Engineering enhanced enzyme variants to boost photosynthesis in C_4 plants



Philippa Sinclair

Wadham College

Department of Biology

University of Oxford

Thesis submitted for the degree of Doctor of Philosophy

Hilary term 2023

Abstract

Major food and biofuel crops maize, sugarcane and sorghum utilise a specialised type of photosynthesis termed C_4 photosynthesis, in which carbon is first fixed into a four-carbon compound by phospho*enol*pyruvate carboxylase (PEPC). The enzymatic rate of PEPC limits the rate of photosynthetic carbon assimilation under conditions of low CO_2 in the mesophyll cell, for example shortly after photosynthetic induction or under stress conditions when stomatal conductance is limited. Here, a system has been established in which the growth rate of *Escherichia coli* is dependent on the enzymatic rate of C_4 PEPC derived from the model C_4 species *Setaria viridis*. Concurrently, a site-saturation mutagenesis technique has been optimised and libraries of *PEPC* variants generated, targeting sites that are likely to be influential for C_4 PEPC kinetics. *E. coli* were transformed with *PEPC* variant libraries, and a screen was undertaken whereby *E. coli* cells that expressed a PEPC variant that conferred a growth advantage to the cell became enriched in the bacterial culture over time. Three PEPC variants with P531A, F794V and N863K mutations emerged as candidates for *in planta* testing and provide insight into the molecular basis of C_4 PEPC kinetics. In summary, the high-throughput screening pipeline established here has accelerated the identification of promising PEPC variants that could increase the rate of photosynthesis and yield in C_4 plants.

Acknowledgements

I wish to thank my family, friends and colleagues who have provided so much support over the last few years. Thank you to Professor Steven Kelly for giving me the opportunity to undertake this DPhil project and to the academics who have provided helpful support, suggestions and advice at each assessment milestone.

My biggest thanks must go to everyone in the Kelly and Langdale labs, past and present, who have become very good friends to me and have provided so much support (technical and emotional) over the last four years. Cake day (not fruit day) will be especially missed. Lab badminton was another real highlight of our time together, as well as having lunch together and going for drinks at various points, pandemic permitting. Thanks also to the lab support staff and administrative staff in the Department and the DTP who enable our research efforts to happen.

Playing badminton has been a welcome distraction from work throughout the DPhil; making brilliant friends at Oxford University Badminton Club and the various BUCS, Varsity and local league matches we have taken part in have been a great source of enjoyment over the last four years. I have also enjoyed, and been grateful for, the catch ups with friends from Wadham and the DTP. We have all been through so much together and sharing stories of these (frankly, at times, unbelievable) experiences has been one of the only things that has kept me sane. Catriona, I hope that one day the equipment you need to take all your measurements finally works. Your strength in the face of adversity is an inspiration to us all.

Lastly, thank you to my friends from home, the University of Bristol, Tom and my family for your unwavering support. Thank you for cheering me on and believing in me. I hope that I can be as supportive to you all in return.

Table of contents

Abstract ..... **i**

Acknowledgements ..... **ii**

Table of contents ..... **iii**

Abbreviations ..... **xii**

Chapter 1: Introduction **1**

1.1 Supercharging C_4 photosynthesis 1

1.1.1 The evolution of C_4 photosynthesis 1

1.1.2 Diversity between C_4 photosynthesis pathways 9

1.1.3 Can diversity between C_4 species be exploited to make improvements to the rate of photosynthesis in C_4 crop species? 10

1.2 The role and evolution of phospho*enol*pyruvate carboxylase across the tree of life 14

1.2.1 The functions and regulation of PEPC 14

1.2.2 The structure and catalytic properties of bacterial and plant PEPC 18

1.2.2.1 The residues within bacterial and plant PEPC involved in quaternary structure and cofactor binding 19

1.2.2.2 The residues within bacterial and plant PEPC involved in substrate binding and catalysis 21

1.2.2.3 The residues within PEPC involved in effector binding 24

1.2.3 The molecular evolution of C_4 PEPCs 27

1.3 Directed evolution of plant enzymes in *Escherichia coli* 33

1.4 Regulation of genome edited crops around the world 39

1.4.1 Regulation of genetically modified and genome edited crop plants in the United Kingdom and European Union 39

1.4.2 Acceptance of genome edited crops around the world 41

1.5 Thesis aims and outline 43

Chapter 2: Materials and methods 45

2.1 Bacterial strains, growth conditions and media 45

2.1.1 Bacterial strains 45

2.1.2 Bacterial growth media 45

2.1.3 Transformation of *E. coli* DH5α competent cells 46

2.1.4 Preparation and transformation of *E. coli* Δppc electrocompetent cells 46

2.1.5 Growth of bacterial strains in M9 glucose 48

2.1.6 Glycerol stocks 48

2.1.7 Set up of the Chi.Bio reactors for bacterial culture 49

2.2 Molecular cloning 50

2.2.1 Primer design 50

2.2.2 Golden Gate cloning 50

2.2.2.1 Cloning PEPC expression vectors 50

2.2.2.2 Variations of cloning the library of *6His::PEPC* variants into the placeholder vector 51

2.2.2.3 Cloning the library of *6His::PEPC* variants into the placeholder vector 51

2.2.3 Colony PCR reactions 51

2.2.4 Gel electrophoresis 52

2.2.5 Plasmid preparation 52

2.2.6 Sanger sequencing 52

2.3 Generation of *6His::PEPC* variants by Darwin Assembly 52

2.3.1 ssDNA generation 53

2.3.2 Primer phosphorylation 53

2.3.3 Isothermal assembly 54

2.3.4 Assembly purification 54

2.3.5 Amplification of *6His::PEPC* variant libraries (outnest PCR) 55

2.4 Analysis of 6His::PEPC expression level in bacteria by immunoblot 56

2.5 Screens of 6His::PEPC variants in *E. coli* 57

2.5.1 Modelling the composition of *E. coli* Δppc strains in bacterial co-culture 57

2.5.1.1 Modelling the co-culture of three bacterial strains with different hypothetical rates of growth 57

2.5.1.2 Modelling the co-culture of three bacterial strains using experimental growth rate data 58

2.5.2 A screen of *E. coli* Δppc strains with different levels of 6His::PEPC expression 58

2.5.3 A screen of 6His::PEPC variants in *E. coli* continuous co-culture 59

2.5.4 Analysis of 6His::PEPC abundance in continuous co-culture: semi-quantitative PCR and next-generation sequencing (NGS) 60

2.5.5 Enzyme kinetics characterisation of P531A, F794V and N863K SvPEPC variants..... 62

2.5.6 Modelling the impact of P531A, F794V and N863K SvPEPC variants on rate of photosynthetic carbon assimilation in *Setaria viridis* 63

Chapter 3: A novel bacterial screen for improved PEPC variants 64

3.1 Introduction 64

3.1.1 Engineering PEPC to improve food security 64

3.1.2 Using *E. coli* to screen 6His::PEPC variants in high-throughput 66

3.1.3 Chapter aims and outline 70

3.2 Results and Discussion..... 71

3.2.1 Installation and testing of Chi.Bio bioreactors in the laboratory..... 71

3.2.2 Construction of *Setaria viridis* C_4 PEPC expression vectors 79

3.2.2.1 Domestication of *Setaria viridis* C_4 *PEPC*..... 79

3.2.2.2 Design of multiple PEPC expression vectors..... 81

3.2.3 6His::PEPC expression level can be finely tuned in *E. coli* Δppc..... 83

3.2.4 Growth rate of *E. coli* Δppc in M9 glucose is dependent on the enzymatic rate of 6His::PEPC 87

3.2.5 Simulating the population dynamics of an *E. coli* Δppc culture that contains bacterial strains with different rates of growth..... 94

3.2.6 Bacteria that have the highest 6His::PEPC expression outcompete bacteria that have lower 6His::PEPC expression in continuous co-culture ... 97

3.3 Chapter summary ... 105

3.4 Authorship statement ... 106

Chapter 4: Generation of *6His::PEPC* variants *via* site saturation mutagenesis ... 107

4.1 Introduction ... 107

4.1.1 Choice of mutagenesis approach to create *6His::PEPC* variant libraries ... 107

4.1.2 Choice of mutagenesis technique to create *6His::PEPC* variant libraries ... 112

4.1.3 Chapter aims and outline ... 120

4.2 Results and Discussion ... 121

4.2.1 There is substantial variation between C_4 species at sites that have evolved under positive selection in PEPC. ... 121

4.2.2 Generation of *6His::PEPC* variant libraries by Darwin Assembly mutagenesis ... 137

4.2.2.1 Generation of the ssDNA template for mutagenesis ... 137

4.2.2.2 Oligonucleotide design ... 140

4.2.2.2.1 Inner oligonucleotide design ... 140

4.2.2.2.2 Boundary and outnest oligonucleotide design ... 145

4.2.2.3 Generation of a single site saturation library ... 147

4.2.2.4 Generation of 21 libraries of single-site enzyme variants 151

4.2.3 Electroporation of *6His::PEPC* into *E. coli* Δppc..... 156

4.2.4 Optimisation of cloning *6His::PEPC* variants into an expression vector 159

4.2.4.1 Development of a single-insert cloning system..... 159

4.2.4.2 Troubleshooting the cloning step using the single-insert cloning system 164

4.2.4.2.1 Testing different Golden Gate reaction conditions 164

4.2.4.2.2 Can increasing the number of digestion and ligation cycles improve Golden Gate efficiency? 171

4.2.4.2.3 Transformation efficiency vs transformation frequency 173

4.2.5 Transformation of *E. coli* Δppc with a library of 6His::PEPC variants . 178

4.3 Chapter summary..... 182

Chapter 5: Identification of enhanced variants of *Seteria viridis* C_4 PEPC . 183

5.1 Chapter introduction..... 183

5.1.1 Development of a novel 6His::PEPC variant screen in *E. coli*..... 183

5.1.2 Establishing a pipeline to identify the best-performing 6His::PEPC variants using continuous co-culture in *E. coli* Δppc 184

5.1.3 Chapter aims and outline 188

5.2 Results and Discussion..... 189

5.2.1 Development of a NGS and bioinformatics pipeline to identify best-performing 6His::PEPC variants..... 189

5.2.1.1 DNA amplification and sample multiplexing prior to sequencing.. 189

5.2.1.2 Bioinformatic analysis of the NGS reads 193

5.2.2 A proof of concept screen of 20 6His::PEPC variants identifies a 6His::PEPC variant that confers a competitive advantage to *E. coli* in continuous co-culture 196

5.2.2.1 A screen of a library of *6His::PEPC* variants mutated at site 780 for improved enzymatic activity in *E. coli* 196

5.2.2.2 A S780G mutation in 6His::PEPC confers a competitive advantage to bacteria dependent on 6His::PEPC for growth 199

5.2.3 Identification of best-performing 6His::PEPC variants mutated at sites under selection in C_4 PEPC 207

5.2.3.1 A screen of 6His::PEPC variants for improved enzymatic activity in *E. coli*..... 207

5.2.3.2 P531A, F794V and N863K mutations confer a competitive advantage to *E. coli* dependent on 6His::PEPC for growth in M9 glucose 211

5.2.3.2.1 A P531A 6His::PEPC variant is most successful in the screen and has increased catalytic capacity..... 219

5.2.3.2.2 F794V and N863K 6His::PEPC variants have substitutions that have occurred repeatedly in C_4 PEPC 236

5.3 Chapter summary..... 245

5.4 Authorship statement 247

Chapter 6: General discussion..... 248

6.1 Overview..... 248

6.2 Exploiting natural variation between C_4 plants to identify PEPC variants with increased catalytic capacity 249

6.3 F794V and N863K variants are candidates for characterisation *in planta*. 250

6.4 Sites under selection as targets for 6His::PEPC mutagenesis..... 252

6.5 The application of directed evolution in bacteria to plant PEPC..... 254

6.6 Differences in PEPC regulation between bacteria and plants..... 256

6.7 The application of directed evolution to crop plant PEPCs 258

6.8 The potential of directed evolution to engineer improved C_4 PEPC variants 261

Appendix 263

A.1 Golden Gate cloning parts..... 263

A.1.1 *Setaria viridis* C_4 PEPC (Sevir.4G143500) codon optimised for expression in *Escherichia coli* (SC module), written 5’ to 3’ 263

A.1.2 Golden Gate parts, written 5’ to 3’..... 265

A.1.3 Placeholder fragment, written 5’ to 3’..... 265

A.1.4 Placeholder vector, written 5’ to 3’ 265

A.1.5 PEPC PCR product fragment, written 5’ to 3’ 267

A.2 Additional Darwin Assembly results 269

A.2.1 Gels of large-scale PCR for libraries..... 269

A.2.1 Variants tested/not tested in the large screen of 347 6His::PEPC variants together (across both biological replicates)..... 271

A.3 Primers..... 272

A.3.1 Darwin Assembly boundary and outnest primers 272

A.3.2 Darwin Assembly mutagenic inner primers 272

A.3.3 Colony PCR primers 274

A.3.4 NGS barcoded primers 274

A.4 Single amplicon of barcoded primer products 276

Bibliography 278

Abbreviations

6His::PEPC	*Setaria viridis* C_4 phospho*enol*pyruvate carboxylase with an N-terminal 6xHis polyhistidine tag
Acetyl-CoA	Acetyl coenzyme A
ALT	Alanine aminotransferase
A_{net}	Net carbon assimilation rate
ANOVA	Analysis of variance
AST	Aspartate aminotransferase
BSA	Bovine serum albumin
CA	Carbonic anhydrase
CAM	Crassulacean acid metabolism
CCM	Carbon concentrating mechanism
CCR	Catalytic cycles until replacement
CDS	Coding sequence
CFU	Colony-forming units
C_m	Intracellular CO_2 concentration
CRISPR	Clustered, regularly interspaced, short palindromic repeats
DEFRA	Department for Environment, Food and Rural Affairs (United Kingdom)
DNA	Deoxyribonucleic acid
dNTPs	Deoxyribonucleotide triphosphates
dsDNA	Double-stranded DNA
EPSPS	5-enolpyruvylshikimate-3-phosphate synthase
EU	European Union

G6P	Glucose 6-phosphate
GAPDH	Glyceraldehyde 3-phosphate dehydrogenase
GC content	Guanine-cytosine content
GE	Genome editing
GFP	Green fluorescent protein
GM	Genetic modification
GMO	Genetically modified organism
GTP	Guanosine 5'-triphosphate
GUS	β-glucuronidase
His-tag	Polyhistidine tag (6xHis)
HSD	Honest significant difference
K_{cat}	Rate of catalytic conversion of substrate into product
K_{cat}/K_m	Specificity constant for a substrate
K_m	Substrate concentration at which the reaction rate is ½ maximum velocity
LB medium	Luria-Bertani medium
LED	Light-emitting diode
NAD-ME	NAD-dependent malic enzyme
NADP-MDH	NADP-dependent malate dehydrogenase
NADP-ME	NADP-dependent malic enzyme
NCBI	National Center for Biotechnology Information (database)
NEB	New England Biolabs
NGS	Next-generation sequencing
OAA	Oxaloacetate
OD	Optical density

P	Promoter (DNA sequence)
PCK	Phospho*enol*pyruvate carboxykinase
PCR	Polymerase chain reaction
PDB	Protein Data Bank
PEP	2-phospho*enol*pyruvate
PEPC	Phospho*enol*pyruvate carboxylase
PPDK	Pyruvate, phosphate dikinase
RBS	Ribosome binding site (DNA sequence)
RNAi	RNA interference
Rubisco	Ribulose 1,5-bisphosphate carboxylase/oxygenase
RuBP	Ribulose 1,5-bisphosphate
SBPase	Sedoheptulose 1,7-bisphosphatase
SDN	Site-directed nuclease
SDS PAGE	Sodium dodecyl sulphate polyacrylamide gel electrophoresis
ssDNA	Single-stranded DNA
SvPEPC	*Setaria viridis* C_4 PEPC
T	Terminator (DNA sequence)
TALENs	Transcription activator-like effector nucleases
TCA cycle	Tricarboxylic acid cycle
V_{max}	Maximum velocity
ZFN	Zinc-finger nucleases

Chapter 1: Introduction

1.1 Supercharging C_4 photosynthesis

1.1.1 The evolution of C_4 photosynthesis

Photosynthesis, the process by which organisms use light energy to convert water and carbon dioxide (CO_2) into oxygen (O_2) and sugars, provides the consumable energy that supports almost all life on Earth. In the majority of photosynthetic organisms, comprising plants, algae and photosynthetic bacteria, CO_2 is first fixed into a three-carbon compound in a pathway called C_3 photosynthesis (Bassham *et al.*, 1956). The primary carbon fixation step in C_3 photosynthesis entails the carboxylation of ribulose 1,5-bisphosphate (RuBP) by ribulose 1,5-bisphosphate carboxylase/oxygenase (Rubisco). However, Rubisco can undertake both carboxylation and oxygenation reactions (Figure 1.1) (Bowes, Ogren and Hageman, 1971). The efficiency of C_3 photosynthesis is reduced when Rubisco binds O_2 instead of CO_2 because this results in fewer carboxylation reactions, requires an energetically expensive photorespiratory pathway to metabolise the toxic by-product 2-phosphoglycolate, and during photorespiration not all fixed CO_2 is re-captured (Tolbert and Yamazaki, 1969; Bowes, Ogren and Hageman, 1971; Sharkey, 1986). Some photosynthetic organisms have evolved a carbon concentrating mechanism (CCM) to increase the concentration of CO_2 around Rubisco and reduce the frequency of this undesirable oxygenation reaction, thereby increasing photosynthetic efficiency (Ehleringer *et al.*, 1991).

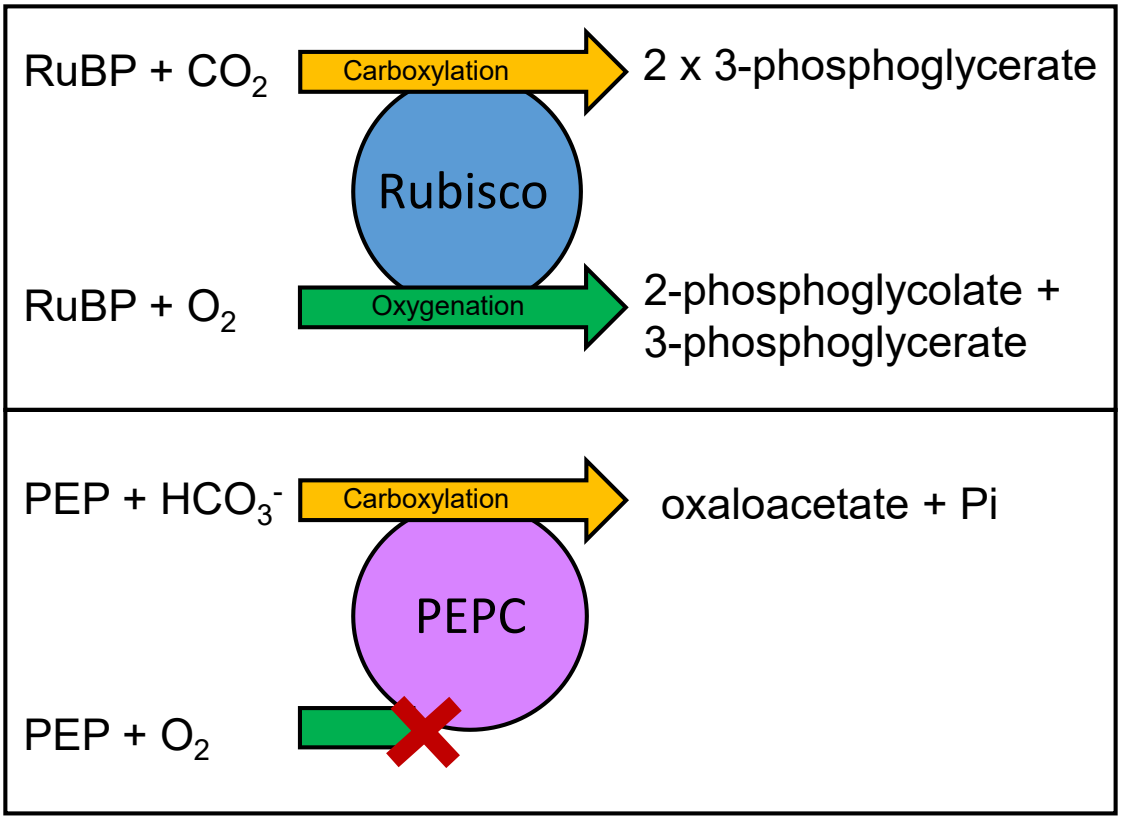


Figure 1.1: Rubisco can catalyse a carboxylation or oxygenation reaction whereas phospho*enol*pyruvate carboxylase (PEPC) only catalyses a carboxylation reaction. RuBP is ribulose 1,5-bisphosphate, Pi is inorganic phosphate.

Three specialised photosynthetic types that concentrate CO_2 around Rubisco have been identified in land plants (Embryophyta): pyrenoids in hornworts, Crassulacean acid metabolism (CAM), and C_4 photosynthesis (Ranson and Thomas, 1960; Hatch and Slack, 1966; Hatch, 1987; Villarreal and Renner, 2012). In the latter two forms of photosynthesis, the primary carbon fixation step is undertaken by phospho*enol*pyruvate carboxylase (PEPC) which, unlike Rubisco, does not bind O_2 and thus is insensitive to the O_2 concentration in the cell (Figure 1.1). This carboxylation reaction produces the four-carbon compound oxaloacetate, which is subsequently converted to either malate or aspartate which are in turn decarboxylated around Rubisco. Rubisco then catalyses a secondary carboxylation of this released CO_2, the product of which enters the Calvin Benson Bassham cycle, as per normal C_3 photosynthesis. Primary and secondary carboxylation reactions are temporally or spatially separated during CAM or C_4 photosynthesis, respectively, creating a high concentration of CO_2 around Rubisco (Ranson and Thomas, 1960; Hatch, 1987). These CCMs are highly effective: the high abundance of C_4 species globally and efficiency of the C_4 photosynthetic pathway together result in 25% terrestrial carbon fixation by species that undertake C_4 photosynthesis, despite the photosynthetic type only occurring in 3% of angiosperm species (Still *et al.*, 2003; Edwards *et al.*, 2010). Moreover, C_4 photosynthesis confers increased radiation, nitrogen and water use efficiencies (Ehleringer and Monson, 1993; Zhu, Long and Ort, 2008; Ghannoum, Evans and von Caemmerer, 2010). Thus, C_4 photosynthesis confers a suite of advantageous agronomical traits in the species in which it is found, including key crop plants maize (*Zea mays*) and sugarcane (*Saccharum officinarum*).

C_4 photosynthesis has evolved over 65 times independently during the last 35 million years and is a canonical example of convergent evolution (Sage, 2016). In the majority of these independent origins of C_4 photosynthesis species compartmentalise the photosynthetic pathway across two different cell types (Sedelnikova, Hughes and Langdale, 2018). Most commonly, these two cell types are mesophyll and bundle sheath cells and the arrangement of these cell types is termed Kranz anatomy, due to the wreath-like appearance of the bundle sheath cells around the veins in a C_4 leaf (Kranz is the German for wreath) (Figure 1.2) (Haberlandt, 1896; Rhoades and Carvalho, 1944; Brown, 1975). Typically, primary carbon assimilation takes place in the mesophyll and photosynthetic carbon reduction takes places in the bundle sheath (Figure 1.2) (Berry, Downton and Tregunna, 1970; Muhaidat, Sage and Dengler, 2007; Rao and Dixon, 2016). However, Kranz anatomy can differ substantially between C_4 species (Edwards and Voznesenskaya, 2011; Sedelnikova, Hughes and Langdale, 2018). Specifically, there is variation in the number of layers of bundle sheath cells around the vascular tissue, the form and positioning of the chloroplasts within different cell types, and the presence or absence of a suberin layer between the mesophyll and bundle sheath cells (Brown, 1975, Hattersley and Browning, 1981; Edwards and Voznesenskaya, 2011). It is also possible to conduct C_4 photosynthesis within a single cell by partitioning the primary carbon assimilation and photosynthetic carbon reduction reactions in different regions of the cell, in the case of some terrestrial plants, or between the cytosol and chloroplast without spatial separation of different cytoplasmic domains, in the case of some aquatic plants (Bowes and Salvucci, 1989; Freitag and Stichler, 2000; Voznesenskaya *et al.*, 2001; von Caemmerer *et al*., 2014).

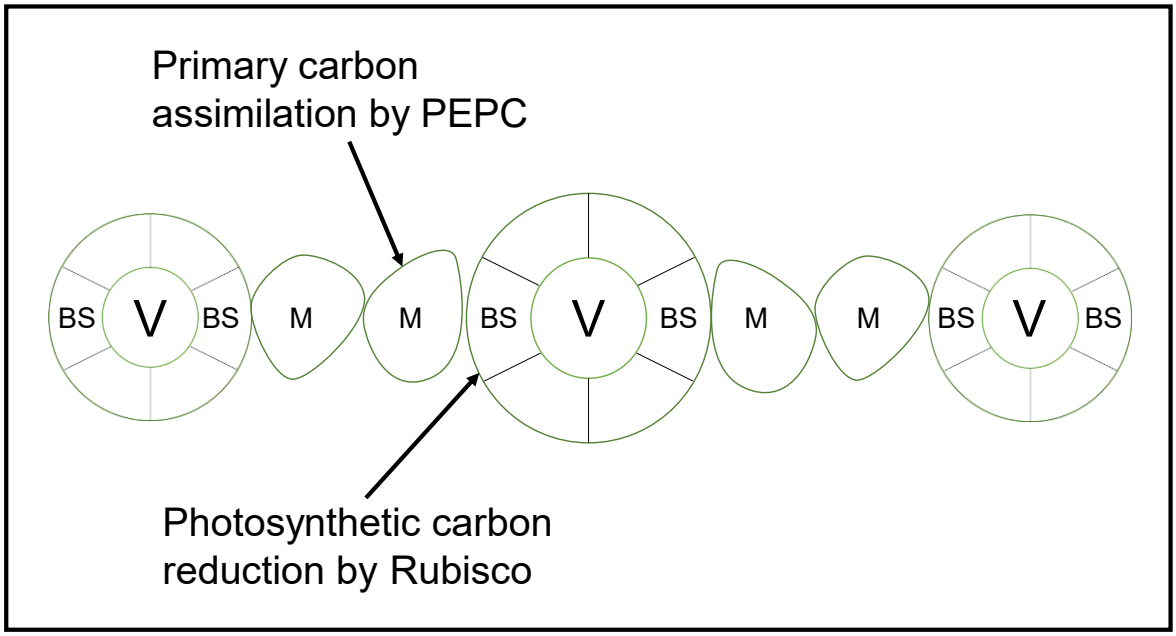


Figure 1.2: Kranz anatomy partitions primary carbon assimilation and reduction in different specialised cell types. This schematic diagram is representative of typical mature C_4 maize leaf organisation. Bundle sheath cells (BS) are arranged concentrically around veins (V). Adjacent veins are separated by mesophyll cells (M) which contain photosynthetic C_4 phospho*enol*pyruvate carboxylase (PEPC).

In addition to the differences in cell types and arrangement of cells around the veins, there are also biochemical differences in the way in which the pathway has been achieved. Although the full extent of the biochemical differences are unknown, C_4 photosynthesis subtypes are generally grouped according to the primary decarboxylase that releases CO_2 around Rubisco; NAD-dependent malic enzyme (NAD-ME), NADP-dependent malic enzyme (NADP-ME) and phospho*enol*pyruvate carboxykinase (PCK) (Figure 1.3) (Kanai and Edwards, 1999; Rao and Dixon, 2016). Thus, although C_4 photosynthesis collectively refers to plants that spatially-partition photosynthesis within a leaf, there are a large number of anatomical and biochemical differences between C_4 species and consequently a large number of variations on the C_4 photosynthesis pathway.

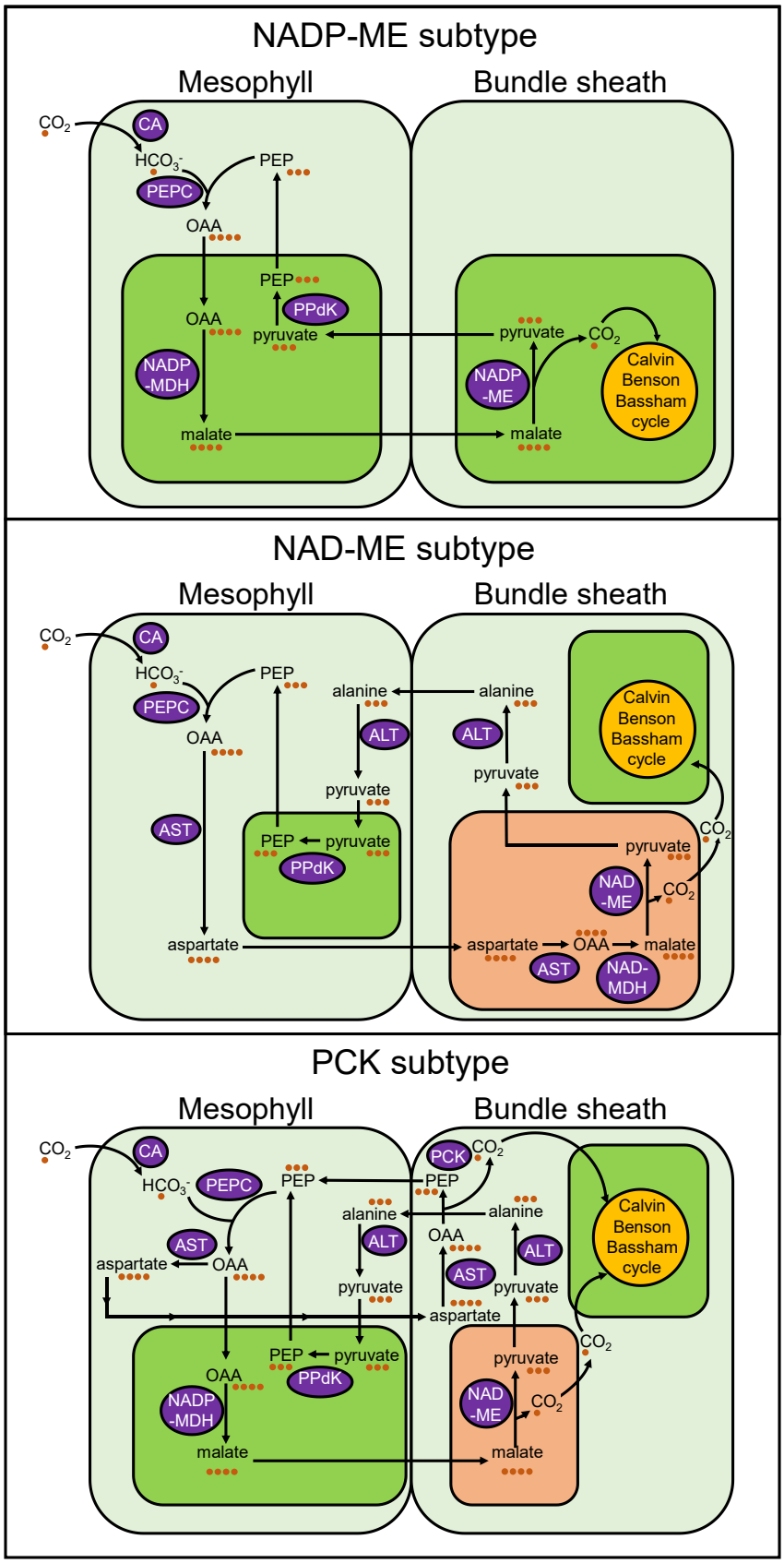


Figure 1.3: C_4 photosynthesis pathways are divided into three main sub-types according to the primary decarboxylase. Figure continued on next page.

Figure 1.3 (continued) Types of C_4 photosynthetic pathway are shown grouped into the three main sub-types. The number of carbon atoms in each compound is represented as a red circle. The dark green and orange compartments are chloroplasts and mitochondria, respectively, and purple oval shapes represent enzymes. Abbreviations: carbonic anhydrase (CA), phospho*enol*pyruvate carboxylase (PEPC), 2-phospho*enol*pyruvate (PEP), oxaloacetate (OAA), NADP-dependent malate dehydrogenase (NADP-MDH), pyruvate, phosphate dikinase (PPDK), NADP-dependent malic enzyme (NADP-ME), aspartate aminotransferase (AST), alanine aminotransferase (ALT), NAD-dependent malic enzyme (NAD-ME), NAD-dependent malate dehydrogenase (NAD-MDH), phospho*enol*pyruvate carboxykinase (PCK). Adapted from Ludwig, 2016.

1.1.2 Diversity between C_4 photosynthesis pathways

In addition to the substantial anatomical and biochemical variation between C_4 species, there is also substantial variation in the age and productivity of C_4 plants (Ehleringer and Pearcy, 1983; Sage, 2016; Niklaus and Kelly, 2019). For example, plants within the *Chloridoideae*, one of the largest subfamilies of grasses, are estimated to be amongst the first to have evolved the C_4 photosynthetic type nearly 30 million years ago, whereas some dicotyledonous C_4 species in the genus *Cleome* and *Flaveria* likely evolved less than 500,000 years ago (Feodorova *et al.*, 2010; Christin *et al.*, 2011; Sage, 2016). As some of the world's most productive crop species belong to the oldest C_4 lineages, *Chloridoideae* and *Andropogoneae* (Sage, 2016; Niklaus and Kelly, 2019), it is tempting to speculate that the oldest C_4 lineages have exploited this additional time to better-adapt their biochemistry and anatomy to improve plant productivity. Furthermore, younger C_4 lineages have had little time for evolution by natural selection to optimise the sequence and abundance of all enzymes required to conduct the C_4 cycle (Niklaus and Kelly, 2019). Thus, we might expect to observe significant diversity in the efficacy and productivity of C_4 pathways between extant C_4 species because of the phylogenetic distance between lineages, and because these pathways have had different amounts of time to optimise C_4 photosynthesis from an ancestral C_3 state (Niklaus and Kelly, 2019). Indeed, there is ample evidence to demonstrate that genes that have become co-opted into C_4 photosynthesis undergo numerous changes to their expression pattern and amino acid sequence to facilitate appropriate expression level, subcellular localisation and adaptation to their new physiological function (Tausta *et al.*, 2002; Engelmann *et al.*, 2003; Gowik *et al.*, 2004; Williams, Aubry and Hibberd,

2012; Ludwig, 2016; Alvarez *et al.*, 2019; DiMario and Cousins, 2019; Niklaus and Kelly, 2019).

1.1.3 Can diversity between C_4 species be exploited to make improvements to the rate of photosynthesis in C_4 crop species?

The biggest challenge of the 21st century is to meet food and fuel requirements for a growing population while limiting agricultural expansion and without placing further pressure on biodiversity. Crop demand in 2050 is predicted to be double the level of demand in 2005 (Tilman *et al.*, 2011). To date, most improvements in crop yields have been achieved through improved disease protection, application of synthetic fertilizers, and breeding desired traits, for example dwarf varieties that produce more grain and are more robust to challenging weather conditions (Zhu, Long and Ort, 2010). However, future crop yield increases are not currently anticipated to keep up with an increasing demand (Ray *et al.*, 2013; Long, Marshall-Colon and Zhu, 2015; Tian *et al.*, 2021). Therefore, improvement to the efficiency and capacity of photosynthesis will be required to meet crop requirements in the future, including improvements to maize and sugarcane, key food and biofuel crops that undertake C_4 photosynthesis (von Caemmerer and Furbank, 2016).

The main targets for enhancing C_4 photosynthesis are: light capture and electron transport; transport and diffusion of metabolites; and enzyme activity for both carbon assimilation (primary and secondary) and regeneration of substrates (von Caemmerer and Furbank, 2016). Limitations to electron transport have been reduced through constitutive overexpression of the Rieske FeS protein, which resulted in an increased rate of photosynthesis in the C_4 grass *Setaria viridis* (Ermakova *et al.*, 2019). It has been suggested that flux of metabolites between cell types is already optimal, although it remains unclear the extent to which movement

of metabolites between cell types and organelles is limiting to C_4 photosynthesis (Hatch and Osmond, 1976; Wang, Long and Zhu, 2014; von Caemmerer and Furbank, 2016). By contrast, carbon assimilation is known to limit C_4 photosynthesis at both high and low intracellular CO_2 concentrations. At high intracellular CO_2 concentrations, carbon assimilation is limited by Rubisco activity and substrate regeneration by pyruvate, phosphate dikinase (PPDK) and sedoheptulose 1,7-bisphosphatase (SBPase) (von Caemmerer *et al.*, 1997; Ölçer, Lloyd and Raines, 2001; Chastain *et al.*, 2011; von Caemmerer and Furbank, 2016; Driever *et al.*, 2017). Conversely, at low CO_2 concentrations, for example under drought stress or in conditions of fluctuating light, carbon assimilation is limited by the catalytic efficiency of carbonic anhydrase and PEPC to deliver CO_2 to Rubisco (von Caemmerer *et al.*, 2004; Studer *et al.*, 2014; von Caemmerer and Furbank, 2016; Pignon and Long, 2020; Wang, Chan and Long, 2021). Therefore, the activity of enzymes involved in carbon assimilation and substrate regeneration are an attractive target for the enhancement of C_4 photosynthesis.

Given the limited time that evolution has had to optimise activity of the enzymes that control flux through the C_4 photosynthetic pathway in different species, it is most parsimonious to assume that in any given species, not every enzyme will be equally optimised for function in a C_4 context (Niklaus and Kelly, 2019). Molecular evolution estimates suggest that if an amino acid substitution can be attained by a single nucleotide change, and this novel variant of an enzyme has a fitness effect of ≥1% in a large population, then this substitution would be expected to have swept to fixation within 5,000 generations and therefore be present in most extant C_4 lineages (Niklaus and Kelly, 2019). For example, the alanine to serine change at site 780 in the peptide sequence of PEPC (*Z. mays* numbering) can be achieved by

a single base change in four different ways. This A780S substitution, which results in a higher affinity of PEPC for the substrate bicarbonate, has been observed in eight independent origins of C_4 photosynthesis in grasses which indicates that there is a strong selective advantage for this amino acid change (Christin *et al.*, 2007; DiMario and Cousins, 2019).

However, 61% of amino acid substitutions require two or more base changes, which are rarer than a single base change, and any intermediate amino acid that does not confer a fitness advantage would be subject to genetic drift or negative/purifying selection (Niklaus and Kelly, 2019). Therefore, amino acid substitutions that require two or more base changes (or transition states) are only likely to have occurred in the oldest lineages, if at all (Niklaus and Kelly, 2019; Phansopa *et al.*, 2020). This is even more pertinent in relation to mutations that are not at a single amino acid locus but instead are beneficial only in combination with other amino acid changes at different sites in the protein that may not yet have evolved. Thus, it is likely that amino acid substitutions that could confer fitness advantages to C_4 cycle enzymes have not evolved in all C_4 lineages, including key C_4 crops, and it could be possible to increase flux through the photosynthetic pathway by introducing improved variants of each enzyme through simple amino acid sequence changes.

A commonality between C_4 photosynthesis pathways is primary carbon fixation by PEPC (O'Leary, 1982; Hatch, 1987; Lepiniec *et al.*, 1994; Hibberd and Covshoff, 2010; Langdale, 2011). Changes to C_4 PEPC expression pattern and kinetics have occurred as the enzyme has been co-opted into the C_4 cycle; expression of C_4 PEPC is high in the mesophyll cells where the primary carbon fixation step occurs and the enzyme has evolved a higher affinity for bicarbonate and lower affinity for PEP than C_3 homologs (Ting and Osmond, 1973; Bläsing *et al.*, 2002; Engelmann,

Bläsing and Westhoff, 2002; Engelmann *et al.*, 2003; Gowik *et al.*, 2004, 2006; Westhoff and Gowik, 2004; Lara *et al.*, 2006; DiMario and Cousins, 2019). Changes to the affinity of PEPC for PEP and bicarbonate are underpinned by changes to the amino acid sequence of the enzyme (Bläsing, Westhoff and Svensson, 2000; Engelmann, Bläsing and Westhoff, 2002; Christin *et al.*, 2007; Rosnow, Edwards and Roalson, 2014; DiMario and Cousins, 2019). The sequence variation between C_4 PEPCs has resulted in variation in C_4 PEPC kinetics across C_4 species, which in turn has a large impact on the modelled rate of C_4 photosynthesis (DiMario *et al.*, 2021). This suggests that there is an opportunity to engineer improved versions of C_4 PEPC, likely with a higher catalytic turnover and/or higher affinity for bicarbonate (DiMario *et al.*, 2021). Thus, PEPC is a key target for engineering improvement in productivity of C_4 crop plants.

1.2 The role and evolution of phospho*enol*pyruvate carboxylase across the tree of life

1.2.1 The functions and regulation of PEPC

Phospho*enol*pyruvate carboxylase (PEPC, EC 4.1.1.31) catalyses the irreversible carboxylation of 2-phospho*enol*pyruvate (PEP) using bicarbonate (HCO_3^-) as the inorganic carbon source to produce oxaloacetate (OAA) and inorganic phosphate. PEPC is widely distributed across the tree of life, with genes encoding the enzyme found in the genomes of many eukaryotes, bacteria and Archaea (Chollet, Vidal and O'Leary, 1996; Ettema *et al.*, 2004; Izui *et al.*, 2004). Within the organisms in which it is found, PEPC primarily provides an anaplerotic function to maintain a pool of tricarboxylic acid cycle intermediates (Izui *et al.*, 2004; Sauer and Eikmanns, 2005). However, PEPC has acquired an additional function in plants that conduct C_4 photosynthesis and CAM, wherein a photosynthetic isoform is solely responsible for primary carbon assimilation (Hatch, 1987; Chollet, Vidal and O'Leary, 1996; Kanai and Edwards, 1999). As discussed above, this new function enables these species to separate the process of primary carbon assimilation from the process of photosynthetic carbon reduction causing a concomitant enhancement of photosynthetic efficiency (Hatch, 1987; Chollet, Vidal and O'Leary, 1996; Kanai and Edwards, 1999).

Plant PEPCs fulfil a wide range of metabolic functions. Non-photosynthetic PEPC isoforms in plants (C_3 PEPC), like bacterial and Archaeal PEPCs, primarily facilitate anaplerotic replenishment of tricarboxylic acid cycle (TCA) cycle intermediates (Plaxton and Podestá, 2006). In addition, C_3 PEPCs fulfil specialised anaplerotic functions in seed development (Sangwan, Singh and Plaxton, 1992), fruit ripening (Blanke and Lenz, 1989), stomatal opening (Outlaw, Jr., 1990; Cousins *et al.*, 2007)

and carbon provision for bacteroids in legume root nodules (Deroche and Carrayol, 1988; Vance *et al.*, 1994). C_4 and CAM plants are uniquely dependent on a PEPC isoform involved in primary carbon fixation (C_4 PEPC), which is essential for these derived photosynthetic types. In order to fulfil this range of functions, PEPC is encoded by a small multigene family in plants (Lepiniec *et al.*, 1994; Chollet, Vidal and O'Leary, 1996; Kanai and Edwards, 1999; Jacobs *et al.*, 2008; O'Leary, Park and Plaxton, 2011).

The enzyme kinetics of bacterial and plant PEPCs are subject to allosteric regulation *in vivo,* in contrast to Archaeal PEPCs which do not appear to be subject to equivalent regulation by effectors. Of bacterial PEPCs, most is known about the *Escherichia coli* PEPC, which is activated by acetyl-CoA, long-chain fatty acids, guanosine 5'-triphosphate (GTP), and hexose- and triose-phosphates, and inhibited by aspartate, malate, citrate and possibly also succinate (Figure 1.4 A) (Nishikido *et al.*, 1965; Silverstein and Willis, 1973; Smith, 1977; Morikawa *et al.*, 1980; McAlister *et al.*, 1981; Izui *et al.*, 1981; Sauer and Eikmanns, 2005). Conversely, Archaeal PEPCs are not activated by typical allosteric regulators such as G6P and acetyl-CoA, and sensitivity to allosteric inhibitors malate and aspartate is not consistent across Archaeal PEPCs (Sako *et al.*, 1996, 1997; Ettema *et al.*, 2004; Patel, Kraszewski and Mukhopadhyay, 2004; Matsumura, Izui and Mizuguchi, 2006). Therefore, different forms of PEPC regulatory control have evolved across different domains of life.

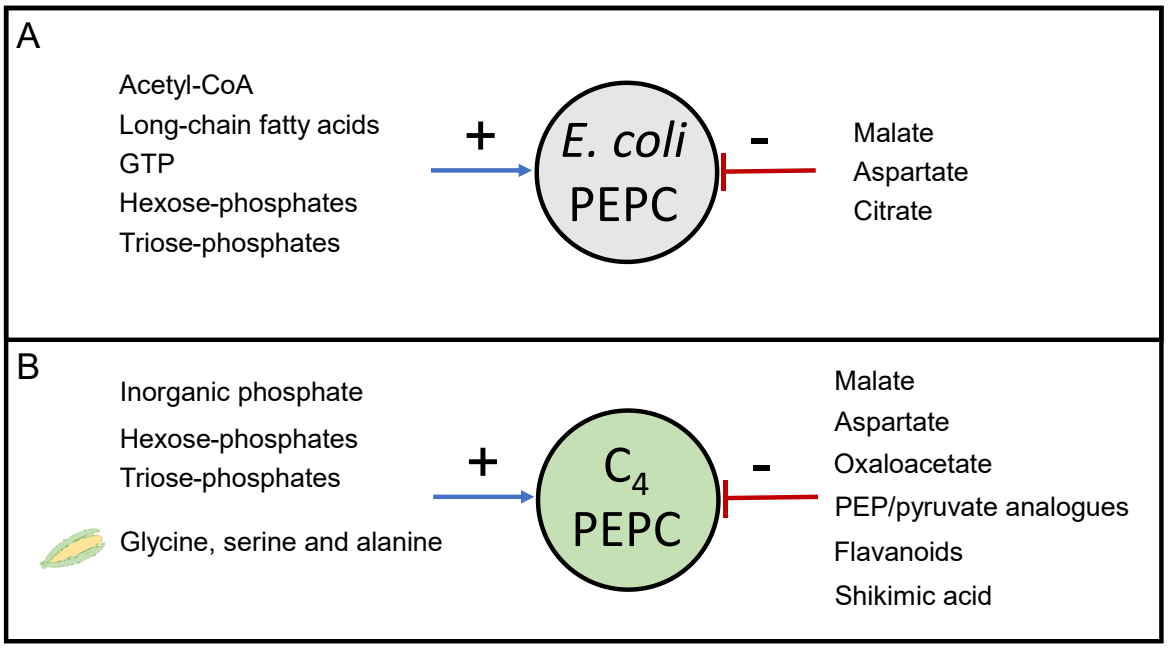


Figure 1.4: *Escherichia coli* and C_4 PEPC are regulated by multiple metabolic effectors. (A) The metabolic effectors that regulate *E. coli* PEPC. (B) The metabolic effectors that regulate C_4 photosynthetic PEPC. Maize icon indicates plants within the Poaceae family. GTP is guanosine 5'-triphosphate.

Plant PEPCs (both C_3 and C_4) are inhibited by aspartate and malate and activated by triose- and hexose-phosphates including glucose 6-phosphate (G6P) (Lepiniec *et al.*, 1994; Chollet, Vidal and O'Leary, 1996; Izui *et al.*, 2004). In addition, C_4 PEPC has been shown to be activated by inorganic phosphate, hexose- and triose-phosphates and inhibited by organic acids, PEP/pyruvate analogues, flavanoids and shikimic acid (Figure 1.4 B) (Lepiniec *et al.*, 1994; Rajagopalan *et al*., 1994; Chollet, Vidal and O'Leary, 1996; Colombo, Pairoba and Andreo, 1996; Pairoba, Colombo and Andreo, 1996; Izui *et al.*, 2004). The amino acids glycine, serine and alanine also activate C_4 PEPC from monocot grasses in the Poaceae family (Figure 1.4 B) (Nishikido and Takanashi, 1973; Doncaster and Leegood, 1987; González-Segura *et al.*, 2018). The extent to which plant PEPCs are affected by these metabolites differs between C_3 and C_4 PEPC isoforms; C_4 PEPCs have evolved greater sensitivity to G6P and tolerance to aspartate and malate compared to C_3 non-photosynthetic homologues (Dong *et al.*, 1998; Bläsing *et al.*, 2002; Svensson, Bläsing and Westhoff, 2003; Jacobs *et al.*, 2008). Thus, C_4 PEPCs have evolved specialised regulatory properties as the enzyme has become co-opted into a photosynthetic role.

In addition to allosteric control, plant PEPCs are regulated by the phosphorylation of a conserved serine residue at the N terminus of the protein, unlike bacterial or Archaeal PEPCs (Chollet, Vidal and O'Leary, 1996; Vidal and Chollet, 1997; Nimmo, 2003). PEPC is phosphorylated by PEPC kinase (at residue S15 in *Z. mays* C_4 PEPC) and de-phosphorylated by protein phosphatase type 2A, but the phosphorylation state of PEPC primarily depends on PEPC kinase activity (Carter *et al.*, 1991; Vidal and Chollet, 1997; Nimmo, 2003; Izui, 2004). Phosphorylation of PEPC alters the enzyme kinetics, increasing the sensitivity of PEPC to activation by

G6P and reducing the sensitivity to inhibitors malate and aspartate, thereby increasing flux through the enzyme *in vivo* (Carter *et al.*, 1991; Chollet, Vidal and O'Leary, 1996; Vidal and Chollet, 1997; Nimmo, 2003). C_3 PEPC and C_4 PEPC in C_4 species and are phosphorylated during the day in response to light, whereas C_4 PEPC in CAM species are phosphorylated during the night in response to the circadian clock (Carter *et al.*, 1991; Jiao *et al.*, 1991; Hartwell *et al.*, 1996, 1999; Smith *et al.*, 1996; Vidal and Chollet, 1997). Phosphorylation of C_4 and CAM PEPCs during the day and night, respectively, corresponds to the time of day when primary photosynthetic carbon fixation, catalysed by PEPC, is taking place (Chollet, Vidal and O'Leary, 1996; Hartwell *et al.*, 1996, 1999; Nimmo, 2003; Aldous *et al.*, 2014). The careful control of PEPC kinase expression in response to light levels in C_3 and C_4 leaves, the plant circadian clock in CAM species and cell metabolites in all species provides an additional level of regulation of PEPC enzyme kinetics in planta (Vidal and Chollet, 1997; Saze *et al.*, 2001; Nimmo, 2003).

