Evolution guided engineering to improve C_4 photosynthesis



Michael Niklaus

Lincoln College

Department of Plant Sciences

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Trinity 2019

Abstract

Anticipated future population growth necessitates that more food is produced using the same, or less, area of land than we use today. One potential way to achieve this is through the improvement of photosynthesis in the crop plants that feed the world. Recently, there has been several successful attempts to improve photosynthesis in C_3 crop plants. However, not all crop plants conduct C_3 photosynthesis, and comparatively little effort has been focussed on improving alternative photosynthetic pathways such as C_4 photosynthesis and CAM. Correspondingly, this DPhil project had two parallel aims. The first was to develop a novel approach to improve photosynthesis in C_4 crop plants. The premise of the approach is that there is substantial natural variation in the catalytic properties of enzymes of the C_4 cycle between different species, and if one could understand the molecular basis for this variation then one could engineer enhanced flux through the C_4 cycle through simple sequence changes. The novel developed in this thesis is divided into three stages. 1) Identification of target sites for enzyme engineering though comparative genomics. 2) Use bacterial systems to assess the impact of variation at these target sites on enzyme function. 3) Engineer the changes required to produce the best performing variant directly in crop plants using genome editing technologies. I achieved step one and came close to achieving step two in during the course of this project. The second overarching aim was to identify the dicarboxylate and monocarboxylate transporters that facilitate malate import and pyruvate export from the bundle sheath cell chloroplast in C_4 plants. The lack of knowledge about the identity of these transporters represents the largest gap in our understanding of C_4 photosynthesis and hinders efforts to engineer the C_4 pathway into C_3 plants. This part of the DPhil project resulted in the identification of a single gene which provided both the malate import and pyruvate export functions of the bundle sheath cell chloroplast in the C_4 plant *Setaria italica,* thus completing the first biochemical map for a C_4 cycle in any species. Together, the results from this thesis provide a novel platform for potentially enhancing C_4 photosynthesis, and for the first time define a complete parts list required to engineer a complete C_4 pathway into C_3 plants.

Acknowledgements

Doing research is challenging, difficult, and frustrating, but it can also be extremely rewarding. Looking back on my DPhil I was lucky enough to have experienced some amazing rewarding moments. I would like to first thank the person who made that all possible, my supervisor Dr S. Kelly, whose expertise was invaluable in formulating the idea behind harnessing the natural variation in C_4 species to identify potential engineering targets. In addition, the support and patience I have received and the countless scientific discussions, which I thoroughly enjoyed, were invaluable. It was also great to see first-hand how a nicely formulated idea was developed into a side project and eventually lead to a great discovery. Thank you for making my journey through my DPhil such a positive experience.

I would like to extend my thanks to my second supervisor, Prof J. Langdale, whose expertise and challenging questioning throughout the DPhil project lead to careful reflections of the planned experiments and therefore to substantial improvements. It made me aware of the importance of critical examination.

I would like to acknowledge all my colleagues from the Kelly & Langdale Lab for their support and collaboration throughout my DPhil. I would particularly like to thank Dana Vlad for all the tips and tricks regarding molecular cloning and for always being available for helpful discussions. Thanks go to David Emms for all the helpful discussions regarding algorithm development and bioinformatics issues, to Ursula Flores-Perez for introducing me into the world of plant protoplasts and all the suggestions and help regarding protein expression. I would also like to thank Julie, Lizzie and Rob, who made working in the laboratory so much easier and who pulled the strings in the background.

I would like to thank my parents Andrea and Hans-Jakob, who are always there to support me and make sure I do not lose touch with reality.

Finally, I would like to give special thanks my wife Riet, who left family and friends behind in Belgium to join me in Oxford and started from scratch. I am proud of you.

Table of contents

Abstract……....i

Acknowledgements……ii

Abbreviations……..xiii

Chapter 1: General Introduction 1

1.1 Introduction into C_4 photosynthesis 1

1.2 C_4 photosynthesis; is age more than a number? 2

1.3 C_4 photosynthesis; equal but not the same 6

1.4 Improving C_4 function by simple molecular sequence changes 11

1.5 Improving C_4 function through gene duplication 14

1.6 Improving C_4 function through recruitment of cis-regulatory networks 16

1.7 Improving C_4 function through subcellular retargeting of proteins and lateral transfer of genetic material 18

1.8 Challenges and opportunities for engineering C_4 photosynthesis 19

1.9 Thesis aims and overview 24

Chapter 2: A novel method for identification of significant differences in aligned amino acid sequences between two sets of species 26

2.1 Chapter introduction 26

2.1.1 Background 26

2.2 Results and discussion 31

2.2.1 The total variation distance metric 31

2.2.2 Definition of groups and subgroups 33

2.2.3 Algorithm overview 33

2.2.4 Statistical assessment of the analysis pipeline to measure the amino acid distribution in a multiple sequence alignment between two groups 35

2.2.5 Comparison between known sites under selection in C_4 PEPC and sites identified by the Total Variation Distance algorithm 43

2.2.6 Extension of the analysis pipeline to consider transcript abundance and dealing with duplicated genes 49

2.3 Chapter conclusion 51

2.3.1 Limitations and potential extension of the Total Variation Distance algorithm 53

2.3.2 Summary 55

2.4 Material and Methods 57

2.4.1 Dataset simulation 57

2.4.2 Analysis pipeline software specification 57

2.4.3 Application of the Total Variation Distance algorithm on the Christin et al. 2007 dataset and output processing 58

Chapter 3: Application of the Total Variation Distance algorithm to a dataset containing C_3 and C_4 plants to identify targets to improve protein function for C_4 photosynthesis 59

3.1 Chapter introduction 59

3.1.1 Background 59

3.1.2 Example rate limiting stages of the C_4 cycle 62

3.1.3 Exploiting natural variation in C_4 plants to identify potential engineering targets to improve photosynthetic efficiency 65

3.1.4 Chapter aims and outline 65

3.2 Results and discussion 66

3.2.1 Processing of the transcriptome dataset and inferring orthogroups 66

3.2.2 Application of the Total Variation Distance algorithm on a set of proteins that are associated with the C_4 cycle 69

3.2.3 Enzymes of the core C_4 cycle with altered amino acid distributions 71

3.2.4 Transporters of the C_4 cycle with altered amino acid distributions 87

3.2.5 Enzymes of the photorespiration cycle 96

3.3 Chapter conclusion 103

3.3.1 Summary 105

3.4 Materials and Methods 106

3.4.1 Dataset 106

3.4.2 Plant growth 108

3.4.3 Isolation of RNA 109

3.4.4 *De novo* assembly of short reads 109

3.4.5 List of genes closely associated to the C_4 cycle 110

3.4.6 Initial processing of *de novo* assembled transcriptomes 113

3.4.7 Generation of multiple sequence alignments 114

3.4.8 Splitting fused orthogroups 114

3.4.9 Constructing the species tree 115

3.4.10 Filtering the output from the Total Variation Distance algorithm 115

Chapter 4: An *in vivo* bacterial system to evaluate the relative function of different variants of proteins involved in C_4 photosynthesis 116

4.1 Chapter introduction 116

4.1.1 Background 117

4.1.2 Chapter aims and outline 125

4.2 Results and discussion 125

4.2.1 Generating *E. coli* BL21 (DE3) Δ*ppc* using recombineering 125

4.2.2 Evaluating whether the growth rate is dependent on the function of PEPC 128

4.2.3 Evaluating whether the growth rate is dependent on the function of PPDK in *E. coli Δ*pps mutants grown in media with pyruvate as sole carbon source 130

4.2.4 Generation of a variant library using site-directed mutagenesis 135

4.3 Chapter conclusion 139

4.3.1 Future work 141

4.3.2 Summary 143

4.4 Material and Methods 145

4.4.1 Electrocompetent cells 145

4.4.2 Genome engineering using recombineering 146

4.4.3 Genotyping of the Keio strain JW1692 151

4.4.4 Whole genome sequencing and read mapping of *E. coli* mutant strains 152

4.4.5 Sequence modifications 153

4.4.6 Golden Gate assembling 154

4.4.7 Growth curve of *E. coli* in M9 with a single carbon source 156

4.4.8 Growth curve assessment 157

4.4.9 Keio Collection strains 158

4.4.10 Sequence variant synthesis using site-directed mutagenesis 158

Chapter 5: Identification and characterisation of the bundle sheath cell chloroplast monocarboxylate and dicarboxylate transporter of *Setaria italica* 162

5.1 Chapter introduction 162

5.2 Background 162

5.2.1 Metabolite transport in NADP-ME C_4 plants 162

5.2.2 Further consideration of putative malate and pyruvate transporters 167

5.2.3 The use of bacteria to characterise chloroplast malate and pyruvate transporters 170

5.2.4 Chapter aims and outlines 171

5.3 Results and discussion 172

5.3.1 Identification of a putative monocarboxylate and dicarboxylate transporter using cell-type specific expression data 172

5.3.2 Expression of pMDT1 in *E. coli* results in accumulation of pyruvate in the cell culture medium 175

5.3.3 pMDT1 facilitates uptake of malate into *E. coli* cells. 179

5.3.4 The putative pyruvate:malate transporter from *Setaria italica* localised to chloroplast envelope when expressed in *Setaria viridis* mesophyll protoplasts 184

5.3.5 Downregulation of the endogenous *Setaria viridis* pMDT1 reduces CO_2 assimilation 187

5.4 Chapter conclusion 200

5.5 Material and Methods 204

5.5.1 Plant growth conditions 204

5.5.2 Generating transgenic *Setaria viridis* ME034V 204

5.5.3 DNA extraction and genotyping of transgenic plants 205

5.5.4 Sequence modifications 206

5.5.5 Gateway cloning to create pMDT1 RNAi constructs 207

5.5.6 Expression of pMDT1 in *E. coli* grown in minimal media with glucose as sole carbon source 209

5.5.7 Pyruvate measurement in cell culture supernatant 209

5.5.8 Metabolite extraction from cell culture supernatant 210

5.5.9 Experimental evaluation for the malate uptake of pMDT1 in *Escherichia coli* Δ*dctA* 210

5.5.10 *Setaria viridis* protoplast generation 211

5.5.11 Confocal analysis of transformed protoplasts 212

5.5.12 Keio *Escherichia coli* strains 213

5.5.13 RT-qPCR to verify the relative transcript abundance of pMDT1 in *Setaria viridis*..... 213

5.5.14 Total RNA extraction from *Setaria viridis* leaf tissue 214

5.5.15 RNAseq data analysis of the *Setaria viridis*..... 215

5.5.16 Measuring carbon assimilation in *Setaria viridis*..... 216

5.5.17 Measuring light response in *Setaria viridis* 217

Chapter 6: General discussion 218

6.1 Exploiting natural variation in C_4 plants to identify function determining sites in proteins 218

6.2 Development of a novel data analysis algorithm to identify sites in a protein alignment that have a different amino acid distribution between two sets of species..... 219

6.3 The identification of function determining sites using a similar approach as the Total Variation Distance analysis 220

6.4 Is it possible to improve the biochemical properties of C_4 proteins? 221

6.5 Alternative application and availability of the Total Variation Distance algorithm 223

6.6 Expression of plant transport proteins in *Escherichia coli* and potential biotechnological applications thereof 224

6.7 The identification and characterisation of the malate importer and pyruvate exporter in *Setaria italica*..... 225

6.8 C_4 photosynthesis: similar but different 227

Table of contents

Appendix A 229

A.1 Total Variation Distance additional information 229

A.2 Media 230

A.2.1 M9 mineral medium 230

A.3 Golden Gate cloning 231

A.3.1 *Setaria viridis* 4G143500 sequence (5'-3') optimized for expression in *E. coli* and golden gate cloning (SC module) 231

A.3.2 *Setaria viridis* 3G253900 PPDK sequence (5'-3') optimised for expression in *E. coli* and golden gate cloning (SC module) 232

A.3.3 *Setaria italica* Seita.4G275500 sequence (5'-3') optimized for expression in *E. coli* and golden gate cloning (SC module) 233

A.3.4 *Setaria italica* Seita.4G275500 sequence (5'-3') optimized for *C*-terminal tagging and golden gate cloning (SC1 module) 233

A.3.5 *Arabidopsis thaliana* AT4G19390 sequence (5'-3') optimized for expression in *E. coli* and golden gate cloning (SC module) 233

A.3.6 *Arabidopsis thaliana* AT4G19390 sequence (5'-3') optimized for *C*-terminal tagging and golden gate cloning (SC1 module) 234

A.3.7 Golden gate parts 234

A.3.8 PEPC construct 1 235

A.3.9 PEPC construct 2 236

A.3.10 PPDK construct 1 237

A.3.11 PPDK construct 2 238

A.3.12 Level1 golden gate construct for the expression of pMDT1 GPP tagged and cytosolic GFP in protoplast 240

A.3.13 Level1 golden gate construct to induced the expression of pMDT1 in *Escherichia coli* 241

A.4 Primers for genotyping of the *E. coli* mutants 242

A.5 Primer for Sanger sequencing of the Gateway cloning plasmid pANIC 12A 243

A.6 Primers for Sanger sequencing of the plasmids used for expression in protoplasts 244

A.7 Oligonucleotides for site directed mutagenesis and sequencing of the variant library 244

A.8 RNAi plant and segregating wild type plants 247

A.9 List of differentially expressed genes between RNAi and segregating wild type plants 248

Appendix B Fragmatic: Automatic identification and resolution of erroneously fragmented genes in *de novo* assembled transcriptomes 252

B.1 Abstract 252

B.2 Introduction 253

B.3 Results 256

B.3.1 Problem definition and approach 256

B.3.2 Simulated fragmented sequences 259

B.3.3 Application of Fragmatic to simulated fragmented sequences 260

B.3.4 Application of Fragmatic to real *de novo* assembled transcriptomes 261

B.4 Discussion 266

B.5 Methods 269

B.5.1 Algorithm overview 269

B.5.2 Simulation of fragmented sequences 271

B.5.3 Prediction of open reading frames in *de novo* assembled transcriptomes 271

B.5.4 Inference of orthogroups using OrthoFinder 272

B.5.5 Evaluation of simulated fragmented sequences 272

B.5.6 Evaluation of real transcriptome sequences 272

B.6 Software availability 273

References 274

Abbreviations

2-PG	2-phosphoglycolate
3-PGA	3-phosphoglycerate
ADP	Adenosine diphosphate
AMP	Adenosine monophosphate
ATP	Adenosine triphosphate
AU	Arbitrary unit
BASS2	chloroplastic bile acid:sodium symporter family protein 2
BS cells	bundle sheath cells
CA	carbonic anhydrase
CAM	Crassulacean acid metabolism
cfu	colony forming unit
dN	non-synonymous nucleotide substitutions per non-synonymous site
dNTPs	deoxynucleotide triphosphate
dS	synonymous nucleotide substitutions per synonymous site
GUS	beta-glucuronidase reporter system
HP	Hydroxypyruvate
kDa	kilo Dalton
M cell	mesophyll cell
NAD	Nicotinamide adenine dinucleotide
NAD-ME	NAD-dependent malic enzyme
NADP	Nicotinamide adenine dinucleotide phosphate
NADP-ME	NADP-dependent malic enzyme
PCR cycle	photosynthetic carbon reduction cycle
PEP	Phosphoenolpyruvate
PEPC	PEP carboxylase
PEPCK	PEP carboxykinase
PEPC-PK	PEPC protein kinase

PPC	PEP carboxylase
PPDK	pyruvate phosphate dikinase
PPS	PEP synthase
PS II	photosystem II
SBPase	sedoheptulose 1,7-bisphosphatase
SD	Standard deviation
Seita	*Setaria italica*
Svir	*Setaria viridis*
TCA cycle	Tricarboxylic acid cycle
TPM	Trancript per million
TVD	Total Variation Distance
TVD BioCh	TVD calculated using the biochemical properties
RBS	Ribosomal Binding Site
RNAi	RNA interference
rpm	revolution per minute
RuBisCO	Rubilose-1,5-Bisphosphate Carboxylase and Oxygenase
RuBP	Rubilose-1,5-bisphosphate
U	Enzyme unit

Chapter 1: General Introduction

Large sections of this introduction chapter were published as a review in the Journal of Experimental Botany: **Niklaus M, Kelly S** (2018) The molecular evolution of C_4 photosynthesis: opportunities for understanding and improving the world's most productive plants doi:10.1093/jxb/ery416

1.1 Introduction into C_4 photosynthesis

Plants convert solar energy to chemical energy through a process known as photosynthesis. In doing so, they alter the chemical composition of the atmosphere and produce the sugars that form the basis of the global food chain. In land plants, an ancestral photosynthetic pathway has been modified through evolution such that the majority of extant species can be classified into the three major photosynthetic types; the ancestral C_3 type and two derived types known as C_4 and CAM. These photosynthetic types share many commonalities, but differ in some aspects of biochemical, temporal and spatial organisation of their carbon fixation pathways. These differences in carbon fixation result in some general functional distinctions between species in which they are manifest. For example, C_4 plants are generally more efficient in capturing CO_2 and creating biomass than C_3 or CAM plants (Sage 2004). Similarly, CAM plants are generally more water use efficient than C_4 or C_3 plants (Black Jr 1973; Han and Felker 1997). However, there are exceptions to this general rule. For example, C_4 plants require additional energy to capture CO_2 than C_3 plants and thus in cold climates C_3 plants are more efficient (Zhu et al. 2010).

Of the three major photosynthetic types in land plants, C_3 plants are the most numerous and our knowledge of their biology, biochemistry, and physiology is more advanced than that of C_4 or CAM plants. Consequently, the majority of attempts to

enhance photosynthesis in crops have focussed on C_3 plants while improvements to CAM or C_4 photosynthesis have received comparatively little attention. This is perhaps due to the fact that these photosynthetic types already possess multiple desirable characteristics that would be beneficial to engineer into C_3 species (Borland et al. 2009; von Caemmerer and Furbank 2016; Zhu et al. 2010). However, future increases in global population necessitate that we also improve the yields of these plants (Long et al. 2015; von Caemmerer and Furbank 2016). Such improvements are perhaps most important for the C_4 crop maize, given its prominent role in food, feedstock, and bioenergy production (von Caemmerer and Furbank 2016). However, C_4 crops such as tef (*Eragrotis tef*), sorghum (*Sorghum bicolor*) and millets (*Setaria sp.*) also play a large role in global food security. Collectively, these other crops occupy approximately the same global area as soybean production, but unlike soybean, they are directly used for human consumption (Naylor et al. 2004). Thus considerable enhancement in global food security could be gained from their improvement (Naylor et al. 2004).

This introduction chapter summarises the insights into C_4 function that have been gained by studying natural variation between C_4 species. Particular attention is paid to molecular evolution studies that have revealed how different proteins have been recruited and adapted for C_4 function. It is also discussed how this functional diversity suggests that there is a large potential operating space for improving C_4 photosynthesis.

1.2 C_4 photosynthesis; is age more than a number?

C_4 photosynthesis is a complex trait that has evolved independently at least 62 times during land plant evolution (Sage 2016). In the context of the 480 million year history of colonisation of the land by plants (Kenrick and Crane 1997), C_4 photosynthesis is relatively a recent innovation (Figure 1.1 A). For the first 450 of those 480 million years

atmospheric CO_2 concentration was generally elevated with respect to today's CO_2 levels (Figure 1.1 B). It is only within the last 30 million years, when atmospheric CO_2 was comparatively low, that each of the independent origins of C_4 photosynthesis have evolved (Christin et al. 2011a; Sage 2004) (Figure 1.1 B). Thus, this drop in atmospheric CO_2 is thought to be one of the key ecological drivers of C_4 evolution (Ehleringer et al. 1997). The suite of anatomical, biochemical, and physiological changes that evolved in response to this driver counteract the reduction in atmospheric CO_2 by increasing CO_2 concentration around RuBisCO (Ehleringer et al. 1991). Ultimately, this results in significant energy, water and nitrogen savings such that C_4 plants can achieve ~50% higher rates of photosynthesis than C_3 plants given the same resources under certain environmental conditions (Evans and von Caemmerer 2000). It should be noted here that it is possible that C_4 photosynthesis evolved indepenently at an earlier timepoint when atmospheric CO_2 concentraiton was low and subsequently went extinct when CO_2 increased (Osborne and Beerling 2006).

Although extant C_4 photosynthesis is a relatively recent invention, the independent evolutionary origins of C_4 photsynthesis did not emerge all at the same time. Instead, there is a large disparity in the ages of different C_4 lineages. Some are as young as a few hundred thousand years (e.g. *Cleome angustifolia* ~400,000 years old +/- 200,000 years (Feodorova et al. 2010), or *Flaveria kochiana* ~300,000 years +/- 200,000 years (Christin et al. 2011a)), while the oldest C_4 lineage the *Chloridoideae* is ~ 30 million years old (Christin et al. 2008a; Christin et al. 2011a; Osborne and Sack 2012; Sage et al. 2011) (Figure 1.1 C). Thus unlike for C_3 photosynthesis, where all extant C_3 plants share a pathway that is equally old, there is up to a 30 million year difference in the amount of time that natural selection has had to adapt the anatomy, physiology, and biochemistry of C_4 function in different C_4 lineages. Although many individual C_4

lineages contain both fast and slow growing species, it is noteworthy in this context that older C_4 lineages contain many of the world's most productive species. For example, the oldest C_4 lineage the Chloridoideae, contains some of the world's most productive fodder and feed species (e.g. *Chloris guyana* and *Cynodon dactylon)* (Sage 2016), and the Andropogoneae (~20 mya) contains some of the world's most productive crop plants (e.g. maize, sorghum, sugar cane) (Grass Phylogeny Working Group II 2012; Sage 2004; Sage 2016). In contrast, some of the more recent C_4 lineages can be less productive than their closest C_3 relatives at current atmospheric CO_2 concentrations (Lundgren et al. 2016; Ripley et al. 2008). Thus, it will be interesting to determine whether the amount of time natural selection has had to act to improve C_4 photosynthesis has resulted in general trends in productivity and efficiency of the C_4 cycle in different C_4 lineages.

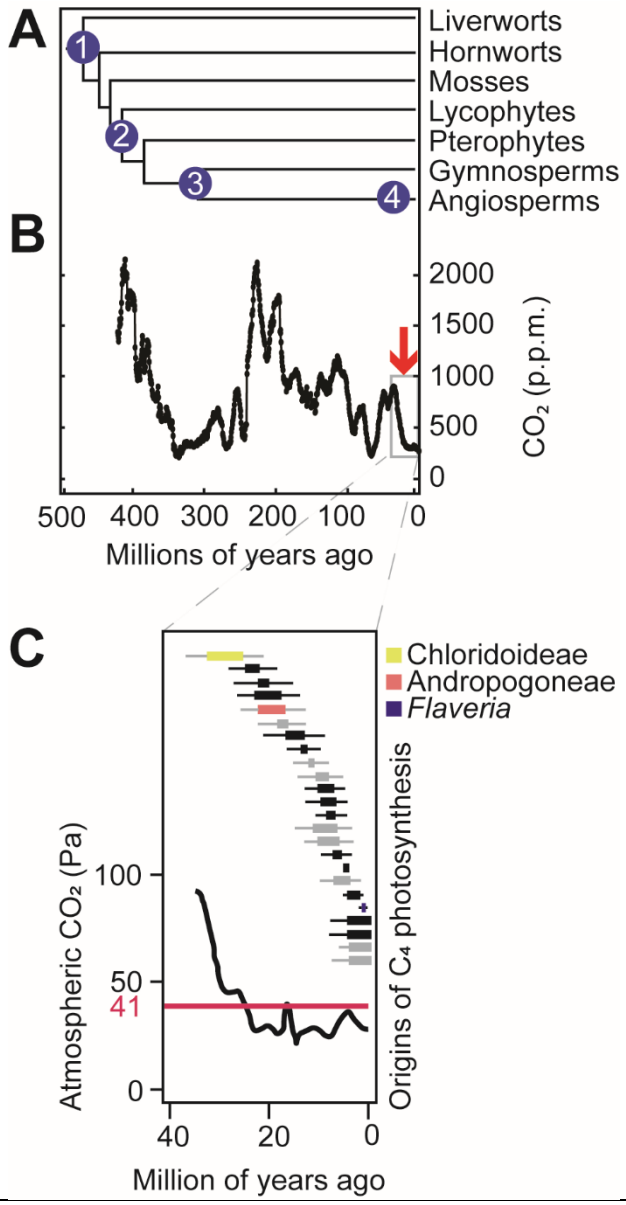


Figure 1.1: The evolution of land plants and the emergence of C_4 photosynthesis. **A**) A simplified phylogenetic tree showing the evolution of land plants over the past 500 million years. Major events are indicated by numbered blue circles: The origin of land plants approximately 475 million years ago (1), origin of vascular plants approximately 420 million years ago (2), origin of seed plants approximately 310 million years ago (3), origin of C_4 photosynthesis about 30 million years ago (4). **B**) Shifts in atmospheric CO_2 concentration in ppm (parts per million) over the last 420 million years (Foster et al. 2017). Red arrow pointing to the drop in atmospheric CO_2 levels **C**) The predicted atmospheric pressure of CO_2 in Pascale (Pa) over the last 40 million years (left hand y-axis) (Beerling and Royer 2011). The current (41 Pa) CO_2 pressure is indicated by a red horizontal line. The inferred emergence

times of C_4 photosynthesis during the last 40 million years. The grey and black bars represent emergence of C_4 photosynthesis in grasses and eudicots, respectively. Error bars represent the uncertainty of the predicted time point of emergence (Christin et al. 2011; Osborne and Sack 2012). Three lineages are highlighted in different colours for illustrative purposes only.

1.3 C_4 photosynthesis; equal but not the same

Given the substantial differences in age between independent C_4 origins, and the large phylogenetic distance between different C_4 lineages (Sage 2004), it is not surprising that there are significant differences in biochemistry, anatomy and physiology between different C_4 species. While there are some species that developed a C_4 pathway within a single cell (Edwards et al. 2004; Voznesenskaya et al. 2002), the vast majority of C_4 plants split the photosynthetic pathway between two cell types. These two cell types are arranged concentrically within the leaf, and comprise an outer layer of photosynthetic carbon assimilation tissue surrounding an inner layer of photosynthetic carbon reduction tissue (Muhaidat et al. 2007; Rao and Dixon 2016). The precise cell type that forms these two layers varies between C_4 lineages. For example, in C_4 grasses, the inner layer photosynthetic carbon reduction tissue can be mestome sheath and/or bundle sheath (Nelson and Langdale 1989). These cell types are both present in C_3 grasses (Kinsman and Pyke 1998), and thus the evolution of C_4 photosynthesis can proceed by recruitment of multiple different cell layers giving rise to a diversity of anatomical forms (Muhaidat et al. 2007; Rao and Dixon 2016) (Figure 1.2 A).

In addition to anatomical disparity between lineages, there are substantial biochemical differences between C_4 plants (Hatch 1987). These biochemical differences are complex and not yet fully understood, but can be generalised according to the decarboxylase localised in the photosynthetic carbon reduction cells that is used to

liberate CO_2 from its C_4 acid (Wang et al. 2014). For this reason, C_4 species are often classified into three semi-distinct subtypes (NADP-malic enzyme, NAD-malic enzyme and phosphoenolpyruvate carboxykinase) according to the activity or expression of this decarboxylase (Figure 1.2 B). More recently, the existence of pure PCK types has been questioned, as it is more often found used in conjunction with one of the other two decarboxylases (Wang et al. 2014).

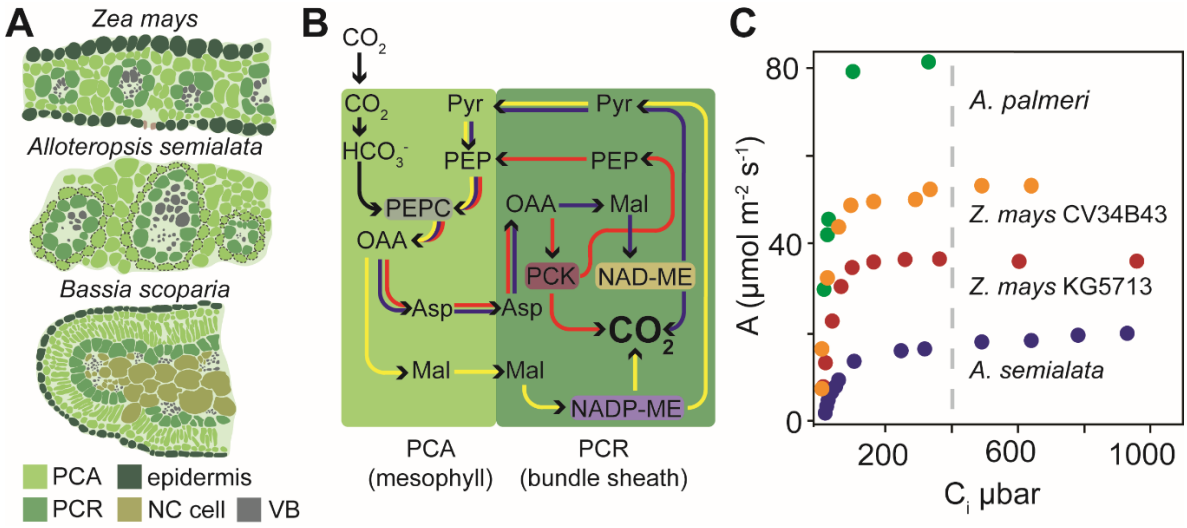


Figure 1.2: Multiple sources of variation present within different C_4 lineages. **A)** Example of anatomical variation between *Zea mays*, *Alloteropsis semialata* and *Bassia scoparia*. Here leaf cross-sections are depicted as drawings to highlight similarities and differences in leaf anatomy between the three example species. The photosynthetic carbon assimilating (PCA) tissue is highlighted in light green and the photosynthetic carbon reduction cells are coloured in dark green. *A. semialata* uses both mesophyll cells (light green) and the outer bundle-sheath cells (light green with dotted rim) as a PCA tissue (Dengler et al. 1985; Ueno and Sentoku 2006). The scale of the illustrations are not the same. Figure abbreviations: NC, non-chlorenchymatous cells; VB, vascular bundle. **B)** Variation in C_4 metabolism between subtypes. C_4 photosynthesis is classified into three subtypes based on the enzyme catalysing the decarboxylation (PCK (red line), NAD-ME (blue line), NADP-ME (yellow line)). Depending on the subtype, different metabolites play key roles in different subcellular compartments in

the C_4 cycle. Pyr is pyruvate, PEPC is phosphoenolpyruvate, Mal is malate, Asp is aspartate, OAA is oxaloacetate. **C)** An A-Ci curve exemplifying the variation in photosynthetic rate among C_4 plants. *A. palmeri* (Ehleringer 1983), *Z. mays* CV34B43 Pioneer Hi-Bred International (Leakey et al. 2006), *Z. mays* KG5713 Sweet Corn, Kelvedon Glory (Sharwood et al. 2016) and *A. semialata* (Lundgren et al. 2016)). Two varieties of maize are shown to demonstrate there is also within species variation.

The differences in biochemistry between different C_4 subtypes is mirrored by differences in transporters required to facilitate this biochemistry (von Caemmerer and Furbank 2003; Weber and von Caemmerer 2010) (Figure 1.3). However, unlike the biochemistry where the complete set of enzymes required to complete a C_4 cycle is known, the complete set of transporters required to conduct a C_4 cycle in any species are unknown. Specifically, in the NADP-ME C_4 subtype the malate uptake and the pyruvate export transporters of the bundle sheath chloroplast are unknown (Schuler et al. 2016). In the NAD-ME C_4 subtype the aspartate import and pyruvate export transporters of the bundle sheath mitochondrion are unknown. In the PEPCK C_4 subtype the malate import and pyruvate export transporters of the bundle sheath mitochondrion are unknown (Weber and von Caemmerer 2010) (Figure 1.3). It is important to highlight that various combinations of the different subtypes exist in nature. As an example, *Zea mays* utilizes a combination of the NADP-ME and PEPCK C_4 photosynthesis and thus many of the transporters required to complete a C_4 cycle in maize are unknown (Weber and von Caemmerer 2010). Furthermore, even with a subtype it is known that species can use different transporters. For example, the BASS2 sodium pyruvate symporter expressed in the mesophyll chloroplasts of many eudicot NADP-ME C_4 (Kelly et al. 2017) species and in the monocot NADP-ME C_4

Setaria italica but is not expressed in maize leaves (John et al. 2014) and thus an alternative pyruvate importer must be used in this species.

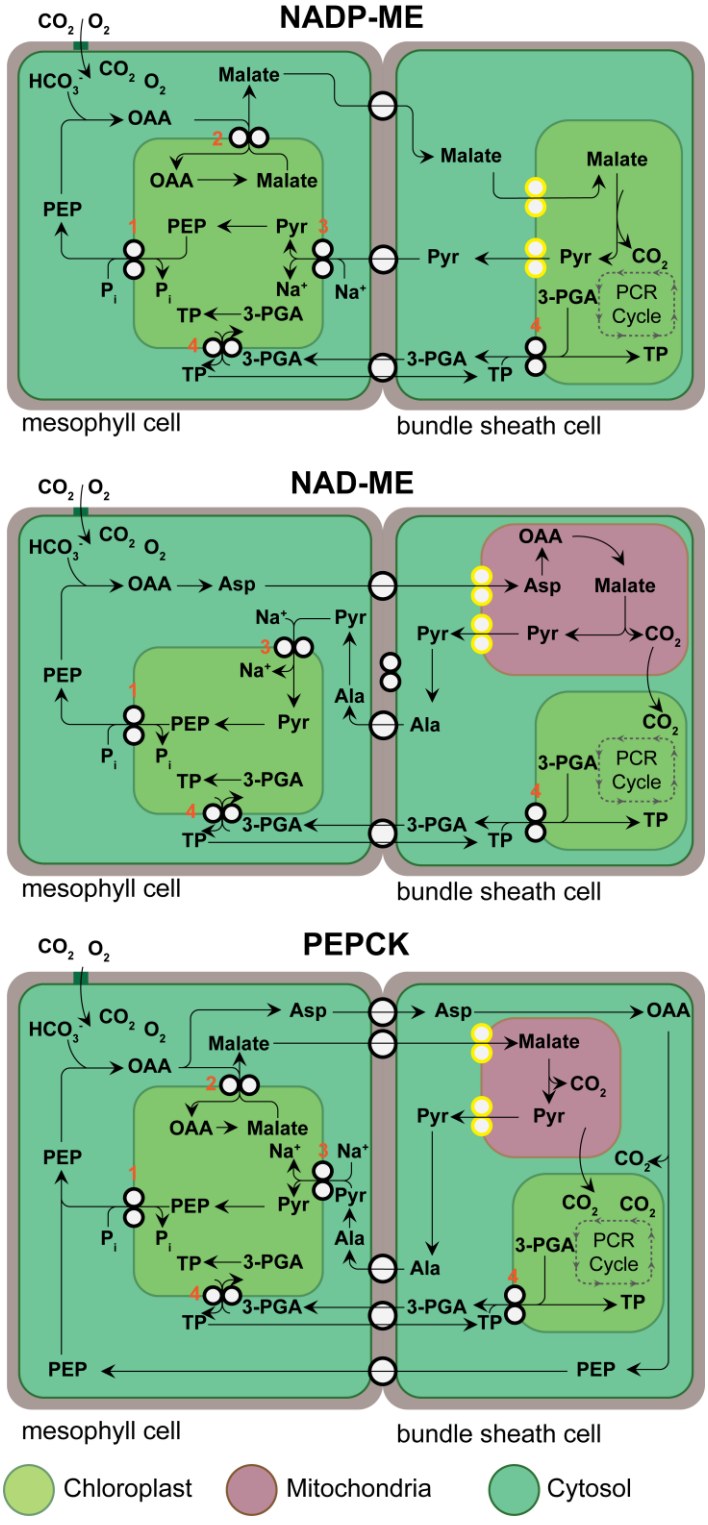


Figure 1.3: The unknown metabolite transporters of the C_4 cycle. Figure adapted from Weber and von Caemmerer 2010 (Schuler et al. 2016; Weber and von Caemmerer 2010). The

unknown transporters are highlighted in yellow. The known transporters are labelled with numbers: **1**) PEP/phosphate translocator (PPT), **2**) dicarboxylate transporter 1 (Dit1), **3**) bile acid:sodium symporter family protein 2 (sodium-dependent pyruvate transporter) (BASS2) **4**) triose phosphate translocator (TPT). Abbreviations of the metabolites: aspartate (Asp), triose phosphate (TP), ortho-phosphate (P_i), phosphoenolpyruvate (PEP) oxaloacetate (OAA), pyruvate (Pyr), 3-phosphoglyceric acid (3-PGA), sodium (Na^+), bicarbonate (HCO_3^-).

Outside of these central enzymes and transporters, other differences in biochemistry between C_3 and C_4 species are continually being identified (Bräutigam et al. 2011; Gowik and Westhoff 2011; Kelly et al. 2017; Lauterbach et al. 2017). The diversity of photosynthetic rates among C_4 species (Figure 1.2 C) collectively results from differences in anatomy, biochemistry, transport, allocation of resources towards leaf photosynthesis (Ghannoum et al. 2005), and the amount of time natural selection has had to adapt genes for C_4 function.

Such differences can be further exacerbated under environmental conditions such as high temperature (Sonawane et al. 2017). Given that photosynthetic rates have been enhanced to different extents in different species, it follows that not all species will have equally improved all of the reactions of their photosynthetic pathways. For example, it has been shown that altered RuBisCO kinetics contribute to variation in photosynthetic nitrogen use efficiency in different subtypes of C_4 photosynthesis (Ghannoum et al. 2005). Thus, it will be instrumental to our understanding of the evolution of C_4 photosynthesis, and how to improve it, to determine the extent to which these differences are attributable to differences in the kinetic (or other) properties of the components of the C_4 cycles in these species.

1.4 Improving C_4 function by simple molecular sequence changes

C_4 photosynthesis, with its multiple independent evolutionary replicates, is an ideal subject for detecting how natural selection has adapted molecular sequences for C_4 function. Accordingly, multiple examples of positive selection have been found in multiple different genes. The first such discovery was made by comparing the sequence of the PEPC gene in C_3 and C_4 *Flaveria* (Bläsing et al. 2000) (Figure 1.4 A). Christin et al. 2007 also identified a strong signal for positive selection at 21 different amino acid positions in the PEPC gene in grasses (Christin et al. 2007). Perhaps the most remarkable change in PEPC that occurred between C_3 and C_4 species was at the amino acid position 780 (position corresponds to *Z. mays* PEPC CAA33317). Here a transition from alanine to serine was observed in eight independent C_4 grass lineages and *Flaveria* (Bläsing et al. 2000; Christin et al. 2007). Interestingly, the alanine to serine substitution increases the K_m(PEP) of the enzyme. It was hypothesized that this increase may play a regulatory role in C_4 photosynthesis because the concentration of PEP is higher around C_4 PEPC than it is around C_3 PEPC (Engelmann et al. 2002; Svensson et al. 2003). Alternatively, it has also been proposed that the higher K_m(PEP) may be coincident with an elevated malate tolerance thus reducing competitive inhibition of enzyme function in C_4 photosynthesis (Svensson et al. 2003).

In a similar approach, the phosphoenolpyruvate carboxy kinase (PCK), NADP-malic enzyme (NADP-ME) (Christin et al. 2009a) and RuBisCO large subunit (rbcL*)* (Christin et al. 2008b) were also found to be subject to positive selection in different C_4 lineages (Christin et al. 2009a; Christin et al. 2008b; Christin et al. 2009b) (Figure 1.4 A). Thus, it appears that several key enzymes in the C_4 cycle have been adapted by positive selection in different C_4 lineages to improve their function for C_4 photosynthesis.

To determine whether other genes outside of the core C_4 pathway also showed evidence of adaptation for C_4 function, a large scale genome-wide scan for positive selection associated with C_4 photosynthesis was conducted (Huang et al. 2017). This identified 88 genes that showed evidence of positive selection in C_4 species from two independent C_4 lineages (Huang et al. 2017) (Figure 1.4 B). This list included genes that had been identified before, but also included genes involved in leaf development (SCARECROW-like 14), vascular patterning (VASCULAR NETWORK DEFECTIVE 1), and other signalling molecules (Huang et al. 2017). The wide-spread extent of positive selection in the ~20 million years following the origin of C_4 in these lineages demonstrate that a large number of factors may influence the fitness of C_4 species once they make the transition from C_3 to C_4 photosynthesis. Moreover, as this study samples only two of the >60 independent lineages of C_4 photosynthesis, and only a small number of species within each lineage, it is likely that increasing the sample size of both the number of species within each lineage and the number of independent C_4 lineages will result in the identification of additional sites and genes under selection. Given the large number of potential sequence variants in genes related to the C_4 cycle, and that different C_4 lineages have had different amounts of time for natural selection to act it is possible that some species have acquired adaptive amino acid changes that other species have not. Identifying these adaptive changes, and species in which they have not yet occurred, could provide clues for where to engineer C_4 proteins for enhanced C_4 function.

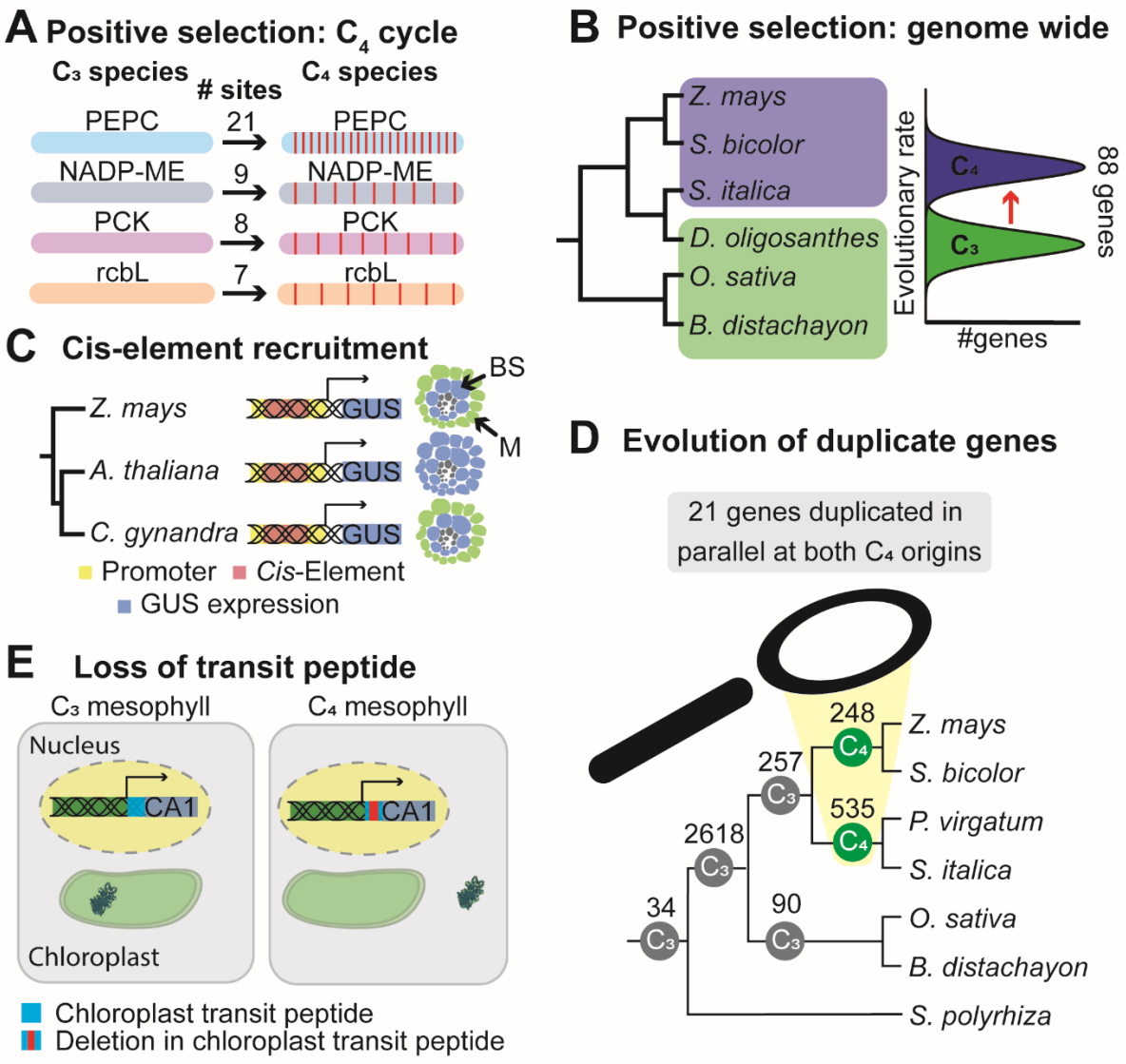


Figure 1.4: The repeated evolution of C_4 photosynthesis has utilised common molecular changes. **A**) Positive selection on sites in genes associated with C_4 photosynthesis was shown in PEPC (21 sites under selection), NADP-ME (9 sites under selection), PCK (8 sites under selection) and rcbL (7 sites under selection). **B**) Genome wide selection scan across C_3 and C_4 species revealed 88 genes under selection leading to a C_4. **C**) Parallel recruitment of regulatory elements in different origins of C_4 photosynthesis. The promotor region of NAD-ME *C. gynadra* NAD-ME1 and AtNAD-ME1 fused to GUS reporter results in BS specific expression in C_4 plants (*Z. mays* and *C. cleome*). In *A. thaliana* the promoter region is fused to a GUS reporter, which does not result in cell type-specific expression (Brown et al. 2011). **D**) 21 genes duplicated in parallel at two independent origins of C_4 photosynthesis (Emms et al. 2016). Independent origins of C_4 photosynthesis are indicated by green circles. The numbers above

branches indicate the number of genes that duplicated along this branch of the species tree. **E**) The loss of the chloroplast targeting peptide of carbonic anhydrase in multiple C_4 lineages moved CA activity to the cytosol where it catalyses the first reaction of the C_4 cycle.

1.5 Improving C_4 function through gene duplication

Positive selection is limited to the analysis of simple sequence changes. However, evolutionary changes to DNA sequences are not limited to changes at single nucleotides. Given the extent of positive selection that is apparent in C_4 genomes it is likely that larger changes in DNA sequences are also numerous. Several studies have looked for larger changes in genome sequences by looking at duplication of whole genomes and of genes involved in the C_4 cycle (Bianconi et al. 2018; van den Bergh et al. 2014; Williams et al. 2012). These studies found that duplication of genes encoding proteins of the core C_4 cycle generally occurred before the origin of C_4 photosynthesis and thus may have been a contributing factor by enabling C_4 evolution through neofunctionalisation and/or dosage effect of gene duplicates. In addition to these studies on C_4 cycle genes, one further study attempted a genome-wide interrogation of gene duplication and the evolution of C_4 photosynthesis (Emms et al. 2016). In this study, an analysis of gene duplication on branches leading to C_4 photosynthesis was conducted in two independent origins of C_4. This revealed that hundreds of new genes were created by gene duplication around the time when C_4 photosynthesis evolved in these lineages, and that a core set of 21 genes was independently duplicated in parallel in both C_4 lineages (Emms et al. 2016). Of particular note was that this study identified the bundle sheath localised sucrose efflux transporter SWEET13 as being recruited in parallel for C_4 function in both C_4 lineages. It was suggested that these transporters function to export sucrose out of the bundle sheath into the vasculature apoplast (Emms et al. 2016), and that this export would be

advantageous for C_4 species as the number of plasmodesmata connections between bundle sheath and vascular parenchyma are reduced in comparison with those between the bundle sheath and mesophyll (Botha 1992; Botha and Van Bel 1992; Danila et al. 2016; Sowiński et al. 2007). This reduction of plasmodesmata connectivity to the vasculature would prevent wasteful diffusion of C_4 metabolites out of the C_4 cycle, but would also prevent sugars being exported into the vasculature. Thus, recruitment of a sucrose efflux transporter would allow plasmodesmata connections to the vasculature to be reduced to maintain C_4 cycle flux between bundle sheath and mesophyll without losing sugar export (Emms et al. 2016). Consistent with this hypothesis, it was subsequently shown that knock-out mutants of SWEET13 in maize were severely stunted with impaired photosynthesis and a high accumulation of free sugar and starch in leaves (Bezrutczyk et al. 2018). As more C_4 genomes become available, it will be interesting to determine the extent to which the same genes are recurrently duplicated and how frequently SWEET13 is recruited for sucrose export in the leaves of C_4 plants.

The study of gene duplication associated with the evolution of C_4 photosynthesis revealed additional surprising findings. First, among these was that C_4 species can switch between paralogs that are used for C_4 photosynthesis. Here, it was revealed that the paralog of the chloroplast localised dicarboxylate transporter used for C_4 photosynthesis in maize today (DCT2/DIT2) was not the isoform that was used when C_4 first evolved in this lineage (Emms et al. 2016). Although the mechanism by which this change occurred is unknown, it is possible that the original gene was lost during fractionation following the recent whole genome duplication in maize. To facilitate this loss there must either have been a period where both copies of the gene were expressed, or there must be some functional redundancy with alternative transport

pathways in maize. Intriguingly, although DCT2 knockouts in maize are stunted they are viable (Weissmann et al. 2016), this lack of lethality may have partially facilitated paralog switching. This example suggests that while there is commonality in the isoforms that are used for C_4 function in some C_4 lineages, paralogs may provide functional redundancy in the case of gene loss. As more data becomes available, it will be interesting to determine the extent to which such paralog switching occurs.

The second additional insight from Emms et al. 2016 revealed that the reason for commonality in the isoforms of genes recruited for C_4 function is simply due to being in the right place at the right time (Emms et al. 2016). That is, the isoforms that were already expressed in leaves when the plants evolved C_4 were those that were recruited for the function in C_4 photosynthesis. This phenomenon was not limited to C_4 cycle genes. Instead, the vast majority (74%) of duplicated genes that are expressed in mature foliar leaves of *Z. mays* and *O. sativa* showed consistent expression whereby the same gene of the duplicate pair was more highly expressed in leaves (Emms et al. 2016). This finding was replicated in a later study that encompassed a larger number of independent origins of C_4 photosynthesis (Moreno-Villena et al. 2017).

1.6 Improving C_4 function through recruitment of cis-regulatory networks

In addition to providing insight into metabolic functions, analyses of duplication events also provide insight into regulatory mechanisms. For example, from analysis of whole genome duplication events, it was concluded that the recruitment of existing genes through alterations of regulatory elements was a key process in the evolution of C_4 photosynthesis (van den Bergh et al. 2014) (Figure 1.4 C). Consistent with this hypothesis, it has recently been reported that a bHLH gene that regulates bundle sheath cell specific expression of the NADP-ME decarboxylase in maize provides the same function in C_3 plants (Borba et al. 2018), and that cell type specific gene

expression patterns in C_4 species utilise the same functional cis-regulatory elements as those in C_3 species (Reyna-Llorens et al. 2018). Given that established regulatory mechanisms were co-opted for control of C_4 gene expression, the elucidation of the cis-elements and trans-factors that control this expression is paramount to our understanding of C_4 evolution and our ability to enhance C_4 photosynthesis. The first trans-factors to be implicated in this process were maize GOLDEN2 and GOLDEN2-LIKE 1 (Hall et al. 1998; Rossini et al. 2001). These transcription factors are regulators of chloroplast development that differentially accumulate in mesophyll and bundle sheath cells in C_4 plants (John et al. 2014; Wang et al. 2013; Waters and Langdale 2009) and it was shown that either gene is capable of inducing chloroplast development in the bundle sheath cells of the C_3 plant rice (Wang et al. 2017). Although few other trans-acting factors have been identified, evidence of parallel recruitment of regulatory elements in C_4 plants has been found. Of particular note was the discovery that bundle sheath cell specific expression of genes in distantly related C_4 species could be mediated by cis-elements present in the coding region of genes (Brown et al. 2011). While exaggeration of ancestral expression patterns appears to be a common way in which many genes are recruited and adapted for C_4, function it is also possible that *ab initio* recruitment has also occurred. An example of such recruitment would be where a gene that is initially not expressed in the right cell and/or at the right time is recruited to C_4 function by activation of expression. Such examples are difficult to prove in the absence of living ancestral species. However, given the expression of the majority of C_4 enzymes and transporters in the leaves of distantly related C_3 plants it seems likely that *ab initio* recruitment was less frequent than exaggeration of ancestral expression patterns during the evolution of C_4 photosynthesis. Thus, it is likely that there is much that could be learned about how C_4 photosynthesis is controlled from

elucidating the gene regulatory networks governing C_3 photosynthesis and cell type specification.

1.7 Improving C_4 function through subcellular retargeting of proteins and lateral transfer of genetic material

Some of the enzymes of the C_4 photosynthetic pathway come from paralogous gene families, where different members of the family exhibit different expression patterns or different subcellular localisations. As stated above, the members of these gene families that were expressed in the right cell type at the right time were those that were recruited for C_4 function, with other members of these families providing redundancy in the case of gene loss (Emms et al. 2016). However, there are examples where proteins change subcellular localisation coincident with the evolution of C_4 photosynthesis.

In C_3 plants, carbonic anhydrase is an abundant enzyme in the chloroplasts of mesophyll cells where it functions to either facilitate the diffusion of CO_2 across the chloroplast envelope or to rapidly convert bicarbonate to CO_2 (Cowan 1986; Price et al. 1994; Reed and Graham 1981; Studer et al. 2014). While C_4 plants also require carbonic anhydrase activity in mesophyll cells to provide bicarbonate for PEPC, they require this activity in the cytosol where PEPC is located rather than the chloroplast. There are two options to solve this localisation problem, either *ab initio* recruit a cytosolic carbonic anhydrase by altering the regulatory sequences controlling its expression, or discard the chloroplast targeting peptide from the gene that is already expressed in the right cell type at the right time. The loss of chloroplast targeting peptides from carbonic anhydrase has occurred in two independent C_4 lineages; once within *Neurachne* (Clayton et al. 2017) and once in *Flaveria* (Tanz et al. 2009). The relative occurrence of transit peptide loss to regulatory recruitment of cytoplasmic CA recruitment is unknown, however transit peptide loss can occur through single (or small

numbers of) point mutations and thus it is tempting to speculate that this may be more likely to occur than acquisition of a functional cis-regulatory elements which may require multiple nucleotide changes (Figure 1.4 E). It will be interesting to determine whether such re-localisation events in other genes also exhibit associations with the evolution of C_4 photosynthesis in multiple lineages.

In addition to the transfer of gene products from one subcellular compartment to another, the transfer of entire genes between species has also been observed in the *Alloteropsis* C_4 lineage (Christin et al. 2012; Dunning et al. 2017; Moreno-Villena et al. 2017). Such events, while likely rare, present an alternative route to acquire molecular sequences that have already been adapted to some extent for C_4 function by natural selection.

1.8 Challenges and opportunities for engineering C_4 photosynthesis

As outlined above, there are multiple ways in which natural selection has modified C_3 photosynthesis to give rise to a myriad of phenotypes that are collectively known as C_4 photosynthesis. Even just considering the two origins of C_4 photosynthesis in the grasses described above (in the Andropogoneae and the Paniceae) there over 100 molecular sequence changes that have been linked to the evolution of C_4 photosynthesis, ranging from single nucleotide substitutions to the evolution of duplicate genes (Figure 1.4). This set of ~100 genes is the set that is in common between both origins and it is apparent from the differences in anatomy, physiology and biochemistry (for example *Z. mays* uses a mix of PCK and NADP-ME decarboxylases while *S. bicolor* and *S. italica* use pure NADP-ME) that the evolution of C_4 photosynthesis in any given lineage involved gene sequence changes that have not been replicated in other lineages. Thus, the true extent of adaptive molecular

sequence changes associated with C_4 photosynthesis in any one species is likely to be larger.

This molecular disparity, coupled with the ~30 million-year difference in the amount of time that natural selection has had to adapt gene sequences for C_4 function, means that it is also likely that not all C_4 cycles are equally well adapted or efficient for either pre- or post-industrial levels of atmospheric CO_2. Specifically, C_4 plants evolved relatively recently such that different C_4 lineages (assuming one generation per year) have existed for between ~10^5 to ~10^7 generations. Where it has been measured, plants have spontaneous mutation rates of between 10^{-10} to 10^{-8} mutations per base per generation (Ossowski et al. 2010; Yang et al. 2017). Thus, using simple Wright–Fisher model (Fisher 1999; Wright 1931) with selection and drift, it is possible to estimate the probability that a variant that is beneficial to C_4 cycle function has arisen and become fixed within any given C_4 lineage as a function of time. For example, if we consider the frequently observed alanine to serine substitution in PEPC (Bläsing et al. 2000; Christin et al. 2007) that can be caused by a single G->T mutation in the first base of the codon, then in a population of 10^6 individual plants there would be a 99% probability of observing this mutation in a single individual plant within ~1800 generations (Figure 1.5 A). For smaller population sizes this time is longer (Figure 1.5 A), however even for very small populations of ~10,000 individuals the mutation would still be expected to emerge once within ~180,000 generations (Figure 1.5 A). As many C_4 species are annual plants a generation equates to a year, however some C_4 species are perennials and thus a generation may equate to 2 to 10 years. Thus, a small population of ~10^4 C_4 perennials (~4-year generation time) may take up to 1 million years to have a 99% probability of evolving the alanine to serine mutation. If this mutation conferred a large fitness advantage then it could sweep to fixation in a

relatively small number of generations that is inversely proportional to population size. For example, if the alanine to serine substitution conferred a 1% fitness advantage then even in large populations of 10^6 individuals one would expect this substitution to reach fixation in ~5,000 generations (Figure 1.5 B). Thus, large effect single nucleotide mutations, such as the A to S substitution is assumed to be, are likely be present in even the youngest C_4 lineages today. However, if the fitness advantage is weaker, then it would take considerably longer to reach fixation. For example, if the fitness advantage is 10-fold weaker, then it would take approximately 10-fold longer to reach fixation (Figure 1.5 C). It is easy to see from these data that variation in population size, selective advantage, and life history strategy could push the number of generations beyond the age range of many C_4 species. For example, in small populations ~10^4 individuals with lower mutation rates ~10^{-10} per base per generation, then it would take 18 million generations for a 99% chance of the alanine to serine mutation to appear in the population. Similarly, for large populations ~10^7 individuals with weak fitness advantages it would take hundreds of thousands of generations for beneficial alleles to become fixed. However, on average, considering all of the factors above, large fitness effect (i.e. ≥0.1% advantage) single nucleotide mutations should be present in most extant C_4 lineages.

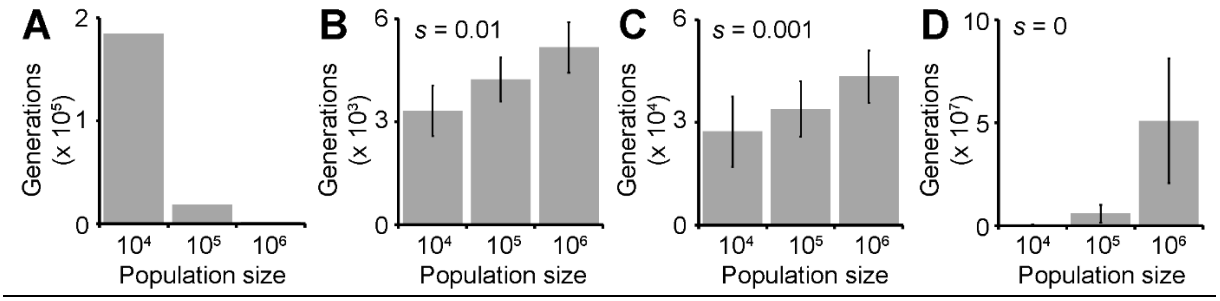


Figure 1.5: Evolutionary dynamics simulations. **A**) The number of generations taken to have a 99% probability of evolving a specific single nucleotide mutation for three different population sizes given a mutation rate of 10^{-8} changes per base per generation. **B**) The average number of generations it would take for a mutation with a fitness advantage of 1% (s = 0.01) to become fixed in the population given three different population sizes. Error bars indicate one standard deviation of the mean. **C**) As in B, but for a ten-fold weaker selective advantage. **D**) The average number of generations it would take for a neutral mutation to become fixed in the population through genetic drift. B, C and D depict the results of 100 replicate simulations considering a diploid genome with h = 0.5 and random mating. https://github.com/SteveKellyLab/Simple_evolutionary_dynamics_simulator.

What is critical to note here is that most amino acid substitutions are not achievable by single nucleotide mutations. Only 150 of the 380 possible amino acid substitutions (39%) can be achieved by such mutations (Wong et al. 2007). Thus, amino acid substitutions that require multinucleotide mutations (61%), where single mutations in isolation have no fitness advantage or are weakly deleterious, are subject to drift or purifying selection respectively. For example, for a population sizes of 10^6 individuals it takes on average ~5.1 x 10^7 generations for a mutation with no fitness advantage to reach fixation through drift (Figure 1.5 D). Thus for 61% of possible amino acid substitutions, even if they confer a strong fitness advantage, it could take tens of millions of generations for the precise series of events required to occur for that substitution to arise. For example, if alanine to tyrosine (requiring two mutations) at the same position in PEPC conferred an even higher fitness advantage than alanine to serine, then it is only likely to have occurred in some species descended from the oldest C_4 lineages. Therefore, fitness advantages caused by amino acid substitutions that require multiple mutations, or that require changes to protein sequences requiring more than one amino acid substitution (i.e. two distant sites in the same gene) take millions of generations to occur and are unlikely to be present in most C_4 lineages today. Thus, it is very unlikely that most single (or multiple) amino acid substitutions that confer fitness advantages have evolved in all C_4 lineages, even given the relaxed purifying selection that species would encounter from the increase in photosynthetic nitrogen use efficiency gained from evolving C_4 photosynthesis (Kelly 2018).

Studying the molecular sequence variation between C_4 species can provide insights into which sites may confer fitness advantages. By focussing on genes of the C_4 cycle, it is likely that these fitness advantages are mediated by modulating flux through the C_4 cycle and thus relate to photosynthetic efficiency. Therefore, sites that show

evidence of positive selection in genes of the core C_4 cycle are likely to be sites that confer improvements to C_4 photosynthetic efficiency. Given the diversity in the age range of different C_4 species, and the time taken for single and multi-site mutations to arise and become fixed in populations, there are likely a large number of sites in C_4 cycle enzymes that have been adapted for C_4 function in some lineages but not in others. Moreover, unless the optimal amino acid substitution was achievable by a single nucleotide mutation, it is likely that the best suited amino acid is not yet present at most sites under positive selection in different C_4 lineages. The same assumptions also hold for single and multi-site mutations required to tailor the abundance of transcripts encoding components of the C_4 cycle to an optimal level for C_4 function. It is therefore also most parsimonious to assume that the relative abundance of proteins in the C_4 pathway are not optimal in the majority of C_4 species. Thus, studying which sites are under positive selection can help target engineering efforts to improve C_4 photosynthesis. Moreover, altering transcript abundance of genes in the core C_4 cycle in C_4 species will also likely lead to improvements in C_4 photosynthetic efficiency. Therefore, studying the molecular basis of the evolution of C_4 photosynthesis is thus not only important for understanding the myriad of ways in which convergent adaptive traits evolve, it holds the potential to unlock new avenues of improvement for multiple important C_4 food, feedstock, and bioenergy crops.

1.9 Thesis aims and overview

The overarching aim of my thesis was to use knowledge of the biochemistry and evolution of C_4 photosynthesis to explore new ways of engineering C_4 plants for enhanced photosynthetic capacity.

To achieve this aim, I decided to search for sites in protein sequences that natural selection had altered to adapt those proteins for function in C_4 photosynthesis. Chapter 2 describes the development of a novel bioinformatic method to identify these sites. The application of this method to data from C_4 species is described in Chapter 3, and the attempts to test the outputs of this method *in vitro* are described in Chapter 4. The ultimate goal was then to engineer advantageous sequence variants identified in the *in vitro* screen from Chapter 4 into the endogenous genes in C_4 plants such as maize using genome editing technologies. However, time constraints did not allow me to get this far.

The final chapter of this thesis (Chapter 5) describes a parallel project that was complimentary to the principal aim of my thesis. In this chapter I focused on addressing one of the largest gaps in our understanding of C_4 biochemistry, the identity of the bundle sheath cell chloroplast malate and pyruvate transporters. In this Chapter I used a combination of bacterial and *in planta* experiments to identify and characterise a single transporter gene that provides both the malate import and pyruvate export functions required for C_4 photosynthesis in the NADP-ME C_4 species *Setaria italica*. This discovery provides the first complete molecular parts list for any C_4 cycle and has formed the basis of a patent application (Niklaus M, AbuJamous B, and Kelly S, Membrane transport protein and uses thereof, Australian Provisional Patent Application No. 2019902940 (patent applied for)).

Chapter 2: A novel method for identification of significant differences in aligned amino acid sequences between two sets of species

2.1 Chapter introduction

The previous chapter outlines the motivation for exploiting natural variation within C_4 plants to create novel protein variants with enhanced biochemical characteristics to potentially improve C_4 photosynthesis. However, specific methods for analysing natural diversity to identify sites within protein sequences for targeted engineering do not exist. Instead, such sites are typically identified through analysis of positive selection, a method designed to detect differences in the relative rate of protein and nucleotide sequence evolution. However, there are limitations to the use of positive selection (discussed in the chapter below), and differences in the relative rate of sequence change do not necessarily equate to differences in sequence. This chapter addresses this methodological gap, and presents a novel method for identification of sites within amino acid sequences for targeted engineering.

2.1.1 Background

Large parts of our current understanding about the evolution of C_4 photosynthesis were generated through molecular evolutionary sequence analysis (Christin et al. 2009a; Christin et al. 2007; Christin et al. 2009b). These analysis utilised likelihood-based phylogenetic methods to identify genes, or sites within genes, that show signatures of selection. One of the most frequently used likelihood methods in the context of C_4 molecular evolution research is the branch-site model (Yang 2007). This model aims to identify genes or sites within genes that have evolved at significantly different rates across a phylogeny. In the case of application to C_4 species, this has been used to

identify genes or sites within genes that have been evolving faster in C_4 species than in C_3 species. Such differences in evolutionary rate between C_3 and C_4 species have been used to propose a series of cases of C_4 specific adaptive evolution (Christin et al. 2008b; Christin et al. 2007; Christin et al. 2009b; Huang et al. 2017).

Conventionally, the rate of protein evolution is measured using a metric called *dN/dS*. The *dN/dS*-value is calculated as the ratio of the number of non-synonymous nucleotide substitutions per non-synonymous site (these are substitutions at sites that result in amino-acid changes) (*dN*) divided by the number of synonymous nucleotide substitutions per synonymous site (these are substitutions at sites that do not result in amino-acid changes) (*dS*) (Pal et al. 2006). A ratio larger than one means that there are more non-synonymous changes than synonymous changes. This metric can indicate that there was some evolutionary pressure to alter and escape the original protein sequence (Pal et al. 2006).

In the context of understanding the convergent evolution of C_4 photosynthesis, it makes intuitive sense to search for genes and sites in genes under positive selection. This is because the evolution of C_4 photosynthesis required a substantial change in biochemistry, and thus it is plausible that such changes are associated with molecular sequence changes. Furthermore, the sites in amino acid sequences that have evolved faster in C_4 species are likely to be those that have some functional significance in the context of C_4 photosynthesis. Thus, if there are sites in sequences that have changed to better suit function in C_4 photosynthesis it is possible that analysis of positive selection could detect these. In addition, looking for evidence of positive selection across the entire genome may also facilitate the identification of whole genes that were previously unknown to play a role in C_4 photosynthesis (Huang et al. 2017).

Although the rate of selection informs us about the evolutionary pressure that acted on molecular sequences, it could potentially ignore sites that are interesting that changed in the protein sequence but whose rate of change relative to the rate of synonymous change is not high. Moreover, *dN/dS* analyses cannot analyse insertions or deletions nor assess whether the changes in amino acid sequence represent a fundamental change in the biochemical properties of the amino acids at a given site between C_3 and C_4 plants. For example, a site in a protein sequence that is alanine in all C_3 plants but is serine in all C_4 plants is a potential key biochemical difference between C_3 and C_4 plants irrespective of the rate at which that change occurred in the protein sequence relative to the rate at which the underlying DNA sequence has evolved. Positive selection will fail to detect such changes if the rate of change of the underlying nucleotide sequence is high. Thus, it is reasonable to propose that, in addition to investigating the relative rates of nucleotide and amino acid sequences in proteins, the sequences of the proteins themselves are directly compared to identify potential differences between C_3 and C_4 versions of genes.

Algorithms such as PAML that try to identify sites under evolutionary selection rely on analysis of single-copy genes and require single contiguous sequences for each gene. This is problematic in situations where two or more copies of a gene exist in a species and/or the underlying data is fragmented. Fragmented, sequence data is very common as RNA sequencing presents an attractive low-cost alternative to access the genetic resources of organisms. Unlike genomic sequences where full length gene sequences are frequently recovered, gene sequences assembled from transcriptome data are often fragmented or incomplete. Such errors are a result of biases in the process of transcriptome sampling (cDNA synthesis and fragmentation) and the expression level of a gene (Surget-Groba and Montoya-Burgos 2010). For example, low expressed

genes can fail to assemble, or assemble into fragments due to low read coverage. Similarly, sequences that are difficult for the reverse transcriptase or DNA polymerase to read (such as GC rich regions) can also cause the read coverage to fail to reach the required minimal cut-off value for assembly. Thus, discrete fragments from the same gene can be found in the final assembly that could lead to misinterpretations such as gene absence, true gene fragmentation, or even gene duplication. These fragmented genes cause issues in conventional software to analyse evolutionary selection such as PAML because fragmented sequences originating from the same species are treated as separate genes.

One way to make use of transcript sequences in a branch-site model style analysis would be to develop a method to identify and resolve fragmented sequences. To this end I developed a novel algorithm called Fragmatic for identifying and resolving fragmented transcripts present in *de novo* assembled transcriptomes. The Fragmatic algorithm relies on evidence from orthologous genes in related organisms to identify fragmented transcripts and merge them together. The applicability and effectiveness of the algorithm was evaluated on both simulated and real data. A complete description of the method including all results and discussion is provided in Appendix B. In brief, on simulated data, the algorithm achieved a 100% precision and ~20% recall rate. However, on real data the estimated precision was between 65% and 90%. Thus, although the method fixed some errors, these improvements were not sufficient to compensate for the additional errors introduced by the method and therefore the method was abandoned and not used further in this study. However, the method is included in the appendix of this thesis for posterity as it represented a substantial body of work. Interestingly, approximately one year after stopping work on the Fragmatic method Piližota et al. 2018 published a very similar approach. This similar method also

suffers from low recall and high rate of erroneous chimera formation (Piližota et al. 2018). Thus, although resolving fragmented sequenced in transcriptome data could improve transcriptome assembly, it is technically challenging.

In addition to being unable to deal with duplicated genes or gene fragments there are also other confounding factors that influence the detection of positive selection. For example, all implementations of the branch-site model assume that mutations occur independently at individual nucleotide sites (Venkat et al. 2018). However, in response to DNA damage, DNA polymerase can create multiple mutations at neighbouring sites during the same replication cycle. These so called multinucleotide mutations occur much more frequently than initially anticipated (Schrider et al. 2011), and violate the assumptions of positive selection analysis leading to false positive detection of positive selection. Furthermore, it was shown that these multinucleotide mutations are the primary source of positive selection identified by the branch-site model (Venkat et al. 2018), and adapting the branch-site model to take into account codons with multiple mutations eliminates the signal of positive selection of sites in the vast majority of genes (Venkat et al. 2018). It is therefore likely that some studies reporting the detection of positive selection based on the branch-site model have suffered from multinucleotide mutation bias and that some sites previously identified as being adaptive for C_4 evolution are false positives.

Given that analysis of positive selection cannot deal with duplicated genes, cannot deal with fragmented transcripts obtained from RNAseq, suffers from a high rate of false positives, and cannot actually assess whether there is a difference in protein sequence between C_3 and C_4 species (but rather assesses whether the underlying sequences have evolved at different rates), we sought to develop an alternative approach to identify sites within protein sequences that are different between C_3 and C_4 species for

targeted engineering. The remainder of this chapter describes the construction and evaluation of this method.

2.2 Results and discussion

2.2.1 The total variation distance metric

The problem to be solved by the desired algorithm is to analyse a multiple alignment of protein sequences to identify sites within that alignment where the identity of the amino acids found in C_3 versions of the protein differ from those found in C_4 versions of the protein. As C_4 photosynthesis evolved from multiple independent origins of C_3 photosynthesis, and as these independent origins have been diversifying for millions of years, it is possible that there will be natural variation in both the C_3 and C_4 versions of the gene. Thus, the method should be able to detect differences in amino acid sequences in the presence of natural variation. More specifically, the method should be designed to detect differences in the distribution of amino acids found in C_3 and C_4 versions of the gene.

A commonly used metric to detect the difference in the distribution of discrete objects between two groups is known as the total variation distance metric. An example of how this metric works is provided below and shown in Figure 2.1.

Let $\boldsymbol{\Omega}$ denote the complete set of amino acid residues observed across all sequences in a given column of a multiple sequence alignment. Let the variable $\boldsymbol{x}$ represents a single amino acid from set $\boldsymbol{\Omega}$. Let $\boldsymbol{\mu(x)}$ represent the probability of observing the amino acid $\boldsymbol{x}$ in the sequences from C_4 species in the given column of the multiple sequence alignment. Let $\boldsymbol{\beta(x)}$ represent the probability of observing the amino acid $\boldsymbol{x}$ in the sequences from C_3 species in the given column of the multiple sequence alignment. Both $\boldsymbol{\mu(x)}$ and $\boldsymbol{\beta(x)}$ are approximated as the number of times amino acid $\boldsymbol{x}$ is observed

as a proportion of the total number of amino acids in that column in the set of C_4 and C_3 species respectively. The TVD between the two probability distributions ***μ(x)*** and ***β(x)*** on a sample space **Ω** can then be defined as follows (Levin et al. 2009) (see Equation 1):

$$\mathbf{TVD} = \frac{\mathbf{1}}{\mathbf{2}} \sum_{\boldsymbol{x}\epsilon\Omega}^{\boldsymbol{n}} |\mu(\boldsymbol{x}) - \boldsymbol{\beta}(\boldsymbol{x})| \qquad \text{(Equation 1)}$$

This metric has an upper limit of 1 and a lower limit of 0. That is, if the distributions of amino acids at a given site ***n*** in a multiple sequence alignment of an orthologous protein between the two groups are identical, the TVD has a value of zero. If, however, the distributions are completely different, a value of one is obtained. This metric can be applied independently to each site ***n*** in a multiple sequence alignment of orthologous proteins.

In order to determine whether or not the difference in the distribution of amino acids is greater than would be expected by chance a permutation test can be conducted. Here, the size of each group is kept the same and the set of amino acids observed in the column are randomly shuffled and the TVD metric is recomputed. This process is repeated 5000 times and the number of times a TVD score greater or equal to the original (unshuffled) TVD score is recorded. This, expressed as a proportion of the number of trials is taken as the probability that such a difference is observed by chance. Sites that are significantly different between C_3 and C_4 species can then be identified as those with a p-value ≤ 0.01.

2.2.2 Definition of groups and subgroups

The aim of the algorithm, in this instance, is to identify sites in multiple sequence alignments that differ between C_3 and C_4 species. Thus, for a given gene (e.g. PEPC) all gene variants that belong to C_3 species belong to the C_3 group and all gene variants that belong to C_4 species belong to the C_4 group. Within each group subgroups are created to account for phylogenetic biases in this data. For example, all sequences belonging to an independent origin of C_4 photosynthesis are considered one subgroup. Similarly, all C_3 species that are an outgroup to that C_4 origin are also considered a single subgroup. All subgroups are weighted equally in the TVD metric. For example, if one independent origin of C_4 photosynthesis has ten representative sequences and a different independent origin has two representative sequences then the values for sequences for origin one are multiplied by 0.1 and the values for the sequences in origin 2 are multiplied by 0.5, such that the total value attributed to the set of sequences is the same. Thus, the proportions per subgroup always sum to one and are independent on the number of species present in the subgroup. The amino acid proportions per subgroup are subsequently integrated for the corresponding trait group (i.e. C_3 or C_4 photosynthesis).

2.2.3 Algorithm overview

The algorithm requires a pre-computed multiple sequence alignment of protein sequences in FASTA format. The algorithm also requires two CSV files that specify the names of the species that are in group one and group two, respectively. The algorithm processes each column of the multiple sequence alignment independently (Figure 2.1). For each column, it calculates the distribution of amino acids for each

group, the similarity between the two groups is then quantified using the total variation distance (Figure 2.1 D and Equation 1).

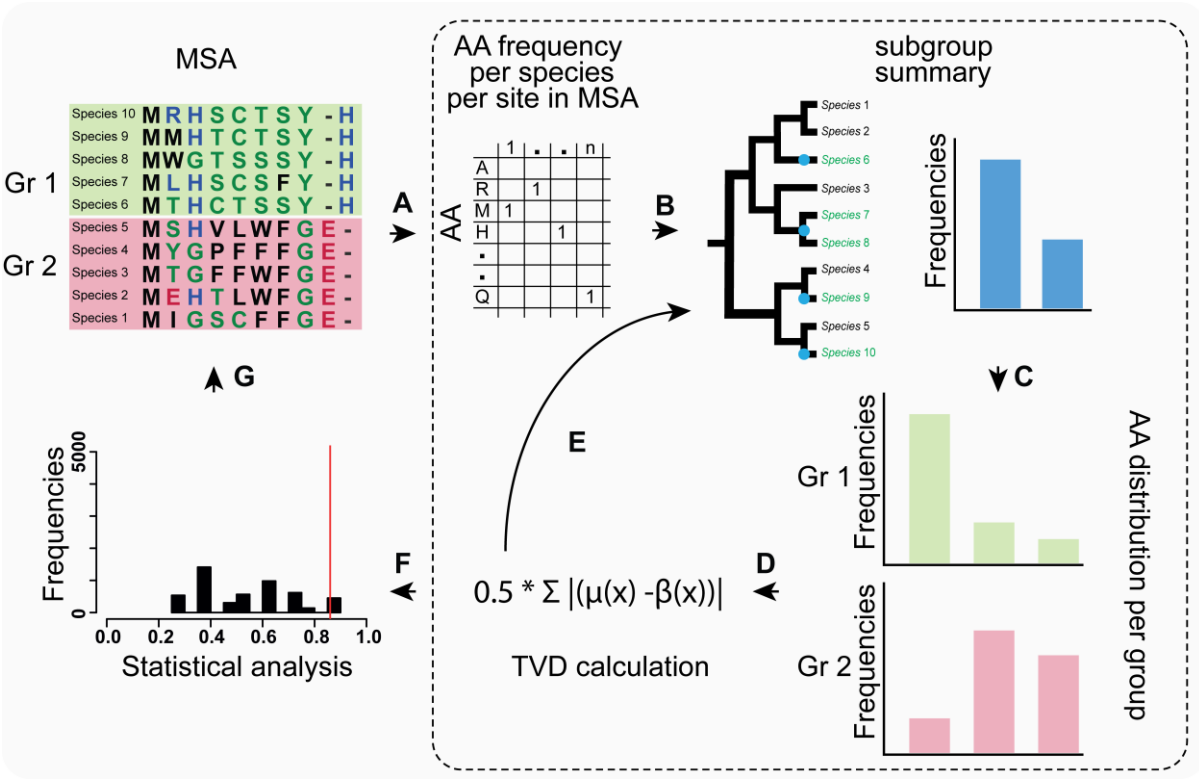


Figure 2.1: Workflow of the Total Variation Distance algorithm **A)** Recording the amino acid frequency per species in the multiple sequence alignment **B)** Summarize the amino acid distribution for each subgroup to take into account the multiple sampling of the same evolution event. The independent evolution of the trait is marked with light blue dots on the phylogenetic tree **C)** Summarize the amino acid distribution for the two pre-specified groups **D)** Use the two distributions to calculate the total variation distance **E)** Permuting across the dataset without replace and recalculate the amino acid distribution for each group **F)** Compare the observed total variation distance to all values obtained during the permutation and compute a p-value **G)** Move to the next column in the multiple sequence alignment file and start the same analysis cycle

The Total Variation Distance algorithm calculates the total variation distance with the conventional amino acid alphabet and an extended amino acid alphabet that considers gapped characters that are flanked both at the *N*-terminus and *C*-terminus with amino acids. Considering so-called bounded gaps as true characters allows the identification of insertions and deletions in the protein sequence between the two groups. A third alphabet is applied in parallel that groups the amino acids according to their biochemical properties (acidic, basic, hydrophobic, neutral and polar) rather than their exact sequence.

