotein translocation to and across the mitochondrial membrane (Neupert and Herrmann 2007). Similarly, elaborate import pathways exist for the other organelles of the cell. For proteins that contain a secretory signal peptide, protein translocation to the endoplasmic reticulum (ER) may occur either during protein translation (co-translational protein import) or after protein translation (post-translational protein import). In the co-translational protein import pathway, a cytosolic GTP-hydrolysing ribonucleoprotein, known as the signal recognition particle (SRP), binds to the signal peptide of a nascent polypeptide as it emerges from the ribosome (Akopian et al. 2013). The SRP mediates the movement of the ribosome-nascent chain complex to the membrane of the ER where the SRP interacts with the GTPase domain of an SRP receptor (Akopian et al. 2013). The interaction with an SRP receptor releases the SRP from the protein's signal peptide and transfers the ribosome-nascent chain complex to the ER membrane pore protein Sec61. Here, the nascent chain is threaded into the Sec61 pore which is then plugged by the ribosome such that protein translation can continue whilst simultaneously transporting the newly synthesised polypeptide through the ER membrane (Akopian et al. 2013). In the case of post-translational ER protein import, the SRP does not bind with the preprotein signal peptide and instead cytosolic chaperones (Hsp70s and Hsp40s) facilitate preprotein transport to the ER (Ngosuwan et al. 2003). At the ER membrane, the preprotein enters Sec61 through

an interaction with the signal peptide and preprotein translocation is facilitated by a series of Hsp70 chaperones in the ER lumen which pull the preprotein through in an ATP-dependent manner (Kunze and Berger 2015). Once in the ER lumen, the signal peptide is cleaved, and the preprotein is folded with the help of further chaperones (Kunze and Berger 2015).

For proteins containing a peroxisomal target signal 1 or 2 (PTS1 or PTS2), protein translocation occurs after protein translation has finished and the protein has been folded in the cytosol (Léon et al. 2006). In the PTS1-mediated pathway, the C-terminal peroxisomal target signal interacts with the receptor protein Pex5 which mediates preprotein translocation to a docking complex in the peroxisomal membrane (made up of approximately eight additional Pex proteins) (Léon et al. 2006). The preprotein, bound to Pex5, moves through this docking complex and into the peroxisomal matrix. Once there, the preprotein is released by an unknown mechanism and the Pex5 receptor is exported back out of the peroxisome in an ATP-dependent manner to be used again in the cytosol. In the PTS2-meditated pathway, protein translocation proceeds in the same way except that the receptor that binds initially to the N-terminal peroxisomal target signal is different. This receptor is Pex7 and alone it cannot initiate protein translocation so it is bound by a second receptor protein, the identify of which varies between species but in plants is the PTS1-receptor Pex5 (Kunze and Berger 2015). Once translocated into the peroxisomal matrix, the PTS2 peptide is cleaved. Proteins containing a nuclear-localisation signal are also transported after protein translation in a pathway mediated by cytosolic receptors and a nuclear-pore complex (Lange et al. 2007). No further details of this pathway are given here because it is less relevant to the contents of this thesis. It should be noted that, once inside an organelle, proteins may possess additional target signals or sequence features that determine their ultimate sub-organellar localisation (Jarvis 2008).

1.4 Organellar proteomes are highly dynamic in time and space

Although many of the functions of different plant organelles are conserved across the plant kingdom, there is evidence of diversity in their form and function between species. For example, the chloroplast and mitochondrion are typically known as the sites of photosynthesis and respiration respectively however both conduct a number of other specialised functions. The chloroplast, for example, can function in sulphate assimilation, fatty acid and lipid synthesis, nitrogen assimilation, amino acid synthesis, and the synthesis of a number of secondary metabolites such as isoprenoids, heme, chlorophyll and carotenoids (Wise 2006; Jensen and Leister 2014). While some of these accessory functions are found in most or all species, others are found in a smaller number of plant lineages (Tielens et al. 2002; Gabaldón and Huynen 2004; Gabaldón and Huynen 2007; Kaur and Hu 2011; Gray 2015; Leister 2016). An extreme but not uncommon example of chloroplast functional diversity are the structurally reduced chloroplasts found in non-photosynthetic plants (e.g. parasites and mycoheterotrophs) and algae (Hadariová, Vesteg, Hampl, & Krajčovič, 2018). While no longer photosynthetic, having lost most or all of their chloroplast encoded photosynthesis genes, these structures are thought to still carry out other housekeeping functions of the chloroplast (Logacheva, Schelkunov, Shtratnikova, Matveeva, & Penin, 2016). For example, the malarial parasite *Plasmodium falciparum* contains a non-photosynthetic chloroplast (known as an apicoplast) derived from the secondary endosymbiosis of a red alga (Keeling 2013). In *Plasmodium falciparum*, and other apicomplexans, the apicoplast still houses vital biochemical pathways for fatty acid and isopentenyl biosynthesis (Ralph et al. 2004).

Although there are several examples of differences in organelle form and function in different eukaryotic lineages, it is not clear to what extent the proteomes of organelles are typically modulated during evolution. While experimental techniques for cataloguing organelle proteomes exist, robust cross-species comparisons of organelle proteomes are difficult due to the high levels of heterogeneity exhibited within individuals. Morphological

and metabolic variation between the same organelles is observed in different cell types and in response to different physiological conditions. For example, in plants that conduct two-celled C_4 photosynthesis, the chloroplasts found in the mesophyll and bundle sheath cells can have significantly different ultrastructure and biochemical functions concomitant with the role of each cell type in concentrating CO_2 around the enzyme RuBisCo (Kanai and Edwards 1999; Edwards et al. 2004; Takabayashi et al. 2005). This dimorphism in chloroplast form and function in C_4 plants has been shown to be the result of the differential accumulation of hundreds of proteins (Majeran et al. 2005). Indeed, such heterogeneity in organelle form and function has been reported across a wide range of plant cell types (reviewed for the chloroplast and mitochondrion in Bowsher and Tobin 2001). Similarly, organelle proteomes have been shown to respond dynamically in response to physiological stress. In *Arabidopsis,* for example, during periods of dehydration the proteomes of the chloroplast, mitochondrion, and peroxisome are all modulated to maintain their energy balance to protect against damage from reactive oxygen species (Gupta et al. 2016). Thus, in the face of significant spatiotemporal heterogeneity, experimental comparisons of organelle proteomes between species are usually limited to a single cell- or tissue-type and are further limited in their ability to detect proteins with low abundance. For some model species larger catalogues exist that collate the results from multiple proteomic experiments to provide a more comprehensive inventory of all *possible* organelle proteins. For example, The Plant Proteome Database (Sun et al. 2009) is a continually updated catalogue of *Arabidopsis* and maize proteins whose subcellular localisations have been experimentally characterised. While useful, the sparsity of species included in these catalogues represents a limitation for investigations into the evolutionary dynamics of organellar proteomes.

1.5 Changes in protein targeting as a mechanism of organellar proteome evolution

Multiple lines of evidence indicate that organellar proteomes are dynamic and changing throughout evolution in eukaryotic lineages. The mechanisms by which organellar proteomes are modulated during evolution is however an area of research that has received little attention. In particular, it is unknown what contribution changes in protein targeting have played in the lineage-specific divergence of organellar proteomes. As mentioned previously, target signals consist of short stretches of amino acids and so mutations in the gene region corresponding to a target signal may affect the efficiency with which a protein is sorted. Indeed, target signals typically lack a consensus sequence and are highly variable in length, and the specificity of a particular organellar target signal is instead defined by the presence or absence of particular structural and chemical features (summarised in Table 1.1 and references therein). A consequence of this is that relatively few amino acid substitutions may be needed to alter the targeting specificity of a particular sequence.

Table 1.1. Features of the canonical target signals for protein translocation to the chloroplast, mitochondrion, peroxisome, secretory pathway, and nucleus. The protein state during translocation (as described in section 1.3) is also detailed.

Target signal	Consensus sequence	Location in protein sequence	Approximate length (amino acids)	Structure	Additional general features	Cleaved upon entry	Protein state during translocation	Co- or post-translational protein import?
Chloroplast transit peptide[a,b]	No	N-terminus	13-146	Amphiphilic α-helix	Abundance of positively charged residues	Yes	Unfolded	Post
Mitochondrial target peptide[c]	No1	N-terminus	13-85	Amphiphilic α-helix	High content of basic and hydroxylated residues	Yes	Unfolded	Post
Peroxisomal target signal 1[d]	Tripeptide: [SA][RK][LM] (w/o AKM and incl. SRI and PRL)	C-terminus	3	Tripeptide	-	No	Folded	Post
Peroxisomal target signal 2[d]	R[LI]X5HL	N-terminus	9	Amphiphilic α-helix	-	Yes	Folded	Post
Signal peptide[e,f]	No	N-terminus	15-45	Hydrophobic α-helix	Three regions with distinct properties; a hydrophobic (H) domain bounded by an N- and C- terminal domain	Yes	Unfolded	Co or post
Nuclear localisation signal[g,h]	No	Not specific	7-16	Intrinsically disordered	Either a monopartite or bipartite motif with particular physicochemical properties	No	Folded	Post

[a](Bruce 2000), [b](Kunze and Berger 2015), [c](Sjöling and Glaser 1998), [d](Reumann 2004), [e](Käll et al. 2004), [f](Bendtsen et al. 2004), [g](Yamagishi et al. 2015), [h](Cokol et al. 2000), [i](Xu et al. 2012)

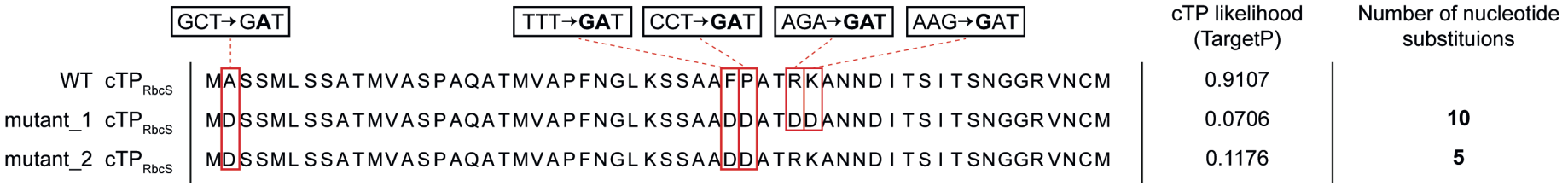


Figure 1.4. As few as five nucleotide substitutions may be sufficient to disrupt chloroplast transit peptide (cTP) functionality. The amino acid sequence of the wild-type cTP of the *Arabidopsis* RuBisCo small subunit (RbcS1A, AT1G67090) is shown (WT cTP_{RbcS}). Two mutant versions of this sequence (mutant_1 and mutant_2) were generated by replacing amino acid sites known to be important for cTP functionality with the acidic residue aspartate. The likelihood (between 0 and 1) that each sequence functions as a cTP is predicted by TargetP v2 (Armenteros et al. 2019). The nucleotide substitutions needed to generate the mutant amino acid sequences are highlighted in bold in the boxes above the sequence and the total number of nucleotide changes is given in the right-most column. The codons shown are those most frequently used in *Arabidopsis* genomic DNA (Athey et al. 2017).

To demonstrate this, two mutant sequences were generated for the chloroplast transit peptide (cTP) of the *Arabidopsis* RuBisCo small subunit protein RbcS1A (AT1G67090). RbcS1A is a nuclear-encoded protein that is targeted to the chloroplast and has a well-characterised cTP. In particular, five amino acid residues in the cTP of RbcS1A (Leu2, Phe33, Pro34, Arg37, and Lys38) have been experimentally shown to be essential for efficient protein targeting: deletion of some or all of these residues led to a decrease and loss of chloroplast targeting efficiency, respectively (Lee et al. 2006). To investigate the number of nucleotide substitutions that might thus be needed to disrupt cTP function, a mutant sequence was generated by substituting each of the five amino acid residues that have been shown to be important for RbcS1A cTP function with aspartate, an amino acid that is typically low in abundance in cTPs (Shen et al. 2017). While the native RbcS1A cTP sequence was shown to be highly likely to contain a cTP by *in silico* prediction (WT, Figure 1.4), the mutant sequence (mutant_1, Figure 1.4) had a very low cTP likelihood. A second mutant sequence containing just three out of the five amino acid substitutions was also found to have a low likelihood of containing a functional cTP (mutant_2, Figure 1.4). Thus, although these were not tested *in vivo*, as few as three amino acid substitutions (corresponding to five nucleotide substitutions) may be sufficient to change the organellar targeting of a protein.

In another example, it has been shown that the loss of a single amino acid residue is sufficient for the loss of mitochondrial targeting and the gain of peroxisomal targeting (Oatey et al. 1996). The protein in this case was alanine-glyoxylate aminotransferase 1 (AGT) which has been shown to be differentially localised to the mitochondrion or peroxisome in different species depending on the diet of the species. The ancestral AGT is predicted to have contained an N-terminal mitochondrial target peptide (mTP) and a C-terminal peroxisomal target signal 1 (PTS1). Despite having both target signals, this ancestral protein was only targeted to the mitochondria due to the hierarchy with which target signals are recognised by cytosolic receptors (as reviewed in Kunze and Berger

2015). The evolution of a peroxisomal AGT in some lineages arose due to the deletion of a translation initiation start codon resulting in a truncated version of the protein only containing the PTS1 (Oatey et al. 1996). Thus, changes in protein targeting may arise through stepwise modifications to the region containing a target signal or through mechanisms that result in the gain or loss of whole target signal regions. While there are several individual examples of protein families whose members are differentially targeted within the cell (for example Gabaldón and Pittis 2015), to date there has been no large-scale assessment as to the frequency with which changes in organellar targeting of proteins occur in eukaryotic lineages or the impact it has had on organelle evolution.

1.6 Computational prediction of target signals

In order to study the gain and loss of organellar protein targeting through evolution it is necessary to identify which proteins contain organellar target signals. As discussed above, few species have had their organelle proteomes experimentally characterised. In any case, experimental proteomics cannot distinguish between proteins that are absent due to not containing a target signal and proteins that are absent due to gene loss or a downregulation of gene expression under the conditions that have been tested (i.e., developmental stage, stress/infection, senescence, tissue type). Thus, while proteomics can provide experimental validation of protein localisation, to study the evolution of organellar protein targeting it is necessary to use computational prediction of organellar target signals.

1.6.1 A brief history of target signal prediction

The first target signal to be experimentally described was the signal peptide (Blobel and Dobberstein 1975a; Blobel and Dobberstein 1975b) and it was not long after its discovery that the first signal peptide prediction method was developed (McGeoch 1985). Based on the limited set of signal peptide sequences known at the time, this first prediction method used two sequence features (max hydrophobicity and length of a hydrophobic region) to

determine the presence or absence of a signal peptide in the N-terminal region of a protein (McGeoch 1985). This method was later incorporated into the first multi-site prediction tool PSORT which expanded the knowledge base of target signal features to a hierarchical decision tree containing 80 nodes (Nakai and Kanehisa 1992). Protein localisation was determined by following a series of “if-then” decisions in the tree, the topology of which was determined by the known or predicted order in which target signals would be processed by cytosolic receptors in the cell (Nakai and Kanehisa 1992). Despite such large databases of target signal features, and a biologically intuitive decision tree, this method struggled to incorporate enough feature information to reliably discriminate between different types of target signals (reviewed in Nielsen et al. 2019). The method was also cumbersome to update as new target signal features were discovered. The PSORT method was however later updated into the multi-compartment prediction methods iPSORT (Bannai et al. 2002) and WoLF PSORT (Horton et al. 2007). These two methods still took a feature-based approach but did not use the detailed hierarchical decision tree that was used in PSORT. Amino acid features such as hydrophobicity, isoelectric point, and net charge were used, and a classification decision was made by either sequentially testing for the features of each type of target signal in a binary yes/no decision (iPSORT, Bannai et al. 2002), or by converting sequence features into numerical vectors and then using a weighted k-nearest neighbour classifier (WoLF PSORT, Horton et al. 2007). In both cases, extensive feature databases had to be curated which had the benefit of allowing a user to extract information about how a classification decision is made but with the drawback that determination of relevant features depends on prior knowledge and thus limits the sequence feature space that can be used. Thus, the majority of predictors developed since use machine learning algorithms which can be trained with databases of experimentally characterised organelle proteins (reviewed in Nielsen et al. 2019).

One of the earliest and most popular of these machine learning algorithms is the neural-network based tool TargetP (Emanuelsson et al. 2000) which can predict the presence of N-terminal chloroplast transit peptides, mitochondrial target peptides, and signal peptides. To date, a plethora of machine learning based methods for target signal prediction havebeen developed. These prediction methods may consider protein targeting to one organelle, multiple organelles, and in some cases to suborganellar structures. As TargetP, iPSORT, and WolfPSORT are used in the analyses conducted in this thesis, their self-reported performance statistics are shown in Table 1.2. It should be noted though that the performance of these tools varies between experimental data sets and that benchmarking target signal prediction methods is notoriously difficult – and few independent attempts have been made. Typically, target signal prediction methods are only compared as part of the publication of a novel method. This may favour the novel method as it will have been trained and tested on more recently updated data sets such that the improvement is due to availability of data rather than the underlying method. Similarly, if during the development of a method, proper homology reduction is not carried out on the testing/training data set, then the classifier may be tested on some sequences that are near-identical to sequences it has been trained on. A full review of the challenges in comparing target signal prediction methods is provided in a recent review paper (Imai and Nakai 2020). In this thesis, the prediction methods in Table 1.2 were used as they predict multiple types of target signal for plant proteins, perform relatively well for each target signal type in the studies that have compared them, and are able to be used on large whole genome sized data sets.

Table 1.2. Self-reported performance statistics for TargetP v1 and v2 (Emanuelsson et al. 2000; Armenteros et al. 2019), iPSORT (Bannai et al. 2002) and WolfPSORT (Horton et al. 2007) on plant proteins. No confusion matrix was provided in the WolfPSORT paper for plant proteins and instead a score of approximately 70% for specificity and sensitivity across target signals was reported for plant proteins. The MCC scores are highlighted as they provide an overall measure of classification quality. **Abbreviations**: Matthews correlation coefficient (MCC), chloroplast transit peptide (cTP), mitochondrial target peptide (mTP), signal peptide (SP).

Algorithm	cTP recall	cTP specificity	cTP MCC	mTP recall	mTP specificity	mTP MCC	SP recall	SP specificity	SP MCC
TargetP v1	0.85	0.69	0.72	0.82	0.86	0.72	0.91	0.96	0.90
TargetP v2	0.86	0.99	0.88	0.85	1.00	0.86	0.98	0.96	0.97
iPSORT	0.68	0.71	0.64	0.84	0.86	0.75	0.91	0.98	0.92
WolfPSORT	0.70	0.70	-	0.70	0.70	-	0.70	0.70	-

As well as mis-prediction, computational methods are also limited in their ability to identify non-canonically targeted proteins. For example, signal-anchored and tail-anchored outer-envelope proteins of the chloroplast and mitochondrion lack cleavable-target signals (Lee et al. 2014). Instead, these proteins possess a single N-terminus or C-terminus transmembrane domain, respectively, which directs protein translocation. It is not clear to what extent organellar proteins are targeted by non-canonical means but the few non-canonical pathways that have been described to date appear to involve particular classes of protein. For example, the outer membrane proteins of the chloroplast and mitochondrion described above, or glycoproteins proteins that are necessarily transported through the ER to their final organellar destination. Thus, while it is currently not possible to accurately identify all organellar proteins by computational prediction alone, it is expected that it is sufficient to detect the majority of canonical target signals and thus could be used to investigate changes in canonical protein targeting during evolution.

1.7 Molecular phylogenies can be used to reconstruct ancestral protein function

The evolutionary history of a given protein in a group of species can be interrogated through its molecular phylogeny. Doing so can provide information about the presence or absence of the gene encoding the protein in different species, changes in the rate of molecular sequence evolution, and the identity of orthologs and paralogs. In addition, molecular phylogenies can be used to reconstruct ancestral protein function. The most common application of this is ancestral sequence reconstruction. In such analyses, ancestral sequences are typically inferred using maximum likelihood methods from the phylogenetic relationship of extant proteins. Subsequently, the proteins of the ancestral sequences can be synthesised, and their function investigated. This method has been used to study, for example, the evolution of protein folding kinetics (Lim et al. 2016) and protein ligand-specificity (Siddiq et al. 2017). In one study, ancestral sequence reconstruction was used to investigate the catalytic properties of an alcohol dehydrogenase (ADH4) during the evolution of primates (Carrigan et al. 2015). In humans ADH4 is involved in ethanol metabolism however in some other primates ADH4 is unable to efficiently metabolise ethanol. Reconstruction of ancestral ADH4 enzymes revealed that a single mutation occurred in our ancestors approximately 10 million years ago that enabled ADH4 to efficiently oxidise ethanol and which was likely an adaptation to an increase in foraging of fermented fruits on the forest floor (Carrigan et al. 2015).

The application of ancestral sequence reconstruction could be similarly used to determine when proteins evolved targeting signals. However, although target signals have common features, the sequences themselves are poorly conserved and there is evidence that they evolve at over twice the rate of mature peptide sequences (Williams et al. 2000). It is unknown therefore if reliable multiple sequence alignments could be generated for target signals over deep evolutionary time. An alternative to ancestral sequence reconstruction is to predict whether target signals are present or absent in extant proteins and in doing

so utilise ancestral character estimation to infer the ancestral targeting of any given protein at any point along its evolution. In this way it would be possible to evaluate on a genome-wide scale how the predicted proteomes of organelles have changed during the evolution of different eukaryotic groups. In this thesis I will use such an approach to investigate organellar proteome evolution during the evolution of a diverse set of land plant species.

1.8 C_4 photosynthesis is a plant trait split between several subcellular compartments

In plants, oxygenic photosynthesis takes place in the chloroplast. Here, quanta of electromagnetic energy (photons) from the sun are used to split water into oxygen, hydrogen ions, and electrons. Photoexcitation of the electrons triggers an electron transport chain that harvests the light energy to produce NADPH and ATP. These reducing and energy equivalents (NADPH and ATP, respectively) are then used by the plant to assimilate carbon dioxide (CO_2) into carbohydrates through the Calvin-Benson cycle. In most plants, CO_2 assimilation occurs in a single cell type of the leaf, the mesophyll cell, and the first product of CO_2 fixation in the Calvin-Benson cycle is the three-carbon compound 3-phosphoglycerate (3-PGA); hence this type of photosynthesis is known as C_3 photosynthesis (Figure 1.5 A). The enzyme that catalyses the carboxylation of CO_2 into 3-PGA is ribulose-1,5-bisphosphate carboxylase-oxygenase (RuBisCo). Although C_3 photosynthesis is widespread in the plant kingdom, CO_2 capture is inefficient in the C_3 pathway because oxygen (O_2) competes with CO_2 for the active site of RuBisCo. The products of the reaction of RuBisCo with O_2 are toxic and must be metabolised by the cell in an energy-costly process known as photorespiration (Bauwe et al. 2010). To overcome the inefficiency of photorespiration, some plants have evolved mechanisms to elevate the concentration of CO_2 around RuBisCo. One such mechanism is C_4 photosynthesis. In plants that carry out C_4 photosynthesis, RuBisCo is not present in the mesophyll cells and is instead restricted to another cell type in the leaf: the bundle sheath cell (Figure 1.5 B) (Hattersley et al. 1977). The concentric arrangement of mesophyll and bundle sheath cells

in the C_4 leaf means CO_2 and O_2 from the air initially diffuses into the mesophyll cells (Sedelnikova et al. 2018). In the C_4 pathway, CO_2 in the mesophyll cells is fixed into a four-carbon compound by the enzyme phosphoenolpyruvate carboxylase (PEPC) which has no affinity for O_2 (Figure 1.5 B). Through a series of biochemical reactions, this four-carbon compound moves into the bundle sheath cells and into the vicinity of RuBisCo where it is decarboxylated to release CO_2, now at high concentrations around RuBisCo, thus limiting photorespiration. A three-carbon compound is also released from this decarboxylation reaction and diffuses back to the mesophyll cells to re-start the cycle (Figure 1.5 B).

As well as being split between two cell types, the biochemical reactions of the C_4 cycle are also split between the cytosol, chloroplast, and mitochondrion within the mesophyll and bundle sheath cells. Thus, the correct intracellular and intercellular localisation of C_4 cycle enzymes and transporters (collectively C_4 proteins) is essential to maintain flux through the C_4 pathway. Given that C_4 photosynthesis has evolved independently over 60 times in plants (Sage 2016), and that many C_4 proteins do not have a photosynthetic function in C_3 plants, there has been a substantial body of research investigating the molecular mechanisms by which proteins are recruited from a C_3 to a C_4 function. For example, the functionalisation of C_4 proteins from a C_3 function has been associated with gene duplication (Emms et al. 2016), changes in gene regulation (Brown et al. 2011), and positive selection within the coding region of genes (Niklaus and Kelly 2019). Given the compartmentalisation of the C_4 cycle, it is also possible that changes in protein targeting were important in the recruitment of proteins during the evolution of C_4 photosynthesis. Indeed, this has been shown to be the case for at least one C_4 enzyme in two independent C_4 lineages (Tanz et al. 2009; Clayton et al. 2017), however to date no genome-wide screen has been conducted to investigate whether other changes in organellar protein targeting were important for the evolution of C_4 photosynthesis.

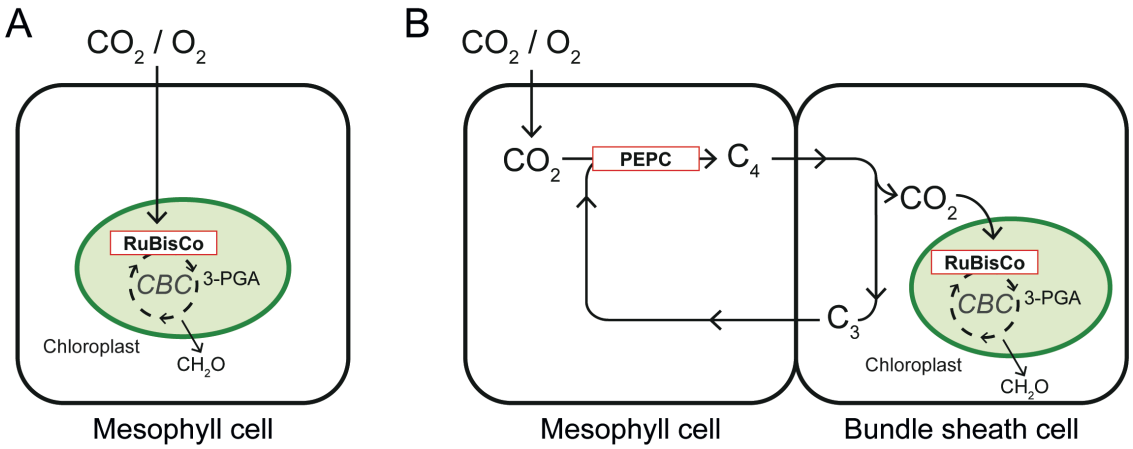


Figure 1.5. A simplified schematic showing photosynthetic carbon assimilation in a C_3 (A) and a C_4 (B) leaf. Abbreviations: carbon dioxide (CO_2), oxygen (O_2), ribulose-1,5-bisphosphate carboxylase-oxygenase (RuBisCo), phosphoenolpyruvate carboxylase (PEPC), 3-phosphoglycerate (3-PGA), representative carbohydrate (CH_2O), four-carbon compound (C_4), three-carbon compound (C_3), Calvin-Benson cycle (CBC).

1.9 Thesis aims and overview

The overarching aim of my thesis was to investigate how changes in protein targeting have shaped organellar proteomes during the evolution of plants and the effect of such changes on a particular plant trait: C_4 photosynthesis.

To achieve this aim, I firstly developed a novel phylogenetic approach to predict when changes in protein targeting occur during evolution. Chapter 2 describes this approach and its application to a large data set of protein families built from a diverse set of plant genomes. The research in this chapter presents a first attempt to assess the extent to which changes in organellar protein targeting have occurred during the evolution of a major group of eukaryotes. The central findings from this chapter have been published in the journal Molecular Biology and Evolution (Costello et al. 2020). In Chapter 3 I use this same approach to interrogate those changes in protein targeting that occurred during the evolution of C_4 photosynthesis in four C_4 grass species. The work in this chapter involved the interrogation of C_4 gene expression and highlighted a potential inconsistency in our current model of C_4 photosynthesis in the species *Zea mays* (maize). Specifically, it is not clear how aspartate (a four-carbon metabolite through which CO_2 is shuttled in the C_4 cycle) is metabolised in the bundle sheath cells of maize. The final results chapter of this thesis (Chapter 4) therefore attempts to address this gap in our understanding of C_4 biochemistry in maize through the characterisation of a novel enzyme identified as potentially carrying out the missing aspartate aminotransferase reaction. The bioinformatic and experimental work conducted in pursuit of this aim is described but unfortunately, I was not able to experimentally resolve the putative enzymatic function of the candidate gene. Although tangential to the main aim of my thesis, this chapter incorporates several of the comparative genomic approaches used in Chapter 2 and Chapter 3 to assess the putative function of the candidate gene. The chapters represent distinct bodies of work and, as such, each contains its own introduction, methods, results, and discussion sections.

Chapter 2: Changes in organellar protein targeting have occurred at a continuous rate throughout plant evolution and the pace of change is accelerated by gene duplication

2.1 Introduction

Organelle biogenesis and function is dependent on the concerted action of both organellar-encoded (if present) and nuclear-encoded proteins. As described in the previous chapter, differences between homologous organelles across the plant kingdom are expected to arise, in part, as a result of differences in the cohort of nuclear-encoded proteins that are targeted to them. However, neither the rate at which differences in protein targeting accumulate nor the evolutionary consequences of these changes are known. The research in this chapter aims to address these unknowns through the development of a phylogenetic method that uses ancestral state estimation to predict when changes in organellar protein targeting have occurred during the evolution of land plants. This analysis presents the first attempt to quantify the extent to which changes in organellar targeting of proteins have occurred during the evolution of a major group of eukaryotes. The results in this chapter have been published in a paper titled "*Gene duplication accelerates the pace of protein gain and loss from plant organelles*" (Costello et al. 2020).

2.1.1 Background

A defining feature of eukaryotic cells is the compartmentalisation of intracellular processes into specialised membrane-bound organelles. Plant cells contain several such organelles, including the nucleus, chloroplast, mitochondrion, peroxisome, Golgi, endoplasmic reticulum, and vacuole. With the exception of the chloroplast and mitochondrion, all organellar proteins are encoded in the nucleus of the cell and must be imported from the cytosol via import channels on the organellar membrane. For both the chloroplast and mitochondrion, a fraction of their respective proteomes are encoded by their own

organellar genomes; however, the vast majority of chloroplast and mitochondrial proteins are encoded in the nucleus (Timmis et al. 2004).

Nuclear-encoded organellar proteins are translocated to and across organellar membranes by means of targeting peptides (hereafter target signals) located within the amino acid sequence of the protein. Although these target signals come in a variety of forms, they are usually located at the N-terminus of the polypeptide chain and cleaved upon entry into the organelle (reviewed in Kunze and Berger 2015). As such, the target signal sequences, once removed, have no impact on the final function of the mature protein. In addition, there is substantial flexibility in the identity and length of target signals (Bannai et al. 2002) such that a large diversity of sequences can function to target proteins to their intended destination.

Early on in the investigation of the protein content of organelles it was noted that many homologous proteins had divergent subcellular localisations, both within and between species. For example, the cytosolic and mitochondrial isoforms of phosphoenolpyruvate carboxykinase proteins in animals (Nordlie and Lardy 1963) or the cytosolic and chloroplastic isoforms of sugar phosphate enzymes in plants (Schnarrenberger et al. 1983). Following the advent of protein, cDNA, and genome sequence data, it was realised that disparate organellar localisation of these proteins was facilitated by differences in the presence of N-terminal target signals, and that such differences occur among many homologous proteins in different species (Marques et al. 2008; Qian and Zhang 2009; Wang et al. 2009; Liu et al. 2014; Ren et al. 2014). Furthermore, larger scale bioinformatic analyses of plant gene families have suggested that changes in protein targeting of homologous proteins may be a common occurrence during plant evolution (Heilmann et al. 2004; Richly and Leister 2004; Gabaldón and Pittis 2015). For example, the computational prediction of organellar targeted proteins in the *Arabidopsis thaliana* genome (using the union of several prediction tools) led to the estimation that 239 *Arabidopsis* protein families contain members that are differentially localised within the cell

(Heilmann et al. 2004). Without taking a phylogenetic approach however, these studies are limited in their ability to estimate when during evolution these changes in protein targeting arise, or the frequency with which target signals are gained and lost. Thus, it is also unknown what impact changes in protein targeting have had on the lineage-specific evolution of organellar proteomes. Lastly, the extent to which the occurrence of changes in organellar targeting is influenced by evolutionary events of the nuclear genome, such as gene and whole-genome duplication, remains an open question (Byun-McKay and Geeta 2007; Marques et al. 2008; McKay et al. 2009; Byun and Singh 2013).

In this chapter I provide answers to these questions. To do this I developed a phylogenomic approach, coupled with ancestral state estimation, to interrogate the evolution of nuclear-encoded organellar proteomes across the land plant phylogeny. This uncovered a pattern of continual change in protein targeting, with ~3.6 changes per million years to the nuclear-encoded proteomes of both the chloroplast and mitochondrion. Functional analysis of the genes encoding these proteins revealed that these changes occurred predominantly to genes with regulatory rather than metabolic functions, indicating that altered regulatory capacity is a major theme of lineage-specific organellar proteome evolution in plants. I further demonstrate that changes in organellar targeting of proteins is associated with an increase in the rate of molecular sequence evolution of the gene encoding the re-targeted protein. Finally, I demonstrate that changes in protein targeting occur at a higher rate following gene duplication, providing evidence that gene and genome duplication are key facilitators of plant organelle evolution.

2.1.2 Author Contributions

Together with my supervisor I conceived the study. David Emms, who is a co-author on the published study, provided the data set of orthogroups and gene trees, as well as advice throughout the project. I conducted all of the analyses presented in this chapter.

2.2 Results

2.2.1 Predicted organellar proteome size correlates with total nuclear proteome size

A bioinformatic approach was taken to identify the cohort of nuclear-encoded organellar targeted proteins of 42 diverse plant species. The organelles included here were the chloroplast, mitochondrion, organelles of the secretory pathway, and the peroxisome. On average across land plants, the predicted chloroplast, mitochondrion, secretory, and peroxisome proteomes comprised 14% (±2%), 14% (±3%), 17% (±2%), and 0.32% (±0.05%) of the total nuclear-encoded proteome, respectively (Figure 2.1). The size of each predicted organellar proteome was found to corelate tightly with total nuclear-encoded proteome size (Figure 2.1). Here, the secretory proteome was considered to comprise of all proteins with a signal peptide. However, it should be noted that the secretory pathway itself is made up of multiple organelles including the endoplasmic reticulum, Golgi apparatus, lysosome and endosome, and the final destination of proteins harbouring a signal peptide may be any of these organelles, the cell wall or the extracellular environment.

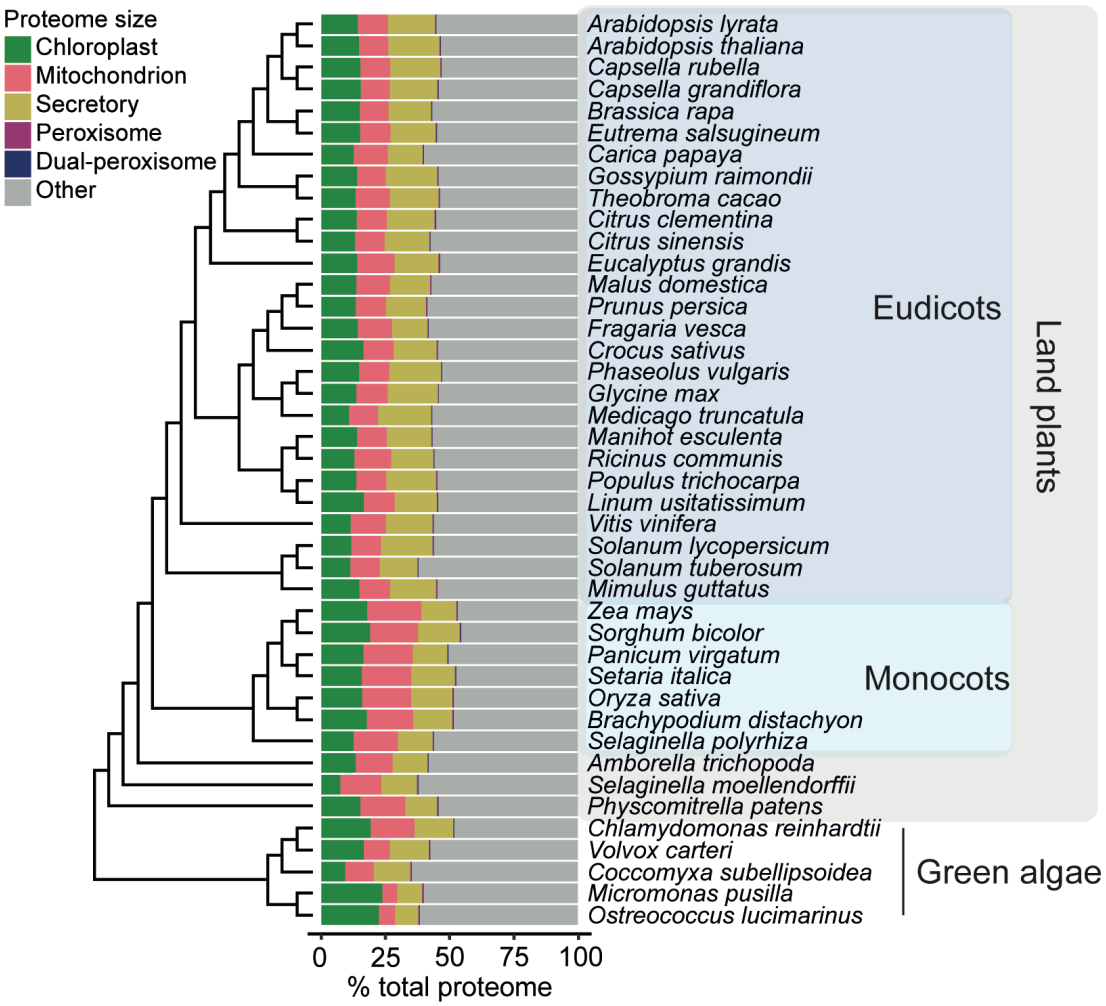


Figure 2.1. Predicted proteome size of different subcellular organelles for the 42 plant species in this data set, shown as a percentage of total proteome size. Chloroplast, mitochondrion, and secretory pathway proteins were predicted using TargetP v1.1 (Emanuelsson et al. 2000) and PredAlgo (Tardif et al. 2012). Peroxisomal proteins were identified by a manual search for the sequence motif of a peroxisomal targeting signal 1 or 2 (PTS1 or PTS2). Proteins predicted to be organellar targeted by TargetP but which also contain a PTS1 or PTS2 are indicated here as dual-localised peroxisomal proteins (N = 2,937). Predicted organellar proteome size correlated with total nuclear proteome size: chloroplast (R^2 = 0.91, P < 0.001), mitochondrion (R^2 = 0.86, P < 0.001), secretory (R^2 = 0.88, P < 0.001), peroxisome (R^2 = 0.94, P < 0.001).

2.2.2 Widespread gain and loss of organellar target signals have occurred throughout the evolution of plants

To identify the changes in nuclear-encoded protein targeting (and therefore organellar proteome content) that have occurred during the evolution of the 42 selected plant species, the predicted targeting of proteins was combined with a complete set of species-tree-reconciled gene trees (N = 18,823) for all orthogroups (gene families) of this data set. Maximum likelihood ancestral state estimation was then performed to predict the organellar targeting of the ancestral proteins represented by each internal branch of each reconciled gene tree. Changes in protein targeting identified by this method were subsequently filtered to only include those changes with high support (i.e., conservation) among extant proteins. This filter required that in order for a change in protein targeting to be considered real, ≥75% of extant proteins descendent from a branch on which a change occurred must display the new organellar targeting state, while ≥75% of proteins descendant from the sister branch must display the ancestral organellar targeting state. A retention score of 75% (rather than 100%) allows for some error in the protein targeting prediction algorithms, and for instances of further change in protein targeting within a gene tree. A worked example showing the use of the retention score in the identification of a change in protein targeting, as well as further justification for the use of a 75% cut-off is provided in Appendix 2.1. All subsequent analyses were conducted with this filtered set of identified changes in protein targeting.

Changes in protein targeting identified along gene-tree branches were then mapped to the corresponding branch of the species tree to infer the number of changes in protein targeting that occurred to each organelle along each branch of the species tree. Two possible approaches can be taken to map gene-tree branches (and thus changes in protein targeting) to branches of the species tree. The first uses a most recent common ancestor approach. Here, a change in targeting on a gene-tree branch is mapped to the species-tree branch that corresponds to the most recent common ancestor of the species

encapsulated by the genes subtending the change. This approach is shown in Figure 2.2 A. The disadvantage of this approach is that in some cases the absence of a gene from a particular clade (either through gene loss or incomplete genome annotation) can lead to uncertainty in the mapping of gene-tree branches to the species tree (this scenario is shown in Figure 2.2 B). The second approach, that takes this uncertainty into account, is proportional mapping. Here, a change in targeting identified on a gene-tree branch that cannot be mapped with certainty is split equally across the branches of the species tree on which it could have occurred (rather than just the most recent branch). To investigate the effect of the mapping approach on the number of identified changes in protein targeting, both approaches were used to map changes to the species tree and the resulting distribution of changes across the species tree was compared. Overall, it was found that the choice of mapping approach made little difference to the global picture of gain and loss in protein targeting across the species tree. Indeed, it was found that there was a tight correlation in the number of changes identified by each approach across branches of the species tree (Appendix 2.2). Thus, for the purposes of the work in this chapter, the results from the most recent common ancestor approach formed the basis of all subsequent analysis. Indeed, this approach is perhaps more appropriate given that many missing genes represent true gene loss and so mapping a change in the ancestor of that gene is not biologically relevant given that the gene was subsequently lost.

Using the most recent common ancestor approach, all identified changes in protein targeting (that passed the 75% filtration step) were mapped to the species tree. In total, across the four organelles, 6,162 gains and 9,058 losses were identified and mapped to internal branches of the species tree (Figure 2.3). Gains and losses in protein targeting were observed along every branch of the species tree, suggesting that changes in organellar protein targeting has been a widespread phenomenon during plant evolution.

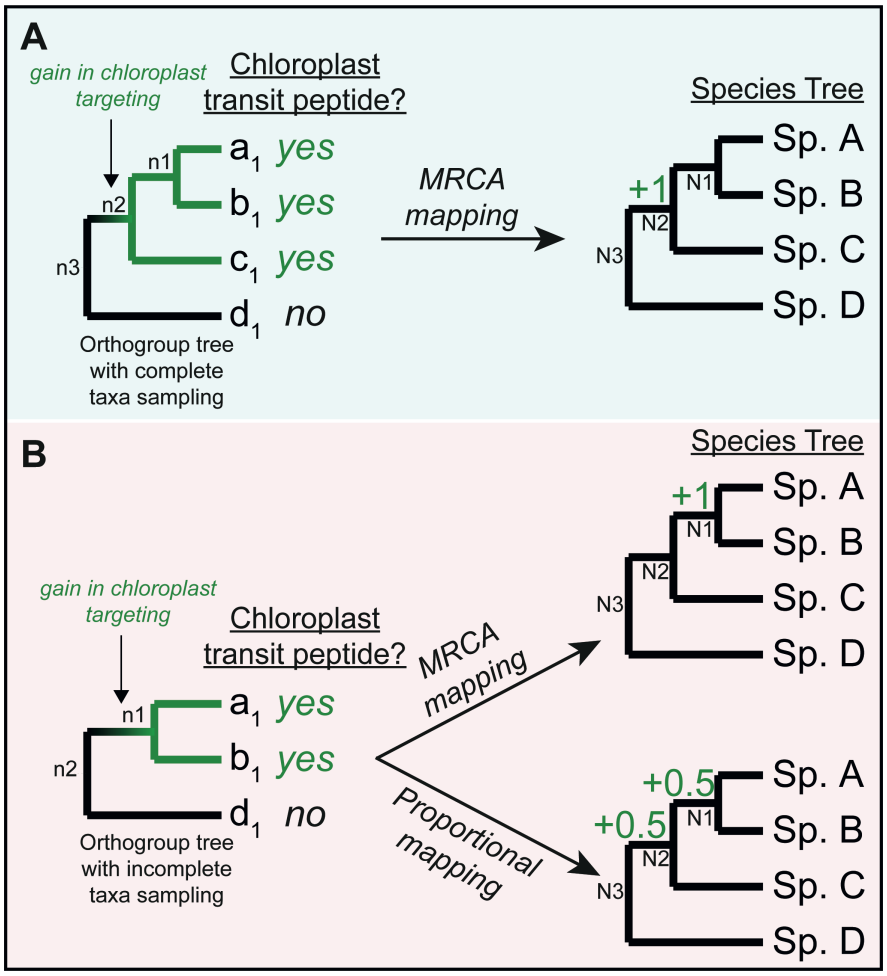


Figure 2.2. The effects of different gene-tree branch to species-tree branch mapping approaches when genes are missing from a particular species. A) Gene-tree branches can be unambiguously mapped to the species tree using a most recent common ancestor (MRCA) mapping approach when there is 100% representation of species within a gene clade. In this example, ancestral state estimation identifies a change in chloroplast targeting on branch n2 of the gene tree which represents the common ancestor of protein a_1, b_1 and c_1. Given this, MRCA mapping would add one (+1) to the tally of changes in chloroplast targeting that occurred in the ancestor of species A, B and C (branch N2 of the species tree). **B)** When a gene is missing such that there is no representation in a clade from one species, it is ambiguous as to which species-tree branch a gene-tree branch should be mapped to. In this gene tree, the gene c1 is not present. Using MRCA mapping, the change in targeting would be added to the tally for the ancestor of species A and B (branch N1 of the species tree). However, without gene c1 it is impossible to know if the change occurred in the ancestor of a1 and b1, or in the ancestor of a1, b1 and c1. A proportional mapping approach would thus map the change evenly to both N1 and N2 of the species tree (+0.5).

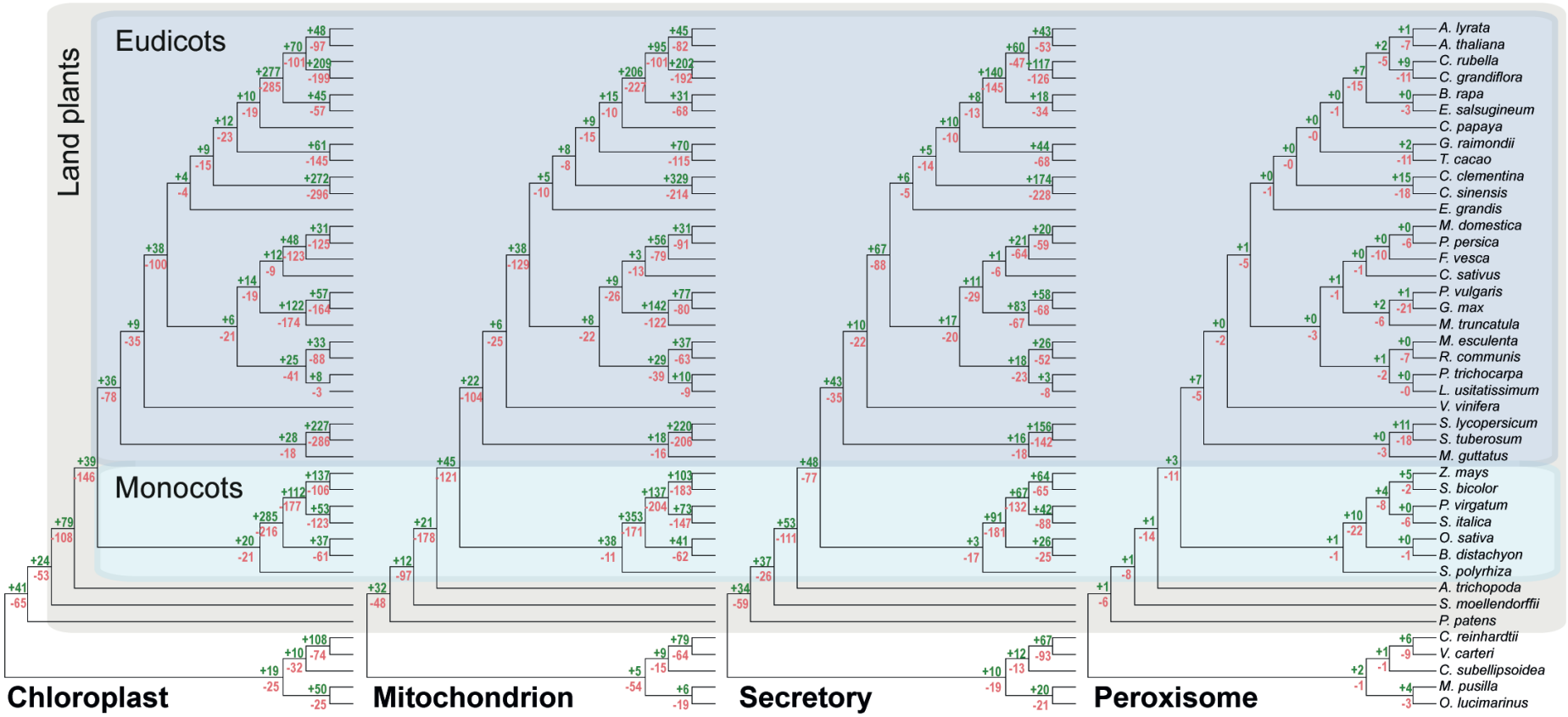


Figure 2.3. The number of gains (green and indicated with a +) and losses (pink and indicated with a -) in organellar protein targeting to the chloroplast, mitochondrion, secretory pathway, and peroxisome identified to have occurred along each non-terminal branch of the species tree encapsulated by the species in this study. Branch lengths shown in the figure are not drawn to scale and do not correspond to evolutionary distances.

2.2.3 Changes to the organellar proteome have occurred continuously throughout plant evolution

To determine whether the rate of change in organellar proteome evolution has been continuous throughout plant evolution, or whether it has proceeded in bursts, the number of changes in protein targeting along each branch of the species tree was compared with the length of the branch, that is, the amount of molecular sequence evolution between species. This comparison revealed that there was a positive linear correlation between the amount of molecular sequence evolution and the number of changes in protein targeting to the chloroplast, mitochondrion, and secretory pathway (Figure 2.4 A-D). Using a time-calibrated species phylogeny, it was possible to estimate that on average 3.5 (1.3 gains and 2.2 losses), 3.6 (1.2 gains and 2.4 losses), 2.4 (0.9 gains and 1.5 losses), and 0.22 (0.05 gains and 0.17 losses) changes in protein targeting to the chloroplast, mitochondrion, secretory pathway, and peroxisome occurred for every million years of land plant evolution, respectively (Figure 2.5 A&B). Thus, organellar protein content has diversified during plant evolution in proportion to molecular evolutionary distance.

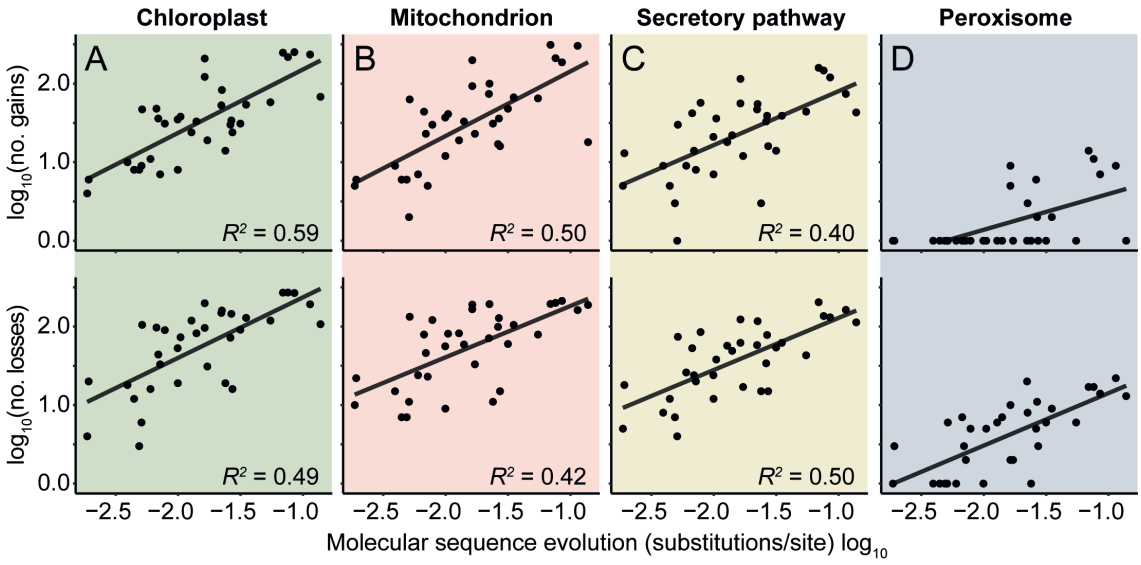


Figure 2.4. The relationship between molecular sequence evolution and organellar proteome evolution. There was a positive relationship (R^2 shown, $P < 0.001$) between species-tree branch length (amino acid substitutions per site) and the number of gains or losses to the **A)** chloroplast, **B)** mitochondrion, **C)** the secretory pathway. **D**) Fewer gains and losses were observed in peroxisomal targeting, with some branches being associated with no peroxisomal changes. The data are shown but no statistical conclusions drawn.

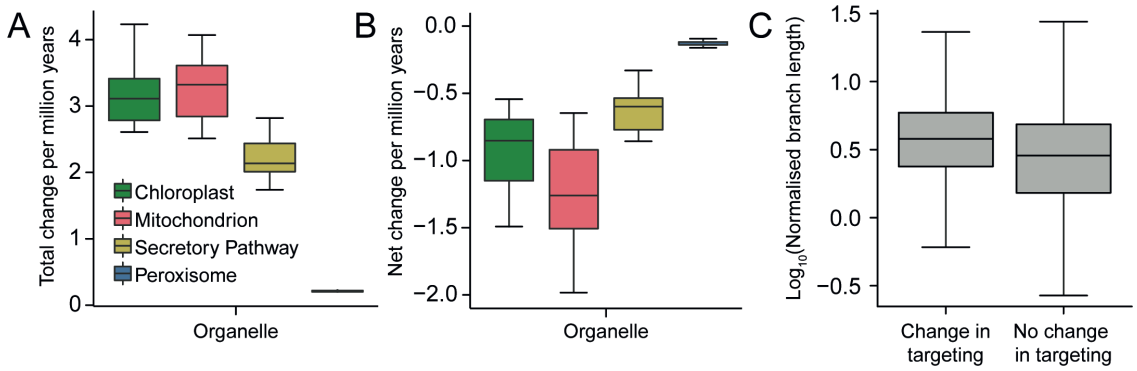


Figure 2.5. The number of changes in organellar targeting per million years for each organelle considered and the evolutionary rate of proteins undergoing a change in organellar targeting. A) Branches (*N* = 10) in the species tree for which divergence time estimates are available were used to produce a time-calibrated phylogeny. The number of changes in organellar targeting was then summed from the origin of the land plants (taken as 450 million years ago) to each of the selected branches and the total number of changes (gains + losses) per million years calculated. **B)** Net change was calculated in the same way except the number of losses was subtracted from the number of gains for each branch. **C)** Gene-tree branch lengths were normalised by their corresponding branch in the species tree. The normalised branch lengths of gene-tree branches associated with (*N* = 6,957) and without (*N* = 264,037) a change in organellar targeting were then compared. On average, branches associated with a change in targeting had a higher normalised branch length (two-tailed t-test, *P* < 0.001).

2.2.4 Changes in organellar protein targeting are associated with an increased rate of molecular sequence evolution

To determine whether changes in organellar targeting of ancestral proteins are associated with an increase in the rate of molecular sequence evolution, the length of the branches in gene trees on which a change in protein targeting occurred was compared with the length of branches on which a change did not occur. To render branch lengths comparable between and across gene trees, each branch in each gene tree was normalised by the length of the corresponding branch in the species tree which was inferred from concatenated single-copy genes (see materials and methods, section 2.4.3). This

normalises gene-tree branches relative to the amount of evolution that occurred during the same period of time for a typical single-copy gene. This comparison revealed that the lengths of branches in gene trees on which a change in organellar targeting occurred are longer than those branches on which a change did not occur (Figure 2.5 C). This difference was not due to a difference in the number or phylogenetic distribution of these branches, as the same difference was also observed if the number and phylogenetic distribution of sampled branches were kept constant between the two groups ($P < 0.001$, Monte Carlo resampling). Thus, changes in organellar protein targeting are associated with an increase in gene evolutionary rate.

2.2.5 Changes in organellar protein targeting occur more frequently following gene duplication

Given that evolution of changes in protein targeting may be prohibited by the detrimental effects caused by the loss of protein function in a native location, it was hypothesised that such changes might be more likely to be retained if they occurred to recently duplicated genes. Furthermore, it has been previously suggested that changes in protein targeting following gene duplication may be an important mechanism of duplicate gene neofunctionalisation (Byun-McKay and Geeta 2007; Marques et al. 2008; Byun and Singh 2013). If these prior hypotheses are correct, it would be expected that changes in protein targeting would occur more frequently following gene duplication events in this data set. To test whether this phenomenon occurred, the association between gene duplication and changes in organellar targeting of proteins was investigated. Across angiosperms, a robust set of 19,353 gene duplication events were identified and the frequency with which changes in protein targeting occurred on either of the two direct descendant child branches of each of these gene duplication events was analysed. This revealed that there was a change in organellar targeting along one of the two immediate child branches for 5.5% of gene duplication events. This frequency was significantly higher than that observed for branches that did not contain a gene duplication event in the same gene trees (2.2%,

hypergeometric test, $P < 0.01$). This phenomenon is observed whether the data set is analysed as a whole or whether gains and losses to individual organelles are analysed separately (Figure 2.6 A). To account for any potential biases in the distribution of gene duplication events in the species phylogeny, an analogous analysis was conducted where the number and phylogenetic position of the non-duplicated branches were randomly sampled so as to be identical to that of the identified gene duplication events (see materials and methods, section 2.4.9). This revealed the same result, whereby there was a higher frequency of change in organellar targeting following gene duplication events (Figure 2.6 B). Thus, overall, the frequency of evolving (and retaining) a change in organellar targeting is higher following gene duplication.

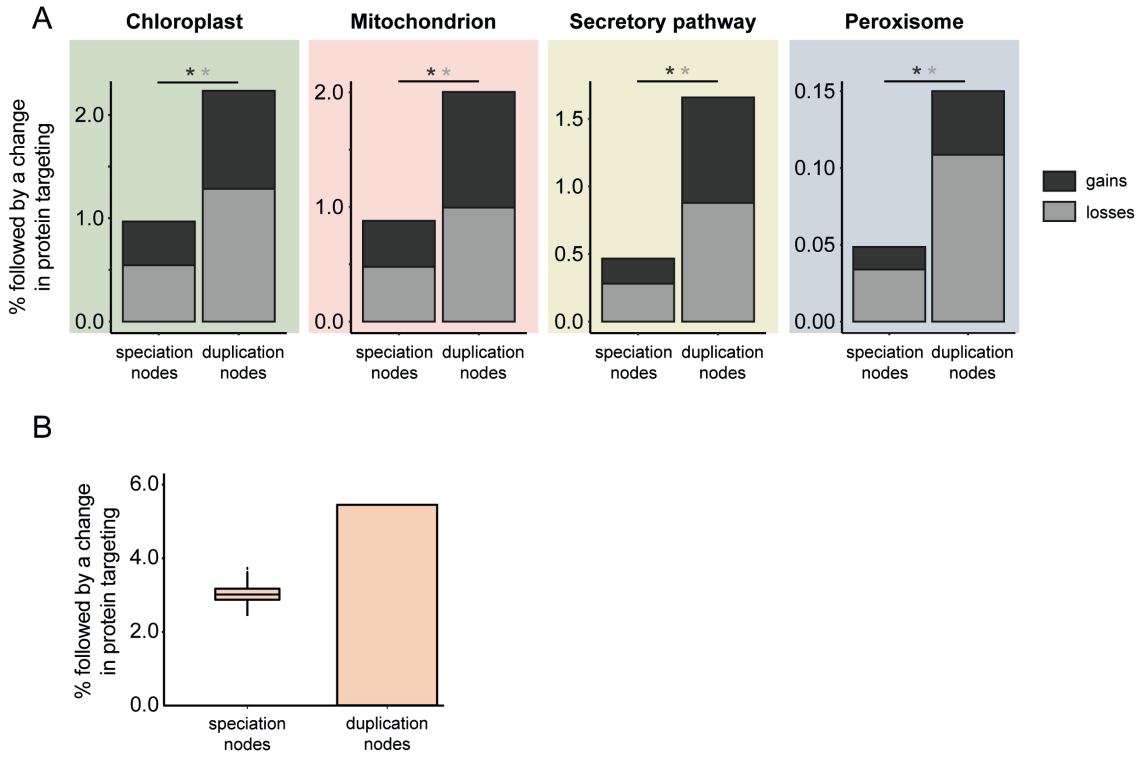


Figure 2.6. Changes in protein targeting occur at a higher rate following gene duplication. A) Comparison of the percentage of changes in protein targeting to the chloroplast, mitochondrion, secretory pathway, and peroxisome following nonduplication (speciation) and gene duplication events. The * signifies a significant difference (hypergeometric test, $P < 0.01$). **B)** The percentage of all changes in protein targeting following gene speciation and duplication events. As there were fewer gene duplication events ($N = 19,352$) than speciation events ($N = 218,410$) across gene trees, a resampling approach was taken to compare a set of speciation branches with the same phylogenetic placement as the gene duplication branches. For each branch in the species tree, gene duplications were analysed, and an identical number of speciation branches mapped to the same branch were sampled at random and analysed. This sampling was repeated 1,000 times to obtain a Monte Carlo resampling *P*-value. The distribution of the percentage of speciation branches followed by a change in protein targeting from these 1,000 samples is shown by the boxplot on the left. The percentage of duplication branches followed by a change in protein targeting is shown by the bar on the right. The proportion of changes in protein targeting was found to be significantly higher following gene duplication (Monte Carlo resampling, $N = 1,000$, $P < 0.001$).

2.2.6 There is no difference in the frequency of changes in organellar protein targeting following single-gene or whole-genome duplications

Gene duplications can arise from single-gene processes (such as tandem duplication), or from whole-genome duplication and triplication (polyploid) events. To investigate whether there was an effect of duplication type on the likelihood of a protein gaining or losing an organellar target signal following gene duplication, all duplicate genes were categorised into one of two sets: set 1 comprised the cohort of gene duplications that originated on branches in the species tree on which a whole-genome duplication (or triplication) event is thought to have occurred; while set 2 comprised the cohort of gene duplicates that originate on branches for which there is no associated polyploidisation event, and thus are assumed to have arisen from single-gene duplication processes. Figure 2.7 shows the species-tree branches associated with a polyploid event. Comparison of these two sets revealed that there was no difference in the likelihood of a change in protein targeting following gene duplication from either single-gene or whole-genome duplication events (Table 2.1, hypergeometric test, $P < 0.05$). The proportion of duplicate proteins from whole-genome or single-gene duplication events that were followed by a change in organellar targeting was 5.59% and 5.45%, respectively (Table 2.1).

Identification of branches in the species tree that are associated with whole-genome duplication allowed the identification of an additional independent set of “fossil gene duplicates”. These are genes that must have existed in a duplicated state for a period of time before returning to single-copy status prior to the subsequent speciation event and thus one duplicate from the pair no longer exists in any extant species. The existence of these fossil duplicates is a corollary of the fact that a whole-genome duplication must have duplicated all of the genes, or else it was a partial genome duplication. Interestingly, this group of fossil gene duplicates also exhibited the same high rate of change in organellar targeting as duplicated genes: the proportion of fossil duplicates followed by a change in targeting was 5.66% (Table 2.1). This suggests that the rate of change in protein targeting

is also elevated for duplicated genes where one duplicate is subsequently lost from the genome.

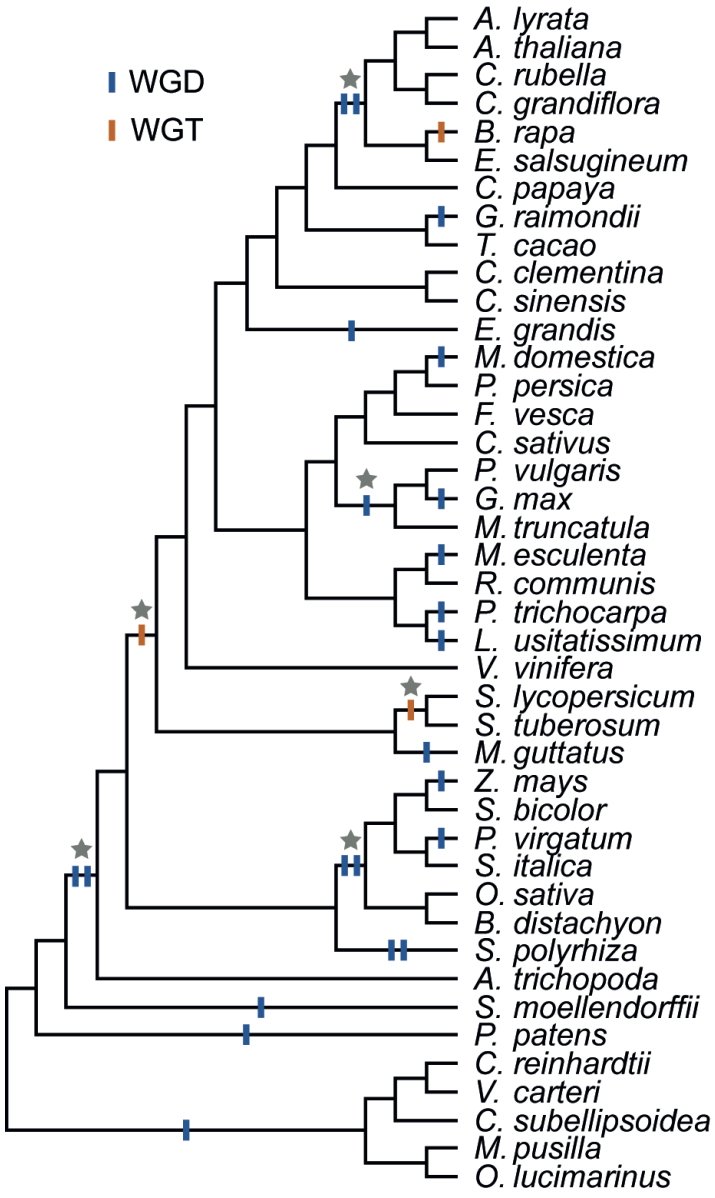


Figure 2.7. The species tree showing the branches associated with a previously reported whole-genome duplication (WGD) and triplication (WGT) event. Polyploid events estimated according to (Jiao et al. 2011; Vanneste et al. 2014; Ren et al. 2018). The non-terminal branches encompassing a polyploid event and used in this study are indicated with stars.

Table 2.1. The proportion of duplication and speciation events immediately followed by a change in organellar protein targeting. Gene-tree branches can represent either gene duplication or non-duplication (speciation) events. As whole-genome duplications (polyploidisation) events involve the duplication of every gene in the genome, gene speciation events mapped to branches of the species tree associated with a polyploid event can be considered unretained duplications. A significantly higher proportion of gene duplications are followed by a change in protein targeting compared to gene speciation events, and this is true for gene duplications arising from polyploidisation and non-polyploidisation (single-gene duplication) events (hypergeometric test, $P < 0.05$). This increase in the proportion of changes following gene duplication was also observed among unretained duplications from the polyploid events. In the table, polyploid and non-polyploid branches refer to gene-tree branches mapped to a polyploid or a non-polyploid branch of the species tree, respectively.

	Polyploid branches		**Non-polyploid branches**	
	Retained duplication	Unretained duplication	Retained duplication	Speciation
A. Occurrences followed by a change in targeting	710	1,271	362	3,451
B. Total occurrences	12,706	22,450	6,647	195,960
Proportion (%) followed by a change in targeting (i.e., A / B)	5.59	5.66	5.45	1.76

2.2.7 Changes in the regulatory machinery are the main changes occurring to the chloroplast and mitochondrial proteomes

To shed light on the functional significance of the identified changes in organellar protein targeting, a functional term enrichment analysis was conducted on the set of proteins that gained or lost a target signal during plant evolution. For this, the functional ontology terms provided by GoMapMan (Ramšak et al. 2014) were used. GoMapMan includes functional terms for plant-specific biological processes, and the functional terms within it belong to a hierarchical tree with around 30 top-level categories such as ‘Development’, ‘Hormone metabolism’ and ‘Stress’. GoMapMan terms were assigned to orthogroups based on the GoMapMan term assigned to the genes of the model organisms (*Arabidopsis* and maize) found within them. For each gene, only one GoMapMan term is assigned, however the

gene will also assume each of the hierarchical terms to which that tern belongs. For example, a gene (and therefore orthogroup) assigned the term 'RNA.regulation of transcription.MYB domain transcription factor family', will also assume the terms 'RNA.regulation of transcription' and 'RNA' (the levels here are demarcated by the full stop '.'). This permits functional enrichment to be tested for both specific and broad biological functions.