1.2.2 The structure and catalytic properties of bacterial and plant PEPC

Most PEPCs can be classified as Archaeal PEPC or a form of PEPC that is highly conserved between bacteria and plants (Matsumura, Izui and Mizuguchi, 2006). Every Archaeal, bacterial and plant PEPC catalyses the same reaction and the catalytic mechanism appears to be conserved, but Archaeal PEPC is smaller than bacterial and plant PEPCs (Matsumura, Izui and Mizuguchi, 2006). Little is known about Archaeal PEPC, but many residues involved in enzyme structure, substrate or effector binding have been identified from bacterial and plant PEPCs which are closely evolutionarily related and structurally similar (Izui *et al.*, 2004; Matsumura, Izui and Mizuguchi, 2006; Paulus, Schlieper and Groth, 2013). The conservation of enzyme structure between bacterial and plant PEPCs have enabled residues and

mobile loops involved in catalysis and effector binding to be identified (Izui *et al.*, 2004). Understanding the structure of bacterial and plant PEPCs, and the enzyme residues involved in substrate and effector binding, provides an insight into enzyme catalytic and regulatory mechanisms and an important context for engineering improved C_4 PEPC variants.

1.2.2.1 The residues within bacterial and plant PEPC involved in quaternary structure and cofactor binding

PEPC in bacteria, equivalent to Class-1 PEPC in plants, is a tetramer composed of four identical monomeric subunits that form a square conformation termed a “dimer-of-dimers” (Figure 1.5 A). A Class-2 hetero-octameric PEPC isoform has also been discovered in plants, which is larger and has very different physical and regulatory properties (reviewed in O’Leary, Park and Plaxton, 2011). However, the work in this thesis is focused on the Class-1 *Setaria viridis* C_4 PEPC isoform and therefore the unusual Class-2 isoform is not considered further. In the tetrameric form of PEPC, salt bridges between E493 and R498 on opposite dimers hold the ‘dimer of dimers’ structure together (residues numbered according to *Z. mays* C_4 PEPC structure) (Figure 1.5 B) (Kai *et al.*, 1999). Each individual monomer comprises approximately 40 α-helices and an eight-stranded β barrel (Figure 1.5 B) (Smith, 1971; Matsumura *et al.*, 2002; Kai, Matsumura and Izui, 2003; Izui *et al.*, 2004). Most of the amino acid residues that are essential for the catalytic activity of the enzyme are located in the region around the C-terminal side of the β barrel which is consequently referred to as the active site (Figure 1.5 C) (Kai *et al.*, 1999; Matsumura *et al.*, 2002). Each monomer requires a divalent metal ion cofactor, typically Mg^{2+} or Mn^{2+}, that binds residues E566 and D603 of the *Z. mays* C_4 PEPC (Figure 1.5 C) (Kai, Matsumura and Izui, 2003).

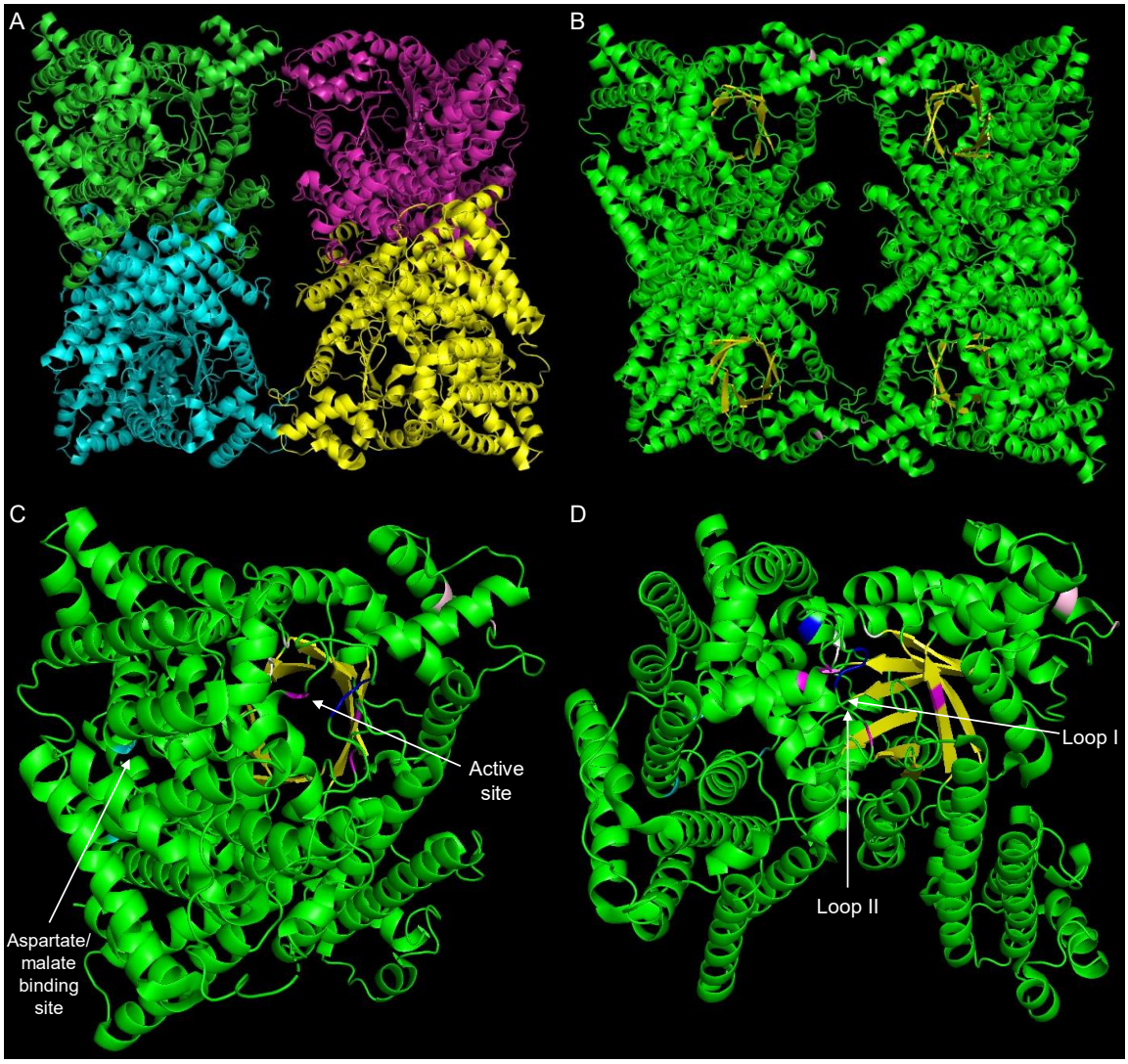


Figure 1.5: The structure of *Z. mays* C_4 PEPC. (A) PEPC has a tetrameric structure. Each monomer is highlighted in a different colour. (B) Each monomer is composed of 42 α-helices (green) and 8 β-strands (yellow), and residues that form salt bridges between dimers are coloured in pale pink. (C) A monomer is shown with cofactor, substrate and aspartate/malate binding residues highlighted. PEP binding residues are coloured in pink, magnesium binding residues in grey, bicarbonate binding residues in dark blue and aspartate/malate binding residues in light blue. (D) Catalytic loops I and II are highlighted, residues coloured as in part C. PDB: 5VYJ.

1.2.2.2 The residues within bacterial and plant PEPC involved in substrate binding and catalysis

PEPC requires the divalent metal ion to be bound for catalysis, followed by a preferential order of substrate binding in which PEP is bound first and bicarbonate second (Janc, O'Leary and Cleland, 1992). A crucial feature of PEPC is the multiple mobile loops that are important for catalysis and regulation and contain PEPC substrate and effector binding residues. For example, PEP interacts with the divalent metal ion and residues R456, R759 and R773 at the catalytic site around the C-terminal side of the β barrel (Figure 1.5 C, Figure 1.6) (Matsumura *et al.*, 2002; Kai, Matsumura and Izui, 2003), but is also thought to bind Loop I. Loop I, comprised of residues 640 to 649, including the functionally essential residue R647 thought to be involved in PEP binding, moves towards the active site when the enzyme is in a high-affinity state for its substrates and is then thought to form a lid to protect reaction intermediates from water during catalysis with another well-conserved mobile loop called Loop II (Figure 1.5 D) (Yano *et al.*, 1995; Kai, Matsumura and Izui, 2003; Izui *et al.*, 2004). Loop II, which extends from K762 to G768, contains residues involved in bicarbonate binding (Figure 1.5 D) (Kai *et al.*, 1999; Kai, Matsumura and Izui, 2003). Residues K606, K762, R763 and R764 appear to be involved in bicarbonate binding, elucidated from the crystal structures and mutagenesis of *E. coli* PEPC and *Z. mays* and *Flaveria trinervia* C_4 PEPC (Figure 1.5 C, D, Figure 1.6) (Gao and Woo, 1995; Kai *et al.*, 1999; Kai, Matsumura and Izui, 2003). Mobile loops I and II contain substrate binding residues and facilitate catalysis by PEPC.

In addition to the sites that have been shown to influence substrate binding, a number of sites have been shown to be important for interaction with reaction

intermediates. Specifically, site-directed mutagenesis has demonstrated that H177 is essential for catalysis and participates in the carboxylation of the catalytic intermediate enolate (Figure 1.6) (Terada and Izui, 1991). This conserved residue is also on a mobile loop, from residue F174 to R184, that moves towards the active site when the enzyme is in a high affinity state (Kai, Matsumura and Izui, 2003). Mutagenesis of H579, the only other conserved histidine residue in *E. coli* PEPC (H639 in *Z. mays* C_4 PEPC), reduced the affinity of PEPC for PEP and reduced the maximum velocity of PEPC, indicating the involvement of this residue in catalytic activity (Figure 1.6) (Terada, Murata and Izui, 1991). Site-directed mutagenesis of conserved residues has led to the identification of many sites involved in substrate binding and catalysis which must be understood to effectively engineer PEPC.

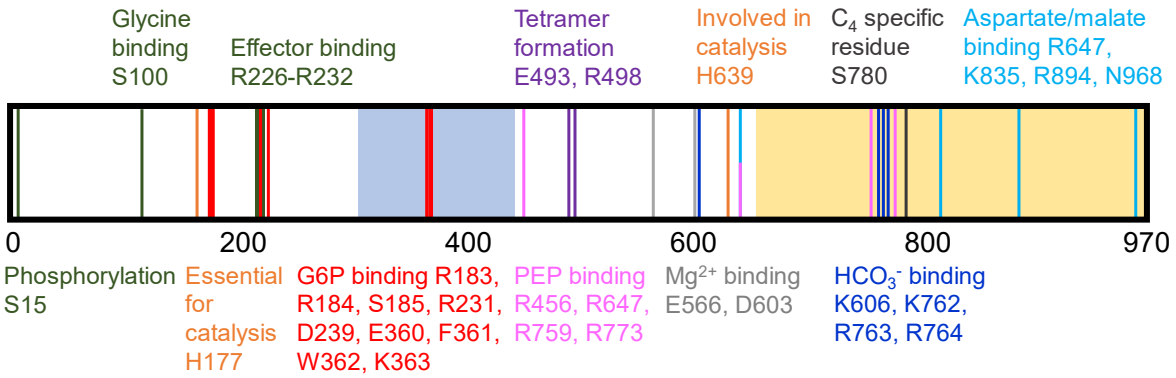


Figure 1.6: A schematic diagram of *Z. mays* C_4 PEPC residues involved in quaternary structure, substrate and/or effector binding or are essential for catalysis. Regions 301 to 443 and 651 to 970 that are influential in C_4 PEPC enzyme kinetics are highlighted in blue and yellow, respectively.

1.2.2.3 The residues within PEPC involved in effector binding

PEPC is subject to tight regulation by metabolic effectors due to the central role that oxaloacetate production and TCA cycle function have in carbon flow and greater cell metabolism. In addition to our detailed understanding of the sites that are involved in substrate binding and catalysis, X-ray crystallography of the *E. coli* PEPC complexed with the allosteric inhibitor aspartate identified that R587, K773, R832 and N881 in *E. coli* (orthologous to R647, K835, R894, N968 in *Z. mays* C_4 PEPC) are involved in aspartate binding (Figure 1.6) (Kai *et al.*, 1999). The allosteric inhibitor malate is thought to bind the same site (Paulus, Schlieper and Groth, 2013), supported by site-directed mutagenesis that showed that mutagenesis of site K835 or R894 in *Z. mays* C_4 PEPC reduces the sensitivity of the enzyme to inhibition by both aspartate and malate (Endo *et al.*, 2008).

Structural studies in both *E. coli* PEPC and *Z. mays* C_4 PEPC show that when aspartate is bound to the PEP binding residue R647 located on Loop I, the residue is moved 20 Å away from the catalytic site providing insight into the mechanism of inhibition (Kai *et al.*, 1999; Matsumura *et al.*, 2002). However, mutagenesis of a conserved lysine residue K620 in *E. coli* (not studied in *Z. mays* C_4 PEPC), reduces inhibition by malate and aspartate despite not being part of the catalytic or inhibitor binding sites (Yano and Izui, 1997). Thus, inhibitor binding residues have been identified, and aspartate binding preventing movement of Loop I appears to be part of the mechanism of inhibition (Izui *et al.*, 2004). However, involvement of residues that do not directly bind aspartate or malate reveal that the mechanism of inhibition is not fully understood, and it is likely that there are more sites involved in inhibition that are yet to be identified.

Crystal structures of *Z. mays* C_4 PEPC have elucidated the binding pocket for allosteric activator G6P. R183, R184, S185, R231, D239, E360, F361, W362 and K363, are involved in allosteric activation of PEPC by G6P, in agreement with previous data showing that the region between residues 301 and 443 is essential for G6P activation (Figure 1.6) (Bläsing, Westhoff and Svensson, 2000; Muñoz-Clares *et al.*, 2020). Moreover, these G6P binding residues are absolutely conserved across C_3 and C_4 PEPCs, apart from lysine at residue 363, which can be arginine in some monocot C_4 species, and glutamic acid at residue 360, which has changed to aspartic acid in some C_3 species (Muñoz-Clares *et al.*, 2020). Very little divergence between species at these residues clearly demonstrates the importance of this allosteric site in the regulation of plant PEPC; residues R183, R184, S185, R231, D239 are highly conserved between bacteria and plants but E360, F361, W362 and K363 do not have equivalent residues in bacterial PEPC as they belong to a plant PEPC-specific loop (Muñoz-Clares *et al.*, 2020). It has also been suggested that the negative charge of residues D504 and D509 in PEPC2 from CAM species *Phalaenopsis equestris* and *Kalanchoë fedtschenkoi*, respectively, equivalent to residue H519 in *Z. mays*, could increase enzyme activity by conferring similar conformational changes to PEPC to those that occur on G6P binding (Yang *et al.*, 2017). The high conservation of residues involved in G6P binding across bacterial and plant PEPCs suggest that activation by G6P has a high level of importance in regulation of PEPC activity, especially in plants.

The glycine binding site, an activator of PEPC in monocot grasses, is in a similar but not identical location to the G6P binding site (González-Segura *et al.*, 2018). Serine at residue 100 in the *Z. mays* C_4 PEPC is essential for glycine binding and simulations suggest that serine and alanine, also activators of PEPC in monocot

grasses, could also bind this site (Figure 1.6) (González-Segura *et al.*, 2018). The *Z. mays* S100K mutant C_4 PEPC had kinetics that were more similar to the dicot *Amaranthus hypochondriacus* (amaranth) C_4 PEPC than the *Z. mays* C_4 PEPC, namely an increased affinity of the enzyme for PEP and G6P and a reduced affinity of the enzyme for malate, demonstrating the importance of this residue in C_4 PEPC kinetics specific to monocot grasses (Güémez-Toro, Mújica-Jiménez and Muñoz-Clares, 2012; González-Segura *et al.*, 2018). Accordingly, monocot sedges and eudicots, including *Flaveria* C_4 PEPC, have a lysine residue conserved at this position (Muñoz-Clares *et al.*, 2020). Moreover, mutagenesis of residues on a loop between R226 and R232 altered the affinity of *Z. mays* C_4 PEPC for both G6P and glycine as well as inhibitors malate and aspartate, highlighting the importance of this loop in regulation by allosteric effectors, despite the loop not containing any known effector binding residues (Figure 1.6) (Yuan *et al.*, 2006). Such site-directed mutagenesis experiments identify residues that confer differences in C_4 PEPC kinetics between species which could be target sites for mutagenesis for crop improvement. Advances in crystallography, along with increasing biochemical evidence, continue to improve our understanding of allosteric control of PEPC.

1.2.3 The molecular evolution of C_4 PEPCs

There are three major C_4 photosynthetic pathways, all of which require PEPC to catalyse the first committed carbon fixation step in the primary carbon assimilation tissue (Figure 1.3). In each lineage that has evolved C_4 photosynthesis, C_4 PEPC has evolved from an ancestral non-photosynthetic C_3 isoform (Kellogg, 1999; Jacobs *et al.*, 2008). In doing so, C_4 PEPCs have evolved altered enzyme kinetics, regulation, and expression patterns compared to ancestral C_4 PEPC homologues (Svensson, Bläsing and Westhoff, 2003; Westhoff and Gowik, 2004; Moody, Christin and Reid, 2020). The multiple independent evolutionary replicates of C_4 PEPC across time and phylogenetic distance provide an opportunity to observe how natural selection has adapted the molecular sequence of the enzyme for its function in the C_4 photosynthetic pathway (Niklaus and Kelly, 2019).

Most kinetic studies of closely related C_3 and C_4 PEPC isoforms from genera including *Flaveria*, *Panicum* and *Alternanthera* show that C_4 PEPCs have evolved a lower affinity for the substrate PEP and a higher affinity for bicarbonate compared to C_3 PEPC, demonstrating a high level of convergent evolution for these kinetic traits (Ting and Osmond, 1973; Bauwe and Chollet, 1986; Svensson, Bläsing and Westhoff, 1997; Bläsing, Westhoff and Svensson, 2000; Gowik *et al.*, 2006; Jacobs *et al.*, 2008; Rosnow, Edwards and Roalson, 2014; DiMario and Cousins, 2019; Moody, Christin and Reid, 2020). Conversely, C_4 PEPCs from the genus *Alloteropsis* appear to have evolved a lower affinity for both substrates PEP and bicarbonate over time, indicating that although C_4 PEPC kinetics have largely converged between different C_4 lineages, there are exceptions (Phansopa *et al.*, 2020). In addition, C_4 PEPCs are less sensitive to malate inhibition, more sensitive to activation by G6P, and demonstrate a higher degree of activation by

phosphorylation compared to C_3 PEPCs (Svensson, Bläsing and Westhoff, 1997; Bläsing, Westhoff and Svensson, 2000; Gowik *et al.*, 2006; Jacobs *et al.*, 2008; Paulus, Schlieper and Groth, 2013; Rosnow, Edwards and Roalson, 2014; Moody, Christin and Reid, 2020). Understanding the molecular changes that underpin the differences between C_3 and C_4 enzyme kinetics could reveal opportunities to engineer C_4 crops to increase flux through the photosynthetic pathway.

In order to identify the molecular changes to C_3 PEPCs that confer these altered characteristics, plants from the genus *Flaveria* have formed a well-studied model system. The *Flaveria trinervia* C_4 PEPC and *Flaveria pringlei* C_3 PEPC orthologs are particularly useful because they share 94% amino acid identity and exhibit typical C_3 and C_4 PEPC kinetics (Jacobs *et al.*, 2008). Thus, most of our knowledge about regions of the enzyme that contribute to C_3 or C_4 enzyme kinetics have been identified through domain swapping between these two orthologous *Flaveria* PEPC variants, as well as site-directed mutagenesis experiments (Svensson, Bläsing and Westhoff, 1997; Bläsing, Westhoff and Svensson, 2000; Engelmann, Bläsing and Westhoff, 2002; Jacobs *et al.*, 2008; Paulus, Schlieper and Groth, 2013; DiMario and Cousins, 2019).

Regions 301 – 443 and 651 – 970 account for the majority of difference in affinity of *Flaveria* C_3 and C_4 PEPCs for PEP (regions numbered according to *Z. mays* C_4 PEPC), with the influence of region 651 – 970 conferred mostly by S780 which is conserved across C_4 PEPCs and required for C_4 PEPC kinetics (Figure 1.6) (Bläsing, Westhoff and Svensson, 2000; Engelmann, Bläsing and Westhoff, 2002; Christin *et al.*, 2007; DiMario and Cousins, 2019). The mutation from serine to alanine at residue 780 made to *F. trinervia* C_4 PEPC, characteristic of C_3 and CAM C_4 PEPCs, conferred an increased affinity for PEP concomitant with a reduced

affinity for bicarbonate, indicating a trade-off between affinity of PEPC for PEP and bicarbonate (Bläsing, Westhoff and Svensson, 2000; DiMario and Cousins, 2019). In addition, in Suaedoideae C_4 species, residue 733 is thought to influence PEP kinetics (Rosnow, Edwards and Roalson, 2014). The significance of other residues may be of some importance but is yet to be fully understood. For example, a comparison of C_3 and C_4 *Flaveria* PEPC sequences and chimeric enzyme kinetics identify a lysine residue at position 347 in *F. trinervia* C_4 PEPC, that is arginine at the same residue in *F. pringlei* C_3 PEPC, in a region of the enzyme that is influential in determining the affinity of each enzyme for PEP (Westhoff and Gowik, 2004). Lysine at this residue is conserved in *Z. mays* C_4 PEPC (K353) so this residue could be essential for C_4 enzyme kinetics, but the significance of the site is still unknown. The lower affinity for PEP that is characteristic of C_4 PEPC could be required for enzyme function at high PEP concentrations in the carbon assimilation tissue, to ensure lower enzyme activity at night, or it could be a necessary trade-off that enables an increased affinity for bicarbonate or reduced sensitivity to malate (Bauwe and Chollet, 1986; Jacobs *et al.*, 2008; DiMario and Cousins, 2019).

In contrast to PEP affinity, a higher affinity for bicarbonate has been strongly selected for in C_4 PEPCs in C_4 species. C_4 PEPC with a higher affinity for bicarbonate results in a higher hypothetical rate of photosynthesis at low CO_2 concentrations where PEPC is known to be rate limiting (DiMario and Cousins, 2019; von Caemmerer, 2021). This suggests that PEPC with a higher affinity for bicarbonate can support a higher rate of carbon assimilation under conditions of heat or drought stress when stomatal conductance is low (DiMario *et al.*, 2021). C_4 PEPC has a higher affinity for bicarbonate compared to C_3 isoforms, typically between 24 – 47 µM compared to 60 – 80 µM, respectively (Bauwe, 1986; DiMario

and Cousins, 2019; Moody, Christin and Reid, 2020; DiMario *et al.*, 2021). The affinity of C_4 PEPC in grass species (Poaceae) for bicarbonate is partially determined by region 528 – 880 and is largely influenced by S780, similar to PEP kinetics (DiMario and Cousins, 2019; DiMario *et al.*, 2021). It appears that bicarbonate binding kinetics are the product of multiple interacting sites in C_4 PEPC, some of which are yet to be identified. The differences in enzyme kinetics between C_3 and C_4 PEPC isoforms likely better adapt each isoform for replenishing the TCA cycle or photosynthetic carbon fixation functions, respectively. Accordingly, C_4 PEPC with a higher affinity for bicarbonate is thought to be desired in the effort to engineer more efficient C_4 photosynthetic pathway (DiMario *et al.*, 2021).

Regions of C_4 PEPC responsible for different regulatory properties relative to C_3 PEPCs have also been identified by characterising the kinetics of enzyme chimera between *F. trinervia* C_4 PEPC and the C_3 *F. pringlei* orthologue. For example, G6P activation appears to be determined by region 301 – 443 in *Flaveria* (*Z. mays* C_4 PEPC numbering) (Bläsing, Westhoff and Svensson, 2000). In addition, C_4 PEPC exhibits a lower sensitivity to inhibitors malate and aspartate, which may have evolved to facilitate efficient PEPC function in the presence of a high concentration of these C_4 photosynthetic intermediates that occurs *in vivo* (Bläsing *et al.*, 2002; Jacobs *et al.*, 2008). Accordingly, it has been shown that a high tolerance to malate is determined by residues 651 – 970 in the C terminus of *F. trinervia* C_4 PEPC, but this was not dependent on residue 780 (Jacobs *et al.*, 2008). In addition, an R884G mutation (G890 in *Z. mays* C_4 PEPC) has occurred multiple times independently as PEPC has become co-opted into the C_4 photosynthesis pathway and accounts for increased tolerance of C_4 PEPC to feedback inhibition by malate and aspartate (Paulus, Schlieper and Groth, 2013). The chimeric and mutated versions of PEPC

characterised in these studies have been expressed in, and purified from, *E. coli* and therefore are not phosphorylated (Bläsing, Westhoff and Svensson, 2000; Bläsing *et al.*, 2002; Jacobs *et al.*, 2008; Paulus, Schlieper and Groth, 2013). It is notable, then, that PEPC kinetics *in planta* are determined by a complex interaction between the amino acid sequence and regulation of PEPC, by both phosphorylation and effectors (Carter *et al.*, 1991; Chollet, Vidal and O'Leary, 1996; Vidal and Chollet, 1997; Nimmo, 2003). Nevertheless, understanding the different ways in which C_3 and C_4 PEPC are regulated by metabolic effectors, and which residues contribute to these differences, will assist in identifying mutations that confer desired changes to enzyme kinetics that could be tested in C_4 crop plants.

Taken together, these enzymatic studies show that PEPC kinetics are determined by the interaction of multiple regions of the enzyme and that many sites influence allosteric control of the enzyme. Regions 301 – 443 and 651 – 970 appear to be particularly important in substrate and effector binding kinetics, however, within those regions of the enzymes little is known about which residues are most important, and therefore there are likely to be a number of residues that are influential that are yet to be discovered. Furthermore, the mutagenesis studies above have repeatedly shown that a single amino acid substitution can confer large changes in PEPC kinetics, for example the influence of site 780 on the affinity of C_4 PEPC for substrates PEP and bicarbonate (Bläsing, Westhoff and Svensson, 2000; DiMario and Cousins, 2019). The size of PEPC is prohibitive for analysis of each site individually in site-directed mutagenesis experiments. However, a comparative genomics approach combined with a high-throughput experimental method could greatly increase the rate at which residues that influence enzyme kinetics are identified. Should any sites confer beneficial characteristics in a C_4 photosynthesis

context, for example a higher catalytic turnover or a higher affinity and/or specificity for bicarbonate, then these sites would then become key targets for potential crop improvement efforts.

1.3 Directed evolution of plant enzymes in *Escherichia coli*

Engineering an enhanced version of an enzyme for a specific function has many applications ranging from the production of precursors to biofuels to the synthesis of pharmaceuticals (Fasan *et al.*, 2011; Lian *et al.*, 2014; Li, Wang and Reetz, 2018). The performance of different enzyme variants is often illustrated as a protein fitness landscape, where sequence space is represented across two dimensions and a third dimension represents the relative 'fitness', or suitability, of the enzyme for a given function (Figure 1.7) (Romero and Arnold, 2009). Experimental improvement of an enzyme for a given function can therefore be considered an 'adaptive walk' on this landscape towards increased fitness (Figure 1.7). The model is inevitably complicated by the vast size of the sequence space and multitude of parameters that contribute to the relative fitness of an enzyme (Romero and Arnold, 2009). Thus, the true possible diversity of any given enzyme can only ever be partially interrogated experimentally. Nevertheless, by screening a large number of enzyme variants using an artificial selection pressure it is possible to make progress towards an improved enzyme for a desired purpose (reviewed in Wang *et al.*, 2021).

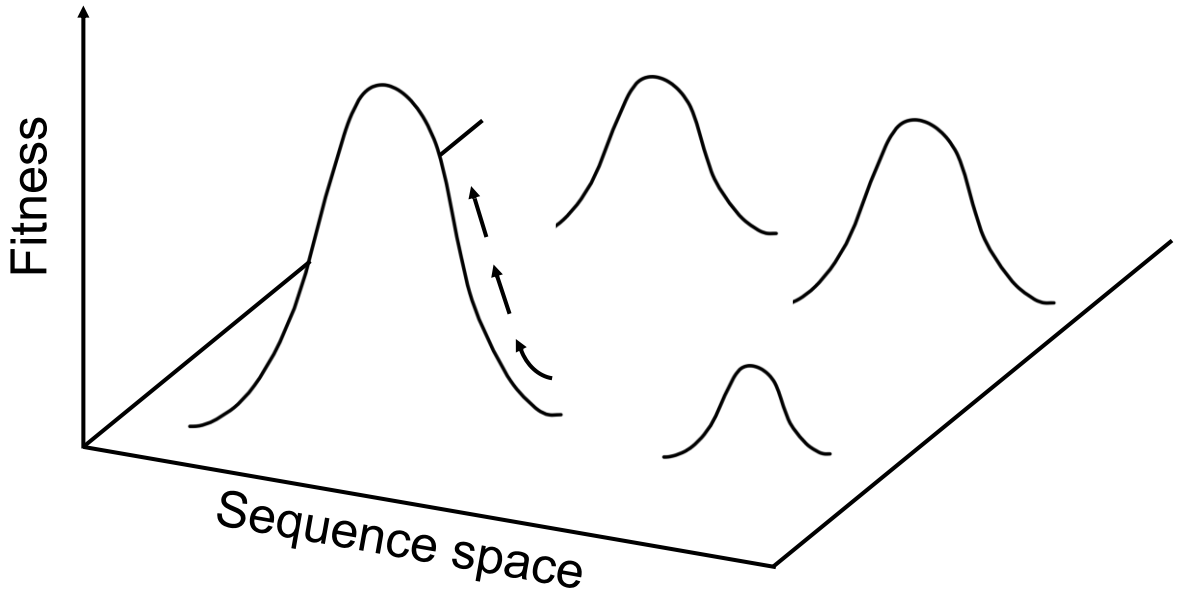


Figure 1.7: A cartoon representation of a simplified protein fitness landscape. Arrows represent an adaptive walk towards the global fitness optimum whereas other peaks represent local fitness optima.

Directed evolution, the process by which protein variants are screened in high throughput to identify mutations that are beneficial for a desired characteristic, has revolutionised protein engineering over the last 30 years (Chen and Arnold, 1993, 2020; Romero and Arnold, 2009). The process employs successive rounds of gene diversification followed by screens of enzyme variants for improved function. Large libraries produced during the gene diversification step, and high throughput screening of variants, aims to interrogate the protein fitness landscape towards a global fitness optimum. Further, the experimenter chooses the selection pressure to create a versatile model for biological optimisation. Previous directed evolution studies have engineered enzymes with improved catalytic activity, stability and substrate selectivity (recently reviewed in Wang *et al.*, 2021). Directed evolution is therefore a powerful tool for rapidly generating and identifying enzymes with desired characteristics that could be applied to confer phenotypic improvements to crop plants (Gionfriddo, De Gara and Loreto, 2019).

Escherichia coli is a popular host for directed evolution experiments. The use of highly competent *E. coli* that have a high transformation efficiency can facilitate screens of large libraries of enzyme variants (Mueller-Cajar and Whitney, 2008). However, there are challenges associated with expressing a plant enzyme in a heterologous host such as *E. coli*. Differences in codon usage, gene expression control mechanisms and post-translational modification between eukaryotic and prokaryotic systems can affect protein folding, function and expression level (Kaur, Kumar and Kaur, 2018; Tripathi and Shrivastava, 2019; Bhatwa *et al.*, 2021). C_4 PEPC is post-translationally phosphorylated, which confers changes to its sensitivity to effectors (reviewed in Chollet, Vidal and O'Leary, 1996; Nimmo, 2003). However, *E. coli* PEPC lacks a phosphorylation domain and *E. coli* lacks PEPC

kinase so *E. coli* PEPC is not regulated by phosphorylation *in vivo* (Macek *et al.*, 2008; Keseler *et al.*, 2021). Therefore, a plant PEPC would not be phosphorylated when subject to heterologous expression in *E. coli*.

Chemical inhibition and RNA interference of PEPC kinase have both been undertaken to interrogate the importance of C_4 PEPC phosphorylation on photosynthetic carbon fixation *in planta*. Chemical inhibition of PEPC kinase in maize and sorghum (*Sorghum vulgare*) resulted in a reduced rate of photosynthetic carbon assimilation in both species, and evidence of a transition to a C_3 mode of carbon fixation in sorghum, which suggests that phosphorylation of C_4 PEPC is essential for a functional C_4 cycle in maize and sorghum (Bakrim *et al.*, 1993). Similarly, reduction of PEPC kinase expression through RNAi in the CAM species *K. fedtschenkoi* greatly reduced the level of phosphorylated PEPC in the dark, when primary photosynthetic carbon fixation catalysed by PEPC occurs in CAM species, and resulted in a concomitant reduction in photosynthetic carbon assimilation rate in the dark (Boxall *et al.*, 2017). However, RNAi knockdown of PEPC kinase in the C_4 species *Flaveria bidentis* also reduced the level of PEPC phosphorylation, but this had no impact on carbon assimilation rate or leaf malate content compared to wild-type *F. bidentis* plants (Furumoto *et al.*, 2007). Thus, it appears that the importance of C_4 PEPC phosphorylation on C_4 cycle function is species-specific. Accordingly, C_4 PEPC variants that have been identified from directed evolution studies of the enzyme in *E. coli* must subsequently be characterised *in planta*.

There are numerous examples of directed evolution of plant enzymes in *E. coli*, despite differences in gene expression systems. Examples include generation of a more soluble hydroxynitrile lyase from cassava or a plant glutathione-S-transferase (GST) with enhanced enzymatic activity (Asano *et al.*, 2011; Chronopoulou *et al.*,

2018). *E. coli* metabolism can also be engineered to create an artificial selection pressure on an enzyme in the bacterial cell; expression of multiple plant genes was required to evolve arogenate dehydratase from *Arabidopsis* that is more resistant to feedback inhibition (Leong and Hanson, 2023). In an alternative approach, multiple bacterial enzymes were subject to directed evolution in *E. coli* and combined to convert glycolate to RuBP, the application of which would be to engineer an alternative photorespiration pathway in plants that would avoid carbon loss and could improve agricultural productivity (Trudeau *et al.*, 2018). These studies have all used *E. coli* to screen for improved variants of plant enzymes, or enzymes with desirable characteristics for future application *in planta*, demonstrating the effectiveness and versatility of directed evolution using *E. coli* for agricultural applications.

Directed evolution studies using *E. coli* have also been shown to produce enzymes that are beneficial when engineered into plants. For example, 5-enolpyruvylshikimate-3-phosphate synthase (EPSPS) variants screened in *E. coli* for increased resistance to the herbicide glyphosate conferred an improved glyphosate-resistance phenotype when engineered into *Arabidopsis thaliana* (Mao *et al.*, 2017). In addition, a thermostable β-glucuronidase (GUS) variant identified in a screen for thermostability using *E. coli* as the host organism demonstrated high levels of activity when expressed in *Arabidopsis* seedlings that were subject to heat treatment up to and including 80 °C (Xiong *et al.*, 2007, 2011). Thermostable GUS could be useful as an improved reporter system for analysis of gene expression in plants exposed to high temperatures, or where protocols include experimental steps at high temperatures prior to gene expression visualisation. These examples demonstrate that despite differences in gene expression between organisms in

different biological kingdoms, use of *E. coli* as a host organism for directed evolution is effective for identifying improved enzyme variants and introducing desired traits into plants.

In the area of photosynthesis research, *E. coli* has been the host organism for a number of studies that have been designed to select for improved carbon fixation by Archaeal Rubisco, for direct application *in planta* (Wilson, Alonso and Whitney, 2016; Wilson *et al.*, 2018; Zhou and Whitney, 2019). Tobacco transformed with Archaeal Rubisco mutant variants that had been selected for in an *E. coli* screen conferred improved plant growth compared to tobacco transformed with the wild-type Archaeal Rubisco, despite the challenges of using heterologous hosts to screen for an improved Archaeal Rubisco in *E. coli* and then engineering promising enzyme variants from the screen into the chloroplast of higher plants (Wilson, Alonso and Whitney, 2016). Taken together, the translation of findings from screening proteins in *E. coli* that confer useful agronomic traits in plants, related to photosynthesis or otherwise, presents a valuable opportunity to engineer enzymes in high-throughput, which has the potential to revolutionise plant research.

1.4 Regulation of genome edited crops around the world

Any mutations that confer an improvement to C_4 PEPC must be introduced to the genome of C_4 crop species in order to translate into higher yielding crop plants. In recent years, a suite of genome editing (GE) techniques have been developed that enable a precise alteration to be made at a specific location in the plant genome and do not require any transgenes to remain in the plant following the edit being made (Metje-Sprink *et al.*, 2019; Gu, Liu and Zhang, 2021; Sprink, Hartung and Metje-Sprink, 2022). Around the world, GE crop plants are subject to equivalent or less cumbersome regulation compared to genetically modified (GM) crop plants (discussed below). Therefore, any improvement that can be made to C_4 PEPC that can be introduced to a plant *via* a GE approach, such as mutating the enzyme at a single site, will assist in the commercialisation of crop plants by comparison to changes that are more substantial or require transgene expression that can only be achieved using a GM approach. The global outlook on current regulation of GE crop plants is considered here.

1.4.1 Regulation of genetically modified and genome edited crop plants in the United Kingdom and European Union

GM crops have been met with resistance from the European community and regulators. The current regulations of genetically modified organisms (GMO) in the European Union (EU) permit a small number of genetically modified crop plants on a case-by-case basis for food and feed (import and consumption) including varieties of cotton, oilseed rape, sugar beet, soybean and maize (European Commission, 2023). However, only one GM crop has been approved for cultivation and not subsequently withdrawn (pending renewal): a variety of insect-resistant maize MON810 (Halford, 2019; European Commission, 2023).

The legislation that underpins the current EU GMO policy, Directive 2001/18/EC of the European Parliament and of the Council and Regulation 1829/2003/EC, defines a GMO as “an organism… in which the genetic material has been altered in a way that does not occur naturally” (European Commission, 2001, 2003). In practice, crop plants in Europe have long been subject to intensive mutagenesis techniques or hybridisation efforts that are not possible without human intervention, so the process by which a mutation has been introduced is used to determine if a plant is a GMO or not (Halford, 2019). For example, plants that have been subject to the introduction of foreign DNA by *Agrobacterium*-mediated transfer, electroporation, or particle bombardment are all considered GMOs (European Commission, 2001; Halford, 2019).

Crops that have been subject to GE are genotypically indistinguishable from plants in which an equivalent mutation occurred naturally, and phenotypically indistinguishable from a plant in which the same trait has been introduced by mutagenesis and breeding that is not subject to GMO regulation (Metje-Sprink *et al.*, 2019; Sowa *et al.*, 2021). Multiple technologies facilitate changes to a plant genome that are considered GE, including clustered, regularly interspaced, short palindromic repeats (CRISPR), transcription activator-like effector nucleases (TALENs) and zinc-finger nucleases (ZFN) (Halford, 2019; Metje-Sprink *et al.*, 2019; Sowa *et al.*, 2021). Indeed, GE can be used to introduce a single targeted mutation to a plant in order to confer a specific trait without the off-target mutations that result from breeding programs that employ mutagenesis for genetic diversification (Holme, Gregersen and Brinch-Pedersen, 2019; Metje-Sprink *et al.*, 2019; Sowa *et al.*, 2021).

The way that GM and GE crops are currently regulated is under scrutiny in the UK and EU. In 2021, a public consultation in the UK by the Department for Environment, Food and Rural Affairs (DEFRA) recommended that legislation should be changed to i) exclude organisms that were produced by GE from the definition of a GMO, and ii) update regulations to be able to bring these crops to market (DEFRA, 2021). These recommendations have since been passed into law with the Genetic Technology (Precision Breeding) Act 2023. Asynchronously, genome edited organisms (GEOs) are currently still regulated in the same way as GMOs in Europe (Confédération paysanne *et al.*, 2018). However, there is an ongoing public consultation by the European Commission on the regulation of plants obtained by targeted mutagenesis, initiated based on evidence that current legislation is 'not fit for purpose' (European Commission, 2021). Together, the Genetic Technology Act 2023 and public consultation in the EU represent progress towards evidence-based agricultural regulatory systems in Britain and the EU, including the approval of GE crops for cultivation.

1.4.2 Acceptance of genome edited crops around the world

GE crops have been approved for commercialisation or are exempt from regulation in many countries that have previously adopted GM crops. Indeed, many countries around the world have already accepted cultivation of crops produced by biotechnologies other than traditional breeding and cultivate GM crops at scale; ~200 million hectares of land is used to grow GM crops globally (ISAAA, 2019; Turnbull, Lillemo and Hvoslef-Eide, 2021). The top five producers of GM crops by land area cultivated were the USA, Canada, Brazil, Argentina and India in 2019, the year for which we have the most recent information (ISAAA, 2019). Thus, the current regulation of GE crops in these five countries will be considered here. All five

countries that cultivate the highest area of land of GM crops currently regulate GE crops within the existing guidelines that apply to GM crops or under less stringent regulation by comparison.

The type of mutation that has been made to a GE crop affects the way in which it is regulated. The three main categories of mutagenesis are characterised by the repair mechanism for a double-strand break introduced to the plant DNA by a site directed nuclease (SDN). The break can be repaired by the plant at random (SDN-1), using a very short repair template (SDN-2), or with a large repair template that remains in the plant (SDN-3) (Turnbull, Lillemo and Hvoslef-Eide, 2021; Sprink, Wilhelm and Hartung, 2022). In the case of SDN-1 or -2, foreign DNA involved in the production of the edit can be removed by breeding and therefore these are generally not considered GMO or subject to GMO regulation in countries where GE crops are accepted (Lema, 2019; Gatica-Arias, 2020; Sprink, Wilhelm and Hartung, 2022).

Brazil, Argentina and India do not consider crops produced by SDN-1 or SDN-2 a GMO and the approval of GE crops is assessed on a case-by-case basis (Lema, 2019; Gatica-Arias, 2020; Sprink, Wilhelm and Hartung, 2022). Further, there is initial evidence that GE crop plants are subject to faster approval compared to GM crops in Argentina (Whelan, Gutti and Lema, 2020). In the US and Canada, GE crops fall within existing frameworks for GM crop regulation (Wolt and Wolf, 2018; Sprink, Wilhelm and Hartung, 2022). Broadly, this means that if GE crops produced by SDN-1 or -2 have equivalent traits to existing crops then these are subject to existing regulation or exempt from requiring authorisation (Wolt and Wolf, 2018; Sprink, Wilhelm and Hartung, 2022). Thus, GE crops are in the process of being adopted in many countries around the world.

1.5 Thesis aims and outline

Food and fuel demands are continuing to increase without a concomitant increase in crop yields (Ray *et al.*, 2013; Long, Marshall-Colon and Zhu, 2015). Gains to crop yields must increase to meet the demands of an increasing global population but yield gains of key agronomic crop plants are slowing or stagnating (Long, Marshall-Colon and Zhu, 2015). It is therefore necessary to develop new methods that will facilitate identification of targets for engineering crop plants to improve photosynthesis and boost crop yields. The aim of this thesis is to develop a high-throughput pipeline to identify mutations to PEPC that could plausibly improve photosynthesis in C_4 plants, which could be introduced by genome editing and thus bypass burdensome regulation applied to GM plants. This will be achieved by screening large libraries of PEPC variants in *E. coli* to identify variants that could be engineered into plants by genome editing. This screen will be undertaken using *Setaria viridis* C_4 PEPC (SvPEPC) as a starting material and a proof-of-concept, but once developed the pipeline could be applied to PEPC from other C_4 crop species.

PEPC catalyses the first committed carbon fixation step in the C_4 cycle and is rate-limiting to photosynthetic flux at low mesophyll cell CO_2 concentrations so is a key target for improving photosynthetic efficiency in C_4 crops. I will target the sites under positive selection in the C_4 isoform of PEPC to generate libraries of mutant variants of C_4 enzymes. I will then perform high-throughput screens of PEPC variants in *E. coli*, where growth of *E. coli* serves as a proxy for enzyme activity, to identify putative improved enzyme variants for C_4 cycle function. Should variants with improved kinetics be identified, then these would be prioritised for engineering into plants. Thus, the outcomes of this thesis will improve our understanding of the molecular

basis of different C_4 PEPC kinetics between species and identify targets sites for engineering crops with the aim of enhancing photosynthetic capacity.

Chapter 2: Materials and methods

2.1 Bacterial strains, growth conditions and media

2.1.1 Bacterial strains

DH5α Competent Cells for Subcloning were obtained from ThermoFisher Scientific.

E. coli K-12 BW25113 and *E. coli* Δppc (strain JW3928-1, ppc-742(del)::kan) were obtained originally from the Keio Collection (Baba *et al.*, 2006).

2.1.2 Bacterial growth media

Bacteria were grown in Luria-Bertani (LB) medium (1% w/v tryptone, 0.5% w/v yeast extract, 0.5% w/v NaCl) or on LB-agar plates (add 1.5% w/v agar) for cloning purposes. Cells were recovered in SOC medium (2% w/v tryptone, 0.5% w/v yeast extract, 10 mM NaCl, 2.5 mM KCl, 10 mM $MgSO_4$, 0.36% w/v glucose, 10 mM $MgCl_2$) after electroporation or heat-shock. Bacteria were grown in M9 glucose minimal medium to establish an expression level of PEPC that is growth rate limiting to *E. coli* Δppc or to undertake screens of C_4 PEPC variants.

M9 glucose minimal medium composition was based on Geerlof, 2010.

Component	Volume	Contents	Final concentration
M9 salt solution (10x) pH 7.2	100 mL	NaH_2PO_4 KH_2PO_4 NaCl NH_4Cl	50 mM 22.0 mM 8.55 mM 9.35 mM
20% glucose	20 mL		0.4% w/v
1 M $MgSO_4$	1 mL		1 mM
1 M $CaCl_2$	0.3 mL		0.3 mM
Biotin (1 mg/mL)	1 mL		1 µg
Thiamin (1 mg/mL)	1 mL		1 µg
Trace elements solution (100x)	10 mL		1X
Water (sterile ddH_2O)	866.7 mL		

Trace elements solution (100X)	EDTA $FeCl_3$-$6H_2O$ $ZnCl_2$ $CuCl_2$-$2H_2O$ $CoCl_2$-$2H_2O$ H_3BO_3 $MnCl_2$-$4H_2O$	13.4 mM 3.1 mM 0.62 mM 76 µM 42 µM 162 µM 8.1 µM

M9 glucose plates were made up to identical concentrations of the components above with 1.6% w/v agar.

Antibiotics were used at the following concentrations:

Antibiotic	Working concentration
Gentamycin	25 µg/mL
Spectinomycin	100 µg/mL
Ampicillin	50 µg/mL
Kanamycin	50 µg/mL or 30 µg/mL when working with *E. coli* Δppc

2.1.3 Transformation of *E. coli* DH5α competent cells

50 µL DH5α Competent Cells for Subcloning were thawed on ice for 15 minutes, then DNA to be transformed (100 pg - 100 ng plasmid DNA or 1-2 µL Golden Gate reaction product) was mixed with the cells and the vial of cells incubated on ice for a further 30 minutes. A heat shock was delivered by exposing the outside of the vial of cells to a 42 ˚C water bath for 45 seconds then immediately returning the cells to the ice for a further two minutes. After the heat shock, 500 µL – 1 mL SOC was added to the vial and the bacterial culture incubated at 37 ˚C, 200 rpm, for 45 minutes. A fraction of the culture was then plated on selective LB/antibiotic plates and incubated overnight at 37 ˚C.

2.1.4 Preparation and transformation of *E. coli* Δppc electrocompetent cells

E. coli Δppc were made electrocompetent according to the Langdale Lab (University of Oxford) protocol, or with minor modifications thereof as described in Chapter 4.

In brief, 50 mL LB was inoculated with *E. coli* (BW25113 or Δppc) and the culture incubated overnight at 37 ˚C, 200 rpm. The following day, four flasks each containing 250 mL LB were inoculated with 3 mL of the overnight culture. All four 250 mL cultures were incubated at 37 ˚C, 200 rpm until the culture reached an OD_{600} of ~0.4 – 0.6.

Once the cultures reached the appropriate OD_{600}, the four flasks were incubated on ice for 30 minutes with occasional swirling. Bacteria were kept on ice or at 4 ˚C from this point onwards and resuspension steps undertaken in the laminar flow hood. Bacteria were harvested by centrifugation for 15 minutes at 4,200 rpm (3,234 rcf). The LB supernatant was discarded then the bacteria washed twice in ice-cold ddH_2O and once in ice-cold 10% (v/v) glycerol, each time harvested by centrifugation as above.

The cells were then resuspended in ice-cold 10% glycerol to a concentration of $1x10^{10}$ cells/mL or higher (an OD_{600} of 1 or above for a 1:100 dilution of the cells in 10% glycerol). 40 µL of the highly concentrated cells were tested for conductivity by transfer into a 1 mm cuvette and application of a 1.8 kV pulse (200 Ω and 25 µF) using the BTX ECM 630. If the aliquot arced, then cells underwent another resuspension step in 10% glycerol. If the aliquot did not arc, the remaining culture was transferred into 1.5 mL Eppendorf tubes in 40 µL aliquots and immediately frozen in liquid nitrogen. Electrocompetent cells were stored at -70 ˚C.

Where cells were resuspended in GYT medium instead of 10% glycerol in the final resuspension step, GYT medium contained: 10% (v/v) glycerol, 0.125% (w/v) yeast extract and 0.25% (w/v) tryptone.