At the end, the user is provided a CSV file that summarizes every column in the multiple sequence alignment. Among the summarized output information are the distributions of the amino acids within each group, the total variation distance, the p-value and the number of informative columns per group. Informative columns are columns that contain either an amino acid or an indel.

2.2.4 Statistical assessment of the analysis pipeline to measure the amino acid distribution in a multiple sequence alignment between two groups

The Total Variation Distance algorithm was tested on a simulated dataset that contained various levels of convergence and noise (see Dataset simulation 2.4.1). The simulated dataset contains ten species and the hypothetical phylogenetic relationship between the ten species is depicted in Figure 2.2. Species 1 to 5 are grouped in the first group and species 6 to 10 are grouped in the second group. Species 1 and species 2 in group 1 and species 7 and species 8 in group 2 have a common origin of the trait of interest and are thus put into the subgroup within their corresponding group (see Algorithm overview 2.2.3). Creating subgroups within groups allows the algorithm to

account for species that obtained the trait of interest at the same evolutionary time point, thus reducing the sampling bias.

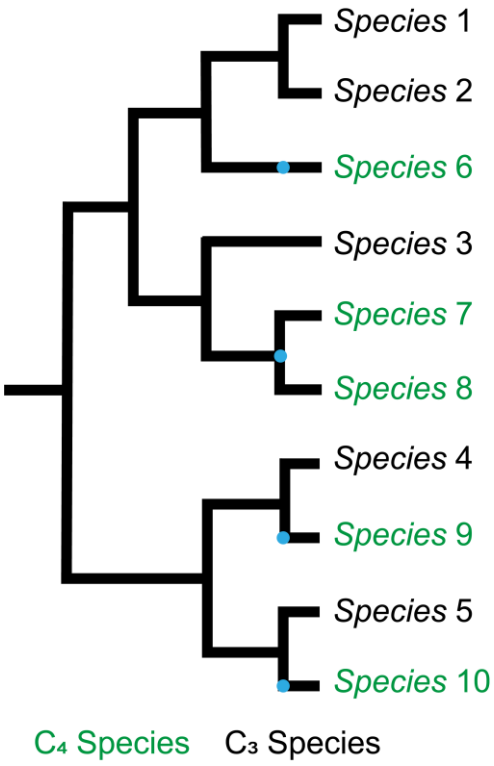


Figure 2.2: Phylogenetic relationship between the simulated plant species belonging either to the group of C_3 or C_4 plants. Species 1-5 have the C_3 trait whereas the Species 6-10 show the C_4 trait. The independent evolution of the C_4 trait is marked with blue dots on the phylogenetic tree.

The calculated total variation distances for each site are shown in Table 2.1. The output of the Total Variation Distance algorithm shown in Table 2.1 is obtained using the amino acid alphabet considering bounded gaps and the biochemical properties. The values of the TVD are bound between zero and one (Equation 1).

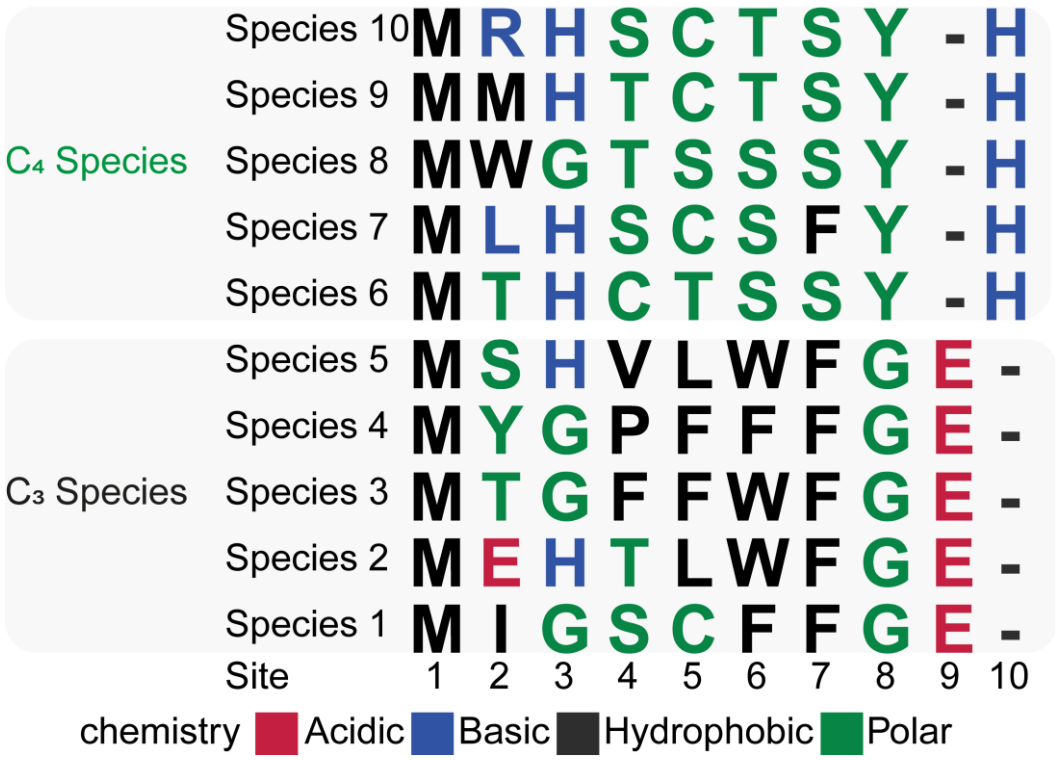


Figure 2.3: Protein alignment of the simulated sequences of the two groups (C_3 and C_4 species) used to evaluate the Total Variation Algorithm.

To highlight the scope of the TVD, a couple of the TVD values should be emphasized (Table 2.1). The resulting TVD value for the first position is 0 indicating that the two groups have an identical amino acid distribution at that site. The second site has a TVD of 0.75 indicating a high difference between the two amino acid distributions. However, upon closer inspection of the amino acid distribution of the second site in the alignment, it becomes apparent that the within group variability is also high. Therefore, it is questionable whether such a high TVD is indicative for biological relevance and whether changes at this site have much impact on the biological function of the protein. Accordingly, it obtains a non-significant p-value in the resampling tests (Table 2.2). Site eight represents a complete switch from glycine in the C_3 group to tyrosine in the C_4 group. Tyrosine and glycine are both polar amino acids and thus the TVD

considering the biochemical property is 0 whereas considering the conventional alphabet the TVD is 1.

The ninth position in the alignment reveals that all C_4 species have a gap character. The developed algorithm considers a gap bounded between amino acids as a true character and thus calculating a TVD of 1. Considering bounded gaps allows the identification of insertion or deletions between the two groups under evaluation. At position ten the gap character is not classified as a true character. Only true characters are considered in the calculation of the total variation distance (see Equitation 1). As a consequence, the resulting TVD value is 0.5.

Table 2.1: The total variation distance for each position in the simulated multiple sequence alignment using the amino acid alphabet considering gapped characters (TVD) and the reduced amino acid alphabet considering the biochemical properties (TVD_{BioCh})

Position	1	2	3	4	5	6	7	8	9	10
TVD	0	0.75	0.5	0.75	0.87	1	0.87	1	1	0.5
TVD_{BioCh}	0	0.62	0.5	0.75	0.87	1	0.87	0	0.5	0.5

It becomes clear that the TVD value in itself is not sufficient to evaluate whether a site in a protein between two groups is different. To evaluate how likely, it is to come across the observed TVD value just by chance, the algorithm is coupled to a permutation test. The permutation test is a resampling technique that creates a distribution based on all the observed values at a given site (see Algorithm overview 2.2.3). If the null hypothesis is true, then random shuffling of the amino acids at a given site should obtain similar TVD scores. If the alternative hypothesis is true, then random shuffling of the data should rarely or never recover a TVD score as high as the one observed in the real data. The number of reshuffle and resampling tests required for robust p-value estimation was determined using simulation by testing various numbers of

permutations including all possible combinations (50, 100, 1000, 2000, 5000, 10 000, 30240) (Table 2.2 and Table 2.3).

Table 2.2: Effect of varying the number of permutations on the estimate of the p-value for each site (1-10) in the multiple sequence alignment. The TVD value for each site was calculated using the extended amino acid alphabet considering bounded gaps.

	Site	1	2	3	4	5	6	7	8	9	10
	TVD	0	0.75	0.5	0.75	0.875	1	0.875	1	1	0.5
Permutations	50	1	1	0.28	0.8	0.12	0.06	0.06	0.06	0.06	0.06
	100	1	1	0.27	0.87	0.09	0.03	0.03	0.03	0.03	0.03
	1000	1	1	0.25	0.83	0.08	0.02	0.022	0.02	0.02	0.02
	2000	1	1	0.27	0.84	0.099	0.03	0.033	0.03	0.03	0.03
	5000	1	1	0.26	0.83	0.09	0.03	0.03	0.03	0.03	0.03
	10000	1	1	0.26	0.83	0.089	0.03	0.031	0.03	0.03	0.03
	30240	1	1	0.25	0.83	0.085	0.03	0.03	0.03	0.03	0.03

Testing different numbers of permutations allows to evaluate the number of re-samplings necessary in order to identify sites in a protein that are most likely to be different between the two groups under evaluation. The p-values obtained with 2000 permutations are a good approximation when compared to the p-values generated with all possible permutations (30240) (Table 2.2 and Table 2.3).

Considering that real life datasets are likely to be larger and noisier it may be necessary to increase the number of permutations. Thus, 5000 permutations was selected. Also, there is not much gain in precision by doubling the number of permutations to 10 000 but the computational time would be twice as long. Thus, the default number of permutations in the Total Variation Distance algorithm is set to 5000. If necessary, the operator has the option to alter the number of permutations.

Table 2.3: Different number of permutations and the corresponding p-value for each site (1-10) in the multiple sequence alignment. The TVD value for each site was calculated using the biochemical characteristics of the amino acids.

	Site	1	2	3	4	5	6	7	8	9	10
	TVD	0.000	0.625	0.500	0.750	0.875	1.000	0.875	0.000	0.500	0.500
Permutations	50	1	0.320	0.240	0.140	0.020	0.020	0.020	1	0.020	0.020
	100	1	0.250	0.230	0.210	0.040	0.040	0.040	1	0.040	0.040
	1000	1	0.269	0.271	0.151	0.038	0.038	0.038	1	0.038	0.038
	2000	1	0.259	0.257	0.148	0.033	0.033	0.033	1	0.033	0.033
	5000	1	0.264	0.256	0.145	0.028	0.028	0.028	1	0.028	0.028
	10000	1	0.255	0.257	0.140	0.028	0.028	0.028	1	0.028	0.028
	30240	1	0.26	0.26	0.14	0.03	0.03	0.03	1	0.03	0.03

To evaluate how the 5000 TVD values generated during the permutation test compare to the actual observed TVD, the frequency of each permutated TVD for each site was plotted (Figure 2.4). The actual observed value for each site is indicated with a red line. None of the TVD values generated for the first site is larger or smaller than the observed value of zero (Figure 2.4 A). The first site in the multiple sequence alignment for each species has a methionine and shuffling the identical amino acid across the two groups does result in an identical TVD.

The sites two to four have varying levels of within and between group variability (Figure 2.4 B-D). All these sites have in common that the observed TVD value is indicating a difference in amino acid distribution between the two groups. The permutation test subsequently reveals that for these sites the within group variation is larger than the difference between the two groups and therefore the difference can be explained just by chance. Thus, these sites are not considered to be statistically different between the two groups.

The sites five to seven have different amino acid distributions between the two groups. Unlike the previous sites, the within group variation is relatively small in these examples. This is reflected in the obtained TVD distribution. Most of the obtained TVD values during the permutation test are smaller than the observed TVD and consequently indicating that the observed difference is likely to be a true difference (Figure 2.4 E-H). The sites nine and ten look similar in the multiple sequence alignment (Figure 2.3) but the algorithm distinguishes between bounded gaps (site nine), which are flanked by other amino acids and terminal gap characters (site 10). The bounded gap is considered as a true character because it could reveal an insertion or deletion. Unbounded gaps could be caused by transcript fragments and thus are ignored. The subsequent permutation test and the plotted distribution for the sites nine and ten reveal that it is unlikely that these differences are explained by chance alone (Figure 2.4 I and J).

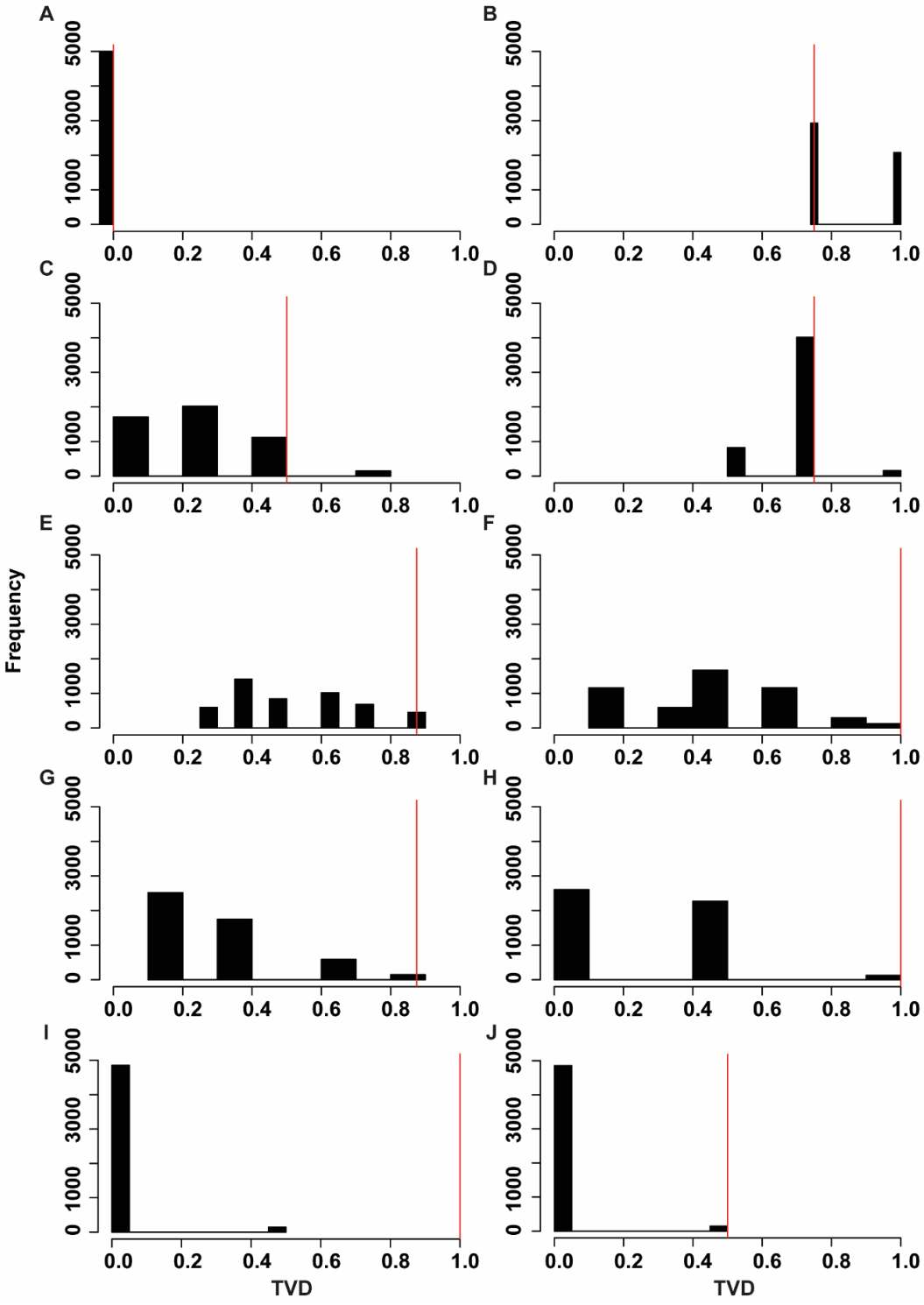


Figure 2.4: Distribution of the TVD values obtained during the permutation test. Sites one to ten of the multiple sequence alignment file correspond to A-J, respectively. Red line indication the actual observed TVD value.

2.2.5 Comparison between known sites under selection in C_4 PEPC and sites identified by the Total Variation Distance algorithm

To test the newly developed algorithm for its capability to identify sites in proteins that are different between C_3 and C_4 species I used a dataset of grass phosphoenolpyruvate carboxylase (PEPC) protein sequences that was published by Christin et al. 2007. The original analysis by Christin et al. 2007 revealed that twenty-one amino acids evolved under positive selection in the majority of the C_4 PEPCs. The publicly available dataset consists of 130 sequences. Among these are fifty-eight C_4 PEPC sequences. These 130 PEPC sequences covered a segment from exon 8 to exon 10, which encompasses a region that was already know to harbour some major C_4 determinants (Bläsing et al. 2000). This segment translates into an amino acid sequence that corresponds to the position 462 to 910 (position corresponds to *Z. mays* PEPC CAA33317).

The Total Variation Distance algorithm identified 48 sites with a p-value ≤ 1% that revealed a different amino acid distribution between the C_3 and C_4 species in the Christin et al. 2007 dataset. To assess the impact of using the TVD as an additional filter criterion the numbers of sites passing the increasing TVD filter criteria were counted (Figure 2.5A). Interestingly, the alanine to serine substitution was the only site that revealed a TVD value greater than 0.9 (Figure 2.5 A). To determine whether there is a difference in the TVD value distribution between the twenty-one sites identified by Christin et al. 2007 and all the significant sites identified by the Total Variation Distance algorithm the corresponding TVD values were recorded and plotted. The average TVD value for the sites identified by Christin et al. 2007 was ~0.7 whereas the average TVD value of all the significant sites of the Total Variation Distance algorithm was ~0.6

(Figure 2.5 B). Thus, for the subsequent analysis and illustrative purposes a TVD cut-off value of 0.7 was implemented.

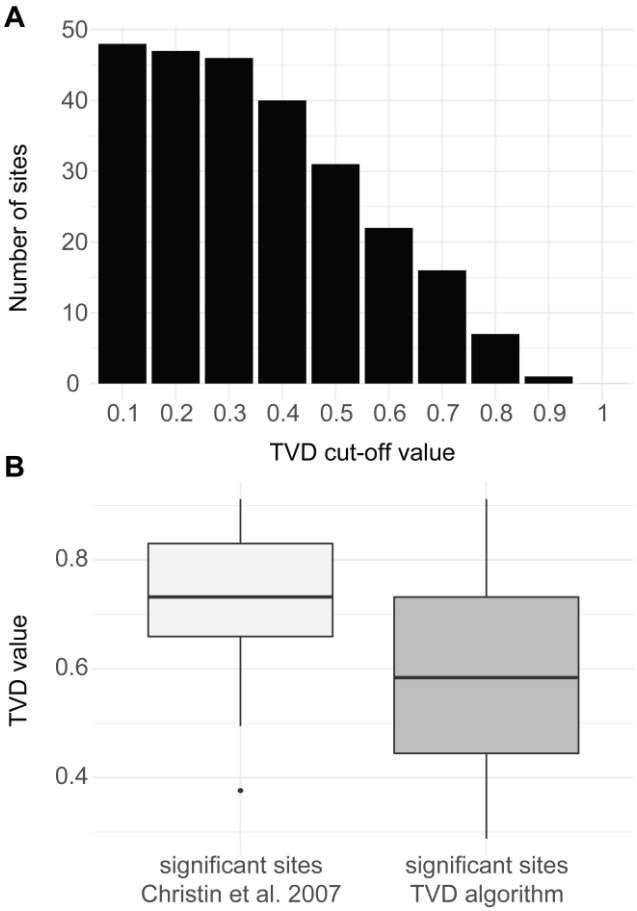


Figure 2.5: Characterisation of the TVD value as an additional filter criterion **A)** Application of various levels of TVD cut-of values and the number of significant sites passing the filter criterion **B)** Boxplot representing the distribution of the TVD values of the significant sites identified by Christin et al. 2007 and all significant sites identified by the Total Variation Distance algorithm.

Applying this additional cut-off value, the Total Variation Distance algorithm identified sixteen sites that have a different amino acid distribution between the two groups (Figure 2.6 A) considering the independent evolution of C_4 photosynthesis in the dataset. Among the sites identified was the site 780 (position corresponds to *Z. mays* PEPC CAA33317) here a change from alanine in C_3 plants to serine was observed in

most C_4 plants (Figure 2.6 A). This site is known to have an impact on malate tolerance of C_4 PEPC (Chapter 1.4). Among the other recognised sites, fourteen sites correspond with the sites previously identified by Christin et al. 2007 (Figure 2.6 B).

None of the identified sites correspond directly to sites that are known to have functional relevance of the C_4 PEPC (Figure 2.6 C). It is important to point out, however, that the identified sites are usually located in close proximity of sites that have a functional relevance. For example, a subset of the identified sites is located in close proximity of the known functional areas that contribute to the PEP binding and the hydrophobic binding pocket. The site 560 reveals a change from an arginine, an amino acid with a basic biochemical characteristic, to mostly proline, a hydrophobic amino acid in the C_4 PEPC. A similar observation was encountered at the site 596 where a switch from an acidic amino acid, glutamic acid, to a mostly neutral glutamine occurred. It seems a functional different amino acid at both the sites (560 and 596) is favoured and the majority of C_4 PEPC variants appear to have made that change. The site 599, which is located next to one of the PEP-binding sites, underwent a change from an isoleucine in the C_3 PEPC to a valine in the C_4 PEPC. Both amino acids are classified as hydrophobic but a potential explanation for the change could be a structural hindrance caused by isoleucine ($C_6H_{13}NO_2$). Valine ($C_5H_{11}NO_2$) is a smaller amino acid and therefore might be more favoured over isoleucine at that site. The site 761 is also in close proximity to various important functional sites such as PEP-binding and HCO_3^- binding. Here, a change from a polar amino acid, serine, in C_3 PEPC to a hydrophobic amino acid, alanine, occurred in the vast majority of C_4 PEPC.

Comparing the output obtained from analysing the PEPC dataset from Christin et al. 2007 with the Total Variation Distance algorithm to sites that have been previously identified in the C_4 PEPC to be under positive selection shows that most of them are

shared by the two analyses (~66%). Furthermore, mapping the identified sites onto the protein sequence and comparing them with the known functional sites makes it clear that these identified sites are clustered closely to sites of known functional relevance. For example, four out of the sixteen identified sites are less than ten amino acids away from residues known to be involved in the PEP-binding site (466,560 ,596 and 599). Similarly, the tetramer, aspartate/malate and the HCO_3^- binding site each have one site in close proximity too (502 and 761, respectively). A permutation test (where the null hypothesis assumed that the sites were randomly distributed) revealed that the probability of finding four sites in proximity of ten or less amino acids to any of the PEP binding sites is 4.54×10^{-4}. Therefore, it is unlikely that the position of these sites is due to chance alone. The concordance of the Total Variation Distance algorithm with previous positive selection analysis gives us confidence that it is able to identify sites that are firstly, potentially functionally relevant, and secondly, are different between the two groups under study.

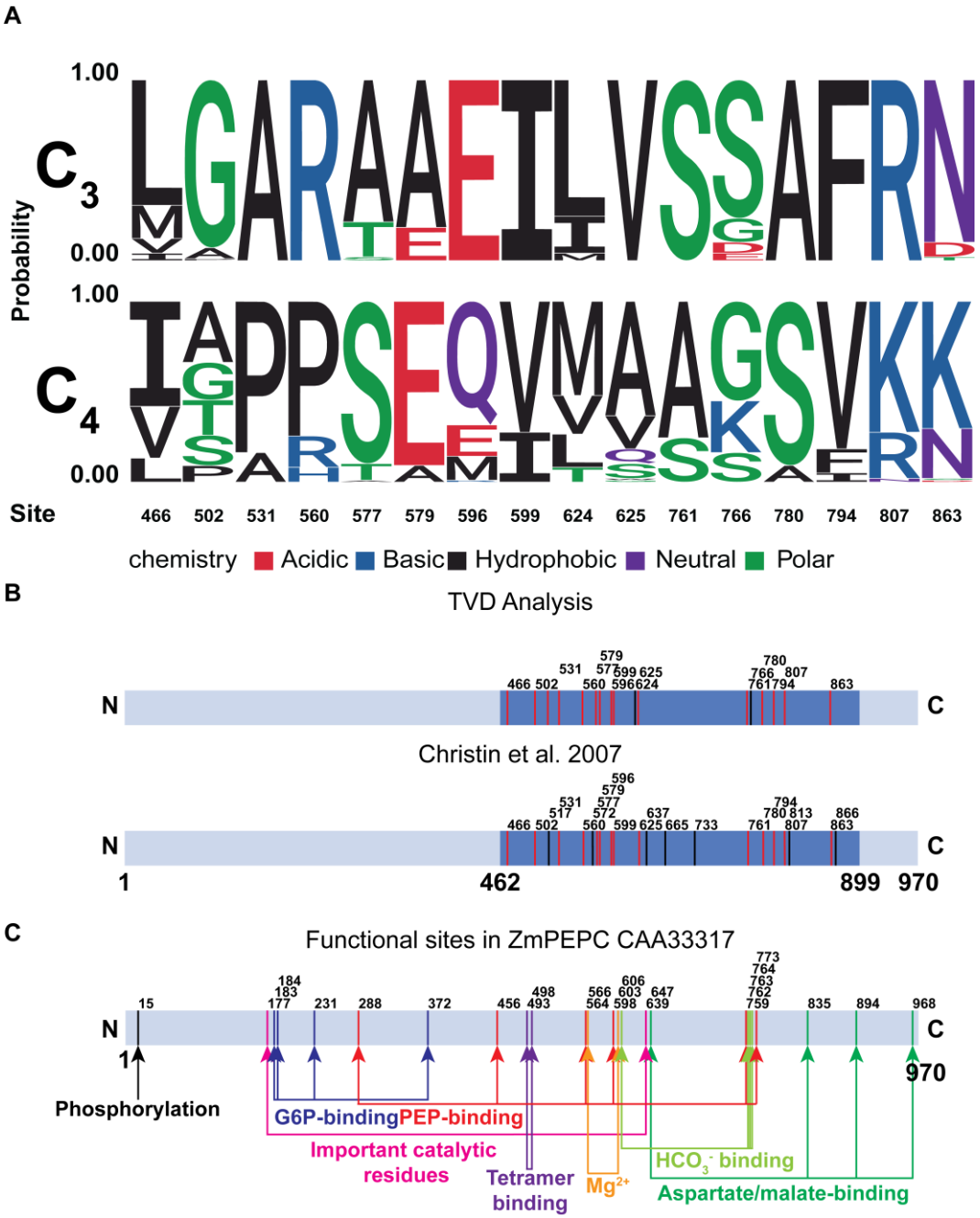


Figure 2.6: Comparison between sites identified as under positive selection (Christin et al. 2007) and sites identified by the Total Variation Distance algorithm relative to known functional sites in C_4 phosphoenolpyruvate carboxylase (PEPC). **A)** Sequence logo showing the diverse amino acid distribution between the two groups (C_3 and C_4 plants) at each significant site in the multiple sequence alignment. **B)** The full length of the PEPC protein consists of 970 amino acids (light violet). The analysis by Christin et al. 2007 evaluated a stretch of the protein

between the position 462 and 910 (dark violet). The Total Variation Distance algorithm identified sixteen sites indicated with red and black vertical bars. Christin et al. 2007 discovered twenty-one sites under positive selection indicated with red and black vertical bars. The sites that agree between the two sites are marked in with red bars and the sites in disagreement are marked with black bars. **C)** Functional sites in ZmPEPC highlighted in colour and labelled correspondingly (Jacobs et al. 2008; Kai et al. 2003; Matsumura et al. 2002; Takahashi-Terada et al. 2005).

The seven sites that were identified as being under positive selection but were not identified by the Total Variation Distance algorithm were further explored. The amino acid distribution of these seven sites reveals changes in the distribution of amino acids between the two groups (Figure 2.7). Moreover, the statistical evaluation of the sites (517, 572, 637, 733 and 813) show that these sites are significantly different between the two groups (p-value < 1%) but crucially the TVD values are smaller than the required cut-off value of ≥ 0.7 (Table 2.4). The sites 665 and 866 are not considered significantly different between the C_3 and C_4 group and have a TVD smaller than the predefined cut-off value (Table 2.4).

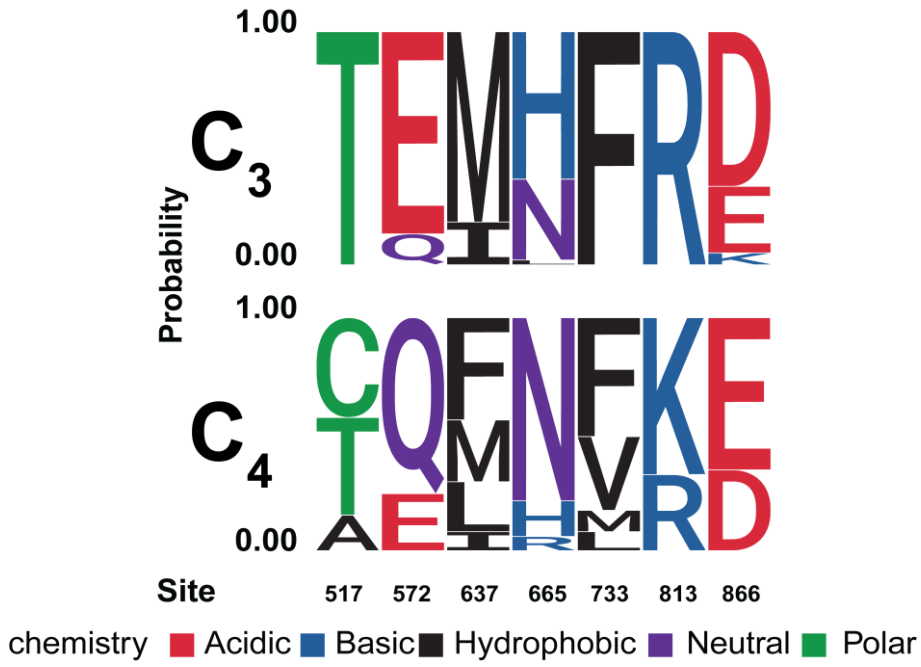


Figure 2.7: Amino acid distribution at sites identified as under positive selection by Christin et al. 2007 but not included by the Total Variation Distance algorithm. The sequence logo reveals the difference in amino acid distributions between the C_3 and C_4 plants.

Table 2.4: The summary of the statistical evaluation of the seven sites that were identified as under positive selection but not identified using the Total Variation Distance algorithm.

Site	**517**	**572**	**637**	**665**	**733**	**813**	**866**
TVD	0.58	0.63	0.66	0.49	0.50	0.68	0.38
p-value	$2.00E^{-04}$	$2.00E^{-04}$	$2.00E^{-04}$	0.0102	$6.00E^{-04}$	$2.00E^{-04}$	0.0498

2.2.6 Extension of the analysis pipeline to consider transcript abundance and dealing with duplicated genes

As the above example was performed on a dataset generated for positive selection analysis it only contained a single representative PEPC gene per species. However, this masks the true situation for most gene families where many species express multiple transcript variants arising either from alternative splicing or from multiple discrete copies of the same gene. As the Total Variation Distance algorithm was built

to have a group structure that can weight sequences to the overall contribution of the group, the method can be easily extended to account for multiple gene variants with different transcript abundance (see Appendix A.1 and Table A.2). The consideration of the transcript abundance is important because duplicated genes played an important role during the evolution of C_4 photosynthesis (see Chapter 1.5). Moreover, these duplicated genes can undergo neofunctionalization whereby one protein variant maybe more abundant and thus responsible for forming the trait of interest (e.g. see Chapter 1.5). Thus, considering all transcript variants and accounting for the expression levels of the individual transcripts (TPM) has the potential to improve the Total Variation Distance algorithm.

The simplest way to incorporate multi gene or multi transcript isoform data is to weight the contribution of each variant in proportion to the transcript abundance of that variant. A worked example is shown in Table A.2. Here, species 8 contains two variants of the same gene called “gene 1” and “gene 2” (Table A.2). In this case, the mRNA for gene 1 is twenty times as abundant as gene 2. This difference in abundance is considered when the frequency of amino acids at position **n** in a multiple sequence alignment is calculated per species (Figure 2.8). If, however, no expression data is present, the analysis pipeline gives equal weight to overlapping transcripts of the same species. The weighted amino acid distribution has an impact on the calculation of the total variation distance. For example, the site six and nine have a higher TVD value (Table 2.5) compared to the same dataset considering equal expression (Appendix A Table A.1). That illustrates the ability of the algorithm to deal with overlapping and duplicated genes and yet using the transcript abundance to control for overestimation or underestimation of a certain protein variant.

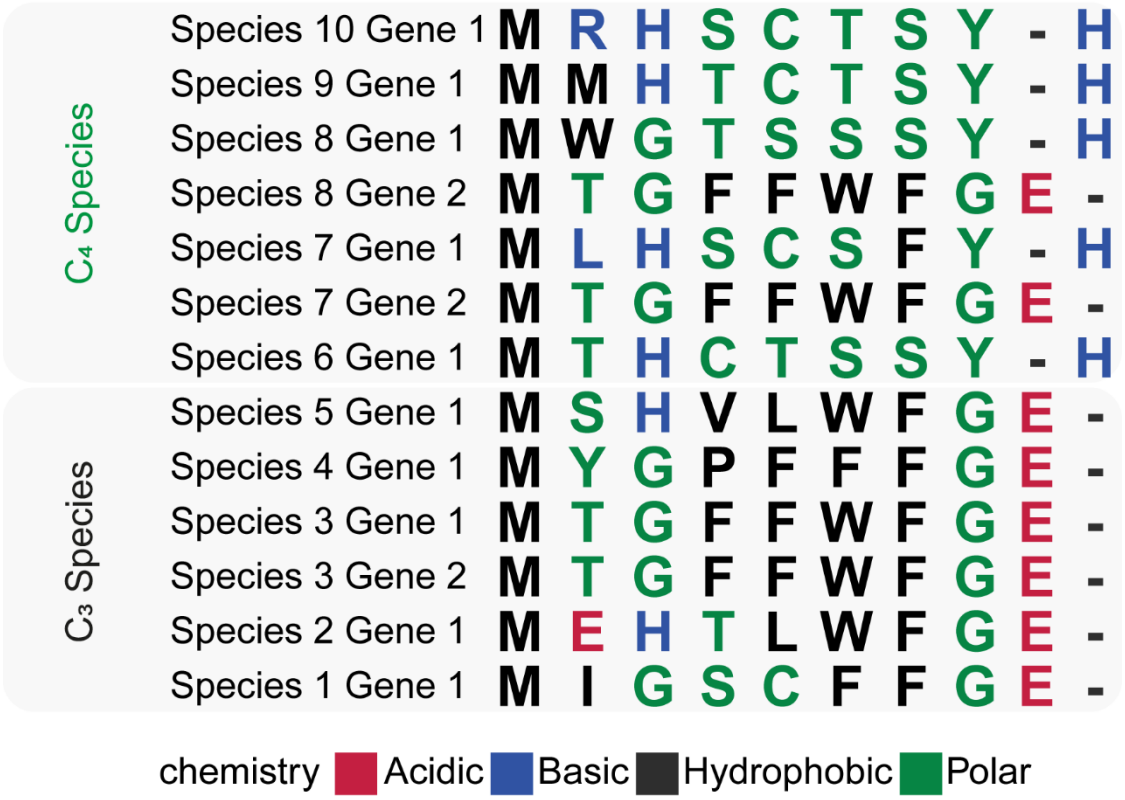


Figure 2.8: Multiple sequence alignment containing overlapping sequences originating from the same species.

Table 2.5: Summary of the TVD algorithm output considering expression levels.

Position	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
TVD	0	0.75	0.49	0.74	0.86	0.99	0.86	0.98	0.98	0.5
p-value	1	1	0.26	0.82	0.09	0.03	0.03	0.03	0.03	0.03
TVD_{BioCh}	0	0.61	0.49	0.74	0.86	0.99	0.86	0	0	0.5
$p\text{-value}_{BioChem}$	1	0.26	0.25	0.14	0.03	0.03	0.03	1	1	0.03

2.3 Chapter conclusion

The overall aim of this chapter was to develop a method that could analyse multiple alignments of protein sequences to identify sites in C_4 enzymes for targeted mutagenesis. The underlying premise for this is that C_4 photosynthesis is a relatively recent trait with multiple independent origins all of which cannot be equally optimised

for C_4 function. Thus, comparison of natural variation in protein sequence alignments of genes in C_3 and C_4 species may identify those sites that are potentially important for C_4 function and could be further improved through targeted mutagenesis approaches. As there is a large resource of transcriptome data for C_3 and C_4 species the method needed to be able to account for fragmented genes, transcript variants and incorporate mRNA abundance.

In this chapter I presented the Total Variation Distance algorithm, a method for identifying sites in multiple sequence alignments that are different between two sets of sequences. The method revealed to be very sensitive and is able to detect rare variants as well as those that are widely occurring. Thus, in addition to the p-value cut-off the introduction of TVD cut-off value was assessed. On a real dataset of direct relevance (the comparison of C_3 and C_4 PEPC genes in grasses) the method identified the majority of sites that had been previously identified through conventional molecular evolution approaches.

The Total Variation Distance algorithm is run as a R script with a multiple sequence alignment file in FASTA format as input. Furthermore, it requires two CSV files that specify the species within each group and within each file the user has the possibility to further specify subgroups to account for multiple sampling of the same evolution event. Optionally, the algorithm can include the transcript abundance for which a path to the corresponding TXT file is required.

2.3.1 Limitations and potential extension of the Total Variation Distance algorithm

Correcting for subgroup size

To correct for multiple sampling from the same evolution event the Total Variation Distance algorithm considers so-called subgroups. A subgroup comprises a group of species that emerged from the same evolution event. The Total Variation Distance algorithm records the amino acid distribution per species and subsequently summarises the distribution for each specified subgroup. Ideally, each subgroup contains the same number of species but that is usually not the case. Unfortunately, the current version of the algorithm does not consider the size of each individual subgroup. The consequence of omitting variation in subgroup size is that an individual species in a large subgroup has less contribution to the overall amino acid distribution recorded than a species belonging to a subgroup with only a few members. This issue could be addressed by using a similar approach to a weighted arithmetic mean to correct for variances in group size.

Considering the branch length of each independent evolution event

Another limitation of the Total Variation Distance algorithm is that it does not consider the estimated time point of the gain of the evolved trait. This is relevant for example in the context of the evolution of C_4 photosynthesis. There is a large discrepancy in the amount of time different C_4 lineages had to evolve and optimize their photosynthesis cycle. Thus, considering a branch length correction of each independent evolution event could refine the output of the Total Variation Distance algorithm. Allocating more weight to lineages that acquired the trait of interest for a longer time period may help identify additional potential sites for mutagenesis. This implementation could be useful

because some of the oldest C_4 lineages do contain some of the most productive plants (see Chapter 1.2)

Implementing the consideration of concurrent amino acid changes

Another limitation of the algorithm that surfaced during the later stage of the development is the inability to highlight whether amino acids that appear to be different between the two groups under evaluation require other amino acids to change concurrently or whether they are mutually exclusive. In other words, the algorithm considers all sites independently. Sites that undergo changes in amino acid distribution that are far apart in the linear sequence could potentially be in close proximity once the protein is folded and assembled. Knowing whether certain types of amino acids change together simplifies the combinatorial complexity later on when investigating different protein variants in an experimental evaluation.

Harnessing the within group variation

A second approach to extend the analysis pipeline is to consider the within group variation. For example, there is large variation of photosynthetic efficiency within C_4 species (Chapter 1 Figure 1.2 C). Thus, an extension of the algorithm that considers the within group variation and allocates different weights to the identified variants at a given (i.e. enzymes from species with higher photosynthetic efficiency could be weighted higher) site could narrow down the number of amino acids to test and thus simplifying the subsequent experimental evaluation.

2.3.2 Summary

This chapter presents a novel algorithm that was developed to identify sites in proteins that are different between two groups. The reason for the development of such an algorithm is because traditional methods that are used to identify sites under selection in protein coding genes have some shortcomings regarding the analysis of fragmented and duplicated proteins. In addition, it appears that some of the underlying assumptions of the statistical framework within these traditional methods do not hold true. Thus, an algorithm was developed that is both insensitive to fragmented and duplicated proteins within an orthogroup, as well as being able to consider the transcript abundance. The algorithm relies on the total variation distance to assess whether the amino acid distribution is different between the two groups under study. The benefit of this statistical metric is that it does not rely on the rate of synonymous and non-synonymous substitutions to assess whether evolution is acting on a given site.

The Total Variation Distance algorithm was evaluated on simulated data and real-life data. The evaluation showed the ability of the algorithm to identify sites between two groups that are different. Also, the evaluation showed that the algorithm is insensitive to sites that appear to be different between the two groups but suffer from a large within group variation. The application of the Total Variation Distance algorithm on a PEPC dataset revealed the capability to identify sites that were previously known to be under positive selection. Although assignment of positive selection has become contentious topic in recent times, this comparison served as a reassuring control, moreover the identified sites appear to be in close proximity to sites that are known to be functionally relevant. A subsequent statistical analysis revealed that it is unlikely to identify sites in such close proximity of functional sites just by chance alone.

The next chapter is going to apply the Total Variation Distance algorithm on a dataset that consists of thirty C_4 plants and seventeen C_3 plants with the aim to identify target sites to improve C_4 protein function through targeted mutagenesis.

2.4 Material and Methods

2.4.1 Dataset simulation

To test the applicability of the Total Variation Distance algorithm an illustrative artificial dataset was generated. The artificial dataset consists of ten species of which species 1 to 5 belong to the first group (C_3 species) and species 6 to 10 to the second group (C_4 species). The phylogenetic relationship between the ten species is depicted in (Figure 2.2). The phylogenetic tree reveals that species 6, species 9 and species 10 are a single representative of an independent evolution event. Whereas, species 7 and species 8 are representatives of the same evolution event. Thus, in total four independent groups acquired the trait of interest (e.g. C_4 photosynthesis).

The simulated multiple sequence alignment contains ten protein coding sequences from ten species and the simulated peptide sequences has the length of ten amino acids. Each position in the multiple sequence alignment (Figure 2.3) is testing the ability of the Total Variation Distance algorithm to identify sites that are different between the two specified groups. The peptide sequence was generated using an R script that randomly samples from a pre-defined amino acid distribution in order to simulate the various different degrees of within and between group variability.

2.4.2 Analysis pipeline software specification

The analysis pipeline was constructed using R (versions 3.5.1, Feather Spray) and RStudio Version 1.2.1335 and relying on following packages (Table 2.6).

Table 2.6: Packages used to develop the Total Variation Distance algorithm.

Package (Version)	Reference
Seqinr (3.4-5)	(Charif and Lobry 2007)
Stringr (1.3.1)	(Wickham 2018)
Biostrings (2.48.0)	(Pagès et al. 2018)
Plyr (1.8.4)	(Wickham 2011)
Parallel (3.5.1)	(Vera et al. 2008)
Stringi (1.1.7)	(Gagolewski and Bartek 2015)

2.4.3 Application of the Total Variation Distance algorithm on the Christin et al. 2007 dataset and output processing

The deposited sequences from AM689877 to AM689901 and from AM690209 to AM690312 by Christin et al. 2007 were downloaded from GenBank. The gene names were used to identify whether it is a sequence belonging to a C_4 PEPC or non-C_4 PEPC. The publicly available dataset consists of seventy-one non-C_4 PEPC and fifty-nine C_4-PEPC sequences. Subsequently, all gene names belonging to the C_4-PEPC or non-C_4 PEPC, respectively were added to the corresponding group file (C_4-PEPC group1.csv and non-C_4 PEPC group2.csv). The grass phylogeny (Grass Phylogeny Working Group II 2012) was used to identify, which species emerged from the same C_4 evolution event and subsequently subgroup structure to account for the sampling bias were created.

The Total Variation Distance algorithm was run using the default mode without considering expression data. To identify sites that reveal differences in the distribution of amino acids between the two groups, following filter parameters were applied: the p-value cut-off was set to 0.01 and the TVD is at least 0.7.

Chapter 3: Application of the Total Variation Distance algorithm to a dataset containing C_3 and C_4 plants to identify targets to improve protein function for C_4 photosynthesis

3.1 Chapter introduction

In general, C_4 crop plants have a number of advantages when compared to C_3 crop plants. These advantages include high productivity and a better water and nitrogen use efficiency. To date, most attempts to enhance the photosynthetic efficiency in crop plants have focused on C_3 plants. This is primarily because there are more C_3 crop plants and C_4 photosynthesis is already considered to be "enhanced" relative to C_3. However, as discussed in the introduction to this thesis and this chapter there are good reasons to believe that C_4 photosynthesis can be improved by relatively simple sequence changes. In this chapter, the novel Total Variation Distance algorithm described in Chapter 2 is applied to a large dataset consisting of thirty C_4 plants and seventeen C_3 plants. The aim is to identify simple sequence changes that could be potential engineering targets in proteins associated with the C_4 cycle to improve the function of those proteins for C_4 photosynthesis.

3.1.1 Background

As stated above, the majority of attempts to enhance photosynthetic efficiency in crops have focused on C_3 plants. Procedures carried out are for example, increasing the metabolic flux through the C_3 cycle through improvements in RuBisCO function (Sharwood 2017) or the overexpressing of the enzyme sedoheptulose 1,7-bisphosphatase that is located at a branch point in the photosynthetic carbon reduction cycle (Lefebvre et al. 2005). Additional strategies to improve the photosynthetic

efficiency in C_3 plants are engineering a bypass of the wasteful process of photorespiration (Peterhansel et al. 2012; South et al. 2019; Trudeau et al. 2018).

Other approaches to improve C_3 photosynthesis include the introduction of the carbon concentrating mechanisms. Such carbon concentrating mechanisms can be found in various lineages. For example, cyanobacteria make use of a structure called carboxysome to elevate the level of CO_2 around RuBisCO (Lin et al. 2014a; Lin et al. 2014b) whereas green alga and some species of hornworts make use of pyrenoids to carry out a similar function (Atkinson et al. 2016; Meyer et al. 2016). These structures are used as a blueprint to create a novel compartment in C_3 plants to increase the CO_2 concentration around RuBisCO and therefore reduce the frequency of photorespiration and thus increase photosynthetic efficiency (Atkinson et al. 2016; Lin et al. 2014a; Lin et al. 2014b; Meyer et al. 2016). A different approach tries to harness the convergent evolution of C_4 photosynthesis and engineer it into the C_3 crop plant rice (Langdale 2011; von Caemmerer et al. 2012; Wang et al. 2016). Unlike the case with pyrenoids or carboxysomes, the engineering of C_4 photosynthesis requires the detailed understanding of the development of the specialized leaf anatomy and the evolution of C_4 photosynthesis (Schuler et al. 2016).

Similarly, improvements to C_3 photosynthesis have been made through enhancement of photoprotection. In full sunlight plant leaves dissipate the excess absorbed light as heat to avoid any damages caused by overexcitation in a process called photoprotection (Li et al. 2009b). Normally plant leaves adjust very slowly to fluctuating light conditions and therefore the maximum efficiency of CO_2 assimilation cannot be achieved until the leaves adjust the level of photoprotection to the amount of light received e.g. in low light conditions (Kromdijk et al. 2016; Li et al. 2009b) . Three key components (photosystem II subunit S, violaxanthin de-epoxidase and zeaxanthin

epoxidase) have been identified that interact between photoprotection and CO_2 assimilation during fluctuating light conditions and were engineered in *Nicotiana tabacum* (Kromdijk et al. 2016). The productivity of field grown lines expressing these components showed an increase in dry-weight by up to 20% when compared to wild type plants (Kromdijk et al. 2016). The photoprotection mechanism is common in all flowering plants and therefore these findings provide a proof of concept that similar approaches may increase productivity in other crop plants (Kromdijk et al. 2016).

In contrast, improvement of C_4 photosynthesis has received almost no attention. This is perhaps due to the fact that the C_4 pathway already possesses multiple beneficial characteristics such as a high photosynthetic rate and high nitrogen and water use efficiency (von Caemmerer and Furbank 2016; Zhu et al. 2010). However, predicted future increases in global population necessitate that we also improve the yields of C_4 plants. Such yield increases are perhaps most important for maize (*Zea mays*), given its prominent role in global food and bioethanol production (von Caemmerer and Furbank 2016). To date only one publication has documented an improvement to C_4 photosynthesis. This was achieved by overexpressing the Rieske FeS protein in *Setaria viridis* to remove limitations in the electron transport and resulted in about 8% higher CO_2 assimilation rate (Ermakova et al. 2019). Thus, although it is not as well studied as C_3 photosynthesis there are likely multiple ways in which C_4 photosynthesis can be improved.

In total, there are ~7500 C_4 plant species that are descendent from 62 independent origins of C_4 photosynthesis (Sage et al. 2011). Within these ~7500 C_4 species there are differences in biochemistry (e.g. decarboxylation enzymes and transporters) and anatomy such as variations in Kranz anatomy (Muhaidat et al. 2007). Even plants with

a single cell C_4 photosynthesis have been discovered (Chuong et al. 2006). Furthermore, there is a large variation in photosynthetic efficiency between C_4 species, and even between closely related species such as *Zea mays* and *Sorghum bicolor* (Ehleringer and Pearcy 1983). In grasses, molecular clock analysis revealed that the earliest C_4 plants evolved around 30 million years ago (Sage 2004). In contrast, the youngest C_4 lineages emerged in the past 2 million years (Christin et al. 2011a). Given that different C_4 lineages have been conducting C_4 photosynthesis for different amounts of time, in different environmental conditions it is unlikely that all ~7500 C_4 plants have equally optimised all rate-limiting steps in the C_4 cycle to the same extent. Thus, some C_4 species must have gene products whose properties are more suited to C_4 function than other C_4 species.

To be able to assess what the effect of the genetic variants could be, a thorough understanding of the regulation of C_4 photosynthetic flux is necessary. However, C_4 photosynthetic flux is complex given the compartmentalisation of the metabolic processes within organelles and the division of labour between the mesophyll and bundle-sheath cells (von Caemmerer and Furbank 2016). The activities of the majority of the enzymes involved in C_4 photosynthesis appear to be closely controlled but their impact on the photosynthetic flux has been a challenge to evaluate (von Caemmerer and Furbank 2016) but a number of them have been identified and are discussed in the section below.

3.1.2 Example rate limiting stages of the C_4 cycle

Carbonic anhydrase (CA), the enzyme that catalyses the conversion of CO_2 and water to bicarbonate and hydronium, is present in abundance at ambient levels of CO_2 in C_4 plants (Studer et al. 2014; von Caemmerer and Furbank 2016). This would

propose that under ideal environmental conditions, when the stomata are open, a reduced CA activity would not have an impact because enough CA activity might be present to sustain photosynthetic flux. However, during heat and water stress, when stomata are closed, an increased CA activity might be beneficial (Studer et al. 2014; von Caemmerer and Furbank 2016) (see Figure 3.1, 1).

Phosphoenolpyruvate carboxylase (PEPC) is the initial enzyme in the C_4 cycle that fixes hydrogen carbonate by attaching it to phosphoenolpyruvate (PEP). It is assumed that the activity of PEPC is tightly controlled by phosphorylation and the levels of metabolites (von Caemmerer and Furbank 2016). Von Caemmerer et al. (2016) hypothesised that PEPC is unlikely to be a limiting step at ambient CO_2 concentration but similar to CA it could have an important function under stress conditions. There is not much known about the biochemical properties of PEPC, such as its affinity for hydrogen carbonate. However, discrepancies amongst different species are known (Bauwe and Chollet 1986). It is questionable whether altering the phosphorylation regulation of PEPC would improve the catalytic efficiency due to the absence of phenotypes observed with loss of function of PEPC protein kinase in *Flaveria bidentis* (Furumoto et al. 2007). Interestingly, *S. viridis* with low PEPC abundance required high CO_2 for growth illustrating the essential role of PEPC in the C_4 cycle (Alonso-Cantabrana et al. 2018). Thus, genetic variation across C_4 species is likely to serve as a blueprint for PEPC with superior catalytic properties (Figure 3.1, 1).

Pyruvate phosphate dikinase (PPDK). Alternatively, to improve the catalytic rate of CA or PEPC, the regeneration of educts such as PEP or RuBP (see Section 3.1.1) can be restrictive factors for CO_2 assimilation and therefore could form interesting targets for increasing photosynthetic rate. PPDK catalyses the step from pyruvate to PEP, for which it consumes one ATP. It is potentially the most rate-limiting step in the

regeneration of PEP (Chastain et al. 2011). The reduction of PPDK in *F. bidentis* using an antisense construct resulted in a considerable decrease in CO_2 assimilation rate (Ashton et al. 1999). This makes PPDK an interesting candidate gene to enhance the assimilation rate. In addition, PPDK is believed to become more limiting under colder temperatures (Wang et al. 2008). Therefore, PPDK with increased activity or a slower dissociation rate might contribute to enhanced photosynthetic rates, also in colder environments (von Caemmerer and Furbank 2016). In many respects, PPDK is similar to SBPase described above, as SBPase is responsible for the regeneration of RuBP, thus improving PPDK may also result in improvements in biomass (Figure 3.1, 2).

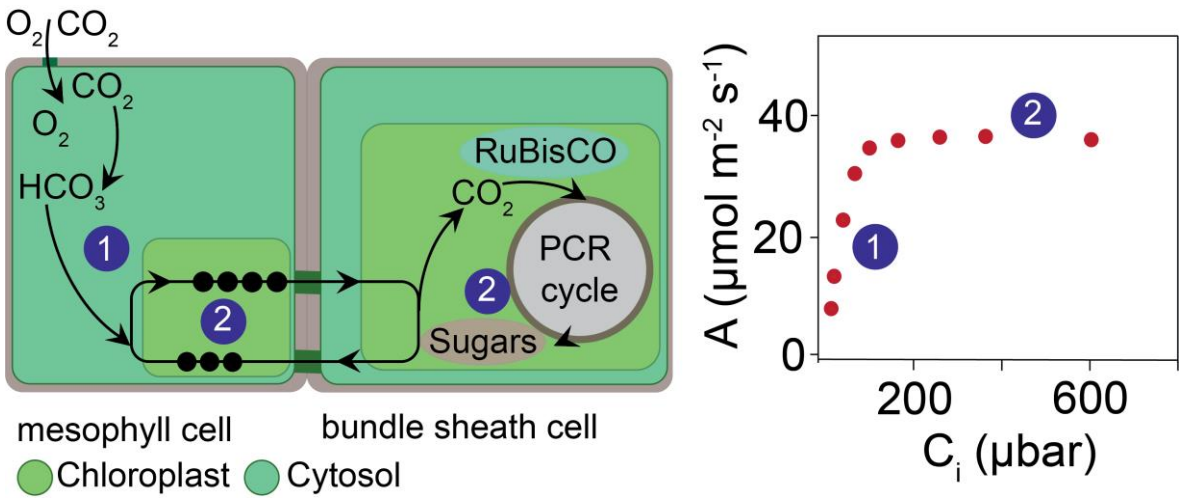


Figure 3.1: Targets to improve the photosynthetic efficiency of C_4 plants **1)** CO_2 delivery and conversion into bicarbonate and carbon fixation of PEPC. The initial rise in CO_2 assimilation is determined by the CO_2 diffusion and CA and PEPC activity. **2)** Substrate regeneration PEP for PEPC and RuBP for RuBisCO could increase the flux through the C_4 cycle. C_4 plants have saturated rates of CO_2 assimilation at around 200 µbar intercellular CO_2, which is determined by the reaction rate of RuBisCO and the rate substrate (PEP and/or RuBP) is regenerated. Figure adjusted from von Caemmerer and Furbank 2016.

3.1.3 Exploiting natural variation in C_4 plants to identify potential engineering targets to improve photosynthetic efficiency

C_4 photosynthesis is a complex trait that evolved multiple times during the last 30 million years in response to a dramatic drop in the concentration of CO_2 in the atmosphere. There are at least 62 independent origins of C_4 photosynthesis across the plant kingdom (Sage 2016). If something as complex as C_4 photosynthesis evolved multiple times in parallel, and has had different amounts of time to optimise their biochemical pathways through natural selection, it seems highly unlikely that all components of the C_4 cycle are equally optimised in every species today (Chapter 1). Moreover, there is ample evidence that variation between species exists. For example, different species use different decarboxylases, have differences in leaf anatomy (Muhaidat et al. 2007), use two-cell C_4 or use single-cell C_4 (Chuong et al. 2006), and importantly have differences in photosynthetic rates (Chapter 1 Figure 1.2).

If such variation in photosynthetic rates exist, then much of this variation must be attributable to variation in the proteins that contribute to C_4 photosynthesis. Therefore, it is possible that variants of genes are present in one lineage that have enhanced properties in a C_4 context compared to the orthologous counterpart in other C_4 species. Identification of these causative variants and transfer of the best performing variants into a crop species has the potential to improve C_4 photosynthesis in that crop species and therefore improve yield.

3.1.4 Chapter aims and outline

Getting transgenes expressed in the right cell type, during the correct developmental stage, at the right level is extremely challenging. However, the development of CRISPR/Cas gene editing has opened new avenues for crop improvement.

Specifically, it is now possible to change a single nucleotide in a plant genome and thus change any amino acid in any protein sequence. Importantly this allows proteins to be altered without altering the endogenous regulatory systems that control their expression. However, given one could change any amino acid in any protein sequence, which one would you chose and for what purpose? One could take the approach of random mutagenesis to explore enzyme variants. However, there are the same number of variants of a 71 amino acid protein as there are particles in the universe (~10^{90}). As random mutagenesis can never explore even a small fraction of the possible protein sequence space, methods that can reduce this space by identifying plausible sites to target for engineering could rapidly accelerate the process of engineering enhanced enzyme variants. Thus, the aim of this chapter is to identify the sites in protein sequences that evolution has adapted for improved C_4 function, and thereby provide a target list of sites that should be changed to optimise proteins for C_4 function (and thereby improve photosynthesis) in C_4 crop plants.

3.2 Results and discussion

3.2.1 Processing of the transcriptome dataset and inferring orthogroups

This chapter utilised a publicly available transcriptomic dataset consisting of 30 C_4 and 17 C_3 species representing eighteen independent evolutionary origins of C_4 photosynthesis (Kelly et al. 2017). This group of species contains examples from all seven of the orders within the eudicots that have evolved C_4 photosynthesis (Kelly et al. 2017). Thus, this dataset encompasses a substantial amount of the diversity that is encapsulated by independently evolved C_4 species (Kelly et al. 2017). Collectively these transcriptome samples comprise 1.5 million contigs of which 963’237 coding sequences were predicted (Table 3.1). The predicted coding sequences were clustered into orthogroups using OrthoFinder (Emms and Kelly 2015) and in total,

14’848 orthogroups were inferred (see Section 3.4.1 to 3.4.6). Out of the 963’237 coding sequences 36’051 were not assigned to an orthogroup.

Table 3.1: Number of inferred coding sequences per species from the transcriptome data

Species	Number of coding sequences	Photosynthesis type	C_4-subtype
Aerva lanata	19065	C_3	-
Aerva persica	23776	C_4	NADP-ME
Allionia incarnata	24720	C_4	NADP-ME
Alternanthera brasiliana	29093	C_3	-
Alternanthera caracasana	26648	C_4	NADP-ME
Alternanthera sessilis	22576	C_3	-
Amaranthus cruentus	22482	C_4	NAD-ME
Amaranthus retroflexus	14607	C_4	NAD-ME
Anticharis glandulosa	15761	C_4	NAD-Me
Arabidopsis thaliana	27416	C_3	-
Atriplex hortensis	19608	C_3	-
Atriplex prostrata	19040	C_3	-
Atriplex rosea	19880	C_4	NAD-ME
Blutaparon vermiculare	17858	C_4	unknown
Boerhavia coccinea	28030	C_4	NADP-ME
Boerhavia dominii	25514	C_4	NADP-ME
Cleome gynandra	21583	C_4	NAD-ME
Cleome violacea	17306	C_3	-
Cleome viscosa	16134	C_3	-
Cypselea humifusa	17030	C_4	NADP-ME
Euphorbia mesembryanthemifolia	23236	C_4	NADP-ME
Flaveria bidentis	22579	C_4	NADP-ME
Flaveria cronquistii	22070	C_3	-
Flaveria kochiana	17055	C_4	NADP-ME
Flaveria pringlei	24370	C_3	-
Flaveria trinervia	22120	C_4	NADP-ME
Heliotropium calcicola	27922	C_3	-
Heliotropium karwinskyi	19208	C_3	-
Heliotropium mendocinum	19673	C_4	NADP-ME
Heliotropium tenellum	20289	C_3	-
Heliotropium tenuifolium	15871	C_4	NADP-ME
Heliotropium texanum	23088	C_4	NADP-ME
Mirabilis jalapa	28532	C_3	-
Mollugo cerviana	15936	C_4	NADP-ME
Mollugo pentaphylla	23276	C_3	-
Portulaca amilis	18858	C_4	NADP-ME
Portulaca grandiflora	13403	C_4	NADP-ME
Portulaca molokiniensis	19107	C_4	NAD-ME
Portulaca oleracea	21139	C_4	NAD-ME
Portulaca pilosa	14799	C_4	NADP-ME
Portulaca suffruticosa	15195	C_4	unknown
Portulaca umbraticola	15518	C_4	NADP-ME
Sesuvium portulacastrum	18473	C_3	-
Sesuvium verrucosum	18336	C_3	-
Trianthema portulacastrum	25037	C_4	NADP-ME
Tribulus eichlerianus	13569	C_4	NADP-ME
Zaleya pentandra	16451	C_4	NADP-ME

3.2.2 Application of the Total Variation Distance algorithm on a set of proteins that are associated with the C_4 cycle

The Total Variation Distance algorithm was applied to the set of genes whose gene products are directly involved in, or closely linked to, the C_4 cycle (in total 77 genes) (see Section 3.4.5). To identify sites in these genes that are different in the distribution of amino acids between the C_3 and C_4 plants the Total Variation Distance algorithm described in Chapter 2 was applied to multiple sequence alignments of these 77 genes from the 47 species of the transcriptome dataset.

This approach identified twelve genes where there were sites in the protein sequence alignment that were significantly different between the set of C_4 and C_3 plants (p-value ≤ 0.01, Monte Carlo resampling test). Figure 3.2 shows how these 12 genes (depicted in dark blue font) fit into our current understanding of C_4 photosynthesis. From this figure it is apparent that the genes showing sites with significant TVD are distributed across the C_4 metabolism.

The genes identified that have a site that is significantly different between the C_3 and C_4 species can be loosely classified into three groups: enzymes of the core C_4 cycle (PEPC, RBCS, PPDK, AMK2, CPGK2), transporters (DiT1, TPT, BASS2, PPT) and enzymes involved in photorespiration (GLDP1, SGAT, HPR2). In total, the analysis pipeline identified 32 sites in these 12 genes using the standard amino acid alphabet.

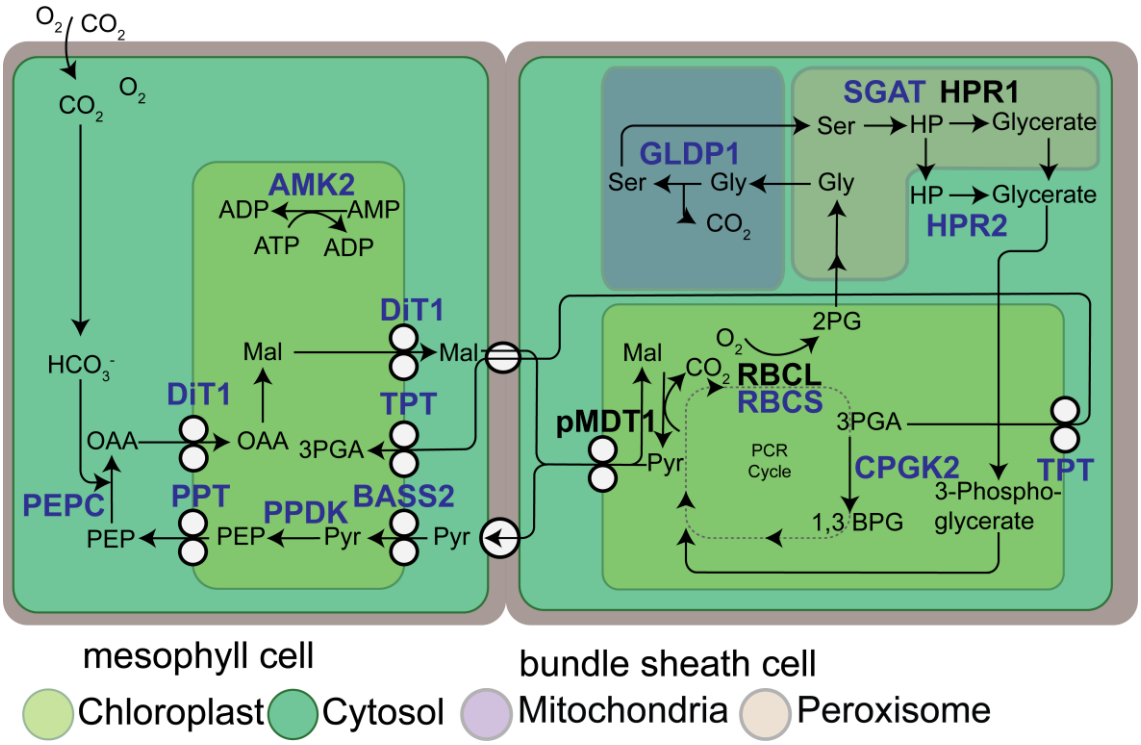


Figure 3.2: Proteins with altered amino acid distributions and their function. Genes highlighted in blue show significant different distribution in amino acid composition between C_3 and C_4 plants. Abbreviations of genes and metabolites as follows: Phosphoenolpyruvate carboxylase (PEPC), dicarboxylate translocator 1 (DiT1), plastid mono and dicarboxylate transporter 1 (pMDT1), RuBisCO small subunit (RBCS), RuBisCO large subunit (RBCL), plastid localized phosphoglycerate kinase (CPGK2), Bile acid:sodium symporter 2 (BASS2), triosephosphate phosphate translocator (TPT), phosphoenolpyruvate phosphate translocator (PPT), adenylate kinase 2 (AMK2), glycine dehydrogenase (GLDP1), serine glyoxylate aminotransferase (SGAT), glyoxylate hydroxy pyruvate reductase 1/2 (HPR1/2). Oxygen (O_2), carbon dioxide (CO_2), hydrogen carbonate (HCO_3^-), oxaloacetate (OAA), malate (Mal), pyruvate (Pyr), 2-phosphoglycolate (2PG), 3-phosphoglycerate (3-PGA), serine (Ser), hydroxypyruvate (HP), glycine (Gly)

3.2.3 Enzymes of the core C_4 cycle with altered amino acid distributions

Among the enzymes that are involved in the C_4 cycle, five were identified that have an altered amino acid distribution between the C_3 and C_4 species in this dataset. Below, the sites identified in these enzymes are presented and discussed.

3.2.3.1 Variation identified in phosphoenolpyruvate carboxylase (PEPC)

Phosphoenolpyruvate carboxylase (PEPC) is one of the first enzymes of the C_4 cycle. It catalyses the carboxylation of PEP using bicarbonate to form oxaloacetate (Westhoff and Gowik 2004). To catalyse that reaction it requires Mg^{2+} as an essential cofactor (Westhoff and Gowik 2004). The function of PEPC has been identified as one of the possible rate limiting steps of the C_4 cycle (von Caemmerer and Furbank 2016) (see Section 3.1.2).

The Total Variation Distance algorithm identified twelve sites that have a different amino acid distribution between the C_3 and C_4 variants of the PEPC gene in the dataset. The sequence logo reveals the composition of the amino acid distribution at each of the twelve identified sites (Figure 3.3 A). Some of the identified sites will be discussed in further detail below because they are close to a known functional site or because they correspond to a site previously marked as being under selection in C_4 plants. In all cases, sites are discussed with reference to the sequence found in *Zea mays*. This is done with a view to identifying those sites in the maize gene that could be changed to potentially improve the maize protein for C_4 function.

The site 228 has a significantly different amino acid distribution between C_3 and C_4 plants and is located next to a glucose-6-phosphate binding site (Figure 3.3). The region where that site is located has previously been identified to be important for the function of C_4 PEPC (Engelmann et al. 2003). Glucose-6-phosphate is considered to

be an allosteric activator of PEPC. The neighbouring arginine at position 227 (Arg-231 in *Z. mays* CAA33317) was suggested to be involved in the binding of the phosphate group of glucose-6-phosphate (Takahashi-Terada et al. 2005). The change that takes place at site 228 in C_4 species, is the emergence of two hydrophobic amino acids, alanine and isoleucine, whereas in the C_3 variants a polar amino acid, threonine, and a neutral one, asparagine, dominate. Interestingly, the maize PEPC variant at position 232 (corresponding to position 228 in the alignment below) has an arginine at that place. Thus, it is possible that changing the maize residue from arginine to alanine or isoleucine may improve the binding affinity of maize PEPC for the activator glucose-6-phosphate.

The site 279 is located four amino acids away from a PEP binding site. Similar to the site 228, the site 279 is located in a known function determining region of C_4 PEPC. The C_3 variants of the protein have a glutamine present whereas the C_4 variants alter the utilise amino acids with basic characteristics such as arginine and lysine (Figure 3.3). The corresponding site in maize PEPC (position 284 in *Z. mays* CAA33317) is also the basic amino acid arginine. As maize has a variant that is common in C_4 species this site would not be a priority for changing in maize.

The sites 319 and 346 are not close to any known functional sites (Figure 3.3). Furthermore, in both cases the corresponding site in maize PEPC (position 324 and 353 in *Z. mays* CAA33317) has one of the variants that is common in C_4 species present namely glutamic acid and lysine, respectively. Thus, this site would not be a priority for change.

Site 361 is located four amino acids away from a known glucose-6-phosphate binding site (Figure 3.3). The biochemical characteristics of the amino acids at that position shift towards polar and basic amino acids in C_4 species. The corresponding site in maize PEPC (position 368 in *Z. mays* CAA33317) has the C_3 variant present namely asparagine. Serine and asparagine have both polar uncharged side chains and therefore similar biochemical characteristics but serine is a smaller amino acid than asparagine. Interestingly, histidine that is present in about 12% of the C_4 species is biochemically and structurally very different from serine and asparagine. Given the variety of residues with contrasting properties at this position this could potentially be an interesting target site to alter the biochemical characteristics and evaluate the impact on the enzyme function in maize.

Site 494 is three amino acids away from the tetramer binding site. From the sequence logo it is clear that the phenylalanine used ubiquitously in C_3 species is less abundant in C_4 species, where it has been replaced either by leucine or isoleucine (Figure 3.3). The maize PEPC at the corresponding position (501 in *Z. mays* CAA33317) has also a leucine. The biochemical characteristics of leucine, isoleucine and phenylalanine are similar, however leucine ($C_6H_{13}NO_2$) and isoleucine ($C_6H_{13}NO_2$) are substantially smaller amino acids compared to phenylalanine ($C_9H_{11}NO_2$). As maize has already switched to a common C_4 variant, and these variants have similar biochemical properties, this site is not a priority for change.

The sites 509 and 510 form an interesting target site because it seems unusual to find two neighbouring sites that have an altered amino acid distribution between the C_3 and C_4 plants (Figure 3.3). At position 509, the majority of C_3 plants possess an aspartic acid whereas the C_4 species has a glycine. The maize PEPC at the corresponding position (516 in *Z.mays* CAA33317) has the C_4 variant glycine present. At the position

510, the polar amino acid threonine is replaced by the hydrophobic alanine in the C_4 group. The maize PEPC at the corresponding site (517 in *Z.mays* CAA33317) has the C_4 variant alanine present and has previously been identified under positive selection (Christin et al. 2007). Thus, maize conforms to the C_4 stereotype and would not need to be altered here.

The site 570 reveals that about half to the C_3 plants in the dataset have an alanine present and the other half of the set of C_3 species have a serine. In the set of C_4 plants this site almost fully converges to a serine (Figure 3.3). This is a well-known substitution that has previously been shown to be under positive selection in grasses (Christin et al. 2007).

Some of the C_4 species in the dataset have changed site 658 to a neutral amino acid (primarily asparagine or glutamine) from the basic and charged amino acid histidine that is present in all C_3 species. A known functional site is seven amino acids away, which is involved in the aspartate/malate binding. Both aspartate and malate are known inhibitors of PEPC (Takahashi-Terada et al. 2005). The corresponding site in maize (position 665 in *Z. mays* CAA33317) has also changed to an asparagine. Interestingly this transition to asparagine has also been identified in many other C_4 monocots (Christin et al. 2007).

The sequence logo at position 773 reveals that at least 60% of the dicot C_4 species in the dataset changed from the ancestral alanine to a serine at that site. In previous analyses, all known C_3 PEPCs have a conserved alanine residue whereas all known C_4 PEPCs converged to a serine (position 780 in *Z. mays* CAA33317) (Bläsing et al. 2000). The alanine to serine conversion is known to have an impact on the catalytic function of PEPC (see Section 1.4). Thus, it is surprising to find some remaining

ancestral alanine signal present at the site 773 in some of the C_4 plants. This could originate from non-photosynthetic PEPCs that were clustered into the same orthogroup. Indeed, from the two C_4 species *Portulaca molokiniensis* and *Alternathera caracasana* transcripts are present in the orthogroup with both alanine and serine at that site. Alternatively, some of the C_4 dicots in the dataset are members of relatively young C_4 lineages. Thus, for these younger lineages there might not have been enough time to change from an alanine to serine at that site. For example, the C_4 species *Heliotropium tenuifolium* and *Heliotropium texanum* have exclusively an alanine present at that site. It remains unclear whether for these two species there was not enough time/selective pressure to change to a serine because the age of the C_4 lineage Boraginaceae remains uncertain (Christin et al. 2011a; Sage 2016).

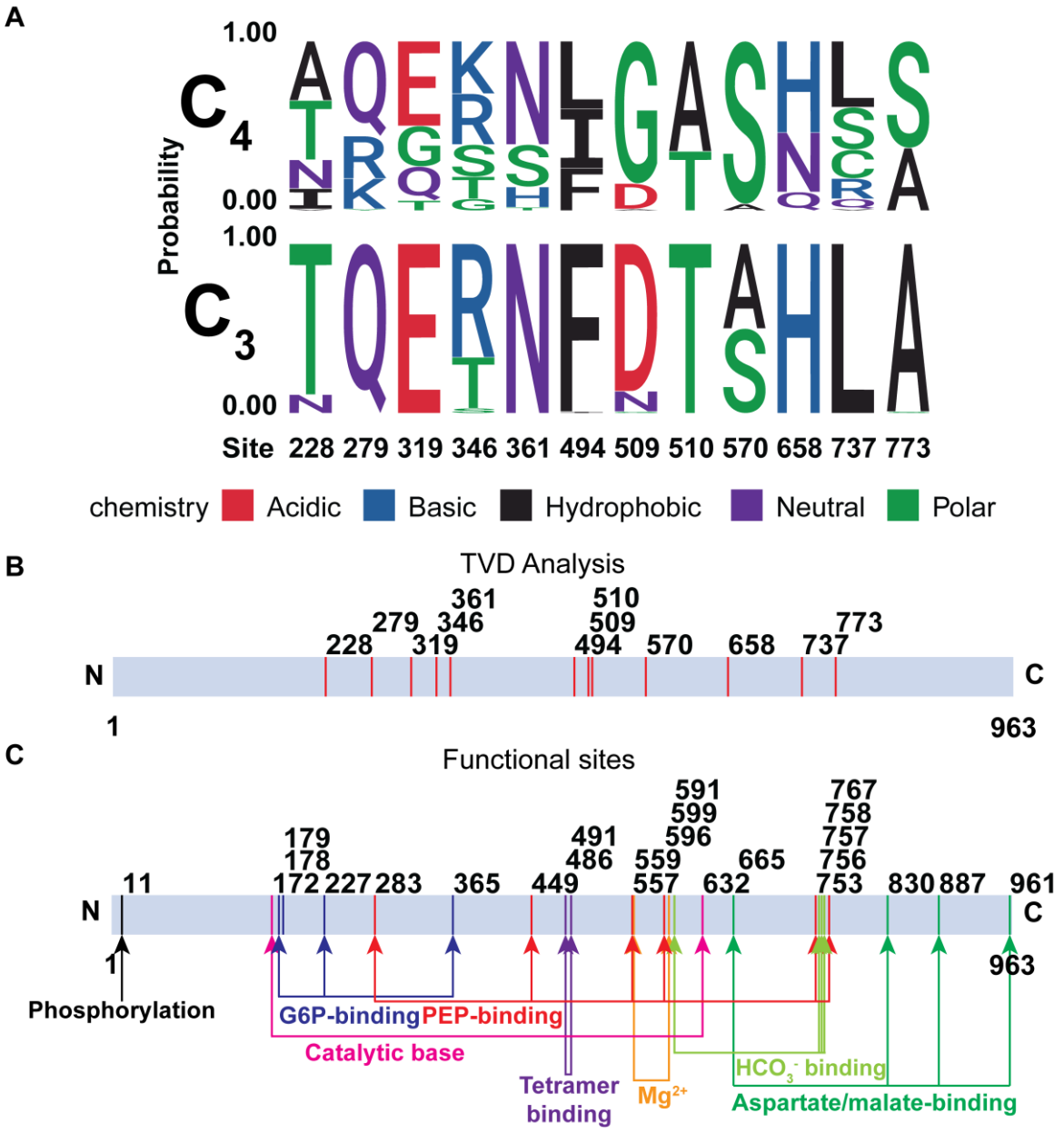


Figure 3.3: Identification of significant sites in Phosphoenolpyruvate carboxylase (PEPC) that are different in C_3 and C_4 species. **A)** Sequence logo revealing the difference in amino acid distribution between C_4 and C_3 plants in the applied dataset **B)** Identified sites were mapped on the *A. thaliana* PEPC protein (AT2G42600) **C)** Mapping the known functional sites onto the *A. thaliana* PEPC protein (AT2G42600) (Jacobs et al. 2008)

To compare these identified sites in PEPC to sites previously identified using positive selection by Christin et al. 2007, the sites were mapped onto the maize PEPC. Out of the twelve sites identified using the Total Variation Distance algorithm, four sites

correspond with the sites previously identified by Christin et al. 2007 (Figure 3.4 A). This is quite remarkable considering that each analysis pipeline applied an entirely different dataset with no sequences in common (Figure 3.4 B). Among the sites in agreement are the threonine to alanine change at site 510 (corresponding to 517 in maize CAA33317), the alanine to serine change at site 570 (corresponding to 577 in maize CAA33317), the transition from histidine to asparagine and glutamine at site 658 (corresponding to 665 in maize CAA33317) and the alanine to serine change at site 773 (corresponding to 780 in maize CAA33317) (Figure 3.3 and Figure 3.4).