For both the chloroplast and mitochondrion, the set of proteins that changed organellar targeting during evolution was found to be enriched for functional terms concerning regulation, both at the transcriptional and posttranscriptional level (Figure 2.8). There was also an enrichment for terms concerning transport, hormone metabolism, secondary metabolism, stress, and development, with few terms related to energy metabolism (see Appendix 2.3 for the full list of enriched terms). Figure 2.8 shows the top 15 most highly enriched terms for proteins gained and lost to the chloroplast and mitochondrion. In both cases the top-level term 'RNA' was one of the most highly enriched terms. Many genes with diverse molecular functions are included in the 'RNA' category, including those which function directly in RNA processing (e.g., ribonucleases and RNA helicases), and those which function to regulate transcription (e.g., transcription factors). Closer inspection of the full list of enriched terms revealed that within the RNA category it is mainly terms related to the regulation of transcription that are enriched among proteins gained and lost to the chloroplast and mitochondrion (see Appendix 2.3). In support of this observation for the chloroplast (overrepresentation of transcriptional regulatory terms among proteins gained and lost), there was also an overrepresentation of proteins that localise to the chloroplast nucleoid, with no statistical overrepresentation of proteins that localise to other chloroplast subcompartments such as the thylakoid, envelope, or stroma. Here, localisation terms were assigned to protein families using the *Arabidopsis* annotations in the Plant Proteome Database (Sun et al. 2009) (full list of terms in Appendix 2.4). These localisation terms similarly form a hierarchical tree structure (for example 'Plastid

Thylakoid.peripheral.stromal side' and thus each hierarchical term was assigned to an orthogroup in the same way as the GoMapMan terms. Together, this analysis suggests that changes to the regulatory landscape of organelles have been the major consequence of changes in protein targeting during the evolution of chloroplasts and mitochondria in land plants.

Consistent with the lack of genetic material, functional terms associated with transcriptional regulatory processes were not observed for proteins gained and lost to the secretory pathway or peroxisome (Appendix 2.3). Instead, among proteins which changed targeting to the secretory pathway there was enrichment for functional terms associated with protein posttranslational modification, signalling, and the cell wall, whereas among proteins which changed targeting to the peroxisome there was enrichment for functional terms associated with metabolism (amino acid, lipid, secondary) and gluconeogenesis (Appendix 2.3).

It has previously been suggested that proteins with core organellar functions are more likely to be sensitive to dosage imbalance following gene duplication. Evidence for this comes from the fact that gene families that tend to maintain single-copy status throughout plant evolution are enriched for core chloroplast and mitochondrion functions, as well as other housekeeping functions (De Smet et al. 2013; Li et al. 2016; Tasdighian et al. 2017). Concomitant with this, it is expected that proteins which are sensitive to dosage imbalance will also be unlikely to undergo changes in protein targeting. To investigate this, a functional term enrichment analysis was conducted for those protein families with no evidence of changes in organellar targeting during their evolution. For protein families with no history of changes in chloroplast and mitochondrion targeting, there was an enrichment for GoMapMan terms related to photosynthesis and mitochondrial electron transport, respectively (Appendix 2.5). There was also an enrichment for functional terms relating to DNA synthesis and chromatin structure suggesting that proteins with core organellar and cellular functions are also less likely to undergo changes in organellar targeting.

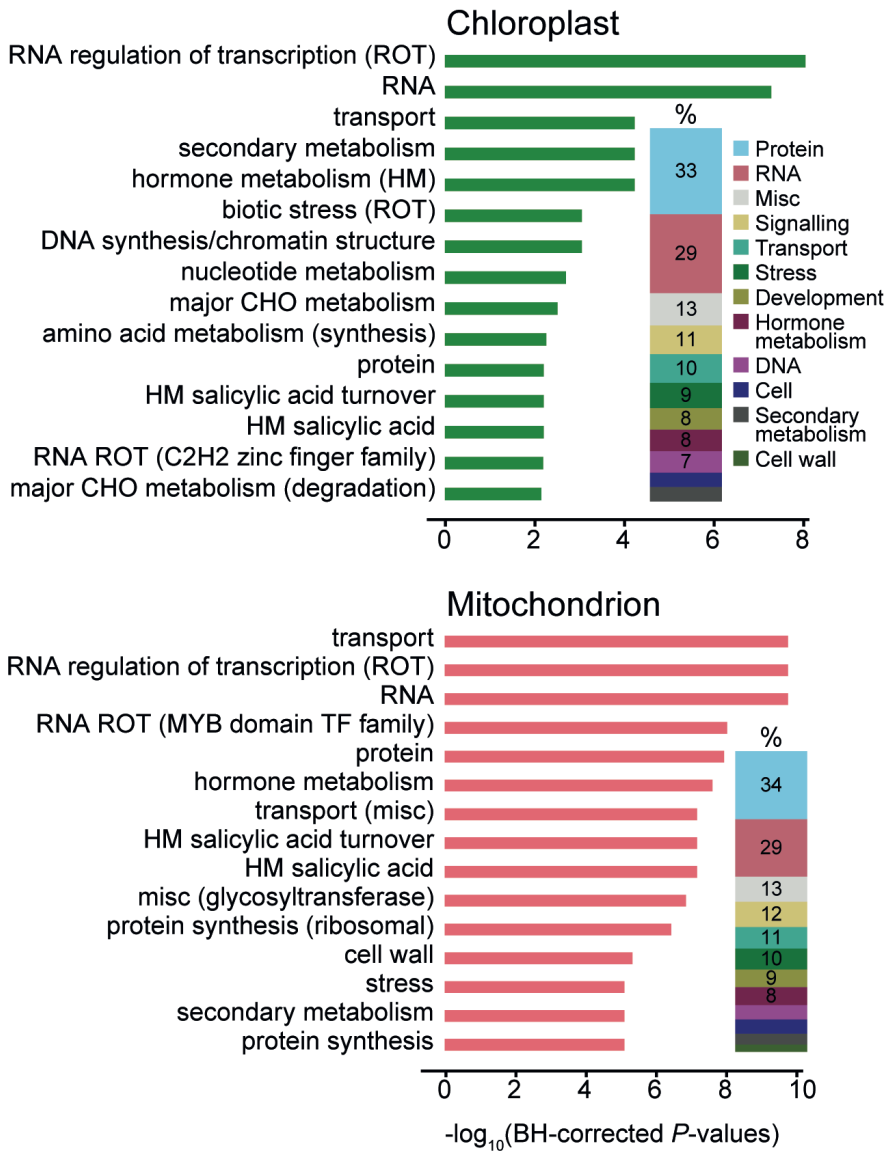


Figure 2.8. Enriched functional terms (GoMapMan) for the set of proteins that gained or lost a chloroplast or mitochondrial target signal during the evolution of the 42 plant species in this data set. Proteins were assigned GoMapMan terms based on the terms assigned to the orthogroup in which they belong. As GoMapMan terms are organised hierarchically, proteins assigned a term also assume the higher-level terms to which that particular term belongs. The horizontal bar graph shows the 15 most highly enriched terms (smallest *P*-values) for proteins that underwent a change in targeting to the chloroplast and mitochondrion. The full list can be found in Appendix 2.3. The horizontal stacked bar plot indicates the proportional representation of top-level MapMan categories assigned to proteins which gained or lost a chloroplast or mitochondrial target signal during evolution. Note, only the top 12 (out of 35) most highly represented top-level terms are shown here.

2.2.8 Using an alternative gene-tree inference method replicates the main findings presented in this chapter

It is expected that all gene trees contain tree inference error, otherwise a correct species tree could be accurately inferred from a gene tree of any single copy gene. The presence of this inference error has been repeatedly shown across all domains of life and in an exemplar analysis of 1,030 gene trees of single copy orthologs from 23 fungi species, all 1,030 gene trees were topologically distinct from each other and from the species tree (Salichos and Rokas 2013). Thus, gene-tree inference error is abundant and will negatively impact on any analysis that relies upon the topology of such trees. To minimise the impact of this inference error in this study, I used a phylogenetic method that takes a joint likelihood approach to infer gene trees. This method incorporates both the multiple sequence alignment of a gene family and the known species tree to infer gene trees. The topology of this 'known' species tree was derived from the current best estimate according to the angiosperm phylogeny working group (Stevens 2001) with branch lengths estimated in this study from a concatenated multiple sequence alignment of all single-copy gene orthogroups that contain sequences from ≥75% of the species (N = 1,230). This approach was taken as it directly addresses the issue of gene-tree inference error and improves the accuracy of gene-tree reconstruction (Szllosi et al. 2015). The software program PHYLDOG (Boussau et al. 2013) was used to infer these species-tree aware gene trees, and these trees were used for the analyses presented in this chapter.

However, to provide a comparison using a different phylogenetic method that does not incorporate the known species tree in the gene-tree inference step, the complete set of gene trees was also inferred from multiple sequence alignments using the software program IQ-TREE (Nguyen et al. 2015). As these gene trees are not reconciled to a species tree, it is unknown which branches in the gene trees correspond to specific branches in the species tree. To enable a collective analysis of these unreconciled gene trees, gene-tree branches were mapped to branches in the species tree using a heuristic

method of most recent ancestor identification (Swenson et al. 2012). This method is akin to the most recent common ancestor mapping used for the placement of changes in protein targeting on the species tree. Following this, ancestral state estimation was conducted to identify changes in organellar protein targeting across this set of gene trees.

Figure 2.9 summarises the results from repeating the main analyses of this study using these species-tree unaware gene trees. It also shows a comparison of the magnitude of changes identified across the species tree compared to the PHYLDOG species-tree aware method. Concomitant with the findings presented using the PHYLDOG method, changes in organellar protein targeting (gains and losses) were identified across the species tree, and the number of changes positively correlated with branch length (Figure 2.9 A). While there was some variation in the total number of gains and losses identified using the PHYLDOG and IQ-TREE data sets, the overall trend (i.e., the amount of change for each species-tree branch) was the same (Figure 2.9 B). It was also found that, as with the results from the PHYLDOG data set, the frequency of evolving a change in organellar targeting is higher following gene duplication events compared to non-duplication (speciation) events (Figure 2.9 C). Thus, whether reconciled or unreconciled trees were used the overall picture of protein gain and loss from these organelles remains the same.

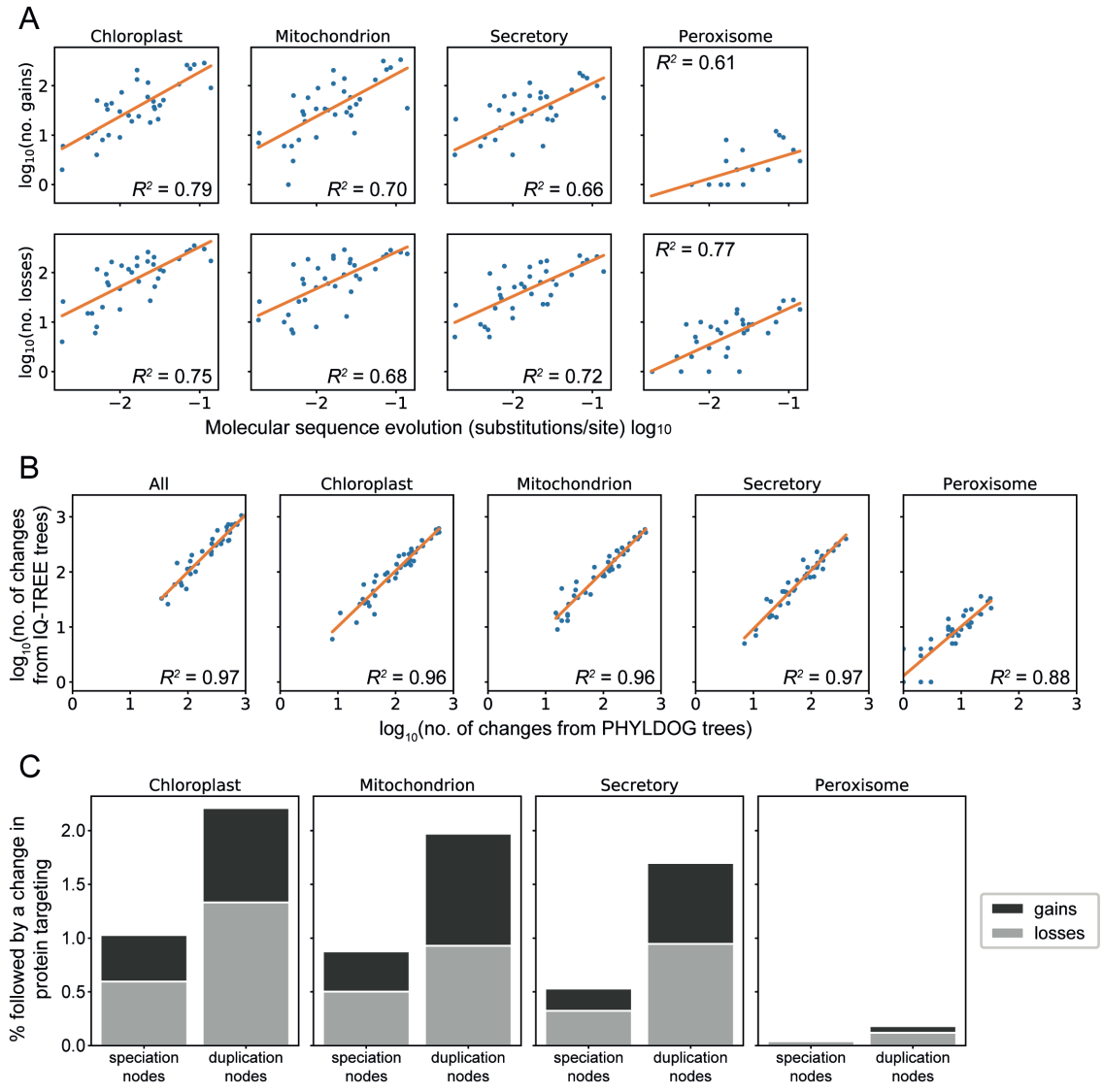


Figure 2.9. Comparison of gene-tree inference method (PHYLDOG vs IQ-TREE) in the identification of changes in organellar protein targeting. A) The relationship between molecular sequence evolution and the number of changes in protein targeting identified along species-tree branches using gene trees inferred with IQ-TREE. There was a positive relationship (R^2 shown, $P < 0.001$) between species-tree branch length (amino acid substitutions per site) and the number of gains and losses to each organelle. **B)** There is a positive correlation between the number of changes in protein targeting identified at each branch of the species tree using gene trees inferred with PHYLDOG and gene trees inferred with IQ-TREE (R^2 shown, $P < 0.001$). **C)** Changes in protein targeting identified using gene trees inferred with IQ-TREE were found to occur more frequently following gene duplication for each organelle (hypergeometric test, $P < 0.01$). This replicates the results found using gene trees inferred with PHYLDOG.

2.2.9 Using alternative organellar target signal prediction tools replicates the main findings presented in this chapter

When this study was conceived, I looked for a target signal prediction tool that could predict protein targeting to multiple organelles, could be run with large amounts of data (1,143,241 protein sequences) and was optimised for plants (i.e., had a plant training set for any machine learning algorithm). TargetP fitted these requirements and its algorithm performs reasonably well for each organellar target signal compared to other methods (Emanuelsson et al. 2000; Armenteros et al. 2019). However, there are several other tools available and a ranked list of several popular predictors can be found as part of the subcellular localisation database for *Arabidopsis* proteins (SUBA) (Hooper et al. 2014). Most of these predictors exist as web tools so are not suitable for the analysis of whole-proteome data sets. Nonetheless, two predictors that can be downloaded and used for large data sets are WoLF PSORT (Horton et al. 2007) and iPSORT (Bannai et al. 2002). To investigate whether the findings presented using TargetP can be replicated using WoLF PSORT and iPSORT I repeated the ancestral state estimation analysis using target signal prediction by these two additional methods for proteins gained and lost to the chloroplast and mitochondrion. As discussed in Chapter 1, WoLF PSORT and iPSORT both employ different strategies in the prediction of organellar target signals to TargetP. WoLF PSORT relies on a set of relatively simple predefined rules to identify target peptides (Horton et al. 2007), while iPSORT implements a k-nearest neighbour classifier for a variety of protein sequence features (Bannai et al. 2002). These methods contrast to TargetP's algorithm that was trained on experimentally validated organellar proteins using a neural network (Emanuelsson et al. 2000).

WoLF PSORT and iPSORT were used to construct two independent data sets of chloroplast and mitochondrial proteins from the same protein sequence data used in the TargetP analysis. Using the predictions made by both tools, the relationship between species-tree branch length and the number of identified changes in protein targeting was

investigated. As with the data set constructed using TargetP, there was a positive linear correlation between the amount of molecular sequence evolution (amino acid substitutions per site) and the number of changes in organellar targeting found to have occurred along each branch of the species tree (Figure 2.10 A&C). Using the WoLF PSORT and iPSORT data set it was also found that changes in protein targeting occur more frequently following gene duplication events compared to non-duplication (speciation) events (Figure 2.10 B&D). Thus, the main findings of the chapter are robust to the method of protein target signal inference that is used.

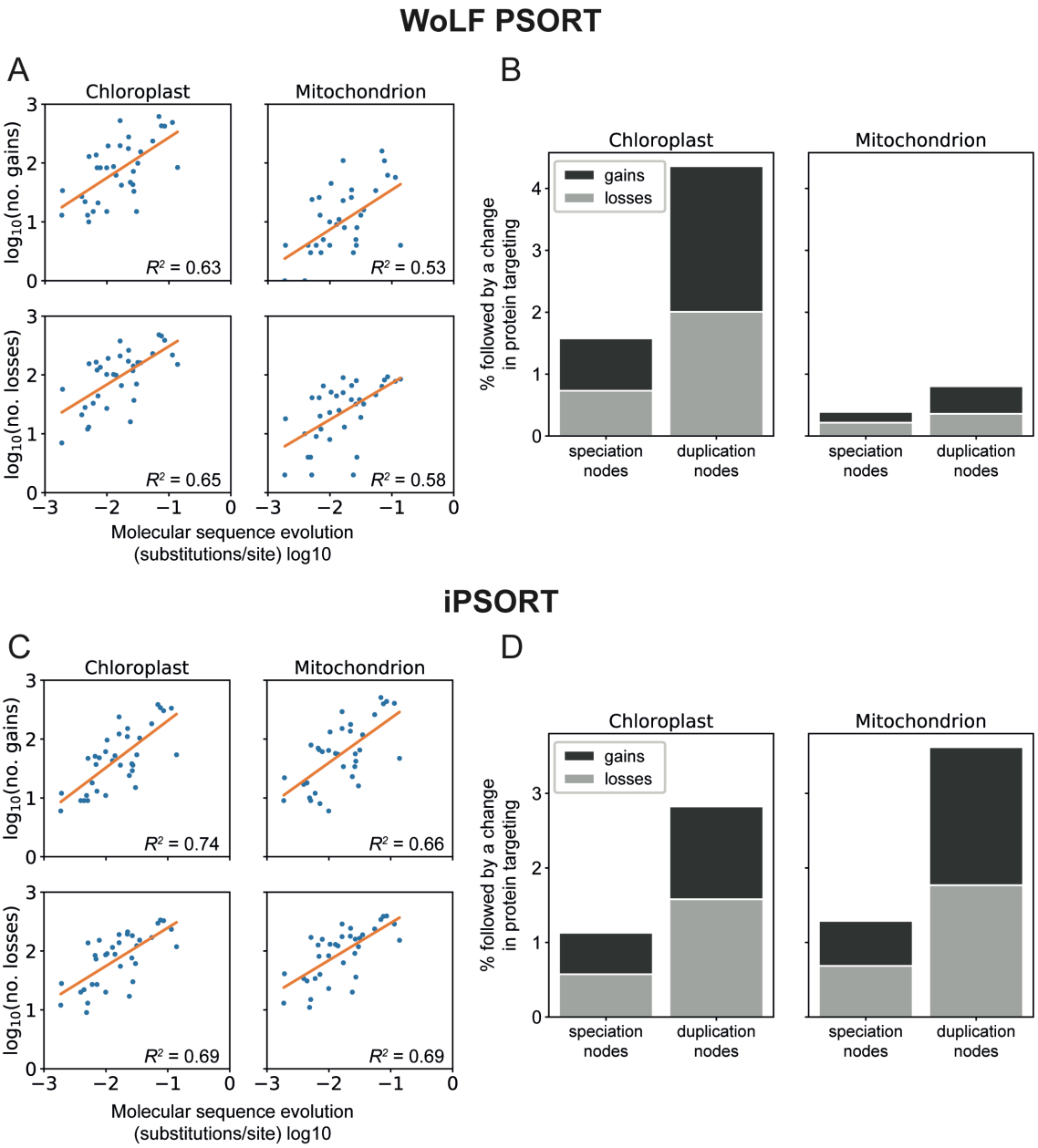


Figure 2.10. Replication of the main findings in this chapter using the alternate target signal prediction methods WoLF PSORT (Horton et al. 2007) and iPSORT (Bannai et al. 2002). A & C) There was a positive relationship (R^2 shown, $P < 0.001$) between species-tree branch length (amino acid substitutions per site) and the number of gains and losses to the chloroplast and mitochondrion when organellar proteins were predicted by WoLF PSORT and iPSORT, respectively. **B & D)** Changes in protein targeting were found to occur more frequently following gene duplication when using WoLF PSORT and iPSORT, respectively (hypergeometric test, $P < 0.01$).

2.3 Discussion

The partitioning of diverse cellular functions into organelles facilitated the evolution of complex multicellular life (Lane and Martin 2010). As a consequence, the origins and early evolution of organelles have been the subject of much research (Mast et al. 2014). Beyond this however, the differential evolution of organelles in different eukaryotic lineages has received much less attention. The work presented in this chapter provides substantial new insight into the dynamics of organellar proteome evolution in land plants. It reveals that there has been continuous change in the nuclear-encoded proteome of organelles since plants colonised the land approximately 450 million years ago. Furthermore, gene duplication has been uncovered as playing a role in accelerating the pace of organellar proteome evolution in plants, revealing a novel mechanism by which evolutionary changes in the nuclear genome impact the evolution of organelles. The implications of this, and the fact that changes in organellar targeting are enriched among proteins with gene regulatory functions, are discussed below.

Although there has been much interest in the comparison of organellar genome content between species (Palmer et al. 2000; Green 2011; Daniell et al. 2016), to date there has been little research into the diversity between the nuclear-encoded organellar proteomes of different eukaryotic species. Furthermore, although it is widely reported that orthologous genes can encode proteins with disparate subcellular localisations, there has been no investigation into the extent to which changes in organellar targeting of proteins occur during evolution. The analyses in this study suggest that there have been substantial changes to the nuclear-encoded proteome of organelles during plant evolution. At least 6,480 and 6,157 gains and losses in protein targeting were identified to have occurred to the chloroplast and mitochondrion, respectively, since the evolution of land plants. This amounts to considerable modulation of the organellar proteomic environment, far greater than that which results from changes in organellar genome content. For instance, comparison of changes in chloroplast gene content among a cohort of plants similar to

that used in this study (covering 64 species) found that most angiosperm plastid genomes contain 113 different genes, and that during the evolution of these species only 62 gene loss events among 38 different genes have occurred (Jansen et al. 2007). Similar reports of genome stability among angiosperms have been made for the mitochondrion genome (Adams et al. 2002). Thus, by comparison, there has been a far greater change (3 orders of magnitude more) to the proteome of the chloroplast and mitochondrion as a result of changes in organellar targeting of nuclear-encoded proteins. This may not, however, reflect relative functional impact given that organellar-encoded proteins represent a highly conserved and specialised suite of proteins that constitute the core functions of organelles. Indeed, it has previously been reported that genetic variation at organellar loci has a disproportionately large effect on phenotype compared with equivalent nuclear loci (Joseph et al. 2013; Dobler et al. 2014). Nonetheless, the findings in this study suggest that considerable modulation of organelle proteomes has occurred during plant evolution via changes in protein targeting of nuclear-encoded genes, and that such changes may be a pervasive mechanism by which proteins acquire new functions. It will be interesting to know whether this phenomenon of continual change in organellar targeting of nuclear-encoded proteins occurs in other eukaryotic lineages.

Gene duplication (whether through individual duplications or polyploid events) is a recurrent theme in eukaryotic evolution (Zhang 2003; Taylor and Raes 2004; Soltis et al. 2015) and has been proposed as a major mechanism by which new genetic material is generated (Ohno 1970; Long et al. 2003; Conant and Wolfe 2008; Magadum et al. 2013; Panchy et al. 2016). It has previously been suggested that gene duplication might facilitate changes in organellar protein targeting due to the fact that, in some instances, gene duplication leads to genetic redundancy and a relaxation of purifying selection on one or both gene copies (Byun-McKay and Geeta 2007; Marques et al. 2008; McKay et al. 2009; Byun and Singh 2013). Under these conditions, the accumulation of the genetic changes necessary for the evolution of a new target signal (or its loss) may be facilitated. Such a

change could then be fixed by either drift (in the case for neutral or weekly deleterious alleles) or selection (for advantageous alleles). The findings presented in this study corroborate this hypothesis, with changes in organellar targeting more likely to occur (or be retained) following gene duplication. As gene duplication facilitates changes in organellar protein targeting, it is thus a key facilitator of lineage-specific organellar evolution.

It has previously been reported that there are biases in the types of genes that are retained following individual and whole-genome duplications (Conant et al. 2014; Freeling et al. 2015; Wendel et al. 2016). Specifically, genes encoding proteins with interconnected functions, such as those that form multiprotein complexes or participate in regulatory networks, have been found to be preferentially retained following whole-genome duplications, whereas the same sets of genes tend not to be retained following individual gene duplications (Maere et al. 2005; Blomme et al. 2006; Freeling 2009; Tasdighian et al. 2017; Liang and Schnable 2018; Wendel et al. 2018). The reason for this is that natural selection acts to maintain stoichiometry among interconnected proteins. Thus, following whole-genome duplication, the duplication of all highly interconnected genes means that the subsequent loss of any one copy will be selected against in order to prevent gene balance perturbations. Concomitantly, individual duplications of the same genes are expected to revert to single copy for the same reason. In this study, no difference was observed in the rate of change in protein targeting following individual gene and whole-genome duplication events, but one interesting avenue of future work would be to investigate whether there is a difference in the types of proteins which undergo a change in protein targeting following these two different duplication events.

Concomitant with the gene-balance following gene duplication hypothesis, one would also expect that selective pressure to maintain stoichiometry would also influence the propensity of a protein to undergo a change in protein targeting more generally. Specifically, one would expect a low rate of change in organellar targeting among highly

interconnected proteins, in particular those that directly interact with several other gene products (RNA or proteins), as the more highly connected a protein is, the less likely it will be able to function in a new location where some or all of its interacting proteins are absent. This idea is analogous to complexity hypothesis which is used to explain why certain proteins are more or less likely to be horizontally transferred between different microbial species. The complexity hypothesis states that genes whose products interact with a large numbers of other gene products are transferred horizontally between species at a lower rate than less connected genes (Jain et al. 1999; Cohen et al. 2011; Novick and Doolittle 2020). Work concerning the complexity hypothesis has demonstrated that it is the connectedness of proteins, as determined by the number of protein-protein (or protein-RNA) interactions, that correlates with the likelihood of being transferred, rather than the particular biological function of the protein (Cohen et al. 2011; discussed in Novick and Doolittle 2020). While the degree of interconnectedness for proteins which have undergone a change in targeting is not investigated directly by the work in this chapter, it is interesting to note that there was no particular enrichment for proteins that form multiprotein complexes. Furthermore, among proteins gained and lost to the chloroplast, mitochondrion, and secretory pathway there was enrichment for proteins with biological functions such as ‘hormone metabolism’ and ‘transport’; categories which mostly include proteins with discrete molecular functions, such as enzymes and transporters, respectively. Such proteins are expected to have low protein-protein interactions, although again this was not tested here. Indeed, within each top-level GoMapMan category it is expected that there will be proteins on a spectrum of interconnectedness and thus an interesting avenue of future work will be to investigate whether the rate of change in protein targeting correlates with the degree of protein connectedness, rather than particular biological functions. Furthermore, it would also be interesting to discover whether the rate at which changes in protein targeting occur is dependent on whether the transfer involves the acquisition of a novel protein in a new location, or whether a paralog for the protein already exists in the new location. Again, this idea is taken from discussions of the

complexity hypothesis which similarly predicts that the complexity barrier for the transfer of proteins to new species may be lower for paralog or xenologous additions, rather than gene family acquisitions (discussed Novick and Doolittle 2020). This is something that could be directly tested for organelle proteome additions using my data set and would provide valuable insight into the selective pressures that govern the likelihood of proteins to be re-targeted during evolution.

Another set of genes which are expected to be resistant to changes in protein targeting are those that function in core organellar functions. Such proteins are typically resistant to gene duplication and consistently return to single-copy status suggesting that the presence of more than one copy is deleterious for the cell (dosage-sensitive). Genes that fall in to this category include those with functions in genome integrity, for example DNA repair, and those with functions in core organellar metabolism, for example genes that encode subunits of the chloroplast photosystem I and II (Blanc and Wolfe 2004; Li et al. 2016). Again, it is expected that this set of dosage-sensitive proteins will also be resistant to changes in organellar targeting. Interrogation of the functions of proteins that did not evolve changes in organellar targeting during plant evolution supported this hypothesis; among proteins that have never undergone a change in targeting there was enrichment for functional terms related to core organellar processes such as photosynthetic and mitochondrial electron transport, and for functions related to DNA synthesis and chromatin structure. Thus, the dosage sensitivity of a protein may affect its predisposition to changes in localisation.

In conclusion, the results from this chapter present a holistic picture of a previously unstudied mechanism of organelle evolution. Moreover, this is the first study to quantify the extent to which changes in organellar protein targeting have occurred during the evolution of a major group of eukaryotes. By using genomic data from a sampling of organisms that span the breadth of the plant kingdom, a picture of a dynamic organelle proteome has emerged which has been shaped by continual changes in the subcellular

targeting of proteins, substantially altering the regulatory landscape of these organelles. Moreover, it has revealed a novel way in which gene and whole-genome duplication play a role in facilitating organellar evolution.

2.4 Materials and methods

2.4.1 Problem definition and approach

In this chapter, I aim to identify the changes that have occurred to plant-cell organellar proteomes following the adaptation of plant to land. To do this, a phylogenomics approach was taken to predict when changes in protein targeting occurred during the evolution of different gene families. Species-tree-reconciled gene trees were inferred from genome data of 42 diverse plant species allowing gains and losses in organellar protein targeting to be mapped across the species phylogeny. These gains and losses were then interrogated to answer questions about the nature of organellar proteome evolution in the plant kingdom and the molecular mechanisms that drive changes in protein targeting. Full details of each step of this approach, as well as the location of the complete data set and all scripts required to repeat the analyses are described below.

2.4.2 Availability of data and material

All data used and generated in this chapter are available in the Zenodo research data archive at the following address: https://doi.org/10.5281/zenodo.1414180. This archive contains the full set of sequences, accession numbers, target signal prediction data, orthogroups, and PHYLDOG-reconciled gene trees for each orthogroup. The archive also contains a data file detailing all the gene duplication events and changes in protein targeting events that were inferred for each orthogroup. A GitHub repository containing all relevant code necessary to repeat the analyses are available at https://github.com/RonaCostello/charting-organelle-protome-evolution.

2.4.3 Construction of orthogroups and inference of species-tree-reconciled gene trees

Protein sequences corresponding to the primary transcripts of 42 fully sequenced plant species were obtained from Phytozome v10 (Goodstein et al. 2012). OrthoFinder (Emms and Kelly 2015) and MAFFT-LINSI v7.058b (Katoh and Standley 2013) were used to infer

orthologous gene groups (orthogroups) and their multiple sequence alignments, respectively. Only orthogroups with at least four genes and representation from more than one species were taken forward for analysis.

PHYLDOG (Boussau et al. 2013) was used to simultaneously infer orthogroup gene trees and reconcile these gene trees to the species tree. PHYLDOG takes a joint likelihood approach to infer gene trees by using both the multiple sequence alignment and the known species tree. This reconciliation method was used to minimise the effects of gene-tree inference error and weakly supported partitions in gene trees. Moreover, PHYLDOG has previously been shown to improve gene tree reconstruction compared with methods that do not take the species tree into account (Szllosi et al. 2015). In all cases, orthogroup multiple sequence alignments were trimmed to remove columns containing more than 66% gap characters prior to PHYLDOG-tree inference using the “LG08” model of sequence evolution. Some of the largest orthogroups were too large to be analysed directly with PHYLDOG (the largest orthogroup contained 12,148 genes). Manual inspection revealed that these large orthogroups were not single orthogroups, but instead were fusions of multiple orthogroups originating from a gene duplication event that preceded the diversification of the species in the analysis. Thus, to enable the analysis of these data, the gene trees for these fused orthogroups were split into correctly circumscribed individual orthogroups at the ancient duplication branch by a process of tree inference and gene-tree–species-tree reconciliation. Each of these disentangled orthogroups was then analysed by PHYLDOG as described above.

To run PHYLDOG on the multiple sequence alignments described above requires a species tree with branch lengths as input. Here, the topology of the species tree was derived from the angiosperm phylogeny working group (Stevens 2001). However, this tree did not contain branch lengths. Thus, to infer branch lengths for this species tree, a concatenated multiple sequence alignment of all single-copy gene orthogroups that contained ≥75% of the species (N = 1,230) was constructed. This concatenated alignment

was subject to phylogenetic-tree inference with the topology constrained to the known species tree using FastTree v2.1.9 (Price et al. 2010) and the command: FastTree -gamma -nome -mllen -intree SpeciesTree.txt ConcatenatedAlignment.al > SpeciesTree_constrained.txt.

To provide a methodologically independent control, and mitigate against any potential overfitting caused by the use of the PHYLDOG method, I also carried out the complete analysis on unreconciled gene trees that were inferred directly from the multiple sequence alignments using IQ-TREE v1.6.6 with the settings –m TEST to automatically select the best-fitting model of sequence evolution for each gene-tree inference (Nguyen et al. 2015). Here, individual gene-tree branches were mapped to branches of the species tree using a heuristic method of last common ancestor identification as described previously (Swenson et al. 2012).

2.4.4 Prediction of organellar-targeted proteins

Of the 42 species included in this study, 37 comprise land plants and five comprise green algae. For each species, the set of proteins predicted to contain a chloroplast transit peptide, mitochondrial target peptide, secretory signal peptide, or the peroxisomal target signal 1 or 2 was identified. For the land plant species, TargetP v1.0 (Emanuelsson et al. 2000) in plant mode with default cut-offs was used to predict chloroplast transit peptides, mitochondrial target peptides, and signal peptides. For the five algal species (*Ostreococcus lucimarinus*, *Micromonas pusilla*, *Coccomyxa subellipsoidea*, *Volvox carteri*, *Chlamydomonas reinhardtii*), prediction was instead carried out with PredAlgo (Tardif et al. 2012) using its default cut-offs. In cases where an amino acid sequence did not meet the minimum length requirement for PredAlgo prediction, the TargetP prediction was taken instead.

The prediction of peroxisomal proteins was carried out by searching for the canonical plant peroxisomal target signals: PTS1 or PTS2 (Reumann 2004). Here, a protein sequence

was classified as having a PTS1 if it had any one of the nine different C-terminal tripeptide sequences: SRL, SRM, SRI, ARL, ARM, PRL, SKL, SKM, AKL. A protein sequence was classified as having a PTS2 peroxisome targeting sequence if it contained either of the PTS2 peptide sequences (R[LI]X5HL) in the N-terminus region of the protein (residues 1-30). TargetP does not take into account cases of dual localisation; however, if a protein was found to have a peroxisomal target signal and a positive TargetP localisation, it was labelled as dual peroxisomal localised.

TargetP was selected as the main target signal predictor as it performs well in benchmarks (Klee and Ellis 2005), is available to download, and has a "plant" mode based on a neural network trained on plant data. To provide additional support for the findings presented in this study, I also ran the complete analysis using two alternative subcellular localisation predictors which take contrasting approaches to target signal prediction: WoLF PSORT (Horton et al. 2007) and iPSORT (Bannai et al. 2002). Here, WoLF PSORT and iPSORT were used to construct two independent data sets of chloroplast and mitochondrion proteins from the same protein sequence data used for TargetP prediction. It should be noted that WoLF PSORT only accepts protein sequences with a length >30 and so all protein sequences shorter than this were labelled as 'other'. Also, WoLF PSORT detected 13 dual localised proteins in the data set and so these proteins were labelled as both chloroplastic and mitochondrial in my data set. Lastly, WoLF PSORT assigns each protein sequence a score for various localisations and the highest scoring compartment was used in the assignment of protein localisation for my data set (winner takes all). As with the TargetP analysis, PredAlgo (Tardif et al. 2012) was used for the prediction of algal proteins.

2.4.5 Ancestral state estimation of organellar protein targeting

Ancestral gains and losses of protein targeting were identified in orthogroups using maximum likelihood ancestral state estimation. For each protein, the presence or absence of a particular organellar target signal was treated as binary trait data and the leaves (i.e.,

extant proteins) of the orthogroup trees assigned “1” or “0” accordingly. Here, each type of target signal was considered separately and each orthogroup tree analysed independently. The presence or absence of a target signal in ancestral protein sequences represented by each internal branch of an orthogroup tree was inferred using the “ACE” function in the R package ape v5.2 (Paradis et al. 2004) using the “all rates different” model for discrete data. The “all rates different” model was selected because the transition probabilities between states (i.e., presence/absence of target signals) are unknown and cannot be assumed to be equal. Furthermore, an “all rates different” model performed better on average than an “equal rates” model as assessed by a chi-squared log likelihood ratio test across all trees.

The ACE (ancestral character estimation) function was used to infer the probability (between 0 and 1) that an ancestral protein sequence (represented by an internal branch in an orthogroup tree) had a specific organellar targeting signal. To identify changes in protein targeting in orthogroup trees, a winner takes all approach was used whereby branches with an ACE score of ≥0.5 were assigned as organellar-targeted proteins and branches with scores of <0.5 were assigned as non-organellar targeted proteins. Further processing and filtration were carried out as described below.

2.4.6 Identification of changes in organellar protein targeting during evolution

The ancestral state estimation data were analysed to identify when changes in organellar targeting occurred during the evolution of a protein as described by its phylogeny. Losses in organellar targeting were identified when there was a transition from a targeted state to a non-targeted state on immediately consecutive branches in an orthogroup tree, and vice versa for a gain. As ancestral state estimation is expected to be sensitive to targeting prediction error or gene-tree error, a stringent filter was imposed for the identification of gains and losses. This filter required that ≥75% of extant proteins descendant from the

branch on which a change is estimated to have occurred retain the changed subcellular state, and >75% of extant proteins descendent from the sister branch maintain the ancestral state. For example, consider an internal bipartition within a gene tree that has two descendant branches X and Y. If a gain in chloroplast targeting is predicted to occur on branch X, then ≥75% of the extant proteins that are descendent from branch X must be predicted to contain a chloroplast transit peptide, whereas ≥75% of the extant proteins that are descendent from branch Y must not be predicted to contain a chloroplast transit peptide. Only if both these criteria are met would a change in chloroplast targeting be assigned to branch X in the orthogroup tree. This requirement meant that the estimation of a change in protein targeting of an ancestral protein was always informed by the predicted targeting of three or more extant proteins. Furthermore, it was also required that protein sequences from two or more species must subtend any branch under consideration. A worked example for the identification of a change in protein targeting in a gene tree is provided in Appendix A2.1.1. Only the changes in organellar targeting that passed this filter were used in subsequent analyses. The branches within orthogroup trees on which these changes occurred were mapped to branches in the species tree using the gene-tree–species-tree reconciliation provided by PHYLDOG (this is analogous to the most recent common ancestor mapping method used in the processing of IQ-TREEs). This enabled the number of gains and losses in protein targeting to each of the four organelles to be tallied for each branch of the species tree.

2.4.7 Comparing the rate of molecular sequence evolution for gene-tree branches associated with and without a change in protein targeting

The normalised rate of molecular sequence evolution for each orthogroup-tree branch was calculated as the length of the branch (substitutions per site) divided by the length of the corresponding branch in the species tree (also substitutions per site). This normalisation was conducted to allow relative rates of molecular sequence evolution to be compared across branches both within and between orthogroup gene trees. To mitigate against error

introduced by ambiguity in gene-tree–species-tree mapping and the effects of gene duplication, this normalised rate of molecular sequence evolution was only calculated for branches which could be unambiguously mapped to the species tree and which were classified as having no evidence of gene duplication. For all qualifying branches, the normalised rate of molecular sequence evolution for those branches associated with a change in organellar targeting was compared with the analogous rate for branches that had no evidence for a change in organellar targeting. To further mitigate against potential biases arising from differences in the phylogenetic distribution of these two sets of gene-tree branches, a random sampling approach was used whereby the number and distribution of branches sampled for both sets were identical. This random sampling process was repeated 1,000 times to obtain a Monte Carlo *P*-value.

2.4.8 Incorporation of divergence time data

To estimate the average rate at which proteins gained or lost organellar target signals during land plant evolution, eight branches were selected from the species tree for which a divergence time is known. The number of gains and losses in targeting to each organelle was then summed for species-tree branches between the branch at the base of the land plants (taken as 450 million years ago (Morris et al. 2018)) and each of these eight dated-branches thereby allowing the number of changes per million years to be calculated. Species-tree branches used in this analysis and references for divergence time estimates are shown in Table 2.2. It should be noted, however, that the eight branches that were selected all share at least part of a common path of evolution and are therefore not independent.

Table 2.2. Divergence times used to produce a time-calibrated phylogeny. The origin of land plants is estimated at 450 million years ago (mya).

Species tree node	Diverging lineages	Estimated divergence time (mya)	Reference for estimated divergence time	Time since land plant origin (mya)
N39	*A. lyrata*, *A. thaliana*	13	(Beilstein et al. 2010)	437
N40	*C. rubella*, *C. grandiflora*	0.04	(Guo et al. 2009)	449.96
N35	*M. domestica*, *P. persica*	92	(Xiang et al. 2017)	358
N31	*P. vulgaris*, *G. max*	19.2	(Schmutz et al. 2014)	430.8
N15	*S. lycopersicum*, *S. tuberosum*	7.3	(Sato et al. 2012)	442.7
N20	*Z. mays*, *S. bicolor*	11.9	(Swigoňová et al. 2004)	438.1
N21	*P. virgatum*, *S. italica*	5	(Bennetzen et al. 2012)	445
N17	*O. sativa*, *B. distachyon*	46.5	(Vogel et al. 2010)	403.5

2.4.9 Identification of changes in organellar protein targeting following gene duplication and non-duplication (speciation) events

To investigate whether changes in organellar protein targeting occur more frequently following gene duplication events or non-duplication (speciation) events, it was necessary to identify branches which correspond to these events in each orthogroup tree. To prevent tree inference error from influencing the results, a stringent filter was applied to enable identification of high-confidence gene duplication branches and speciation branches in each orthogroup tree. High-confidence gene duplication branches were defined as branches for which a gene duplication event was retained in all descendant species of both child branches subtending the duplication event. Similarly, a high-confidence speciation branch was selected as a branch that has no evidence of gene duplication and from which there was no subsequent gene loss in any of the descendant species. Furthermore, the speciation branch had to encompass an entire speciation event in the species tree (i.e., the parent and child branches are mapped to a different species tree branch). In both cases (duplication and speciation branches), complete retention of all genes in all descendant species is required and thus the gene sets can be considered

equivalent. A corollary of this stringent selection criterion is that the branches which pass this filter are also unambiguously placed within the species tree.

A change in organellar targeting immediately following a gene duplication event was identified when a gain or loss in targeting could be placed on one of the two direct child branches subtending the duplication. A change in organellar targeting immediately following a gene speciation event was identified when a gain or loss in targeting could be placed on a speciation branch. Changes in more distant branches (i.e., grandchild branches or great-grandchild branches, etc.) were not considered in this analysis. Thus, only changes in protein targeting that occurred before the next speciation event or gene duplication event were analysed.

To mitigate against potential biases arising from differences in the phylogenetic distribution of the set of duplication branches and the set of speciation branches, a random sampling approach was also conducted whereby the number and phylogenetic distribution of branches sampled from both sets were identical. This random sampling process was repeated 1,000 times to obtain the Monte Carlo *P*-value.

2.4.10 Functional term enrichment analysis

Orthogroups (protein families) were assigned MapMan terms and sub-chloroplast localisation terms by inheriting the terms associated with the genes found within them. MapMan terms were taken from the GoMapMan webpage (Ramšak et al. 2014) and sub-chloroplast terms were assigned using the hierarchical structure provided on the Plant Protein Database (Sun et al. 2009) using only experimentally validated proteins. To test for enrichment, the hypergeometric test was performed, and *P*-values corrected for multiple testing using the Benjamini-Hochberg correction. The aim was to identify functional enrichment among orthogroups whose proteins are differentially localised. To avoid simply identifying functional terms that are enriched in organellar-targeted gene

families, the background population for this test was orthogroups with at least one predicted organellar-targeted protein.

2.4.11 Monte Carlo resampling

To account for differences in sample size and phylogenetic distribution of different events in gene trees (e.g., duplication and speciation events), Monte Carlo resampling was used to assess the significance of differences identified between groups. A Monte Carlo *P*-value was calculated as

$$n + 1/N + 1$$

where *n* is the number of simulations where the simulated value did not agree with the prior expectation and *N* is the total number of simulations.

Chapter 3: Identification of changes in protein targeting associated with the evolution of C_4 photosynthesis

3.1 Introduction

The previous chapter analysed the changes in organellar protein targeting that occurred during the 450-million-year evolution of land plants. This presented an overview of the global patterns of change in protein targeting, providing new insight into the evolution of the protein repertoire of a number of plant cell organelles and how the evolution of this repertoire has been accelerated by gene and genome duplication. In this chapter I use the same methodological approach but instead of providing a global analysis across the entire phylogeny, I focus on key branches in the species tree to discover and analyse the changes in protein targeting that occurred during the evolution of a particular plant trait: C_4 photosynthesis. This derived type of photosynthesis has evolved independently in several plant lineages and is characterised by a metabolic cycle that functions to concentrate carbon dioxide (CO_2) around the photosynthetic carbon fixation enzyme ribulose-1,5-bisphosphate carboxylase-oxygenase (RuBisCo), thus increasing its efficiency.

Several enzymes and transporters function in the C_4 photosynthetic cycle and there has been substantial prior research into the molecular sequence changes associated with the recruitment of these proteins from their ancestral function in a C_3 plant. One aspect of molecular sequence evolution that has received little attention however is the gain and loss of target signals. The C_4 cycle operates across several subcellular compartments so changes in protein targeting may have been important in the adaptation of proteins to a C_4-function during evolution. Precedent for this exists for at least one enzyme, carbonic anhydrase, which lost a chloroplast transit peptide coincident with the evolution of C_4 photosynthesis in both a eudicot and monocot C_4 lineage; a change that is necessary for the cytosolic function of this protein in the C_4 cycle (discussed in Niklaus and Kelly 2019). In this chapter I investigate whether other such changes in protein localisation are

associated with the evolution of C_4 photosynthesis by interrogating the changes in protein targeting that occurred during the evolution of C_4 photosynthesis in four C_4 monocot species.

3.1.1 Background

C_4 photosynthesis is a paradigm for the convergent evolution of complex traits. There are thought to be over 60 independent origins of C_4 photosynthesis distributed across the angiosperm phylogeny (Sage 2016) and its evolution involves substantial changes to leaf anatomy, development, and biochemistry (Hatch 1987; Gowik and Westhoff 2011). The function of C_4 photosynthesis is to concentrate carbon dioxide (CO_2) around the primary photosynthetic carbon fixation enzyme in the leaf, RuBisCo, thus limiting the energy-costly side reaction that RuBisCo catalyses with oxygen (O_2) (Bauwe et al. 2010). The repeated evolution of C_4 photosynthesis is thought to have occurred primarily as an adaptation to a drop in atmospheric CO_2 concentration that occurred approximately 35 million years ago (Ehleringer et al. 1997). Together the biochemical and anatomical features of C_4 photosynthesis result in significant energy (Xin-Guang Zhu et al. 2008) water (Ghannoum et al. 2010) and nitrogen savings (Ghannoum et al. 2010), and today C_4 species represent some of the world's most productive crops (von Caemmerer and Furbank 2016). Concerns about future food security has also motivated interest in introducing C_4 photosynthesis into C_3 crops such as rice to increase their yield potential (Hibberd et al. 2008; Von Caemmerer et al. 2012). This is an ambitious goal and understanding the genetic changes that distinguish C_4 species from their C_3 relatives holds potential to help guide these efforts.

As introduced in Chapter 1, most plants carry out an ancestral type of photosynthesis known as C_3 photosynthesis, named as such because in C_3 plants CO_2 is directly fixed by RuBisCo in the chloroplasts of the leaf mesophyll cells to produce the *three*-carbon compound 3-phosphoglycerate (3-PGA) – this being the first step in the Calvin-Benson cycle (Figure 3.1 A). A problem lies herein in that O_2, which is abundant in the atmosphere and thus also present in the intracellular space of the leaf, competes with CO_2 for the

active site of RuBisCo. Compared to the carboxylation reaction, the oxygenation reaction catalysed by RuBisCo produces half the amount of 3-PGA and a toxic by-product that has to be processed in an energy-consuming process known as photorespiration (reviewed in Bauwe et al. 2010). To overcome this limitation of RuBisCo, several plants have evolved strategies to concentrate CO_2, and thus displace O_2, in the vicinity of RuBisCo. One such strategy is C_4 photosynthesis. In plants that carry out C_4 photosynthesis CO_2 is initially fixed into a four-carbon compound by a different enzyme, phosphoenolpyruvate carboxylase (PEPC), which has no affinity for O_2 (Figure 3.1 B). This fixation reaction by PEPC is spatially separated within the leaf from the CO_2 reduction activity of RuBisCo and the rest of the Calvin-Benson cycle. In most C_4 plants this spatial separation is achieved by splitting the C_4 photosynthetic pathway between two specialised cell types in the leaf: typically, the mesophyll and bundle sheath cells (Muhaidat et al. 2007; Sedelnikova et al. 2018). These cells are arranged in a concentric wreath-like pattern with the mesophyll cells forming an outer layer and the bundle sheath cells forming an inner layer that then surrounds the vascular tissue (Figure 3.1 B). To ensure an efficient C_4 photosynthetic cycle is established throughout the leaf, vein density is also typically increased in C_4 leaves such that almost every mesophyll cell is in contact with a bundle sheath cell (Figure 3.1 B). This pattern of venation in the C_4 leaf is known as Kranz anatomy (Kranz being the German word for ‘wreath’, named as such by the German botanist Gottlieb Haberlandt) (Haberlandt 1918; Sedelnikova et al. 2018). Initial carbon fixation by PEPC occurs in the outer layer of mesophyll cells while carbon reduction by RuBisCo is restricted to the inner layer of bundle sheath cells (Hattersley et al. 1977). The establishment of a diffusion gradient between these two cell-types results in CO_2, locked up in a four-carbon compound, moving from the mesophyll to the bundle sheath cells through the cytoplasmic channels (plasmodesmata) that connect the two cell types. Once in the bundle sheath cell the four-carbon compound is enzymatically decarboxylated thus releasing CO_2 at high concentrations around RuBisCo. A three-carbon compound is also released from this decarboxylation reaction and diffuses back to the mesophyll cells to re-start the cycle.

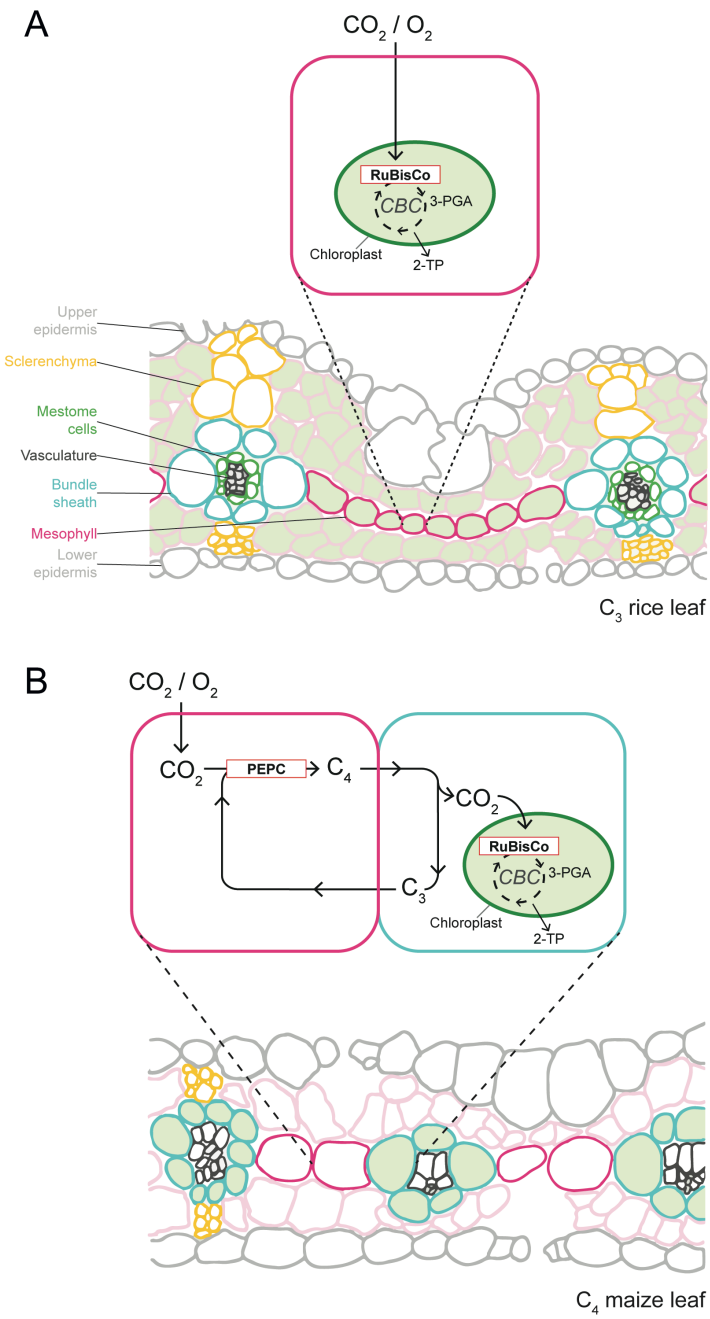


Figure 3.1. Comparison of leaf anatomy and localisation of carbon reduction in a C_3 (A) and C_4 (B) leaf. Transverse cross sections are shown for a C_3 rice (*Oryza sativa*) and C_4 maize (*Zea mays*) leaf. The localisation of RuBisCO, and thus CO_2 reduction, is shown by the light green fill. A simplified schematic of the C_3 and C_4 carbon assimilation pathway is also shown. This figure was adapted from (Sedelnikova et al. 2018). **Abbreviations**: carbon dioxide (CO_2), oxygen (O_2), ribulose-1,5-bisphosphate carboxylase-oxygenase (RuBisCo), phosphoenolpyruvate carboxylase (PEPC), 3-phosphoglycerate (3-PGA), 2-triose phosphate (2-TP), four-carbon compound (C_4), three-carbon compound (C_3).

Several biochemical reactions constitute the C_4 pathway and, in addition to being split between two cell types, these reactions are partitioned between multiple *sub*cellular compartments of the cell. The extent of this partitioning is reflected by the fact that an operational two-cell C_4 cycle requires 10-12 intracellular transmembrane transport steps for every molecule of CO_2 that is assimilated (Weber and von Caemmerer 2010). In the C_3 cycle on the other hand the only transmembrane step is the export of triose-phosphate (TP) produced in the Calvin-Benson cycle out of the chloroplast, and that happens at most once for every three molecules of CO_2 that are assimilated (Weber and von Caemmerer 2010). The exact number of biochemical reactions and transmembrane steps that constitute the C_4 pathway depends on the C_4 subtype under consideration. Typically, C_4 species are categorised into one of three subtypes depending on the C_4 decarboxylation enzyme present in the bundle sheath cells: either NADP-dependent malic enzyme (NADP-ME), NAD-dependent malic enzyme (NAD-ME) or phosphoenolpyruvate carboxykinase (PEP-CK). These three decarboxylating enzymes vary in substrate, product and subcellular localisation and so their associated cycles have a different complement of enzymes and transporters needed to maintain carbon flux through them (a diagram of carbon flow through each of these pathways in shown in Figure 3.2).

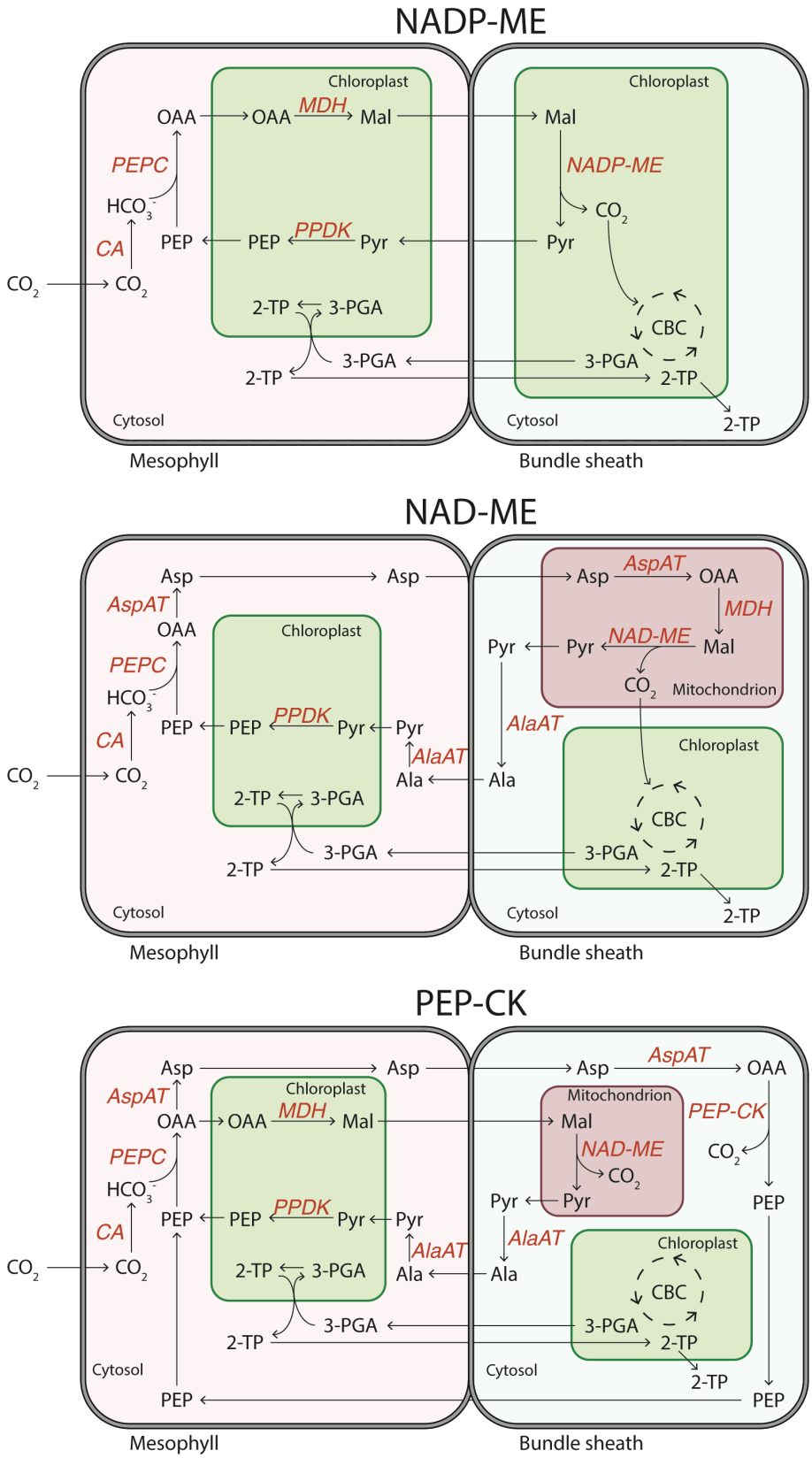


Figure 3.2. Schematic showing the route of metabolite flow for the three main subtypes of C_4 photosynthesis. Metabolite flow is shown with arrows. Where an arrow represents the conversion of one metabolite to another, the enzyme catalysing the

reaction is shown in red italics. It should be noted that variations on each of these pathways exist in nature and that this figure serves to illustrate the number and subcellular distribution of the reactions that constitute a typical C_4 pathway. The transporters that facilitate the transmembrane transport of metabolites are not shown. **Abbreviations**: carbon dioxide (CO_2), bicarbonate (HCO_3^-), oxaloacetate (OAA), aspartate (Asp), malate (Mal), pyruvate (Pyr), phosphoenolpyruvate (PEP), alanine (Ala), 3-phosphoglycerate (3-PGA), 2-triose phosphate (2-TP), Calvin Benson cycle (CBC), carbonic anhydrase (CA), phosphoenolpyruvate carboxylase (PEPC), aspartate aminotransferase (AspAT), malate dehydrogenase (MDH), NADP-dependent malic enzyme (NADP-ME), NAD-dependent malic enzyme (NAD-ME), phosphoenolpyruvate carboxykinase (PEP-CK), alanine aminotransferase (AlaAT), pyruvate phosphate dikinase (PPDK).

All of the enzymes and metabolite transporters that function in typical C_4 pathways have been shown to be present in C_3 plants albeit with different tissue specificities or expression patterns in the leaf (Häusler et al. 2002; Aubry et al. 2011; Gowik and Westhoff 2011; Reyna-Llorens and Hibberd 2017). The recruitment of these proteins from an ancestral function in a C_3 plant to their function in the C_4 pathway thus involved changes in gene regulation through the recruitment of trans- and cis-regulatory elements (Hibberd and Covshoff 2010; Niklaus and Kelly 2019) and molecular sequence adaptation of the proteins themselves, for example single nucleotide mutations leading to altered enzyme kinetics (Niklaus and Kelly 2019). In addition to C_4 genes being expressed at the right time and place, it was also necessary that their protein products were localised to the correct subcellular compartment. This could be achieved either through the recruitment of proteins already localised to the correct compartment in a C_3 ancestor or through molecular sequence evolution leading to changes in organellar protein targeting. The recruitment of proteins already localised to the correct compartment may have been facilitated by the fact that many proteins of the C_4 pathway come from large paralogous gene families whose

members may have already been differentially localised in a C_3 ancestor. This has shown to be the case for the enzyme aspartate aminotransferase (AspAT) which in the C_4 pathway is differentially localised depending on the C_4 subtype under consideration. In NAD-ME species, AspAT is localised to the cytosol in the mesophyll cells and the mitochondrion in the bundle sheath cells, while in PEP-CK species AspAT is localised to the cytosol in both cell types (Figure 3.2 B&C). Phylogenetic analysis of the AspAT gene family showed that following a series of ancient duplications three distinct clades of this protein emerged which differed in their subcellular localisations such that today most plant genomes possess at least one AspAT from each clade (Schlüter et al. 2019). It was found that across C_4 species the identity of the AspAT isoforms recruited to the C_4 pathway was strongly influenced by C_4 subtype: in NAD-ME species AspAT isoforms were recruited from both the mitochondrial and cytosolic clade, while in PEP-CK species an AspAT isoform was recruited from just the cytosolic clade (Schlüter et al. 2019).

As mentioned, the recruitment of proteins to the correct subcellular compartment for C_4 photosynthesis could also occur through changes in protein targeting. This is thought to have occurred for the enzyme carbonic anhydrase (CA) in at least two independent C_4 lineages. In C_3 plants the functions of CA are diverse and include roles in photosynthetic and non-photosynthetic tissue, with CA isoforms having been described that localise to the chloroplast, cytosol, mitochondrion, and plasma membrane (Badger and Price 1994; Aubry et al. 2011; DiMario et al. 2017). The predominant CA in the photosynthetic tissue of C_3 plants has been suggested to be localised to the chloroplast (Badger and Price 1994). In the leaves of C_4 species however, the protein encoded by the most highly expressed CA gene is found in the cytosol of mesophyll cells (Hatch and Burnell 1990; Badger 2003), where it functions to provide a source of HCO_3^- for primary carbon assimilation mediated by PEPC (Figure 3.2). The recruitment of CA to a C_4 function could thus arise either through changes in the regulatory sequences of a non-photosynthetic cytosolic isoform of CA or through the loss of the chloroplast transit peptide of the already

highly expressed isoform in the leaf. To date, this phenomenon has only been investigated in the C_4 species *Flaveria bidentis* (Tanz et al. 2009) and *Neurachne munroi* (Clayton et al. 2017), both of which have closely related C_3 and C_3-C_4 intermediates making them good models to study the C_3 to C_4 transition. In both cases, acquisition of high levels of cytosolic CA in the mesophyll cell occurred by the loss of a chloroplast transit peptide from the already highly expressed leaf protein (rather than by upregulation of a distinct cytosolic isoform). Thus, in two independent origins of C_4 photosynthesis, the evolution of the C_4 pathway was facilitated by a change in organellar targeting of a protein. It is unknown whether the loss of the transit peptide from the chloroplast localised CA is a common phenomenon in other C_4 lineages.

It is unclear to what extent changes in protein targeting during the evolution of C_4 photosynthesis may have been important in the recruitment of other proteins to a C_4 function. To date, there has been no genome-wide investigation into the changes in protein targeting that occurred during the evolution of C_4 photosynthesis in any C_4 lineage. Indeed, until now the identification and evolutionary interrogation of changes in protein targeting has been limited to studies looking at individual or just a few protein families. The work presented in Chapter 2 described the first methodological approach to identify changes in protein targeting at a genome-wide scale during the evolution of a large eukaryotic clade. In this chapter I use the same phylogenetic approach to interrogate those changes in protein targeting which occurred during the evolution of C_4 photosynthesis in four C_4 monocot species. The focus of this work was to identify whether any functionally relevant changes in protein targeting had occurred among known enzymes and transporters of the C_4 pathway concomitant with the evolution of C_4. In addition, I used several previously published transcriptomic data sets to allow for the discovery of changes in protein targeting among proteins whose genes exhibit expression patterns indicative of a role in C_4 photosynthesis.

3.2 Results

3.2.1 Identification of the changes in protein targeting that occurred concurrent with the evolution of C_4 photosynthesis in two monocot lineages

Ancestral state estimation was used to identify the changes in protein targeting that occurred concurrent with the evolution of four C_4 monocot species. These species comprise representatives from two independent origins of C_4 photosynthesis: *Zea mays* and *Sorghum bicolor* – which are descendent from an origin of C_4 in the Andropogoneae (Aliscioni et al. 2012) – and *Setaria italica* and *Panicum virgatum* – which are descendent from an origin of C_4 in the Paniceae (Aliscioni et al. 2012). Although these species were included in the data set of changes in protein targeting used in Chapter 2, a new data set was constructed in this chapter to make use of the latest gene model predictions for these species and an updated version of TargetP (Armenteros et al. 2019). For this new data set, proteome data were collected for the four C_4 species, an additional 16 plant species comprising C_3 angiosperms, and several outgroup species from the bryophyte, lycophytes, and green algae. A phylogeny for these species is shown in Figure 3.3. Orthogroups (*N* = 21,789) and species-tree-reconciled gene trees for each orthogroup were inferred from the proteome data. TargetP v2.0 was then used to predict which proteins in this data set contained a chloroplast transit peptide (cTP), mitochondrial target peptide (mTP) or a signal peptide (SP). A search for proteins containing a peroxisomal target signal (PTS) was conducted by manually searching for PTS motifs in the amino acid sequences of proteins. Orthogroup trees and target signal predictions were then used to infer ancestral changes in protein targeting using the maximum likelihood ancestral state estimation approach described in Chapter 2. As in Chapter 2, a 75% retention score was applied to only identify those changes in protein targeting which had been retained among extant proteins. By mapping gene-tree branches to the species tree, ancestral changes in protein targeting that occurred along branches of the species tree encompassing an origin of C_4 photosynthesis (N17 and N18, Figure 3.3) could be identified. In total, 101 and 145

changes (gains and losses) in protein targeting were mapped to branches N17 and N18 of the species tree, respectively. A breakdown of the number of gains and losses identified to each organelle is shown in Appendix Table A3.1, column A. It is possible that some changes in protein targeting may have occurred in parallel during the evolution of C_4 photosynthesis in the Andropogoneae and Paniceae clades. Such changes would be mapped to the ancestral branch of these two clades in the species tree (N14) and thus changes mapped to this branch were also considered in further analyses. In total, 211 changes (gains and losses) in protein targeting were mapped to branch N14 of the species tree.