To calculate the transformation efficiency of the electrocompetent cells, 2 µL commercial pUC19 plasmid (50 pg/µL, NEB) was mixed with one 40 µL aliquot of cells that had been defrosted on ice. The cells were transferred into a 1 mm cuvette and a 1.8 kV pulse applied as above, then 960 µL SOC was added immediately and the cells incubated at 37 ˚C, 200 rpm for 45 minutes. A dilution series was carried out, plated on LB-agar with ampicillin in triplicate and incubated at 37 ˚C overnight. The following day, colony forming units were counted and transformation efficiency (transformants/µg DNA) calculated.

2.1.5 Growth of bacterial strains in M9 glucose

A single colony on a selective LB-agar plate was picked to inoculate 5 mL LB medium containing the appropriate antibiotic. The culture was incubated at 37 ˚C, 200 rpm for approximately 16 hours. Bacteria were then harvested by centrifugation for 15 minutes at 4,200 rpm (3,234 rcf) and resuspended in 5 mL M9 glucose (without an antibiotic). This centrifugation and M9 glucose re-suspension step was repeated twice more, then the OD_{600} of the bacterial culture was measured and fresh M9 glucose medium was inoculated to an OD_{600} of 0.1 and final volume of 25 mL with the appropriate antibiotic. The 25 mL cultures were incubated in the Chi.Bio at 37 ˚C with constant stirring (see section 2.1.7). The growth rate in the logarithmic phase of growth was calculated as the time taken for the culture to reach OD_{600} = 0.8 from OD_{600} = 0.4 (Kobayashi *et al.*, 2006).

2.1.6 Glycerol stocks

Glycerol stocks were prepared by adding 600 µL overnight bacterial culture to 200 µL sterile 100% (v/v) glycerol and stored at -70 ˚C.

2.1.7 Set up of the Chi.Bio reactors for bacterial culture

The control computer, pump boards and reactors were purchased from LabMaker. 30 mL bottles and 12 mm magnetic stir bars were used for bacterial culture (both Fisher Scientific) with 4.5mm/2.5mm outer/inner diameter laboratory silicone tubing for media in/out (Altec). The Chi.Bios were set up according to the manuals for software download and operation (Steel, 2023). Briefly, the Chi.Bio operating system was downloaded to an SD card and then copied onto the control computer. Lengths of silicone tubing were fed through pumps 1 and 2 (separately) on the pump board so that media could be fed in and out of the reactor. The reactors were each connected to the control computer, a pump board and a power supply and raised to a height that would remain above the fresh/waste media liquid level throughout the experiments.

Media bottles, magnetic stirrers and media input/output tubes were sterilized with 70% ethanol. Tubes were sterilized by pumping 70% ethanol through for ~10-20 seconds, then the media input tubes were flushed through with fresh media containing the antibiotic (usually M9 glucose and gentamycin). The lid of the reactor was wiped with 70% ethanol and the hole that was not in use for tubing was closed with micropore tape.

A 25 mL volume of fresh media was used to calibrate each reactor before an experiment began: the OD absorbance was measured and then this value used to set the OD to 0. Bacterial cultures were then placed in the reactors and connected to the media input/output tubes. The media input tubes were connected to a fresh media supply (containing the antibiotic) and fed in to the bottle in the bioreactor *via* the lid. The waste media tube was also fed in to the bottle in the bioreactor through the lid and was connected to a waste media bucket. Once everything was in place,

the reactors were started which incubates the culture at 37 ˚C with constant stirring (setting 0.5 on the user interface). Following reactor set up, OD regulation was switched on if required through the input of an OD set point to the user interface.

2.2 Molecular cloning

Unless otherwise specified, all restriction endonucleases and reaction buffers were obtained from New England Biolabs (NEB).

2.2.1 Primer design

Primers for cloning, Sanger sequencing and colony PCR were designed according to the following specifications: T_m 57-63 ˚C, length 18-24 nucleotides and GC content 40-60%. Primer sequences were also checked for likelihood of secondary structure formation (hairpin, self- and hetero-dimerisation) using the OligoAnalyzer™ Tool (IDT, https://eu.idtdna.com/pages/tools/oligoanalyzer). A full list of primer sequences used in this research can be found in Appendix 3.

2.2.2 Golden Gate cloning

2.2.2.1 Cloning PEPC expression vectors

A typical level 1 vector cloning reaction consisted of the following: 20 fmol plasmid backbone (pOGG024), 40 fmol other parts (P, U, SC and T modules, sequences can be found in Appendix 1), 1 µL T4 DNA ligase (400 U/µL), 2 µL T4 DNA ligase buffer (10X) and 1 µL BsaI-HF (20 U/µL) made up to a 20 µL final reaction volume with ddH_2O. The reaction was incubated for 30 cycles of 37 ˚C for 5 minutes, then 16 ˚C for 5 minutes, followed by 37 ˚C for 10 minutes and 80 ˚C for 20 minutes.

2.2.2.2 Variations of cloning the library of *6His::PEPC* variants into the placeholder vector

Five different Golden Gate reaction compositions described in Chapter 4 (Table 4.5) were tested and transformed into *E. coli* Δppc. The reactions were incubated for 25 cycles of 37 ˚C for 3 minutes then 16 ˚C for 4 minutes, followed by 37 ˚C for 5 minutes and 65 ˚C for 20 minutes.

2.2.2.3 Cloning the library of *6His::PEPC* variants into the placeholder vector

Cloning reactions of *6His::PEPC* variants into the placeholder vector consisted of the following: 50 fmol plasmid backbone (placeholder vector), 50 fmol insert (PEPC variants: Darwin Assembly mutagenesis outnest PCR product), 1.3 µL T4 DNA ligase (400 U/µL), 2 µL T4 DNA ligase buffer (10X), 2 µL bovine serum albumin (BSA) (1 mg/mL) and 1.3 µL BpiI (BbsI) (ThermoFisher Scientific) (10 U/µL) made up to a 20 µL final reaction volume with ddH_2O. The reaction was incubated for 65 cycles of 37 ˚C for 3 minutes then 16 ˚C for 4 minutes, followed by 37 ˚C for 5 minutes and 65 ˚C for 20 minutes.

2.2.3 Colony PCR reactions

Colonies were picked and separately resuspended in 10 µL ddH_2O. This solution comprised the PCR template. Where plasmid DNA was used as a control, < 0.5 µL plasmid DNA was resuspended in 10 µL ddH_2O, and ddH_2O was used as the template for a negative control PCR reaction. PCR reactions were prepared as follows: 4 µL Green GoTaq Flexi Buffer (5X) (Promega), 1 µL $MgCl_2$ (25 mM), 1 µL dNTPs (10 mM), 1 µL forward primer (10 µM), 1 µL reverse primer (10 µM), 0.2 µL GoTaq G2 Flexi DNA polymerase (5 U/ µL) (Promega), 4 µL template and 7.8 ddH_2O for a final reaction volume of 20 µL. Cycling parameters were 95 ˚C for 3 minutes, then 30 cycles of 95 ˚C for 1 minute, X ˚C for 1 minute and 72 ˚C for 3

minutes, then a final 72 ˚C extension step for 10 minutes. X was calculated based on the primer T_m, typically 55 - 65 ˚C.

2.2.4 Gel electrophoresis

DNA fragments were visualised by gel electrophoresis. Heat was used to dissolve 1 % (w/v) agarose in TAE buffer (5 mM NaOH, 40 mM Tris, 20 mM acetic acid, 1. mM EDTA). Subsequently, SYBR™ Safe DNA Gel Stain (ThermoFisher Scientific) was added at a 1:10,000 dilution volume. DNA samples were mixed 5:1 with 6X TriTack DNA Loading Dye (ThermoFisher Scientific) before loading on the gel, and GeneRuler 1 kb DNA Ladder (ThermoFisher Scientific) was loaded alongside the DNA samples. Samples were run on the gel at 100 V until separated, typically 30 – 60 minutes. Nucleic acid gels were imaged and analysed using an Invitrogen iBright CL1500.

2.2.5 Plasmid preparation

A single colony on a selective LB-agar plate was picked to inoculate 5 mL LB containing the appropriate antibiotic. The culture was incubated at 37 ˚C, 200 rpm for approximately 16 hours overnight. Plasmid DNA was subsequently prepared using the QIAprep Spin Miniprep Kit (Quiagen).

2.2.6 Sanger sequencing

Sanger sequencing was undertaken by Source BioScience (UK) to verify the correct assembly of DNA constructs.

2.3 Generation of *6His::PEPC* variants by Darwin Assembly

The protocol is based on the Darwin Assembly publication (Cozens and Pinheiro, 2018) and details outlined here are additional to those described in Chapter 3.

Setaria viridis C_4 PEPC Sevir.4G143500.1 (SvPEPC) and 6His::PEPC were aligned with *Z. mays* C_4 PEPC CAA33317.1 to map the sites under selection in C_4 PEPC (described in Christin *et al.*, 2007) onto SvPEPC and 6His::PEPC.

Phusion High-Fidelity DNA Polymerase was used as the Darwin Assembly polymerase so the T_m for Darwin Assembly primers were calculated using the NEB T_m Calculator (https://tmcalculator.neb.com/#!/main).

2.3.1 ssDNA generation

Single-stranded DNA (ssDNA) generation is described in detail in Chapter 4. The *6His::PEPC* sequence can be found in Appendix 1. DNA was made single stranded by incubating 1 µg 5.8 kb 6His::PEPC in pUC57 with 10 U/µg Nt.BbvCI and 20 U/µg ExoIII in NEBuffer3.1 at 37 °C for 2 hours then 80 °C for 20 minutes in a 10 µL total reaction volume.

2.3.2 Primer phosphorylation

A list of the mutagenic and boundary primer sequences can be found in Appendix 3. Phosphorylation of each of the mutagenic primer libraries was as follows:

Reagent	Stock concentration	Reaction concentration	Volume
Primer solution*		17 µM	6.46 µL
CutSmart buffer	10X	1X	2 µL
ATP	10 mM	1 mM	2 µL
T4 PNK	10 U/µL		0.4 µL
ddH_2O			9.14 µL

*Primer solutions were set up as follows: 2.04 µL NDT (100 µM), 1.02 µL VMA (100 µM) and 1.7 µL both ATG and TGG (diluted to 10 µM), total volume 6.46 µL.

The 3’ boundary primer was phosphorylated as per the publication: 1 µL 100 µM primer, 1 µL CutSmart buffer (10X), 1 µL ATP (10 mM), 0.2 µL T4 PNK and 6.8 µL ddH_2O.

The reactions were incubated at 37 °C for 2 hours then 80 °C for 20 minutes.

2.3.3 Isothermal assembly

0.15 pmol ssDNA generated as described in 2.3.1 was mixed directly (without purification) with 1.5 pmol 3’ and 5’ boundary oligonucleotides and 30 pmol inner oligonucleotides. The phosphorylated 3’ oligonucleotide was diluted 1:2 to 5 µM and the biotinylated 5’ oligonucleotide was diluted from 100 µM to 5 µM, then 0.3 µL of each boundary oligonucleotide was added to 5 µL (0.15 pmol) ssDNA. 1.76 µL phosphorylated oligonucleotide mixture (17 µM) was also added to the ssDNA. The DNA was mixed well then frozen at -20 °C for 60 minutes before a further incubation at room temperature for 15 minutes. 7.36 µL Darwin Assembly mix (2X) was added to the ssDNA/oligonucleotide mixture and the reaction mixture was incubated at 50 °C for 50 minutes. Darwin Assembly mix (2X) is: 0.05 U/µL Phusion High-Fidelity DNA Polymerase, 8 U/µL Taq DNA ligase, 2 mM NAD+, 0.4 mM each dNTP, 10% (w/v) PEG 8000, 2 mM DTT and 1× CutSmart buffer.

2.3.4 Assembly purification

5 µL Dynabeads (ThermoFisher Scientific) were blocked in 300 µL 2X BWBS-T (20 mM Tris–HCl pH 7.4, 2 M NaCl, 0.2% (v/v) Tween-20, 2 mM EDTA) for 75 minutes on a rotator for continuous mixing. Beads were then isolated on a magnetic stand and then resuspended in 50 µL 2X BWBS-T.

The isothermal assembly reaction volume was adjusted to 50 µL with ddH_2O and added to the blocked Dynabeads. The bead and assembly mixture was incubated

at room temperature for 3 hours with occasional gentle mixing. When the 3-hour incubation period was complete, Dynabeads were isolated as above and then resuspended twice in 200 µL 37°C 30 mM NaOH, once in 200 µL EB-T (10 mM Tris–HCl pH 8.8, 0.1 mM EDTA, 0.01% Tween-20) and finally in 10 µL EB (10 mM Tris–HCl pH 8.8).

2.3.5 Amplification of *6His::PEPC* variant libraries (outnest PCR)

Initial amplification reactions were undertaken on both the Dynabeads and 1:10 diluted Dynabeads (in EB) on a small scale (25 µL). PCR conditions were as follows: 5 µL Phusion HF buffer (10X), 0.5 µL dNTPs (10 mM), 1.25 µL forward outnest primer, 1.25 µL reverse outnest primer, 0.25 µL Phusion High-Fidelity Polymerase, 1 µL template (assembly purified using Dynabeads) and 15.75 µL ddH_2O. Primer sequences are available in Appendix 3. Cycling parameters were 98 ˚C for 30 seconds, then 25 cycles of 98 ˚C for 10 seconds, 60 ˚C for 30 seconds and 72 ˚C for 1 minute, then a final 72 ˚C extension step for 10 minutes.

If the small-scale outnest PCR reaction was successful, it was scaled up by 3 – 5 X and subject to the same cycling parameters. 10 µL PCR reaction product was sent for PCR cleanup and Sanger sequencing (as per 2.2.6) and the remaining reaction product loaded across several wells on a 1% agarose gel according to 2.2.4. Once visualised to confirm presence of a PCR product of the correct size, the bands were excised from the agarose gel, subject to gel extraction using the Monarch DNA Gel Extraction Kit and eluted in 15 µL 60 ˚C EB (supplied).

2.4 Analysis of 6His::PEPC expression level in bacteria by immunoblot

A single colony was picked for untransformed *E. coli* Δppc and BW25113 and each strain of *E. coli* Δppc transformed with three different 6His::PEPC expression constructs described in Chapter 3, and cultured first in LB then M9 glucose as per 2.1.5. The culture was held at an OD_{600} of 0.63 in the Chi.Bio overnight, then 0.5 mL culture was centrifuged at 13,300 rpm (17,000 rcf) for 2 minutes to pellet the bacteria and 450 µL supernatant was discarded. Multiple samples were taken per strain and stored at -70 ˚C.

Samples were thawed, then 50 µL Sample Buffer, Laemmli (2X) (Sigma-Aldrich) was added and the samples were boiled at 98 ˚C for 10 minutes. The samples were centrifuged for 5 minutes at 13,300 rpm (17,000 rcf), then 16 µL each sample was loaded in duplicate onto an 12% Mini-PROTEAN® TGX™ Precast Protein Gel (Bio-Rad) with 5 µL PageRuler™ Prestained Protein Ladder (ThermoFisher Scientific). The samples were run in running buffer (25 mM Tris base, 190 mM glycine, 0.1% SDS) at 50 V, 400 mA for 5 minutes then the voltage increased to 100 V for 1 hour.

One protein gel was incubated overnight in QuickBlue Protein Stain (LubioScience) and de-stained in water the next day. The proteins on the second gel were transferred onto a PVDF membrane in the Mini Transfer Pack (Bio-Rad) using the Bio-Rad Turbo Transfer System (according to the Standard SD protocol: 25 V, 30 minutes, up to 1.0 A). The membrane was blocked in blocking buffer (TBS-T: 10 mM Tris, pH 8, 150 mM NaCl and 0.05% Tween-20, with the addition of 4% (w/v) skimmed milk powder) with gentle shaking overnight at 4 ˚C. The membrane was then washed in TBS-T three times by gentle shaking for 5 minutes at room

temperature. Then, the membrane was incubated in a Monoclonal Anti-polyHistidine antibody (Sigma H1029, 1:3000 dilution in blocking buffer) for 3 hours with gentle shaking at room temperature. Following this incubation, the membrane was washed three times as above in TBS-T. The membrane was incubated in a secondary antibody solution (7.5 µL Sigma Anti-Mouse AP conjugate (A3562) in 25 mL blocking buffer) with gentle shaking for 90 minutes at room temperature.

To visualise the secondary antibody, the membrane was washed in TBS-T as above, then incubated for five minutes in 15 mL detection buffer (1.21 g Tris-HCl, 0.58 g NaCl, pH 9.5, made up to 100 mL) at room temperature. Subsequently, the membrane was placed on the Invitrogen iBright CL1500 imager and a detection solution (1 mL detection buffer and 6 µL CDP-Star™ Substrate, T2304) was added evenly to the top of the membrane. The membrane was covered in a plastic sheet, chemiluminescence imaged and band intensity quantified using the Invitrogen iBright CL1500.

2.5 Screens of 6His::PEPC variants in *E. coli*

2.5.1 Modelling the composition of *E. coli* Δppc strains in bacterial co-culture

2.5.1.1 Modelling the co-culture of three bacterial strains with different hypothetical rates of growth

The population composition for a co-culture containing bacterial strains with fastest, medium and slowest relative rates of growth was modelled using equations $y=1.1^x$, $y=1.05^x$ and $y=1.025^x$, respectively (Figure 3.11 A). Here, x represents the passing of time in arbitrary units. This simulates the equal inoculation of each strain in the initial culture (when x=0), then describe how the strains grow exponentially over time at different rates. The fastest growing strain doubles at twice the rate of the

medium growing strain and four-times the rate of the slowest growing strain. The abundance of each strain was then calculated as a proportion of the total bacterial abundance in the culture over time. The process was repeated for Figure 3.11 B, however, the bacterial strain with the fastest rate of growth was modelled by equation $y=1.2^x$, while medium and slowest growing strains were modelled by equations $y=1.05^x$ and $y=1.025^x$, respectively, as in 3.11 A.

2.5.1.2 Modelling the co-culture of three bacterial strains using experimental growth rate data

The equation $y = 0.5e^{(0.6931x)}$ describes an exponential growth curve that simulates bacterial growth with a fixed (arbitrary) population doubling time (1, 2, 4, 8…). This equation can be used to model the relative abundance of the high, medium and low 6His::PEPC expression strains using the average population doubling time for each strain (114, 136 and 238 minutes, respectively) (Figure 3.12). The model reflects equal inoculation of each strain and the exponential increase in abundance of each strain over 24-h by the following equation, $y = 0.5e^{(0.6931(TE/PD))}$. Time elapsed since inoculation (TE) and population doubling time (PD) are input into the equation in minutes. The abundance of each strain was then calculated as a proportion of the total bacterial abundance in the culture over time.

2.5.2 A screen of *E. coli* Δppc strains with different levels of 6His::PEPC expression

Three strains of *E. coli* Δppc transformed with high, medium and low strength of expression vectors (described in Chapter 3) were inoculated first in LB and then M9 glucose as per 2.1.5. However, although the strains were each washed individually in M9 glucose, a single culture was then made whereby M9 glucose was inoculated with all three strains simultaneously, each to an individual OD_{600} of 0.05 in a total

volume of 80 mL including gentamycin. This culture was then divided into 20 mL cultures between three sterile bottles for incubation in the Chi.Bios, and plasmids were extracted from the remaining 20 mL volume as per 2.2.5. The Chi.Bios were set up as per 2.1.7 to hold the cultures at a constant OD_{600} of 0.63 with fresh M9 glucose and gentamycin medium. 5 mL samples were taken after 24-h and miniprepped as per 2.2.5.

The PCR to test the primers that amplify one of the low, medium and high for exclusivity was undertaken using GoTaq as per 2.2.3 but 0.1 µL GoTaq Flexi DNA polymerase, 1 µL template DNA (1 ng/µL) and 10.9 µL ddH_2O. Cycling parameters were 94 ˚C for 5 minutes, then 15 cycles of 94 ˚C for 1 minute, 55 ˚C for 1 minute and 72 ˚C for 1 minute, then a final 72 ˚C extension step for 5 minutes. The PCR to test amplification of different quantities of DNA template was the same as above but 1 µL template DNA contained plasmid DNA at a concentration of 0.01, 0.1 or 1 ng/µL. Cycling parameters as previous. The PCR to identify change in abundance over 24-h was undertaken as above but 1 µL template DNA contained DNA at 0.5 ng/µL. Cycling parameters as previous.

2.5.3 A screen of 6His::PEPC variants in *E. coli* continuous co-culture

A Golden Gate reaction was undertaken between *6His::PEPC* variants and the placeholder vector as per 2.2.2.3 for every variant library generated by Darwin Assembly. The outnest PCR for the variant library at site 780 was gel-extracted and then used in the Golden Gate reaction straight away, whereas all 20 other libraries were gel-extracted and then subject to DNA cleanup using the Monarch® PCR & DNA Cleanup Kit.

Electrocompetent *E. coli* Δppc were prepared, transformed with Golden Gate reaction product and recovered in SOC as per 2.1.4. For the screen of 6His::PEPC variants mutated at site 780 only (the single library screen), 1 µL Golden Gate reaction product was electroporated into *E. coli* Δppc 5 times independently, then all the SOC cultures were combined after the outgrowth step into a 250 mL LB culture with gentamycin and cultured at 37 ˚C, 200 rpm overnight. For the larger screen of all possible 6His::PEPC variants, one electroporation was performed per site library (21 total) using 2 µL Golden Gate reaction product then all the SOC cultures were combined after the outgrowth step as above.

After overnight growth in LB, 50 mL bacteria were pelleted and washed in M9 glucose as per 2.1.5. For the single library screen, a 300 mL M9 glucose (T_0) culture was inoculated to an OD_{600} of 0.1 with gentamycin, divided between 3 sterile bottles in volumes of 25 mL and plasmids extracted from the remaining T_0 culture. For the screen of all possible variants, a 400 mL M9 glucose (T_0) culture was inoculated to an OD_{600} of 0.1 with gentamycin, divided between 4 sterile bottles in volumes of 25 mL and plasmids extracted from the remaining T_0 culture. The Chi.Bios held the cultures at an OD_{600} of 0.43 over the course of 68 hours (single library screen) or 72 hours (screen of all possible variants) (for Chi.Bio set up, see 2.1.7). 5 mL samples were harvested periodically from each replicate then the bacteria pelleted, supernatant removed and samples stored at -70 ˚C. Plasmids were extracted from each sample using the QIAprep Spin Miniprep Kit (Quiagen).

2.5.4 Analysis of 6His::PEPC abundance in continuous co-culture: semi-quantitative PCR and next-generation sequencing (NGS)

Primers for NGS were ordered from Eurogentec at RP-Cartridge purification specification, <0.1% cross-contamination and QC with MALDI-TOF MS. Primer

melting temperatures were 65-70 ˚C (calculated using the NEB T_m Calculator, https://tmcalculator.neb.com/#!/main), length 18-24 nucleotides and GC content 40-60%. Primer sequences were also checked for likelihood of secondary structure formation (specifically hairpin, self- and hetero-dimerisation) using OligoAnalyzer™ Tool (IDT, https://eu.idtdna.com/pages/tools/oligoanalyzer). A list of the primer sequences can be found in Appendix 3.

Unique sets of primers were used to amplify four different fragments in the *6His::PEPC* gene by PCR. Together, the forward and reverse primer barcodes encode information about time point and replicate number for a given sample that is subject to amplification. PCR reactions were prepared as follows: 10 µL Phusion HF buffer (5X), 1 µL dNTPs (10 mM), 2.5 µL forward primer (10 µM), 2.5 µL reverse primer (10 µM), 0.5 µL Phusion High-Fidelity DNA Polymerase and 10 ng template DNA made up to a 50 µL reaction using ddH_2O. Cycling parameters were 98 ˚C for 30 seconds, then 20 cycles of 98 ˚C for 10 seconds, 62-63 ˚C for 30 seconds and 72 ˚C for 1 minute, then a final extension step at 72 ˚C for 10 minutes. The annealing temperature depended on the fragment: fragment 1-3 primers were annealed at 63 ˚C and fragment 4 primers were annealed at 62 ˚C (fragments described in detail in Chapter 5). The PCR reaction products were cleaned up using the Monarch® PCR & DNA Cleanup Kit, combined into one sample at approximately equal concentrations according to the required specifications for Microbial whole genome library (PCR-free 350 bp) sequencing: >1 µg total DNA, > 20 µL sample volume and > 10 ng/µL sample concentration. Sequencing was undertaken by Novogene.

2.5.5 Enzyme kinetics characterisation of P531A, F794V and N863K SvPEPC variants

Enzyme kinetics measurements were undertaken according to methods published in detail in DiMario *et al*., 2019 and DiMario *et al*., 2021. In brief, wild-type SvPEPC and the three SvPEPC variants (without a His-tag) were each expressed in and extracted from the PEPC-deficient *E. coli* strain *PCR1*, purified by hydrophobic interaction chromatography, desalted and stored in 20% (v/v) glycerol at –80 °C.

Membrane inlet mass spectrometry was used to quantify PEPC activity at 7 different $NaHCO_3^-$ concentrations (50, 100, 200, 350, 500, 750, and 1000 µM) assayed independently at 25 °C in a 600 µL cuvette. The assay buffer contained 100 mM HEPES-KOH (pH 7.6); 10 mM $MgCl_2$, 1 mM DTT, 50 µg ml^{-1} carbonic anhydrase, 5 mM glucose 6-phosphate and 10–15 µg PEPC extract protein. 5 mM PEP was added to initiate the reaction and after a 20 second calibration period a five-second running average of the change in mass 44 was used to calculate PEPC activity at an averaged HCO_3^- concentration.

PEPC activity in response to [PEP] was measured spectroscopically at 340 nm by coupling to malate dehydrogenase. The assay buffer contained 100 mM HEPES-KOH (pH 7.6), 10 mM $MgCl_2$, 1 mM DTT, 0.2 mM NADH, 10 mM $NaHCO_3$, 5 mM glucose 6-phosphate, 7.2 U ml^{-1} malate dehydrogenase and 10–15 µg PEPC extract protein. PEP was added to initiate the reaction. Enzyme activity was measured at 10 PEP concentrations (20, 50, 100, 200, 350, 500, 1000, 2000, 3500 and 5000 µM) at 25 °C in a 600 µL cuvette. Data points for each PEP concentration were taken during the linear part of the reaction over approximately 20-second intervals.

Kinetic parameters V_{max}, $K_mHCO_3^-$ and K_mPEP were evaluated using the Excel Solver function by fitting the Hill equation to the response of PEPC activity to changing [HCO_3^-] or [PEP] as described in DiMario *et al*., 2019 and DiMario *et al*., 2021. K_{cat} was calculated using V_{max} obtained from the MIMS measurements (HCO_3^- response curve) and estimated PEPC concentration. Specificity for bicarbonate was calculated for each SvPEPC variant by dividing K_{cat} by $K_mHCO_3^-$.

2.5.6 Modelling the impact of P531A, F794V and N863K SvPEPC variants on rate of photosynthetic carbon assimilation in *Setaria viridis*

The von Caemmerer (2021) model of C_4 photosynthesis was used to predict the photosynthetic carbon assimilation rate of *S. viridis* (Figure 5.20) under different intracellular CO_2 concentrations. The published dataset was downloaded and model values V_{pmax} and K_p (V_{max} of PEPC and affinity of PEPC for bicarbonate, respectively), in the 'A vs cm' tab, were adjusted relative to the differences between K_{cat} and $K_mHCO_3^-$ values measured for each SvPEPC variant compared to SvPEPC, respectively (Figure 5.18). In the model, V_{pmax} is 200 µmol m^{-2} s^{-1} and Kp is 82 µbar CO_2. For example, estimated V_{pmax} for P531A variant is 200*(212/89) = 476 µmol m^{-2} s^{-1}. An equivalent enzyme concentration is therefore assumed for V_{pmax}. All other input values in the model remained identical.

	V_{pmax}		Kp	
	K_{cat} (s^{-1}) (measured)	V_{pmax} (µmol m^{-2} s^{-1}) (estimated)	$K_mHCO_3^-$ (µM) (measured)	Kp (µbar) (estimated)
SvPEPC	89	N/A	28	N/A
P531A variant	212	476	58	170
F794V variant	73	164	60	176
N863K variant	96	216	81	237

Chapter 3: A novel bacterial screen for improved PEPC variants

3.1 Introduction

3.1.1 Engineering PEPC to improve food security

Maize and sugarcane, which both undertake C_4 photosynthesis, are major food and biofuel crops (Byrt, Grof and Furbank, 2011; van der Weijde *et al.*, 2013; Wang, Chan and Long, 2021). Maize, sugarcane and other C_4 crops such as sorghum and millet comprise over 25% of the land used by crops and sugarcane and maize were the two most highly produced crops in 2020 with 1,870 million tonnes and 1,162 million tonnes harvested globally, respectively (FAO, 2022). The next two most widely-produced crops, the C_3 crops wheat and rice, were produced on a smaller scale of 761 million tonnes and 757 million tonnes, respectively (FAO, 2022). Thus, C_4 crops are among the most important crops for food, feed and fuel production and occupy a substantial component of the land used for agriculture.

In C_4 photosynthesis, carbon from the atmosphere is first fixed into oxaloacetate by PEPC (O'Leary, 1982; Hatch, 1987; Lepiniec *et al.*, 1994; Rajagopalan, Devi and Raghavendra, 1994). Oxaloacetate will rapidly and spontaneously decarboxylate to pyruvate under physiological conditions (Tsai, 1967; Wolfenden, Lewis and Yuan, 2011), and thus in C_4 photosynthesis oxaloacetate is stabilised by enzymatic conversion to either aspartate or malate (or both) (Ting and Osmond, 1973; Furbank, 2011; Rao and Dixon, 2016). Once stabilised, aspartate or malate (or both) are transported to the site where Rubisco is located before being decarboxylated to release CO_2 (Rao and Dixon, 2016). This decarboxylation creates a supra-atmospheric concentration of CO_2 around Rubisco, reducing the occurrence of

Rubisco oxygenation and the concomitant loss of energy *via* photorespiration (Sage, Sage and Kocacinar, 2012; Bräutigam and Gowik, 2016).

Although C_4 plants are extremely important to humanity as domesticated crops, efforts to increase crop productivity to date have focused primarily on making improvements to C_3 photosynthesis or engineering C_4 traits into C_3 plants (Ku *et al.*, 1999; Parry *et al.*, 2013; Long, Marshall-Colon and Zhu, 2015; Ort *et al.*, 2015; Kromdijk *et al.*, 2016; Ermakova *et al.*, 2020; Burgess *et al.*, 2022; Sakoda *et al.*, 2022). However, there is also an urgent need to improve C_4 photosynthesis in order to meet crop improvement targets to feed a population of an estimated 10 billion people by 2050. A number of targets for improving C_4 photosynthesis have been identified (von Caemmerer and Furbank, 2016; Wang, Chan and Long, 2021). These include engineering improved variants of the C_4 photosynthesis enzymes PEPC and pyruvate orthophosphate dikinase (PPDK), as well as engineering improvements to Rubisco and its regulatory proteins, and enhancing the flow of electrons through the chloroplast (von Caemmerer and Furbank, 2016; Wang, Chan and Long, 2021).

In normal conditions, the operation of the C_4 cycle means that Rubisco-mediated CO_2 fixation is not limited by CO_2 supply. However, the rate of Rubisco-mediated carbon fixation in C_4 plants is limited by the enzymatic rate of PEPC under conditions that result in a low CO_2 concentration in the mesophyll cell (Kiirats *et al.*, 2002; von Caemmerer and Furbank, 2016; Zhou, Akçay and Helliker, 2019; von Caemmerer, 2021), for example if the plant is subject to drought stress (Pignon and Long, 2020) or during photosynthetic induction in fluctuating light conditions (Wang, Chan and Long, 2021). In support of this observation, it has also been noted that when plants have evolved C_4 photosynthesis, the kinetics of PEPC have also

evolved to favour a higher affinity and specificity for bicarbonate and lower affinity for PEP (Moody, Christin and Reid, 2020; DiMario *et al.*, 2021). Accordingly, one potential way in which it might be possible to increase the yield of C_4 plants could be through the engineering of improvements to PEPC such that the enzymatic rate is increased, defined as the moles of substrate converted per unit time *in vivo*, for example under low CO_2 availability (von Caemmerer and Furbank, 2016).

Maize, sugarcane and sorghum all predominantly do a subtype of C_4 photosynthesis called NADP-ME, named after the primary decarboxylating enzyme in the pathway (Wang, Chan and Long, 2021). *Setaria viridis* is a model organism that also undertakes the NADP-ME subtype of C_4 photosynthesis (Brutnell *et al.*, 2010; von Caemmerer, 2021). *S. viridis* has a small stature and short generation time that makes it practical to grow in limited space and with limited resources in a research setting, so protocols have been developed to genetically engineer *S. viridis* in the laboratory (Brutnell *et al.*, 2010). The work presented here therefore focuses on engineering SvPEPC so that enzyme variants that are identified as having potentially improved kinetics in this work can be subsequently validated *in planta* in the laboratory.

3.1.2 Using *E. coli* to screen 6His::PEPC variants in high-throughput

Despite C_4 photosynthesis being a highly efficient form of photosynthesis, it is likely that not all species have adapted the photosynthetic form of PEPC to the same extent and that not all possible variants of the enzyme have been explored in nature, given that C_4 photosynthesis is a relatively recent evolutionary trait (Niklaus and Kelly, 2019). Thus, there are likely to be mutations that could be made to the enzyme to enhance the enzymatic rate of PEPC which could potentially enhance yield under stress conditions or in fluctuating light when the enzymatic rate of PEPC

limits the rate of photosynthesis (Pignon and Long, 2020; Wang, Chan and Long, 2021). Protein engineering approaches such as random mutagenesis, rounds of directed evolution or rational mutagenesis can generate a large number of enzyme variants to be screened. However, screening large numbers of PEPC variants for improved kinetics *in planta* is impractical because of the time, resources and space required to conduct plant transformation and phenotyping. Selection in microbes can facilitate screening capacities of up to 10^{11} enzyme variants (Lassner and Bedbrook, 2001), therefore it is necessary to set up a screen for PEPC activity using bacteria in order to screen a large number of enzyme variants.

PEPC fulfils an anaplerotic function in *E. coli*, supplying carbon to the tricarboxylic acid cycle in the form of oxaloacetate (Ashworth and Kornberg, 1966) (Figure 3.1). PEPC activity is required for the growth of *E. coli* on glucose as the sole carbon source in order to maintain a functioning TCA cycle for two reasons. First, *E. coli* lacks pyruvate carboxylase which could otherwise supply oxaloacetate in the absence of PEPC (Ashworth and Kornberg, 1966; Gokarn *et al.*, 2001). Second, although there is an alternative anaplerotic pathway, the glyoxylate bypass (Figure 3.1) (Kornberg and Krebs, 1957), this pathway is shut down during growth of *E. coli* on glucose because of catabolite repression and high levels of isocitrate dehydrogenase activity which removes isocitrate as a substrate for the glyoxylate bypass reactions (Cozzone, 1998). Thus, *E. coli* Δppc that lack an endogenous PEPC cannot grow on glucose as the sole carbon source.

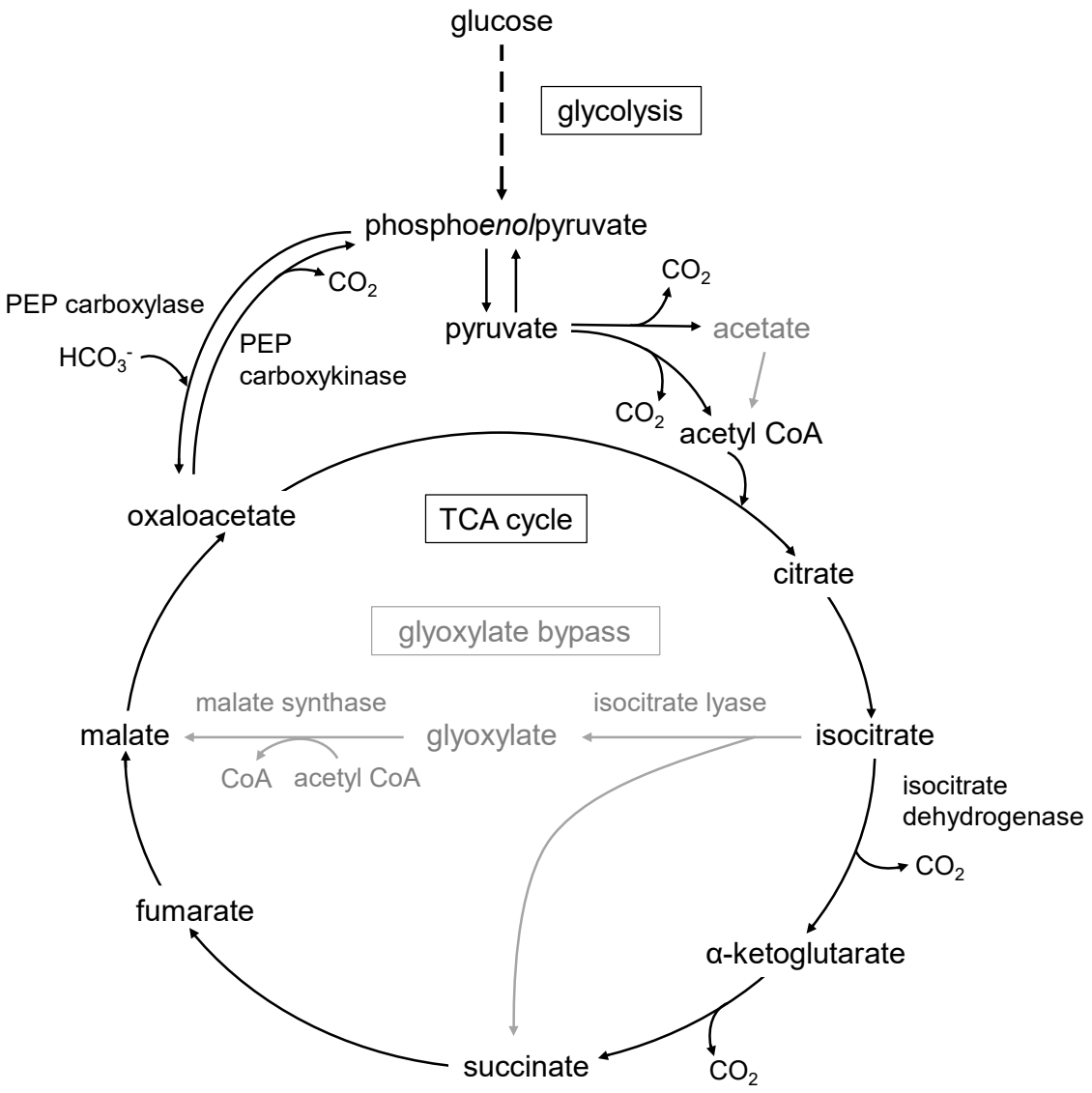


Figure 3.1: Glycolysis, the TCA cycle and the glyoxylate bypass are central to *E. coli* carbon metabolism. The key reactions in the TCA cycle (black arrows) and glyoxylate bypass (grey arrows) are illustrated.

It could be agronomically advantageous to develop a system by which it is possible to generate, screen and evaluate large numbers of PEPC variants prior to confirmation of function of one or few variants *in planta.* To achieve this, the dependence of *E. coli* on PEPC for growth on glucose can be exploited to develop high-throughput screens for increased enzymatic rate of PEPC. Here, we show that *E. coli* Δppc, a mutant strain that lacks the endogenous PEPC, cannot grow on minimal media when supplemented with glucose as a sole carbon source. However, growth can be restored when cells heterologously express *S. viridis* C_4 PEPC. In this chapter, I show that *E. coli* can be engineered such that bacterial growth rate can be made directly dependent on the expression level of *S. viridis* C_4 PEPC. In doing so, I establish an experimental system to enable high-throughput screening of hundreds of *S. viridis* PEPC variants to facilitate the discovery of variants with an increased enzymatic rate *in vivo*.

3.1.3 Chapter aims and outline

The aim of the work described in this chapter was to establish a high-throughput system in which an increased rate of catalysis by PEPC could be screened for using *E. coli* Δppc. This was achieved in three steps. First, the use of a new low-cost, small volume bioreactor was established in the laboratory for *E. coli* culture and continuous co-culture. Second, an expression vector suitable for a screen of PEPC variants in *E. coli* was identified. Third, a proof-of-concept screen was undertaken using three strains of *E. coli* Δppc that had different levels of *S. viridis* C_4 PEPC expression.

3.2 Results and Discussion

3.2.1 Installation and testing of Chi.Bio bioreactors in the laboratory

Chi.Bio bioreactors are an open-source system that facilitates inexpensive continuous co-culture of bacteria, with a system small enough to fit on a benchtop (Figure 3.2) (Steel *et al.*, 2020). The culture volume of 25 mL per bioreactor is maintained at 37 °C with constant stirring, and the machines have the capacity to continuously measure the culture optical density (OD) (the component parts that facilitate these features are illustrated in Figure 3.3). OD measurements are taken every minute by default which enables the continuous and accurate characterisation of bacterial growth in the bioreactors over time. In addition, the Chi.Bio bioreactors can use the OD measurements to control the rate at which cell culture volume is replaced with fresh medium and thus can also act as a turbidostat, where the culture is held at a constant OD over time.

For turbidostat function of the Chi.Bio bioreactors, when an OD measurement is taken the value becomes part of a feedback loop to a computer that compares it to a set point that the operator has chosen. If the OD value is above the set point, then a small volume of culture media is pumped out of the bioreactor and replaced by fresh media which is pumped in. This process will be repeated until the culture OD is returned to the set point. If the OD value is below the set point, then no media exchange will occur and thus the culture is left to grow until the set point is reached. In this way, it is possible to hold bacterial cultures in the logarithmic phase of growth indefinitely in these machines by connecting them to a fresh media supply and setting them up to hold the cultures at an OD between 0.4 and 0.8 (Kobayashi *et al.*, 2006).

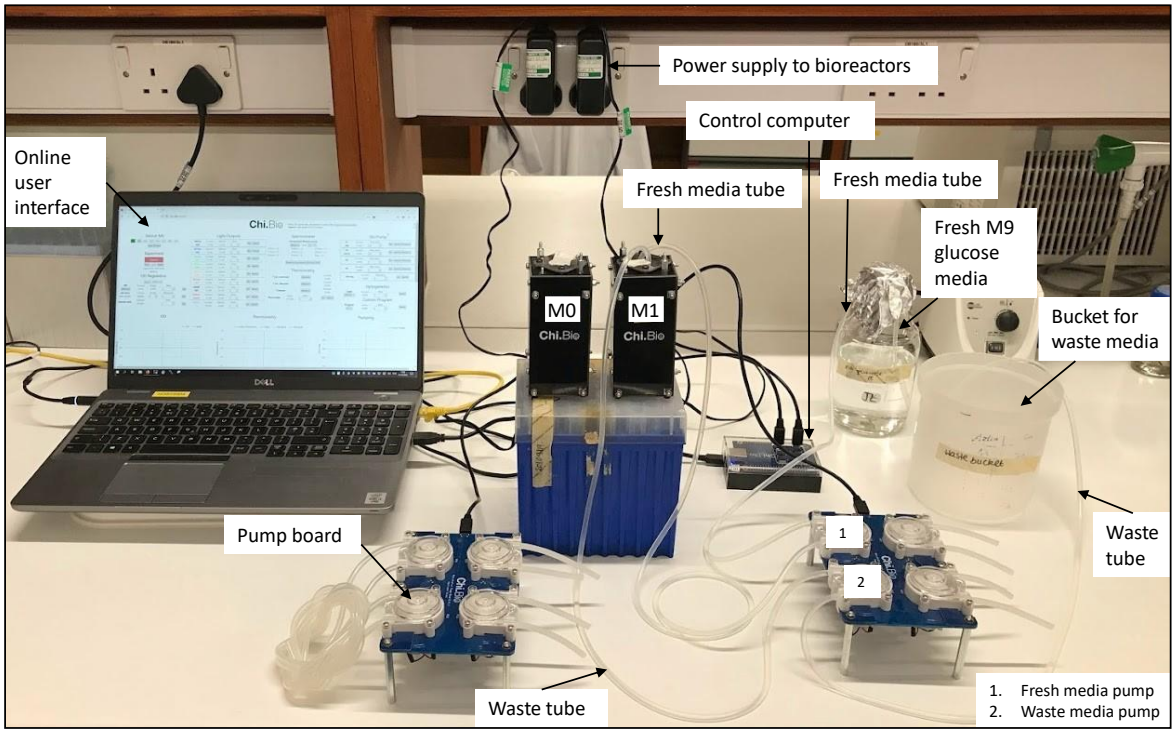


Figure 3.2: Two Chi.Bio bioreactors (M0 and M1) set up for bacterial culture. M0 is set up to facilitate unrestricted bacterial growth and M1 is set up to facilitate bacterial culture with turbidostat function.

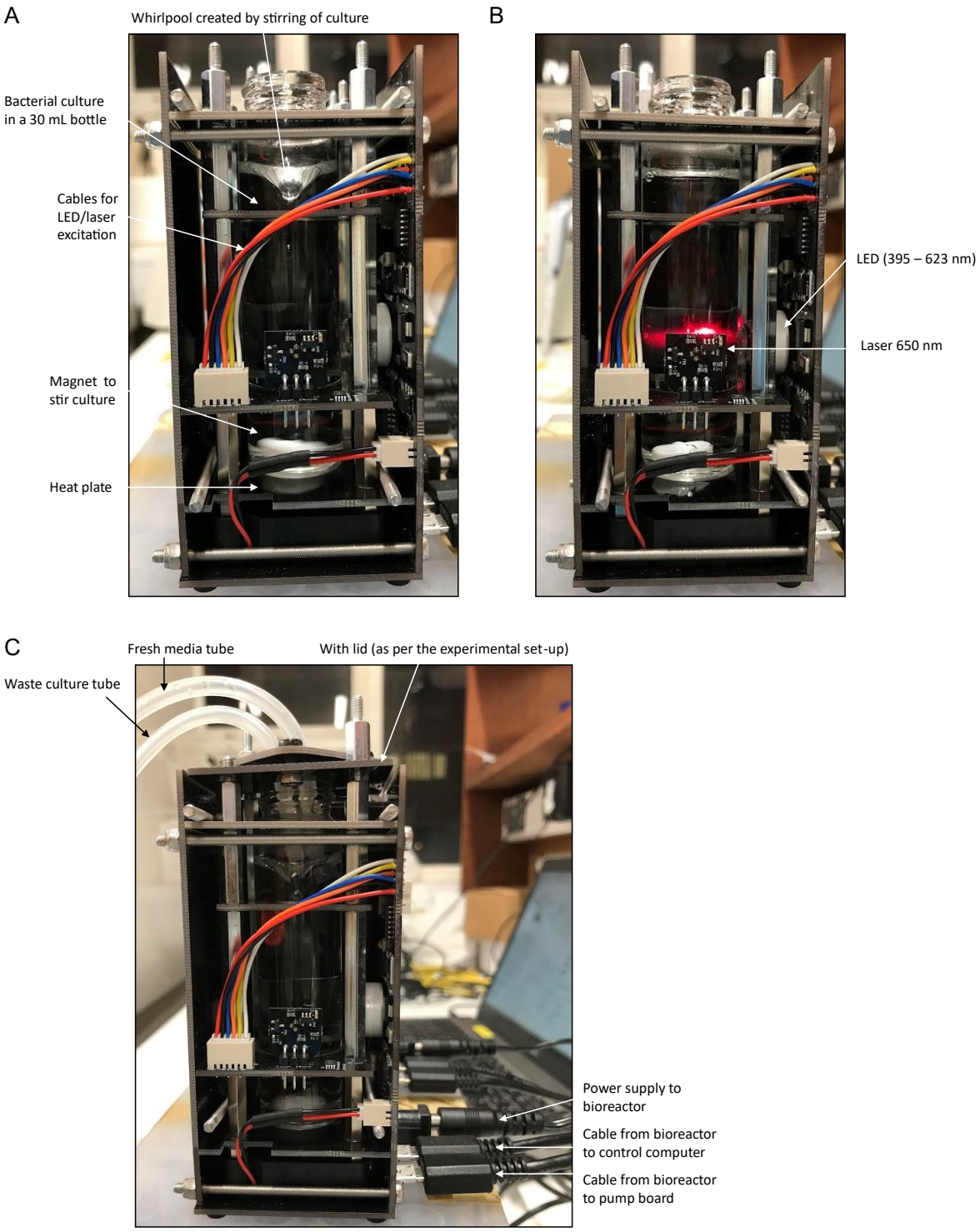


Figure 3.3: The features of a Chi.Bio bioreactor containing a bacterial culture that facilitate bacterial culture and OD measurement. Figure legend continued on the next page.

Figure 3.3 (continued) (A) A whirlpool is created when a white magnet stirs a 25 mL culture. The culture is sitting on top of a heat plate that is used to maintain the culture temperature at 37 °C. (B) An LED connected to multi-coloured cables enables illumination of the culture by a range of wavelengths between 395 and 623 nm, and a 650 nm laser works in conjunction with a spectrophotometer opposite (not labelled) to measure OD of the culture. (C) When the bioreactor is set up as a turbidostat, two tubes sit inside the neck of the culture bottle and the action of the pumps that these tubes are connected to (not shown here) provides fresh media or removes waste culture to hold the bacterial culture at a set OD indefinitely.

Test *E. coli* cultures were successfully grown in the Chi.Bio bioreactors once the bioreactors were set up in the laboratory. However, the OD measurement that each Chi.Bio machine displayed was approximately double the value expected. For example, the Chi.Bio bioreactors would consistently register the OD of the bacterial culture as ~0.2 at the start of each test run despite bacterial cultures being inoculated to an OD of 0.1 (as calculated using a benchtop spectrophotometer). Also, each culture would reach a maximum OD reported to be ~4 if left to grow unrestricted. It was therefore necessary to calibrate the Chi.Bio bioreactors to the benchtop spectrophotometer (Figure 3.4).