To identify which region of the C_4 PEPC protein contributes most to the C_4 characteristics, Engelmann et al. 2002 defined five regions along the PEPC sequence. They found these regions by swapping segments of the amino acid sequence between the PEPC variants of *Flaveria pringlei* (C_3 plant) and *Flaveria trinervia* (C_4 plant). Here, they were able to identify that the largest C_4 functional determination is caused by region two (Figure 3.4 C) and the alanine to serine substitution in region five at site 774 (correspond to site 780 in maize CAA33317) (Engelmann et al. 2002). The Total Variation Distance algorithm identified three sites in region 2 and these are thus candidates for the causative sites for functional differentiation between C_3 and C_4 variants of PEPC (Figure 3.4 C).

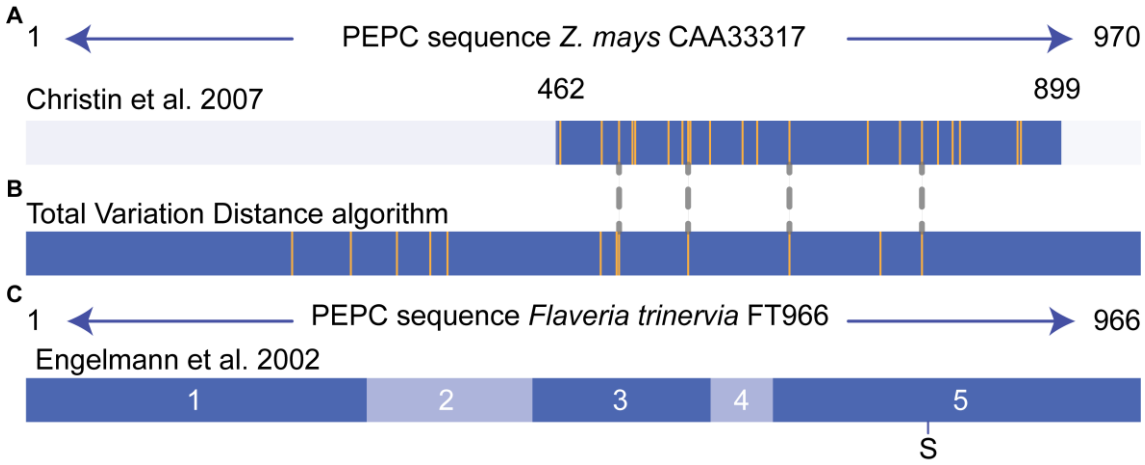


Figure 3.4: Comparison between previously identified sites under selection and functional determining regions of Phosphoenolpyruvate carboxylase (PEPC). **A**) Twenty-one sites under positive selection identified in a monocot dataset highlighted in yellow (Christin et al. 2007) **B)** Twelve sites with a different amino acid distribution identified using the Total Variation Distance algorithm (highlighted with yellow vertical bars). Sites that correspond between the positive selection analysis and the Total Variation Distance algorithm are connected with grey dotted bars. **C)** Designated region for the swapping experiment between C_3 and C_4 *Flaveria* PEPC to identify the functional determining region for the C_4 function (Engelmann et al. 2002)

3.2.3.2 Variation identified in RBCS (AT5G38410)

RuBisCO is one of the most abundant proteins in the world (Ellis 1979). It has a dual catalytic site, catalysing both the CO_2 fixation in the first step of the photosynthetic carbon reduction cycle and the oxygenation reaction of RuBP in the first step of photorespiration. RuBisCO is a large enzyme complex consisting of eight large chloroplast-encoded subunits (each 55 kDa) and eight small nuclear encoded subunits (each 13 kDa) (Dean et al. 1989; Suzuki and Makino 2012). Four large subunit dimers form two active sites (Andersson et al. 1989; Spreitzer 2003). Although the small subunits of RuBisCO are not directly involved in catalysing the oxygenation or carboxylation reaction, it is known that they can alter the catalytic efficiency and specificity (Spreitzer 2003).

The Total Variation Distance algorithm identified a single site at a small RuBisCO subunit (Figure 3.5 A). The site 168 is located next to a multimerization site (Figure 3.5 B and C). The majority of the C_4 species have a valine at this site. In contrast, the vast majority of the C_3 plants have an isoleucine present (90%) at this site while valine is rare (10%). Functionally, isoleucine and valine are not very different but structurally valine ($C_5H_{11}NO_2$) is slightly smaller than isoleucine ($C_6H_{13}NO_2$). Given the location of the site, the best guess is that the change may impact on the stability of the multimeric complex (Figure 3.5). The site in maize is also a valine (corresponds to position 159 in *Z. mays* GRMZM2G083344) and thus not a priority for engineering enhanced C_4 photosynthesis in maize.

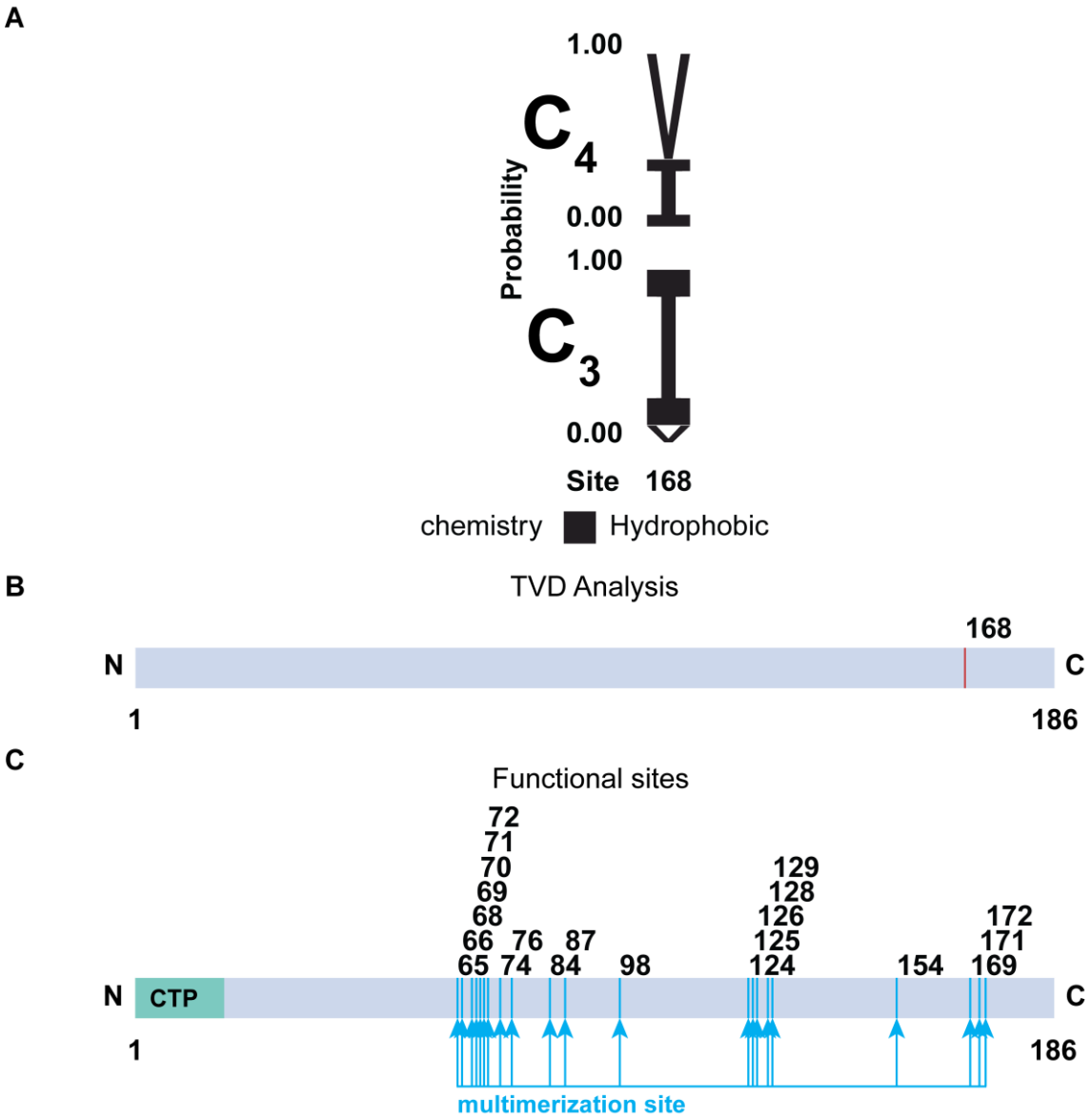


Figure 3.5: Identification of sites that have an altered amino acid distribution between the sampled C_3 and C_4 plants in RuBisCO small subunit (RBCS, AT5G38410). **A)** Sequence logo revealing the difference in amino acid distribution at the significant site. **B)** Identified site was mapped onto the *A. thaliana* RBCS protein sequence. **C)** Functional sites were mapped onto the RBCS protein sequence (InterPro B3H5S2). Chloroplast transit peptide (CTP)

3.2.3.3 Variation identified in PPDK (AT4G15530)

Pyruvate phosphate dikinase (PPDK) is another key enzyme of the C_4 cycle. It is responsible for the regeneration of PEP which is the primary CO_2 acceptor in the chloroplast stroma in the mesophyll cells. PPDK is a highly abundant protein in the leaves of C_4 plants where it amounts to about 10% of the total soluble protein (Chastain 2011). It is catalytically active as a homotetramer and each subunit is about 95 kDa in size (Chastain 2011). The amino acids involved in the catalytic activities of PPDK are conserved across C_3 and C_4 plants. A key event for the recruitment of PPDK as a C_4 enzyme was the modification of the promoter region to achieve cell type-specific expression in the mesophyll chloroplasts (Chastain et al. 2011) (for other examples see Chapter 1.6 and 1.7).

The Total Variation Distance algorithm identified two sites that have a different amino acid distribution between the C_3 and C_4 species in the dataset (Figure 3.6 A). Considering that PPDK is highly conserved across C_3 and C_4 plants it is perhaps less surprising to find only a limited number of sites to be different (Figure 3.6 B). It is, however, worth highlighting that the sites identified are located towards the C-terminus of the protein, where all of the catalytic residues are present (Figure 3.6 B and C). In addition, for both the sites a large proportion of the C_4 species in the dataset still have the ancestral C_3 amino acid present.

Site 737 in C_3 species is always glutamic acid. However, in C_4 species there is a diversity of residues at this site. Here, maize (709 in *Z. mays* GRMZM2G306345) has changed to a glutamine, a residue that is very rare in C_4 eudicot species. Thus, it may be interesting to explore the alternative residues commonly used in C_4 species at this site, namely lysine and alanine. The second site that is detected as different between

C_3 and C_4 species is site 894. Here, the ancestral residue is a leucine and about 50% of C_4 species have changed this leucine to either a phenylalanine or isoleucine or other residues. Maize has retained the ancestral leucine residue (site 868 in *Z. mays* GRMZM2G306345). Given the central position of PPDK in the C_4 cycle, the small number of sites identified in this gene, and their proximity to the pyruvate binding sites, it would be tempting to test out alternative C_4 variants at these positions in the maize gene to investigate whether these changes affected enzyme function.

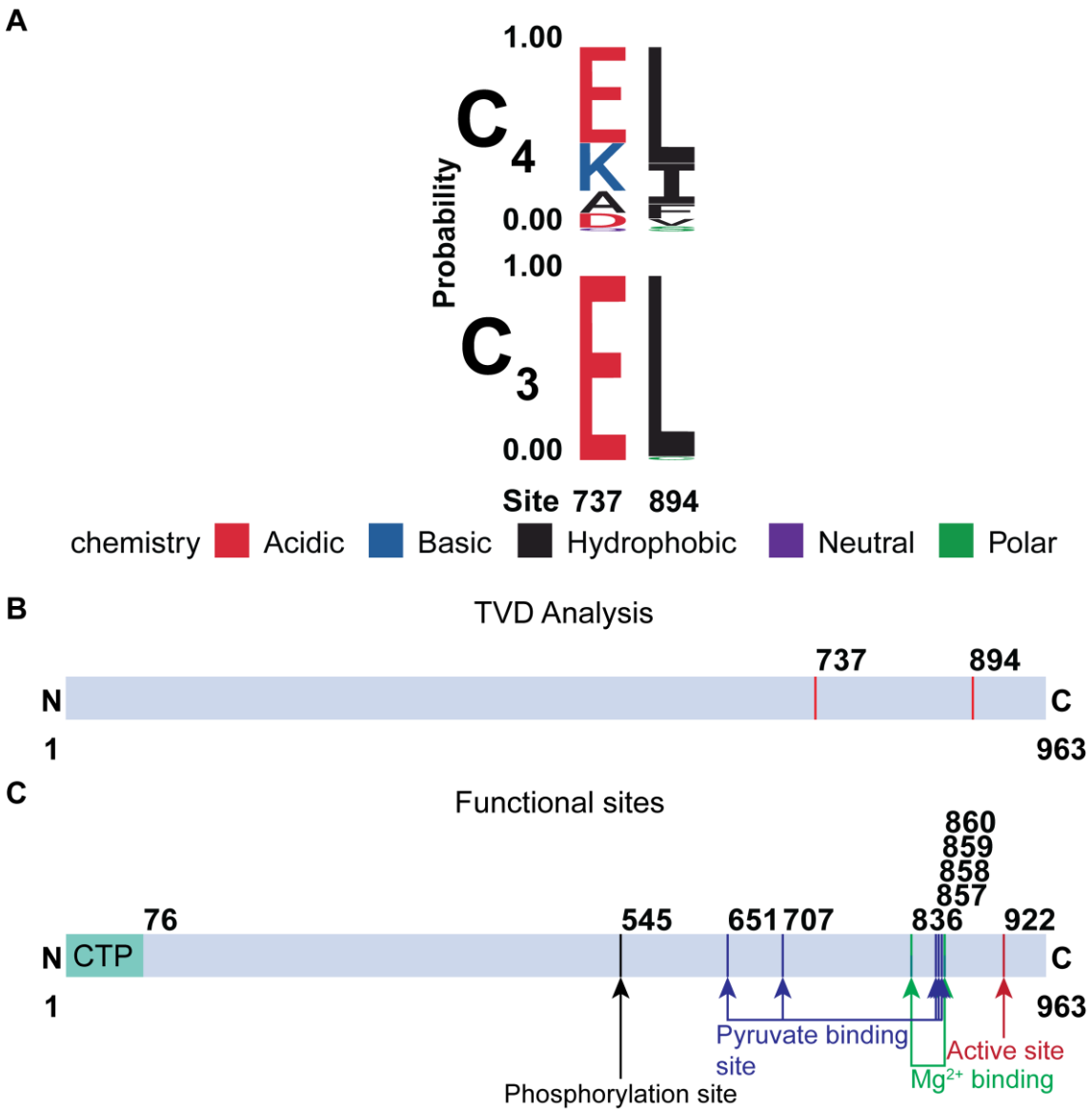


Figure 3.6: Identification of significant sites in pyruvate phosphate dikinase (PPDK, AT4G15530) that are different between the C_3 and C_4 species. **A)** Sequence logo revealing the difference in amino acid distribution between C_4 and C_3 plants. **B)** Identified sites were mapped onto the *A. thaliana* PPDK protein sequence. **C)** Functional sites were mapped onto the PPDK protein sequence (uniprot O23404). Chloroplast transit peptide (CTP)

3.2.3.4 Variation identified in AMK2 (AT5G47840)

Adenosine monophosphate kinase 2 (AMK2) catalyses the generation of two ADP by transferring a phosphate group obtained from ATP to AMP in the chloroplast stroma. AMK2 plays a central role in the energy homeostasis and adenine nucleotide metabolism because nucleotides are essential metabolites in various biochemical and developmental processes (Lange et al. 2008). The Total Variation Distance algorithm identified four sites that have an altered amino acid distribution between C_3 and C_4 species in the dataset (Figure 3.7 A).

The first site at position 2 in the alignment is located within the predicted chloroplast transit peptide region (Figure 3.7 B and C). The vast majority of the C_4 plants in the dataset have an alanine present whereas the C_3 plants have a more diverse amino acid composition. Chloroplast transit peptides are generally variable across different species and thus it is unlikely that this change is functionally relevant (Figure 3.7 B and C). It may perhaps improve its effectiveness or specificity as a transit peptide.

At the site 173, the majority of C_4 species have a leucine, which is also present to a large extent in the C_3 species (73%). A similar phenomenon is observed for site 182. Perhaps the site of most interest is the one at position 232 which is located four amino acids upstream of the AMP binding site (Figure 3.7 B and C). The frequency of tyrosine at that site in C_4 species is increased compared to the amino acid distribution in the C_3 species. Tyrosine is a special amino acid because it consists of a large hydrophobic chain that has a polar hydroxy group and thus can display a dual biochemical role. Given that the enzyme is peripheral to the C_4 cycle and that the sites are broadly similar in C_3 and C_4 species, these sites would not be a high priority for engineering enhanced C_4 function. Moreover, the sites in maize (which correspond to positions 13, 177, 186

and 236 in GRMZM2G178192, respectively) are also the most common sites found in C_4 species and thus not a priority for engineering enhanced C_4 photosynthesis in maize.

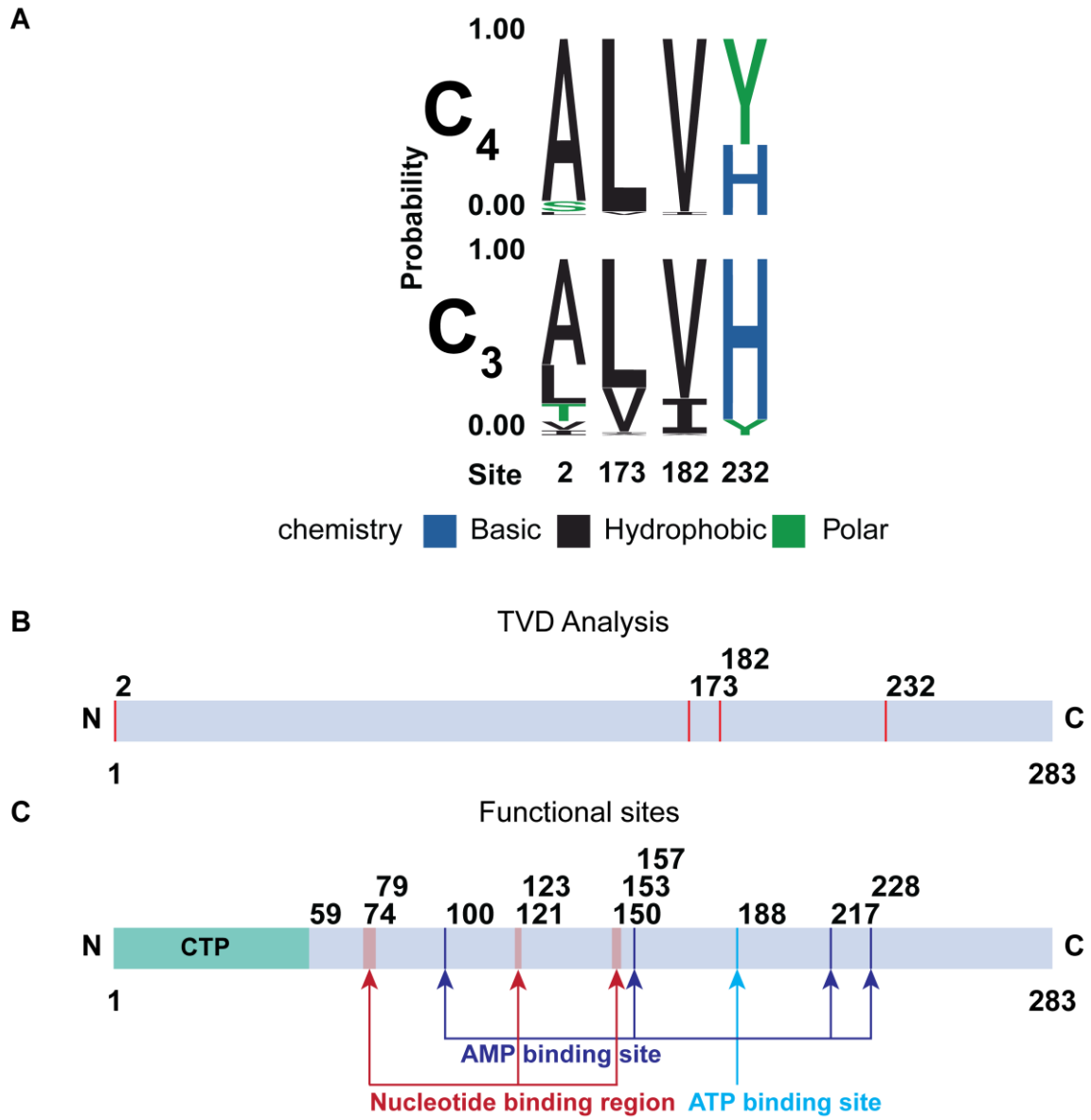


Figure 3.7: Identification of significant sites in adenosine monophosphate kinase 2 (AMK2, AT5G47840) that are different between the sampled C_3 and C_4 species. **A)** Sequence logo revealing the difference in amino acid distribution at each significant site. **B)** Identified sites were mapped onto the *A. thaliana* AMK2 protein sequence. **C)** Functional sites and domains were mapped onto the AMK2 protein sequence (uniprot Q9FIJ7). Chloroplast transit peptide (CTP).

3.2.3.5 Variation identified in CPGK2 (AT1G56190)

The chloroplast localised phosphoglycerate kinase participates in the photosynthetic carbon reduction cycle where it catalyses the reversible reaction of ATP and 3-phosphoglycerate to obtain ADP and glycerate 1,3 -diphosphate (KEGG K00927).

The Total Variation Distance algorithm identified a single site in CPGK2 that has a different distribution between the C_3 and C_4 species in the dataset (Figure 3.8 A). The C_4 species all have a serine at this site whereas the C_3 species has either a serine or threonine present. Threonine and serine have similar biochemical characteristics. Nonetheless, it seems that the C_4 species in the dataset prefers serine to threonine at that site (Figure 3.8 B and C). The site is not particularly close to the known substrate binding sites (16 amino acids away from the closest). The site in maize is also a serine (correspond to site 134 in GRMZM2G089136) and thus not a priority for engineering enhanced C_4 photosynthesis in maize.

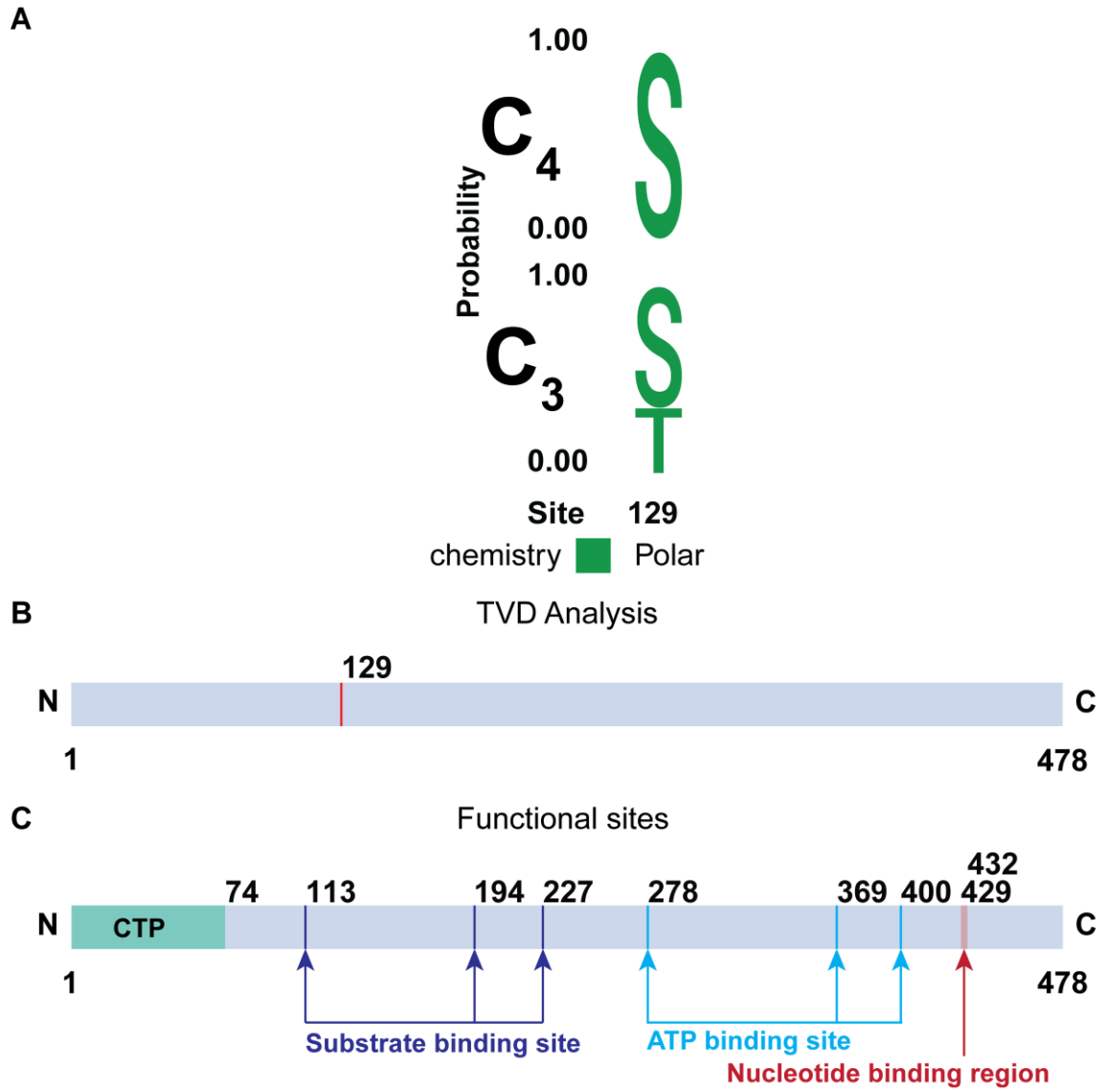


Figure 3.8: Identification of significant sites in the chloroplast localised phosphoglycerate kinase (CPGK2, AT1G56190) that are different between the sampled C_3 and C_4 species **A)** Sequence logo revealing the difference in amino acid distribution the significant site **B)** Identified site was mapped onto the *A. thaliana* CPGK2 protein sequence **C)** Functional sites were mapped onto the CPGK2 protein sequence (uniprot P50318)

3.2.4 Transporters of the C_4 cycle with altered amino acid distributions

The carbon concentrating mechanism in C_4 plants takes place, with few exceptions (Voznesenskaya et al. 2001), in two distinct cell types requiring a flux of metabolites between organelles and the cytosol and between mesophyll and bundle sheath cells

(Bräutigam et al. 2008). For example, in the C_4 NADP-ME there are at least ten transport steps necessary to assimilate one molecule of CO_2 (Weber and von Caemmerer 2010). This is in stark contrast to carbon assimilation in C_3 plants that require one transport process per three molecules of carbon assimilated (Weber and von Caemmerer 2010). The Total Variation Distance algorithm identified four transporters that had a different amino acid distribution between C_3 and C_4 species in the dataset.

3.2.4.1 Variation identified in DiT1 (AT5G12860)

DiT1 is a dicarboxylate antiporter located at the inner chloroplast envelop. In C_3 plants DiT1 exchanges 2-oxoglutarate for malate and is therefore involved in the re-assimilation of ammonia generated during the photorespiration cycle (Kinoshita et al. 2011; Weber et al. 1995). In C_4 plants, DiT1 is present in the chloroplasts in the mesophyll cells where it exchanges oxaloacetate for malate (Schuler et al. 2016). DiT1 has twelve transmembrane domains and a molecular weight of 59.2 kDa.

The Total Variation Distance algorithm identified a single site at position 248 that is different between C_3 and C_4 species (Figure 3.9 A). During the evolution of C_4 photosynthesis the alanine became more abundant than it is in the set of C_3 plants in the dataset. The site is located at the interface of a transmembrane domain and a non-cytoplasmic domain (Figure 3.9 B and C). Alanine and valine are similar amino acids regarding their biochemical properties and size. It is, thus, not obvious what the effect of this change in C_4 plants would be. Given the similar biochemical properties of the amino acids and that the site in maize is also the most commonly occurring site in C_4 species i.e. alanine (correspond to site 258 in GRMZM2G383088), this site is not a priority for engineering enhanced C_4 photosynthesis in maize.

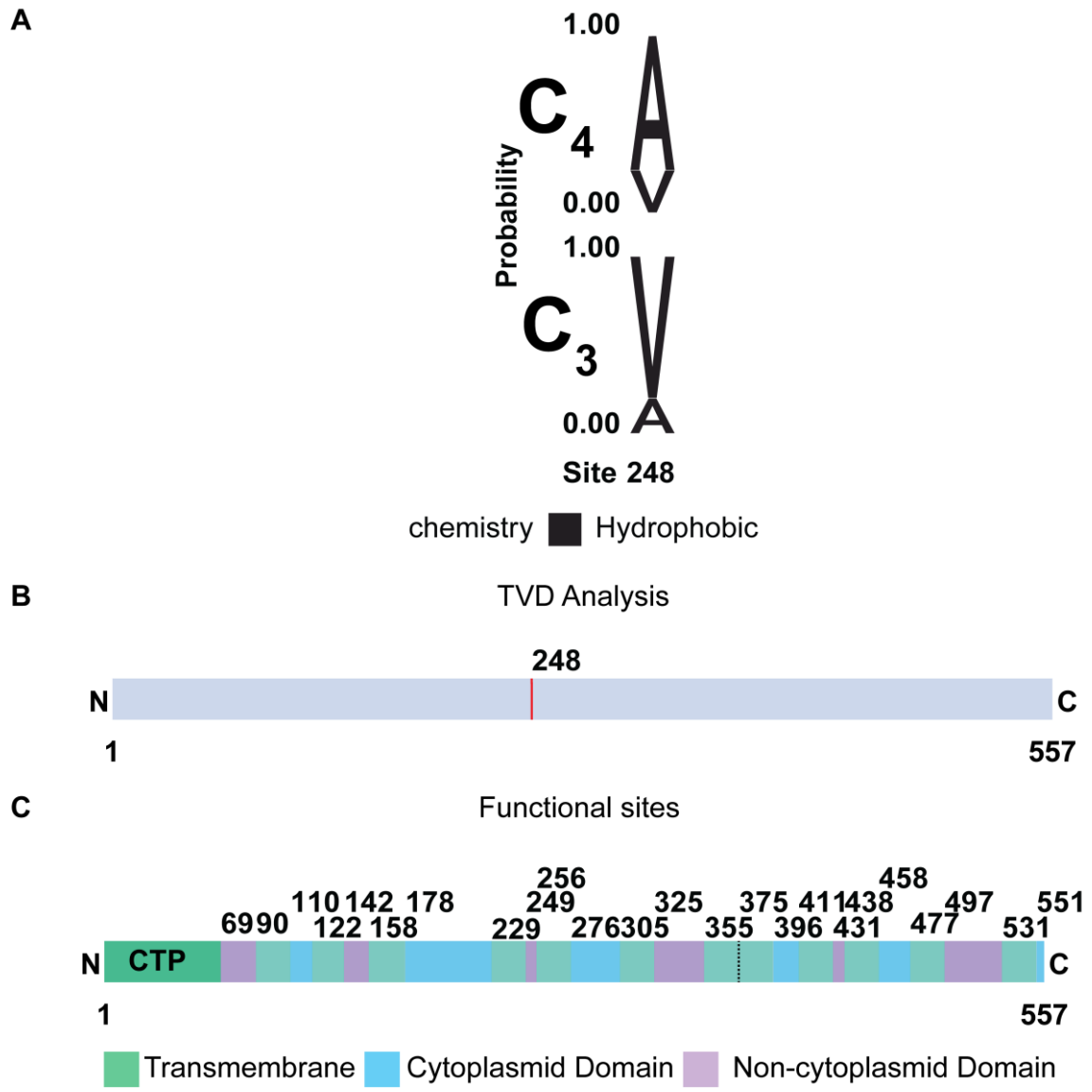


Figure 3.9: Identification of the significant site in dicarboxylate transporter 1 (DiT1, AT5G12860) that are different between the tested C_3 and C_4 species. **A)** Sequence logo revealing the difference in amino acid distribution at the significant site. **B)** Identified site was mapped onto *A. thaliana* DiT1 protein sequence. **C)** Functional domains were mapped onto the DiT1 protein sequence. Dotted line indicating the edge of two separate adjacent transmembrane domains (uniprot and InterPro Q9LXV3). Chloroplast transit peptide (CTP).

3.2.4.2 Variation identified in TPT (AT5G46110)

The triose phosphate/phosphate translocator (TPT) is required in some C_4 plants, mainly in the NADP-ME subtype because they have lower photosystem II (PS II)

activity in the bundle sheath cells. Thus, there is not enough reducing power accessible to reduce all the 3-PGA generated during the photosynthetic carbon reduction cycle to triose, a three-carbon sugar. 3-PGA is therefore transported across to the mesophyll cells where it is reduced to triose. Subsequently, triose is transported back to participate in the photosynthetic carbon reduction cycle to eventually regenerate to RuBP. These various import and export processes, into and from the mesophyll and bundle sheath chloroplast, are carried out by TPT (Weber and von Caemmerer 2010).

The Total Variation Distance algorithm identified three sites that have a different amino acid distribution between the C_3 and C_4 species (Figure 3.10 A). Site 202 is located in a non-cytoplasmic domain and the C_4 all have an isoleucine at this position. However, the majority of C_3 species also have isoleucine at this position (Figure 3.10 B and C).

Site 275 is also located in a non-cytoplasmic domain and changes between the C_3 and the C_4 species. Over 90% of C_3 species have an aspartic acid at this position, whereas 59% of C_4 species have a glycine at this position (Figure 3.10 B and C). These two amino acids differ significantly in their mass and biochemical properties.

Site 370 is located in a transmembrane helix and both C_3 and C_4 species have hydrophobic amino acids at that site. The C_3 variants have a more heterogeneous composition with valine, isoleucine and alanine whereas the C_4 variants almost fully converged to valine (Figure 3.10 B and C).

The sites in maize are isoleucine, aspartic acid and alanine (corresponding to site 187,260 and 355 in GRMZM2G083344 respectively) and are thus quite different to the commonly occurring variants observed in other C_4 species. These sites could form an interesting engineering target because it is known from C_3 plants that TPT operates

close to saturation and therefore could become flux limiting under ideal photosynthetic conditions (Weber and von Caemmerer 2010).

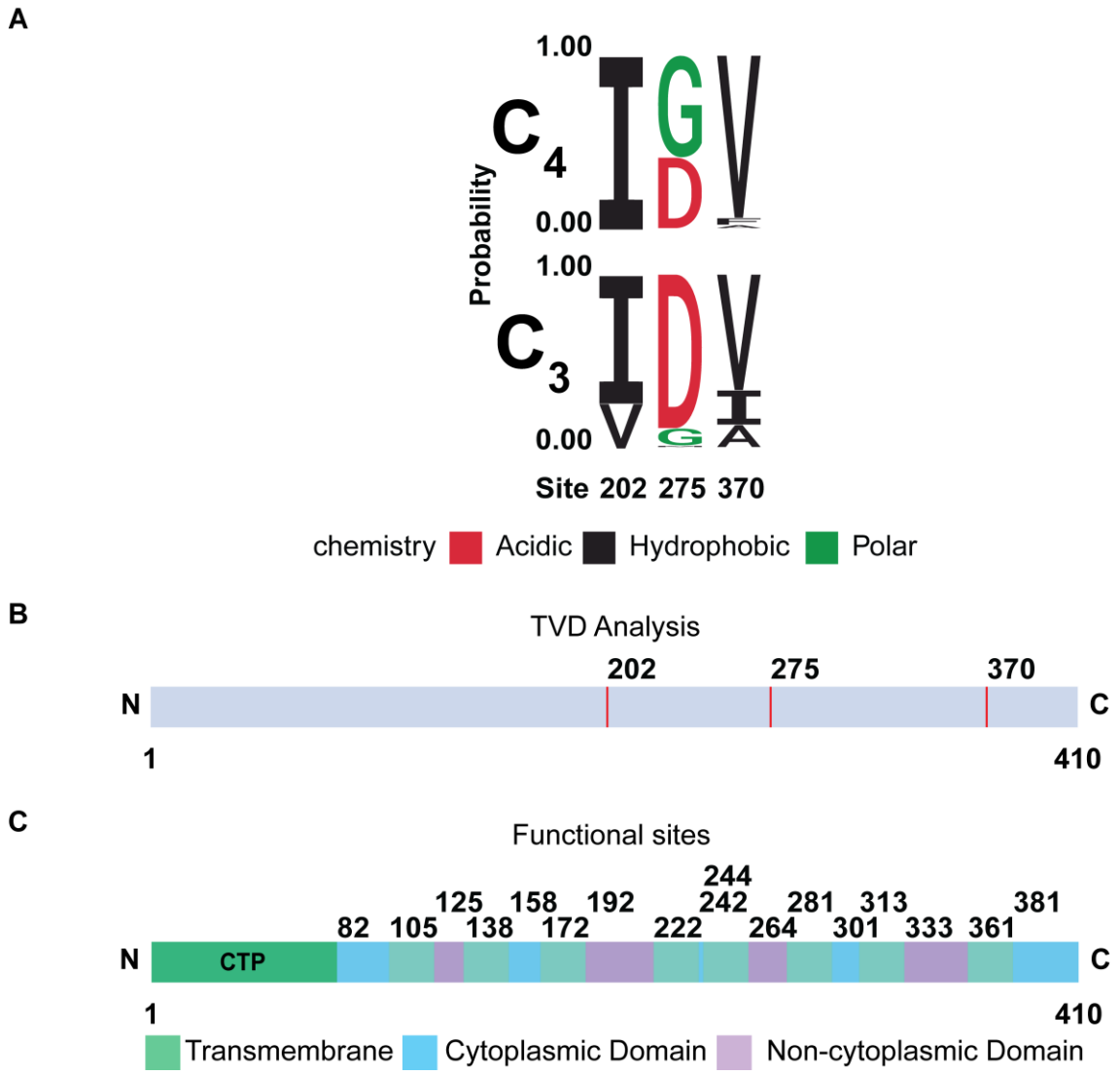


Figure 3.10: Identification of significant sites in the triose phosphate / phosphate translocator (TPT, AT5G46110) that are different between the sampled C_3 and C_4 species. **A)** Sequence logo revealing the difference in amino acid distribution at each significant site. **B)** Identified sites were mapped onto the *A. thaliana* TPT protein sequence. **C)** Functional domains were mapped onto the TPT protein sequence (uniprot Q9ZSR7). Chloroplast transit peptide (CTP).

3.2.4.3 Variation identified in BASS2 (AT2G26900)

Bile acid:sodium symporter family protein 2 (BASS2) is localized at the chloroplast envelop where it enables the sodium dependent pyruvate import into the mesophyll cell chloroplasts (Furumoto et al. 2011). The primary structure of BASS2 is highly conserved across C_3 and C_4 plants with the exception for the chloroplast targeting peptide (Furumoto et al. 2011). It is 43.6 kDa in size and has nine transmembrane domains (Figure 3.11).

The Total Variation Distance algorithm identified two sites in BASS2 that have a different amino acid distribution between C_3 and C_4 species (Figure 3.11 A). Site 177 is located within the third transmembrane domain (counting from the *N*-terminus).

The C_3 species all have a phenylalanine at that position whereas in the C_4 variants the proportion of phenylalanine decreases and new amino acids such as leucine and tyrosine emerge. Tyrosine is structurally very similar to phenylalanine but possesses an additional hydroxy group at the phenyl group. Thus, tyrosine can be classified both as hydrophobic and polar. Site 177 is located between the transmembrane domain and the non-cytoplastic domain (Figure 3.11 C). Therefore, tyrosine is potentially preferred because of these dual biochemical properties at the interface of the two domains (Figure 3.11 C). The orthologous sequence in maize GRMZM2G157912 is missing the corresponding domain and thus no corresponding site was identified through sequence alignment.

The majority of the C_3 species have a hydrophobic alanine (75%) at the position 271. During the evolution of the C_4 trait, the majority of the C_4 variants have changed to a leucine (74%) at this position, which is present to a small degree in the C_3 species (7%) (Figure 3.11 B and C). Site 271 is located between the transmembrane domain and

the cytoplastic domain (Figure 3.11 C). The *Z. mays* BASS2 ortholog seems heavily truncated and has a valine at the corresponding site (site 2 in GRMZM2G157912). This could potentially explain why BASS2 is not highly expressed in *Z. mays* mesophyll cells (Huang et al. 2017; John et al. 2014). Therefore, these identified engineering targets do not receive top priority to alter.

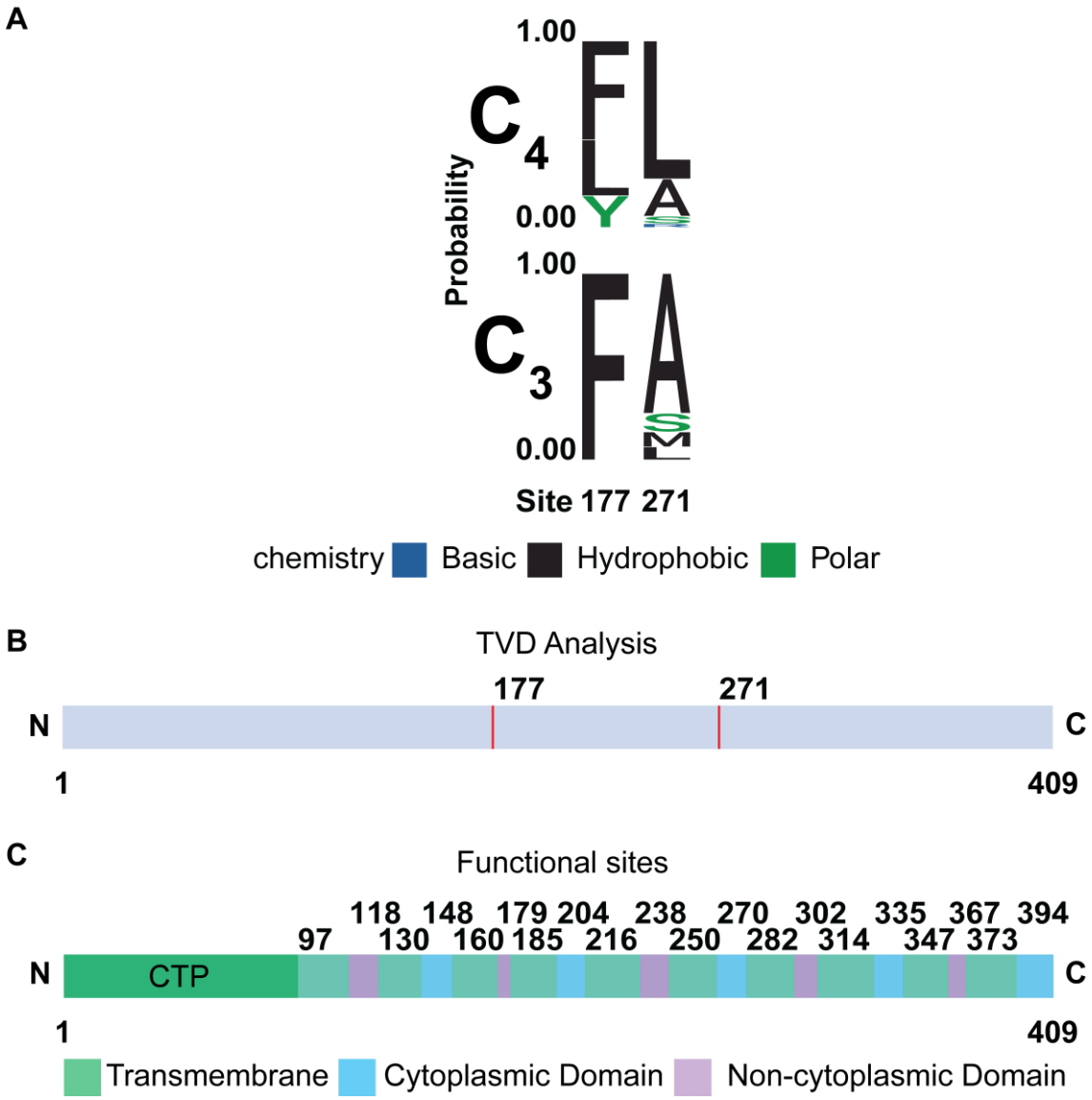


Figure 3.11: Identification of significant sites in bile acid:sodium symporter family protein 2 (BASS2, AT2G26900) that are different between the tested C_3 and C_4 species. **A)** Sequence logo revealing the difference in amino acid distribution at each significant site. **B)** Identified

sites were mapped onto the *A. thaliana* BASS2 protein sequence. **C)** Functional domains were mapped onto the BASS2 protein sequence (InterPro Q1EBV7). Chloroplast transit peptide (CTP).

3.2.4.4 Variation identified in PPT (AT5G33320)

The phosphoenolpyruvate phosphate translocator (PPT) exports PEP in exchange for a phosphate. In C_3 plants the function of PPT is to supply the plastidic metabolism with PEP where it can enter the shikimate pathway as a substrate (Fischer et al. 1997). In C_4 plants, the export of PEP into the cytoplasm is part of the C_4 cycle where it serves a primary carbon acceptor (Bräutigam et al. 2008).

The Total Variation Distance algorithm identified two sites that have a different amino acid distribution between C_3 and C_4 species (Figure 3.12 A). The site 134 is located within a non-cytoplasmic domain and the C_4 species have proline (33.5%), histidine (32.5%) and glutamine (22.5%) present. The C_3 species predominantly have histidine (87%) available at that site. It is interesting to note, however, that proline and glutamine are also present in the C_3 species but at a much lower frequency, 10% and 3% respectively. *Z. mays* has at the corresponding position (site 115 in GRMZM2G047404) the C_4 variant proline present.

The site 330 is placed in a transmembrane domain. The amino acid distribution at that site reveals a convergence to mostly leucine in the C_4 species whereas in the C_3 species a more heterogeneous distribution of leucine, isoleucine and valine is present (Figure 3.12 B and C). The corresponding site in *Z. mays* has an isoleucine present (site 311 in GRMZM2G047404). As all of these residues have similar properties and

lie within a transmembrane helix. Overall none of the sites in this gene are considered priorities for engineering enhanced C_4 photosynthesis in maize.

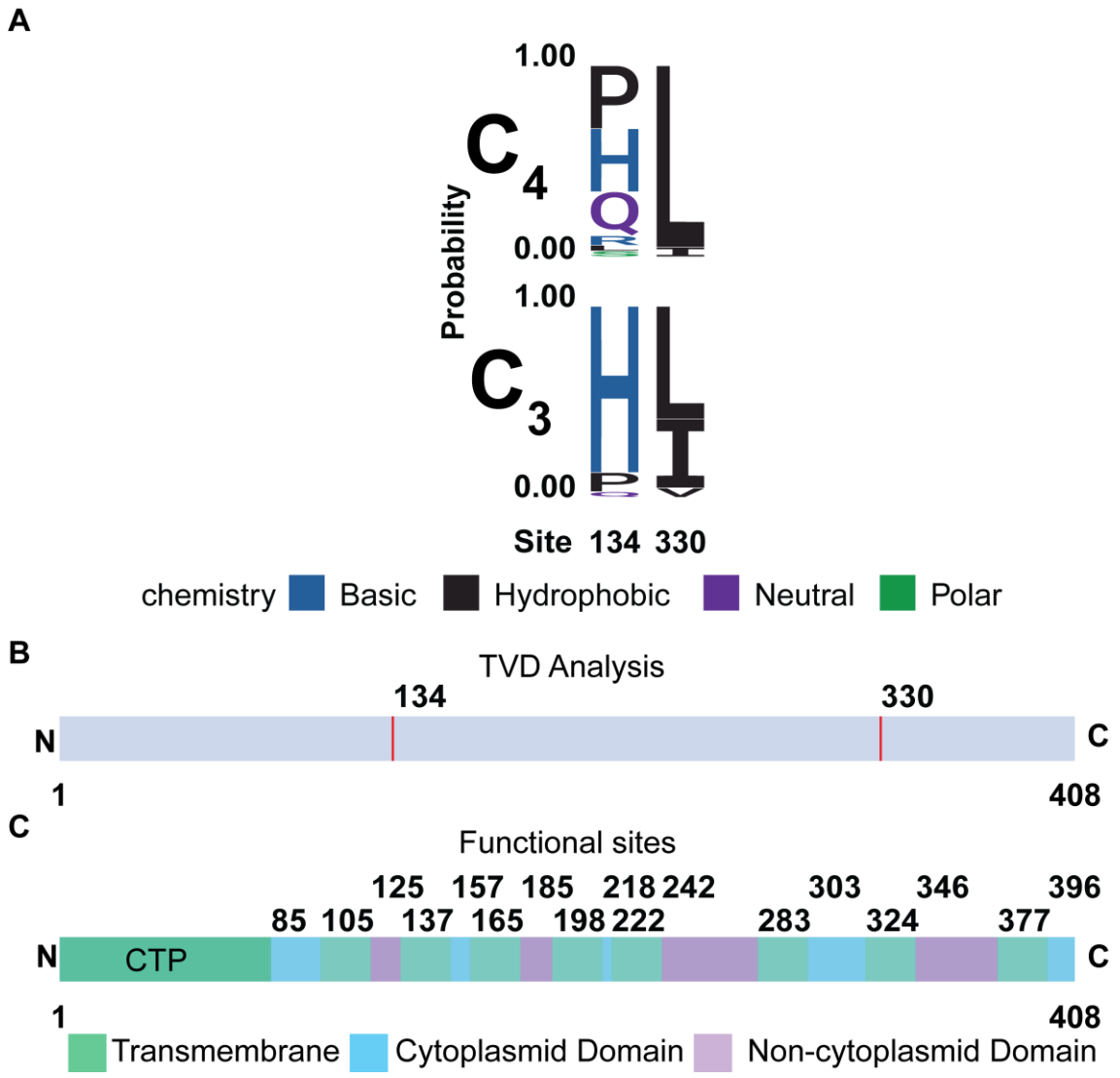


Figure 3.12: Identification of significant sites in Phosphoenolpyruvate phosphate translocator (PPT, AT5G33320) that have a different amino acid distribution between C_3 and C_4 plants. **A)** Sequence logo depicting the amino acid distribution at the significant sites. **B)** Identified sites were mapped onto the *A. thaliana* PPT protein sequence. **C)** Functional domains were mapped onto the PPT protein sequence (uniprot and InterPro Q88RXN3). Chloroplast transit peptide (CTP).

3.2.5 Enzymes of the photorespiration cycle

RuBisCO is a dual catalytic enzyme with carboxylase and oxygenase activity. The carboxylation reaction results in two molecules of 3-phosphoglycerate (3-PGA). If RuBisCO, however, catalyses a reaction with oxygen, the reactions yields one molecule of 3-PGA and one molecule of 2-phosphoglycolate (2-PG). 2-PG is toxic to plant cells and requires detoxification via a multi-step process called photorespiration (Peterhansel et al. 2010). This multi-step process consists of sixteen enzymes spread across the chloroplasts, mitochondria and peroxisome (Douce et al. 2001). The enzymes that have a different amino acid distribution between the C_3 and C_4 species in the dataset are GLDP1, SGAT and HPR2. It is interesting to point out that these three enzymes catalyse successive reactions in different cellular compartments. GLDP1 is located in the mitochondria where it catalyses the decarboxylation of glycine yielding a serine. This serine is transported across to the peroxisome where SGAT catalyses the transamination reaction of serine yielding a hydroxypyruvate. Hydroxypyruvate remains in the peroxisome or enters the cytosol where HPR2 reduces it to glycerate (Buchanan et al. 2000). Below, the sites identified in the enzymes involved in photorespiration are presented and discussed.

3.2.5.1 Variation identified in GLDP1 (AT4G33010)

The P protein (GLDP1) is part of the glycine decarboxylase (GLD) multi enzyme complex. It is localised in the mitochondria where it catalyses the decarboxylation of glycine that is exported from the peroxisomes during the photorespiration cycle. The multi enzyme complex consists of four different enzymes labelled P-protein (homodimer, 200 kDa), H-Protein (monomeric, 14 kDa), T-protein (monomeric, 41 kDa) and L-protein (homodimer, 100 kDa) (Douce et al. 2001; Engel et al. 2007).

The Total Variation Distance algorithm identified a single site that was different between the C_3 and C_4 species. At site 739, the C_4 species in the dataset all have an aspartic acid whereas the C_3 species is more heterogeneous and have aspartic acid (75%), glutamic acid (24%) and a valine (1%) present at the same site (Figure 3.13 A).

The site 739 is not particularly close to any known functional sites of GLDP1, but is within ten amino acids of a site that is involved in the binding of pyridoxal 5-phosphate (Figure 3.13 B and C). The corresponding site in *Z. mays* has a glutamic acid present (site 740 in GRMZM2G104310). Given that the rate of photorespiration in C_4 plants is reduced this site is low priority for engineering enhanced C_4 photosynthesis in maize.

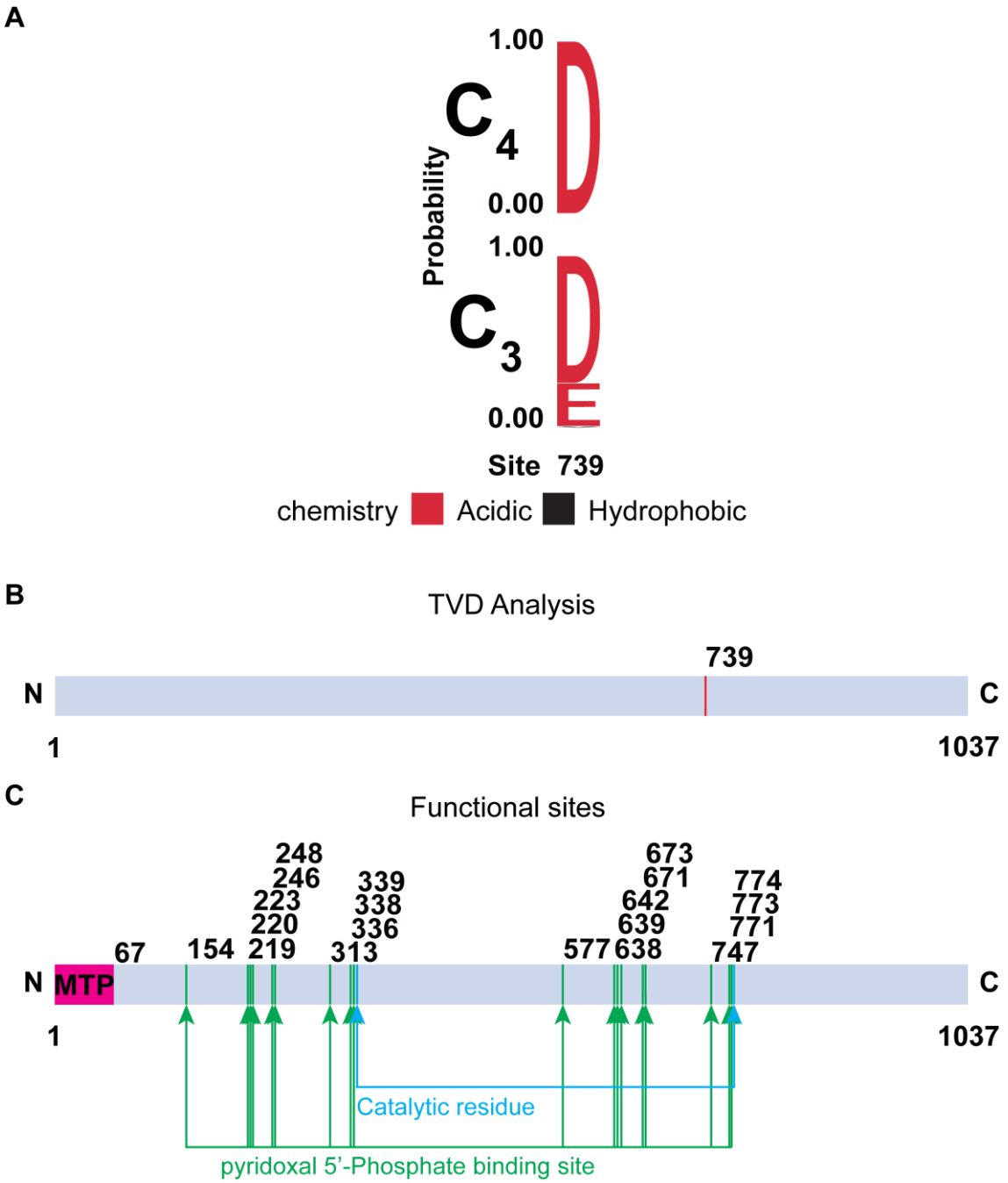


Figure 3.13: Identification of the significant sites in glycine decarboxylase P protein 1 (GLDP1, AT4G33010) that are different between the C_3 and C_4 species. **A)** Sequence logo revealing the difference in amino acid distribution at the significant site. **B)** Identified site was mapped onto the *A. thaliana* GLDP1. **C)** Functional sites were mapped onto the *A. thaliana* GLDP1 (InterPro Q94B78). Mitochondrial targeting peptide (MTP).

3.2.5.2 Variation identified in SGAT (AT2G13360)

Serine:glyoxylate aminotransferase (SGAT) is located in the peroxisome where it catalyses the transamination reaction of serine and glyoxylate to hydroxypyruvate and glycine (Zhang et al. 2013; Zhang et al. 2015). The size of SGAT can vary considerably across different species ranging from 85 kDa to 185 kDa (Zhang et al. 2015). In addition, the catalytic properties of the enzyme can differ among species (Zhang et al. 2015).

The Total Variation Distance algorithm identified a single site at position 240 that exhibited a difference in amino acid distribution between C_3 and C_4 species (Figure 3.14 A). All C_3 species in the dataset have a phenylalanine at this position, while many C_4 species have substituted this residue for a histidine or serine. Both histidine and serine have different biochemical properties compared to the ancestral phenylalanine. The corresponding site in in *Z. mays* has the C_3 variant phenylalanine present (site in 242 in GRMZM2G137868). The identified site is nine amino acids away from two homodimer binding sites. It is, however, given that photorespiration is generally reduced in C_4 plants compared to C_3 plants, problematic to speculate about a potential improvement of the enzymatic function through the altered amino acids at that site (Figure 3.14 B and C). Thus, this site would be low priority for engineering enhanced C_4 photosynthesis in *Z. mays*.

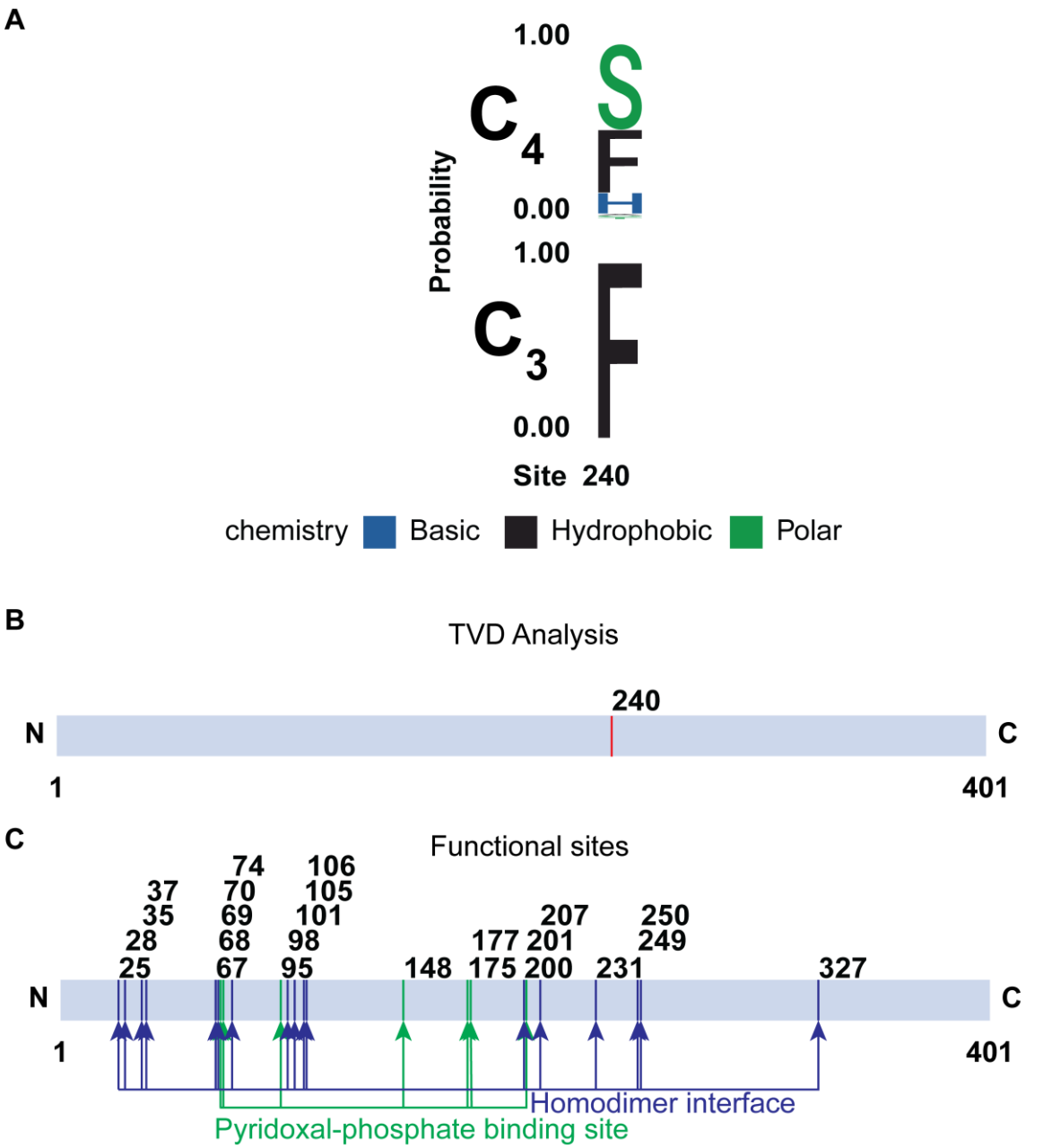


Figure 3.14: Identification of significant sites in Serine:glyoxylate aminotransferase (SGAT AT2G13360) that are different between the sampled C_3 and C_4 species. **A)** Sequence logo revealing the difference in amino acid distribution at the significant site. **B)** Identified site was mapped on the *A. thaliana* SGAT protein sequence. **C)** Functional sites mapped onto SGAT protein sequence (InterPro Q56YA5).

3.2.5.3 Variation identified in HPR2 (AT1G79870)

During the photorespiratory cycle, hydroxypyruvate originating from the transamination of serine remains in the peroxisome where it is reduced to glycerate consuming one NADP. This reaction is mostly catalysed by the peroxisomal enzyme hydroxypyruvate reductase 1 (HPR1) and to a smaller degree by the cytosolic HPR2 (Timm et al. 2011; Timm et al. 2008).

Both proteins HPR1 and HPR2 were analysed by the Total Variation Distance algorithm. One site in HPR2 was identified as being different between the two groups of C_3 and C_4 species (Figure 3.15 A). The C_3 species have three different amino acids present at site 52 in the multiple sequence alignment, asparagine, serine and aspartic acid. The most frequently occurring residue is asparagine. However, the C_4 species are all asparagine at this site (Figure 3.15 A). The site is located towards the *N*-terminus of the protein (Figure 3.15 B) and the functional relevant sites are located towards the *C*-terminus of the protein (Figure 3.15 C). Moreover, the corresponding site in *Z. mays* (site 52 in GRMZM2G159587) has a glycine at that site. Again, considering that photorespiration is reduced in C_4 plants compared to C_3 plants and that HPR2 is involved only to a small degree in the reduction of hydroxypyruvate it is likely that the identified sites have only a minor impact on the overall flux through the photorespiration cycle. It is difficult to speculate on what the functional significance of this might be and therefore this site would be low priority.

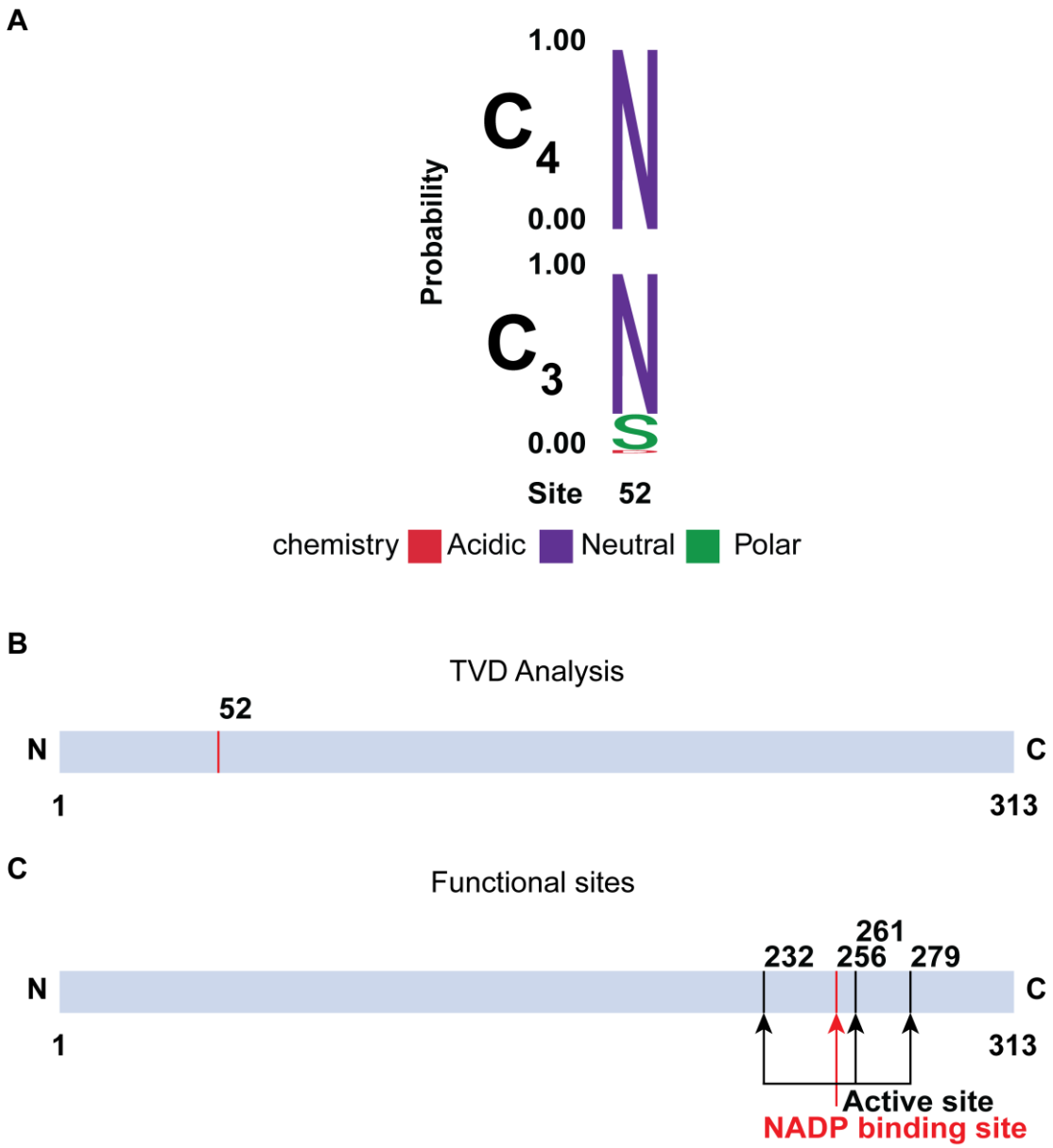


Figure 3.15: Identification of significant sites in hydroxypyruvate reductase 2 (HPR2, AT1G79870) that have a different amino acid distribution between the sampled C_3 and C_4 species. **A)** Sequence logo revealing the composition of amino acids at the significant site. **B)** Identified site was mapped onto the *A. thaliana* HPR2 protein sequence. **C)** Functional sites mapped onto the HPR2 protein sequence (uniprot Q9CA90).

3.3 Chapter conclusion

The Total Variation Distance algorithm was applied to a large transcriptomic dataset consisting of 30 C_4 plants and 17 C_3 plants. The analysis focused on proteins linked to C_4 photosynthesis. Out of 77 proteins analysed, 12 proteins had sites that had a different amino acid distribution between the C_4 and C_3 species present in the dataset. The proteins with significantly different sites can roughly be classified into three groups: proteins involved in photorespiration, transport proteins and enzymes of the core C_4 cycle.

For experimental assessment, the proteins with significant sites belonging either to the group of transport proteins or the enzymes of the core C_4 cycle are of most interest to follow up on. In the context of improving C_4 photosynthesis, the enzymes belonging to the photorespiration cycle are targets of lesser importance. It was previously estimated that the rate of photorespiration in C_4 plants e.g. maize is about 2% compared to the rate of photorespiration in tobacco (Zelitch 1968). Although photorespiration is heavily reduced, it still takes place in C_4 plants and interestingly maize plants with a non-functional glycolate oxidase are not viable at ambient CO_2 levels (Zelitch et al. 2009). Thus, the reduction of the rate of photorespiration in C_4 plants could have reduced the evolutionary pressure on the enzymes associated with it and the sites that have a different amino acid distribution between C_3 and C_4 species might be explained through relaxed selection in the C_4 plants.

The transport of metabolites in the C_4 cycle is usually modelled as exceeding the rate of RuBisCO carbon fixation (Bellasio 2016; Laisk and Edwards 2000). This assumption may or may not be correct, as transport rates in plants are difficult to measure and rates are estimated from *in vitro* assays or assays in exogenous systems such as *E. coli* or yeast. Thus, it may be that changes in amino acid sequences of the

transporter genes alter the transport rate. It may also be that these changes alter the substrate specificity making them more specific for the substrates of the C_4 cycle (indeed some evidence for both of these effects is provided in Chapter 5 of this thesis where I characterise different transport rates and activities of a C_3 and a C_4 variant of the same transporter gene). Thus, there are multiple ways in which engineering the transporter might impact on the overall photosynthetic efficiency in C_4 plants.

From the perspective of improving photosynthesis in maize and other C_4 crop plants, the most promising candidates to engineer are likely to be PEPC and PPDK. PEPC catalyses the primary carboxylation and it is thought to be rate limiting in a scenario where the stomata are not fully open (Section 3.1.2). Thus, perhaps not surprisingly, PEPC had the largest number of sites (twelve) with a different amino acid distribution between C_3 and C_4 species in this dataset. The extent of the change that is observed is indicative that substantial improvement of PEPC for function in C_4 photosynthesis may be possible. Interestingly, out of the twelve sites identified by the Total Variation Distance algorithm, four sites correspond to sites identified by Christin et al. 2007. The convergence on these four sites is interesting as Christin et al. 2007 utilised a dataset consisting entirely of monocots whereas the dataset applied in this study consists only of dicots.

PPDK catalyses the step from pyruvate to PEP and thus, provides the substrate for PEPC. The substrate regeneration is thought to be rate limiting (Section 3.1.2) and increasing the abundance of PEP might contribute to a faster overall flux through the C_4 cycle. Furthermore, *Miscanthus* × *giganteus*, a C_4 plant that is closely related to maize (*Z. mays*), remains photosynthetically active 6°C below the minimum growing temperature for maize (Wang et al. 2008). It achieves this through increasing the expression of PPDK. The increase in expression is necessary to keep the PPDK

homotetramer functional because PPDK dissociates at a temperature between 10°C to 15°C (Wang et al. 2008). Thus, creating PPDK variants that required less activation energy to form the homotetramer protein complex could lead to cold-tolerant maize that does not need to increase the expression of PPDK to achieve a similar cold tolerance as *Miscanthus* × *giganteus*. Thus, in addition to potentially altering the kinetics of the enzyme, it may be that the sites identified in this study help promote stability of the homotetramer complex. Both of these properties could be investigated in follow up studies.

3.3.1 Summary

This chapter has identified a set of sites in amino acid sequences of genes related to photosynthetic function in C_4 plants for targeted engineering. The next chapter describes the attempt that was made to engineer these sites and test their functional significance.

3.4 Materials and Methods

This chapter relied on previously published transcriptome assembly files and transcript abundance files that are available for download at Zenodo https://doi.org/10.5281/zenodo.162327 (Kelly et al. 2017). The original work for this chapter started with the prediction of open reading frames in the *de novo* assembled transcriptomes (Section 3.4.6). In order to provide a coherent materials and methods section, the reasoning behind the species selection and the generation of the dataset are outlined below.

3.4.1 Dataset

This chapter utilises data from a transcriptomic dataset that comprises 30 C_4 and 17 C_3 species representing 18 independent origins of C_4 photosynthesis (Figure 3.16). This group of species includes examples from all seven of the orders within the eudicots that have evolved the C_4 trait (*Zygophyllales*, *Malpighales*, *Brassicales*, *Caryophyllases*, *Boraginales*, *Laminales* and *Asteriales*). Thus, this dataset contains a considerable amount of the diversity that is encapsulated by C_4 species (Kelly et al. 2017). Out of the 30 C_4 species, 20 species belong to the C_4 subtype NADP-ME and 8 to the C_4 subtype NAD-ME (Table 3.2). The exact C_4 subtype for the species *Portulaca* suffruticosa and *Blutaparon vermiculare* are not fully identified.

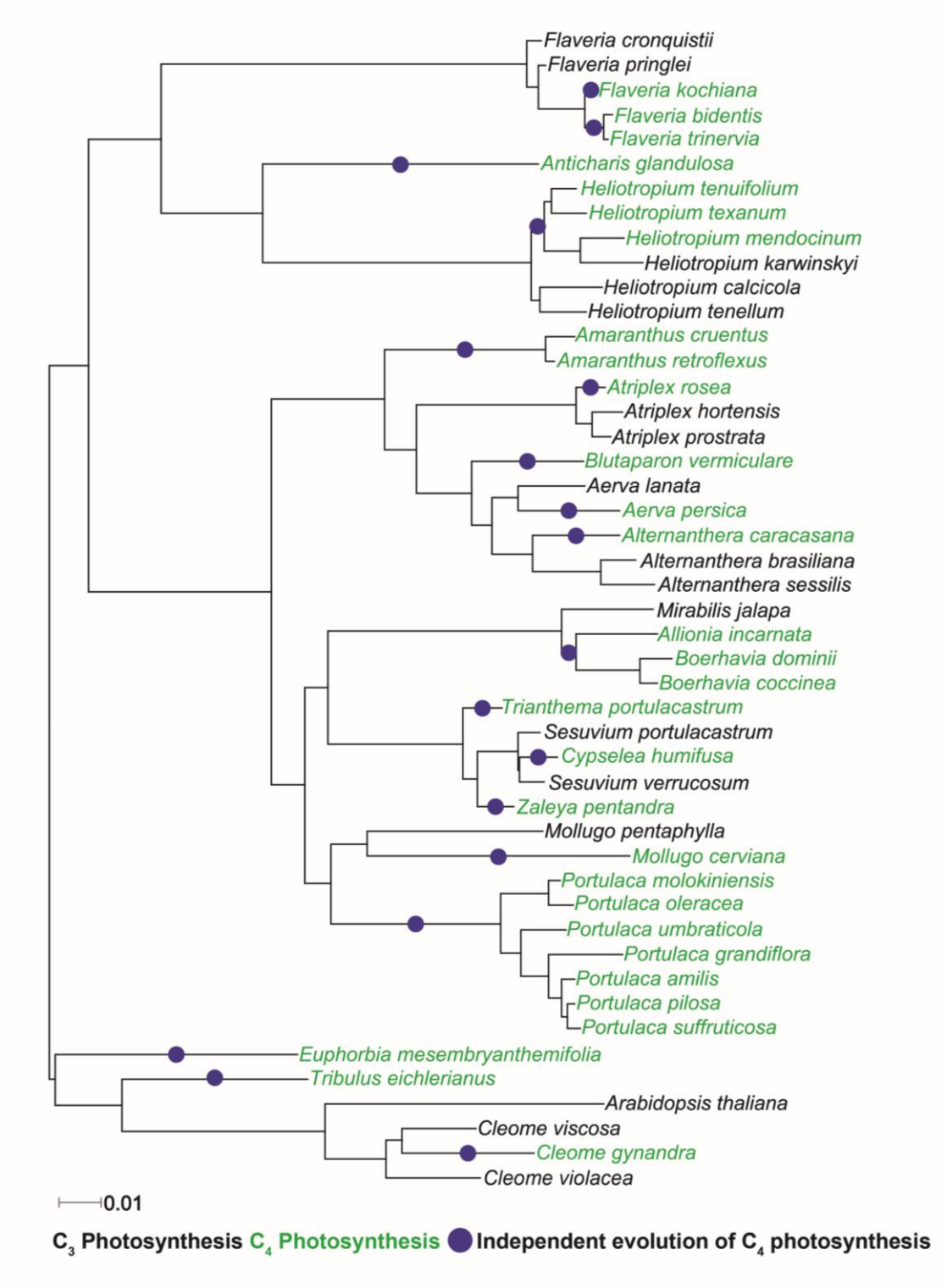


Figure 3.16: Species tree of the species present in the dataset. In black are the plants classified as C_3 photosynthetic plants and in green are the plants classified as C_4 photosynthetic.

Table 3.2: The thirty C_4 plant species from the applied dataset and their C_4 photosynthesis subtype.

Species	C_4 subtype	Reference
Portulaca amilis	NADP-ME	(Voznesenskaya et al. 2010)
Portulaca grandiflora	NADP-ME	(Voznesenskaya et al. 2010)
Portulaca pilosa	NADP-ME	(Voznesenskaya et al. 2010)
Portulaca umbraticola	NADP-ME	(Voznesenskaya et al. 2010)
Portulaca molokiniensis	NAD-ME	(Voznesenskaya et al. 2010)
Portulaca oleracea	NAD-ME	(Voznesenskaya et al. 2010)
Portulaca suffruticosa	unknown	
Mollugo cerviana	NADP-ME	(Christin et al. 2011b)
Cypselea humifusa	NADP-ME	(Muhaidat et al. 2007)
Zaleya pentandra	NAD-ME	(Muhaidat et al. 2007)
Trianthema portulacastrum	NADP-ME	(Hassan et al. 2005)
Boerhavia coccinea	NADP-ME	(Douglas and Manos 2007)
Boerhavia dominii	NADP-ME	(Douglas and Manos 2007)
Allionia incarnata	NADP-ME	(Douglas and Manos 2007)
Alternanthera caracasana	NADP-ME	(Sage et al. 2007)
Blutaparon vermiculare	unknown	
Aerva persica	NADP-ME	(Sage et al. 2007)
Amaranthus cruentus	NAD-ME	(Sage et al. 2007)
Amaranthus retroflexus	NAD-ME	(Sage et al. 2007)
Atriplex rosea	NAD-ME	(Kadereit et al. 2010)
Heliotropium mendocinum	NADP-ME	(Sage et al. 2011)
Heliotropium tenuifolium	NADP-ME	(Sage et al. 2011)
Heliotropium texanum	NADP-ME	(Sage et al. 2011)
Anticharis glandulosa	NAD-ME	(Kellogg 1999)
Flaveria bidentis	NADP-ME	(McKown et al. 2005)
Flaveria trinervia	NADP-ME	(McKown et al. 2005)
Flaveria kochiana	NADP-ME	(McKown et al. 2005)
Cleome gynandra	NAD-ME	(Feodorova et al. 2010)
Euphorbia mesembryanthemifolia	NADP-ME	(Steinmann and Porter 2002)
Tribulus eichlerianus	NADP-ME	(Sheahan and Chase 1996)

3.4.2 Plant growth

The plants used in this study were grown as described previously (Sage et al. 2013). In short, all plants were grown in 20 litre pots containing a sandy-loam soil and placed in a greenhouse at the University of Toronto. The average temperature in the

greenhouse during the day and night time were 28°C and 23°C respectively. The plants were watered daily and fertilised weekly (Kelly et al. 2017; Sage et al. 2013).