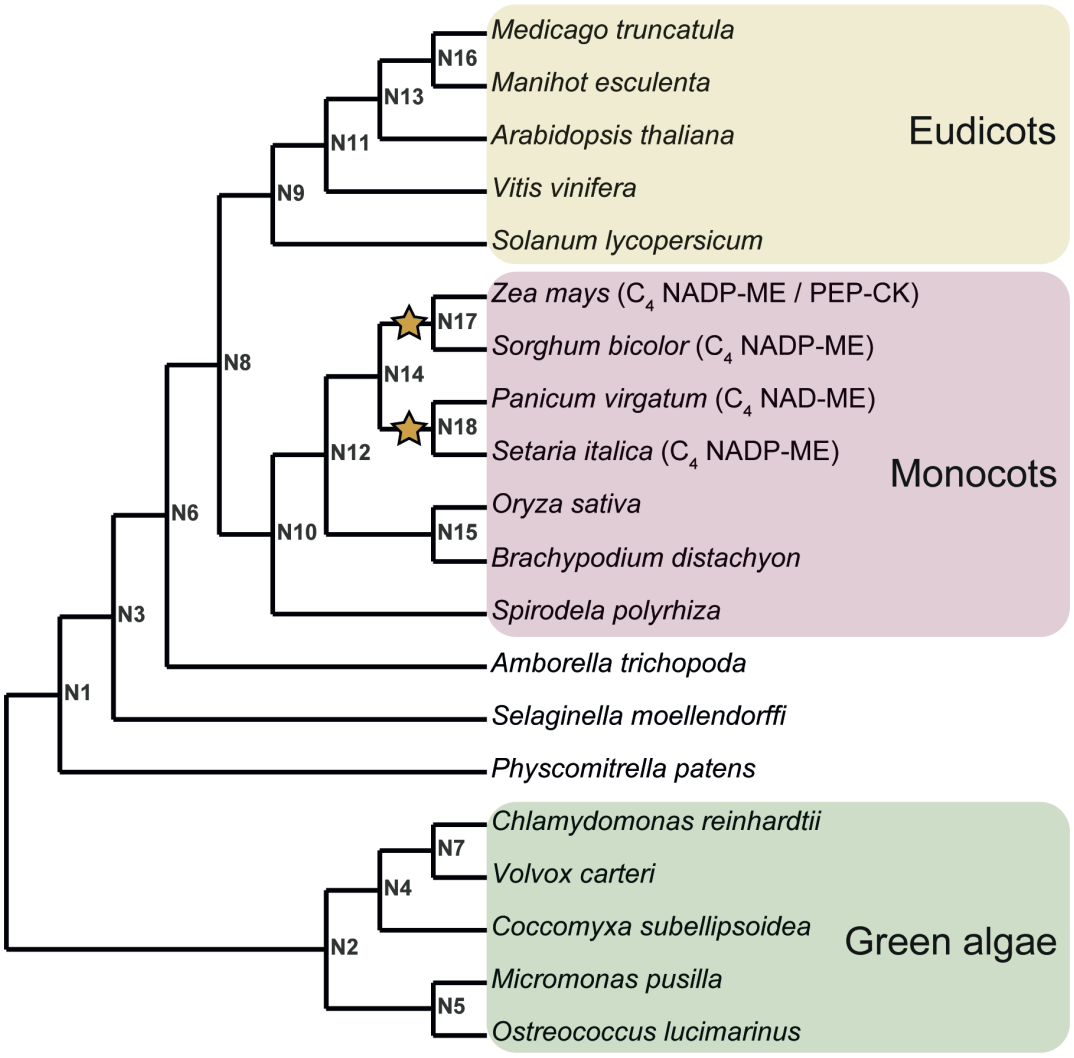


Figure 3.3. Phylogenetic tree of the species used for the construction of orthogroups and gene trees in this chapter. The yellow stars indicate branches of

the species tree associated with an origin of C_4 photosynthesis. The C_4 species in the tree are indicated and the C_4 subtype shown in the parentheses. N17 corresponds to the Andropogoneae lineage and N18 corresponds to the Panicae lineage. **Abbreviations**: NADP-dependent malic enzyme (NADP-ME), NAD-dependent malic enzyme (NAD-ME), phosphoenolpyruvate carboxykinase (PEP-CK).

3.2.2 Species-tree branches encompassing an origin of C_4 are associated with an elevated rate of change in protein targeting to the chloroplast and secretory pathway

To investigate whether changes in protein targeting occurred at a higher rate on branches that encompass a C_4 origin, the number of changes in protein targeting mapped to each branch of the species tree was compared. As each branch in the species tree represents varying amounts of evolutionary distance, the number of changes along each branch was normalised by the length of the branch. Comparison of this normalised number of changes across the tree revealed that the rate of change in protein targeting has been steady (within one order of magnitude) throughout the evolution of angiosperms (Figure 3.2 A). This is concomitant with the finding in Chapter 2 that the rate of change in protein targeting has been constant throughout plant evolution. Nonetheless, the species-tree branch encompassing an origin of C_4 photosynthesis in the Paniceae (N18) was associated with more changes in protein targeting per unit molecular distance than any other branch in the species tree (Figure 3.2 A). Indeed, it was found that the branches of the species tree encompassing C_4 origins (N14, N17, and N18, collectively 'C_4 branches') were, on average, associated with more change in protein targeting than other species-tree branches in the angiosperm clade (two-tailed t-test, $P < 0.05$). To investigate this further, the amount of change along each branch of the species tree was broken down by gains and losses in protein targeting to each organelle. This revealed that C_4 branches of the species tree are associated with a higher rate of chloroplast transit peptide loss, and signal peptide gain and loss (Figure 3.2 B). In this analysis the ancestral branch to the two origins

of C_4 photosynthesis (N14) was included as a 'C_4 branch' for the reason described in the previous section. If branch N14 is not considered in the analysis, then only chloroplast losses were found to occur at a statistically higher rate for the C_4 branches.

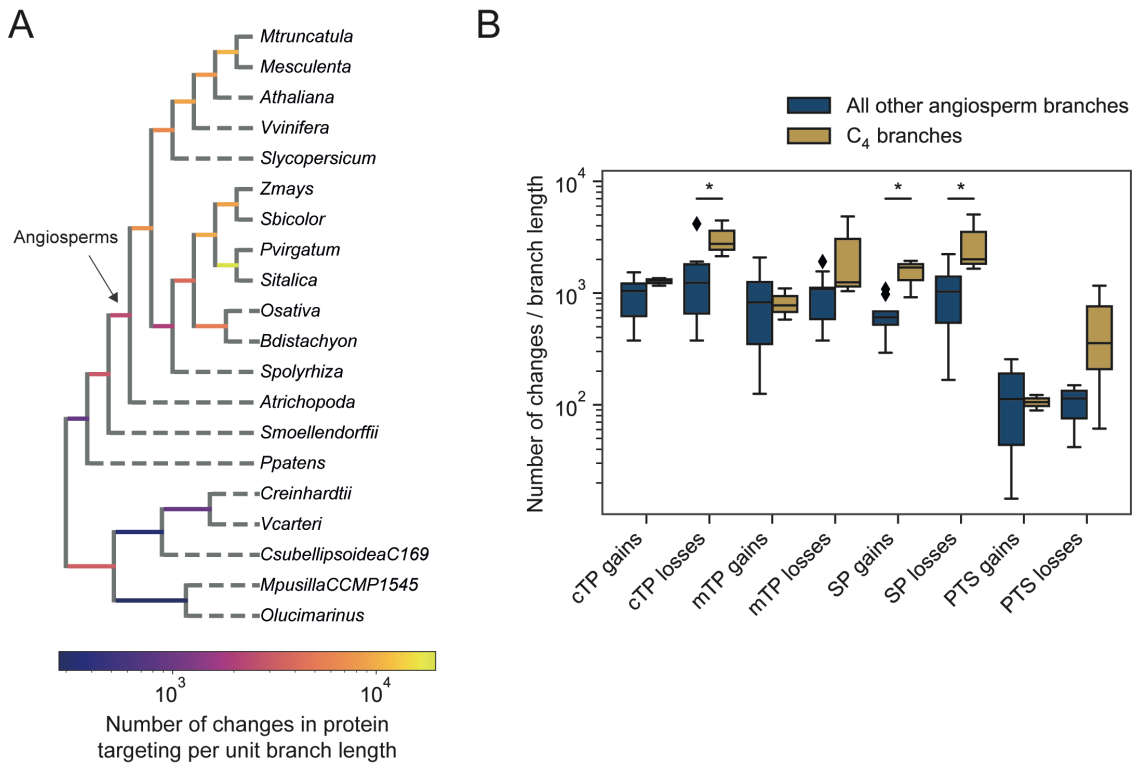


Figure 3.2. Comparison of the number of changes in protein targeting along each branch of the species tree as a function of phylogenetic distance. A) The number of changes in protein targeting along each branch of the species tree, normalised by the length of the branch (amount of molecular sequence evolution). This normalised amount of change for each branch is shown on a log scale. Note here that branch lengths in the tree do not represent phylogenetic distance. **B)** A comparison of the number of changes in protein targeting per unit branch length is broken down by the type of target signal change and compared between branches associated with the evolution of C_4 photosynthesis (N14, N17 and N18 in Figure 3.1) and all other branches in the angiosperm clade. Asterisk indicates a significant difference (two-tailed t-test, $P < 0.05$). **Abbreviations**: chloroplast transit peptide (cTP), mitochondrial target peptide (mTP), signal peptide (SP), peroxisomal target signal (PTS).

3.2.3 No ancestral changes in protein targeting were identified among known C_4-cycle proteins coincident with the evolution of C_4 photosynthesis

Among the changes in protein targeting identified concomitant with the evolution of C_4 photosynthesis in this dataset, none were found to have occurred in the ancestor of proteins known to be directly involved in the C_4 pathway in any of the four C_4 species. A corollary of this is that no change in protein targeting was identified for a carbonic anhydrase (CA) concomitant with the evolution of C_4 photosynthesis in this data set. Manual inspection of the orthogroup gene tree for CA, along with the predicted targeting of extant CA proteins, revealed that the mesophyll-cell expressed CA isoforms in the four C_4 species form a monophyletic clade (Figure 3.3). Of these proteins, the most highly expressed isoform in *Z. mays*, *S. bicolor* and *S. italica* had no target signal prediction in agreement with the cytosolic function of CA in the C_4 pathway. In *P. virgatum* on the other hand, the two *CA* isoforms with highest mRNA expression were predicted to contain a chloroplast transit peptide (Figure 3.3). The localisation of these CA proteins has not been experimentally confirmed in *P. virgatum,* but it is assumed that at least one isoform here has been mis-predicted given that across C_4 species described so far CA functions in mesophyll cell cytosol to convert carbon dioxide to bicarbonate (Ludwig 2016). The gene tree in Figure 3.3 shows that the highly expressed C_4 *CA* genes are found within a wider clade of C_3 and C_4 monocot genes, the majority of which are predicted to encode a cytosolic protein. This suggests that the ancestral CA protein for this clade was cytosolic and thus an already cytosolic CA was recruited to the C_4 pathway in *Z. mays*, *S. bicolor*, *S. italica*, and *P. virgatum*.

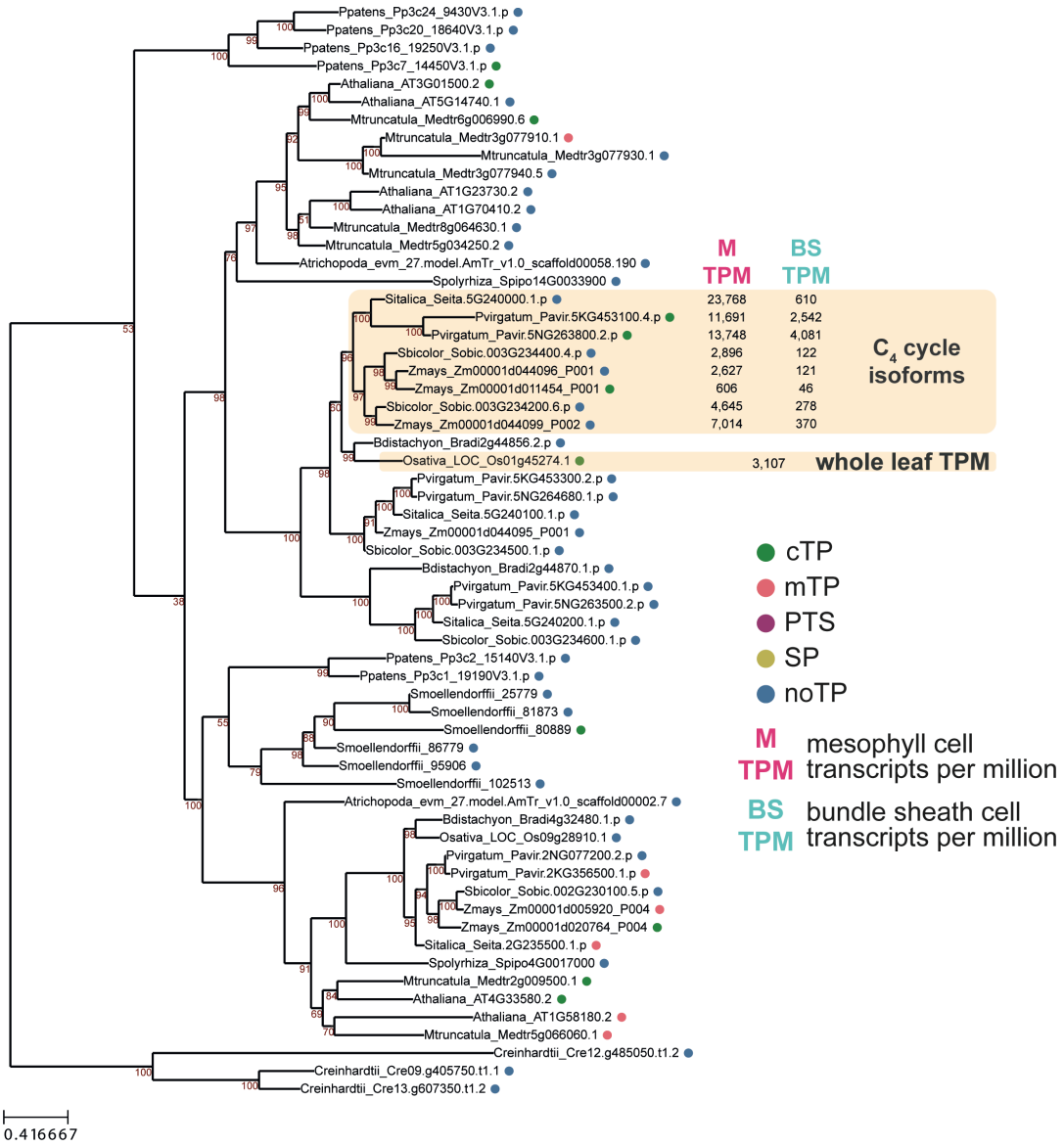


Figure 3.3. Maximum likelihood phylogenetic tree for the beta-carbonic anhydrase (CA) orthogroup containing the C_4-cycle CA isoforms. The tree was constructed with IQ-TREE using the substitution model WAG+I+G4, and bootstrap values are from UFBoot. The scale bar represents amino acid substitutions per site. Previously published transcriptome data sets were used to investigate CA transcript abundance in the C_4 species *Z. mays* (Chang et al. 2012), *S. bicolor* (Emms et al. 2016), *S. italica* (John et al. 2014), and *P. virgatum* (Rao et al. 2016). Mesophyll and bundle sheath cell expression (transcripts per million, TPM) are shown only for transcripts with high mesophyll cell expression (>500 TPM). Whole leaf TPM is also shown for the ortholog of these C_4-cycle isoforms in the C_3 species *O. sativa* (van Campen et al. 2016). **Abbreviations**: chloroplast transit peptide (cTP), mitochondrial target peptide (mTP), signal peptide (SP), peroxisomal target signal (PTS), no predicted target signal (noTP). Scale bar represents amino acid substitutions per site.

3.2.4 Incorporation of transcriptomic data sets to identify changes in protein targeting relevant to the evolution of C_4 photosynthesis

Genes which encode proteins that function in the C_4 pathway are often characterised by preferential accumulation of their corresponding transcripts in either bundle sheath or mesophyll cells (for example: John et al. 2014; Emms et al. 2016; Rao et al. 2016; Denton et al. 2017). While only some of these differentially expressed genes encode proteins that function directly in the C_4 photosynthetic pathway, many are presumed to provide the correct metabolic and cellular context to enable the C_4 cycle to function efficiently. Thus, to identify other changes in protein targeting that might have had functional consequences for the evolution of C_4 photosynthesis, previously published transcriptomic data sets were leveraged to identify the set of changes in targeting that occurred during C_4 evolution in the ancestor of extant genes with differential expression in mesophyll and bundle sheath cells. In this analysis, mesophyll and bundle sheath cell specific transcriptomic data sets were used for *Z. mays*, *S. bicolor*, *S.viridis*, and *P. virgatum* (details about these data sets can be found in Appendix 3.2). Cell-type enrichment of gene expression was conducted by differential expression analysis ($P < 0.05$) and only those genes expressed in the leaf (whole-leaf expression > 10 transcripts per million) were considered in further analyses. Using this transcript data, the set of gene-tree branches that represent a change in protein targeting and which are subtended by a mesophyll or bundle sheath cell enriched gene were identified. In total, 62, 17, and 80 changes in protein targeting mapped to branches N16, N17, and N14 of the species tree, respectively, were found to meet this criterion. A breakdown of the number of gains and losses identified to each organelle for this set of changes is shown in in Appendix Table A3.1, column B.

Among the proteins identified to have changed targeting concomitant with C_4 evolution, and in the ancestor of a mesophyll or bundle sheath cell enriched gene, few had any known or previously hypothesised C_4-related function. Indeed, the proteins that were identified in this set possessed a wide range of biological functions with no clear functional

enrichment. For example, of the proteins lost to the chloroplast there was gene ontology (GO) term enrichment for vitamin binding, endonuclease activity, coenzyme binding, cofactor binding and hydrolase activity on ester bonds. On the other hand, among proteins gained to the chloroplast there was no enrichment for any GO term. Similarly, no GO term enrichment was identified among proteins gained to the mitochondrion. Among proteins lost to the mitochondrion however there was enrichment for the following three catabolic processes: glycine catabolic process, neurotransmitter catabolic process, and α–amino acid catabolic processes. Of particular note was the identification of a loss of chloroplast targeting among four different pentatricopeptide repeat (PPR) proteins. In plants, PPR proteins function almost exclusively as RNA-binding proteins for the modulation of gene expression in the chloroplast and mitochondrion, although some PPR proteins are also known to function in the nucleus (Gutmann et al. 2020). In one of these PPR protein families a loss of chloroplast targeting was identified in the ancestor of the four C_4 species. Interestingly, this ancestral protein which had lost a chloroplast transit peptide is orthologous to the *Arabidopsis* protein DEFECTIVELY ORGANISED TRIBUTARIES 4 (DOT4). Molecular and transgenic work in *Arabidopsis* have established a function for DOT4 (AT4G18750) as a chloroplast-localised C-to-U RNA editing factor involved in regulating leaf vasculature patterning (Petricka et al. 2008; Hayes et al. 2013; Hein et al. 2019). In the context of C_4 evolution, a change in targeting of this protein might be of interest given that vascular patterning is substantially altered in C_4 species (Sedelnikova et al. 2018). The function of DOT4 proteins in the C_4 species used in this study are however unknown. It is also interesting to note that, in *Arabidopsis,* DOT4 activity is specific for a C-to-U editing site (C488) in the chloroplast RNA polymerase gene rpoC1, however it has been reported that among Poaceae (grass) species this editing site has been lost from rpoC1 due to a C-to-T conversion in the gene (Hein et al. 2019). Thus, it is unknown if DOT4 will similarly function in vascular patterning in C_3 and C_4 grasses and perhaps a loss of chloroplast localisation is associated with a new function or specificity for this gene. The DOT4 gene tree and expression data are provided in Appendix 3.3 (Figure A3.3.1).

As discussed previously, among the other proteins identified in this analysis there was a plethora of different protein functions represented, but few of these proteins could be readily attributed as having a C_4-related function based on prior knowledge of C_4 photosynthesis. Thus, a second screen was conducted to identify ancestral changes in targeting among proteins with higher leaf expression. The results from this screen will be discussed in the following section (3.2.5). In addition, a change in peroxisomal targeting was identified for an aspartate aminotransferase (AspAT). This change did not occur in the ancestor of the paralog that was subsequently recruited to the C_4 cycle in any of the four C_4 species, but examination of the gene expression data for aspartate aminotransferase highlighted a potential missing link in our current understanding of C_4 photosynthesis in maize. This will be discussed in a subsequent result section (3.2.7).

3.2.5 Investigation of ancestral changes in protein targeting among extant genes with high leaf-expression

To investigate the ancestral changes in organellar protein targeting that occurred among proteins encoded by highly expressed genes in the C_4 species in this data set, a whole-leaf expression threshold of ≥500 transcripts per million (TPM) was employed. This threshold was chosen as it is expected to encompass any enzymes or transporters functioning in the C_4 cycle (the lowest expressed enzyme of the core C_4 cycle in maize for example is NADP-ME at 2,987 TPM). Only six changes in protein targeting were identified in this screen, however after manual inspection of target signal prediction, two of these changes were dismissed as false positives based on previously published experimental localisation data. The remaining four changes (Figure 3.4 A) were found to have occurred to proteins with no previous functional annotation in any C_4 cycle. Indeed, three of these proteins – ROOT UVB SENSITIVE (RUS6), TETRATRICOPEPTIDE REPEAT-like SUPERFAMILY PROTEIN (TPR-like) and TonB-DEPENDENT HEME RECEPTOR A (TDHA) – have no previous functional characterisation in plants, but interrogation of gene expression data showed them to be expressed to an appreciable level in at least two of

the C_4 species. The fourth protein was a beta-glucosidase (BGLU15) and will be discussed in more detail in the next section (3.2.6). Additional transcriptomic data sets were employed in this analysis to compare C_4 gene expression with that in the C_3 leaf of *Arabidopsis thaliana* and *Oryza sativa* (rice).

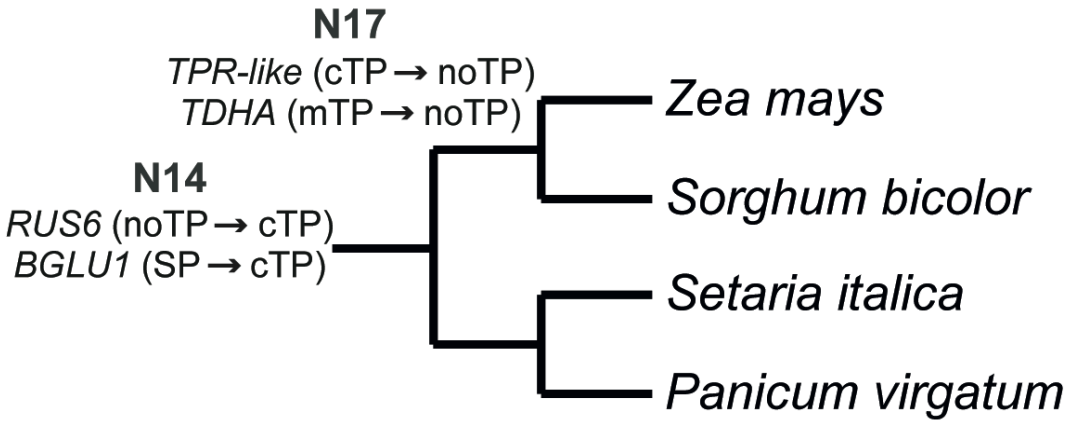


Figure 3.4. Identification of the changes in protein targeting that occurred during the evolution of C_4 photosynthesis and in the ancestor of highly expressed genes. In total 211, 101, and 145 changes in targeting were identified to branches N14, N18, and N17 of the species tree, respectively. Of these, four changes in protein targeting (two mapped to N14 and two mapped to N17) were identified as occurring in the ancestor of at least one gene with high whole-leaf expression (≥500 TPM). The proteins are TETRATRICOPEPTIDE REPEAT-like SUPERFAMILY PROTEIN (TPR-like), TonB-DEPENDENT HEME RECEPTOR A (TDHA), ROOT UVB SENSITIVE (RUS6), and BETA-GLUCOSIDASE 1 (BGLU1). The direction of change in each case is indicated in parentheses. **Abbreviations**: chloroplast transit peptide (cTP), mitochondrial target peptide (mTP), signal peptide (SP), no N-terminal target signal identified (noTP).

RUS6 proteins contain a domain of unknown function (DUF647). While there has been no previous characterisation of any RUS6 protein, other homologous RUS proteins (specifically, RUS1 and RUS2) have been implicated in a developmental response to ultraviolet B (UV-B) light; *rus2-1 Arabidopsis* seedling mutants exhibit a developmental block when grown under non-photodamaging levels of UV-B light and the function of these proteins appears to be dependent on a protein-protein interaction between RUS1 and RUS2 (Leasure et al. 2009). In the data set presented here, a RUS6 paralog was identified as having gained a chloroplast transit peptide in the ancestor of the four C_4 species used in this study (Figure 3.5 A). This chloroplast-targeted RUS6 paralog was found to be upregulated in the C_4 leaf compared to the non-chloroplast-targeted paralog (Figure 3.5 B).

For the TonB-DEPENDENT HEME RECEPTOR A (TDHA), a loss of mitochondrial targeting was identified for a TDHA paralog in the ancestor of maize and *Sorghum* (Figure 3.6 A). Similar to RUS6, this non-mitochondrial targeted paralog was then upregulated in the leaf of these two C_4 species compared to the mitochondrial paralogs (Figure 3.6 B). There is little information about this protein in plants but TonB-receptors in bacteria are well characterised outer-membrane receptors that assist in the active transport of organic iron chelators, such as heme, across the membrane and into the periplasmic space (Noinaj et al. 2010). Concomitant with this, the plant TDHA proteins in this orthogroup are predicted to contain a single transmembrane domain suggesting they may also function in the transport of heme, or other organic molecules, across membranes. It is currently not clear however what the functional significance is of a loss of mitochondrial targeting for the two TDHA proteins in maize and *Sorghum*.

Lastly, a loss of chloroplast targeting was identified in the ancestor of a highly expressed tetratricopeptide repeat like (TPR-like) protein in maize and *Sorghum bicolor* (Figure 3.7 A). In this case, only a single copy of this gene was found to be present in each species such that maize and *S.bicolor* possess a single non-chloroplast targeted isoform of this

protein while *S. viridis* and *P. virgatum* possess a single chloroplast targeted isoform (Figure 3.7 B). Proteins containing TPRs typically mediate protein-protein interactions, and a vast array of proteins have been identified which contain this structural motif (Sharma and Pandey 2016). The TPR-like protein identified in this study is encoded by a gene with moderate expression in each of the NADP-ME subtype species (maize, *S. bicolor*, *S. viridis*) (Figure 3.7 B). No functional annotation is available for members of this TPR-like family however and so it is not clear what the functional consequence is of the predicted change in localisation.

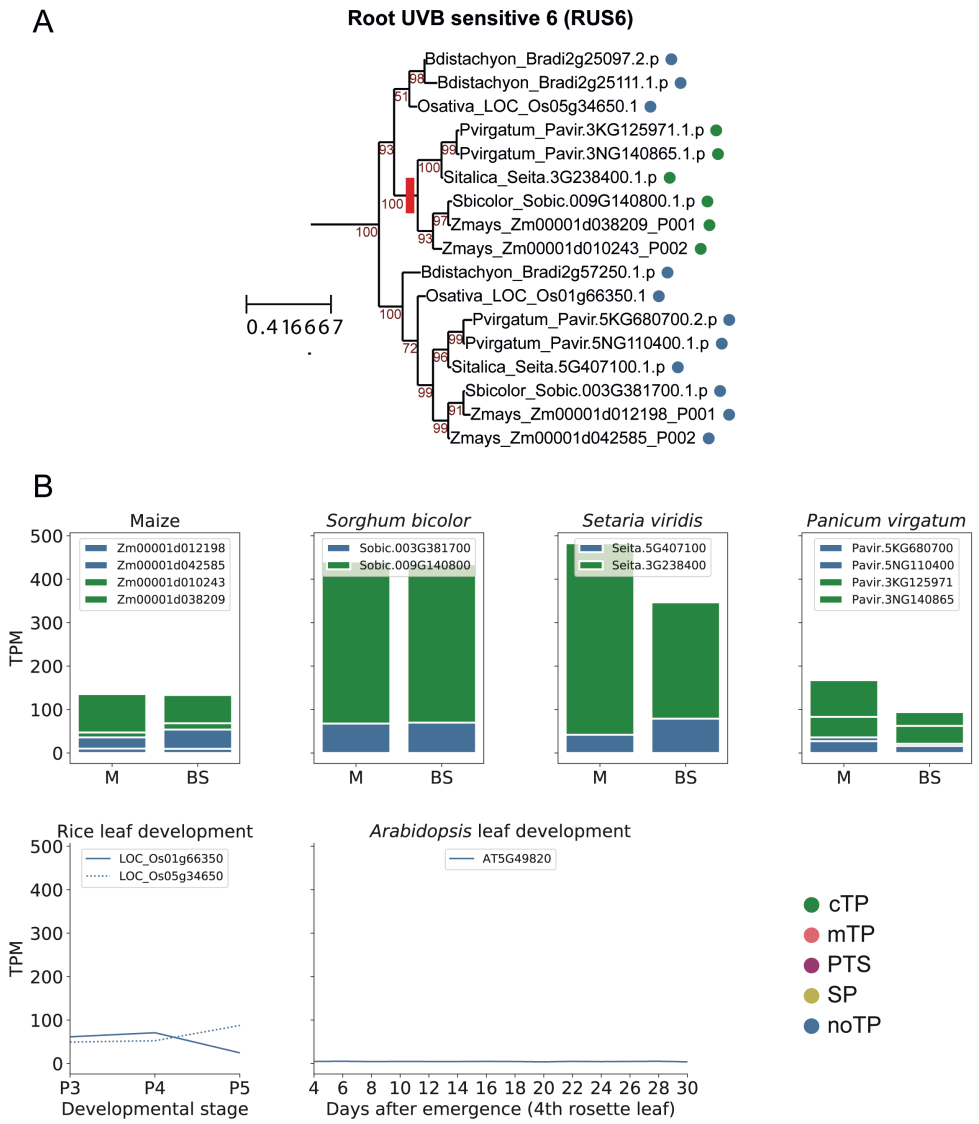


Figure 3.5. Identification of a gain in chloroplast targeting for a RUS6 protein in an ancestor of maize, *Sorghum bicolor*, *Setaria italica,* and *Panicum virgatum*. A) Maximum-likelihood phylogenetic tree of the RUS6 gene family showing the monocot clade containing the identified change in protein targeting. The branch along which a change was estimated to have occurred is shown by the red bar. The tree was constructed with IQ-TREE using the substitution model JTT+G4, and bootstrap values are from UFBoot. The scale bar represents amino acid substitutions per site. The full gene tree can be found in Appendix 3.3. **B)** Expression profile of *RUS6* in selected data sets: mesophyll (M) and bundle sheath (BS) cell expression is shown for *Zea mays* (maize), *Sorghum bicolor*, *Setaria viridis*, and *Panicum virgatum*. Expression along a leaf developmental gradient is shown for C_3 species *Oryza sativa* (rice) and *Arabidopsis thaliana*. **Abbreviations**: chloroplast transit peptide (cTP), mitochondrial target peptide (mTP), signal peptide (SP), peroxisomal target signal (PTS), plastochron stage 3-5 (P3-P5), transcripts per million (TPM).

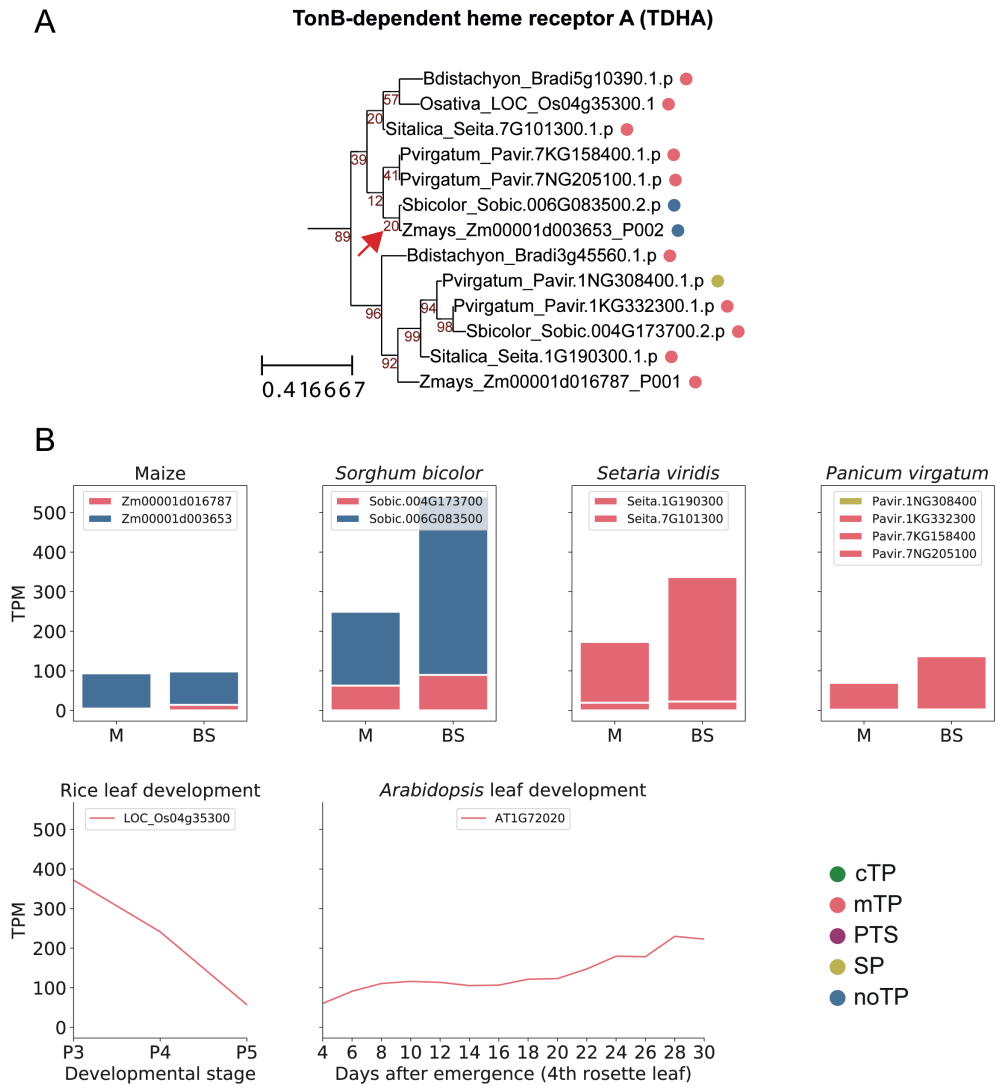


Figure 3.6. Identification of a loss in mitochondrial targeting for a TDHA protein in an ancestor of maize and *Sorghum bicolor*. A) Maximum-likelihood phylogenetic tree of the TDHA gene family showing the monocot clade containing the identified change in protein targeting. The branch along which a change was estimated to have occurred is shown by the red arrow. The tree was constructed with IQ-TREE using the substitution model LG+G4, and bootstrap values are from UFBoot. The scale bar represents amino acid substitutions per site. The full gene tree can be found in Appendix 3.3. **B)** Expression profile of *TDHA* in selected data sets: mesophyll (M) and bundle sheath (BS) cell expression is shown for *Zea mays* (maize), *Sorghum bicolor*, *Setaria viridis*, and *Panicum virgatum*. Expression along a leaf developmental gradient is shown for C_3 species *Oryza sativa* (rice) and *Arabidopsis thaliana*. **Abbreviations**: chloroplast transit peptide (cTP), mitochondrial target peptide (mTP), signal peptide (SP), peroxisomal target signal (PTS), plastochron stage 3-5 (P3-P5), transcripts per million (TPM).

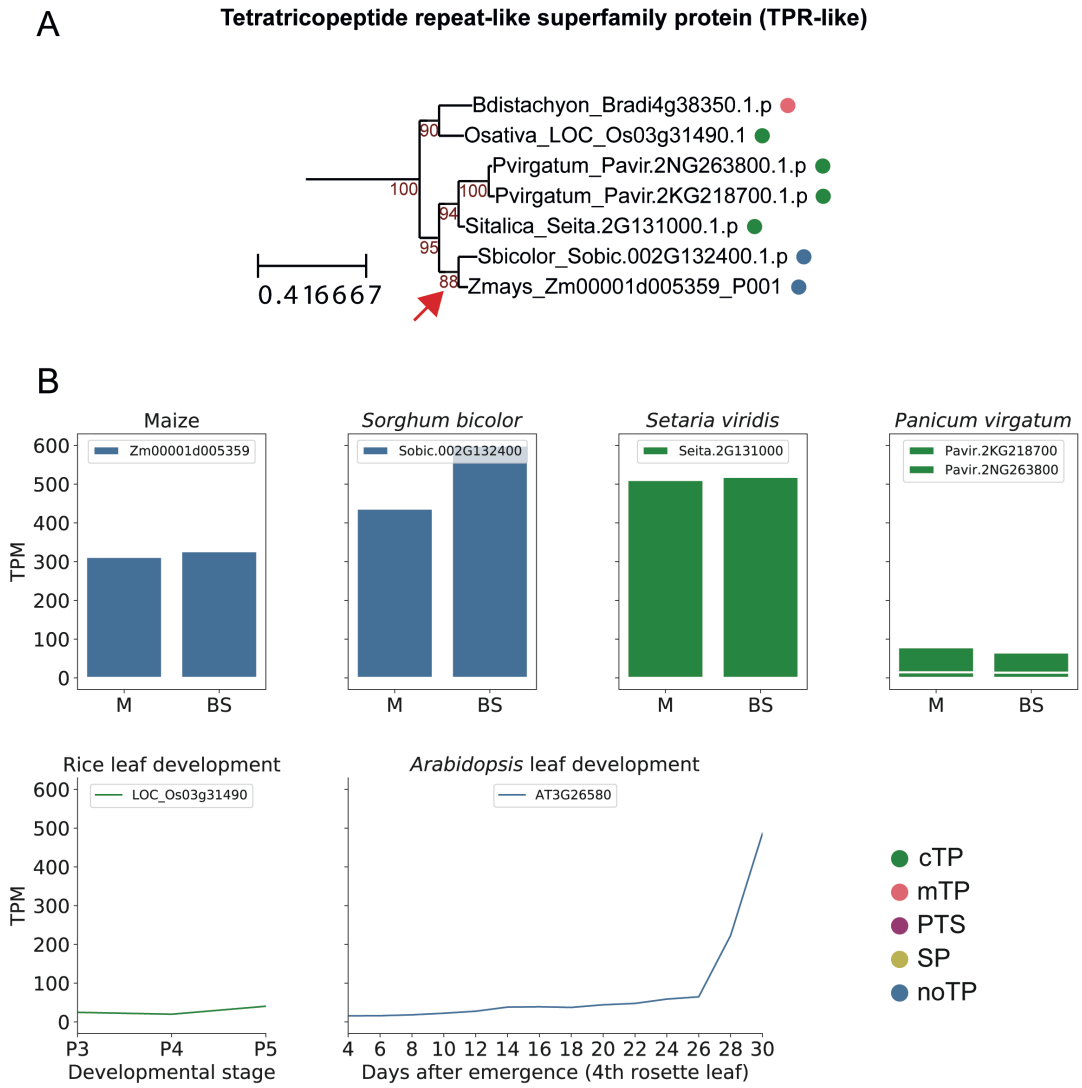


Figure 3.7. Identification of a loss in chloroplast targeting for a TPR-like protein in an ancestor of maize and *Sorghum bicolor*. A) Maximum-likelihood phylogenetic tree of the TPR-like gene family showing the monocot clade containing the identified change in protein targeting. The branch along which a change was estimated to have occurred is shown by the red arrow. The tree was constructed with IQ-TREE using the substitution model JTTDCMut+I+G4, and bootstrap values are from UFBoot. The scale bar represents amino acid substitutions per site. The full gene tree can be found in Appendix 3.3. **B)** Expression profile of *TPR-like* in selected data sets: mesophyll (M) and bundle sheath (BS) cell expression is shown for *Zea mays* (maize), *Sorghum bicolor*, *Setaria viridis*, and *Panicum virgatum*. Expression along a leaf developmental gradient is shown for C_3 species *Oryza sativa* (rice) and *Arabidopsis thaliana*. **Abbreviations**: chloroplast transit peptide (cTP), mitochondrial target peptide (mTP), signal peptide (SP), peroxisomal target signal (PTS), plastochron stage 3-5 (P3-P5), transcripts per million (TPM).

3.2.6 A putative C_4-related change in protein targeting was identified for a beta-glucosidase involved in cytokinin signalling in maize

The fourth change in protein targeting identified among the highly expressed genes was for a beta-glucosidase in the ancestor of the four C_4 species (maize, *Sorghum bicolor*, *Setaria viridis*, and *Panicum virgatum*) in this data set. Beta-glucosidases are a large and diverse group of enzymes found in all domains of life (Ketudat Cairns and Esen 2010). The catalytic activity of beta-glucosidases is in the hydrolysis of beta-glycosidic bonds present in glycosides and oligosaccharides (Ketudat Cairns and Esen 2010). Upon hydrolysis of these bonds, glucose in released as well as the functional group or compound to which the glucose was bound. In plants, beta-glucosidases have been implicated in a wide range of biological functions that are reflective of the diversity of substrates on which these enzymes act. Such functions include the release of chemical defences to biotic stress, the formation of intermediates in cell wall lignification, cell wall catabolism, and the activation of phytohormones (reviewed in Morant et al. 2008).

In the analysis presented in this chapter, an ancestral change in protein targeting from the secretory pathway to the chloroplast was identified for a beta-glucosidase (Figure 3.8) that in maize has been shown to catalyse the hydrolysis of cytokinin-O-glucosides; biologically inactive conjugates of cytokinin. Upon hydrolysis, free and active cytokinin is released from the conjugation which can be used directly by the plant (Brzobohatý et al. 1993). Although it is well described in plants that cytokinin-O-glucosides are an important and abundant storage form of cytokinin – in the maize leaf >80% of cytokinin in the leaf has been estimated to occur in the O-glucoside form (Gajdošová et al. 2011; Hošek et al. 2020) – to date only one gene has been identified as encoding a beta-glucosidase with substrate activity for cytokinin-O-glucosides. This gene is the maize beta-glucosidase *Zm*BGLU1 (also known as Zm-p60.1) (Brzobohatý et al. 1993). No other plant beta-glucosidase has so far been identified as capable of catalysing this reaction and releasing biologically active cytokinin from cytokinin-glucose conjugates.

The maize beta-glucosidase *Zm*BGLU1 (Zm00001d023994) is descendent from the identified change in protein targeting shown in Figure 3.8. In addition, several other paralogous proteins from maize, *Sorghum bicolor* and *Panicum virgatum* are also descendant from this gain in chloroplast targeting. A chloroplast localisation has been confirmed for at the maize paralog with a characterised function (*Zm*BGLU1, Kiran et al. 2006) and in addition several of the other maize paralogs in this clade have a chloroplast-localised annotation in the plant proteome database (Sun et al. 2009). From the gene tree in Figure 3.8, a change in protein targeting from the secretory pathway to the chloroplast is estimated to have occurred in the ancestor of the four C_4 species. The acquisition of a chloroplast transit peptide by this protein may be associated with a divergence in function: while the chloroplast-targeted beta-glucosidase in maize (*Zm*BGLU1) has been shown to have substrate specificity for cytokinin-glucosides (Brzobohatý et al. 1993; Kiran et al. 2006), one of the *Arabidopsis* orthologs of this protein (AT2G44450, Figure 3.8) has been shown to instead have specific activity for the hydrolysis of flavonol glucosides (Roepke et al. 2017). In addition, this *Arabidopsis* protein has been shown to localise to the apoplast (Borderies et al. 2003; Roepke and Bozzo 2015; Roepke et al. 2017), thus confirming the prediction that this protein is targeted to the secretory pathway.

To investigate how a change in protein targeting from the apoplast to the chloroplast might have arisen for this beta-glucosidase, the N-terminal region of the predicted chloroplast- and secretory-targeted beta-glucosidases shown in Figure 3.8 were compared. Analysis of the target signal cleavage site predicted by TargetP revealed that the predicted chloroplast transit peptides were on average approximately 30 amino acids longer than the predicted signal peptides (Figure 3.8). Comparison of the multiple sequence alignment for these proteins showed that this difference in target signal length was mostly due to an insertion of approximately 30 amino acids in the N-terminal region of the predicted chloroplast-targeted proteins. Representative sequences from this alignment are shown in Figure 3.8. Thus, the predicted change in targeting of an ancestral beta-glucosidase from

the secretory pathway to the chloroplast appears to have involved an insertion in, and thus an extension of, the N-terminal gene region.

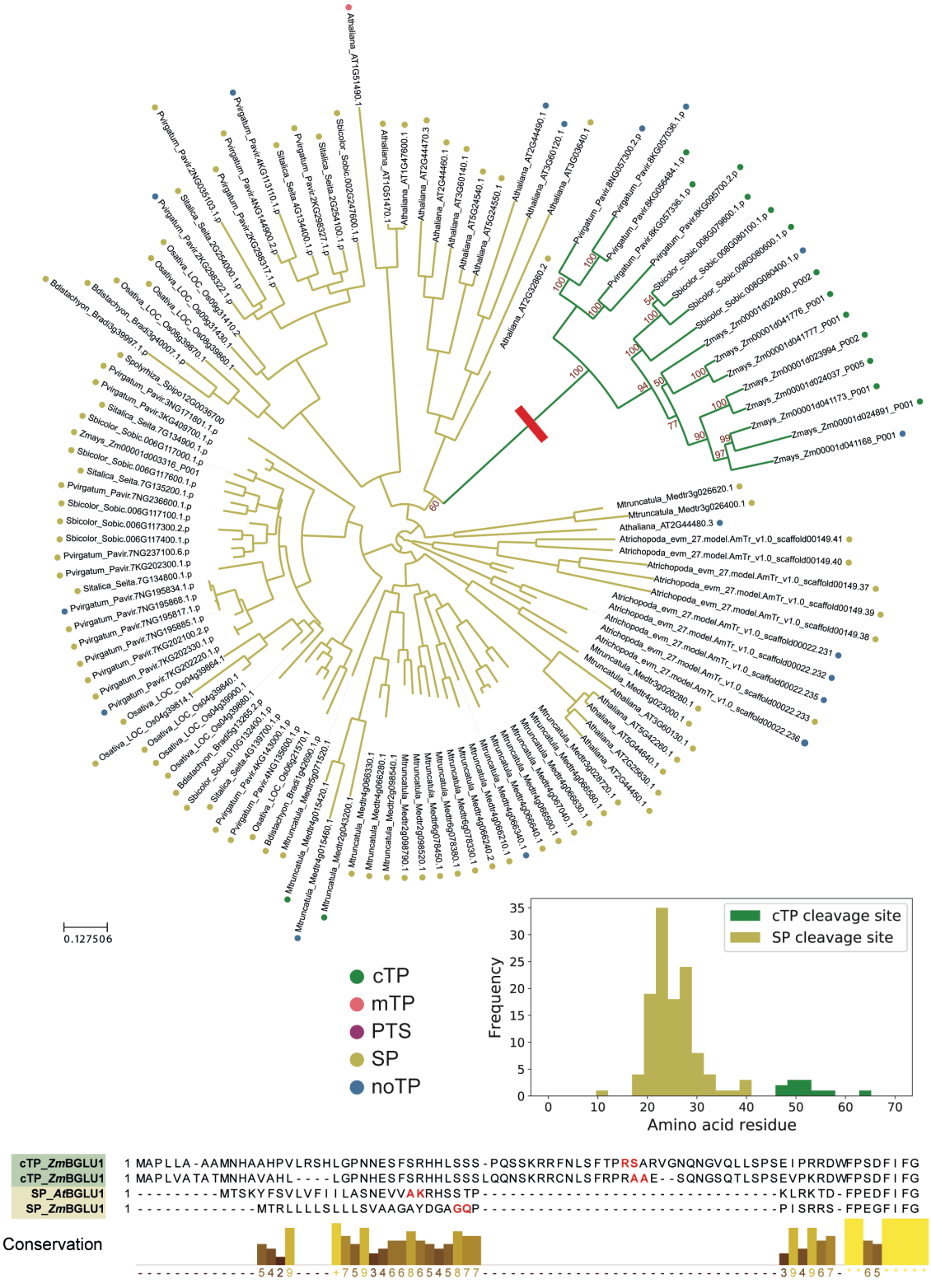


Figure 3.8. Maximum likelihood phylogenetic tree of the beta-glucosidase gene family containing *ZmBGLU1*. The red bar indicates the branch along which a change in protein targeting was identified to have occurred. The coloured circles next to gene names denote the target signal prediction: chloroplast transit peptide (cTP), mitochondrial target peptide (mTP), peroxisomal target signal (PTS), signal peptide

(SP), no N-terminal target signal identified (noTP). The branches are coloured to help differentiate between the clade of mostly chloroplast-targeted proteins and the clade of mostly secretory-targeted proteins. The tree was constructed with IQ-TREE using the substitution model JTTDCMut+I+G4, and bootstrap values (shown for the clade of interest) are from UFBoot. The scale bar represents amino acid substitutions per site. The histogram compares the predicted target signal cleavage site (amino acid residue) for the chloroplast BGLUs descendent from the change in protein targeting and the secretory BGLUs in the rest of the tree. Target signal cleavage site was predicted by TargetP v2.0. The multiple sequence alignment shows the N-terminal region of two representative amino acid sequences with a cTP (cTP_*Zm*BGLU1, Zm00001d023037 and Zm00001d041173) and two representative, and orthologous, amino acid sequences with an SP (SP_*At*BGLU1, AT2G25630 and SP_*Zm*BGLU1, Zm00001d003316). The red residues indicate the predicted target signal cleavage site. The histogram underneath the multiple sequence alignment shows the conservation of the physico-chemical properties for each column of the alignment (see materials and methods, section 3.4.1 for details).

The transcript for one of the chloroplast-targeted BGLU1 proteins descendant from the gain in chloroplast targeting identified for in this family was found to be abundant in the bundle sheath cells of the mature maize leaf (Figure 3.9 A). This single transcript belongs to the gene Zm00001d02403 which is also the most closely related paralog to *ZmBGLU1* (Zm00001d023994), the previously characterised BGLU1 with cytokinin-O-glucoside specific activity. Negligible expression of any transcript from this beta-glucosidase gene family was found in the mesophyll or bundle sheath cells of *Sorghum bicolor*, *Setaria viridis,* or *Panicum virgatum* (Figure 3.9 B-D). Given the demonstrated function of *ZmBGLU1* in the release of active cytokinin from inactive cytokinin-conjugates, and the function cytokinin has in plant development, cell division and chloroplast biogenesis, *BGLU1* gene expression was investigated in two further transcriptomic data sets of the maize leaf along a developmental gradient. The first of these data sets compares

mesophyll and bundle sheath cell gene expression at three points along the developmental gradient of the maize leaf. Analysis of transcript abundance in this data set again confirmed that expression of this beta-glucosidase is restricted to the bundle sheath cells of the leaf and that expression is made up by the same single transcript (Figure 3.9 E). In addition, expression of the bundle sheath transcript was found to be highest in the leaf section closest to the ligule (Figure 3.9 E). Leaf tissue from this region is considered to be photosynthetically immature and is expected to be a sink rather than a source of photosynthetic metabolites. This is illustrated by the fact that previous analysis of transcripts expressed in this region of the leaf blade shows an upregulation of transcripts from genes which function in chloroplast biogenesis and development (Tausta et al. 2014).

Comparison of beta-glucosidase gene expression in a second maize leaf developmental data set showed that earlier in leaf development there is expression of several of the other chloroplast targeted *BGLU1* genes (Figure 3.9 F). In this data set, transcript expression is compared between the maize husk and foliar leaves. These two leaf types have a distinctly different pattern of leaf venation: in the maize foliar leaf, veins are spaced close together thus ensuring the typical Kranz anatomy arrangement of mesophyll and bundle sheath cells, while in the husk leaf veins are spaced further apart leading to a mesophyll and bundle sheath cell arrangement more similar to that found in a C_3 monocot leaf (Wang, Kelly, et al. 2013). Beta-glucosidase expression was found to peak in both the foliar and husk leaf at the stage of development just prior to the emergence of the leaf blade from the sheath of the preceding leaf (Figure 3.9 F, plastochron 5 or P5). This suggests that the chloroplast-targeted beta-glucosidase isoform may function in the development of both the husk and foliar leaf in maize. It should be noted however that the relative abundance of beta-glucosidase transcripts at P5 is different between the husk and foliar leaf, and that in both cases the most highly expressed transcript at this stage is not the same transcript that is highly expressed in the mature leaf blade (Figure 3.9 F). Inferences about the cell-type specificity of the transcript expressed at P5 can therefore not be made. Expression

of the bundle-sheath specific transcript expressed in the mature leaf (shown in Figure 3.9 E) does however become apparent later in the development of the foliar leaf, where it peaks at the immature foliar leaf stage (FI, Figure 3.9 F). Negligible expression of this transcript was observed in the husk leaf.

The expression of a bundle sheath cell specific beta-glucosidase which may function to liberate cytokinin in the developing maize leaf is interesting given that cytokinin is a positive regulator of chloroplast biogenesis and one of the hallmark features of the C_4 leaf is an abundance of chloroplasts in the bundle sheath cells. Indeed, the concentration of chloroplasts in a typical C_4 leaf far exceeds that found in a typical C_3 leaf concomitant with the restriction of RuBisCo and other Calvin cycle enzymes to the bundle sheath cells of C_4 leaves (Leegood 2008). It is well established that cytokinin stimulates chloroplast development, and several cytokinin responsive transcription factors such as CYTOKININ RESPONSIVE GATA FACTOR 1 (CGA1) have been shown to promote and regulate chloroplast biogenesis (Chiang et al. 2012; Hudson et al. 2013; Cortleven and Schmülling 2015; Wang, Hendron, et al. 2017). In maize specifically, the transcription factors GOLDEN2 and GOLDEN2-LIKE (together GLK genes) have been identified to promote chloroplast development in a cell-type specific manner (Rossini et al. 2001; Wang, Fouracre, et al. 2013) and in *Arabidopsis* these GLK genes have been shown to function as part of a cytokinin signalling pathway (Kobayashi et al. 2012). Thus, in maize at least, the activity of this bundle-sheath-cell-specific, chloroplast-targeted, beta-glucosidase may positively regulate a cell-type specific cytokinin signalling pathway in the bundle sheath cells. It is interesting to note here that there was very little expression of the bundle-sheath-cell-specific beta-glucosidase transcript in the developing husk leaf and accordingly it has been previously reported that chloroplast density is much lower in the husk leaf bundle sheath cells compared to those of the foliar leaf (Pengelly et al. 2011; Wang, Kelly, et al. 2013). Thus, the beta-glucosidases identified in this study present a promising candidate for future work aiming to understand how chloroplast development is regulated in the

maize C_4 leaf. The change in targeting of this protein during the evolution of C_4 photosynthesis may also be associated with the neofunctionalisation of this protein given the differing substrate specificities of the maize and *Arabidopsis* orthologs and the fact that no gene expression was identified in the developing leaf of *Arabidopsis* and rice (Figure 3.9 G&H).

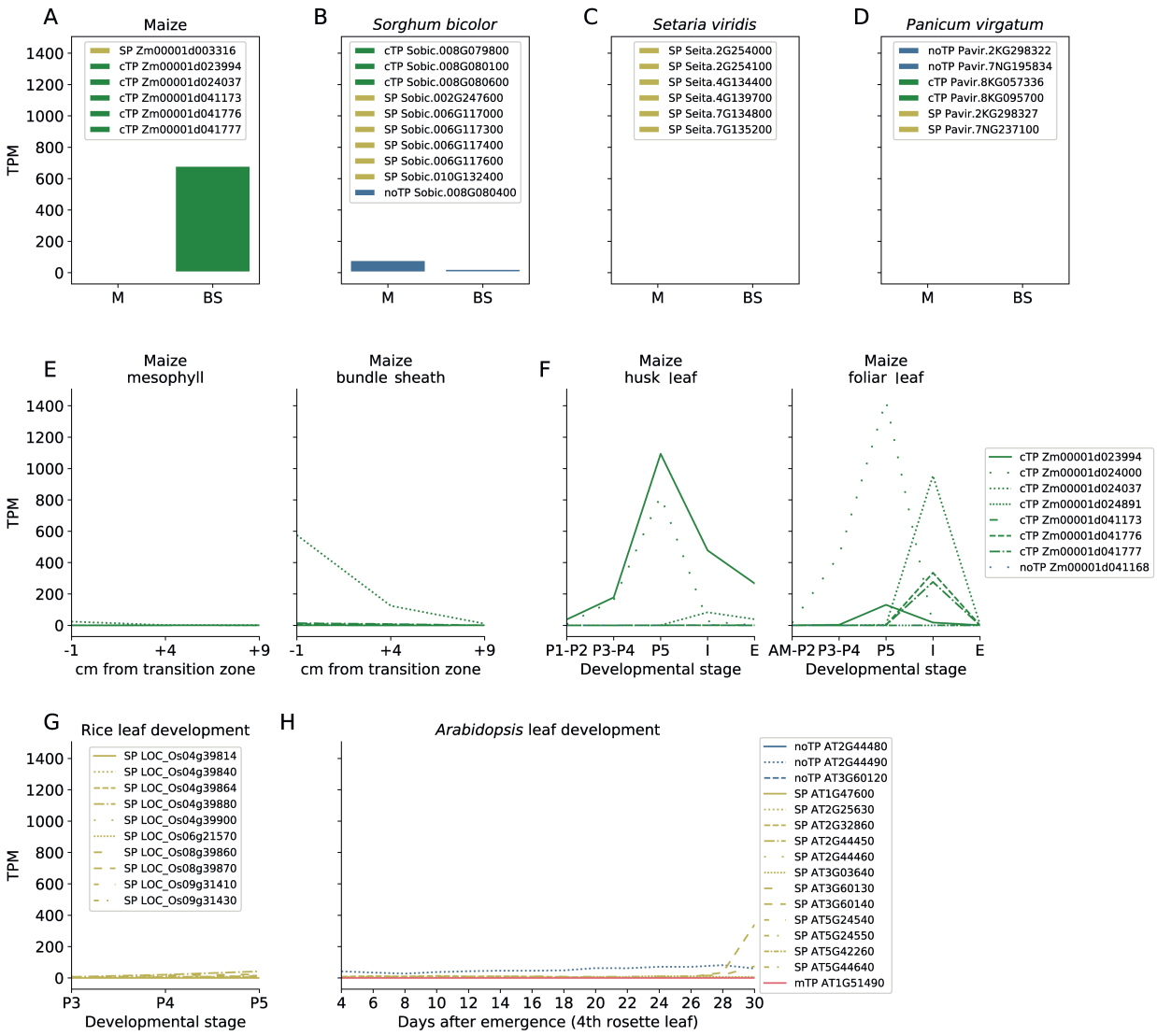


Figure 3.9. Figure legend on the next page.

Figure 3.9. Transcript expression for members of the beta-glucosidase gene family containing *ZmBGLU1* in selected transcriptomic data sets. Due to the size of the gene family, only transcripts with a mean leaf expression of >1 transcript per million (TPM) are shown. Details of the transcriptome data sets used can be found in Appendix 3.2. Transcript amounts in **A)** mesophyll (M) and bundle sheath (BS) cells from mature maize leaves, **B)** mesophyll (M) and bundle sheath (BS) cells from mature *Sorghum bicolor* leaves, **C)** mesophyll (M) and bundle sheath (BS) cells from mature *Seteria viridis* leaves, **D)** mesophyll (M) and bundle sheath (BS) cells from mature *Panicum virgatum* leaves, **E)** mesophyll (M) and bundle sheath (BS) cells along a developmental gradient (the developmental gradient denotes distance from the source-sink transition zone of the maize leaf with -1cm representing the least mature tissue and +9cm representing the most mature tissue), **F)** developing maize husk and foliar leaves (AM = apical meristem, P = plastochron, I = immature leaf, E = fully expanded leaf), **G)** developing rice leaves (P3-P5 = plastochron stage 3-5), and **H)** developing *Arabidopsis* rosette leaves. The colour of the bars and lines indicates the predicted targeting of the encoded protein. **Abbreviations**: chloroplast transit peptide (cTP), mitochondrial target peptide (mTP), peroxisomal target signal (PTS), signal peptide (SP), no N-terminal target signal identified (noTP), transcripts per million (TPM).

3.2.7 Analysis of a change in targeting for an aspartate aminotransferase highlights a gap in our understanding of C_4 photosynthesis in maize

A gain in peroxisomal targeting was identified to have occurred for an aspartate aminotransferase (AspAT) concomitant with the evolution of C_4 photosynthesis (Figure 3.10 A). Although AspAT is a core C_4-cycle enzyme, the identified change in targeting occurred in the ancestor of an AspAT isoform with low leaf expression and which is thus presumably not functioning in the C_4 pathway. Interrogation of gene expression for this family did however highlight an inconsistency between AspAT transcript abundance in the maize leaf and the expected activity of AspAT given our current model of C_4 photosynthesis in this species. In maize, the predominant C_4 pathway in operation is an

NADP-ME pathway. In this pathway malate is used as the four-carbon compound to move CO_2 from the mesophyll cells to the bundle sheath cells (Figure 3.2). In addition however, a secondary C_4 cycle also operates in maize via PEP-CK. In this case, aspartate is used as the four-carbon transfer compound (Figure 3.2) and it has been shown that in maize as much as 14-25% of carbon that moves from the mesophyll to the bundle sheath cells does so as aspartate (Chapman and Hatch 1981; Arrivault et al. 2017). For carbon flux via aspartate to be sustained in the maize leaf through a PEP-CK pathway, it is expected that AspAT activity will be present in both the mesophyll and bundle sheath cells (Figure 3.2). Analysis of AspAT gene expression in maize however highlighted that while there is high AspAT expression in the mesophyll cell, the expression of AspAT in the bundle sheath cells is negligible (Figure 3.10 B). This incongruence in the expression, and activity, of AspAT between the mesophyll and bundle sheath cells of maize has been noted before (Schlüter et al. 2019) and remains an open question in our understanding of C_4 photosynthesis in this species. One possible explanation is that in maize, aspartate in the bundle sheath cell is converted to oxaloacetate through an as of yet undiscovered pathway. The identification and characterisation of a candidate aspartate-metabolising enzyme in this regard is the focus of the next chapter.

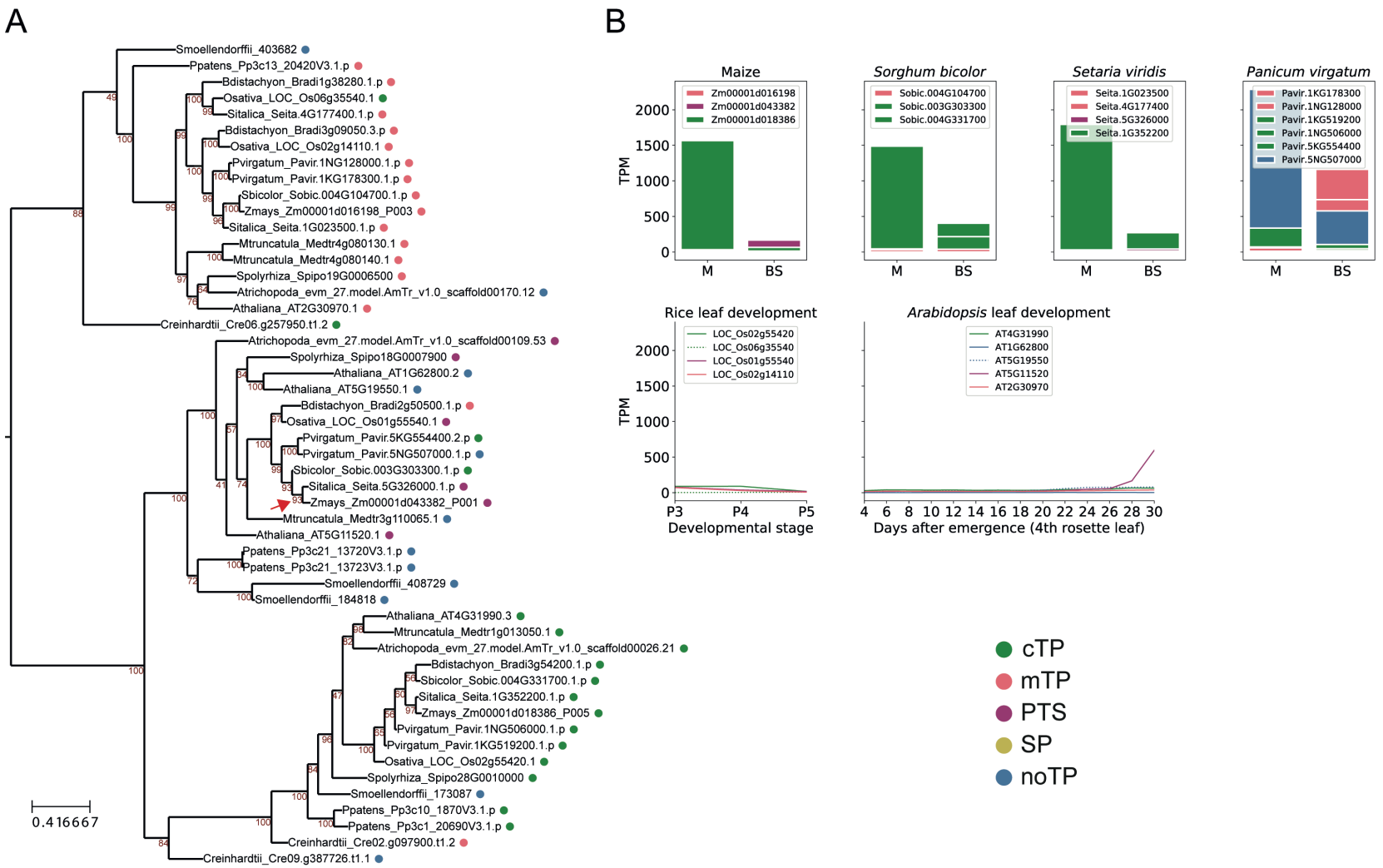


Figure 3.10. Analysis of a change in targeting within a family of plant AspATs highlights that in maize AspAT gene expression is much lower in the bundle sheath cells than is expected for this C_4 subtype. Figure legend continued on the next page.

Figure 3.10. (continued) A) Maximum likelihood phylogenetic tree for the AspAT gene family. The red arrow indicates the branch along which a gain in peroxisomal targeting was identified to have occurred. The tree was constructed with IQ-TREE using the substitution model LG+I+G4, and bootstrap values are from UFBoot. The scale bar represents amino acid substitution per site. **B)** Expression of AspAT in selected transcriptomic data sets (detailed in Table A3.001). Transcript expression in maize, *Sorghum bicolor*, *Setaria viridis*, and *Panicum virgatum* is shown for mesophyll (M) and bundle sheath (BS) cells. In rice, expression is shown along the plastochron (P) developmental gradient. In *Arabidopsis*, expression is shown along a developmental gradient of leaf maturity. **Abbreviations**: chloroplast transit peptide (cTP), mitochondrial target peptide (mTP), peroxisomal target signal (PTS), signal peptide (SP), no N-terminal target signal identified (noTP), transcripts per million (TPM).

3.3 Discussion

The overall aim of this chapter was to investigate the changes in organellar protein targeting that occurred during the evolution of C_4 photosynthesis and identify whether any of them may have been important in the evolution of this plant trait. The underlying premise for this work is that C_4 evolution involves the concerted recruitment of enzymes and transporters from ancestral C_3 functions often unrelated to photosynthesis (Aubry et al. 2011). As discussed in Chapter 2, changes in protein targeting present a mechanism by which gene neofunctionalisation may occur, especially following gene duplication. Indeed, hundreds of gene duplication events have been estimated to have occurred on the branches encompassing an origin of C_4 photosynthesis in the C_4 species considered in this data set (Emms et al. 2016). Although there have been several studies investigating the molecular sequence changes that led to the recruitment of proteins to a C_4 function, to-date none have focused on identifying whether changes in organellar protein targeting may have been important.

In this chapter I used the ancestral state estimation method presented in Chapter 2 to identify the changes in protein targeting that occurred during the evolution of C_4 photosynthesis in four C_4 monocot species. These species were chosen because there are mesophyll and bundle sheath cell-specific transcriptome data sets available for them and so identified changes in targeting could be screened for those that occurred in the ancestor of proteins known or predicted to have a C_4-function. In the end, although many changes in protein targeting were identified to have occurred during the evolution of C_4 photosynthesis in these species, none were found among the enzymes and transporters that are known to function directly in the C_4 pathway.

Although no changes in targeting were identified among core C_4 pathway proteins, transcriptome data sets were employed to allow for the discovery of changes in proteins whose extant members have a putative role in C_4 photosynthesis. Of particular interest

was the predicted change in targeting of a beta-glucosidase from the secretory pathway to the chloroplast in the ancestor of the four C_4 species in this data set. This beta-glucosidase protein was found to belong to a large gene family containing many paralogous genes. In maize, one of the chloroplast-targeted beta-glucosidase paralogs has been shown to function in the release of active cytokinin from the inactive cytokinin-conjugates (cytokinin-O-glucoside). In the experiments demonstrating this, tobacco protoplasts were induced to divide by cytokinin-O-glucosides only when expressing the maize beta-glucosidase *ZmBGLU1* (Brzobohatý et al. 1993). In an additional study, stable tobacco transformants overexpressing the same maize *ZmBGLU1* were found to exhibit a cytokinin hypersensitive response when grown on excess cytokinin (Kiran et al. 2006).