A simple linear model was fitted to determine the relationship between Chi.Bio OD and benchtop spectrophotometer OD. There was a strong linear relationship between the OD measured by the Chi.Bio and the OD measured by the benchtop spectrophotometer ($p<0.001$, $R^2 = 0.996$, $F(1,8) = 2281$). The regression model was:

$$\text{True OD} = 0.50523 \times (\text{Chi.Bio OD measurement}) - 0.07078 \quad \text{(Equation 2.1)}$$

The path length in the Chi.Bio spectrophotometer was approximately twice the path length of the benchtop spectrophotometer and Equation 2.1 appropriately corrected for this longer path length and any other differences in spectrophotometer configuration in the Chi.Bio. From this point onwards, all OD data from Chi.Bio bioreactors presented in the thesis has been subject to calibration by Equation 2.1.

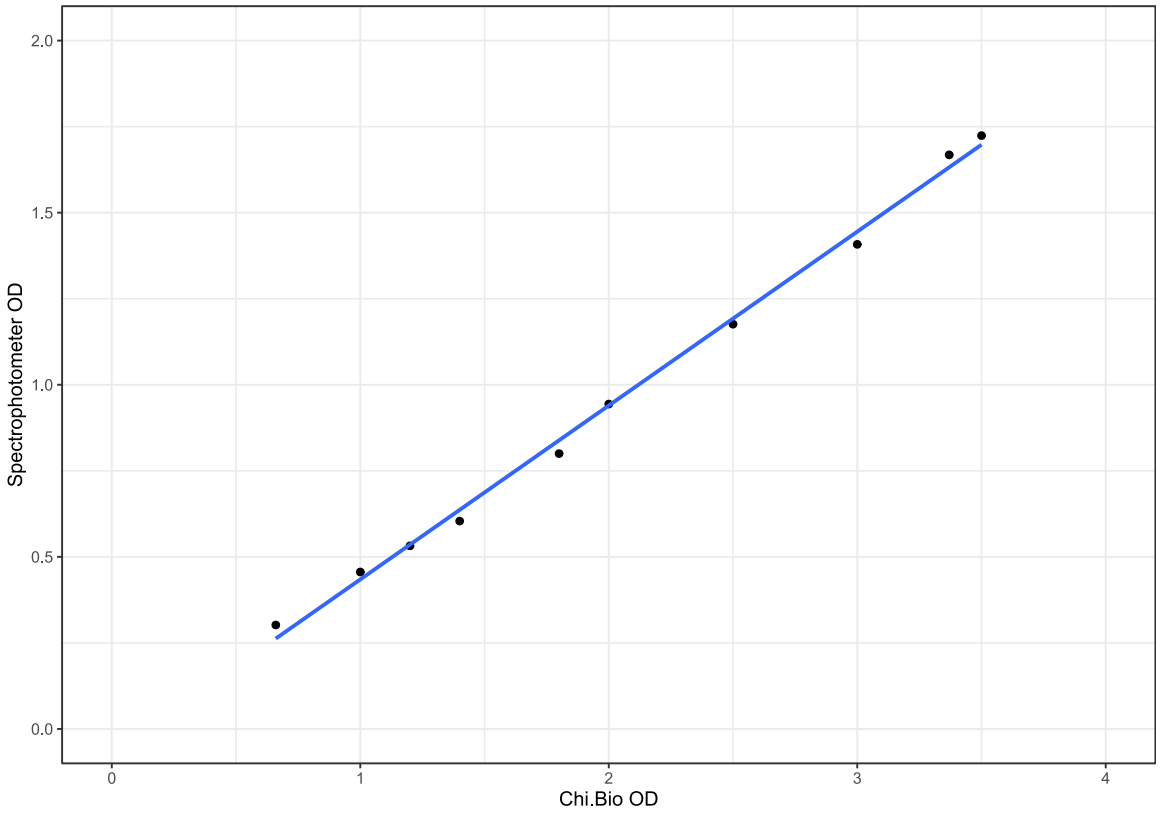


Figure 3.4: The OD registered by a benchtop spectrophotometer was approximately half the value reported by the Chi.Bio machines. To correct for an inflated OD measured by the Chi.Bio bioreactors, OD = 0.50523(Chi.Bio OD measurement) – 0.07078, R^2 = 0.9961, $p<0.001$.

The Chi.Bio bioreactors can either support unrestricted bacterial growth or hold the culture at a specified OD (Figure 3.5). The strain of K-12 bacteria BW25113 was used to illustrate unrestricted growth in M9 glucose minimal medium (Figure 3.5, A-B) or a culture held in the logarithmic phase of growth (Figure 3.5, C-D). The bacteria showed typical growth in M9 glucose minimal media in the Chi.Bio bioreactors (Figure 3.5 A). When the Chi.Bio bioreactors were not set as turbidostats, there was no fresh or waste media exchange and therefore no action of Pump 1 to provide fresh media to the culture (Figure 3.5 B).

The Chi.Bio operated as a turbidostat to hold the culture at a constant OD of 0.43 (Figure 3.5 C), which was facilitated by fresh media being pumped in to the bioreactor (Figure 3.5 D) and an equivalent volume of waste culture being pumped out simultaneously. The culture took five hours to reach an OD of 0.43, then it was held at this OD for 19 hours, during which time fresh M9 glucose was being pumped in at a rate of 15.5 mL $hour^{-1}$. Waste culture was being pumped out at an equivalent volume per hour in order to maintain the culture at a constant volume of 25 mL.

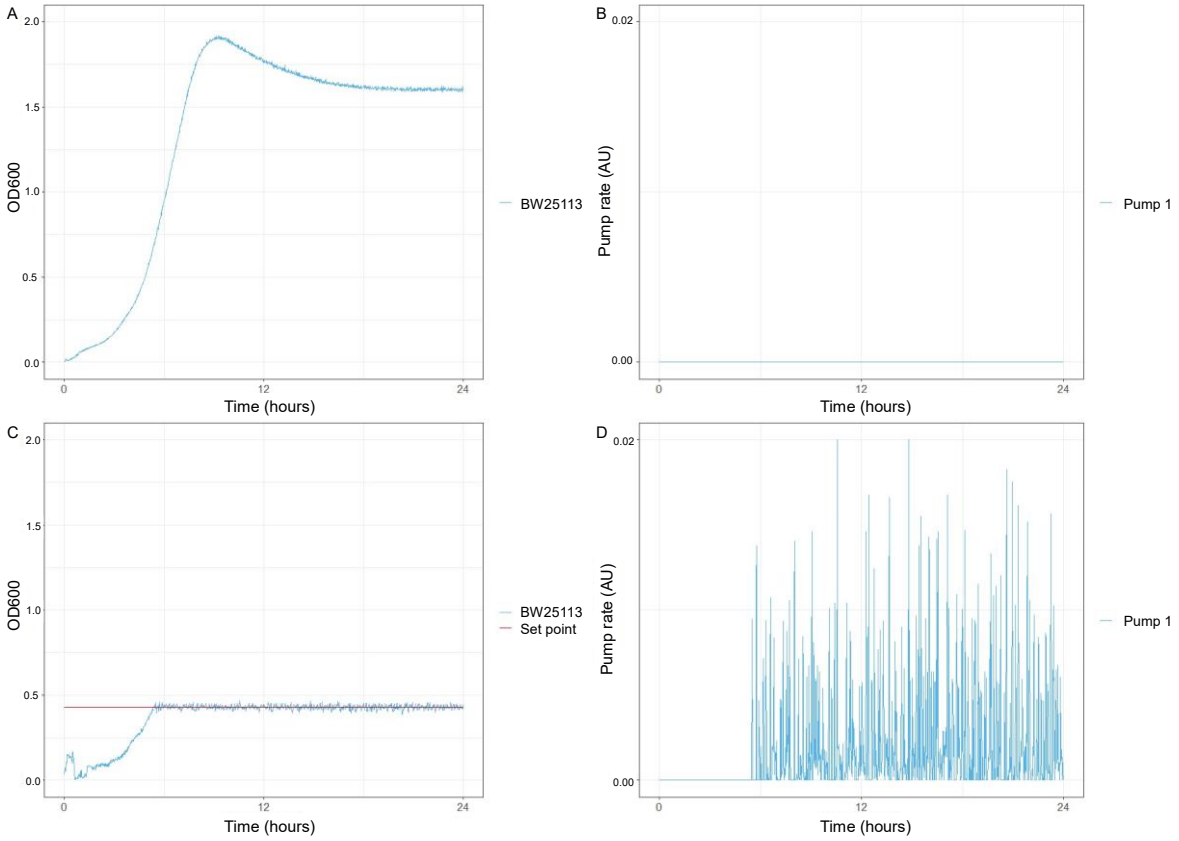


Figure 3.5: Chi.Bio bioreactors can support unlimited bacterial growth or continuous bacterial co-culture in the logarithmic phase of growth. The optical density and pump rate of fresh media for a bioreactor set to grow in an unrestricted manner (A-B) or set to hold the bacterial culture in the logarithmic phase of growth (C-D). A) The OD of the bacterial culture increased as the culture was grown in the media when bacteria were left to grow unrestricted. B) No fresh media was pumped in to the bioreactor when the culture was grown in an unrestricted manner. C) The OD of a Chi.Bio bioreactor that was acting as a turbidostat remained constant over time at the set point (here, OD = 0.43). D) Pump 1 acted to pump fresh media in to the culture in order to maintain a low OD to support continuous bacterial co-culture in the logarithmic phase of growth.

3.2.2 Construction of *Setaria viridis* C_4 PEPC expression vectors

3.2.2.1 Domestication of *Setaria viridis* C_4 *PEPC*

Setaria viridis C_4 PEPC (Sevir.4G143500.1, SvPEPC) is 964 amino acids long, with a tetrameric quaternary structure in a complex with Mg^{2+}. The *SvPEPC* gene (without introns) is 2895 bp. The gene sequence was altered in the following ways prior to expression in *E. coli* in order to accommodate the expression of a plant protein in a bacterial system. First, the sequence was codon optimised for expression in *E. coli* using OPTIMIZER (Puigbò *et al.*, 2007). Second, the sequence was domesticated to remove any remaining BsaI or BpiI recognition sites within the gene to make the coding sequence compatible with Golden Gate Cloning using BsaI and BpiI. Third, a His-tag (6xHis) followed by a glycine linker sequence (2xGly) was included at the N terminus of the protein between the start codon and the next amino acid in the PEPC sequence in anticipation of downstream His-tag purification requirements. This amino acid sequence is henceforth referred to as 6His::PEPC. The DNA sequence design work was undertaken by Michael Niklaus, a previous PhD student in the Kelly lab. I conducted all subsequent cloning and experiments.

A comparison between the N terminus of the SvPEPC coding sequence and the 6His::PEPC coding sequence used in this thesis is shown in Figure 3.6. Inclusion of an N-terminal His-tag has been shown to make little to no difference to bacterial or plant PEPC enzyme kinetics (Dong, Hata and Izui, 1997; Xu *et al.*, 2018; Moody, Christin and Reid, 2020) and is common to facilitate downstream PEPC purification prior to undertaking enzyme kinetics analysis (Dharmarajan *et al.*, 2009; Masumoto *et al.*, 2010; Ying *et al.*, 2017). Therefore, it was not predicted to impact enzyme activity during expression in *E. coli*.

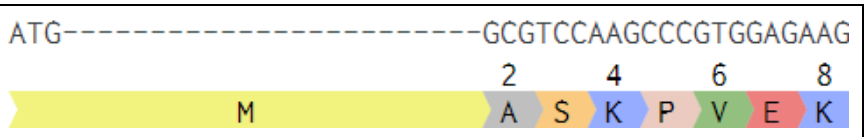
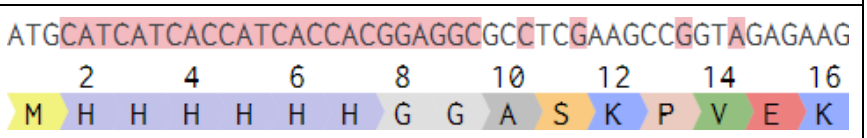
SvPEPC Sevir.4G143500.1	
6His::PEPC	

Figure 3.6: A comparison of the notable differences between the SvPEPC coding sequence and 6His::PEPC altered for *E. coli* expression. DNA bases sit above the amino acid translation of the original plant PEPC coding sequence (above) and the 6His::PEPC sequence optimised for expression in bacteria (below). DNA bases highlighted in red in the lower sequence show where the coding sequence differs between the *6His::PEPC* sequence optimised for expression in *E. coli* compared to the original *SvPEPC* sequence. The sequence for expression in *E. coli* has an N-terminal His-tag (6xHis), a glycine linker (2xGly) and has been codon optimised for expression in *E. coli*, hence the way that an amino acid is encoded may be different in the lower sequence compared to the sequence above despite the amino acid sequence remaining the same. This illustration of codon optimisation for expression in *E. coli* for a subsection of the PEPC sequence is representative of the whole gene.

3.2.2.2 Design of multiple PEPC expression vectors

E. coli require PEPC to grow in M9 glucose minimal medium, in which glucose is the sole carbon source (Cozzone, 1998; Sauer and Eikmanns, 2005). *E. coli* Δppc from the Keio collection (Baba *et al.*, 2006) no longer express the native bacterial PEPC so it is not possible for these bacteria to grow in M9 glucose. In order to screen for improved 6His::PEPC activity in the *E. coli* Δppc mutant strain, bacterial growth rate must be dependent on 6His::PEPC enzyme activity.

The first step in the development of a screen of 6His::PEPC in *E. coli* was to demonstrate that *E. coli* Δppc could not grow in M9 glucose unless the bacterial growth phenotype was rescued by heterologous 6His::PEPC expression. It was also important to demonstrate that the growth rate of *E. coli* Δppc in M9 glucose was dependent on the enzymatic rate of the introduced 6His::PEPC. This is because in order for the screen to work, there had to be a growth rate advantage for the cells that had a higher flux through 6His::PEPC so that these increased in relative abundance in the population. Finally, there must also be sufficient room for improvement of growth to allow for sensitive detection of variants with different degrees of kinetic improvement.

To achieve the three requirements above, a suite of three expression vectors were made that were designed to confer high, medium and low expression of 6His::PEPC respectively (Figure 3.7). To minimise the potential sources of variation, the vector backbone, promoter, 6His::PEPC and terminator sequences were kept the same in all three 6His::PEPC expression vectors. The estimated growth rate of *E. coli* Δppc transformed with 6His::PEPC-high, 6His::PEPC-med and 6His::PEPC-low in M9 glucose during the logarithmic phase of growth were expected to be fastest, second fastest and slowest, respectively.

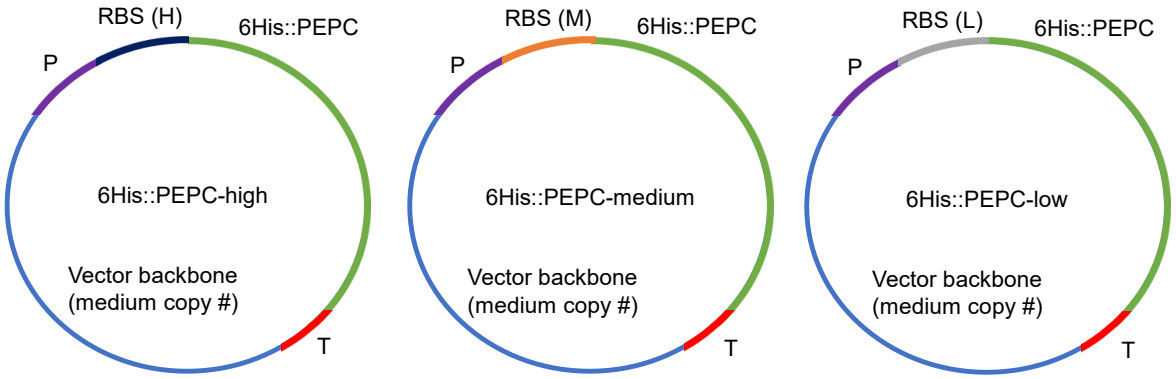


Figure 3.7: Vectors designed to confer a range of 6His::PEPC expression levels in *E. coli* Δppc. Promoter (P), ribosome binding sites (RBS) that confer predicted high (H), medium (M) and low (L) strengths of expression, terminator (T) all shown.

3.2.3 6His::PEPC expression level can be finely tuned in *E. coli* Δppc

The three constructs expected to confer high, medium and low levels of 6His::PEPC respectively (Figure 3.7) were separately transformed into *E. coli* Δppc to create high, medium and low 6His::PEPC expression strains. The activity of each ribosome binding site was previously determined using GFP (Haskett *et al.*, 2022), but had not been previously characterised in *E. coli* Δppc or with the 6His::PEPC coding sequence. Thus, the level of 6His::PEPC expression in each of the strains was validated by immunoblot in triplicate. In order to identify 6His::PEPC variants that perform well in the bacterial screen, the growth rate of *E. coli* Δppc must be dependent upon 6His::PEPC activity. Thus, the first step was to determine the expression level of 6His::PEPC in the three *E. coli* Δppc strains. The second step was to validate that *E. coli* Δppc growth rate is dependent upon level of 6His::PEPC expression. The final step was to restrict the expression level of 6His::PEPC variants, such that any increase in enzymatic rate *in vivo* would lead to an increased growth rate of *E. coli* Δppc transformed with this variant.

Each bacterial strain was held in the logarithmic phase of growth (OD = 0.63) in liquid culture using the Chi.Bio bioreactors, then a sample of identical volume was taken from each bacterial culture in order to harvest an equivalent mass of cells for each strain. Immunoblots were subsequently performed to quantify the expression level of 6His::PEPC in each strain. The expression level of 6His::PEPC in each strain was in agreement with the expected strength of the RBS (Figure 3.8 A, B). It is usual to include a loading control for each sample to confirm equivalent levels of protein loading between samples. Glyceraldehyde 3-phosphate dehydrogenase (GAPDH) protein can be used as a loading control for bacterial samples (Wu *et al.*, 2012), however, the role of GAPDH in glycolysis makes it an unsuitable control

because glycolysis and the TCA cycle are so closely interlinked (Figure 3.1) so any change to PEPC expression level could impact GAPDH regulation. Instead of using a loading control, a duplicate SDS PAGE gel was loaded and stained with QuickBlue Protein Stain to demonstrate equivalent protein loading (Figure 3.8 C).

Each blot was imaged and the band intensities quantified (Figure 3.8 D). There was a statistically significant difference between 6His::PEPC expression level in each strain ($p<0.001$) (Figure 3.8 B, Table 3.1). On average, 6His::PEPC abundance was 1.9- and 2.7-fold higher, respectively, in the cell lines containing the medium and high expression vectors relative to the cell line containing the low expression vector (Table 3.1).

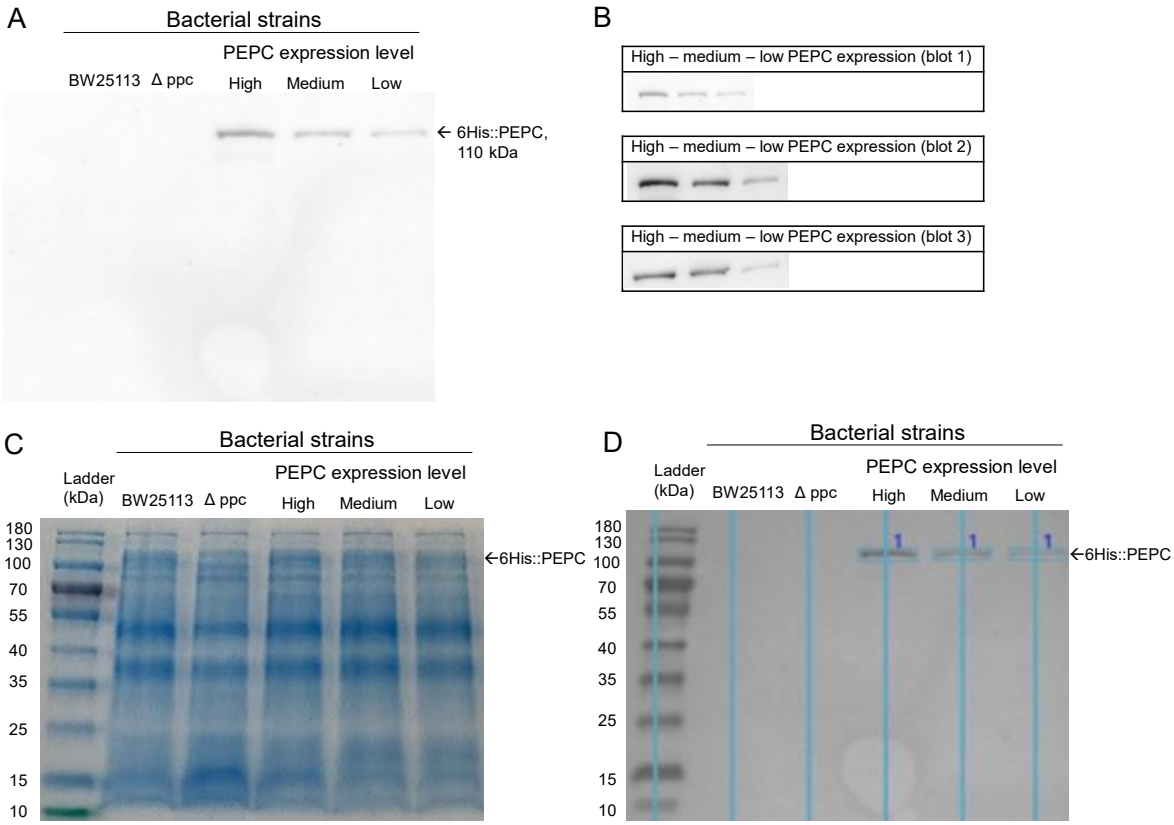


Figure 3.8: Expression level of PEPC in *E. coli* Δppc was proportional to predicted strength of the ribosome binding site. A) Detection of 6His::PEPC expression in *E. coli* Δppc high, medium and low 6His::PEPC expression strains compared to two control strains that were not transformed with 6His::PEPC. B) Detection of 6His::PEPC expression in the bacterial strains was undertaken in triplicate through three immunoblots undertaken independently. C) A replica SDS-PAGE gel was stained with QuickBlue Protein Stain to demonstrate the level of protein loading in each lane. D) Band identification of the immunoblot in part A prior to band intensity quantification.

Table 3.1: The strength of PEPC expression in *E. coli* Δppc 6His::PEPC strains was proportional to the predicted strength of the RBS in the PEPC expression construct.

Immunoblot	Bacterial strain	Band intensity value	Fold change over lowest value in each replicate
1	High	2.480e+06	2.95
	Medium	1.237e+06	1.47
	Low	840,827	Reference value
2	High	6.830e+06	2.76
	Medium	4.915e+06	1.98
	Low	2.478e+06	Reference value
3	High	4.742e+06	2.43
	Medium	4.179e+06	2.14
	Low	1.954e+06	Reference value

Average relative expression levels of 6His::PEPC in the high and medium expression strains were 2.17- and 1.86-fold higher than in the lowest expression strain, respectively. A one-way ANOVA was performed to compare the effect of RBS on strength of 6His::PEPC expression (n=3 per strain). There was a statistically significant difference in 6His::PEPC expression level between the strains ($F(2, 6) = 90.44$, $p<0.001$). Tukey's HSD Test for multiple comparisons found that the expression level of the predicted high and medium expression strains was significantly higher than the predicted low expression strain ($p<0.0001$ and $p<0.001$, respectively), and were significantly different to one another ($p<0.05$).

3.2.4 Growth rate of *E. coli* Δppc in M9 glucose is dependent on the enzymatic rate of 6His::PEPC

Plant C_4 PEPCs have frequently been expressed in *E. coli* prior to X-ray crystallography studies or enzyme assays, including C_4 PEPCs from maize, sorghum, *F. trinervia* and 20 different C_4 grasses (Crétin *et al.*, 1991; Wang *et al.*, 1992; Duff *et al.*, 1993; Bläsing, Westhoff and Svensson, 2000; Endo *et al.*, 2008; Jacobs *et al.*, 2008; Paulus, Schlieper and Groth, 2013; González-Segura *et al.*, 2018; DiMario and Cousins, 2019; Moody, Christin and Reid, 2020; Muñoz-Clares *et al.*, 2020; DiMario *et al.*, 2021). In these studies, multiple different *E. coli* strains have been used for heterologous C_4 PEPC expression and purification, including PEPC-deficient strains XH11 (Crétin *et al.*, 1991) and PCR1 (Sabe *et al.*, 1984), and commercial strains derived from BL21(DE3). Thus, heterologous expression of C_4 PEPC in multiple *E. coli* strains has resulted in a functional enzyme. In addition, it has been shown that sorghum C_4 PEPC rescues growth of PEPC-deficient strain XH11 in M9 glucose (Crétin *et al.*, 1991). However, no studies have been published to date that have expressed and purified *S. viridis* C_4 PEPC from *E. coli*; *S. viridis* C_4 PEPC kinetics have previously been characterised using the enzyme purified from leaf discs (Boyd, Gandin and Cousins, 2015). Given the number of C_4 PEPCs that have been successfully expressed in *E. coli*, it appeared to be likely that expression of 6His::PEPC in *E. coli* Δppc would result in a functional enzyme. However, it was not known whether the expression of 6His::PEPC in *E. coli* Δppc would rescue growth of the bacteria in M9 glucose, consistent with the results obtained using the sorghum C_4 PEPC, or the level of *6His::PEPC* expression that would be required to achieve this.

As discussed above, PEPC-deficient *E. coli* are unable to grow in M9 glucose, where glucose is the sole carbon source, because PEPC is obligatory for anaplerosis to maintain TCA cycle function (Figure 3.9) (Ashworth and Kornberg, 1966; Cozzone, 1998; Sauer and Eikmanns, 2005; Tong *et al.*, 2020). Indeed, PEPC-deficient *E. coli* Δppc was unable to grow on M9 glucose (Figure 3.9). However, growth of *E. coli* Δppc was restored on expression of 6His::PEPC (Figure 3.9), demonstrating that functional 6His::PEPC protein was being produced and was sufficient to restore bacterial growth on M9 glucose. The average doubling time in the logarithmic phase of growth was 1.90, 2.26, and 3.97 hours for the high, medium, and low 6His::PEPC expressing lines, respectively (Figure 3.10). Thus, the highest expressing 6His::PEPC line grew at over twice the rate of the slowest growing *E. coli* Δppc::6His::PEPC strain in the logarithmic phase of growth. This was similar to the difference in protein abundance between the two strains - 2.1 fold faster growth and 2.7 fold higher protein abundance (Table 3.1, Figure 3.10).

The growth rate of each *E. coli* Δppc::6His::PEPC strain in the logarithmic phase of growth was a function of the predicted strength of PEPC expression (Figure 3.10). Indeed, the *E. coli* Δppc::6His::PEPC strain transformed with an expression vector designed to confer the lowest level of PEPC expression grew significantly slower in the logarithmic phase of growth than the *E. coli* Δppc::6His::PEPC strains transformed with expression vectors designed to confer higher levels of PEPC expression and the two control strains BW25113 and BW25113 + pOGG024 ($p<0.001$) (Figure 3.10). Given that the enzyme sequences were identical in each line, this indicates that the growth rate was dependent on, and proportional to, 6His::PEPC expression level and therefore also the enzymatic rate, in terms of the moles of substrate converted per unit time *in vivo*, by 6His::PEPC.

The parent strain to the Keio collection (BW25113, derived from *E. coli* K-12) could grow in M9 glucose because the endogenous bacterial PEPC was providing anaplerotic function in these bacteria (Figure 3.9). The average doubling time of BW25113 in the logarithmic phase of growth in M9 glucose in the Chi.Bio bioreactors was 75 minutes (Figure 3.10), whereas previous doubling times reported for *E. coli* K-12 or K-12 derived strains in M9 glucose were between 40 minutes and 1 hour (Reshes *et al.*, 2008; Jain and Srivastava, 2009; Thakur *et al.*, 2010). Thus, the growth rate of BW25113 was consistently slower than published results but very similar within replicates. Interestingly, the growth rate was similar to that reported of *E. coli* grown anaerobically on glucose (Gonzalez, Long and Antoniewicz, 2017).

During anaerobic growth on glucose, *E. coli* are still dependent on PEPC to supply carbon to the TCA cycle since there is no carbon flux through the glyoxylate bypass (Gonzalez, Long and Antoniewicz, 2017). However, carbon flux is altered including a low rate of flux through the TCA cycle which becomes bifurcated at oxaloacetate (OAA), so OAA is a substrate for the production of citrate but also for the production of malate which is subsequently converted into succinate (Gonzalez, Long and Antoniewicz, 2017). If the cultures were not sufficiently aerated, this could have been impacting flux through core carbon metabolism. However, the Chi.Bio bioreactors created a whirlpool which aerated the bacterial culture for growth. A more likely cause of the discrepancy between the published growth rate of *E. coli* K-12 and the growth rate of BW25113 measured in this study is that the OD is measured every minute, at which point the Chi.Bio paused stirring before taking a suite of measurements over a few seconds. A faster rate of bacterial growth may have been observed if the Chi.Bios were programmed to take OD measurements

less frequently with more constant culture agitation. Despite the potential to further optimise bacterial culture in the Chi.Bios, it was concluded that the bacteria were growing aerobically and that the system was still suitable for use to screen for variants in 6His::PEPC kinetics as growth rate was demonstrated to be dependent on level of 6His::PEPC expression.

In addition to testing the growth rate of *E. coli* Δppc and the parent strain, BW25113, BW25113 transformed with the original backbone plasmid pOGG024 were grown in M9 glucose to approximate the impact of maintaining the plasmid on the growth of the *E. coli* Δppc mutant strain (Figure 3.9). When the pOGG024 vector was introduced to BW25113 cells, the growth rate in the logarithmic phase of growth slowed, increasing the doubling time to an average of 1.72 hours (Figure 3.10). The slower growth of BW25113 + pOGG024 compared to BW25113 in M9 glucose demonstrated that maintaining the plasmid and producing the proteins required to confer antibiotic resistance to gentamycin resulted in a slower growth rate of BW25113 + pOGG024. Interestingly, all three *E. coli* Δppc strains expressing 6His::PEPC grew more slowly than BW25113 + pOGG024 in M9 glucose in the logarithmic phase of growth (Figure 3.9, Figure 3.10). This suggests that although 6His::PEPC can restore growth, the exogenous transgene under the control of a constitutive promoter was not as well adapted to function in *E. coli* as the native *E. coli* PEPC orthologue under the control of its own promoter.

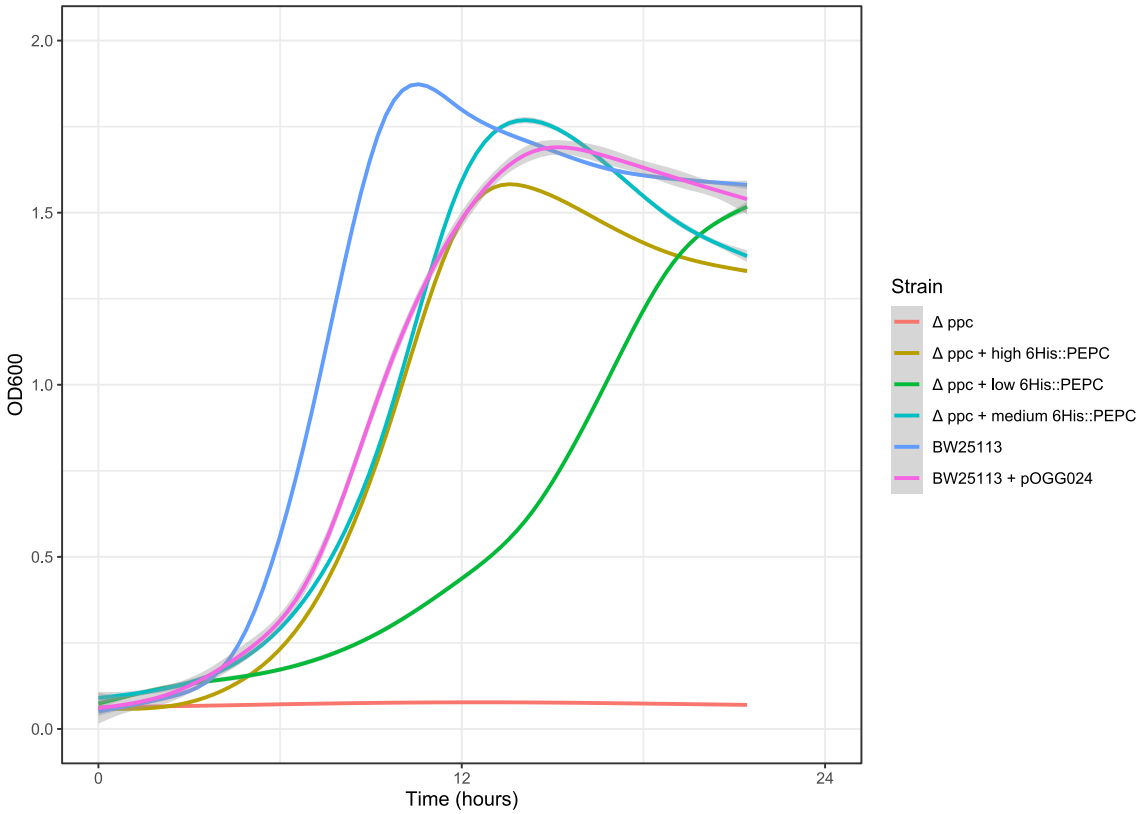


Figure 3.9: Growth of *E. coli* Δppc is restored in M9 glucose when the bacteria express 6His::PEPC. Data are mean ± 95% shaded CI (n=3) of six bacterial strains in M9 glucose in the Chi.Bio bioreactors. Δppc refers to *E. coli* Δppc that lack an endogenous PEPC. BW25113 was the parent strain for generating *E. coli* Δppc. BW25113 + pOGG024 was transformed with the backbone used to create the 6His::PEPC expression vectors. *E. coli* Δppc was transformed with vectors designed to confer a high, medium and low level of 6His::PEPC expression.

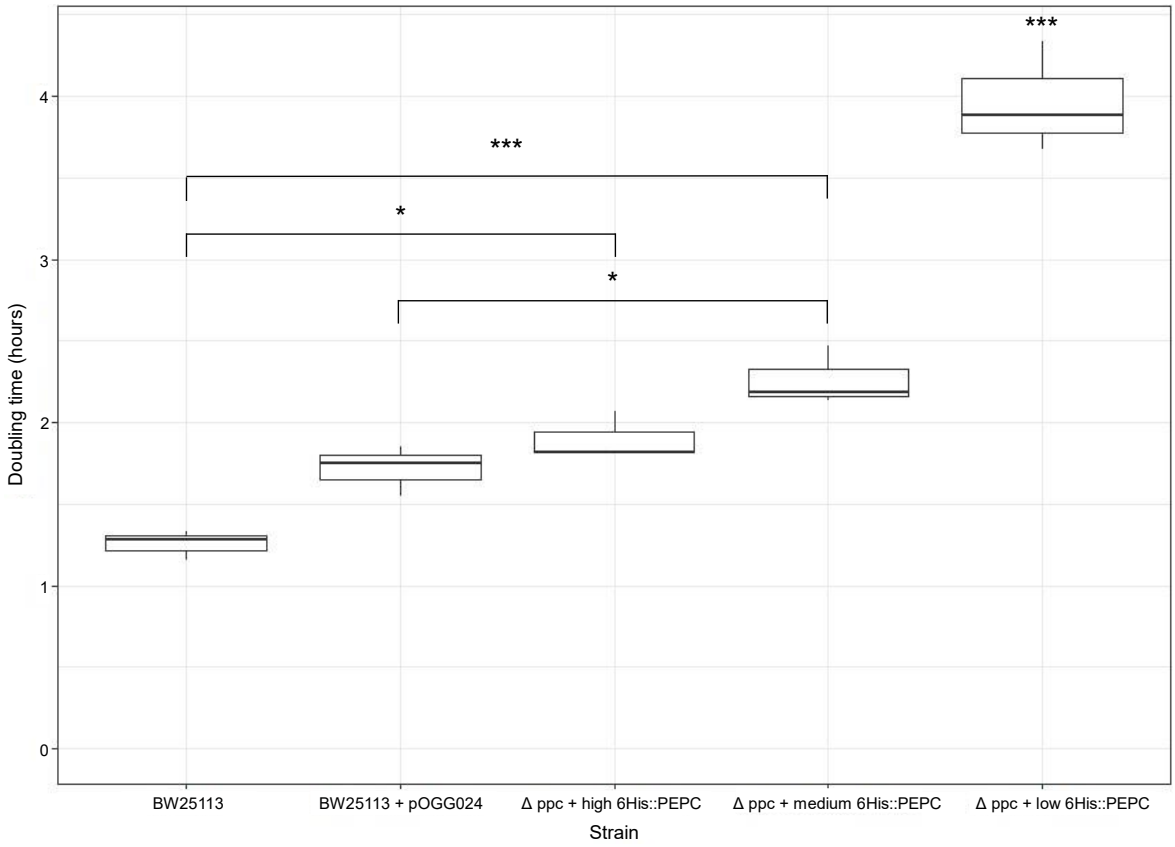


Figure 3.10: Growth rate of *E. coli* Δppc was proportional to 6His::PEPC expression level in the logarithmic phase of growth in M9 glucose. BW25113 is the parent strain of *E. coli* Δppc. BW25113 + pOGG024 was transformed with the plasmid backbone used to create 6His::PEPC expression vectors. For the other strains, *E. coli* Δppc was transformed with vectors designed to confer a high, medium and low level of 6His::PEPC expression. Statistical significance in growth rate between strains is indicated above the plot as determined by a one-way ANOVA and Tukey's HSD Test for multiple comparisons, n=3 per strain, *** $p<0.001$, ** $p<0.01$, * $p<0.05$. The mean doubling time of the lowest PEPC expression strain was significantly lower than that of all other strains.

The results above demonstrated that growth rate of *E. coli* Δppc was dependent on the expression level, and therefore enzymatic rate, of 6His::PEPC when the bacterial culture was grown in M9 glucose. A screen of 6His::PEPC activity could therefore be established by measuring changes to bacterial population dynamics over time. Here, 6His::PEPC variants could be cloned into the expression vector that confers the lowest level of 6His::PEPC expression and the vector subsequently transformed into *E. coli* Δppc. As each 6His::PEPC variant would be expressed at an equivalent low level, the growth rate of each transformant would be dependent on, and limited by, the enzymatic rate of 6His::PEPC *in vivo*. Thus, a transformant cell expressing a 6His::PEPC variant with a higher enzymatic rate should grow faster than a transformant expressing a 6His::PEPC variant with a lower enzymatic rate, and therefore become more highly represented in the population over time during the screen. In this way, libraries of PEPC variants can be screened simultaneously using individual cell growth rate as a proxy for 6His::PEPC enzymatic rate. Moreover, given that cells containing the low expression vector grew twice as slowly as cells containing the high expression vector, the screen has the capacity to detect up to a two-fold increase in the enzymatic rate of 6His::PEPC.

3.2.5 Simulating the population dynamics of an *E. coli* Δppc culture that contains bacterial strains with different rates of growth

Although the outcomes of a real screen for variation in 6His::PEPC kinetics were unknown, it was considered valuable to model the population dynamics of *E. coli* Δppc containing 6His::PEPC variants in order to simulate how the abundance of different transformants may change over the course of a screen. First, a continuous co-culture of three different strains was simulated whereby the growth rate of the medium and fastest growth strains were double and quadruple the slowest growth strain, respectively (Figure 3.11A). The relative abundance of the fastest growth strain increased rapidly over time such that in the time it would take for the fastest growth strain population to double 15 times, >99% of the culture is the fastest growth strain (Figure 3.11 A). This phenomenon became more pronounced the larger the difference in growth rate. For example, if the growth advantage of the fastest and medium growth strains were 8-times and double the slowest growth strain, respectively, then the strain that was growing fastest was predicted to reach >99% of the population in the time it would take for the fastest growth strain population to double 10 times (Figure 3.11 B). Thus, should a mutation to 6His::PEPC confer an equivalent increase to bacterial growth rate as simulated here then the abundance of *E. coli* Δppc expressing the best-performing variant should substantially increase over a small number of generations.

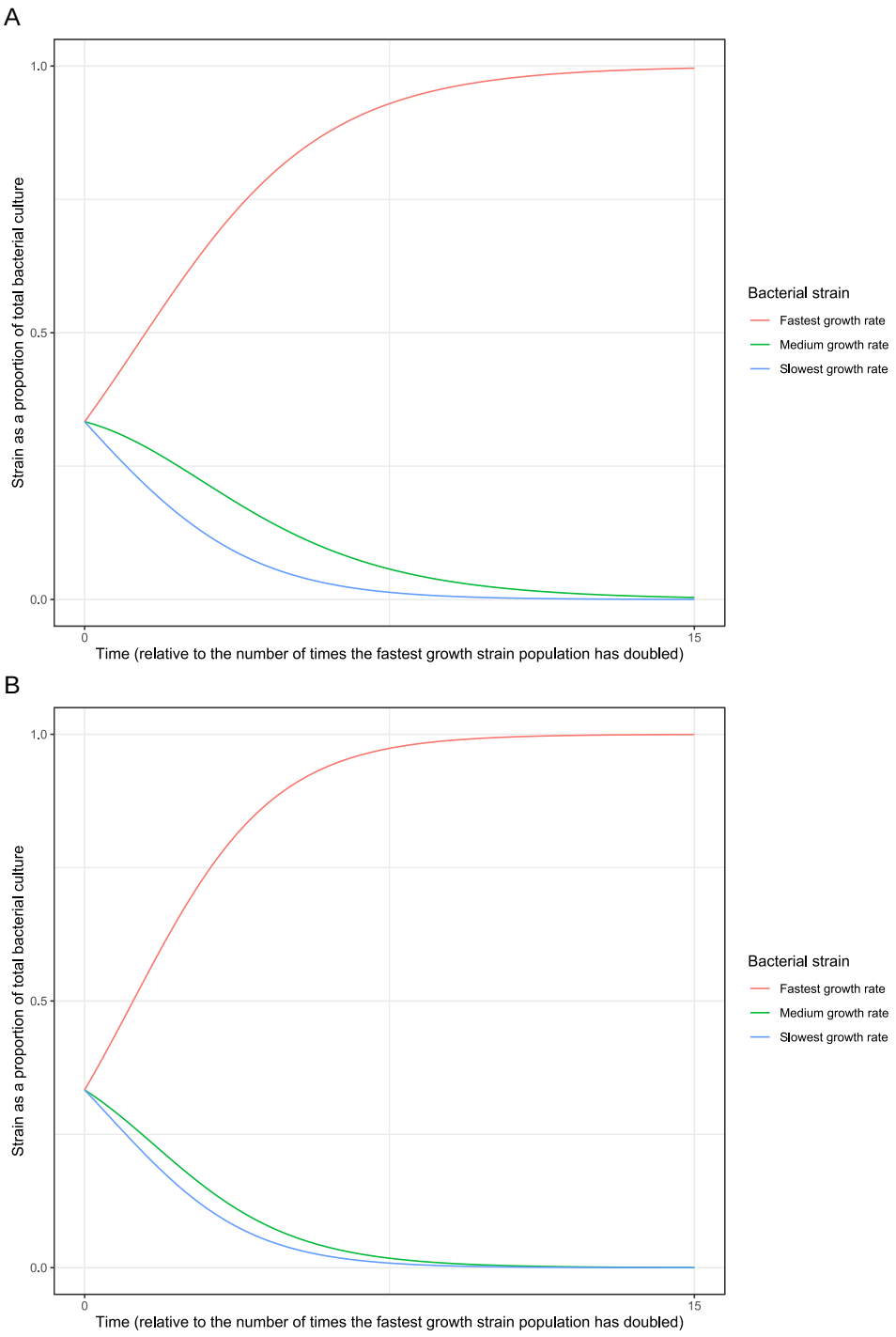


Figure 3.11: Computational modelling predicts that *E. coli* Δppc with a faster growth rate rapidly outcompete *E. coli* Δppc with a slower growth rate, therefore, if bacterial growth rate is proportional to the enzymatic rate of 6His::PEPC, bacteria that express the best-performing 6His::PEPC variant(s) will become highly represented over time in continuous co-culture. Figure legend continued on the next page.

Figure 3.11 (continued) A) Population dynamics of bacteria that have quadruple or double the growth rate compared to the slowest growing strain (fastest and medium growth rates, respectively). B) Population dynamics of bacteria that have 8-times or double the growth rate compared to the slowest growing strain (fastest and medium growth rates, respectively).

3.2.6 Bacteria that have the highest 6His::PEPC expression outcompete bacteria that have lower 6His::PEPC expression in continuous co-culture

A proof of concept experiment was designed to compete the strains of *E. coli* Δppc expressing high, medium or low levels of 6His::PEPC (Figure 3.7) against each other in continuous co-culture. The three strains were inoculated in equal proportion in M9 glucose and held in the logarithmic phase of growth for 24-h in the Chi.Bio bioreactors.

First, the previously characterised growth rates of each strain (Figure 3.10) were used to model the changes in population composition over time (Figure 3.12). In 24-h, the expected number of generations of the high, medium and low PEPC expression strains were 12.63, 10.59 and 6.05, respectively. Figure 3.12 showed that after 24-h the relative proportion of cells from the high, medium and low expression strains are expected to be 79.8%, 19.4% and 0.8%, respectively, from each strain being equally represented at the start of the co-culture assay. Large changes to relative abundance of each strain are predicted to occur within this time frame, suggesting that the experimental application of this approach should be a powerful method to quickly screen a library of 6His::PEPC variants simultaneously for increased enzymatic rate.

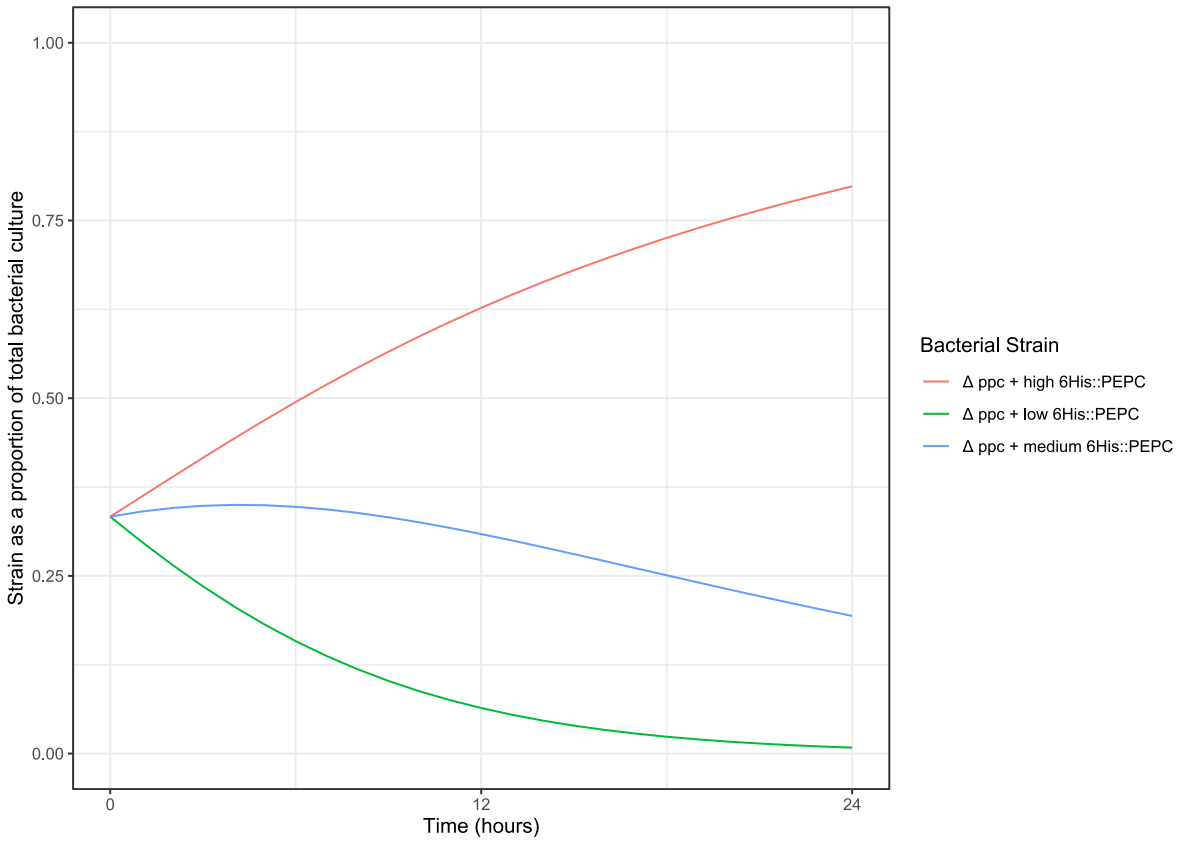


Figure 3.12: Predicted population dynamics of three *E. coli* Δppc strains that grow at different rates conferred by different levels of 6His::PEPC expression held in continuous co-culture.

A semi-quantitative PCR step was used in order to track the relative abundance of each *E. coli* Δppc strain expressing high, medium or low levels of 6His::PEPC over time in order to validate the model in Figure 3.12. Three sets of primers pairs were designed that each amplify one of the three strains (Figure 3.13, Table 3.2). The forward primer was unique to each expression vector and the reverse primer was the same in all reactions. In this way, each primer pair was specific to one of the three expression vectors. The primer pairs were validated on plasmid DNA to ensure that each primer pair was specific to its target expression vector (Figure 3.14 A). The number of cycles was minimised in order to maintain the semi-quantitative aspect of the PCR. It was established that amplification of 0.1-1 ng of template DNA was detectable after 15 cycles of PCR amplification with all three primer pairs (Figure 3.14 B). Therefore, 0.5 ng template DNA and 15 cycles were used to amplify DNA extracted from the bioreactors at the start and end of a 24-hour screen of the three strains in order to observe changes to the relative abundance of each strain over time.