3.4.3 Isolation of RNA

Plant material was collected as described previously (Johnson et al. 2012) in June 2010 between 09:00 and 13:00. Leaves were collected from two to four plant species to obtain a pooled sample between 0.1 g to 1.0 g of leaf tissue. Mature leaves were sampled from the most recent fully expanded leaves. Juvenile leaf samples were collected from fully emerged but not fully expanded leaves. RNA was extracted from the pooled leaf samples following standard RNA isolation protocols (Johnson et al. 2012; Lyu et al. 2015). Subsequently, the RNA samples were shipped to BGI for paired-end sequencing (Kelly et al. 2017).

3.4.4 *De novo* assembly of short reads

Raw reads of plant species that had multiple libraries such as juvenile and mature leaf samples were merged preceding to the *de novo* assembly of the short reads. Raw reads were quality trimmed by applying trimmomatic (Bolger et al. 2014) to remove bases with low quality and adaptor sequences. The settings used for the raw read processing by trimmomatic were: `LEADING: 10 TRAILING: 10 SLIDINGWINDOW:5:15 MINLEN:50`. In addition to the quality filtering, reads originating from ribosomal RNA were removed using SortMeRNA (Kopylova et al. 2012) applying the default settings. Subsequently, BayesHammer (Nikolenko et al. 2013) was applied with the default settings to perform a read error correction. The read numbers were normalised using khmer (Crusoe et al. 2015) with the following settings: `-C 20 -k 21 -M 8e9`. Next, the overlapping paired-end reads were concatenated applying ALLPATHS-LG (Gnerre et al. 2011). All reads, including the non-

concatenated once, were used as an input for a preliminary *de novo* assembly with SGA (Simpson and Durbin 2010) and requiring a minimum overlap size of 80 nucleotides without any mismatches. The obtained contigs underwent four additional rounds of assembly and for each additional round the number of minimum overlap size was reduced by 10 nucleotides. The trimmed and normalised reads were subsequently mapped to the *de novo* assembly using bowtie2 (Langmead and Salzberg 2012). Reads that are missing in the assembly were identified and added to the assembly file. Subsequently, the unmapped reads and the assembled contigs underwent four rounds of assembly using SGA starting with an overlap size of 70 and reducing the overlap size by 10 nucleotides for the following round. Next, SSPACE (Boetzer and Pirovano 2014) was used with following settings, `-k 10 -a 0.7 -n 50 -o 20` to create a scaffold out of the generated contigs together with the full set of non-ribosomal, trimmed, normalised paired-end reads. The obtained scaffolds underwent gap filling using the SGA gap filling function. As a final step, the assembled contigs underwent a base-error correction applying Pilon (Walker et al. 2014) with the default parameters (Kelly et al. 2017).

3.4.5 List of genes closely associated to the C_4 cycle

The set of genes that are closely linked to the C_4 cycle were obtained from the same study, and comprised those genes that were consistently upregulated or downregulated across all C_4 plants in the dataset and that were connected, or peripheral to the C_4 cycle in some way (Kelly et al. 2017) (Table 3.3 to Table 3.5).

Table 3.3: List of genes and corresponding accession number of genes analysed with the Total Variation Distance algorithm.

Gene	Accession Number	Protein	Cell compartment
GLDT1	AT1G11860	Aminomethyltransferase	Mitochondria
ALAAT1	AT1G17290	Alanine aminotransferase 1	Mitochondria
GLYR2	AT1G17650	Glyoxylate/succinic semialdehyde reductase 2	Chloroplast
ATCAT3	AT1G20620	Catalase-3	Peroxisome
PGM3	AT1G23190	Probable phosphoglucomutase	Cytosol
GGT1	AT1G23310	Glutamate--glyoxylate aminotransferase 1	Peroxisome
PLGG1	AT1G32080	Plastidal glycolate/glycerate translocator 1	Chloroplast
GDCH2	AT1G32470	Glycine cleavage system H protein 3	Mitochondria
MTLPD1	AT1G48030	Dihydrolipoyl dehydrogenase 1	Mitochondria
GLN1;5	AT1G48470	Glutamine synthetase cytosolic isozyme 1-5	Cytosol
ALATS	AT1G50200	Alanine-tRNA ligase	Mitochondria
PEPC1	AT1G53310	Phosphoenolpyruvate carboxylase 1	Chloroplast
CPGK2	AT1G56190	Phosphoglycerate kinase 2	Chloroplast
RBCS1A	AT1G67090	Ribulose bisphosphate carboxylase small chain 1A	Chloroplast
HPR1	AT1G68010	Glycerate dehydrogenase HPR	Peroxisome
GGT2	AT1G70580	Glutamate--glyoxylate aminotransferase 2	Peroxisome
ALAAT2	AT1G72330	Alanine aminotransferase 2	Mitochondria
ENO1	AT1G74030	Enolase 1	Chloroplast
LL-DAP-AT	AT1G77670	Pyridoxal phosphate (PLP)-dependent transferases	Chloroplast
PGK3	AT1G79550	phosphoglycerate kinase	Cytosol
ME4	AT1G79750	NADP-dependent malic enzyme 4	Chloroplast
HPR2	AT1G79870	Glyoxylate/hydroxypyruvate reductase A	
GLYK	AT1G80380	D-glycerate 3-kinase	Chloroplast
GAD4	AT2G02010	Glutamate decarboxylase 4	Cytosol
SGAT	AT2G13360	Serine--glyoxylate aminotransferase	Peroxisome
TPI	AT2G21170	Triosephosphate isomerase	Chloroplast
GLDP2	AT2G26080	Glycine dehydrogenase (decarboxylating) 2	Mitochondria
BASS2	AT2G26900	Sodium/pyruvate cotransporter BASS2	Chloroplast
ENOC	AT2G29560	Cytosolic enolase 3	Cytosol

Table 3.4: Continuation of table 3.3

ASP1	AT2G30970	Aspartate aminotransferase	Mitochondria
GDH2	AT2G35120	Glycine cleavage system H protein 2	Mitochondria
GDCH1	AT2G35370	Glycine cleavage system H protein 1	Mitochondria
ENO2	AT2G36530	Bifunctional enolase 2/transcriptional activator	Cytosol, Mitochondria, Nucleus
AGT3	AT2G38400	Alanine--glyoxylate aminotransferase 2 homolog 2, mitochondrial	Mitochondria
PEPC2	AT2G42600	Phosphoenolpyruvate carboxylase 2	Cytosol
PPT2	AT3G01550	Phosphoenolpyruvate/phosphate translocator 2	Chloroplast
PDK	AT3G06483	Pyruvate dehydrogenase (acetyl-transferring)] kinase	Mitochondria
HAOX1	AT3G14130	Peroxisomal (S)-2-hydroxy-acid oxidase	Peroxisome
GLO2	AT3G14415	2-hydroxy-acid oxidase GLO2	Peroxisome
GLO1	AT3G14420	2-hydroxy-acid oxidase GLO1	Chloroplast, Cytosol, Nucleus, Peroxisome
PEPC3	AT3G14940	phosphoenolpyruvate carboxylase 3	Chloroplast
ASPGB1	AT3G16150	Probable isoaspartyl peptidase/L-asparaginase 2	
NHD1	AT3G19490	Sodium/proton antiporter 1	Chloroplast
POP2	AT3G22200	Gamma-aminobutyrate transaminase POP2	Mitochondria
PK	AT3G22960	Plastidial pyruvate kinase 1	Chloroplast
GLYR1	AT3G25530	Glyoxylate/succinic semialdehyde reductase 1	Cytosol
CSR1	AT3G48560	Acetolactate synthase	Chloroplast
PUMP1	AT3G54110	Mitochondrial uncoupling protein 1	Mitochondria
CTIMC	AT3G55440	Triosephosphate isomerase	Cytosol
NDH com I	AT4G02580	NADH dehydrogenase flavoprotein 2	Mitochondria
PPDK	AT4G15530	Pyruvate, phosphate dikinase 1	Chloroplast
PPDK-rp	AT4G21210	Pyruvate, phosphate dikinase regulatory protein 1	Chloroplast
GLDP1	AT4G33010	Glycine dehydrogenase (decarboxylating) 1	Mitochondria
CA5	AT4G33580	Beta carbonic anhydrase 5	Chloroplast
CAT	AT4G35090	Catalase-2	Peroxisome
SHM1	AT4G37930	Serine hydroxymethyltransferase 1	MT

Table 3.5: Continuation of table 3.3

AGT2	AT4G39660	Alanine--glyoxylate aminotransferase 2 homolog 1	Mitochondria
GLU1	AT5G04140	Ferredoxin-dependent glutamate synthase 1	Chloroplast, Mitochondira
GLCDH	AT5G06580	D-lactate dehydrogenase	Mitochondria
ASPGA1	AT5G08100	L-asparaginase 1	Cytosol
ASN3	AT5G10240	Asparagine synthetase [glutamine-hydrolyzing] 3	Cytosol
ASP3	AT5G11520	Aspartate aminotransferase 3	Chloroplast
DIT1	AT5G12860	Dicarboxylate transporter 1	Chloroplast
ASP2	AT5G19550	Aspartate aminotransferase	Cytosol
SHM2	AT5G26780	Serine hydroxymethyltransferase 2	Mitochondria
CUE1, PPT1	AT5G33320	Phosphoenolpyruvate/phosphate translocator 1	Chloroplast
GS2	AT5G35630	Glutamine synthetase	Chloroplast, Mitochondria
PGLP1	AT5G36700	Phosphoglycolate phosphatase 1A	Chloroplast
RBCS3B	AT5G38410	Ribulose bisphosphate carboxylase small chain	Chloroplast
RBCS2B	AT5G38420	Ribulose bisphosphate carboxylase small chain 2B	Chloroplast
RBCS1B	AT5G38430	Ribulose bisphosphate carboxylase small chain 1B	Chloroplast
TPT	AT5G46110	Triose phosphate/phosphate translocator TPT	Chloroplast
BOU	AT5G46800	Mitochondrial carnitine/acylcarnitine carrier-like protein	Mitochondria
ATPK5	AT5G47760	Phosphoglycolate phosphatase 2	Chloroplast
AMK2	AT5G47840	Adenylate kinase 2	Chloroplast
PGMP	AT5G51820	Phosphoglucomutase	Chloroplast
MDH	AT5G58330	Malate dehydrogenase	Chloroplast
UCP2	AT5G58970	Mitochondrial uncoupling protein 2	Mitochondria
DIT2.2	AT5G64280	Dicarboxylate transporter 2.2	Chloroplast
DIT2.1	AT5G64290	Dicarboxylate transporter 2.1	Chloroplast

3.4.6 Initial processing of *de novo* assembled transcriptomes

Open reading frames in the *de novo* assembled transcriptomes were predicted using computational methods (TransDecoder (Haas et al. 2013) Version 2.0.1 and Prodigal (Hyatt et al. 2010b) Version 2.6.3). In cases where both methods predicted an open reading frame for a transcript then the priority was given to the prediction made by TransDecoder. TransDecoder was run in default mode. Prodigal was run with the genetic code set to eukaryotic default. Transcripts without predictions by TransDecoder or Prodigal were subsequently subject to an additional round of gene

prediction using Prodigal with the genetic code set to bacterial to facilitate the identification of genes from the chloroplast and mitochondrion. Subsequently, the predicted amino acid sequences from each species were subject to orthogroup inference using OrthoFinder (Emms and Kelly 2015) (VERSION 1.0.6) using the default program settings. Orthogroups containing four or more sequences were kept for further analysis.

3.4.7 Generation of multiple sequence alignments

Amino acid sequences of all the species belonging to a given orthogroup were aligned with mafft (Katoh and Standley 2013; Katoh and Toh 2010). If there were fewer than 200 sequences to be aligned then the settings used for mafft were: `-- reorder -- retree 2 -- thread 10`. If there were more than 200 sequences to be aligned then the settings used for mafft were "`- - reorder - -thread 12`.

3.4.8 Splitting fused orthogroups

Phylogenetic tree of each of the selected orthogroups were inspected to evaluate whether they were a true orthogroup or whether they comprised a fused orthogroup. An orthogroup is the complete set of genes descended from a single copy gene in the most recent common ancestor of the set of species under consideration. OrthoFinder sometimes makes mistakes and accidentally fuses two or more closely related orthogroups together. Phylogenetic analysis of all orthogroups used for this chapter was thus conducted to search for and resolve such fused orthogroups.

3.4.9 Constructing the species tree

A concatenated multiple sequence alignment was constructed to infer a phylogentic tree for the species used in this analysis. Here, orthogroup alignments were considered if they contained a single gene for at least twenty-four out of the forty-seven (>50%) species. In addition to that orthogroups with at least twenty-five phylogenetically informative columns were considered. A column of a multiple sequence alignment was considered as phylogenetically informative if it contained at least two different amino acids in three or more species. Columns containing gaps were not considered as inforamtive as they are not used for tree inference by maximum likelihood methods. Subsequently, informative columns were sampled randomly with replacement and added to the concatenated multiple sequence alignment. In total 1000 such concatenated multiple sequence alignments were created, and for each of these alignments a phylogenetic tree was constructed using FastTree (Price et al. 2010). A consensus tree was created using RAxML (Stamatakis 2014) applying the majority rule method (Aberer et al. 2010).

3.4.10 Filtering the output from the Total Variation Distance algorithm

To identify sites that reveal differences in the distribution of amino acids between the C_3 and C_4 plants in this dataset, the following filter parameters were applied: at least five species in both groups had to have an informative character present, the TVD score had to be ≥0.2 (i.e. at least 20% different between C_3 and C_4 species) and the p-value cut-off set to 0.01. I.e. the probability of observing such a 20% difference by random sampling of the same set of sequences was ≤ 0.01.

Chapter 4: An *in vivo* bacterial system to evaluate the relative function of different variants of proteins involved in C_4 photosynthesis

4.1 Chapter introduction

Out of the twelve proteins that showed sites that have a different amino acid distribution between C_3 and C_4 species, PEPC and PPDK were most interesting and also predicted to be most likely rate limiting (Chapter 3.1.2). Thus, these two enzymes were selected for experimental evaluation. Initially, the idea was to engineer the identified variants directly in to *Setaria viridis*, a C_4 model plant, to test the impact of the individual protein sequence changes on the photosynthetic rate. *Setaria viridis* was chosen here instead of maize as the intention was to ultimately provide proof of concept through CRISPR guided engineering, and thus use of *Setaria* instead of maize would dramatically reduce the time and resources required to do this. If it worked in *Setaria* then the goal would be to transfer the same changes into maize.

Combining the number of identified sites with the hypothesis that the best protein variant has perhaps not evolved yet (Section 1.8) leads to a large combinatorial complexity and a large number of protein variants to test. This large number is not feasible to test in plants. For example, in the case of PEPC, with twelve sites that have an altered amino acid distribution between C_3 and C_4 plants, if we just consider amino acid variants observed in the C_4 species in the dataset that combinatorial complexity translates into 259’200 possible protein variants to engineer into the C_4 model plant *S. viridis*. Even in the case of PPDK, that only had two sites that showed a significant difference in C_4 species has 25 different protein variants to evaluate. Even 25 different transgenic plants is too large a number to consider for experimental evaluation.

Furthermore, changing only a few sites in an enzyme might have merely a small effect on the overall photosynthetic efficiency under greenhouse conditions. Therefore, it would require large sample sizes to detect the impact of such changes. Consequently, an alternative approach to screen the protein variants before the *in planta* engineering is necessary. To address the issue of combinatorial complexity, a bacterial test system was developed to identify the most promising protein variants prior to engineering them into plants.

4.1.1 Background

4.1.1.1 Testing plant proteins in a bacterial system

One important consideration when testing eukaryotic proteins in a prokaryotic system is the lack of post-translational protein modifications that can have an impact on the protein folding and function. Although, it becomes more and more clear that prokaryotes have the capabilities of post-translational protein modifications, these systems are different (and interpreted differently by the cell) to those in eukaryotes (Cain et al. 2014). The low capacity of post-translational protein modification in a prokaryotic host can cause various obstacles for the expression of heterologous proteins (Cain et al. 2014; Tokmakov et al. 2012).

In the context of PEPC and PPDK this is relevant because both enzymes undergo post-translational protein modifications in plants. PEPC is phosphorylated by the light-dependent activity of PEPC protein kinase (PEPC-PK) and is dephosphorylated by the protein phosphatase 2A (Furumoto et al. 2007; Vidal and Chollet 1997). To assess how crucial the phosphorylation of PEPC is to carry out its function in the C_4 cycle, *Flaveria bidentis* RNAi lines were created that harboured a construct targeting the downregulation PEPC-PK (Furumoto et al. 2007). The downregulation of PEPC-PK

resulted in lower levels of phosphorylated C_4-PEPC compared to wild type plants during the light (Furumoto et al. 2007). In an *in vitro* assay, PEPC extracted from RNAi lines displayed a lower malate tolerance compared to wild type *F. bidentis* PEPC extracted in the light (Furumoto et al. 2007). Interestingly, the level of phosphorylation had no impact on the CO_2 assimilation rate in *F. bidentis* (Furumoto et al. 2007). Furthermore, the leaf malate content was found not to be different between *F. bidentis* wild type plants and *F. bidentis* RNAi line. Therefore, given that C_4 PEPC is functional without post-translational phosphorylation, it might be possible to express functional plant C_4-PEPC in a prokaryotic host to screen large numbers of PEPC protein variants.

Similar to PEPC, PPDK is also regulated in a light dependent manner. However, in contrast to PEPC, PPDK is dephosphorylated (active) in the light and phosphorylated (inactive) in the dark (Chastain et al. 2002). The regulatory phosphorylation of C_4-PPDK is carried out by the dual function kinase/phosphatase protein called PPDK regulatory protein (Chastain et al. 2002) (Figure 4.1). Therefore, to generate a catalytically active PPDK in a prokaryotic system the protein does not require any post-translational modifications.

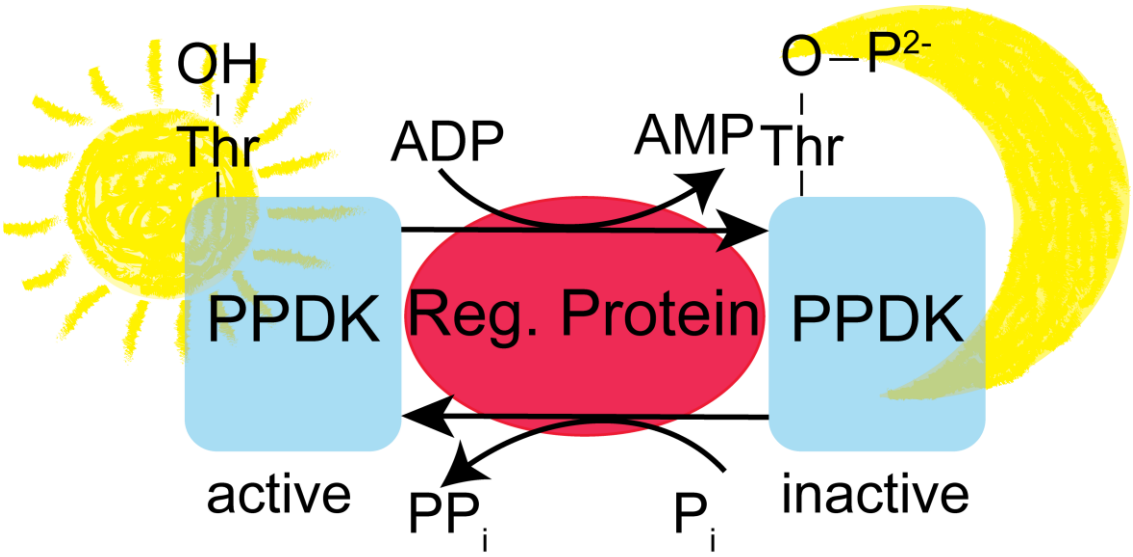


Figure 4.1: A simplified model of the regulation of C_4 PPDK by the dual functional regulatory protein. C_4 PPDK is located in the chloroplast of the mesophyll cells. During the dark the concentration of ADP increases and the direction of reaction is favoured towards phosphorylation of PPDK. During light, the concentration of ADP decreases and the dephosphorylation reaction is favoured. Figure adjusted from Chastain et al. 2002 (Chastain et al. 2002)

To test a large number of different protein variants and identify the variants with the most beneficial characteristics, a method to discriminate between the various protein variants is necessary. There are too many possible variants to produce individually and screen using *in vitro* enzyme assays. Thus one possibility to discriminate between various protein variants is to make the growth rate of the prokaryotic host organism dependent on the function of the protein as it would function in the C_4 cycle. If the growth rate is dependent on the function of the protein, the protein variants with better characteristics will display a higher growth rate compared to the protein variants with worse characteristics. As a consequence, in a bacterial population where all variants are present, the cells expressing proteins variants with better characteristics will

increase in relative abundance over time because these variants give the cells a growth advantage.

Another important consideration when designing such a growth assay is the direction of the reaction of interest. For example, PPDK catalyses the reversible phosphorylation of pyruvate to PEP (Pocalyko et al. 1990):

$$pyruvate + ATP + P_i \overset{PPDK}{\longleftrightarrow} PEP + AMP + PP_i$$

The reaction that is relevant for C_4 photosynthesis is the phosphorylation of pyruvate to PEP. Thus, to identify the best performing protein variant in a prokaryotic test system the direction of reaction has to be the same as the one in the C_4 cycle.

4.1.1.2 Engineering the flux through the central metabolism of *E. coli* to test the function of enzymes involved in C_4 photosynthesis

Glycolysis and the tricarboxylic acid (TCA) cycle form the core of central metabolism. During glycolysis, glucose is oxidised in a multi-step process to pyruvate and yields two ATP and two NADH. The TCA cycle catalyses the oxidation of acetyl-CoA to CO_2 and yields ATP, NADPH and also provides precursor metabolites for amino acids (reviewed in Sauer and Eikmanns 2005). Glycolysis and the TCA cycle are linked through the so-called PEP-pyruvate-oxaloacetate node. This is a highly relevant hub for the carbon flux distribution in the central metabolism (Figure 4.2) (Sauer and Eikmanns 2005).

Under aerobic growth conditions with glucose as sole carbon source, glucose is broken down to PEP and pyruvate. Pyruvate dehydrogenase complex (PDHC) catalyses the

oxidation of pyruvate to acetyl-CoA, where it enters the TCA cycle (Figure 4.2). Additionally, PEP can be carboxylated by the PEP carboxylase (PEPC) to yield one oxaloacetate that enters the TCA cycle. This latter reaction is required to replenish TCA cycle intermediates that were withdrawn from the cycle and used as precursor metabolites elsewhere (Sauer and Eikmanns 2005). PEP is an essential building block that is required in two metabolic processes in the TCA cycle and the Shikimate pathway to create e.g. aromatic amino acids (Figure 4.2). Thus, *Escherichia coli* (*E. coli*) can regenerate PEP from pyruvate through the function of PEP synthase (PPS) (Figure 4.2).

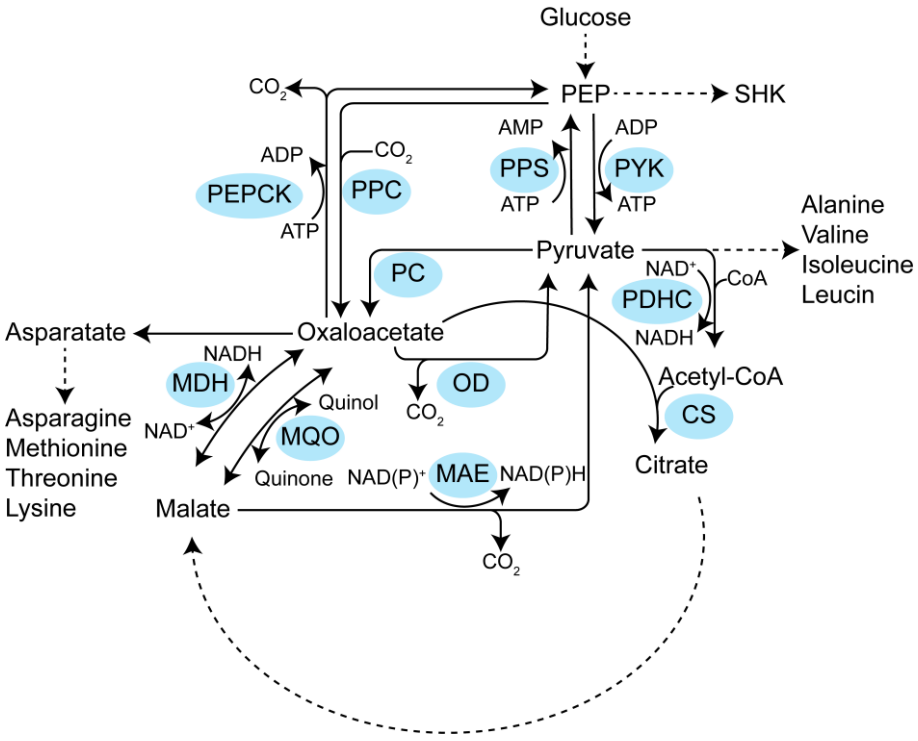


Figure 4.2: The PEP-pyruvate-oxaloacetate node in bacteria grown under aerobic conditions. Abbreviations: malic enzyme (MAE), malate dehydrogenase (MDH), malate quinone oxidoreductase (MQO), oxaloacetate decarboxylase (OD), pyruvate carboxylase (PC), pyruvate dehydrogenase complex (PDHC), PEP carboxykinase (PEPCK), PEP carboxylase (PPC), PEP synthase (PPS), pyruvate kinase (PYK). Note *E. coli* does not have an

oxaloacetate decarboxylase. Figure adapted from Sauer and Eikmanns 2005 (Sauer and Eikmanns 2005)

Some of the reactions that take place at the PEP-pyruvate-oxaloacetate node (e.g. PPS, PPC) are identical to the reactions catalysed by PPDK and PEPC in the C_4 cycle. The bacterial PPC catalyses the carboxylation of PEP to obtain oxaloacetate. This is the identical reaction that takes place in mesophyll cells of C_4 plants catalysed by PEPC. In addition, it is known that *E. coli* mutants in *ppc* are strictly auxotroph for TCA cycle intermediates (Kornberg 1966; Sauer and Eikmanns 2005). In other words, *E. coli ppc* mutants fed with glucose as sole carbon source are unable to grow because they are unable to replenish oxaloacetate in the TCA cycle. Glucose as sole carbon source requires the function of PPC and therefore the growth on glucose can potentially be limited by the abundance and activity of the PPC enzyme.

The bacterial PPS catalyses the reaction from pyruvate to PEP (Figure 4.2). This is the identical reaction of PPDK in the chloroplast of the mesophyll cells in C_4 plants. Feeding *E. coli* with pyruvate as sole carbon source would require the function of PPS to generate PEP, otherwise the cell is unable to grow. In *E. coli* there is not an alternative route to generate PEP because *E. coli* does not contain a pyruvate carboxylase that is present in other prokaryotes (Figure 4.2) (Sauer and Eikmanns 2005). Therefore, *E. coli* relies on the function of PPS to generate PEP when fed with pyruvate as a sole carbon source and thus the growth can potentially be controlled by the abundance and activity of PPS.

E. coli Δ*ppc* or *E. coli* Δ*pps* are unable to grow in media where the sole carbon source is glucose or pyruvate, respectively. These no-growth phenotypes could potentially be restored by expressing either the C_4-PEPC in the *E. coli* Δ*ppc* or the C_4-PPDK in *E. coli* Δ*pps* mutant. The reaction of either of these enzymes is catalysing the equivalent

reaction in the C_4 cycle. The abundance of PPDK and PEPC is potentially growth limiting in the growth-media with either glucose or pyruvate as sole carbon source. If the growth is limited by the abundance of say C_4-PEPC, then an *E. coli* cell that receives a C_4-PEPC variant with an improved characteristic will grow faster than the wild type C_4-PEPC. Thus, the growth rate would serve as a proxy for the enzyme function and cells growing quicker are likely to have a superior enzyme variant.

This approach can be expanded into a competition assay (Van den Bergh et al. 2018). Here, variation is introduced at the previously identified sites (see Chapter 3) through targeted mutagenesis of the coding sequence. Ideally, the targeted mutagenesis approach introduces all possible sequence variants and subsequently gives rise to a large variant library. The variant library is transformed into *E. coli* knockout mutants that are unable to grow in the absence of the gene encoded in the variant library (Figure 4.3 A). The protein sequence encoded by the plasmid thus restores growth of the *E. coli* knockout mutants (Figure 4.3 B). The *E. coli* population can then be cultivated for multiple subsequent generations via sub-culturing (Figure 4.3 C). If the growth rate is limited by the function of the protein encoded on the plasmid, the abundance of cells harbouring and expressing a superior protein variant will become more abundant over time. The change in abundance of each variant could, for example, be followed through sequencing of the plasmid population over successive sub-cultures of the population (Figure 4.3 D). The change in abundance of gene encoding an enzyme variant over time would reveal the variants contributing to the fastest growth rate and thus having the most beneficial characteristics (Figure 4.3 E). Screening a large number of protein variants using this bacterial competition assay system could enable the identification of the best variant to be tested *in planta* (Figure 4.3 F).

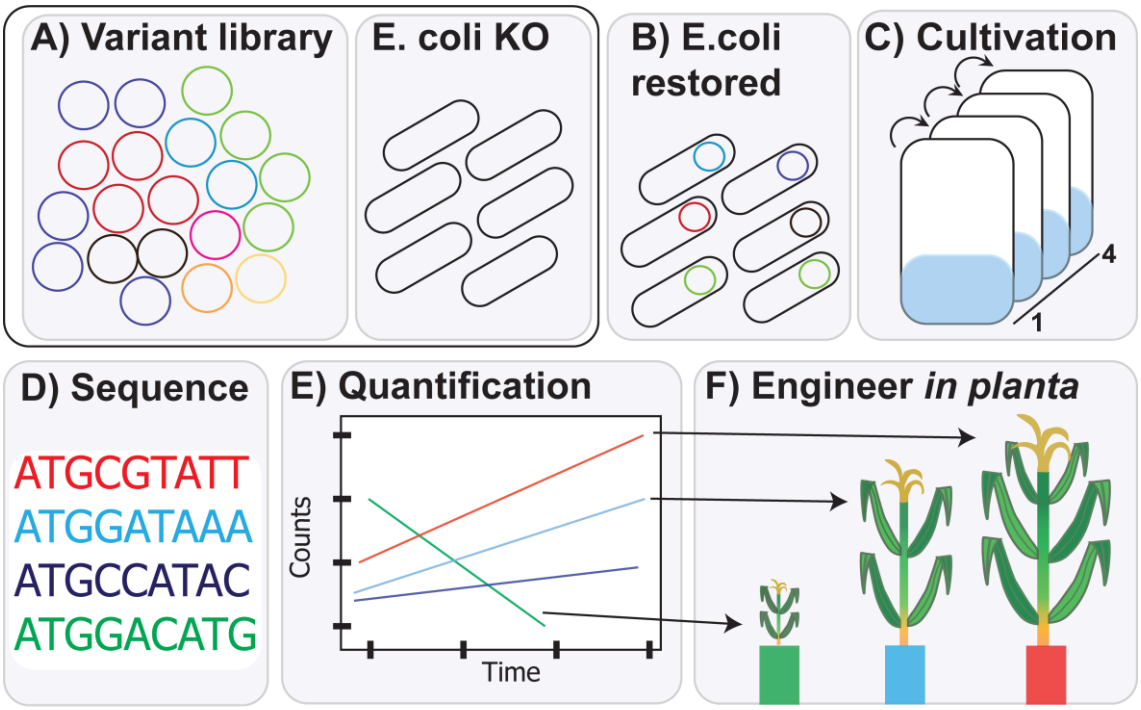


Figure 4.3: Competition assay to identify the protein variant with beneficial functional characteristics. **A)** Generation of a variant library using site-directed mutagenesis. Different coloured plasmids indicate different sequence variants. This plasmid library is transformed into a population of bacteria (e.g. *E. coli*) that lack the gene under investigation but require it for growth in a synthetic media with a single carbon source **B)** Due to the presence of the protein of interest (e.g. PEPC) the bacterial cells harbouring the plasmid are able to growth in a synthetic media with a single carbon source (e.g. glucose) **C)** Multiple subsequent batch cultivation of the bacterial population allows the enrichment of the cells that harbour an above average protein variant **D)** Sequencing of a sample of the bacterial population allows the identification and quantification of variants present at a given time point **E)** Plotting the quantified variants over time allows the identification of the most promising variant **F)** These identified variants can subsequently be engineered into a C_4 plant where the best variant (red) has the largest impact on the photosynthetic efficiency and the worst variants (green) results in a reduction of the photosynthetic efficiency compared to the wild type variant (blue).

4.1.2 Chapter aims and outline

The aim of the project presented in Chapter 4 was to experimentally validate whether the identified sites in the previous chapter have an impact on the protein function through the use of bacterial growth competition assays. The experimental evaluation was limited to the two potentially rate-limiting enzymes of the C_4 cycle PPDK and PEPC. In order to be able to test a large number of sequence variants, the first aim was to set-up a high-throughput test system in bacteria where the growth rate is dependent on the enzymatic function. The second aim was to establish an approach to exhaustively change codons at the previously identified sites (see Chapter 3) to all possible codons and thus create a variant library with a large number of sequence variants. Subsequently, the bacterial growth competition assay would be used to identify the most promising sequence variants.

4.2 Results and discussion

4.2.1 Generating *E. coli* BL21 (DE3) Δ*ppc* using recombineering

To generate an *E. coli* Δ*ppc*, the endogenous *ppc* was replaced with an ampicillin resistance gene cassette (*ampR*) using recombineering (Section 4.4.2). The *ampR* gene confers resistance to the antibiotic ampicillin and allows the cells to grow under antibiotic selection. To target and replace *ppc*, the recombineering method relies on homologous sites flanking the *ampR* gene that correspond to the *ppc* gene. Thus, a primer pair was designed to target that homologous region that is flanking the introduced *ampR* gene and corresponded to the first and last 50 bp of the endogenous *ppc* gene. Using this primer pair to amplify the endogenous *ppc* gene the resulting fragment size was 2627 bp, whereas in the *ppc* knockout mutant containing the ampicillin gene the fragment size was 1599 bp. Thus, a colony polymerase chain

reaction (PCR) enabled a screen for *ppc* mutants using the obtained fragment size to discriminate between wild type cells and *ppc* mutant cells (Figure 4.4 A).

As anticipated, the *E. coli* BL21 (DE3) *ppc* knockout mutant, hereafter called *E. coli Δppc*, was able to grow in rich media (LB) but not in M9 with glucose as sole carbon source (Figure 4.4 B). Growth was restored through the expression of the *Setaria viridis* PEPC (Sevir.4G143500) in *E. coli* Δ*ppc*. Thus, the PEPC gene from the C_4 plant *Setaria viridis* can provide the function of endogenous *E. coli* PPC. To rule out that the transformation had an impact on the growth phenotype, *E. coli Δppc* were transformed with an empty pUC19 and the no-growth phenotype remained present (Figure 4.4 C).

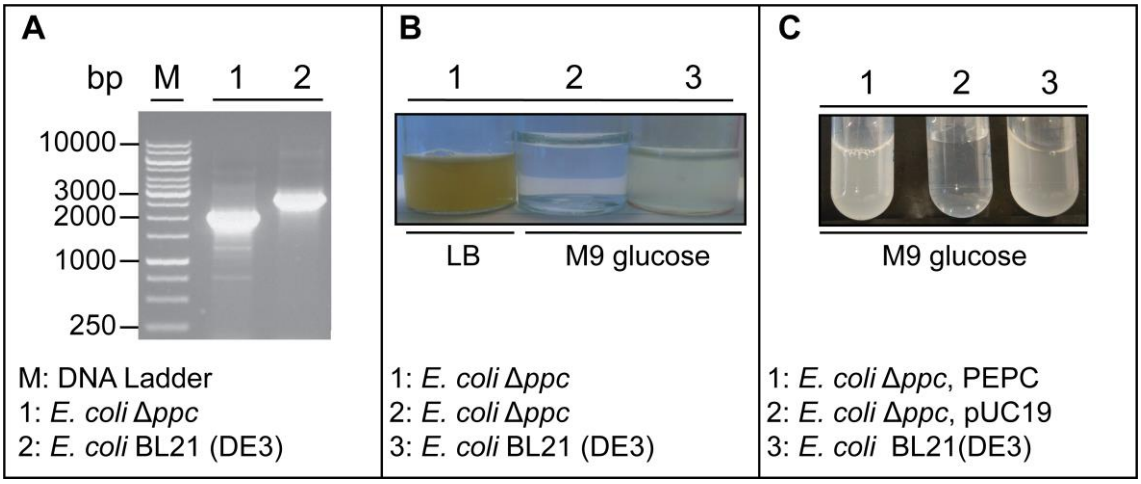


Figure 4.4: Generation of an *E. coli* BL21(DE3) *Δppc* mutant **A)** Agarose gel electrophoresis showing the different PCR fragment sizes. The PCR product encompassing the region containing endogenous *E. coli ppc* gene has a fragment size of 1599 bp when *ppc* is replaced with *ampR* (1) and 2627 bp when the native gene is not deleted (2). **B)** *E. coli* Δ*ppc* was able to grow on rich media such as LB (1) but the cells were unable to grow in media containing glucose as sole carbon source (2). *E. coli* BL21 (DE3) showed normal growth without media with glucose as sole carbon source (3) **C)** *E. coli Δppc* expressing *S. viridis* PEPC

(Sevir.4G143500) was able to grow in media with glucose as sole carbon source (1). *E. coli Δppc* harbouring an empty pUC19 was unable to grow in media with glucose as sole carbon source (2). *E. coli* BL21 (DE3) were able to grow in media with glucose as sole carbon source (3).

To have an additional verification whether the PPC knockout was successful, the whole genome of the generated *E. coli Δppc* was sequenced. The mean coverage was 107 reads per base and the total number of reads was 1'183'818. The paired-end reads were mapped once to the wild type *E. coli* BL21 (DE3) genome and a second time to an *in silico* generated version of the *E. coli* BL21 (DE3) genome where *ppc* gene was replaced with the *ampR* gene (see Section 4.4.4). Visualizing the mapped reads to the wild type genome revealed that there were no reads that could map where the *ppc* gene is located on the chromosome. However, when the reads were mapped onto the *in silico* generated genome, reads mapped to the inserted *ampR* gene (Figure 4.5). Thus, the *ppc* gene has been fully removed from the genome and there is no evidence for any sequence corresponding to the native *E. coli ppc* gene in the *E. coli Δppc* knockout line.

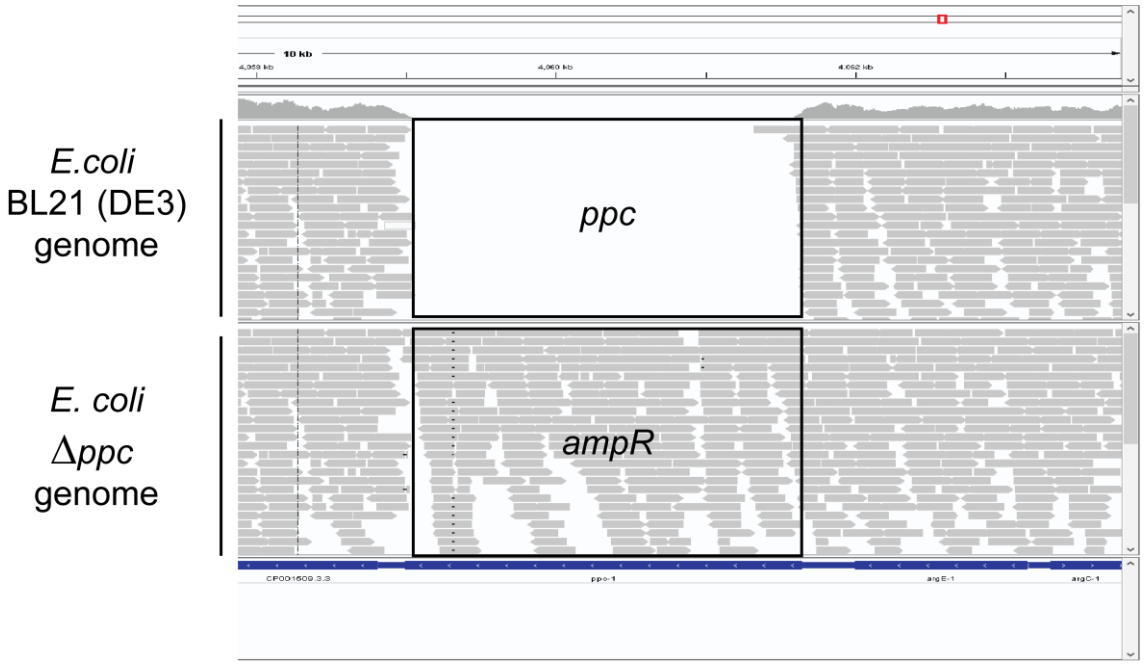


Figure 4.5: Whole genome sequencing of the *E. coli* Δ*ppc* mutant. Top panel shows the paired-end reads generated from the *Δppc* mutant mapped to the *E. coli* BL21 (DE3) reference genome. Mapped reads are represented as grey bars. The region surrounding the endogenous *ppc* gene is shown (framed with solid line). No reads corresponding to DNA from this region of the genome were present in the Δ*ppc* mutant line. The lower panel shows the paired-end reads mapped to the *in silico* generated genome where the *ppc* gene was replaced with *ampR* gene. The reads generated from sequencing the *Δppc* mutant mapped to this region as expected indicating that the *ampR* gene was present where the *ppc* gene was located.

4.2.2 Evaluating whether the growth rate is dependent on the function of PEPC

To evaluate whether the growth rate could be engineered so that it was dependent on the function of exogenously expressed PEPC (Sevir.4G143500), various constructs were created with different ribosomal bindings sites (RBS) that control the expression of PEPC (Sevir.4G143500). This strategy of modifying the RBS was taken to minimise the number of variants to be tested. The expression constructs were identical regarding the backbone plasmid and the strength of the promoter and thus they only vary in the

strength of the ribosomal binding site (RBS) (see Section 4.4.6 and Appendix A.3). Both constructs carry a promoter characterized with a strength of 4095.5 AU (Arbitrary units) (Anderson). Construct 1 carries an RBS sequence with a characterized strength of 925 AU and construct 2 has an RBS sequence with a characterized strength of 135 AU. If the growth rate is dependent on the function of PEPC, the cells harbouring the PEPC construct 1 should grow faster than the cells harbouring construct 2.

The assessment of the growth was carried out in Erlenmeyer flasks in triplicates for each of the cell lines. The estimated growth rate for the PEPC construct 1 was 1.003 h^{-1} ± 0.069 (SD, N=3) and the estimated generation was time 41.6 min ± 2.9 (SD, N=3). The estimated growth rate for the PEPC construct 2 was 0.963 h^{-1} ± 0.085 (SD, N=3) and the estimated generation time was 43.3 min ± 4.1 (SD, N=3) (Figure 4.6). The *E. coli* Δ*ppc* mutants expressing PEPC with the weaker RBS sequence have a slightly lower growth rate and longer doubling time than the *E. coli ppc* mutants expressing PEPC with a stronger RBS sequence. However, this difference was not significant (p-value > 0.05) as there is a large variation associated with the estimated growth characteristics.

To fully verify whether the growth rate is dependent on the abundance of PEPC (Sevir.4G143500) additional PEPC constructs need to be evaluated with weaker promoter and RBS parts. For a potential competition assay, having weaker expression would benefit to discriminate between the created PEPC variants with beneficial functional characteristics and variants with negative functional characteristics. However, I did not have time to complete these tests as I prioritised other experiments. This step would be important to optimise should the work be continued.

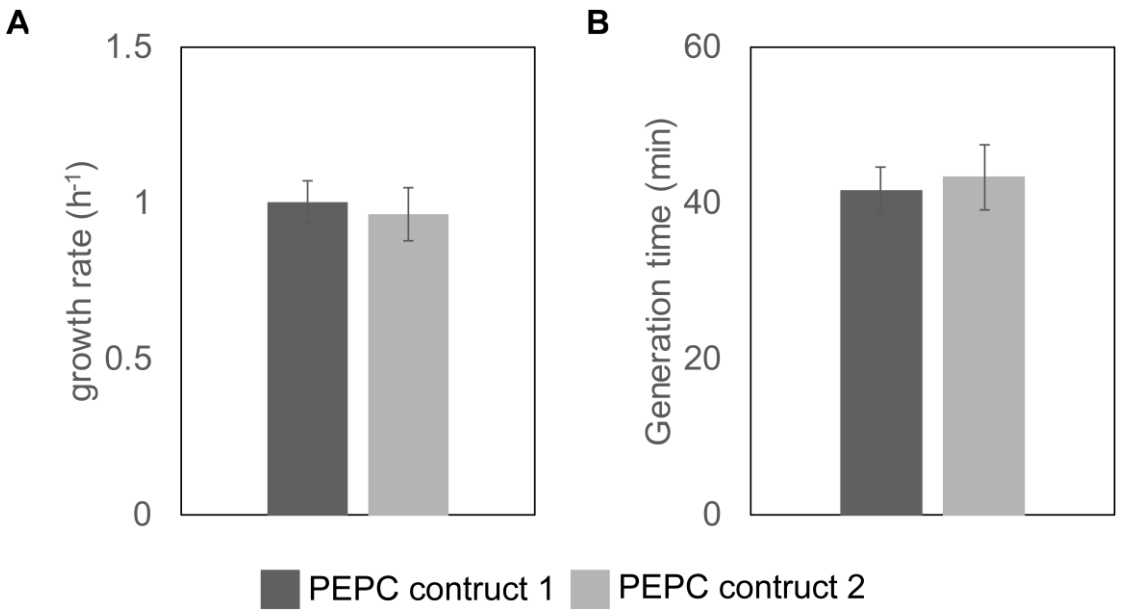


Figure 4.6: Comparison of the growth characteristics in function of the expression of Svir PEPC. **A)** Growth rate measurement of the *E. coli Δppc* expressing Sevir.4G143500. *E*. *coli Δppc* harbouring PEPC construct 1 had an estimated growth rate of 1.003 h^{-1} ± 0.069 (SD, N=3) and *E*. *coli Δppc* harbouring PEPC construct 2 had an estimated growth rate of 0.963 h^{-1} ± 0.085 **B)** The generation time for the *E*. *coli Δppc* harbouring PEPC construct 1 and PEPC construct 2 were estimated to be 41.6 min ± 2.9 (SD, N=3) and 43.3 min ± 4.1 (SD, N=3), respectively. The generation time was not significantly different (t-test, p-value > 0.05).

4.2.3 Evaluating whether the growth rate is dependent on the function of PPDK in *E. coli Δ*pps mutants grown in media with pyruvate as sole carbon source

The generation of *E. coli* knockout mutants using recombineering can sometimes be difficult to achieve. Multiple attempts to generate a Δ*pps* line failed and thus *E. coli* Δ*pps* mutants (JW1692) were obtained from the Keio collection (Baba et al. 2006). The Keio collection provides single-gene deletions of all nonessential genes in the *E. coli* strain BW25113 (Baba et al. 2006). The approach by Baba et al. 2006 relied on a homologous recombination method to delete the target gene. The gene deletions were generated by replacing the endogenous gene with a kanamycin expression cassette that was flanked upstream and downstream with 50 nt homologous ends for targeting

the gene deletion. These homologous ends were targeted with a specific primer pair to test whether the received strain was indeed an *E. coli Δpps* mutant. The expected PCR fragment sizes with the *pps* specific primers were 2444 bp for the *E. coli* wild type BW25113 and 1367bp for the *E. coli Δpps* mutant 1367 bp (see Section 4.4.3). Thus, using a conventional PCR it was possible to confirm the genotype of the different strains and classify them based on the obtained fragment size (Figure 4.7 A). In addition, a further test was performed to confirm that activity of PPS was absent. Here, M9 media with pyruvate as a sole carbon source was inoculated separately with wild type and Δ*pps* mutant cells. The overnight culture revealed that the Δ*pps* mutants were unable to utilize pyruvate as expected and thus were not able to grow (Figure 4.7 B) (see Section 4.1.1.2).

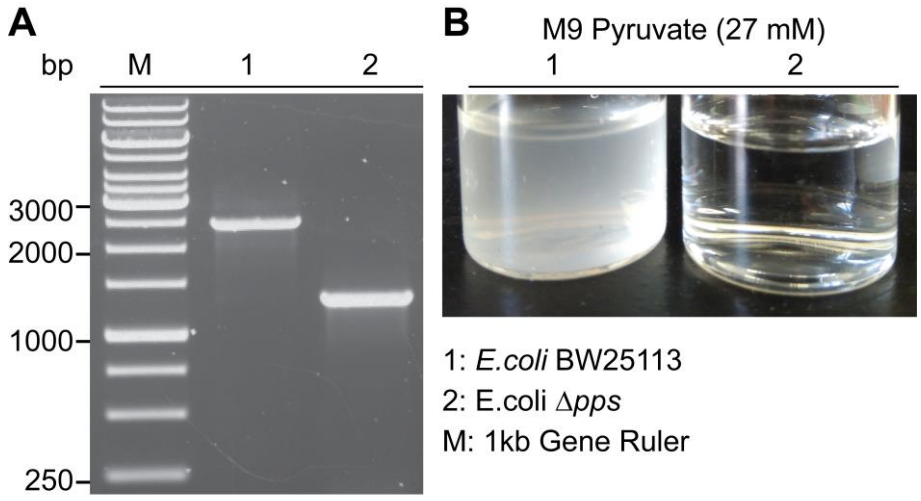


Figure 4.7: Confirmation of the *pps* gene knockout in the Keio strain JW1692: **A)** Colony PCR of *E coli* BW25113 and *E. coli* Δ*pps* JW1692 targeting with *pps* specific primers. The differences in fragment size 2444 bp and 1367 bp allows to discriminate between wild type colonies and Δ*pps* mutant colonies **B)** *E. coli* BW25113 and *E. coli* Δ*pps* JW1692 were incubated overnight in M9 with pyruvate (27 mM) as sole carbon source. The wild type cells were able to grow whereas the Δ*pps* mutants were not.

To further verify whether the received Δ*pps* mutant cells were indeed *pps* knockout mutants, the whole genome was sequenced. The mean coverage obtained was 54 reads per base and the total number of reads was 620'319. The paired-end reads generated from the Δ*pps* strain were mapped against the wild type *E. coli* BW25113 genome and a second time to an *in silico* generated version of the BW25113 genome where the *pps* gene was replaced with the kanamycin expression cassette (see Section 4.4.4). The visualization of the mapped reads revealed that there were no reads that could map to the *pps* gene. However, reads mapped to the kanamycin resistance cassette as expected, confirming that the *pps* gene was fully lost from the genome and was replaced by the kanamycin expression cassette (Figure 4.8).

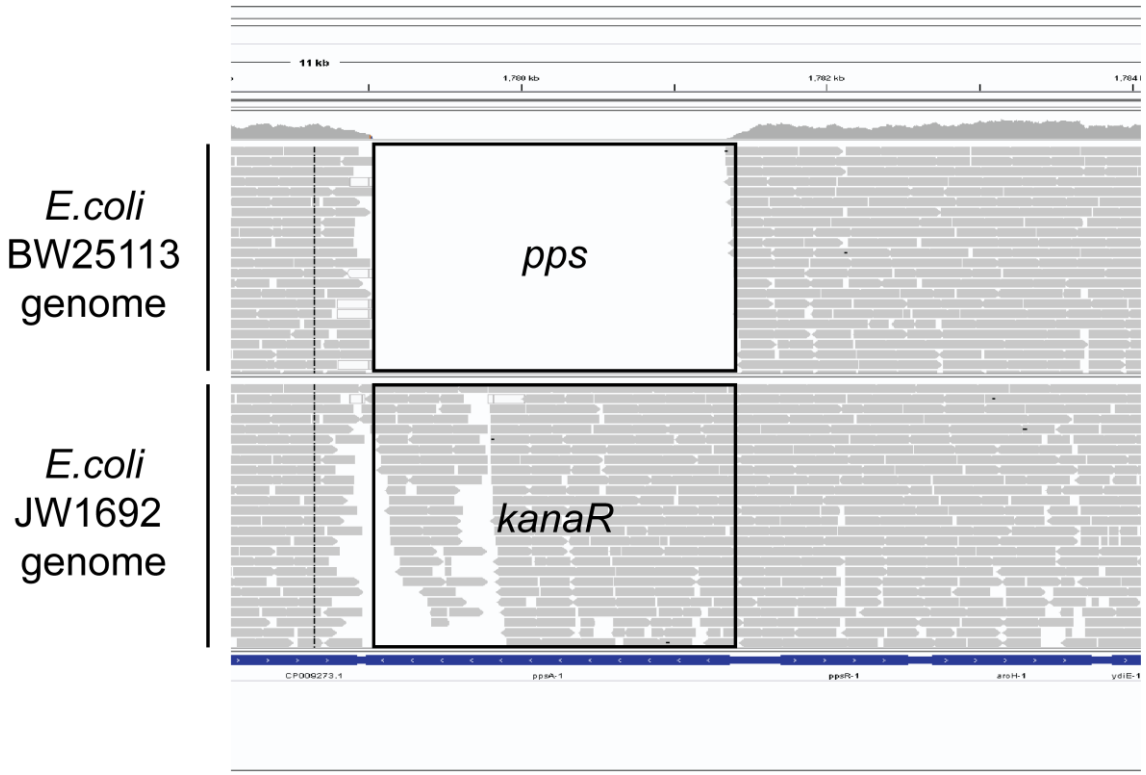


Figure 4.8: Whole genome sequencing of *E. coli* JW1692 (Δ*pps*). The top panel shows the reads from the Δ*pps* line mapped to the wildtype *E. coli* BW25113 reference genome. No reads were mapped to the region where *pps* should be located on the chromosome indicating that *pps* was not present. Mapped reads are indicated as grey bars. Lower panel shows the reads mapped to the *in silico* generated *E. coli* BW25113 genome were *pps* was replaced with the kanamycin expression cassette.

To assess whether the growth of *E. coli Δpps* can be restored in M9 with pyruvate as sole carbon source by the expression of PPDK from a C_4 plant, two plasmids were created that express the *S. viridis* PPDK (Sevir.3G253900) (see Appendix A.3). The evaluated constructs varied only in the RBS sequence. The backbone, promoter, PPDK sequence, and terminator remained identical. The expression of the PPDK was controlled by a promoter characterized with a strength of 4095.5 AU (Anderson) (see Section 4.4.6 and Appendix A.3). The RBS sequence of the construct PPDK 1 was characterized with a strength of 6437 AU and the RBS sequence of the construct

PPDK 2 was characterized with a strength of 135 AU. Growth analysis of the *E. coli* Δ*pps* mutant expressing PPDK at various strengths was carried out in a 96 well plate with eight replicates per construct. The growth curve revealed that the PPDK construct 1 enabled the *E. coli Δpps* to have a growth rate of 0.209 h^{-1} ± 0.011 (SD, N=8) and a generation time of 3.3 h ± 0.297 (SD, N=8) whereas the PPDK construct 2 facilitated a growth rate of 0.123 h^{-1} ± 0.008 (SD, N=8) and a generation time of 5.6 h ± 0.277 (SD, N=8) (Figure 4.9 A and B). The obtained generation time between the two PPDK construct was significantly different (t-test, p-value < 0.0001).

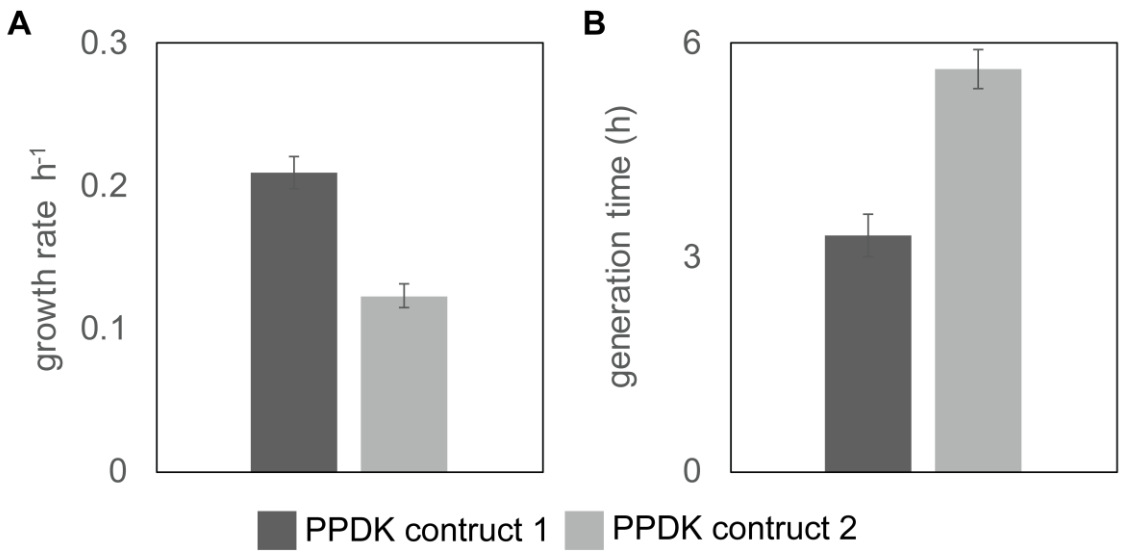


Figure 4.9: Comparison of the growth characteristics of *E. coli Δpps* expressing *S. viridis* PPDK (Sevir.3G253900) on M9 media containing pyruvate as a sole carbon source. **A)** Estimated growth rate: PPDK construct 1 had a growth rate of 0.209 h^{-1} ± 0.011 (SD, N=8) and PPDK construct 2 had a growth rate of 0.123 h^{-1} ± 0.008 (SD, N=8) **B)** Estimated generation time: PPDK construct 1 had a generation time of 3.3 h ± 0.297 (SD, N=8) and PPDK construct 2 had a generation time of 5.6 h ± 0.277 (SD, N=8). Generation time between the two constructs was significantly different (t-test, p-value < 0.0001).

The *E. coli* Δ*pps* expressing the PPDK with weaker RBS had a longer generation time when compared to the cells with stronger RBS when grown in M9 with pyruvate as sole carbon source. From this it follows that the growth rate in the *E. coli Δpps* was dependent on the abundance of the exogenous PPDK enzyme and thus, dependent on the function of PPDK.

For potential competition assays the PPDK construct 2 would be an ideal plasmid to generate large number of variants and eventually identify a PPDK protein variant with improved characteristics.

4.2.4 Generation of a variant library using site-directed mutagenesis

To establish an approach to generate a large variant library using site-directed mutagenesis the PEPC construct 2 was used to carry out the experimental evaluation. The identified sites in PEPC that revealed a different amino acid distribution between the C_3 and C_4 species in the dataset were identified in the *S. viridis* (Sevir.4G143500) gene. For each of these sites degenerated primers were designed to introduce site-specific degenerated codons (Appendix A.7) to create large numbers of sequence variants (Method Section 4.4.10).

In brief, targeted mutagenesis was carried out using the QuikChange Multi Site-Directed Mutagenesis Kit (see Section 4.4.10). The template plasmid harboured the *S. viridis* PEPC sequence (Sevir.4G143500) and in a first attempt, five sites (228, 362,495, 695, 774) were targeted with degenerated primers. The obtained plasmid library was sequenced to assess the degree of degenerate codons introduced into the PEPC coding sequence at the targeted sites. The Sanger sequence chromatogram revealed that the targeted sites successfully introduced novel codons (Figure 4.10. A). Out of the five sites targeted, each site revealed degenerate codons, indicating that

the targeted mutagenesis approach was successful (Figure 4.10. B). Inspection of the sequencing chromatograms showed that the libraries were sometimes biased towards some nucleotides at different codon positions (Figure 4.10 B).

To provide a positive control for the targeted mutagenesis approach a plasmid harbouring a non-functional GFP was used. The non-functioning GFP has a tyrosine to histidine mutation at position 66 and is therefore not fluorescent. Targeting that mutated codon and changing the histidine to a tyrosine restores the GFP fluorescence. About 30% of the colonies that grew on the plate appeared to have fully restored the GFP function. The equal amount of template plasmid was used to transform *E. coli* cells without undergoing site directed mutagenesis and appear as white colonies (Figure 4.10 C). The positive control reveals that although restoring the fluorescence of the GFP was achieved, the number of functional plasmids obtained at the end of the site directed mutagenesis was low compared to the amount that was used as input (Figure 4.10 C).

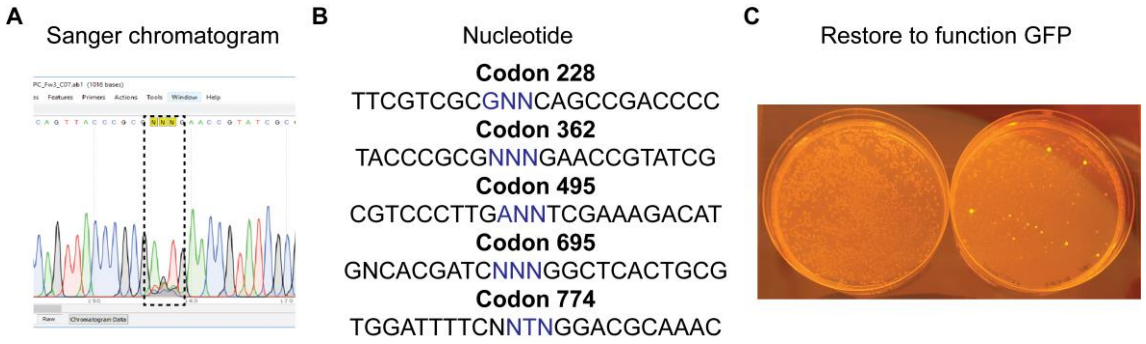


Figure 4.10: Verification of the site directed mutagenesis. **A)** An example Sanger chromatogram for one selected site revealing that multiple nucleotides are detected at each position in the codon that has been selected for site saturating mutagenesis. The nucleotides upstream and downstream of the targeted site remain unchanged **B)** The base calling output of the Sanger sequencing run indicating that the targeted sites (1-5) have undetermined nucleotides. **C)** Positive control to evaluate the success of the site-directed mutagenesis reaction. Here, the site-directed mutagenesis restores the fluorescence of GFP. Left plate: equal amount of plasmid that did not undergo site directed mutagenesis was transformed into *E. coli* resulting in white colonies. Right plate: plasmids that underwent the site directed mutagenesis with restored GFP function appeared green.

After the verification that the introduction of degenerated codons worked, the next step was to evaluate the number of variants that were obtained after the site directed mutagenesis approach. To obtain a large number of transformed cells carrying different PEPC variants, electrocompetent *E. coli* with a transformation efficiency of 1.5 -2 × 10^{10} colonies per 1 µg of plasmid were generated (see Section 4.4.1).

The default procedure for the site-directed mutagenesis requires 100 ng of plasmid as a template input. To assess the impact of the amount of template on the number of viable colonies obtained, site directed mutagenesis was carried out with 100 ng, 10 ng

and 1 ng of plasmid template. The created variant libraries were all separately ethanol precipitated and eluted in 4 µl TE buffer prior to the electroporation. All of the eluted mutagenesis reaction was used to transform into the electrocompetent cells that were subsequently plated on LB plates containing appropriate antibiotic selection. The number of colony forming units (cfu) were counted as a proxy for the number of viable transformed cells. The number of viable cells for the 100 ng template was 256 cfu ± 33 (SD, N = 6), for 10 ng 254 cfu ± 24 (SD, N = 6) and for 1 ng 164 cfu ± 65 (SD, N = 6). To increase the number of variants, and subsequently the number of cfus, the manufacturer recommendations were followed. The recommendations were to increase the reaction volume from 25 µl to 100 µl and to use 300 ng of plasmid as a template. Increasing the reaction volume and the template concentration resulted in 267 colonies ± 45 (SD, N = 3).

Given that all possible sequence variants for five sites encompasses 3.2×10^6 variants, the numbers of obtained colonies was far from sampling even 1% of all possible variants of the five site library. One potential limitation could be caused by the plasmid structure. At the end of the site-directed mutagenesis procedure the plasmid was present as a single stranded plasmid rather than double stranded. The transformation efficiency for single stranded DNA is much lower than double stranded DNA. Thus, three different approaches were taken to attempt to convert the single stranded library to a double stranded library. The first approach consisted of adding a primer for the second strand synthesis together with the QuikChange Multi Site-directed enzyme mix. The second approach consisted of adding primer for second strand synthesis without any additional enzyme. The idea was to test whether an annealed primer at the single stranded plasmid would be sufficient to increase the number of cfus. The third approach included adding a primer for second strand

synthesis and the DNA Polymerase I Large (Klenow) Fragment (see Section 4.4.10). Subsequently, the reactions were ethanol precipitated and eluted in 4 µl of TE buffer and the entire eluent was transformed into electrocompetent *E. coli*. The reaction with primer only resulted in fewer than ten colonies whereas the two other approaches resulted in about 200 colonies. This experiment was carried out once but it seemed that trying to synthesize a second strand was not a successful avenue to proceed with to increase the size of the variant library. This should be repeated and other approaches to boost colony number tested.

4.3 Chapter conclusion

In this chapter I presented the first steps towards the development of a novel method to evaluate the relative performance characteristics of enzymes of the C_4 cycle in a bacterial test system. The method was created to test the changes in protein function in a higher throughput than one would achieve by altering the gene of interest *in planta,* or by testing the enzyme variants using conventional enzyme assays. The method was tested using PEPC and PPDK that are considered to be rate limiting in the C_4 cycle (Section 3.1.2) and for which engineering targets have been identified in previous chapters of this thesis (Section 3.2.3.1 and 3.2.3.3). The method relies on the metabolic flux through central metabolism in *E. coli*.

The reaction that is catalysed by PPC in a bacterial cell is identical to the reaction that is catalysed by PEPC in the mesophyll cells of a C_4 plant, and it proceeds in one direction only. Thus, the bacterial competition assays in cells fed on glucose as a sole carbon source test the function of the enzyme operating as it would in the mesophyll cell of a C_4 plant. In contrast, the *pps* gene in *E. coli* and the PPDK gene in plants catalyse reactions that can occur in either direction. However, in the mesophyll cell of a C_4 plant, the PPDK enzyme catalyses the conversion of pyruvate to PEP and not

vice versa. To ensure that the reaction proceeds in the same direction in the bacterial competition assays as it does in the C_4 cycle requires changing the carbon source from glucose to pyruvate. Thus, although, plant PPDK could function in both directions in *E. coli*, the direction of conversion of pyruvate to PEP is enforced by feeding cells on pyruvate as a sole carbon source. Thus, growth is limited by the ability of the cells to convert pyruvate to PEP so that energy can be generated through the TCA cycle and that amino acids and fatty acids can be generated from TCA cycle intermediates. This is an important point as incorrect experimental set up, such as feeding cells on glucose, could lead to testing the function of the PPDK enzyme operating in the wrong direction.

The experimental set up described in this chapter was designed to be able to select for several desirable traits of the enzymes in question without having to measure them directly. Such traits include, an increased in turnover rate (k_{cat}), and increase in substrate specificity, a reduction in competitive inhibition, enhanced half-life. However, it is also likely that false positive variants could also be selected. Here, a false positive variant is a variant which confers enhanced growth in *E. coli*, but would not confer enhanced growth in plants. Such false positive variants could include variants with enhanced solubility in *E. coli* leading to enhanced protein accumulation and thus enhanced flux. Similarly, there could be variants with enhanced translational efficiency in *E. coli* through more optimal codon usage leading to enhanced protein levels and enhanced flux. Thus, it would be beneficial to perform a second screening step on variants after the competition assay before attempting to engineer these variants into plants. Such a second screening step could involve conducing enzyme assays on protein produced in a yeast or cell-free (plant or animal) protein translation system to see if the enzymes produced have altered kinetic parameters when compared to the original/wild type sequence.

To create large number of sequence variants a site-directed mutagenesis approach was used to introduce multiple codons at the previously identified engineering targets in PEPC. Sequencing of the generated variant library revealed that each of the targeted sites had a mix of various codons present. Revealing that the site-directed mutagenesis strategy worked successfully. The introduction of the variant library into an *E. coli* population revealed that only a limited number of sequence variants were recovered. The numbers obtained even after multiple rounds of optimization were far below what was required to do a meaningful competition assay in order to identify a protein variant with improved characteristics. Thus, future work is required in this area to improve the number of variants that can be screened in a single competition assay.

4.3.1 Future work

Although the results were promising, this work was not seen through to completion due to time constraints. Specifically, this project was put on hold as the project described in Chapter 5 was proving more successful and was thus allocated a higher priority. Unfortunately, there was not enough time to complete both projects.

The major bottleneck encountered in Chapter 4 was the generation of large numbers of sequence variants required to construct the variant library. Multiple methods were tested of which the commercial QuikChange kit was the most successful. However, since completing this work other promising methods have been developed. For example, the Darwin assembly method (Cozens and Pinheiro 2018) is reported to be able to deliver libraries with more than 1×10^8 variants and targeting more than 10 sites that can be far apart on the sequence of interest. There was not time to test this method, but it could solve the problem of generating sufficiently large libraries. If it proves impossible to generate libraries of sufficient depth, then an alternative solution

would be to just use random mutagenesis. However, the combinatorial complexity of such libraries is far larger than the number of particles in the universe and it does not take advantage of knowledge gained from evolutionary analyses. Thus, I believe it is still worthwhile to pursue the approach outlined in this chapter.

The bacterial test system where the growth rate is limited by the abundance of the protein of interest was achieved for PPDK. However, the assay for PEPC was not fully established at the time of writing up. Plant PEPC was shown to be function in *E. coli* and was able to restore growth in *E. coli Δppc* when the cells are fed with glucose as sole carbon source. However, it remained unclear whether the growth rate in the restored *E. coli Δppc* expressing plant PEPC was limited by the abundance of PEPC or not. To carry out a competition assay with PEPC some additional expression vectors would need to be generated to verify that growth rate limitation due to PEPC abundance was occurring. This could be done by including some additional promoter and RBS sequences with weaker expression than previously carried out.

The established bacterial test system relied on the idea of limiting the growth rate by the abundance of the protein of interest and subsequently identifying variants with more optimal characteristics. For future work the idea of the bacterial test system could be expanded to evaluate other components of the C_4 cycle. For example, transporters of the C_4 cycle such as BASS2. The Total Variation Distance algorithm has identified sites that are different between the C_3 and C_4 species (see Section 3.2.4.3). In the C_4 cycle BASS2 facilitates the transport of pyruvate from the cytoplasm into the chloroplast of bundle sheath cells. To identify a BASS2 variant that has a more favourable characteristics, a growth test in a media with pyruvate as sole carbon source could be carried out and the pyruvate uptake into the cells is performed by BASS2. To launch such an experiment all pyruvate uptake systems need to be

removed. Currently, there are three known pyruvate uptake systems *YfeH,* cstA (Hwang et al. 2018) and btsT (Kristoficova et al. 2018). In addition, there are at least two additional potential uptake systems yhjX (Fried et al. 2013) and YnfM (Göhler et al. 2011) that might also be involved. Successive knockouts of these transporters would need to be made to generate a cell line that was unable to import pyruvate. This makes the design of such an experiment more challenging.

4.3.2 Summary

This chapter illustrates the development of a bacterial systems to comparatively evaluate protein variants for potential improved suitability for function in the C_4 cycle. The reason for developing this new test system emerged because the number of sequence variants previously identified was too large to carry out *in planta* directly, or using conventional enzyme assays. Thus, a process that allows us to narrow down the numbers and screen for the potentially best sequence variants was required. The method developed relies on the fact that certain biochemical reactions that take place in *E. coli* and C_4 plants are identical. *E. coli* fed on synthetic media with either glucose or pyruvate relies on the function of PPC or PPS, respectively. The absence of PPC in cells grown in the media with glucose as sole carbon source or in the absence of PPS in cells grown in the media with pyruvate as sole carbon source results in a no-growth phenotype. Growth can be restored through expressing the plant enzymes PEPC that catalyses the equivalent reaction to PPC, and PPDK that carries out the equivalent reaction of PPS. To test large number of sequence variants, the growth rate has to be limited by the activity of the introduced enzymes. If growth is not limited by enzyme activity then there is no capacity for enhanced enzyme variants to result in enhanced growth. To achieve such a dependency, expression vectors were created with various levels of expression strength. For PPDK, the levels of expression were reduced

sufficiently such that growth of *E. coli* was limited by PPDK abundance. Such dependency was not achieved for PEPC. Here, the low and medium abundance constructs resulted in grow rates that were nearly identical. Thus, a construct that resulted in lower expression is required so that there is capacity to increase growth rate by improving the properties of the enzyme. To create a sequence variant library a site-directed mutagenesis approach was used. The approach was able to introduce degenerate codons at the sites previously identified. Unfortunately, the number of variants obtained was too low to make a meaningful competition assay. Although the project was not taken through to completion, several important steps have been optimised and several roadblocks have been overcome.

4.4 Material and Methods

4.4.1 Electrocompetent cells

E. coli BL21(DE3), *E. coli* BW25113 and *E. coli* JW1692 were made electrocompetent by following the method described by Sambrook and Russell (Volume 1, Chapter 1, Protocol 26) (Sambrook and Russell 2006). The process carried out is described in short below. A single colony was picked from a fresh LB plate grown over night and used to inoculate 50 ml of LB in a 250 ml Erlenmeyer flask. The culture was grown over night at 37°C with 250 rpm shaking in an incubator.

The next day, two aliquots of 500 ml LB in 2000 ml Erlenmeyer flasks were prewarmed to 37°C and subsequently inoculated with 25 ml of the overnight culture. The newly inoculated Erlenmeyer flasks were incubated at 37°C with vigorous shaking at 250 rpm until the culture reached an optical density measured at 600 nm (OD_{600}) of 0.4. Once the culture reached an OD_{600} of 0.4, the Erlenmeyer flasks were immediately transferred to an ice-water bath and incubated for 30 min while occasionally swirling to chill the bacterial suspension.

Subsequently, the cell suspensions were transferred to a pre-cooled centrifugation bottle and the cells were harvested by centrifugation at 1000 g for 15 min at 4°C in a Sorval Evolution RC with the F10 6x500Y rotor. The supernatant was discarded and the cell pellet was resuspended in 250 ml ice-cold water containing 10% glycerol. Next, the cell suspension was centrifuged at 1000 g for 20 min at 4°C. The resuspension and collection of the cells in ice-cold water 10% glycerol was carried out twice.

After the second centrifugation step the OD_{600} of a 1:100 dilution of the cell suspension was measured and diluted to a concentration of ~ 2 to 3×10^8 cells ml^{-1} (an OD_{600} of 1.0 is ~ 2.5×10^8 cells ml^{-1}) with ice-cold GYT (10% (v/v) glycerol, 0.125% (w/v) yeast

extract, 0.25% (w/v) tryptone). From that cell suspension 50 µl were transferred to an ice-cold electroporation cuvette (0.1 cm Invitrogen P41050) and tested whether electrical arcing occurs when an electrical discharge was applied (Bio Rad Gene Pulser was set to 25 µF, 2.5 kV and 200 Ω). If arcing occurred the cells were washed a second time in ice-cold GYT. Otherwise 50 µl of the cell suspension was transferred to small Eppendorf tubes that were kept on ice and subsequently frozen in liquid nitrogen. The 50 µl aliquots were store until used in the -80°C freezer.

The competency of the electrocompetent cells was evaluated in triplicates with the pUC19 plasmid (100 pg μl^{-1}) using electroporation cuvettes (0.1 cm Invitrogen P41050) and applying an electrical discharge (Bio Rad Gene Pulser) set to 25 µF, 2.5 kV and 200 Ω. For each transformation a dilution of 1:1000 was carried out and 20-100 µl from each transformation were spread on LB plates containing 100 µg ml^{-1} ampicillin. The plates were incubated at 37°C overnight and the cfu were counted the next day to assess the transformation efficiency using equation 2.

$$\frac{number\ of\ colonies}{pg\ of\ pUC19\ DNA\ (250\ pg)} \times \frac{10^6 pg}{\mu g} \times \frac{volume\ of\ transformation\ (1000\ \mu l)}{volume\ plated\ (20{-}100\ \mu l)} \qquad \text{(Equation 2)}$$

4.4.2 Genome engineering using recombineering

To create *E. coli* knock-out mutants, a published method for genetic engineering was utilised (Sharan et al. 2009). The method relies on homologous recombination to target the sequence of interested (Sharan et al. 2009). Below the method carried out to create an *E. coli* BL21 (DE3) *Δppc* mutant is described in detail.

Preparing linear dsDNA

To create an ampicillin resistance selection marker (*ampR*) to insert into the bacterial chromosome the linearized plasmid pEU-E01-His-tev-mcs-n1 (CFS cell-free Sciences Co., Ltd Matsuyama, Japan) was used as a template. The linearization of the plasmid was carried out through a restriction digestion with *EcoRI* (New England Bio Labs) for 2 h at 37°C with 300 ng template plasmid and subsequently an inactivation step for 20 min at 65°C.

The linearized plasmid was diluted 1:1000 in H_2O and used as a template to amplify the selectable marker *ampR*. The *ampR* gene was amplified, using following primer pair: forward primer 5'-ATGAACGAACAATATTCCGCATTGCGTAGTAATGTCAGTATGCTCGGCAATCTTCCGCTTCCTCGCTCA-3' and reverse primer 5'-TTAGCCGGTATTACGCATACCTGCCGCAATCCCGGCAATAGTGACCATTAGTGCGCGGAACCCCTATTT-3'. The underlined part of the primer corresponds to the primer that targets the ampR expression cassettes whereas the remaining parts correspond to the 50 base pair homologous sites (upstream and downstream of the *ppc* gene) required to replace the endogenous *ppc* gene in *E. coli* during recombineering. Amplification was carried out in a 50 µl reaction volume consisting of 24.6 µl nuclease free water, 10 µl 5x green GoTaq Flexi buffer (Promega), 5 µl $MgCl_2$ (25 mM), 1 µl dNTPs (each 10 mM), 4 µl forward primer (10 mM), 4 µl reverse primer (10 mM), 0.4 µl GoTaq G2 Flexi DNA polymerase (100 U) (Promega) and 1 µl template. The reaction mixture was amplified by a 2720 Thermal cycler (Applied Biosystems). The thermal cycler was programmed for 5 mins at 95°C as initial denaturing step, followed by 25 steps of 15 s at 95°C denaturing, 20 s at 60°C for primer annealing, 2 min at 72°C for extension, and final extension for 10 s at 72°C. The final hold was set to 15°C. The

PCR products were evaluated through gel electrophoresis to verify the correct product size.

The PCR product was ethanol precipitated to remove salts from the solution. To the PCR reaction mixture, 0.1 volume of 3 M sodium acetate and 3 volume of 100% ice cold ethanol was added, thoroughly mixed and place into a -20°C for 1 h. After the precipitation step, the tube was centrifuged at full speed at 4°C for 30 min. The supernatant was removed and discarded. The DNA pellet was washed in 1 ml ice cold 75% ethanol and centrifuged again for 10 min at 4°C. The ethanol solution was removed and the DNA pellet was air dried overnight. The next day, the pellet was suspended in 15 µl nuclease free water and the concentration was verified and diluted to 100 ng/ µl.

Provision of the λ Red recombination genes

To provide the *E. coli* BL21(DE3) with the necessary λ Red recombination genes, the pSIM5 plasmid (gifted by Donald L. Court) was transfected into electrocompetent *E. coli* BL21(DE3) (Note: pSIM5 has a temperature sensitive origin of replication, thus to maintain the plasmid in the bacterial cells do not incubate > 33°C). 10-50 ng of pSIM5 and 50 µl of electrocompetent *E. coli* BL21(DE3) were pipetted into ice-cold 0.1 cm electroporation cuvettes. The electroporation was carried out using a Bio-Rad Gene Pulser applying following settings 25 µF, 2.5 kV, 200 Ω.

After the electric pulse, 500 µl of LB (room temperature) was added to the cuvette and the solution was transferred to a 2 ml Eppendorf tube and incubated for 1 h at 28°C. After the outgrowth period, 75 µl of the bacterial solution was plated on LB plates containing 25 µg ml^{-1} chloramphenicol and grown overnight at 32°C. The next day a

single colony was picked and grown overnight in 5 ml LB containing 25 µg ml^{-1} chloramphenicol and grown will shaking at 200 rpm at 32°C.