Cytokinins are a class of plant hormones that have a wide range of biological functions related to growth and development (Hwang et al. 2012). After synthesis, the level of biologically active cytokinin in the cell is regulated by processes of cytokinin degradation and conjugation. The conjugation of active cytokinin to glucose is generally thought to be reversible (*O*-glucosides) or irreversible (*N*-glucosides), although there is increasing evidence that plants can actually hydrolyse several forms of the 'irreversible' *N*-glucosides but the pathways are less well described (Hošek et al. 2020; Hoyerová and Hošek 2020). In both monocot and dicots cytokinin-*O*-glucosides have been shown to be abundant in the leaf tissue (Gajdošová et al. 2011). In maize for example over 80% of cytokinin in the leaf has been estimated to occur in the inactive *O*-glucoside form (Gajdošová et al. 2011). The conjugation of cytokinin into *O*-glucosides is catalysed by *O*-glucosyltransferases and several genes encoding these enzymes with this activity have been described (Rodo et al. 2008). The reverse of this reaction is the hydrolysis of cytokinin-*O*-glucosides to release active forms of cytokinin, a reaction that is expected to be catalysed by a beta-glucosidase (Hošek et al. 2020). To date however only one gene has been discovered as encoding a beta-glucosidase with this activity. This gene is *ZmBGLU1* and it belongs to the clade of chloroplast-targeted beta-glucosidases identified in this chapter. Very few of the beta-

glucosidases in this gene family have been functionally characterised but one of the *Arabidopsis* members in this group is an ortholog of *ZmBGLU1* has been shown to hydrolyse flavonol-glucosides, as opposed to cytokinin-glucosides, in the apoplast (Borderies et al. 2003; Roepke and Bozzo 2015; Roepke et al. 2017). It is tempting to speculate therefore that the change in beta-glucosidase targeting in the monocot clade is associated with the neofunctionalisation of this enzyme to a role in the metabolism of cytokinin conjugates. Further phylogenetic analysis and characterisation of this beta-glucosidase gene family is needed to establish if this is the case. Furthermore, although chloroplast-targeted beta-glucosidases were identified in maize, *Sorghum* and *Panicum*, gene expression for this enzyme was only found in the leaf tissue of maize. It would be interesting to determine however whether the orthologs of *ZmBGLU1* in *Sorghum* and *Panicum* are expressed at earlier stages of leaf development given that in maize *BGLU1* expression peaked during the immature stages of leaf development. Unfortunately, RNAseq data sets for mesophyll and bundle sheath cell specific expression don't yet exist for *Sorghum* and *Panicum*.

In maize, transcript expression of the chloroplast-targeted beta-glucosidase isoform was found to be highest in the developing leaf and transcript accumulation for one maize beta-glucosidase paralog was found to be restricted to the bundle sheath cells of the developing foliar leaf. Given that this bundle sheath cell specific gene is most closely related paralog to the *ZmBGLU1*, it is presumed that it also encodes a beta-glucosidase with substrate activity for cytokinin-*O*-glucosides. The high activity of a beta-glucosidase releasing active cytokinin into the bundle sheath cell during development is interesting given that cytokinin signalling has been implicated in the differentiation of bundle sheath cells for their role in C_4 photosynthesis (Ermakova et al. 2020). In particular, bundle sheath cells of C_4 plants have an overabundance of chloroplasts compared to bundle sheath cells of C_3 plants and this 'photosynthetic functionalisation' of the bundle sheath is thought to have been a crucial step in the evolution of C_4 photosynthesis (Wang, Khoshravesh, et al. 2017; Ermakova et

al. 2020). Cytokinin has been demonstrated to promote chloroplast development and biogenesis (reviewed in Mähönen et al. 2006; Cortleven and Schmülling 2015; Wang, Khoshravesh, et al. 2017; Skalák et al. 2019), and in *Arabidopsis* it has been estimated that over 100 photosynthetic genes are cytokinin responsive (Wang, Hendron, et al. 2017). Furthermore, transcription factors such as the GLK genes, which are thought to act downstream of cytokinin, have been associated with bundle sheath differentiation in maize and there is much interest in using such transcription factors to induce chloroplast development in the bundle sheath cells of C_3 plants (Rossini et al. 2001; Kobayashi et al. 2012; Wang, Fouracre, et al. 2013). Expression of maize GLK genes in rice led to a C_4-like functionalisation of the bundle sheath cells including the activation of chloroplast biogenesis and the proliferation of plasmodesmata connections between neighbouring cells (Wang, Khoshravesh, et al. 2017). Thus, a hypothesis is that in maize, during leaf development, this chloroplast-targeted beta-glucosidase is activated in the bundle sheath cells to ensure sufficient levels of active cytokinin for the photomorphogenesis of this cell-type. If so, then this beta-glucosidase could be a promising target for engineering strategies looking to prime bundle sheath cells in C_3 plants for a C_4 pathway. While the expression of individual chloroplast biogenesis transcription factors ectopically in the vascular tissue of C_3 plants has been successful in promoting chloroplast biogenesis, these efforts are still a long way from recreating the chloroplast abundance seen in the bundle sheath cells of C_4 plants (Ermakova et al. 2020). Given the expression pattern of these beta-glucosidases in maize, transferring the activity of this enzyme into the bundle sheath cells of a C_3 leaf may be one way to activate a cytokinin signalling pathway at a higher level and thus could be a more successful strategy for promoting the features of a C_4 bundle sheath cell in a C_3 plant.

Lastly in this chapter, a change in protein targeting of an aspartate aminotransferase (AspAT) was highlighted. This change is not thought to have been important in the recruitment of AspAT to a C_4 function, but rather analysis of AspAT expression called

attention to a gap in our understanding of maize C_4 photosynthesis that will be investigated in the next chapter (Chapter 4).

3.4 Materials and methods

3.4.1 Orthogroup and phylogenetic tree inference

Proteome data (primary protein transcripts) for plant species was obtained from Phytozome v13 (Goodstein et al. 2012). Orthogroups, orthogroup gene trees, and the species tree were inferred using OrthoFinder v2.4.0 (Emms and Kelly 2015) with the settings 'M -msa' to infer rooted gene trees from orthogroup multiple sequence alignments using FastTree v2.1.9 (Price et al. 2010). These orthogroup gene trees were subsequently used for ancestral state estimation. Each internal branch of each gene tree was reconciled to a branch in the species tree according to the last common ancestor of the genes subtending the gene-tree branch (most recent common ancestor mapping described in Chapter 2).

For orthogroups of interest identified during this analysis (including those discussed in this chapter) a second gene-tree-inference method was used to confirm the topology of the gene tree and the placement of an ancestral change in protein targeting. Here, orthogroup multiple sequence alignments were trimmed using trimAL v1.4 with the settings '-gt 0.25' (Capella-Gutiérrez et al. 2009) and gene trees inferred from the trimmed alignment using IQ-TREE v.2.0.3 (Nguyen et al. 2015) with the settings '-m TEST' to automatically select the best-fitting model of sequence evolution for each gene-tree inference, and '-B 1000' to obtain bootstrap support values using UFBoot (Hoang et al. 2018). These IQ-TREE gene trees were then rooted using the OrthoFinder script 'resolve.py' with the gene tree and rooted species tree as input. This script roots the gene trees using information about the species topology (i.e., outgroups) in the rooted species tree. For one orthogroup (OG0000024) inspection of outgroup placement in the phylogeny revealed that the orthogroup was in fact a fusion of six individual orthogroups and so it was split up as such and gene trees were re-inferred for each new group. Phylogenetic trees were inspected using Dendroscope (Huson et al. 2007) and visualisations created with the ETE toolkit (Huerta-Cepas et al. 2016). Visualisations of multiple sequence alignments were created

using Jalview v2.11.0 (Waterhouse et al. 2009). The conservation histogram created by Jalview shows the conservation of the physico-chemical properties for each column of the alignment with identical residues at a site scoring highest, and amino acid substitutions in the same physico-chemical class scoring next highest. The score for each column is shown below the histogram; conserved columns with a score of 11 are indicated by '*' and columns with only mutations in the same physico-chemical class are indicated with a '+'.

3.4.2 Prediction of protein targeting

Prediction of protein targeting to the chloroplast, mitochondrion and secretory pathway was conducted using TargetP v2.0 (Armenteros et al. 2019) in plant mode with default cut-offs. Peroxisome proteins were identified by the method described in Chapter 2. Briefly, a protein was assigned as targeted to the peroxisome if it contained a peroxisomal targeting signal 1 (PTS1) at the C-terminus or a peroxisomal targeting signal 2 (PTS2) at the N-terminus of the protein sequence. PTS1 motifs were identified as one of the following amino acid tripeptides located at the C-terminus of the protein immediately before the stop codon: SRL, SRM, SRI, ARL, ARM, PRL, SKL, SKM, and AKL. PTS2 motifs were identified by the amino acid sequence R[LI]X5HL located in the first 30 residues of the protein sequence. Where a peroxisomal target signal was identified in addition to another target signal, the protein sequence was assigned as targeted to both compartments for subsequent ancestral state estimation.

3.4.3 Ancestral state estimation

Ancestral changes in protein targeting to the chloroplast, mitochondrion, secretory pathway and peroxisome were identified in orthogroup gene trees by the ancestral state estimation method described in Chapter 2. Briefly, each organelle was considered independently and the presence or absence of an organelle target signal in extant protein sequences was considered as binary trait data. Using this trait data and orthogroup gene trees, maximum likelihood ancestral character estimation (ACE) was conducted using the

"ACE" function in the R package ape 5.2 (Paradis et al. 2004) with the "all rates different" model for discrete data. The ACE method assigns a probability score (between 0 and 1) that an ancestral protein sequence (represented by an internal gene-tree branch) had a specific organellar target signal. Branches were thus assigned using a winner takes all approach as either representing an ancestral protein with (probability score ≥0.5) or without (probability score <0.5) a target signal. Changes in protein targeting between consecutive internal branches of a gene tree could then be identified. As in the analysis in Chapter 2, a filter was applied to identify only those changes with reasonable support from the predicted targeting of extant proteins. Here, ≥75% of the extant proteins' descendent from a branch on which a change is estimated to have occurred must be predicted to be targeted to the changed state, while ≥75% of extant proteins' descendent from the sister branch maintain the ancestral state.

3.4.4 Analysis of gene expression data

RNA-seq data sets used in this chapter were previously published and available in NCBI SRA. A list of accession numbers and study details is provided in Appendix Table A3.001. In most cases, read data for these studies was processed in-house to standardise the analysis method across data sets and thus produce comparable TPM values. In addition, this allowed for the quality of read data to be evaluated. In cases where bespoke postprocessing of reads had been carried out in a study, for example cross-species mapping of reads or deconvolution of reads to specify cell type, then count data were taken as presented in that study (details in Appendix Table A3.001). For those data sets processed in this chapter, raw read (fastq) files were downloaded from NCBI SRA and trimmed with Trimmomatic v0.39 (Bolger et al. 2014), with settings 'LEADING: 20, TRAILING: 20, SLIDINGWINDOW: 5:20, MINLEN: 35'. Transcript counts and effective lengths were then quantified using the gene models from the relevant species (see Appendix Table A3.001) using Salmon v0.14.1 (Patro et al. 2017) with settings (-l A --validateMappings). Transcript counts from all gene models corresponding to the same

gene were summed to generate abundance estimates for the primary transcript of each gene. In addition, raw counts from each expression data set listed above were analysed using DESeq2 v1.22.1 (Love et al. 2014) to identify differentially expressed mesophyll and bundle sheath genes. Only transcripts with a base mean of normalized counts >2 were included for differential expression analysis.

3.4.5 Functional term enrichment analysis and prediction of transmembrane helices

Functional term enrichment analysis was conducted using STRING v11.0 (Szklarczyk et al. 2019) with default settings. Here, the *Arabidopsis* ortholog to a clade containing a change in protein targeting in the test set was used as input. If there was more than one *Arabidopsis* ortholog then a single one was randomly selected. Enrichment was tested for against the background of all predicted proteins in the *Arabidopsis thaliana* genome. A false discovery rate stringency was set at 0.05. The prediction of transmembrane helices in proteins was conducted using TMHMM v.2.0 (Möller et al. 2001).

Chapter 4: Bioinformatic and molecular characterisation of a putative aspartate aminotransferase localised in the bundle sheath cells of maize

4.1 Introduction

In the C_4 photosynthetic pathway, carbon dioxide (CO_2) is initially sequestered in the mesophyll cells of the leaf in a four-carbon compound which then diffuses into the bundle sheath cells and is subsequently decarboxylated to release CO_2 at high concentrations around the enzyme ribulose-1,5-bisphosphate carboxylase-oxygenase (RuBisCo). While the identity of the carbon metabolites that flow between the mesophyll and bundle sheath cells have been well characterised, not all the enzymes and transporters that facilitate this carbon flux in all C_4 species have been identified. In the C_4 species *Zea mays* (maize) for example, aspartate is one of the four-carbon compounds shuttled from the mesophyll to the bundle sheath cells where it is thought to be converted to oxaloacetate prior to decarboxylation. However, the aminotransferase enzyme responsible for catalysing this conversion of aspartate to oxaloacetate has not been identified (highlighted in Schlüter et al. 2019). This chapter aims to address this knowledge gap by identifying the enzyme responsible for this aminotransferase reaction in maize. The work in this chapter builds upon bioinformatic work previously conducted in the Kelly Lab that identified candidate genes in this regard. From this work, the gene *AILP1* was identified as a strong candidate for carrying out (in part or in full) the deamination of aspartate to oxaloacetate. To date however, there has been no experimental characterisation of *AILP1* genes in any species. The work in this chapter therefore details a bioinformatic and experimental investigation into the putative function of *AILP1* in plants and its potential role in the maize C_4 photosynthetic cycle.

4.1.1 Background

It is just over 50 years since the first comprehensive biochemical elucidation of a C_4 photosynthetic cycle (Hatch and Slack 1966). Shortly after its discovery in the leaves of sugarcane, it became clear that C_4 photosynthesis is widely distributed across the plant phylogeny and that several biochemical variations of the cycle exist (Edwards et al. 1971; Hatch et al. 1975; Hatch and Kagawa 1976). In all variations that have been described, the C_4 photosynthetic cycle functions to concentrate carbon dioxide (CO_2) around the enzyme ribulose-1,5-bisphosphate carboxylase-oxygenase (RuBisCo) which catalyses the fixation of CO_2 into organic molecules. Under saturating CO_2 concentrations, RuBisCo operates efficiently to catalyse a reaction with CO_2 and ribulose 1,5-bisphosphate (RuBP) to form two molecules of 3-phosphoglycerate (3-PGA). Under non-saturating CO_2 concentrations however (such as in the atmosphere) oxygen can also enter the active site of RuBisCo and the reaction with RuBP leads to the formation of a toxic intermediate, 2-phosphoglycolate (2-PG). Detoxification and recovery of carbon from 2-PG requires energy in the form of ATP and reducing power, while also resulting in the loss of an ammonia ion that must be re-assimilated. Collectively, the set of reactions that carry out this recovery process are known as photorespiration. For C_3 plants without a CO_2 concentration mechanism the cost of photorespiration can comprise a substantial part of the daily operating energy budget. For example, in warmer climates photorespiration has been estimated to consume 30% of the cellular energy budget (Sage 2004), with annual losses to on-farm wheat and soybean yield in the USA estimated at 20% and 36% respectively (Walker et al. 2016). Temperature is important here as the affinity of RuBisCo for oxygen increases with temperature. Thus, as temperature rises, the negative consequences of RuBisCo promiscuity in C_3 plants become more pronounced (Ehleringer et al. 1991). The consequence of an increase efficiency in CO_2 sequestration in C_4 plants also means that they have a higher water-use and nitrogen-use efficiency compared to their C_3 relatives as result of being able to keep their stomata more closed and needing less total investment in leaf RuBisCo, respectively (Gowik and Westhoff 2011).

In C_4 photosynthesis a CO_2 concentration gradient is set up between two distinct compartments in the leaf. In most C_4 plants these are the mesophyll cells and bundle sheath cells, though several variations exist in different C_4 lineages (Nelson and Langdale 1992; Edwards 2011). While in the ancestral C_3 pathway, CO_2 entering the mesophyll cells of the leaf diffuses straight to the active site of RuBisCo in the chloroplasts, in C_4 plants RuBisCo is not present in the chloroplast of the mesophyll cell and instead CO_2 is initially fixed into a four-carbon compound oxaloacetate by the enzyme phosphoenolpyruvate carboxylase (PEPC) which has no affinity for molecular oxygen. As oxaloacetate is inherently unstable, it is converted into a more stable four-carbon compound – either malate or aspartate – in the mesophyll cell. Malate or aspartate then diffuses into the bundle sheath cells, whose chloroplasts contain RuBisCo, and is subsequently decarboxylated to release CO_2. Chloroplast positioning and/or the presence of a suberised cell wall surrounding the bundle sheath cells prevents diffusion of gasses back out of the cell and thus CO_2 accumulates at high concentrations in the vicinity of RuBisCo (Von Caemmerer and Furbank 2003).

Historically C_4 plants have been categorised into one of three subtypes depending on the decarboxylation enzyme present in the bundle sheath cells: either NADP-dependent malic enzyme (NADP-ME), NAD-dependent malic enzyme (NAD-ME) or phosphoenolpyruvate carboxykinase (PEP-CK). These three decarboxylases are differentially localised within the bundle sheath cell and each catalyses a reaction with different substrates and products. Concomitantly, the route of metabolite flow associated with each decarboxylating enzyme is different (as shown in Figure 3.2, Chapter 3). In the case of NADP-ME and NAD-ME, malate is used as the substrate for decarboxylation which takes place in the chloroplast and mitochondrion, respectively. The malate used in this reaction may come directly from the mesophyll cell or, alternatively, aspartate from the mesophyll cell may be converted to malate in the bundle sheath cell. In the case of PEP-CK, oxaloacetate is the substrate and is typically supplied through the conversion of aspartate

to oxaloacetate by an aspartate aminotransferase in the bundle sheath cell. Despite these subtype distinctions, it has become increasingly apparent during the last 50 years of C_4 research that the three C_4 subtype pathways do not always operate in isolation (Furbank 2011; Wang et al. 2014; Rao and Dixon 2016). In particular, all species classically attributed to the PEP-CK subtype have now been found to have at least some activity of NADP-ME or NAD-ME (Bräutigam et al. 2014). Similarly, several NADP-ME and NAD-ME species have been found to possess some secondary activity of PEP-CK. This has led to the suggestion that there are in fact only two main subtypes of C_4 photosynthesis (NADP-ME and NAD-ME) which may possess differing amounts of PEP-CK activity (Rao and Dixon 2016). Figure 4.1 shows a schematic of these two malic enzyme pathways (NADP-ME and NAD-ME) alongside a PEP-CK cycle. No mixed malic enzyme subtypes have been found to occur to date. Typically, in species that conduct a mixed-subtype C_4 pathway, one decarboxylation cycle is the primary cycle and dominates over the other in terms of total carbon flux (although detailed flux analyses have only been conducted for a handful of species). It should also be noted that the subcellular pathway of metabolite flow through each of these pathways can vary between species of the same subtype. In particular, the aspartate and alanine aminotransferase reactions that typically support an NAD-ME or PEP-CK cycle have been shown to be differentially localised within the mesophyll and bundle sheath cells depending on the C_4 species under consideration (Hatch and Mau 1973; Chapman and Hatch 1983; Sommer et al. 2012).

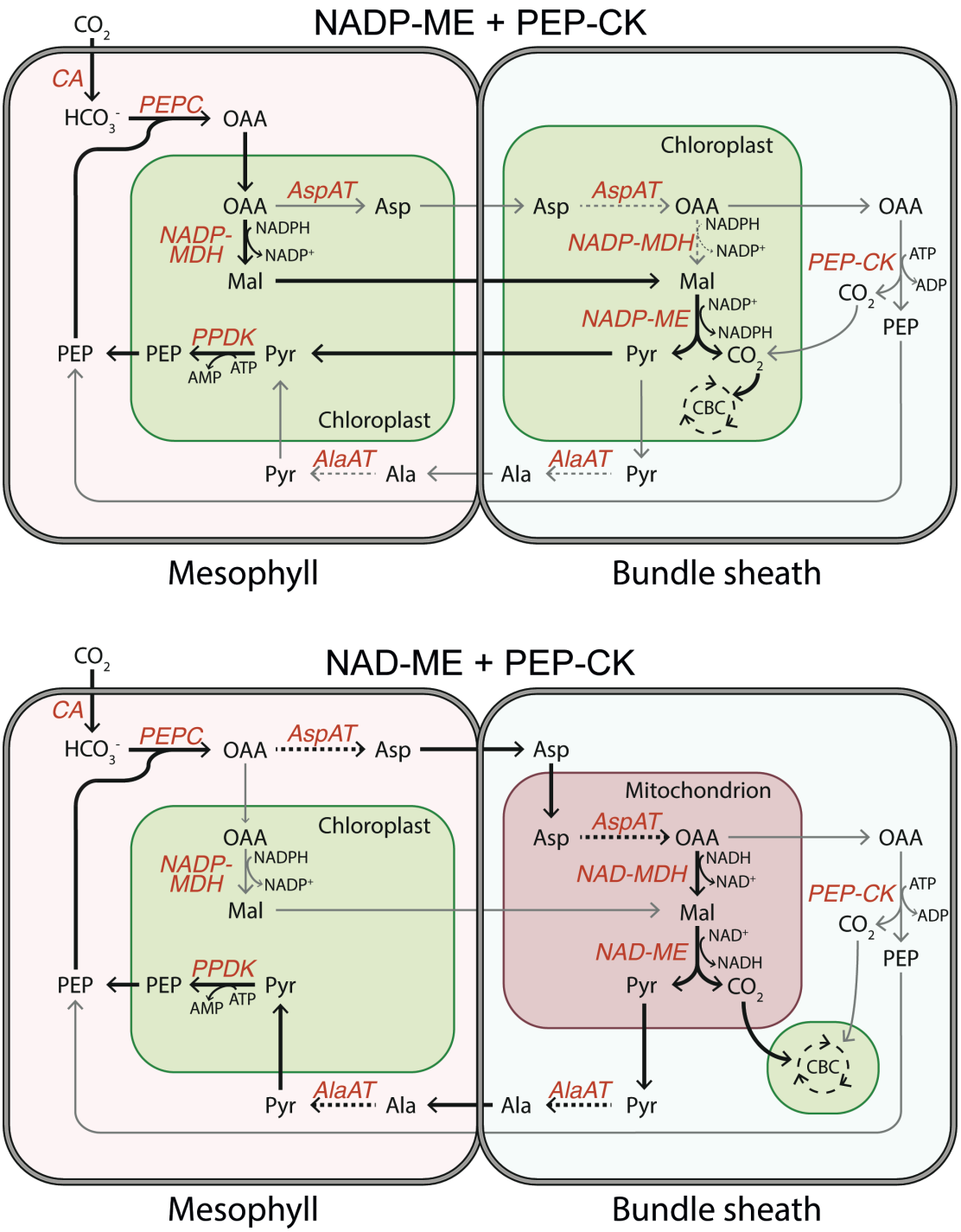


Figure 4.1. Schematic of the NADP-ME and NAD-ME C_4 photosynthetic pathways with a secondary PEP-CK pathway mapped on. The canonical malic enzyme pathway (NADP-ME or NAD-ME) is indicated by black, bold lines. The PEP-CK cycle (and other alternative routes of metabolite flow) are shown by the grey, narrow lines. Where an arrow represents the conversion of one metabolite to another, the enzyme catalysing the reaction is shown in red italics. Dashed lines indicate an enzymatic reaction whose subcellular localisation has been shown to vary between species.

Reactions involving energy or reducing equivalents (ATP, NADPH, or NADH) are indicated. The transporters that facilitate the transmembrane transport of metabolites are not shown. **Abbreviations:** bicarbonate (HCO_3^-), oxaloacetate (OAA), aspartate (Asp), malate (Mal), pyruvate (Pyr), phosphoenolpyruvate (PEP), alanine (Ala), Calvin Benson cycle (CBC), carbonic anhydrase (CA), phosphoenolpyruvate carboxylase (PEPC), aspartate aminotransferase (AspAT), NADP-dependent malate dehydrogenase (NADP-MDH), NAD-dependent malate dehydrogenase (NAD-MDH), NADP-dependent malic enzyme (NADP-ME), NAD-dependent malic enzyme (NAD-ME), phosphoenolpyruvate carboxykinase (PEP-CK), alanine aminotransferase (AlaAT), pyruvate phosphate dikinase (PPDK).

The grass species *Zea mays* (maize) is often categorised as an NADP-ME C_4 but there is now substantial evidence that a secondary PEP-CK cycle also operates in this species (Hatch 1971; Chapman and Hatch 1981; Walker et al. 1997; Wingler et al. 1999; Pick et al. 2011; Arrivault et al. 2017). Carbon labelling experiments in maize have demonstrated that between 10 and 25% of CO_2 is incorporated into aspartate in the mesophyll cells rather than malate (Hatch 1971; Arrivault et al. 2017). Although it was initially thought that aspartate must be converted to malate in the bundle sheath cells for decarboxylation by NADP-ME (Chapman and Hatch 1981), it was later demonstrated that the bundle sheath cells have appreciable levels of PEP-CK activity (Walker et al. 1997) and that isolated bundle sheath strands can decarboxylate aspartate under conditions supporting PEP-CK activity (Wingler et al. 1999). More recently, detailed cell-specific carbon labelling experiments confirmed the presence of a photosynthetically active pool of aspartate and PEP (the product of PEP-CK decarboxylation) in maize and demonstrated that there is a sufficient concentration gradient of aspartate and PEP between the mesophyll and bundle sheath cells to drive intercellular diffusion of these metabolites (Arrivault et al. 2017). While the NADP-ME pathway in maize has been well characterised, it is less clear how metabolite flux operates through this secondary PEP-CK cycle.

A more detailed schematic of the enzymatic reactions in the maize C_4 cycle are shown in Figure 4.2. Here, CO_2 diffuses into the mesophyll cells and is initially converted to bicarbonate by the enzyme carbonic anhydrase before being combined with three-carbon PEP to form four-carbon oxaloacetate by PEPC. From here the route of carbon flow splits. Following the predominant NADP-ME cycle, oxaloacetate formed from the reaction catalysed by PEPC is transported into the chloroplast where it is reduced to malate by NADP-dependent malate dehydrogenase. Malate is then transported out of the chloroplasts and diffuses into the bundle sheath cell through the plasmodesmata that connect the two cells. In the bundle sheath cell, malate is transported into the chloroplast again and decarboxylated directly by NADP-ME. The product of this decarboxylation reaction is a molecule of CO_2 and a molecule of pyruvate. The released CO_2 enters the Calvin cycle via RuBisCo and the pyruvate is transported out of the chloroplast and diffuses into the mesophyll cells where it is converted back to PEP by pyruvate phosphate dikinase (PPDK). A fraction of the pyruvate in the bundle sheath cell is converted to alanine which diffuses to the mesophyll cells before being converted back to pyruvate. This ensures that nitrogen is balanced between the two cell types by bringing back the amino-group that is transported to the bundle sheath cell as part of aspartate when aspartate is used as the four-carbon transfer metabolite. As NADP-ME species typically have lower activities of the ATP-producing photosystem II in the bundle sheath cells, the energy-intensive reactions of the Calvin cycle are exported to the mesophyll cells in a pathway known as the triose-phosphate pathway (Arrivault et al. 2017). Here, 3-PGA moves from the mesophyll cells to the bundle sheath cells where it is reduced to triosephosphate which is then transported back to the bundle sheath cells to continue the regeneration of RuBP in the Calvin cycle.

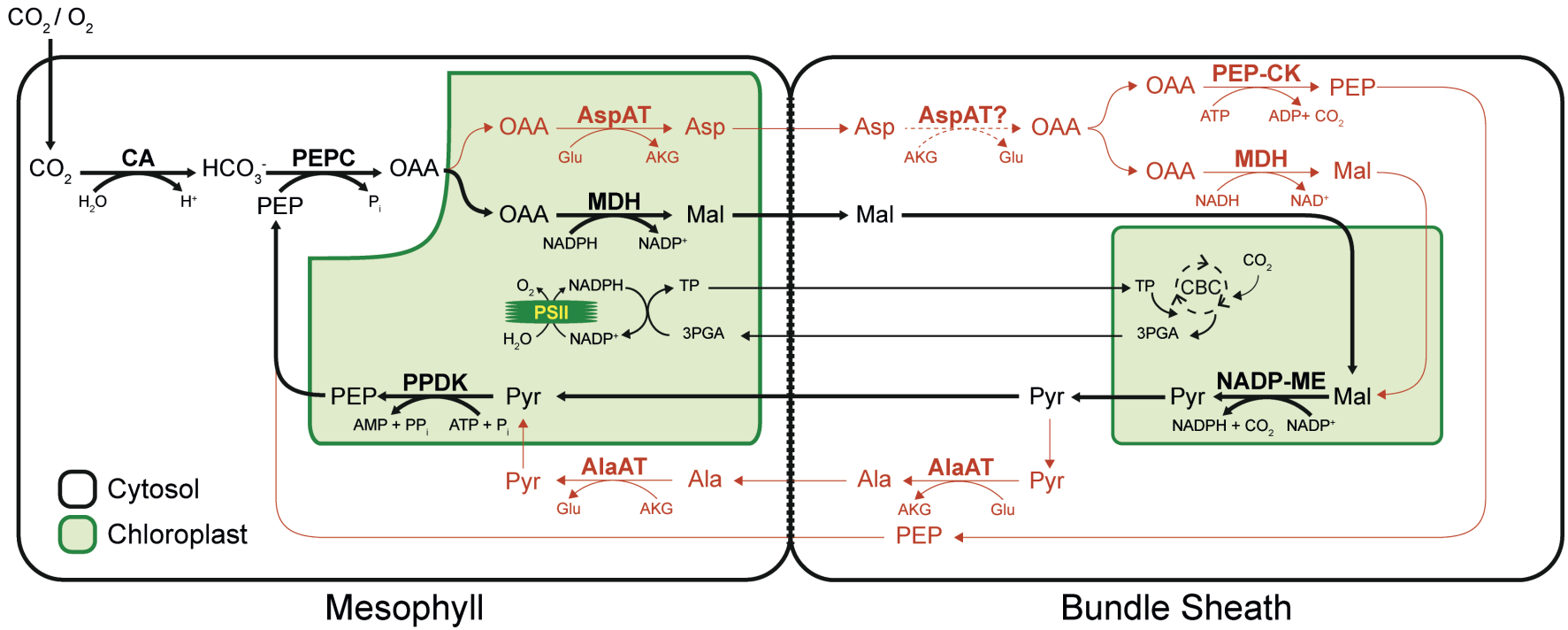


Figure 4.2. A more detailed schematic showing metabolite flow and the main enzymatic reactions of the C_4 pathway in maize. The black, bold arrows show the primary NADP-ME decarboxylation pathway. The thinner red arrows show the secondary PEP-CK decarboxylation pathway. The dotted red line indicates the metabolism of aspartate to oxaloacetate in the bundle sheath however the enzyme that catalyses this reaction has not been identified. In addition, the pathway of 3-PGA reduction through the triose-phosphate pathway is shown by the thin black arrows in the centre. **Abbreviations:** oxaloacetate (OAA), aspartate (Asp), malate (Mal), pyruvate (Pyr), phosphoenolpyruvate (PEP), alanine (Ala), glutamate (Glu), α-ketoglutarate (AKG), 3-phosphoglyceric acid (3PGA), triose phosphate (TP), Calvin Benson cycle (CBC), photosystem II (PSII), carbonic anhydrase (CA), phosphoenolpyruvate carboxylase (PEPC), aspartate aminotransferase (AspAT), NADP-dependent malate dehydrogenase (MDH), NADP-dependent malic enzyme (NADP-ME), NAD-dependent malic enzyme (NAD-ME), phosphoenolpyruvate carboxykinase (PEP-CK), alanine aminotransferase (AlaAT), pyruvate phosphate dikinase (PPDK).

As previously mentioned, in addition to the transfer of malate from the mesophyll to the bundle sheath cells, aspartate has also been shown to act as a transfer metabolite of CO_2 in maize C_4 photosynthesis (Chapman and Hatch 1981; Wingler et al. 1999; Arrivault et al. 2017). Following the proposed route of aspartate decarboxylation, oxaloacetate in the mesophyll cell chloroplasts is converted to aspartate by the enzyme aspartate aminotransferase (AspAT). This transamination reaction involves the transfer of an amino group from glutamine to oxaloacetate, thus forming aspartate and α-ketoglutarate. Aspartate is then transported out of the chloroplast and diffuses into the bundle sheath cells where it can either be converted back to oxaloacetate for decarboxylation by PEP-CK in the cytosol (as in a canonical PEP-CK pathway) or converted to malate for decarboxylation by NADP-ME in the chloroplast. In either of these decarboxylation pathways, aspartate must be converted to oxaloacetate in the bundle sheath. However, in maize it has been previously noted that the activity of AspAT is substantially lower in the bundle sheath cells compared to the mesophyll cells (Majeran et al. 2005; Friso et al. 2010; Chang et al. 2012; Tausta et al. 2014; Denton et al. 2017; Schlüter et al. 2019). Although several of these studies infer activity from mRNA expression data, in one study the enzyme activity of AspAT was reported to be 10-fold lower in the bundle sheath cells compared to the mesophyll cells (Denton et al. 2017). This same study also showed that while AspAT activity in the mesophyll cells increases along a leaf developmental gradient from least to most photosynthetically mature tissue, in the bundle sheath cells this pattern of AspAT activity is reversed with the lowest AspAT activity detected in the mature leaf tissue (Denton et al. 2017). Thus, it is currently not clear how aspartate is metabolised as part of the C_4 pathway in the bundle sheath cells of maize (see Table 4.1 for a summary of the evidence supporting this notion).

Table 4.1. Main lines of evidence supporting the notion that aspartate is decarboxylated by PEP-CK in the bundle sheath cells of maize, and that the enzyme activity of aspartate aminotransferase (AspAT) is not sufficient to support this reaction.

Year	Evidence for aspartate decarboxylation, PEP-CK activity, or AspAT activity	Experiment	Main finding
(Hatch 1971)	Aspartate decarboxylation	Radiolabelled CO_2 pulse-chase experiment in maize	Metabolite analysis of leaf tissue following exposure to $^{14}CO_2$ found that 25% and 75% of the carbon label is present in aspartate and malate, respectively.
(Chapman and Hatch 1981)	Aspartate decarboxylation	Monitoring of radiolabelled aspartate and malate decarboxylation in maize bundle sheath strands	Bundle sheath cells demonstrated to be able to decarboxylate aspartate. This reaction is dependent on α-ketoglutarate as would be expected in a typical AspAT reaction.
(Walker et al. 1997)	PEP-CK activity	Immunolocalisation of PEP-CK and measurement of enzyme activity in leaves of different C_4 plants	Demonstration that there is PEP-CK protein and enzyme activity localised in the bundle sheath cells of the maize.
(Wingler et al. 1999)	Aspartate decarboxylation, PEP-CK activity, and AspAT activity	Immunolocalisation of PEP-CK and AspAT, activity assays for PEP-CK and AspAT, and monitoring of radiolabelled aspartate and malate in isolated bundle sheath strands taken from the maize leaf	Demonstration that: ▪ PEP-CK protein and enzyme activity is present in maize bundle sheath strands. ▪ Aspartate can be decarboxylated by the bundle sheath strands under conditions that support PEP-CK activity. ▪ PEP is the main product of aspartate decarboxylation. ▪ No cytosolic AspAT protein detected in bundle sheath cells.
(Pick et al. 2011)	PEP-CK activity	Transcriptomic and metabolomic analysis along a developmental gradient of the maize leaf	PEP-CK activity along the developmental gradient of the maize leaf is shown to be similar to other core-C_4 enzymes.
(Arrivault et al. 2017)	Aspartate decarboxylation	Radiolabelled CO_2 pulse-chase experiment combined with mesophyll and bundle sheath cell metabolite analysis (in maize)	Metabolite labelling kinetics demonstrate that the total aspartate pool in the leaf participates in the C_4 cycle and that there is a sufficient concentration gradient of aspartate to support its movement from the mesophyll to the bundle sheath cells.
(Denton et al. 2017)	AspAT activity	Transcriptomic analysis of maize mesophyll and bundle sheath cells and AspAT enzyme activity assays	Demonstration that: ▪ AspAT transcript expression is enriched in mesophyll cells. ▪ There is a significant enrichment of AspAT enzyme activity in the mesophyll cells, with very little activity found in the bundle sheath cells.
(Schlüter et al. 2019)	AspAT activity	Phylogenetic analysis of AspAT and analysis of gene expression across a range of C_4 transcriptomic data sets.	Highlights that in two representative NAD-ME C_4 species, AspAT gene expression is evenly distributed between the mesophyll and bundle sheath cells. In NADP-ME maize however AspAT gene expression is restricted to the mesophyll cells.

CO_2 labelling studies have demonstrated that in maize at least some CO_2 is incorporated into aspartate in the mesophyll cells and that this aspartate moves to the bundle sheath cells and can supply CO_2 as part of the C_4 cycle (Chapman and Hatch 1981; Wingler et al. 1999; Arrivault et al. 2017). For aspartate to be decarboxylated by PEP-CK however, it must first be converted to oxaloacetate as this is the substrate of PEP-CK (Walker et al. 1997). It is possible that some aspartate is also decarboxylated by NADP-ME although this has not been shown to-date. In this case aspartate would still need to be converted to malate (which is the substrate of NADP-ME) and this conversion usually proceeds by first converting aspartate to oxaloacetate, and then subsequently converting oxaloacetate to malate. As described above however, the identity of the enzyme that catalyses the aspartate to oxaloacetate reaction in maize is unknown. It is also unknown in which subcellular compartment this reaction might occur. One possibility is that an as yet unidentified, non-canonical, aspartate aminotransferase exists to convert aspartate to oxaloacetate in the bundle sheath cell. Another possibility is that aspartate is converted to oxaloacetate, or malate, via a different biochemical pathway. An example of one such pathway is shown in Figure 4.3. In this figure, reactions 1 and 2 show the pathway of aspartate metabolism already described in which aspartate is converted to oxaloacetate or malate for decarboxylation by PEP-CK or NADP-ME. An alternative route is suggested by reactions 3 and 4. Here, aspartate is deaminated by an aspartate ammonia-lyase (reaction 3) forming the metabolic intermediate fumarate which could then be hydrolysed to form malate (reaction 4) by the enzyme fumerase. Malate could then be decarboxylated by NADP-ME or converted to oxaloacetate by malate dehydrogenase (reaction 2). Fumarase enzymes have been characterised in several plants including maize and function as part of the tricarboxylic acid (TCA) cycle in the mitochondria. To-date no plant enzyme has been described with aspartate ammonia-lyase activity, although such enzymes are known to exist in bacteria and animal species (EC 4.3.1.1). The pathway shown by reaction 3 and 4 in Figure 4.3 serves to illustrate that aspartate could be

metabolised via an alternative route by as yet undiscovered enzymes such as an aspartate ammonia-lyase.

To discover the enzyme(s) responsible for aspartate metabolism in the bundle sheath cells of maize, a bioinformatic screen had previously been conducted in the Kelly Lab. In this screen transcriptome data sets were leveraged, and a comparative genomics approach taken, to identify uncharacterised genes which encode proteins with possible aminotransferase function, and which are highly expressed in the bundle sheath cells of C_4 plants. This work led to the identification of the maize gene *ALUMINIUM-INDUCED-LIKE PROTEIN HOMOLOG 1* (*AILP1*) and the aim of this chapter therefore was to try and validate the function of *AILP1* and investigate its putative role in maize C_4 photosynthesis.

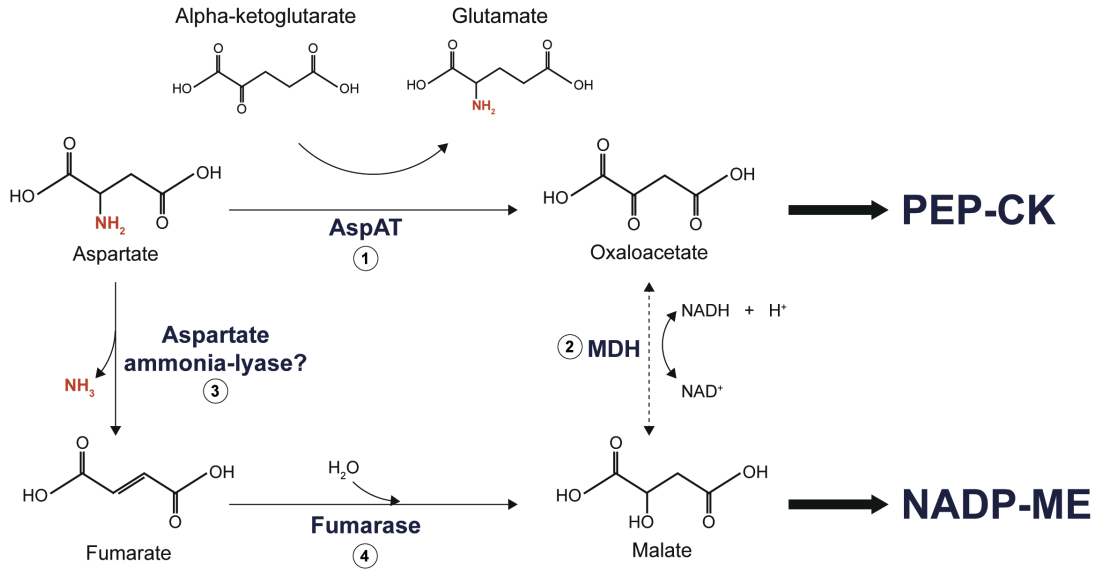


Figure 4.3. Canonical and alternative routes of aspartate metabolism. Scheme showing the canonical metabolic route by which aspartate is converted to oxaloacetate for decarboxylation by PEP-CK (reaction 1) or NADP-ME (reaction 2). A hypothetical alternative pathway is given by reactions 3 and 4. To date, no plant enzyme has been described with aspartate-ammonia lyase activity. **Abbreviations:** aspartate aminotransferase (AspAT), phosphoenolpyruvate carboxykinase (PEP-CK), NADP-dependent malic enzyme (NADP-ME), ammonia (NH_3), water (H_2O).

4.1.2 Aims and outline

The overarching aim of the project presented in this chapter was to experimentally characterise a putative aspartate aminotransferase enzyme (*AILP1*) that was identified through bioinformatic analysis in the Kelly Lab, and to determine whether it may function to process aspartate in the bundle sheath cells of maize. The first aim was to verify and extend the bioinformatic investigations already conducted into *AILP1*. The second aim was to verify the enzyme activity of recombinant AILP1 *in vitro*. The third aim was to verify the subcellular localisation of the AILP1 because this information is crucial in understanding how AILP1 may fit into the current model of C_4 photosynthesis in maize. Finally, the fourth aim was to investigate the function of AILP1 in the model plant *Arabidopsis* using a gene-silencing approach. Although it would have been preferable to pursue a transgenic approach in maize, the timeline of this project was not suitable for such experiments. With no prior knowledge of *AILP1* function however, *Arabidopsis* offered the shortest and least expensive route to acquiring insight about the function of this novel gene in plants.

4.2 Results

4.2.1 A bioinformatic search for candidate aminotransferases operating in maize bundle sheath cells identifies the gene *AILP1*

In our current model of C_4 photosynthesis in maize a fraction of the CO_2 that enters the mesophyll cell is combined with three-carbon PEP to form four-carbon aspartate. This aspartate moves to the bundle sheath cells where it must be converted to either oxaloacetate or malate for decarboxylation by PEP-CK or NADP-ME respectively. The enzyme responsible for the catalysis of either of these reactions in maize has however not been identified. To address this, a bioinformatic screen was conducted previously in the Kelly Lab to identify novel enzymes which could catalyse the necessary aminotransferase or ammonia-lyase reaction with aspartate in the bundle sheath cells of maize.

This bioinformatic work led to the identification of the maize gene *ALUMINIUM-INDUCED-LIKE PROTEIN HOMOLOG 1* (*ZmAILP1*, Zm00001d023376). This gene was found to be preferentially expressed in bundle sheath cells with expression levels comparable to other core C_4 genes (Table 4.2). Another two genes from the same gene family (hereafter referred to as the *AILP1* gene family) were also found to be expressed to appreciable levels in the bundle sheath cells (Table 4.2). In addition to enrichment of gene expression in the bundle sheath, *AILP1* gene sequences were found to be most closely related to those of asparagine synthetase (Figure 4.4 A), an enzyme that catalyses the ATP-dependent conversion of aspartate into asparagine using glutamine or ammonia as the donor of an amino group (EC:6.3.5.4). Thus, the closest relatives of the *AILP1* gene family bind to and process aspartate.

Table 4.2. mRNA abundance and cell type preferential expression for *AILP1* and two other leaf-expressed *AILP1* paralogs. For comparison, mRNA data are provided for two core-C_4 enzymes (PEPC and NADP-ME) as well as the most highly expressed isoform of alanine aminotransferase (AlaAT) and aspartate aminotransferase (AspAT) in the maize leaf. mRNA levels are from an RNAseq experiment conducted in Chang et al. 2012. **Abbreviations:** transcripts per million (TPM), mesophyll cell (M), bundle sheath cell (BS), phosphoenolpyruvate carboxylase (PEPC), NADP-dependent malic enzyme (NADP-ME), aspartate aminotransferase (AspAT), alanine aminotransferase (AlaAT).

Gene	Accession	M mRNA (TPM)	BS mRNA (TPM)	Significant difference ($P < 0.01$) in M/BS expression?
AILP1	Zm00001d023376	2,549	9,522	Yes
AILP1 paralog	Zm00001d026632	346	786	No
AILP1 paralog	Zm00001d033983	15	706	Yes
PEPC	Zm00001d046170	8,419	444	Yes
NADP-ME	Zm00001d000316	66	11,499	Yes
AspAT	Zm00001d018386	1,533	53	Yes
AlaAT	Zm00001d030557	303	203	Yes

To date, no *AILP1* gene from any species (plant or otherwise) has been experimentally characterised; the only available annotation of the protein sequence is that it contains a single domain of unknown function (DUF3700, Pfam domain ID PF12481) that encompasses the majority of the protein sequence (Figure 4.4 B). This Pfam domain is however listed as being closely related to glutamine amidotransferase enzyme domains and thus it was inferred that *AILP1* may also encode an aminotransferase enzyme. Concomitantly, the closely related asparagine synthetase enzyme (reference *Arabidopsis* sequence: AT3G47340) catalyses the removal of an amido group from glutamine and the subsequent transfer of that group to aspartate thus creating asparagine. Characterised asparagine synthetases contain two domains: a glutamine amidotransferase domain (GATase_7, Pfam domain ID PF13537) encoded by the N-terminal half of the protein sequence and an asparagine synthetase domain (Ans_synthase, Pfam domain ID

PF00733) encoded by the C-terminal half (Figure 4.4 B). These two domains work independently of each other; the GATase_7 domain catalyses the liberation of an amido group from glutamine and the Ans_synthase domain catalyses the ATP-dependent transfer of that amido group onto aspartate (Lomelino et al. 2017). Further phylogenetic interrogation of the relationship between AILP1 and asparagine synthetase protein sequences revealed that AILP1 shares homology with only the N-terminal glutamine amidotransferase domain of asparagine synthase (Figure 4.4 B). This homology suggests that AILP1 may function to catalyse the removal of ammonia groups (as an ammonia-lyase) from amino acids, rather than to transfer ammonia groups between amino acids and other organic compounds (as in amino- and amido- transferases). If this is the case, it is not clear what the substrate of AILP1 would be given that the amino acid residues required for glutamine binding in the GATase7 domain of asparagine synthase are not conserved in AILP1 (Figure 4.4 C). Therefore, if AILP1 does function to liberate ammonia derivatives from amino acids, glutamine may not be the donor.

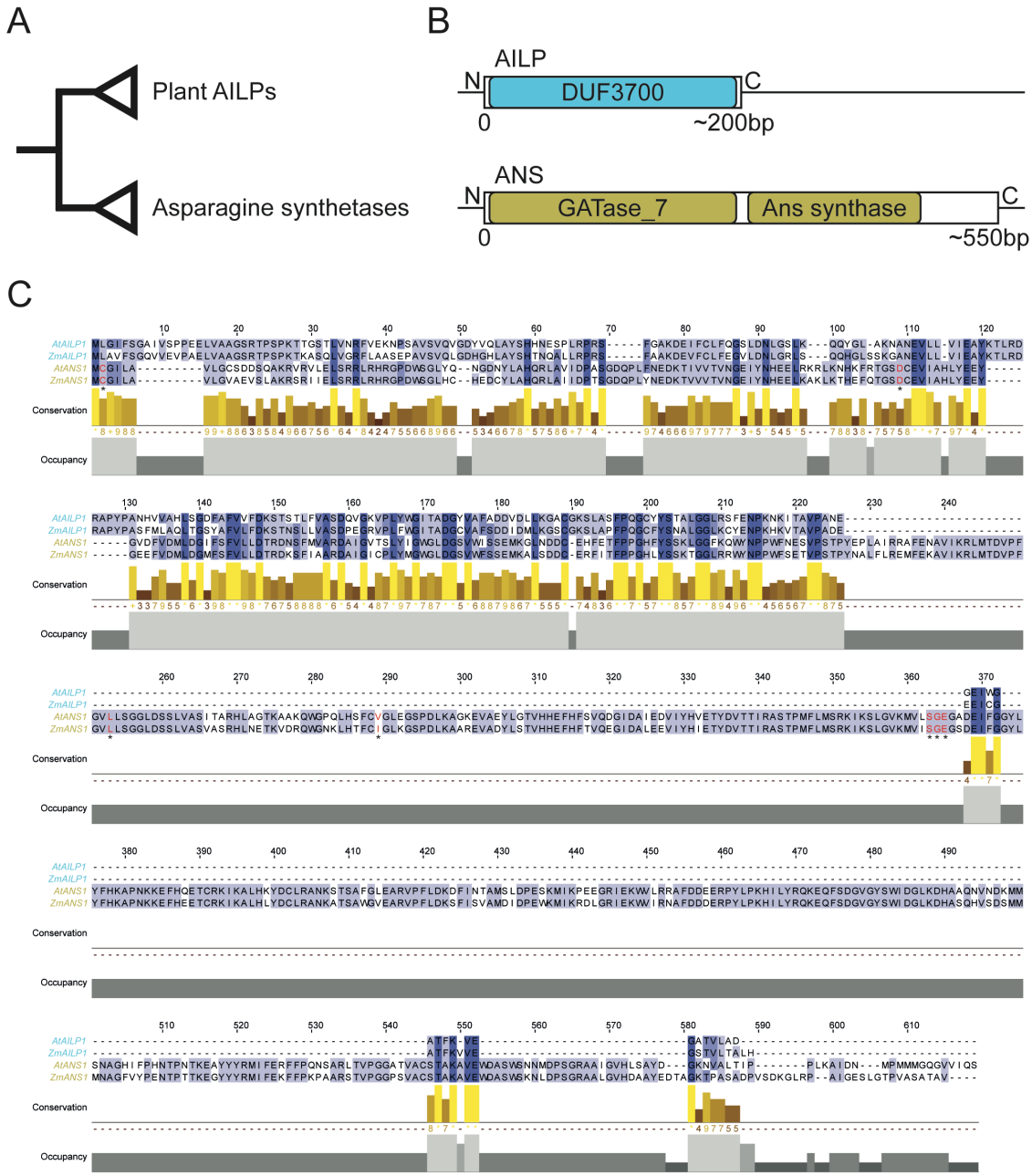


Figure 4.4. Protein sequences of ALUMINIUM INDUCED PROTEIN 1 (AILP1) are homologous to the N-terminal half of asparagine synthetase (ANS). A) A simplified phylogenetic tree illustrating that the *AILP1* gene family is most closely related to a gene family of asparagine synthetases. **B)** Alignment of the protein sequences and functional domains of AILP1 and ANS illustrating the homologous protein region. Members of the AILP1 protein family contain a domain of unknown function (DUF3700) which is homologus to the glutamine amidotransferase domain (GATase_7) of ANS. **C)** Multiple sequence alignment created in Jalview v2.11.0 of two AILP1 proteins from *Arabidopsis* (*At*AILP1, AT5G19140) and maize (*Zm*AILP1, Zm00001d023376) and two asparagine synthetase proteins from *Arabidopsis* (*At*ANS1, AT3G47340) and maize (*Zm*ANS1,

Zm00001d045675). The top two protein sequences in the alignment are AILP1 sequences and the bottom two are ANS sequences. The amino acid residues are shaded according to the percentage conservation at each site, with the darkest shading representing 100% conservation. Characterised functional residues of the two asparagine synthetase sequences are coloured red and indicated with a star below the residues at that position. The conservation histogram reflects the conservation of the physico-chemical properties for each column of the alignment with identical residues at a site scoring highest, and amino acid substitutions in the same physico-chemical class scoring next highest. The score for each column is shown below the histogram: conserved columns score of 11 and are indicated by '*', while columns with mutations in the same physico-chemical class are indicated with a '+'. The occupancy histogram reflects the number of residues aligned at each column in the multiple sequence alignment.

4.2.2 *ZmAILP1* belongs to a highly conserved plant specific gene family containing many paralogous genes

To investigate the evolution of *AILP1* in the plant kingdom a phylogenetic analysis was undertaken using the genomes of 54 diverse plant species, spanning the breadth of the plant kingdom, including green algae (the phylogenetic tree for this set of species is shown in Appendix 4.1). This analysis found *AILP1* genes in every genome investigated, with most species possessing several copies of the gene (in total 347 *AILP1* genes were found across the 54 species). Construction of a phylogenetic tree from AILP1 protein sequences (primary protein transcripts) showed three distinct clades of AILP1 proteins in land plants which are estimated to have arisen from a series of gene duplications after the divergence of land plants and green algae (Figure 4.5). The majority of land plants were found to possess at least one protein-encoding gene from each of these three clades.

Despite the ubiquitous occurrence in plants, no *AILP1* genes were found in the genomes of species from outside of the Viridiplantae (plants and green algae). A BLAST search of

an *Arabidopsis AILP1* (AT5G19140) against all genomes *not including Viridiplantae* only returned asparagine synthase genes. Similarly, a hidden Markov model generated from an alignment of AILP1 protein sequences from the 54 plant species was used to conduct a search for homologous proteins in the genomes of the following non-plant model species: *Drosophila melanogaster*, *Homo sapiens*, *Mus musculus, Saccharomyces cerevisiae*, *Schizosaccharomyces pombe* and *Escherichia coli*. In all genomes the only significant hits returned were for asparagine synthetase sequences. Together therefore AILP1 appears to be a Viridiplantae-specific protein which evolved from a duplication or partial duplication of asparagine synthetase in an ancestor of land plants and green algae.

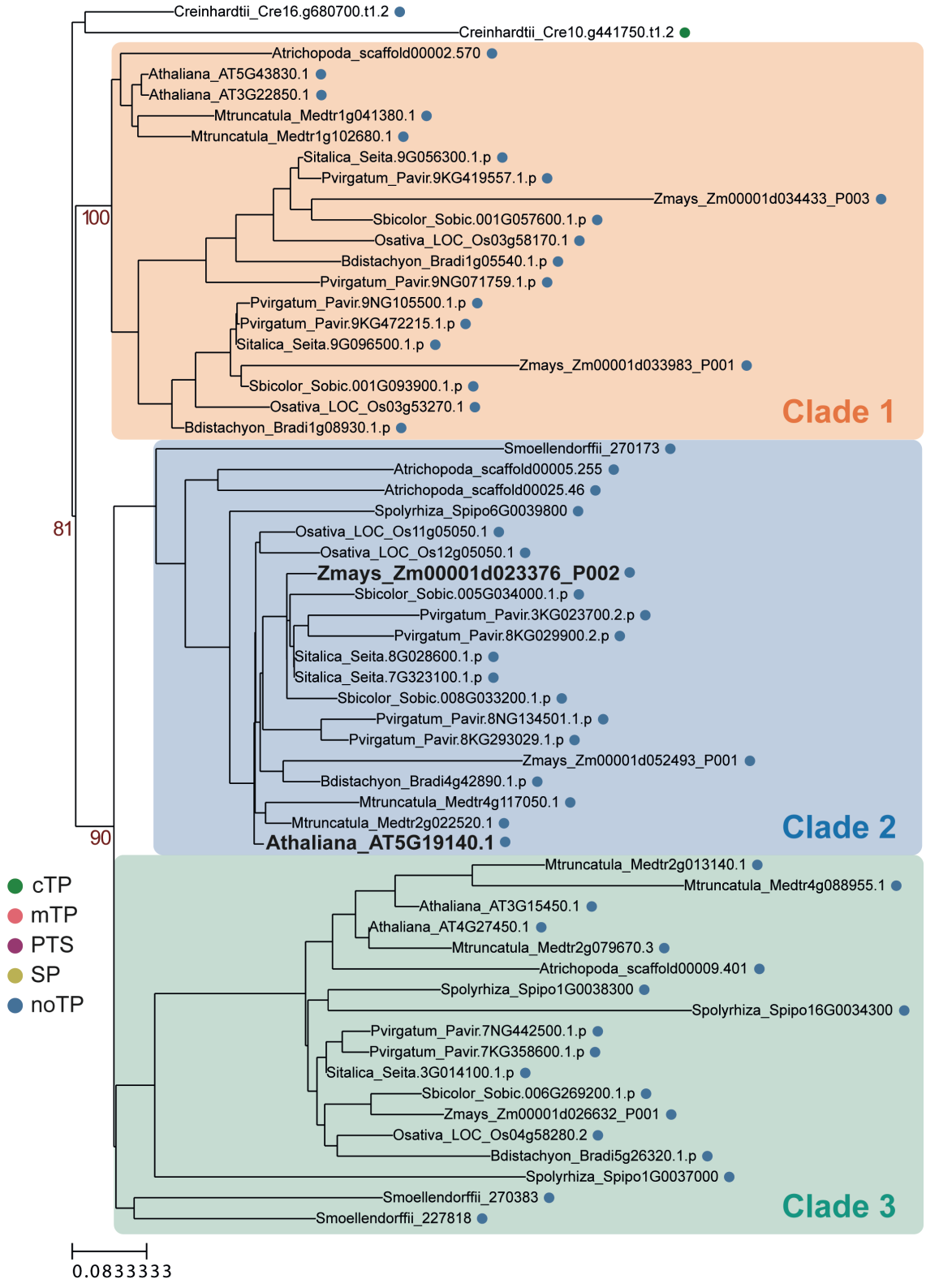


Figure 4.5. Maximum likelihood phylogenetic tree of the *AILP1* gene family. The coloured circles next to gene names denote subcellular localisation as predicted by TargetP v2.0. The coloured boxes denote distinct clades of *AILP1* genes in land plants. The *AILP1* gene with the highest leaf expression in maize and *Arabidopsis* is present in

clade 2 and indicated by larger text and bold font. The tree was constructed with IQ-TREE using the substitution model JTTDCMut+I+G4, and bootstrap values (shown for the major bipartitions of the three clades) are from UFBoot. Scale bar represents amino acid substitutions per site. **Abbreviations**: chloroplast transit peptide (cTP), mitochondrial target peptide (mTP), secretory peptide (SP), peroxisomal target signal (PTS), no predicted target signal (noTP).

4.2.3 Investigation of *AILP1* gene expression across species shows upregulation in the photosynthetically active tissue of C_4 species

The initial bioinformatic screen for maize genes encoding enzymes with functional domains related to those domains found in aminotransferases identified *ZmAILP1* in part because this gene has high mRNA expression in the bundle sheath cells of maize (Table 4.2 and Figure 4.6 A). In order to verify this finding, *AILP1* expression was investigated in another maize transcriptome data set that compares gene expression between mesophyll and bundle sheath cells along a developmental gradient of the maize foliar leaf (Tausta et al. 2014). The leaf sections sampled in this data set fall either side of the leaf sink-source transition zone so that the expression of genes can be compared between immature, maturing and mature photosynthetic tissue. As expected for a gene actively involved in photosynthesis the expression of *ZmAILP1* (Zm00001d023376) increased along the photosynthetic developmental gradient with gene expression peaking in the most photosynthetically active tissue (Figure 4.6 B). This pattern of expression is observed in both mesophyll and bundle sheath cells, but with higher overall expression in the bundle sheath cells (Figure 4.6 B). The finding that *ZmAILP1* expression increases along a photosynthetic developmental gradient was further corroborated in another transcriptomic data set which compared gene expression between nascent maize leaves, still in the primordia stage of development, and mature leaves which had fully expanded (Wang et al. 2013). In this study, samples were taken from both the foliar and husk leaves of maize.

The reason for studying expression between these two leaf types is that while the foliar leaf in maize carries out C_4 photosynthesis, the husk leaf does not possess Kranz anatomy and thus primarily carries out C_3 photosynthesis. As there is a smaller concentration of photosynthetically active bundle sheath cells in the husk leaf, differential expression between the foliar and husk leaf is an indicator of differential relative abundance in bundle sheath and mesophyll cells. In both leaf types, *ZmAILP1* expression was again found to peak in the most photosynthetically active tissue (Figure 4.6 C), however expression was around 2-fold higher in the foliar compared to the husk leaf. Together, the expression profile of *ZmAILP1* in these three maize leaf data sets (Figure 4.6 A-C) is compatible with a potential role of *Zm*AILP1 in the bundle sheath cells as part of C_4 photosynthesis. Furthermore, these data sets show that the majority of *AILP1* expression in the maize leaf is attributable to a single AILP1 gene (*ZmAILP1*, Zm00001d023376) and therefore this isoform was chosen to be the focus of *in vitro* assays to determine AILP1 enzyme activity.

The potential role of AILP1 in C_4 photosynthesis can be further explored by leveraging transcriptomic data sets that specifically compare leaf gene expression between C_3 and C_4 species. In one such data set, leaf gene expression was compared between closely related C_3, C_3-C_4 intermediate and C_4 species in the genus *Flaveria* (Gowik et al. 2011). Investigation of *AILP1* expression in this data set revealed a higher abundance of *AILP1* transcripts in the C_4 *Flaveria* species compared to the C_3 *Flaveria* species, with an intermediate level of expression in the C_3-C_4 intermediate species (Figure 4.6 D). This pattern of expression among *Flaveria* species is commonly seen for genes known to have been recruited to a C_4 function during the evolution of C_4 photosynthesis. Indeed, in the original analysis of this *Flaveria* transcriptomic data set, *AILP1* was identified as part of a list of significantly upregulated C_4-related genes (Gowik et al. 2011). In another comparative transcriptome analysis of 17 C_3 and 30 C_4 eudicots (Kelly et al. 2017) *AILP1* transcripts were also found to be more highly abundant in the leaves of the C_4 species compared to C_3 species (Figure 4.6 E). This pattern is seen whether all *AILP1* transcripts

are considered or just those that are orthologous to the highly expressed *ZmAILP1* transcript (clade 2 in Figure 4.5). The pattern of expression is lost and overall read counts are much lower across species when only those transcripts that are *not* orthologous to *ZmAILP1* are considered. This suggests that *AILP1* genes from clade 2 have been upregulated in the leaf among different C_4 species. It should be noted that caution needs to be applied in the interpretation of these expression data sets; while they offer clues about the relevance of AILP1 function to C_4 photosynthesis, nothing is known about the post-translational modification of AILP1, and expression levels may not equate to levels of functional protein. Having said this, to see a C_4-like pattern of expression for *AILP1* across multiple data sets was promising for a putative C_4 function for this gene and warranted further investigation of its encoded protein.

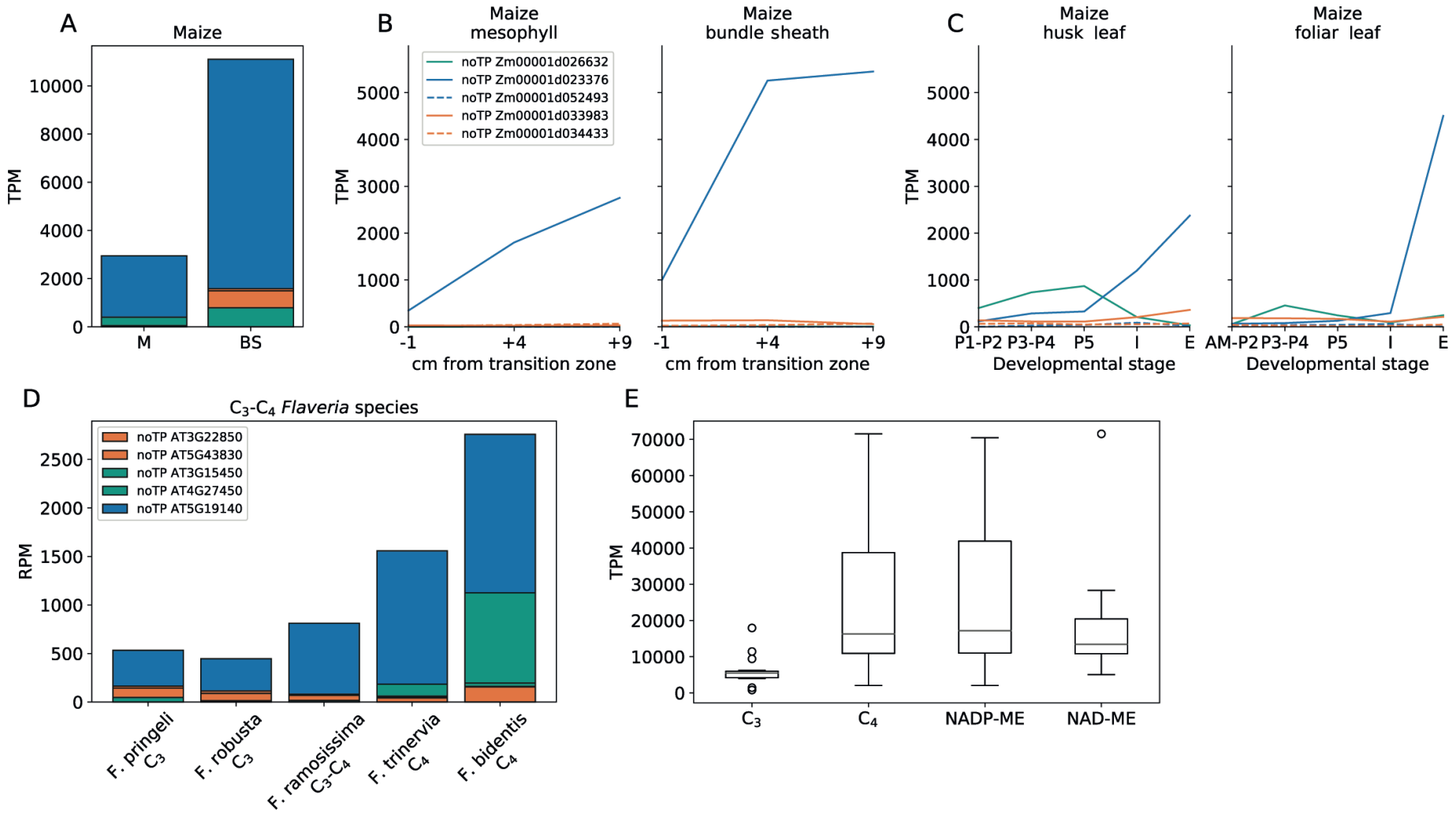


Figure 4.6. Expression of *AILP1* in selected transcriptome data sets. Figure legend continued on the next page.

Figure 4.6. (continued) The colours used in the plots indicate which clade the gene is from in the phylogenetic tree in Figure 4.5. Note that the legend is shared for the maize plots (A-C). Transcript amounts in **A)** mesophyll (M) and bundle sheath (BS) cells from mature maize leaves (Chang et al. 2012), **B)** mesophyll and bundle sheath cells along a developmental gradient of the maize leaf (Tausta et al. 2014), here the x-axis denotes distance from the source-sink transition zone of the maize leaf with -1cm representing the least mature tissue and +9cm representing the most mature tissue, and **C)** developing maize husk and foliar leaves (Wang, Kelly, et al. 2013), **D)** Comparison of *AILP1* expression in the leaves of C_3, C_3-C_4 intermediate, and C_4 *Flaveria* species (Gowik et al. 2011). Transcripts in this data set were mapped to the *Arabidopsis* genome hence the *Arabidopsis* gene accessions are shown in the legend. **E)** Comparison of total *AILP1* expression in selected C_3 (N=17) and C_4 (N=30) eudicot species (Kelly et al. 2017). Also shown is the expression in C_4 species broken down by C_4 subtype: NADP-ME and NAD-ME. Transcripts were mapped to the *Arabidopsis* genome and so for each species the total TPM was summed for each *Arabidopsis AILP1* gene. **Abbreviations:** transcripts per million (TPM), reads per million mapped reads (RPM), no predicted target signal according to TargetP v2.0 (noTP), NADP-dependent malic enzyme (NADP-ME), NAD-dependent malic enzyme (NAD-ME), apical meristem (AM), plastochron (P), immature leaf (I), fully expanded leaf (E).