The strains that expressed higher levels of 6His::PEPC outcompeted lower expressing strains over time in continuous co-culture (Figure 3.15). All the expression strains were detected in the initial culture (Figure 3.15 A). After 24-h of co-culture, the high and medium strains were still detectable, but the low expression vector was significantly less detectable than at T_0 (Figure 3.15 B). This experimental data is aligned with the model of population change over a 24-h period for these *E. coli* Δppc strains, which predicted that after 24-h in co-culture, the lowest PEPC expression strain would be less than 1% of the total culture volume, reduced from 33% at T_0 (Figure 3.12). There were insufficient data points to detect a significant difference between amplicon brightness at T_0 and T_{24} for the high PEPC expression

vector (T_0, n=1 and T_{24}, n=3) (Figure 3.15 B), but there appeared to be a small increase in brightness after 24-h (Figure 3.15 A).

This experiment was a proof of concept that *E. coli* Δppc strains that expressed high levels of 6His::PEPC and therefore had a high growth rate outcompeted other strains over time. However, the way of visualising the abundance of each expression strain used here by loading PCR product on a gel is a much less sensitive way of determining relative abundance compared to other methods of quantification, for example sending samples for next-generation sequencing (NGS). Employing methods such as NGS to quantify 6His::PEPC variant abundance over time should facilitate detection of changes in abundance with greater sensitivity.

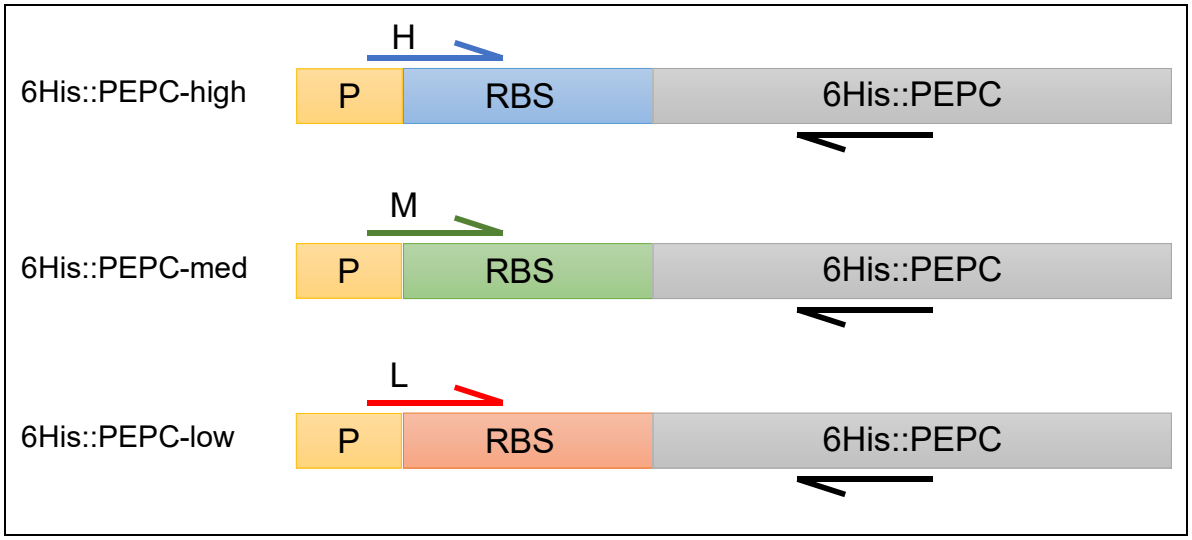
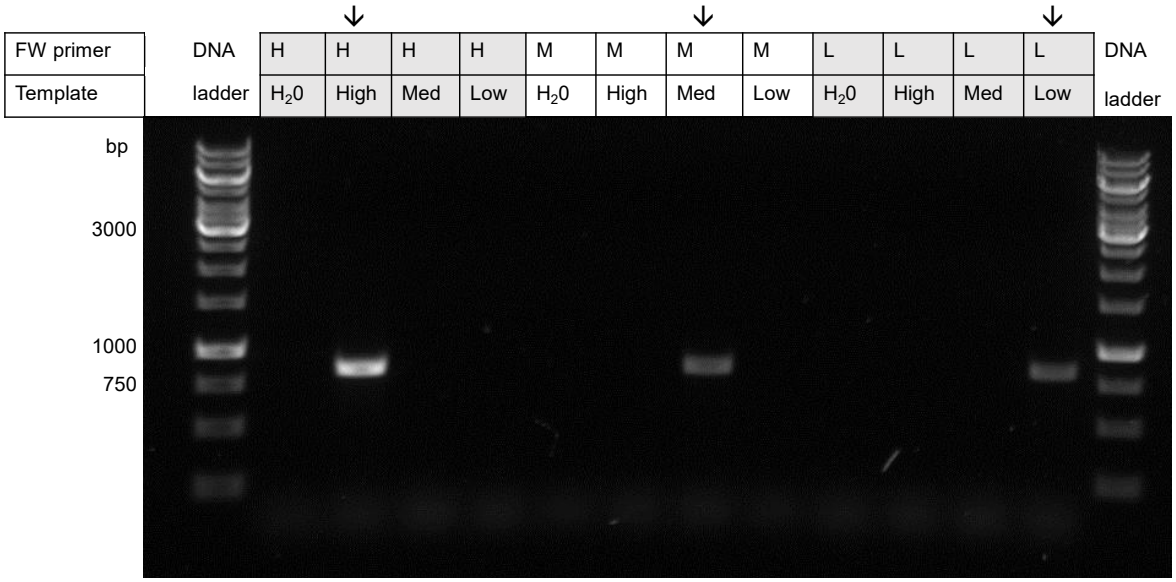


Figure 3.13: Primer design for a semi-quantitative PCR that discriminated between three *6His::PEPC* constructs. In order to achieve amplification of each expression vector separately, the forward primers were different and could anneal only to one of the three ribosome binding sites used in the suite of expression vectors. The reverse primer anneals to *6His::PEPC*, which remained the same between all three expression vectors, so the reverse primer was the same in each primer pair.

Table 3.2: Sequences for three sets of primers that each amplify a fragment from one of each of the high, medium and low PEPC expression vectors.

Vector	Forward primer	Reverse primer
High	CTCAGTCCTAGGTATAGTGCTAGCTACTCTAGAATTAAAG	TTCCAATGTTTTTCAGGGCCGT
Medium	GCTACTCTAGAGCTAATCTTCGCGTACTAAAA	TTCCAATGTTTTTCAGGGCCGT
Low	GCTACTCTAGAGCTAATCTTCGCGTACTAA**GT**	TTCCAATGTTTTTCAGGGCCGT

A



B

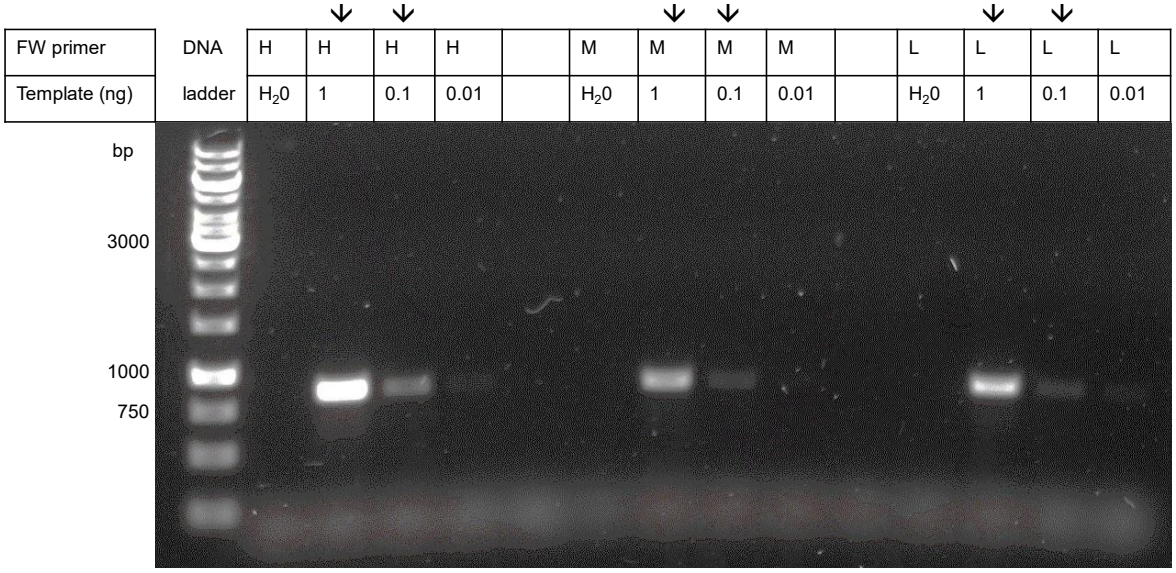


Figure 3.14: Three primer pairs designed to amplify DNA from the high, medium and low expression vectors were specific to one of the three vectors.

A) Three sets of primers (H, M and L) were tested on each of three templates (High, Medium and Low). Arrows indicate conditions in which amplification was successful. B) The amount of template required to achieve amplification of DNA after 15 cycles of PCR for all three primer-template combinations was 0.1 ng.

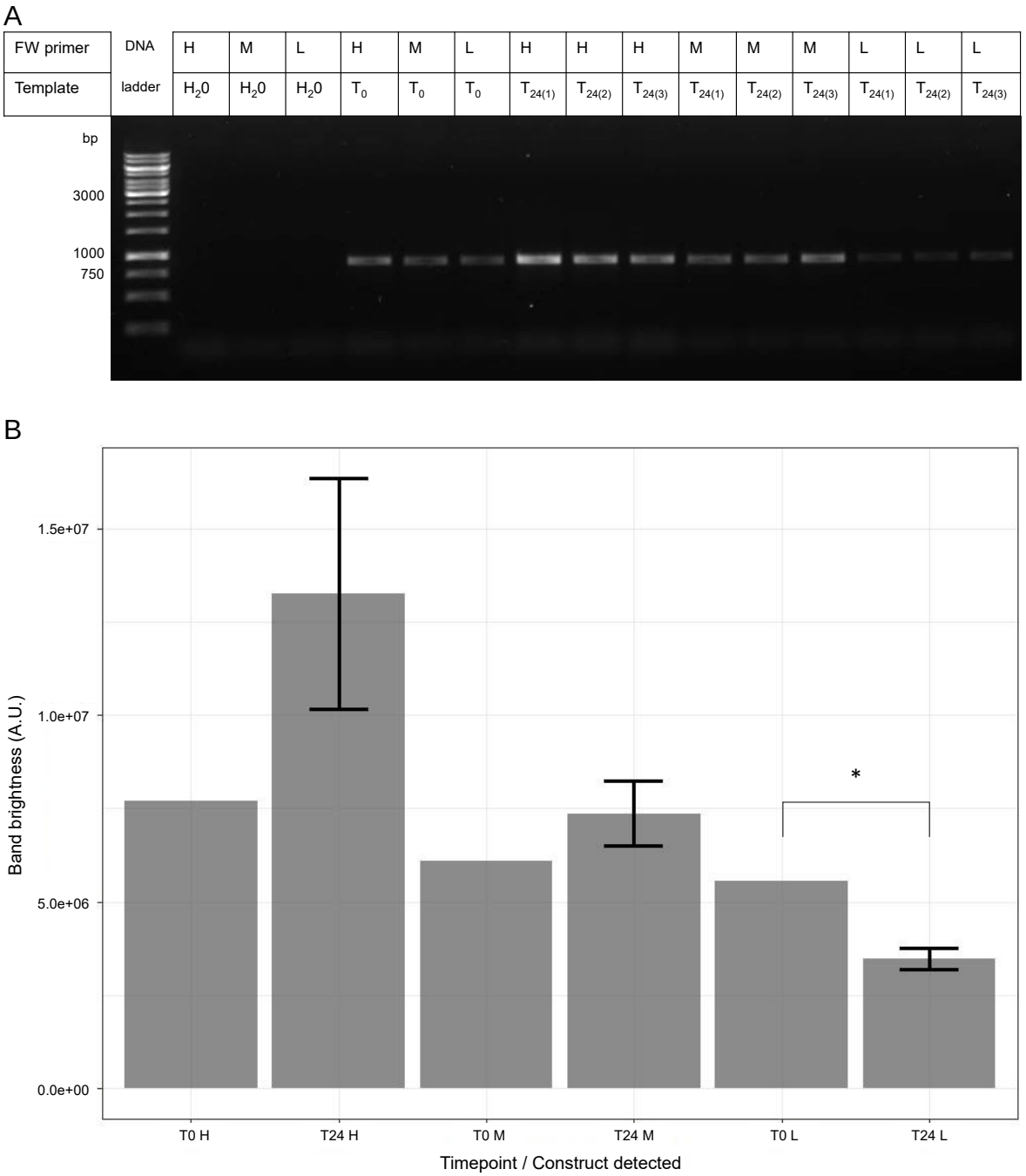


Figure 3.15: A semi-quantitative PCR demonstrated that strains with higher heterologously-expressed PEPC abundance outcompeted strains with lower PEPC abundance over time. Figure legend continued on the next page.

Figure 3.15 (continued) A) A semi-quantitative PCR was performed on DNA extracted from a culture of *E. coli* Δppc expressing 6His::PEPC at high, medium and low abundance inoculated to the same density in an initial co-culture (T_0) and prepared from a sample of the co-culture in each bioreactor after 24-h. Experiment carried out in triplicate. H, M and L indicate lanes in which high, medium and low expression constructs are detectable. H_20 lanes are three PCR controls without any template DNA added. B) There was a significant difference between the band intensity of the low expression vector amplicon band after 24-h compared to at T_0 ($p < 0.5$). Data is the mean brightness of the three PCR bands per condition ±SD.

3.3 Chapter summary

This chapter described the design and validation of a novel high-throughput system to screen for increased enzymatic activity of 6His::PEPC in *E. coli*. *E. coli* require PEPC to grow in M9 glucose minimal medium and here it has been demonstrated that i) heterologous expression of 6His::PEPC in *E. coli* that lack an endogenous PEPC (Δppc) restored growth of the bacteria in M9 glucose, ii) the growth rate of *E. coli* was limited by the expression level and the enzymatic rate of 6His::PEPC, and iii) there is capacity to double the growth rate of *E. coli* in M9 glucose through improvements in the kinetics of 6His::PEPC. Finally, a pilot proof of concept test of the system was performed to screen for differences in growth rate between strains. Overall, this chapter provided the technology to turn variation in 6His::PEPC enzyme kinetics into variation in cell growth rate. The next step in the creation of a fully working screen was to generate a large library of 6His::PEPC enzyme variants to test in this system. The next chapter describes generation of this library of variants and the optimisation of *E. coli* Δppc transformation.

3.4 Authorship statement

The domestication of *Setaria viridis* C_4 PEPC described in Chapter 3.2.2.1 was undertaken by Michael Niklaus, a previous DPhil student under the supervision of Professor Steven Kelly. All other material is my own.

Chapter 4: Generation of *6His::PEPC* variants *via* site saturation mutagenesis

4.1 Introduction

4.1.1 Choice of mutagenesis approach to create *6His::PEPC* variant libraries

In order to screen for 6His::PEPC enzymes with enhanced kinetics it was first necessary to generate a library of enzyme variants. This can be done in a random or rational manner and both approaches have previously been employed to generate variants with desired functionality for multiple different enzymes (Chen and Arnold, 1993; Zhao, reviewed in Chockalingam and Chen, 2002; Chockalingam *et al.*, 2005; reviewed in Cobb, Chao and Zhao, 2013; reviewed in Wang *et al.*, 2021). Traditional directed evolution employs random mutagenesis to generate libraries of enzyme variants (Chen and Arnold, 1993; Wang *et al.*, 2021) and this approach has been successful in generating enzyme variants that exhibit novel or desirable characteristics for a diverse range of applications, from metalloenzymes for chemical synthesis (Yang *et al.*, 2018) to antifungal antibiotics (Cheng *et al.*, 2020).

The major benefit of employing random mutagenesis to generate enzyme variants for directed evolution is that no prior knowledge of the structure-function relationship of the protein is required in order to generate putative improved enzyme variants (Chica, Doucet and Pelletier, 2005; Wang *et al.*, 2021). Moreover, the random nature of mutations that are introduced can lead to the identification of unexpected mutations conferring changes to enzyme kinetics at residues that may not have been targeted in a rational approach (Zhao and Arnold, 1997). However, the number of possible SvPEPC variants is greater than the number of atoms in the universe; there are a possible 20 to the power 964 variants (~1.6 x 10^{1254}) of SvPEPC

compared to 10^{82} atoms in the observable universe so a library of variants generated through random mutagenesis is only a sparse sample of the total possible 6His::PEPC sequence space. Further, the majority of random mutations are likely to be deleterious or neutral to enzyme function and therefore large populations, a strong selective pressure and often multiple rounds of screening are required to identify protein variants that have improved function (Lassner and Bedbrook, 2001).

Since random mutagenesis can only explore a small fraction of the possible sequence space of a given protein, it is desirable to develop a mutagenesis strategy that reduces the number of possible enzyme variants while continuing to interrogate residues that are likely to be influential on enzyme function. This has resulted in attempts to create more compact 'smart' libraries that target structural or functional information about the enzyme and are more practical to screen *in vitro* (Chica, Doucet and Pelletier, 2005). These rational mutagenesis approaches can be employed when residues have been identified that appear to influence protein function in order to reduce the sequence space to be explored and therefore increase the chance of discovering a highly optimised variant for given selection parameters (Chica, Doucet and Pelletier, 2005).

Given the importance of PEPC in C_4 photosynthesis, the enzyme is well-characterised and residues involved in catalysis and allosteric regulation have been identified (Figure 4.1) (Bläsing, Westhoff and Svensson, 2000; Matsumura *et al.*, 2002; Endo *et al.*, 2008; Jacobs *et al.*, 2008; Paulus, Schlieper and Groth, 2013; González-Segura *et al.*, 2018; DiMario and Cousins, 2019; Muñoz-Clares *et al.*, 2020; DiMario *et al.*, 2021). The residues in and around these functional sites are therefore potential targets for rational mutagenesis. However, there are a large

number of potential residues that meet these criteria. In addition, although functional sites are hotspots for mutations that confer improvements to enzyme activity, mutations to sites throughout an enzyme can alter enzyme kinetics in a favourable way (Savile *et al.*, 2010; Jiménez-Osés *et al.*, 2014; Yang *et al.*, 2018).

An alternative approach would be to direct the mutagenesis approach towards the sites under selection in C_4 PEPC that have been identified in previous bioinformatic work (Christin *et al.*, 2007). Targeting these sites for mutagenesis would focus the exploration space of the mutagenesis assay on sites that we can infer are of likely importance to the evolution of PEPC following its co-option into the C_4 cycle evolution (Christin *et al.*, 2007). In total, 21 such sites in the PEPC sequence have been identified as evolving under positive selection in C_4 plants in the C_4 isoform of PEPC (Figure 4.1).

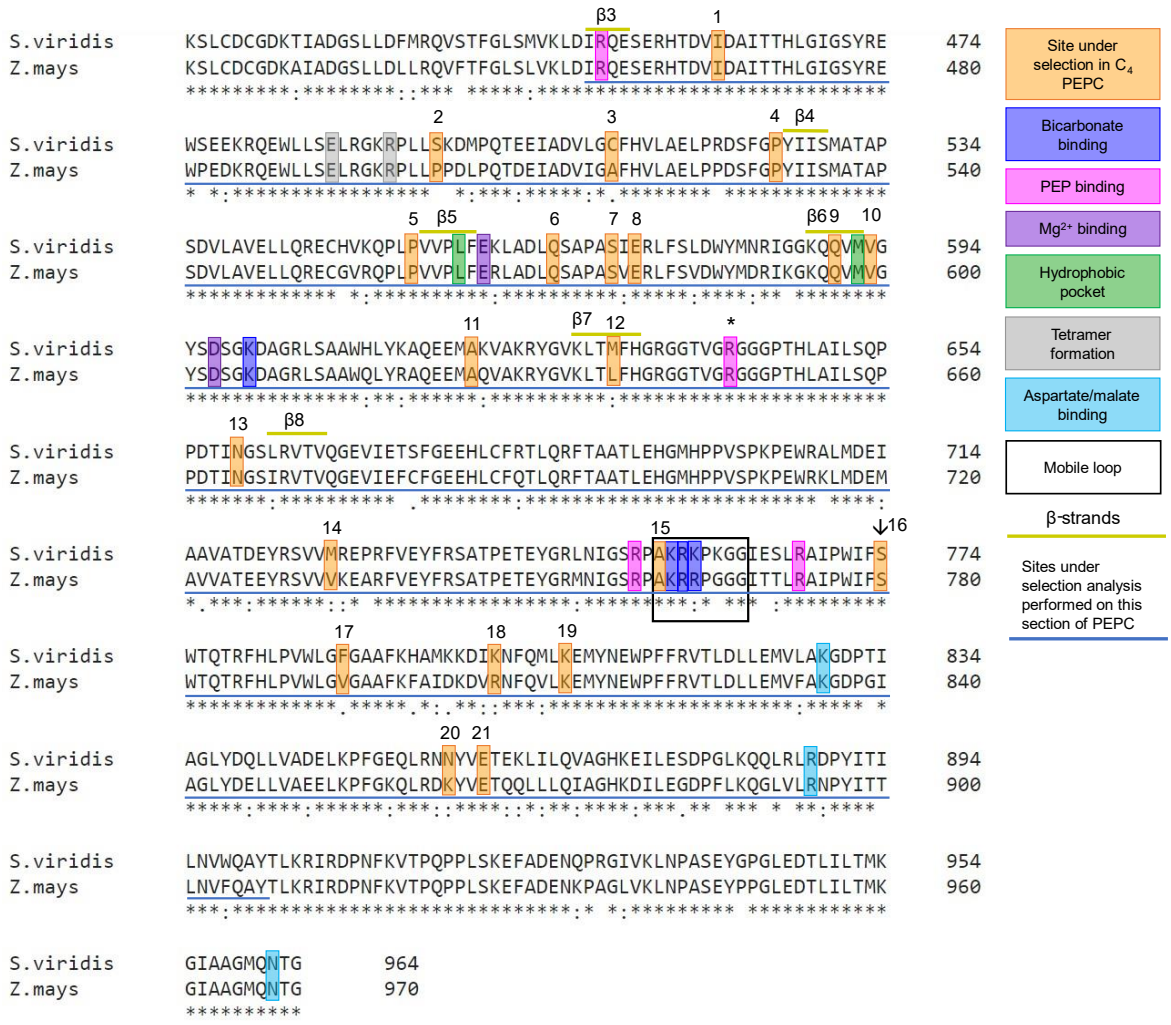


Figure 4.1: Sites under selection in C_4 PEPC highlighted amongst functional residues in SvPEPC and *Z. mays* C_4 PEPC. The 21 sites under selection in C_4 PEPC identified by Christin *et al.* (2007) are numbered and highlighted in yellow. R647 (*Z. mays* numbering) is both a PEP and aspartate/malate binding residue denoted by an asterisk. The serine residue at site 780 that is essential for C_4 PEPC enzyme kinetics and a site under selection in C_4 PEPC is denoted by an arrow. No other site categories overlap. β-strands are numbered and denoted above the corresponding residue.

Sites under selection can be identified in C_4 enzymes due to the many independent origins of C_4 photosynthesis; it is possible to identify sites evolving at a faster rate in C_4 compared to C_3 species by comparison of the DNA sequences that encode C_4 enzymes with the sequences that encode their C_3 homologues (Christin *et al.*, 2007). These sites are likely to contribute to advantageous changes to the enzyme kinetics in the C_4 pathway (Christin *et al.*, 2007), underpin the differences in enzyme kinetics between the C_3 and C_4 homologues and potentially give rise to differences in photosynthetic rate between C_4 species (Niklaus and Kelly, 2019). For example, one site that has been identified in the selection analysis is an alanine to serine substitution at site 780 (*Z. mays* numbering) that is known to be an essential residue for C_4-specific PEPC enzyme kinetics (Engelmann, Bläsing and Westhoff, 2002; Svensson, Bläsing and Westhoff, 2003; Christin *et al.*, 2007; Endo *et al.*, 2008; DiMario and Cousins, 2019).

Residues 455 – 907 of PEPC were selected for analysis of sites under selection because many sites known to contribute to C_4-specific PEPC kinetics are located in this region of the enzyme (Bläsing, Westhoff and Svensson, 2000; Christin *et al.*, 2007) (Figure 4.1). The bioinformatic analysis was blind to structural and functional information so although many residues that greatly influence C_4 PEPC kinetics are contained within the PEPC sequences that have been included in the analysis, residues have been identified irrespective of proximity to functional sites. An advantage of this approach, and of targeting all 21 sites under selection in C_4 PEPC for mutagenesis, is that these residues are being acted on by natural selection in multiple C_4 plants even if it is not obvious why this is the case from structural analysis, so it is possible that the screen in *E. coli* will identify residues that were previously unknown to influence enzyme kinetics.

By focussing on the 21 sites under selection in C_4 PEPC, and aiming to identify single amino acid substitutions that that confer a competitive advantage to 6His::PEPC in a bacterial screen, the resulting variants can be tested *in planta* through a GE approach (Weiss *et al.*, 2020) as opposed to requiring genetic modification of the plant. GE crop plants are already permitted in Asia, Australasia and the Americas, and the law has recently changed to accommodate growth of GE crops in the U.K. (Turnbull, Lillemo and Hvoslef-Eide, 2021; Sprink, Wilhelm and Hartung, 2022). In countries that have not yet approved GE crops, including in Europe, the GE legislative landscape is either under discussion or appears to be moving in the direction of approving GE crops for cultivation (European Commission, 2021; Turnbull, Lillemo and Hvoslef-Eide, 2021). By contrast, GM crop plants continue to be met with resistance in Europe and cultivation is de facto banned in 15 EU member states including France and Germany (Halford, 2019). In order for the results obtained in this thesis to be translated into practical application in C_4 crop plants it is important to design the mutagenesis approach such that variants that perform well can be tested *in planta via* GE methods. Targeting only the sites under selection in C_4 PEPC for mutagenesis will ensure that characterisation of promising mutations *in planta* can be achieved by GE technology which should ensure that crop plants engineered with putative improved C_4 PEPC meet the least resistance to commercialisation around the world.

4.1.2 Choice of mutagenesis technique to create *6His::PEPC* variant libraries

There are multiple ways of performing site saturation mutagenesis to generate libraries of enzyme variants. Although the methods vary, the principle is similar: a set of mutagenic oligonucleotides designed to encode every possible amino acid at a specific site are incorporated into the gene of interest to produce a set of coding

sequence variants. Early technologies include cassette mutagenesis where mutagenic DNA cassettes up to 25 base pairs long are integrated into a gene (Wells, Vasser and Powers, 1985) or PCR-based mutagenesis methods including iPCR (Stemmer and Morris, 1992) and commercial site-directed mutagenesis kits, for example QuikChange (Aligen) or Q5® Site-Directed Mutagenesis Kit (NEB). Cassette or PCR-based mutagenesis methods are efficient for mutagenesis at a single site or sites in close proximity that can be encoded on the same cassette or oligonucleotide, but mutating multiple distant sites on the same sequence either requires multiple rounds of mutagenesis, or are achieved in one step but at low efficiencies (Hogrefe *et al.*, 2002).

New multi-site saturation mutagenesis methods have been developed that facilitate quick, highly efficient mutagenesis of one or more sites including Golden Mutagenesis (Püllmann *et al.*, 2019), Darwin Assembly (Cozens and Pinheiro, 2018) and OmniChange (Dennig *et al.*, 2011). All three techniques are based on the incorporation of mutagenic oligonucleotides similar to previous PCR-based mutagenesis methods but can achieve mutation of multiple sites simultaneously with higher efficiencies. Golden Mutagenesis and OmniChange can facilitate mutation of up to five sites simultaneously (Dennig *et al.*, 2011; Püllmann *et al.*, 2019), while Darwin Assembly has been employed to achieve mutagenesis of KOD DNA polymerase at 19 codons simultaneously (Cozens and Pinheiro, 2018). Therefore, Darwin Assembly was selected as the mutagenesis method of choice here as it is quick, versatile and can be rapidly scaled up to target a large number of sites simultaneously (Cozens and Pinheiro, 2018). Further, it has been used to simultaneously mutate 8 sites on a gene of a similar size to PEPC: the TgoT DNA polymerase gene is 2.4 kb (Cozens and Pinheiro, 2018) and *6His::PEPC* is 2.9 kb.

Mutagenesis by Darwin Assembly is achieved through incorporation of a pool of mutagenic oligonucleotides into the target gene to create a library of coding sequence variants. First, a ssDNA template is generated when one of the two DNA strands in a plasmid containing the target gene is nicked and digested (Figure 4.2). Mutagenic ‘inner’ oligonucleotides anneal to the ssDNA template either side of the codon chosen for mutagenesis and encode a mutation at the codon (Figure 4.2). Boundary oligonucleotides anneal to the template at either end of the target gene (Figure 4.2). The full-length coding sequence of the target gene is then reconstituted during an isothermal assembly step when the 3’ end of each oligonucleotide becomes an initiation site for Phusion DNA polymerase to synthesise the complementary strand of DNA to the ssDNA template and Taq DNA ligase joins each fragment together, during which time the mutagenic oligonucleotides that have annealed become incorporated into variants of the gene (Figure 4.2). The isothermal assembly reaction is shown in more detail in Figure 4.3.

The 5’ boundary oligonucleotide has a biotin label, so any isothermal assembly products where the 5’ boundary oligonucleotide has been incorporated can be easily isolated from the mixture of half-products, dNTPs, enzymes and ssDNA template that are not biotinylated using streptavidin bead purification (Figure 4.2). Subsequent amplification of the enzyme variants using ‘outnest’ oligonucleotides that anneal only to the boundary oligonucleotides that were incorporated on either side of the target gene ensure that only full-length isothermal assembly products become amplified. In addition, restriction endonuclease sites can be incorporated into these ‘outnest’ oligonucleotides which facilitates easy downstream cloning (Figure 4.2).

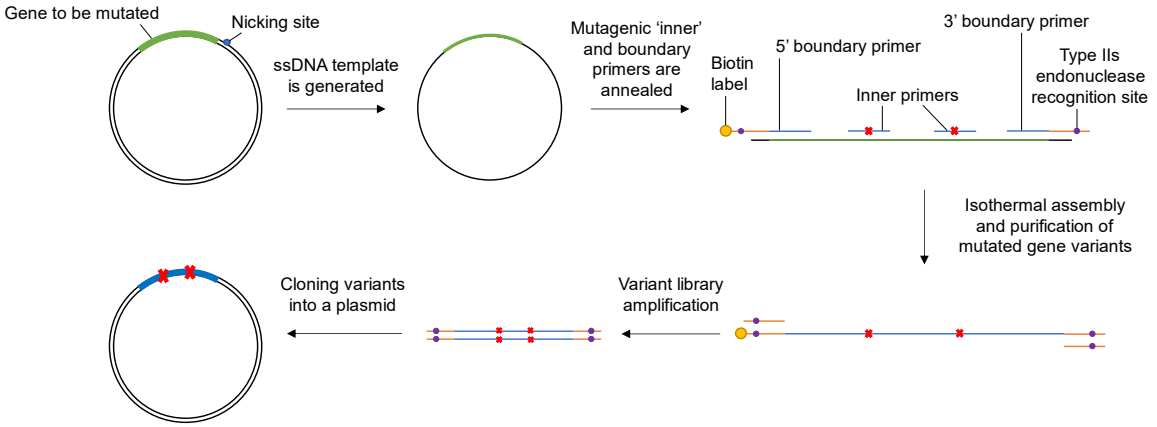


Figure 4.2: Darwin Assembly generates libraries of enzyme variants that can be cloned into a vector. One strand of plasmid DNA is nicked and digested to generate a ssDNA template to which oligonucleotides can anneal. Boundary oligonucleotides (orange and blue) anneal to the 5' and 3' ends of the gene of interest and both boundary oligonucleotides have a tag that contains a Type IIS endonuclease recognition site (purple) for downstream cloning steps. The 5' boundary oligonucleotide also has a 5' biotin label (yellow) for downstream purification. Mutations are introduced when a mutagenic inner oligonucleotide anneals to the template DNA either side of the codon to be mutated. Boundary and inner oligonucleotides anneal to the ssDNA template, then the action of a DNA polymerase and a DNA ligase incorporates the oligonucleotides into the full-length gene in an isothermal assembly step. The biotinylated isothermal assembly products are bound to streptavidin-coated beads and purified by washing away the template and other non-biotinylated DNA. The biotinylated product that remains acts as a template for an outnest PCR step using oligonucleotides that anneal to the ends of the isothermal assembly product (orange). This amplifies a variant library that can be easily cloned into a vector using the Type IIS recognition sites.

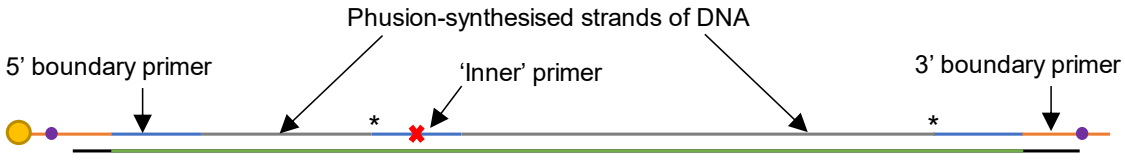


Figure 4.3: Isothermal assembly results in mutagenic inner oligonucleotides becoming incorporated into the *6His::PEPC* gene. The ssDNA template is represented in green. A biotin label is denoted by a yellow circle and Type IIS restriction endonuclease recognition sites are denoted by small purple circles. Taq DNA ligase catalyses formation of phosphodiester bonds (*) between the end of the newly synthesised DNA strand (grey) and the adjacent oligonucleotide (blue).

Darwin Assembly site saturation mutagenesis can be used to mutate any codon along the length of target gene. For each site, a set of 20 inner oligonucleotides that between them encode every amino acid at the position of the codon to be mutated are added to the ssDNA template (Figure 4.4). During Darwin Assembly, the mutagenic oligonucleotides anneal to target gene template and in doing so subsequently become incorporated into the full-length versions of the target gene. This creates a library of up to 20 enzyme variants, each with a different amino acid at the site of mutagenesis (Figure 4.4). When site saturation mutagenesis is undertaken at multiple sites simultaneously, the size of the variant library that can be achieved increases rapidly to thousands or millions of variants (Table 4.1). These large enzyme variant libraries have been shown to be achievable. For example, mutagenesis of eight residues of TgoT DNA polymerase resulted in 2.25×10^8 transformants mutated across all of the residues targeted with a maximum of 1.95% wild-type sequence contamination at any given site (Cozens and Pinheiro, 2018).

The 21 sites under selection in C_4 PEPC will be targeted for site saturation mutagenesis using Darwin Assembly (Figure 4.2). Using this recently published mutagenesis approach that has not been used to generate PEPC variant libraries before will require careful optimisation of the protocol. First, the protocol will be optimised for site saturation mutagenesis at a single site and then this will be repeated for all other sites under selection. This will result in a set of 21 libraries that between them encode up to 420 6His::PEPC variants that can be taken forward to the next step of the screen for increased enzymatic rate of 6His::PEPC. Once the screen of 6His::PEPC variants in *E. coli* has been established, then the scope for each screen can be increased to include variants produced by multi-site saturation mutagenesis.

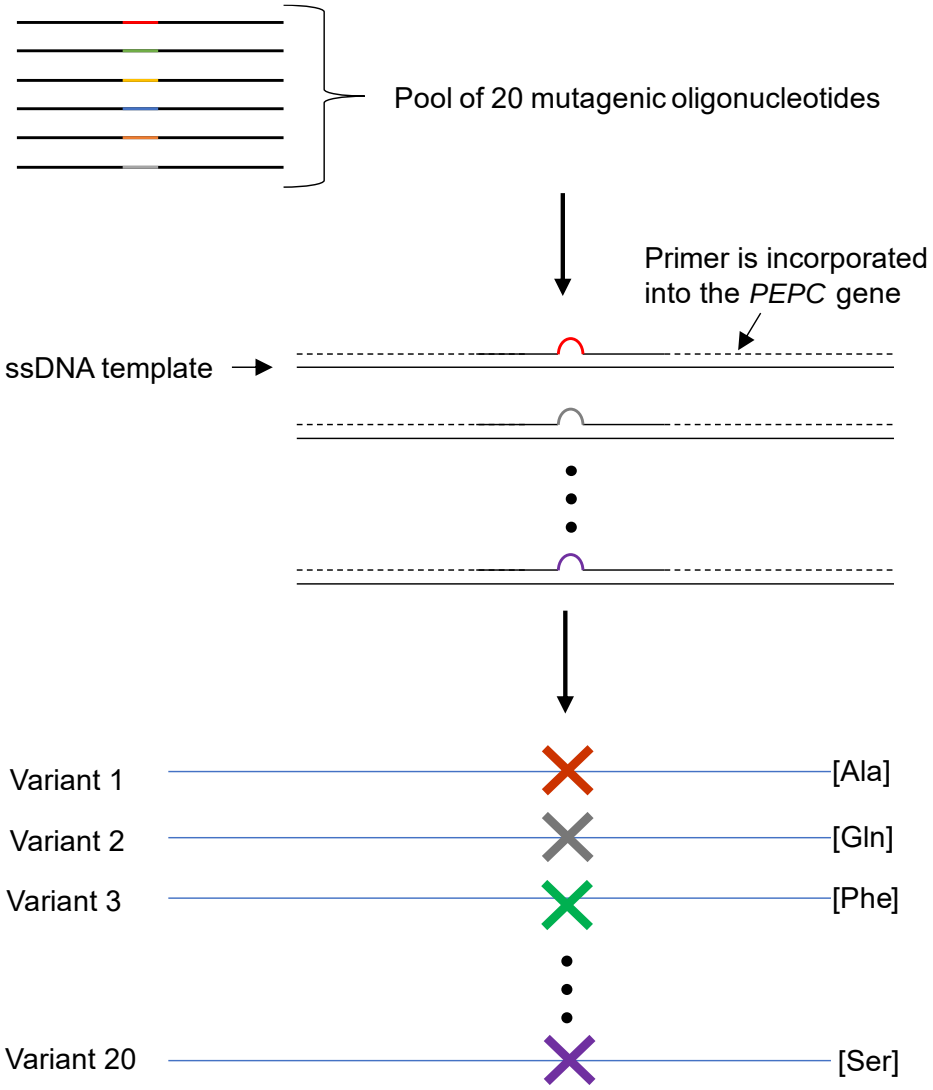


Figure 4.4: The integration of mutagenic oligonucleotides into the full-length *6His::PEPC* gene should result in the generation of enzyme variant libraries. Oligonucleotide sequences that anneal to the ssDNA template are black and different colours at the site of mutagenesis indicate different amino acids. The part of the oligonucleotide that does not anneal to the template is indicated by a loop. Crosses in different colours indicate different amino acids have been incorporated at the site of mutagenesis.

Table 4.1: The number of possible enzyme variants in a variant library increases with the number of sites that are mutated simultaneously.

Number of sites mutated simultaneously	Theoretical size of enzyme variant library
1	20
2	400
3	8,000
5	3,200,000
10	1.024×10^{13}

4.1.3 Chapter aims and outline

There were two main aims that underpinned the work described in this chapter. The first aim was to identify and implement a suitable mutagenesis approach to generate libraries of *6His::PEPC* variants. The second aim was to optimise the transformation of these libraries into an *E. coli* strain in which the endogenous *PEPC* gene has been deleted (*E. coli* Δppc). In combination, the completion of these two aims will enable the high throughput screening of 6His::PEPC variants to identify and prioritise amino acid changes for engineering into plants.

4.2 Results and Discussion

4.2.1 There is substantial variation between C_4 species at sites that have evolved under positive selection in PEPC.

The sites that were detected as under selection in C_4 PEPC are located throughout the 3D structure of the enzyme, with the majority located either in or near to the β barrel, or on the surface of the enzyme closest to the β barrel (Figure 4.5). None of the sites under selection in C_4 PEPC are known functional residues involved in substrate, inhibitor or activator binding, however, some sites have been identified as important in C_4 enzyme kinetics. For example, site 733 is thought to influence PEP kinetics in the Suaedoideae subfamily (Rosnow, Edwards and Roalson, 2014) and site 780 has been shown to be influential in PEP and bicarbonate binding and is thought to be required for C_4 PEPC kinetics, namely a higher affinity for bicarbonate and a lower affinity for PEP (Bläsing, Westhoff and Svensson, 2000; Engelmann *et al.*, 2003; Westhoff and Gowik, 2004; DiMario and Cousins, 2019). Moreover, the majority of sites are within region 528 – 880 which contributes to the affinity of C_4 PEPC for bicarbonate in the Poaceae family, and within region 652 – 970 which is involved in PEP and malate binding (Table 4.2) (Bläsing, Westhoff and Svensson, 2000; Jacobs *et al.*, 2008; DiMario *et al.*, 2021). Thus, although the contribution of the majority of sites under selection to C_4 PEPC kinetics has not been previously characterised, most are in a region of the enzyme known to influence the affinity of the enzyme for substrates PEP and bicarbonate and the inhibitor malate and are therefore well-placed to confer changes to enzyme kinetics.

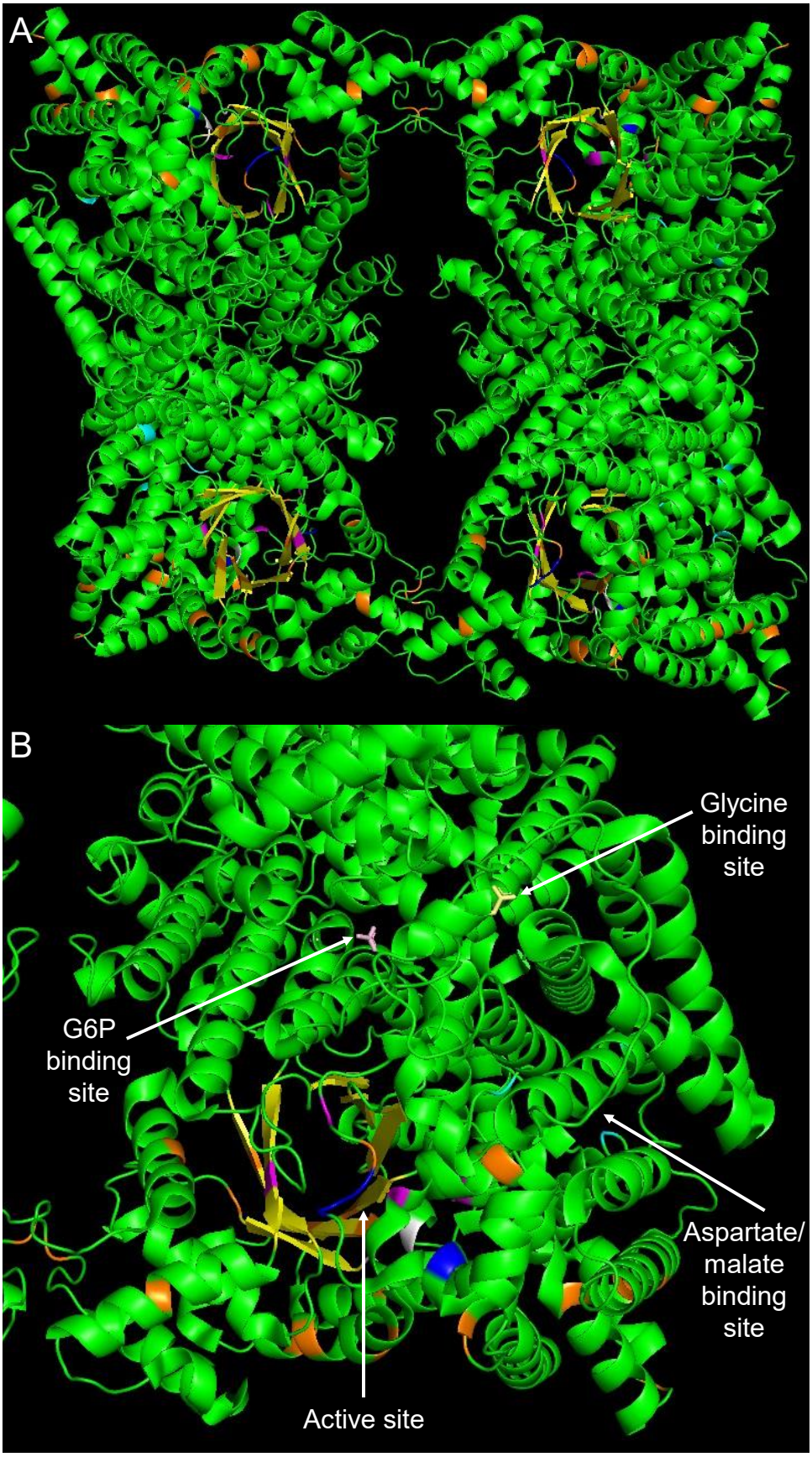


Figure 4.5: The locations of sites under selection in C_4 PEPC. Figure legend continued on the next page.

Figure 4.5 (continued) (A) PEPC is tetrameric in structure composed of α-helices (green) and β-strands (yellow) that form a β barrel. Sites under selection are coloured in orange. PEP binding residues coloured in pink, magnesium binding residues in grey, bicarbonate binding residues in dark blue and aspartate/malate binding residues in light blue. (B) The majority of the sites under selection are located on the part of the enzyme closest to the β barrel. Acetate coloured in light pink is located at the putative glucose 6-phosphate binding site and glycine coloured in light green is located at the glycine binding site. All other colours as in part A. PDB ID: 5VYJ.

Table 4.2: Sites under selection described for C_3 PEPCs, C_4 PEPCs, *Z. mays* C_4 PEPC and SvPEPC.

	Site number (aligned to *Z. mays* C_4 PEPC)	Amino acids found in C_3 PEPCs	Amino acids found in C_4 PEPCs	Amino acid at the residue in *Z. mays* C_4 PEPC	Amino acid at the equivalent residue in SvPEPC (SvPEPC codon numbers)
1	466	L	IV	I	I (460)
2	502	G	APSTV	P	S (496)
3	517	T	AC	A	C (511)
4	531	A	P	P	P (525)
5	560	R	PH	P	P (554)
6	572	EVI	Q	Q	Q (566)
7	577	A	ST	S	S (571)
8	579	A	E	E	E (573)
9	596	E	QMR	Q	Q (590)
10	599	IT	V	V	V (593)
11	625	V	ASQGWL	A	A (619)
12	637	M	FLI	L	M (631)
13	665	HL	NR	N	N (659)
14	733	FL	VM	V	M (727)
15	761	S	A	A	A (755)
16	780	A	S	S	S (774)
17	794	F	VI	V	F (788)
18	807	R	KN	R	K (801)
19	813	R	K	K	K (807)
20	863	NEDS	K	K	N (857)
21	866	D	EQ	E	E (860)

Table footnote available on the following page.

Table 4.2 (footnote) This table collates the changes described in Christin et al., 2007 between PEPC sequences in C_3 (ppc-B2) and C_4 (ppc-C_4) groups and displays the equivalent amino acid at each site in C_4 PEPCs in *Z. mays* and *Setaria viridis* for comparison. Cells highlighted in green in the *Z. mays* C_4 PEPC column are residues that differ between *Z. mays* and *S. viridis* C_4 PEPC. Cells highlighted in grey in the SvPEPC column are residues that have the same amino acid in SvPEPC as C_3 PEPCs.

Of the 21 sites under selection, most is known about the influence of site 780 on C_4 PEPC kinetics. Site 780 is located close to the active site of the enzyme (Figure 4.6) and it has been established that the change from alanine to serine at site 780 in C_4 PEPC is a pre-requisite for C_4 PEPC kinetics (Bläsing, Westhoff and Svensson, 2000) so it is axiomatic that it has been identified as a site under selection in C_4 PEPC. When the opposite S780A mutation was introduced to C_4 *F. trinervia* PEPC, the enzyme had a reduced affinity for bicarbonate and increased affinity for PEP, demonstrating a transition towards typical C_3 PEPC (Bläsing, Westhoff and Svensson, 2000; DiMario and Cousins, 2019). However, serine at site 780 in C_4 PEPC is not sufficient to solely account for PEP and bicarbonate kinetics of C_4 PEPC, so it is possible that other sites under selection are also contributing to the altered kinetics (Bläsing, Westhoff and Svensson, 2000; Engelmann, Bläsing and Westhoff, 2002; DiMario and Cousins, 2019).

Near to site 780, site 761 is located next to bicarbonate binding residues on a mobile loop that is essential for catalysis and is located only 5.3 Å away from a PEP binding residue on an adjacent alpha helix (Figure 4.6). It has been suggested that the serine to alanine substitution at site 761 maintains a putative N-myristylation site while precluding the possibility of a casein kinase II phosphorylation site, and that the availability of these sites influences the regulation of C_4 PEPC, however, this has not been formally tested (Christin *et al.*, 2007). The significance of repeated amino acid substitution to alanine at site 761 as PEPC is co-opted into C_4 photosynthesis remains uncharacterised. However, the close proximity to substrate binding residues concomitant with a change in polarity as the residue changes from a serine to an alanine residue in C_4 PEPC could be influencing the changes in affinity of PEPC for substrates bicarbonate and PEP in C_4 PEPC.