Induction of the λ recombination genes

From the overnight culture, 0.25 ml were transferred to 17.5 ml LB containing 25 µg ml^{-1} chloramphenicol in a 250 ml Erlenmeyer flask. The cells were grown while shaking at 200 rpm at 32°C until the cells reached an OD_{600} between 0.4-0.6. To induce the expression of the λ Red system, the mid-log phase bacterial culture was incubated in a 42°C water bath while shaking for 15 min.

Preparation of the electrocompetent cells and electroporation of the linear dsDNA

Following the incubation at 42°C the cell suspension was immediately chilled in a water-ice bath and left on ice for 10 min. The suspension was transferred to a 50 ml centrifuge tubes and centrifuged at 4600 g for 10 min at 4°C. The supernatant was poured off and 1 ml of ice-cold sterile water was added to the cell pellet. The cell pellet was gently suspended using a 5 ml pipette tip. Afterwards, another 30 ml of ice-cold water was added and gently mixed. The cells are collected again by centrifuging at 4600 g for 10 min at 4°C (Beckman GS-15 R). The supernatant was quickly removed while not disturbing the cell pellet. Again, the pellet was suspended in 1 ml ice-cold sterile water. The 1 ml cell suspension was transferred to a 1.5 ml pre-chilled Eppendorf tube and centrifuged for 30 s at 10 000 g at 4°C. The supernatant was carefully removed and 200 µl of ice-cold sterile water was used to re-suspend the cell pellet. The cells were kept on ice until used.

To introduce the selection marker, 50 µl of competent cells and 100 ng of PCR fragment (see Preparing linear dsDNA) were mixed and pipetted into a chilled

electroporation cuvette 0.1 cm (ThermoFisher). The Bio Rad Gene Pulser was set to 25 µF, 2.5 kV and 200 Ω. After the electric pulse 500 µl of LB were added to the electroporation cuvette and transferred to a 1.5 ml Eppendorf tube. The cells were let to recover for 1 h at 32°C while shaking at 150 rpm.

Outgrowth following electroporation and identification and confirmation of the recombinant clones

To outgrow and screen possible recombinant clones, the cells were incubated for another 2 h at 32°C while shaking at 150 rpm to allow for the expression of the *ampR* resistance gene. Following, the recovery period, a 1:10 dilution in LB was carried out. Subsequently, from the undiluted and diluted cell suspension 50 µl were pipetted on LB plates containing 100 µg ml^{-1} ampicillin.

The plates were incubated at 32°C for 24 h. The potential recombinant cfu appearing on the LB plate were assessed using colony PCR. To prepare a colony PCR, colonies were picked with a pipette tip and each colony was placed into an individual well in a 96-well plate containing LB and the same pipette tip was used to inoculate the PCR reaction mix where the genomic DNA of the cells served directly as template.

To screen for the presence of the *ampR* gene or the native *ppc* following primer pair was used: forward primer 5'-CCGCATTGCGTAGTAATGTCA-3' and reverse primer 5'-ATTACGCATACCTGCCGCAA-3'. The primer pair targets a region within the first and last 50 bp of the *ppc* gene. This is the same region used for the flanking homology site for the *ampR* gene. Thus, the same primer pair was used to screen for colonies carrying the *ampR* gene (1599 bp) or maintained the endogenous *ppc* (2627 bp). The difference in amplicon size was used to discriminate between positive and negative colonies. Amplification was carried out in a 50 µl reaction volume consisting of 25.6 µl

nuclease free water, 10 µl 5x green GoTaq Flexi buffer (Promega), 5 µl $MgCl_2$ (25 mM), 1 µl dNTPs (each 10 mM), 4 µl forward primer (10 mM), 4 µl reverse primer (10 mM), 0.4 µl GoTaq G2 Flexi DNA polymerase (100 U) (Promega).

The reaction mixture was incubated in a 2720 Thermal cycler (Applied Biosystems). The thermal cycler was programmed for 5 min at 95°C as initial denaturing step, followed by 25 steps of 15 s at 95°C denaturing, 20 s at 60°C for primer annealing, 2 min at 72°C for extension, and final extension for 10 s at 72°C. The final hold was set to 15°C and the PCR products were evaluated through gel electrophoresis to verify the correct product size.

4.4.3 Genotyping of the Keio strain JW1692

To genotype the *E. coli* JW1692 a primer pair was designed that targets the first and last 50 bp of the *pps* gene (forward primer 5'-CGCCGCCGGGGATTTATTTT-3' reverse primer 5'- AGAAATGTGTTTCTCAAACCGTTCA-3'). These targeted 50-bp were present in the mutant strain as well because they were used as homologous sites flanking the kanamycin expression cassette (Baba et al. 2006). The expected product size for the wild type *E. coli* was 2444 bp whereas the *E. coli* JW1692 was expected to have a PCR fragment size of 1367 bp. Thus, the difference in amplification product size allowed to discriminate between the wild type and the mutant strain.

Amplification was carried out in a 30 µl reaction volume consisting of 19.68 µl nuclease free water, 6 µl 5x green GoTaq Flexi buffer (Promega), 3 µl $MgCl_2$ (25 mM), 0.6 µl dNTPs (each 10 mM), 0.24 µl forward primer (100 mM), 0.24 µl reverse primer (100 mM), 0.24 µl GoTaq G2 Flexi DNA polymerase (100 U) (Promega). The reaction mixture was incubated in 2720 Thermal cycler (Applied Biosystems). The thermal cycler was programmed for 5 min at 95°C as initial denaturing step, followed by

25 steps of 15 s at 95°C denaturing, 20 s at 60°C for primer annealing, 1 min at 72°C for extension, and final extension for 1 min at 72°C. The final hold was set to 15°C and the PCR products were evaluated through gel electrophoresis to verify the correct product size.

4.4.4 Whole genome sequencing and read mapping of *E. coli* mutant strains

In addition to the genotyping by PCR, the genomes of all *E. coli* mutant strains were fully sequenced. The *E. coli* strains were shipped according to MicrobesNG (http://www.microbesng.uk) instructions for whole genome sequencing of microbes. MicrobesNG provides llumina 250 bp paired end sequencing with 30× coverage using MiSeq or HiSeq and is supported by the BBSRC (grant number BB/L024209/1). Around five weeks after shipping the strains to MicrobesNG, the raw reads and the already quality trimmed reads were available for download. To carry out the read mapping the provided quality trimmed reads were used. The genome sequence for *E. coli* BW25113 version CP009273.1 (Grenier et al. 2014) and *E. coli* BL21 (DE3) version CP001509.3 (Jeong et al. 2009) were obtained from NCBI (Coordinators 2016). The corresponding genome assembly file ASM75055v1 for *E. coli* BW25113 and GCA 00002266v1 for *E. coli* BL21 (DE3) were obtained from Ensemble Bacteria (Zerbino et al. 2017) and the European Nucleotide Archive (https://www.ebi.ac.uk/ena/data/view/GCA_000022665.1), respectively.

To map the reads to the corresponding genome, the genome files were first indexed using `bowtie2-build` and the genome sequence files were provided as FASTA file (Langmead and Salzberg 2012). The quality trimmed paired reads were mapped to the indexed genome file and outputted as SAM file using `bowtie2 –x –1 trimmed_1.fastq.gz -2 trimmed_2.fastq.gz –S mapped.sam` (Langmead

and Salzberg 2012). The output file was converted from the SAM format to the BAM format using `samtools view -b -S mapped.sam > mapped.bam` (Li et al. 2009a). Subsequently, the BAM file was sorted and outputted as sorted BAM file using `samtools sort mapped.bam sorted_mapped.bam`. Finally an index of the sorted BAM file was created using `samtools index sorted_mapped.bam` (Li et al. 2009a). The mapped reads were visualized using IGV (Robinson et al. 2011) by importing the genome file, the corresponding assembly file together with the mapped reads. Conveniently, IGV has a search function that allows to located specific genes in the genome such as *ppc* or *pps*.

To create the *in silico* altered genomes, the target gene *pps* was located in the genome file of *E. coli* BW25113 and replaced with the kanamycin expression cassette. Similarly, *ppc* was located in the genome file of *E. coli* BL21 (DE3) and replaced with the introduced *ampR* gene. Subsequently, the read mapping was repeated as described above using the *in silico* altered genome files as input. The visualization with IGV allowed the verification whether the reads mapped to the expected selectable marker gene, and that the deleted gene was truly deleted.

4.4.5 Sequence modifications

The *S. viridis* coding sequences (without any introns) for PPDK (Sevir.3G253900) and PEPC (Sevir.4G143500) were modified to facilitate the expression of a plant gene in a bacterial cell. Firstly, any targeting sequences were removed that could interfere with the targeting or folding of the protein e.g. the chloroplast transit peptide was removed from PPDK. Secondly, the codon optimization of PEPC and PPDK for the *E. coli* host system was carried out using the publicly available web-tool OPTIMIZER (Puigbò et

al. 2007). To make the sequences compatible for golden gate (Engler et al. 2008) assemblies the nucleotide sequence was further modified such as the removal of Bsa I restriction sites and replacing the start and stop codon with adapter sequences. The adapter sequences were added to the 5’ and 3’ end of the coding sequences (Table 4.1).

Table 4.1: Adapter sequences added to the 5’ and 3’ end of the coding sequences of PEPC and PPDK to make the sequences compatible for golden gate cloning. The start and stop codon in the adapter sequence are highlighted in capital bold letters and the golden gate SC module extensions are underlined (Engler et al. 2008).

Adapter	5’-> 3’ sequence
5’ extension	cactctgtggtctcaa**ATG**catcatcaccatcaccacggaggc
3’ extension	**TAA**cctcagcgctttgagaccacgaagtg

The 5’ extension contained after the SC module sequence a 6× His tag followed by a glycine linker to allow the purification and/or the visualization using a nickel column or an anti-His-tag antibody, if required. The 3’ extension contained after the stop codon a BbCI site to use nicking nuclease, if required, followed by the SC module sequence (Table 4.1). Finally, the modified sequences were sent as FASTA files to Gene Universal Inc. that synthesised and ligated the sequences into a pUC57 plasmid.

4.4.6 Golden Gate assembling

Golden Gate assemblies were carried out using 50 ng destination plasmid (pOGG024 Figure A.1) (Geddes et al. 2019), 60 ng of each pre-cloned DNA part (see Appendix A.3) and added to 1 µl Cut Smart buffer (10x) (NEB), 1 µl ATP (10 mM), BsaI 10 U (NEB), and T4 DNA ligase 400 U (NEB) in a final volume of 10 µl. The reaction tubes

were kept chilled on ice until transferred to a 2720 Thermal cycler (Applied Biosystems). The assembly reactions were incubated for 7 min at 37°C, followed by thirty cycles consisting of 7 min at 37°C and 7 min at 16°C, followed by a 10 min at 37°C and an inactivation step for 15 min at 65°C then a final hold at 15°C prior to transformation into *E. coli* DH5α.

The golden gate assembly reaction was added to heat competent *E. coli* DH5α and incubated on ice for 1 h before incubating for 25 s in a 42°C water bath. Subsequently, the cells were recovered by adding 1 ml of LB and incubated for 1 h at 37°C while shaking at 220 rpm. The *E. coli* cells were plated on LB containing gentamycin 12.5 mg l^{-1}, IPTG 0.1 mM and xGal 40 mg l^{-1} and were incubated overnight at 37°C. The IPTG and xGal allowed an initial screen because positive cfu appeared white whereas cfu harbouring empty plasmids appeared blue. Thus, about five white colonies were picked and grown overnight in 5 ml liquid LB at 37°C while shaking at 250 rpm. That overnight culture was used to harvests the cells and isolate the plasmids using a plasmid purification kit (miniprep Qiagen).

To confirm the correct assembly of the golden gate construct a PCR was carried out that amplified a fragment from the plasmid backbone across the promoter and RBS sequence into the sequence of PEPC and PPDK, respectively. The PCR was carried out as described in Section 4.4.3 and the primer pair pOGG024.Fw1 5'-GGCCATCGTCCACATATCCA-3' and Lv1.PEPC.Rev1 5'-TGTAAGTAAGTGACGCGAGC-3' were used to evaluate all the constructs with PEPC with an expected product size of 1474 bp. All pOGG024 constructs harbouring PPDK were evaluated using the primer pair pOGG024.Fw1 5'-GGCCATCGTCCACATATCCA-3' and Lv1.PPDK.Rev 5'-CCATCTTGATTGCTTGGCGG-3' and the expected product size was 1451 bp. To

verify that the PCR product had the correct fragment size and free of potential contamination a gel electrophoresis was done. In addition, plasmids were Sanger sequenced using pOGG024.Fw1 5'-GGCCATCGTCCACATATCCA-3' and pOGG024.Rv1 5'-AACAACCGCTTCTTGGTCG-3'.

4.4.7 Growth curve of *E. coli* in M9 with a single carbon source

Growth curves for the *E. coli Δppc* expressing *Setaria viridis* PEPC (Sevir.4G143500) (see Appendix A.3.1) were carried out in M9 glucose (see Appendix A.2.1). Growth curves for the *E. coli* JW1692 (Δ*pps*) expressing *Setaria viridis* PPDK (Sevir.3G253900) were carried out in M9 pyruvate (see Appendix A.2 and A.3). Growth curve characterization of *E. coli* BL21 (DE3) Δ*ppc* expressing *Setaria viridis* PEPC (Sevir.4G143500) were carried out by picking multiple colonies from a fresh LB plate and used to inoculate 5 ml of liquid LB containing 12.5 mg l^{-1} gentamycin. The LB culture was grown over night at 37°C with vigorous shaking at 250 rpm. The next day the OD_{600} was measured for each culture and 5 ml M9 glucose were inoculated at equal density (OD_{600} ~0.05) and grown overnight to adjust the cells to the synthetic media at 37°C with vigorous shaking at 250 rpm.

The next day the cells were harvested by centrifugation for 2 min at 12000 g and the supernatant was discarded and washed in M9 glucose. This step was included to remove all traces of the remaining LB that remained in the culture. Subsequently, the OD_{600} was measured and again 5-7 ml M9 glucose containing 12.5 mg l^{1} gentamycin was inoculated at equal density (OD_{600} ~0.05) at 37°C with vigorous shaking at 250 rpm.

The growth curve in Erlenmeyer flasks was carried out using 25 ml M9 glucose containing 12.5 mg l^{-1} gentamycin in 500 ml flasks. For each tested plasmid three

independent measurements were set up and the cells were inoculated at a density of OD_{600} 0.03. Subsequently, the cells were incubated at 37°C at 220 rpm in a horizontal shaker. The OD_{600} was measured and recorded every ~ 60 min.

The growth curve characterization of the *E. coli* JW1692 (Δ*pps*) expressing *Setaria viridis* PPDK (Sevir.3G253900) (see Appendix A.3.2) was carried out in a similar fashion as described above with the exception that the carbon source was pyruvate instead of glucose. The growth curve was recorded in a 96-well plate with eight independent measurements per plasmid tested. The well contained 100 µl M9 with pyruvate as a sole carbon source together with the initial population density of OD_{600} 0.01. The 96-well plate was sealed with a gas permeable membrane (4titude). Next, the growth curve was recorded using a plate reader (FLUOstar Omega BMG Labtech) and the OD_{595} absorbance was measured every 30 min and incubated at 37°C at 500 rpm (double orbital shaking).

4.4.8 Growth curve assessment

Growth curves were assessed using the R package Growthcurver (Sprouffske and Wagner 2016). The package allows to calculate metrics to summarize growth curves such as growth rate and generation time. The parameters were estimated through fitting the observed data to a standard form of logistic equation that is defined with following parameters, N_0 population at time point 0, N_t population at timepoint t, K carrying capacity and r the growth rate (see Equation 3).

$$N_t = \frac{K}{1+\left(\frac{K-N_0}{N_0}\right)e^{-rt}} \qquad \text{(Equation 3)}$$

4.4.9 Keio Collection strains

To characterized whether PPDK was functional in bacteria the *E. coli* Δ*pps* mutant (JW1692) was obtained from the Keio Collection (Baba et al. 2006). The *E. coli* parental strain BW25113 and was therefore used as controls and as a source of material.

4.4.10 Sequence variant synthesis using site-directed mutagenesis

The reagents used for the site-directed mutagenesis were purchased from Agilent unless stated otherwise. The mutagenic oligonucleotides were designed to target the previously identified sites in the rate limiting enzymes of the C_4 cycle such as e.g. PEPC (see Section A.7). The design of the mutagenic oligonucleotides was carried out using the QuikChange Primer Design Tool provided by Agilent Technologies (see Section A.7 for full list of degenerated primers). Primers for sequencing were ordered from Eurogentec. The template plasmids were maintained in *E. coli* DH5α and purified plasmid was isolated using a Qiafilter plasmid midi kit (Qiagen).

The site-directed mutagenesis was carried out using the PEPC construct 2 (see Section 4.4.6) by following the method described by the QuikChange Multi Site-Directed Mutagenesis Kit Instruction Manual (Agilent Technologies). The procedure conducted is described below.

For simultaneous mutagenesis of multiple sites, the oligonucleotides were required to be present in equimolar amounts. To calculate the amount of each oligonucleotide equation 4 was used:

$$\frac{50\ ng\ of\ oligonucleotide}{330\ x\ length\ of\ oligonucleotide} = pmoles\ of\ oligonucleotide \quad \text{(Equation 4)}$$

The mutant strand synthesis reaction was carried out in 25 µl reaction volume consisting of 2.5 µl 10x QuikChange Multi reaction buffer, water to a final volume of 25 µl, 100 ng template plasmid, mutagenic primers 50 ng each. 1 µl dNTP mix and 1 µl QuikChange Multi enzyme blend.

The thermal cycler was programmed for 1 min at 95°C as initial denaturing step, followed by 30 steps of 1 min at 95°C denaturing, 1 min at 55°C for oligo annealing, 12 min at 65°C for extension and finally the reaction was cooled down on ice. Subsequently, the initial template vector was digested by adding 1 µl of DpnI (10 U μl^{-1}) to each mutagenesis reaction. The reaction mixture was gently mixed and centrifuged for 20 s and incubated at 37°C for 1 h. The reaction was cleaned up using ethanol precipitation (0.1 volume of 3 M sodium acetate and 3 volume of 100% ice cold ethanol) mixed and place into a -20°C for 1 h. The supernatant was removed and the reaction tube was let to dry.

Sanger sequencing of the generated variant library

To evaluate whether the site-directed mutagenesis was able to generate a library of PEPC variants that was degenerate at targeted sites, the obtained variant library was Sanger sequenced. To have double stranded plasmid and to increase the amount of plasmid, the variant library was transformed into electrocompetent *E. coli* using a Bio Rad Gene Pulser with following settings 25 µF, 2.5 kV, 200 Ω. The transformed *E. coli* population was recovered in LB for 1 h at 37°C while shaking at 150 rpm. Subsequently, the cell suspension was transferred to 5 ml LB containing gentamycin 12.5 mg l^{-1} and grown overnight at 37°C while shaking at 250 rpm. The following day, the cells were collected and the plasmid was purified using a miniprep kit (Qiagen).

The purified plasmid together with the Sanger sequencing primes (see Appendix Table A.18) were ship for sequencing to Source BioScience.

Estimation of the number of obtained sequence variants

To estimate the number of sequence variants, the variant library was eluted in 4 µl TE buffer and transformed into 50 µl electrocompetent *E. coli* BL21 (DE3) Δ*ppc* using an ice-cold electroporation cuvette (0.1 cm Invitrogen P41050) cuvette. The electroporation took place by applying an electric pulse with a Bio Rad Gene Pulser with following settings 25 µF, 2.5 kV, 200 Ω. Directly after the electric pulse, 500 µl of room temperature LB was added and the cell suspension was transferred to a 2 ml Eppendorf tube. The cell suspension was incubated to recover for 1 h at 37°C while shaking at 150 rpm. After the incubation step, the cells were collected through centrifugation for 2 min at 12000 g and suspended in 75 µl LB. The entire reaction was plated onto pre-warmed LB plates containing 12.5 mg l^{-1} gentamycin and incubated overnight at 37°C. The following day the number of cfu were counted and used as an estimate of the number of sequence variants created.

Second strand synthesis of the mutated strand

After the DpnI digestion the site-directed mutated plasmids were single stranded. In an attempt to improve the transformation efficiency, the second strand was synthesised by adding a complementary primer (5'-cctgctctgcgaggctggccgtagg-3'). Thus, two different approaches were explored.

The second strand was synthesised by means of repeating the QuikChange Multi Site-Directed Mutagenesis approach with the difference that the used primer was complementary to the single stranded plasmid present and it contained no mutated sites.

Thus, the second strand synthesis reaction was carried out in 25 µl reaction volume consisting of 2.5 µl 10x QuikChange Multi reaction buffer, 15.3 µl water, ~1 µg purified variant library (5 µl), 0.2 µl primer for second strand synthesis (100 µM). 1 µl dNTP mix and 1 µl QuikChange Multi enzyme blend. Subsequently, the reaction was placed into a thermal cycler that was programmed for 1 min at 95°C as initial denaturing step, followed by 1 step of 1 min at 95°C denaturing, 1 min at 55°C for oligonucleotide annealing, 12 min at 65 °C for extension and finally the reaction was cooled down to 10°C. An almost identical reaction was set up with the only difference that the enzyme blend was omitted and the amount of water was adjusted.

An alternative second strand synthesis reaction was carried out using DNA polymerase I, large (Klenow) fragment (New England Biolabs). The reaction volume was 25 µl and consisted of 2.5 µl NEB2 buffer, 15.3 µl water, ~1 µg purified variant library (5 µl), 0.2 µl primer for second strand synthesis (100 µM), 0.5 µl dNTP (each 10 mM), 1.5 µl (7.5 U) Klenow fragment. The reaction was mixed and incubated for 1.5 h at room temperature and subsequently moved for 2 h at 37°C.

Each of the second strand synthesis approaches were purified using ethanol precipitation and were electroporated into *E. coli* BL21 (DE3) Δ*ppc* as described above (see Section 4.4.10).

Chapter 5: Identification and characterisation of the bundle sheath cell chloroplast monocarboxylate and dicarboxylate transporter of *Setaria italica*

5.1 Chapter introduction

Multiple transmembrane transport steps are required to facilitate C_4 photosynthesis. While the transporters that catalyse many of these transport steps have now been characterised, there are still several transporters that have escaped experimental detection. Foremost amongst these missing transporters are the transporters that facilitate the malate uptake and pyruvate export reactions of the bundle sheath cell chloroplast. In this chapter, a novel monocarboxylate and dicarboxylate transporter is characterised that provides both the malate uptake and pyruvate export function in the C_4 plant *Setaria italica*. The identification of this transporter completes the molecular pathway for the core C_4 cycle of this species.

5.2 Background

5.2.1 Metabolite transport in NADP-ME C_4 plants

C_4 photosynthesis was first described over 50 years ago (Hatch and Slack 1966). Since then, substantial insights into this photosynthetic pathway have been gained by studying its evolution, ecology and biochemistry (Christin et al. 2015; Osborne and Sack 2012; Sage 2016; Westhoff and Gowik 2004). These studies have revealed that C_4 photosynthesis has evolved independently more than 60 times, and that the plants descendant from these independent evolutionary events can exhibit substantial disparity in their biochemistry, anatomy and physiology. Although there are substantial differences between these independently evolved C_4 species, it is possible categorize them based on shared aspects of their anatomy (Muhaidat et al. 2007; Rao and Dixon

2016) or biochemistry (Wang et al. 2014). For example, C_4 species are typically categorised into three distinct subtypes by the major decarboxylase enzyme that is expressed in the bundle sheath cells (von Caemmerer and Furbank 2003) (also see Chapter 1). Within each subtype, it is thought that species share similarities in the enzymes and transporters that provide the substrate and recycle the end product for this major decarboxylase (von Caemmerer and Furbank 2003).

The function of the C_4 pathway is to increase CO_2 concentration around RuBisCO (Weber and von Caemmerer 2010)(Chapter 1). This enhanced CO_2 concentration results in significant energy, water and nitrogen savings such that C_4 plants can achieve ~50% higher rates of photosynthesis than C_3 plants given the same resources (Evans and von Caemmerer 2000). Thus, C_4 photosynthesis is a desirable trait for introduction into C_3 crop plants, such as rice, to increase yield (Hibberd et al. 2008). However, in order to engineer the pathway, one needs to have a complete list of components required for the pathway to function.

In C_3 plants, CO_2 enters the mesophyll cells by diffusion. It subsequently enters the chloroplast by diffusion and then the entire CO_2 assimilation pathway occurs within the chloroplast. (Figure 5.1 A) (Weber and von Caemmerer 2010). Inside the chloroplast, the photosynthetic carbon reduction cycle produces one triose phosphate for every three CO_2 molecules that are reduced. This triose phosphate either leaves the chloroplast through the triose phosphate translocator (TPT) in exchange for orthophosphate or it stays in the chloroplast (Flügge 1999). If it remains within the chloroplast it can take two different routes. Either it completes the photosynthetic carbon reduction cycle and regenerates to RuBP, or it is converted into starch (Figure 5.1 A). Thus, in C_3 plants a maximum of one transport step is required for every three molecules of CO_2 that are assimilated (Weber and von Caemmerer 2010).

In contrast to C_3 plants, the pathway for CO_2 fixation in most C_4 plants is distributed across two distinct cells, namely the mesophyll and bundle sheath cells. Furthermore, within each cell the metabolic reactions are further partitioned between subcellular organelles and the cytosol, such that completion of the C_4 pathway requires movement of metabolites both between cells and into and out of organelles within these cells. Therefore, C_4 plants require many more transport steps than C_3 plants to facilitate CO_2 assimilation (Bräutigam et al. 2008).

In NADP-ME type C_4 photosynthesis the primary carboxylation takes place in the cytosol of the mesophyll cell. Here, phosphoenolpyruvate (PEP) is carboxylated by PEPC yielding one molecule of the C_4 compound oxaloacetate (Weber and von Caemmerer 2010). Next, oxaloacetate is transported into the chloroplast (Figure 5.1 B Step 1) where it is reduced to malate by malate dehydrogenase and transported back to the cytosol (Figure 5.1 B Step 2). Both the oxaloacetate import and malate export steps are mediated by the oxoglutarate/malate transporter DiT1/OMT1. This transporter is a C_4 dicarboxylate antiporter that exchanges one C_4 dicarboxylate for another, thus maintaining stoichiometry. Malate then diffuses from the mesophyll cell to the bundle sheath cell via the plasmodesmata that connect the two cells (Weber and von Caemmerer 2010). This diffusion is driven by a concentration gradient and facilitated by the abundant plasmodesmata present at the interface of mesophyll and bundle sheath cells (Stitt and Heldt 1985). Once in the bundle sheath, malate is transported into the chloroplast (Figure 5.1 B Step 3). It has previously been proposed that DCT2 is the transporter that facilitates this transport for the C_4 cycle (Weissmann et al. 2016). However, this is impossible, as DCT2 (like OMT1) is a four carbon dicarboxylate antiporter that exchanges one four carbon dicarboxylate for another. This transport reaction does not result in a net gain of carbon in the chloroplast

(4C – 4C = 0), and therefore cannot support the C_4 cycle. Thus, a different and as yet unknown transporter must provide net import of malate. Once inside the chloroplast malate is decarboxylated by the NADP-ME enzyme. This reaction yields one molecule of CO_2 and one molecule of pyruvate. It also converts one molecule of $NADP^+$ to NADPH and in doing so transfers reducing power to the bundle sheath from the mesophyll. The resulting pyruvate is subsequently transported from the chloroplast to the cytosol of the bundle sheath cell by another unknown transporter (Figure 5.1 B Step 4). Pyruvate diffuses to the mesophyll cell via the plasmodesmata where it is transported into the chloroplast (Figure 5.1 B Step 5). In some C_4 species, this pyruvate transporter is thought to be pyruvate sodium symporter BILE ACID:SODIUM SYMPORTER 2 (BASS2). Transcript abundance data suggests that this is the pyruvate transporter used in a large number C_4 species within the eudicots (Kelly et al. 2017). In addition, cell type specific transcript abundance data also suggests that this is the pyruvate transporter used in the mesophyll cells of *Setaria italica* (John et al. 2014). However, this gene is not expressed in sorghum or maize and thus at least one additional pyruvate import system must also exist in C_4 plants. Once inside the chloroplast, pyruvate is phosphorylated to phosphoenolpyruvate by pyruvate phosphate dikinase (PPDK). Finally, phosphoenolpyruvate is transported to the cytosol by the phosphoenolpyruvate/phosphate translocator PPT1 (Figure 5.1 B Step 6), completing the C_4 cycle (Weber and von Caemmerer 2010). Thus, to carry out the simplest possible NADP-ME C_4 cycle, a plant has to carry out six transport processes for every single CO_2 molecule assimilated. Out of these six transport steps, two are carried out by transporters that are yet to be characterised (Figure 5.1 B).

The assimilation of one molecule of CO_2 in the photosynthetic carbon reduction cycle in the bundle sheath chloroplast requires two NADPH. This is necessary because

every assimilated CO_2 yields two molecules of 3-phosphoglycerate that require reduction to triose phosphate. Only one of the two NADPH is generated from the oxidative decarboxylation of malate (see above). While some additional reducing power is generated by photosystem I in the bundle sheath, it is thought that there is generally not enough reducing power available to complete the photosynthetic carbon reduction cycle (Majeran et al. 2008). As a consequence, one molecule of 3-phosphoglycerate is transported from the bundle sheath chloroplasts (Figure 5.1 Step 7), and diffuses to the mesophyll cell where it is transported into the mesophyll cell chloroplast (Figure 5.1 Step 8). Both the export from the bundle sheath chloroplast and the import to the mesophyll cell chloroplast are mediated by the triose-phosphate/phosphate translocator (TPT) where it is reduced to glyceraldehyde phosphate by the glyceraldehyde phosphate dehydrogenase. Glyceraldehyde phosphate is transported out of the mesophyll chloroplast presumably via the TPT transporter (Figure 5.1 Step 9 and 10). It then diffuses back to the bundle sheath cell where it is transported into the chloroplast presumably by TPT. Glyceraldehyde phosphate then enters the carbon reduction cycle or is converted into starch (Weber and von Caemmerer 2010). This additional cycle to generate reducing power adds four transport steps for every single molecule of CO_2 that is assimilated. Thus, in total there are ten transport processes required to fix a single molecule of CO_2 using C_4 photosynthesis.

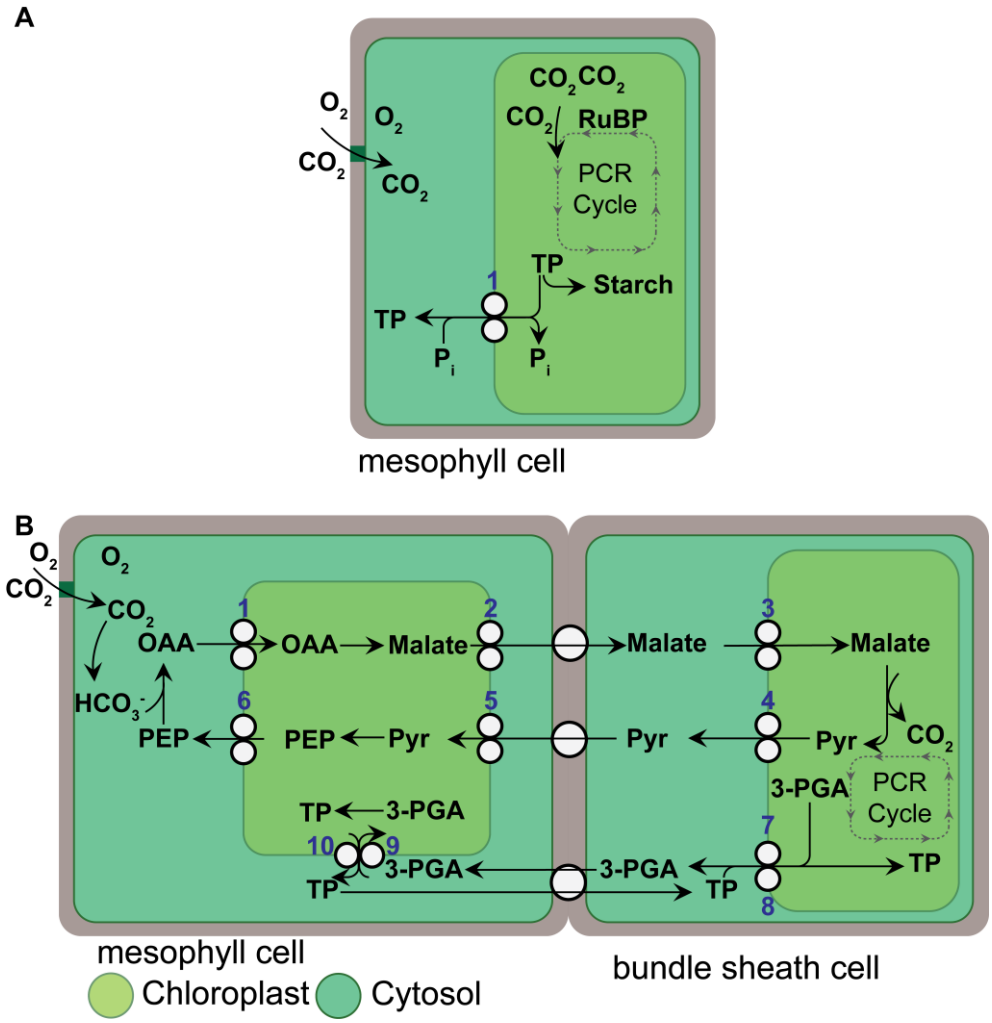


Figure 5.1: The number of transport steps required in C_3 and C_4 plants to assimilate CO_2 **A)** In C_3 plants carbon diffuses into the chloroplast where it is being fixed in the photosynthetic carbon reduction cycle (PCR). For every three molecules of carbon one transport step is required. **B)** The carbon concentrating mechanism in C_4 plants requires at least ten transport steps for every single molecule of CO_2 assimilated. Abbreviations: ribulose 1,5-bisphosphate (RuBP), triose phosphate (TP), ortho-phosphate (P_i), phosphoenolpyruvate (PEP) oxaloacetate (OAA), pyruvate (Pyr), 3-phosphoglyceric acid (3-PGA), bicarbonate (HCO_3^-)

5.2.2 Further consideration of putative malate and pyruvate transporters

A potential malate uptake system DICARBOXYLIC ACID TRANSPORTER 2 (DCT2) has been described in reconstituted liposomes (Taniguchi et al. 2004) and *Z. mays*

knockout plants (Weissmann et al. 2016). Given the severe phenotype observed in *Z. mays* knockout plants, DCT2 must be carrying out a crucial function (Weissmann et al. 2016). However, the functional characterisation of DCT2 showed that it exchanges malate for oxaloacetate, 2-oxoglutarate, glutamate and glutamine (Taniguchi et al. 2004). Thus, DCT2 exchanges malate, a C_4 carbon compound, with either another C_4 carbon compound (oxaloacetate) or a C_5 compound (2-oxoglutarate, glutamate, glutamine). From this it follows that with DCT2 alone, there is no net gain in carbon in the bundle sheath chloroplast (Figure 5.2). In contrast, it is possible that DCT2 facilitates net carbon export from the bundle sheath chloroplast given that DCT2 exports C_5 compounds in exchange for cytosolic malate. Thus, other as yet unknown uptake systems must be present to supply malate to the bundle sheath chloroplasts and ensure that there is a net gain in CO_2 around RuBisCO.

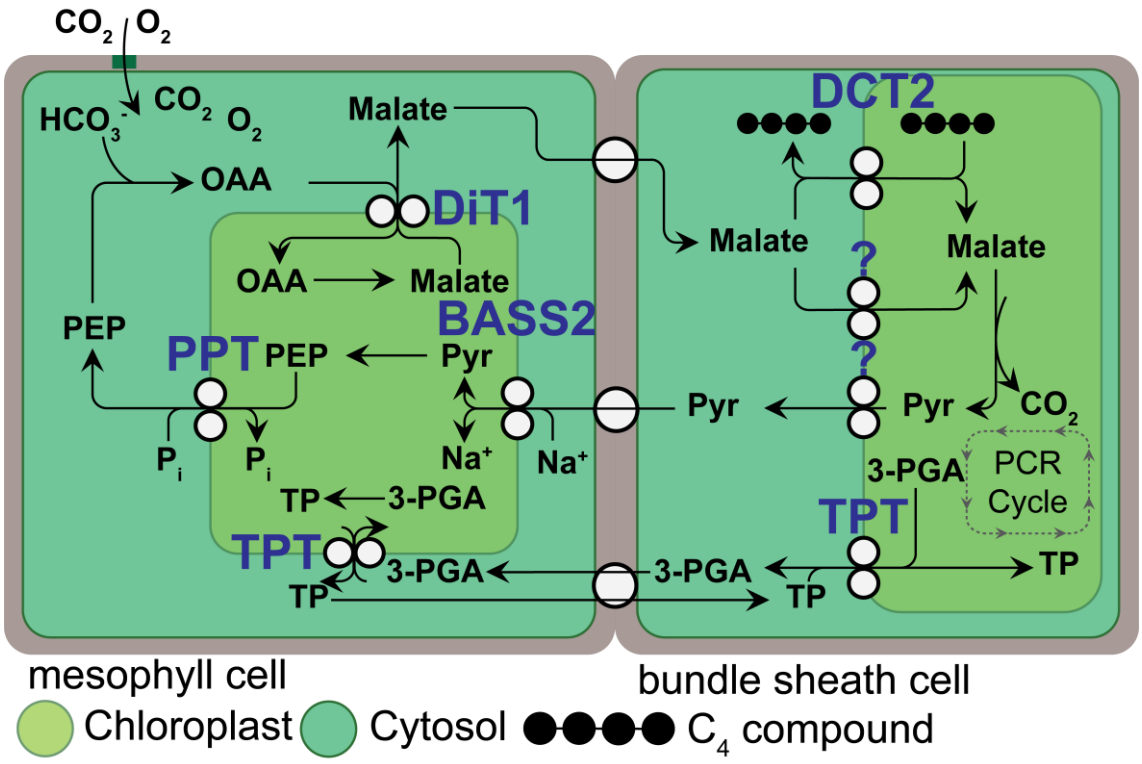


Figure 5.2: The known transporters of the NADP-ME C_4 cycle. Abbreviations of the transporters: PEP/phosphate translocator (PPT), dicarboxylate transporter 1 (Dit1), dicarboxylic acid transporter 2 (DCT2), bile acid:sodium symporter family protein 2 (sodium-dependent pyruvate transporter) (BASS2), triose phosphate translocator (TPT). Abbreviations of the metabolites: triose phosphate (TP), ortho-phosphate (P_i), phosphoenolpyruvate (PEP) oxaloacetate (OAA), pyruvate (Pyr), 3-phosphoglyceric acid (3-PGA), Sodium (Na^+), bicarbonate (HCO_3^-)

Multiple pyruvate export systems have been suggested base on analysis of cell-type-specific expression of transporter-like genes in C_4 plants (John et al. 2014; Majeran and van Wijk 2009). For example, the preferential expression of the multi-transmembrane domain protein MEP3a (Seita.3G258000.1) in the bundle sheath cells in *Setaria viridis* and *Zea mays* lead to the hypothesis that MEP3a is the pyruvate exporter in the bundle sheath chloroplast (John et al. 2014). However, the function of MEP3a has yet to be experimentally characterised. Thus, a pyruvate exporter also remains uncharacterised.

5.2.3 The use of bacteria to characterise chloroplast malate and pyruvate transporters

Malate and pyruvate are key metabolites in primary metabolism and also serve as the fundamental building blocks for amino acids and fatty acids (Figure 5.3). The metabolites are not unique to plants, but occupy a central position in metabolism of almost all organisms. Furthermore, they are continuously produced during cellular growth and operation. However, because they serve as fundamental building blocks, or precursors thereof, they are continuously consumed through respiration and the synthesis of amino acids and fatty acids. Therefore, with some exceptions, in a normal cell pyruvate and malate do not usually accumulate to high levels. Moreover, discarding these compounds is energetically wasteful and thus transports that mediate net efflux of these metabolites in microbial systems have not been characterised. As both pyruvate and malate occupy key positions in the primary metabolism of *E. coli*, and as chloroplasts are domesticated bacteria, this chapter sets out to use *E. coli* to accelerate the discovery of potential chloroplast pyruvate and malate transporters. In brief, *E. coli* lacking the dicarboxylate importer dctA (*ΔdctA*) is unable to grow on malate as a sole carbon source (Davies et al. 1999). Thus, to screen for malate import activity, it was thought to simply express candidate transporter genes in a *ΔdctA* strain of *E. coli* (JW3496-1) (Baba et al. 2006) and test for the ability of the transporter to restore growth on malate. To measure pyruvate export activity, we simply overexpress the candidate transporter in wild type *E. coli* cells grown on glucose as a sole carbon source and measure pyruvate accumulation in the cell culture medium. Here, *E. coli* grown on glucose as a sole carbon source breaks down each molecule of glucose into two molecules of pyruvate. Thus, if glucose is in excess, and the putative transporter exports pyruvate, then pyruvate should accumulate in the cell culture medium. This

should occur even though there are five transporters in *E. coli* that have been shown, or are predicted, to import pyruvate (*YfeH* (InterPro/UniProtKB: P39836,), *CstA* (Schultz and Matin 1991), *BtsT* (Kristoficova et al. 2018), *YnfM* (Göhler et al. 2011), *YhjX* (Zhao et al. 2017)).

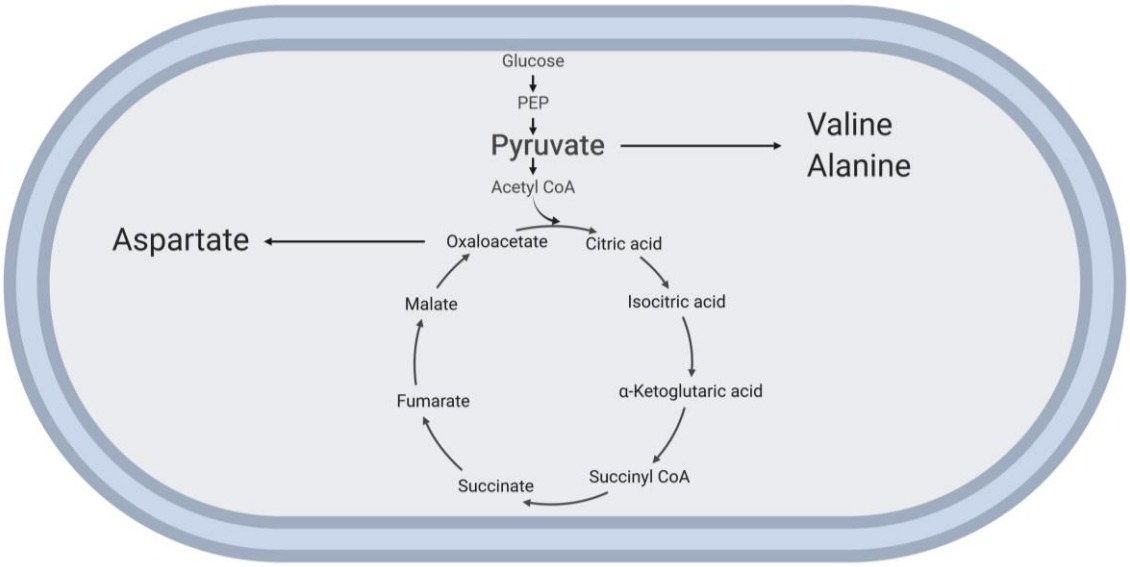


Figure 5.3: Taking advantage of the metabolic flux through *E. coli* grown in minimal media with glucose as the sole carbon source. *E. coli* oxidises each glucose into two molecules of pyruvate, thus in this particular set up there is only one defined carbon source present in the supernatant and pyruvate, the metabolite of interest, is present in abundance in the cell.

5.2.4 Chapter aims and outlines

The aim of the project presented in Chapter 5 was to experimentally characterise a putative metabolite transporter that was identified through bioinformatic analysis in Steven Kelly's lab, and determine whether it could provide the function of the bundle sheath chloroplast pyruvate export or malate import transporter in the NADP-ME C_4 subtype. The first aim was to experimentally validate the target gene in a quick and

simple system. Therefore, the bacterial test system was used. In case the target gene did not possess the predicted function, the fast turnaround of the bacterial test system would allow to go after the next best candidate gene. The second aim was to downregulate the target gene in the C_4 model plant *Setaria viridis* once either the pyruvate or malate transport activity was confirmed. The downregulation of a crucial metabolite transporter in the C_4 cycle was expected to have an impact on the carbon assimilation rate if no other redundant genes with similar functions are expressed.

5.3 Results and discussion

5.3.1 Identification of a putative monocarboxylate and dicarboxylate transporter using cell-type specific expression data

To identify potential candidate genes with either monocarboxylate or dicarboxylate transport activity, a bioinformatics analysis was carried out by Steven Kelly and Basel Abu Jamous. The analysis leveraged cell type specific transcript abundance data together with protein topology to shortlist a single candidate gene Seita.4G275500.1. It was predicted that it may be involved in the uptake of malate and/or the export of pyruvate in the bundle sheath cells of NADP-ME C_4 plants. Therefore, this candidate gene was taken further to functionally characterise in the laboratory and was named putative monocarboxylate and/or dicarboxylate transport 1 (pMDT1). The gene belongs to an uncharacterised gene family that can be found in land plants, algae and bacteria (Figure 5.4). In plants, two separate gene duplication events prior to the divergence of the angiosperms have given rise to three distinct clades of genes in modern day land plants. The candidate gene selected for this chapter is from clade III (Figure 5.4). To provide additional insight into the function of the gene, and to

determine whether there are any potential differences between the function of the gene in C_3 and C_4 species, the *Arabidopsis thaliana* ortholog of the pMDT1 (AT4G19390) was also selected for functional characterisation.

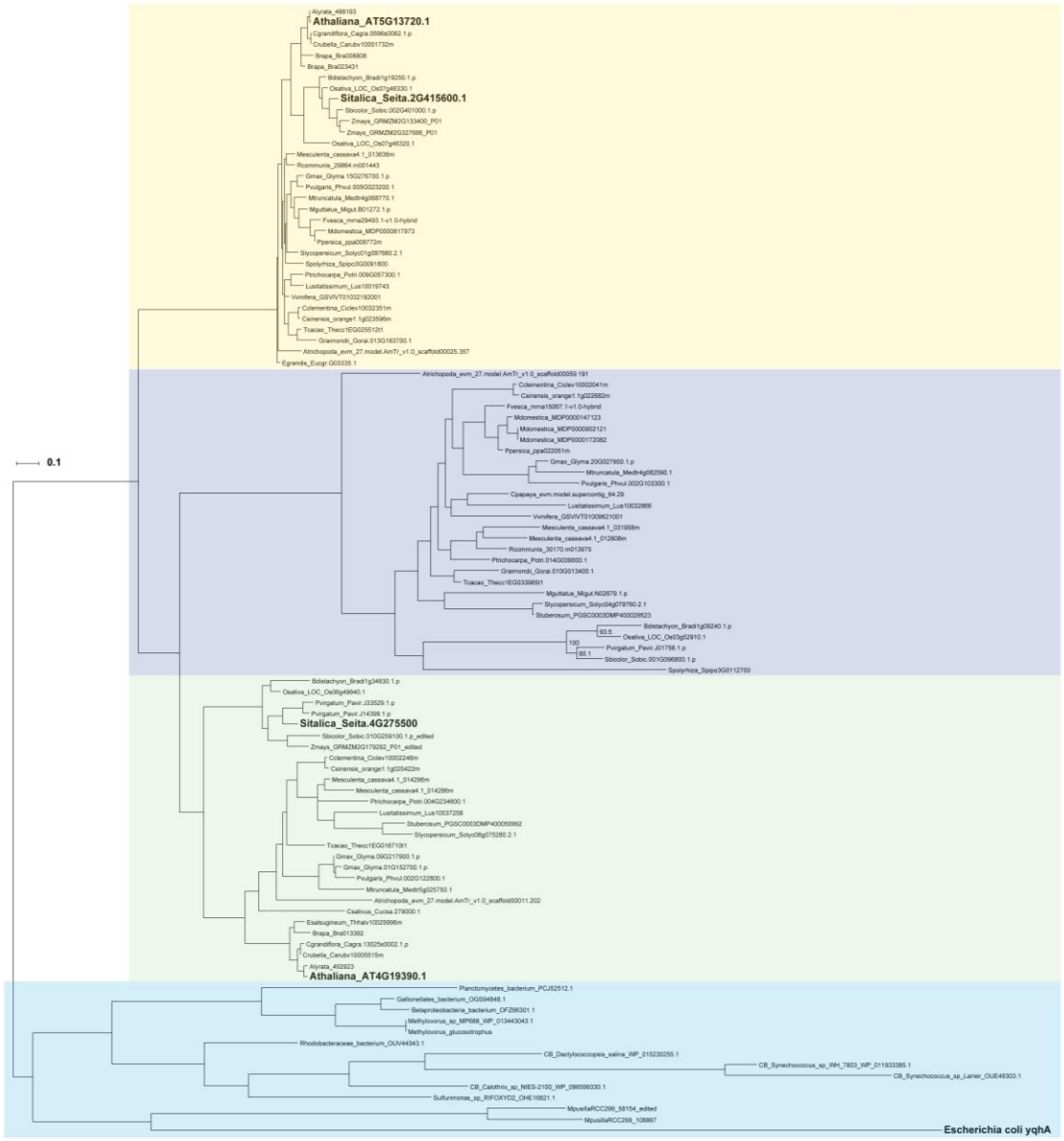


Figure 5.4: Maximum likelihood phylogenetic tree depicting the relationship between examples of genes belonging to the uncharacterized protein family 0114. The gene tree contains examples from algal, bacterial and plant orthologous from the UPF0114 gene family.

Analysis of the protein sequence of pMDT1 revealed that both the *Setaria italica* and *Arabidopsis thaliana* versions of the gene have a chloroplast transit peptide and four transmembrane domains (Figure 5.5). The predicted number of transmembrane domains is low when compared to known malate or pyruvate transporters such as OMT1 (13 transmembrane helices), DCT2 (13 transmembrane helices) or BASS2 (9 transmembrane helices, Figure 3.11). pMDT1 also has fewer transmembrane helices than PPT1 (7 transmembrane helices) and TPT (8 transmembrane helices) which transport phosphoenolpyruvate and 3-phophoglycerate respectively. Thus, it is possible that if the protein is indeed a pyruvate or malate transporter it may function as a multimer e.g. a homodimer consisting of 8 transmembrane helices.

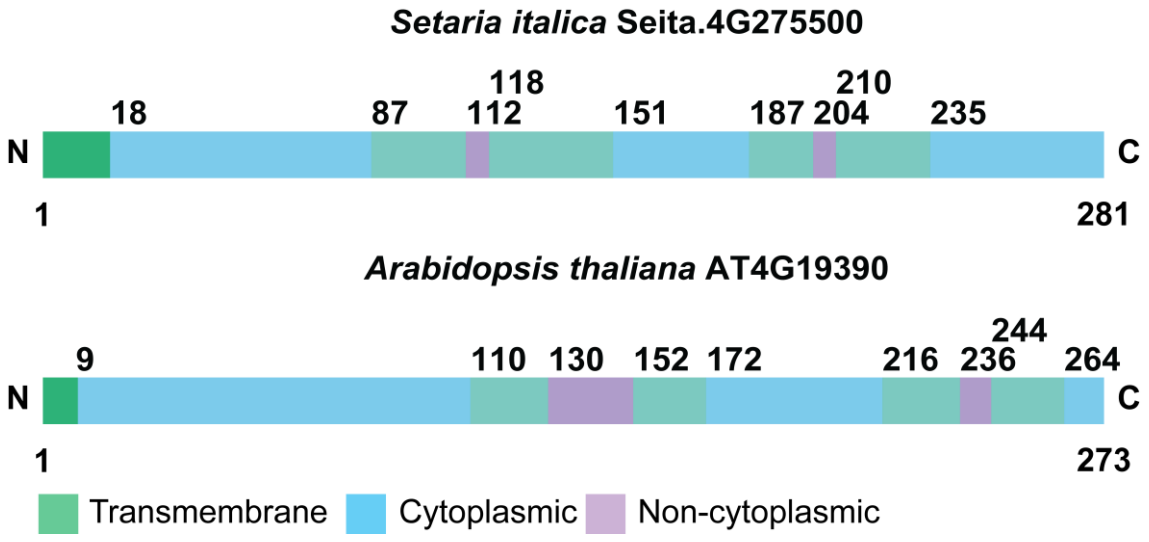


Figure 5.5: Domain organisation of the *Setaria italica* pMDT1 and its *Arabidopsis thaliana* ortholog. The protein domains are highlighted in colours. Chloroplast targeting peptide in dark green, cytoplasmic domain in light blue, transmembrane domain in light green and the non-cytoplasmic domain in purple. Domain information was obtained from InterPro K3XZ06 for *S. italica* Seita.4G275500 and from aramemnon (Schwacke et al. 2003) for *A. thaliana* AT4G19390

Regarding the localization of pMDT1, both variants have a predicted chloroplast transit peptide. This is in line with the required localization to transport monocarboxylates or dicarboxylates across the chloroplast envelop. Furthermore, the *Arabidopsis* protein was previously identified to be part of the chloroplast envelope proteome (Ferro et al. 2010). However, no functional characterisation of the *A. thaliana* ortholog of pMDT1 has been conducted.

5.3.2 Expression of pMDT1 in *E. coli* results in accumulation of pyruvate in the cell culture medium

To evaluate whether pMDT1 or it's *A. thaliana* ortholog have pyruvate export activity, a codon modified coding sequence of each gene was synthesised without the chloroplast transit peptide or introns (see Section 5.5.4). These sequences were cloned into an inducible expression vector and transformed into *E. coli* BW25113. *E. coli* harbouring either the *S. italica or A. thaliana* variant of pMDT1 were grown in M9 minimal medium with glucose as sole carbon source and induced once the optical density (OD_{600}) reached 0.2 (Figure 5.6 A). The non-induced control harbouring the *S. italica* gene variant displayed an exponential growth characteristic as expected. However, *E. coli* cell cultures expressing either the *A. thaliana* or the *S. italica* variant of pMDT1 stopped growing as soon as expression was induced (Figure 5.6 A).

To examine whether the growth arrest could have been caused by pyruvate export, the pyruvate concentration in the supernatant was measured. The amount of pyruvate in the cell culture medium of the non-induced *E. coli* cells was below the detection limit of the assay (Figure 5.5). However, the amount of pyruvate in cell culture medium of the induced *E. coli* cells expressing either the *S. italica* or *A. thaliana* gene increased

over time such that large quantities of pyruvate accumulated outside of the cells (Figure 5.6). To help put this export rate in context, the cells expressing the *S. italica* version of the gene exported an equivalent of ~1.6 times their total dry weight cell mass in pyruvate during the 180 minutes following induction (4.4×10^{-4} g of pyruvate from 2.8×10^{-4} g dry weight of cells). Thus, although the cells did not increase in number following induction, they continuously took up glucose and exported it back to the cell culture medium in the form of pyruvate. Under normal aerobic conditions wild type *E. coli* has been measured to consume glucose at a maximum rate of 20 mmol per gram dry weight per hour (Varma et al. 1993). Thus, during the induction the cells should have consumed a maximum of 16.8 µmol of glucose, which could be converted into a maximum of 33.6 µmol of pyruvate assuming a 100% conversion of glucose to pyruvate. Therefore, as the cells exported 5 µmol of pyruvate to the cell culture medium, assuming that glucose consumption was at its maximum, ~15% of the carbon taken up as glucose was exported to the cell culture medium as pyruvate. Thus, as cells did not grow either glucose uptake was lower in these cells, or other metabolites were also being exported.

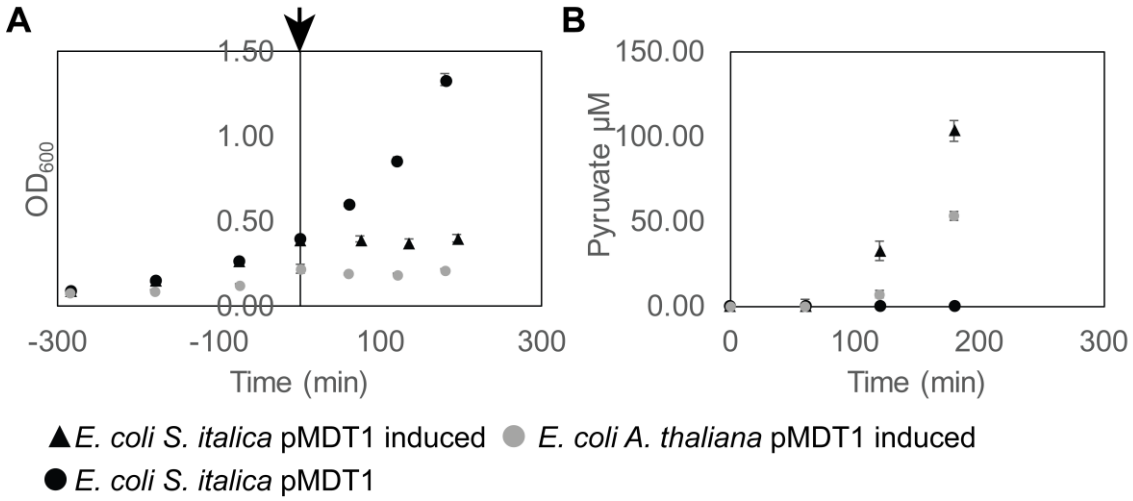


Figure 5.6: Expression of the pMDT1 in *E. coli* BW25113. **A)** The C_4 and C_3 variant of the pMDT1 were induced in *E. coli*. The moment of induction was considered as timepoint zero

and is marked with a black arrow. As non-induced control *E. coli* harbouring the C_4 variant was used. **B)** The accumulation of pyruvate in the supernatant since induction in the induced and non-induced *E. coli* cultures. Error bars indicate the standard deviation of the samples (N=3).

To evaluate whether the two variants of pMDT1 also exported other metabolites, all organic acids were extracted using a methanol-chloroform extraction (see Material and Methods 5.5.8). The extracted metabolites were analysed by mass spectrometry in collaboration with John Lunn and Regina Feil at the Max Planck Institute for Plant Physiology in Golm. The metabolite screen revealed that both of the tested transporters predominantly exported pyruvate (Figure 5.7). However, in both cases significant quantities of other metabolites were also detected (Figure 5.7). For example, in the *E. coli* cell lines expressing the *S. italica* transporter there was significant accumulation of succinate, malate, citrate, 3-phosphoglycerate and fumarate in the cell culture medium. Interestingly, the total mass contained within these 6 metabolites is exactly equivalent to the total mass of glucose that could have been maximally consumed in the time since induction (i.e. 16.8 µmol). Thus, it appears that there is ~100% conversion and export of consumed glucose as monocarboxylates and dicarboxylates following expression of this transporter. However, it should be noted that glucose consumption was not measured directly and could have been higher than the expected value for wild type *E. coli* cells. Nonetheless, it appears that the putative transporter from *S. italica* is efficient at exporting pyruvate and other carboxylates from the cell and as export was equal to the anticipated glucose consumption rate there is unlikely to be other metabolites not detected here that are also exported by the transporter. Similarly, the *A. thaliana* variant of the gene exported the equivalent of ~7 µmol of glucose and thus ~42% of the total maximal consumed glucose.

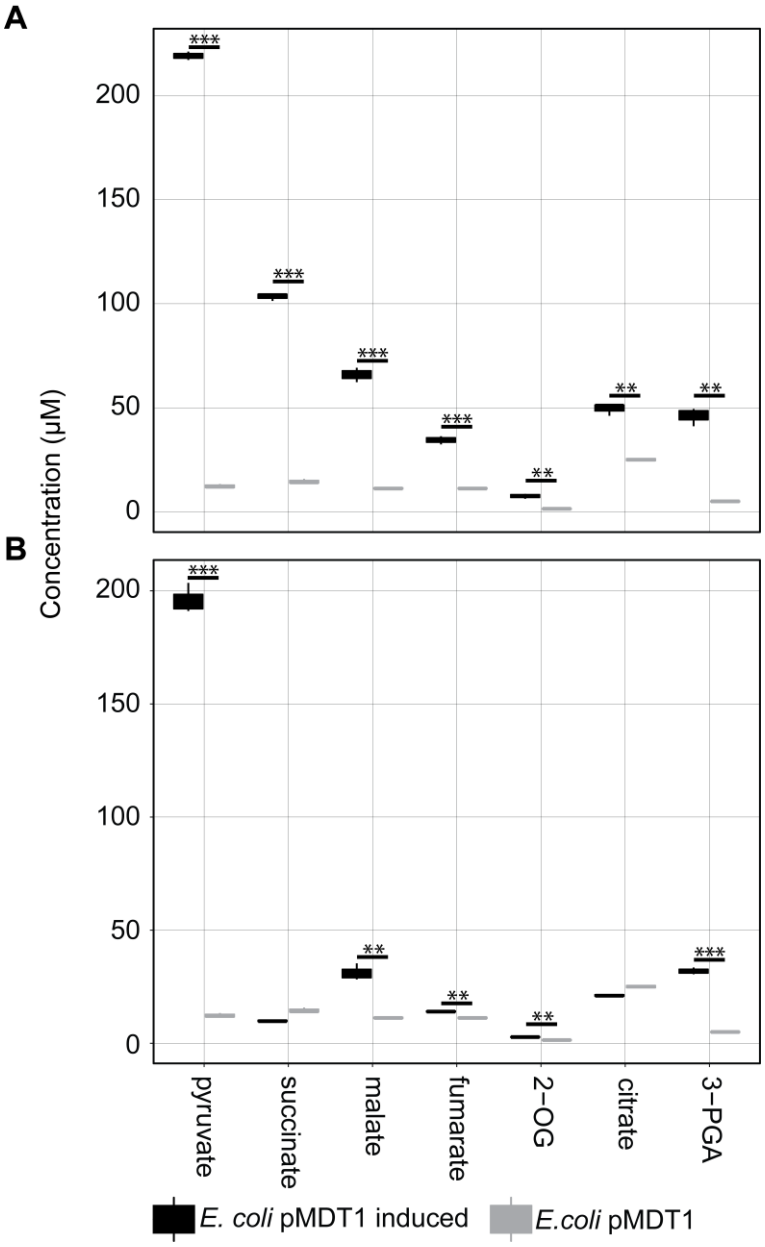


Figure 5.7: Boxplot representing concentrations of metabolites measured in the supernatant after 3 h of induction and the none induced control. **A)** *E. coli* expressing the *S. italica* pMDT1 and **B)** *E. coli* expressing the *A. thaliana* pMDT1. The measured metabolites can be classified into monocarboxylates (pyruvate), dicarboxylates (succinate, malate, fumarate, 2-oxoglutarate), tricarboxylates (citrate) and phosphorylated carboxylates (3-phosphoglyceric acid (3-PGA)). Stars indicating significant difference between the induced and none-induced control (t-test ** for p-value ≤ 0.01 and *** for p-value ≤ 0.001)

The metabolite measurements revealed that both pMDT1 variants facilitated the export of large quantities of pyruvate out of the *E. coli* cells. This is compatible with the hypothesis that this transporter facilitates pyruvate export from the bundle sheath cell chloroplast in the C_4 cycle of *S. italica*. However, given that the *S. italica* version of the transporter appeared to be more permissive to C_4 dicarboxylates (i.e. there were substantially higher levels of succinate, malate and fumarate in the cell culture medium of the cell lines expressing the *S. italica* transporter than in the cell lines expressing the *A. thaliana* version of the same gene), this raised the question whether the transporter may also facilitate influx of dicarboxylates. Specifically, could it facilitate the import of malate if malate was supplied as a sole carbon source?

5.3.3 pMDT1 facilitates uptake of malate into *E. coli* cells.

To evaluate whether expression of pMDT1 is sufficient to enable *E. coli* cells to take up malate from the cell culture medium, both the *S. italica* and *A. thaliana* versions of the gene were expressed in the *E. coli* malate uptake null mutant Δ*dctA* (JW3496) (Baba et al. 2006). If the transporter is capable of facilitating malate uptake, then expression of the transporter in cells grown on malate as a sole carbon source should result in export of pyruvate to the cell culture medium. Expression of the transporter is not expected to restore growth, as it will be expected that any malate that is taken up by the cells will be exported back to the cell culture medium in the form of pyruvate and other carboxylates (Figure 5.8).

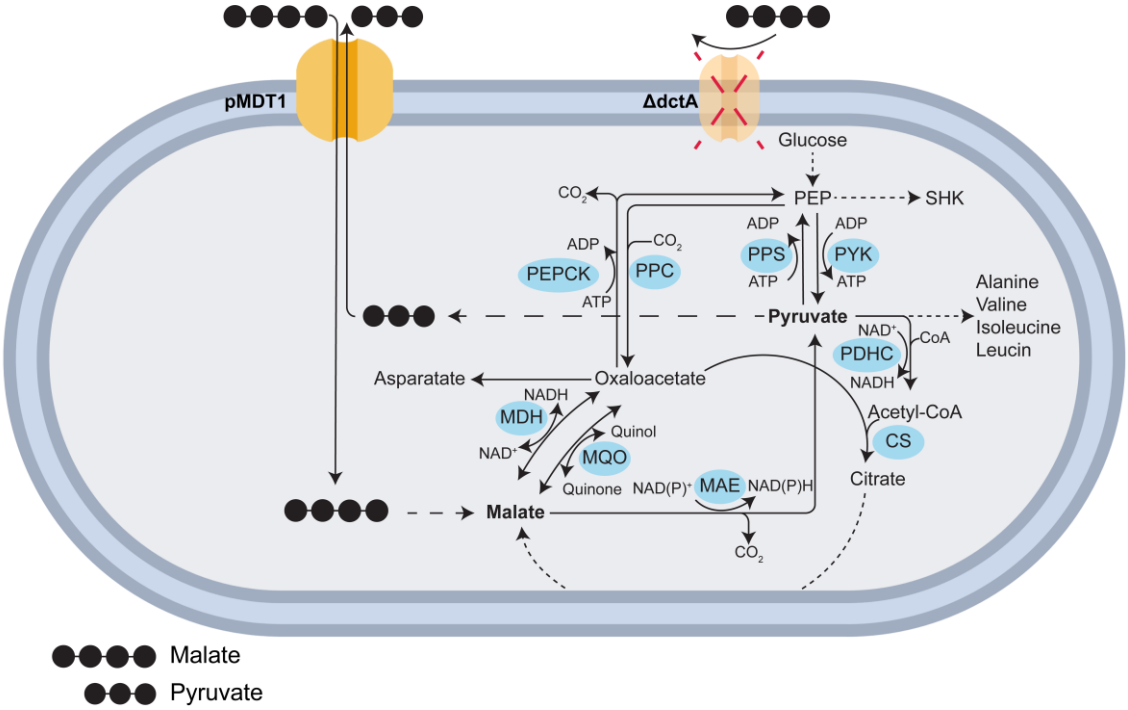


Figure 5.8: Schematic representation of an *E. coli ΔdctA* expressing the putative monocarboxylate and dicarboxylate transporter 1 (pMDT1). *E. coli ΔdctA* grown in media with malate as sole carbon source and expression pMDT1 may uptake malate that would eventually be converted to pyruvate. Pyruvate is exported through the previously identified characteristics of pMDT1. Abbreviations: MAE, malic enzyme; MDH, malate dehydrogenase; MQO, malate quinone oxidoreductase, PC, pyruvate carboxylase, PDHC, pyruvate dehydrogenase complex, PEPCK, PEP carboxykinase, PPC, PEP carboxylase, PPS, PEP synthase, PYK, pyruvate kinase. Figure adapted from Sauer and Eikmanns 2005 (Sauer and Eikmanns 2005)

E. coli Δ*dctA* are unable to grow in media with malate as sole carbon source because DctA is required for the uptake of malate (Davies et al. 1999) (Section 5.2.3). To verify whether the obtained *E. coli* JW3496 strain was indeed a *dctA* knockout mutant, the whole genome was sequenced to confirm the absence of the gene (Figure 5.9). The paired-end reads were mapped against the wild type *E. coli* BW25113 genome and to the *in silico* created version of the BW25113 genome where *dctA* was replaced with

the kanamycin expression cassette (Figure 5.9). The visualization of the mapped reads revealed that there were no reads present that mapped to the *dctA* gene in the Δ*dctA*. However, reads mapped to the kanamycin resistance cassette that was located in the position of *dctA* in the genome. This confirmed that the *E. coli* strain JW3496 was deficient in the *dctA* gene.

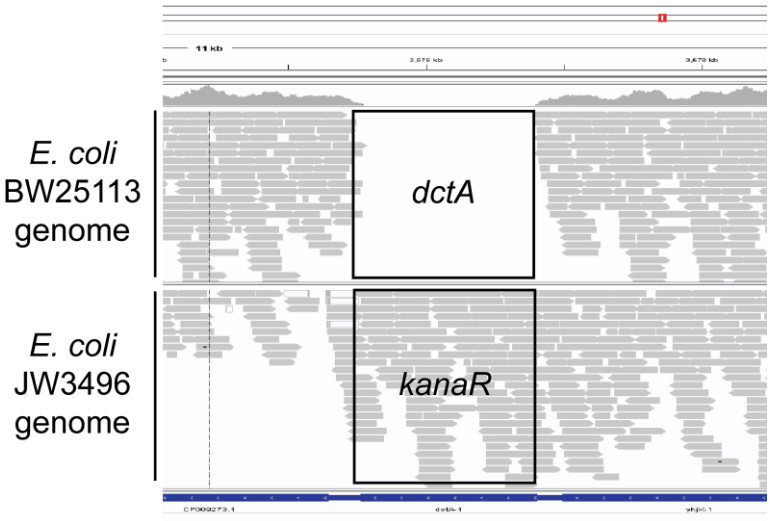


Figure 5.9: Whole genome sequencing of *E. coli* JW3496 (Δ*dctA*). Mapped reads are indicated as grey bars. The top panel shows the reads from the Δ*dctA* line mapped to the wild type *E. coli* BW25113 reference genome. No reads were mapped to the region where *dctA* should be located on the chromosome indicating that *dctA* was not present. Lower panel shows the reads mapped to the *in silico* generated *E. coli* BW25113 genome were *dctA* was replaced with the kanamycin expression cassette. Reads mapped continuously across the kanamycin gene as expected from a direct replacement of *dctA* with kanamycin resistance gene.

The Δ*dctA* cells harbouring an inducible expression construct containing either the *S. italica* or the *A. thaliana* version of pMDT1 were pre-grown in minimal media with glucose as sole carbon source (as Δ*dctA* cells cannot grow on malate). Next, cells

were washed and transferred to new minimal media with glucose as sole carbon source and the expression of the transporter genes was induced for 1 h to provide the cells enough time and resources to express the transporter proteins. This test was also conducted to confirm that the transporter was expressed and that pyruvate was being exported to the cell culture medium as expected (Figure 5.10 A). Subsequently, the cells were washed in minimal media without any carbon source to remove all traces of glucose and all pyruvate from the supernatant. The washed cells were then transferred to minimal media with malate as sole carbon source and the concentration of pyruvate in the supernatant was determined (Figure 5.10 B). *E. coli* Δ*dctA* cultures expressing either the *A. thaliana* or the *S. italica* version of pMDT1 both exported pyruvate to the cell culture medium, whereas the non-induced cell cultures did not (Figure 5.10 B). As the amount of pyruvate exported substantially exceeds the dry weight of the cells present in the cell culture medium the only plausible mechanism to explain the pyruvate accumulation is that the transporter also facilitates the influx of malate into the bacterial cell so that it is subsequently metabolised to pyruvate and exported back through the same transporter. To determine whether the transporter allowed any C_4 dicarboxylate into the cell an identical experiment was conducted this time with succinate as the sole carbon source. Feeding cells with succinate did not result in any detectable levels of pyruvate export and thus the transporter facilitates the import of certain dicarboxylates but not others.

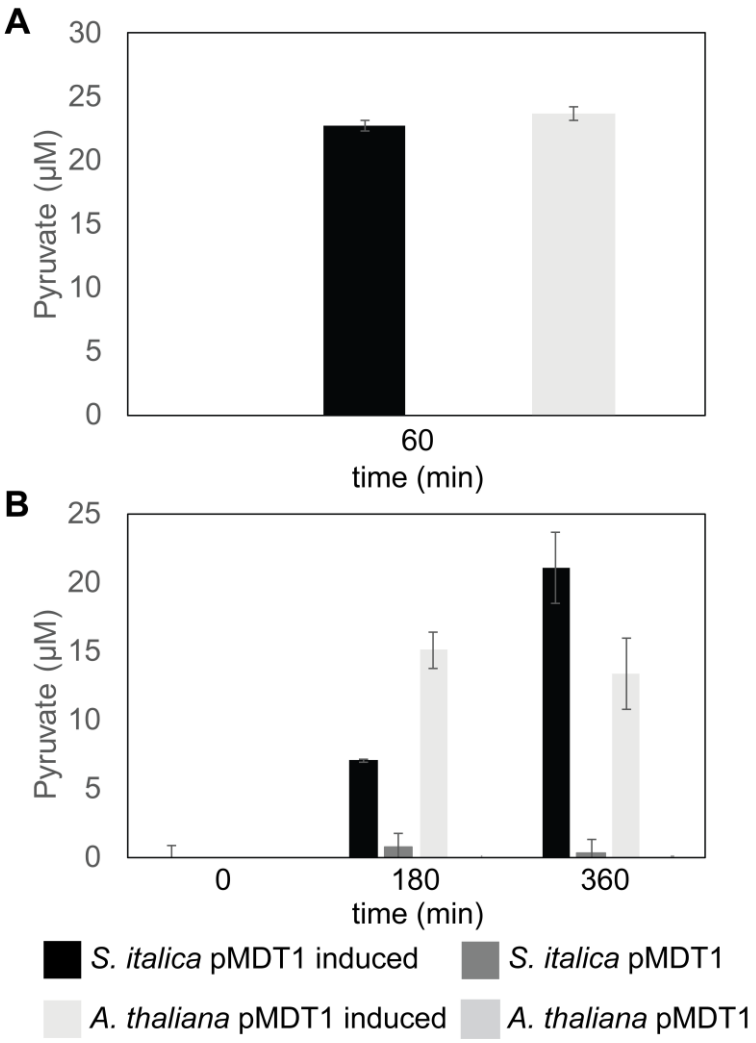


Figure 5.10: Pyruvate measurement in the supernatant of *E. coli* Δ*dctA* harbouring either the *S. italica* pMDT1 (black bars) or the *A. thaliana* pMDT1 (grey bars) pMDT1 variant grown in M9 malate **A)** Pyruvate concentration after 60 min in media with glucose as sole carbon source either non-induced or induced **B)** Pyruvate concentration measured in media with malate as sole carbon source at timepoint 0, 180, 360 min in non-induced and induced cells. Measurements were performed in triplicates and the error bars represent the standard deviation (N=3).

5.3.4 The putative pyruvate:malate transporter from *Setaria italica* localised to chloroplast envelope when expressed in *Setaria viridis* mesophyll protoplasts

The *A. thaliana* version of the pMDT1 protein has previously been shown to be found in the chloroplast envelop proteome of *A. thaliana* (Ferro et al. 2010). Therefore, it was expected that the *S. italica* ortholog should also similarly be localised to the chloroplast envelop. Moreover, both genes have strongly predicted chloroplast transit peptides. However, to confirm this localisation, a *C*-terminal GFP fusion of the full length pMDT1 protein was expressed transiently in isolated *Setaria viridis* mesophyll protoplasts.

As shown in Figure 5.11, the transporter-GFP fusion protein localised to punctate foci on the periphery of the chloroplasts and also to the plasma membrane. Such punctate foci at the periphery of chloroplasts is normal for a transmembrane transporter (Zhang et al. 2018), however the plasma membrane signal was not expected. The Seita.4G275500::GFP fusion protein was approximately twice the size (53.95 kDa) compared to the untagged Seita.4G275500 protein (27.05 kDa). Doubling the size of the protein may have interfered with the ability of the protein to fold correctly, or localise to its correct location within the cell. Moreover, as the protein is preferentially expressed in bundle sheath cells, it could be that the enforced expression in leaf mesophyll cells also caused an experimental artefact. Interestingly, AT4G19390::GFP expressed in *A. thaliana* mesophyll protoplasts resulted in a similar patchy fluorescent signal found in the chloroplast, but no signal was detected along the plasma membrane (Figure 5.12).

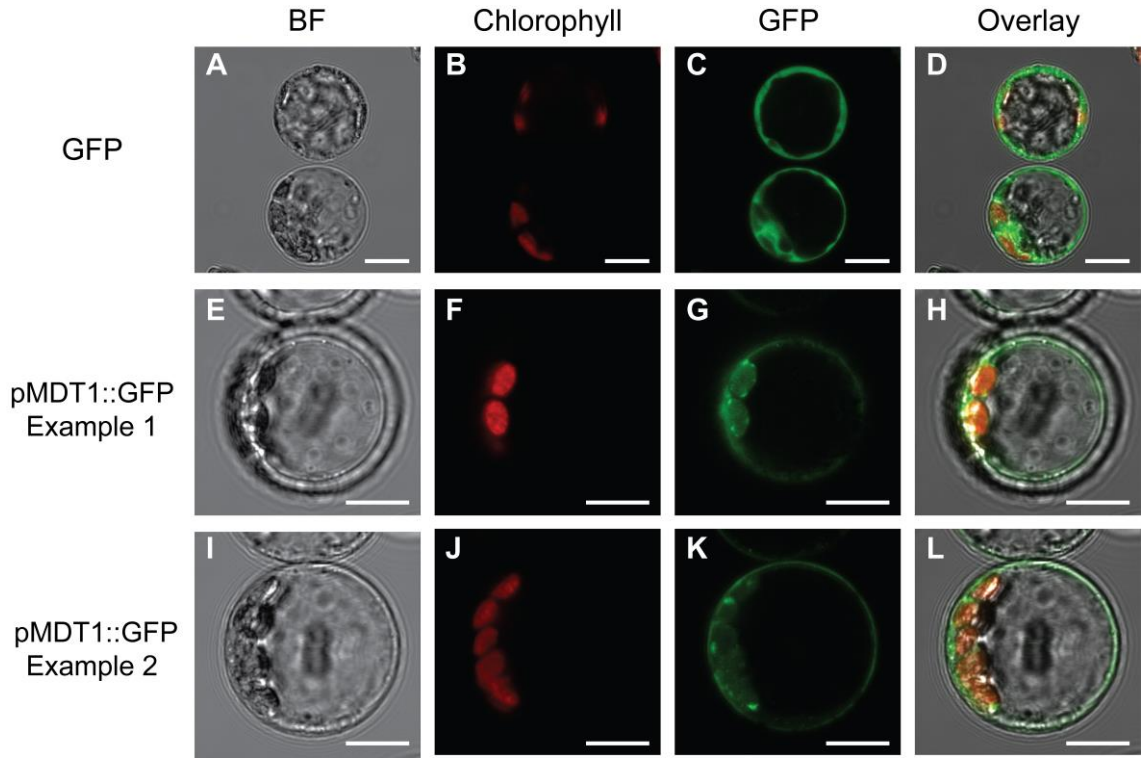


Figure 5.11: Sub-cellular localization of the *Setaria italica* pMDT1 variant fused to GFP (Seita.4G275500::GFP) in *S. viridis* protoplast. Representative results are shown for the cytosolic expressed GFP (**A-D**) and the *S. italica* pMDT1 variant (**E-L**). Images in the first column show the bright 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.