4.2.4 No amino-lyase enzyme activity could be detected for purified recombinant *Zm*AILP1

The AILP1 protein sequence is homologous to the glutamine amidotransferase (GATase) domain of asparagine synthetase which catalyses the removal of the ammonia-dervied side group (amido group) from the amino acid glutamine. The homology of AILP1 to this GATase domain suggests AILP1 may also function to catalyse the removal of ammonia-derived groups from amino acids. In particular, I wanted to know if AILP1 could catalyse the removal of an amino group (yielding an ammonia molecule) from aspartate which would support a role for AILP1 in the metabolism of aspartate to oxaloacetate in the bundle sheath cells of maize. To do this, purified *Zm*AILP1 was incubated with a range of amino

acids and an ammonia assay kit was used to provide a quantitative measure of ammonia liberation. A commercial ammonia assay kit (Sigma #AA0100) was used for this. The main reagent of the kit contained α-ketoglutaric acid (KGA), NADPH and L-glutamate dehydrogenase (GDH). When this reagent is mixed with a sample containing ammonia (NH_4^+), NADPH is oxidised and this oxidation can be measured in a spectrophotometer as a decrease in absorbance at 340 nm. This reaction is as follows:

$$NH_4^+ + KGA + NADPH \xrightarrow{GDH} L\text{-Glutamate} + NADP^+ + H_2O$$

The decrease in absorbance at 340 nm is proportional to the ammonia concentration in the sample. An assay was thus set up in which purified recombinant AILP1 protein was incubated with different amino acids in an assay buffer similar to that previously used in enzyme assays of plant asparagine synthetases (Todd et al. 2008; Duff et al. 2011; Xu et al. 2018). After incubation for approximately 30 minutes, the substrates of the ammonia assay kit were added (not including GDH). An absorbance reading at 340 nm was taken (T0 measurement) and then GDH was added. After another five minute incubation, a second absorbance reading was taken (T1 measurement). As a positive control for this assay, recombinant *Zm*AILP1 was replaced by the enzyme glutaminase (Sigma #G8880) which catalyses the removal of ammonia from glutamine and therefore provides a positive control for the ability of this assay to detect the enzymatic liberation of ammonia from amino acids.

In order to carry out this enzyme assay I sought to express *Zm*AILP1 in *Escherichia coli* and purify the protein using affinity chromatography. In the first instance, I created an *E. coli* expression construct containing the *ZmAILP1* gene sequence fused to a C-terminal polyhistidine-tag (His-tag). The expressed protein from this construct (*Zm*AILP1::Hisx6) was however found to accumulate entirely in the insoluble fraction of the cell lysate (Figure

4.7) indicating that *Zm*AILP1 protein was aggregating into inclusion bodies. To try and increase protein solubility a range of *E. coli* expression conditions were trialed, including lower temperatures and lower concentrations of the chemical inducer isopropylthio-β-galactoside (IPTG). In the end however, the solubility could not be improved. Thus, a second expression construct was created containing the *ZmAILP1* gene sequence fused at the N-terminal to a gene sequence encoding a maltose binding protein (MBP). The MBP is a much larger tag than a His-tag (383 amino acids compared to 6 amino acids) and is commonly used to increase protein solubility in *E. coli* expression systems and enhance protein production (Lebendiker and Danieli 2011). Using the MBP-fusion construct (MBP::*Zm*AILP1) I was able to express soluble protein (Figure 4.7). Subsequently the protein could be purified by affinity chromatography using an amylose resin which the maltose binding protein has affinity for. Elution of the protein from the resin was conducted using 10 mM maltose and successful elution was confirmed by sodium dodecyl sulphate polyacrylamide gel electrophoresis (SDS-PAGE) (Figure 4.8). The elution fractions were then pooled and used directly in the enzyme assay described above using 2 mM glutamine and 2 mM aspartate as a substrate. In the end however, I did not detect any ammonia liberation (Figure 4.9) and thus the function of *Zm*AILP1 remains inconclusive.

A disadvantage of using an MBP-fusion is that the MBP may disrupt protein function due to its size which in this case is larger than the *Zm*AILP1 protein itself. Thus, an additional strategy was taken to try and express the soluble *Zm*AILP1 bound to the His-tag (*Zm*AILP1::Hisx6). This strategy made use of a cell-free expression system (myTXTL, Arbor Biosciences #505024). As protein production often proceeds at a slower rate in cell-free systems, protein folding is promoted and the formation of inclusion bodies prevented (Fakruddin et al. 2013; Smolskaya et al. 2020). The myTXTL expression system comes supplied with a master mix containing *E. coli* cell extract, amino acids and an energy buffer, and protein production is initiated with the addition of plasmid DNA. Using this approach I was able to express and purify soluble recombinant *Zm*AILP1 protein (*Zm*AILP1::Hisx6).

Unfortunately, the protein was consistently lost (as monitored by SDS-PAGE) during buffer exchange prior to the enzyme assay. Buffer exchange was conducted using spin columns and it is likely that the protein became ‘stuck’ to the column membrane due to the physico-chemical properties of the protein or aggregation of the protein following purification. No such issue was observed when using the maltose-binding protein fusion.

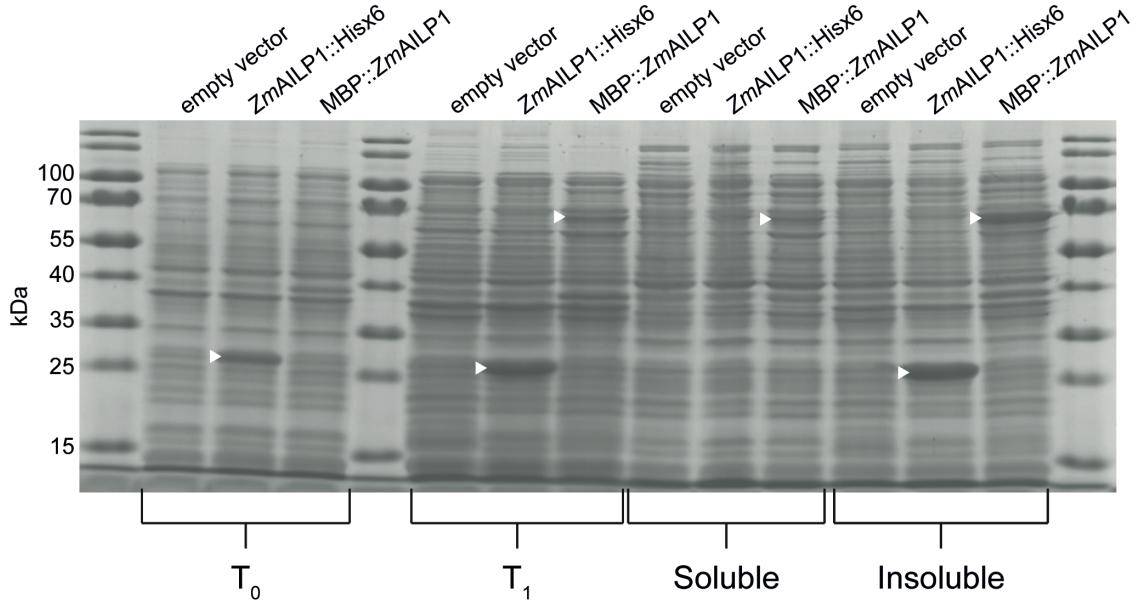


Figure 4.7. SDS-PAGE of protein fractions from *E. coli* expressing *Zm*AILP1 fused to a C-terminal His-tag (*Zm*AILP1::Hisx6, expected protein size 27 kDa) or an N-terminal maltose-binding protein (MBP::*Zm*AILP1, expected protein size 68 kDa). An empty His-tag vector was used as a negative control (empty vector). Fractions are shown prior to induction (T0) and 4-5 hours after induction (T1) with IPTG. The fraction at T1 was then separated by centrifugation into the soluble and insoluble fraction. The white arrows indicate visible fusion protein. kDa = kilodaltons.

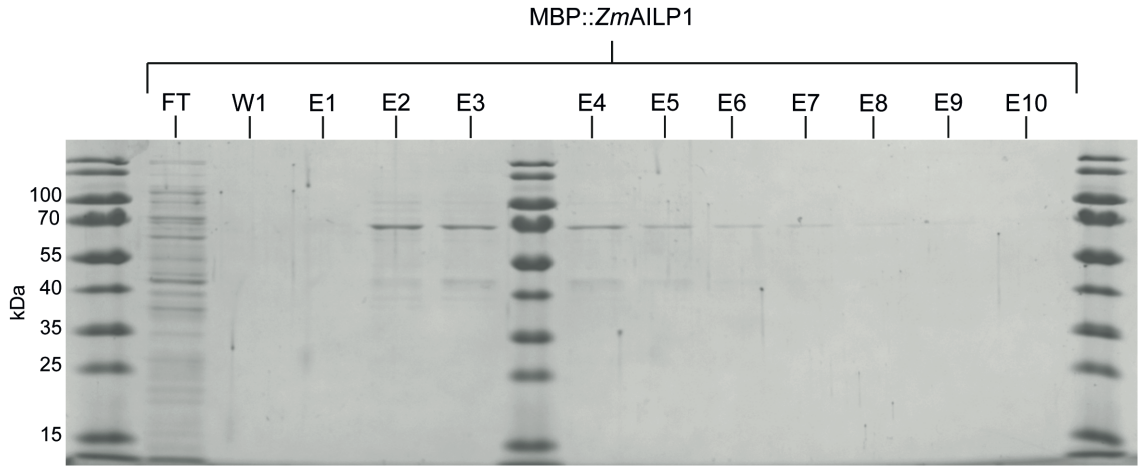


Figure 4.8. Protein purification of MBP::*Zm*AILP1 monitored on an SDS-PAGE. The soluble protein fraction from *E. coli* was loaded onto an amylose column and, after incubation, an initial flow through (FT) fraction was collected. The column was then washed to remove unbound protein and the wash (W) fraction collected. Lastly, protein bound to the column was eluted (E) over several fractions in a buffer containing 10mM maltose. Expected protein size of MBP::*Zm*AILP1 is 88 kilodaltons (kDa).

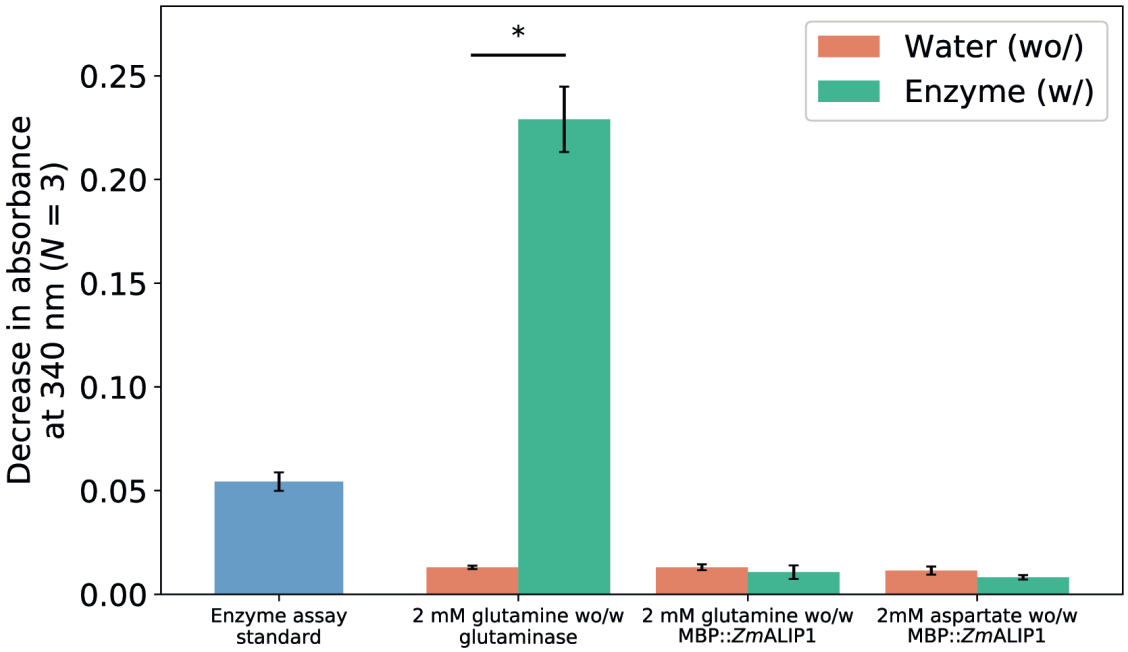


Figure 4.9. Ammonia liberation was not detected after incubation of MBP::*Zm*AILP1 with glutamine or aspartate. Enzyme (glutaminase or MBP::*Zm*AILP1) or water was added to an assay buffer containing 2 mM glutamine or 2 mM aspartate and incubated for approximately 15 minutes. The reagents of the ammonia assay kit were then added *except* for glutamate dehydrogenase and an absorbance reading at 340 nm was taken. Glutamate dehydrogenase was then added, and another absorbance reading taken. The change in absorbance is shown in this plot. A decrease in absorbance indicates ammonia is present in the sample. The asterisk indicates a significant difference between water and enzyme samples (two-tailed t-test, $P < 0.05$). The enzyme assay standard is the positive control provided with the ammonia assay kit.

4.2.5 *Zm*AILP1 and *At*AILP1 localise to the cytosol in *Arabidopsis* protoplasts

As well as being split between two different cell types, the C_4 photosynthetic pathway is also split between different organelles of the cell. If *Zm*AILP1 is involved in aspartate metabolism in the bundle sheath cell it will be important to know where in the cell the enzyme is localised so that the route of C_4 metabolite flow can be deciphered. Target signal prediction using TargetP did not identify an organellar target signal in the protein sequence of *Zm*AILP1 suggesting that the protein is localised to the cytosol. To confirm or refute this predicted targeting, the subcellular localisation of *Zm*AILP1 was assessed in *Arabidopsis* protoplasts. As the work in this chapter also sought to investigate AILP1 function in *Arabidopsis*, the subcellular localisation of the protein encoded by the *Arabidopsis* AILP1 gene targeted by gene-editing was also assessed. To do this, AILP1 protein from *Maize* and *Arabidopsis* was fused at the C-terminus to a green fluorescent protein (GFP) and transiently expressed in *Arabidopsis* mesophyll protoplasts. As shown in Figure 4.10 and Figure 4.11, both fusion proteins (*Zm*AILP1::GFP and *At*AILP1::GFP) were found to accumulate in the cytosol of protoplasts, confirming the predicted cytosolic localisation of the *Zm*AILP1 and *At*AILP1.

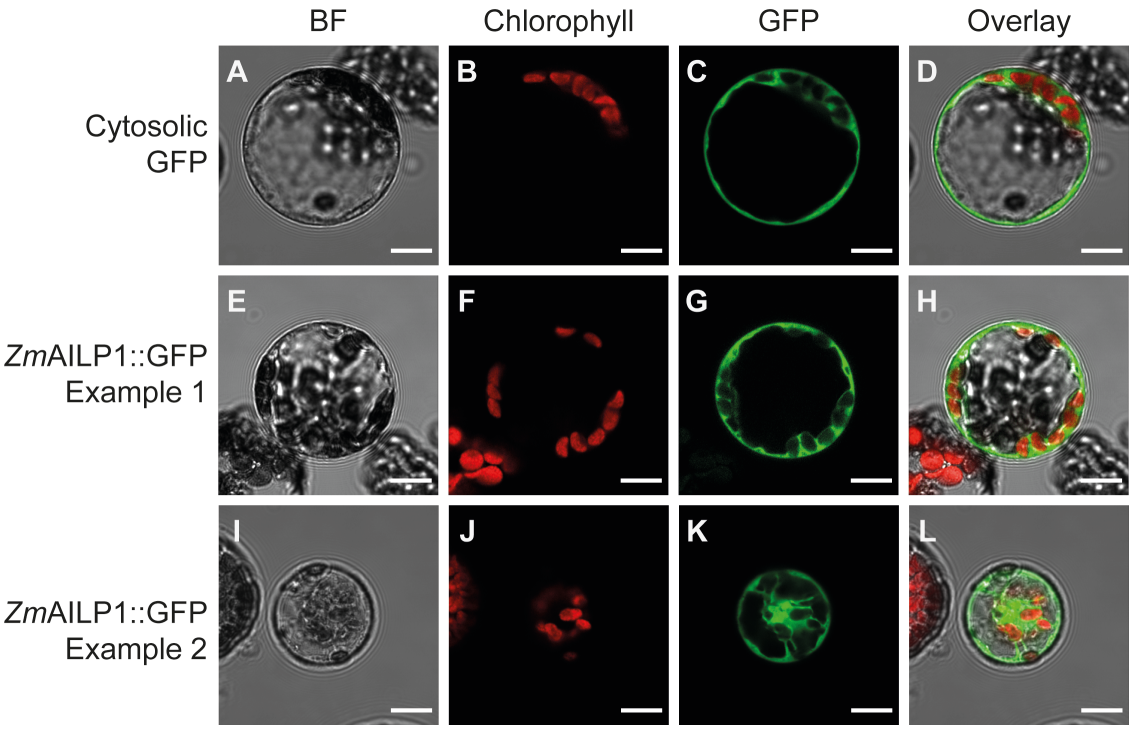


Figure 4.10. Subcellular localisation of *Zm*AILP1 (Zm00001d023376) fused to a green fluorescent protein (*Zm*AILP1::GFP) in *Arabidopsis thaliana* protoplasts. A-D) Representative localisation of a cytosolic GFP. **E-L)** Localisation of *Zm*AILP1::GFP. Images in the first column show the bright field image (**A, E, I**), the images in the second column show the chlorophyll autofluorescence (**B, F, J**), images in the third column show GFP fluorescence (**C, G, K**) and images in the fourth column show the overlaid images corresponding to bright field chlorophyll and GFP images (**D, H, L**). Scale bar represents 10 µm.

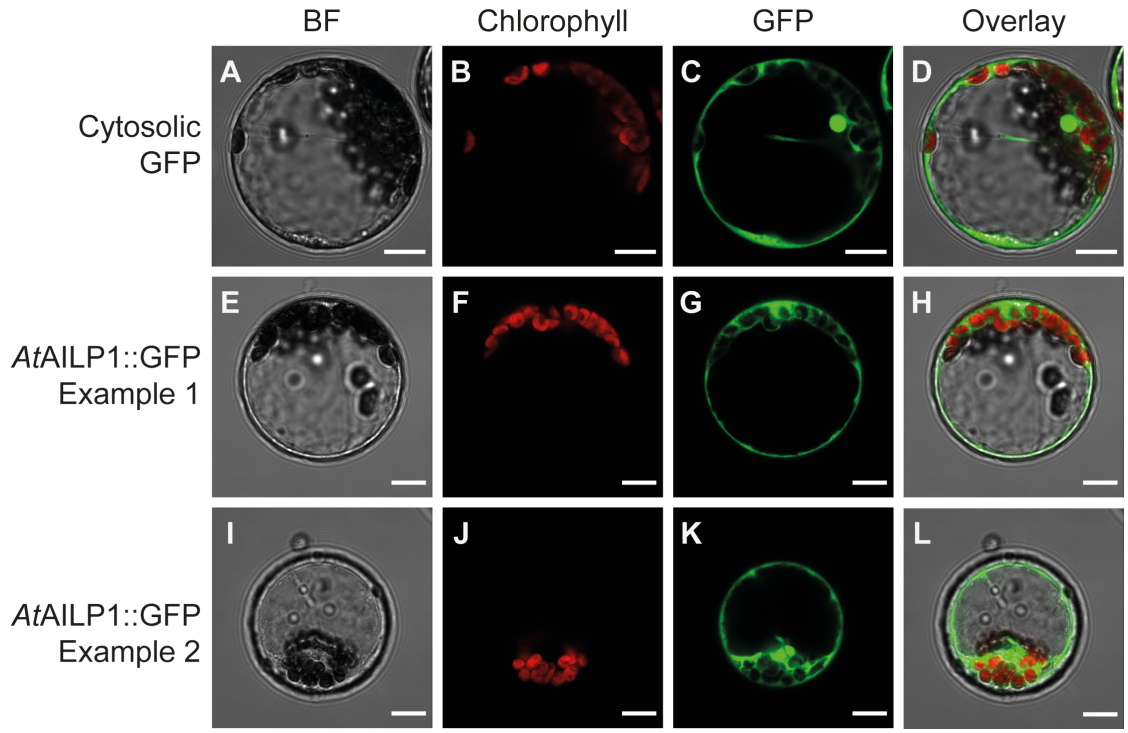


Figure 4.11. Subcellular localisation of *At*AILP1 (AT5G19140) fused to a green fluorescent protein (*At*AILP1::GFP) in *Arabidopsis thaliana* protoplasts. A-D) Representative localisation of a cytosolic GFP. **E-L)** Localisation of *At*AILP1::GFP. Images in the first column show the bright field image (**A, E, I**), the images in the second column show the chlorophyll autofluorescence (**B, F, J**), images in the third column show GFP fluorescence (**C, G, K**) and images in the fourth column show the overlaid images corresponding to bright field chlorophyll and GFP images (**D, H, L**). Scale bar represents 10 µm.

4.2.6 The generation of loss-of-function *ailp1* mutants in Arabidopsis using CRISPR-Cas9

There has been no previous experimental characterisation of any *AILP1* gene in any species and at the time of writing there were no suitable tDNA insertion lines for *AILP1* in the relevant maize and *Arabidopsis* stock centres. Thus, to investigate the function of *AILP1* more generally in plants a gene editing approach was taken to disrupt gene function in the model plant *Arabidopsis thaliana* (*Arabidopsis*). In *Arabidopsis* there are five *AILP1* genes but only one, AT5G19140 (hereafter called *AtAILP1*), is an ortholog of the maize *AILP1* with high bundle sheath cell expression. Interrogation of an *Arabidopsis* leaf transcriptome data set showed that this *Arabidopsis* ortholog, *AtAILP1*, is highly expressed in the mature leaf (Figure 4.12 A). Similarly to the maize ortholog, expression was also found to increase with leaf (and therefore photosynthetic) maturity, although expression also continued to increase into scenescence (Figure 4.12 A). In the mature leaf tissue, the majority of total *AILP1* gene expression is attributable to the one gene, *AtAILP1* (AT5G19140), and thus this gene was chosen to be the focus of mutagenesis work. Interestingly, in a second *Arabidopsis* transcriptomic data set comparing whole-leaf and bundle sheath cell specific gene expression, *AtAILP1* expression was found to be significantly higher in the bundle sheath cells (Figure 4.12 B). *AtAILP1* expression was still appreciable in the whole-leaf sample of cells but this finding suggests that the preferential accumulation of *AtAILP1* transcripts in the bundle sheath cells may be evolutionarily conserved between monocots and eudicots.

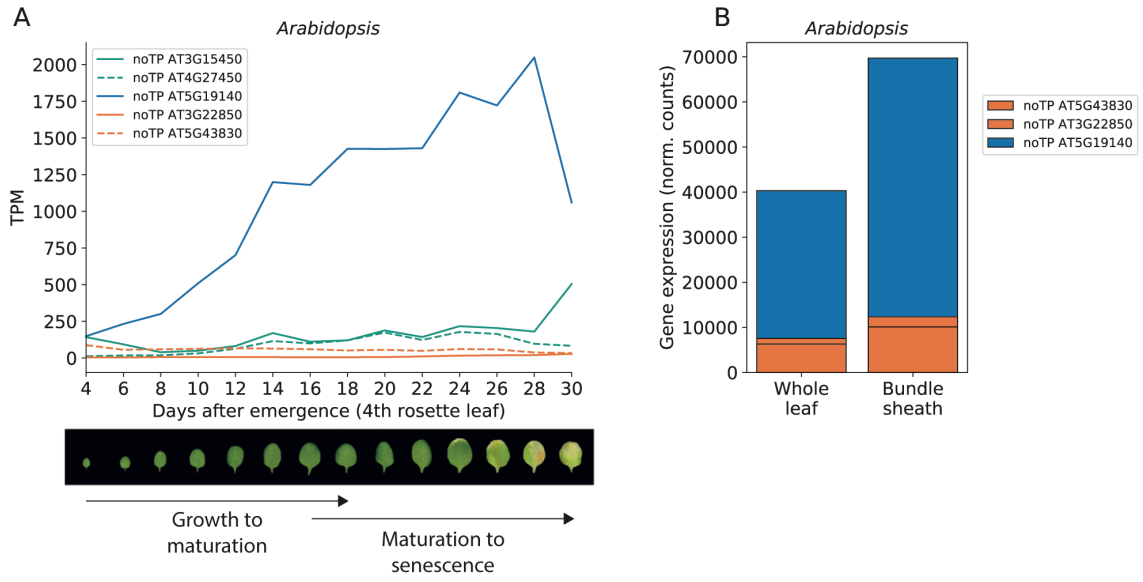


Figure 4.12. mRNA abundance of *AILP1* genes in *Arabidopsis* leaves. The colours used in the plots indicate which clade the gene is from in the phylogenetic tree in Figure 4.5. **A)** Transcript abundance in rosette leaves along a developmental gradient (Woo et al. 2016). Leaf photos taken from (Woo et al. 2016). **B)** Comparison of normalised (norm.) mRNA abundance between whole leaf and bundle sheath cell samples taken from the mature *Arabidopsis* leaf. In this study, whole-leaf and bundle sheath cell localised ribosomes were isolated and resident mRNA abundance measured (Aubry et al. 2014). **Abbreviations:** transcripts per million (TPM), no predicted target signal according to TargetP v2.0 (noTP).

To generate loss-of-function mutations in *AtAILP1* a CRISPR-Cas9 approach was taken. Here, a construct was created containing three short guide RNAs (sgRNAs) each driven by the *Arabidopsis RNA polymerase III* promoter (*AtU6-26*). This construct also contained the *Cas9* gene driven by the promoter of the egg-specific *Arabidopsis YAO* gene (Yan et al. 2015) and the *bar* gene to enable selection of transformants using the herbicide Basta. A map of the parts in this construct is shown in Figure 4.13 A. The three sgRNAs target the first and second exon of *AtAILP1*. A simplified illustration of the gene region targeted by each sgRNA is given in Figure 4.13 B and the specific location is shown in Appendix 4.2.

Approximately 45 *Arabidopsis* Col-0 plants were transformed by floral dipping with the CRISPR construct described above. Successfully transformed T1 generation plants (T1 plants) were identified through Basta selection and *AtAILP1* mutations screened by sanger sequencing of genomic DNA (gDNA) extracted from leaf tissue. Despite being driven by an egg-specific promoter, the expression of *Cas9* was found to be leaky; almost all sanger sequence chromatograms contained several traces suggesting that gDNA was extracted from a chimeric population of cells. Promising T1 plants were therefore chosen based on whether a predominant sanger sequence trace could be identified for a mutation at one of the sgRNA sites (suggesting that the mutation is in the germline). Determination of this was aided by the 'inference of CRISPR edits' web tool developed by synthego (Synthego 2019). Out of 17 T1 plants screened, four were identified as putatively harbouring a heritable out-of-frame mutation. From each of these four plants, 32 T2 generation progeny (T2 plants) were initially screened by PCR and gel electrophoresis to identify those plants for which the transgene (i.e., the T-DNA harbouring the CRISPR construct) had segregated out (Figure 4.13 C). Then from the transgene-free T2 plants, 24 were randomly selected (six from each T1 parental plant) for sanger sequencing of the *AtAILP1* gene to identify plants homozygous for an out-of-frame mutation in the *AILP1* gene sequence. Four such T2 plants were identified and their sanger sequence traces are shown in Figure 4.13 D. Two of the plants (1.10.18 and 1.10.19) contained a single nucleotide insertion near the binding site of the first sgRNA (Figure 4.13 D). Plant 1.10.32 contained a single nucleotide insertion at the first and second sgRNA site and plant 1.12.2 contained a large deletion between the first and second sgRNA site (Figure 4.13 D). These lines have subsequently been labelled Lines 1-4 (see Figure 4.13 D). For each line, the introduced mutations are expected to result in an out-of-frame amino acid sequence and the introduction of a stop codon in the first or second exon of *AILP1*. Progeny from these four T2 lines were used for phenotyping.

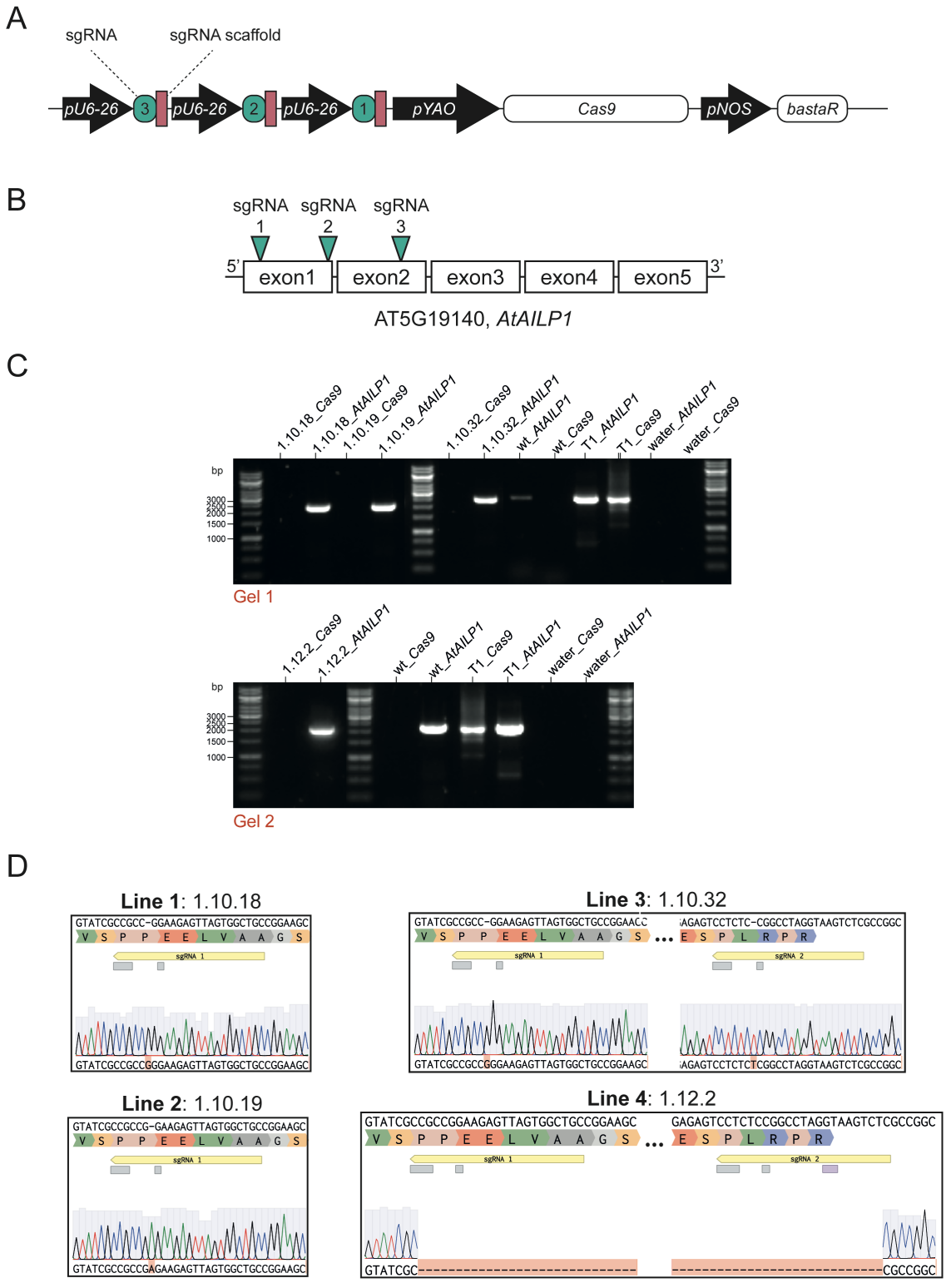


Figure 4.13. CRISPR/Cas9-mediated gene editing was used to create transgene free *Arabidopsis* plants with an out-of-frame mutation in *AtAILP1*. A) A map of the CRISPR-Cas9 construct used to transform *Arabidopsis* plants. Expression of each short guide RNA (sgRNA) is driven by the *Arabidopsis* ubiquitous promoter, *pU6-26*. Expression of the *Cas9* gene is driven by the *Arabidopsis* egg-cell specific promoter

pYAO. Each sgRNA is separated by a scaffold to prevent read-through expression. **B)** Diagram showing the regions of *AtAILP1* targeted by the three sgRNAs. **C)** Agarose gel electrophoresis showing the PCR product from an amplification of transgene (*Cas9*) and *AtAILP1* DNA from selected T2 lines. The *Cas9* fragment has an expected size of 2,328 base pairs (bp) and the wild type *AtAILP1* fragment has an expected size of 2,336 bp. T1 = the parental T1 plant. Note that the lanes shown have been spliced together (from the same gel). The full gel of gel 1 and gel 2 is shown in Appendix 4.3. **D)** Sanger sequence chromatogram showing homozygous mutations in *AtAILP1* in the four T2 lines. These lines were taken forward for phenotyping in the T3 generation. The chromatogram is aligned to the wild type *AtAILP1* sequence. The ellipses indicate a section of the gene sequence not shown. Note that these T2 plants have been renamed as Lines 1-4 for further reference. **Abbreviations:** short guide RNA (sgRNA), *Arabidopsis RNA polymerase III* promoter (*pU6-26*), *Arabidopsis nopaline synthase* promoter (*p*YAO), *CRISPR associated protein 9* (Cas9), Basta resistance gene (*bastaR*).

4.2.7 *ailp1* mutants display a growth rate and flowering time phenotype indistinguishable from wild type plants

The growth and development of *ailp1* mutants was monitored. During rosette growth arial photos were taken regularly and the visible rosette leaf area recorded. After bolting, maximum plant height was measured. No consistent difference in growth rate was found between the four *ailp1* lines and wild type (Figure 4.14 A&B). Only Line 1 was found to have a different growth phenotype compared to wild type; this line has a slightly reduced rosette growth rate with a smaller final rosette size (Figure 4.14 A) and plant height (Figure 4.14 B). Flowering time was measured as the number of days after stratification when the first anther opens to release its pollen. Again, a significant difference from wild type was only observed for Line 1 which on average had a later flowering date (Figure 4.15 A), although no difference in the number of rosette leaves at the date of flowering was observed (Figure 4.15 B). When plants from all four *ailp1* lines were considered together

however, *ailp1* plants were found to flower significantly later than wild type plants (two-tailed t-test, $P < 0.001$). No difference was found for the number of rosette leaves at the onset of flowering suggesting that *ailp1* plants flower at the same developmental stage as wild type plants but that the rate of development is slightly slower.

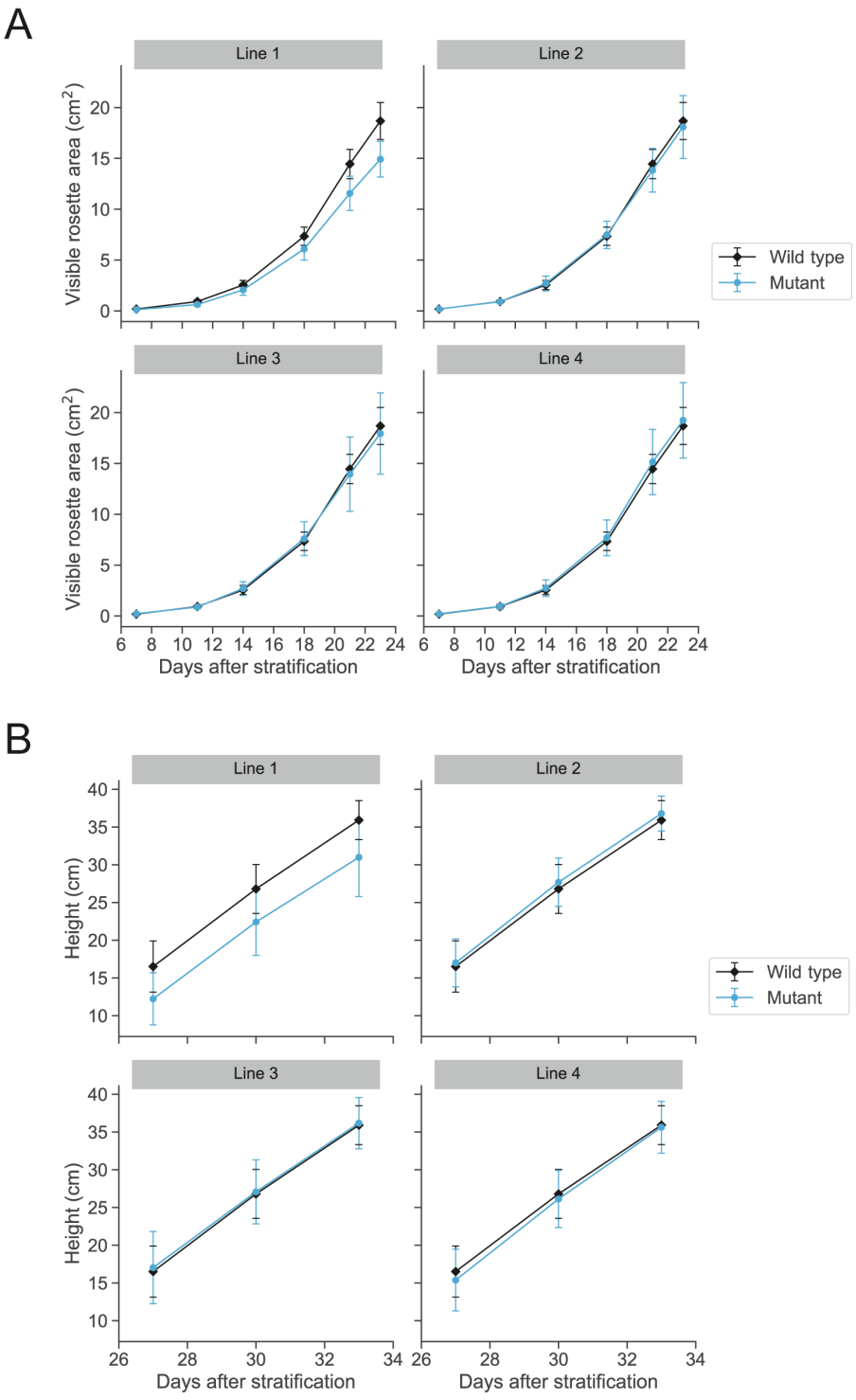


Figure 4.14. There is no consistent difference in growth rate between *ailp1* and wild type *Arabidopsis* plants. A) Vegetative growth rate. A significant difference in final rosette area was only found for Line 1 (two-tailed t-test, $P < 0.001$) **B)** Growth rate after flowering measured as maximum plant height. Again, a significant difference in plant height for each measurement day was only observed for Line 1 (two-tailed t-test, $P < 0.01$). $N = 15$ for all lines.

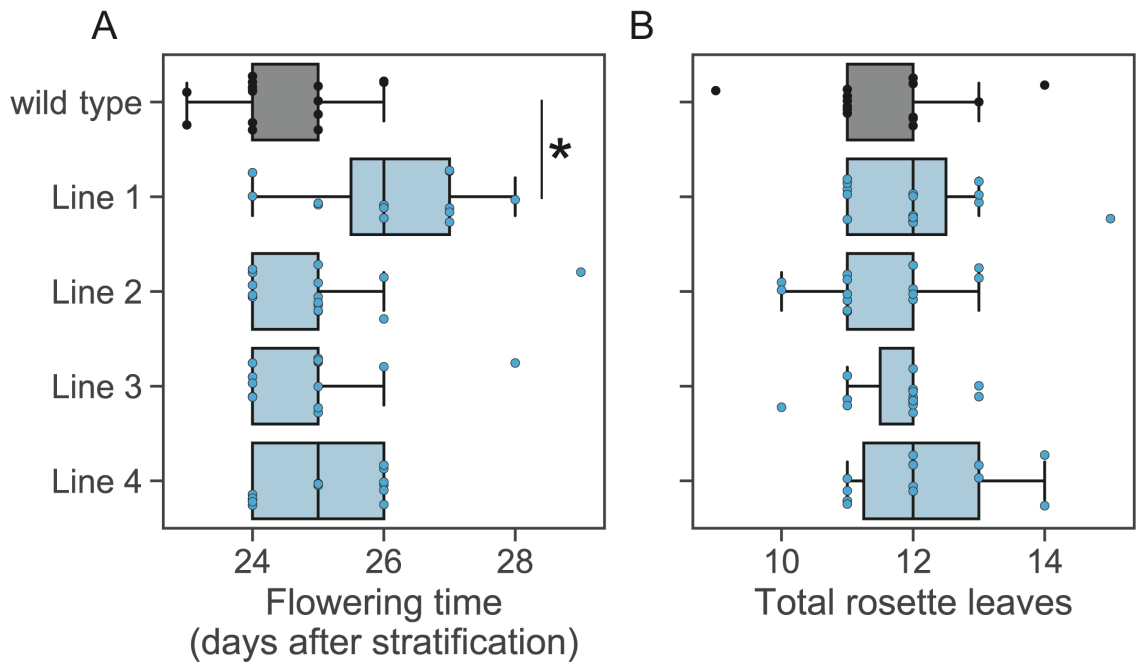


Figure 4.15. There is no consistent difference in flowering time between *ailp1* and wild type *Arabidopsis* plants. A) Flowering time measured as the number of days after stratification when the first anther opens to release its pollen. **B)** The number of fully expanded rosette leaves at the onset of flowering. Asterisk indicates a significant difference between an *ailp1* line and wild type (two-tailed t-test, $P < 0.01$). $N = 15$ for all lines.

4.2.8 Despite *AtAILP1* being highly expressed in the leaf, no perturbation in photosynthesis was found for *ailp1* plants

The expression of *AtAILP1* increases during leaf development suggesting that it might be important for photosynthesis. To test whether photosynthesis is affected in *ailp1* lines a range of photosynthetic measurements were taken using a Licor LI-6800 portable photosynthesis system. No significant difference in carbon assimilation (measured over a range of CO_2 concentrations, Figure 4.16 A) was found between *ailp1* lines and wild type plants although stomatal conductance was found to be higher at moderate CO_2 concentrations for Line 3 and Line 4 (Figure 4.16 B). Similarly, no differences were found

in carbon assimilation, the efficiency of photosystem II, or electron transport rate for *ailp1* plants over a range of light intensities (Figure 4.17 A-C). Interestingly it was found that measurements of non-photochemical quenching (NPQ) were lower at all light intensities in *ailp1* plants (Figure 4.17 D). Unfortunately, the Licor did not successfully take NPQ measurements for all plants and thus the sample size for some lines (including wild type) was reduced and no significant differences were found. There was not enough time to repeat this measurement in this experiment, but it would be interesting to know if NPQ is reduced in *ailp1* plants and, if so, how these plants respond under fluctuating light conditions.

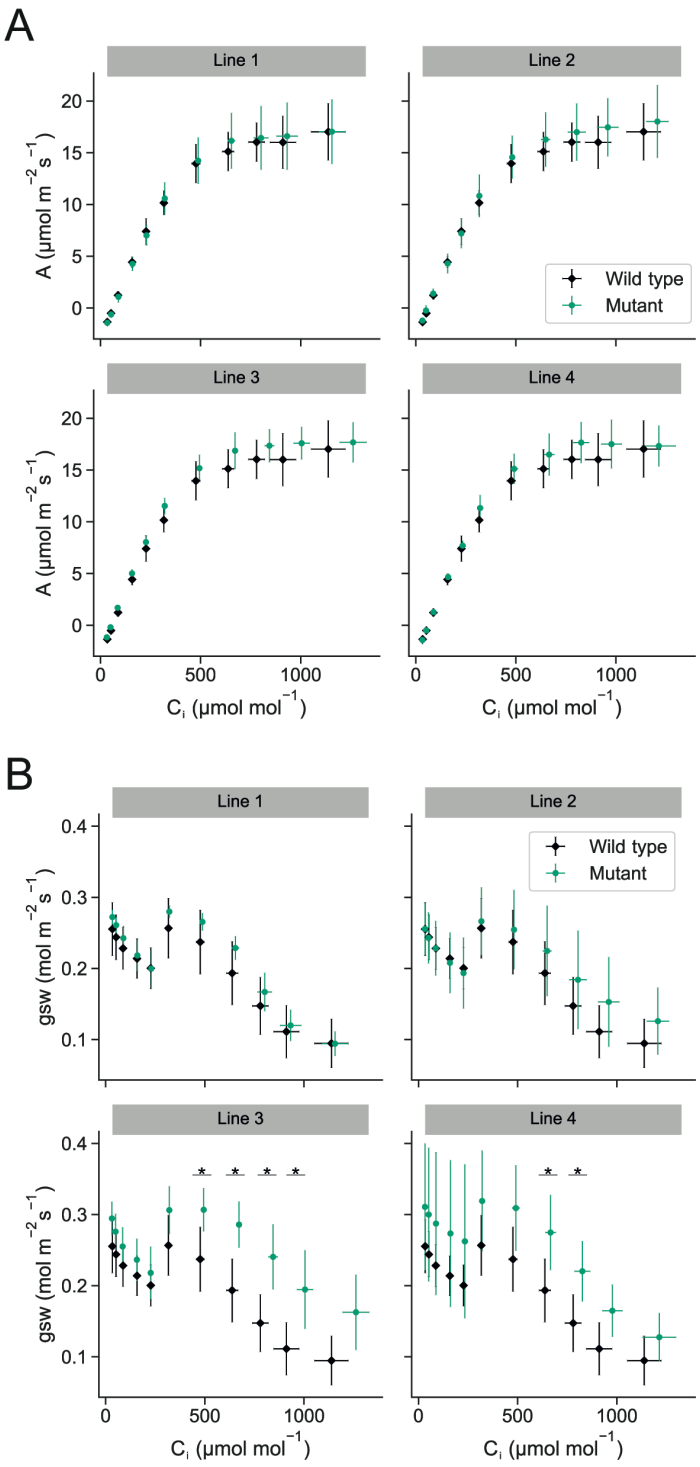


Figure 4.16. Characterisation of photosynthesis in *Arabidopsis ailp1* and wild type plants. A) Comparison of carbon assimilation (A) in mutant lines 1-4 and segregating wild type plants over a range of intercellular carbon concentrations (C_i). **B)** Stomatal conductance (gsw) was also measured concurrently with the assimilation measurements. Error bars represent standard deviation. The asterisks indicate a significant difference between an *ailp1* line and wild type plants at a given C_i value (two-tailed t-test, $P < 0.01$). For each line and wild type $N = 5$.

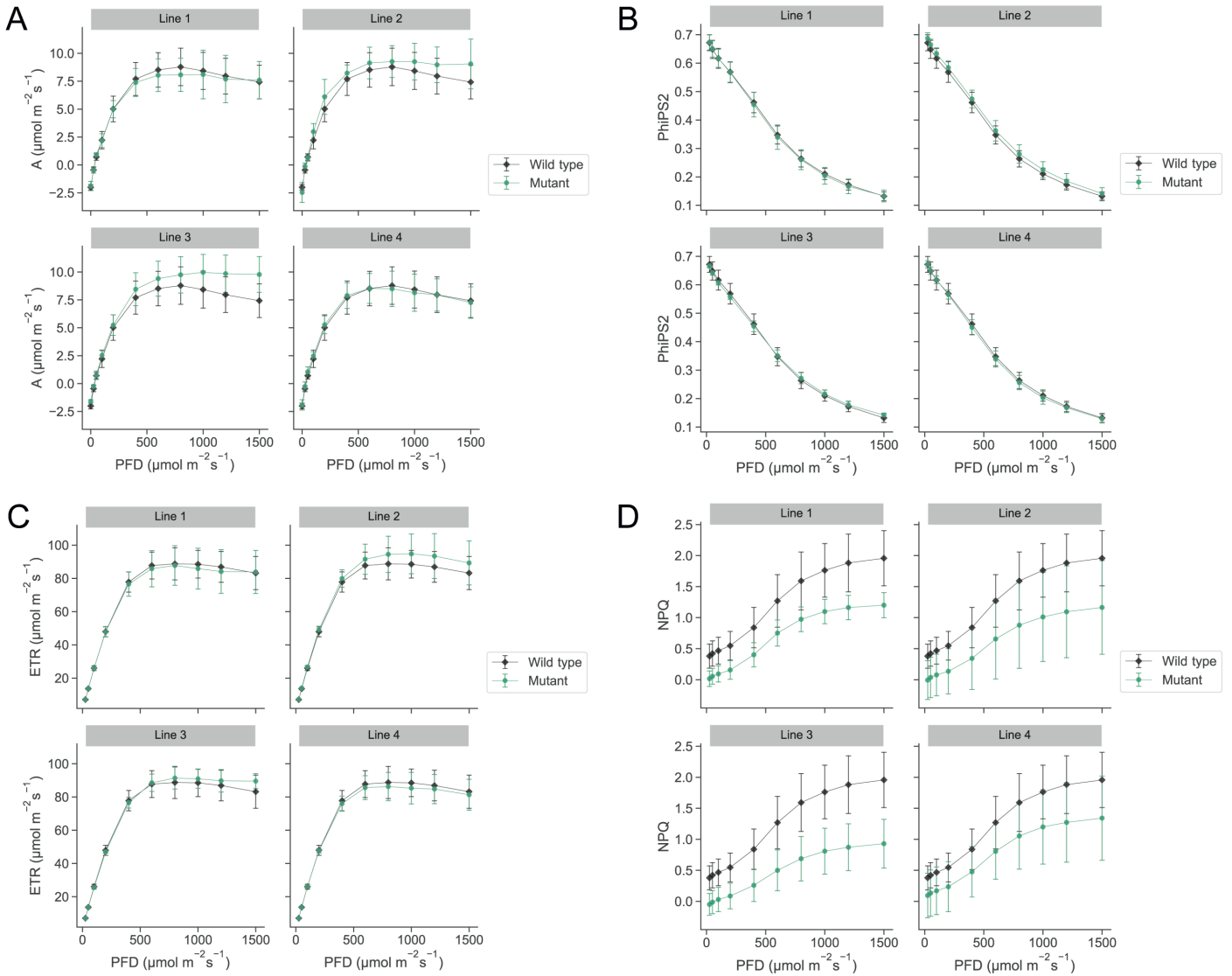


Figure 4.17. Figure legend on next page.

Figure 4.17. Measurement of various photosynthetic parameters over a range of PFDs (photon flux densities). A) CO_2 assimilation rate (A), **B)** the operating efficiency of photosystem II (PhiPS2), **C)** electron transport rate (ETR), **D)** non-photochemical quenching (NPQ). *N* = 5 for each line and wild type in plots A-C. For plot D, *N* values for each group are in parentheses as follows: wild type (3), Line 1 (5), Line 2 (5), Line 3 (2), Line 4 (4).

4.2.9 Leaves from *ailp1* plants have a higher nitrogen content than wild type plants

In plants, amino- and amido-transferases function in the metabolism of amino acids and thus play an important role in the assimilation of carbon and nitrogen into different organic compounds. If AILP1 does function as an amino- or an amido-transferase (or an amino-lyase) then it would be expected that a loss of function in *AtAILP1* activity would lead to perturbations in the balance of carbon and nitrogen in the leaf. To investigate this, the total carbon and nitrogen content of leaves from *ailp1* and wild type plants was measured and compared. For all four *ailp1* lines the carbon-to-nitrogen ration (C:N) was found to be lower than in the wild type plants, although this was only significant for Line 1 (Figure 4.18 A). Pooling the data for *ailp1* plants however revealed that overall, the C:N ratio of leaf tissue from *ailp1* plants is significantly lower (Figure 4.18 B). This suggests that the *AILP1* mutations (and presumably AILP1 loss of function) in these plants have resulted in an increase in leaf nitrogen content relative to their carbon content.

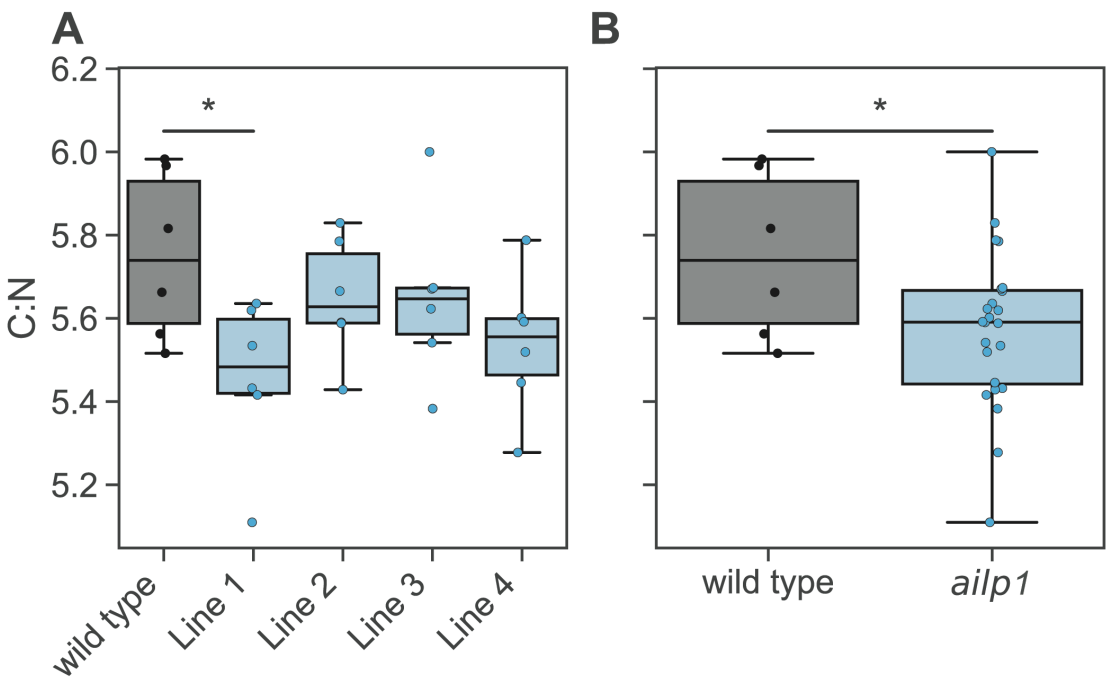


Figure 4.18. Comparison of leaf carbon-to-nitrogen ratio (C:N) in wild type and *ailp1* plants. A) Data for each individual *ailp1* line (*N* = 6). **B)** Data for each *ailp1* line is pooled and compared to wild type. The asterisk indicates a significant difference between *ailp1* lines and wild type (t-test, $P < 0.05$).

4.3 Discussion

In the maize C4 photosynthetic cycle it is predicted that aspartate is converted to oxaloacetate in the bundle sheath, however the enzyme that catalyses this reaction has not been identified. A bioinformatic screen previously conducted in the Kelly Lab to try and identify the enzyme(s) responsible for this transamination reaction led to the identification of the maize protein 'aluminium-induced-like protein homolog 1' (AILP1) as a promising candidate. In this chapter I present the work conducted to try and further characterise the function of this protein and assess its possible role in C_4 photosynthesis.

Analysis of *AILP1* expression in maize and other C_4 species supports the hypothesis that this gene may function in C_4 photosynthesis. In maize, *AILP1* expression is highest in the most photosynthetically active part of the leaf and *AILP1* transcripts preferentially accumulate in the bundle sheath cells. In other transcriptomic data sets that include both C_3 and C_4 species *AILP1* also has a pattern of expression that would be expected from a gene involved in C_4 photosynthesis. Among different *Flaveria* species, *AILP1* expression increases from the C_3 to C_3-C_4 intermediates to the C_4 species. The highest *AILP1* expression among *Flaveria* species was found in the C_4 NADP-ME subtype *Flaveria bidentis*. Although *F. bidentis* is thought to only conduct NADP-ME decarboxylation (Gowik et al. 2011), it has been shown that in this species both malate and aspartate are used as the C_4 transfer compound (Meister et al. 1996; Gowik et al. 2011). The aspartate in this cycle is thus converted in the bundle sheath cells to oxaloacetate by aspartate aminotransferase and then to malate by NADP-malate dehydrogenase for decarboxylation by NADP-ME (Meister et al. 1996). Although aspartate aminotransferase activity has been demonstrated in the bundle sheath cells of *F. bidentis*, it was found to be less than half of the activity reported in the mesophyll cell (Meister et al. 1996). If AILP1 does function to metabolise aspartate, then in *F. bidentis* it is possible that it functions alongside the canonical aspartate aminotransferase.

There has been no previous experimental characterisation of any *AILP1* gene but the partial homology of AILP1 to asparagine synthetase led to the hypothesis that AILP1 may also catalyse a reaction with aspartate in an aminotransferase reaction. Specifically, AILP1 is homologous to the glutamine amidotransferase domain of asparagine synthetase which functions to hydrolyse glutamine to yield ammonia and glutamate (Lomelino et al. 2017). As the conversion of aspartate to oxaloacetate similarly involves the removal of a nitrogen group from aspartate (thus forming oxaloacetate), I looked to investigate whether AILP1 is able to catalyse the removal of ammonia from glutamine or aspartate using an enzyme activity assay. Unfortunately, the purification of AILP1 proved challenging as when expressed in *E. coli* AILP1 protein fused to a His-tag accumulated entirely in inclusion bodies. The reasons for this are unclear but plant asparagine synthetases have similarly been reported to accumulate in inclusion bodies when expressed in *E. coli* (Todd et al. 2008; Duff et al. 2011; Xu et al. 2018) suggesting it might have something to do with the physico-chemical properties of these two homologous proteins. While it is possible to purify active protein from inclusion bodies, doing so involves the use of strong denaturants such as urea or guanidinium hydrochloride to break up the protein aggregate and furthermore proteins have to be refolded after purification. Without a routine kinetic assay for AILP1 function, it is impossible to assess the success of such a strategy in the purification of biologically active protein. Thus, a purification strategy using a solubility tag was favoured as it involved fewer handling steps. Although this was successful in producing soluble protein, no amino-lyase activity was detected.

At the inception of the work in this thesis chapter it was hoped that activity of AILP1 would be elucidated before conducting any phenotypic characterisation of *Arabidopsis ailp1* mutants. Nonetheless, given the hypothesis that AILP1 may function to process the nitrogen groups of amino acids, and thus play a role in amino acid metabolism, I sought to investigate whether the carbon-to-nitrogen balance in *ailp1* leaves was perturbed. Indeed, it was found that the carbon to nitrogen ratio of *ailp1* leaf tissue differed from wild type

plants suggesting amino acid metabolism may be affected. Amino acid metabolism plays an important role in the assimilation and recycling of nitrogen in the leaf and in matur, photosynthetically active leaf tissue amino acid catabolism dominates over amino acid biosynthesis as nutrients are reallocated from source to sink leaves (Hildebrandt et al. 2015). The finding that mature rosette leaves from *ailp1* plants have a higher nitrogen content (relative to carbon) suggests amino acid catabolism might be affected in these plants. As amino acid catabolism involves the removal of nitrogen groups from amino acids through deamination reactions, this would support the hypothesised activity of AILP1. At the moment though this is just speculation, and a more detailed analysis is needed to assess the relative amino acid content of the source and sink tissues of *ailp1* plants.

There have been several suggestions as to the possible benefits of operating a secondary aspartate/PEP-CK shuttle alongside an NADP-ME or NAD-ME C_4 pathway. For example it has been suggested that the utilisation of more than one transfer metabolite lowers the concentration needed of any one metabolite to maintain sufficient flux through the cycle (Pick et al. 2011). Alternatively, the relative use of each C_4 subtype cycle has been linked to the physiology of the mesophyll and bundle sheath cells in different species (Gutierrez et al. 1974; Walker et al. 1997; Furbank 2011) or under different environmental conditions (Khamis et al. 1992; Bellasio and Griffiths 2014; Sharwood et al. 2014; Cacefo et al. 2019). This is because the energetic demands (in the form of ATP and reducing power) on the mesophyll and bundle sheath cells varies between the different C_4 pathways (Munekage and Taniguchi 2016). For instance, malate decarboxylation by NADP-ME produces NADPH in the bundle sheath cells while decarboxylation of aspartate by PEP-CK doesn't. Thus, although NADP-ME species like maize typically have low PSII activity (and thus NADPH-producing linear electron transport) in the bundle sheath, a correlation has been observed in NADP-ME species between the amount of PEP-CK activity in the bundle sheath cell (which leads to a greater deficit of NADPH) and the amount of PSII activity

(Gutierrez et al. 1974; Walker et al. 1997; Furbank 2011). No doubt the evolutionary drivers of a mixed-subtype C_4 cycle are likely to be multifaceted and to vary between species.

In maize a number of open questions remain to be answered before we can better understand the implications of a mixed-subtype C_4 pathway in this species. As well as the unknown route by which aspartate is converted to oxaloacetate, it is also not clear what the relative contribution of PEP-CK and NADP-ME is to the decarboxylation of aspartate. The conversion of aspartate (via oxaloacetate) to malate in the bundle sheath cells requires the activity of a malate dehydrogenase (MDH). While the expression of the chloroplast-localised NADP-dependent MDH is low in the bundle sheath cells of maize there is appreciable expression of a cytosolic NAD-dependent MDH and it is unknown if some oxaloacetate may be converted to malate through the activity of this enzyme (cytosolic MDH gene expression in the maize leaf is shown in Appendix 4.4). Furthermore, in the bundle sheath cells the transport of malate across the chloroplast membrane appears to be inextricably linked to aspartate as the bundle sheath cell localised transporter DCT2 has been proposed to transport malate into the chloroplast in exchange for aspartate (Weissmann et al. 2015). Although there are a number of aspects of this transport mechanism yet to be elucidated, the finding that aspartate facilitates malate decarboxylation even in the absence of PEP-CK activity (Wingler et al. 1999) may be explained by this requirement for aspartate to facilitate malate transport into the chloroplast. Empirical evidence for this is however lacking. Thus, there is still a lot to be learned about C_4 photosynthesis in maize. In this chapter I have attempted to fill one of the gaps in our knowledge about how the C_4 cycle operates in this species through the discovery of the enzyme that metabolises aspartate in the bundle sheath cells. Although the definitive function of AILP1 could not be elucidated, some promising results concerning the carbon to nitrogen ratio in *ailp1* plants hint that the gene may play a role in amino acid metabolism. This, combined with its high expression in bundle sheath cells of C_4 plants,

means that further work is warranted to discover whether this protein provides a role in C_4 photosynthesis.

4.4 Materials and methods

4.4.1 Phylogenetic analyses

Orthogroups were generated using Orthofinder v2.3.11 (Emms and Kelly 2015) from proteome data (primary protein transcripts) of 54 plant species spanning the phylogeny of the plant kingdom (a list of species, proteome versions used, and download location is given in Appendix A4.1). The orthogroup for aluminium-induced-like protein homolog 1 (AILP1) was identified and a multiple sequence alignment generated from the orthogroup amino acid sequence using MAFFT-LINSI v7.470 (Katoh and Standley 2013). Manual inspection of the *AILP1* multiple sequence alignment was conducted to remove erroneous sequences a gene tree was inferred using IQ-TREE v2.0.3 (Nguyen et al. 2015) with the settings -m TEST to automatically select the best-fitting model of sequence evolution. The tree was then manually rooted. For visualisation, ETE3 v3.1.1 (Huerta-Cepas et al. 2016) was used to prune the tree to a smaller number of species while maintaining tree topology. The subcellular localisation of AILP1 proteins was predicted using TargetP v2.0 (Armenteros et al. 2019).

To investigate whether AILP1 proteins are present in the proteomes of species outside of the plant kingdom, a sequence similarity search was conducted using HMMER v3.1b2 (Finn et al. 2011). Here, a hidden Markov model (hmm) was generated from the AILP1 orthogroup multiple sequence alignment using hmmbuild. Then, using hmmsearch with the default per-sequence and per-domain E-value cut-off of 0.01, the AILP1 hmm was used to search for similar sequences in target proteomes.

4.4.2 Analysis of gene expression data

RNA-seq data sets used in this chapter were previously published and available in NCBIs short read archive. A list of accession numbers and study details is provided in Appendix 3.2. The data were processed as described in Chapter 3.

4.4.3 Strains, growth conditions and media

All *Escherichia coli* strains used were grown on Lysogeny broth (LB: 1% tryptone, 0.5% yeast, 0.5% NaCl, plus 1.5% agar for solid media) and cultured overnight at 37°C unless otherwise stated. Liquid cultures were grown with continuous shaking at 200 rpm. Plasmid construction and manipulation was conducted in *Escherichia coli* strain DH5α. Protein expression was conducted in *Escherichia coli* strain BL21(DE3). *Agrobacterium tumefaciens* strain gv3101 was used for plant transformation by floral dipping. *Agrobacterium tumefaciens* was grown at 28°C for 2 d on LB plates or in liquid LB with continuous shaking at 150 rpm. LB for *Agrobacterium* growth always contained gentamycin 30 μg ml^{-1} and rifampicin 10 μg ml^{-1}. Where appropriate, bacterial cultures were screened with antibiotics at the following concentrations: 100 μg ml^{-1} ampicillin, 50 μg ml^{-1} kanamycin, 50 μg ml^{-1} spectinomycin. Where used, X-Gal was added directly to media at a concentration of 20 μg ml^{-1}.

4.4.4 Plant material and growth conditions

Arabidopsis thaliana ecotype Columbia-0 seeds were grown under long-day conditions (16 h of light, 8 h of darkness) in either the greenhouses or in controlled growth chambers at the Department of Plant Sciences, University of Oxford. Seeds were germinated on soil: 3:1 ratio of Levington Advance F2 compost and Sinclair Pro medium (2.0–5.0 mm) vermiculite, supplemented with Exemptor insecticide at the recommended dose (4 g per litre of soil).

4.4.5 Modification of protein sequences for cloning and expression

AILP1 gene sequences from *Arabidopsis* (*AtAILP1*, AT5G19140) and maize (*ZmAILP1*, Zm00001d023376) were modified to facilitate cloning and expression in either bacterial or plant cells. In all cases, to make the sequences compatible for Golden Gate cloning (Engler et al. 2008) they were modified to remove BsaI and BpiI restriction enzyme sites. In addition, sequences were extended with *either* an SC1 or C Golden Gate adaptor to

allow for cloning into a level 1 Golden Gate vector with an N- or C- terminal tag, respectively (Table 4.3). For bacterial expression, sequences were codon-optimised for *E. coli* expression using the web-tool OPTIMIZER (Puigbò et al. 2007). For plant expression, sequences were generated using the native codons. If a gene sequence was to be cloned into a construct with a C-terminal tag, then the stop codon was omitted from sequence. Sequences generated were sent as FASTA files to Gene Universal Inc. where they were synthesised and ligated into a pUC57 plasmid.

Table 4.3. Adapter sequences for Golden Gate cloning added to the 5' and 3' end of the coding sequences of *AILP1.* The adaptor sequences are in lower case. The first codon (5' extension) or last codon (3' extension) in the *AILP1* sequence that flanks the adaptor is highlighted in capital and bold. Note, the SC1 3'extension does not contain a stop codon to allow for cloning into a level 1 vectors with a C-terminal tag.

Level 0 golden gate part	Adapter	5' -> 3' sequence
SC1	5' extension	cactctgtggtctcaa**ATG**
SC1	3' extension	**CAT**ggtgtgagaccacgaagtg
C	5' extension	cactctgtggtctcaaggt**ATG**
C	3' extension	**TAA**gctttgagaccacgaagtg

4.4.6 Generation of constructs for protein expression of *ZmAILP1* in *E. coli*

For the expression of *ZmAILP1* in *E. coli* two constructs were generated. The first contained *ZmAILP1* fused to a C-terminal polyhistidine-tag (His-tag) and the second contained *ZmAILP1* fused to an N-terminal maltose-binding protein (MBP). For the C-terminal His-tag the expression vector pOPINE (Berrow et al. 2007) was used (provided by Professor Ray Owens at the University of Oxford). For the N-terminal MBP the expression vector pLOU3 was used (provided by Professor Phil Poole's lab at the University of Oxford). *ZmAILP1* was subcloned into these vectors by restriction enzyme

digest using the multiple cloning sites available on each (vector maps can be found in Appendix 4.5). To generate the appropriate restriction enzyme sites at the 5' and 3' end of the *ZmAILP1* sequence for this cloning, primers containing adaptor sequences were used to amplify the gene sequence from pUC57 by PCR. These primers and the introduced restriction enzyme sites are detailed in Table 4.4. The resulting PCR product was purified using a commercial PCR clean up kit and then digested in a single reaction with Esp3I and HindIII followed by another clean up step. Note that the type II restriction enzyme Esp3I was used here to generate overhangs compatible with those generated by the restriction enzyme NcoI. To create compatible sticky ends the pOPINE and pLOU3 plasmids were digested, separately, in a double digest with NcoI and HindIII. In this digestion reaction the 5' DNA was dephosphorylated using a Quick Dephosphorylation Kit (NEB #M0508). The resulting digested plasmid was run on an agarose gel and extracted using a commercial gel extraction kit. The digested PCR product was then ligated with the digested destination plasmids in a T4 DNA ligase (NEB #M0202) reaction containing a 1:3 recipient plasmid to insert ratio. After ligation, the reaction was heat inactivated for 10 min at 65°C and the resulting product transformed into *E. coli* and plated on LB plates containing ampicillin. Plasmid preparation was conducted for several transformed colonies and correctly ligated plasmids identified by sanger sequencing. Primers used for sanger sequencing of pOPINE ligations were 5'-GGGAATTGTGAGCGGATAAC-3' (forward) and 5'-GGGAATTGTGAGCGGATAAC-3' (reverse). Primers used for sanger sequencing of pLOU3 ligations were 5'-TCGTCAGACTGTCGATGAAG-3' (forward) and 5'-CCTCTTCGCTATTACGCCAG-3' (reverse).