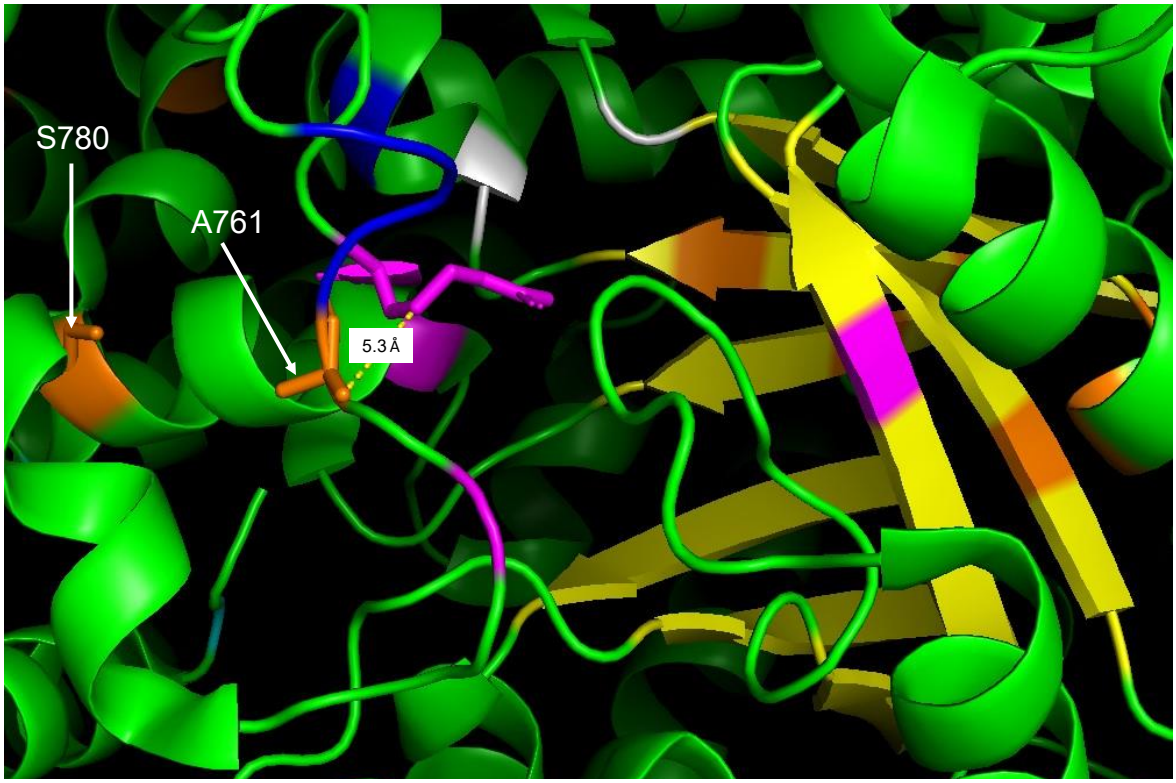


Figure 4.6: Residue 761 is located adjacent to substrate binding residues and residue 780 is near to the catalytic site of the enzyme. β barrel represented in yellow, PEP binding residues coloured in pink, magnesium binding residues in grey, bicarbonate binding residues in dark blue, aspartate/malate binding residues in light blue and sites under selection in orange. PDB: 5VYJ.

The three sites 531, 560 and 596 are also located near to the active site of the enzyme and close to one another in or near the β barrel. The R group in $RCH(NH_3^+)CO_2^-$ of three amino acids all face towards the centre of the β barrel and are very close in proximity, within 6 Å of one another in the structure of glycine activated *Z. mays* C_4 PEPC (Figure 4.7). It is surprising that proline has been repeatedly selected for in place of alanine at site 531 given that proline residues have restricted flexibility and cannot complete the hydrogen bonding network and therefore are not common in β sheets (Bhattacharjee and Biswas, 2010). Conversely, the repeated replacement of arginine with proline at site 560 is less surprising in the disordered region at the N terminus of the β barrel where there is more flexibility and proline appears to induce a turn (Figure 4.7). Finally, the change from glutamic acid to glutamine at site 596 results in the loss of a negative charge within the β barrel which could increase the affinity of the enzyme for bicarbonate. Given the proximity of these residues to the active site of the enzyme and to one another, the residues are well-positioned to influence PEPC kinetics and it is tempting to speculate that one or more or the amino acids could be coevolving.

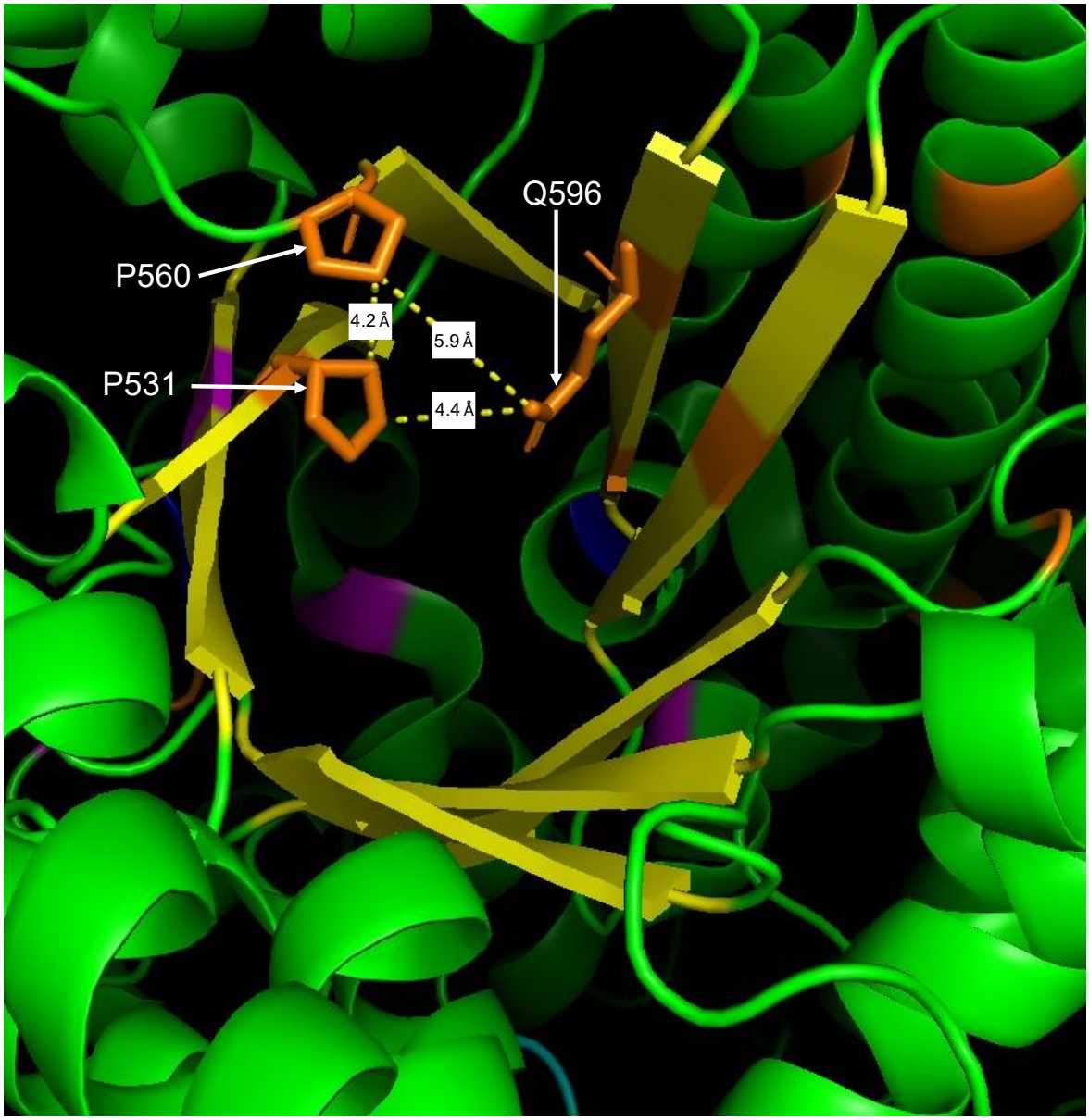


Figure 4.7: Sites 531, 560 and 596 are located in close proximity towards the N terminus of the β barrel. β barrel represented in yellow, PEP binding residues coloured in pink, bicarbonate binding residues in dark blue, aspartate/malate binding residues in light blue and sites under selection in orange. PDB: 5VYJ.

Another set of sites under selection that are all in a similar region of PEPC are 733, 794, 807, 813, 863 and 866 (Figure 4.8 A). These sites are situated mostly on the outside of the enzyme in a region between the active site and the aspartate/malate binding pocket (Figure 4.8 B). The biochemistry of the amino acids at these sites has remained largely similar as the enzyme has been co-opted into the C_4 photosynthesis pathway. However, a notable change has occurred at site 863 where selection for positively charged lysine in C_4 PEPC, in place of polar, uncharged or negatively charged amino acids in C_3 PEPC, enables the formation of a salt bridge with E727 on an adjacent surface alpha helix (Figure 4.8 A). Salt bridge formation could result in conformational changes to the enzyme that alters the affinity of the enzyme for substrate, activators or inhibitors, or influences enzyme stability. Interestingly, although *Z. mays* C_4 PEPC has evolved a lysine at position 863, *S. viridis* has an uncharged asparagine which is more typical of C_3 PEPC. The influence of these sites on PEPC kinetics is challenging to predict from the 3D structure of the enzyme.

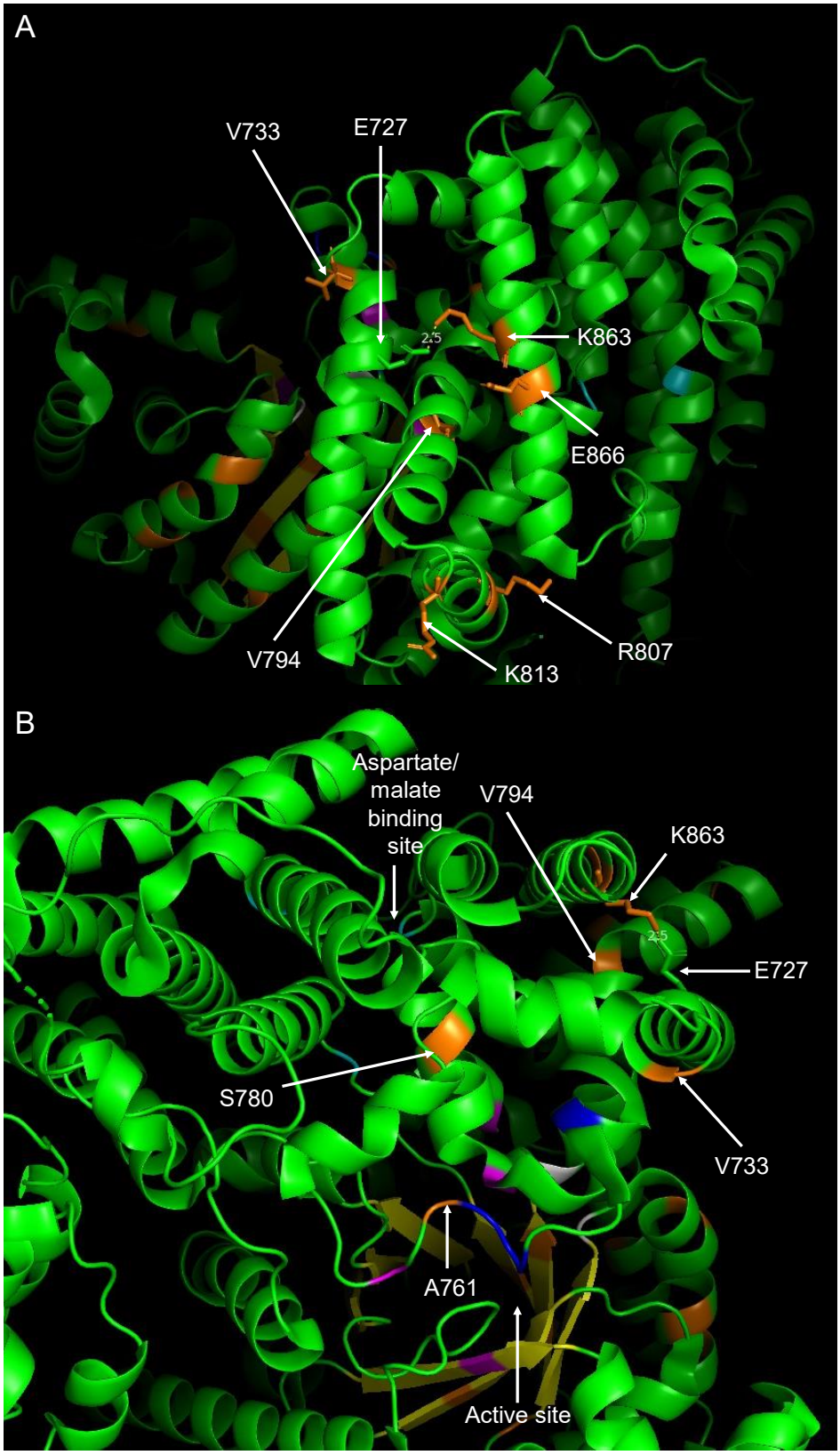


Figure 4.8: Sites 733, 794, 807, 813, 863 and 866 are situated between the PEPC active site and aspartate/malate binding site. Figure legend continued on the next page.

Figure 4.8 (continued) (A) K863 can form a salt bridge with E727 (2.5 Å distance shown as yellow dotted lines). (B) The sites under selection are shown in relation to the active site and aspartate/malate binding pocket, but sites 807, 813 and 866 are not visible in this orientation. Sites 761 and 780 are also shown. β barrel represented in yellow, PEP binding residues coloured in pink, magnesium binding residues in grey, bicarbonate binding residues in dark blue, aspartate/malate binding residues in light blue and sites under selection in orange. PDB: 5VYJ.

Aforementioned sites 794 and 863 are two of only three sites under selection in SvPEPC that do not have an amino acid that has been typically selected for in C_4 PEPC and instead the amino acid at these residues is more typical of C_3 PEPC (Table 4.2). The third residue is site 637 which has converged from methionine to leucine, isoleucine or phenylalanine in C_4 PEPCs, but is methionine in SvPEPC. The significance of residues 637, 794 and 863 retaining an amino acid residue typical of C_3 PEPC in SvPEPC is unknown.

Site 637 is also in a cluster of residues in a similar region of the enzyme that includes 572, 577, 579, 599, 625 and 665 (Figure 4.9). Aside from site 637, which has converged on amino acids with similar properties in C_4 PEPC compared to C_3 PEPC, most of these residues have undergone changes in charge, polarity and/or size as the C_4 residues have evolved. These biochemical changes could confer an allosteric effect on enzyme kinetics given that none of the residues are known to be involved in substrate or effector binding. In accordance with this theory, the transition from histidine to asparagine at site 665 has been suggested to be involved in protein folding and could be conferring a structural change to the enzyme that alters kinetics (Christin *et al.*, 2007). Additionally, some sites have diverse biochemical properties amongst C_3 PEPCs, for example site 572, or amongst C_4 PEPCs, for example site 625 (Table 4.2). It would be interesting to understand the contribution of each of the sites in this group to C_4 PEPC kinetics and the significance of differences in amino acid biochemistry between C_4 species at sites such as 625.

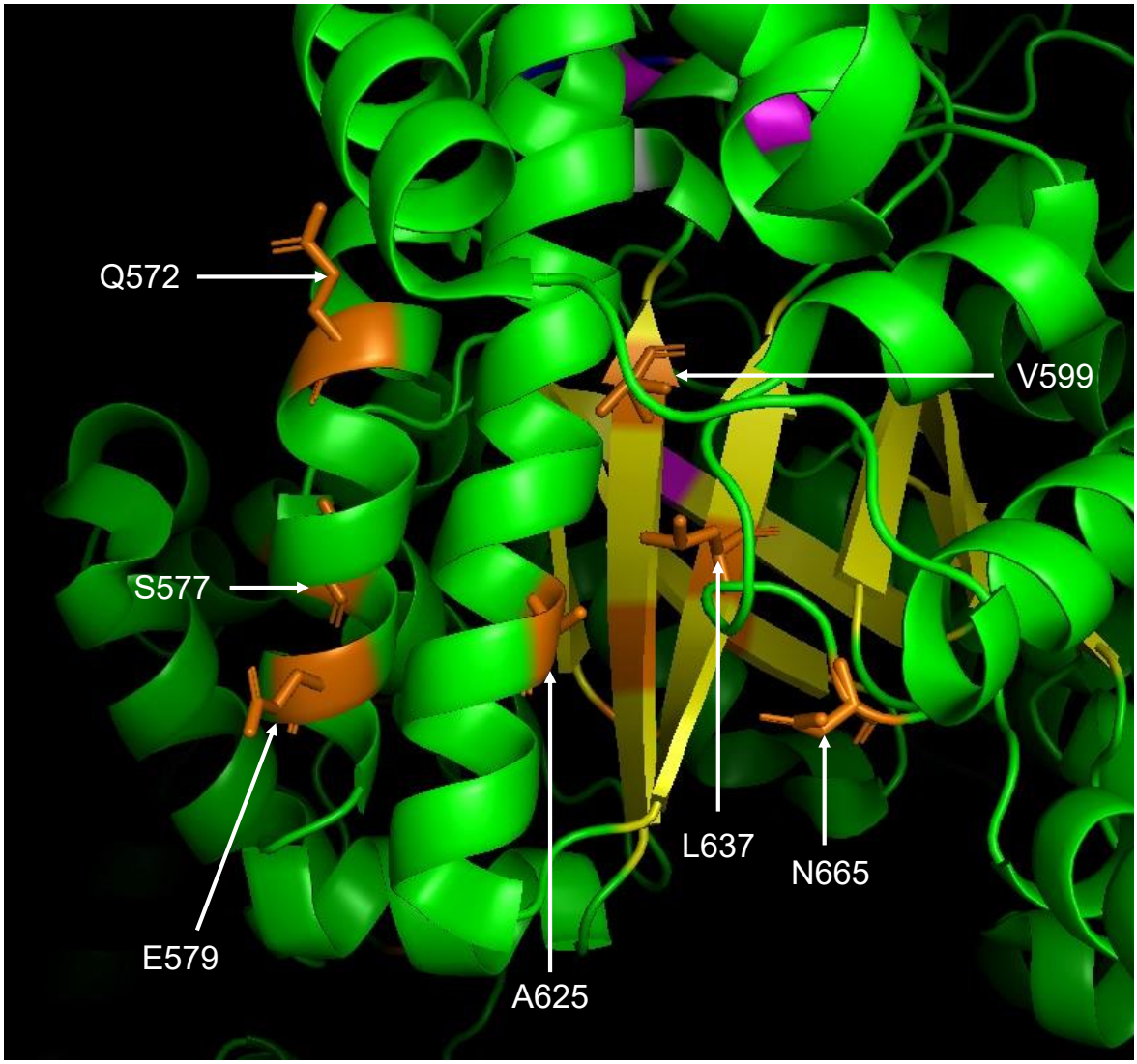


Figure 4.9: Sites 572, 577, 579, 599, 625, 637 and 665 are close to each other on adjacent α-helices and β-strands. β barrel represented in yellow, PEP binding residues coloured in pink, magnesium binding residues in grey and sites under selection in orange. PDB: 5VYJ.

A final cluster of sites, 466, 502 and 517, are situated between the active site and the interface between two monomers (Figure 4.10). Site 502 is closest to this interface, on a mobile loop near to residues E493 and R498 that are involved in tetramer formation (Figure 4.10). The evolution of site 502 from glycine in C_3 PEPCs to proline in *Z. mays* C_4 PEPC is a significant shift in terms of the movement constraints of the mobile loop which could be having an impact on the tetramer assembly of the enzyme. Conversely, site 466 has evolved to have a very similar amino acid in C_4 PEPCs compared to C_3 PEPCs and site 517 has evolved from a threonine to a similar polar amino acid cysteine in some C_4 species including *S. viridis* but to non-polar alanine in others including *Z. mays* (Table 4.2). In summary, analysis of the local 3D environment of each site that has evolved under positive selection in PEPC does not give clear or direct evidence for their role in the kinetic evolution of the enzyme. However, many are near the active site of the enzyme or could be structurally significant, and the repeated evolution of the sites under selection in different C_4 lineages concomitant with changes to C_4 PEPC kinetics suggests that some or all of these sites have been influential in the adaptation of PEPC to the C_4 photosynthesis pathway.

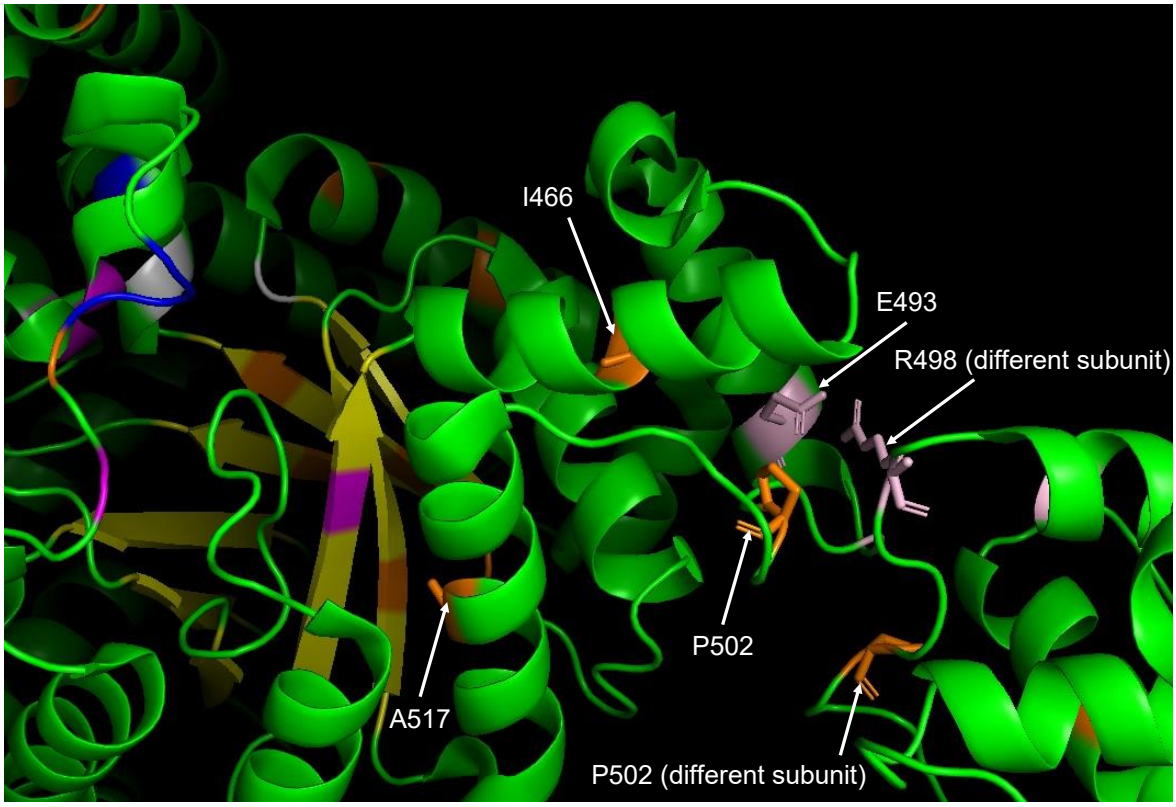


Figure 4.10: Sites 466, 502 and 517 are near the interface of two PEPC monomers. β barrel represented in yellow, PEP binding residues coloured in pink, magnesium binding residues in grey, bicarbonate binding residues in dark blue and sites under selection in orange. PDB: 5VYJ.

4.2.2 Generation of *6His::PEPC* variant libraries by Darwin Assembly mutagenesis

Given that the analysis of positive selection was conducted on less than 1% of C_4 species, the full extent of sequence variation at any particular site under selection in all C_4 species is unknown. Thus, it was decided to perform site saturating mutagenesis at each site that was identified as evolving under positive selection. This approach was taken so that all amino acid variants could be tested at each site, including variants that are difficult to evolve as they require multi-nucleotide mutations, or transitions *via* potentially deleterious amino acid residues, and thus may not have evolved in nature (Niklaus and Kelly, 2019).

4.2.2.1 Generation of the ssDNA template for mutagenesis

The *6His::PEPC* gene (Chapter 3.2.2.1) in the plasmid pUC57 was used to generate the template for Darwin Assembly. The first step of Darwin Assembly is to nick and digest a strand of the plasmid DNA to generate a ssDNA template for oligonucleotide annealing (Figure 4.2). When *6His::PEPC* was cloned into pUC57, a recognition site for the nicking endonuclease Nt.BbvCI was included downstream of the 6His::PEPC CDS. Therefore, on incubation with Nt.BbvCI a single nick should be made to one strand of the plasmid DNA at the recognition site downstream of the 6His::PEPC CDS and only the DNA of the nicked strand digested away, resulting in a ssDNA template for subsequent steps of Darwin Assembly. The ssDNA generation step had to be optimised for a different nicking endonuclease and different plasmid to the published Darwin Assembly protocol due to sequence incompatibilities that could not be avoided.

In the Darwin Assembly protocol 3 – 10 U/µg nicking endonuclease and 40 – 60 U/µg ExoIII are recommended for nicking the DNA and digesting away the nicked

strand, respectively. ExoIII catalyses the digestion of nucleotides initiated from a nicked site in double-stranded DNA (dsDNA) (Sayers, Schmidt and Eckstein, 1988). Thus, incubation of the plasmid dsDNA with Nt.BbvCI should result in one strand of plasmid DNA becoming nicked at the BbvCI recognition site, then ExoIII should digest the nicked strand completely and the strand that is not nicked will remain intact and will become the ssDNA template.

The reaction conditions, including the concentration of plasmid, buffer and concentration of each enzyme used, needed to be optimised in order to achieve nicking and digestion activity that resulted in ssDNA. Ten U/µg Nt.BbvCI consistently nicked the plasmid DNA, but the quantity of ExoIII was varied to find the optimal concentration. The optimal reaction conditions were 1 µg 5.8 kb plasmid DNA incubated with 10 U/µg Nt.BbvCI and 20 – 40 U/µg ExoIII in NEBuffer3.1 at 37 °C for 2 hours then 80 °C for 20 minutes to inactivate the enzymes, in a 10 µL reaction volume (Figure 4.11 A). Twenty U/µg ExoIII was used in all future experiments to keep the concentration of glycerol in each reaction to a minimum. These reaction conditions were consistently successful in reactions of the same volume and when the reaction was scaled up to 50 µL (Figure 4.11 B). ssDNA has higher mobility than nicked or linearised DNA during gel electrophoresis, so these conditions were loaded onto each gel alongside for comparison by incubating the plasmid dsDNA with Nt.BbvCI only or BamHI, respectively.

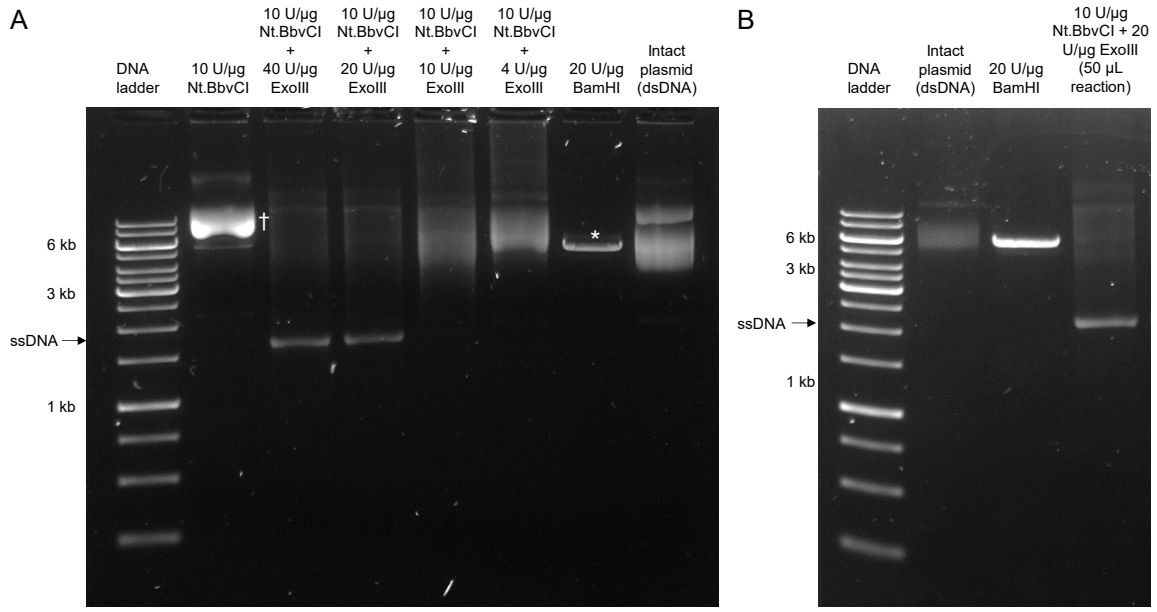


Figure 4.11: An ssDNA template was generated for Darwin Assembly mutagenesis of *6His::PEPC*. A range of ExoIII concentrations were tested to determine the optimal conditions to generate a ssDNA template from *6His::PEPC* in pUC57 plasmid. One strand of the plasmid was cut (nicked) by Nt.BbvCI and the nicking site became the initiation point for digestion of the nicked strand by ExoIII. (A) 10 U/µg Nt.BbvCI and 20 – 40 U/µg ExoIII were optimal for ssDNA generation in 10 µL reactions. Plasmid that had been nicked by Nt.BbvCI (†), linearised by BamHI (*) and the plasmid incubated without enzymes were used as comparisons to determine the activity of the nicking enzyme, size of the linearised plasmid and the appearance of the undigested plasmid DNA on the gel, respectively. Single stranded plasmid DNA was identifiable as it ran furthest on a gel compared to nicked, linear and supercoiled DNA. (B) ssDNA was generated successfully in five reactions of 10 U/µg Nt.BbvCI and 20 U/µg ExoIII that had been scaled up to 50 µL, one of which is shown.

4.2.2.2 Oligonucleotide design

4.2.2.2.1 Inner oligonucleotide design

The Darwin Assembly protocol introduces mutations to the gene of interest at a specific amino acid residue through the addition of a library of mutagenic inner oligonucleotides to the ssDNA template. The success of this mutagenesis approach therefore depends on suitable mutagenic oligonucleotide design. Once the sites under selection identified in Christin et al. (2007) had been mapped to the 6His::PEPC sequence used in this study, mutagenic oligonucleotides were designed to anneal to the *6His::PEPC* gene (Table 4.3). At the site of mutagenesis a 'small-intelligent' library design was implemented where each amino acid is equally represented in the oligonucleotide set for a specific residue in an attempt to ensure that every amino acid would be equally represented at that position in the final variant library (Tang *et al.*, 2012). To achieve this, the oligonucleotides that were designed for every amino acid residue were ordered with "NDT", "VMA", "ATG" and "TGG" in place of the wild-type codon, where N=A, C, G or T, D=A, G or T, V=A, C or G and M=A or C (a full list of mutagenic primers is available in Appendix 3). These combinations of bases encoded all the amino acids on the basis of one codon per amino acid. Accordingly, the oligonucleotides were mixed in a 12:6:1:1 molar ratio prior to oligonucleotide annealing in order to achieve equal representation of codons encoding every amino acid in the pool of oligonucleotides (Figure 4.4). This method avoids stop codons, is optimised for *E. coli* codon usage and facilitated library generation that was not biased towards amino acids that are encoded most frequently in the genetic code.

Mutagenic inner oligonucleotides were designed to anneal to the 6His::PEPC CDS either side of an amino acid residue under selection in C_4 PEPCs in Christin et al.,

2007. Oligonucleotides used in this study were between 17 and 28 nucleotides in length and had 40-60% GC content, acceptable levels of self-dimerisation (weaker than ΔG -9 kcal/mol) and melting temperatures (T_m) between 58 and 63 °C as far as possible, assuming no mismatches (Table 4.3). This was in-keeping with the published protocol which was successful using primers with an annealing temperature between 55-60 °C, approximately equivalent to a T_m between 60-65 °C (Cozens and Pinheiro, 2018). The most challenging aspect of mutagenic oligonucleotide design was the requirement to bind exactly to the sequence adjacent to the codon to be mutated because these sequences could have high GC content, be prone to self-dimerisation or have a high T_m even as a short sequence. Accordingly, compromises had to be made that were inevitably likely to affect library assembly.

As an example of the multiple trade-offs involved in inner primer design, the codon optimised *6His::PEPC* sequence either side of site 637 contained a repetitive A sequence on one side of the codon and a GC rich sequence on the other side (Figure 4.12). Extension of the oligonucleotide by 11 nucleotides on either side, as the Darwin Assembly protocol recommends, resulted in a oligonucleotide that has a high T_m (65 °C), a higher T_m for sequences with NDT and VMA at the position of the codon (62-66 °C) and self-dimerisation of -16.5 kcal/mol, far lower than Integrated DNA Technologies-recommended minimum -9 kcal/mol (IDT, 2023), making this oligonucleotide unworkable. The oligonucleotide that was used for mutagenesis had 13 nucleotides before and 8 nucleotides after the methionine codon that was the mutagenesis target, resulting in a 24-nucleotide long oligonucleotide with a T_m of 60 °C and a self-dimerisation predication of -5.36 kcal/mol assuming no mismatches, within the ranges of acceptable specifications

(Figure 4.12). In future work, codon optimisation could be considered concurrently with mutagenic primer design, or the *6His::PEPC* gene sequence could be mutated prior to Darwin Assembly, for example to substitute one codon for an equivalent, frequently used codon in *E. coli* at a specific position in the sequence in order to avoid primer design challenges. In this study, trade-offs between oligonucleotide length, T_m and GC content had to be made because the design of mutagenic oligonucleotides was constrained by the DNA sequence adjacent to each site under selection.

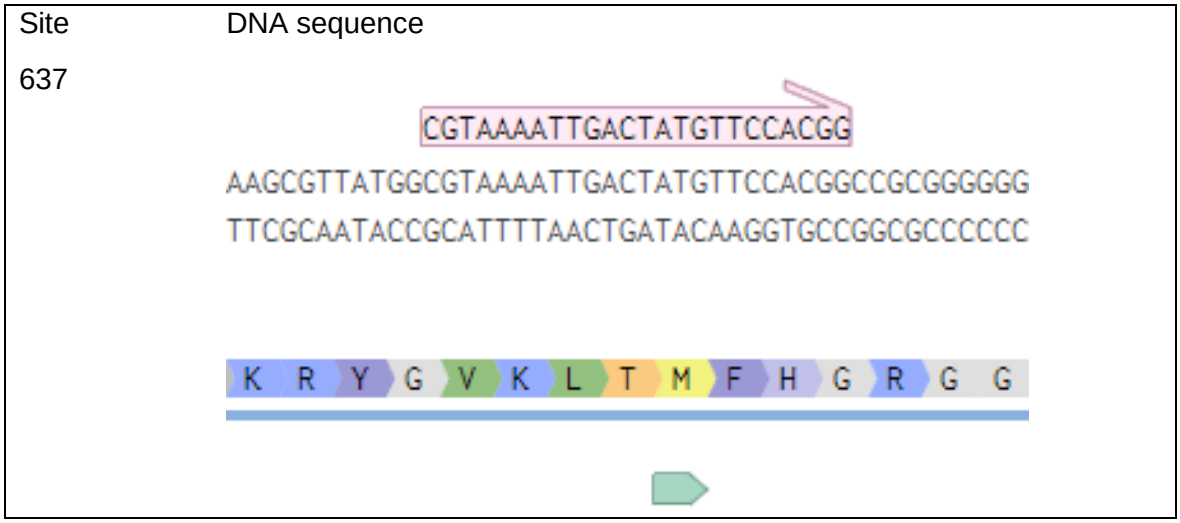


Figure 4.12: An example inner oligonucleotide designed to anneal either side of a site under selection (site 637). The inner oligonucleotide is in pink above the gene sequence. The site under selection that was the target of mutagenesis is denoted with a green arrow beneath the amino acid sequence.

Table 4.3: Primers for mutagenesis of each site under selection.

Site under selection	Primer sequence (assuming no mismatches)
466	CATACTGATGTG**ATT**GACGCTATC
502	GTCCCTTGTTA**TCG**AAAGACATGC
517	GTGCTTGGA**TGT**TTTCACG
531	CGATTCTTTTGGG**CCG**TATATTATCTCG
560	CCCTTA**CCG**GTTGTTCCAC
572	GACTTG**CAG**TCAGCACCTG
577	ACCTGCC**TCA**ATCGAGC
579	CCTCAATC**GAG**CGTTTATTCTC
596	GAAGCAA**CAG**GTGATGGTAGG
599	CAGGTGATG**GTA**GGATATAGTG
625	GAGGAAATG**GCA**AAGGTTGC
637	CGTAAAATTGACT**ATG**TTCCACGG
665	GACACGATC**AAT**GGCTCACTG
733	CTGTCGTC**ATG**CGCGAG
761	CGTCCT**GCC**AAACGCAAG
780	GTGGATTTTC**TCT**TGGACGC
794	TGGCTGGGC**TTT**GGCGCAGCGTTT
807	GAAGGATATT**AAG**AACTTTCAGATGC
813	CAGATGCTT**AAA**GAAATGTATAATGAG
863	CAATTACGCAAC**AAT**TACGTAGAAAC
866	CAATTACGTA**GAA**ACCGAAAAGC

Codons that are targets for mutagenesis are shown in bold.

4.2.2.2.2 Boundary and outnest oligonucleotide design

The 5’ and 3’ boundary oligonucleotides were designed to anneal on either side of the *6His::PEPC* gene, with an additional ‘tag’ sequence that is not complementary to *6His::PEPC* and contains a BpiI (Type IIS) restriction endonuclease recognition site (Table 4.4). The outnest oligonucleotides were designed to anneal to the ‘tag’ sequences in the boundary oligonucleotides that became part of the purified assembly products. In this way, the outnest primers were used in an outnest PCR step to amplify libraries of enzyme variants captured during the isothermal assembly process. The BpiI recognition site was required in both 5’ and 3’ boundary and outnest primers so that as the oligonucleotide was incorporated into the Darwin Assembly product the sites became available for downstream Golden Gate cloning (Figure 4.2). The requirement for the tag sequence to contain a restriction endonuclease site constrained the sequence design and the sequence must also be compatible for annealing of outnest primers that have an appropriate T_m for the subsequent amplification step. In addition, the 3’ boundary oligonucleotide was designed to have an optional 3'-3' dT (3' inverted dT) to protect the oligonucleotide during the isothermal assembly step from degeneration by ExoIII carried over from template generation, and therefore to target improved capture of full length *6His::PEPC* sequences. Together, the boundary and outnest primers were unique to this study and required careful design optimisation in order to generate *6His::PEPC* variant libraries.

Table 4.4: The boundary and outnest primers designed for amplification of *6His::PEPC* variant libraries and introduction of Type IIS restriction endonuclease sites for downstream cloning.

Darwin Assembly primer	Primer sequence (assuming no mismatches)
5' boundary	GGAGAAGACCAAATGCATCATCACCATCACC with a 5’ Biotin-TEG label
5’ outnest	GGACTGGAGAAGACCAAATGCATC
3' boundary	GAATGCAGAACACTGGCTAACCTCAACGCTTTAGTCTTCGG*T 3’ modification is an inverted 3’ base dT
3’ outnest	ACCGAAGACTAAAGCGTTGAGGTTAG

Type IIS restriction endonuclease recognition site highlighted in grey and cut site underlined.

4.2.2.3 Generation of a single site saturation library

The variants that were generated using oligonucleotides that target position 780 were amplified from DNA that was bound to the magnetic beads. This resulted in a 3 kb fragment that was the correct size for the amplified 6His::PEPC gene (Figure 4.13 A). An aliquot of the PCR product was sequenced by Sanger sequencing to get an indication of the extent of variation that was introduced at site 780. The sequencing chromatogram showed that multiple bases were detected at each position in the codon indicating that multiple variants were generated. No other bases were mutated along the length of the sequencing product. Thus, these results demonstrated that targeted mutagenesis to site 780 was successful and that no off target mutations were induced in the flanking regions of the site (Figure 4.13 B).

Amplification of the Darwin Assembly template was scaled up (Figure 4.14 A) and the scaled-up PCR product also showed that mutations had been introduced (Figure 4.14 B). The scaled-up PCR product was gel-extracted from the agarose gel in Figure 4.14 A; yield of the *6His::PEPC* variant library was increased by scaling up the outnest PCR reaction 5-fold, heating the elution buffer to 60 ˚C prior to application on the DNA-binding column and a subsequent ten-minute incubation at room temperature prior to centrifugation to elute, resulting in a gel extraction product concentrated at 53.1 ng/µL in 15 µL elution buffer.

The codons called in place of the wild-type TCT codon were TNT in the small-scale PCR (Figure 4.13 B) and TGT when it had been scaled up (Figure 4.14 B), suggesting that during isothermal assembly oligonucleotides with T at the first or last position in the codon targeted for mutagenesis may have annealed preferentially to the template. The complementarity between the template DNA sequence AGA and the oligonucleotides with T at the first or third position in the

codon at the equivalent position could have made this more likely to occur and could have introduced codon bias towards codons that are similar to the wild-type sequence into the library.

In previous published work, a weak correlation was identified between codon frequency at the site of mutagenesis and the similarity of the codon to the wild-type sequence at the same site in an enzyme variant library generated by Darwin Assembly, but this correlation alone did not account for all the codon bias observed (Cozens and Pinheiro, 2018). The remaining possible causes of codon bias have not been formally identified (Cozens and Pinheiro, 2018), however, it appears that oligonucleotide kinetics (T_m, self-dimerisation and heterodimerisation) can vary considerably between oligonucleotides in a oligonucleotide pool and therefore these factors combined could have contributed to frequency biases in the final library.

In addition to potential biases in the library introduced at the isothermal assembly step, the small differences in the abundance of each base in the codon in the sequencing chromatograms of the PCR reaction product before and after scaling up suggested that the variants that became most abundantly amplified from the isothermal assembly product by chance differ between PCR reactions. There was strong evidence that Darwin Assembly has generated enzyme variants at the target residue, but it is likely that some variants are present at a lower abundance in the variant pool than others and therefore that these codons would be difficult to detect using Sanger sequencing. Taken together, all these results are strong evidence for mutagenesis at the target site and therefore the generation of a library of *6His::PEPC* variants.

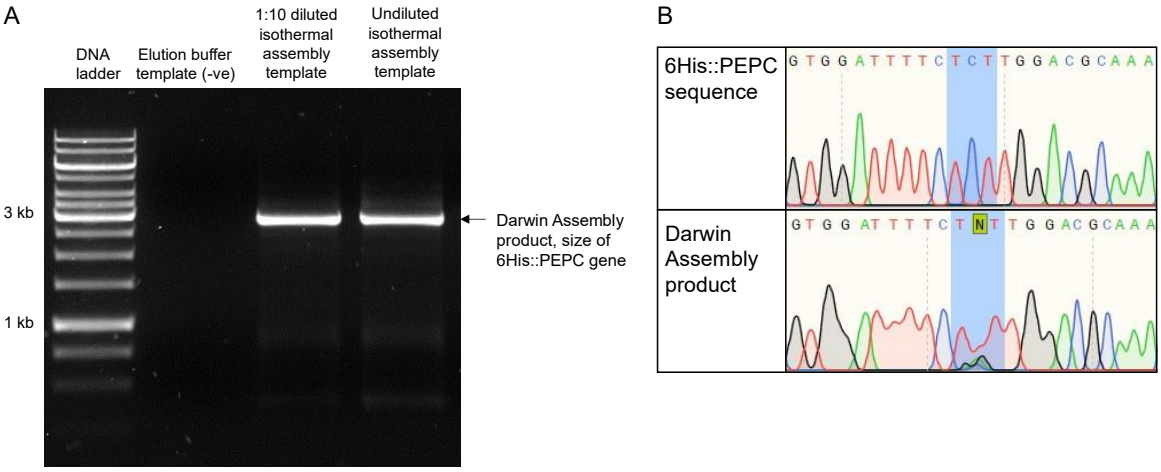


Figure 4.13: Example sequencing result from Darwin Assembly mutagenesis performed on *6His::PEPC* showed that mutagenesis was successful. (A) Amplification of a variant library resulted in a product the size of the full length 6His::PEPC gene. (B) A sequencing chromatogram of the PCR reaction product indicated that the wild-type sequence was TCT at site 780 (above) and that multiple bases were detected at all three positions in the codon at site 780 following mutagenesis.

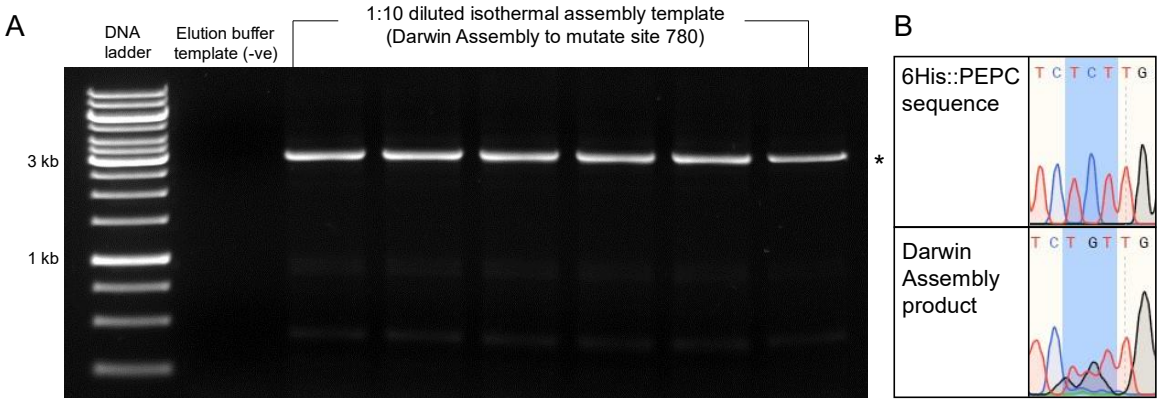


Figure 4.14: *6His::PEPC* variant library amplification can be scaled up in order to produce a high yield for downstream processes. (A) Large scale amplification of a variant library results in a product the size of the full length *6His::PEPC* gene (*). (B) A sequencing chromatogram of the large-scale PCR reaction product indicates that the wild-type sequence was TCT at site 780 (above) and that multiple bases were detected at all three positions in the codon at site 780 following mutagenesis.

4.2.2.4 Generation of 21 libraries of single-site enzyme variants

Enzyme variant libraries were generated for all 21 sites under selection in 6His::PEPC. Small scale PCRs were undertaken to confirm successful amplification of a PCR product (Figure 4.15 A), and then amplification was scaled up to generate sufficient quantity of the variant libraries for downstream cloning and transformation into *E. coli* Δppc (Figure 4.15 B). Some non-specific PCR products were generated as part of the amplification process but only at very small quantities and the gel extraction step prior to cloning ensured that only the full *6His::PEPC* variant sequences were used for subsequent cloning and transformation. *6His::PEPC* variant libraries were generated at all 21 sites (all large-scale reaction products included in Appendix 2).

There was some evidence of mutagenesis in the Sanger sequencing chromatogram of the scaled-up amplicon for 16 of the 21 variant libraries (Figure 4.14 B, Figure 4.16). This was apparent when the sequencing chromatogram appeared to show evidence of non wild-type bases or multiple bases at one or more positions in the codon. It was not possible to determine how many of up to 20 mutant variants were present in the amplicon pool for each single site library using these chromatograms. However, given the 'small-intelligent' oligonucleotide design approach where each amino acid was encoded by a single codon, a sequencing result containing evidence of multiple bases at one or more positions in the codon indicated that multiple *6His::PEPC* variants encoding different amino acids at a specific residue were present in the sequencing reaction and therefore that libraries of *6His::PEPC* variants have been generated.

Five of the 21 sites (517, 625, 637, 733 and 866) showed limited or no evidence of mutagenesis from the Sanger sequencing chromatograms (Figure 4.16) so an

alternative method of oligonucleotide binding was implemented to try and improve the mutagenesis result at these sites. The oligonucleotide and template mixtures were heated to 95 °C and then cooled at a rate of 0.1 °C/s to 4 °C, but this method of oligonucleotide annealing had no effect on the effectiveness of the mutagenesis approach as judged by the sequencing chromatogram of the amplicons. Oligonucleotide phosphorylation was performed in batches using a master mix of reagents each time, so poor performance of some oligonucleotides and not others phosphorylated within the same batch could not have been due to a problem with the phosphorylation step.

It is likely that the effectiveness of mutagenesis was negatively impacted during the assembly of some libraries by limitations to mutagenic inner oligonucleotide design, given the requirement for each oligonucleotide to span a site under selection in C_4 PEPC. For example, the self-dimerisation score for the oligonucleotide sequence at site 517 assuming no mismatches was ΔG -3.61 which is weak, so oligonucleotide dimers are unlikely to cause a problem during library generation. However, there were two oligonucleotides within the pool of 20 mutagenic oligonucleotides for this site with AAT and AAA at the site of mutagenesis, respectively (as per the 'small-intelligent' library design) that had heterodimerisation score of ΔG = -14.08 kcal/mol which are strong dimerization values. This could have biased the final library to include fewer mutant variants that encoded lysine and asparagine in place of the wild-type codon at site 517. Despite these few exceptions, there was evidence of successful mutagenesis at the majority of sites (Figure 4.16) and this technique resulted in *6His::PEPC* variants to screen in *E. coli*.

Although variation at some sites was not detected above, it is likely that some variation was captured but due to low abundance and the integrative nature of

Sanger sequencing, that they failed to be detected by this sequencing method. A more accurate estimate of the true number of variants present at each of the 21 sites was only possible when DNA from screens of 6His::PEPC in *E. coli* were sent for NGS (in Chapter 5). In conclusion, mutagenesis of *6His::PEPC* by Darwin Assembly created 21 libraries of enzyme variants and the majority showed strong evidence of mutagenesis, suggesting a large diversity of possible 6His::PEPC variants were captured.