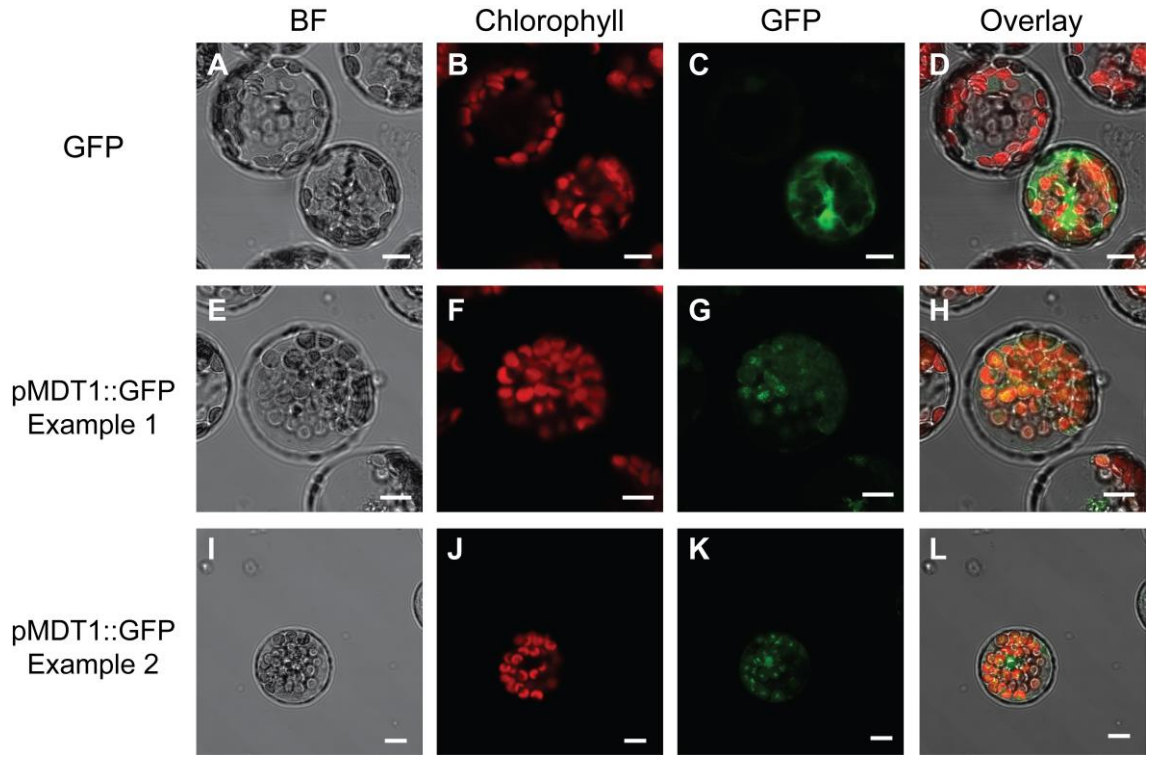


Figure 5.12: Sub-cellular localization of the *Arabidopsis thaliana* pMDT1 variant fused to GFP (AT4G19390::GFP) expressed in *A. thaliana* protoplast. Representative results are shown for the cytosolic expressed GFP (**A**-**D**) and the *A. thaliana* pMDT1 variant (**E**-**L**). Images in the first 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.

The results obtained from the subcellular localisation made us confident that both variants are localised to the chloroplast envelope, as was predicted by the presence of a chloroplast transit peptide and the presence of the protein in chloroplast envelope proteome of *A. thaliana*. However, these results did not contribute to our understanding whether the transporter participates in the C_4 cycle. To evaluate this an additional experiment was conducted to downregulate the abundance of the gene in the closely related C_4 model plant *Setaria viridis*.

5.3.5 Downregulation of the endogenous *Setaria viridis* pMDT1 reduces CO_2 assimilation

Setaria viridis, also known as wild foxtail millet, is a sub-species of *Setaria italica*. It has a smaller stature than *S. italica*, has a faster generation time and is readily amenable to genetic manipulation (Brutnell et al. 2010). Thus, *S. viridis* was chosen as the model species in which to conduct all transgenic experiments and investigate the function of pMDT1 in C_4 plants. Two distinct RNAi constructs were created to knock down the expression of the pMDT1 transporter in *S. viridis*, and 31 independent transformation events resistant to hygromycin were obtained using these constructs.

To determine the level of downregulation of pMDT1 transcripts caused by the RNAi construct, RT-qPCR was performed in the T_0 generation of the hygromycin positive *S. viridis* regenerating plants. The transcript levels of the lines harbouring an RNAi construct were compared to plants that were generated during the same transformation batch but did not express the RNAi fragment. The percentage reduction in pMDT1 transcript levels in the RNAi lines varied between 40% and 99% when compared to expression levels in control plants. Three independent RNAi lines (1_35, 1_31, 1_30) with a strong knock-down effect were selected from this set of regenerating plants for further analysis in the T_1 generation. All three lines were expressing the same RNAi construct (Figure 5.13).

The progeny of the three independent T_0 transformation events were genotyped, and for each independent line one hygromycin-positive plant (harbouring an RNAi construct) and a corresponding segregating wild type plant were selected. To confirm that the RNAi construct still achieved knock down of pMDT1 transcripts in the T_1 generation, transcript abundance was measured again by RT-qPCR. As in the T_0

generation, transcripts encoding pMDT1 were substantially reduced in the RNAi lines when compared to segregating azygous wild type sibling plants (Ø: 88% knockdown, p-value < 0.0001). This result indicated the RNAi construct was stable, and was still active in the T_1 generation causing a reduction in the transcript abundance (Figure 5.14 A).

To validate that the plants were still conducting C_4 photosynthesis, dried leaf material was collected from mature T_1 plants and wild type segregants and subject to dry matter carbon isotope analysis. On average a C_3 plant that fixes CO_2 via RuBisCO has a ratio of ^{13}C to ^{12}C between -33 to -24 ‰ in dried leaf material. In contrast, a plant that conducts C_4 photosynthesis first fixes CO_2 using the enzyme PEPC, this enzyme does not discriminate between ^{12}C and ^{13}C and this causes the carbon isotope ratio to be less skewed. On average, C_4 plants have a ratio between ^{13}C and ^{12}C of -16 to -10 ‰ (O'Leary 1988). The plants expressing the RNAi construct had the same carbon isotope signature as the segregating wild type plants, both of which are typical for a C_4 plant (Figure 5.14 B). Thus, the C_4 pathway was still operating in these RNAi plants.

Furthermore, the ratio of carbon to nitrogen and the number of tillers remained unaltered between the RNAi lines and the segregating azygous plants (Figure 5.14 C and D). Thus, RNAi did not cause the C_4 plants to revert to C_3 photosynthesis, nor did it affect leaf nitrogen content, nor did it cause the plants to grow obviously slower under greenhouse conditions (also Appendix A Figure A.6).

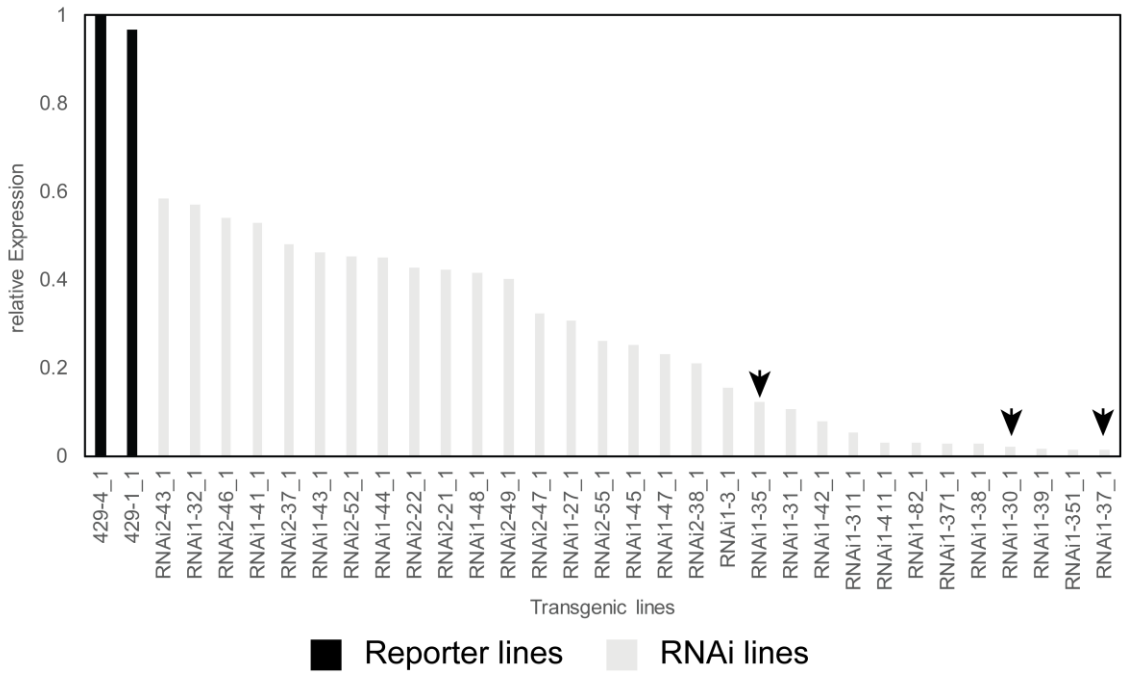


Figure 5.13: Screening the transcript abundance of the RNAi T_0 plants in comparison to parallel transformed reporter lines (black). Lines selected for further analysis in the T_1 generation were marked with a black arrow.

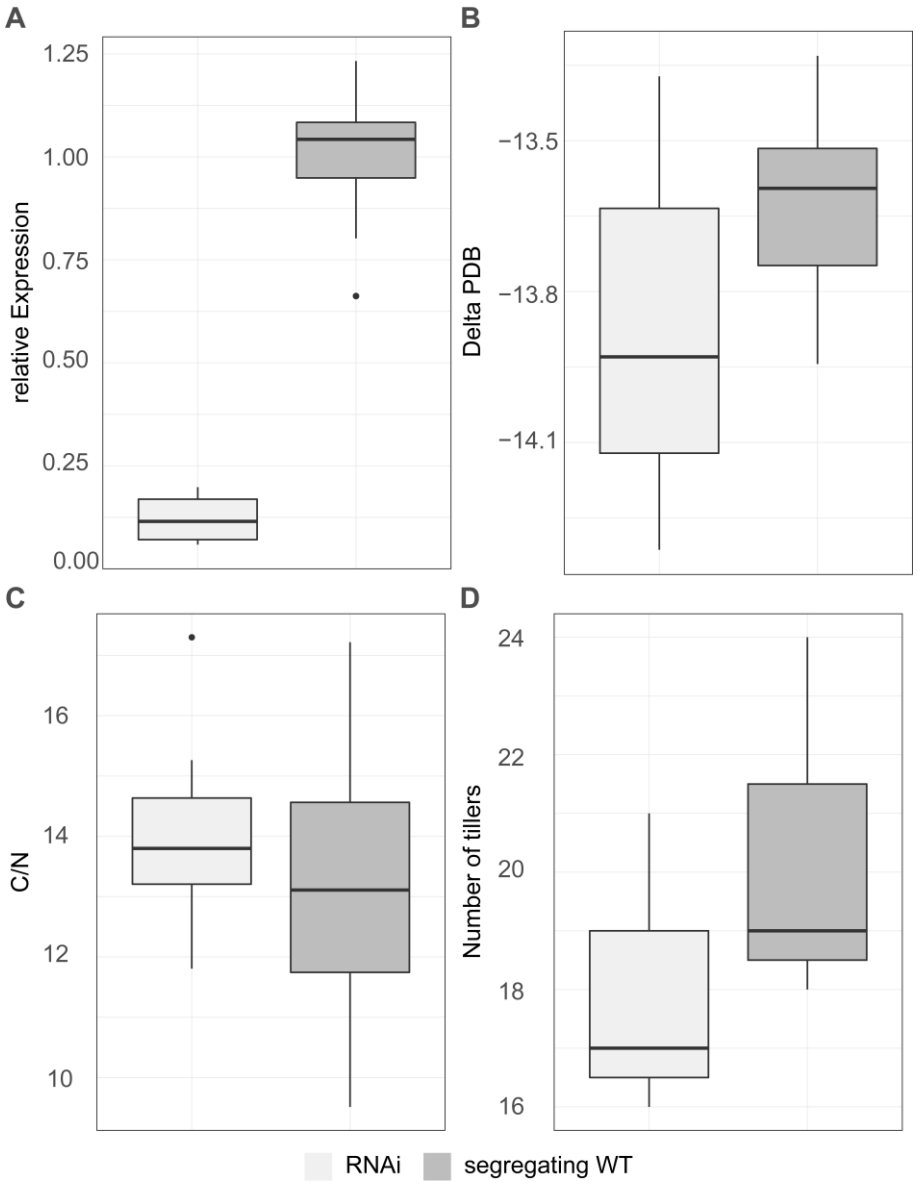


Figure 5.14: Boxplot depicting various measured characteristics of the RNAi lines and segregating wild type plants. **A)** Relative transcript abundance measured by RT-qPCR per plant three independent RNA extraction were performed **B)** $\delta^{13}C$ ratio calculated according to the Pee Dee Belemnite (PDB) standard. Per plant ten samples were taken across the whole plant **C)** carbon nitrogen ratio per plant ten samples were taken across the whole plant **D)** number of tillers counted for the RNAi lines and segregating wild type plants

To verify whether the downregulation of pMDT1 had an impact on the carbon assimilation (A) in the plants expressing the RNAi construct, the photosynthetic rate was measured using a LI-6800 at various carbon concentrations and then compared

to the segregating azygous sibling plants (Figure 5.15 A) (see Section 5.5.16). The maximum rate of carbon assimilation was reduced in plants expressing the RNAi construct compared to segregating wild type plants. However, the expression of the RNAi construct did not have a significant impact on the stomatal conductance (gsw) (Figure 5.15 B).

The carbon assimilation rate in C_4 plants has a clear characteristic in response to increasing CO_2 concentrations. The initial slope of the carbon assimilation rate is determined by the diffusion of atmospheric carbon and the activity of PEPC (V_{pmax}) (von Caemmerer and Furbank 2016). Due to the carbon concentrating mechanism, the carbon assimilation rate saturates between 200 and 400 µmol mol^{-1} in C_4 plants. At higher carbon concentrations, the saturated rates of carbon assimilation in C_4 plants are determined by the activity of RuBisCO (V_{cmax}) and the regeneration of PEP (von Caemmerer and Furbank 2016).

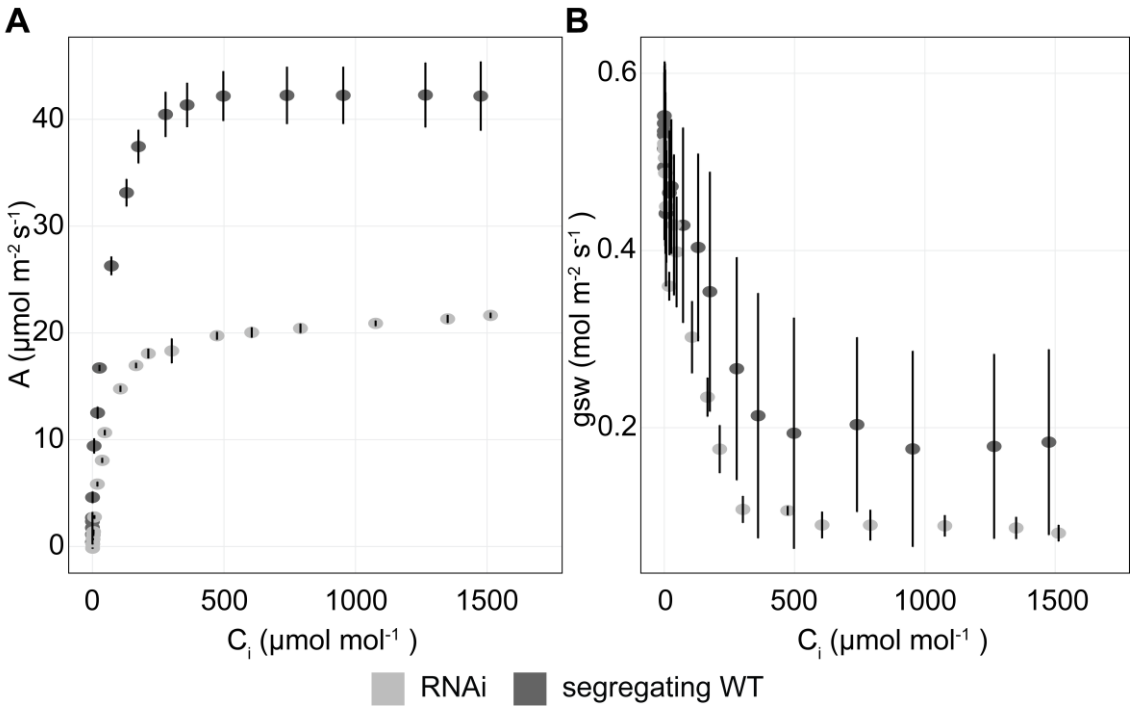


Figure 5.15: Characterisation of photosynthesis in the RNAi plants and the segregating wild type plants **A**) Carbon assimilation of RNAi and corresponding segregating wild type plants over a range of intercellular carbon (C_i). Plotted are the averages of three T_1 plants for the RNAi and segregating wild types. The error bars represent the associated standard deviation. **B**) Stomatal conductance (gsw) over the same range of C_i were measured concurrently as the A-C_i measurements. Each point represents the average of three independent T_1 plants RNAi and segregating wild types respectively. The error bars represent the standard deviation (N=3)

To identify the impact of the downregulation of pMDT1 the observed data of the segregating wild type and RNAi lines was fitted with the model of C_4 photosynthesis (Von Caemmerer 2000). The C_4 model enables the estimation of the maximum PEPC activity (V_{pmax}) and RuBisCO activity (V_{cmax}) (Figure 5.16). In the plant lines expressing the RNAi construct, the maximum PEPC activity was reduced (Ø: 68.92 µmol m^{-2} s^{-1}± 6.47 (SD, N = 3)) compared to the segregating wild type plants (Ø: 163.26 µmol m^{-2} s^{-1} ± 37.64 (SD, N = 3)) (Figure 5.16 A). This difference in the maximum PEPC activity was significantly different (p-value < 0.05). Moreover, the estimated maximum RuBisCO activity in the RNAi expressing plant lines was reduced as well (Ø: 21.68

µmol m^{-2} s^{-1} ± 0.54, (SD, N = 3)) compared to the segregating wild type plants (Ø: 37.01 µmol m^{-2} s^{-1} ± 2.58 (SD, N = 3) (Figure 5.16 B). The difference in the estimated maximum RuBisCO activity was significantly different (p-value < 0.01).

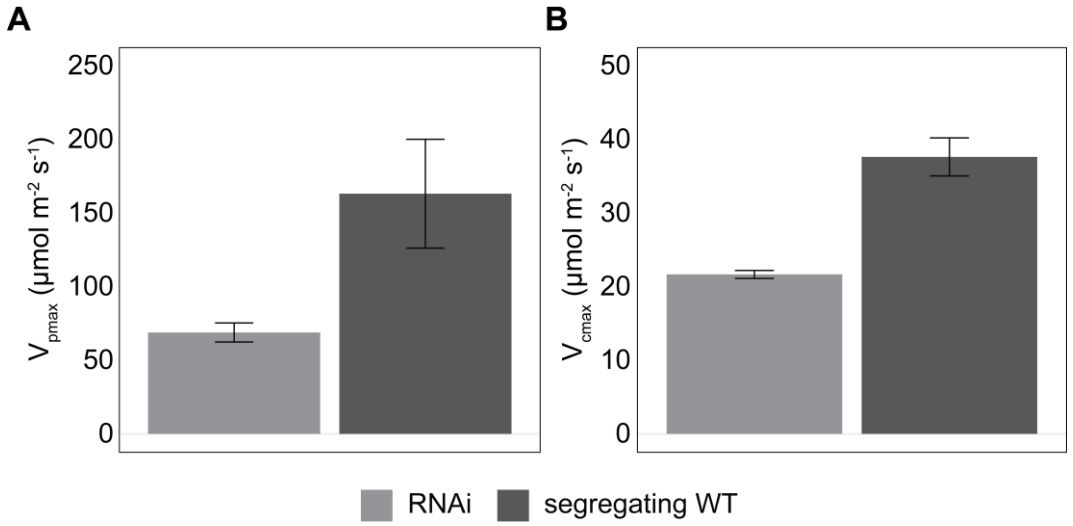


Figure 5.16: Estimating the maximum rate of the two carboxylating reactions in the C_4 cycle **A**) the maximum PEPC activity in the RNAi lines and the segregating wild type plants was estimated to be 68.92 µmol m^{-2} s^{-1} ± 6.47 (SD, N = 3), 163.26 µmol m^{-2} s^{-1} ± 37.64 (SD, N = 3), respectively (t-test, p-value ≤ 0.05). **B**) the maximum RuBisCO activity in the RNAi lines and segregating wild type plants was estimated to be 21.68 µmol m^{-2} s^{-1} ± 0.54 (SD, N = 3) and 37.01 µmol m^{-2} s^{-1} ± 2.58 (SD, N = 3), respectively (t-test, p-value ≤ 0.01).

The reduction of the maximum PEPC activity as well as the reduction of maximum RuBisCO activity are in line with the proposed function of pMDT1. The initial slope of an A-C_i curve is determined by the activity of PEPC and the delivery of CO_2. Thus, if there is less PEP available in the mesophyll cytoplasm caused by the reduction of pyruvate return flow, it makes intuitive sense that the maximum reaction activity of PEPC is reduced because of the diminished substrate availability. Subsequently, this has an impact on the amount of CO_2 fixed and therefore the number of molecules of

oxaloacetate and malate generated. Malate delivery to the bundle sheath chloroplast is interrupted twice. Firstly, the generation of malate caused by the reduced flux through the C_4 cycle, and secondly by the reduction in transport into the bundle sheath chloroplasts due to reduced expression of pMDT1. As a consequence, there is less substrate (CO_2) around RuBisCO, which is reflected in the reduction of both A_{max} and V_{cmax} in the RNAi plants when compared to the segregating azygous sibling plants.

Although there was a difference in carbon assimilation between the RNAi plants and the segregating azygous sibling plants, there was no obvious growth defect. Thus, a light response curve was also conducted. The light response curve revealed that at lower light concentration (comparable to the greenhouse conditions ~ 200-400 µmol m^{-2} s^{-1}) there was little or no difference in assimilation rate. However, at the higher light intensities the maximum assimilation rate in the RNAi lines was about 70% compared to the control plants (Figure 5.17 B). Fitting a non-linear regression that was accounting for the two groups revealed that the light response between the RNAi plants and the segregating azygous siblings was significantly different (p-value ≤ 0.001).

During the light response curve, it was noted that the RNAi plants took significantly longer to adapt to higher light concentrations. To illustrate this defect leaves were illuminated at 300 µmol m^{-2} s^{-1} for 10 minutes and then stepwise increases in PPFD were made and the extent of increase in assimilation rate was recorded after one minute. In response to the increasing light intensities, the RNAi plants were unable to increase the assimilation rate as fast as the segregating azygous sibling plants (Figure 5.17 A). A linear model accounting for the different genotypes revealed that the increase in assimilation in response to the light quantities was significantly different (p-value ≤ 0.01). However, it was very surprising to find that the maximum assimilation rate between the A-C_i and the light response curve revealed that much of a

discrepancy. One potential explanation might be that the plants used for the light response curve were a different batch than the plants used for the A-C_i measurement that were grown in summer while the plants used for the A-Ci curves were grown in winter/spring.

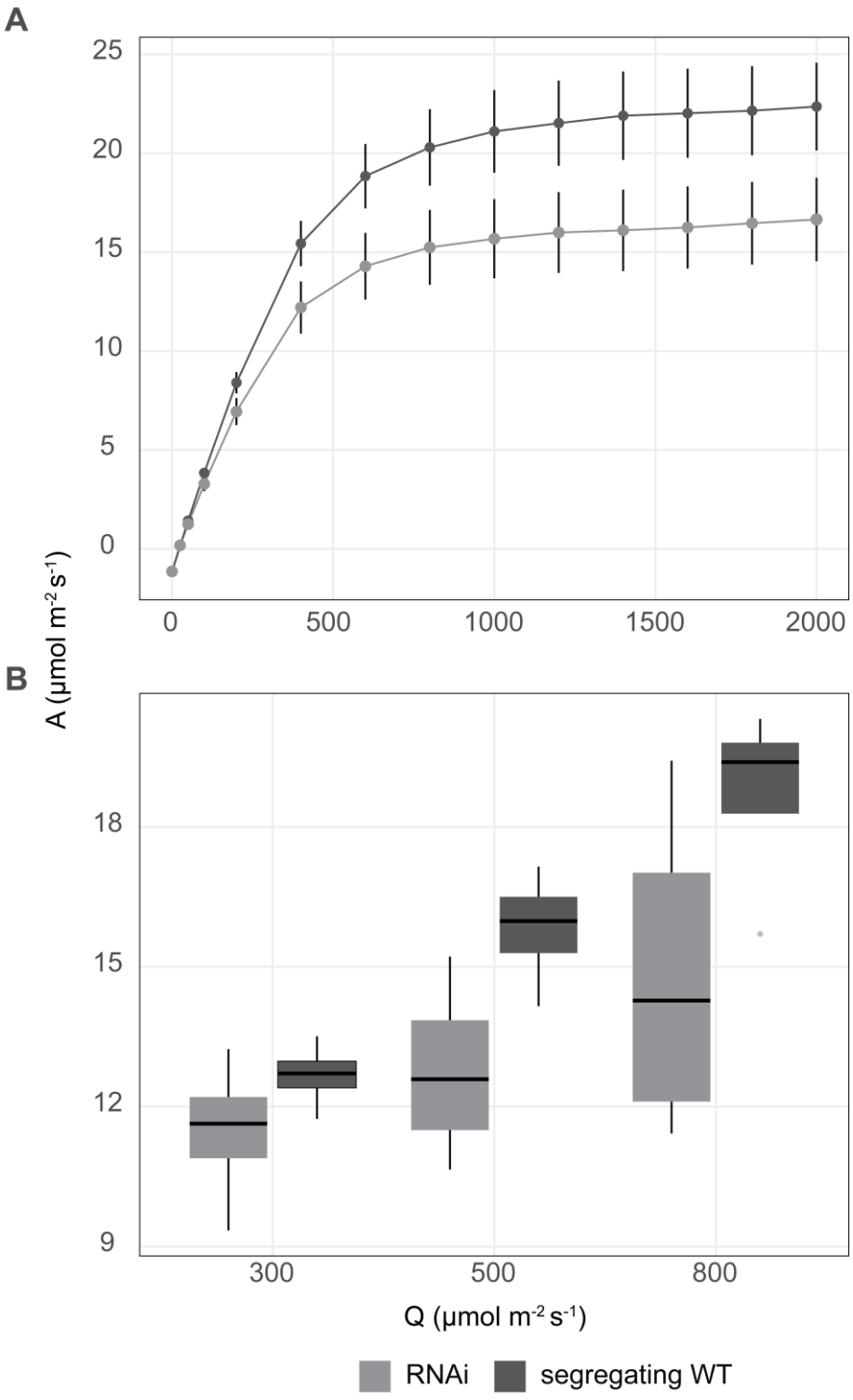


Figure 5.17: Response to increasing amount of light. **A)** Light response curve measuring the assimilation rate at 2000, 1800, 1600, 1400, 1200, 100, 800, 600, 400, 200, 100, 50, 0 µmol m^{-2} s^{-1}. Error bars indicate the standard error (N=5) **B)** Boxplot of the assimilation rate measured 1 minute after increasing the light intensity from 300 to 500 µmol m^{-2} s^{-1} and from 500 to 800 µmol m^{-2} s^{-1} (N=4).

To verify whether the downregulation of pMDT1 in the *S. viridis* plants altered the expression of other genes, RNAseq analysis was performed. The RNAseq analysis compared the transcript abundance in whole leaves of the RNAi plants with the segregating wild type plants (List of differentially expressed genes in Appendix A.9). The RNAseq analysis revealed that in total, 97 genes displayed an altered expression pattern ($p\text{-value}_{adj} \leq 0.05$) in response to the perturbation caused by pMDT1. Among these 97 genes with altered expression patterns was pMDT1 ($p\text{-value}_{adj} = 1.03 \times 10^{-6}$) confirming the findings obtained with RT-qPCR (Figure 5.14 A). To determine whether the differentially expressed genes have a cell-type specific expression pattern, we identified the differentially expressed genes in the *S. viridis* cell-type specific expression dataset by John et al. 2014 (John et al. 2014). Out of the 97 differentially expressed genes, 38 and 36 were identified to be preferentially expressed in the mesophyll cells and bundle-sheath cells, respectively. The remaining 23 differentially expressed genes in the RNAi lines did not display a significant preference for either of the two cell types. A subsequent hypergeometric test revealed that the set of differentially expressed genes is enriched 1.64-fold ($p\text{-value} = 8.5 \times 10^{-4}$) for bundle sheath specific expressed genes and 1.84-fold ($p\text{-value} = 4.6 \times 10^{-5}$) for mesophyll specific expressed genes (Figure 5.18). Thus the set of genes whose mRNA abundance changed in response to downregulation of the bundle sheath chloroplast malate:pyruvate transporter was enriched for those that preferentially accumulated in bundle sheath or mesophyll cells.

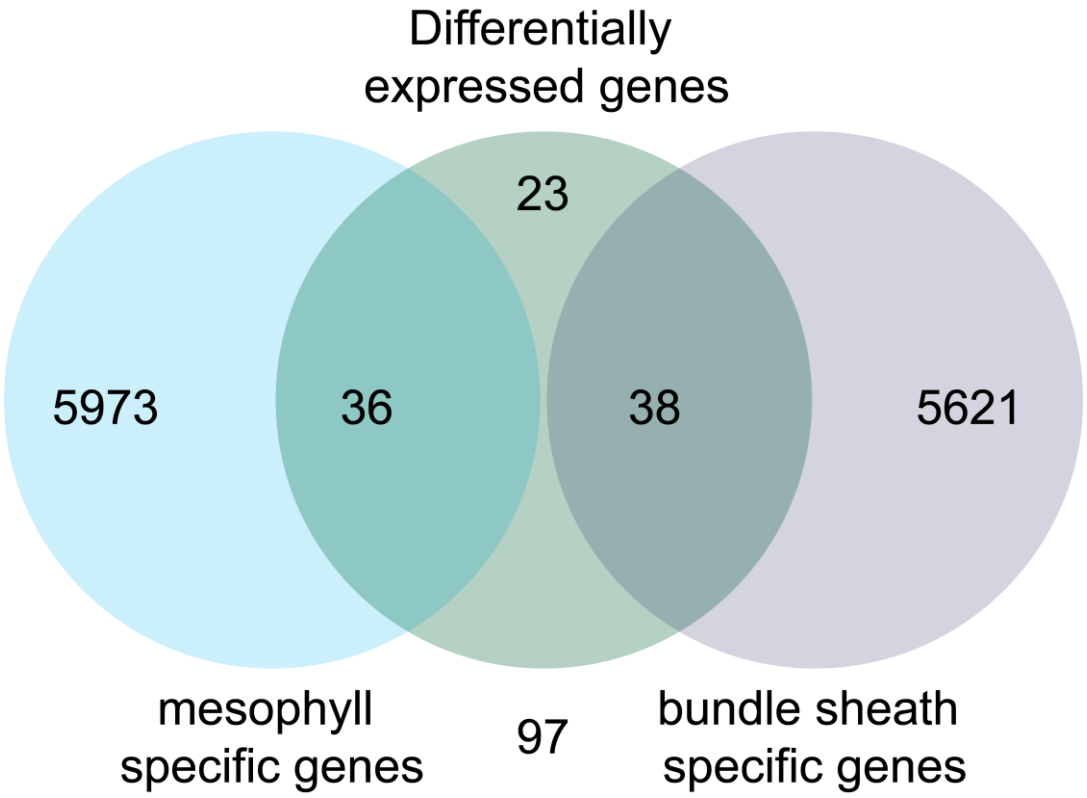


Figure 5.18: Venn diagram representing the set of cell-type specific expressed genes. Out of the 97 differentially expressed genes, 36 are known to be among the set of genes that are preferentially expressed in the mesophyll cells. Similarly, 38 out of the 97 differentially expressed genes were identified to belong to the set of genes that revealed a preferential expression in the bundle sheath cells.

The protein sequence of the set of differentially expressed genes was used to predict a targeting peptide potentially present at the *N*-terminus. Out of the 97 differentially expressed genes, 35 did not have any targeting peptides, 31 were marked as having a chloroplast transit peptide, 17 gene products were predicted to have a signal peptide for the secretory pathway and 14 were predicted to be targeted to the mitochondria. Interestingly, among the chloroplast targeted proteins, 15 had a lower expression whereas 16 had an increased expression in response to the perturbation caused by the RNAi construct. Thus, there was no general upregulation of chloroplast protein in response to the downregulation of pMDT1. Moreover, the paralogous gene of pMDT1

(Sevir.2G427600 also see Figure 5.4) did not reveal an altered expression between the RNAi lines and the segregating wild type. The average transcript per million values obtained were about 160 and 197 for the RNAi lines and segregating wild types, respectively. Thus, there was no obvious compensation for the loss of the transporter through the upregulation of paralogous genes.

The assessment of the functional characteristics of the differentially expressed genes revealed some surprising findings. The differentially expressed genes were ordered according to the absolute $\log_2$(Fold-change). The most upregulated gene is an uncharacterised gene whose orthologs are predicted to be involved in thiamine biosynthesis. Interestingly, among the top 12 most upregulated genes, two additional genes are also associated with thiamine (Sevir.9G138500) and thiazole biosynthesis (Sevir.4G208400). Thiamine is required as a cofactor during glycolysis, the pentose phosphate pathway and the TCA cycle (Goyer 2010). Thiamine is also required in the response to abiotic and biotic stress (Goyer 2010). In plant chloroplasts, thiamine diphosphate serves as a cofactor for the pyruvate dehydrogenase and pyruvate decarboxylase (Tovar-Méndez et al. 2003). Pyruvate dehydrogenase yields acetyl-coenzyme A and NADH whereas pyruvate decarboxylase yields one molecule of CO_2 and one molecule of acetaldehyde (Goyer 2010; Rioval et al. 1990). Thus, although the pyruvate dehydrogenase and pyruvate carboxylase enzymes did not change in abundance themselves, the increase in thiamine biosynthesis could be suggestive of an increase flux through these enzymes. Though this is only speculation.

The THI4 thiazole synthase enzyme (Sevir.4G208400) that was upregulated in the RNAi lines is an unusual enzyme because it participates in the reaction itself rather than catalysing it (Chatterjee et al. 2011). After every reaction the protein has to be degraded and resynthesized (Chatterjee et al. 2011). The estimated energy required

for THI4 is between 2-10% of the total maintenance respiration (Amthor et al. 2019). Thus, replacing a non-suicidal THI4 in crop plants is predicted to improve crop yields (Amthor et al. 2019). Taken together, it is possible that the increase in thiamine/thiazole biosynthesis genes is a signature of an increase in respiration of C_4 cycle intermediates.

5.4 Chapter conclusion

In this Chapter, I presented the characterisation of a putative monocarboxylate and/or dicarboxylate transporter (pMDT1) of NADP-ME C_4 plants (Seita.4G275500.1) and the orthologous gene in the C_3 model organism *Arabidopsis thaliana* (AT4G19390). The gene was chosen for characterisation through a bioinformatic screen that aimed to identify potential bundle sheath chloroplast malate and/or pyruvate transporters. The initial characterisation was performed in *Escherichia coli* liquid cultures in minimal media. These experiments revealed that both *S. italica* and *A. thaliana* protein variants facilitated the export of pyruvate from the *E. coli* cell into the cell culture medium. Subsequently, experiments also confirmed that the transporters also facilitated the influx of malate into a bacterial cell while simultaneously exporting pyruvate. Thus, there was experimental evidence that the chosen candidate transporter gene facilitates the influx of the dicarboxylate malate and efflux of the monocarboxylate pyruvate.

A subcellular localization study was performed using *S. viridis* and *A. thaliana* protoplasts to verify whether the putative transporter proteins are targeted to the chloroplast. The subcellular localization in the protoplast showed that the protein variants are indeed directed to the chloroplast envelope, in accordance with their predicted chloroplast transit peptides and the proteomic evidence obtained from isolated chloroplast envelope proteomes in *A. thaliana*. Thus, the experimentally gathered information convinced me to perform the experimental evaluation *in planta* to

verify what the biological role of that transporter is and whether it is associated with the C_4 cycle.

To evaluate whether pMDT1 is part of the C_4 cycle, I created an RNAi construct to downregulate the abundance of pMDT1 in the C_4 model plant *S. viridis*. I chose to do RNAi as *a priori* one might expect that a knockout of this transporter would be lethal or very sick if it was unable to conduct C_4 photosynthesis. By using RNAi, it would be possible to generate mutants that still had some transporter expression and thus should be able to still conduct C_4 photosynthesis, albeit at a reduced rate. The genetically engineered plants showed stable and heritable reduction in transcript abundance of pMDT1 in both the T_0 and T_1 generation. However, much to my surprise the plants looked absolutely fine in the greenhouse with no obvious greening or growth defect.

To assess what the impact of the reduced abundance of pMDT1 is on the C_4 cycle, the photosynthetic rate was measured for segregating azygous wild type plants and plants expressing the RNAi construct. The carbon assimilation rate in the RNAi plants was substantially reduced compared to the segregating azygous wild type sibling plants. The observed reduction in carbon assimilation rate was consistent with the knockdown of the transporter and a reduced flux through the C_4 cycle. The extent of this reduction in assimilation rate was very surprising given that the plants had no obvious growth defect. A subsequent light response measurement revealed that at lower light intensities there was no difference in assimilation rate between the RNAi lines and the segregating azygous sibling plants.

The characterisation of the bundle sheath chloroplast malate and pyruvate transporter provides the final missing piece of the C_4 cycle in *Setaria italica*. In doing so, this provides the first complete C_4 cycle in NADP-ME plants (Figure 5.19). For the first time, all enzyme and transporters required to facilitate the flux of metabolites through the mesophyll and bundle sheath cells are known in the C_4 NADP-ME subtype. However, this finding is limited to plants that express a pMDT1 and it appears that neither *Sorghum bicolor* nor *Zea mays* express this transporter in their leaves. Thus, alternative malate and pyruvate transport proteins must also be used in other C_4 lineages. The experimental strategy used in this chapter could be used to identify the missing transporters of other C_4 cycles.

Although I have shown that pMDT1 facilitates the export of pyruvate and import of malate, that it localises to the chloroplast envelope, and that knockdown causes a reduction in flux through the C_4 cycle, the exact mechanism how the transport step takes place remains elusive. It could be that pMDT1 functions as a selective channel that facilitates bidirectional movement of specific metabolites along a concentration gradient, or it could be an antiporter that facilitates a direct exchange of malate for pyruvate. In the case of the latter the antiport reaction would likely be [malate + H^+ <=> pyruvate], whereby the H^+ offsets the negative charge imbalance caused by counter exchanging a monocarboxylate for a dicarboxylate. Here, both pyruvate and malate are deprotonated at the physiological pH of the cytosol or chloroplast stroma, and thus both exists as carboxylate anions wherein pyruvate has a single negatively charged carboxylate group while malate has two. If the protein acts as a channel, then other endogenous transporters would have to mediate H^+ efflux to maintain charge balance of the cell. Experiments with purified protein reconstituted in liposomes are currently underway through collaboration with Prof. Simon Newstead in the Department of

Biochemistry at the University of Oxford. At time of writing, pMDT1 protein has been produced and purified but not yet assessed for transporter mechanism. A provisional patent application has been filed (P0010239AU) for the use of this transporter to engineer a C_4 cycle into a C_3 plant.

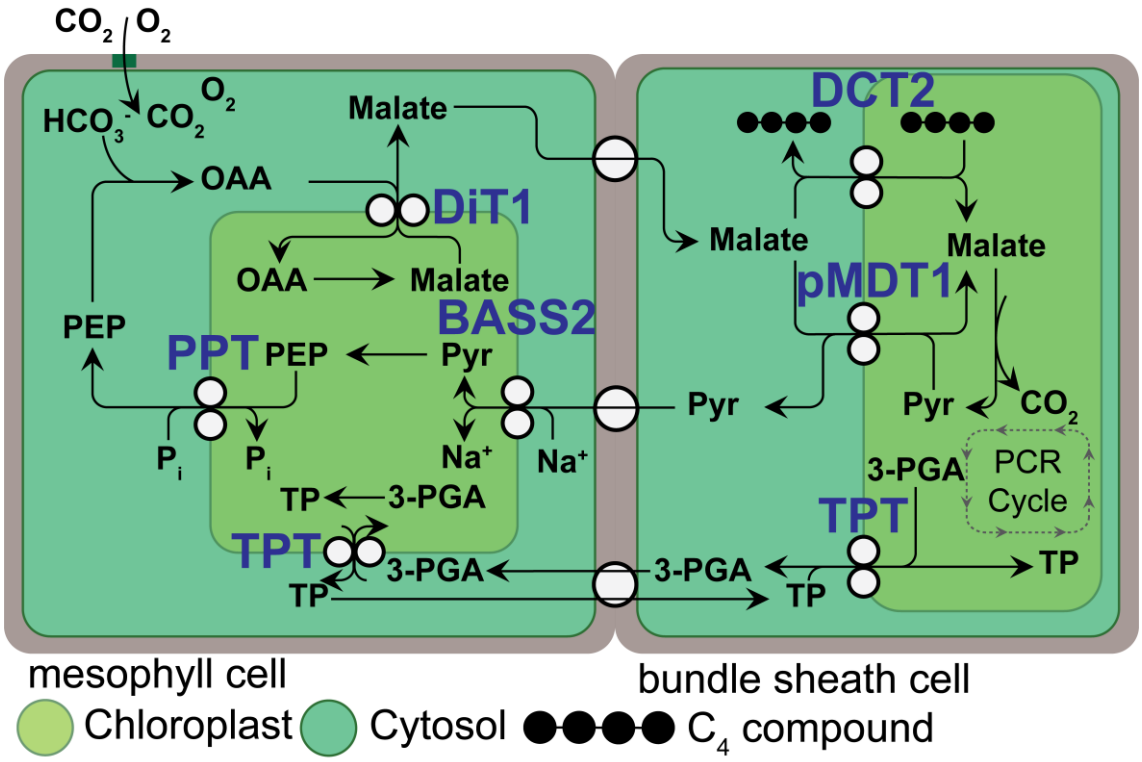


Figure 5.19: A new model describing the flux through the C_4 cycle of the NADP-ME subtype. This model is valid for C_4 NADP-ME plants expressing pMDT1. Abbreviations of the transporters: PEP/phosphate translocator (PPT), dicarboxylate transporter 1 (Dit1), dicarboxylic acid transporter 2 (DCT2), bile acid:sodium symporter family protein 2 (sodium-dependent pyruvate transporter) (BASS2), triose phosphate translocator (TPT). Abbreviations of the metabolites: ribulose 1,5-bisphosphate (RuBP), triose phosphate (TP), ortho-phosphate (P_i), phosphoenolpyruvate (PEP) oxaloacetate (OAA), pyruvate (Pyr), 3-phosphoglyceric acid (3-PGA), Sodium (Na^+)

5.5 Material and Methods

5.5.1 Plant growth conditions

T_1 seeds were incubated in moist moss (Spagmoss) at 4°C for three weeks to promote germination. After three weeks the seeds were dried at 37°C overnight. The following day the seeds were germinated on wet tissue paper in a growth incubator at 30°C with 16 h photoperiod. After 2-3 days the germinated seedlings were transferred to pots containing soil (Levington Advance Pot and Bedding M2 compost) and transferred to the greenhouse facility 28°C / 20°C (day/night) with long photoperiod (16 hr light and 8 hr dark). The plants were watered daily and treated against spider mites with 2% Savona.

5.5.2 Generating transgenic *Setaria viridis* ME034V

The confirmed pANIC12A plasmids harbouring either the RNAi_1 or the RNAi_2 were transformed into *Setaria viridis* ecotype ME034V by Julia Lambret-Frotte and Georgia Smith. To generate transgenic *S. virids* an *Agrobacterium tumefaciens*-mediated transformation protocol was used. In short, Stable transformation of *S. viridis* was carried out as described by Joyce van Eck (Version January 2014, unpublished). To improve the germination the seed coat was removed. Subsequently, the seeds were surface sterilized using a mixture of bleach (10%, v/v) and Tween 20 (0.1%, v/v) for 3 min. For callus induction the seeds were placed embryo facing upwards on CIM (4.3 g l^{-1} Murashige and Skoog salts, 10 ml l^{-1} 100×MS vitamin stock, 40 g l^{-1} maltose, 35 mg l^{-1} $ZnSO_4 \times 7H_2O$, 0.6 mg l^{-1} $CuSO_4 \times 5H_2O$, pH 5.8, 4 g l^{-1} Gelzan, after autoclaving 0.5 mg l^{-1} kinetin and 2 mg l^{-1} 2,4-D were added) and incubated for 3-4 weeks in the dark at 24°C. Calli structures were removed from soft tissue and transferred to fresh CIM and incubated for 2 weeks in the dark. The callus was

multiplied by dividing it into pieces of 2-mm diameter and plated on fresh CIM. The transformation was carried out 7 days later.

Agrobacterium tumefaciens AGL1 harbouring the construct of interest (Section 5.5.5) were grown in the presence of 50 µg ml^{-1} kanamycin and 50 µg ml^{-1} rifampicin at 28°C until an OD_{600} of 0.6 was reached. The cells were collected and resuspended in 50 ml of liquid CIM containing acetosyringone (200 mM) and synperonic (0.01%, w/v). The calli were transferred into the *Agrobacterium* suspension and incubated for 5 min at room temperature. Subsequently, the calli were removed from the solution and dried on sterile filter paper and transferred to selective CIM plates containing hygromycin 40 µg ml^{-1} and 150 µg ml^{-1} timetin and incubated in the dark at 24°C for 16 days. Next, calli were transferred to selective plant regeneration media (4.3 g l^{-1} Murashige and Skoog salts, 10 ml l^{-1} 100×MS vitamins, 20 g l^{-1} sucrose, pH 5.8, 7 g l^{-1} Phytoblend, after autoclaving 2 mg l^{-1} kinetin, 150 µg ml^{-1} timetin and 15 µg ml^{-1} hygromycin) and incubated at 24°C under long photoperiod (16 hr light and 8 hr dark).

Emerging shoots were transferred to rooting media (2.15 g l^{-1} Murashige and Skoog salts, 10 ml l^{-1} 100×MS vitamins, 30 g l^{-1} sucrose, pH 5.7, 7 g l^{-1} Phytoblend, after autoclaving 150 mg l^{-1} timetin and 20 µg ml^{-1} hygromycin). Shoots that developed root were genotyped (Section 5.5.3) and positive plants were transferred to soil (Section 5.5.1).

5.5.3 DNA extraction and genotyping of transgenic plants

Leaf tissue (1-2 cm^2) was collected and placed into a 1.5 ml Eppendorf tube containing a steel ball. The tube containing the leaf was flash frozen in liquid nitrogen. The frozen tissue was disrupted using a TissueLyser (Qiagen) while keeping the samples frozen. Next, 400 µl DNA extraction Buffer (200 mM Tris pH 7.5, 250 mM NaCl, 25 mM EDTA,

0.5% SDS) was added to each tube and thoroughly mixed. The tubes were centrifuged at 15000 rpm for 10 min and 350 µl of the supernatant was transferred into a fresh Eppendorf tube containing 350 µl isopropanol. The solution was mixed and incubated at room temperature for 10 min to precipitate the DNA. The DNA was collected by centrifugation at 15000 rpm for 10 min and the supernatant was discarded. The DNA pellet was washed with 600 µl 70% ethanol and allowed to air dry. Subsequently, the DNA pellet was resuspended in 100 µl of H_2O.

To genotype the plants the primer pair (forward and reverse) 5'-caaccaagctctgatagagt-3', 5'-gaagaatctcgtgctttca-3', were used to amplify the hygromycin resistance gene with an expected fragment size of 725 bp and the primer pair (forward and reverse) 5'-cagcaagccgcctatatggag-3', 5'-tcgtctcaggagtggccaagt-3' were used to amplify Sevir.2G209800 (C_2H_2-type zinc finger) as a positive control for the DNA extraction with an expected product size of 539 bp. The reaction was set-up as previously the number of cycles was 35 and the annealing temperature was set to 58°C.

5.5.4 Sequence modifications

The coding sequences for the two gene variants AT4G19390, Seita.4G275500 were obtained from Phytozome (Goodstein et al. 2012). For the experimental work carried out in protoplast the native gene sequence without the introns was used. However, the stop codon was removed to enable *C*-terminal tagging with GFP. Furthermore, the sequences were extended with the golden gate SC1 adapters for subsequent cloning into a level 1 vector (Engler et al. 2008). The sequences used for the experiments in *Escherichia coli* were codon optimized using OPTIMIZER (Puigbò et al. 2007) and extended with the golden gate SC module for subsequent cloning into a level 1 vector. Furthermore, all BsaI sites where removed by synonymous codon replacement to

avoid any interference with the subsequent cloning steps. The sequences were synthesized by Gene Universal and delivered in a pUC57 cloning vector. The full sequences submitted to Gene Universal can be found in the Appendix A.3.3 to A.3.6 .

Golden gate reactions were carryout out as described previously (in Section 4.4.6) and the components for the plasmid construct used in protoplast experiments can be found in Appendix A.3.12. The components for the plasmid used for the bacterial experiments can be found in the Appendix A.3.13.

5.5.5 Gateway cloning to create pMDT1 RNAi constructs

Two RNAi constructs were created using gateway cloning (Hartley et al. 2000). Therefore, two primer pairs were designed containing the required attB sites. The first primer pair was designed to create a RNAi construct targeting the 5'-end of the gene whereas the second primer pair was designed to target the 3'-end of the gene (Table 5.1). A PCR reaction was performed to obtained fragments with the primers. As template the plasmid harbouring the native *Setaria italica* pMDT1 was used (Appendix A.3.4) The PCR reactions and the temperature time settings was carried out as described in Section 4.4.3 with the adjustment of the elongation time to 30 s. The PCR product was purified as described in Section 4.4.2.

To carry out the BP reaction the Gateway BP Clonase II kit (Invitrogen) was used according to the included instructions. In short, 150 ng of purified PCR fragment were mixed with 150 ng pDONOR 207 and 2 µl 5x BP Clonase reaction buffer. Next, the volume was adjusted to 8 µl with TE buffer (pH 8). To the mixture 2 µl BP Clonase II enzyme mix was added, mixed and incubated at 25°C for 1 h. After the incubation step 1 µl of proteinase K solution was added to each sample and incubated at 37°C for 10 min to terminate the reaction. The reaction was transformed into heat shock competent

E. coli DH5α and grown on LB plates containing gentamycin 12.5 mg l^{-1}. About five colonies were picked and grown in liquid culture overnight. Subsequently, the plasmid was purified and evaluated through PCR using the primers in Table 5.1 without the Gateway adapters.

Table 5.1: Primers containing the attB sites to amplify a fragment of the *Setaria italica* gene. The underlined part of the primer sequence is specific to pMDT1. The position in the site column revealed the target position on the Seita.4G275500 gene sequence.

ID	Primers with attB Gateway adapters (5'->3'):	Site
RNAi_1	GGGGACAAGTTTGTACAAAAAAGCAGGCTTTATGAAGCTCAGGCCTCTCACT GGGGACCACTTTGTACAAGAAAGCTGGGTAACAGCTCATACAAGCCCGT	1-461
RNAi_2	GGGGACAAGTTTGTACAAAAAAGCAGGCTTT CTTCATGACCTTCCTCGGCA GGGGACCACTTTGTACAAGAAAGCTGGGT TTGATGCAGCGAAGCAAAGG	258-654

To carry out the LR reaction the Gateway LR Clonase II kit (Invitrogen) was used and performed according to the included manual. In short, 150 ng of pDONOR 207 harbouring the RNAi gene fragment was added to 150 ng gateway-compatible destination vector pANIC 12A (Mann et al. 2012) and 2 µl 5x LR Clonase reaction buffer. The reaction volume was adjusted to 8 µl with TE buffer (pH 8.0) and 2 µl of the LR Clonase enzyme mix was added to the solution. The reaction was incubated at 25°C for 1 h and the reaction was terminated as described above in the BP reaction. The LR reaction was transformed into heat shock competent *E. coli* DH5α and selected in LB plate containing kanamycin 50 mg l^{-1}. Again, a handful of colonies were picked and grown overnight in liquid LB and the plasmid was purified using a miniprep Kit (Qiagen). The purified plasmid was analysed using Sanger sequencing using the primers in the Appendix Table A.16

5.5.6 Expression of pMDT1 in *E. coli* grown in minimal media with glucose as sole carbon source

Three independent colonies of *Escherichia coli* BW25113 were selected either harbouring the *A. thaliana* or the *S. italica* version of pMDT1 on an inducible vector construct (see Section A.3.13). The pre-growth was carried out in LB and M9 glucose as described previously in Section 4.4.7. To test the transport characteristics of the two pMDT1 variants the cultures were inoculated at an optical density (OD_{600}) of 0.1. The growth was monitored by measuring the OD_{600} and the expression of pMDT1 was induced with 50 µM 2, 4-Diacetylphloroglucinol (DAPG) once the cells reached an OD_{600} of ~0.2. Samples of the supernatant were taken every hour. Cell suspension was collected in 1.5 ml Eppendorf tube and centrifuged for 5 min at 4°C at 20'000 g. Subsequently, the supernatant was sterile filtered using a 0.22 µm filter (Whatman Puradisc) and added into a fresh 1.5 ml Eppendorf tube. The supernatant was flash frozen in liquid nitrogen and subsequently stored at -80°C until the day of pyruvate measurement. The same supernatant was used to extract the organic acids (see Section 5.5.8) for metabolite screen.

5.5.7 Pyruvate measurement in cell culture supernatant

The concentration of pyruvate in the culture supernatant was assessed using a colorimetric assay (abcam ab65342) following the manufacturer's instructions. The assay relies on the oxidation of pyruvate by pyruvate oxidase that generates colour that is detectable at λ = 570 nm. The pyruvate standards (0, 2, 4, 6, 8, 10 nmol $well^{-1}$) were prepared in duplicates and prepared fresh. The supernatant samples were prepared by mixing 50 µl of supernatant with 50 µl of reaction mix (containing, 46 µl pyruvate assay buffer, 2 µl pyruvate probe and 2 µl enzyme mix) in a 96-well plate.

After an incubation at room temperature for 30 min in the dark, the OD at 570 nm was measured using a plate reader (FLUOstar Omega, BMG Labtech).

5.5.8 Metabolite extraction from cell culture supernatant

The organic acids were extracted as described by Lunn et al. 2006 (Lunn et al. 2006) with minor modifications. The procedure is described below. The day before the extraction the Eppendorf tubes containing the frozen cell culture supernatant samples were transferred to the -20°C freezer and the next morning the tubes were thawed on ice. 20 µl of the supernatant was added to 350 µl of ice-cold chloroform methanol (3:7, v/v) and mixed on a e.g. vortex. Next, the tubes were placed in the -20°C freezer for 2 h with occasional mixing. After the incubation period 350 µl of ice-cold water was added. The solution was mixed and allowed to reach a temperature of 4°C. To extract the metabolites, the tubes were centrifuged at 13'000 g at 4°C for 10 min and the aqueous methanol-water phase was transferred to a fresh 1.5 ml Eppendorf tube. The chloroform phase was re-extracted by adding 300 µl of ice-cold water and centrifuged as before. Again, the aqueous phase was removed and added to the first aqueous extract. Next the Eppendorf tubes containing the aqueous extract were dried in a centrifugal vacuum dryer (Savant speed vac) at 35°C until all solution evaporated.

5.5.9 Experimental evaluation for the malate uptake of pMDT1 in *Escherichia coli* Δ*dctA*

Three independent colonies of *Escherichia coli* Δ*dctA* (JW3496) were selected either harbouring the *A. thaliana* or the *S. viridis* pMDT1 variant on an inducible expression construct (Section A.3.13). The pre-growth was carried out in M9 minimal medium with glucose as sole carbon source as described previously in Section 4.4.7. To induce the expression of pMDT1, the cells were transferred to a fresh batch of 50 ml M9 glucose

at an OD_{600} of 0.5 containing 50 µM DAPG and 12.5 mg l^{-1} gentamycin. The procedure for the non-induced control was the same however, no DAPG was added to the media. The cultures were kept in 250 ml Erlenmeyer flask at 37°C while shaking at 250 rpm. After 1 h the cell cultures were collected by centrifugation at 13'000 g for 5 min and washed twice in M9 without any carbon source (but included 12.5 mg l^{-1} gentamycin and ± DAPG as appropriate). Subsequently, the cells were resuspended in M9 malate (including 12.5 mg l^{-1} gentamycin and ± DAPG as appropriate) at an OD_{600} of 0.2. Samples of the supernatant were collected every two hours, sterile filtered and flash frozen in liquid nitrogen. The amount of pyruvate was assessed as described in Section 5.5.7.

5.5.10 *Setaria viridis* protoplast generation

S. viridis protoplasts were isolated relying on a modified method used for protoplast generation from *Arabidopsis thaliana* leaves (Wu et al. 2009). Approximately, 2 g leaf material of 3-4 week-old greenhouse grown *S. virdis* plants was collected. *S. viridis* protoplasts were isolated by cutting the leaf perpendicular to the leaf veins into strips of ~1 mm width. The sliced leaf material was transferred into to a petri dish (Ø 90 mm) and soaked with ~25 ml enzyme solution (2% cellulose R-10 (Yakult Honsha. Co.), 0.08% macerozyme (Yakult Honsha Co), 500 mM mannitol,1 mM $CaCl_2$,1 gl^{-1} BSA, 9.4 mM MES, pH 5.6) for about 2 h while shaken (30 rpm on platform shaker) at room temperature. During the incubation step, the petri dish was covered from light. After the initial incubation step, the released protoplasts turn the solution into a mildly green suspension. The suspension was filtered using a nylon mesh with a pore size of 60 µm to remove unwanted tissue debris. The cell suspension was collected in a 50 ml tube and the protoplasts were collected by centrifugation for 5 min, 60 g at 4°C (Beckman GS-15 centrifuge). The supernatant was removed without disturbing the cell pellet and

washed with 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). The wash step was repeated and after the second wash step ~15 ml of the 25 ml W5 supernatant was removed. The protoplasts were carefully suspended in the remaining ~10 ml W5 and incubated on ice for 30 min. In the meantime, the cell density was determined with an e.g. Neubauer hemocytometer. All pipetting with protoplasts was carried out with cut pipette tips to reduce the amount of shear force the protoplasts were exposed to. Subsequently, the protoplasts were centrifuged for 2 mins, 60 g at 4°C and re-suspended in MaMg solution (400 mM mannitol, 15 mM $MgCl_2$, 5mM MES, pH 5.6) to a final concentration of $1.5x10^6$ protoplasts per 1000 µl.

Approximately $1.5x10^5$ protoplasts in 100 µl MaMg solution were mixed with 10 or 20 µg of plasmid DNA at room temperature. Subsequently, 110 µl of PEG solution (400 mM mannitol, 100 mM $Ca(NO_3)_2$, 40% PEG (MW4000)) was added to the suspension and incubated for 5 min at room temperature. After the incubation the protoplasts were carefully washed with 750 µl W5 and collected by centrifuge at 50 g for 1 min (Sigma 1-14 micro centrifuge). This wash step was done twice. After the second wash step, the protoplasts were suspended in 1 ml of W5 (room temperature) and incubated in 24-well plates at 28°C covered from light for 16 hr.

5.5.11 Confocal analysis of transformed protoplasts

Protoplasts were examined with a Zeiss SLM880 Axio Imager 2 confocal microscope using the objective C-Apochromat 40x using oil emersion. The laser wavelength was 514 nm and the GFP channel detected wavelength between 520-580 nm. The channel for the detection of chlorophyll autofluorescence was set to 650-735 nm. For both

channels a detector gain of 700 was set. Subsequently, the recorded images were processed with ImageJ 1.52h.

5.5.12 Keio *Escherichia coli* strains

The growth experiments were carryout with the *E. coli* genotypes as indicated. The wild type *E. coli* BW25113 and the *E. coli* Δ*dctA* mutant (JW3496) were obtained from the Keio Collection (Baba et al. 2006).

5.5.13 RT-qPCR to verify the relative transcript abundance of pMDT1 in *Setaria viridis*

To evaluate the relative transcript abundance of pMDT1 in *Setaria viridis* RNAi and segregating wild type plants a RT-qPCR was performed. For each line three independent RNA extractions were carried out. RNA was extracted using the RNAeasy plant tissue kit (Qiagen) according to the manufacturer's instructions and the RNA was eluted in 40 µl of H_2O (DEPC treated). Subsequently, any DNA contamination was removed by DNAase treatment (Turbo DNA free Kit Thermo Fisher) according to the manufacturer's instructions. Next, cDNA was synthesised using the Maxima first strand cDNA synthesis kit (Thermo Fisher). One reaction consisted of 1 µg RNA and 4 µl of 5x reaction buffer, 1 µl enzyme and water to 20 µl. The reaction was incubated for 10 min at 25°C followed by 20 min at 50°C and then 5 min at 85°C. Subsequently, the obtained cDNA was diluted 1:10 in water.

The reaction mix for the RT-qPCR was carried out in 50 µl final volume consisting of 25 µl master mix (SYBR Green real-time PCR master mix (Thermo Fisher)), 5 µl (1:10 diluted) cDNA, 1 µl primer mix (10 µM each) and 19 µl H_2O. Every cDNA sample was measured in three technical replicates. The RT-qPCR run was carried out in the real-time PCR system (StepOnePlus ThermoFisher) with following settings: 95°C for 10

min, followed by 40 cycles at 95°C for 15 s, 60°C for 10 s, 72°C for 15 s followed by a melting curve. To quantify the transcript abundance of pMDT1 following primer pair was used 5'-CCCTGCTTCCTCAAGGGATG-3', 5'-TCCGGAAGACTGAACAAGCC-3' (forward and reverse). Moreover, the abundance of two housekeeping genes, folylpolyglutamate synthase and phosphoglucomutase were measured for normalization with following primer pairs 5'-AAGTTTTTGCGTCACTTCCACT-3' 5'-CAAATTTCTGCCCTCGCTAAT-3' (forward and reverse) and 5'-ACGAGAAGGATTCATCCAAGAC-3', 5'-CGTGTACTCTTGCATCTTGGAG-3' (forward and reverse), respectively (Lambret-Frotté et al. 2015). The baseline-corrected normalised fluorescent reporter signal (ΔRn) were used to calculate the C_T-values using the algorithm Miner (Zhao and Fernald 2005). The received C_T-values were used as an input to calculate the relative expression using the R Bioconductor package EasyqPCR (Version 3.9). In case the C_T values of the three replicates differed in more than 0.5 C_T the value was discarded and the RT-qPCR for that sample repeated.

5.5.14 Total RNA extraction from *Setaria viridis* leaf tissue

For each individual plant (e.g. segregating wild type or RNAi line) a pooled sample of four mature leaves was collected and flash frozen in liquid nitrogen. The frozen tissue was grinded making use of a Tissuelyser (Qiagen TissueLyser II) while keeping the samples frozen. The frozen tissue was suspended in 1 ml TRIzol reagent (Thermo Fisher) reagent using a vortex until the powder was fully suspended. Subsequently, the sample was incubated at room temperature for 5 min. Next, the cell debris were collected through centrifugation at 15'000 g at 4°C for 10 min.

The supernatant (about 700 µl) was transferred to a 1.5 ml Eppendorf tube containing 200 µl chloroform and thoroughly mixed. The solution was centrifuged at 15'000 g at 4°C for 15 min and the upper aqueous phase was transferred to a fresh 1.5 ml Eppendorf tube containing 500 µl isopropanol. The solution was mixed and kept at room temperature for 10 min. Next, the extracted RNA was collected by centrifugation at 15'000 g at 4°C for 10 min. The supernatant was poured off and the RNA pellet was washed through resuspension in 75% ethanol and collected using a centrifuge at 15'000 g at 4°C for 5 min. The supernatant was removed and the wash step was repeated. Next, the RNA pellet was resuspended in 50 µl water (diethylpyrocarbonate treated) and the remaining DNA was removed using TurboDNAse (Thermo Fisher) treatment. The RNA was shipped to BGI for single-end transcriptome sequencing.

5.5.15 RNAseq data analysis of the *Setaria viridis*

The raw single-end reads were downloaded from BGI and processed as follows. Trimmomatic v0.39 (Bolger et al. 2014) was used to trim of the sequencing adapters and the low quality bases and the settings were: `LEADING:20 TRAILING:20 SLIDINGWINDOW:5:20 HEADCROP:1 MINLEN:35`. To evaluate the change in quality before and after the raw read processing the FastQC v0.11.5 (http://www.bioinformatics.babraham.ac.uk/projects/fastqc) was used. The most recent version of all *Setaria viridis* transcripts (Sviridis_500_v2.1.cds.fa.gz) was obtained from Phytozome V13 (Goodstein et al. 2012) and indexed using salmon v0.8.1 (Patro et al. 2017). The RNAi construct was added to the *S. viridis* transcript file to remove the reads obtained through the expression of the RNAi construct. The *Setaria italica* template used for the construction of the RNAi construct had two nucleotides that were different from the *Setaria viridis* pMDT1 and therefore enable the discrimination of reads originating from the RNAi construct from reads originating from

the endogenous gene. Subsequently, the trimmed reads were mapped to the indexed *S. viridis* transcript file using salmon (Patro et al. 2017). The mapping of the reads provided a quantification file that contains the read counts per transcript. The read counts for multiple isoforms of the same gene were added together to produce a single transcript abundance estimate value per gene. These counts were used as input to test for differential expression using DESeq2 v3.9 (Love et al. 2014) in R (versions 3.5.1, Feather Spray) and RStudio Version 1.2.1335. Genes were considered to be differentially expressed between the segregating wild type and RNAi lines, if the Benjamini-Hochberg adjusted p-value was ≤ 5%.

The *S. viridis* mesophyll and bundle sheath specific raw reads (John et al. 2014) were obtained from the European Nucleotide Archive project number PRJEB5074 (mesophyll cell samples: ERR385861, ERR385862, ERR385863, bundle sheath samples: ERR385864, ERR385865, ERR385866). The raw reads were processed as described above to produce analogous and comparable expression values.

5.5.16 Measuring carbon assimilation in *Setaria viridis*

The carbon assimilation (A) was measured over a range of intercellular carbon concentrations (C_i) on the fully expanded leaf located four leaves below the flag leaf of 5-week-old *S. viridis* using a LI-COR 6800 (LI-COR Bioscience). The measurements were performed after the leaf had equilibrated at 400 µmol mol^{-1} CO_2, flow rate of 500 µmol s^{-1}, 0.1kPa over pressure, 65% relative humidity, leaf temperature at 28°C, and irradiance 1500 µmol photons m^{-2} s^{-1} (90% red light 10% blue light). The carbon assimilation curves were subsequently measured in intervals of 3 min at following CO_2 concentrations: 400, 50, 20, 10, 8, 5, 2, 0, 75, 100, 200, 300, 400, 600, 800,1000, 1200,

1500, 1800, 2000 µmol mol^{-1}. During the measurements the irradiance, relative humidity and leaf temperature were kept constant.

5.5.17 Measuring light response in *Setaria viridis*

The light response was measured over a range of different light quantities on the fully expanded leaf located four leaves below the flag leaf of 5-week-old *S. viridis* using a LI-COR 6800 (LI-COR Bioscience). The measurements were performed using following parameters: 400 µmol mol^{-1} CO_2, flow rate of 500 µmol s^{-1}, 0.1 kPa over pressure, 65% relative humidity, leaf temperature at 28°C. To equilibrated the leaf to the high light the light intensities was increase in a step-wise manner (300, 500, 800, 1500, 2000 µmol photons m^{-2} s^{-1} 90% red light 10% blue light). Subsequently, the light response curved was recorded at following light quantities: 2000, 1800, 1600, 1400, 1200, 100, 800, 600, 400, 200, 100, 50, 0 µmol photons m^{-2} s^{-1} (90% red light 10% blue light). Throughout the measurements the relative humidity and leaf temperature were kept constant.

Chapter 6: General discussion

6.1 Exploiting natural variation in C_4 plants to identify function determining sites in proteins

The multiple independent evolution of C_4 photosynthesis from ancestral C_3 plants required some structural and biochemical adaption of the proteins and enzymes recruited into this complex trait. Perhaps one of the best-studied enzymes of the C_4 cycle is PEPC, which catalyses the primary carboxylation in most of the C_4 plants. In C_4 grasses it was shown that 21 amino acids evolved under positive selection (Christin et al. 2007). In addition, the C_4 model organism *Flaveria trinervia* has closely related C_3 and C_{3-4} species and multiple detailed analyses were carried out to study the molecular sequence changes in PEPC necessary to move from a PEPC with C_3 characteristics to a PEPC with C_4 characteristics (Bläsing et al. 2000; Engelmann et al. 2003; Engelmann et al. 2002; Jacobs et al. 2008; Svensson et al. 2003). Moreover, across different C_4 species there are known kinetic differences in key enzymes of the C_4 cycle. For example, variation in k_{cat} of RuBisCO in various C_4 species (Sage 2002), variation in $V_{max\ PPDK}$ between *Miscanthus* × *gigantheus* and *Zea mays* (Wang et al. 2008), and variation in the kinetics of PEPC in C_4 and C_4-like *Flaveria* (Engelmann et al. 2003).

If such kinetic variations in key enzymes of the C_4 cycle exist, then much of this variation must be attributable to differences in the protein sequence. Therefore, it is possible that variants of genes are present in one lineage that have enhanced properties in a C_4 context compared to the orthologous counterpart in other C_4 species. Additionally, the best possible enzyme variant for the particular function in the C_4 cycle might not have evolved yet because of the limited amount of time since the emergence of C_4 photosynthesis. Thus, the overarching goal of the first part of my DPhil project

was to harness the natural variation in a transcriptomic dataset consisting of 30 C_4 species and 17 C_3 species. To identify the molecular basis of these differences in kinetics I developed a novel analysis algorithm. The novel algorithm was developed because traditional methods that identify sites under selection in proteins have some limitations and are not built to analyse data consisting of fragmented and duplicated genes, nor can they account for gene expression data.

6.2 Development of a novel data analysis algorithm to identify sites in a protein alignment that have a different amino acid distribution between two sets of species

The developed data analysis algorithm makes use of the total variation distance metric to score the difference in the amino acid distribution of a protein alignment containing C_3 and C_4 orthologous proteins. The so-called Total Variation Distance algorithm revealed in a simulation experiment that it is able to overcome the problems of fragmented and duplicated sequences. Moreover, the analysis algorithm incorporates the transcript abundance into the total variation distance metric. Using the transcript abundance enables to account for duplicated genes that are clustered in the same orthogroup. Subsequently, the transcript abundance is used to allocate more weight to the higher expressed variant that therefore has a larger contribution to the observed trait. To have an estimation regarding the ability of the algorithm to identify sites that have an altered amino acid distribution between C_3 and C_4 species, previously known sites under positive selection in the PEPC sequence served as a positive control. Applied to the same dataset as used by Christin et al. 2007, the majority of the sites predicted to be under positive selection in PEPC were also highlighted by the Total Variation Distance algorithm. Interestingly, on a completely different dataset consisting of dicotyledonous plants rather than monocotyledonous plants four sites corresponded

to the sites identified by Christin et al. 2007 showing a prime example of convergent evolution on the molecular level between distant species. Thus, this outcome underlined the ability of the Total Variation Distance to identify potential difference that mediate biochemical variation within a set of two defined groups.

The subsequent analysis using the transcriptome data from 30 C_4 and 17 C_3 species was restricted to proteins that are closely linked to the C_4 cycle. The analysis revealed 12 proteins with in total 32 sites that have a different amino acid distribution between the C_3 and C_4 plants in the dataset. The twelve proteins can be loosely classified into three groups: enzymes of the core C_4 cycle, transporters and enzymes involved in photorespiration. None of the identified sites correspond directly to known sites involved in the catalytic function of the proteins. This is perhaps not surprising because changing components involved in the catalytic function might more often than not have a detrimental impact on the protein function. However, a subsequent analysis of the sites identified in PEPC that were located in close proximity to the PEP binding sites revealed that these sites are not randomly distributed in the protein sequence but are closer to known functional sites than you would expect by chance alone.

6.3 The identification of function determining sites using a similar approach as the Total Variation Distance analysis

In a recent study, Alvarez et al. 2019 published a similar approach to the Total Variation Distance algorithm that lead to the identification of the C_4 function determining sites of NADP-ME. The analysis relied on the well-curated NADP-ME isoforms of *Zea mays*, *Sorghum bicolor, Setaria italica*, and *Oryza sativa*. Importantly, the analysed dataset contained the non-C_4-functioning and C_4 isoforms of the C_4 plants and the ortholog from *O. sativa*. The developed analysis algorithm filtered for sites with strict differential conservation in the C_4 isoforms (Alvarez et al. 2019). This analysis approach revealed

in total 20 sites that passed that filter criteria. Subsequently, experimental evaluation using targeted mutagenesis revealed that four of the identified sites were major contributor to the acquired C_4 function (Alvarez et al. 2019). Interestingly, the Total Variation Distance algorithm did not identify any sites in either decarboxylating enzyme (NADP-ME or NAD-ME). One explanation is that the dataset consists of species of the NADP-ME and NAD-ME C_4 subtypes and therefore, it is perhaps not surprising that the Total Variation Distance algorithm did not detect differences in the amino acid distribution. However, given that a similar approach compared to the Total Variation Distance algorithm was able to identify sites that turned out to be major factors of altered enzyme kinetics for the function in C_4 plants illustrates that exploiting natural variation within C_4 plants is a promising approach. Moreover, it would make sense to re-analyse the NADP-ME gene sequence only including NADP-ME species in the C_4 group while NAD-ME species should be considered in the C_3 group.

Unlike the approach from Alvarez et al. 2019 the Total Variation Distance algorithm does not require fully curated sequences and deals with the different isoforms by considering the transcript abundance. It seems that the Total Variation Distance algorithm is able to deal with untidy data better than the strict differential conservation approach and is thus able to process a larger sample size.

6.4 Is it possible to improve the biochemical properties of C_4 proteins?

Instead of limiting the discovery to the major sequence determinants responsible for the changes in biochemical characteristics, the hypothesis was that the ideal enzyme has not evolved just yet because of the limited amount of time since C_4 photosynthesis emerged. However, evolution is indicating which sites we should target. Thus, the approach tried to carry out targeted mutagenesis at the previously identified sites to create a large variant library that ideally contains all possible sequence variants.

However, due to time constrains establishing a method to create sufficiently large variant libraries was prevented.

The eventual goal would have been to identify a superior protein variant using bacterial competition assays. I have shown that it is possible to use a bacterial test system to interrogate the function of plant enzymes. Moreover, the combination of the low expression level of the gene of interest and the supply of a single carbon source showed that the growth rate can be made dependent on the function of the gene of interest. Therefore, it is possible to use such a growth assay to run a competition assay with an entire population of cells each harbouring different sequence variants of the gene of interest. After repeated sub-culturing, the cells that harbour a sequence variant with improved biochemical characteristics compared should be enriched. Moreover, such a variant would be unlikely to have evolved by chance during a conventional plant breeding approach. The final aim would be to engineer that improved variant of a key C_4 cycle protein into a C_4 crop plant using CRISPR/Cas to improve the photosynthetic efficiency. C_4 crop plants already possess a variety of beneficial traits compared to C_3 crop plants. Thus, this approach would be able to improve the photosynthetic rate without having to introduce an entirely new metabolic network and changes in the leaf development as it is the case for projects like C_4 rice. In addition, a crop improvement approach that makes use of the CRISPR/Cas gene editing technology can be marketed in some countries without any regulation (Waltz 2018) which facilitates the distribution and a broad acceptance in the public.

6.5 Alternative application and availability of the Total Variation Distance algorithm

The Total Variation Distance algorithm is so far not published or publicly available. In order to make the algorithm fit for public use, it is currently being implemented as a stand-alone user-friendly Python script, which will be made available on GitHub. This gives me the opportunity to address some of the limitations raised in Chapter 2 (Section 2.4.1), such as the consideration of unequal group sizes. The vision for alternative applications are for example the analysis of other convergent traits such as CAM photosynthesis or any other dataset that can be grouped into two distinct groups e.g. salt tolerance in plants. Similar to the application presented in this DPhil, the comparison of protein alignments that consist of salt tolerant and intolerant species could identify sites in proteins that contribute to this trait.

One particular area where the applicability of the Total Variation Distance algorithm would be interesting to test in the future is as a subsequent analysis of a genome wide association study. Genome wide association studies are usually already grouped into two distinct groups (e.g. cases versus controls) and are tested for whether the association of single nucleotide polymorphisms (SNP) occurs more frequently in the cases than in the control group (Pearson and Manolio 2008). Thus, the outcome of the genome wide association study is usually whether a certain SNP is associated with the cases but it does not provide any evidence which of the genes in the associated region are causing the trait (Pearson and Manolio 2008). This is where the Total Variation Distance algorithm could support a subsequent analysis by characterising the amino acid distribution between the cases and the controls of all the proteins that are associated with the SNP. The discovery of changes in the amino acid distribution between the two groups may lead to the causative gene. In addition, if the species

under study have a well-annotated genome, the analysis of *cis*-regulatory elements would certainly be of interest too. However, the analysis of nucleotide sequences is currently not implemented yet and would require major changes such as a different scoring metric.

6.6 Expression of plant transport proteins in *Escherichia coli* and potential biotechnological applications thereof

The interdisciplinary approach of a bioinformatics analysis and a plant protein screen in bacteria lead to the preliminary characterization of one of the last missing steps of metabolite transport in a small group of NADP-ME C_4 plants. The expression of the plastid monocarboxylate and dicarboxylate transporter 1 (pMDT1) in *E. coli* grown in minimal media with glucose as sole carbon source resulted in the accumulation of pyruvate and other metabolites such as succinate. Interestingly, the succinate export was only detected for the *S. italica* variant of pMDT1 and not the *A. thaliana* variant. Succinate is an interesting compound that is used as a precursor for biodegradable polymers and was listed as an important platform chemical from biomass (Meng et al. 2016). There are already biotechnological methods available to produce succinate in large quantities using cell factories (Meng et al. 2016). Nevertheless, the genes in the UPF0114 gene family display a so far untapped diversity that could be used at least two-fold. Firstly, all genes of the UPF0114 should be characterised and the metabolite export should be closely investigated. Perhaps, one of the gene products already specifically exports a value-added compound that has a direct biotechnological application. Secondly, pMDT1 could be used as a chassis protein that undergoes directed evolution and the impact on the export characteristics are observed. Ideally, this would happen in a host organism that was engineered to produce a value-added compound (e.g. a compound that is of interest for industry or pharmaceutical

applications). The potential benefit to the production process would emerge if that compound has to be purified from the cell broth and using a transport protein that exports this compound into the culture supernatant could simplify the downstream purification process and therefore make it more economically feasible.

6.7 The identification and characterisation of the malate importer and pyruvate exporter in *Setaria italica*

The putative monocarboxylate and dicarboxylate transporter 1 was initially characterised in *E. coli*. These experiments in the bacterial system revealed that pMDT1 exports not only pyruvate but also malate, which is relevant in the context of C_4 photosynthesis. The accumulation of malate in the supernatant posed the question whether pMDT1 would also facilitate a malate influx if the concentration gradient would favour such a transport direction. Therefore, a subsequent characterization in *E. coli* Δ*dctA* mutant (which is unable to import malate) in minimal media with malate as sole carbon source revealed that pMDT1 enables the influx of malate into the bacterial cell while maintaining pyruvate export.