Table 4.4. Primers used to amplify *ZmAILP1* from pUC57 and introduce the necessary restriction enzyme sites needed for cloning into pOPINE or pLOU3. Capitalised and bold residues in the primer sequence represent the homologous sequence to the template DNA, the lower-case residues show the introduced overhangs, and the italicised residues show the introduced restriction enzyme recognition sites.

Destination plasmid	*ZmAILP1* template sequence taken from the following level 0 golden gate part	PCR Primer	Primer sequence (5' – 3')	Introduced restriction site
pOPINE	SC1	Forward	aaaa*cgtctccc***ATGCTGGCAGTTTTTAGCGGTCA**	Esp3I
pOPINE	SC1	Reverse	acc*aagctta*t**ATGCAGGGCGGTCAGAACG**	HindIII
pLOU3	C	Forward	aaaa*cgtctccc***ATGCTGGCAGTTTTTAGCGGTCA**	Esp3I
pLOU3	C	Reverse	aac*aagc**TT***AATGCAGGGCGGTCAG**	HindIII

4.4.7 Expression of recombinant *ZmAILP1* in *E. coli*

The two expression constructs described above were transformed into *E. coli* BL21 (DE3). In both cases, a single colony of transformed BL21 cells was then used to inoculate a starter culture of 10 ml LB (supplemented with ampicillin and 1% glucose [v/v]) which was grown for 4-5 h at 37°C with continuous shaking at 200 rpm. 1 ml of this starter culture was then used to inoculate 50 ml of LB (supplemented with ampicillin and 1% glucose [v/v]) which was then incubated at 37°C with continuous shaking at 200 rpm until the cells reached as OD_{600} of approximately 0.6. At this point, a 1 ml sample was taken for analysis by SDS-PAGE (T_0 sample) and the remaining cell culture was cooled on ice for 15 min before inducing expressing with the addition of isopropylthio-β-galactoside (IPTG) to a final concentration of 0.5 mM. The cells were then incubated overnight at room temperature with continuous shaking at 200 rpm. The next morning, another 1 ml sample was taken for SDS-PAGE analysis (T_1 sample) and the remaining cell culture were placed on ice for 5 min to arrest growth. The cells were then harvested by centrifugation at 4000 x g for 20

min at 4°C. The resulting cell pellet was then placed on ice and immediately used for downstream protein purification.

4.4.8 Protein purification of recombinant *Zm*AILP1

Working on ice, the cell pellet collected from the expression culture was resuspended in either 5 ml of cold lysis buffer (for purification of the His-tag fusion, *Zm*AILP1::Hisx6) or 4 ml of cold column buffer (for purification of the maltose-binding protein fusion, MBP::*Zm*AILP1). The lysis buffer contained 50 mM NaH_2PO_4, 300 mM NaCl, 10 mM imidazole, adjusted to pH 8.0 and supplemented with 1 mg ml^{-1} lysozyme (Sigma #L6876) and a protease inhibitor cocktail (Roche #11836170001). The column buffer contained 20 mM Tris-HCl, 200 mM NaCl, 1 mM EDTA, 10 mM 2-mercaptoethanol, adjusted to pH 7.4 and supplemented with a protease inhibitor cocktail (Roche #11836170001). In both cases, the resuspended cell pellet was then subject to mechanical cell lysis; first, the resuspended cells were split between four 1.5 ml microcentrifuge tubes and placed in a dry-ice ethanol bath for 20 min and then thawed at room temperature before being placed back on ice. The samples were then added to 2 ml lysing matrix B tubes (MP Bio) and further homogenised using a ribolyser (MP Bio, FastPrep-24) in pulses of 30 s at a speed of 6 ms^{-1} followed by 2 min on ice for a total of six pulses. To remove the beads and cells debris, the tubes were then centrifuged at 20,000 x g for 20 min at 4°C and the supernatant of each tube decanted and pooled (soluble fraction) and saved for further protein purification. One of the remaining pellets (insoluble fraction) was resuspended in 1 ml of lysis buffer (for the His-tag protein) or column buffer (or the MBP-tag protein). A sample from both the soluble and insoluble fraction was then loaded onto an SDS-PAGE gel to assess protein solubility. This method of protein expression and cell lysis was sufficient to yield soluble MBP::*Zm*AILP1 however the His-tag fusion protein *Zm*AILP1::Hisx6 remained insoluble. Modifications made to the expression and cell lysis conditions to try and increase protein solubility included lowering the IPTG concentration to 0.1 mM, growing all *E. coli* cultures at room temperature, supplementing the lysis buffer with

stabilising reagents (10% glycerol [v/v], 1 mM EDTA and 20 mM 2-mercaptoethanol) and conducting cell lysis at lower intensities and by using alternate lysis methods such as by using a sonicator. These modifications did not however result in the production of soluble *Zm*AILP1::Hisx6. Thus, further protein purification is only described for MBP::*Zm*AILP1.

MBP::*Zm*AILP1 recombinant protein was purified from the soluble fraction of the cell lysate using affinity chromatography. Firstly a 5 ml gravity flow column (Qiagen #34964) was prepared by washing the column membrane with 5 ml of column buffer and then capping the column and adding 2 ml bed volume of amylose resin (NEB #E8021S). The resin was then equilibrated by washing with 5 ml of column buffer. At all stages, care was taken not to let the resin in the column run dry. The cell lysate containing MBP::*Zm*AILP1 was diluted 1:6 [v/v] in column buffer and then gently loaded into the column, taking care not to disrupt the packed resin. The flow-through was collected and a 5 µl sample taken for SDS-PAGE (flow through sample). The column was then washed with 24 ml of column buffer, again saving a 1 µl sample of the flow through for SDS-PAGE (wash sample). The protein was then eluted from the column in 15x 500 µl fractions by washing the column with column buffer supplemented with 10 mM maltose. Each fraction was saved and immediately placed put in the fridge (4°C). A 5µl sample from each fraction was taken for analysis by SDS-PAGE (elution samples). Fractions containing purified protein were then pooled and used immediately in enzyme assays.

4.4.9 Enzyme assay to assess the ability of AILP1 to liberate ammonia from amino acids

To assess whether AILP1 can catalyse the removal of an ammonia group from glutamine or aspartate an enzyme assay was set up. The eluted and pooled sample of purified recombinant AILP1 was concentrated using an Amicon Ultra-0.5 10 KDa centrifugal filters (Merck # UFC501024) following the manufacturer's instructions. The concentrated sample was then either used directly in enzyme assays or transferred by buffer exchange into a

devised assay buffer containing 50 mM HEPES, 2.5 mM $MgCl_2$, 0.5 mM DTT, 20% glycerol, adjusted to pH 7.6. In the latter case, buffer exchange was conducted using an Amicon Ultra-0.5 10 KDa centrifugal filters by conducting three rounds of successive sample dilution (filling the centrifugal unit) and concentration in the assay buffer. The time and speed for this buffer exchange was that recommended for the size of the AILP1 protein in the manufacturer's instructions. Protein recovery from the centrifugal units was monitored by SDS-PAGE or a Bradford assay. For the enzyme assay, 10 µl of protein concentrate or 10 µl water was added to a well of 96-well plate containing either 10 µl of 2 mM L-glutamine or 10 µl of 2 mM L-aspartate. A positive control reaction contained 10 µl of reconstituted lysozyme (Sigma #L6876) at a concentration of 10 units μl^{-1}, and 10 µl of 2 mM L-glutamine. All reactions were conducted in triplicate. The 96-well plate was then incubated for 15-30 min at room temperature after which ammonia liberation was monitored using an ammonia assay kit (Sigma #AA0100) following the manufacturer's instructions. Briefly, 200 µl of the ammonia assay reagent was added to each sample in the 96-well plate and the absorbance at 340 nm was measured in a plate reader. 2 µl of the assay's L-glutamate dehydrogenase solution was then added to each sample and incubated for 5 min with gentle mixing. After incubation the absorbance at 340 nm was then measured again.

4.4.10 Generation of constructs for protoplast expression

Golden Gate assembly (Engler et al. 2008) was used to generate protoplast expression constructs containing either *ZmAILP1* or *AtAILP1* fused to a C-terminal GFP-tag. The parts used in this assembly are shown in Appendix 4.6, Table A4.6.1. Golden Gate reactions contained 40 fmol of each part (Table A4.6.1), 1 µl T4 DNA ligase (2,000,000 units μl^{-1}), 1 µl BsaI, 1.5 µl T4 DNA ligase buffer, 1.5 µl 2 mg ml^{-1} BSA in a final reaction volume of 15 µl. The assembly reactions were incubated in a thermocycler for 25 cycles consisting of 3 min at 37°C and 4 min at 16°C, followed by 1 cycle consisting of 5 min at 50°C and 5 min at 80°C. Reaction tubes were cooled and then 1 µl of the assembly reaction was used to

transform 10 µl of heat-shock competent *E. coli* (DH5α). The transformed cells were plated on LB containing ampicillin and X-Gal and incubated overnight at 37°C. Two white colonies for each construct were picked and grown overnight in 5 ml liquid LB at 37°C. Subsequently plasmid preparation was conducted, and correctly assembled constructs screened by restriction enzyme digest. The restriction enzyme reaction contained 0.5 µg plasmid DNA, 2.5 µl 10x CutSmart Buffer (NEB), 1 µl DraIII-HF, 1 µl SacII in a final reaction volume of 25 µl. The digest was carried out at 37°C for 15 min and gel electrophoresis used to verify the correct fragment size of the digested product.

4.4.11 *Arabidopsis thaliana* protoplast generation and transformation

Arabidopsis thaliana Col-0 protoplasts were isolated using the 'Tape-*Arabidopsis* Sandwich' method (Wu et al. 2009). Briefly, 10 leaves from 3-4 week-old greenhouse grown *Arabidopsis* plants were collected. The upper epidermal surface of leaves was affixed to strip of electrical insulation tape (AT7 PVC Electrical Insulation Tape #AT0007) and then a strip of magic tape (Scotch® Magic™ Tape) was affixed to the lower epidermal layer. The blunt edge of curved tweezers was used to ensure even fixation of magic tape to the lower epidermal layer. The magic tape was pulled back to remove the lower epidermal layer and expose the mesophyll cells underneath. To isolate protoplasts leaves were transferred into a petri dish (Ø 90 mm), lower epidermal side down, containing ~25 ml of enzyme solution (1% [w/v] Cellulase Onozuka R-10 [Yakult Honsha Co.], 0.25% [w/v] maceroenzyme [Yakult Honsha Co.], 500 mM mannitol, 1 mM $CaCl_2$, 1 mg ml^{-1} BSA, 9.4 mM MES, pH 5.6) and incubated for ~2 h while shaken (40 rpm on platform shaker) at room temperature. During this incubation step the petri dish was covered from light. After incubation, the majority of leaf mesophyll protoplasts were released into the enzyme solution. Using tweezers, the leaf material was very gently shaken to remove any last protoplasts before removing the remaining leaf debris from the petri dish. The cell suspension was then collected in a 50 ml tube on ice and protoplasts were collected by centrifugation at 60 x g for 5 min at 4°C. The supernatant was removed without disturbing

the cell pellet and the cell pellet resuspended in 25 ml pre-chilled W5 solution (154 mM NaCl, 125 mM $CaCl_2$, 5 mM KCl, 5 mM glucose, 1.5 mM MES, pH 5.6) before being centrifuged at 60 x g for 5 min at 4°C. Again, the supernatant was discarded, and a second wash step was carried out in which the cell pellet was resuspended in 25 ml pre-chilled W5 solution and then filtered using a nylon mesh with a pore size of 60 μl to remove tissue debris. The suspension was centrifuged again at 60 x g for 5 min at 4°C. Next, 15 ml of the supernatant was removed, and the cell pellet carefully resuspended in the remaining supernatant and incubated on ice for 30 mins. After incubation, the protoplasts had settled to the bottom of the tube and the supernatant was carefully removed. Protoplasts were then gently resuspended in 2 ml of MaMg solution (400 mM mannitol, 15 mM $MgCl_2$, 5mM MES, pH 5.6) and 100μl of cell suspension was mixed with 5-10 μg of plasmid DNA at room temperature. Note that all pipetting steps with protoplasts were carried out with cut pipette tips to reduce cell shearing. Next, 100 μl of PEG solution (600 mM mannitol, 150 mM $Ca(NO_3)_2$, 40% PEG [w/v]) was added to the suspension and incubated for 5 min at room temperature. After the incubation the protoplasts were carefully washed by adding 1.5 ml of room temperature W5 solution and collected by centrifugation at 40 x g for 1 min at 4°C. The supernatant was removed, and the wash step repeated once. After the second wash step, 1 ml of the supernatant was removed, and the protoplast cell pellet was resuspended in the remaining supernatant and transferred to a well of a 24-well plate and incubated for 16 hr at 24°C covered from the light.

4.4.12 Confocal analysis of transformed protoplasts

Protoplasts were examined with a Zeiss LSM 510 laser scanning microscope using the objective C-Apochromat 40x with water emersion. The laser wavelength was 514 nm and the GFP channel detected wavelengths between 500-551 nm. The channel for the detection of chlorophyll autofluorescence was set to 650-750 nm. For both channels the gain was adjusted to limit fluorescence saturation. The recorded images were processed with the Fiji distribution of ImageJ v1.52p (Schindelin et al. 2012).

4.4.13 CRISPR-Cas9 vector construction

Short guide RNAs (sgRNAs) targeting the *Arabidopsis AILP1* gene (AT5G19140) were designed using the online tool CRISPOR (Concordet and Haeussler 2018). The sgRNA sequences (without the protospacer adjacent motif) were extended with adaptors containing Esp3I restriction enzyme sites to facilitate cloning into a level 1 Golden Gate vector. The resulting sequences, and their reverse complement sequences, were ordered as oligonucleotides. The ordered oligonucleotide sequences are shown in Table 4.5. The complementary oligonucleotides (forward and reverse complement sequences) were annealed in a 10 µl reaction containing 1 µl of each oligonucleotide and 8 µl of water by incubation in a thermocycler at 95°C for 5 min after which they were left to cool to room temperature for 1 h. Each annealed oligonucleotide was diluted in water (1:250 [v/v]) and cloned into a specific level 1 Golden Gate vector (specified in Table 4.5) in a reaction containing 0.1 µg destination vector DNA, 10 µl of the diluted oligonucleotides, 1 µl Esp3I, 0.5 µl T4 DNA ligase (400,000 U ml^{-1}), 2 µl 10x Tango Buffer (Thermo Scientific), 2 µl 10 mM ATP in a final reaction volume of 20 µl. Details of the destination vectors is given in Appendix 4.6, Table A4.6.2. The assembly reactions were incubated in a thermocycler for 3 cycles consisting of 20 min at 55°C and 10 min at 16°C, followed by 1 cycle of 10 min at 55°C and 20 min at 80°C. Reaction tubes were cooled and 1 µl of the assembly reaction used to transform 4 µl of heat-shock competent *E. coli*. The transformed cells were plated on LB containing ampicillin and X-Gal and incubated overnight at 37°C. A white colony for each construct was picked and grown overnight in 5 ml liquid LB at 37°C with continuous shaking at 200 rpm after which plasmid preparation was conducted.

Table 4.5. Ordered short guide RNA (sgRNA) oligonucleotide sequences. Upper-case and bold residues represent the sgRNA sequences homologous to *AILP1*, lower-case residues represent the introduced adaptor sequences, and italicised residues show Esp3I recognition sites.

sgRNA	Ordered oligonucleotide	sgRNA sequences + adaptors (5' – 3')	Destination vector ID
1	Forward	aatt*cgtctc*aattg**GCAGCCACTAACTCTTCCGG**gttta*gagacg*aatt	EC17777
	Reverse complement	aatt*cgtctc*taaac**CCGGAAGAGTTAGTGGCTGC**caatt*gagacg*aatt	
2	Forward	aatt*cgtctc*aattg**GAGACTTACCTAGGCCGGAG**gttta*gagacg*aatt	EC17778
	Reverse complement	aatt*cgtctc*taaac**CTCCGGCCTAGGTAAGTCTC**caatt*gagacg*aatt	
3	Forward	aatt*cgtctc*aattg**CCGGGTAAGGAGCTCTGTCA**gttta*gagacg*aatt	EC17779
	Reverse complement	aatt*cgtctc*taaac**TGACAGAGCTCCTTACCCGG**caatt*gagacg*aatt	

Level 2 Golden Gate cloning (Engler et al. 2008) was used to assemble a binary vector for the delivery of the three sgRNAs and *Cas9* gene into *Arabidopsis*. The parts used for the assembly of this vector are detailed in TableA4.006.2 and a vector map of the final construct is shown in Figure A4.006. The assembly of the vector was conducted in a Golden Gate reaction containing 40 fmol of each part (Table A4.006.2), 1.5 µl T4 DNA ligase (400,000 units μl^{-1}), 1 µl BpiI, 1.5 µl T4 DNA ligase buffer, and 1.5 µl 2 mg ml^{-1} BSA in a final reaction volume of 15 µl. The assembly reactions were incubated in a thermocycler for 25 cycles consisting of 3 min at 37°C and 4 min at 16°C, followed by 1 cycle of 5 min at 40°C and 5 min at 80°C. Reaction tubes were cooled and 1 µl of the reaction mix used to transform 10 µl of heat-shock competent *E. coli*. The transformed cells were plated on LB containing kanamycin and X-Gal and incubated overnight at 37°C. Four blue colonies for were picked and grown overnight in 5 ml liquid LB at 37°C. Subsequently plasmid preparation was conducted, and correctly assembled constructs screened in a restriction enzyme digest reaction containing 0.2 µg plasmid DNA, 1 µl AflII, and 1 µl Buffer 3.1 (NEB) in a final reaction volume of 10 µl. The digest was carried out at 37°C for 15 min and gel electrophoresis used to verify the correct fragment size of the digested product.

4.4.14 *Agrobacterium*-mediated *Arabidopsis* transformation and identification of *ailp1* mutants

Agrobacterium tumefaciens cells (strain gv3101, lab stock) were transformed with the final CRISPR-Cas9 construct by electroporation (50 µl of cells were transformed with ~100 ng of plasmid DNA). After electroporation, the cells were incubated in 750 µl of SOC for 2 h with continuous shaking at 150 rpm at 28°C and then plated onto LB plates containing gentamycin, rifamycin and ampicillin. Plated cells were incubated for two days at 37°C after which a colony was picked and grown for 24 h in 5 ml LB at 28°C with continuous shaking at 200 rpm. After 24 h, 500 µl of this starter culture was used to inoculate 200 ml of fresh LB containing gentamycin, rifamycin and ampicillin. The remaining starter culture was used for plasmid preparation and plasmid verification by digest and gel electrophoresis. The 200 ml subcultures were grown overnight at 28°C with continuous shaking at 200 rpm. The next day, the cells were collected by centrifugation at 3000 x g for 20 min at 4°C. The cell pellet was resuspended in 200 ml of infiltration media (10 mM $MgCl_2$, 5% sucrose [w/v], 1x Gamborg's vitamin solution [v/v] and 100 ng ml^{-1} benzylaminopurine) and placed on a rotary shaker (150 rpm) at 28°C for ~30 min. Just before *Arabidopsis* infiltration 0.3% [v/v] Silwet 77 was added to the infiltration media/cell mixture. *Arabidopsis thaliana* plants were selected at the point of floral emergence (approximately 4 weeks old). Some individuals were set aside and propagated to generate wild type progeny to use as control plants. The rest were transformed by floral dipping into the infiltration medium for ~20 s. Dipped plants were then sealed in a plastic bag and placed at 4°C overnight after which they were moved to the greenhouse.

T1 seeds were sown on soil and stratified for five days at 4°C. Once the plants had germinated (cotyledons visible) they were screened for transgene insertion by the application (spraying) of basta herbicide (530 µg l^{-1}) every two days for six days. Identified T1 transformants were transplanted to larger pots and genomic DNA (gDNA) was extracted from leaf tissue of each. Sanger sequencing was used to screen for mutations

in the *AILP1* gene using the primers *AtAILP1*_fwd (5'-TCTCTCCTTTGTCTCGTCTTCC-3') and *AtAILP1*_rev (5'-GGCTACAAACAGAGTGGAGGT-3'). The web tool ICE (inference of CRISPR edits) developed by synthego (Synthego 2019) was used to help identify those plants with heritable out-of-frame mutations. Four T1 lines were identified and approximately 30 seeds from each were sown on soil and stratified for five days at 4°C. From the resulting T2 lines, gDNA was extracted from the leaf tissue of each and those lines without the Cas9 T-DNA insertion were identified by PCR and gel electrophoresis using the primers Cas9_fwd (5'-TACCAAAACCGACCGACCAG-3') and Cas9_rev (5'-CCTGGGGTGCCTAATGAGTG-3'). Homozygous *AILP1* mutants in transgene-free lines were then identified using Sanger sequencing with the primers *AtAILP1*_fwd and *AtAILP1*_rev.

4.4.15 Phenotypic characterisation of *Arabidopsis* CRISPR mutants

Seeds from *Arabidopsis* plants with homozygous CRISPR-induced mutations and the segregating wild type were sown on soil and stratified at 4°C for five days before being moved to a controlled growth room (25°C, 200 µmol m^{-2} s^{-1} light). One plant occupied one pot (15 pots to a tray) and plants from the same line were distributed randomly across trays in column blocks of five. The position of trays in the growth room was also rotated every two days to minimise positional effects.

While in the rosette stage, the growth of *Arabidopsis* plants was monitored by taking aerial photos every 3-4 days. From photos, the visible rosette area of each plant was measured using Image J with the downloaded macro 'Measure Rosette Area Tool' (http://dev.mri.cnrs.fr/projects/imagej-macros/wiki/Measure_Rosette_Area_Tool).
Settings for this macro were kept the same for all photos taken on the same day. The rosette area detected by the macro was manually checked for each photo and any erroneously detected leaf area was removed. Similarly, it was checked that the tool captured the entire leaf area for each plant. Once the plants had bolted, plant height

(maximal extension between the soil level and the tips of shoots) was measured with a ruler every 3-4 days until the first siliques began to dry. Flowering time for each plant was measured as the date when the first anther opens, and the pollen is visible. On this date the number of fully expanded rosette leaves was also counted.

Photosynthetic measurements were taken for the youngest fully expanded rosette leaf on *Arabidopsis* plants just before the onset of flowering (~4-week-old). Carbon assimilation (A) was measured over a range of intercellular carbon concentrations (C_i) using a LI-COR 6800 (Li-COR Bioscience). Chamber conditions kept constant were as follows: 300 µmol s^{-1} flow rate, 50% relative humidity, 10,000 rpm fan speed, 25°C, and a photosynthetic photon flux density of 1,500 µmol m^{-2} s^{-1}. The carbon assimilation measurements were performed after the leaf had equilibrated to 400 µmol mol^{-1} CO_2. Measurements were then taken in intervals of ~3 min at the following CO_2 concentrations: 400, 300, 200, 100, 50, 25, 400, 400, 600, 800, 1000, 1200, 1500 µmol mol^{-1}. The leaf was then equilibrated back at a CO_2 concentration of 400 µmol mol^{-1} and used for light response measurements. Here, the CO_2 concentration was held constant at 400 µmol mol^{-1} and carbon assimilation measured at ~3 min intervals at the following light intensities: 1500, 1200, 1000, 800, 600, 400, 200, 100, 50, 25, 0 µmol photons m^{-2} s^{-1} (90% red light, 10% blue light).

The composition of carbon and nitrogen in biomass was determined by the stable isotope lab at the School of Archaeology at the University of Oxford. Sample preparation was conducted as follows: *Arabidopsis* plants at the onset of flowering (~4-weeks old) were selected and biomass collection was conducted in the middle of the photoperiod on a single day. For each plant, the two youngest, fully expanded rosette leaves were quickly excised using a razor blade and deposited in a single 2 ml Eppendorf tube before being flash frozen in liquid nitrogen. The samples were freeze-dried (Christ, Alpha 2-4 LD) for three days and then homogenised in a tissue lyser (Thermo Scientific, TissueLyser II) and weighed into tin capsules. The samples were then sent to the School of Archaeology for analysis.

Chapter 5: General discussion

"Novelties come from previously unseen association of old material. To create is to recombine."

Francis Jacob – Evolution and Tinkering, Science, 1977
(Jacob 1977)

5.1 A first attempt to identify the changes in organellar protein targeting that occurred during the evolution of a diverse set of eukaryotic species

Prior to the work conducted in this thesis, studies investigating the evolution of organellar proteomes were limited to comparisons of organellar protein content between a small number of species. Such studies, both experimental and bioinformatic, have shed light on the commonalities and differences in the protein repertoire of organelles from different species (Richly and Leister 2004; Gabaldón et al. 2006; Gabaldón and Huynen 2007; Szklarczyk and Huynen 2010; Gray 2012; Schaeffer et al. 2014; Leister 2016), and have been used to estimate the minimal protein content of organelles in the last eukaryotic common ancestor (Gabaldón and Huynen 2003; Gabaldón and Huynen 2007; Gray 2015). Together these studies have demonstrated that there is diversity in organellar protein content across species and suggest that organellar proteomes have not been static since their inception in the eukaryotic cell.

The majority of organellar proteins are encoded in the nucleus and thus organellar proteomes will be shaped by molecular sequence evolution in the nucleus and evolutionary events such as gene duplication and gene loss. In some cases, changes in the sequence composition of a gene may give rise to an alteration in the subcellular targeting of the encoded protein and it has been posited that this may be a mechanism by

which organellar proteomes are modified throughout evolution in a lineage-specific manner (Byun-McKay and Geeta 2007; Marques et al. 2008; Gabaldón and Pittis 2015). However, few studies have investigated the extent to which changes in protein targeting occur during evolution, and those that have are limited to the identification of paralogous proteins with divergent subcellular localisations (Marques et al. 2008; Byun and Singh 2013; Liu et al. 2014). While these studies provide evidence that changes in protein targeting do occur, they rely on BLAST scores rather than phylogenetic trees to identify paralogs and, without interrogation of a gene tree, it is not possible to ascertain when during evolution a change in protein targeting occurred or whether a particular organellar target signal was gained or lost. In addition, they fail to identify changes in protein targeting independently from gene duplication events.

Considering this, the aim of Chapter 2 was to develop a phylogenetic method to identify when organellar target signals were gained and lost from proteins during the evolution of a large and diverse set of eukaryotic species. To do this I co-opted a method (ancestral state estimation) that is typically used for estimating the evolution of phenotypic traits across a species phylogeny. This appears to be the first time that maximum likelihood ancestral state estimation has been used to predict genome-wide changes in protein targeting across gene-tree phylogenies.

5.2 Two recent studies investigating chloroplast transit peptide gains corroborate several findings in this thesis and provide insight into the origin of novel transit peptides

Since the publication of the work in Chapter 2, two additional studies have been published (by the same group) that similarly use gene trees and protein target signal prediction to identify changes in chloroplast targeting during the evolution 15 diverse plant species (Christian, Hewitt, Roalson, et al. 2020; Christian, Hewitt, Nelson, et al. 2020). Both studies focus specifically on the identification of gains in chloroplast targeting which were only

identified in gene-trees if all extant proteins with a predicted chloroplast transit peptide formed a monophyletic group. This strategy is like that used in Chapter 2 to identify highly supported changes in protein targeting except that in these two later studies gene-trees containing multiple gains, or losses, were not considered. The overarching aims of these two studies were different to that presented in this thesis: the first sought to optimise the prediction workflow for the identification of chloroplast-targeted proteins and describe a novel gene-clustering (i.e., orthogroup inference) method (Christian, Hewitt, Roalson, et al. 2020) while the second used this data set to investigate the diversity in chloroplast transit peptide composition and the mechanisms by which novel chloroplast transit peptides are acquired (Christian, Hewitt, Nelson, et al. 2020). Nonetheless, supportive of the findings in this thesis, an abundance of proteins were identified to have gained a chloroplast transit peptide during plant evolution (Christian, Hewitt, Roalson, et al. 2020). In addition, it was similarly found that proteins that had gained a chloroplast transit peptide were enriched for functional terms related to the regulation of gene expression, protein binding & regulation, and metabolism & biosynthesis (Christian, Hewitt, Nelson, et al. 2020).

In addition to the identification of gains in chloroplast targeting during the evolution of multiple plant species, the authors of these studies also constructed gene trees from the pan-genomes of *Arabidopsis thaliana* and *Brachypodium distrachyon* to look for more recent acquisitions of chloroplast targeting. A large number of gains in chloroplast targeting were identified in the pan-genome data sets (180 in *Arabidopsis* and 2,272 in *B. distrachyon*) further supporting the notion that changes in protein targeting are rife during evolution (Christian, Hewitt, Nelson, et al. 2020). Unlike the multi-species data set that only identified gains in chloroplast targeting, in the pan-genome data sets losses in chloroplast targeting were also identified and reported to occur approximately four times more frequently than gains (Christian, Hewitt, Nelson, et al. 2020). A higher rate of target signal loss is a trend that was also observed in the data set in Chapter 2 of this thesis. As

hypothesised in Chapter 2, this may be because the accumulation of mutations leading to target signal loss are more likely to arise than mutations leading to target signal gain. Alternatively, this pattern may be the result of some inherent bias in the prediction of target signals (i.e., the number of false negatives and false positives) that then biases the estimation of ancestral states in gene trees. It is difficult to assess or model the impact of such misprediction bias however as reports on the false positive and false negative rates of TargetP come from tests with highly curated data sets which are not reflective of the whole genome and there is disagreement between studies as to whether the false positive or false negative rate is higher (e.g., Emanuelsson et al. 2000; Christian, Hewitt, Roalson, et al. 2020). Nonetheless, the finding in Chapter 2 that for each of the different types of organellar target signal (the prediction for which will have different inherent biases) there was a higher rate of loss lends support to this being a biological phenomenon. This is further corroborated by the findings in the pan-genome data set from (Christian, Hewitt, Nelson, et al. 2020) that use a different target signal prediction pipeline and phylogenetic method for the identification of target signal gains and losses to the one used in this thesis. It is a shame that Christian, Hewitt, Nelson, et al. 2020 did not report on the rate of losses in the cross-species data set as it would have been interesting to know if the phenomenon holds true over a similar species phylogeny to the one used in Chapter 2.

Lastly, Christian, Hewitt, Nelson, et al. 2020 also investigated the relative contribution of different types of molecular sequence mutations to the evolution of novel chloroplast transit peptides. Interestingly, while transit peptide origins identified at the multi-species scale were found to have originated predominantly from large N-terminal insertions, at the pan-genome scale almost all origins were the result of simple nucleotide substitutions (Christian, Hewitt, Nelson, et al. 2020). This suggests that the likelihood of a novel chloroplast transit peptide being retained over long evolutionary timescales may be biased based on the mechanism by which it arose. One hypothesis to explain this is that the insertion of a sequence encoding all or part of a target signal is less likely to disrupt protein

function than the substitution of existing residues. If this were the case, then a change in protein targeting originating from a DNA insertion would be more likely to be fixed in a population than a change arising from DNA substitutions. Further support for this notion comes from a study that compared mutations in the N-terminal region of paralogous proteins with divergent predicted targeting in the genomes of several plant species (Byun and Singh 2013). Here it was also found that complete and partial insertions/deletions accounted for a much higher proportion of the N-terminal mutations leading to a change in targeting than single nucleotide substitutions (Byun and Singh 2013), although it is not entirely clear in this study how much of the N-terminal region was taken into account.

In summary, these two recent papers also paint a dynamic picture of gains and losses to the chloroplast proteome during evolution and provide interesting insights into how chloroplast transit peptides arise during evolution. In the future it will be interesting to establish whether patterns of target signal gain and loss are replicated among other major groups of eukaryotes. Indeed, it is my supposition that the evolutionary forces leading to changes in organellar protein targeting will be distinctly different in other groups of eukaryotes. For one, most other eukaryotes don't possess a second endosymbiotic energy organelle and thus the expansion and complexity of organellar regulatory proteins will be profoundly different. Furthermore, genome architecture varies between different eukaryotic groups in ways that would be expected to affect the propensity for target signals to be gained or lost based on the findings in this thesis. The most notable facet of this being the varied rate of polyploidisation, and concomitantly gene duplicate persistence, between different groups of eukaryotes (Lynch and Conery 2000; Leitch et al. 2004).

5.3 Changes in protein targeting following gene duplication may arise through mechanisms of gene neofunctionalisation or subfunctionalisation

It was Ohno who first theorised that gene duplication, which gives rise to genetic redundancy, allows for the accumulation of otherwise 'forbidden' mutations at genetic loci and is thus a major driver for the evolution of new gene functions (Ohno 1970). Since Ohno's seminal work, gene duplication has been established as an important, but not exclusive, mechanism by which genes may acquire new functions (neofunctionalisation) (Kaessmann 2010). It was for this reason that we hypothesised in Chapter 2 that changes in protein targeting may occur at a higher rate following gene duplication. Under the gene neofunctionalisation model duplicate genes diverge in function such that one paralog retains the ancestral function, while the other paralog develops a new function (Panchy et al. 2016). The changes in protein targeting following a gene duplication identified in Chapter 2 and Chapter 3 were thus considered to be neofunctionalisation events (Figure 5.1 A). It should also be considered however that changes in protein targeting following gene duplication could occur though mechanisms of subfunctionalisation rather than neofunctionalisation. In the subfunctionalisation model, ancestral gene function gets partitioned between duplicates such that there is no overall change in ancestral functionality (reviewed in Panchy et al. 2016). Thus, if an ancestral protein were targeted to more than one location in the cell (i.e., dual-targeted) then, following gene duplication, the paralogs may diverge so that the ancestral localisation becomes partitioned between the paralogs (Figure 5.1 B). The targeting phenotype of extant proteins will be the same whether divergence in targeting of paralogs occurred through mechanisms of neofunctionalisation or subfunctionalisation (bottom panel of Figure 5.1). Thus, from the gene trees analysed in Chapter 2 it is not actually possible to discern whether paralog divergence occurred through the acquisition of novel targeting or through the splitting of ancestral targeting.

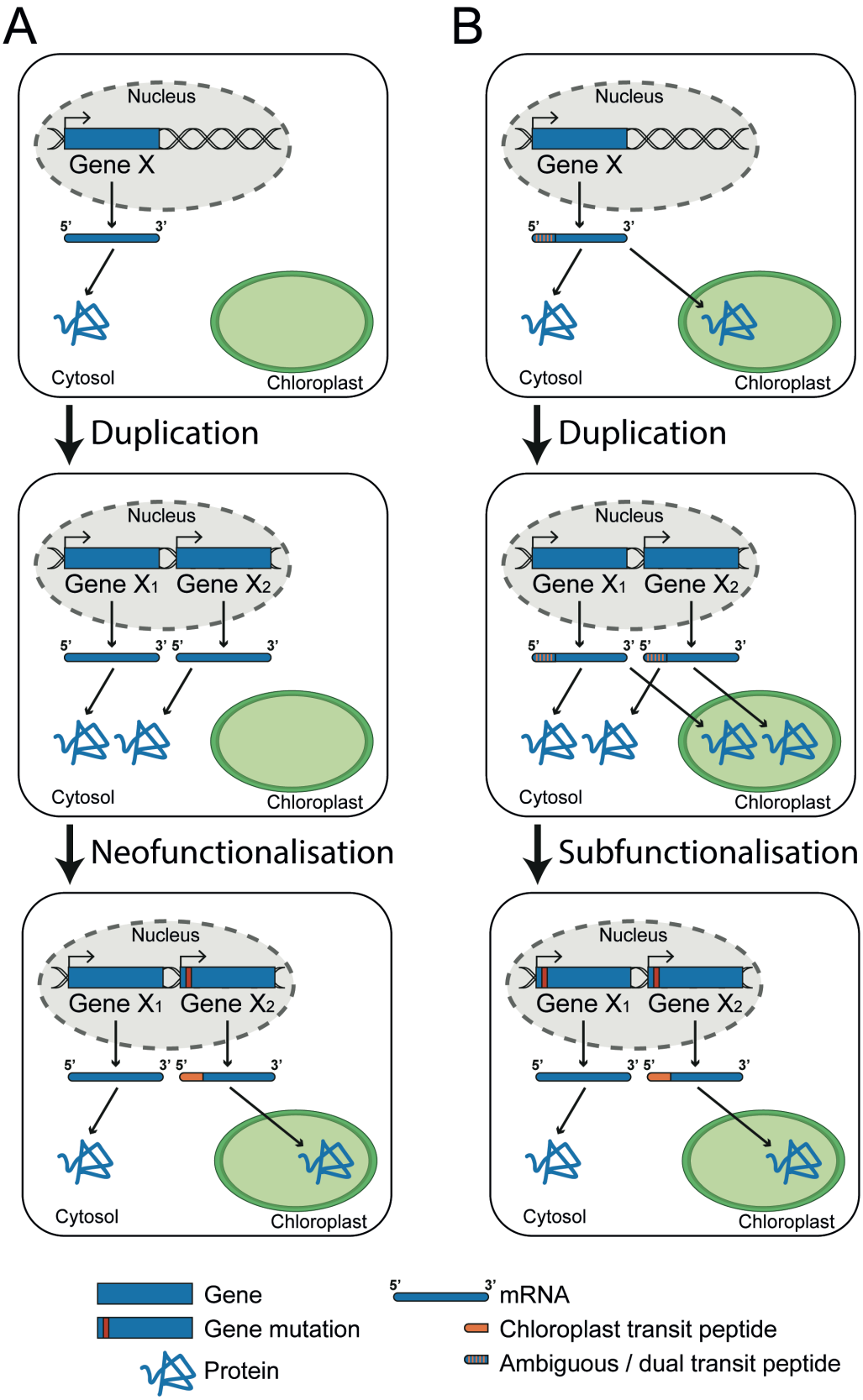


Figure 5.1. Changes in protein targeting following gene duplication can occur through mechanisms of neofunctionalisation or subfunctionalisation. A) In this

scenario an ancestral gene encodes a cytosolic protein. Following gene duplication, mutations in the N-terminal region of one of the paralogs leads to the acquisition of a chloroplast transit peptide (neofunctionalization). **B)** In this scenario an ancestral gene encodes a protein that is dual targeted to the cytosol and chloroplast. Following gene duplication, mutations in the N-terminal region of both paralogs leads to the partitioning of the ancestral localisation between the two paralogs such that one is localised to the cytosol, and one is localised to the chloroplast (subfunctionalisation).

There has been much theorising as to the relative importance of different adaptive and nonadaptive forces in the short- and long-term retention of duplicate genes (for example Force et al. 1999; Bergthorsson et al. 2007; Innan and Kondrashov 2010; Assis and Bachtrog 2013; Copley 2020). Because the accumulation of several mutations may be needed before gene neofunctionalisation can occur, it has been suggested that other mechanisms of paralog retention, such as gene dosage and subfunctionalisation, may be more important in the short-term retention of duplicate genes (Ohno 1970; Lynch and Conery 2000). A question thus arises as to whether subfunctionalisation of protein targeting may be a more common occurrence than neofunctionalisation of protein targeting for those changes in targeting that were identified to have occurred immediately after a gene duplication event (i.e., before the next speciation event in the gene tree). Currently it is very difficult to investigate this question because the extent of dual targeting among extant proteins is not clear. Although there are several examples of dual-targeted proteins, experimentally identifying all dual-targeted proteins in a proteome is challenging given that the multiple localisations of a protein might only become apparent in certain tissue types or under certain physiological conditions. Furthermore proteome-wide experimental studies identifying dual-targeted proteins often don't differentiate between protein isoforms encoded by multiple genes (i.e., paralogs) or by a single gene (what I consider here as dual-targeted). For example, a recent study that mapped the subcellular

distribution of 12,003 human proteins using antibody-based immunofluorescence found that 51% of proteins could be detected in more than one subcellular compartment (Thul et al. 2017). The antibodies that were used in this study however were designed to target as many isoforms of a given protein as possible and thus the proportion of individual genes from a single locus that encode a dual-targeted protein is still unknown. Two computational methods have been developed to date that attempt to predict dual-targeted proteins in plants, both of which employ machine-learning methods to predict ambiguous target signals that could direct a protein to both the chloroplast and mitochondrion (Mitschke et al. 2009; Sperschneider et al. 2017). Using one of these approaches it was estimated that approximately 500 proteins in *Arabidopsis* are dual targeted to the chloroplast and mitochondrion (Mitschke et al. 2009). Currently though these methods are not sufficient to estimate all types of dual-targeted proteins that may be present in the proteome of a species.

Future improvements in the prediction of dual-targeted proteins and the expansion of species-specific databases with highly curated protein localisations over multiple tissue types and conditions may allow for the phylogenetic identification of changes in protein targeting by sub- and neo-functionalisation. Indeed, a recent phylogenetic analysis of a plant ascorbate peroxidase combined with experimental validation of protein targeting across several species has provided evidence for divergence in protein targeting by subfunctionalisation (Xu et al. 2013). For now, however, a genome-wide phylogenetic study like the one conducted in this thesis is limited to the identification of changes in protein targeting as binary gains/losses.

5.4 The ancestral state estimation method used is expected to underestimate the total number of changes in protein targeting that occurred during plant evolution

The work presented in Chapter 2 identified changes in protein targeting and revealed that hundreds of proteins have undergone a change in targeting in both ancient and more recent plant lineages. The changes in protein targeting that were identified are however expected to underestimate the total number of changes that have occurred during the 450 million years of land plant evolution that is represented by this species set. Firstly, I applied a stringent criterion for the identification of changes that meant that any changes in protein targeting that were not conserved during evolution, either because of subsequent changes in targeting or gene loss, would not be identified.

Secondly, the target signal prediction tools I used in this study are only able to predict the canonical N-terminal target signals of chloroplast, mitochondrion and secretory pathway proteins and the canonical motifs of peroxisomal target signals. Several proteins are however known to be sorted in the cell by non-canonical targeting pathways. Many of these non-canonical pathways have been described and typically involve certain classes of proteins. For example, signal-anchored and tail-anchored proteins of the chloroplast and mitochondrion outer membrane don't contain an N-terminal target signal and instead targeting is dependent on the presence of a single, moderately hydrophobic transmembrane domain, although it is not currently clear how the cell differentiates between chloroplast- and mitochondrial-anchored proteins (Rossig et al. 2013). In another example, some chloroplast proteins have been shown to be targeted to the chloroplast via the endomembrane system and thus contain a signal peptide at their N-terminus (Reyes-Prieto et al. 2007; Li and Chiu 2010). It has been found that chloroplast proteins destined for post-translational glycosylation are transported in this way because glycosylation takes place in the endoplasmic reticulum or Golgi apparatus (Reyes-Prieto et al. 2007; Li and

Chiu 2010). Once glycosylated, these glycoproteins are transported via vesicles through the endomembrane system and onto the plastid envelope (Baslam et al. 2016). The mechanism of glycoprotein transport through the plastid membrane is however currently unclear. Non-canonical targeting processes have also been described for peroxisomal proteins. In particular, some peroxisomal proteins have been found to possess non-canonical C-terminal peroxisomal targeting signal 1 (PTS1) tripeptides (Chowdhary et al. 2012). In these cases, the non-canonical PTS1 sequences are only functional when they exist in combination with auxiliary enhancer residues located between positions -4 and -12 to the C-terminal tripeptide (Chowdhary et al. 2012). Together these examples serve to illustrate that several non-canonical organellar targeting pathways are known to exist and thus changes in targeting among proteins that use these pathways will not be captured in Chapter 2. Hopefully, continued characterisation of the protein signatures that delineate protein targeting through these non-canonical pathways will enable the prediction of such proteins by future target signal prediction methods.

Lastly, it is expected that several proteins will be dual targeted in the cell (although as mentioned previously the proportion that are is unclear) and thus the changes in protein targeting among such proteins will be missed or misidentified by the methods in this thesis. Dual targeting may arise due to a protein containing an ambiguous target signal (as in Figure 5.1) or through the production of differently targeted protein isoforms from a single gene (reviewed in Kunze and Berger 2015; Sharma et al. 2018). This latter case may occur due to alternative splicing, exon skipping, alterative transcription initiation, or alternative translation initiation. The frequency with which alternative transcript isoforms give rise to differentially targeted proteins is unknown. Increasingly however, high-throughput RNA (Law et al. 2015) and 5' mRNA (Wang et al. 2016) transcriptome data sets, as well as new tools such as multi-colour translation reading frame imaging (Boersma et al. 2019), will permit the genome-wide identification of genes that produce differentially targeted protein isoforms. Lastly, in some rare cases protein targeting has been shown to be post-

translationally regulated. One of the best characterised examples of this is for the single stranded DNA binding protein WHIRLY (WHY1) that can be targeted to either the chloroplast or nucleus depending on the phosphorylation state of the protein (Ren 2017). WHY1 functions in chloroplast retrograde signalling and protein phosphorylation of WHY1 is regulated by environmental and developmental cues (Ren 2017). In this case, the WHY1 protein contains both a chloroplast transit peptide and a nuclear localisation signal, with the latter only being recognised by the nuclear-import machinery after protein phosphorylation in the cytosol (Ren et al. 2017; Lin et al. 2019). It is currently not clear however if WHY1 targeting to the nucleus occurs after the protein has already been localised to the chloroplast and thus lost its transit peptide, or if phosphorylation of WHY1 results in the nuclear-localisation signal overriding the chloroplast transit peptide.

In summary, protein targeting to organelles can be highly dynamic and thus, by only looking for the presence/absence of canonical N-terminal target signals, it is not possible to capture the full complement of changes to organellar proteomes during evolution. Currently it is presumed that the majority of organellar proteins are targeted to a single organellar compartment by canonical targeting pathways, however more genome-wide analyses are needed to assess the prevalence of proteins that don't meet this canon.

5.5 What is the functional significance of identified changes in protein targeting?

In Chapter 2 I conducted a broad functional characterisation of the proteins that had been gained and lost to the different organelles during plant evolution. This revealed that many of the proteins gained and lost to the chloroplast and mitochondrion during evolution had functions related to gene and protein regulation. This may not be entirely surprising given that in plants regulatory proteins often belong to large paralogous gene families (Seoighe and Gehring 2004) and thus the rate of change in protein targeting could be expected to be higher among such proteins (although I also found that there was an enrichment, as

well as an abundance, of changes among regulatory proteins). A number of RNA-binding protein families are known to contain members with differential localisation between the chloroplast, mitochondrion, and nucleus (Lee and Kang 2020). In particular, several differentially localised proteins have been described among the RNA-binding pentatricopeptide repeat (PPR) proteins (Manna 2015). PPR proteins function to regulate chloroplast and mitochondrial (and occasionally nuclear) gene expression and are a highly expanded gene family in plants with over 400 PPR genes identified in *Arabidopsis* alone (Lurin et al. 2004; Hammani et al. 2011; Gutmann et al. 2020). Given that PPR proteins belong to one of the largest gene families in plants, many of the changes in targeting identified in Chapter 2 could have occurred among PPR proteins. This was not something I looked at, but it would be an interesting avenue of future investigation. Although regulatory proteins such as PPR proteins are known to be differentially localised to the endosymbiotic organelles and the nucleus, the analysis in Chapter 2 is the first to show that there has been continued widespread modulation of the regulatory proteins targeted to these organelles throughout plant evolution.

The analysis in Chapter 2 also sought to describe the general patterns of target signal gain and loss during plant evolution. To make interpretations as to the functional significance of identified changes in protein targeting it is necessary to incorporate gene expression and/or protein activity data as was done to investigate changes during the evolution of C_4 photosynthesis in Chapter 3. It is hoped however that the data set constructed in Chapter 2 (or a reiteration of the methods) could be used to study the changes in protein targeting that occurred during the evolution of other plant lineages or plant traits.

5.5.1 Investigation of the changes in protein targeting that occurred during the evolution of C_4 photosynthesis

In Chapter 3 of this thesis, I used the ancestral state estimation method developed in Chapter 2 to identify the changes in protein targeting that occurred on branches of the

species tree along which C_4 photosynthesis evolved. Hundreds of evolutionarily conserved changes in protein targeting were identified coincident with evolution of C_4 photosynthesis. Many of these changes are however likely to be associated with cellular functions and plant traits not related to C_4 photosynthesis. Therefore, to identify those changes that may have been associated with C_4 photosynthesis, I incorporated into the analysis transcriptome data for extant genes descendent from a change in protein targeting. Although no changes in protein targeting were identified among known C_4-cycle proteins concomitant with the evolution of C_4 photosynthesis, several other changes were identified among proteins whose extant members have gene expression patterns suggestive of a C_4-function (high leaf expression and/or differential mesophyll and bundle sheath expression). Of particular note was the identification of a beta-glucosidase (B-GLU1) that was estimated to have been relocalised from the secretory pathway to the chloroplast in an ancestor of the four C_4 grass species. Transcript analysis suggests that B-GLU1 may be involved in leaf-development in maize and concomitantly previous characterisation of a chloroplast-localised B-GLU1 paralog from maize suggests this protein may function to increase levels of cytokinin in the cell (Brzobohatý et al. 1993). As several chloroplast biogenesis transcription factors operate downstream of cytokinin, this protein could be a promising target for engineering efforts that seek to promote chloroplast development in the vascular tissue of C_3 plants in order to 'prime' them for the instalment of a C_4 pathway. The biggest of these engineering strategies is the C_4 rice project, and indeed the need to 'functionalise' rice bundle sheath cells may pose an 'anatomical roadblock' to the success of this project (Ermakova et al. 2020).

Several other changes in protein targeting were identified in the analysis in Chapter 3 that may warrant further investigation, but that could not be immediately attributed to a function in C_4 photosynthesis. Although these proteins were expressed in the leaf, and in some cases more abundantly in the mesophyll or bundle sheath cells, most lacked any functional characterisation and so making inferences about the importance of these proteins to a

functioning C_4 cycle is difficult. Given more time, one way this analysis could have been improved is through the inclusion of C_3 grass species from the PACMAD clade that would phylogenetically intervene the Andropogoneae C_4 clade and the Paniceae C_4 clade. With the inclusion of such species it would have been possible to discern any changes in targeting that occurred in parallel concomitant with the origin of C_4 in two independent lineages (or indeed several C_4 lineages given the multiple origins of C_4 in the PACMAD clade). Thus, such changes would have a second piece of supporting evidence, along with the transcript analysis, to suggest a change in protein targeting was linked to the evolution of C_4. A similar approach has been used previously to identify genes that were independently duplicated in parallel during the evolution of the same two C_4 lineages investigated in this thesis (Emms et al. 2016). In doing so, (Emms et al. 2016) identified 21 genes that duplicated in parallel during the evolution of C_4 and whose expression suggested a parallel recruitment into C_4 photosynthesis.

5.6 More work is needed to identify the function of AILP1

AILP1 was identified to putatively encode an enzyme that might metabolise aspartate in the bundle sheath cells of maize based on two main lines of evidence: the shared ancestry of *AILP1* with the gene sequence for the enzyme asparagine synthetase and the preferential expression of *AILP1* in the bundle sheath cells of maize. The *AILP1* gene sequence is homologous to the domain in asparagine synthase that catalyses the hydrolysis (deamination) of glutamine to yield ammonia. As the canonical aspartate aminotransferase reaction similarly involves the deamination of aspartate, it was hypothesised that AILP1 may function to catalyse the deamination of aspartate as part of a non-canonical aspartate aminotransferase reaction or as part of another metabolic pathway that would process aspartate to oxaloacetate. An initial assay was thus devised to test for this activity but unfortunately AILP1 was not amenable to the chosen purification strategy. With no prior experimental characterisation or established assay for AILP1 function, designing a purification strategy is challenging. Ideally when optimising a

purification strategy for an enzyme, an assay for enzyme function would be conducted at every step in the purification protocol to ensure that it is suitable for the recovery of biologically active enzyme (as asserted by Burgess 2009). Without a known activity assay, then the success of any given purification strategy cannot be assessed. Indeed, if no biological activity is detected at the end of the assay, it could be due to either a problem with the purification conditions or that the wrong assay is being used. Nonetheless, small scale protein purification in *E. coli* or a cell-free system can be a quick way to obtain sufficient protein to carry out functional assays and I hoped that with a rapid purification pipeline I could test for AILP1 activity over a range of assay conditions in a short time. Unfortunately, however, I was not able to purify soluble protein using this system. An alternative strategy to purifying heterologous protein from a non-native expression system is to purify native protein from plant tissue by precipitating out a protein sample and conducting ion-exchange chromatography. Without an assay this would still be a challenging endeavour however and thus more *in planta* work may be preferable to try and narrow down the putative function of the AILP1.

The perturbation in the carbon to nitrogen ratio of *ailp1* knockout lines in *Arabidopsis* suggests that AILP1 may function in nitrogen metabolism. In the short-term, this could be further investigated by conducting a metabolite screen of leaf tissue from *ailp1* plants to ascertain the relative abundances of different free amino acids in the leaf. The enrichment or depletion of particular amino acids in the mutant plants may offer clues as to the amino acid metabolic pathway, if any, AILP1 functions within. In the long-term it would be ideal to generate *AILP1* knockout mutants in maize so that the requirement of AILP1 in maize C_4 photosynthesis could be directly tested. Curiously, AILP1 has recently been identified in a screen of proteins with a high protein turnover rate in *Arabidopsis* leaves (Li et al. 2017). In this study AILP1 was identified in the set of 20 most rapidly degraded proteins in the leaf (Li et al. 2017). I am not sure what the functional implications of this are, and AILP1 is not discussed in the text. Other proteins in this top-20 set include other

aminotransferases such as methionine aminotransferase (BCAT4) and gamma-aminobutyrate transaminase (GABA-T), and high rates of protein degradation have been linked to a multitude of protein attributes such as aggregation propensity, membership of multi-subunit protein complexes, location in the cell, and the presence of particular protein domains (Li et al. 2017). This feature of AILP1 may be something to bear in mind for future work on this protein.

5.7 It is still unclear what happens to aspartate in the bundle sheath cells of maize

The analysis of aspartate aminotransferase (AspAT) transcript abundance in previously published data sets (Chapter 3) supports previous reports that there is significantly less AspAT mRNA in the bundle sheath cells of maize compared to the mesophyll cells (Tausta et al. 2014; Denton et al. 2017; Schlüter et al. 2019). It raises the question therefore as to whether there is sufficient AspAT activity in the bundle sheath cells to support flux of aspartate through the maize C_4 cycle. The attempted characterisation of AILP1 in Chapter 4 of this thesis was conducted based on the hypothesis that there may be another enzyme in the bundle sheath cells of maize that is metabolising all or part of the aspartate pool. The alternative hypothesis is that there is sufficient activity of AspAT in the bundle sheath. From the transcriptomic evidence this seems unlikely given the low expression of AspAT in the bundle sheath (<200 TPM in the bundle sheath - see Chapter 3) and that transcript abundance of C_4 proteins generally correlates well with protein abundance (Ponnala et al. 2014). Nonetheless, further investigation is needed to establish if this is the case. To date, there has only been one study (Denton et al. 2017) that has investigated the enzyme activity of AspAT extracted from maize leaf bundle sheath cells. In this study it was reported that, across mature leaf sections, between 0 and 20% of leaf AspAT activity was attributable to enzyme from bundle sheath cells. Unfortunately, however, it is not clear in this study which AspAT isozyme was extracted, the direction AspAT activity was assayed in, or the conditions of the assay. Further activity assays are thus needed to confirm

whether if there really is little to no AspAT activity in maize bundle sheath cells and to establish the kinetics of any enzyme that is there.

There have been a couple of maize leaf bundle sheath chloroplast proteome studies and these have provided evidence that there is at least some AspAT protein in the bundle sheath chloroplasts (Majeran et al. 2005; Friso et al. 2010). In one of these studies, protein abundance of AspAT in the bundle sheath cells was found to be 50% of that in mesophyll cells as estimated by mass spectrometry (Friso et al. 2010) and in the second study protein abundance of AspAT in the bundle sheath was estimated to be on average 20% of that in the mesophyll as estimated by a range of proteomic techniques including mass spectrometry and two-dimensional gel electrophoresis (Majeran et al. 2005). Thus, it seems that there is at least some AspAT protein in the chloroplasts of the bundle sheath in maize. It is also possible that AspAT may be present in other subcellular compartments, although again mRNA abundance suggests that there would not be much. In my transcriptomic analysis in Chapter 3, one RNAseq data set (Chang et al. 2012) showed a predicted peroxisomal AspAT to have the highest transcript abundance in the bundle sheath (approximately 100 TPM) while in another RNAseq data set (Tausta et al. 2014) a predicted chloroplast AspAT had the highest transcript abundance (approximately 200 TPM). It is not clear yet if this peroxisome AspAT functions in the C_4 cycle and again, confirmatory AspAT activity assays are needed.

It is possible that the AspAT present in the bundle sheath cell of maize is sufficient for aspartate flux through the C_4 cycle and that AspAT is necessarily higher in the mesophyll cell for some other reason. For instance, it could be that the kinetics of the AspAT enzyme mean that less enzyme is needed in the bundle sheath cells or that there is a higher substrate to product ratio in the bundle sheath. The former of these may not be likely though given that all other C_4 AspATs that have been kinetically characterised show max velocities in the direction of oxaloacetate utilisation and thus aspartate production (i.e. the mesophyll reaction) (Hatch 1973; Taniguchi and Sugiyama 1990; Taniguchi et al. 1995).

Although to the best of my knowledge only AspAT kinetics from NAD-ME species have been characterised as such. The substrate to product ratio for AspAT in the two cell types has not been determined; aspartate concentration is higher in the mesophyll cell concomitant with the diffusion gradient needed for aspartate movement to the bundle sheath (Arrivault et al. 2017) but, because oxaloacetate is unstable and thus difficult to measure, the relative abundance of oxaloacetate is unknown. Another suggestion for a higher mesophyll content of AspAT is that some aspartate that diffuses into the bundle sheath cells does not contribute to C_4 metabolism but rather is used for the synthesis of amino acids. If this is the case then it is expected that aspartate would be rapidly converted into these amino acids because the kinetics of radiolabelled CO_2 incorporation into aspartate suggests that there is no photosynthetically inactive pool of aspartate in the mesophyll cell (Arrivault et al. 2017). The amino acids synthesised from aspartate in the plastid are isoleucine, tyrosine, methionine, phenylalanine, threonine, lysine, and isoleucine (De La Torre et al. 2014). Recent radiolabelled CO_2 metabolite studies in maize (Arrivault et al. 2017; Denton et al. 2017) did not measure the incorporation of radiolabel into these amino acids and thus doing so would be an informative avenue of future research.

In summary, the fate of aspartate is a piece of the puzzle yet to be solved in our understanding of C_4 photosynthesis in maize. Current evidence suggests that the activity of aspartate aminotransferase in the bundle sheath cells of maize may not be compatible with our current model of C_4 photosynthesis in this species. While it is possible that an as of yet uncharacterised enzyme is the missing piece in this puzzle (as was explored in this thesis) there is also a number of outstanding lines of enquiry that may provide insight to the relative roles of aspartate aminotransferase in the maize mesophyll and bundle sheath cells. Furthermore, it is currently not clear to what extent aspartate is used as a C_4 transfer metabolite in other closely related C_4 NADP-ME species such as *Sorghum bicolor* and *Setaria viridis*. Quantifying aspartate flux through these species, especially *S. bicolor*

which shares a C_4 origin with maize, may provide new insight into the route of aspartate flux in NADP-ME species that use aspartate as a secondary C_4 transfer metabolite.

Appendices

Appendix 2.1: Worked example of, and justification for, the 75% retention score used in the identification of changes in protein targeting

Figure 2.3 in Chapter 2 shows the species tree with the number of changes in organellar targeting mapped to each internal branch. These numbers were obtained following a data filtration step. This filter required that in order for a change in protein targeting to be considered real, ≥75% of the extant proteins descendant from a branch on which a change occurred must display the changed organellar targeting state, while ≥75% of genes descendant from the sister branch must display the ancestral state.

As a worked example of this, Figure A2.1.1 shows the orthogroup tree for orthogroup OG0002487. This is the orthogroup containing the enzyme 2-phosphoglycolate phosphatase (PGLP), the first enzyme in the photorespiratory pathway. In *Arabidopsis* there are two genes in this orthogroup, the chloroplast-localised PGLP1 gene which is involved in photorespiration, and a second gene called PGLP2 which has been shown to be localised to the cytosol and currently has an unknown function (Schwarte and Bauwe 2007). In Figure A2.1.1 the orthogroup tree for PGLP is shown. As in *Arabidopsis*, most angiosperm species have at least one copy of PGLP1 and PGLP2 and prediction of chloroplast targeting among extant proteins also suggests that most angiosperm species have a chloroplast-targeted and a cytosolic-targeted PGLP protein (Figure A2.1.1). At each internal branch in the tree, the likelihood that the ancestral protein represented by that branch has a chloroplast transit peptide is given (Figure A2.1.1, green pie chart). These likelihood scores were computed using maximum likelihood ancestral state estimation given the predicted targeting of extant PGLP proteins in the tree. A winner takes all approach was used such that internal branches are denoted as ‘chloroplast targeted’ if their likelihood of being so was ≥50%. By this approach, three occurrences of a change in

chloroplast targeting can be identified in the tree which are indicated by an asterisk or a cross. However, only one of these changes passes the 75% retention score in the extant proteins (indicated by a red asterisk) and so is the only one that would have been considered in the genome-wide comparative analysis presented in this Chapter 2.

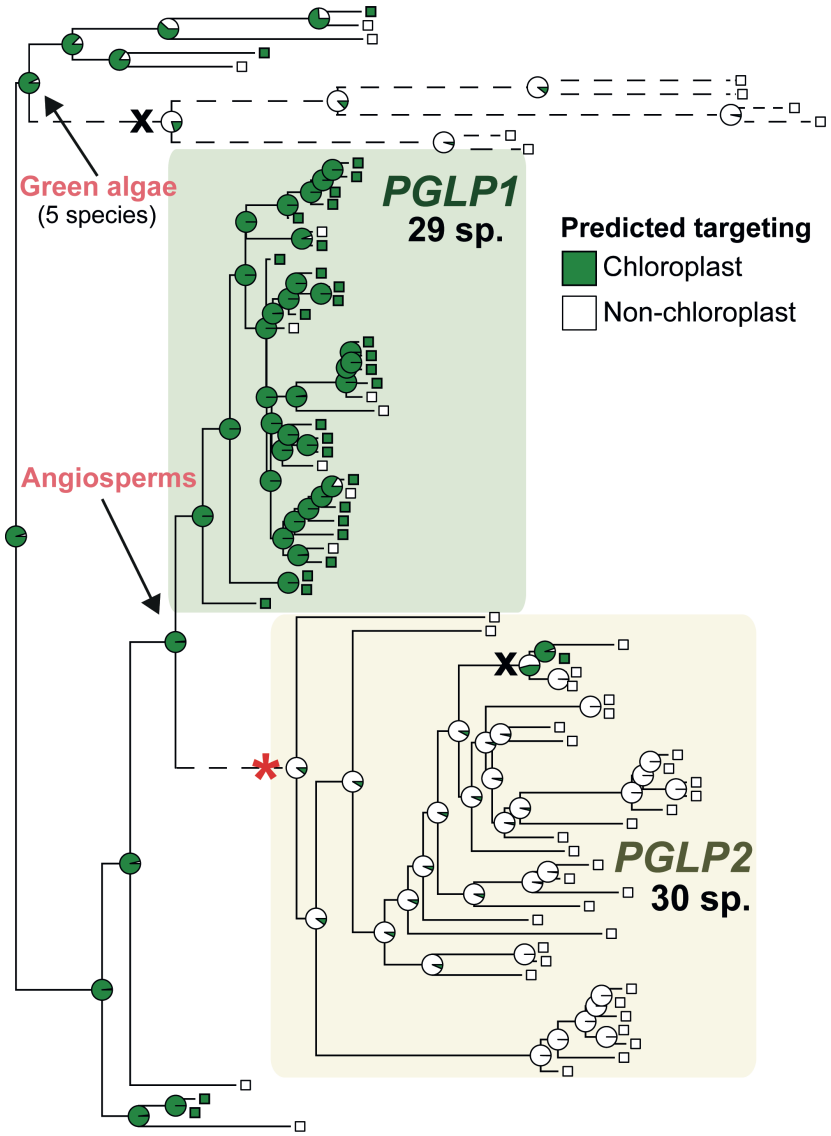


Figure A2.1.1. Gene tree of 2-phosphoglycolate phosphatase (PGLP) showing the results of ancestral state estimation and the branches along which a change in chloroplast targeting has been identified. The method of ancestral state estimation is illustrated as follows: for a given target signal, the presence/absence of that target signal is predicted for extant proteins (boxes at tips of branches), and then ancestral state estimation uses this information to estimate the likelihood that each ancestral protein (represented by internal branches) contains that target signal. The likelihood is

between 0 and 1 and is shown by the pie charts for internal branches of the tree. A winner takes all approach is used to assign an ancestral protein as chloroplast targeted (likelihood score ≥0.5) or non-chloroplast targeted (likelihood score <0.5). Thus, a change in protein targeting is identified when the estimated targeting changes between two consecutive branches in the tree. Such changes are indicated by the red star and black crosses in the gene tree of PGLP. A filtration step is then applied to consider only those changes which are well supported by the extant protein data. Here, ≥75% of the extant proteins which subtend a change in protein targeting must possess the derived targeted state. Similarly, ≥75% of the extant proteins which subtend the sister branch to that on which a change was identified must maintain the ancestral targeted state. With this filter applied, only one change in protein targeting is considered to have occurred in this gene tree of PGLP. This change is a loss of chloroplast targeting in the ancestor of the angiosperm species and is indicated by the red asterisk. The retention score for the clade which had lost chloroplast targeting is 97% (29 out of 30) while the score for the sister clade which had retained the ancestral targeting is 90% (26 out of 29).

This example also demonstrates that after a change in organellar targeting occurs, it is possible that further changes might happen among more recent branches in the tree. It is for this reason that I chose a relaxed retention score of 75%. A score of 100% would have made identification of changes in organellar targeting at deeper branches in orthogroup trees difficult due to the increased chance that there would be further changes in targeting as you move through the tree from the root to the tips. The fact that some proteins may undergo multiple rounds of changes in organellar targeting will inevitably make resolving those changes more difficult. A 75% retention score on the other hand provides a stringency to filter for what appear to be well-supported changes in protein targeting (like that identified for PGLP) while providing some flexibility to minor prediction errors and multiple rounds of change in protein targeting.

In selection of a suitable retention score, the number of identified changes in protein targeting was investigated for a range of possible retention scores. As expected, the number of changes identified for each branch of the species tree increased as the cut-off was relaxed from 100 to 0 (Figure A2.1.2). For the majority of branches in the species tree, there was not a large difference in the number of changes in organellar targeting identified using a retention score of 100% and 75%. However, for the deeper branches in the species tree a relaxation in the retention score did result in more changes in organellar targeting being identified. This is was as expected given the logic set out in the previous paragraph, i.e., that changes among deeper branches are more likely to fail to meet a retention score due to the increased time in which subsequent changes in targeting may occur in the tree. Further relaxation of the cut-off (50% and 25%) resulted in far more changes being identified for all branches (Figure A2.1.2). Thus, a retention score of 75% was selected but it is acknowledged that there is a trade-off here between the confidence of identified changes and the sensitivity of their detection.