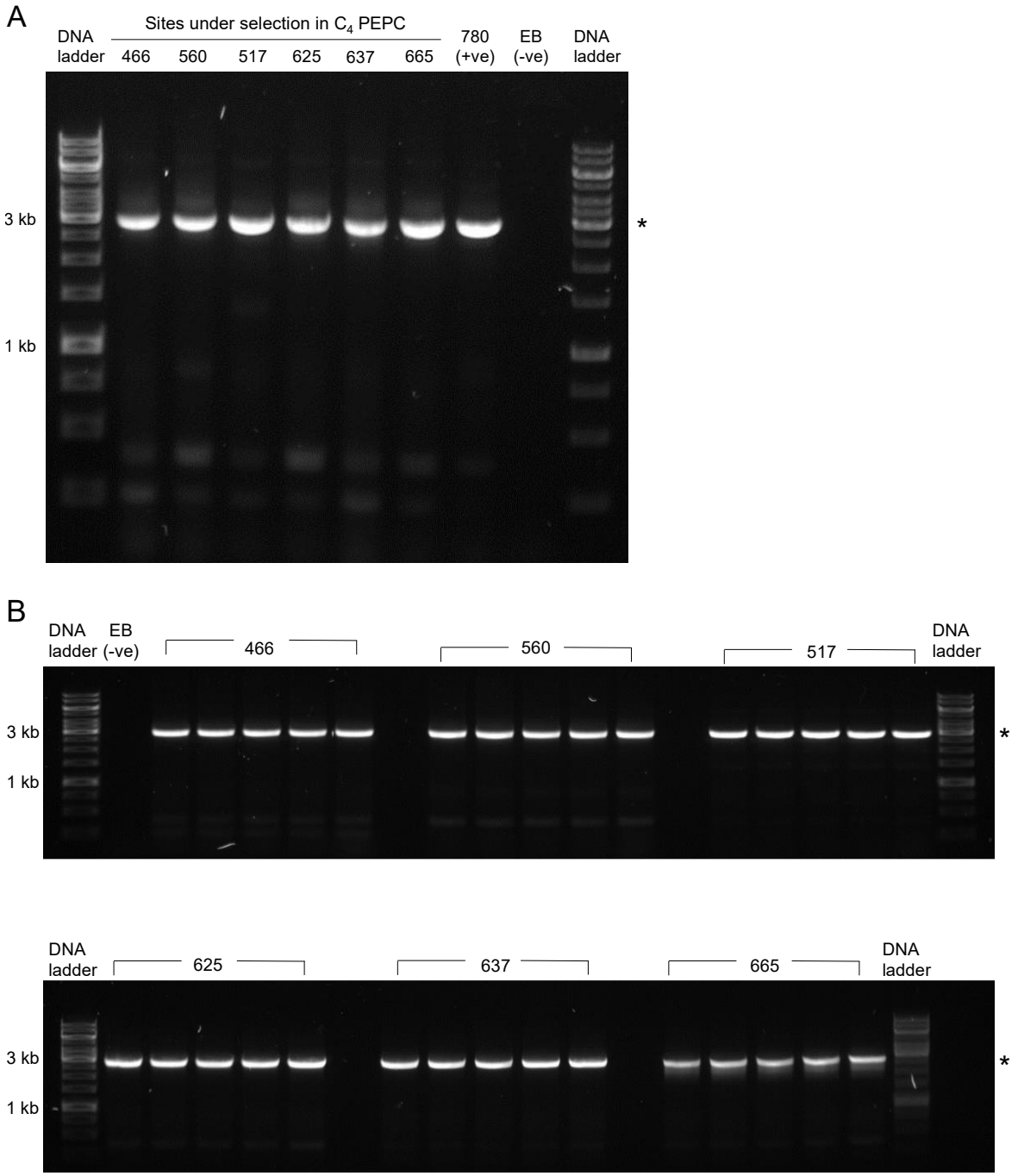


Figure 4.15: Darwin Assembly mutagenesis resulted in an amplicon the correct size for full-length *6His::PEPC* in both small- and large-scale PCR reactions. Six variant libraries (sites 466-665) were amplified from template DNA in small (A) and large (B) scale reactions, representative of this procedure undertaken for all 21 sites. The asterisk on the right hand side of the gel electrophoresis denotes the expected size of the 6His::PEPC gene.

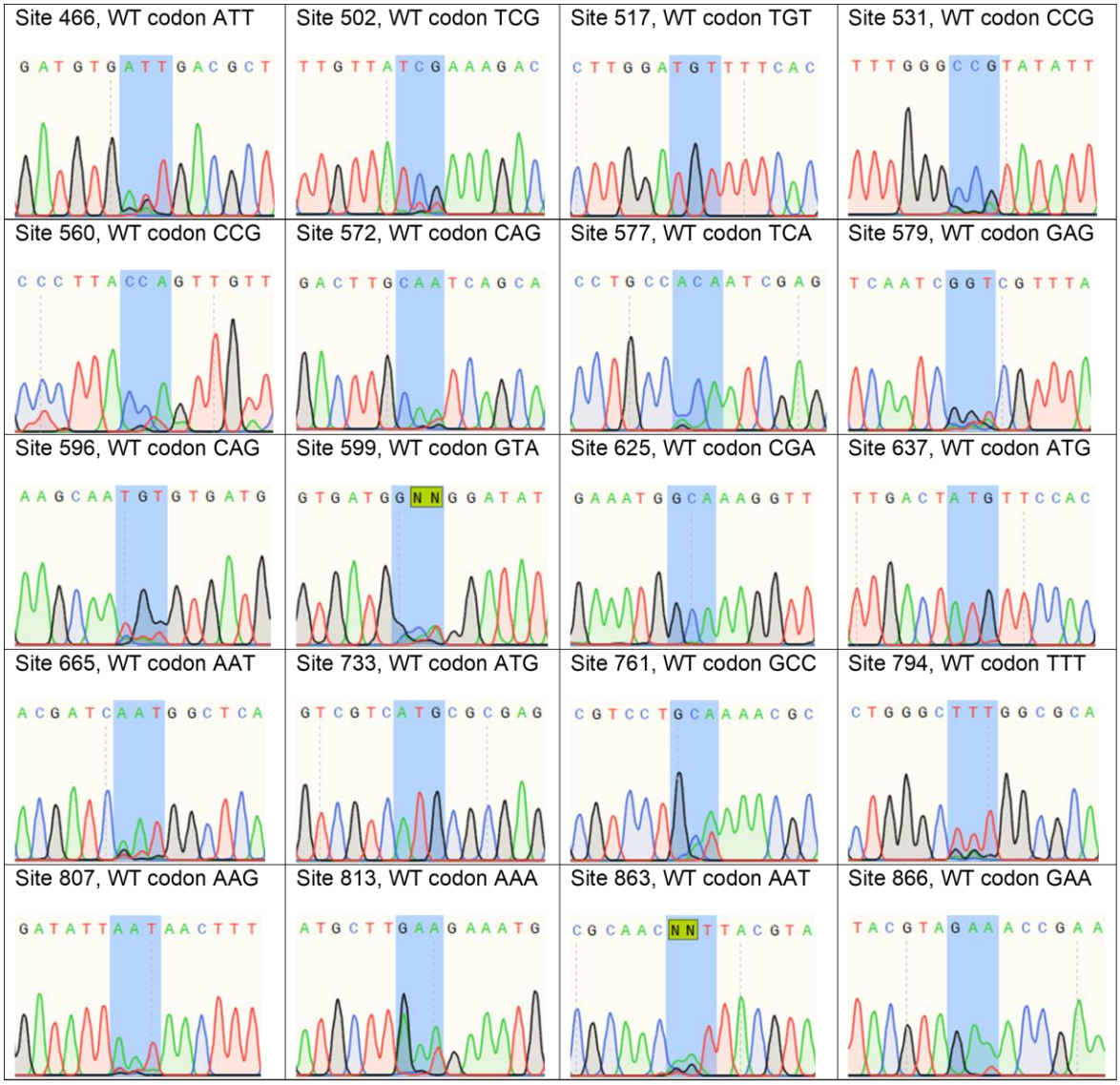


Figure 4.16: Sequencing results indicated successful mutagenesis of 15 out of 20 sites under selection in 6His::PEPC. Mutagenesis was demonstrated by the evidence of multiple bases at one or more positions in the codon when the Darwin Assembly product has been sequenced. Equivalent evidence of mutagenesis at site 780 is presented in previous Figure 4.13.

4.2.3 Electroporation of *6His::PEPC* into *E. coli* Δppc

The efficiency of *6His::PEPC* library transformation into *E. coli* Δppc will limit the number of variants that can be simultaneously analysed in any subsequent screen. Generation of highly competent *E. coli* Δppc will facilitate transformation of large *6His::PEPC* variant libraries with high library coverage. The two most reliable and effective methods of *E. coli* transformation are electroporation and delivering a 'heat shock' to heat-shock competent cells. Methods of preparation and transformation of heat-shock competent cells can yield *E. coli* with transformation efficiencies of up to 10^8 or 10^9 transformants/µg DNA but this can vary between batches (Hanahan, 1983). In contrast, it is possible to generate electrocompetent *E. coli* that consistently have transformation efficiencies of up to 10^9 or 10^{10} transformants/µg DNA (Dower, Miller and Ragsdale, 1988). Thus, electrocompetent *E. coli* Δppc were prepared to maximise the potential variant library coverage (Hanahan, Jessee and Bloom, 1991; Aune and Aachmann, 2010).

Protocols published by the company BTX, Langdale Lab (University of Oxford) and Green and Sambrook (2020) were each followed in order to establish the optimal method of preparing electrocompetent *E. coli* Δppc with the highest transformation efficiency. The wash step procedure of the protocol published by Green and Sambrook was adjusted to include two resuspensions in ice-cold water and only one in 10% glycerol, as opposed to the opposite number of each wash step described in the protocol to make it more similar to Dower, Miller and Ragsdale (1988) that reported very high transformation efficiencies, up to 10^{10} colony-forming units (CFU)/µg DNA. The three protocols were similar in scope: culture the bacteria in LB, harvest the bacteria at an OD of 0.4-0.5 (early to mid-logarithmic phase), keep the cells cold from harvesting onwards and perform wash steps in ice-cold

deionised water and ice-cold glycerol (10-15%) (Dower, Miller and Ragsdale, 1988; Green and Sambrook, 2020; BTX, 2022; Langdale, 2022).

The two major differences between the three protocols were the concentration of glycerol solution used during the final wash step and the chemical composition of the solution that the bacteria were resuspended in during the final resuspension step. Both the BTX and Langdale Lab protocols dictated final resuspension in 15% and 10% ice-cold glycerol, respectively, whereas Green and Sambrook (2020) dictated resuspension in GYT medium (10% glycerol with addition of yeast extract and tryptone). *E. coli* Δppc prepared according to the Langdale Lab protocol had the highest transformation efficiency, 6.1 $x10^9$ CFU/µg DNA compared to 4.1 $x10^6$ CFU/µg and 1.28 x 10^7 CFU/µg for competent cells prepared according to the BTX or Green and Sambrook (2020) protocols, respectively. The high transformation efficiency of bacteria prepared according to the Langdale Lab protocol was in line with maximum published values (Dower, Miller and Ragsdale, 1988).

Transformation efficiency is typically calculated using a small plasmid that is easy to transform such as pUC18 or pUC19 (Dower, Miller and Ragsdale, 1988; BTX, 2022) which are both 2.7 kb (Norrander, Kempe and Messing, 1983), but the plasmid size of the *6His::PEPC* variant library is 5.8 kb. Therefore, the most highly competent electrocompetent cells prepared using the Langdale Lab protocol were transformed with the 6His::PEPC expression vector that confers a low level of 6His::PEPC expression that has been previously assembled and characterised in *E. coli* Δppc (Chapter 3.2.2) that *6His::PEPC* variant library assembly will be based on. Electroporation of this 6His::PEPC expression vector into electrocompetent *E. coli* Δppc prepared using the Langdale Lab protocol (the same batch previously tested with pUC19), resulted in a transformation efficiency of 1.3 $x10^8$ CFU/µg DNA.

The reduction in transformation efficiency using the pre-assembled 6His::PEPC vector was because there were fewer than half as many plasmids per µg DNA compared to pUC19 and because as the size of DNA molecule used to transform *E. coli* increases, the transformation efficiency decreases (Miller and Nickoloff, 1995; Sheng, Mancino and Birren, 1995). Despite this, this transformation efficiency was more than sufficient for obtaining the 200 colonies required for up to 10-fold coverage of a single site library of up to 20 *6His::PEPC* variants. Thus, the transformation efficiency of electrocompetent *E. coli* Δppc will not be limiting to transformation with a *6His::PEPC* single site variant library.

4.2.4 Optimisation of cloning *6His::PEPC* variants into an expression vector

4.2.4.1 Development of a single-insert cloning system

6His::PEPC mutagenesis by Darwin Assembly results in the generation of enzyme variant libraries with Type IIS endonuclease recognition sites either side of the *6His::PEPC* gene to facilitate easy Golden Gate cloning into a compatible vector (Cozens and Pinheiro, 2018). The Type IIS endonucleases used in Golden Gate assembly cleave DNA outside of the endonuclease recognition sites which confers several advantages; multiple fragments can be ligated in a particular order, the assembly is 'seamless' because the cut and recognition sites are not the same and both digestion and ligation can occur simultaneously because the intended product does not contain the Type IIS recognition sites so this cloning method can be highly efficient even with complex assemblies (Engler, Kandzia and Marillonnet, 2008; Weber *et al.*, 2011; Pryor *et al.*, 2020).

An expression vector that conferred a low level of 6His::PEPC expression that was growth rate-limiting in *E. coli* Δppc in M9 glucose had previously been generated and characterised (Chapter 3). This expression vector was originally cloned as a four-fragment Golden Gate assembly using the Type IIS endonuclease BsaI (four fragments: promoter, ribosome binding site, 6His::PEPC gene and terminator sequences). It was possible to design outnest oligonucleotides for Darwin Assembly with appropriate BsaI sites so that 6His::PEPC variants could be cloned into an expression vector with all the same component parts *via* an equivalent four-fragment assembly as in Chapter 3.2.2. However, single insert cloning is more efficient than cloning multiple inserts (Kucera and Cantor, 2018; Marillonnet and Grützner, 2020; Pryor *et al.*, 2020). Therefore, a new vector was generated to facilitate single-insert assembly.

This new vector, termed the placeholder vector, was designed to be identical to the expression vector that confers the lowest level of 6His::PEPC expression in Chapter 3, except the 6His::PEPC CDS is replaced by a placeholder sequence of DNA. The DNA fragment used as a placeholder contained BsaI recognition sites that face inwards which were required in the primary round of cloning to create the placeholder vector (Figure 4.17 A). The fragment also contained BpiI sites that face outwards that were required for the second round of cloning where *6His::PEPC* variants were cloned in to the placeholder vector (Figure 4.17 A). The placeholder fragment was based on the *lacZ*α fragment (from the lactose operon) because it is widely used in vector construction and is useful in blue/white selection of colonies in tandem with compatible *E. coli* strains (Norrander, Kempe and Messing, 1983; Kovach *et al.*, 1994; Juers, Matthews and Huber, 2012).

To make the placeholder vector, the original expression vector that conferred the lowest level of 6His::PEPC expression in *E. coli* Δppc (Chapter 3) was re-assembled in a Golden Gate reaction using all the same component parts (promoter, ribosome binding site, terminator and a medium copy number vector backbone) but a fragment of DNA acting as a placeholder instead of the 6His::PEPC CDS (Figure 4.17 B). The BsaI sites were cut away from the placeholder fragment during this initial cloning step, but the BpiI sites remained within the placeholder fragment in the placeholder vector. On digestion of the Darwin Assembly product and placeholder vector with BpiI, the BpiI recognition sites were cut away leaving compatible sticky ends between *6His::PEPC* and the expression vector. On ligation with Taq DNA ligase, the *6His::PEPC* variant was cloned into the expression vector where the placeholder DNA fragment once was. In this way, *6His::PEPC* was cloned

into a pre-assembled expression vector in a single-fragment assembly (Figure 4.18).

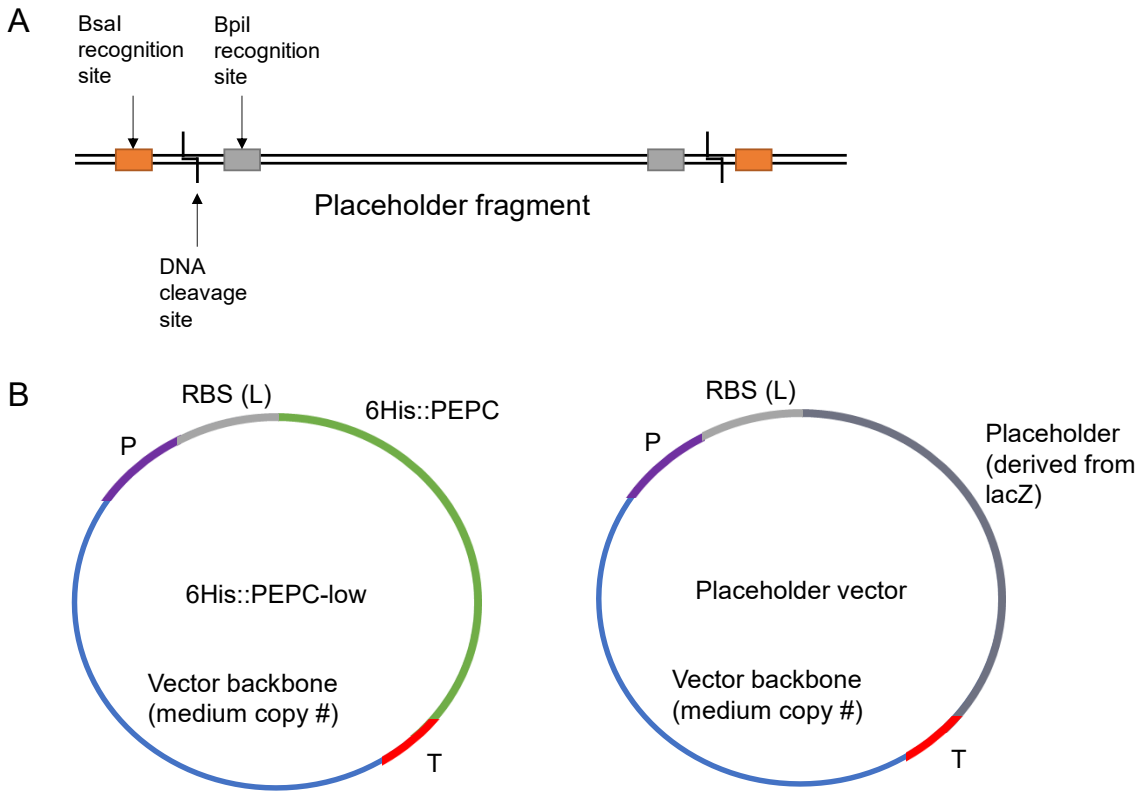


Figure 4.17: Design of the placeholder fragment and placeholder vector used for single-fragment Golden Gate assembly of *6His::PEPC* into a vector suitable for expression in *E. coli* Δppc. (A) The cut site of both restriction endonucleases BsaI and BpiI remained the same. (B) A cartoon of the expression vector containing the 6His::PEPC gene next to the placeholder vector where the placeholder fragment was cloned in between the promoter and ribosome biding site (P and RBS L) and terminator (T) sequences.

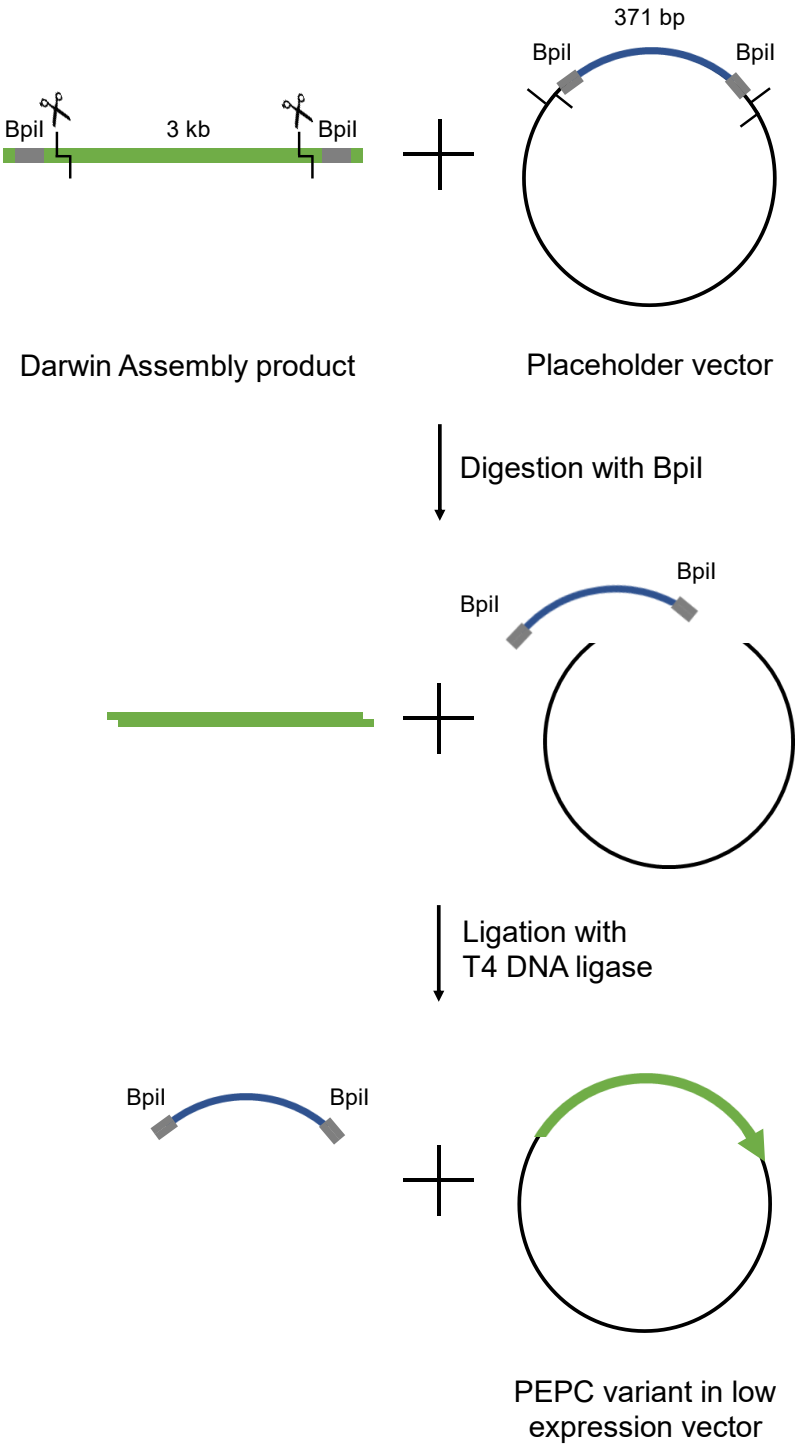


Figure 4.18: A single-fragment Golden Gate assembly of *6His::PEPC* variants into an expression vector. Digestion of both the 6His::PEPC variant library and placeholder vector with BpiI resulted in compatible complementary (sticky) ends of DNA for the Darwin Assembly product to be seamlessly cloned into the pre-assembled low-expression vector in a single-insert assembly.

4.2.4.2 Troubleshooting the cloning step using the single-insert cloning system

4.2.4.2.1 Testing different Golden Gate reaction conditions

Multiple Golden Gate reactions were performed, each with different reaction conditions, to test the efficiency of cloning a *6His::PEPC* variant library into the placeholder vector (Figure 4.18), and to try to find optimal reaction conditions. All the reactions included the placeholder vector, BpiI as the restriction endonuclease and a *6His::PEPC* variant library equivalent. To generate a *6His::PEPC* fragment equivalent to the one that will be produced after mutagenesis, *6His::PEPC* was amplified using PCR oligonucleotides that introduced equivalent sequences on either side of the *6His::PEPC* gene to those that were introduced after boundary and outnest PCRs in Darwin Assembly (Figure 4.19, oligonucleotide sequences in Appendix 3). The PCR product therefore had BpiI recognition sites on either side of the *6His::PEPC* gene that were suitable for cloning the *6His::PEPC* gene into the placeholder vector. Phusion high-fidelity DNA polymerase was used to ensure the *6His::PEPC* gene and sequences introduced either side were amplified with minimal errors and to conform with the Darwin Assembly protocol that has been optimised (Chapter 4.2.2). The *6His::PEPC* PCR product was then separated by gel electrophoresis and gel extracted to ensure only full-length *6His::PEPC* products were isolated before being used in all the test Golden Gate reactions.

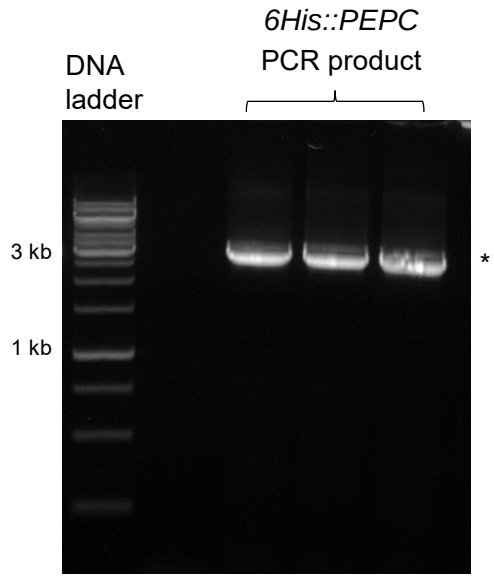


Figure 4.19: *6His::PEPC* PCR product with BpiI cloning sites either side of the CDS to simulate the variant library amplification after Darwin Assembly. The asterisk on the right hand side of the gel electrophoresis denotes the expected size of the *6His::PEPC* gene.

To optimise cloning *6His::PEPC* variant libraries into the placeholder vector, five different Golden Gate reaction conditions were tested between the simulated *6His::PEPC* variant library and the placeholder vector (Table 4.5). Each set of reaction conditions was similar but one or more of the following components varied between reactions: the molar ratio of insert to backbone, the buffer used and the quantity of each enzyme (BpiI and T4 DNA ligase). One µL of each Golden Gate reaction product (10% total reaction product) was transformed into chemically competent *E. coli* Δppc. Chemically competent *E. coli* Δppc were used initially as a quick way of establishing whether or not any of the Golden Gate reactions were successful.

Table 4.5: Five different reaction conditions were tested to improve the efficiency of ligation of *6His::PEPC* variants into the pre-assembled placeholder vector.

Component	Details	1	2	3	4	5
Placeholder vector	54.5 ng/µL 3331 bp	10 fmol	24 fmol	10 fmol	25 fmol	25 fmol
6His::PEPC PCR product	118 ng/µL 2958 bp	20 fmol	55 fmol	20 fmol	25 fmol	25 fmol
ATP	10 mM	1 µL	1 µL			1 µL
BSA	NEB, 1 mg/mL			1 µL	1 µL	
BpiI	ThermoFisher, 10 U/µL	0.5 µL	0.5 µL	0.65 µL	0.65 µL	0.5 µL
T4 DNA ligase	400 U/µL	0.25 µL	0.25 µL	0.65 µL	0.65 µL	0.25 µL
Buffer G (10X)	ThermoFisher	1 µL	1 µL			1 µL
T4 DNA ligase buffer (10X)				1 µL	1 µL	
Water	ddH_2O	To 10 µL	To 10 µL	To 10 µL	To 10 µL	To 10 µL

Placeholder vector and PEPC PCR product sequences are available in Appendix 1.

BSA is bovine serum albumin.

The transformations were plated on LB-agar plates and M9 glucose plates. The rationale behind plating on M9 glucose as well as LB-agar plates is that given that *E. coli* Δppc cannot grow in M9 glucose without expression of the 6His::PEPC (Chapter 3.2.4), only *E. coli* Δppc transformed with *6His::PEPC* correctly ligated in the expression vector would grow on the M9 glucose plates. In theory, it would therefore be possible to distinguish between cells that have taken up the placeholder vector compared to the *6His::PEPC* in low expression vector ligation product, because cells that have taken up the placeholder vector can only grow on LB and not in M9 glucose. Transformants were plated on LB following an outgrowth step in SOC, then all the remaining cells were re-suspended in M9 glucose liquid culture before a fraction of the remaining cells were plated on M9 glucose.

For all five Golden Gate reactions tested and transformed into *E. coli* Δppc, the largest number of transformants were observed on the LB-agar plates (Figure 4.20), but these could be *E. coli* Δppc transformed with either the correct ligation product or the placeholder vector. Two of the five Golden Gate reactions (4 and 5) that both contained equal molar ratio of plasmid backbone and *6His::PEPC* insert consistently resulted in a higher number of CFU on both the LB and M9 glucose plates (Figure 4.20). The estimated number of CFU per transformation of chemically competent *E. coli* Δppc with 1 µL Golden Gate reaction product given the number of CFU in M9 glucose plates was 56 and 75 for the best two performing transformations, respectively (Figure 4.20). These colonies were thought to be transformed with correct ligation products because heterologous expression of 6His::PEPC must be required for the bacteria to survive on plates where glucose is the sole carbon source (Cozzone, 1998; Sauer and Eikmanns, 2005). The colonies took longer to appear on the M9 glucose plates (3 days) compared to overnight on

the LB-agar plates. Given that *6His::PEPC* has been cloned into a low expression vector that confers a slow rate of growth and that the cells are dependent on 6His::PEPC activity to grow on the M9 glucose plates, this was not considered unusual.

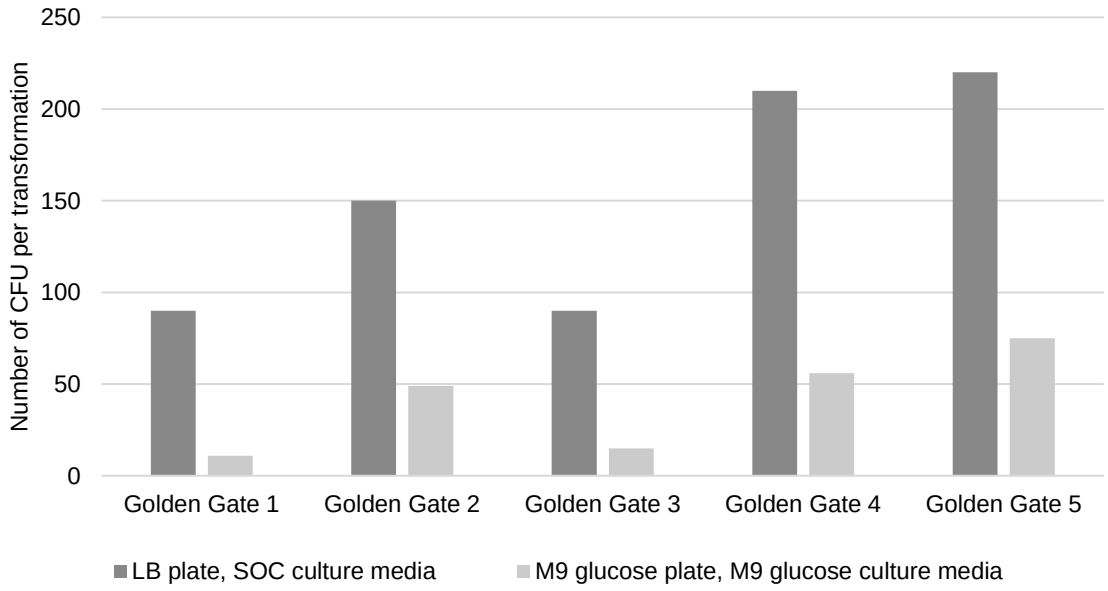


Figure 4.20: The number of CFU after five Golden Gate reactions were undertaken under different reaction conditions revealed that different reaction conditions resulted in different numbers of transformants when the DNA parts remained the same. Five reactions were undertaken that were different in one or more of the following ways: the molar ratio of *6His::PEPC* to backbone, the buffer used and the quantity of each enzyme (BpiI and T4 DNA ligase). The CFU presented here are representative of a single transformation of 1 µL (10%) Golden Gate reaction product into 40 µL chemically competent *E. coli* Δppc (set up of reactions 1 – 5 described in Table 4.5).

4.2.4.2.2 Can increasing the number of digestion and ligation cycles improve Golden Gate efficiency?

To try and improve cloning efficiency, the two best-performing Golden Gate reactions – Golden Gate reactions 4 and 5 – were repeated with either 25 cycles of digestion and ligation as before, or 65 cycles of digestion and ligation. The reaction products were transformed into electrocompetent *E. coli* Δppc in order to achieve the maximum number of transformants per transformation reaction. 1 µL of each Golden Gate reaction product was transformed into electrocompetent *E. coli* Δppc and plated on LB and M9 glucose plates as before. Transformation was undertaken without de-salting the Golden Gate reaction product in order to minimise the number of steps in the mutagenesis, cloning and *E. coli* transformation pipeline and maintain a low risk of co-transformation of bacteria with two plasmids that encode different 6His::PEPC variants (Tomoiaga *et al.*, 2022).

The reaction conditions for Golden Gate 4 with either 25 or 65 cycles of digestion and ligation steps were the best-performing reaction out of the four tested here (Figure 4.21). T4 DNA ligase buffer contains ATP required for the DNA ligase to catalyse ligation between DNA fragments and BSA was added to stabilise the enzyme because it is present in the buffer recommended for BpiI digestion (New England BioLabs, 2023; ThermoFisher Scientific, 2023). In addition, all Golden Gate reactions were held at 65 °C for 20 minutes after the digestion and ligation cycles to inactivate BpiI and DNA ligase and consequently increase the efficiency of transformation into electrocompetent *E. coli* Δppc (Ymer, 1991).

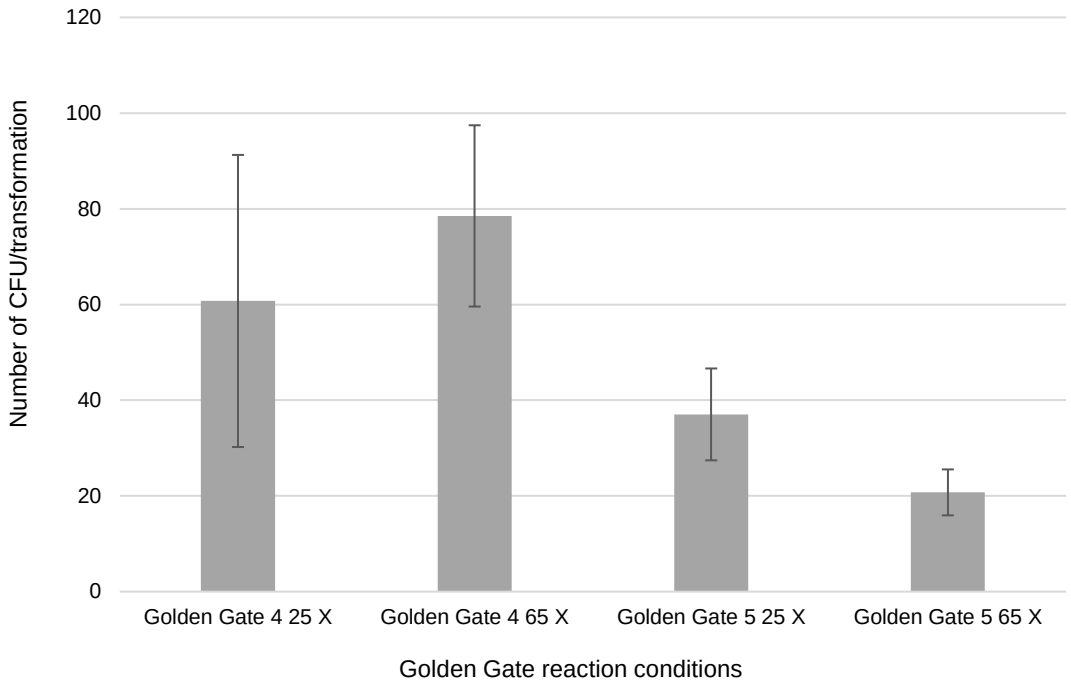


Figure 4.21: The number of CFU observed in M9 glucose plates after electrocompetent *E. coli* Δppc were transformed with 1 µL (10%) product of Golden Gate reactions that differed either by composition (numbered 4 and 5) or number of digestion/ligation cycles (25 or 65 cycles (X)). Data is the mean of three *E. coli* Δppc transformations ±SD.

One µL reaction product from a Golden Gate reaction undertaken using the above reaction conditions and 65 cycles of digestion and ligation transformed into electrocompetent *E. coli* Δppc resulted in 78 CFU on average per transformation (Figure 4.21). This was greater than the average number of colonies observed when using fewer digestion and ligation cycles or when different reaction conditions were used and the resulting transformation was plated in M9 glucose (Figure 4.21). The number of transformants from the same reaction plated on LB was 2,550. This suggested that there may have been a problem with the cloning step since it appears that only 3% of the bacteria had been transformed with the correctly ligated product and were therefore able to grow in M9 glucose. Given that the electrocompetent cells were highly competent (Chapter 4.2.3), the step that is limiting the number of transformants appeared to be the Golden Gate cloning step.

Despite difficulties optimising the cloning step, given that there were approximately 78 CFU after transformation of 1 µL Golden Gate reaction in 1 vial of electrocompetent cells (Figure 4.21), then it was expected that 780 transformants could have been obtained from transformation of all 10 µL of the Golden Gate reaction product in 10 separate transformations into electrocompetent cells. Only three separate transformations would therefore have been necessary to achieve over 10-fold coverage of a single site library of 20 variants. Thus, despite the numbers being much lower than expected, it was possible to continue with transformation of *E. coli* Δppc and achieve good coverage of the single site libraries with the results demonstrated here.

4.2.4.2.3 Transformation efficiency vs transformation frequency

A final investigation was performed to identify whether an increase in the quantity of Golden Gate reaction product used in each transformation resulted in a higher

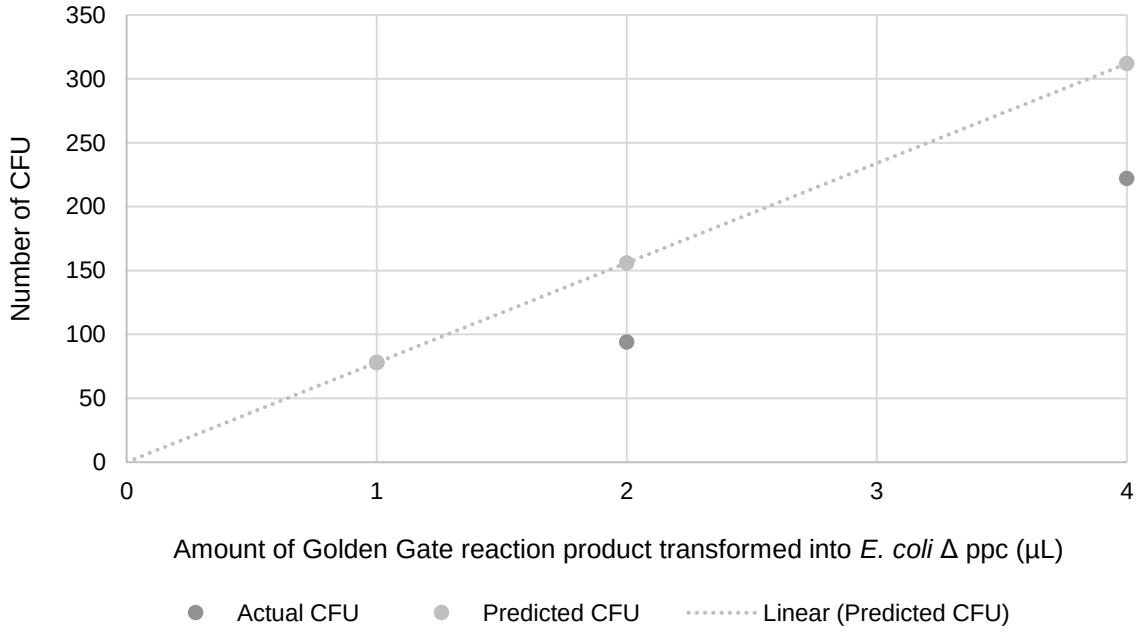
number of CFU because an increase in the transformation frequency (number of transformants per electroporation) could have reduced the number of transformations that were required for equivalent library coverage of a single site library. The transformation frequency has been shown to increase when greater quantities of DNA are used per *E. coli* transformation, even if these results represent a decrease in transformation efficiency (CFU/ µg) (Dower, Miller and Ragsdale, 1988; Miller and Nickoloff, 1995).

The Golden Gate reaction that yielded the most transformants in previous experiments, Golden Gate 4 reaction conditions subject to 65 digestion and ligation cycles, resulted in an estimated 78 CFU after transformation with 1 µL reaction product. Electrocompetent *E. coli* Δppc were transformed with 2 µL and 4 µL of the same Golden Gate reaction product for comparison and plated on M9 glucose plates. Both of the transformations resulted in a higher number of CFU per transformation, 94 and 222 on average respectively calculated from CFU in M9 glucose plates (Figure 4.22) but not to the equivalent of double or quadruple the number of CFU, respectively. Conversely, when the volume of reaction product was increased, this resulted in a lower transformation efficiency (CFU per µL Golden Gate product) (Figure 4.22).

These experiments demonstrated that by increasing the amount of Golden Gate reaction product per electroporation undertaken, fewer vials of electrocompetent cells would be necessary to gain enough transformants for sufficient library coverage of a single-site library. By contrast, to maximise the number of transformants per volume cloning reaction product it would be necessary to use 1 µL reaction product in each electroporation. Crucially, for a library screen it is preferable to keep transformation efficiency high to reduce the likelihood of co-

transformation events, even if this is at the expense of CFU frequency per transformation (Dower, Miller and Ragsdale, 1988; Miller and Nickoloff, 1995; Green and Sambrook, 2020).

A



B

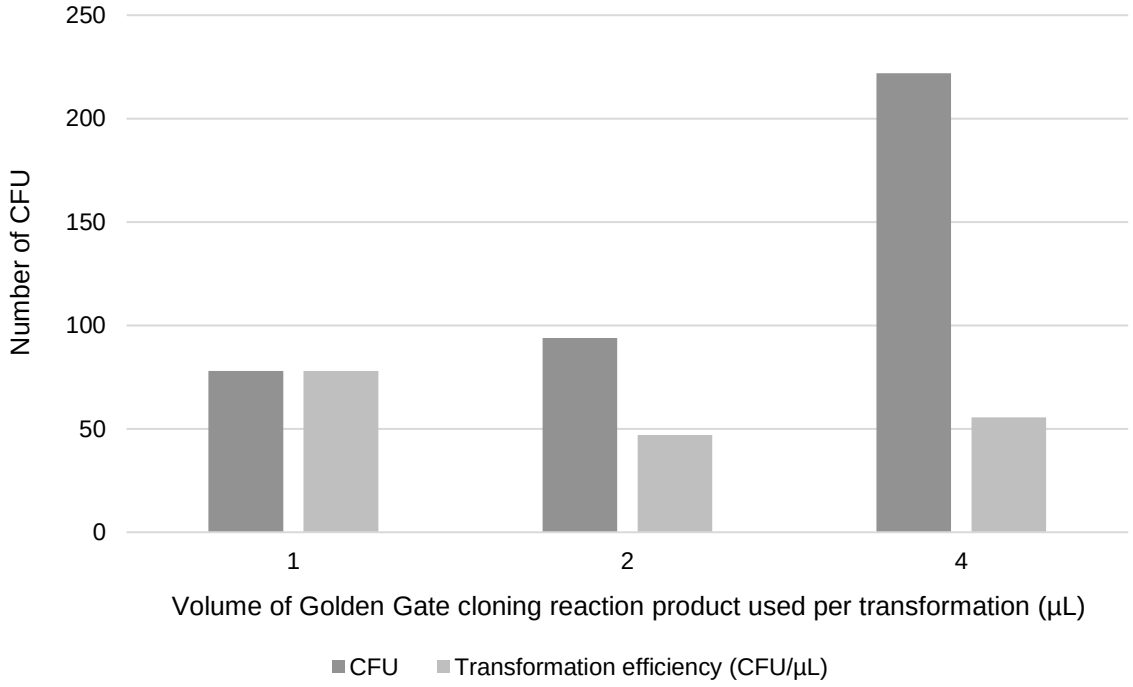


Figure 4.22: The number of CFU increased but transformation efficiency decreased as the volume of Golden Gate reaction product used for a single electroporation increased. Figure legend continued on the next page.

Figure 4.22 (continued) A) The number of CFU that resulted from 1 µL Golden Gate reaction product transformed into 40 µL electrocompetent *E. coli* Δppc compared to the number of CFU when the same volume of cells was transformed with double or quadruple the volume of reaction product. B) The same data presented as a chart to illustrate CFU per transformation.

4.2.5 Transformation of *E. coli* Δppc with a library of 6His::PEPC variants

A library of *6His::PEPC* variants with mutations at site 780 were assembled into the placeholder vector and electroporated into *E. coli* Δppc. 40 CFU were observed on the LB plates after 10% outgrowth culture was plated but only an equivalent of 8.8 CFU per M9 glucose plate (average number of CFU across three plates). However, a colony PCR indicated that the generation, cloning and transformation of a variant library into *E. coli* Δppc was highly successful. All 20 colonies analysed (of 40 on the LB plate) contained the 6His::PEPC expression vector as opposed to the placeholder vector (Figure 4.23). This suggested that there were ~400 transformants in the outgrowth culture from a single electroporation. A single electroporation using 1 µL of the Golden Gate reaction product could therefore confer ~200-fold library coverage of a single-site library.

To get an idea of the variation in the library, plasmid preparations were isolated for 14 out of the 40 transformants on the LB plate and the DNA was sent for sequencing. Four different amino acids were observed at the codon that was subject to mutagenesis, none of which were the wild-type serine codon TCT (Table 4.6). This demonstrated the success of the mutagenesis, cloning and transformation pipeline resulting in *E. coli* Δppc containing different *6His::PEPC* variants. For each *6His::PEPC* variant, the mutated codon started or ended with a T, which was consistent with the idea that there could have been an element of codon bias within the *6His::PEPC* variant library depending on which oligonucleotides were most similar to the wild-type codon suggested in section 4.2.2.3.

Given that there was approximately 20 times library coverage from a single electroporation it was likely that *6His::PEPC* variants encoding a different amino acid at site 780 to those observed here were also transformed into *E. coli* Δppc

(Table 4.6) even if these other *6His::PEPC* variants appeared at a lower frequency by comparison. The true diversity of *6His::PEPC* variants screened in *E. coli* Δppc could only be detected when samples underwent NGS which is covered in the next chapter. The sequencing result for all 14 transformants also demonstrated a single codon sequence at the site of mutagenesis which suggested that bacteria were not subject to co-transformation with multiple plasmids.

In conclusion, a pipeline was established here whereby a library of *6His::PEPC* variants were generated, assembled into a suitable expression vector and then the vector transformed into *E. coli* Δppc with a sufficient level of library coverage. This success laid the foundation for screens of 6His::PEPC variants in *E. coli* Δppc.

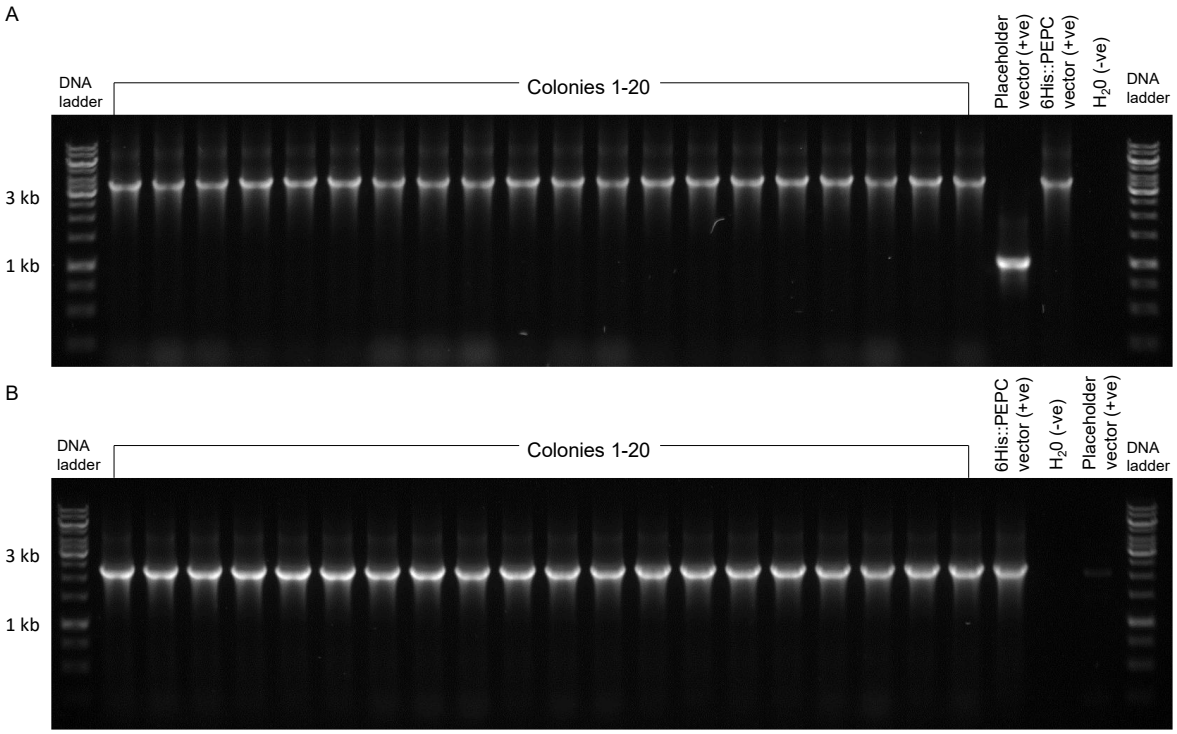


Figure 4.23: A high proportion of *E. coli* Δppc were transformed with a correct Golden Gate assembly between *6His::PEPC* and the placeholder vector. Amplicons from CFU on selective plates were compared to amplicons from the backbone vector and the 6His::PEPC pre-assembled vector. Where ligation of the *6His::PEPC* product into the expression vector was successful, the amplicon is the same between the CFU and 6His::PEPC pre-assembled vector. Here, two sets of oligonucleotides were used that either anneal either side of the cloning sites (A) or to the *6His::PEPC* gene and the backbone (B).

Table 4.6: Four different *6His::PEPC* variants were observed in *E. coli* Δppc that had been transformed with the *6His::PEPC* library.

Codon	**Amino acid encoded**	**Frequency observed**
TCT (wild-type)	**Serine**	**0**
TTT	Phenylalanine	5 (36%)
TAT	Tyrosine	3 (21%)
AGT	Serine	4 (29%)
TGT	Cysteine	2 (14%)

4.3 Chapter summary

In this chapter, *6His::PEPC* mutagenesis, cloning and *E. coli* transformation were optimised to develop a pipeline that could underpin high-throughput screens to identify 6His::PEPC variants that confer a relative increase in bacterial growth rate. Darwin Assembly was found to be a powerful multi-site saturation mutagenesis technique that could be used to generate libraries of *6His::PEPC* variants mutated at a single site or multiple sites simultaneously. Transformation of *E. coli* Δppc with a simulated *6His::PEPC* variant library did not result in a large number of transformants that would ensure high library coverage of multi-site mutagenesis libraries due to cloning limitations. However, the aim of this project was to identify single amino acid changes to 6His::PEPC that could be delivered *via* gene editing, so the assembly of *6His::PEPC* variants into the expression vector would not limit coverage of single-site libraries based on the results presented here.