In addition, subcellular localization studies in *Arabidopsis thaliana* and *Setaria viridis* protoplasts revealed that the corresponding pMDT1 variant fused to a GFP localises to the chloroplast envelope. However, the transport characterization in the bacterial test system as well as the subcellular localisation in plant protoplasts did not confirm the involvement in the C_4 cycle. To verify whether pMDT1 participates in the C_4 cycle, an RNAi construct was designed to downregulate the transcript in the C_4 model plant *S. viridis*. The successful downregulation of pMDT1 was confirmed in the T_1 generation by RT-qPCR as well as by RNAseq. The plant expressing the RNAi construct revealed a reduced PEPC activity as well as reduced RuBisCO activity compared to the control plants. Although the maximum photosynthetic rate was reduced in the RNAi plants

compared to the control plants there was surprisingly no obvious difference in biomass accumulation. The reason that there was no obvious growth difference could have been caused by the low light conditions at which the plants were kept in the greenhouse, therefore no difference in assimilation rate resulted. Therefore, it would be interesting to grow some of the T_1 RNAi plants at high light conditions to see whether the biomass accumulation is impaired compared to control plants. Moreover, of these plants grown under high light conditions it is likely that the metabolite composition in the leaf is different between the RNAi plants and the segregating wild type plants. Therefore, measuring the leaf metabolites might contribute to our understanding regarding the perturbation the downregulation of pMDT1 is causing and how the plant is coping on a metabolite level.

The impact on the PEPC activity and RuBisCO activity caused by the reduction of pMDT1 in *S. viridis* was only likely if no other redundant genes with similar functions were expressed that compensated for reduction in metabolite flux. Thus, I am confident that the identified and characterised pMDT1 is the malate import and pyruvate export system in *S. italica* and other NADP-ME C_4 plants that express that gene in the bundle sheath cells. For the first time a full parts list of all transporters required in a C_4 plant is available.

Although the function and contribution of pMDT1 was characterised, the actual transport mechanism remains elusive. There are a few hypotheses about the potential transport mechanism of pMDT1 but without experimental proof these hypotheses remain speculations. To find an answer to this question, a collaboration with Prof. Newstead was initiated to carry out experiments of purified reconstituted proteins in liposomes.

6.8 C_4 photosynthesis: similar but different

Although for the first time a full parts list exists to create a synthetic NADP-ME C_4 cycle it remains elusive how for example *Zea mays* is importing malate and exporting pyruvate into and from the bundle sheath chloroplast. The characterised pMDT1 from *Setaria italica* Seita.4G275500 and *Arabidopsis thaliana* AT4G19390 are not orthologous to a *Z. mays* or *Sorghum bicolor* gene. There are, however, paralogs present for both the C_4 crops in the UPF0114 gene family. It is known that C_4 species can switch between paralogs that are used for C_4 photosynthesis, as it was the case with DCT2 in *Z. mays* (Emms et al. 2016). Therefore, it would be interesting to evaluate whether any of the *Z. mays* or *S. bicolor* paralogs in the UPF0114 gene family are perhaps carrying out the monocarboxylate and dicarboxylate transport function. The bacterial test system used to characterize the pMDT1 from *S italica* Seita.4G275500 and *A thaliana* AT4G19390 would be directly applicable to test these paralogs.

The entire gene family UPF0114 consists of so far uncharacterised genes and therefore, it would be highly interesting to know what the other genes in this gene family are doing. For example, Sevir.2G427600 (corresponding to Seita2G415600), which is in clade I, was expressed in the RNAi and control plants but it was not upregulated in response to the downregulation of Sevir.4G287300 (corresponding to Seita.4G275500 pMDT1). It is possible that the gene in clade I would have the same function as pMDT1. Interestingly, a close inspection of the branch length in clade I reveals that the protein sequence is not very diverse among the plant species and therefore it is likely that the function is potentially conserved. However, in clade II the branch length is longer indicating large sequence diversity and therefore an interesting source of alternative functions might have emerged. This diversity could be interesting to exploit for biological curiosity and potential biotechnological applications.

As described in Chapter 1, there are other missing transporters within the NADP-ME subtype as well as in the two other NAD-ME and PEPCK subtype. The bacterial assay used to characterise pMDT1 would be an ideal tool to go after the other missing transporters. Therefore, potentially contributing step by step to a better understanding of C_4 photosynthesis, a truly fascinating trait.

Appendix A

A.1 Total Variation Distance additional information

Table A.1: Calculated TVD and the corresponding p-value of the simulated dataset considering duplicated gens from the same species. The analysis assumed equal expression values for each transcript.

Position	1	2	3	4	5	6	7	8	9	10
TVD	0	0.75	0.44	0.63	0.75	0.875	0.81	0.87	0.87	0.5
p-value	1.00	1.00	0.25	0.91	0.11	0.06	0.03	0.03	0.03	0.03
TVDBioCh	0.00	0.50	0.44	0.63	0.75	0.88	0.81	0.00	0.00	0.50
$p\text{-value}_{BioChem}$	1.00	0.37	0.26	0.14	0.06	0.03	0.03	1.00	1.00	0.03

Table A.2: Simulated transcript abundance of the genes in the multiple sequence alignment.

Name	TPM
Spec1_Gene_1	100
Spec2_Gene_1	100
Spec3_Gene_1	100
Spec4_Gene_1	100
Spec5_Gene_1	100
Spec6_Gene_1	100
Spec7_Gene_1	100
Spec8_Gene_1	100
Spec9_Gene_1	100
Spec10_Gene_1	100
Spec3_Gene_2	50
Spec7_Gene_2	5
Spec8_Gene_2	5

A.2 Media

A.2.1 M9 mineral medium

For the preparation of M9 glucose add the following components to 867 ml H_20 (Table A.3) (Geerlof 2010).

Table A.3: Composition of M9 glucose

Volume	Stock Solutions	Content	Final concentration
100 ml	M9 salt solution (10x) pH 7.2	Na_2HPO	33.7 mM
		KH_2PO_4	22.0 mM
		NaCl	8.55 mM
		NH_4Cl	9.35 mM
20 ml	20% glucose	glucose	22.2 mM (0.4%)
1 ml	1 M $MgSO_4$	$MgSO_4$	1 mM
0.3 ml	1 M $CaCl_2$	$CaCl_2$	0.3 mM
1 ml	Biotin (1 mg ml^{-1})	Biotin	1 µg
1 ml	Thiamin (1 mg ml^{-1})	Thiamin	1 µg
10 ml	Trace elements (100x)	Trace elements	1 x

Preparation 100x trace elements:

Table A.4: Composition 100x trace elements

100x trace elements	Compound:	Concentration:
	EDTA	13.4 mM
	$FeCl_3$-$6H_2O$	3.1 mM
	$ZnCl_2$	0.62 mM
	$CuCl_2$-$2H_2O$	76 µM
	$CoCl_2$-$2H_2O$	42 µM
	H_3BO_3	162 µM
	$MnCl_2$-$4H_2O$	8.1 µM

For experiments with pyruvate as sole carbon source in M9 media, the final concertation of pyruvate was 27.2 mM as previously described in *E. coli* feeding experiments (Paliy and Gunasekera 2007). For experiments with malate as sole carbon source the final concentration in the media was 20.4 mM, which is equivalent the equivalent to the amount of carbon in the pyruvate media.

A.3 Golden Gate cloning

A.3.1 *Setaria viridis* 4G143500 sequence (5’-3’) optimized for expression in *E. coli* and golden gate cloning (SC module)

5’cactctgtggtctcaaatgCATCATCACCATCACCACggaggcGCCTCGAAGCCGGTAGAGAAGCACCACA
GCATTGATGCTCAATTACGTTTATTGGCTCCGGGTAAGGTGTCGGAAGATGACAAATTGGTCGAG
TACGATGCTTTACTTATCGACCGTTTCTTGGATATCTTCCAAGACCTTCACGGTCCGAACATTCGT
GAATTCGTCCAGGAATGTTACGAAGTCGCTGCGGAATATGAGCGTGACCGCGATGCCGCAAAAT
TATCCGAGCTGGGATCACGCTTGACAAAGCTGTCCCCAAACGACGCAATCGTGGTAGCTTCAAG
CTTCAGTAATATGCTTAACTTAACAAACCTTGCCGAGGAGGTACAACTGGCTCATCTTCGTCGTAA
CAAATTAAAGCGTGGGGACTTCGCTGACGAGGGGTTTGCCGCCACGGAAAGTGACATcGAaGAaA
CCTTAAAACGCCTGGTGACTGAATTGGGCAAGTCGAAAGAAGAAGTCTTTGATGCCCTGAAAAAC
CAAACTGTTGACTTAGTATTCACAGCTCATCCCACACAATCCGTCCGCCGTTCCCTGTTGCAAAA
GCATGCGCGCATCCGTAATTGCCTTACACAATTATACGCAAAAGACATTACTGAGGACGACAAAC
AAGAATTGGATGAGGCTCTTCAGCGCGAAATTCAGGCGTGTTTTCGTActgatgagattcgtcgcGCGcag
ccgacccccaagaCGAAATGCGTTATGGTATGTCCTACTTtCAcGAaACGATCTGGAAGGGCGTTCCA
AAGTTCCTTCGCCGTGTAGATACGGCCCTGAAAAACATTGGAATTGACGAACGTTTGCCGTACAA
CGCGCCGCTTATCCAGTTTAGCAGTTGGATGGGaGGcGACCGTGACGGGAACCCCCGCGTAACC
CCCGAAGTGACACGCGACGTTTGCCTTTTAGCGCGTATGATGGCTAGTAACTTGTACTTTTCTCA
GATCGAAGAGCTTATGTTCGAATTAAGTATGTGGCGCTGCAATGACGAGCTGCGCGCgaGaGCaG
AGGAGCTGCATGCCGCTTCGCAGAAAATCTCTAAACATTATATCGAGTTTTGGCGCCAGTTACCC
GCGAGTGAACCGTATCGCGTGGTCCTGGGTTATGTCCGCGACAAGTTATATAGCACCCGCGAAC
GCTCGCGTCACTTACTTACAAGTGGCTTTAGCGATATCCCAGAAGATAGCGCTTTTAAAAATGTTG
AAGAGTTCCTGGAGCCGCTGGAATTATGCTATAAGTCTTTATGCGATTGTGGCGATAAGACCATT
GCCGATGGCTCGTTACTTGACTTTATGCGTCAAGTAAGCACCTTCGGTCTTTCTATGGTCAAACTG
GATATTCGTCAGGAATCTGAACGTCATACTGATGTGATTGACGCTATCACAACCCATCTTGGCATT
GGGTCCTATCGCGAGTGGTCGGAAGAAAAACGTCAGGAGTGGTTATTATCCGAACTGCGTGGAA
AGCGTCCCTTGTTATCGAAAGACATGCCTCAGACAGAGGAAATCGCTGACGTGCTTGGATGTTTT
CACGTGCTGGCAGAATTACCTCGCGATTCTTTTGGGCCGTATATTATCTCGATGGCGACGGCTCC
GTCCGACGTTTTAGCAGTCGAGCTGTTACAGCGTGAGTGTCACGTCAAGCAGCCCTTACCGGTT
GTTCCACTTTTTGAGAAATTAGCTGACTTGCAGTCAGCACCTGCCTCAATCGAGCGTTTATTCTCA
CTGGACTGGTATATGAATCGTATCGGCGGGAAGCAACAGGTGATGGTAGGATATAGTGATTCTG
GTAAGGATGCCGGACGCCTGTCGGCGGCATGGCATCTGTATAAAGCTCAAGAGGAAATGGCAAA
GGTTGCAAAGCGTTATGGCGTAAAATTGACTATGTTCCACGGCCGCGGGGGGACTGTCGGACGT
GGTGGGGGTCCTACTCATCTGGCTATTCTGTCCCAACCCCCTGACACGATCAATGGCTCACTGC
GTGTGACTGTTCAAGGGGAAGTAATCGAGACTAGCTTCGGGGAGGAACATCTTTGTTTCCGCACT
TTACAACGTTTCACAGCTGCCACGCTTGAACATGGAATGCAtCCtCCaGTaTCGCCAAAACCTGAGT
GGCGTGCGTTAATGGATGAAATCGCCGCTGTGGCCACTGACGAGTATCGCTCTGTCGTCATGCG
CGAGCCGCGCttcgtggagtacttccgcAGCgcgactcccgaaactgagTACGGCCGCCTTAATATTGGATCGCG
TCCTGCCAAACGCAAGCCTAAGGGAGGGATCGAAAGCCtgcgtgccattccgtggattttcTCTtggacgcaaacc
cgcTTTCACCTTCCTGTCTGGCTGGGCTTTGGCGCAGCGTTTAAGCATGCGATGAAGAAGGATATT
AAGAACTTTCAGATGCTTAAAGAAATGTATAATGAGTGGCCGTTTTTCCGTGTTACTCTTGACTTA
CTTGAAATGGTCTTAGCAAAGGGTGATCCGACCATCGCTGGATTGTATGACCAACTGCTTGTAGC
TGACGAGCTTAAGCCTTTTGGCGAACAATTACGCAACAAtTAcGTaGAaACcGAAAAGCTTATCCTT
CAGGTCGCCGGACATAAAGAGATTCTTGAGAGTGATCCTGGGCTTAAGCAGCAGCTGCGTTTAC
GTGATCCATATATCACTATCCTGAATGTTTGGCAAGCTTATACCTTAAAGCGCATCCGTGACCCTA
ATTTCAAAGTAACTCCGCAACCGCCTTTGAGCAAGGAGTTTGCAGATGAGAATCAACCGCGCGGT
ATTGTGAAATTGAATCCTGCCAGCGAATATGGACCgGGaCTaGAgGAtACCCTTATTTTAACCATGA
AGGGTATCGCTGCAGGAATGCAGAACACTGGCTAAcctcagcgctttgagaccacgaagtg-3’

A.3.2 *Setaria viridis* 3G253900 PPDK sequence (5’-3’) optimised for expression in *E. coli* and golden gate cloning (SC module)

5’cactctgtggtctcaaatgCATCATCACCATCACCACggaggcGCCGTAGCAGACGTGGCCCCGATCCAAG
CAACCAAAAAGCGCGTATTCCACTTTGGAAAGGGCAAATCTGAAGGGAATAAGACGATGAAGGAA
CTTCTTGGCGGAAAAGGAGCAAATTTGGCCGAAATGTCGTCCATCGGTCTTTCCGTTCCCCCGG
GCTTCACTGTATCGACCGAGGCCTGCCAACAGTATCAGGAAGCAGGACGCTCATTACCGCAAGG
CCTGTGGGATGAGATTATTGATGGCTTAGCTTGGGTTGAGGAAGAGATGGGCGCCCGTTTAGGC
GACCCCCAACGTCCCCTGTTGTTAAGTGTGCGTTCGGGGGCCGCTGTCTCTATGCCAGGGATGA
TGGACACTGTATTAAACCTGGGACTGAATGATGAGGTAGCAGCTGGTTTGGGAGCTAAGAGTGG
GGAGCGTTTTGCGAATGACAGCTATCGTCGCTTCCTTGATATGTTTGGCAATGTAGTCATGGATAT
TCCTCACGCTTTGTTTGAAGAAAAATTGGAAGCAATGAAAGAGGCAAAAGGTGTTAAAAATGATAA
CGACCTGACGGCATCGGATCTGAAAGAATTGGTCGCGCAATACAAAGAAGTCTACGTGGAGGCC
AAAGGCGAGCCGTTTCCATCGGATCCTAAGAAGCAGCTTGAGCTTGCCGTTCTTGCTGTGTTCAA
CTCCTGGGACTCGCCCCGCGCCAATAAGTACCGTTCCATCAATCAAATCACAGGGCTTAAGGGG
ACCGCGGTAAATGTACAGTCCATGGTATTTGGTAATATGGGCGATACCAGCGGAACCGGCGTCC
TTTTCACTCGCAACCCCTCGACAGGGGAAAAAAAGCTGTACGGCGAGTTCCTGGTAAACGCACAA
GGCGAAGATGTGGTCGCCGGTATCCGCACTCCAGAAGATCTGGATGCAATGAAAGCACAGATGC
CTGATGCGTATGTTGAACTGGTTGAGAACTGTAAGATTTTGGAGTCCCATTATAAGGAGATGATG
GACATTGAGTTCACAGTCCAGGAAAATCGTTTATGGATGCTTCAGTGTCGCAGCGGCAAGCGCAC
GGGCCAGGGGGCCGTCAAGATCGCGGTCGACATGGTTAATGAGGGCCTTGTCGAGCGCCGCCA
AGCAATCAAGATGGTCGAGCCGGGGCACTTAGATCAGCTGCTTCACCCACAATTCGAGAACCCTT
CGGCATACAAGGATCAGGTCATTGCCACTGGACTGCCCGCAAGTCCTGGAGCTGCAGTCGGCCA
AATcGTaTTtACcGCaGAgGAcGCcGAaGCaTGGCATGCCCAAGGGAAAGCAGCTATTTTGGTTCGT
ACTGAAACCTCtCCaGAgGAtGTcGGCGGGATGCATGCCGCGGTAGGTATTCTGACCGCCCGTGG
CGGCATGACCTCTCACGCAGCCGTAGTTGCACGCGGTTGGGGTAAGTGCTGTGTAAGCGGATGC
TCGTCCGTATTGGTCAACGATGCAGAAAAATCCGTAGTTATTGGAGATAAAGTAATGCACGAAGG
TGAGTGGCTGAGCTTAAATGGGAGTACGGGTGAAGTCATCGTCGGCAAACAACCGCTTTCGCCG
CCGGTCTTAAGTGGCGACCTGGGCACCTTCATGTCGTGGGTGGATGAAGTCCGCCAACTTAAGG
TGTTGGCAAATGCCGATACACCGGAGGATGCTCTTACGGCACGCAATAAcGGaGCtGAaGGTATT
GGATTGTGCCGCACGGAACACATGTTTTTTGCCTCAGATGAACGTATCAAGGCTGTGCGCCAGAT
GATCATGGCCCCAACTTTGGAGCTTCGTCAGAAAGCTCTGGACCGCCTGTTGCCCTATCAACGTT
CAGATTTTGAGGGCATCTTCCGCGCTATGGATGGCTTGTCTGTGACAATCCGCCTGTTAGATCCA
CCGCTGCATGAATTCTTACCAGATGGTAATGTCGAGGACATTGTGCGTGAATTGTGCTCTGAGAC
AGGGGCTAACCAGGAAGAAGCATTAGCCCGCATCGAGAGCCTTAGTGAAGTGAACCCAATGCTT
GGGTTTCGTGGCTGCCGTCTGGGCATTTCGTATCCCGAATTAACGGAGATGCAGGCCCGTGCGA
TTTTTGAAGCAGCTATTGCCATGACTAATCAGGGCGTGCAAGTTTTTCCCGAAATCATGGTTCCCC
TTGTTGGGACACCACAGGAACTGGGCAACCAGGTCGCACTTATCCGTGAAACGGCAAATAAGGT
CTTTGCTGCGTTAGGGAAAACAATTGACTACAAGATTGGAACTATGATCGAGATTCCCCGCGCAG
CACTTGTTGCCGATGAGATTGCGGAACAGGCAGAATTTTTCTCGTTCGGTACGAACGATCTGACA
CAGATGACCTTTGGCTACTCGCGTGACGACGTTGGCAAGTTCTTGCCCATTTACTTATCCCAGGG
GATTTTACAGCACGACCCGTTTGAGGTCCTGGATCAACGTGGAGTTGGCGAATTGGTAAAATTTG
CAACTGAACGCGGTCGCAAAGCTCGCCCGAACTTAAAGGTTGGGATCTGCGGTGAACATGGAGG
CGAACCCTCCTCGGTAGCATTTTTTGCAAAATCCGGTCTTGACTATGTCTCGTGTAGCCCGTTTC
GCGTGCCTATTGCCCGTTTAGCGGCCGCTCAGGTCTTAGTTTAAcctcagcgctttgagaccacgaagtg-3’

A.3.3 *Setaria italica* Seita.4G275500 sequence (5’-3’) optimized for expression in *E. coli* and golden gate cloning (SC module)

CACTCTGTGGTCTCAAATGCGCGAGGTTTTATTCCGTCCAGTAGGGCTGCCGACGGAGACTAAG
TTTGGTGCTGGTTTAGAGGATCGCATCGAGAAAGTCATCTGCGCCTGCCGTTTCATGACGTTTTT
GGGAATCGGAGGGCTGTTGGCAGGGTGTGTACCTTGCTTCTTAAAGGGGTGCGTCTATGTCATG
GACGCCTTTGTAGAATACTACCTGCACGGAGGCGGTATGTTGATCCTTATGCTTTTAGAGGCGAT

TGATATGTTCCTTATTGGAACCGTTATGTTTGTCTTTGGGACTGGCCTGTACGAACTTTTCATCTC
GGAAATGGATATGTCTTATGGGTCTAATCTTTTTGGACTGTTTTCGTTGCCAGAACGTCCTAAGTG
GCTTGTGATCCAGAGTGTTAACGATCTTAAAACGAAACTGGGGCACGTTATTGTGATGTCCCTTTT
GGTAGGTATTTTTGAGAAAAGTTGGCGTGTCACGATTACCTCCTGTACGGATTTATTGTGCTTTGC
GGCCTCCATTTTTTTGTCGAGCGGGTGTCTGTACCTGTTATCACGTTTGAGTAACACAAAAGGTG
GATCTCACACTTGAGCTTTGAGACCACGAAGTG

A.3.4 *Setaria italica* Seita.4G275500 sequence (5'-3') optimized for *C*-terminal tagging and golden gate cloning (SC1 module)

CACTCTGTGGTCTCAAATGAAGCTCAGGCCTCTCACTTGCGTGGCGGCGGGGTGCGCCGGGTG
GGCGTGGAGGCCGAGGTCGCGCGTGCGGTCAGAGGCGGTGTCACCCAAGCGTTCCCACGCGG
CAGCGGCGGCGGCGGGCGCGGTTCATTCGGAGGAGCACCGCCGCGGCGGCATGCGCGAGGT
GCTCTTCCGCCCGGTGGGGCTGCCCACCGAGACGAAGTTCGGGGCGGGGCTGGAGGATCGGA
TCGAGAAGGTCATCTGCGCCTGCCGCTTCATGACCTTCCTCGGCATCGGCGGCTTGCTCGCCGG
CTGCGTCCCCTGCTTCCTCAAGGGATGCGTTTATGTGATGGACGCCTTCGTCGAGTACTACCTGC
ACGGCGGTGGAATGCTCATCCTAATGTTGCTTGAAGCCATTGACATGTTTCTCATTGGAACGGTC
ATGTTTGTATTCGGGACGGGCTTGTATGAGCTGTTCATCAGTGAAATGGACATGTCTTATGGCTC
CAACTTGTTTGGCTTGTTCAGTCTTCCGGAACGACCCAAGTGGCTGGTAATCCAGTCGGTGAATG
ATCTTAAGACAAAGCTGGGCCATGTCATTGTCATGAGTCTACTGGTTGGCATCTTTGAGAAGAGC
TGGAGAGTGACCATTACATCCTGTACTGACCTCCTTTGCTTCGCTGCATCAATCTTCCTCTCCTCA
GGTTGCCTCTACCTACTTTCCAGGCTCAGTAACACCAAAGGAGGGAGCCATACCGGTGTGAGAC
CACGAAGTG

A.3.5 *Arabidopsis thaliana* AT4G19390 sequence (5'-3') optimized for expression in *E. coli* and golden gate cloning (SC module)

CACTCTGTGGTCTCAAATGAGTACCAACCGTTTTGAAGCCTTAGAGGAAGGGATTGAAAAAGTTA
TTTATTCGTGTCGTTTTATGACGTTCTTAGGTACACTGGGGTCCTTGTTAGGTAGCGTGCTGTGTT
TCATCAAGGGCTGTATGTATGTTGTAGATTCTTTTCTTCAATATTCTGTCAATCGCGGGAAGGTTA
TTTTCCTGTTGGTCGAGGCCATTGATATTTATTTGTTGGGAACCGTTATGTTAGTGTTTGGACTGG
GCCTGTACGAGCTGTTCATCTCGAATCTGGATACTTCTGAGAGCCGCACCCACGACATCGTTTCT
AATCGCTCATCCTTGTTTGGTATGTTCACCTTGAAGGAGCGCCCCCAATGGCTTGAAGTAAAATC
GGTGAGCGAGCTGAAAACGAAACTGGGTCACGTAATTGTTATGTTGTTACTGATCGGGTTATTTG
ATAAGTCTAAACGTGTTGTTATCACCAGTGTTACGGACCTGTTATGCATTAGTGTAAGCATCTTCT
TCAGCTCAGCATGTCTGTTCTTGTTAAGCCGTCTTAACGGCAGCCACTGAGCTTTGAGACCACGA
AGTG

A.3.6 *Arabidopsis thaliana* AT4G19390 sequence (5'-3') optimized for *C*-terminal tagging and golden gate cloning (SC1 module)

CACTCTGTGGTCTCAAATGACGACGCCGTGCCGGACCATTAACGCTAATGCTATCGCCGCACCG
TCGCCTTCGGGATTGATCTTTAACGGGTTTCGTGATTTCGTCCCGATTGAAAAGAGGCTTGTTATA
TCAAGCTTTCGTGGCTTGAAACTACCGTCGAGGACGACGAAAACTATTACGAGCTCTGACTGGAG
TTGGAGTTATAGATCTCCGGGAAGATTGGCGTCTGCGTCTACGTCTACGTCTGCGTCTACGTCTA
CGTCTGCGGCTGTAACATCGAATTCTACGAATAGATTTGAAGCTTTAGAAGAGGGGATTGAGAAG
GTCATTTATAGTTGTCGATTTATGACGTTTCTTGGAACATTAGGATCTTTGCTCGGATCGGTTCTCT
GTTTCATCAAGGGATGTATGTATGTTGTAGACTCGTTTTTGCAGTATTCGGTGAATCGTGGGAAAG

TGATATTCCTTTTGGTTGAGGCCATAGATATATATCTCCTAGGAACTGTGATGTTAGTCTTTGGATT
GGGACTTTACGAGTTGTTCATCAGCAATCTCGATACTTCAGAGTCGCGCACACACGATATTGTTTC
CAATAGGTCGAGCCTCTTTGGCATGTTCACCTTAAAGGAGCGACCTCAATGGTTAGAGGTGAAAT
CGGTAAGCGAGCTGAAAACAAAGTTGGGACATGTAATAGTGATGCTACTATTGATTGGTTTGTTTG
ACAAGAGCAAGAGAGTAGTCATAACTTCTGTCACGGATTTGCTATGTATCTCTGTTTCTATCTTTTT
TTCTTCTGCTTGTCTCTTCTTGCTCTCTAGGCTCAATGGTTCACACGGTGTGAGACCACGAAGTG

A.3.7 Golden gate parts

Table A.5: Golden gate parts: promoters and RBS sequences gifted by the Poole Lab

Level 0 part	Sequence 5'-3'	Source
pOGG121	ggagtttacggctagctcagtcctaggtatagtgctagc agtagctagcactatacctaggactgagctagccgtaaa	(Grant 2019)
pOGG143	tacttctagaattaaagaggagaaattaaccaatg cattggttaatttctcctctttaattctagaagta	(Grant 2019)
pOGG145	tactctagagctaatcttcgcgtactaaaaaggcaagt cattacttgccttttagtacgcgaagattagctctag	(Grant 2019)
pOGG146	tactctagagctaatcttcgcgtactaagtaggcaagt cattacttgcctacttagtacgcgaagattagctctag	(Grant 2019)
pOGG156	gcttaacgcatgagaaagccccggaagatcaccttccgggggctttttattgcgctcctt agcgaaggagcgcaataaaaaagcccccggaaggtgatcttccgggggctttctcatgcgtt	(Grant 2019)

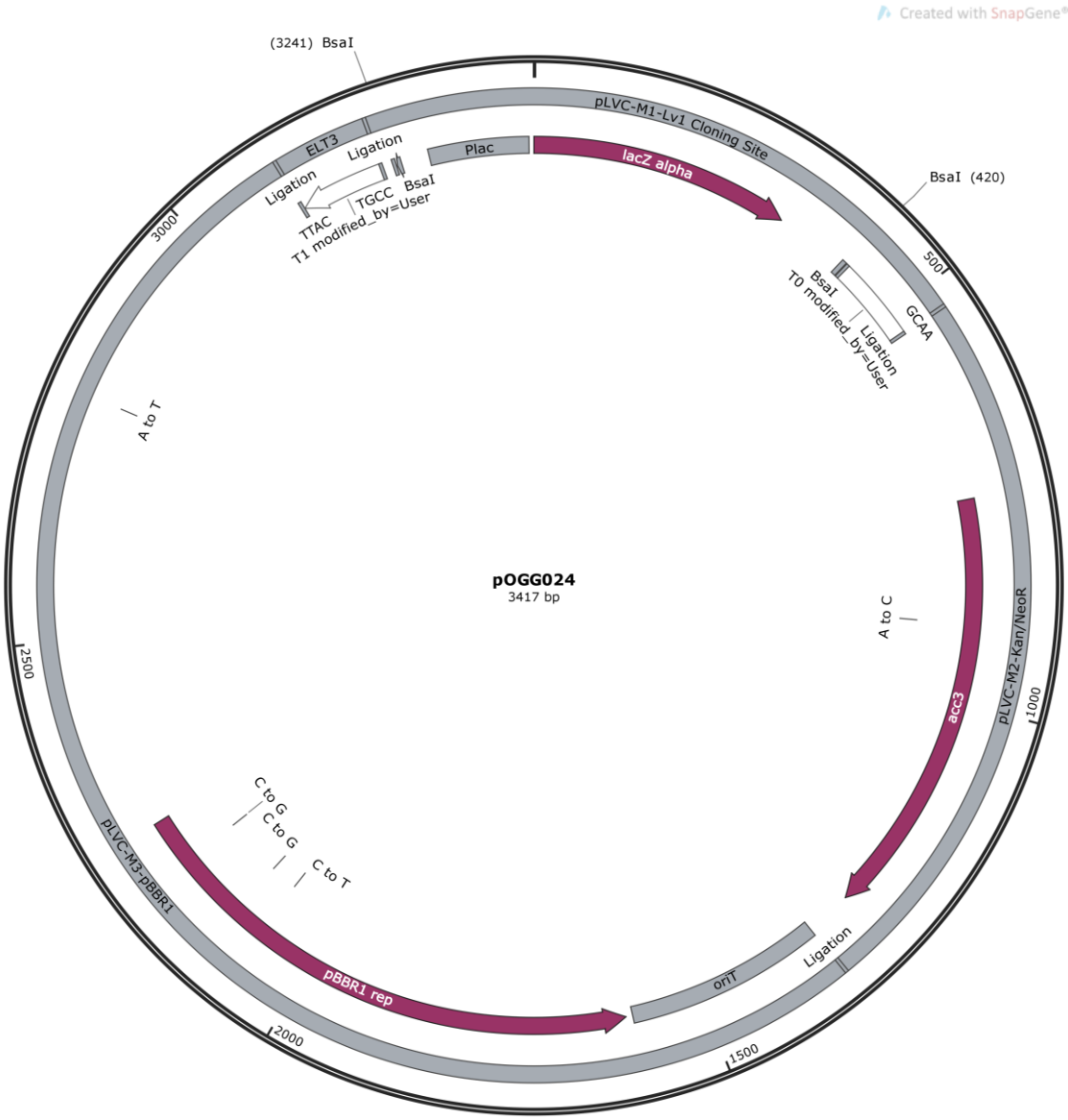


Figure A.1: Plasmid map pOGG024 (Geddes et al. 2019)

A.3.8 PEPC construct 1

Table A.6: Parts list PEPC construct 1

Level 0 part	ID	Selection	Strength AU
Backbone	pOGG024	GentR	-
Promotor	POGG121	SpecR	4095.5
RBS	pOGG145	SpecR	925
CDS	Sevir.4G143500	AmpR	-
Terminator	pOGG156	SpecR	-

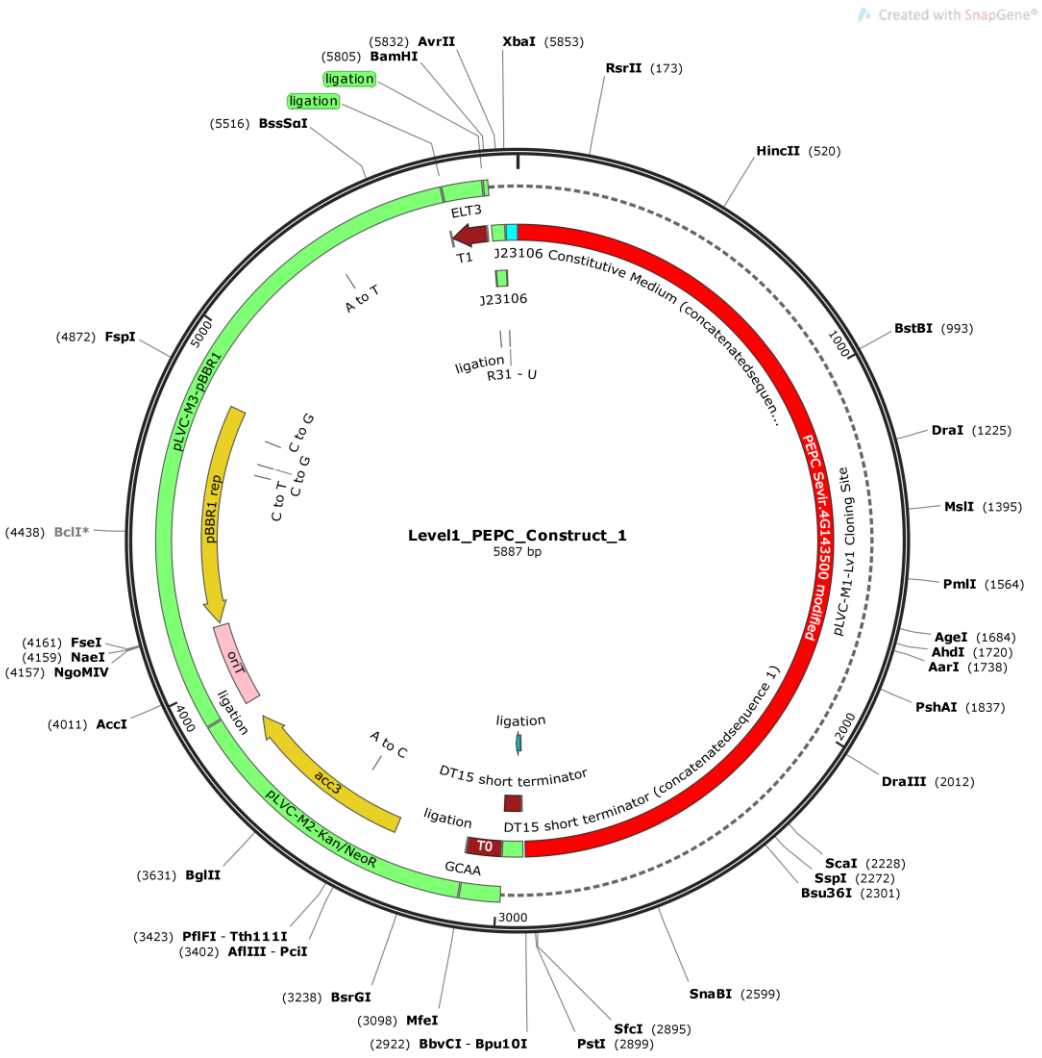


Figure A.2: Plasmid map PEPC construct 1

A.3. 9PEPC construct 2

Table A.7: Parts list of PEPC construct 2

Level 0 part	ID	Selection	Strength AU
Backbone	pOGG024	GentR	-
Promotor	POGG121	SpecR	4095.5
RBS	pOGG146	SpecR	135
CDS	Sevir.4G143500	AmpR	-
Terminator	pOGG156	SpecR	-

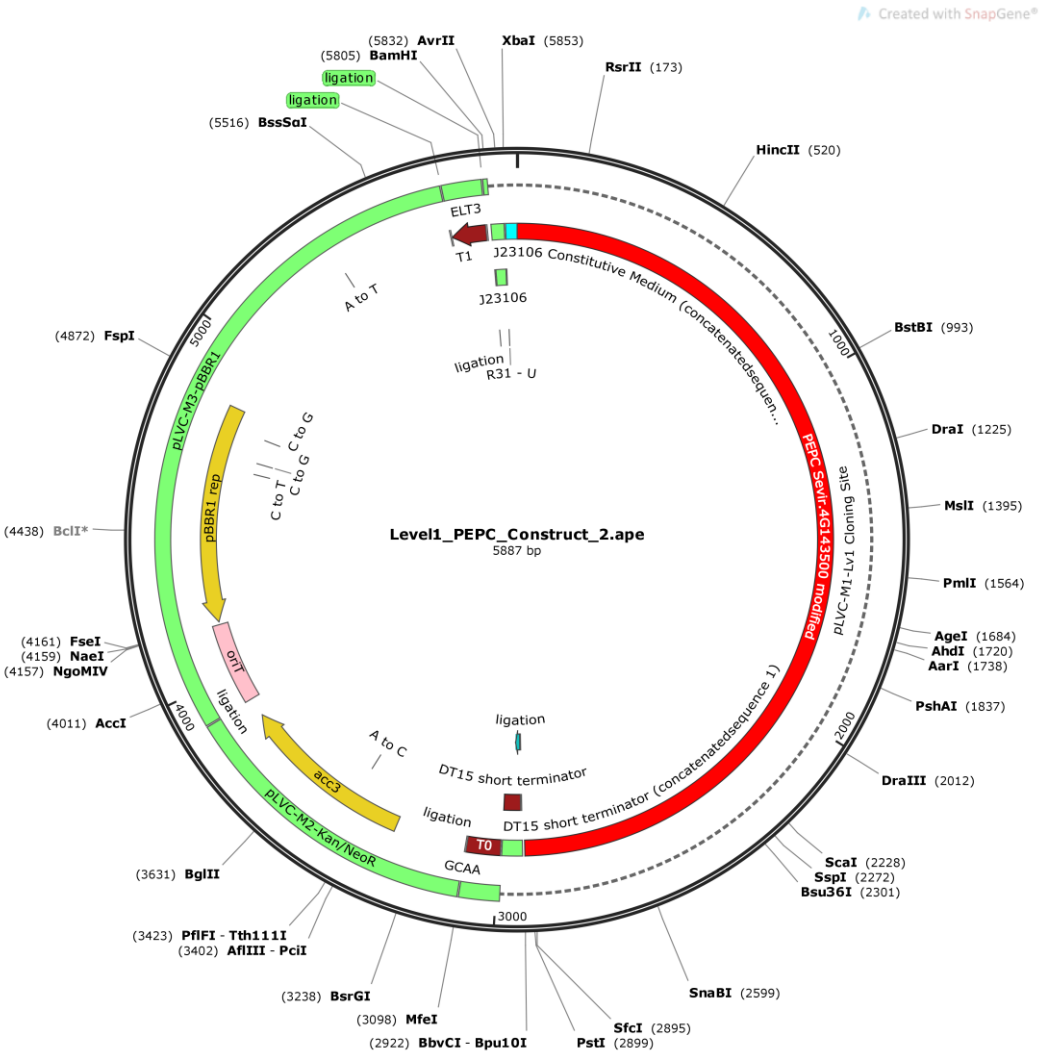


Figure A.3: Plasmid map PEPC construct 2

A.3.10 PPDK construct 1

Table A.8: Parts list of PPDK construct 1

Level 0 part	ID	Selection	Strength AU
Backbone	pOGG024	GentR	-
Promotor	POGG121	SpecR	4095.5
RBS	pOGG143	SpecR	6437
CDS	Sevir.3G253900	AmpR	-
Terminator	pOGG156	SpecR	-

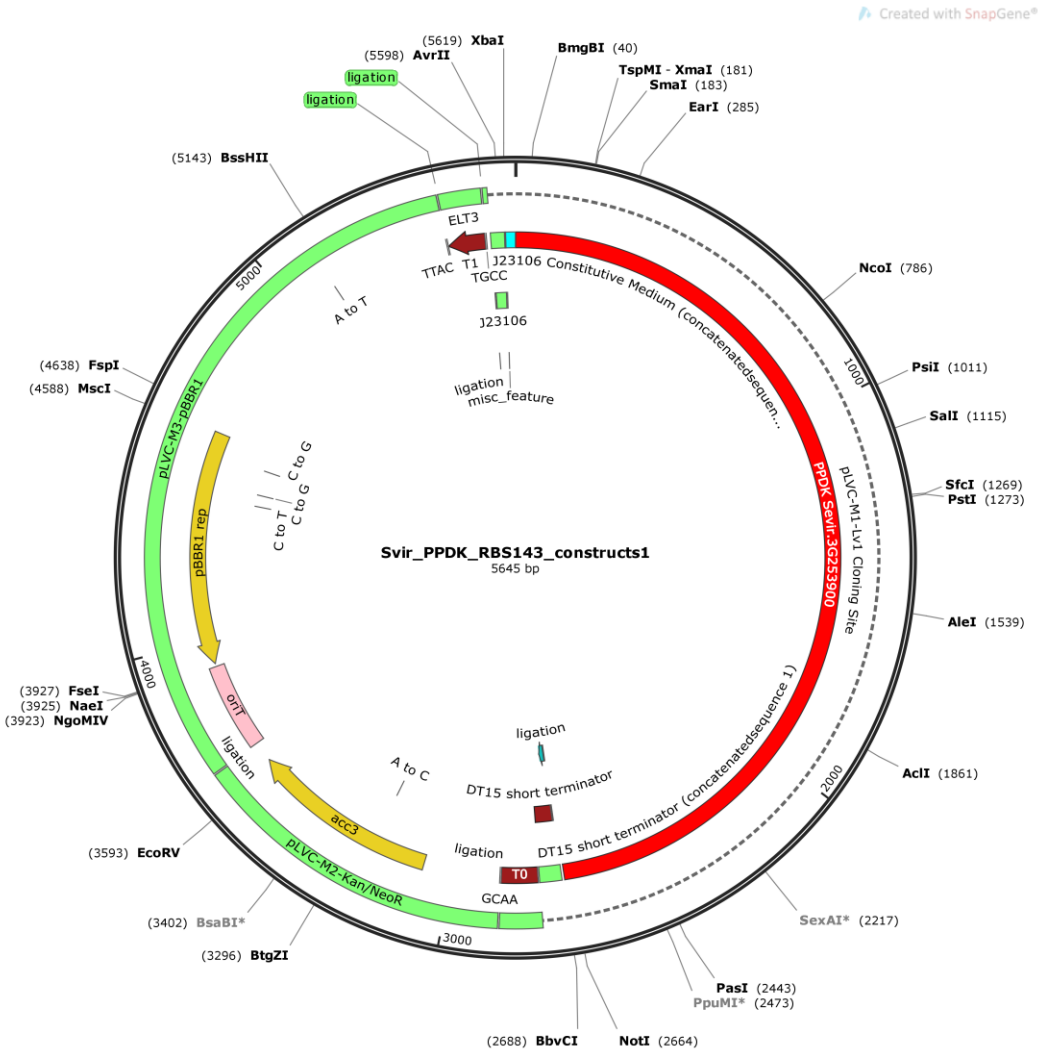


Figure A.4: Plasmid map of PPDK construct 1

A.3.11 PPDK construct 2

Table A.9: Parts list of PPDK construct 2

Level 0 part	ID	Selection	Strength AU
Backbone	pOGG024	GentR	-
Promotor	pOGG121	SpecR	4095.5
RBS	pOGG146	SpecR	135
CDS	Sevir.3G253900	AmpR	-
Terminator	pOGG156	SpecR	-

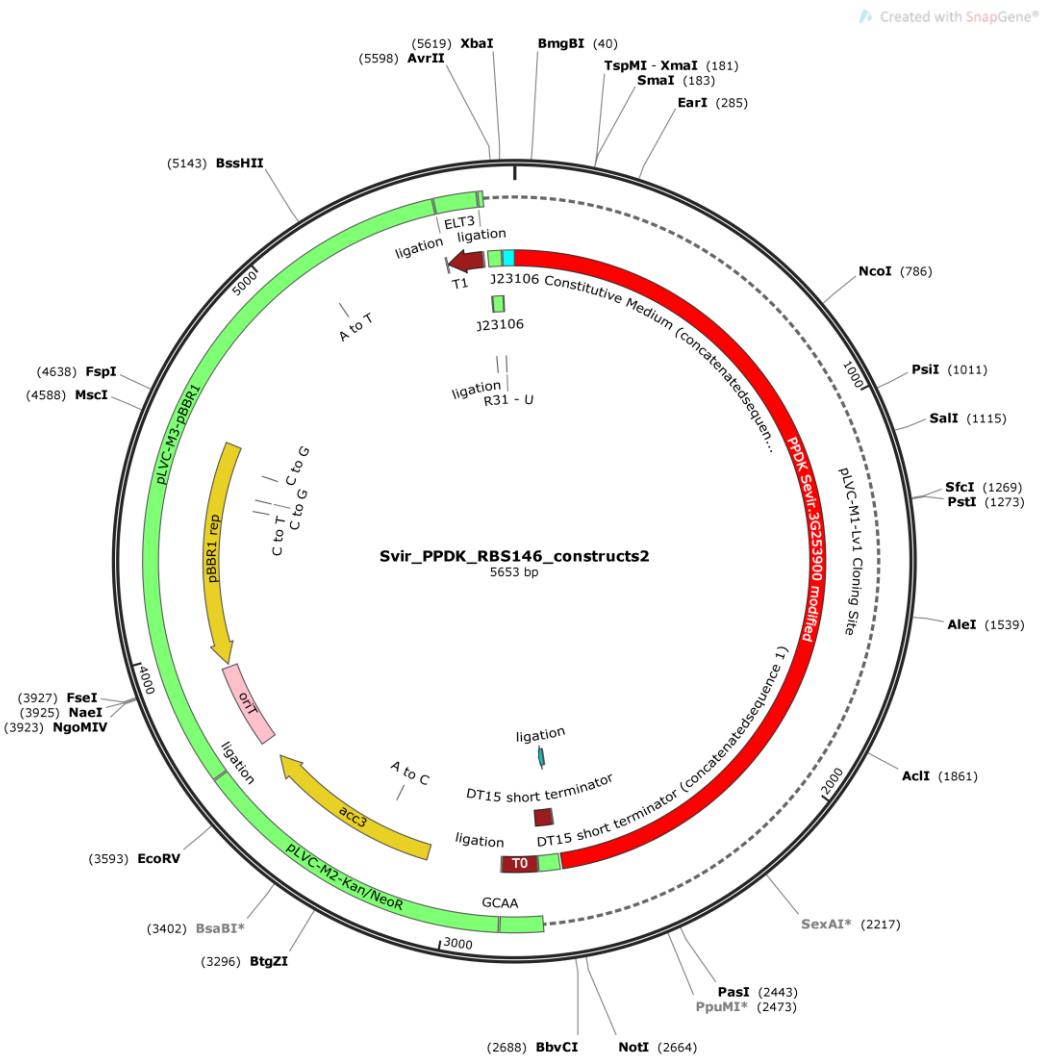


Figure A.5: Plasmid map PPDK construct 2

A.3.12 Level1 golden gate construct for the expression of pMDT1 GPP tagged and cytosolic GFP in protoplast

Table A.10: Level 1 golden gate construct components to express *Setaria italica* pMDT1 tagged with an eGFP at the *C*-terminus in protoplast

Level 0 part	ID	Selection	Reference
Backbone	pJK001c	Kan	(Grosse-Holz et al. 2018)
2x 35s promoter	pICH51288	AmpR	(Engler et al. 2014)
eGFP	EC15095	AmpR	Gifted by Giles Oldroyd, (Cambridge), ENSA group (Cambridge) and Ben Miller (UEA)
CDS	pUC57_Seita.4G275500_noStop	AmpR	This study
35s terminator	pICH41414	AmpR	Weber et al 2011

Table A.11: Level 1 golden gate construct components to express expressing *Arabidopsis thaliana* pMDT1 tagged with an eGFP at the *C*-terminus in protoplast

Level 0 part	ID	Selection	Reference
Backbone	pJK001c	Kan	(Grosse-Holz et al. 2018)
2x 35s promoter	pICH51288	AmpR	(Engler et al. 2014)
eGFP	EC15095	AmpR	Gifted by Giles Oldroyd, (Cambridge), ENSA group (Cambridge) and Ben Miller (UEA)
CDS	pUC57_AT4G19390_noStop	AmpR	This study
35s terminator	pICH41414	AmpR	(Weber et al. 2011)

Table A.12: Level 1 golden gate construct components to express cytosolic eGFP as a control of transfection in protoplast experiments

Level 0 part	ID	Selection	Reference
Backbone	pJK001c	Kan	(Grosse-Holz et al. 2018)
2x 35s promoter	pICH51288	AmpR	Engler et al. 2014
eGFP	EC15112	AmpR	Gifted by Giles Oldroyd, (Cambridge), ENSA group (Cambridge) and Ben Miller (UEA)
Terminator	pICH41414	AmpR	Weber et al 2011

A.3.13 Level1 golden gate construct to induced the expression of pMDT1 in *Escherichia coli*

Table A.13: Parts list of the inducible vector system for the expression of the *E. coli* optimized C_4 pMDT1 variant

Level 0 part	ID	Selection	Reference
Backbone	pOGG024	GentR	(Geddes et al. 2019)
PU	pOGG137	SpecR	(Grant 2019)
SC	pUC57_Seita.4G275500	AmpR	This study
Terminator	pOGG156	SpecR	(Grant 2019)

Table A.14: Parts list of the inducible vector system for the expression of the *E. coli* optimized C_3 pMDT1 variant

Level 0 part	ID	Selection	Reference
Backbone	pOGG024	GentR	(Geddes et al. 2019)
PU	pOGG137	SpecR	(Grant 2019)
SC	pUC57_AT4G19390	AmpR	This study
Terminator	pOGG156	SpecR	(Grant 2019)

A.4 Primers for genotyping of the *E. coli* mutants

Table A.15: Primer for genotyping of the E. coli PPC and PPS mutant strains

Primer ID	Sequence 5' to 3'	Length	Tm	GC%	product size	
					WT	KO
ppc_Fw1	CCGCATTGCGTAGTAATGTCA	21	59.	47	2627	1599
ppc_Rv2	ATTACGCATACCTGCCGCAA	20	60	50		
pps_Fw1	CGCCGCCGGGGATTTATTTT	20	61	55	2444	1367
pps_Rv1	AGAAATGTGTTTCTCAAACCGTTCA	25	59	36		

A.5 Primer for Sanger sequencing of the Gateway cloning plasmid pANIC 12A

Table A.16: Sanger sequencing primers for pANIC12A construct

Primer ID	Sequence (5'->3')	Length	Tm	GC%	Product size
Fw4_pANIC	CTGCTGCTGTCGGCTTTAAC	20	59.83	55	1572
Rv4_pANIC	TAAAACGACGGCCAGTGCC	19	60.96	57.89	
Fw5_pANIC	ACCGAATACGGCGTGGATAC	20	59.97	55	1121
Rv5_pANIC	GTAAAACGACGGCCAGTGC	19	59.79	57.89	
Fw6_pANIC	GCATAGGGTTTGGTTTGCCC	20	59.75	55	1484
Rv6_pANIC	GCGAGGTACGGTAGGAGTTG	20	59.9	60	
Fw7_pANIC	TGCACTTGTTTGTCGGGTCA	20	60.39	50	
Rv7_pANIC	AGTCGAGCATCTCTTCAGCG	20	59.9	55	1469

A.6 Primers for Sanger sequencing of the plasmids used for expression in protoplasts

Table A.17: Primers for Sanger sequencing

Primer ID	Sequence (5'->3')	Length	Tm	GC%	Product Size bp		
					Seita.4G275500	AT4G19390	EC15112
Fw_pJK001_1	ATCCTTCGCAAGACCCTTCC	20	59	55	1598	1676	851
Rv_pJK001_1	GTCCATGCCGAGAGTGATCC	20	60	60			
Fw_pJK001_2	CCACTATCCTTCGCAAGACCC	21	60	57	1730	1808	983
Rv_pJK001_2	TGCTCAACACATGAGCGAAAC	21	59	47			

A.7 Oligonucleotides for site directed mutagenesis and sequencing of the variant library

Table A.18: Oligonucleotides for site-directed mutagenesis

Gene	Primer sequence 5'to3'	Site DNA			Site AA	Length	Tm
Sevir.4G143500	5'-tctgaggcatgtctttcgannncaagggacgctttccacgc-3'	1485	1484	1483	495	41	78.7
	5'-gcgaggtaattctgccagcacgtgaaannnnnnaagcacgtcagcgatttcctctgtctg-3'	1530	1529	1528	510	60	79.4
	5'-cgagtttgggtccannnaaagatccaggggatagcacgaaggct-3'	2322	2321	2320	774	44	79.8
	5'-actcgatataatgtttagagatnnnctgcgaagcggcatgcagctc-3'	1041	1040	1039	347	46	78.1
	5'-cgtattcagtctctggtgtcgcnnngcgaaaatattcaacaaagcgc-3'	2214	2213	2212	738	47	79.1
	5'-ccacatacttaattcgaacataagctcnnngatctgagaaaagtacaagttactagc-3'	957	956	955	319	57	78.8
	5'-cctcccatccaactgctaaannngataagcggcgcgttgtacg-3'	837	836	835	279	43	79.8
	5'-cagtgagaataaacgctcgatnnnggcaggtgctgactgcaagtc-3'	1713	1712	1711	571	45	79.9
	5'-cttgaggggtcggctgnnngcgacggatctcgtcc-3'	684	683	682	228	35	79.4
	5'-ccacgcgatacggttcnnncgcgggtaactggcgc-3'	1086	1085	1084	362	35	79.4
	5'-gtcacacgcagtgagccnnngatcgtgtcaggggtt-3'	1977	1976	1975	659	37	78.3

Table A.19: Primers for Sanger sequencing of the variant library. Number in the primer ID indicate the position the primer is located on the PEPC sequence.

ID	Name ID	Sequence 5'-3'
FW1	Sseq_PEPC_37_F	5'-ATGCATCATCACCATCACCA-3'
FW2	Sseq_PEPC_417_F	5'-CAAATTAAAGCGTGGGGACT-3'
FW3	Sseq_PEPC_949_F	5'-AAGTGACACGCGACGTTTG-3'
FW4	Sseq_PEPC_1440_F	5'-TGACGCTATCACAACCCATC-3'
FW5	Sseq_PEPC_1918_F	5'-AAGGTTGCAAAGCGTTATGG-3'
FW6	Sseq_PEPC_2438_F	5'-AGCATGCGATGAAGAAGGAT-3'
RV1	Sseq_PEPC_300_R	5'-TCCCAGCTCGGATAATTTTG-3'
RV2	Sseq_PEPC_1286_R	5'-GGCTCCAGGAACTCTTCAAC-3'
RV3	Sseq_PEPC_2297_R	5'-CCGTACTCAGTTTCGGGAGT-3'
RV4	Sseq_PEPC_2793_R	5'-GAAATTAGGGTCACGGATGC-3'
RV5	Sseq_PEPC_2932_R	5'-CAGCGATACCCTTCATGGTT-3'

A.8 RNAi plant and segregating wild type plants



Figure A.6: Segregating wild type or RNAi line of *Setaria viridis* 34 days after germination. White bar on the right indicates 10 cm.

A.9 List of differentially expressed genes between RNAi and segregating wild type plants

Table A.20: Differentially expressed genes between RNAi and segregating wild type plants. Positive log_2 FoldChange value indicate a down regulation the RNAi lines compared to the WT

Gene	log2FoldChange	padj	Gene function	Athaliana Acession	Celltype
Sevir.3G146800	-3.04053	0.016537	In operons/gene neighbourhoods with thiamine biosynthesis genes in cyanobacteria		M
Sevir.4G287300	2.892165	1.03E-06	Uncharacterised protein family (UPF0114)	AT4G19390.1	BS
Sevir.3G084000	-2.81044	0.036492	tonoplast intrinsic protein 4;1	AT2G25810.1	M
Sevir.8G033100	2.781714	0.004375	Major facilitator superfamily protein	AT2G16980.2	
Sevir.9G138500	-2.54349	0.000693	thiaminC THIC	AT2G29630.1	M
Sevir.5G272100	-2.39892	0.003275	Component on NADH-dehydrogenase-like complex. Involved in cyclic electron transport in PSI	AT1G51100.1	
Sevir.1G332800	2.238194	0.034633	Phosphate-responsive 1 family protein	AT4G08950.1	M
Sevir.4G031300	2.115552	0.003652	sulfate transporter 3;4	AT3G15990.1	BS
Sevir.2G437601	-2.02182	0.042567	GTP1/OBG family protein	AT5G18570.1	M
Sevir.6G243400	1.834122	0.01593	cytochrome P450, family 78, subfamily A, polypeptide 6	AT2G46660.1	BS
Sevir.5G236000	-1.79452	0.000693	P-loop containing nucleoside triphosphate hydrolases superfamily protein	AT5G08610.1	M
Sevir.4G208400	-1.78963	0.008726	thiazole biosynthetic enzyme, chloroplast (ARA6) (THI1) (THI4)	AT5G54770.1	
Sevir.7G154203	-1.73478	0.001409	terpene synthase 21	AT5G23960.1	NA
Sevir.1G030600	1.708158	0.011439	Cysteine proteinases superfamily protein	AT3G49340.1	
Sevir.9G182200	1.665593	0.018821	MATE efflux family protein	AT3G21690.1	BS
Sevir.5G014000	1.617586	0.00775	manganese tracking factor for mitochondrial SOD2	AT4G27940.1	BS
Sevir.9G056400	-1.5611	0.008726	Leucine-rich repeat (LRR) family protein	AT3G22800.1	M
Sevir.2G239600	1.561002	0.017645	Remorin family protein	AT1G45207.2	M
Sevir.1G338700	1.559015	0.019204	alpha/beta-Hydrolases superfamily protein	AT5G42930.1	NA
Sevir.7G049900	1.544942	0.023471	basic helix-loop-helix (bHLH) DNA-binding superfamily protein	AT5G57150.1	NA
Sevir.5G405200	1.533914	0.02082	F-box family protein	AT5G07610.1	
Sevir.4G257900	1.515595	0.005526	P-loop containing nucleoside triphosphate hydrolases superfamily protein	AT3G28580.1	BS
Sevir.9G247800	-1.5059	0.038547	chlorophyllase 2	AT5G43860.1	
Sevir.5G420500	1.489686	0.038844	Protein of unknown function (DUF1685)	AT2G31560.2	
Sevir.3G072500	-1.42443	0.043389	12-oxophytodienoate reductase 1	AT1G76680.1	
Sevir.2G007600	1.398475	0.003625	Flavin-binding monooxygenase family protein	AT5G61290.1	BS

Table A.21: continue from table A.20

Gene	log2FoldChange	padj	Gene function	Athaliana Acession	Celltype
Sevir.9G106000	1.360337	0.021766	ethylene-dependent gravitropism-deficient and yellow-green-like 3	AT1G17870.1	M
Sevir.1G014600	1.356278	0.01153			M
Sevir.5G326200	1.343305	4.03E-05	RING/U-box superfamily protein	AT2G42360.1	BS
Sevir.1G352900	-1.31085	0.028909	glutamate receptor 2.7	AT2G29120.1	
Sevir.9G081600	-1.29921	0.01593	cytochrome P450, family 81, subfamily D, polypeptide 5	AT4G37320.1	M
Sevir.9G281600	1.28413	0.04838	Protein of unknown function (DUF179)	AT3G29240.1	
Sevir.2G313300	1.272361	0.002866	RING/U-box superfamily protein	AT1G20823.1	BS
Sevir.9G298700	-1.25117	0.019514	NB-ARC domain-containing disease resistance protein	AT3G14470.1	M
Sevir.9G186500	1.239437	0.016657			BS
Sevir.3G418200	1.238236	0.010261	NB-ARC domain-containing disease resistance protein	AT3G46730.1	M
Sevir.5G217000	-1.20175	0.028909	aldehyde dehydrogenase 2C4	AT3G24503.1	BS
Sevir.3G063100	1.177828	0.002449	myb domain protein 48	AT3G46130.1	BS
Sevir.7G142100	1.173502	0.016537			BS
Sevir.9G206900	1.159006	0.011145	alpha-galactosidase 1	AT5G08380.1	BS
Sevir.6G022800	-1.15092	0.001816	cysteine-rich RLK (RECEPTOR-like protein kinase) 34	AT4G11530.1	M
Sevir.7G132400	-1.11565	0.011439	Ribosomal protein L34e superfamily protein	AT1G26880.1	
Sevir.9G028200	-1.11303	0.042567	plastid transcriptionally active 13	AT3G09210.1	M
Sevir.7G171900	1.110592	0.011439	inositol transporter 2	AT1G30220.1	BS
Sevir.3G111700	1.088461	0.046787	HPT phosphotransmitter 4	AT3G16360.2	BS
Sevir.9G350700	-1.08384	0.02082	glutathione S-transferase TAU 16	AT1G59700.1	M
Sevir.6G215300	1.071407	0.031069	hydroxy methylglutaryl CoA reductase 1	AT1G76490.1	M
Sevir.1G118400	-1.05005	0.018821	wall-associated kinase 2	AT1G21270.1	
Sevir.2G071600	-1.03617	0.047431	Tetratricopeptide repeat (TPR)-like superfamily protein	AT1G59720.1	M
Sevir.2G007500	1.027801	0.004375	Flavin-binding monooxygenase family protein	AT5G61290.1	BS
Sevir.3G186200	1.008163	0.029092			
Sevir.4G032400	1.006962	0.02828	HXXXD-type acyl-transferase family protein	AT5G23940.1	BS

Table A.22: continue from table A.20

Gene	log2FoldChange	padj	Gene function	Athaliana Acession	Celltype
Sevir.9G350500	-0.99792	0.001939	glutathione S-transferase TAU 18	AT1G10360.1	M
Sevir.9G082000	-0.98915	0.040048	cytochrome P450, family 81, subfamily D, polypeptide 4	AT4G37330.1	M
Sevir.4G271300	-0.98303	0.015234	Translation protein SH3-like family protein	AT5G54600.1	M
Sevir.1G251100	0.978792	0.042567	GRAM domain-containing protein / ABA-responsive protein-related	AT5G23370.1	BS
Sevir.9G309000	0.964895	0.017645	serine carboxypeptidase-like 7	AT3G10450.1	BS
Sevir.8G107800	0.963772	0.038844	Protein of unknown function (DUF1685)	AT2G31560.1	
Sevir.3G241900	0.961889	0.011439	amino acid permease 2	AT5G09220.1	BS
Sevir.6G193200	-0.95896	0.00569	RNA-binding (RRM/RBD/RNP motifs) family protein	AT4G09040.1	M
Sevir.6G028400	-0.94516	0.004375			M
Sevir.9G055800	0.941799	0.011439	Aluminium induced protein with YGL and LRDR motifs	AT3G22850.1	BS
Sevir.3G355000	-0.93748	0.038844	nucleoside diphosphate kinase 2	AT5G63310.1	M
Sevir.1G329400	0.933107	0.036064	Glycosyl hydrolase family protein	AT1G78060.1	BS
Sevir.9G142700	0.915942	0.006585	SBP (S-ribonuclease binding protein) family protein	AT3G12920.1	BS
Sevir.8G092300	-0.90093	2.53E-05	Eukaryotic aspartyl protease family protein	AT2G03200.1	M
Sevir.9G566000	-0.88455	0.038844	nascent polypeptide-associated complex subunit alpha-like protein 2	AT3G49470.1	M
Sevir.3G228000	0.879156	0.01153	HXXXD-type acyl-transferase family protein	AT2G39980.1	BS
Sevir.7G280900	0.863697	0.038844	serine carboxypeptidase-like 33	AT3G17180.1	BS
Sevir.7G276001	0.843875	0.042567	N-terminal nucleophile aminohydrolases (Ntn hydrolases) superfamily protein	AT3G16150.1	BS
Sevir.1G184900	-0.82878	0.030117	Ribosomal protein S11 family protein	AT3G52580.1	
Sevir.5G235300	-0.82633	0.030117	Mog1/PsbP/DUF1795-like photosystem II reaction center PsbP family protein	AT5G11450.1	M

Table A.23: continue from table A.20

Gene	log2FoldChange	padj	Gene function	Athaliana Acession	Celltype
Sevir.7G113300	0.809164	0.004375	T-complex protein 11	AT1G22930.1	BS
Sevir.2G434900	0.785929	0.031069	sodium hydrogen exchanger 2	AT3G05030.1	M
Sevir.5G273300	-0.77744	0.040048	chloroplast 30S ribosomal protein S20, putative	AT3G15190.1	M
Sevir.7G171300	-0.76213	0.046622	P-loop containing nucleoside triphosphate hydrolases superfamily protein	AT3G16840.1	
Sevir.9G416000	-0.75589	0.003202	Uroporphyrinogen decarboxylase	AT2G40490.1	
Sevir.2G395900	-0.75441	0.000443	P-loop containing nucleoside triphosphate hydrolases superfamily protein	AT3G18600.1	
Sevir.9G435600	0.749361	0.046622	acyl-activating enzyme 7	AT3G16910.1	BS
Sevir.3G295100	0.747685	0.000201	WRKY DNA-binding protein 70	AT3G56400.1	M
Sevir.1G028900	-0.74339	0.007807	tonoplast monosaccharide transporter2	AT4G35300.4	M
Sevir.1G260900	0.733132	0.004631	alpha/beta-Hydrolases superfamily protein	AT4G16820.1	BS
Sevir.3G051900	-0.7267	0.018353	AMP-dependent synthetase and ligase family protein	AT2G47240.2	M
Sevir.5G359700	-0.72439	0.011145	Ribonuclease III family protein	AT4G37510.1	M
Sevir.6G047100	0.717502	0.017641	Protein of unknown function (DUF1005)	AT5G17640.1	M
Sevir.7G283600	0.699535	0.004731	RING/U-box superfamily protein	AT2G15530.2	BS
Sevir.4G202100	0.68965	0.03207	CONSTANS-like 5	AT5G57660.1	BS
Sevir.1G062100	-0.65627	0.043312	Uncharacterised BCR, YbaB family COG0718	AT2G24020.1	M
Sevir.9G103050	0.65485	0.036064			
Sevir.7G267200	0.653822	0.018821	TGACG motif-binding factor 4	AT5G10030.1	BS
Sevir.6G064000	0.641495	0.04838	myb-like transcription factor family protein	AT5G47390.1	BS
Sevir.6G237900	-0.63321	0.04838	Ribosomal protein L30/L7 family protein	AT2G01250.1	M
Sevir.3G268300	0.62545	0.017645			BS
Sevir.1G336100	-0.61836	0.037911	phenylalanyl-tRNA synthetase, putative / phenylalanine--tRNA ligase, putative	AT4G39280.2	
Sevir.8G065100	0.610833	0.018821	NAD kinase 2	AT1G21640.1	BS
Sevir.7G092800	-0.52612	0.036064	Protein of unknown function (DUF1624)	AT5G47900.1	M
Sevir.1G072500	-0.51622	0.02612	hydroxymethylbilane synthase	AT5G08280.1	M

Appendix B

Fragmatic: Automatic identification and resolution of erroneously fragmented genes in *de novo* assembled transcriptomes

B.1 Abstract

High throughput RNA sequencing allows investigating organisms for which a reliable reference genome is not available yet. In order to make sense of the data gathered during such an RNA-seq experiment, a process called *de novo* assembly is required to reconstruct the entire set of transcripts present. Due to several methodological reasons, including insufficient sequencing depth or variation in coverage, transcripts are often incompletely reconstructed. This incomplete reconstruction results both in the presence of both partial genes and fragmented genes in *de novo* assembled transcriptome. Such fragments are artefacts, which can substantially change the interpretation and conclusion of the experiment. Here, we present Fragmatic; a tool for identifying and resolving fragmented transcripts present in *de novo* assembled transcriptomes. Fragmatic uses evidence from orthologous genes in related organisms to identify and resolve fragmented transcripts. Resolution of fragmented genes improves the quality of the dataset reducing error in downstream applications such as phylogenetic inference. The precision and utility of Fragmatic was demonstrated on both real and simulated data. On simulated data Fragmatic achieved 100% precision and 18.75% recall. On real data we estimate the precision of Fragmatic to be between 66% and 89%. These results indicate that Fragmatic, with some further testing and development, will likely provide a useful tool for improvement of *de novo* assembled transcriptomes.

B.2 Introduction

Next generation RNA sequencing has dramatically changed our capacity to evaluate the underlying genetic and quantitative basis of numerous biological traits. For many species the cost and computational complexity of obtaining a genome sequence is still prohibitive and RNA sequencing presents an attractive low-cost alternative to access the genetic resources of those organisms. The most widely used technologies for high throughput RNA sequencing generate large amounts of short read data (50 to 300 base pairs in length). If a reference genome is available, these short reads can be mapped to this reference and further downstream analysis such as differential gene expression can be performed. If, however, the organism under study is lacking a suitable reference genome, then transcript sequences must be reconstructed from these short reads through a process known as *de novo* transcriptome assembly. These assembled transcriptomes can then be used for gene discovery, expression level quantification and differential gene expression testing.

Given the importance of *de novo* transcriptome assembly to comparative biological research in non-model species, multiple algorithms have been developed to produce *de novo* transcriptome assemblies from short read data (Smith-Unna et al. 2016). At present the most commonly used algorithms are Trinity (Grabherr et al. 2011), Oases (Schulz et al. 2012), Trans-ABySS (Robertson et al. 2010), IDBA-tran (Peng et al. 2013) and SOAPdenovo-Trans (Xie et al. 2014). These algorithms each approach the challenge of *de novo* transcriptome assembly in different ways. Additionally, each of these algorithms have a number of tuneable input parameters, that can be modified by the user to adjust for differences in RNA sequencing library construction, coverage depth and other features. Thus, the same input read data can be used to generate substantially different transcriptome assemblies, depending on the algorithm and

parameter sets that are used. These differences can lead to different biological interpretations such as differences in gene presence/absence. However, it is currently unknown to what extend fragmentation of transcripts occurs in a *de novo* assembled transcriptomes generated by different methods.

Unlike genomic sequences where sequencing coverage is uniform across the genome, the sequencing coverage of a transcriptome is variable as it is a function of the expression level of a gene (Surget-Groba and Montoya-Burgos 2010). *de novo* assembly methods recognise overlapping reads as continuous (and therefore as part of a larger transcript) if the number of overlapping reads exceed an operator-selected minimal cut-off value. Under normal circumstances highly expressed genes reach this minimal cut-off value and thus a fully assembled transcript is obtained (Figure B.1 A). However, low expressed genes or genes with an area that is difficult to sequence (such as GC rich regions) can cause the read coverage to fail to reach the required minimal cut-off value (Figure B.1 B). The transcript regions for which the read count is above this specified cut-off value will be assembled transcripts. However, if that coverage drops below the threshold the transcript may become fragmented or incompletely assembled. Thus, discrete fragments from the same gene can be found in the final assembly that could lead to misinterpretations such as gene absence, true gene fragmentation, or even gene duplication. One might argue that reducing the coverage cut-off required would solve this problem. However, reducing the cut-off value in an uninformed manner leads to assemblies containing many erroneously chimeric sequences.

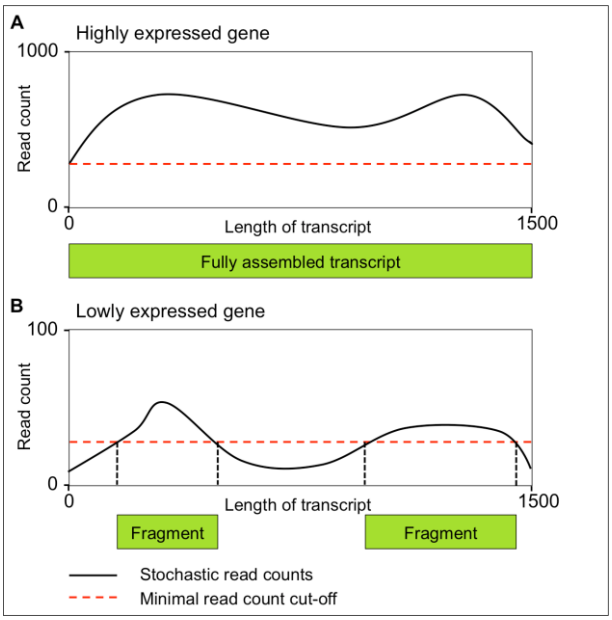


Figure B.1: Schematic representation of how assembly fragmentation occurs. The reads are stochastically sampled leading to stochastic variation in coverage depth. **A)** The coverage of a highly expressed gene is above the minimal read count cut-off required by the *de novo* assembly tool. Enough coverage is present across the entire length of the gene to produce fully assembled transcript. **B)** The variation in read depth for a lowly expressed gene can result in coverage dropping below the minimal cut-off value resulting in the assembly of the gene in multiple discontinuous fragments.

Given the difficulty in accurately assembling a transcriptome in the absence of any genome reference several attempts have been made to assist the process by leveraging information from related species. For example, in plants the reference genome of the model plant *Arabidopsis thaliana* was used to assist in the assembly of the transcriptome of a related species *Pachycladon enysii,* using a blend of read mapping against the *Arabidopsis* reference and *de novo* assembly of the un-mappable reads (Lesley et al. 2008). In a similar fashion the transcriptome assembly of the social wasp, *Polistes metricus,* was complemented using the reference genome of the honey bee *Aphis mellifera*, which diverged 100 to 150 million years ago (Toth et al. 2007). This approach has utility only for those species that have a closely related species with

a reference genome and is negatively affected by evolutionary distance. This is due to the fact that differences in nucleotide sequences between species accumulate as a function of evolutionary distance resulting in the inability to map reads from one species to the genome of another. However, discrepancies in the amino acid sequence accumulate much slower and therefore analysing the predicted amino acid sequence of transcripts against the reference proteome might be a valid procedure to enhance the *de novo* assemblies of non-model species, even in the case that the reference model is distantly related (Surget-Groba and Montoya-Burgos 2010).

Here we present Fragmatic, a new algorithm for the identification and resolving of fragmented transcripts in *de novo* assembled transcriptomes. Fragmatic requires a single sequence file containing multiple nucleotide or amino acid sequences that contain potential gene fragments as well as genes related to those putative gene fragments from other species. Fragmatic generates and analyses multiple sequence alignments and phylogenetic trees inferred from these sequences to evaluate whether or not subsets of the sequences in the file are likely discrete fragments generated from the same gene. We demonstrate the utility of Fragmatic on real and simulated sequence data.

B.3 Results

B.3.1 Problem definition and approach

The goal of *de novo* transcriptome assembly is to precisely recreate the entire set of transcripts that are present in the short-read data in the absence of a suitable reference genome. There are multiple elements that negatively influence the accuracy of this recreation process. Among these are sequencing errors, partial coverage of the transcript caused by inadequate sequencing depth, and biological variation such as

spliced variants. Assembly errors can also be caused by algorithmic simplifications such as reduced k-mer length and increasing the number of permitted mismatches between sequences. These errors result in assembly artefacts including assembly of gene families, transcript fusion, unjustified insertions in contigs, and structural abnormalities such as fragmentation.

Fragmatic aims to detect fragmented transcripts and resolve them, thus improving the quality and completeness of a *de novo* assembled transcriptome. Fragmatic obtains evidence of fragmentation by comparing potential fragmented genes to orthologous genes from closely related species. Fragmatic proceeds by creating a multiple sequence alignment and inferring a phylogenetic tree from this alignment (Figure B.2). Next, it assesses the gene tree by identifying monophyletic groups of sequences originating from the same species. These groups of sequences either arose through gene duplication since the divergence of the species in question from the closest relative in the tree, or they contain fragmented genes (or both). Fragmatic analyses the multiple sequence alignments for these single species monophyletic groups of sequences to identify those that are potentially fragments of the same gene. If potential fragments are found they are merge together and the process is repeated until no further putative fragmented genes can be identified or resolved (Figure B.2). This figure outlines the steps taken by the Fragmatic algorithm. The entire process is iterative and continues until the topology of the tree that is inferred is identical to the topology of the tree that was inferred in the previous round. While it is very likely to find monophyletic groups of sequences that arose through gene duplication it is very unlikely to find that these duplicated genes will be fragmented in such a way that they contain no overlapping regions or identical sequences, which are shorter than the minimal overlap required by the *de novo* assembly tool.

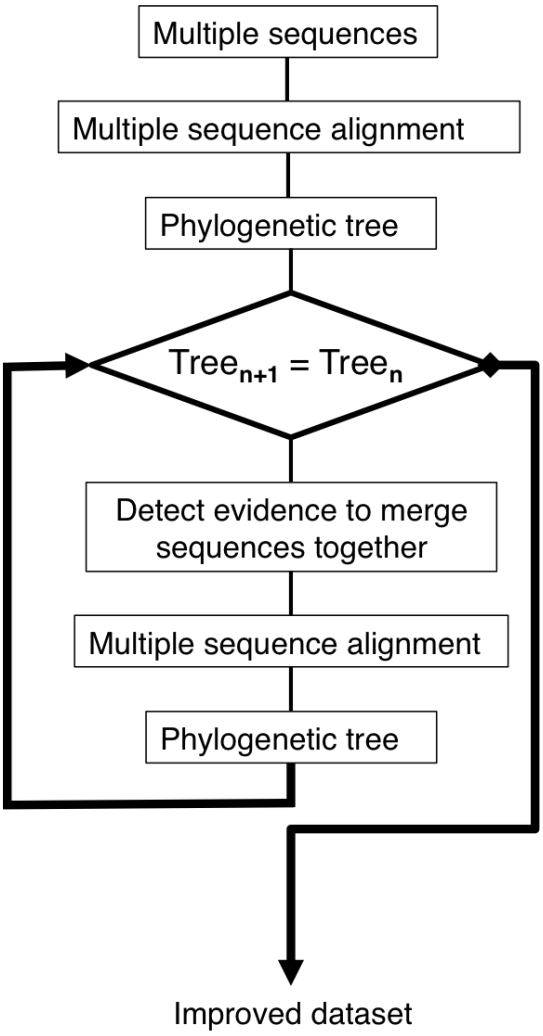


Figure B.2: Schematic workflow of Fragmatic: As input Fragmatic takes a multiple sequence file. It creates a multiple sequence alignment and infers a phylogenetic tree. It enters a while loop and identifies monophyletic groups of sequences originating from the same species. It merges together potential fragmented genes under the condition that there is enough evidence to do so. A new phylogenetic tree is inferred based on the improved dataset. The while loop ends once the topology of the previous tree is equal to the topology of the current tree.

B.3.2 Simulated fragmented sequences

To assess the ability of Fragmatic to detect and resolve fragmented genes a simulated dataset of *in silico* fragmented genes was generated. These simulated fragmented

genes were generated by taking a known set of genes from a given species and *in silico* fragmenting a random subset of 1000 of these genes according to a set of simple assumptions regarding how these fragmented genes arise during the assembly process. Firstly, a *de novo* assembly algorithm will fail to assemble two fragments of a gene together if the length of the overlapping sequence shared by the two gene fragments is less than the required overlap set by the *de novo* assembler operator. Thus, simulated fragmented genes could contain up to a 39 nucleotide (13 amino acid) overlap as this is the default value used in many *de novo* transcriptome assemblers. Secondly, fragmented genes will arise if insufficient sequencing depth causes the *de novo* assembly algorithm to fail to connect two ends of a transcript. As the minimum assembled contig length of many *de novo* transcriptome assemblers is 200 nucleotides (66 amino acids) simulated fragmented sequences contained at least two fragments of 200 nucleotides separated by any number of un-assembled nucleotides.

In Figure B.3 A shows the distribution of the length of the overlap region for those simulated fragmented sequences where the ends overlap. Figure 8.3 B shows the distribution of the length of the gap between gene fragments for those fragmented sequences whose ends did not overlap.