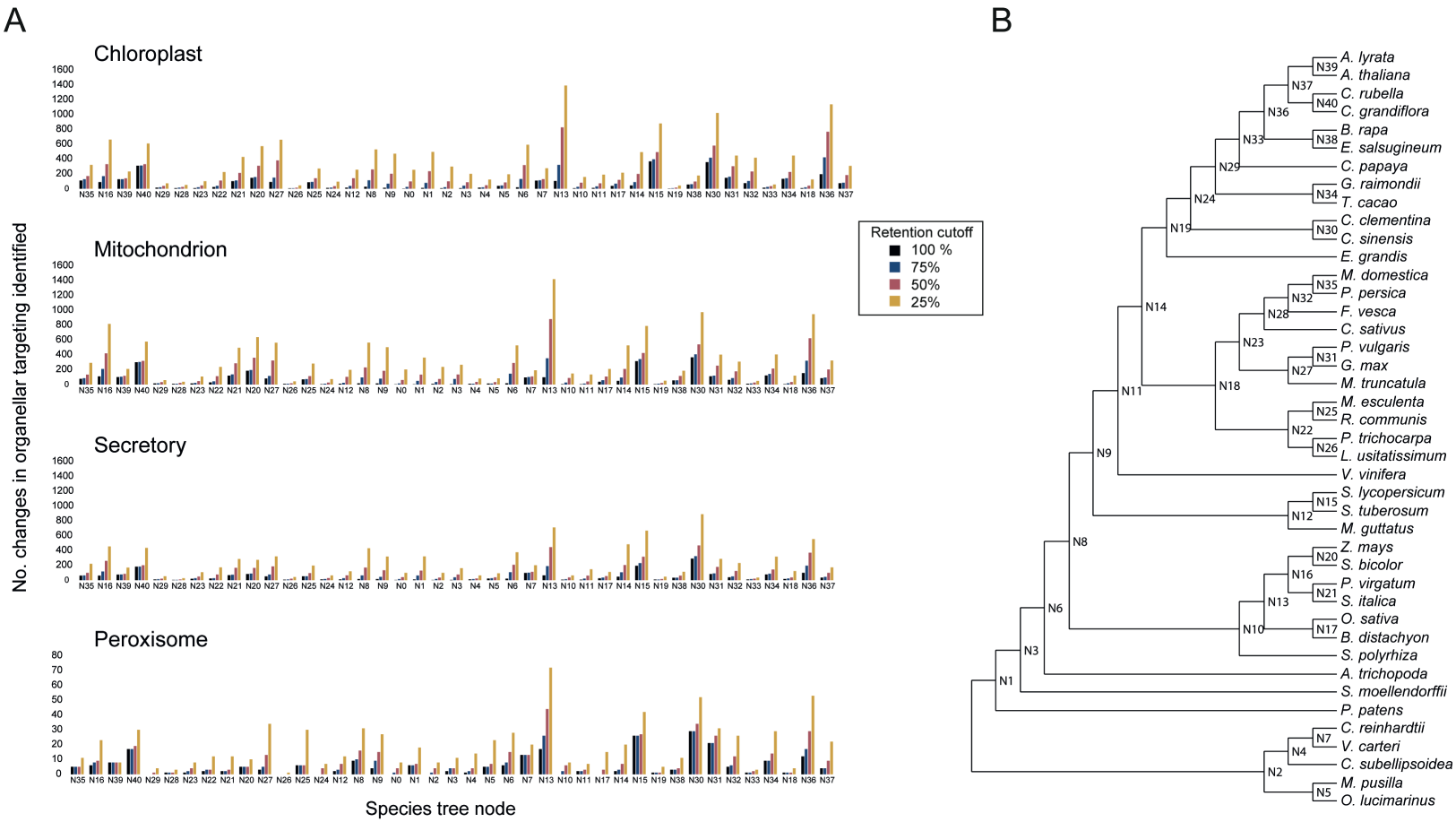


Figure A2.1.2. The number of changes in organellar targeting identified along each branch of the species tree using different retention scores (i.e., the % of extant proteins subtending a change in a gene tree that support the change). A) At its strictest, the retention score was 100%. Relaxation of the retention score to 75%, 50% and 25% resulted in more changes in organellar targeting being identified each time. This is shown as the number of changes in organellar targeting identified for each branch of the species tree using the different retention cut-offs. **B)** The species tree with branch labelled to match those shown in panel A.

Appendix 2.2: Comparison of gene-tree–species-tree mapping approaches

In order to map changes in organellar targeting identified along gene-trees branches to branches in the species tree it was necessary to reconcile each gene-tree branch to the species-tree phylogeny. This mapping was done using a most recent common ancestor approach employed by PHYLDOG (Boussau et al. 2013). In some cases, the absence of a gene from a particular clade in the gene tree can lead to uncertainty in the mapping of gene-tree branches to the species tree using a most recent common ancestor approach. Thus, the number of mapped changes in protein targeting affected by this uncertainty was investigated and the overall pattern of gain and loss across the species tree investigated using an alternative mapping approach.

The number of changes in protein targeting mapped to the species tree using a most recent common ancestor mapping approach is shown in Figure A2.2 A (this is the same result shown in Figure 2.3, Chapter 2). For each branch of the species tree, I identified the number of changes in organellar targeting that were mapped to it but that could have also occurred on the consecutive preceding branch (i.e., ambiguous mapping). The percentage of changes affected in this way for each branch of the species tree in shown in Figure A2.2 B. In total, 21% of gains and 17% of losses occur on gene-tree branches that correspond to multiple branches of the species tree. The results of Figure A2.2 B show that some branches of the species tree are more affected by this uncertainty in mapping than others. For example, less than 50% of the changes mapped to the ancestral branch of *M. domestica*, *P. persica* and *F. vesca* could be unambiguously placed on that branch. This susceptibility is likely caused by the fact that there is only one species (*C. sativus*) in the sister clade to this branch and a single missing gene from this species can result in mapping uncertainty. If there were two species in the sister clade then both species would have to be missing the gene for the same uncertainty to arise (which is more unlikely).

An alternative approach to mapping changes in organellar targeting to the species tree involves splitting changes across the branches of the species tree on which they could have occurred. This time, when a change in targeting occurs on a gene-tree branch that maps to multiple consecutive branches in the species tree, the gain or loss is distributed equally between those branches in the species tree (rather than just the most recent branch). This proportional mapping is demonstrated in Figure 2.2 B of Chapter 2. The placement of gains and losses on the species tree using a proportional mapping approach was investigated and is shown in Figure A2.2 C. Overall more gains and losses were identified using this mapping approach due to a proportion of changes previously mapped to terminal branches now being represented on internal branches of the tree. The number of gains and losses identified for each branch of the species tree by these two mapping approaches (most recent common ancestor and proportional mapping) was found to be very tightly correlated (Figure A2.2 D). I therefore do not think that the choice of mapping approach greatly affects the global picture of gain and loss across the species tree. A most recent common ancestor approach is perhaps more appropriate given that many missing genes represent true gene loss and so mapping a change in the ancestor of that gene is not biologically relevant given that the gene was subsequently lost.

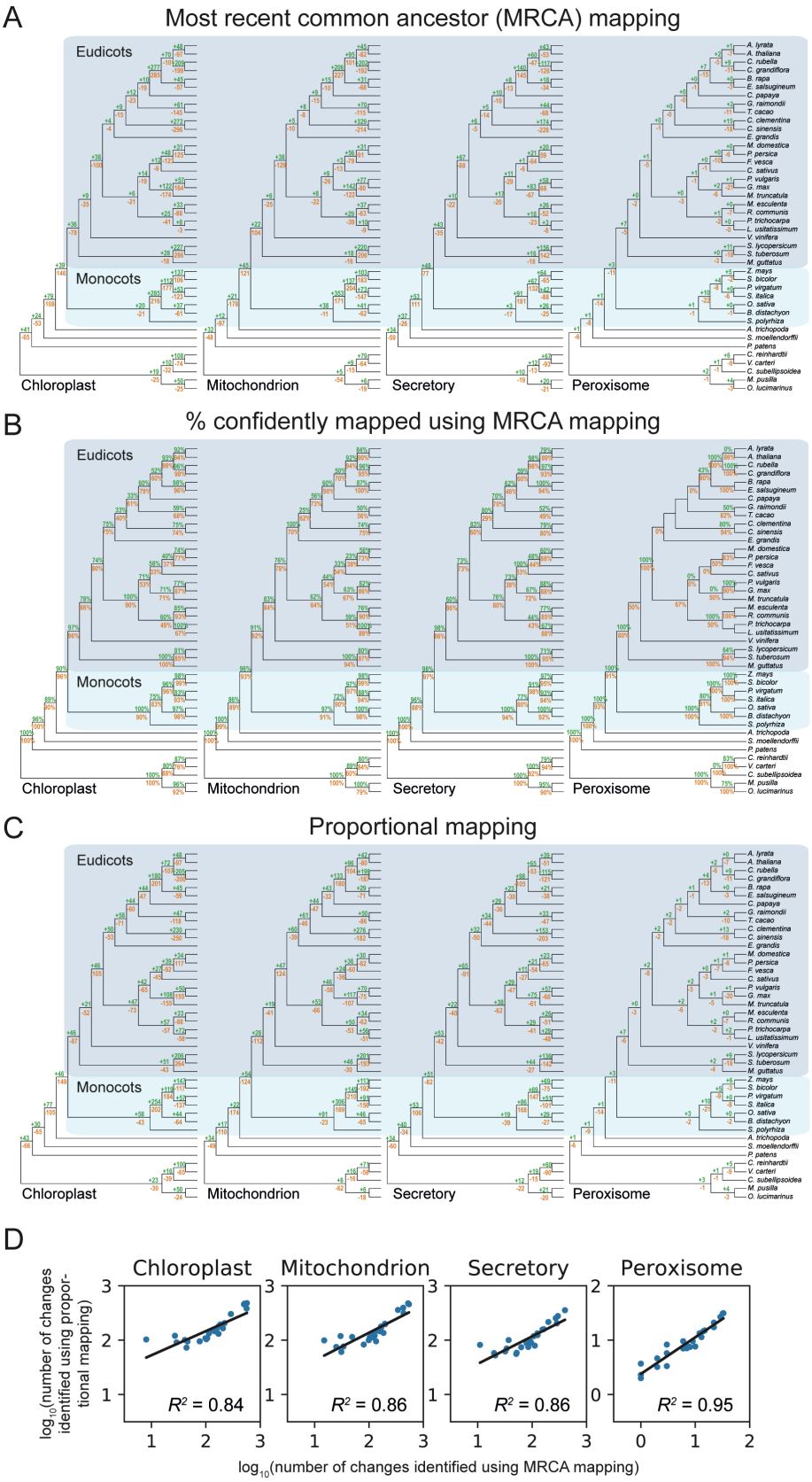


Figure A2.2. Comparison of the number of changes in organellar protein targeting identified along each branch of the species tree using a most recent common

ancestor (MRCA) or proportional mapping approach. A) Changes in protein targeting mapped to the species tree using a MRCA approach. **B)** The percentage of changes mapped using the MRCA approach that can be confidently mapped to a single branch on the species tree. **C)** Changes in protein targeting mapped to the species tree using a proportional mapping approach. **D)** The number of changes identified by both mapping approaches is strongly correlated across the species tree (R^2 values shown, $P < 0.001$).

Appendix 2.3: Functional terms (MapMan) enriched among proteins that underwent a change in organellar targeting during plant evolution.

Significantly enriched functional terms ($P < 0.01$) are shown for proteins gained and lost (re-targeted) to each organelle. In each case, the background population is all proteins with at least one family member targeted to the organelle under consideration. The test population was all proteins which had been gained or lost to that organelle at least once during plant evolution. The *P*-values shown have been corrected for multiple testing using the Benjamini–Hochberg (BH) correction.

A) Among proteins gained/lost to the chloroplast:

MapMan term	MapMan ID	Total occurrences	Sample occurrences	BH corrected p-value
RNA.regulation_of_transcription	27.3	2172	1155	1.63E-08
RNA	27	2656	1384	8.94E-08
hormone_metabolism	17	712	399	7.97E-05
secondary_metabolism	16	427	251	7.97E-05
transport	34	923	505	7.97E-05
stress.biotic.regulation_of_transcription	20.1.5	61	46	0.001098145
DNA.synthesis/chromatin_structure	28.1	437	250	0.001098145
nucleotide_metabolism	23	148	95	0.002422284
major_CHO_metabolism	2	186	115	0.00367503
amino_acid_metabolism.synthesis	13.1	201	122	0.006420741
hormone_metabolism.salicylic_acid	17.8	132	84	0.007337195
hormone_metabolism.salicylic_acid.synthesis-degradation	17.8.1	132	84	0.007337195
protein	29	3175	1576	0.007337195
RNA.regulation_of_transcription.C2H2_zinc_finger_family	27.3.11	77	53	0.007486278
major_CHO_metabolism.degradation	2.2	151	94	0.008268803
misc	26	1196	621	0.008302973

B) Among proteins gained/lost to the mitochondrion:

MapMan term	MapMan ID	Total occurrences	Sample occurrences	BH corrected p-value

RNA	27	2793	1353	0
RNA.regulation_of_transcription	27.3	2278	1129	1.00E-40
transport	34	980	520	4.78E-10
RNA.regulation_of_transcription.MYB_domain_transcription_factor_family	27.3.25	272	168	2.14E-08
protein	29	3427	1609	2.61E-08
hormone_metabolism	17	750	401	5.39E-08
hormone_metabolism.salicylic_acid	17.8	138	94	1.41E-07
hormone_metabolism.salicylic_acid.synthesis-degradation	17.8.1	138	94	1.41E-07
transport.misc	34.99	288	173	1.41E-07
misc.UDP_glucosyl_and_glucoronyl_transferases	26.2	251	153	2.82E-07
protein.synthesis.ribosomal_protein	29.2.1	287	170	7.06E-07
cell_wall	10	431	237	8.02E-06
protein.synthesis	29.2	748	386	1.33E-05
stress	20	936	473	1.33E-05
secondary_metabolism	16	438	239	1.33E-05
development	33	800	409	1.78E-05
misc	26	1282	627	4.03E-05
protein.postranslational_modification	29.4	956	477	6.48E-05
protein.synthesis.ribosome_biogenesis	29.2.2	453	241	0.00013424
stress.abiotic	20.2	435	232	0.000156376
development.unspecified	33.99	673	341	0.000422186
mitochondrial_electron_transport_/_ATP_synthesis	9	180	104	0.001458294
DNA.synthesis/chromatin_structure	28.1	449	232	0.002429025
protein.synthesis.ribosomal_protein.prokaryotic	29.2.1.1	143	84	0.003564603
protein.synthesis.ribosomal_protein.prokaryotic.unknown_organellar.50S_subunit	29.2.1.1.3.2	31	24	0.004576258
protein.synthesis.ribosomal_protein.eukaryotic.60S_subunit	29.2.1.2.2	95	59	0.004728449
RNA.processing	27.1	435	223	0.005471282
hormone_metabolism.brassinosteroid.synthesis-degradation	17.3.1	119	71	0.005651233
cell	31	647	320	0.006656735
lipid_metabolism	11	414	212	0.007716035
cell_wall.degradation	10.6	85	53	0.007716035
mitochondrial_electron_transport_/_ATP_synthesis.cytochrome_c_reductase	9.5	10	10	0.008530616
micro_RNA,_natural_antisense_etc	32	145	83	0.009293591
secondary_metabolism.phenylpropanoids	16.2	119	70	0.009293591
protein.synthesis.ribosomal_protein.eukaryotic	29.2.1.2	137	79	0.009293591

C) Among proteins gained/lost to the secretory pathway:

MapMan term	MapMan ID	Total occurrences	Sample occurrences	BH corrected p-value
protein.postranslational_modification.kinase.receptor_like_cytoplasmatic_kinase_VII	29.4.1.57	254	160	0

protein.postranslational_modification	29.4	869	393	0
signalling	30	1169	513	0
misc	26	1242	594	6.28E-11
cell_wall	10	441	213	1.01E-10
protein	29	3068	1163	1.01E-09
signalling.receptor_kinases.leucine_rich_repeat_XI	30.2.11	195	111	1.01E-09
transport	34	906	404	1.01E-09
protein.postranslational_modification.kinase	29.4.1	307	182	1.29E-09
signalling.receptor_kinases	30.2	645	325	1.66E-09
signalling.receptor_kinases.DUF_26	30.2.17	201	114	2.22E-09
misc.myrosinases-lectin-jacalin	26.16	95	63	2.22E-09
hormone_metabolism	17	657	291	1.92E-08
stress	20	966	406	2.13E-08
misc.UDP_glucosyl_and_glucoronyl_transferases	26.2	220	117	2.42E-08
hormone_metabolism.brassinosteroid	17.3	166	93	4.37E-08
signalling.receptor_kinases.S-locus_glycoprotein_like	30.2.24	110	67	1.02E-07
cell_wall.hemicellulose_synthesis	10.3	32	26	1.82E-06
stress.biotic.receptors	20.1.2	222	112	2.36E-06
development	33	712	300	3.43E-06
secondary_metabolism	16	398	181	4.07E-06
signalling.receptor_kinases.leucine_rich_repeat_III	30.2.3	61	40	9.97E-06
stress.biotic	20.1	599	255	1.16E-05
hormone_metabolism.cytokinin	17.4	54	36	2.06E-05
protein.degradation	29.5	1304	506	4.13E-05
transport.peptides_and_oligopeptides	34.13	138	73	4.27E-05
lipid_metabolism	11	381	169	6.29E-05
redox	21	221	106	8.79E-05
development.unspecified	33.99	588	245	0.000121411
signalling.receptor_kinases.leucine_rich_repeat_XII	30.2.12	82	47	0.000193159
cell_wall.hemicellulose_synthesis.glucuronoxylan	10.3.2	11	11	0.000203754
mitochondrial_electron_transport_/_ATP_synthesis.cytochrome_c_reductase	9.5	11	11	0.000203754
cell_wall.degradation	10.6	85	48	0.000241334
misc.oxidases_-_copper,_flavone_etc	26.7	139	71	0.000241334
hormone_metabolism.cytokinin.signal_transduction	17.4.2	23	18	0.000399957
misc.cytochrome_P450	26.1	147	73	0.000610437
misc.short_chain_dehydrogenase/reductase_(SDR)	26.22	67	39	0.00061864
protein.degradation.ubiquitin	29.5.11	904	353	0.000873716
signalling.receptor_kinases.leucine_rich_repeat_VII	30.2.7	34	23	0.001189587
mitochondrial_electron_transport_/_ATP_synthesis	9	166	79	0.001732542
stress.abiotic	20.2	381	160	0.003004169
hormone_metabolism.brassinosteroid.induced-regulated-responsive-activated	17.3.3	16	13	0.003140261
micro_RNA,_natural_antisense_etc	32	136	66	0.003140261
protein.degradation.ubiquitin.E3.RING	29.5.11.4.2	466	191	0.003152533
hormone_metabolism.brassinosteroid.signal_transduction	17.3.2	42	26	0.003152533
misc.beta_1,3_glucan_hydrolases	26.4	32	21	0.00434069

signalling.receptor_kinases.leucine_rich_repeat_X	30.2.10	24	17	0.004451016
hormone_metabolism.brassinosteroid.synthesis-degradation	17.3.1	108	54	0.004451016
signalling.receptor_kinases.misc	30.2.99	75	40	0.005310953
protein.degradation.aspartate_protease	29.5.4	52	30	0.005310953
misc.GDSL-motif_lipase	26.28	55	31	0.007300044
redox.glutaredoxins	21.4	31	20	0.007796522
protein.degradation.ubiquitin.E3	29.5.11.4	725	281	0.007822404
hormone_metabolism.brassinosteroid.signal_transduction.BRI	17.3.2.1	29	19	0.007822404
signalling.calcium	30.3	171	78	0.007822404
amino_acid_metabolism.degradation.glutamate_family	13.2.2	21	15	0.008166909

D) Among proteins gained/lost to the peroxisome:

MapMan term	MapMan ID	Total occurrences	Sample occurrences	BH corrected p-value
amino_acid_metabolism.degradation	13.2	30	21	0.000406283
lipid_metabolism.lipid_degradation.beta-oxidation	11.9.4	20	15	0.002514571
secondary_metabolism.phenylpropanoids.lignin_biosynthesis.4CL	16.2.1.3	11	10	0.003207324
secondary_metabolism.phenylpropanoids.lignin_biosynthesis	16.2.1	28	18	0.004421625
lipid_metabolism	11	96	43	0.005943128
amino_acid_metabolism.degradation.aspartate_family	13.2.3	16	12	0.007802286

Appendix 2.4: Localisation terms (from the Plant Proteome Database) enriched among proteins that underwent a change in organellar targeting during plant evolution

Table A2.4. The total number of proteins (total occurrences) and the number of proteins which have undergone a change in protein targeting (sample occurrences) which are annotated to localise to each of the sub-chloroplast compartments given in the Plant Proteome Database (PPDB). A hypergeometric test was conducted to test for the enrichment of each term among proteins which have undergone a change in organellar protein targeting during plant evolution (*N* = 1,656) compared to the population of all proteins (*N* = 2,466). Note that only those proteins with an annotated *Arabidopsis* member in the same protein family could be used in this analysis. Data for experimentally characterised proteomes taken from the PPDB taken on 13th March 2018.

PPDB compartment	Total occurrences	Sample occurrences	Raw *P*-value	BH corrected *P*-value
plastid-unspecified	600	437	5.74E-05	0.001319451
plastid-nucleoid	117	92	0.003142117	0.036134342
plastid-stroma	506	360	0.009872957	0.075692668
plastoglobules	28	23	0.061702006	0.354786532
envelope-inner	99	73	0.089502043	0.366792156
envelope-inner-peripheral	75	56	0.09568491	0.366792156
dual-targeted-mitochondria	63	45	0.277258883	0.884854908
dual-targeted	81	57	0.30777562	0.884854908
envelope-inner-membrane-space	2	2	0.450890643	0.99924049
dual-targeted-plasma-membrane	2	2	0.450890643	0.99924049
envelope	164	111	0.477897626	0.99924049
dual-targeted-nucleus	7	5	0.582036129	1
dual-targeted-cytosol	9	6	0.662709099	1
plastid-ribosome	63	41	0.694031284	1
thylakoid-peripheral-lumenal-side	53	34	0.736844184	1
envelope-outer	30	19	0.745153093	1
thylakoid-peripheral	102	62	0.9357873	1
thylakoid-peripheral-stromal-side	49	28	0.95065191	1
envelope-inner-peripheral-stromal-side	16	8	0.955350108	1
dual-targeted-peroxisome	5	2	0.957406613	1
envelope-outer-integral	4	1	0.988469554	1

thylakoid	283	151	0.99999998	1
thylakoid-integral	108	41	1	1

Appendix 2.5: Functional terms (MapMan) enriched among proteins that did not undergo a change in protein targeting to A) the chloroplast or B) the mitochondrion during plant evolution.

Here, the sample population (sample occurrences) was all protein families with no history of changes in protein targeting to the chloroplast (A) or mitochondrion (B) during plant evolution. The background population was all protein families from the sample population *and* all protein families with a well-supported change in protein targeting (≥75% retention score) during plant evolution. Significantly enriched functional terms are shown ($P < 0.01$) where the functional term was assigned to at least 10 protein families in the background population.

A)

MapMan term	MapMan ID	Total occurrences	Sample occurrences	BH corrected p-value
amino_acid_metabolism.synthesis	13.1	214	49	5.81E-14
nucleotide_metabolism	23	161	35	1.79E-12
major_CHO_metabolism	2	201	48	1.01E-11
not_assigned.no_ontology_assigned	35.2	843	261	6.37E-11
major_CHO_metabolism.degradation	2.2	160	37	1.27E-10
amino_acid_metabolism	13	322	89	5.41E-10
hormone_metabolism.salicylic_acid	17.8	143	34	5.52E-09
nucleotide_metabolism.salvage	23.3	31	3	5.52E-09
DNA.synthesis/chromatin_structure	28.1	491	147	6.12E-09
nucleotide_metabolism.synthesis	23.1	34	4	2.02E-08
minor_CHO_metabolism.myo-inositol	3.4	41	6	6.11E-08
secondary_metabolism	16	493	155	4.80E-06
hormone_metabolism.abscisic_acid	17.1	51	11	2.33E-05
DNA	28	721	236	2.43E-05
glycolysis.plastid_branch	4.2	15	2	5.75E-05
protein.assembly_and_cofactor_ligation	29.8	35	7	6.99E-05
protein.aa_activation	29.1	95	25	7.48E-05
misc.UDP_glucosyl_and_glucoronyl_transferases	26.2	269	83	0.000169948
PS	1	386	123	0.000181871
TCA_/_organic_transformation	8	66	17	0.000381409

TCA_/_organic_transformation.other_organic_acid_transformations	8.2	17	3	0.000432337
transport.calcium	34.21	17	3	0.000432337
fermentation	5	21	4	0.000436099
secondary_metabolism.isoprenoids	16.1	170	51	0.000534991
signalling.in_sugar_and_nutrient_physiology	30.1	48	12	0.000861151
N-metabolism.misc	12.4	20	4	0.000881243
misc.rhodanese	26.23	16	3	0.000886967
transport.ABC_transporters_and_multidrug_resistance_systems	34.16	86	24	0.000976721
major_CHO_metabolism.synthesis	2.1	44	11	0.001194185
lipid_metabolism.FA_synthesis_and_FA_elongation	11.1	185	57	0.001319678
polyamine_metabolism.synthesis	22.1	19	4	0.001717874
Co-factor_and_vitamine_metabolism	18	78	22	0.002046011
PS.lightreaction	1.1	282	91	0.002725816
TCA_/_organic_transformation.TCA	8.1	52	14	0.002898916
cell_wall.precursor_synthesis	10.1	185	58	0.00315903
misc.acyl_transferases	26.5	10	2	0.003184274
secondary_metabolism.phenylpropanoids	16.2	136	42	0.005286871
misc.cytochrome_P450	26.1	159	50	0.006488283
Co-factor_and_vitamine_metabolism.pantothenate	18.4	13	3	0.007964982
polyamine_metabolism	22	20	5	0.009758717

B)

MapMan term	**MapMan ID**	**Total occurrences**	**Sample occurrences**	**BH corrected p-value**
hormone_metabolism.salicylic_acid	17.8	143	21	1.27E-20
misc.UDP_glucosyl_and_glucoronyl_transferases	26.2	269	61	1.62E-16
transport.misc	34.99	312	83	1.14E-10
not_assigned.no_ontology_assigned	35.2	843	263	7.86E-09
nucleotide_metabolism.degradation	23.2	42	6	6.67E-08
protein.synthesis	29.2	802	254	3.14E-07
mitochondrial_electron_transport_/_ATP_synthesis	9	190	51	8.01E-07
N-metabolism.misc	12.4	20	2	1.33E-06
cell_wall.degradation	10.6	91	21	1.80E-06
cell_wall.hemicellulose_synthesis	10.3	37	6	2.27E-06
RNA.processing	27.1	473	147	1.11E-05
nucleotide_metabolism	23	161	44	1.36E-05
amino_acid_metabolism	13	322	97	1.95E-05
PS.photorespiration	1.2	37	7	3.08E-05
mitochondrial_electron_transport_/_ATP_synthesis.F1-ATPase	9.9	33	6	3.48E-05
RNA.silencing	27.5	39	8	9.04E-05

mitochondrial_electron_transport_/_ATP_synthesis.NADH-DH_(type_I)	9.1	46	10	9.69E-05
N-metabolism	12	48	11	0.000230815
amino_acid_metabolism.synthesis	13.1	214	64	0.000281243
misc.short_chain_dehydrogenase/reductase_(SDR)	26.22	77	20	0.000364943
transport.calcium	34.21	17	3	0.000581921
TCA_/_organic_transformation.other_organic_acid_transformations	8.2	17	3	0.000581921
mitochondrial_electron_transport_/_ATP_synthesis.cytochrome_c	9.6	17	3	0.000581921
TCA_/_organic_transformation	8	66	17	0.000620332
tetrapyrrole_synthesis.ALA_dehydratase	19.4	66	17	0.000620332
protein.glycosylation	29.7	49	12	0.000784406
DNA.unspecified	28.99	156	46	0.00079269
mitochondrial_electron_transport_/_ATP_synthesis.cytochrome_c_reductase	9.5	12	2	0.000811929
TCA_/_organic_transformation.TCA	8.1	52	13	0.000898619
transport	34	1093	371	0.001611502
transport.ABC_transporters_and_multidrug_resistance_systems	34.16	86	24	0.001678046
misc.acid_and_other_phosphatases	26.13	54	14	0.001678046
amino_acid_metabolism.degradation	13.2	117	34	0.001678046
tetrapyrrole_synthesis	19	151	46	0.003936504
secondary_metabolism.simple_phenols	16.1	10	2	0.003994452
transport.metabolite_transporters_at_the_mitochondrial_membrane	34.9	71	20	0.004534669
misc.cytochrome_P450	26.1	159	49	0.005183502
redox.glutaredoxins	21.4	35	9	0.005700702
signalling.phosphinositides	30.4	35	9	0.005700702
protein.assembly_and_cofactor_ligation	29.8	35	9	0.005700702
misc.gluco-,_galacto-_and_mannosidases	26.3	60	17	0.008374695
tetrapyrrole_synthesis.glu-tRNA_synthetase	19.1	9	2	0.008374695
development.squamosa_promoter_binding_like_(SPL)	33.3	9	2	0.008374695
minor_CHO_metabolism.galactose	3.8	13	3	0.009293374
major_CHO_metabolism	2	201	64	0.009539029

Appendix 3.1: Breakdown of the number of gains and losses in protein targeting identified to each organelle along branches of the species tree associated with the evolution of C_4 photosynthesis

Table A3.1. The number of changes in protein targeting identified along branches of the species tree associated with the evolution of C_4 photosynthesis. Column A shows the total number of changes in protein targeting to each organelle. Column B shows the number of those changes which occurred in the ancestor of at least one gene that is expressed in the leaf (total leaf TPM > 10) and is differentially expressed between the mesophyll (M) and bundle sheath (BS) cells ($P < 0.05$).

Organelle	**A.** The total number of changes (gains/losses) in protein targeting identified	**B.** The number of changes (gains/losses) that occurred in the ancestor of a leaf-expressed gene with differential M/BS expression
	Andropogoneae C_4 origin (N17)	
Chloroplast	65 (19/46)	37 (13/24)
Mitochondrion	35 (18/17)	14 (8/6)
Peroxisome	3 (2/1)	0
Secretory Pathway	42 (15/27)	11 (5/6)
	Paniceae C_4 origin (N18)	
Chloroplast	30 (7/23)	9 (2/7)
Mitochondrion	29 (4/25)	6 (1/5)
Peroxisome	6 (0/6)	0
Secretory Pathway	36 (10/26)	2 (0/2)
	Ancestral to both Andropogoneae and Paniceae C_4 origins (N14)	
Chloroplast	78 (29/49)	46 (19/27)
Mitochondrion	41 (13/28)	14 (6/8)
Peroxisome	10 (2/8)	5 (1/4)
Secretory Pathway	82 (37/45)	15 (4/11)

Appendix 3.2: Details about the transcriptomic data sets used in Chapter 3 and Chapter 4 of this thesis

Table A3.2. Details of transcriptome data sets used in Chapter 3 and Chapter 4. In most cases the transcript data were processed as described in the Materials and Methods section of Chapter 3. For data sets where significant post-processing was required, for example cross-species mapping of reads, then read data were taken directly from the supplemental data in the study reference.

Study reference	SRA accession	Species, tissue/cell type	Sample collection details	Reads processed in this study?	Phytozome13 file used to map reads
(Chang et al. 2012)	SRA047278	*Zea mays*, mesophyll and bundle sheath cells	Enzymatic and mechanical separation of mesophyll cells and bundle sheath strands, respectively	Yes	Zmays_493_RefGen_V4.transcript.fa.gz
(Tausta et al. 2014)	SRP035577	*Zea mays*, mesophyll and bundle sheath cells	Laser microdissection of mesophyll and bundle sheath cells at three positions along the developmental gradient of a mature leaf	Yes	Zmays_493_RefGen_V4.transcript.fa.gz
(Wang, Kelly, et al. 2013)	SRP028231	*Zea mays*, leaf foliar blade and husk leaf sheath	RNA isolated for each leaf type at 5 developmental stages (from primordia to fully expanded leaf)	Yes	Zmays_493_RefGen_V4.transcript.fa.gz
(Emms et al. 2016)	ERP013053	*Sorghum bicolor*, mesophyll and bundle sheath cells	Leaf rolling and mechanical separation of mesophyll and bundle sheath cells, respectively	Yes	Sbicolor_454_v3.1.1.transcript.fa.gz
(John et al. 2014)	ERP004434	*Setaria viridis*, mesophyll and bundle sheath cells	Leaf rolling and mechanical separation of mesophyll and bundle sheath cells, respectively	Yes	Sitalica_312_v2.2.transcript.fa.gz
(Rao et al. 2016)	SAMN04002964	*Panicum virgatum*, mesophyll and bundle sheath cells	Manual dissection of mesophyll and bundle sheath cells using a scalpel blade under a fluorescence microscope	Yes	Pvirgatum_516_v5.1.transcript.fa.gz
(van Campen et al. 2016)	SRP062323	*Oryza sativa*, leaf tissue	RNA isolated at different leaf developmental stages (from primordia to the mature leaf blade)	Yes	Osativa_323_v7.0.transcript.fa.gz
(Woo et al. 2016)	SRP018034	*Arabidopsis thaliana*, 4th rosette leaf	RNA isolated from the fourth rosette leaf at different plant ages	Yes	Athaliana_167_TAIR10.annotation_info.txt
(Gowik et al. 2011)	-	*Flaveria* species, leaf	RNA isolated from the second and fourth visible leaves at the top of plants	No	-
(Kelly et al. 2017)	-	30 C_4 and 17 C_3 eudicot species	RNA isolated from leaves of all species	No	-
(Aubry et al. 2014)	-	*Arabidopsis thaliana* whole-leaf and bundle sheath cell tissue	mRNA isolated from whole-leaf and bundle sheath cell specific ribosomes. Mature *Arabidopsis* leaf used.	No	-

Appendix 3.3: Gene trees and expression data supplementary to the main text in Chapter 3

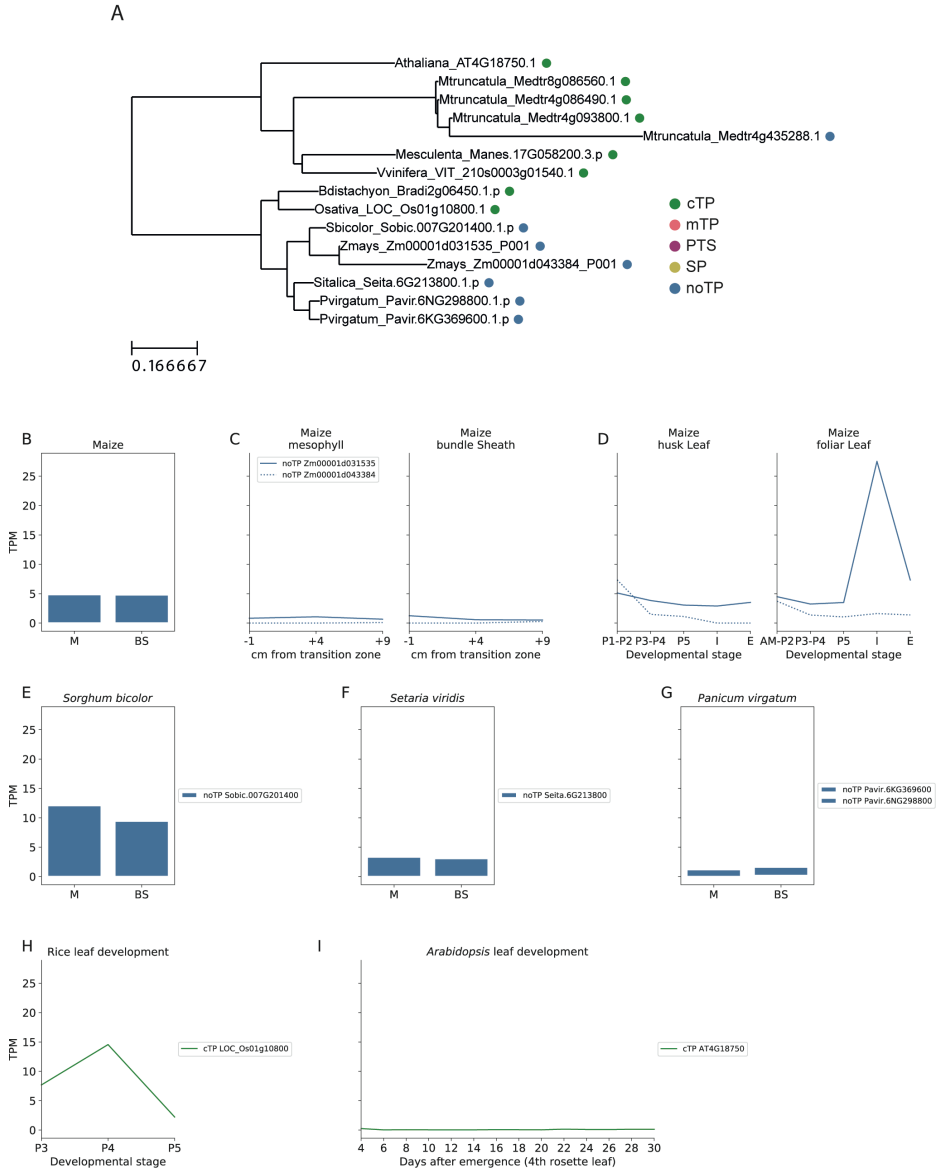


Figure A3.3.1. Phylogenetic gene tree and transcript expression data for the orthogroup containing the *Arabidopsis* gene *DOT4*. A) Phylogeny showing all members of the orthogroup containing DOT4. **B-I)** Expression in selected transcriptomic data sets: **B** (Chang et al. 2012), **C** (Tausta et al. 2014), **D** (Wang, Kelly, et al. 2013), **E** (Emms et al. 2016), **F** (John et al. 2014), **G** 2016), **H** (van Campen et al. 2016), **I** (Woo et al. 2016). **Abbreviations**: mesophyll (M), bundle sheath (BS), chloroplast transit peptide (cTP), mitochondrial target peptide (mTP), secretory peptide (SP), peroxisomal target signal (PTS), no N-terminal target signal identified (noTP), transcripts per million (TPM).

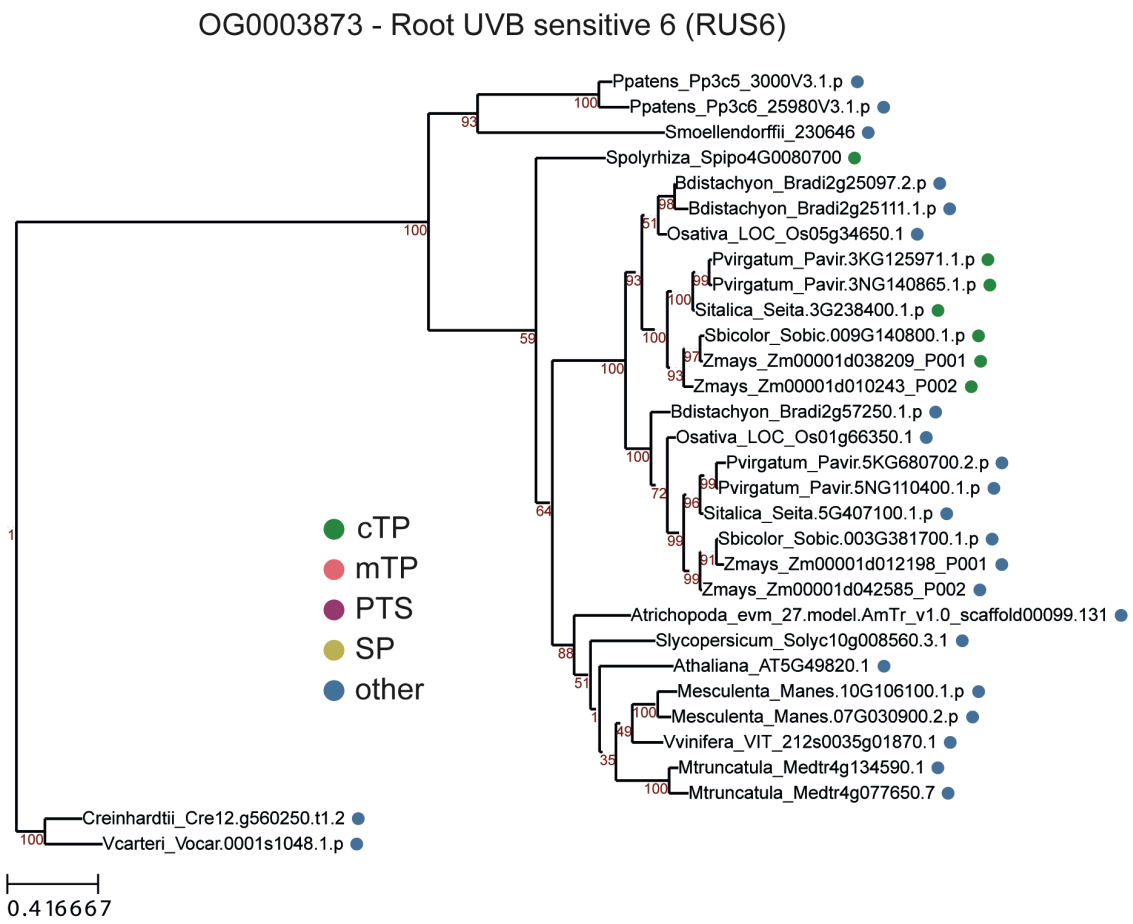


Figure A3.3.2. Phylogenetic gene tree for all members of the orthogroup containing Root UVB sensitive 6 (RUS6). Abbreviations: chloroplast transit peptide (cTP), mitochondrial target peptide (mTP), secretory peptide (SP), peroxisomal target signal (PTS), no N-terminal target signal identified (noTP), transcripts per million (TPM).

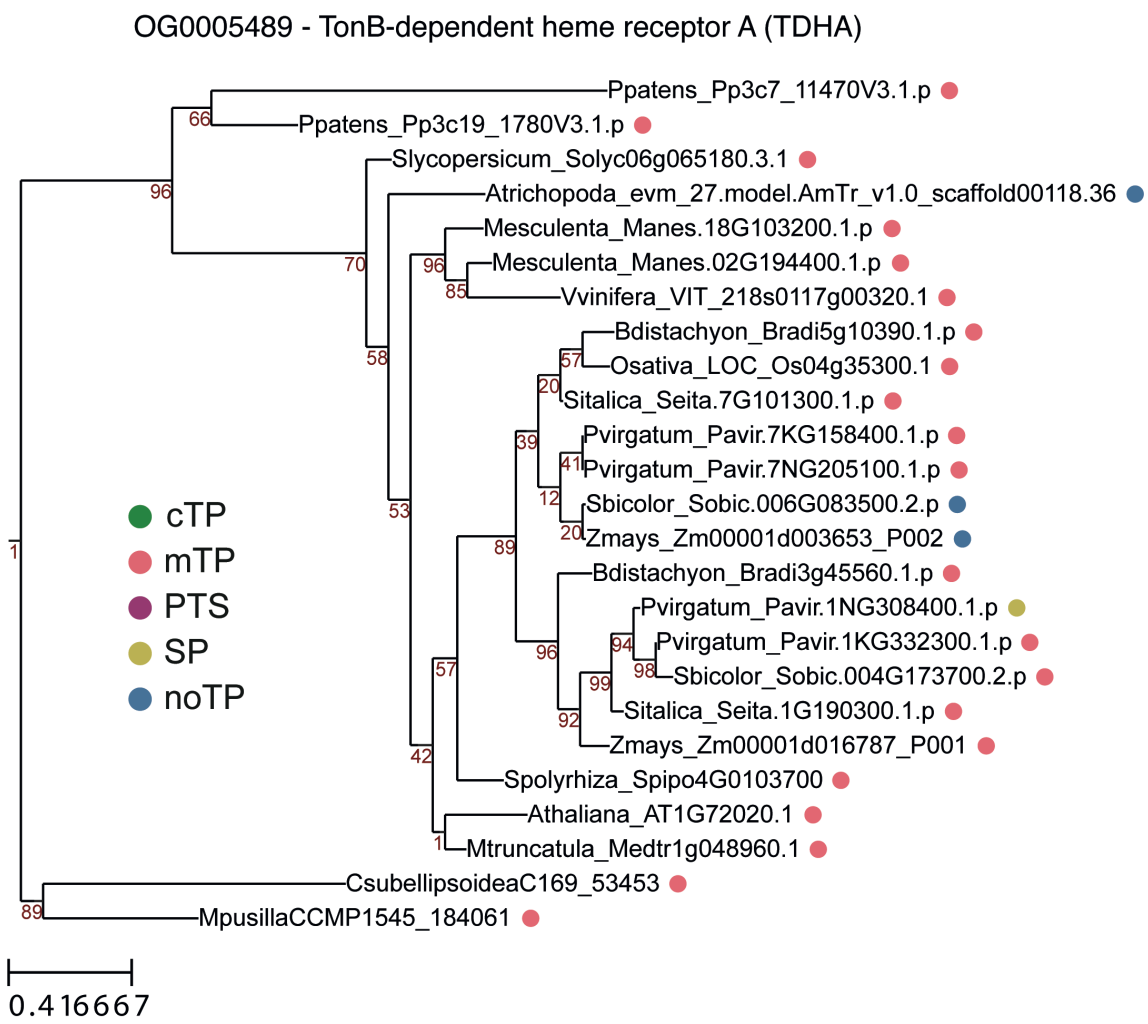


Figure A3.3.3. Phylogenetic gene tree for all members of the orthogroup containing TonB-dependent heme receptor A (TDHA). Abbreviations: chloroplast transit peptide (cTP), mitochondrial target peptide (mTP), secretory peptide (SP), peroxisomal target signal (PTS), no N-terminal target signal identified (noTP), transcripts per million (TPM).

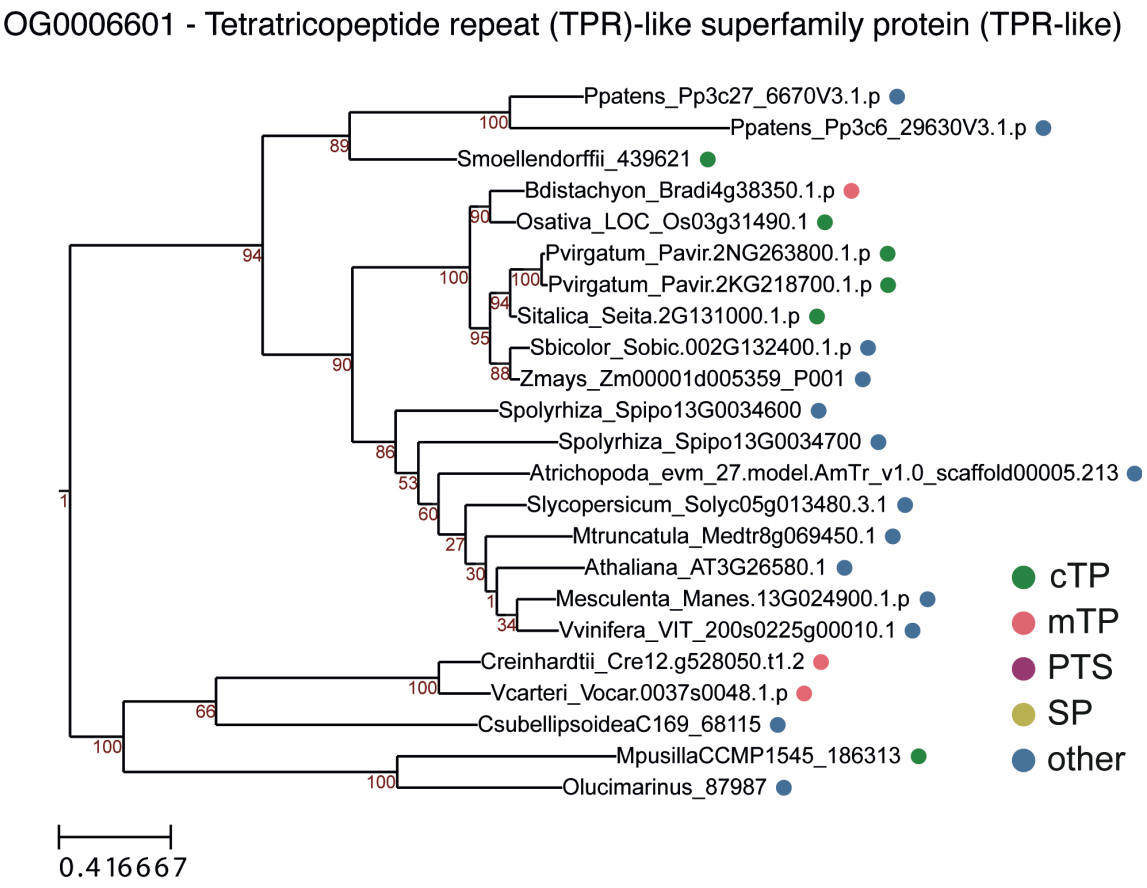


Figure A3.3.4. Phylogenetic gene tree for all members of the orthogroup containing Tetratricopeptide repeat (TPR)-like superfamily protein. Abbreviations: chloroplast transit peptide (cTP), mitochondrial target peptide (mTP), secretory peptide (SP), peroxisomal target signal (PTS), no N-terminal target signal identified (noTP), transcripts per million (TPM).

Appendix 4.1: The set of species (and their phylogenetic relationship) used in the construction of the AILP1 orthogroup and gene tree

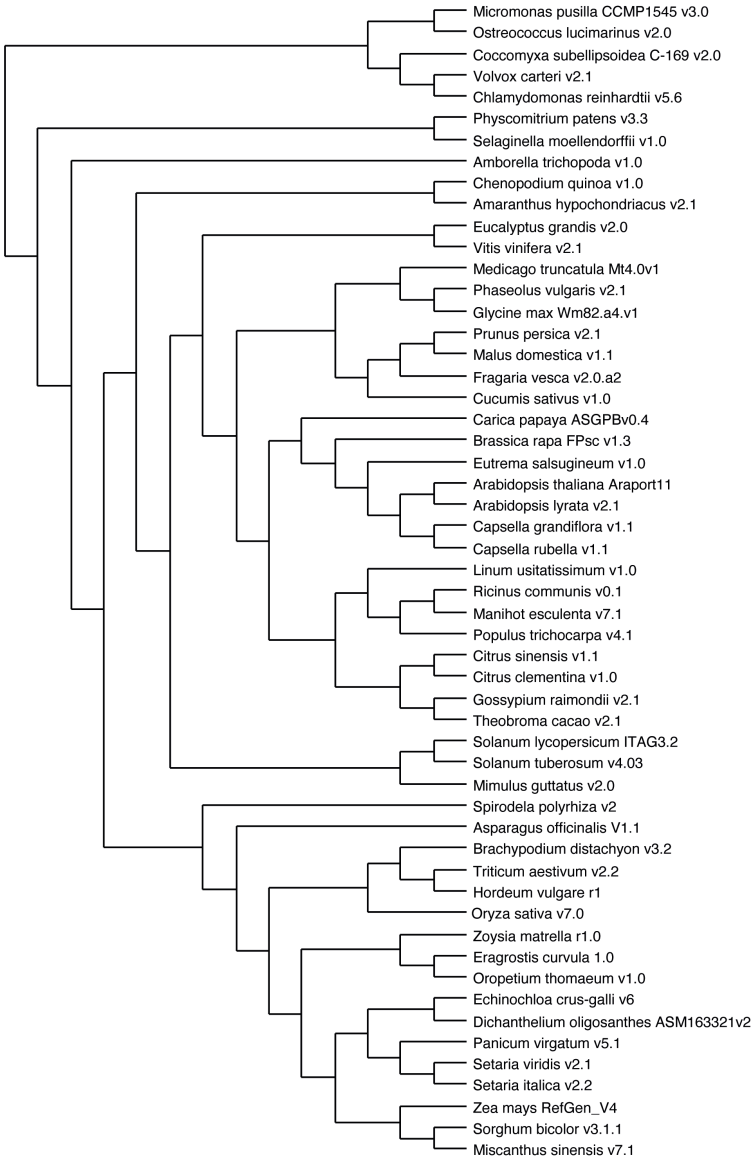


Figure A4.1. Species tree constructed by OrthoFinder for the 54 plant proteomes used to construct orthogroups in Chapter 4. The majority of proteome data sets were obtained from Phytozome v13 and the proteome version is shown after the species name. For those proteome data sets not acquired from Phytozome, the species name and locations of the data sets are as follows: *Zoysia matrella* (Zoysia Genome Database, http://zoysia.kazusa.or.jp), *Echinochloa crus-galli* (RiceWeedomes, http://ibi.zju.edu.cn/RiceWeedomes/Echinochloa/), *Eragrostis curvula* (NCBI Genome 83361), Dichanthelium oligosanthes (NCBI Genome 44711).

Appendix 4.2: Binding and cleavage sites of the sgRNAs used in the editing of the *Arabidopsis AILP1* gene

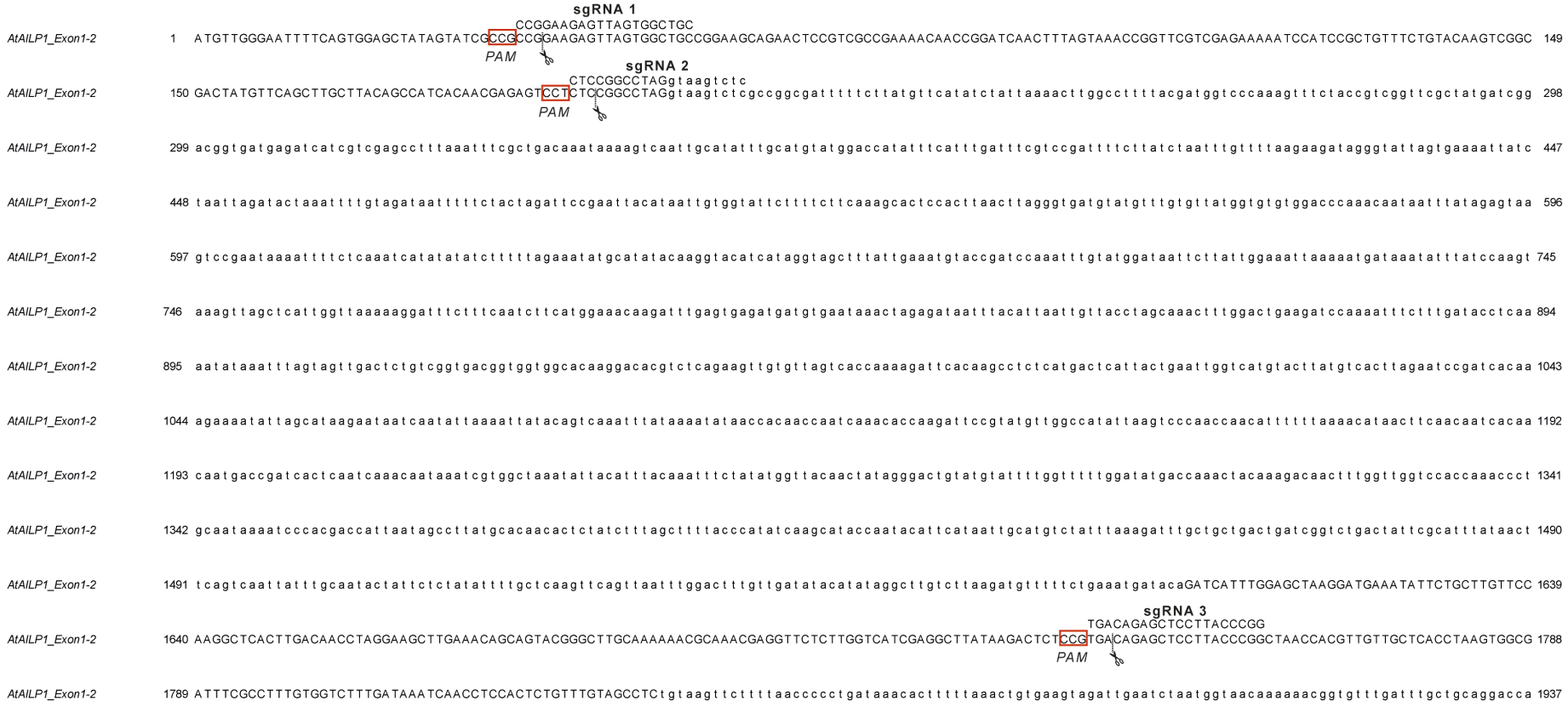


Figure A4.2. Gene sequence of the N-terminal region from *AtAILP1* (AT5G19140.1) showing the short guide RNA (sgRNA) sites used to direct the Cas9 endonuclease. Nucleotides in capital and lower case denote nucleotides from exons and introns, respectively. The protospacer adjacent motif (PAM) is highlighted by the red box and the expected Cas9 cut site is shown by the dashed line and scissor icon.

Appendix 4.3: Gel electrophoresis from screening T2 CRISPR edited plants

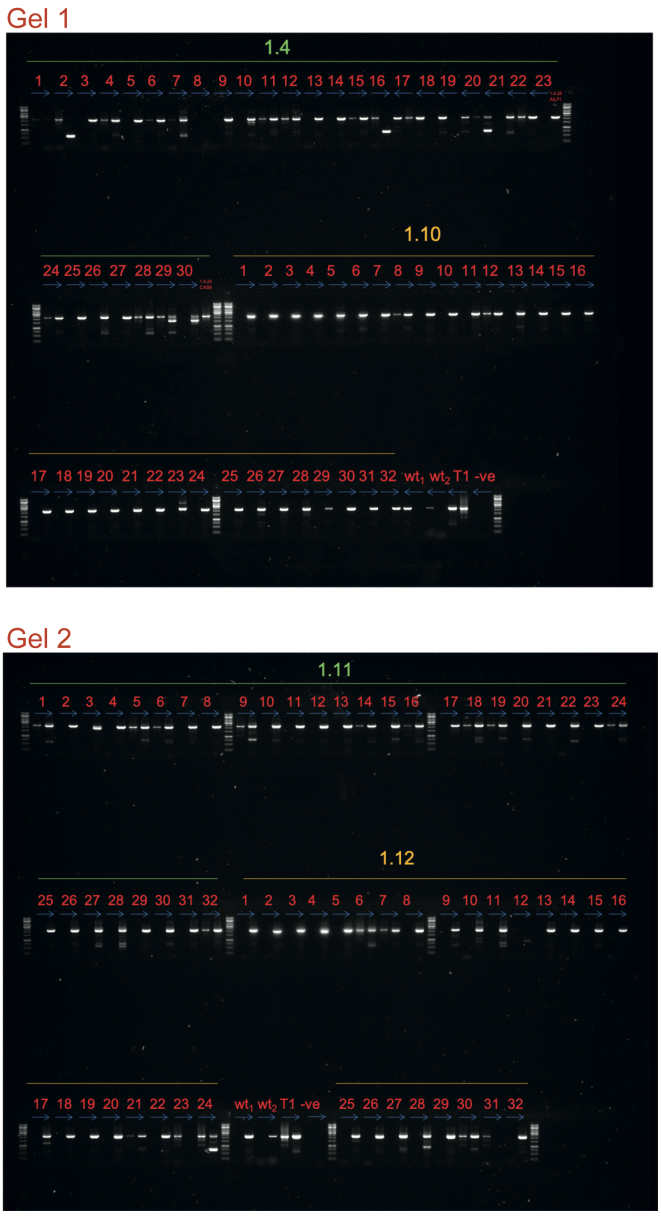
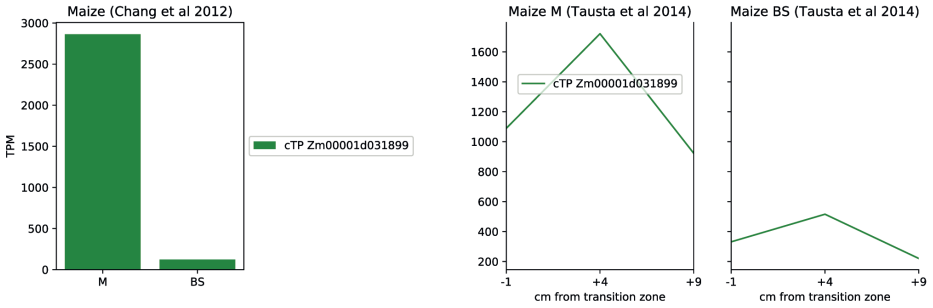


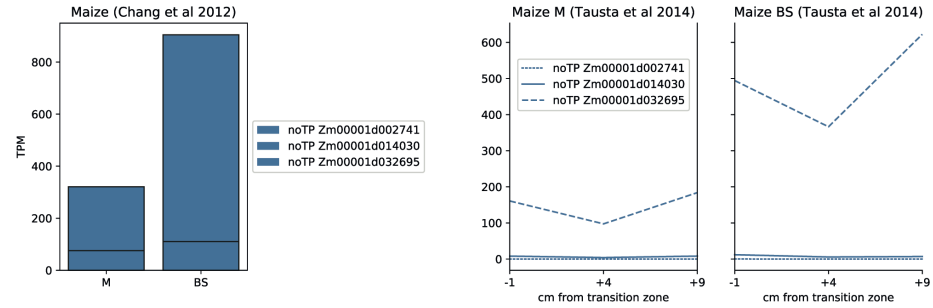
Figure A4.3. Agarose gel electrophoresis showing the presence or absence of *At*AILP1 and the CRISPR/Cas9 construct in each T2 plant. T2 plants are grouped by the parental line they came from: 1.4, 1.10, 1.11 and 1.12. Plants are numbered and two lanes are shown for each: an arrow going from left to right indicates that the PCR product from the amplification of the transgene (*Cas9*) is on the left (size 2328 bp) and the PCR product from the amplification of *AtAILP1* is on the right (size 2336 bp). For arrows going from right to left, the opposite is true. The ladder used is GeneRuler 1 kb DNA ladder from ThermoScientific (#SM0311).

Appendix 4.4: Expression of different malic dehydrogenase proteins in maize

OG0004873: NADP-MDH



OG0002526: cyNAD-MDH



OG0000524: NAD-MDH

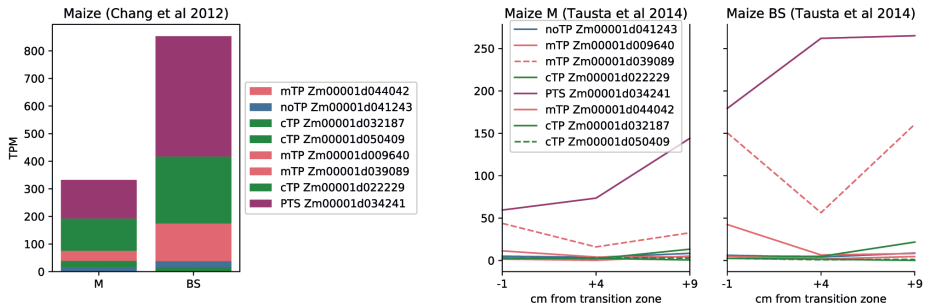


Figure A4.4. Expression of malate dehydrogenase in maize. Three malate dehydrogenase orthogroups were identified in the data set used in this study. Transcript amounts for the genes in each of these orthogroups is shown for mesophyll (M) and bundle sheath (BS) cells from mature maize leaves (Chang et al. 2012) and for mesophyll and bundle sheath cells along a developmental gradient (Tausta et al. 2014). The developmental gradient denotes distance from the source-sink transition zone of the maize leaf with -1cm representing the least mature tissue and +9cm representing the most mature tissue.

Appendix 4.5: Vector maps of the protein expression constructs used to express recombinant *Zm*AILP1 in *E. coli*

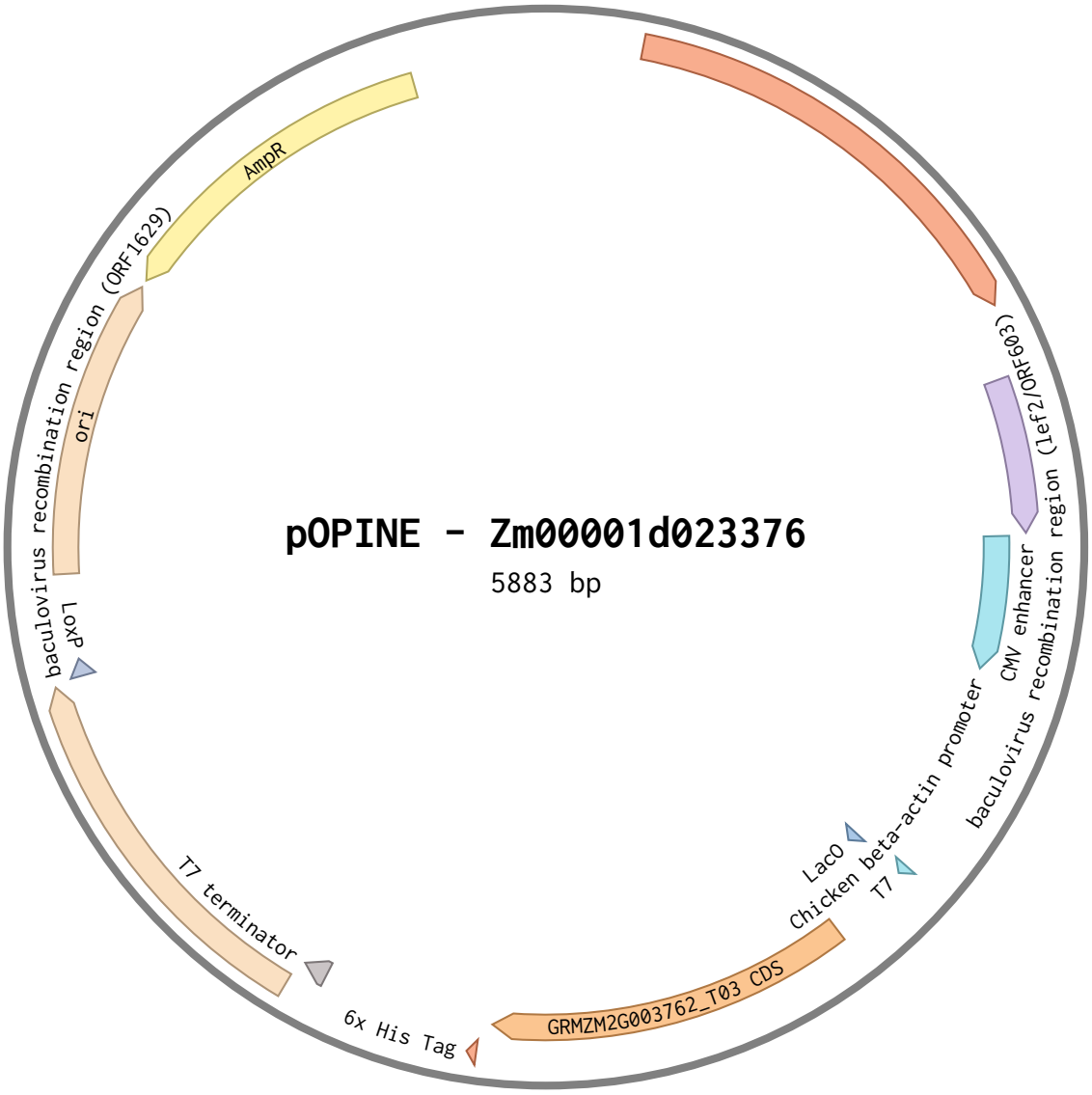


Figure A4.5.1. Vector map of the expression construct containing *ZmAILP1* fused to a C-terminal His-tag.

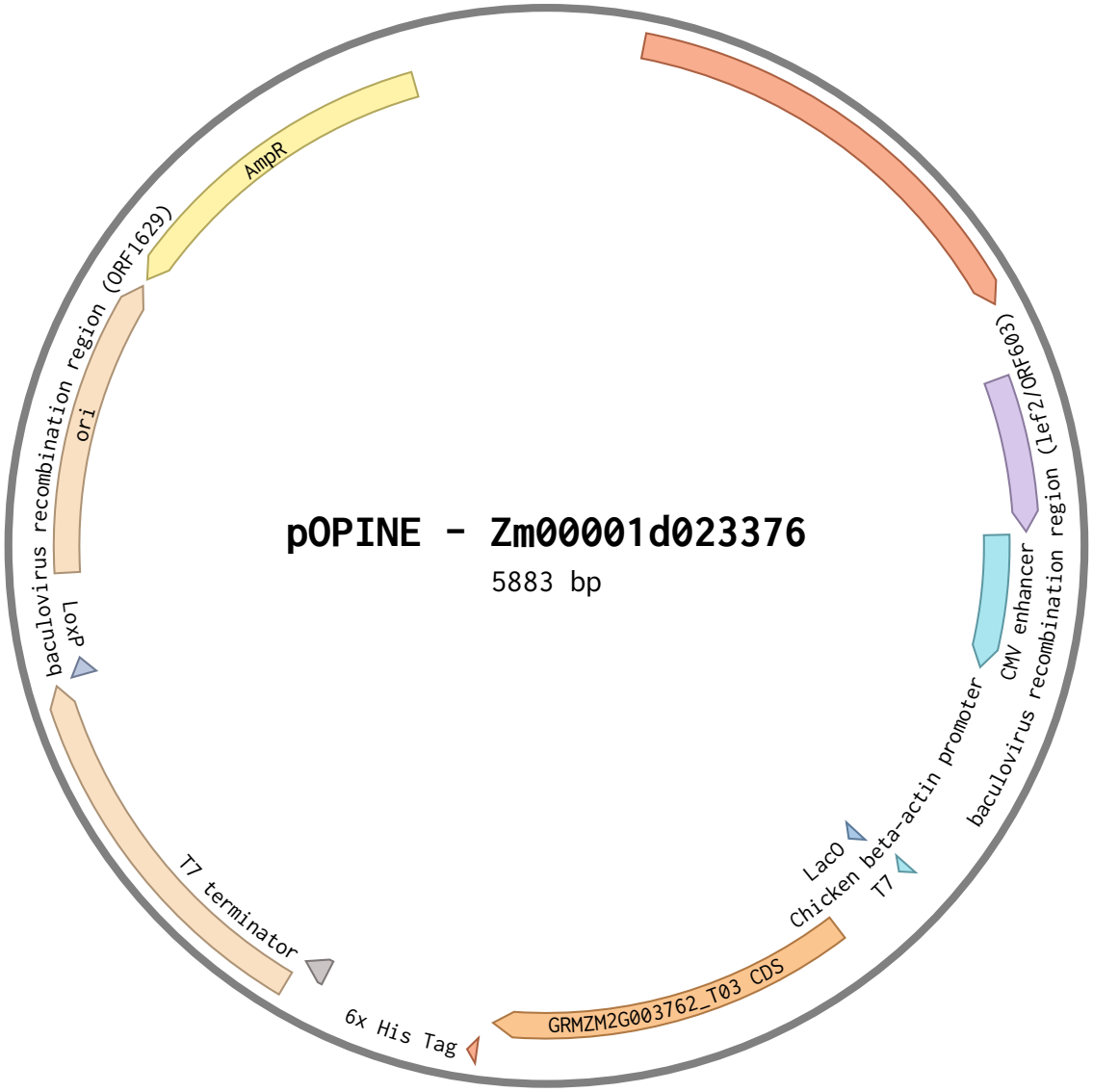


Figure A4.5.2. Vector map of the expression construct containing *ZmAILP1* fused to a C-terminal His-tag.

Appendix 4.6: Golden gate parts used for plant transformation work

Table A4.6.1. Golden gate parts for the generation of protoplast expression constructs

Level 0 golden gate part	ID	Description	Selection	Source
Backbone	EC47742	Level 1 vector	Amp (+ blue/white selection)	Sylvestre Marillonette (Weber et al. 2011)
PU	EC15058	35S promoter	Amp	Ben Miller, UEA
SC1	AT5G19140_SC1 Zm00001d023376_SC1	CDS with SC1 adaptors	Amp	This study
C2	EC15095	eGFP	Amp	Ben Miller, UEA
T	EC41421	Nos terminator	Amp	Sylvestre Marillonette (Weber et al. 2011)

Table A4.6.2. Golden Gate destination vectors for the short guide RNA (sgRNA) sequences used for *At*AILP1 gene disruption. The short guide RNAs (sgRNAs) were cloned into one of three level 1 Golden Gate vectors (EC17777, EC17778 or EC17779). Each contained the *U6-26* promoter which the sgRNA is cloned downstream of.

ID	Description	Selection	Source
EC17777	Compatible for cloning a sgRNA into the **first** sgRNA position in the final CRISPR-Cas9 construct	Amp (+blue/white selection)	Langdale lab
EC17778	Compatible for cloning a sgRNA into the **second** sgRNA position in the final CRISPR/Cas9 construct	Amp (+blue/white selection)	Langdale lab
EC17779	Compatible for cloning a sgRNA into the **third** sgRNA position in the final CRISPR/Cas9 construct	Amp (+blue/white selection)	Langdale lab

Table A4.6.3. Golden gate parts used in the assembly of the CRISPR/Cas9 plasmid for AtAILP1 gene disruption.

Assembly position	ID	Description	Selection	Source
Backbone	pAGM4723	Vector backbone for Level 2 Golden Gate cloning (adapted from addgene plasmid #48015)	Kan (+blue/white selection)	The Sainsbury Laboratory (Mark Youles and Laurence Toimlinson)
1	EC17703	*bar* gene for basta herbicide resistance	Amp	Langdale lab
2	EC17739	*Cas9p* gene	Amp	Langdale lab
3	EC17777	sgRNA1 driven by the *U6-26* promoter	Amp	Langdale lab
4	EC17778	sgRNA2 driven by the *U6-26* promoter	Amp	Langdale lab
5	EC17779	sgRNA3 driven by the *U6-26* promoter	Amp	Langdale lab
6	EC49299	Endlinker	Amp	Langdale lab



Anakin: Master, sir, I heard Yoda talking about midichlorians. I've been wondering, what are midichlorians?

Qui-Gon Jinn: Midichlorians are a microscopic life form that resides within all living cells.

Anakin: They live inside me?

Qui-Gon Jinn: Inside your cells, yes. And we are symbionts with them.

Anakin: Symbionts?

Qui-Gon Jinn: Life forms living together for mutual advantage. Without the midichlorians, life could not exist and we would have no knowledge of the Force. They continually speak to us telling us the will of the Force.

Star Wars: Episode I — The Phantom Menace (1999), written by George Lucas

References

Adams KL, Qiu Y-L, Stoutemyer M, Palmer JD. 2002. Punctuated evolution of mitochondrial gene content: High and variable rates of mitochondrial gene loss and transfer to the nucleus during angiosperm evolution. Proc. Natl. Acad. Sci. 99:9905–9912.

Akopian D, Shen K, Zhang X, Shan SO. 2013. Signal recognition particle: An essential protein-targeting machine. Annu. Rev. Biochem. 82:693–721.

Aliscioni S, Bell HL, Besnard G, Christin PA, Columbus JT, Duvall MR, Edwards EJ, Giussani L, Hasenstab-Lehman K, Hilu KW, et al. 2012. New grass phylogeny resolves deep evolutionary relationships and discovers C4 origins. New Phytol. 193:304–312.

Archibald JM. 2015a. Endosymbiosis and eukaryotic cell evolution. Curr. Biol. 25:R911–R921.

Archibald JM. 2015b. Genomic perspectives on the birth and spread of plastids. Proc. Natl. Acad. Sci. U. S. A. 112:10147–10153.

Armenteros JJA, Salvatore M, Emanuelsson O, Winther O, Von Heijne G, Elofsson A, Nielsen H. 2019. Detecting sequence signals in targeting peptides using deep learning. Life Sci. Alliance 2:e201900429.

Arrivault S, Obata T, Szecówka M, Mengin V, Guenther M, Hoehne M, Fernie AR, Stitt M. 2017. Metabolite pools and carbon flow during C4 photosynthesis in maize: 13CO 2 labeling kinetics and cell type fractionation. J. Exp. Bot. 68:283–298.

Assis R, Bachtrog D. 2013. Neofunctionalization of young duplicate genes in Drosophila. Proc. Natl. Acad. Sci. U. S. A. 110:17409–17414.

Athey J, Alexaki A, Osipova E, Rostovtsev A, Santana-Quintero L V., Katneni U, Simonyan V, Kimchi-Sarfaty C. 2017. A new and updated resource for codon usage tables. BMC Bioinformatics 18:1–10.

Aubry S, Brown NJ, Hibberd JM. 2011. The role of proteins in C3 plants prior to their recruitment into the C4 pathway. J. Exp. Bot. 62.

Aubry S, Smith-Unna RD, Boursnell CM, Kopriva S, Hibberd JM. 2014. Transcript residency on ribosomes reveals a key role for the Arabidopsis thaliana bundle sheath in sulfur and glucosinolate metabolism. Plant J. 78:659–673.

Badger M. 2003. The roles of carbonic anhydrases in photosynthetic CO2 concentrating mechanisms. Photosynth. Res. 77.

Badger MR, Price GD. 1994. The Role of Carbonic Anhydrase in Photosynthesis. Annu. Rev. Plant Physiol. Plant Mol. Biol. 45:369–392.

Bannai H, Tamada Y, Maruyama O, Nakai K, Miyano S. 2002. Extensive feature detection of N-terminal protein sorting signals. Bioinformatics 18:298–305.

Barbrook AC, Howe CJ, Kurniawan DP, Tarr SJ. 2010. Organization and expression of organellar genomes. Philos. Trans. R. Soc. B Biol. Sci. 365:785–797.

Baslam M, Oikawa K, Kitajima-Koga A, Kaneko K, Mitsui T. 2016. Golgi-to-plastid trafficking of proteins through secretory pathway: Insights into vesicle-mediated import toward the plastids. Plant Signal. Behav. 11:1–5.

Bauwe H, Hagemann M, Fernie AR. 2010. Photorespiration: players, partners and origin. Trends Plant Sci. 15:330–336.

Beilstein MA, Nagalingum NS, Clements MD, Manchester SR, Mathews S. 2010. Dated molecular phylogenies indicate a Miocene origin for Arabidopsis thaliana. Proc. Natl. Acad. Sci. 107:18724–18728.

Bellasio C, Griffiths H. 2014. The operation of two decarboxylases, transamination, and partitioning of C4 metabolic processes between mesophyll and bundle sheath cells allows light capture to be balanced for the maize C4 pathway. Plant Physiol. 164:466–480.

Bendtsen JD, Nielsen H, Von Heijne G, Brunak S. 2004. Improved prediction of signal peptides: SignalP 3.0. J. Mol. Biol. 340:783–795.

Bennetzen JL, Schmutz J, Wang H, Percifield R, Hawkins J, Pontaroli AC, Estep M, Feng L, Vaughn JN, Grimwood J, et al. 2012. Reference genome sequence of the model plant Setaria. Nat. Biotechnol. 30:555–561.

Bergthorsson U, Andersson DI, Roth JR. 2007. Ohno's dilemma: Evolution of new genes under continuous selection. Proc. Natl. Acad. Sci. U. S. A. 104:17004–17009.

Berrow NS, Alderton D, Sainsbury S, Nettleship J, Assenberg R, Rahman N, Stuart DI, Owens RJ. 2007. A versatile ligation-independent cloning method suitable for high-throughput expression screening applications. Nucleic Acids Res. 35:e45.

Blanc G, Wolfe KH. 2004. Functional divergence of duplicated genes formed by polyploidy during arabidopsis evolution. Plant Cell 16:1679–1691.

Blobel G, Dobberstein B. 1975a. Transfer of proteins across membranes. I. Presence of proteolytically processed and unprocessed nascent immunoglobulin light chains on membrane-bound ribosomes of murine myeloma. J. Cell Biol. 9:835–851.

Blobel G, Dobberstein B. 1975b. Transfer of proteins across membranes: II. reconstitution of functional rough microsomes from heterologous components. J. Cell Biol. 67:852–862.

Blomme T, Vandepoele K, De Bodt S, Simillion C, Maere S, Van de Peer Y. 2006. The gain and loss of genes during 600 million years of vertebrate evolution. Genome Biol. 7:R43.

Boersma S, Khuperkar D, Verhagen BMP, Sonneveld S, Grimm JB, Lavis LD, Tanenbaum ME. 2019. Multi-Color Single-Molecule Imaging Uncovers Extensive Heterogeneity in mRNA Decoding. Cell 178:458-472.e19.

Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: A flexible trimmer for Illumina sequence data. Bioinformatics 30:2114–2120.

Borderies G, Jamet E, Lafitte C, Rossignol M, Jauneau A, Boduart G, Monsarrat B, Esquerré-Tugayé MT, Boudet A, Pont-Lezica R. 2003. Proteomics of loosely bound cell wall proteins of Arabidopsis thaliana cell suspension cultures: A critical analysis. Electrophoresis 24:3421–3432.

Boussau B, Szollosi GJ, Duret L, Gouy M, Tannier E, Daubin V. 2013. Genome-scale coestimation of species and gene trees. Genome Res. 23:323–330.

Bowsher CG, Tobin AK. 2001. Compartmentation of metabolism within mitochondria and plastids. J. Exp. Bot. 52:513–527.

Bräutigam A, Schliesky S, Külahoglu C, Osborne CP, Weber APM. 2014. Towards an integrative model of C4 photosynthetic subtypes: Insights from comparative transcriptome analysis of NAD-ME, NADP-ME, and PEP-CK C4 species. J. Exp. Bot. 65:3579–3593.

Brown NJ, Newell CA, Stanley S, Chen JE, Perrin AJ, Kajala K, Hibberd JM. 2011. Independent and parallel recruitment of preexisting mechanisms underlying C4 photosynthesis. Science (80-.). 331:1436–1439.

Bruce BD. 2000. Chloroplast transit peptides: Structure, function and evolution. Trends Cell Biol. 10:440–447.

Brzobohatý B, Moore I, Kristoffersen P, Bako L, Campos N, Schell J, Palme K. 1993. Release of active cytokinin by a β-glucosidase localized to the maize root meristem. Science (80-.). 262:1051–1054.

Burgess RR. 2009. Chapter 4 Preparing a Purification Summary Table. 1st ed. Elsevier Inc.

Byun-McKay SA, Geeta R. 2007. Protein subcellular relocalization: a new perspective on the origin of novel genes. Trends Ecol. Evol. 22:338–344.

Byun SA, Singh S. 2013. Protein subcellular relocalization increases the retention of eukaryotic duplicate genes. Genome Biol. Evol. 5:2402–2409.

Cacefo V, Ribas AF, Zilliani RR, Neris DM, Domingues DS, Moro AL, Vieira LGE. 2019. Decarboxylation mechanisms of C4 photosynthesis in Saccharum spp.: Increased PEPCK activity under water-limiting conditions. BMC Plant Biol. 19:1–14.

von Caemmerer S, Furbank RT. 2016. Strategies for improving C4 photosynthesis. Curr. Opin. Plant Biol. 31:125–134.

Von Caemmerer S, Furbank RT. 2003. The C4 pathway: An efficient CO2 pump. Photosynth. Res. 77:191–207.

Von Caemmerer S, Quick WP, Furbank RT. 2012. The development of C4 rice: Current progress and future challenges. Science (80-.). 336:1671–1672.

van Campen J, Yaapar MN, Narawatthana S, Lehmeier C, Wanchana S, Thakur V, Chater C, Kelly S, Rolfe SA, Quick WP, et al. 2016. Combined Chlorophyll Fluorescence and Transcriptomic Analysis Identifies the P3/P4 Transition as a Key stage in Rice Leaf Photosynthetic Development. Plant Physiol. .

Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. 2009. trimAl: A tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics 25:1972–1973.

Carrigan MA, Uryasev O, Frye CB, Eckman BL, Myers CR, Hurley TD, Benner SA. 2015. Hominids adapted to metabolize ethanol long before human-directed fermentation. Proc. Natl. Acad. Sci. U. S. A. 112:458–463.

Chaikeeratisak V, Nguyen K, Khanna K, Brilot AF, Erb ML, Coker JKC, Vavilina A, Newton GL, Buschauer R, Pogliano K, et al. 2017. Assembly of a nucleus-like structure during viral replication in bacteria. Science (80-.). 355:194–197.

Chang YM, Liu WY, Shih ACC, Shen MN, Lu CH, Lu MYJ, Yang HW, Wang TY, Chen SCCSMSCCSM, Chen SCCSMSCCSM, et al. 2012. Characterizing regulatory and functional differentiation between maize mesophyll and bundle sheath cells by transcriptomic analysis. Plant Physiol. 160:165–177.

Chapman K, Hatch M. 1981. Aspartate Decarboxylation in Bundle Sheath Cells of Zea mays and Its Possible Contribution to C3 Photosynthesis. Funct. Plant Biol. 8:237–248.

Chapman KSR, Hatch MD. 1983. Intracellular location of phosphoenolpyruvate carboxykinase and other C4 photosynthetic enzymes in mesophyll and bundle sheath protoplasts of Panicum maximum. Plant Sci. Lett. 29:145–154.

Chiang YH, Zubo YO, Tapken W, Kim HJ, Lavanway AM, Howard L, Pilon M, Kieber JJ, Eric Schaller G. 2012. Functional characterization of the GATA transcription factors GNC and CGA1 reveals their key role in chloroplast development, growth, and division in Arabidopsis. Plant Physiol. 160:332–348.

Chowdhary G, Kataya ARA, Lingner T, Reumann S. 2012. Non-canonical peroxisome targeting signals: identification of novel PTS1 tripeptides and characterization of enhancer elements by computational permutation analysis. BMC Plant Biol. 12:1–14.

Christian RW, Hewitt SL, Nelson G, H RE, Dhingra A. 2020. Plastid transit peptides—where do they come from and where do they all belong? Multi-genome and pan-genomic assessment of chloroplast transit peptide evolution. PeerJ 8.

Christian RW, Hewitt SL, Roalson EH, Dhingra A. 2020. Genome-Scale Characterization of Predicted Plastid-Targeted Proteomes in Higher Plants. Sci. Rep. 10:1–22.

Clayton H, Saladié M, Rolland V, Sharwood R, Macfarlane T, Ludwig M. 2017. Loss of the Chloroplast Transit Peptide from an Ancestral C3 Carbonic Anhydrase Is Associated with C4 Evolution in the Grass Genus Neurachne. Plant Physiol. 173:1648–1658.

Cohen O, Gophna U, Pupko T. 2011. The complexity hypothesis revisited: Connectivity Rather Than function constitutes a barrier to horizontal gene transfer. Mol. Biol. Evol. 28:1481–1489.

Cokol M, Nair R, Rost B. 2000. Finding nuclear localization signals. EMBO Rep. 1:411–415.

Conant GC, Birchler JA, Pires JC. 2014. Dosage, duplication, and diploidization: Clarifying the interplay of multiple models for duplicate gene evolution over time. Curr. Opin. Plant Biol. 19:91–98.

Conant GC, Wolfe KH. 2008. Turning a hobby into a job: How duplicated genes find new functions. Nat. Rev. Genet. 9:938–950.

Concordet JP, Haeussler M. 2018. CRISPOR: Intuitive guide selection for CRISPR/Cas9 genome editing experiments and screens. Nucleic Acids Res. 46:W242–W245.

Copley SD. 2020. Evolution of new enzymes by gene duplication and divergence. FEBS J. 287:1262–1283.

Cortleven A, Schmülling T. 2015. Regulation of chloroplast development and function by cytokinin. J. Exp. Bot. 66:4999–5013.

Costello R, Emms DM, Kelly S. 2020. Gene Duplication Accelerates the Pace of Protein Gain and Loss from Plant Organelles. Mol. Biol. Evol. 37:969–981.

Cui J, Liu J, Li Y, Shi T. 2011. Integrative identification of Arabidopsis mitochondrial proteome and its function exploitation through protein interaction network. PLoS One 6.

Dacks JB, Field MC. 2007. Evolution of the eukaryotic membrane-trafficking system: origin, tempo and mode. J. Cell Sci. 120:2977–2985.

Daniell H, Lin CS, Yu M, Chang WJ. 2016. Chloroplast genomes: Diversity, evolution, and applications in genetic engineering. Genome Biol. 17.

Denton AK, Maß J, Külahoglu C, Lercher MJ, Bräutigam A, Weber APM. 2017. Freeze-quenched maize mesophyll and bundle sheath separation uncovers bias in previous tissue-specific RNA-Seq data. J. Exp. Bot. 68.

DiMario RJ, Clayton H, Mukherjee A, Ludwig M, Moroney J V. 2017. Plant Carbonic Anhydrases: Structures, Locations, Evolution, and Physiological Roles. Mol. Plant 10:30–46.

Dobler R, Rogell B, Budar F, Dowling DK. 2014. A meta-analysis of the strength and nature of cytoplasmic genetic effects. J. Evol. Biol. 27:2021–2034.

Duff SMG, Qi Q, Reich T, Wu X, Brown T, Crowley JH, Fabbri B. 2011. A kinetic comparison of asparagine synthetase isozymes from higher plants. Plant Physiol. Biochem. 49:251–256.

Edwards GE. 2011. C4 Photosynthesis: Kranz forms and single-cell C4 in terrestrial plants. In: Raghavendra AS, Sage RF, editors. C4 photosynthesis and related CO2 concentrating mechanisms. Springer. p. 29–61.

Edwards GE, Franceschi VR, Voznesenskaya E V. 2004. Single-cell C4 photosynthesis versus the dual-cell (Kranz) paradigm. Annu. Rev. Plant Biol. 55:173–196.

Edwards GE, Kanai R, Black CC. 1971. Phosphoenolpyruvate carboxykinase in leaves of certain plants which fix CO2 by the C4-dicarboxylic acid cycle of photosynthesis. Biochem. Biophys. Res. Commun. 45:278–285.

Ehleringer JR, Cerling TE, Helliker BR. 1997. C4 photosynthesis, atmospheric CO2, and climate. Oecologia 112:285–299.

Ehleringer JR, Sage RF, Flanagan LB, Pearcy RW. 1991. Climate change and the evolution of C4 photosynthesis. Trends Ecol. Evol. 6:95–99.

Emanuelsson O, Nielsen H, Brunak S, von Heijne G. 2000. Predicting Subcellular Localization of Proteins Based on their N-terminal Amino Acid Sequence. J. Mol. Biol. 300:1005–1016.

Emms DM, Covshoff S, Hibberd JM, Kelly S. 2016. Independent and Parallel Evolution of New Genes by Gene Duplication in Two Origins of C4 Photosynthesis Provides New Insight into the Mechanism of Phloem Loading in C4 Species. Mol. Biol. Evol. 33:1796–1806.

Emms DM, Kelly S. 2015. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. Genome Biol. 16:157.

Engler C, Kandzia R, Marillonnet S, Marsischky G, LaBaer J, Marillonnet S, Giritch A, Gils M, Kandzia R, Klimyuk V, et al. 2008. A one pot, one step, precision cloning method with high throughput capability.El-Shemy HA, editor. PLoS One 3:e3647.

Ermakova M, Danila FR, Furbank RT, von Caemmerer S. 2020. On the road to C4 rice: advances and perspectives. Plant J. 101:940–950.

Fakruddin M, Mohammad Mazumdar R, Bin Mannan KS, Chowdhury A, Hossain MN. 2013. Critical Factors Affecting the Success of Cloning, Expression, and Mass Production of Enzymes by Recombinant E. coli. ISRN Biotechnol. 2013.

Finn RD, Clements J, Eddy SR. 2011. HMMER web server: Interactive sequence similarity searching. Nucleic Acids Res. 39.

Force A, Lynch M, Pickett FB, Amores A, Yan YL, Postlethwait J. 1999. Preservation of duplicate genes by complementary, degenerative mutations. Genetics 151:1531–1545.

Freeling M. 2009. Bias in plant gene content following different sorts of duplication: Tandem, whole-genome, segmental, or by transposition. Annu. Rev. Plant Biol. 60:433–453.

Freeling M, Scanlon MJ, Fowler JF. 2015. Fractionation and subfunctionalization following genome duplications: Mechanisms that drive gene content and their consequences. Curr. Opin. Genet. Dev. 35:110–118.

Friso G, Majeran W, Huang M, Sun Q, van Wijk KJ. 2010. Reconstruction of metabolic pathways, protein expression, and homeostasis machineries across maize bundle sheath and mesophyll chloroplasts: Large-scale quantitative proteomics using the first maize genome assembly. Plant Physiol. 152:1219–1250.

Furbank RT. 2011. Evolution of the C4 photosynthetic mechanism: Are there really three C4 acid decarboxylation types? J. Exp. Bot. 62:3103–3108.

Gabaldón T, Huynen MA. 2003. Reconstruction of the proto-mitochondrial metabolism. Science (80-.). 301:609.

Gabaldón T, Huynen MA. 2004. Shaping the mitochondrial proteome. Biochim. Biophys. Acta -

Bioenerg. 1659:212–220.

Gabaldón T, Huynen MA. 2007. From endosymbiont to host-controlled organelle: The hijacking of mitochondrial protein synthesis and metabolism. PLoS Comput. Biol. 3:2209–2218.

Gabaldón T, Pittis AA. 2015. Origin and evolution of metabolic sub-cellular compartmentalization in eukaryotes. Biochimie 119:262–268.

Gabaldón T, Snel B, van Zimmeren F, Hemrika W, Tabak H, Huynen MA. 2006. Origin and evolution of the peroxisomal proteome. Biol. Direct 1.

Gajdošová S, Spíchal L, Kamínek M, Hoyerová K, Novák O, Dobrev PI, Galuszka P, Klíma P, Gaudinová A, Žižková E, et al. 2011. Distribution, biological activities, metabolism, and the conceivable function of cis-zeatin-type cytokinins in plants. J. Exp. Bot. 62:2827–2840.

Ghannoum O, Evans JR, von Caemmerer S. 2010. Nitrogen and Water Use Efficiency of C4 Plants. In: C4 Photosynthesis and Related CO2 concentrating mechanisms. p. 129–146.

Goodstein DM, Shu S, Howson R, Neupane R, Hayes RD, Fazo J, Mitros T, Dirks W, Hellsten U, Putnam N, et al. 2012. Phytozome: A comparative platform for green plant genomics. Nucleic Acids Res. 40:D1178–D1186.

Gowik U, Bräutigam A, Weber KL, Weber APM, Westhoff P. 2011. Evolution of C4 photosynthesis in the genus flaveria: How many and which genes does it take to make C 4? Plant Cell 23:2087–2105.

Gowik U, Westhoff P. 2011. The path from C3 to C4 photosynthesis. Plant Physiol. 155:56–63.

Gray MW. 2012. Mitochondrial evolution. Cold Spring Harb. Perspect. Biol. 4.

Gray MW. 2015. Mosaic nature of the mitochondrial proteome: Implications for the origin and evolution of mitochondria. Proc. Natl. Acad. Sci. 112:10133–10138.

Gray MW, Burger G, Lang BF. 1999. Mitochondrial evolution. Science (80-.). 283:1476–1481.

Gray MW, Doolittle WF. 1982. Has the endosymbiont hypothesis been proven? Microbiol. Rev. 46:1–42.

Gray MW, Lang BF. 1998. Transcription in chloroplasts and mitochondria: a tale of two polymerases. Trends Microbiol. 6:1–3.

Green BR. 2011. Chloroplast genomes of photosynthetic eukaryotes. Plant J. 66:34–44.

Guo Y-L, Bechsgaard JS, Slotte T, Neuffer B, Lascoux M, Weigel D, Schierup MH. 2009. Recent speciation of Capsella rubella from Capsella grandiflora, associated with loss of self-incompatibility and an extreme bottleneck. Proc. Natl. Acad. Sci. 106:5246–5251.

Gupta DB, Rai Y, Gayali S, Chakraborty S, Chakraborty N. 2016. Plant Organellar Proteomics in Response to Dehydration: Turning Protein Repertoire into Insights. Front. Plant Sci. 7:460.

Gutierrez M, Gracen VE, Edwards GE. 1974. Biochemical and cytological relationships in C4 plants. Planta 119:279–300.

Gutmann B, Royan S, Schallenberg-Rüdinger M, Lenz H, Castleden IR, McDowell R, Vacher MA, Tonti-Filippini J, Bond CS, Knoop V, et al. 2020. The Expansion and Diversification of Pentatricopeptide Repeat RNA-Editing Factors in Plants. Mol. Plant 13:215–230.

Haberlandt G. 1918. Physiologische pflanzenanatomie. (Engelman W, editor.). Leipzig, Germany

Hammani K, Gobert A, Hleibieh K, Choulier L, Small I, Giegé P. 2011. An Arabidopsis dual-localized pentatricopeptide repeat protein interacts with nuclear proteins involved in gene expression regulation. Plant Cell 23:730–740.

Hatch M, Kagawa T, Craig S. 1975. Subdivision of C4-Pathway Species Based on Differing C4 Acid Decarboxylating Systems and Ultrastructural Features. Funct. Plant Biol. 2:111–128.

Hatch MD. 1971. The C4-pathway of photosynthesis. Evidence for an intermediate pool of carbon dioxide and the identity of the donor C4-dicarboxylic acid. Biochem. J. 125:425–432.

Hatch MD. 1973. Separation and properties of leaf aspartate aminotransferase and alanine aminotransferase isoenzymes operative in the C4 pathway of photosynthesis. Arch. Biochem. Biophys. 156:207–214.

Hatch MD. 1987. C4 photosynthesis: a unique blend of modified biochemistry, anatomy and ultrastructure. Biochim. Biophys. Acta - Rev. Bioenerg. 895:81–106.

Hatch MD, Burnell JN. 1990. Carbonic anhydrase activity in leaves and its role in the first step of C4 photosynthesis. Plant Physiol. 93:825–828.

Hatch MD, Kagawa T. 1976. Photosynthetic activities of isolated bundle sheath cells in relation to differing mechanisms of C4 pathway photosynthesis. Arch. Biochem. Biophys. 175:39–53.

Hatch MD, Mau SL. 1973. Activity, location, and role of asparate aminotransferase and alanine aminotransferase isoenzymes in leaves with C4 pathway photosynthesis. Arch. Biochem. Biophys. 156:195–206.

Hatch MD, Slack CR. 1966. Photosynthesis by sugar-cane leaves. A new carboxylation reaction and the pathway of sugar formation. Biochem. J. 101:103.

Hattersley P, Watson L, Osmond C. 1977. In situ Immunofluorescent Labelling of Ribulose-1,5-Bisphosphate Carboxylase in Leaves of C3 and C4 Plants. Funct. Plant Biol. 4:523–539.

Häusler RE, Hirsch HJ, Kreuzaler F, Peterhänsel C. 2002. Overexpression of C4-cycle enzymes in transgenic C3 plants: A biotechnological approach to improve C3-photosynthesis. J. Exp. Bot. 53:591–607.

Hayes ML, Giang K, Berhane B, Mulligan RM. 2013. Identification of two pentatricopeptide repeat genes required for rna editing and zinc binding by c-terminal cytidine deaminase-like domains. J. Biol. Chem. 288:36519–36529.

Heilmann I, Pidkowich MS, Girke T, Shanklin J. 2004. Switching desaturase enzyme specificity by alternate subcellular targeting. Proc. Natl. Acad. Sci. U. S. A. 101:10266–10271.

Hein A, Brenner S, Knoop V, Pritham E. 2019. Multifarious Evolutionary Pathways of a Nuclear RNA Editing Factor: Disjunctions in Coevolution of DOT4 and Its Chloroplast Target rpoC1eU488SL. Genome Biol. Evol. 11:798–813.

Hibberd JM, Covshoff S. 2010. The regulation of gene expression required for C4 photosynthesis. Annu. Rev. Plant Biol. 61:181–207.

Hibberd JM, Sheehy JE, Langdale JA. 2008. Using C4 photosynthesis to increase the yield of rice-rationale and feasibility. Curr. Opin. Plant Biol. 11:228–231.

Hildebrandt TM, Nunes Nesi A, Araújo WL, Braun HP. 2015. Amino Acid Catabolism in Plants. Mol. Plant 8:1563–1579.

Hoang DT, Chernomor O, Von Haeseler A, Minh BQ, Vinh LS. 2018. UFBoot2: Improving the ultrafast bootstrap approximation. Mol. Biol. Evol. 35:518–522.

Hooper CM, Tanz SK, Castleden IR, Vacher MA, Small ID, Millar AH. 2014. SUBAcon: A consensus algorithm for unifying the subcellular localization data of the Arabidopsis proteome. Bioinformatics 30:3356–3364.

Horton P, Park KJ, Obayashi T, Fujita N, Harada H, Adams-Collier CJ, Nakai K. 2007. WoLF PSORT: Protein localization predictor. Nucleic Acids Res. 35:W585–W587.

Hošek P, Hoyerová K, Kiran NS, Dobrev PI, Zahajská L, Filepová R, Motyka V, Müller K, Kamínek M. 2020. Distinct metabolism of N-glucosides of isopentenyladenine and trans-zeatin determines cytokinin metabolic spectrum in Arabidopsis. New Phytol. 225:2423–2438.

Hoyerová K, Hošek P. 2020. New Insights Into the Metabolism and Role of Cytokinin N-Glucosides in Plants. Front. Plant Sci. 11:5–11.

Hudson D, Guevara DR, Hand AJ, Xu Z, Hao L, Chen X, Zhu T, Bi YM, Rothstein SJ. 2013. Rice cytokinin GATA transcription factor1 regulates chloroplast development and plant architecture. Plant Physiol. 162:132–144.

Huerta-Cepas J, Serra F, Bork P. 2016. ETE 3: Reconstruction, Analysis, and Visualization of Phylogenomic Data. Mol. Biol. Evol. 33:1635–1638.

Huson DH, Richter DC, Rausch C, Dezulian T, Franz M, Rupp R. 2007. Dendroscope: An interactive viewer for large phylogenetic trees. BMC Bioinformatics 8.

Hwang I, Sheen J, Müller B. 2012. Cytokinin signaling networks. Annu. Rev. Plant Biol. 63:353–380.

Imai K, Nakai K. 2020. Tools for the recognition of sorting signals and the prediction of subcellular localization of proteins from their amino acid sequences. Front. Genet. 11:1491.

Innan H, Kondrashov F. 2010. The evolution of gene duplications: classifying and distinguishing between models. Nat. Rev. Genet. 11:97–108.

Jacob F. 1977. Evolution and tinkering. Science (80-.). 196.

Jain R, Rivera MC, Lake JA. 1999. Horizontal gene transfer among genomes: The complexity hypothesis. Proc. Natl. Acad. Sci. U. S. A. 96:3801–3806.

Jansen RK, Cai Z, Raubeson LA, Daniell H, DePamphilis CW, Leebens-Mack J, Muller KF, Guisinger-Bellian M, Haberle RC, Hansen AK, et al. 2007. Analysis of 81 genes from 64 plastid genomes resolves relationships in angiosperms and identifies genome-scale evolutionary patterns. Proc. Natl. Acad. Sci. 104:19369–19374.

Jarvis P. 2008. Targeting of nucleus-encoded proteins to chloroplasts in plants. New Phytol. 179:257–285.

Jensen PE, Leister D. 2014. Chloroplast evolution, structure and functions. F1000Prime Rep. 6:1–14.

Jiao Y, Wickett NJ, Ayyampalayam S, Chanderbali AS, Landherr L, Ralph PE, Tomsho LP, Hu Y, Liang H, Soltis PS, et al. 2011. Ancestral polyploidy in seed plants and angiosperms. Nature 473:97–100.

John CR, Smith-Unna RD, Woodfield H, Covshoff S, Hibberd JM. 2014. Evolutionary convergence of cell-specific gene expression in independent lineages of C4 grasses. Plant Physiol. 165:62–75.

Joseph B, Corwin JA, Li B, Atwell S, Kliebenstein DJ. 2013. Cytoplasmic genetic variation and extensive cytonuclear interactions influence natural variation in the metabolome. Elife 2:e00776.

Kaessmann H. 2010. Origins, evolution, and phenotypic impact of new genes. Genome Res. 20:1313–1326.

Käll L, Krogh A, Sonnhammer ELL. 2004. A combined transmembrane topology and signal peptide prediction method. J. Mol. Biol. 338:1027–1036.

Kanai R, Edwards G. 1999. The Biochemistry of C4 Photosynthesis. In: C4 Plant Biology. San Diego: Elsevier Science & Technology. p. 49–87.

Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. Mol. Biol. Evol. 30:772–780.

Kaur N, Hu J. 2011. Defining the Plant Peroxisomal Proteome: From Arabidopsis to Rice. Front. Plant Sci. 2:103.

Keeling PJ. 2013. The number, speed, and impact of plastid endosymbioses in eukaryotic evolution. Annu. Rev. Plant Biol. 64:583–607.

Kelly S, Covshoff S, Wanchana S, Thakur V, Quick WP, Wang Y, Ludwig M, Bruskiewich R, Fernie AR, Sage RF, et al. 2017. Wide sampling of natural diversity identifies novel molecular signatures of C4 photosynthesis. bioRxiv.

Kerfeld CA, Erbilgin O. 2015. Bacterial microcompartments and the modular construction of microbial metabolism. Trends Microbiol. 23:22–34.

Ketudat Cairns JR, Esen A. 2010. β-Glucosidases. Cell. Mol. Life Sci. 67:3389–3405.

Khamis S, Lamaze T, Farineau J. 1992. Effect of nitrate limitation on the photosynthetically active pools of aspartate and malate in maize, a NADP malic enzyme C4 plant. Physiol. Plant. 85:223–229.

Kiran NS, Polanská L, Fohlerová R, Mazura P, Válková M, Šmeral M, Zouhar J, Malbeck J, Dobrev PI, Macháčková I, et al. 2006. Ectopic over-expression of the maize β-glucosidase Zm-p60.1 perturbs cytokinin homeostasis in transgenic tobacco. J. Exp. Bot. 57:985–996.

Klee EW, Ellis LBM. 2005. Evaluating eukaryotic secreted protein prediction. BMC Bioinformatics BMC Bioinf:256.

Kobayashi K, Baba S, Obayashi T, Sato M, Toyooka K, Keränen M, Aro EM, Fukaki H, Ohta H, Sugimoto K, et al. 2012. Regulation of root greening by light and auxin/cytokinin signaling in Arabidopsis. Plant Cell 24:1081–1095.

Koonin E V. 2010. The origin and early evolution of eukaryotes in the light of phylogenomics. Genome Biol. 11:1–12.

Koonin E V. 2015. Archaeal ancestors of eukaryotes: Not so elusive any more. BMC Biol. 13.

Kunze M, Berger J. 2015. The similarity between N-terminal targeting signals for protein import into different organelles and its evolutionary relevance. Front. Physiol. 6.

De La Torre F, Cañas RA, Pascual MB, Avila C, Cánovas FM. 2014. Plastidic aspartate aminotransferases and the biosynthesis of essential amino acids in plants. J. Exp. Bot. 65:5527–5534.

Lane N, Martin W. 2010. The energetics of genome complexity. Nature 467:929–934.

Lange A, Mills RE, Lange CJ, Stewart M, Devine SE, Corbett AH. 2007. Classical nuclear localization signals: Definition, function, and interaction with importin α. J. Biol. Chem. 282:5101–5105.

Law MY, Childs KL, Campbell MS, Stein JC, Olson AJ, Holt C, Panchy N, Lei J, Jiao D, Andorf CM, et al. 2015. Automated update, revision, and quality control of the maize genome annotations using MAKER-P improves the B73 RefGen_v3 gene models and identifies new genes1. Plant Physiol. 167:25–39.

Leasure CD, Tong H, Yuen G, Hou X, Sun X, He ZH. 2009. Root UV-B sensitive2 acts with root UV-B sensitive1 in a root ultraviolet B-sensing pathway. Plant Physiol. 150:1902–1915.

Lebendiker M, Danieli T. 2011. Purification of Proteins Fused to Maltose-Binding Protein. In: Walls D, Loughran ST, editors. Protein Chromatography: Methods and Protocols. Totowa, NJ: Humana Press. p. 281–293.

Lee DW, Lee S, Lee GJ, Lee KH, Kim S, Cheong GW, Hwang I. 2006. Functional characterization of sequence motifs in the transit peptide of arabidopsis small subunit of Rubisco. Plant Physiol. 140:466–483.

Lee J, Kim DH, Hwang I. 2014. Specific targeting of proteins to outer envelope membranes of endosymbiotic organelles, chloroplasts, and mitochondria. Front. Plant Sci. 5:173.

Lee K, Kang H. 2020. Roles of organellar rna-binding proteins in plant growth, development, and abiotic stress responses. Int. J. Mol. Sci. 21:1–17.

Leegood RC. 2008. Roles of the bundle sheath cells in leaves of C3 plants. J. Exp. Bot. 59:1663–1673.

Leister D. 2016. Towards understanding the evolution and functional diversification of DNA-containing plant organelles. F1000Research 5:330.

Leister D, Kleine T. 2008. Towards a comprehensive catalog of chloroplast proteins and their interactions. Cell Res. 18:1081–1083.

Leitch EAR, Soltis DE, Soltis PS, Leitch IJ, Pires JC, Mable BK. 2004. Biological relevance of polyploidy : ecology to genomics ‘ Why polyploidy is rarer in animals than in plants ’: myths and mechanisms. Society:453–466.

Léon S, Goodman JM, Subramani S. 2006. Uniqueness of the mechanism of protein import into the peroxisome matrix: Transport of folded, co-factor-bound and oligomeric proteins by shuttling receptors. Biochim. Biophys. Acta - Mol. Cell Res. 1763:1552–1564.

Li H, Chiu C-C. 2010. Protein Transport into Chloroplasts. Annu. Rev. Plant Biol. 61:157–180.

Li L, Nelson CJ, Trösch J, Castleden I, Huang S, Millar AH. 2017. Protein degradation rate in Arabidopsis thaliana leaf growth and development. Plant Cell 29:207–228.

Li Z, Defoort J, Tasdighian S, Maere S, Van de Peer Y, De Smet R. 2016. Gene duplicability of core genes is highly consistent across all Angiosperms. Plant Cell 28:326–344.

Liang Z, Schnable JC. 2018. Functional divergence between subgenomes and gene pairs after whole genome duplications. Mol. Plant 11:388–397.

Lim SA, Hart KM, Harms MJ, Marqusee S. 2016. Evolutionary trend toward kinetic stability in the folding trajectory of RNases H. Proc. Natl. Acad. Sci. U. S. A. 113:13045–13050.

Lin Wenfang, Huang D, Shi X, Deng B, Ren Y, Lin Wenxiong, Miao Y. 2019. H2O2 as a Feedback Signal on Dual-Located WHIRLY1 Associates with Leaf Senescence in Arabidopsis. Cells 8:1–20.

Lindsay MR, Webb RI, Strous M, Jetten MSM, Butler MK, Forde RJ, Fuerst JA. 2001. Cell compartmentalisation in planctomycetes: Novel types of structural organisation for the bacterial cell. Arch. Microbiol. 175:413–429.

Liu S-L, Pan AQ, Adams KL. 2014. Protein subcellular relocalization of duplicated genes in Arabidopsis. Genome Biol. Evol. 6:2501–2515.

Lomelino CL, Andring JT, McKenna R, Kilberg MS. 2017. Asparagine synthetase: Function, structure, and role in disease. J. Biol. Chem. 292:19952–19958.

Long M, Betrán E, Thornton K, Wang W. 2003. The origin of new genes: Glimpses from the young and old. Nat. Rev. Genet. 4:865–875.

Love MI, Huber W, Anders S. 2014. Moderated estimation of fold change and dispersion for RNA-

seq data with DESeq2. Genome Biol. 15.

Ludwig M. 2016. Evolution of carbonic anhydrase in C4 plants. Curr. Opin. Plant Biol. 31:16–22.

Lurin C, Andrés C, Aubourg S, Bellaoui M, Bitton F, Bruyère C, Caboche M, Debast C, Gualberto J, Hoffmann B, et al. 2004. Genome-wide analysis of arabidopsis pentatricopeptide repeat proteins reveals their essential role in organelle biogenesis. Plant Cell 16:2089–2103.

Lynch M, Conery JS. 2000. The evolutionary fate and consequences of duplicate genes. Science (80-.). 290:1151–1155.

Mackenzie S, McIntosh L. 1999. Higher plant mitochondria. Plant Cell 11:571–585.

Maere S, De Bodt S, Raes J, Casneuf T, Van Montagu M, Kuiper M, Van De Peer Y. 2005. Modeling gene and genome duplications in eukaryotes. Proc. Natl. Acad. Sci. U. S. A. 102:5454–5459.

Magadum S, Banerjee U, Murugan P, Gangapur D, Ravikesavan R. 2013. Gene duplication as a major force in evolution. J. Genet. 92:155–161.

Mähönen AP, Bishopp A, Higuchi M, Nieminen KM, Kinoshita K, Törmäkangas K, Ikeda Y, Oka A, Kakimoto T, Helariutta Y. 2006. Cytokinin signaling and its inhibitor AHP6 regulate cell fate during vascular development. Science (80-.). 311:94–98.

Majeran W, Cai Y, Sun Q, Van Wijk KJ. 2005. Functional differentiation of bundle sheath and mesophyll maize chloroplasts determined by comparative proteomics. Plant Cell 17:3111–3140.

Manna S. 2015. An overview of pentatricopeptide repeat proteins and their applications. Biochimie 113:93–99.

Marques AC, Vinckenbosch N, Brawand D, Kaessmann H. 2008. Functional diversification of duplicate genes through subcellular adaptation of encoded proteins. Genome Biol. 9:1–12.

Martin W, Herrmann RG. 1998. Gene Transfer from Organelles to the Nucleus: How Much, What Happens, and Why? Plant Physiol. 118:9–17.

Martin W, Rujan T, Richly E, Hansen A, Cornelsen S, Lins T, Leister D, Stoebe B, Hasegawa M, Penny D. 2002. Evolutionary analysis of Arabidopsis, cyanobacterial, and chloroplast genomes reveals plastid phylogeny and thousands of cyanobacterial genes in the nucleus. Proc. Natl. Acad. Sci. 99:12246–12251.

Mast FD, Barlow LD, Rachubinski RA, Dacks JB. 2014. Evolutionary mechanisms for establishing eukaryotic cellular complexity. Trends Cell Biol. 24:435–442.

McGeoch DJ. 1985. On the predictive recognition of signal peptide sequences. Virus Res. 3:271–286.

McKay SAB, Geeta R, Duggan R, Carroll B, McKay SJ. 2009. Missing the Subcellular Target: A Mechanism of Eukaryotic Gene Evolution. In: Pontarotti P, editor. Evolutionary Biology. Berlin, Heidelberg: Springer Berlin Heidelberg. p. 175–183.

Meister M, Agostino A, Hatch MD. 1996. The roles of malate and aspartate in C4 photosynthetic metabolism of Flaveria bidentis (L.). Planta 199:262–269.

Millar AH, Sweetlove LJ, Giegé P, Leaver CJ. 2001. Analysis of the Arabidopsis mitochondrial proteome. Plant Physiol. 127:1711–1727.

Mitschke J, Fuss J, Blum T, Höglund A, Reski R, Kohlbacher O, Rensing SA. 2009. Prediction of dual protein targeting to plant organelles: Methods. New Phytol. 183:224–236.

Möller S, Croning MDR, Apweiler R. 2001. Evaluation of methods for the prediction of membrane spanning regions. Bioinformatics 17:646–653.

Morant AV, Jørgensen K, Jørgensen C, Paquette SM, Sánchez-Pérez R, Møller BL, Bak S. 2008. β-Glucosidases as detonators of plant chemical defense. Phytochemistry 69:1795–1813.

Morris JL, Puttick MN, Clark JW, Edwards D, Kenrick P, Pressel S, Wellman CH, Yang Z, Schneider H, Donoghue PCJ. 2018. The timescale of early land plant evolution. Proc. Natl. Acad. Sci. 115:201719588.

Muhaidat R, Sage RF, Dengler NG. 2007. Diversity of Kranz anatomy and biochemistry in C4 eudicots. Am. J. Bot. 94:362–381.

Munekage YN, Taniguchi YY. 2016. Promotion of Cyclic Electron Transport Around Photosystem i with the Development of C4 Photosynthesis. Plant Cell Physiol. 57:897–903.

Nakai K, Kanehisa M. 1992. A knowledge base for predicting protein localization sites in eukaryotic cells. Genomics 14:897–911.

Nelson T, Langdale JA. 1992. Developmental Genetics of C4 Photosynthesis. Annu. Rev. Plant Biol. 43:25–47.

Neupert W, Herrmann JM. 2007. Translocation of proteins into mitochondria. Annu. Rev. Biochem. 76:723–749.

Ngosuwan J, Wang NM, Fung KL, Chirico WJ. 2003. Roles of cytosolic Hsp70 and Hsp40 molecular chaperones in post-translational translocation of presecretory proteins into the endoplasmic reticulum. J. Biol. Chem. 278:7034–7042.

Nguyen LT, Schmidt HA, Von Haeseler A, Minh BQ. 2015. IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol. Biol. Evol. 32:268–274.

Nielsen H, Tsirigos KD, Brunak S, von Heijne G. 2019. A Brief History of Protein Sorting Prediction. Protein J. 38:200–216.

Niklaus M, Kelly S. 2019. The molecular evolution of C 4 photosynthesis: Opportunities for understanding and improving the world's most productive plants. J. Exp. Bot. 70:795–804.

Noinaj N, Guillier M, Barnard TJ, Buchanan SK. 2010. TonB-dependent transporters: Regulation, structure, and function. Annu. Rev. Microbiol. 64:43–60.

Nordlie RC, Lardy HA. 1963. Mammalian liver phosphoneolpyruvate carboxykinase activities. J. Biol. Chem. 238:2259–2263.

Novick A, Doolittle WF. 2020. Horizontal persistence and the complexity hypothesis. Biol. Philos. 35:1–22.

Oatey PB, Lumb MJ, Danpure CJ. 1996. Molecular basis of the variable mitochondrial and peroxisomal localisation of alanine-glyoxylate aminotransferase. Eur. J. Biochem. 241:374–385.

Ohno S. 1970. Evolution by Gene Duplication. 1st ed. Springer, Berlin, Heidelberg

Palmer JD, Adams KL, Cho Y, Parkinson CL, Qiu YL, Song K. 2000. Dynamic evolution of plant mitochondrial genomes: mobile genes and introns and highly variable mutation rates. Proc. Natl. Acad. Sci. U. S. A. 97:6960–6966.

Panchy N, Lehti-Shiu MD, Shiu S-H. 2016. Evolution of gene duplication in plants. Plant Physiol. 171:2294–2316.

Paradis E, Claude J, Strimmer K. 2004. APE: Analyses of phylogenetics and evolution in R language. Bioinformatics 20:289–290.

Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C. 2017. Salmon provides fast and bias-aware quantification of transcript expression. Nat. Methods 14:417–419.

Pengelly JJL, Kwasny S, Bala S, Evans JR, Voznesenskaya E V., Koteyeva NK, Edwards GE, Furbank RT, von Caemmerer S. 2011. Functional analysis of corn husk photosynthesis. Plant Physiol. 156:503–513.

Petricka JJ, Clay NK, Nelson TM. 2008. Vein patterning screens and the defectively organized tributaries mutants in Arabidopsis thaliana. Plant J. 56:251–263.

Pick TR, Bräutigam A, Schlüter U, Denton AK, Colmsee C, Scholz U, Fahnenstich H, Pieruschka R, Rascher U, Sonnewald U, et al. 2011. Systems analysis of a maize leaf developmental gradient redefines the current C4 model and provides candidates for regulation. Plant Cell 23:4208–4220.

Ponnala L, Wang Y, Sun Q, Van Wijk KJ. 2014. Correlation of mRNA and protein abundance in the developing maize leaf. Plant J. 78:424–440.

Price MN, Dehal PS, Arkin AP. 2010. FastTree 2 - Approximately maximum-likelihood trees for large alignments. PLoS One 5:e9490.

Puigbò P, Guzmán E, Romeu A, Garcia-Vallvé S. 2007. OPTIMIZER: A web server for optimizing the codon usage of DNA sequences. Nucleic Acids Res. 35:W126–W131.

Qian W, Zhang J. 2009. Protein Subcellular Relocalization in the Evolution of Yeast Singleton and Duplicate Genes. Genome Biol. Evol. 1:198–204.

Ralph SA, van Dooren GG, Waller RF, Crawford MJ, Fraunholz MJ, Foth BJ, Tonkin CJ, Roos DS, McFadden GI. 2004. Metabolic maps and functions of the Plasmodium falciparum apicoplast. Nat. Rev. Microbiol. 2:203–216.

Ramšak Ž, Baebler Š, Rotter A, Korbar M, Mozetič I, Usadel B, Gruden K. 2014. GoMapMan: Integration, consolidation and visualization of plant gene annotations within the MapMan

ontology. Nucleic Acids Res. 42:D1167–D1175.

Rao X, Dixon RA. 2016. The differences between NAD-ME and NADP-ME subtypes of C4 photosynthesis: More than decarboxylating enzymes. Front. Plant Sci. 7.

Rao X, Lu N, Li G, Nakashima J, Tang Y, Dixon RA. 2016. Comparative cell-specific transcriptomics reveals differentiation of C4 photosynthesis pathways in switchgrass and other C4 lineages. J. Exp. Bot. 67:1649–1662.

Ren L-L, Liu Y-J, Liu H-J, Qian T-T, Qi L-W, Wang X-R, Zeng Q-Y. 2014. Subcellular relocalization and positive selection play key roles in the retention of duplicate genes of Populus class III peroxidase family. Plant Cell 26:2404–2419.

Ren R, Wang H, Guo C, Zhang N, Zeng L, Chen Y, Ma H, Qi J. 2018. Widespread Whole Genome Duplications Contribute to Genome Complexity and Species Diversity in Angiosperms. Mol. Plant 11:414–428.

Ren Y, Li Y, Jiang Y, Wu B, Miao Y. 2017. Phosphorylation of WHIRLY1 by CIPK14 Shifts Its Localization and Dual Functions in Arabidopsis. Mol. Plant 10:749–763.

Reumann S. 2004. Specification of the peroxisome targeting signals type 1 and type 2 of plant peroxisomes by bioinformatics analyses. Plant Physiol. 135:783–800.

Reyes-Prieto A, Weber APM, Bhattacharya D. 2007. The origin and establishment of the plastid in algae and plants. Annu. Rev. Genet. 41:147–168.

Reyna-Llorens I, Hibberd JM. 2017. Recruitment of pre-existing networks during the evolution of C4 photosynthesis. Philos. Trans. R. Soc. B Biol. Sci. 372:2–7.

Richly E, Leister D. 2004. An improved prediction of chloroplast proteins reveals diversities and commonalities in the chloroplast proteomes of Arabidopsis and rice. Gene 329:11–16.

Rodo AP, Brugière N, Vankova R, Malbeck J, Olson JM, Haines SC, Martin RC, Habben JE, Mok DWS, Mok MC. 2008. Over-expression of a zeatin O-glucosylation gene in maize leads to growth retardation and tasselseed formation. J. Exp. Bot. 59:2673–2686.

Roepke J, Bozzo GG. 2015. Arabidopsis thaliana β-glucosidase BGLU15 attacks flavonol 3-O-β-glucoside-7-O-α-rhamnosides. Phytochemistry 109:14–24.

Roepke J, Gordon HOW, Neil KJA, Gidda S, Mullen RT, Freixas Coutin JA, Bray-Stone D, Bozzo GG. 2017. An apoplastic β-glucosidase is essential for the degradation of flavonol 3-O-β-Glucoside-7-O-α-rhamnosides in arabidopsis. Plant Cell Physiol. 58:1030–1047.

Roger AJ, Muñoz-Gómez SA, Kamikawa R. 2017. The Origin and Diversification of Mitochondria. Curr. Biol. 27:R1177–R1192.

Rossig C, Reinbothe C, Gray J, Valdes O, Von Wettstein D, Reinbothe S. 2013. Three proteins mediate import of transit sequence-less precursors into the inner envelope of chloroplasts in arabidopsis thaliana. Proc. Natl. Acad. Sci. U. S. A. 110:19962–19967.

Rossini L, Cribb L, Martin DJ, Langdale JA. 2001. The maize Golden2 gene defines a novel class of transcriptional regulators in plants. Plant Cell 13:1231–1244.

Sage RF. 2004. The evolution of C4 photosynthesis. New Phytol. 161:341–370.

Sage RF. 2016. A portrait of the C4 photosynthetic family on the 50th anniversary of its discovery: Species number, evolutionary lineages, and Hall of Fame. J. Exp. Bot. 67:4039–4056.

Salichos L, Rokas A. 2013. Inferring ancient divergences requires genes with strong phylogenetic signals. Nature 497:327–331.

Sato S, Tabata S, Hirakawa H, Asamizu E, Shirasawa K, Isobe S, Kaneko T, Nakamura Y, Shibata D, Aoki K, et al. 2012. The tomato genome sequence provides insights into fleshy fruit evolution. Nature 485:635–641.

Schaeffer S, Harper A, Raja R, Jaiswal P, Dhingra A. 2014. Comparative analysis of predicted plastid-targeted proteomes of sequenced higher plant genomes.Theg SM, editor. PLoS One 9:e112870.

Schatz G, Dobberstein B. 1996. Common Principles of Protein Translocation Across Membranes. Science (80-.). 271:1519–1526.

Schindelin J, Arganda-Carreras I, Frise E, Kaynig V, Longair M, Pietzsch T, Preibisch S, Rueden C, Saalfeld S, Schmid B, et al. 2012. Fiji: An open-source platform for biological-image analysis. Nat. Methods 9:676–682.

Schlüter U, Bräutigam A, Droz JM, Schwender J, Weber APM. 2019. The role of alanine and

aspartate aminotransferases in C4 photosynthesis. Plant Biol. 21:64–76.

Schmutz J, McClean PE, Mamidi S, Wu GA, Cannon SB, Grimwood J, Jenkins J, Shu S, Song Q, Chavarro C, et al. 2014. A reference genome for common bean and genome-wide analysis of dual domestications. Nat. Genet. 46:707–713.

Schnarrenberger C, Herbert M, Kruger I. 1983. Intracellular compartmentation of isozymes of sugar phosphate metabolism in green leaves. Isozymes Curr Top Biol Med Res 8:23–51.

Sedelnikova O V., Hughes TE, Langdale JA. 2018. Understanding the genetic basis of c4 Kranz anatomy with a view to engineering c3 crops. Annu. Rev. Genet. 52:249–70.

Seoighe C, Gehring C. 2004. Genome duplication led to highly selective expansion of the Arabidopsis thaliana proteome. Trends Genet. 20:461–464.

Sharma M, Bennewitz B, Klösgen R. 2018. Rather rule than exception? How to evaluate the relevance of dual protein targeting to mitochondria and chloroplasts. Photosynth Res 138:335–343.

Sharma M, Pandey GK. 2016. Expansion and function of repeat domain proteins during stress and development in plants. Front. Plant Sci. 6:1–15.

Sharwood RE, Sonawane B V., Ghannoum O. 2014. Photosynthetic flexibility in maize exposed to salinity and shade. J. Exp. Bot. 65:3715–3724.

Shen B-R, Zhu C-H, Yao Z, Cui L-L, Zhang J-J, Yang C-W. 2017. An optimized transit peptide for effective targeting of diverse foreign proteins into chloroplasts in. Nat. Publ. Gr. 7:1–12.

Siddiq MA, Hochberg GK, Thornton JW. 2017. Evolution of protein specificity: insights from ancestral protein reconstruction. Curr. Opin. Struct. Biol. 47:113–122.

Sjöling S, Glaser E. 1998. Mitochondrial targeting peptides in plants. Trends Plant Sci. 3:136–140.

Skalák J, Vercruyssen L, Claeys H, Hradilová J, Černý M, Novák O, Plačková L, Saiz-Fernández I, Skaláková P, Coppens F, et al. 2019. Multifaceted activity of cytokinin in leaf development shapes its size and structure in Arabidopsis. Plant J. 97:805–824.

De Smet R, Adams KL, Vandepoele K, Van Montagu MCE, Maere S, Van de Peer Y. 2013. Convergent gene loss following gene and genome duplications creates single-copy families in flowering plants. Proc. Natl. Acad. Sci. 110:2898 –2903.

Smith DR, Keeling PJ. 2015. Mitochondrial and plastid genome architecture: Reoccurring themes, but significant differences at the extremes. Proc. Natl. Acad. Sci. U. S. A. 112:10177–10184.

Smolskaya S, Logashina YA, Andreev YA. 2020. Escherichia coli extract-based cell-free expression system as an alternative for difficult-to-obtain protein biosynthesis. Int. J. Mol. Sci. 21.

Soltis PS, Marchant DB, Van de Peer Y, Soltis DE. 2015. Polyploidy and genome evolution in plants. Curr. Opin. Genet. Dev. 35:119–125.

Sommer M, Bräutigam A, Weber APM. 2012. The dicotyledonous NAD malic enzyme C 4 plant Cleome gynandra displays age-dependent plasticity of C 4 decarboxylation biochemistry. Plant Biol. 14:621–629.

Sperschneider J, Catanzariti A-M, DeBoer K, Petre B, Gardiner DM, Singh KB, Dodds PN, Taylor JM. 2017. LOCALIZER: subcellular localization prediction of both plant and effector proteins in the plant cell. Sci. Rep. 7:44598.

Stevens PF. 2001. Angiosperm Phylogeny Website. Available from: http://www.mobot.org/MOBOT/research/APweb/

Sun Q, Zybailov B, Majeran W, Friso G, Olinares PDB, van Wijk KJ. 2009. PPDB, the plant proteomics database at Cornell. Nucleic Acids Res. 37:D969-74.

Swenson KM, Doroftei A, El-Mabrouk N. 2012. Gene tree correction for reconciliation and species tree inference. Algorithms Mol. Biol.

Swigoňová Z, Lai J, Ma J, Ramakrishna W, Llaca V, Bennetzen JL, Messing J. 2004. Close split of sorghum and maize genome progenitors. Genome Res. 14:1916–1923.

Synthego. 2019. Synthego Performance Analysis, ICE Analysis v2.0.

Szklarczyk D, Gable AL, Lyon D, Junge A, Wyder S, Huerta-Cepas J, Simonovic M, Doncheva NT, Morris JH, Bork P, et al. 2019. STRING v11: Protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. Nucleic Acids Res. 47:D607–D613.

Szklarczyk R, Huynen MA. 2010. Mosaic origin of the mitochondrial proteome. Proteomics 10:4012–4024.

Szllosi GJ, Tannier E, Daubin V, Boussau B. 2015. The inference of gene trees with species trees. Syst. Biol. 64:e42–e62.

Takabayashi A, Kishine M, Asada K, Endo T, Sato F. 2005. Differential use of two cyclic flows around photosystem I for driving CO2-concentration mechanism in C4 photosynthesis. Proc. Natl. Acad. Sci. U. S. A. 102:16898–16903.

Taniguchi M, Kobe A, Kato M, Sugiyama T. 1995. Aspartate aminotransferase isozymes in Panicum miliaceum L., an NAD-malic enzyme-type C4 plant: Comparison of enzymatic properties, primary structures, and expression patterns. Arch. Biochem. Biophys. 318:295–306.

Taniguchi M, Sugiyama T. 1990. Aspartate aminotransferase from Eleusine coracana, a C4 plant: Purification, characterization, and preparation of antibody. Arch. Biochem. Biophys. 282:427–432.

Tanz SK, Tetu SG, Vella NGF, Ludwig M. 2009. Loss of the Transit Peptide and an Increase in Gene Expression of an Ancestral Chloroplastic Carbonic Anhydrase Were Instrumental in the Evolution of the Cytosolic C4 Carbonic Anhydrase in Flaveria. Plant Physiol. 150:1515–1529.

Tardif M, Atteia A, Specht M, Cogne G, Rolland N, Brugière S, Hippler M, Ferro M, Bruley C, Peltier G, et al. 2012. Predalgo: A new subcellular localization prediction tool dedicated to green algae. In: Molecular Biology and Evolution. Vol. 29. p. 3625–3639.

Tasdighian S, Van Bel M, Li Z, Van de Peer Y, Carretero-Paulet L, Maere S. 2017. Reciprocally retained genes in the Angiosperm lineage show the hallmarks of dosage balance sensitivity. Plant Cell 29:2766–2785.

Tausta LS, Li P, Si Y, Gandotra N, Liu P, Sun Q, Brutnell TP, Nelson T. 2014. Developmental dynamics of Kranz cell transcriptional specificity in maize leaf reveals early onset of C4-related processes. J. Exp. Bot. 65:3543-3555.

Taylor JS, Raes J. 2004. Duplication and divergence: The evolution of new genes and old ideas. Annu. Rev. Genet. 38:615–643.

Thomson SM, Pulido P, Jarvis RP. 2020. Protein import into chloroplasts and its regulation by the ubiquitin-proteasome system. Biochem. Soc. Trans. 48:71–82.

Thul PJ, Akesson L, Wiking M, Mahdessian D, Geladaki A, Ait Blal H, Alm T, Asplund A, Björk L, Breckels LM, et al. 2017. A subcellular map of the human proteome. Science (80-.). 356.

Tielens AGM, Rotte C, Van Hellemond JJ, Martin W. 2002. Mitochondria as we don't know them. Trends Biochem. Sci. 27:564–572.

Timmis JN, Ayliffe MA, Huang CY, Martin W. 2004. Endosymbiotic gene transfer: organelle genomes forge eukaryotic chromosomes. Nat. Rev. Genet. 5:123–135.

Todd J, Screen S, Crowley J, Peng J, Andersen S, Brown T, Qi Q, Fabbri B, Duff SMG. 2008. Identification and characterization of four distinct asparagine synthetase (AsnS) genes in maize (Zea mays L.). Plant Sci. 175:799–808.

Vanneste K, Baele G, Maere S, Van de Peer Y. 2014. Analysis of 41 plant genomes supports a wave of successful genome duplications in association with the Cretaceous-Paleogene boundary. Genome Res. 24:1334–1347.

Vogel JP, Garvin DF, Mockler TC, Schmutz J, Rokhsar D, Bevan MW, Barry K, Lucas S, Harmon-Smith M, Lail K, et al. 2010. Genome sequencing and analysis of the model grass Brachypodium distachyon. Nature 463:763–768.

Walker BJ, Vanloocke A, Bernacchi CJ, Ort DR. 2016. The costs of photorespiration to food production now and in the future. Annu. Rev. Plant Biol. 67:107–129.

Walker RP, Acheson RM, Técsi LI, Leegood RC. 1997. Phosphoenolpyruvate carboxykinase in C4 plants: Its role and regulation. Aust. J. Plant Physiol. 24:459–468.

Wang P, Fouracre J, Kelly S, Karki S, Gowik U, Aubry S, Shaw MK, Westhoff P, Slamet-Loedin IH, Quick WP, et al. 2013. Evolution of GOLDEN2-LIKE gene function in C3 and C4 plants. Planta 237:481–495.

Wang P, Hendron RW, Kelly S. 2017. Transcriptional control of photosynthetic capacity: conservation and divergence from Arabidopsis to rice. New Phytol. 216:32–45.

Wang P, Kelly S, Fouracre JP, Langdale JA. 2013. Genome-wide transcript analysis of early maize leaf development reveals gene cohorts associated with the differentiation of C4 Kranz

anatomy. Plant J. 75:656–670.

Wang P, Khoshravesh R, Karki S, Tapia R, Balahadia CP, Bandyopadhyay A, Quick WP, Furbank R, Sage TL, Langdale JA. 2017. Re-creation of a Key Step in the Evolutionary Switch from C3 to C4 Leaf Anatomy. Curr. Biol. 27:3278-3287.e6.

Wang X, Hou J, Quedenau C, Chen W. 2016. Pervasive isoform-specific translational regulation via alternative transcription start sites in mammals. Mol. Syst. Biol. 12:875.

Wang X, Huang Y, Lavrov D V., Gu X. 2009. Comparative study of human mitochondrial proteome reveals extensive protein subcellular relocalization after gene duplications. BMC Evol. Biol. 9:1–11.

Wang Y, Bräutigam A, Weber APM, Zhu XG. 2014. Three distinct biochemical subtypes of C4 photosynthesis? A modelling analysis. J. Exp. Bot. 65:3567–3578.

Waterhouse AM, Procter JB, Martin DMA, Clamp M, Barton GJ. 2009. Jalview Version 2-A multiple sequence alignment editor and analysis workbench. Bioinformatics 25:1189–1191.

Weber APM, von Caemmerer S. 2010. Plastid transport and metabolism of C3 and C4 plants-comparative analysis and possible biotechnological exploitation. Curr. Opin. Plant Biol. 13:256–264.

Weber E, Engler C, Gruetzner R, Werner S, Marillonnet S. 2011. A modular cloning system for standardized assembly of multigene constructs.Peccoud J, editor. PLoS One 6:e16765.

Weissmann S, Ma F, Furuyama K, Gierse J, Berg H, Shao Y, Taniguchi M, Allen DK, Brutnell TP. 2015. Interactions of C4 Subtype Metabolic Activities and Transport in Maize Are Revealed through the Characterization of DCT2 Mutants. Plant Cell 28:466–484.

Wendel JF, Jackson SA, Meyers BC, Wing RA. 2016. Evolution of plant genome architecture. Genome Biol. 17.

Wendel JF, Lisch D, Hu G, Mason AS. 2018. The long and short of doubling down: polyploidy, epigenetics, and the temporal dynamics of genome fractionation. Curr. Opin. Genet. Dev. 49:1–7.

van Wijk KJ, Baginsky S. 2011. Plastid proteomics in higher plants: Current state and future goals. Plant Physiol. 155:1578–1588.

Williams EJB, Pal C, Hurst LD. 2000. The molecular evolution of signal peptides. Gene 253:313–322.

Wingler A, Walker RP, Chen ZH, Leegood RC. 1999. Phosphoenolpyruvate carboxykinase is involved in the decarboxylation of aspartate in the bundle sheath of maize. Plant Physiol. 120:539–546.

Wise R. 2006. The Diversity of Plastid Form and Function. In: The structure and function of plastids. Vol. 23. p. 3–26.

Woo HR, Koo HJ, Kim JJHJ, Jeong H, Yang JO, Lee IH, Jun JH, Choi SH, Park SSJ, Kang B, et al. 2016. Programming of plant leaf senescence with temporal and inter-organellar coordination of transcriptome in arabidopsis. Plant Physiol. 171:452–467.

Wu FH, Shen SC, Lee LY, Lee SH, Chan MT, Lin CS. 2009. Tape-arabidopsis sandwich - A simpler arabidopsis protoplast isolation method. Plant Methods 5.

Xiang Y, Huang CH, Hu Y, Wen J, Li S, Yi T, Chen H, Xiang J, Ma H. 2017. Evolution of Rosaceae Fruit Types Based on Nuclear Phylogeny in the Context of Geological Times and Genome Duplication. Mol. Biol. Evol. 34:262–281.

Xin-Guang Zhu, Stephen P Long, Donald R Ort. 2008. What is the maximum efficiency with which photosynthesis can convert solar energy into biomass? Curr. Opin. Biotechnol. 19:153–159.

Xu D, Farmer A, Collett G, Grishin N V., Chook YM. 2012. Sequence and structural analyses of nuclear export signals in the NESdb database. Mol. Biol. Cell 23:3677–3693.

Xu H, Curtis TY, Powers SJ, Raffan S, Gao R, Huang J, Heiner M, Gilbert DR, Halford NG. 2018. Genomic, biochemical, and modeling analyses of asparagine synthetases from wheat. Front. Plant Sci. 8:1–12.

Xu L, Carrie C, Law SR, Murcha MW, Whelan J. 2013. Acquisition, conservation, and loss of dual-targeted proteins in land plants. Plant Physiol. 161:644–662.

Yamagishi R, Okuyama T, Oba S, Shimada J, Chaen S, Kaneko H. 2015. Comprehensive analysis of the dynamic structure of nuclear localization signals. Biochem. Biophys. Reports 4:392–

396.

Yan L, Wei S, Wu Y, Hu R, Li H, Yang W, Xie Q. 2015. High-Efficiency Genome Editing in Arabidopsis Using YAO Promoter-Driven CRISPR/Cas9 System. Mol. Plant 8:1820–1823.

Yeates TO, Thompson MC, Bobik TA. 2011. The protein shells of bacterial microcompartment organelles. Curr. Opin. Struct. Biol. 21:223–231.

Zhang J. 2003. Evolution by gene duplication: An update. Trends Ecol. Evol. 18:292–298.