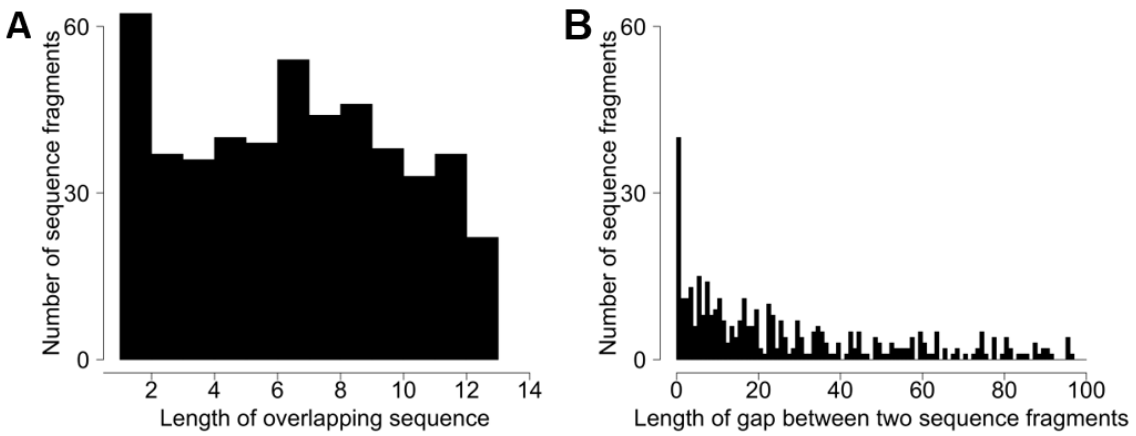


Figure B.3: Distribution of the simulated (**A**) fragmented sequences where the ends overlap and (**B**) fragmented sequences whose ends did not overlap sequences generated using the publicly available *Oryza sativa* proteome.

B.3.3 Application of Fragmatic to simulated fragmented sequences

To assess whether Fragmatic was capable of identifying and resolving these simulated fragmented genes the publicly available proteome of four closely related monocotyledons, *Oryza sativa, Brachypodium distachyon, Setaria italica,* and *Sorghum bicolor,* were used to create 1000 fragmented coding sequences for each of the species (see Table B.1). These fragmented coding sequences together with the non-fragmented sequences from the reference proteomes were clustered with OrthoFinder (Emms and Kelly 2015) to form orthogroups. This resulted in the construction of 17’911 orthogroups that contained 3007 of the 4000 fragmented sequences (Table B.1). Of these 3007 clustered fragments only 443 we clustered such that all fragments of the same gene were place in the same orthogroup. Being clustered in the same orthogroup is necessary for Fragmatic to be able to analyse and resolve these simulated fragments. Thus, the accuracy of Fragmatic is assessed on these 443 genes rather than the total set of 4000. As the reduction from 4000 to 443 is a limitation of the orthogroup clustering algorithm.

Table B.1: Summary statistics of the simulated dataset. Fragments (# Frag.) created for each of the species. The number of Fragments clustered in an orthogroup (# Frag. in cluster). Number of fragments clustered in the same orthogroup as its fragmented counterpart (# paired Frag. same cluster). How many of the these were identified and merged by Fragmatic (# merg. Frag.). The proportion of merged fragments (Recall rate). The proportion of correctly merged fragments (Precision). The false positive rate (FPR).

Organisms	**# Frag.**	**# Frag. in clusters**	**# paired Frag. same cluster**	**# merg. Frag.**	**Recall rate**	**Precision**	**FPR**
O. sativa	1000	722	110	21	19 %	100 %	0 %
B. distachyon	1000	756	91	22	24 %	100 %	0 %
S. italica	1000	781	129	19	14 %	100 %	0 %
S. bicolor	1000	748	113	21	18%	100 %	0 %
Total	4000	3007	443	83			

Out of these 443 correctly clustered fragments Fragmatic merged between 19 and 22 fragments together, which resulted in a total of 83 merged fragments. All 83 merged fragments were correct and no chimeric sequences were created. Thus, with this simulated dataset Fragmatic has a final recovery rate of approximately ~18% with a precision of 100% (Table B.1). It important to note that among the 83 merged sequences constituted of fragments with an overlapping sequence.

B.3.4 Application of Fragmatic to real *de novo* assembled transcriptomes

To demonstrate the utility of Fragmatic on real sequence data, Fragmatic was applied to publically available *Oryza sativa* transcriptome assemblies generated using three of the most widely used *de novo* assemblers, Oases, Soap, and Trinity. To provide the orthogroups on which Fragmatic was run the predicted proteins from each of these *de novo* assembled transcriptomes were clustered independently with the reference proteomes from three closely related monocotyledons *(Brachypodium distachyon, Seteria italica,* and *Sorghum bicolor)* using OrthoFinder. Fragmatic was run on each

orthogroup and the performance of Fragmatic was assessed by BLASTing all of sequence fragments that were joined together by Fragmatic against the genome derived reference proteome for *Oryza sativa* and assessing whether the joined gene fragments each identified the same reference gene as their respective top BLAST hit.

Each of the *de novo* assembled transcriptomes was generated from the same input data. However, the number of transcripts in each of the assemblies varied considerably between assembly methods, with Trinity having ~2.4 times the number of transcripts as to Oases (Table B.2). Although the number of transcripts varied the number of orthogroups inferred from these genes was very similar (13’8100 and 14’442). This discrepancy suggests that variance in the number of assembled contigs is either due to redundancy (multiple assembled variants for each gene) or fragmentation (multiple discontinuous fragments for each gene).

When applied to the complete set of orthogroups inferred for each dataset Fragmatic identified and resolved between 21 and 56 putative fragmented genes. When these resolved genes were assessed by analysis of the top BLAST hits for each fragment that was joined it was revealed that unlike the simulated analysis only 66% to 89% of those joined sequences were correct (Table B.2).

Table B.2: Summary statistics of the real *de novo* assembled transcriptome. Number of transcript present in the *de novo* assembled transcriptome (# Transcripts). Number of orthogroups inferred with OrthoFinder (# Orthogroups). Number of merged fragments (# merged). The number of merged fragments obtained the same BLAST top hit (# true) or not (# false). Proportion of correctly merged fragments (Precision).

Specie ***Oryza sativa***	# Transcripts	# Ortho-groups	# merged	# true	# false	Precision
Oases	58 152	14 029	56	46	10	82 %
Soap	81 206	13 882	21	14	7	66 %
Trinity	142 419	14 442	49	44	5	89 %

One example of a correctly resolved sequence fragment by Fragmatic is depicted in Figure B.4. To simplify the figure the full multiple sequence alignment is not shown, but rather just the two sequence fragments that were joined together by Fragmatic and the closest outgroup sequence in the tree. The two *Oryza sativa* transcripts (named *O. sativa*_1 and *O. sativa*_2 convenience) are shown aligned to the orthologous gene from *B. distachyon* (named *B. distachyon*_4). Fragment *O. sativa*_1 aligns from position 52 to position 190 and fragment *O. sativa*_2 aligns from position 190 to position 346 (Figure 8.4 A). The two fragments have a single overlapping alanine residue at position 190. Fragmatic used the evidence from the alignment together with the evidence obtained from the phylogenetic tree (Figure B.4 B) to join these two sequences together. The joined sequences are shown in Figure B.4 C, and the new phylogenetic tree is shown in Figure B.4 D. Both of the joined fragments hit LOC_Os09g28100.1 (Rhomboid-like protein 10) as their top BLAST hit and are thus fragments of the same gene.

The cases where Fragmatic joined two sequences together that had different best BLAST hits in the reference proteome were investigated. These errors could be

classified into two separate categories. 1) Sequence pairs that had little or no support in the phylogenetic tree. 2) Sequence pairs with erroneous alignments. An example of an erroneous alignment is shown in Figure 8.5 A. here two gene fragments (*O. sativa*_7 and *O. sativa*_5) align to the orthologous gene of *Brachypodium distachyon*. At position 447 and 448, however, a gapped alignment between the two fragments *O. sativa*_7 and *O.sativa*_5 becomes apparent. Modification of the alignment rules to prohibit gapped overlaps could lead to improvements in the accuracy of Fragmatic.

In Figure B.5 B, a phylogenetic tree with the corresponding node support value is illustrated. The two fragments *Oryza sativa 8* and *Oryza sativa 9* were merged together and did not obtain the same BLAST top hit. The two sequences as such did pass all the filtering implemented in Fragmatic. However, when considering the phylogenetic tree, it was immediately clear that the support value for this monophyletic group was low. Thus, the partitioning at this given node should not be considered as reliable. Utilisation of these support values could lead to improvements in the accuracy of Fragmatic.

If we assume that the orthogroup clustering in the real sequence dataset had similar properties to the simulated sequence dataset, it is estimated that there are somewhere between 1050 to 2800 fragmented genes in each of the real-life datasets. It will be interesting to see whether further improvement of Fragmatic could identify and resolve more of these fragmented genes.

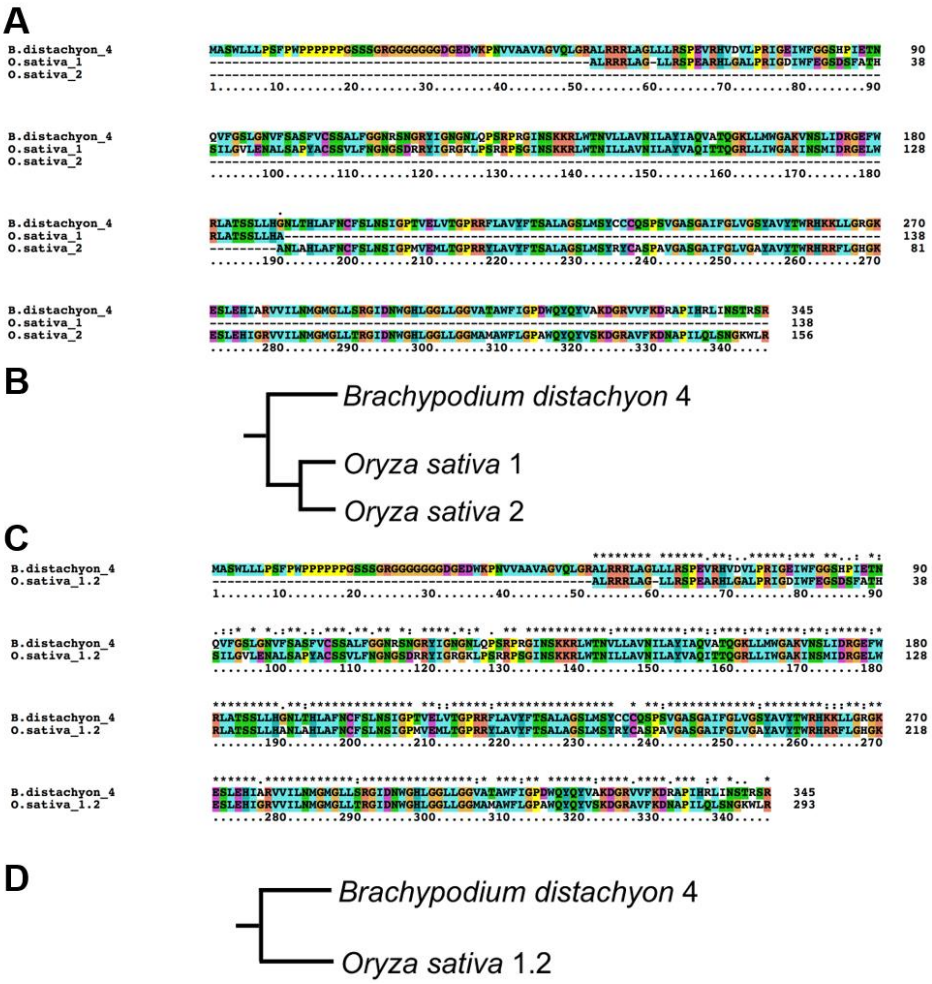


Figure B.4: Successful example of Fragmatic merging together two fragmented sequences with an overlapping alanine at 190. A) Multiple sequence alignment of *B. distachyon*_4, *O. sativa*_1 and *O. sativa*_2 B) Phylogenetic tree of *B. distachyon*_4, *O. sativa*_1 and *O. sativa*_2 C) Multiple sequence alignment of *B. distachyon*_4, *O. sativa*_1.2 D) Phylogenetic tree of *B. distachyon*_4, *O. sativa*_1.2

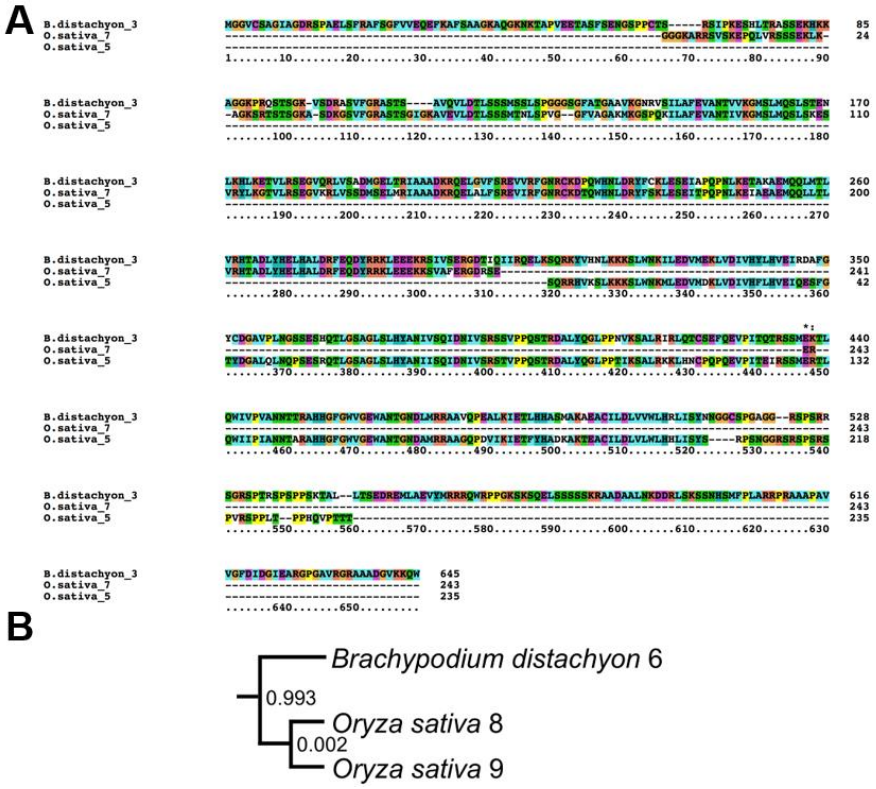


Figure B.5: Two examples of erroneous merging of fragmented sequences. **A)** Error introduced in the multiple sequence alignment at position 447/448. **B)** Error introduced due to low support value at a paired node.

B.4 Discussion

Here I present Fragmatic a novel method for identifying and resolving erroneously fragmented genes in *de novo* assembled transcriptomes. Fragmatic analyses multiple sequence alignments and phylogenetic trees inferred from orthogroups containing putative gene fragments to evaluate whether or not subsets of the sequences in the orthogroup are likely to be fragments generated from the same gene. The performance of Fragmatic was evaluated using both simulated and real sequence data. In tests using simulated sequence data Fragmatic performed very well with 100% precision (correctly joining real fragmented sequences together). However, Fragmatic suffered

on average from very low recall rate of 18.75% (missing many simulated fragmented sequences). Fragmatic also suffered due to errors in the orthogroup inference step where the fragments of real genes failed to be clustered together in the same orthogroup and thus Fragmatic could not detect them. The majority of the low recall rate can be explained by error in tree inference. The simulated fragments did not appear as a monophyletic group and thus Fragmatic did not process them.

Analysis of the simulated fragmented genes revealed that Fragmatic was successful only in the instances where the simulated gene fragments overlapped in the alignment. Simulated fragments that did not overlap were not joined together. There are two possible explanations for this. Firstly, tree inference tools require an overlapping sequence between two fragments to estimate the evolutionary distance without making further assumptions. Secondly, it is possible that there was an error in the Fragmatic code that prevented these genes from being resolved. Had more time been available this would have been pursued to identify the precise reason for this.

Fragmatic was also evaluated using real *de novo* assembled transcriptome data generated using different assembly methods. On these real datasets Fragmatic achieved on average a precision of ~ 79%. The reason for the reduction in precision is caused by the introduction of erroneous merging of fragments. Those errors happened because Fragmatic did not account for gapped alignments between the fragments and did not filter for low support values. These additional checks would have been implemented next to determine their effect on the precision on the method.

Given the number of gene fragments that were identified, if we assume that Fragmatic has the same rate of recall as for the simulated data then we can extrapolate there are

between 1050 and 2800 fragmented genes in each of the *de novo* assembled transcriptomes.

If time had allowed, there are several additional analyses that would have been done. For example, in the analysis presented only 4 species were clustered with the simulated and real assembly data. This number was chosen to be small for algorithm development purposes and hence it would be important to determine whether the performance of the method is influenced (positively or negatively) by the number of species used in the analysis. I anticipate that increasing the number of species will increase the information available to infer the phylogenetic tree and result in more sequences being successfully joined together. It would be necessary to determine the role of both species number and phylogenetic distance on the accuracy of Fragmatic.

Another area where improvement would be necessary is in the simulation of fragmented sequences. It is clear from the comparison of the simulated and real data that Fragmatic performs better on simulated data than it does on real data. This suggests that my data simulation method is too simplistic to capture the full range of possibilities encountered in a *de novo* assembled transcriptome. A more realistic model of sequence fragmentation including allowing more than two fragments per gene may result in convergence in performance characteristics of Fragmatic on real and simulated data.

Other filters should also be tested. For example, filtering the phylogenetic trees such that only monophyletic groups that obtain greater than some minimal support value are then further investigated as putative fragments. In addition, the effect of different multiple sequence alignment and tree inference methods should be tested.

From the simulated dataset we know that 443 fragments were clustered in the same orthogroup however, in the end only 83 were merged together. The reason for this is that the unidentified 360 (approximately 80%) fragment pairs did not appear as a monophyletic group in the phylogenetic tree. Thus, utilising the phylogenetic tree may be too stringent a criterion. Therefore, it might be worthwhile to investigate the performance characteristics of a “tree free” version of Fragmatic. In this version of the algorithm the evidence used to decide whether fragments should be joined together would come solely from the multiple sequence alignment. While it is likely that this method would have a better recall, it may be that the precision of the method is dramatically worse.

B.5 Methods

B.5.1 Algorithm overview

Fragmatic requires as input data a file containing multiple related sequences that may or may not contain erroneously fragmented genes (Figure B.6 Panel A). It creates an initial multiple sequence alignment using MAFFT. If there are 200 or more sequences it uses the FFT-NS-2 method and if there are less than 200 sequences it uses L-INS-i (Figure B.6 Panel B). Fragmatic infers a phylogenetic tree from this alignment using FastTree (Figure B.6 Panel C). Fragmatic then traverses the tree and starts to identify monophyletic groups of sequences where each member of the group comes from the same species (Figure B.6 Panel D; A1.1.1 and A1.1.2, C2.1.1 and C2.1.2). Once such a group is found Fragmatic, evaluates the relationship between these sequences in the multiple sequence alignment. There are two possibilities that could give rise to such monophyletic groups. 1) The genes are recent duplicates that diverge from each other after the split of the species in question from its closest relative in the tree. 2) The genes are fragments of the same gene. In the case of the former the genes will overlap

along their length and those overlapping regions may or may not be identical. In the case of the latter the genes may have a region of overlap whose length is less than the overlap length used in the assembly process, or may occupy different non-overlapping locations in the multiple sequence alignment. If monophyletic sequences are found that satisfy the criteria for being putative gene fragments then they are stitched together and the tree regenerated. The process is repeated until no further fragmented genes are identified (Figure B.6 Panel E).

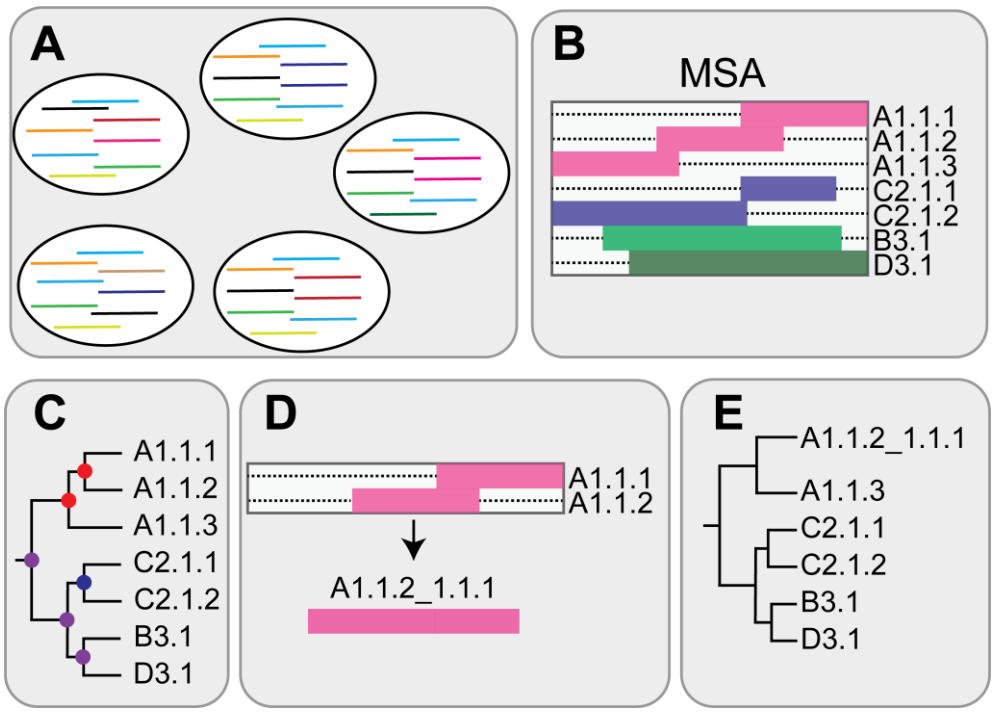


Figure B.6: Workflow of Fragmatic depicted in five steps: **A**) Input data consisting of clustered sequences **B**) Creating multiple sequence alignment file **C**) Infer phylogenetic tree **D**) Collecting evidence to merge multiple sequences together **E**) Iterate as long as the topology of the tree changes. Output the simplified and improved data set.

B.5.2 Simulation of fragmented sequences

To create simulated fragmented sequences, the primary transcript sequences on Phytozome V11 of following species were considered: *Oryza sativa* (323 V7), *Brachypodium distachyon* (314 V3.1), *Setaria italic* (312 V2.2), and *Sorghum bicolor* (313 V3.1). Fragmented sequences were simulated in two different ways. Firstly, as fragmented with an overlapping sequence shorter than the sequence length required by the *de novo* assembler to put two sequences together, and secondly as fragments containing a gap between the sequences. For each of the datasets a total of 1000 sequences were randomly sampled which were required to be longer than 140 amino acids long. About half of the sequences sampled was used to create overlapping fragments, the other half was used to create gapped fragments. Overlapping sequences were constructed so that the minimal sequence length is at least 74 amino acids long. The length of the overlapping sequence was randomly selected as an integer number between 1 and 13. Gapped sequences were created in such a way that both the sequences would have a length of at least 74 amino acids. Thus, the gap size could be any random integer number between 1 and the total sequence length subtracted by twice the minimal fragment length. Fragmented sequences were labelled in such a way that in case two sequences were merged together, it was immediately clear whether the sequences originated from the same sequence and whether the fragments were correctly merged.

B.5.3 Prediction of open reading frames in *de novo* assembled transcriptomes

Open reading frames in the *de novo* assembled transcriptomes were predicted using TransDecoder (Haas et al. 2013) (Version 2.0.1) and Prodigal (Hyatt et al. 2010a) (Version 2.6.3). In cases where both methods predicted an open reading frame for a

transcript then priority was given to the prediction made by TransDecoder. TransDecoder was run in default mode. Prodigal was run with the genetic code set to eukaryotic default.

B.5.4 Inference of orthogroups using OrthoFinder

Predicted amino acid sequences from each species were subject to orthogroup inference using OrthoFinder(Emms and Kelly 2015) (v 0.4.0) using the default program settings. Orthogroups containing four or more sequences (the minimal number of sequences required to infer a phylogenetic tree) were subject to analysis using Fragmatic.

B.5.5 Evaluation of simulated fragmented sequences

In order to evaluate whether the simulated fragments were merged together correctly, a labelling system was put in place. This labelling system indicates from which transcript the simulated fragments originate, whether it is the fragment at the 5’ or 3’ end and the size of the overlapping or gapped region between the two fragments. This additional information was added to the FASTA header of the fragments. This allowed for easy evaluation of the performance of Fragmatic by analysing the sequence names of the sequences that were joined together.

B.5.6 Evaluation of real transcriptome sequences

Three different publicly available *de novo* transcriptome assemblies of the *Oryza sativa* transcriptome were downloaded. Each of these assemblies was generated from the same input read data using different assembly methods: Trinity (Grabherr et al. 2011), Oases (Schulz et al. 2012), and SOAPdenovo-Trans (Xie et al. 2014).

For the assemblies generated using Oases only transcripts that had a confidence score of 1.0 were selected for further analysis. A BLAST database was created using the *Oryza sativa* reference proteome (Phytozome V11 323 Version 7.0 primary transcripts only).

B.6 Software availability

Fragmatic was written in python 2.7 and relies on Biopython(Cock et al. 2009) and ETE 3(Huerta-Cepas et al. 2016) a python toolkit to work with phylogenetic trees. Additional software implemented were MAFFT(Katoh and Standley 2013) (v7.058b) and LINSI(Katoh and Standley 2013) (v7.058b) for creating the multiple sequence alignment. For tree inference FastTree(Price et al. 2010) (v2.1.4) was implemented.

References

Aberer AJ, Pattengale ND, Stamatakis A (2010) Parallelized phylogenetic post-analysis on multi-core architectures Journal of Computational Science 1:107-114 doi:http://dx.doi.org/10.1016/j.jocs.2010.03.006

Alonso-Cantabrana H, Cousins AB, Danila F, Ryan T, Sharwood RE, von Caemmerer S, Furbank RT (2018) Diffusion of CO_2 across the Mesophyll-Bundle Sheath Cell Interface in a C_4 Plant with Genetically Reduced PEP Carboxylase Activity Plant Physiology 178:72-81 doi:10.1104/pp.18.00618

Alvarez CE et al. (2019) Molecular adaptations of NADP-malic enzyme for its function in C_4 photosynthesis in grasses Nature Plants 5:755-765 doi:10.1038/s41477-019-0451-7

Amthor JS, Bar-Even A, Hanson AD, Millar AH, Stitt M, Sweetlove LJ, Tyerman SD (2019) Engineering Strategies to Boost Crop Productivity by Cutting Respiratory Carbon Loss The Plant Cell 31:297-314 doi:10.1105/tpc.18.00743

Anderson PJC Registry of Standard Biological Parts Anderson promoter collection Available: http://parts.igem.org/Promoters/Catalog/Anderson. Accessed 26. February 2018 2018

Andersson I, Knight S, Schneider G, Lindqvist Y, Lundqvist T, Brändén C-I, Lorimer GH (1989) Crystal structure of the active site of ribulose-bisphosphate carboxylase Nature 337:229-234 doi:10.1038/337229a0

Ashton AR, Furbank RT, Trevanion SJ (1999) Antisense RNA inhibition of pyruvate, orthophosphate dikinase and NADP malate dehydrogenase in the C_4 plant *Flaveria bidentis*: analysis of plants with a mosaic phenotype Functional Plant Biology 26:537-547

Atkinson N et al. (2016) Introducing an algal carbon-concentrating mechanism into higher plants: location and incorporation of key components Plant Biotechnology Journal 14:1302-1315 doi:10.1111/pbi.12497

Baba T et al. (2006) Construction of *Escherichia coli* K-12 in-frame, single-gene knockout mutants: the Keio collection Mol Syst Biol 2:2006.0008-2006.0008 doi:10.1038/msb4100050

Bauwe H, Chollet R (1986) Kinetic properties of phosphoenolpyruvate carboxylase from C_3, C_4, and C_3-C_4 intermediate species of *Flaveria* (Asteraceae) Plant Physiology 82:695-699

Bellasio C (2016) A generalized stoichiometric model of C_3, C_2, C_2+C_4, and C_4 photosynthetic metabolism Journal of Experimental Botany 68:269-282 doi:10.1093/jxb/erw303

Bezrutczyk M et al. (2018) Impaired phloem loading in zmsweet13a,b,c sucrose transporter triple knock-out mutants in *Zea mays* New Phytologist 218:594-603 doi:doi:10.1111/nph.15021

Bianconi ME, Dunning LT, Moreno-Villena JJ, Osborne CP, Christin P-A (2018) Gene duplication and dosage effects during the early emergence of C_4 photosynthesis in the grass genus *Alloteropsis* Journal of Experimental Botany 69:1967-1980 doi:10.1093/jxb/ery029

Black Jr CC (1973) Photosynthetic carbon fixation in relation to net CO_2 uptake Annual Review of Plant Physiology 24:253-286

Bläsing OE, Westhoff P, Svensson P (2000) Evolution of C_4 Phosphoenolpyruvate Carboxylase in *Flaveria*, a Conserved Serine Residue in the Carboxyl-terminal Part of the Enzyme is a Major Determinant for C_4-specific Characteristics Journal of Biological Chemistry 275:27917-27923

Boetzer M, Pirovano W (2014) SSPACE-LongRead: scaffolding bacterial draft genomes using long read sequence information BMC Bioinformatics 15:211

Bolger AM, Lohse M, Usadel B (2014) Trimmomatic: a flexible trimmer for Illumina sequence data Bioinformatics 30:2114-2120

Borba AR et al. (2018) Synergistic Binding of bHLH Transcription Factors to the Promoter of the Maize NADP-ME Gene Used in C_4 Photosynthesis is Based on an Ancient Code

Found in the Ancestral C_3 State Molecular Biology and Evolution 35:1690-1705 doi:10.1093/molbev/msy060
Borland AM, Griffiths H, Hartwell J, Smith JAC (2009) Exploiting the potential of plants with crassulacean acid metabolism for bioenergy production on marginal lands Journal of Experimental Botany 60:2879-2896
Botha C (1992) Plasmodesmatal distribution, structure and frequency in relation to assimilation in C_3 and C_4 grasses in southern Africa Planta 187:348-358
Botha C, Van Bel A (1992) Quantification of symplastic continuity as visualised by plasmodesmograms: diagnostic value for phloem-loading pathways Planta 187:359-366
Bräutigam A, Hoffmann-Benning S, Weber APM (2008) Comparative Proteomics of Chloroplast Envelopes from C_3 and C_4 Plants Reveals Specific Adaptations of the Plastid Envelope to C_4 Photosynthesis and Candidate Proteins Required for Maintaining C_4 Metabolite Fluxes Plant Physiology 148:568-579 doi:10.1104/pp.108.121012
Bräutigam A et al. (2011) An mRNA Blueprint for C_4 Photosynthesis Derived from Comparative Transcriptomics of Closely Related C_3 and C_4 Species Plant Physiology 155:142-156 doi:10.1104/pp.110.159442
Brown NJ, Newell CA, Stanley S, Chen JE, Perrin AJ, Kajala K, Hibberd JM (2011) Independent and Parallel Recruitment of Preexisting Mechanisms Underlying C_4 Photosynthesis Science 331:1436-1439 doi:10.1126/science.1201248
Brutnell TP et al. (2010) *Setaria viridis*: a model for C_4 photosynthesis The Plant Cell 22:2537-2544
Buchanan BB, Gruissem W, Jones RL (2000) Biochemistry & molecular biology of plants. . American Society of Plant Physiologists,
Cain JA, Solis N, Cordwell SJ (2014) Beyond gene expression: The impact of protein post-translational modifications in bacteria Journal of Proteomics 97:265-286 doi:https://doi.org/10.1016/j.jprot.2013.08.012
Charif D, Lobry JR (2007) SeqinR 1.0-2: A Contributed Package to the R Project for Statistical Computing Devoted to Biological Sequences Retrieval and Analysis. In: Bastolla U, Porto M, Roman HE, Vendruscolo M (eds) Structural Approaches to Sequence Evolution: Molecules, Networks, Populations. Springer Berlin Heidelberg, Berlin, Heidelberg, pp 207-232. doi:10.1007/978-3-540-35306-5_10
Chastain CJ (2011) Chapter 15 Structure, Function, and Post-translational Regulation of C_4 Pyruvate Orthophosphate Dikinase. In: Raghavendra AS, Sage RF (eds) C_4 Photosynthesis and Related CO_2 Concentrating Mechanisms. Springer Netherlands, Dordrecht, pp 301-315. doi:10.1007/978-90-481-9407-0_15
Chastain CJ, Failing CJ, Manandhar L, Zimmerman MA, Lakner MM, Nguyen THT (2011) Functional evolution of C_4 pyruvate, orthophosphate dikinase Journal of Experimental Botany 62:3083-3091 doi:10.1093/jxb/err058
Chastain CJ et al. (2002) Pyruvate,orthophosphate dikinase in leaves and chloroplasts of C_3 plants undergoes light-/dark-induced reversible phosphorylation Plant physiology 128:1368-1378 doi:10.1104/pp.010806
Chatterjee A et al. (2011) *Saccharomyces cerevisiae* THI4p is a suicide thiamine thiazole synthase Nature 478:542 doi:10.1038/nature10503
Christin P-A, Arakaki M, Osborne CP, Edwards EJ (2015) Genetic Enablers Underlying the Clustered Evolutionary Origins of C_4 Photosynthesis in Angiosperms Molecular Biology and Evolution 32:846-858 doi:10.1093/molbev/msu410
Christin P-A, Besnard G, Samaritani E, Duvall MR, Hodkinson TR, Savolainen V, Salamin N (2008a) Oligocene CO_2 decline promoted C_4 photosynthesis in grasses Current Biology 18:37-43
Christin P-A et al. (2012) Adaptive evolution of C_4 photosynthesis through recurrent lateral gene transfer Current Biology 22:445-449

Christin P-A, Osborne CP, Sage RF, Arakaki M, Edwards EJ (2011a) C_4 eudicots are not younger than C_4 monocots Journal of Experimental Botany 62:3171-3181 doi:10.1093/jxb/err041
Christin P-A, Petitpierre B, Salamin N, Büchi L, Besnard G (2009a) Evolution of C_4 Phosphoenolpyruvate Carboxykinase in Grasses, from Genotype to Phenotype Molecular Biology and Evolution 26:357-365 doi:10.1093/molbev/msn255
Christin P-A, Salamin N, Muasya AM, Roalson EH, Russier F, Besnard G (2008b) Evolutionary Switch and Genetic Convergence on rbcL following the Evolution of C_4 Photosynthesis Molecular Biology and Evolution 25:2361-2368 doi:10.1093/molbev/msn178
Christin P-A, Salamin N, Savolainen V, Duvall MR, Besnard G (2007) C_4 Photosynthesis Evolved in Grasses via Parallel Adaptive Genetic Changes Current Biology 17:1241-1247 doi:http://dx.doi.org/10.1016/j.cub.2007.06.036
Christin P-A, Samaritani E, Petitpierre B, Salamin N, Besnard G (2009b) Evolutionary Insights on C_4 Photosynthetic Subtypes in Grasses from Genomics and Phylogenetics Genome Biology and Evolution 1:221-230 doi:10.1093/gbe/evp020
Christin PA, Sage TL, Edwards EJ, Ogburn RM, Khoshravesh R, Sage RF (2011b) Complex evolutionary transitions and the significance of C_3–C_4 intermediate forms of photosynthesis in *Molluginaceae* Evolution: International Journal of Organic Evolution 65:643-660
Chuong SDX, Franceschi VR, Edwards GE (2006) The Cytoskeleton Maintains Organelle Partitioning Required for Single-Cell C_4 Photosynthesis in *Chenopodiaceae* Species The Plant Cell 18:2207-2223 doi:10.1105/tpc.105.036186
Clayton H, Saladié M, Rolland V, Sharwood R, Macfarlane T, Ludwig M (2017) Loss of the Chloroplast Transit Peptide from an Ancestral C_3 Carbonic Anhydrase Is Associated with C_4 Evolution in the Grass Genus *Neurachne* Plant Physiology 173:1648-1658 doi:10.1104/pp.16.01893
Cock PJA et al. (2009) Biopython: freely available Python tools for computational molecular biology and bioinformatics Bioinformatics 25:1422-1423 doi:10.1093/bioinformatics/btp163
Coordinators NR (2016) Database resources of the National Center for Biotechnology Information Nucleic Acids Res 44:D7-D19 doi:10.1093/nar/gkv1290
Cowan I Economics of carbon fixation in higher plants. In: On the economy of plant form and function: proceedings of the Sixth Maria Moors Cabot Symposium, Evolutionary Constraints on Primary Productivity, Adaptive Patterns of Energy Capture in Plants, Harvard Forest, August 1983, 1986. Cambridge [Cambridgeshire]: Cambridge University Press, c1986.,
Cozens C, Pinheiro VB (2018) Darwin Assembly: fast, efficient, multi-site bespoke mutagenesis Nucleic Acids Res 46:e51-e51 doi:10.1093/nar/gky067
Crusoe MR et al. (2015) The khmer software package: enabling efficient nucleotide sequence analysis F1000Research 4
Danila F, Quick WP, White RG, Furbank RT, von Caemmerer S (2016) The metabolite pathway between bundle sheath and mesophyll: quantification of plasmodesmata in leaves of C_3 and C_4 monocots The Plant Cell:tpc. 00155.02016
Davies SJ et al. (1999) Inactivation and Regulation of the Aerobic C_4-Dicarboxylate Transport (dctA) Gene of *Escherichia coli* Journal of Bacteriology 181:5624-5635
Dean C, Pichersky E, Dunsmuir P (1989) Structure, evolution, and regulation of RbcS genes in higher plants Annual review of plant biology 40:415-439
Douce R, Bourguignon J, Neuburger M, Rébeillé F (2001) The glycine decarboxylase system: a fascinating complex Trends in Plant Science 6:167-176 doi:https://doi.org/10.1016/S1360-1385(01)01892-1
Douglas NA, Manos PS (2007) Molecular phylogeny of Nyctaginaceae: taxonomy, biogeography, and characters associated with a radiation of xerophytic genera in North America American Journal of Botany 94:856-872

Dunning LT et al. (2017) Introgression and repeated co-option facilitated the recurrent emergence of C_4 photosynthesis among close relatives Evolution 71:1541-1555 doi:doi:10.1111/evo.13250
Edwards GE, Franceschi VR, Voznesenskaya EV (2004) Single-cell C_4 photosynthesis versus the dual-cell (Kranz) paradigm Annu Rev Plant Biol 55:173-196
Ehleringer J, Pearcy RW (1983) Variation in quantum yield for CO_2 uptake among C_3 and C_4 plants Plant Physiology 73:555-559
Ehleringer JR, Cerling TE, Helliker BR (1997) C_4 photosynthesis, atmospheric CO_2, and climate Oecologia 112:285-299
Ehleringer JR, Sage RF, Flanagan LB, Pearcy RW (1991) Climate change and the evolution of C_4 photosynthesis Trends in Ecology & Evolution 6:95-99 doi:https://doi.org/10.1016/0169-5347(91)90183-X
Ellis RJ (1979) The most abundant protein in the world Trends in Biochemical Sciences 4:241-244 doi:https://doi.org/10.1016/0968-0004(79)90212-3
Emms DM, Covshoff S, Hibberd JM, Kelly S (2016) Independent and Parallel Evolution of New Genes by Gene Duplication in Two Origins of C_4 Photosynthesis Provides New Insight into the Mechanism of Phloem Loading in C_4 Species Molecular Biology and Evolution 33:1796-1806 doi:10.1093/molbev/msw057
Emms DM, Kelly S (2015) OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy Genome biology 16:1-14
Engel N et al. (2007) Deletion of Glycine Decarboxylase in Arabidopsis Is Lethal under Nonphotorespiratory Conditions Plant Physiology 144:1328-1335 doi:10.1104/pp.107.099317
Engelmann S, Bläsing OE, Gowik U, Svensson P, Westhoff P (2003) Molecular evolution of C_4 phosphoenolpyruvate carboxylase in the genus *Flaveria*-a gradual increase from C_3 to C_4 characteristics Planta 217:717-725 doi:10.1007/s00425-003-1045-0
Engelmann S, Bläsing OE, Westhoff P, Svensson P (2002) Serine 774 and amino acids 296 to 437 comprise the major C_4 determinants of the C_4 phosphoenolpyruvate carboxylase of *Flaveria trinervia* FEBS Letters 524:11-14 doi:10.1016/S0014-5793(02)02975-7
Engler C, Kandzia R, Marillonnet S (2008) A One Pot, One Step, Precision Cloning Method with High Throughput Capability PLOS ONE 3:e3647 doi:10.1371/journal.pone.0003647
Engler C et al. (2014) A golden gate modular cloning toolbox for plants ACS Synth Biol 3:839-843 doi:10.1021/sb4001504
Ermakova M, Lopez-Calcagno PE, Raines CA, Furbank RT, von Caemmerer S (2019) Overexpression of the Rieske FeS protein of the Cytochrome b6f complex increases C_4 photosynthesis in *Setaria viridis* Communications Biology 2:314 doi:10.1038/s42003-019-0561-9
Evans J, von Caemmerer S (2000) Would C_4 rice produce more biomass than C_3 rice? In: Studies in Plant Science, vol 7. Elsevier, pp 53-71
Feodorova TA, Voznesenskaya EV, Edwards GE, Roalson EH (2010) Biogeographic Patterns of Diversification and the Origins of C_4 in Cleome (Cleomaceae) Systematic Botany 35:811-826 doi:10.1600/036364410X539880
Ferro M et al. (2010) AT_CHLORO, a Comprehensive Chloroplast Proteome Database with Subplastidial Localization and Curated Information on Envelope Proteins Molecular & Cellular Proteomics 9:1063-1084 doi:10.1074/mcp.M900325-MCP200
Fischer K, Kammerer B, Gutensohn M, Arbinger B, Weber A, Häusler RE, Flügge UI (1997) A new class of plastidic phosphate translocators: a putative link between primary and secondary metabolism by the phosphoenolpyruvate/phosphate antiporter The Plant Cell 9:453-462 doi:10.1105/tpc.9.3.453
Fisher RA (1999) The genetical theory of natural selection: a complete variorum edition. Oxford University Press,
Flügge U-I (1999) Phosphate translocators in plastids Annual review of plant biology 50:27-45

Fried L, Behr S, Jung K (2013) Identification of a Target Gene and Activating Stimulus for the YpdA/YpdB Histidine Kinase/Response Regulator System in *Escherichia coli* Journal of Bacteriology 195:807-815 doi:10.1128/jb.02051-12
Furumoto T, Izui K, Quinn V, Furbank RT, von Caemmerer S (2007) Phosphorylation of phosphoenolpyruvate carboxylase is not essential for high photosynthetic rates in the C_4 species *Flaveria bidentis* Plant Physiology 144:1936-1945
Furumoto T et al. (2011) A plastidial sodium-dependent pyruvate transporter Nature 476:472
Gagolewski M, Bartek T (2015) R package stringi: Character string processing facilities doi:10.5281/zenodo.19071
Geddes BA, Mendoza-Suárez MA, Poole PS (2019) A Bacterial Expression Vector Archive (BEVA) for Flexible Modular Assembly of Golden Gate-Compatible Vectors Frontiers in Microbiology 9 doi:10.3389/fmicb.2018.03345
Geerlof A (2010) M9 mineral medium. https://www.helmholtz-muenchen.de/fileadmin/PEPF/Protocols/M9-medium_150510.pdf. Accessed 20 June 2019
Ghannoum O, Evans JR, Chow WS, Andrews TJ, Conroy JP, von Caemmerer S (2005) Faster Rubisco Is the Key to Superior Nitrogen-Use Efficiency in NADP-Malic Enzyme Relative to NAD-Malic Enzyme C_4 Grasses Plant Physiology 137:638-650 doi:10.1104/pp.104.054759
Gnerre S et al. (2011) High-quality draft assemblies of mammalian genomes from massively parallel sequence data Proceedings of the National Academy of Sciences 108:1513-1518
Göhler A-K et al. (2011) More than just a metabolic regulator--elucidation and validation of new targets of PdhR in Escherichia coli BMC Syst Biol 5:197-197 doi:10.1186/1752-0509-5-197
Goodstein DM et al. (2012) Phytozome: a comparative platform for green plant genomics Nucleic Acids Res 40:D1178-D1186 doi:10.1093/nar/gkr944
Gowik U, Westhoff P (2011) The Path from C_3 to C_4 Photosynthesis Plant Physiology 155:56-63 doi:10.1104/pp.110.165308
Goyer A (2010) Thiamine in plants: Aspects of its metabolism and functions Phytochemistry 71:1615-1624 doi:https://doi.org/10.1016/j.phytochem.2010.06.022
Grabherr MG et al. (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome 29:644-652
Grant KC (2019) Engineering rhizobacteria as synthetic biology chassis. University of Oxford
Grass Phylogeny Working Group II (2012) New grass phylogeny resolves deep evolutionary relationships and discovers C_4 origins New Phytologist 193:304-312
Grenier F, Matteau D, Baby V, Rodrigue S (2014) Complete Genome Sequence of Escherichia coli BW25113 Genome Announc 2:e01038-01014 doi:10.1128/genomeA.01038-14
Grosse-Holz F et al. (2018) Three unrelated protease inhibitors enhance accumulation of pharmaceutical recombinant proteins in *Nicotiana benthamiana* Plant biotechnology journal 16:1797-1810 doi:10.1111/pbi.12916
Haas BJ et al. (2013) *De novo* transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis Nature protocols 8:1494-1512
Hall LN, Rossini L, Cribb L, Langdale JA (1998) GOLDEN 2: a novel transcriptional regulator of cellular differentiation in the maize leaf The Plant Cell 10:925-936
Han H, Felker P (1997) Field validation of water-use efficiency of the CAM plant *Opuntia ellisianain* south Texas Journal of Arid Environments 36:133-148
Hartley JL, Temple GF, Brasch MA (2000) DNA Cloning Using In Vitro Site-Specific Recombination Genome Research 10:1788-1795 doi:10.1101/gr.143000
Hassan N, Thiede J, Liede-Schumann S (2005) Phylogenetic analysis of *Sesuvioideae* (Aizoaceae) inferred from nrDNA internal transcribed spacer (ITS) sequences and morphological data Plant Systematics and Evolution 255:121-143
Hatch MD (1987) C4 photosynthesis: a unique elend of modified biochemistry, anatomy and ultrastructure Biochimica et Biophysica Acta (BBA)-Reviews on Bioenergetics 895:81-106

Hatch MD, Slack CR (1966) Photosynthesis by sugar-cane leaves: a new carboxylation reaction and the pathway of sugar formation Biochemical Journal 101:103
Hibberd JM, Sheehy JE, Langdale JA (2008) Using C_4 photosynthesis to increase the yield of rice rationale and feasibility Current opinion in plant biology 11:228-231
Huang P, Studer AJ, Schnable JC, Kellogg EA, Brutnell TP (2017) Cross species selection scans identify components of C_4 photosynthesis in the grasses Journal of Experimental Botany doi:10.1093/jxb/erw256
Huerta-Cepas J, Serra F, Bork P (2016) ETE 3: Reconstruction, analysis and visualization of phylogenomic data Molecular Biology and Evolution doi:10.1093/molbev/msw046
Hwang S, Choe D, Yoo M, Cho S, Kim SC, Cho S, Cho B-K (2018) Peptide Transporter CstA Imports Pyruvate in Escherichia coli K-12 Journal of Bacteriology 200:e00771-00717 doi:10.1128/jb.00771-17
Hyatt D, Chen G-L, LoCascio PF, Land ML, Larimer FW, Hauser LJ (2010a) Prodigal: prokaryotic gene recognition and translation initiation site identification BMC Bioinformatics 11:1-11 doi:10.1186/1471-2105-11-119
Hyatt D, Chen GL, Locascio PF, Land ML, Larimer FW, Hauser LJ (2010b) Prodigal: prokaryotic gene recognition and translation initiation site identification BMC Bioinformatics 11 doi:10.1186/1471-2105-11-119
Jacobs B, Engelmann S, Westhoff P, Gowik U (2008) Evolution of C_4 phosphoenolpyruvate carboxylase in *Flaveria*: determinants for high tolerance towards the inhibitor L-malate Plant, Cell & Environment 31:793-803 doi:10.1111/j.1365-3040.2008.01796.x
Jeong H et al. (2009) Genome Sequences of *Escherichia coli* B strains REL606 and BL21(DE3) Journal of Molecular Biology 394:644-652 doi:https://doi.org/10.1016/j.jmb.2009.09.052
John CR, Smith-Unna RD, Woodfield H, Covshoff S, Hibberd JM (2014) Evolutionary convergence of cell-specific gene expression in independent lineages of C_4 grasses Plant Physiology 165:62-75
Johnson MT et al. (2012) Evaluating methods for isolating total RNA and predicting the success of sequencing phylogenetically diverse plant transcriptomes PloS one 7:e50226
Kadereit G, Mavrodiev EV, Zacharias EH, Sukhorukov AP (2010) Molecular phylogeny of *Atripliceae* (*Chenopodioideae, Chenopodiaceae*): implications for systematics, biogeography, flower and fruit evolution, and the origin of C_4 photosynthesis American Journal of Botany 97:1664-1687
Kai Y, Matsumura H, Izui K (2003) Phosphoenolpyruvate carboxylase: three-dimensional structure and molecular mechanisms Archives of biochemistry and biophysics 414:170-179 doi:10.1016/s0003-9861(03)00170-x
Katoh K, Standley DM (2013) MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability Molecular Biology and Evolution 30:772-780 doi:10.1093/molbev/mst010
Katoh K, Toh H (2010) Parallelization of the MAFFT multiple sequence alignment program Bioinformatics 26:1899-1900 doi:10.1093/bioinformatics/btq224
Kellogg EA (1999) Phylogenetic aspects of the evolution of C_4 photosynthesis C_4 plant biology 12:411-444
Kelly S (2018) The Amount of Nitrogen Used for Photosynthesis Modulates Molecular Evolution in Plants Molecular Biology and Evolution 35:1616-1625 doi:10.1093/molbev/msy043
Kelly S et al. (2017) Wide sampling of natural diversity identifies novel molecular signatures of C_4 photosynthesis bioRxiv doi:10.1101/163097
Kenrick P, Crane PR (1997) The origin and early evolution of plants on land Nature 389:33
Kinoshita H et al. (2011) The chloroplastic 2-oxoglutarate/malate transporter has dual function as the malate valve and in carbon/nitrogen metabolism The Plant Journal 65:15-26 doi:10.1111/j.1365-313X.2010.04397.x
Kinsman E, Pyke K (1998) Bundle sheath cells and cell-specific plastid development in *Arabidopsis* leaves Development 125:1815-1822

Kopylova E, Noé L, Touzet H (2012) SortMeRNA: fast and accurate filtering of ribosomal RNAs in metatranscriptomic data Bioinformatics 28:3211-3217
Kornberg HL (1966) The role and control of the glyoxylate cycle in Escherichia coli Biochem J 99:1-11 doi:10.1042/bj0990001
Kristoficova I, Vilhena C, Behr S, Jung K (2018) BtsT, a Novel and Specific Pyruvate/H+ Symporter in *Escherichia coli* Journal of Bacteriology 200:e00599-00517 doi:10.1128/jb.00599-17
Kromdijk J, Głowacka K, Leonelli L, Gabilly ST, Iwai M, Niyogi KK, Long SP (2016) Improving photosynthesis and crop productivity by accelerating recovery from photoprotection Science 354:857-861 doi:10.1126/science.aai8878
Laisk A, Edwards GE (2000) A mathematical model of C_4 photosynthesis: The mechanism of concentrating CO_2 in NADP-malic enzyme type species Photosynthesis Research 66:199 doi:10.1023/a:1010695402963
Lambret-Frotté J, de Almeida LCS, de Moura SM, Souza FLF, Linhares FS, Alves-Ferreira M (2015) Validating Internal Control Genes for the Accurate Normalization of qPCR Expression Analysis of the Novel Model Plant *Setaria viridis* PloS one 10:e0135006-e0135006 doi:10.1371/journal.pone.0135006
Langdale JA (2011) C_4 Cycles: Past, Present, and Future Research on C_4 Photosynthesis The Plant Cell 23:3879-3892 doi:10.1105/tpc.111.092098
Lange PR, Geserick C, Tischendorf G, Zrenner R (2008) Functions of Chloroplastic Adenylate Kinases in Arabidopsis Plant Physiology 146:492-504 doi:10.1104/pp.107.114702
Langmead B, Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2 Nature methods 9:357
Lauterbach M, Schmidt H, Billakurthi K, Hankeln T, Westhoff P, Gowik U, Kadereit G (2017) *De novo* Transcriptome Assembly and Comparison of C_3, C_3-C_4, and C_4 Species of Tribe *Salsoleae* (*Chenopodiaceae*) Frontiers in Plant Science 8:1939 doi:10.3389/fpls.2017.01939
Lefebvre S, Lawson T, Fryer M, Zakhleniuk OV, Lloyd JC, Raines CA (2005) Increased sedoheptulose-1, 7-bisphosphatase activity in transgenic tobacco plants stimulates photosynthesis and growth from an early stage in development Plant Physiology 138:451-460
Lesley JC, Patrick JB, Claudia V, Simon J (2008) An Approach to transcriptome analysis of non-model organisms using short-read sequences. In: Genome Informatics 2008. IMPERIAL COLLEGE PRESS, pp 3-14
Levin DA, Peres Y, Wilmer EL (2009) Markov Chains and Mixing Times. American Mathematical Society,
Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N (2009a) The Sequence Alignment/Map format and SAMtools Bioinformatics 25 doi:10.1093/bioinformatics/btp352
Li Z, Wakao S, Fischer BB, Niyogi KK (2009b) Sensing and Responding to Excess Light Annual Review of Plant Biology 60:239-260 doi:10.1146/annurev.arplant.58.032806.103844
Lin MT, Occhialini A, Andralojc PJ, Devonshire J, Hines KM, Parry MA, Hanson MR (2014a) β-Carboxysomal proteins assemble into highly organized structures in *Nicotiana* chloroplasts The Plant Journal 79:1-12
Lin MT, Occhialini A, Andralojc PJ, Parry MAJ, Hanson MR (2014b) A faster Rubisco with potential to increase photosynthesis in crops Nature 513:547-550 doi:10.1038/nature13776
Long SP, Marshall-Colon A, Zhu X-G (2015) Meeting the global food demand of the future by engineering crop photosynthesis and yield potential Cell 161:56-66
Love MI, Huber W, Anders S (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2 Genome Biology 15:550 doi:10.1186/s13059-014-0550-8
Lundgren MR et al. (2016) Evolutionary implications of C_3–C_4 intermediates in the grass *Alloteropsis semialata* Plant, cell & environment 39:1874-1885

Lunn JE et al. (2006) Sugar-induced increases in trehalose 6-phosphate are correlated with redox activation of ADPglucose pyrophosphorylase and higher rates of starch synthesis in *Arabidopsis thaliana* Biochemical Journal 397:139-148

Lyu M-JA et al. (2015) RNA-Seq based phylogeny recapitulates previous phylogeny of the genus *Flaveria* (*Asteraceae*) with some modifications BMC Evolutionary Biology 15:116

Majeran W, van Wijk KJ (2009) Cell-type-specific differentiation of chloroplasts in C_4 plants Trends in Plant Science 14:100-109 doi:https://doi.org/10.1016/j.tplants.2008.11.006

Majeran W, Zybailov B, Ytterberg AJ, Dunsmore J, Sun Q, van Wijk KJ (2008) Consequences of C_4 differentiation for chloroplast membrane proteomes in maize mesophyll and bundle sheath cells Molecular & Cellular Proteomics 7:1609-1638

Mann DGJ et al. (2012) Gateway-compatible vectors for high-throughput gene functional analysis in switchgrass (Panicum virgatum L.) and other monocot species Plant Biotechnology Journal 10:226-236 doi:10.1111/j.1467-7652.2011.00658.x

Matsumura H et al. (2002) Crystal Structures of C_4 Form Maize and Quaternary Complex of E. coli Phosphoenolpyruvate Carboxylases Structure 10:1721-1730 doi:http://dx.doi.org/10.1016/S0969-2126(02)00913-9

McKown AD, Moncalvo JM, Dengler NG (2005) Phylogeny of *Flaveria* (*Asteraceae*) and inference of C_4 photosynthesis evolution American Journal of Botany 92:1911-1928

Meng J, Wang B, Liu D, Chen T, Wang Z, Zhao X (2016) High-yield anaerobic succinate production by strategically regulating multiple metabolic pathways based on stoichiometric maximum in Escherichia coli Microbial Cell Factories 15:141 doi:10.1186/s12934-016-0536-1

Meyer MT, McCormick AJ, Griffiths H (2016) Will an algal CO_2-concentrating mechanism work in higher plants? Current opinion in plant biology 31:181-188 doi:10.1016/j.pbi.2016.04.009

Moreno-Villena JJ, Dunning LT, Osborne CP, Christin P-A (2017) Highly expressed genes are preferentially co-opted for C_4 photosynthesis Molecular Biology and Evolution:msx269-msx269 doi:10.1093/molbev/msx269

Muhaidat R, Sage RF, Dengler NG (2007) Diversity of Kranz anatomy and biochemistry in C_4 eudicots American Journal of Botany 94:362-381

Naylor RL, Falcon WP, Goodman RM, Jahn MM, Sengooba T, Tefera H, Nelson RJ (2004) Biotechnology in the developing world: a case for increased investments in orphan crops Food Policy 29:15-44 doi:https://doi.org/10.1016/j.foodpol.2004.01.002

Nelson T, Langdale JA (1989) Patterns of leaf development in C_4 plants The Plant Cell 1:3

Nikolenko SI, Korobeynikov AI, Alekseyev MA BayesHammer: Bayesian clustering for error correction in single-cell sequencing. In: BMC genomics, 2013. vol 1. BioMed Central, p S7

O'Leary MH (1988) Carbon Isotopes in Photosynthesis: Fractionation techniques may reveal new aspects of carbon dynamics in plants BioScience 38:328-336 doi:10.2307/1310735

Osborne CP, Beerling DJ (2006) Nature's green revolution: the remarkable evolutionary rise of C_4 plants Philosophical Transactions of the Royal Society B: Biological Sciences 361:173-194 doi:10.1098/rstb.2005.1737

Osborne CP, Sack L (2012) Evolution of C_4 plants: a new hypothesis for an interaction of CO_2 and water relations mediated by plant hydraulics Phil Trans R Soc B 367:583-600

Ossowski S et al. (2010) The rate and molecular spectrum of spontaneous mutations in Arabidopsis thaliana science 327:92-94

Pagès H, Aboyoun P, Gentleman R, DebRoy S (2018) Biostrings: Efficient manipulation of biological strings doi:10.18129/B9.bioc.Biostrings

Pal C, Papp B, Lercher MJ (2006) An integrated view of protein evolution Nat Rev Genet 7:337-348

Paliy O, Gunasekera TS (2007) Growth of *E. coli* BL21 in minimal media with different gluconeogenic carbon sources and salt contents Applied Microbiology and Biotechnology 73:1169-1172 doi:10.1007/s00253-006-0554-8

Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C (2017) Salmon provides fast and bias-aware quantification of transcript expression Nature methods 14:417-419 doi:10.1038/nmeth.4197
Pearson TA, Manolio TA (2008) How to Interpret a Genome-wide Association Study JAMA 299:1335-1344 doi:10.1001/jama.299.11.1335
Peng Y, Leung HCM, Yiu S-M, Lv M-J, Zhu X-G, Chin FYL (2013) IDBA-tran: a more robust de novo de Bruijn graph assembler for transcriptomes with uneven expression levels Bioinformatics 29:i326-i334
Peterhansel C, Blume C, Offermann S (2012) Photorespiratory bypasses: how can they work? Journal of Experimental Botany 64:709-715 doi:10.1093/jxb/ers247
Peterhansel C, Horst I, Niessen M, Blume C, Kebeish R, Kürkcüoglu S, Kreuzaler F (2010) Photorespiration *Arabidopsis* Book 8:e0130-e0130 doi:10.1199/tab.0130
Piližota I, Train C-M, Altenhoff A, Redestig H, Dessimoz C (2018) Phylogenetic approaches to identifying fragments of the same gene, with application to the wheat genome Bioinformatics 35:1159-1166 doi:10.1093/bioinformatics/bty772
Pocalyko DJ, Carroll LJ, Martin BM, Babbitt PC, Dunaway-Mariano D (1990) Analysis of sequence homologies in plant and bacterial pyruvate phosphate dikinase, enzyme I of the bacterial phospoenolpyruvate:sugar phosphotransferase system and other PEP-utilizing enzymes. Identification of potential catalytic and regulatory motifs Biochemistry 29:10757-10765 doi:10.1021/bi00500a006
Price GD et al. (1994) Specific reduction of chloroplast carbonic anhydrase activity by antisense RNA in transgenic tobacco plants has a minor effect on photosynthetic CO_2 assimilation Planta 193:331-340
Price MN, Dehal PS, Arkin AP (2010) FastTree 2–approximately maximum-likelihood trees for large alignments PLoS One 5 doi:10.1371/journal.pone.0009490
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 doi:10.1093/nar/gkm219
Rao X, Dixon RA (2016) The differences between NAD-ME and NADP-ME subtypes of C_4 photosynthesis: more than decarboxylating enzymes Frontiers in plant science 7:1525
Reed M, Graham D (1981) Carbonic anhydrase in plants: distribution, properties and possible physiological roles Progress in phytochemistry
Reyna-Llorens I et al. (2018) Ancient duons may underpin spatial patterning of gene expression in C_4 leaves Proceedings of the National Academy of Sciences 115:1931-1936 doi:10.1073/pnas.1720576115
Rioval J, Ricard B, Pradet A (1990) Purification and partial characterization of pyruvate decarboxylase from Oryza sativa L European journal of biochemistry 194:791-797
Ripley BS, Abraham TI, Osborne CP (2008) Consequences of C_4 photosynthesis for the partitioning of growth: a test using C_3 and C_4 subspecies of *Alloteropsis semialata* under nitrogen-limitation Journal of experimental botany 59:1705-1714
Robertson G et al. (2010) *De novo* assembly and analysis of RNA-seq data 7:909-912
Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP (2011) Integrative genomics viewer Nat Biotechnol 29:24-26 doi:10.1038/nbt.1754
Rossini L, Cribb L, Martin DJ, Langdale JA (2001) The maize golden2 gene defines a novel class of transcriptional regulators in plants The Plant Cell 13:1231-1244
Sage RF (2002) Variation in the kcat of Rubisco in C_3 and C_4 plants and some implications for photosynthetic performance at high and low temperature Journal of Experimental Botany 53:609-620 doi:10.1093/jexbot/53.369.609
Sage RF (2004) The evolution of C_4 photosynthesis New Phytologist 161:341-370 doi:10.1111/j.1469-8137.2004.00974.x
Sage RF (2016) A portrait of the C_4 photosynthetic family on the 50th anniversary of its discovery: species number, evolutionary lineages, and Hall of Fame Journal of Experimental Botany 67:4039-4056 doi:10.1093/jxb/erw156
Sage RF, Christin P-A, Edwards EJ (2011) The C_4 plant lineages of planet Earth Journal of Experimental Botany doi:10.1093/jxb/err048

Sage RF, Sage TL, Pearcy RW, Borsch T (2007) The taxonomic distribution of C_4 photosynthesis in *Amaranthaceae sensu stricto* American Journal of Botany 94:1992-2003
Sage TL et al. (2013) Initial events during the evolution of C_4 photosynthesis in C_3 species of Flaveria Plant Physiology 163:1266-1276
Sambrook J, Russell DW (2006) Transformation of *E. coli* by Electroporation Cold Spring Harbor Protocols 2006:pdb.prot3933 doi:10.1101/pdb.prot3933
Sauer U, Eikmanns BJ (2005) The PEP-pyruvate-oxaloacetate node as the switch point for carbon flux distribution in bacteria: We dedicate this paper to Rudolf K. Thauer, Director of the Max-Planck-Institute for Terrestrial Microbiology in Marburg, Germany, on the occasion of his 65th birthday FEMS Microbiology Reviews 29:765-794 doi:10.1016/j.femsre.2004.11.002
Schrider DR, Hourmozdi JN, Hahn MW (2011) Pervasive Multinucleotide Mutational Events in Eukaryotes Current biology : CB 21:1051-1054 doi:10.1016/j.cub.2011.05.013
Schuler ML, Mantegazza O, Weber APM (2016) Engineering C_4 photosynthesis into C_3 chassis in the synthetic biology age The Plant Journal 87:51-65 doi:10.1111/tpj.13155
Schultz JE, Matin A (1991) Molecular and functional characterization of a carbon starvation gene of *Escherichia coli* Journal of Molecular Biology 218:129-140 doi:https://doi.org/10.1016/0022-2836(91)90879-B
Schulz MH, Zerbino DR, Vingron M, Birney E (2012) Oases: robust *de novo* RNA-seq assembly across the dynamic range of expression levels Bioinformatics 28:1086-1092 doi:10.1093/bioinformatics/bts094
Schwacke R et al. (2003) ARAMEMNON, a novel database for *Arabidopsis* integral membrane proteins Plant physiology 131:16-26 doi:10.1104/pp.011577
Sharan SK, Thomason LC, Kuznetsov SG, Court DL (2009) Recombineering: a homologous recombination-based method of genetic engineering Nature Protocols 4:206 doi:10.1038/nprot.2008.227
Sharwood RE (2017) Engineering chloroplasts to improve Rubisco catalysis: prospects for translating improvements into food and fiber crops New Phytologist 213:494-510
Sheahan MC, Chase MW (1996) A phylogenetic analysis of Zygophyllaceae R. Br. based on morphological, anatomical and rbcL DNA sequence data Botanical Journal of the Linnean Society 122:279-300
Simpson JT, Durbin R (2010) Efficient construction of an assembly string graph using the FM-index Bioinformatics 26:i367-i373
Smith-Unna R, Boursnell C, Patro R, Hibberd J, Kelly S (2016) TransRate: reference free quality assessment of de novo transcriptome assemblies Genome Research doi:10.1101/gr.196469.115
Sonawane BV, Sharwood RE, von Caemmerer S, Whitney SM, Ghannoum O (2017) Short-term thermal photosynthetic responses of C_4 grasses are independent of the biochemical subtype Journal of Experimental Botany 68:5583-5597 doi:10.1093/jxb/erx350
South PF, Cavanagh AP, Liu HW, Ort DR (2019) Synthetic glycolate metabolism pathways stimulate crop growth and productivity in the field Science 363:eaat9077 doi:10.1126/science.aat9077
Sowiński P, Bilska A, Barańska K, Fronk J, Kobus P (2007) Plasmodesmata density in vascular bundles in leaves of C_4 grasses grown at different light conditions in respect to photosynthesis and photosynthate export efficiency Environmental and experimental botany 61:74-84
Spreitzer RJ (2003) Role of the small subunit in ribulose-1,5-bisphosphate carboxylase/oxygenase Archives of Biochemistry and Biophysics 414:141-149 doi:https://doi.org/10.1016/S0003-9861(03)00171-1
Sprouffske K, Wagner A (2016) Growthcurver: an R package for obtaining interpretable metrics from microbial growth curves BMC Bioinformatics 17:172 doi:10.1186/s12859-016-1016-7

Stamatakis A (2014) RAxML Version 8: A tool for Phylogenetic Analysis and Post-Analysis of Large Phylogenies Bioinformatics doi:10.1093/bioinformatics/btu033
Steinmann VW, Porter JM (2002) Phylogenetic relationships in Euphorbieae (Euphorbiaceae) based on ITS and ndhF sequence data Annals of the Missouri Botanical Garden:453-490
Stitt M, Heldt HW (1985) Generation and maintenance of concentration gradients between the mesophyll and bundle sheath in maize leaves Biochimica et Biophysica Acta (BBA) - Bioenergetics 808:400-414 doi:https://doi.org/10.1016/0005-2728(85)90148-3
Studer AJ, Gandin A, Kolbe AR, Wang L, Cousins AB, Brutnell TP (2014) A Limited Role for Carbonic Anhydrase in C_4 Photosynthesis as Revealed by a ca1ca2 Double Mutant in Maize Plant Physiology 165:608-617 doi:10.1104/pp.114.237602
Surget-Groba Y, Montoya-Burgos JI (2010) Optimization of de novo transcriptome assembly from next-generation sequencing data Genome Research 20:1432-1440
Suzuki Y, Makino A (2012) Availability of Rubisco Small Subunit Up-Regulates the Transcript Levels of Large Subunit for Stoichiometric Assembly of Its Holoenzyme in Rice Plant Physiology 160:533-540 doi:10.1104/pp.112.201459
Svensson P, Bläsing OE, Westhoff P (2003) Evolution of C_4 phosphoenolpyruvate carboxylase Archives of Biochemistry and Biophysics 414:180-188 doi:https://doi.org/10.1016/S0003-9861(03)00165-6
Takahashi-Terada A, Kotera M, Ohshima K, Furumoto T, Matsumura H, Kai Y, Izui K (2005) Maize Phosphoenolpyruvate Carboxylase: mutations at the putative binding site for the glucose 6-phosphate caused desensitization and abolished responsiveness to regulatory phosphorylation Journal of Biological Chemistry 280:11798-11806 doi:10.1074/jbc.M408768200
Taniguchi Y, Nagasaki J, Kawasaki M, Miyake H, Sugiyama T, Taniguchi M (2004) Differentiation of dicarboxylate transporters in mesophyll and bundle sheath chloroplasts of maize Plant and Cell Physiology 45:187-200
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 C_4 Carbonic Anhydrase in Flaveria Plant Physiology 150:1515-1529 doi:10.1104/pp.109.137513
Timm S, Florian A, Jahnke K, Nunes-Nesi A, Fernie AR, Bauwe H (2011) The Hydroxypyruvate-Reducing System in Arabidopsis: Multiple Enzymes for the Same End Plant Physiology 155:694-705 doi:10.1104/pp.110.166538
Timm S et al. (2008) A Cytosolic Pathway for the Conversion of Hydroxypyruvate to Glycerate during Photorespiration in *Arabidopsis* The Plant Cell 20:2848-2859 doi:10.1105/tpc.108.062265
Tokmakov AA, Kurotani A, Takagi T, Toyama M, Shirouzu M, Fukami Y, Yokoyama S (2012) Multiple post-translational modifications affect heterologous protein synthesis J Biol Chem 287:27106-27116 doi:10.1074/jbc.M112.366351
Toth AL et al. (2007) Wasp Gene Expression Supports an Evolutionary Link Between Maternal Behavior and Eusociality Science 318:441-444
Tovar-Méndez A, Miernyk JA, Randall DD (2003) Regulation of pyruvate dehydrogenase complex activity in plant cells European Journal of Biochemistry 270:1043-1049
Trudeau DL et al. (2018) Design and in vitro realization of carbon-conserving photorespiration Proceedings of the National Academy of Sciences 115:E11455-E11464 doi:10.1073/pnas.1812605115
Van den Bergh B, Swings T, Fauvart M, Michiels J (2018) Experimental Design, Population Dynamics, and Diversity in Microbial Experimental Evolution Microbiology and Molecular Biology Reviews 82:e00008-00018 doi:10.1128/mmbr.00008-18
van den Bergh E, Külahoglu C, Bräutigam A, Hibberd JM, Weber APM, Zhu X-G, Eric Schranz M (2014) Gene and genome duplications and the origin of C_4 photosynthesis: Birth of a trait in the Cleomaceae Current Plant Biology 1:2-9 doi:https://doi.org/10.1016/j.cpb.2014.08.001

Varma A, Boesch BW, Palsson BO (1993) Stoichiometric interpretation of *Escherichia coli* glucose catabolism under various oxygenation rates Appl Environ Microbiol 59:2465-2473
Venkat A, Hahn MW, Thornton JW (2018) Multinucleotide mutations cause false inferences of lineage-specific positive selection Nature Ecology & Evolution 2:1280-1288 doi:10.1038/s41559-018-0584-5
Vera G, Jansen RC, Suppi RL (2008) R/parallel--speeding up bioinformatics analysis with R BMC Bioinformatics 9:390-390 doi:10.1186/1471-2105-9-390
Vidal J, Chollet R (1997) Regulatory phosphorylation of C_4 PEP carboxylase Trends in Plant Science 2:230-237 doi:https://doi.org/10.1016/S1360-1385(97)89548-9
Von Caemmerer S (2000) Biochemical models of leaf photosynthesis. Csiro publishing,
von Caemmerer S, Furbank RT (2003) The C_4 pathway: an efficient CO_2 pump Photosynthesis Research 77:191
von Caemmerer S, Furbank RT (2016) Strategies for improving C_4 photosynthesis Current Opinion in Plant Biology 31:125-134 doi:https://doi.org/10.1016/j.pbi.2016.04.003
von Caemmerer S, Quick WP, Furbank RT (2012) The development of C_4 rice: current progress and future challenges science 336:1671-1672
Voznesenskaya EV, Franceschi VR, Kiirats O, Artyusheva EG, Freitag H, Edwards GE (2002) Proof of C_4 photosynthesis without Kranz anatomy in *Bienertia cycloptera* (Chenopodiaceae) The Plant Journal 31:649-662
Voznesenskaya EV, Franceschi VR, Kiirats O, Freitag H, Edwards GE (2001) Kranz anatomy is not essential for terrestrial C_4 plant photosynthesis Nature 414:543-546 doi:10.1038/35107073
Voznesenskaya EV, Koteyeva NK, Edwards GE, Ocampo G (2010) Revealing diversity in structural and biochemical forms of C_4 photosynthesis and a C_3-C_4 intermediate in genus *Portulaca L.* (*Portulacaceae*) Journal of experimental botany 61:3647-3662 doi:10.1093/jxb/erq178
Walker BJ et al. (2014) Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement PloS one 9:e112963
Waltz E (2018) With a free pass, CRISPR-edited plants reach market in record time Nat Biotechnol 36:6-7 doi:10.1038/nbt0118-6b
Wang D, Portis AR, Moose SP, Long SP (2008) Cool C_4 Photosynthesis: Pyruvate P(i) Dikinase Expression and Activity Corresponds to the Exceptional Cold Tolerance of Carbon Assimilation in *Miscanthus×giganteus* Plant Physiology 148:557-567 doi:10.1104/pp.108.120709
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 C_4 Kranz anatomy The Plant Journal 75:656-670
Wang P et al. (2017) Re-creation of a Key Step in the Evolutionary Switch from C_3 to C_4 Leaf Anatomy Current Biology 27:3278-3287. e3276
Wang P, Vlad D, Langdale JA (2016) Finding the genes to build C_4 rice Current Opinion in Plant Biology 31:44-50 doi:https://doi.org/10.1016/j.pbi.2016.03.012
Wang Y, Bräutigam A, Weber AP, Zhu X-G (2014) Three distinct biochemical subtypes of C_4 photosynthesis? A modelling analysis Journal of experimental botany 65:3567-3578
Waters MT, Langdale JA (2009) The making of a chloroplast The EMBO journal 28:2861-2873
Weber A, Menzlaff E, Arbinger B, Gutensohn M, Eckerskorn C, Fluegge U-I (1995) The 2-oxoglutarate/malate translocator of chloroplast envelope membranes: molecular cloning of a transporter containing a 12-helix motif and expression of the functional protein in yeast cells Biochemistry 34:2621-2627
Weber APM, von Caemmerer S (2010) Plastid transport and metabolism of C_3 and C_4 plants-comparative analysis and possible biotechnological exploitation Current Opinion in Plant Biology 13:256-264 doi:https://doi.org/10.1016/j.pbi.2010.01.007
Weber E, Engler C, Gruetzner R, Werner S, Marillonnet S (2011) A Modular Cloning System for Standardized Assembly of Multigene Constructs PLOS ONE 6:e16765 doi:10.1371/journal.pone.0016765

Weissmann S et al. (2016) Interactions of C_4 Subtype Metabolic Activities and Transport in Maize Are Revealed through the Characterization of DCT2 Mutants The Plant Cell 28:466-484 doi:10.1105/tpc.15.00497
Westhoff P, Gowik UDO (2004) Evolution of C_4 Phosphoenolpyruvate Carboxylase. Genes and Proteins: a Case Study with the Genus *Flaveria* Annals of Botany 93:13-23 doi:10.1093/aob/mch003
Wickham H (2011) The Split-Apply-Combine Strategy for Data Analysis 2011 40:29 doi:10.18637/jss.v040.i01
Wickham H (2018) stringr: Simple, Consistent Wrappers for Common String Operations R package version 1.3.1
Williams BP, Aubry S, Hibberd JM (2012) Molecular evolution of genes recruited into C_4 photosynthesis Trends in Plant Science 17:213-220 doi:https://doi.org/10.1016/j.tplants.2012.01.008
Wong TS, Wong TS, Roccatano* D, Schwaneberg U (2007) Challenges of the genetic code for exploring sequence space in directed protein evolution Biocatalysis and Biotransformation 25:229-241 doi:10.1080/10242420701444280
Wright S (1931) Evolution in mendelian populations Genetics 16:97-159
Wu F-H, Shen S-C, Lee L-Y, Lee S-H, Chan M-T, Lin C-S (2009) Tape-*Arabidopsis* Sandwich - a simpler *Arabidopsis* protoplast isolation method Plant Methods 5:16 doi:10.1186/1746-4811-5-16
Xie Y et al. (2014) SOAPdenovo-Trans: *De novo* transcriptome assembly with short RNA-Seq reads Bioinformatics doi:10.1093/bioinformatics/btu077
Yang N et al. (2017) Contributions of *Zea mays* subspecies mexicana haplotypes to modern maize. Nature communications 8:1874. doi:10.1038/s41467-017-02063-5
Yang Z (2007) PAML 4: Phylogenetic Analysis by Maximum Likelihood Molecular Biology and Evolution 24:1586-1591 doi:10.1093/molbev/msm088
Zelitch I (1968) Investigation on photorespiration with a sensitive 14C-assay Plant physiology 43:1829-1837
Zelitch I, Schultes NP, Peterson RB, Brown P, Brutnell TP (2009) High Glycolate Oxidase Activity Is Required for Survival of Maize in Normal Air Plant Physiology 149:195-204 doi:10.1104/pp.108.128439
Zerbino DR et al. (2017) Ensembl 2018 Nucleic Acids Res 46:D754-D761 doi:10.1093/nar/gkx1098
Zhang B et al. (2018) Inner Envelope CHLOROPLAST MANGANESE TRANSPORTER 1 Supports Manganese Homeostasis and Phototrophic Growth in *Arabidopsis* Molecular Plant 11:943-954 doi:https://doi.org/10.1016/j.molp.2018.04.007
Zhang Q, Lee J, Pandurangan S, Clarke M, Pajak A, Marsolais F (2013) Characterization of *Arabidopsis* serine:glyoxylate aminotransferase, AGT1, as an asparagine aminotransferase Phytochemistry 85:30-35 doi:https://doi.org/10.1016/j.phytochem.2012.09.017
Zhang Z, Mao X, Ou J, Ye N, Zhang J, Peng X (2015) Distinct photorespiratory reactions are preferentially catalyzed by glutamate:glyoxylate and serine:glyoxylate aminotransferases in rice Journal of Photochemistry and Photobiology B: Biology 142:110-117 doi:https://doi.org/10.1016/j.jphotobiol.2014.11.009
Zhao C, Lin Z, Dong H, Zhang Y, Li Y (2017) Reexamination of the Physiological Role of PykA in *Escherichia coli* revealed that It negatively regulates the intracellular ATP levels under anaerobic conditions Applied and Environmental Microbiology 83:e00316-00317 doi:10.1128/aem.00316-17
Zhao S, Fernald RD (2005) Comprehensive algorithm for quantitative real-time polymerase chain reaction J Comput Biol 12:1047-1064 doi:10.1089/cmb.2005.12.1047
Zhu X-G, Long SP, Ort DR (2010) Improving photosynthetic efficiency for greater yield Annual review of plant biology 61:235-261