The work in this chapter culminated in generation of 21 libraries of *6His::PEPC* variants by mutating sites that have been identified as being under selection in C_4 PEPC. This included a library of *6His::PEPC* variants mutated at a residue under selection in C_4 PEPC (equivalent to site 780 in *Z. mays*) that is highly influential in C_4 PEPC enzyme kinetics. Variants with mutations at site 780 were subsequently cloned into a growth-rate limiting expression vector and transformed into *E. coli* Δppc with high library coverage. The *6His::PEPC* variants generated at all 21 sites under selection were assembled into a growth-rate limiting expression vector and taken forwards for a high-throughput screen of 6His::PEPC in *E. coli* in the next chapter.

Chapter 5: Identification of enhanced variants of *Seteria viridis* C_4 PEPC

5.1 Chapter introduction

5.1.1 Development of a novel 6His::PEPC variant screen in *E. coli*

The results outlined in the previous two chapters describe the development of a pipeline for screening 6His::PEPC variants in bacterial co-culture such that cellular growth rate of *E. coli* Δppc is dependent on 6His::PEPC activity. In Chapter 3, it was demonstrated that heterologous expression of 6His::PEPC in *E. coli* that lacks endogenous PEPC (Δppc) restores bacterial growth in M9 glucose minimal medium and that bacterial growth rate was limited by the expression level of 6His::PEPC. In Chapter 4, a mutagenesis, cloning and *E. coli* Δppc transformation pipeline was established whereby *6His::PEPC* variants were generated by site saturation mutagenesis and transformed into *E. coli* Δppc on a plasmid vector that confers a low level of 6His::PEPC expression. Given that the level of 6His::PEPC expression was fine-tuned such that 6His::PEPC expression level limited the growth rate of *E. coli* Δppc in M9 glucose, any relative improvement in the rate of catalysis by a 6His::PEPC variant in *E. coli* Δppc should confer a faster rate of bacterial growth to the host cell and become enriched in the population during a co-culture assay.

The aim of the research described in this chapter was to identify single amino acid substitutions that improved enzyme activity of 6His::PEPC such that they could in future be engineered into *S. viridis* for *in planta* characterisation through a gene editing approach. Although this chapter is focussed on C_4 PEPC from *S. viridis*, the screening pipeline could be extended to C_4 PEPCs from crop plants including maize, sugarcane and sorghum.

5.1.2 Establishing a pipeline to identify the best-performing 6His::PEPC variants using continuous co-culture in *E. coli* Δppc

The *6His::PEPC* gene is encoded on a plasmid, so it is possible to identify 6His::PEPC variants that are increasing in relative abundance in the culture over time by simply performing a plasmid miniprep and then sequencing the plasmid DNA that is isolated from each sample. However, the cost of sequencing many samples using NGS quickly becomes prohibitive. For example, a single experiment in which a population is sampled for sequencing 6 times over the course of 72 hours with 3 biological replicates would require 18 samples to be sequenced. However, given that Illumina sequencing can produce a large number of reads, it is unnecessary to sequence each library independently. Instead, it is possible to take advantage of the high sequencing depth to sequence multiple barcoded libraries in each sequencing run.

In order to develop a barcoding and multiplexing strategy for Illumina sequencing it was necessary to consider several factors. First, the co-cultivation assay was conducted at a low cell density so that the culture was maintained in the logarithmic phase of growth (Kobayashi *et al.*, 2006). Second, the expression plasmid had a medium copy number origin of replication (pBBR1) and thus was present at only 5-10 copies per cell (Jahn *et al.*, 2016; Geddes, Mendoza-Suárez and Poole, 2019). Third, the total culture volume was only 25 mL, so samples taken from the bioreactors had to be small in order to avoid the creation of population bottlenecks that could adversely affect variant frequency. Together, these constraints meant that the total plasmid DNA prepared from the sample taken at each time point was limited and therefore had to be amplified by PCR in order to produce enough DNA to satisfy Illumina library preparation sample requirements. This PCR step facilitated

the addition of DNA “barcodes” (unique short nucleotide sequences) to each PCR product. The barcodes allowed each sample to be uniquely identified and thus enabled DNA samples to be pooled and processed in a single NGS run. The resulting sequence data could then be demultiplexed computationally to identify the precise replicate and time point that each read came from.

The barcoding PCR step was designed for 150 bp paired-end reads. Here, four different PCR amplicons were designed such that all 21 sites that were subject to mutagenesis were contained within one of the four amplicons (Figure 5.1 A). The PCR amplicons were designed such that every site was within 150 bp from the 5’ end of a primer in the minimum number of fragments (Figure 5.1 B). The four amplicons contained 4, 9, 6 and 2 sites, respectively (Figure 5.1).

The PCR primers were barcoded through the addition of two bases at the 5’ end of the primer; the barcodes were either AC, AG, CA, CT, GA, GT, TC or TG. The barcode was preceded by a four base ‘buffer’ so that in the case of incomplete sequencing from the 5’ end of DNA, the barcode will remain intact (Figure 5.2 A). Using barcodes, eight different forward primers and eight different reverse primers can be used to create 64 different barcode combinations, thus enabling 64-way multiplexing of any given sequencing reaction (Figure 5.2 B). All barcoded primer pairs were checked for secondary structures and only primer pairs that passed common metrics such as heterodimerisation scores were taken forward for use (all NGS primers listed in Appendix 3). All primers were synthesised to “Next-generation sequencing grade”, which guarantees <0.1% cross-contamination in order to reduce the risk of incorrect barcodes being amplified during the PCR step.

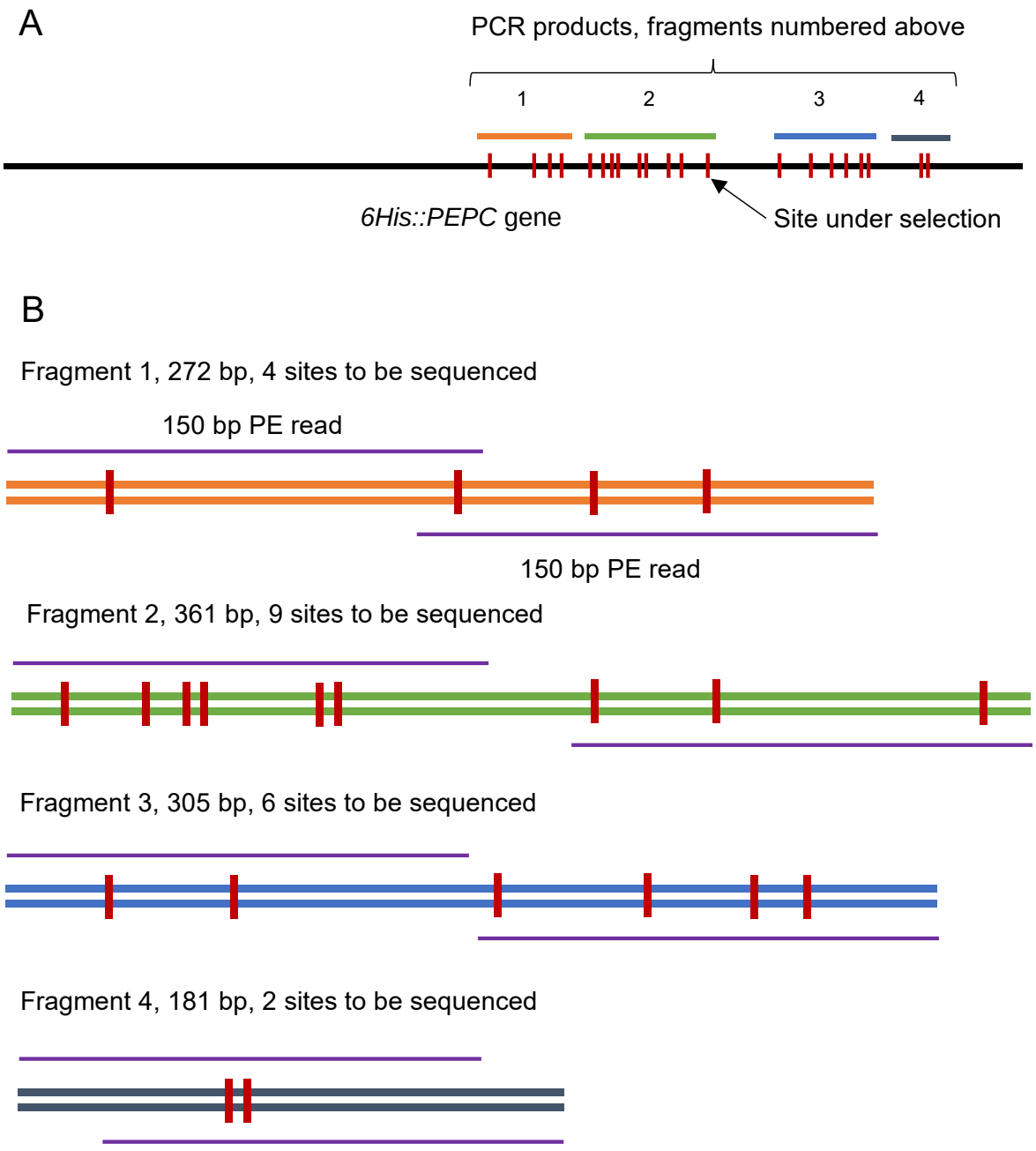


Figure 5.1: The amino acid encoded at each site under selection could be identified by paired-end 150 base pair NGS of appropriately positioned PCR products that between them contained every site. PCR products are denoted in orange, green, blue and grey, respectively. Sites under selection are represented in red.

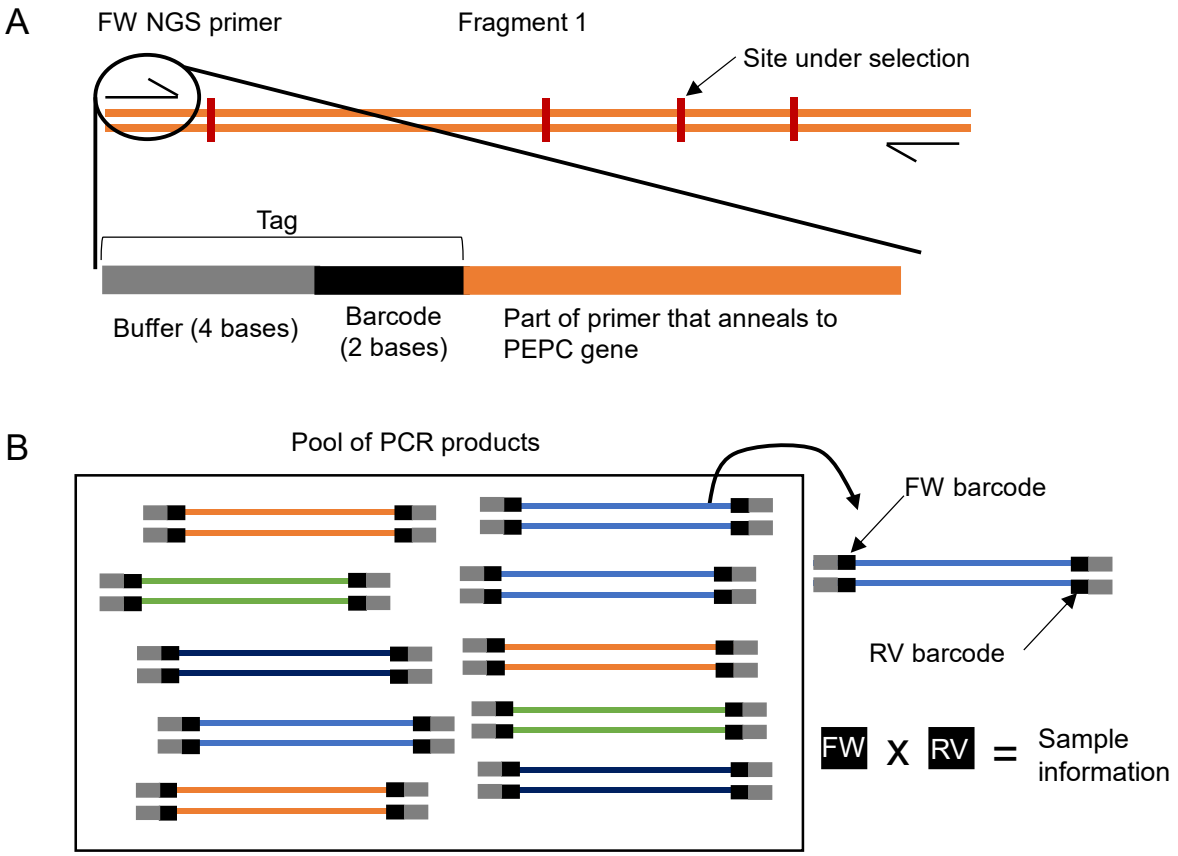


Figure 5.2: All of the primers used in PCR amplification contained a barcode which was used to encode replicate and time point data. PCR products are denoted in orange, green, blue and grey, respectively and sites under selection are represented in red. (A) Forward and reverse primers each contained a barcode and a buffer sequence that are two and four bases long, respectively, added as a tag at the 5' end of the primer. (B) The forward and reverse barcodes in combination from the paired-end read sequencing data facilitated identification of the original sample information that the PCR product was amplified from.

5.1.3 Chapter aims and outline

The aims of this chapter were three-fold. The first aim was to complete the development of the screening pipeline such that 6His::PEPC variants are competed in *E. coli* Δppc continuous co-culture and the best-performing variants identified through a NGS and bioinformatics component of the pipeline. This was achieved by screening up to 20 6His::PEPC variants in *E. coli* that have been mutated at a site that is known to be influential in C_4 PEPC kinetics. Once this had been successful, the second aim was to extend the screen to up to 420 6His::PEPC variants that were all mutated at a single site in the enzyme and could therefore be introduced into a plant *via* a gene editing approach. The third aim was to evaluate the best-performing variants across the two screens which will inform subsequent plant editing strategies.

5.2 Results and Discussion

5.2.1 Development of a NGS and bioinformatics pipeline to identify best-performing 6His::PEPC variants

5.2.1.1 DNA amplification and sample multiplexing prior to sequencing

In the preceding chapters, a system was developed to generate *6His::PEPC* variants and screen for improved 6His::PEPC activity in *E. coli*. This pipeline is flexible and scalable; mutations can be introduced at any site(s) of interest in 6His::PEPC using the Darwin Assembly mutagenesis approach and libraries of enzyme variants can screened simultaneously by continuous co-culture. Using such a large library approach required the development of a bioinformatics pipeline that allowed the determination of both which variants were present, and how each variant performed relative to all the others over the course of the co-culture experiment. To do this, the bacterial culture was sampled periodically over the duration of the assay and these samples were processed to track the relative abundance of each 6His::PEPC variant over time and therefore identify best-performing enzyme variants.

The PCR to amplify plasmid DNA isolated from a sample taken at a specific time point during a screen must be semi-quantitative so that the relative abundance of each variant in the amplicon is representative of the abundance of each variant within the isolated sample PCR template DNA. To achieve this, the maximum amount of template and the minimum number of PCR cycles were used and the PCR was halted during the exponential phase (Figure 5.3 A). Evaluation of multiple parameters identified that 20 PCR cycles from 10 ng DNA template yielded sufficient amplicon DNA to meet sequencing requirements and was also within the

exponential phase of the PCR, identified by a lower signal intensity of the amplicon after 20 cycles compared to 40 cycles (Figure 5.3 B, C).

Each combination of barcoded primers that was ordered resulted in a single amplicon product (evidenced in Appendix 4). Once the barcoded primers had all been tested, DNA extracted from the bacterial cultures at every time point from each screen was amplified using a unique set of forward and reverse primers. In total, 44 unique forward and reverse primer combinations were used to amplify DNA that was sampled from 7 continuous co-cultivation assays. This comprised four technical replicates of the first biological replicate and three technical replicates of the second biological replicate of a 6His::PEPC variant screen that encompassed variants from all 21 single site mutagenesis libraries together. The PCR product was purified and equal quantities of DNA from each PCR were pooled and sent for sequencing.

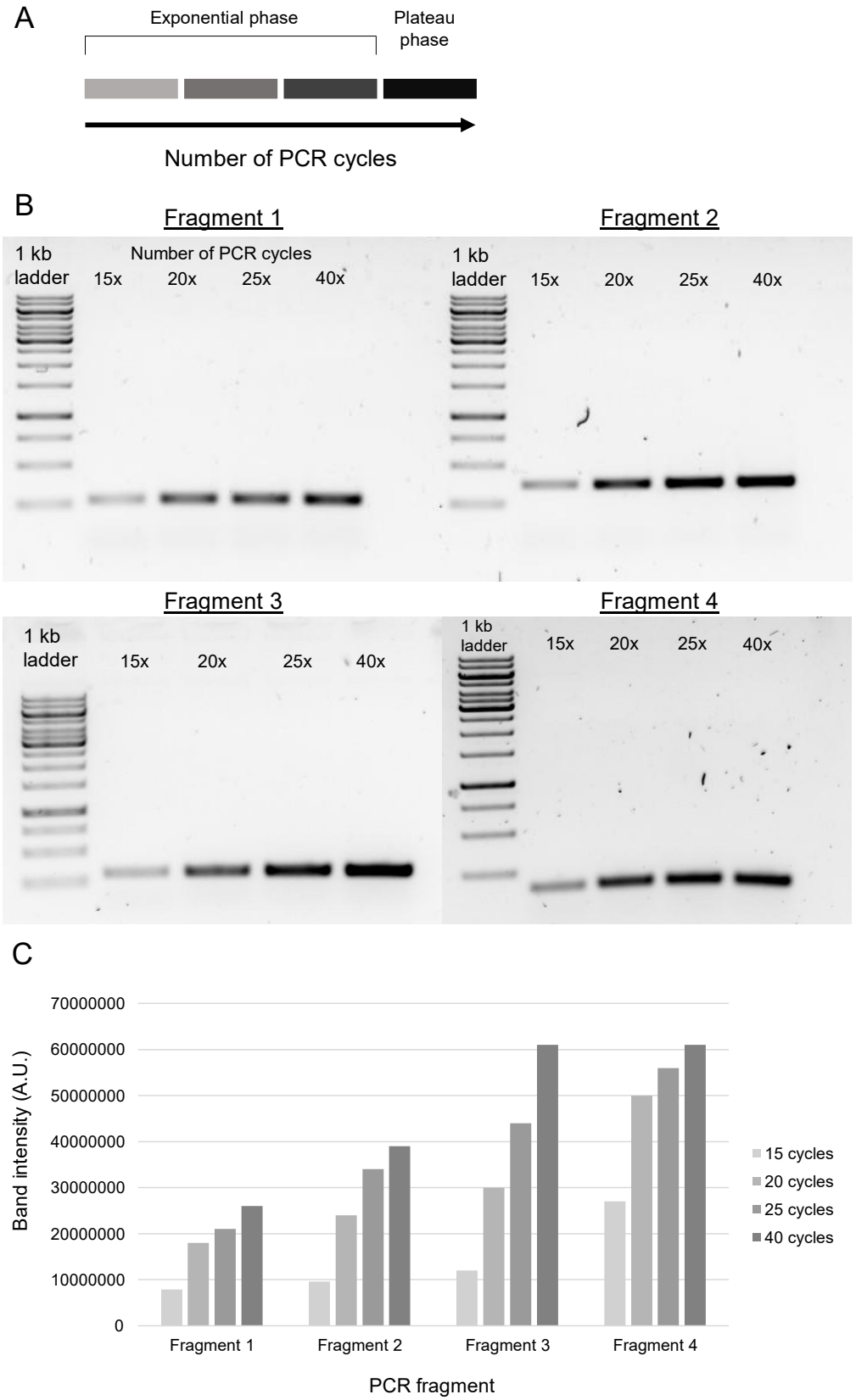


Figure 5.3: PCRs for each fragment of DNA established the number of cycles that could be undertaken before the PCR is fully saturated. Figure legend continued on the next page.

Figure 5.3 (continued) (A) A cartoon representation of PCR amplicon band intensity, as visualised on a gel electrophoresis, increasing with an increasing number of PCR cycles until a plateau is reached where for every additional cycle the band intensity remains constant. (B) Gel electrophoresis results of PCRs undertaken for fragments 1 to 4 that between them encompassed all 21 sites under selection in C_4 PEPC along the *6His::PEPC* gene sequence, undertaken for an increasing number of cycles. (C) The band intensity of each band in the gel electrophoreses in part (B) quantified by the Invitrogen iBright CL1500.

5.2.1.2 Bioinformatic analysis of the NGS reads

The process by which information was extracted from the sequencing reads is described in brief below. First, as there were only four PCR amplicons, each read was uniquely mapped to its amplicon by substring matching to the first 18 nucleotides of the amplicon (Figure 5.4, part 1). The codons at the sites under evaluation in each amplicon were then extracted (Figure 5.4, part 2). The barcode information encoded in the forward and reverse primers was then used to assign each codon variant to a particular time point and replicate number (Figure 5.4, part 3), and the information collated such that the abundance of all variants could be evaluated (Figure 5.4, part 4). In total, 4.8 million read pairs were processed in this manner, which resulted in a total count of 20,843,805 codons at the 21 sites that were subject to mutagenesis. The relative abundance of each variant at each position was then evaluated (Figure 5.4, parts 5 and 6).

6His::PEPC variants could be tracked over time in *E. coli* culture and changes in the relative abundance of an individual variant quantified, even in a complex bacterial culture containing hundreds of variants, through the sequencing protocol outlined above (Figure 5.5).

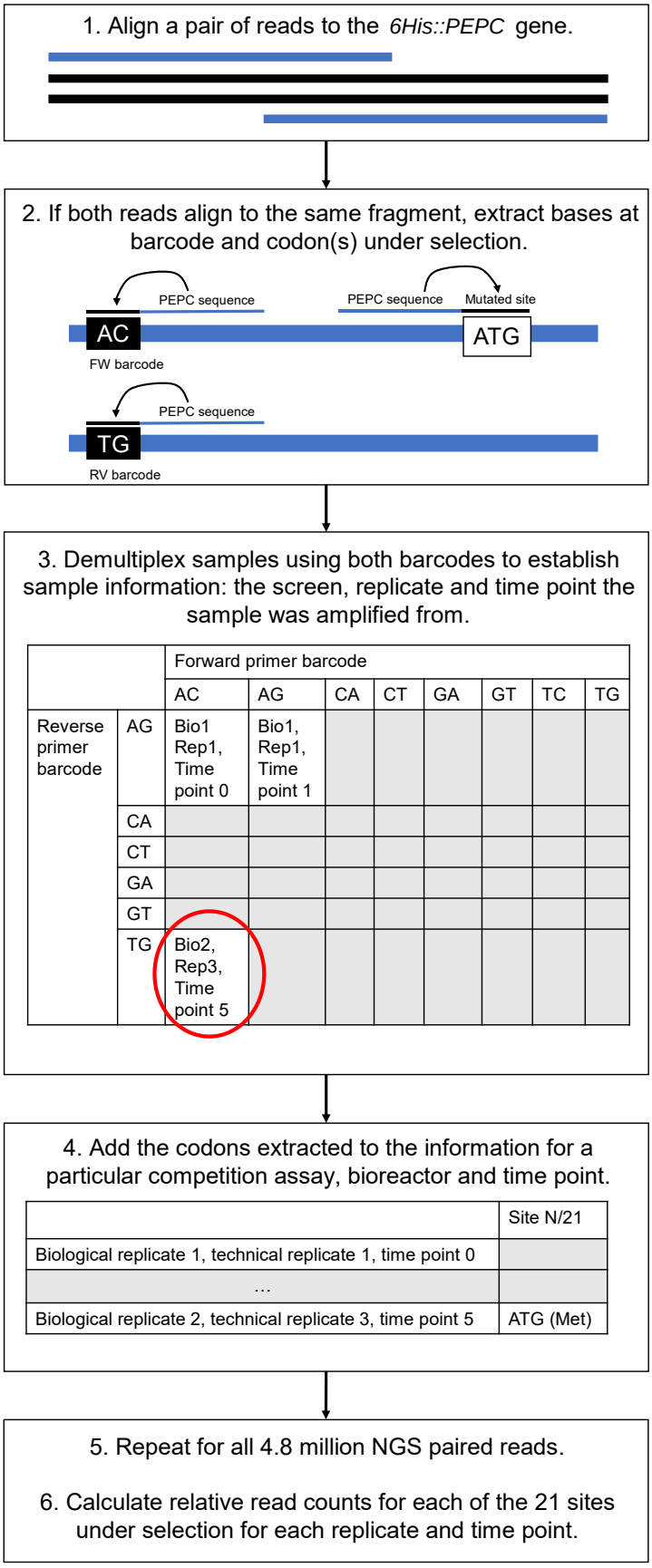


Figure 5.4: The bioinformatic pipeline developed to process NGS reads. Bio1 denotes biological replicate 1. Rep1 denotes technical replicate 1.

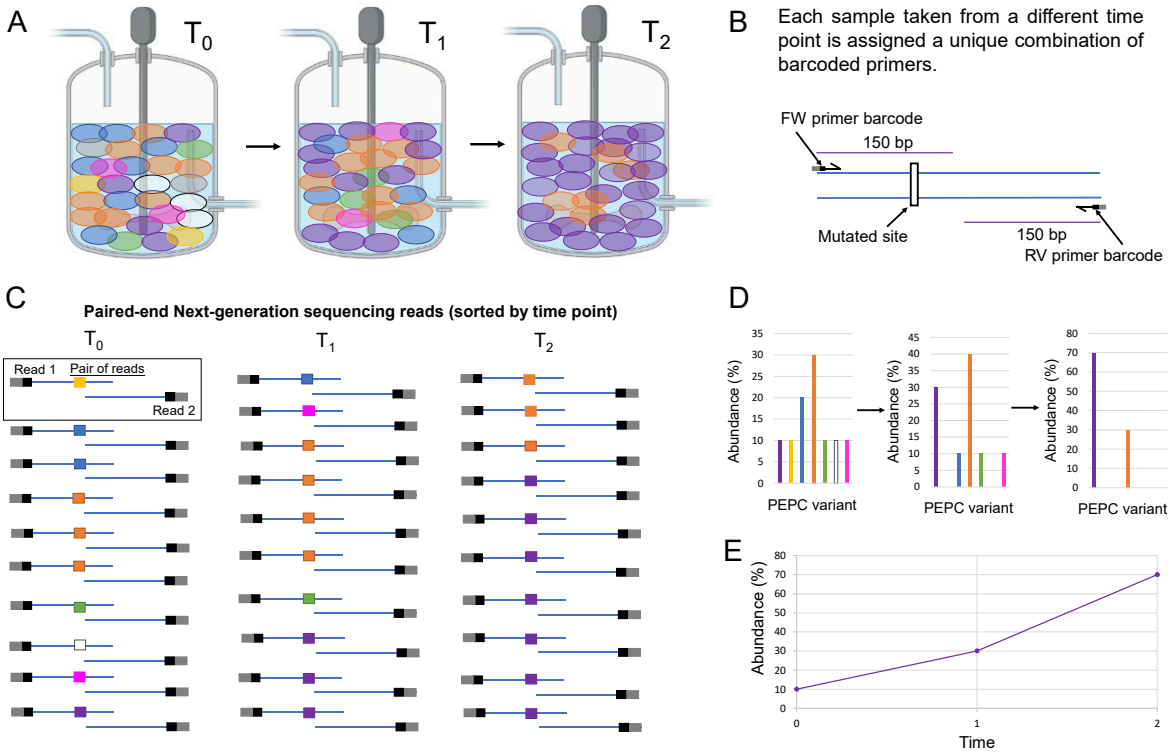


Figure 5.5: 6His::PEPC variant abundance can be tracked over time *via* an amplification, NGS and analysis pipeline. (A) Bacteria that express one of many 6His::PEPC variants are co-cultured (different variants having been transformed into bacteria indicated by different colour ovals). (B) DNA prepared from a sample taken at a specific time point during the co-culture is amplified using barcoded primers. Amplicons contain a site(s) under selection and the barcodes denote the sample metadata. (C) Sequencing data can be assigned to the time point that the sample DNA that the PCR product was amplified from was taken. The abundance of multiple 6His::PEPC variants (D), or a single 6His::PEPC variant (E) can be tracked in the population over time.

5.2.2 A proof of concept screen of 20 6His::PEPC variants identifies a 6His::PEPC variant that confers a competitive advantage to *E. coli* in continuous co-culture

5.2.2.1 A screen of a library of *6His::PEPC* variants mutated at site 780 for improved enzymatic activity in *E. coli*

A single site was selected to validate the library sequencing approach described above in a complete end-to-end implementation of a co-cultivation assay. Here, site 780 (*Z. mays* numbering) was selected as it is known to be a major determinant of C_4 PEPC kinetics. Most C_3 PEPCs have an alanine at this position whereas all known C_4 PEPCs have a serine at this position and the transition from alanine to serine at this position is required for C_4 PEPC kinetics (Bläsing, Westhoff and Svensson, 2000). Mutagenesis experiments have shown that C_4 PEPC subject to the mutation S780A transition towards more C_3-like PEPC kinetics, namely a lower affinity for bicarbonate in *Flaveria trinervia* C_4 PEPC (residue 774 in *F. trinervia*) (DiMario and Cousins, 2019) and a higher affinity for PEP in *Z. mays* and *F. trinervia* C_4 PEPC (Bläsing, Westhoff and Svensson, 2000; Engelmann, Bläsing and Westhoff, 2002; Endo *et al.*, 2008). Thus, it was expected that if the screen is selecting for optimal PEPCs with selection pressures similar to those *in planta* then S780 and A780 variants would become the more abundant during the experiment since these amino acids have been selected for at site 780 in C_3 and C_4 PEPC. Furthermore, it was unknown whether any other residues would also be functional at this position, or how they would perform relative to alanine or serine.

E. coli Δppc was transformed with a library of *6His::PEPC* variants generated at site 780 and held in the early logarithmic phase of growth in M9 glucose liquid culture where growth rate dependent on the enzymatic rate of 6His::PEPC (Figure 5.6).

The first sample (T_0) was prepared from the culture used to inoculate all three technical replicates of the screen. Subsequent 5 mL samples were taken 20, 30, 43, 53 and 68 hours post inoculation of the co-culture and the samples were replaced with an equivalent volume of fresh M9 glucose media. Fresh M9 glucose with the selective antibiotic was supplied continuously to maintain an OD of 0.43 and therefore maintain the cells in the logarithmic phase of growth over the duration of the screen. The sampled cultures were subject to plasmid DNA extraction and processed according to sample preparation and bioinformatic analysis pipeline described above.

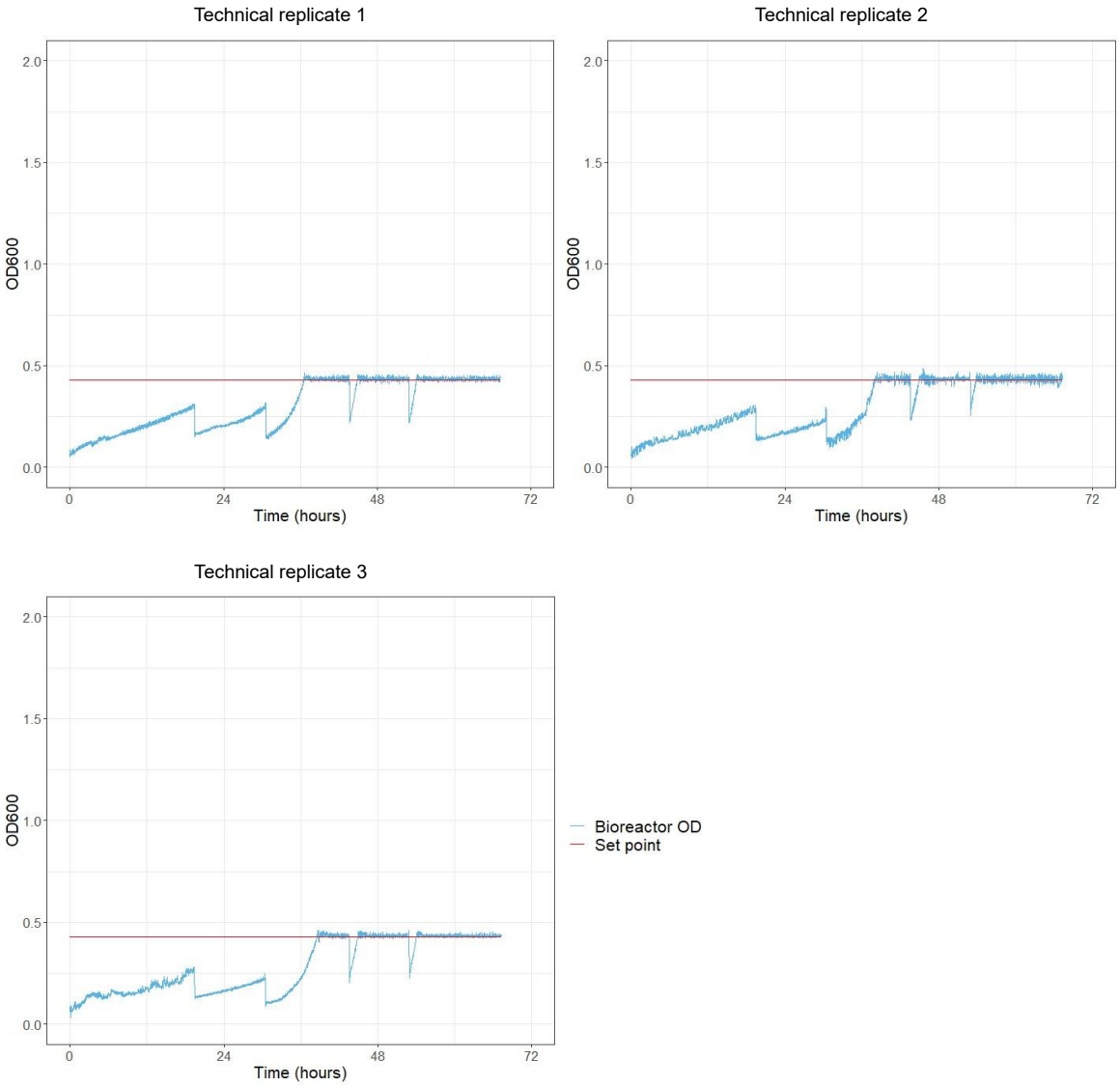


Figure 5.6: The three technical replicates of *E. coli* ppc transformed with *6His::PEPC* variants mutated at site 780 that were held indefinitely in the logarithmic phase of growth. Samples were taken at 20, 30, 43, 53 and 68 hours and the sample volume replaced by fresh media.

5.2.2.2 A S780G mutation in 6His::PEPC confers a competitive advantage to bacteria dependent on 6His::PEPC for growth

Following site saturating mutagenesis of site 780, 19 out of a possible 20 variants were present as determined by NGS (Figure 5.7). The only possible variant missing from the initial sample of *6His::PEPC* variants was 6His::PEPC with glutamine at site 780. It appears that either the mutagenesis approach did not result in a glutamine (Q) variant being generated, or that it was eliminated during the pre-culture step in conditions where 6His::PEPC activity should not have been limiting. It should be noted that none of the available plant PEPC sequences on NCBI have a glutamine residue at this position (Figure 5.8). However, some species of bacteria including *Methylobacterium*, *Erythrobacter* and *Acetobacter* do have glutamine at this position in PEPC, and these sequences account for ~2.8% of the 39,581 unique PEPC sequences from across the tree of life that are available from NCBI. Thus, glutamine at site 780 is likely not prohibitive for PEPC function and instead it is likely that the mutagenesis approach was unsuccessful in generating this 6His::PEPC variant. Taken together, the screening pipeline has resulted in successful generation and transformation of 95% possible *6His::PEPC* variants mutated at a single site into *E. coli* Δppc to participate in the screen, including variants that have alanine and serine at position 780.

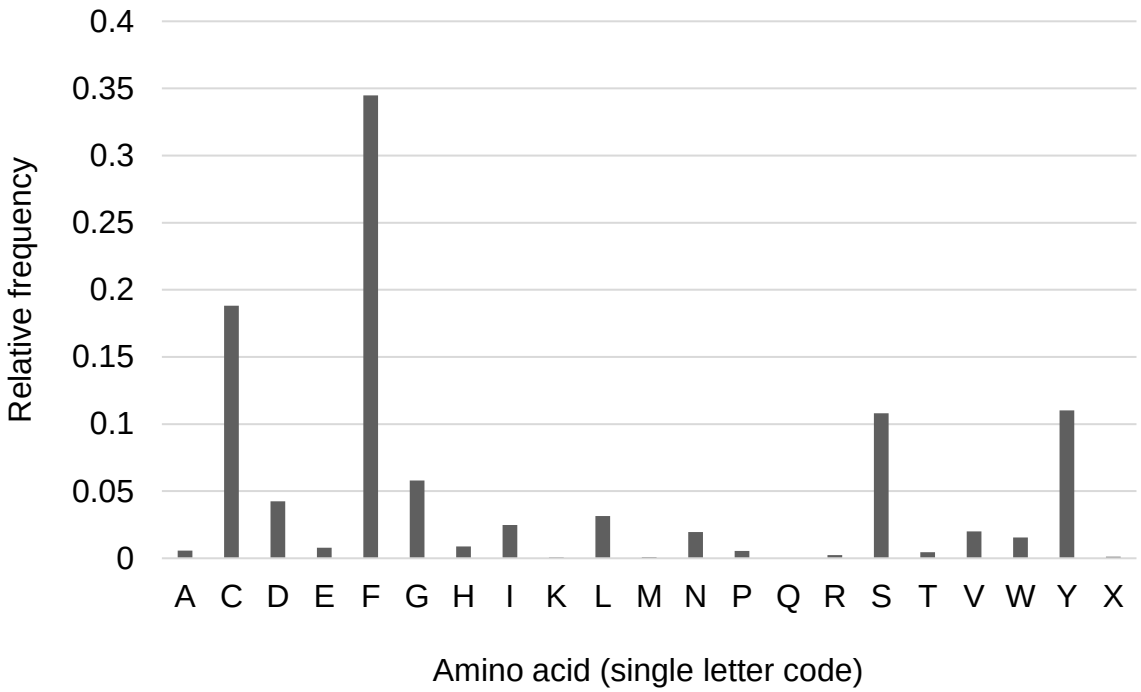


Figure 5.7: 6His::PEPC variants generated at site 780 (*Z. mays* numbering) that were present at the start of the *E. coli* screen. Each variant is represented by the single letter code of the amino acid at position 780. X represents a stop codon.

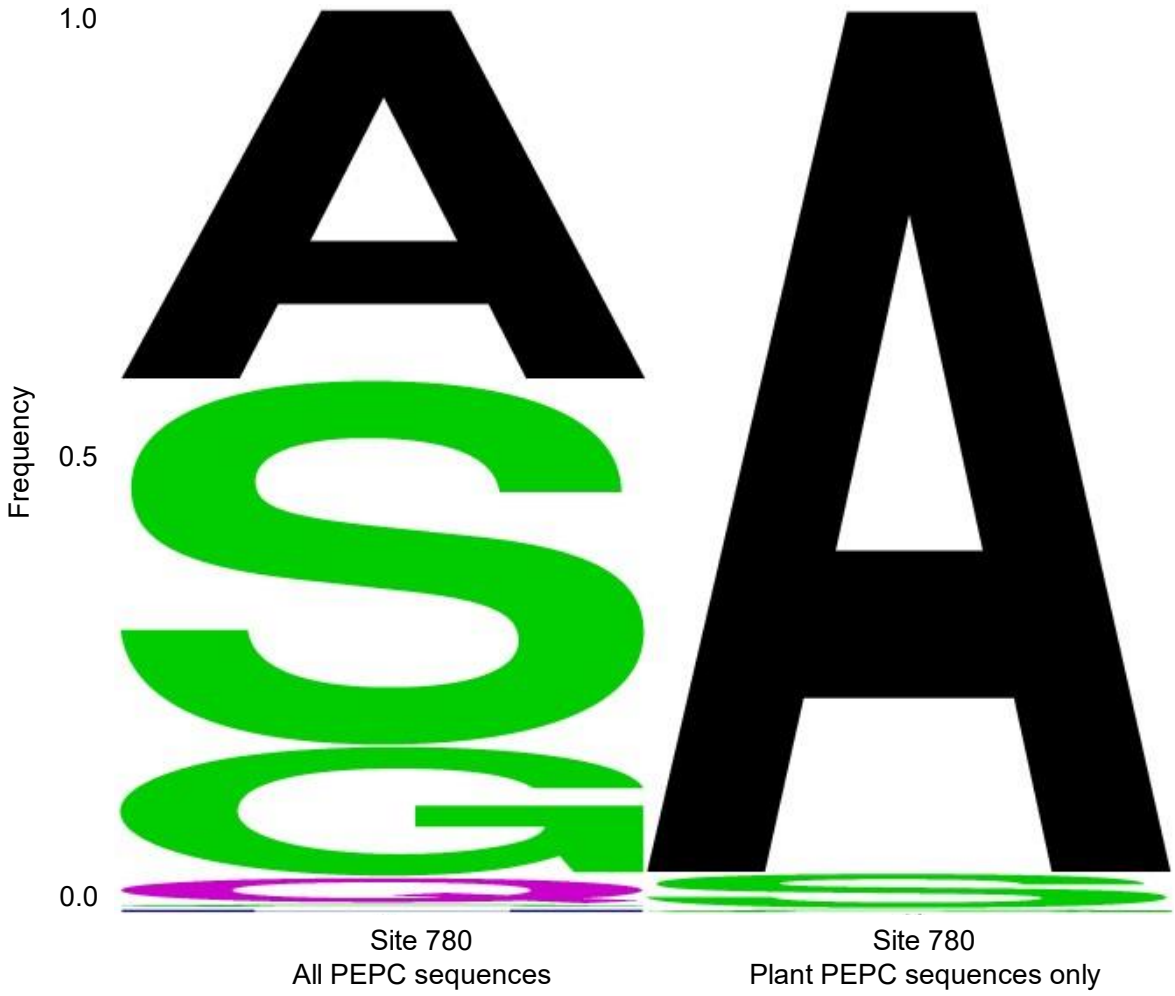


Figure 5.8: Frequency of amino acids at site 780 (*Z. mays* numbering) from PEPC sequences across the tree of life and from plant PEPC sequences. Plant PEPC sequences are a subset of all the sequences. Polar amino acids and glycine are represented in green, basic amino acids in blue, neutral amino acids in purple, acidic amino acids in red and hydrophobic amino acids in black.

Variants of 6His::PEPC that have glycine, alanine and serine at the equivalent residue to site 780 performed well in the screen and all other variants decreased in relative abundance over the course of the screen (Figure 5.9). This result was observed in all three technical replicates that were inoculated from the same T_0 culture (Figure 5.9). *E. coli* PEPC has alanine at this position, not glycine, so the screen was not selecting for amino acids identical to *E. coli* PEPC within the context of the *Setaria* C_4 PEPC background for the mutation. Alanine and serine are well-characterised amino acids at site 780 in C_4 PEPC (Bläsing, Westhoff and Svensson, 2000; Engelmann, Bläsing and Westhoff, 2002; Endo *et al.*, 2008; DiMario and Cousins, 2019), but the effect of glycine in C_4 PEPC has not been previously tested so it would be interesting to discover the effect of glycine at this site in SvPEPC on enzyme kinetics.

In the 39,581 unique PEPC sequences sampled from across the tree of life, alanine, serine and glycine were the three most abundant amino acids at position 780, these correspond to 41.0%, 40.5% and 14.6% of the sequences, respectively (Figure 5.8). Of the 1,873 unique plant PEPC sequences in the dataset, 92.5% of sequences were alanine, 4.1% were serine and only 0.6% sequences were glycine at this position. Plant species that have a glycine residue at site 780 are two different species of *Marchantia* (liverworts), two species of moss (*Physcomitrium patens* and *Ceratodon purpureus*), and a charophytic alga *Klebsormidium nitens*. It is interesting that glycine performed well within the background of the 6His::PEPC enzyme sequence because it is very rare in plant sequences.

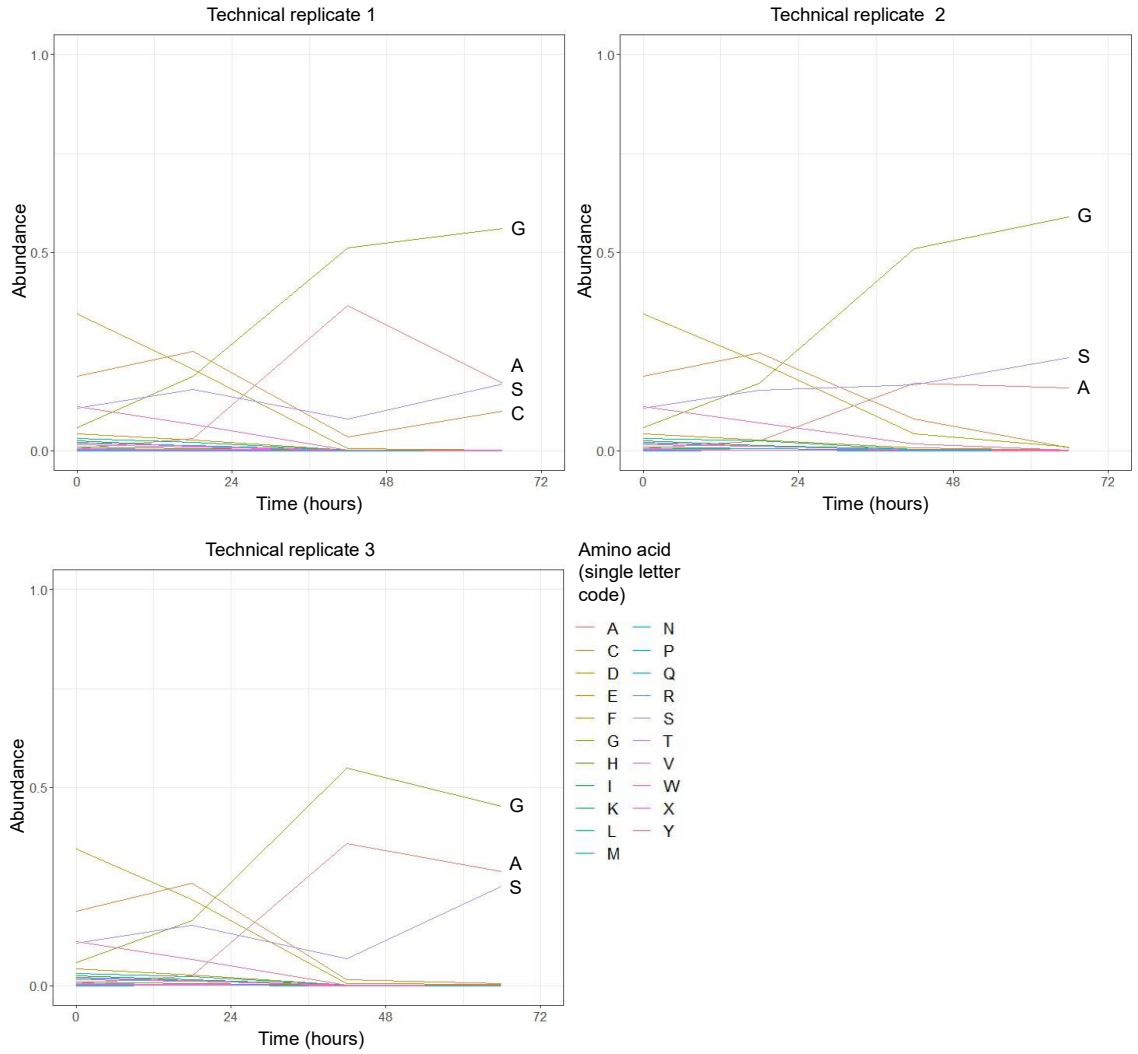


Figure 5.9: The abundance of 6His::PEPC mutated at position 780 in the enzyme (*Z. mays* numbering) changed over time in continuous *E. coli* co-culture. Three technical replicates inoculated from the same starting culture are shown. X represents a stop codon.

Alanine, serine and glycine together comprise 96.1% of the variation of sequences analysed across the tree of life at site 780 (Figure 5.8). The small size of these three amino acids suggests that the size of amino acid at this residue could be influential in residue biochemistry. The site is deep within the protein and near to another alpha helix, so perhaps size of the amino acid side-chain is restrictive (Figure 5.10). The presence of glycine at residue 780 in nature is surprising given that the residue is in an alpha helix and the helical penalty of having glycine in an alpha helix is 1 kcal/mol, whereas the helical penalty of an alanine or serine are 0 or 0.5, kcal/mol respectively (Pace and Scholtz, 1998). Thus, glycine is predicted to be disruptive to the alpha helix structure compared to alanine or serine (López-Llano *et al.*, 2006).

Alanine and glycine are hydrophobic, non-polar amino acids, whereas serine has a polar R group, so potentially in C_4 biochemistry this polarity is influential in the enzyme kinetics, including the affinity of PEPC for bicarbonate which increases as the A780S mutation is made (DiMario and Cousins, 2019). In the crystal structure of the activated enzyme (PDB: 5VYJ), the O-H in the serine R group does not appear to be involved in hydrogen bonding to any other R groups so it could be influential in catalysis instead. Given that changing this site to serine increases the affinity of the enzyme for bicarbonate in C_4 PEPC, it is possible that the hydrogen bonding occurs with another residue during catalysis conformational change. It would therefore be interesting to investigate the impact of non-polar glycine at residue 780 of SvPEPC on enzyme kinetics, including catalytic capacity (K_{cat}) and the affinity of the enzyme for bicarbonate, to gain further insight into why this enzyme variant was selected for in the screen and predict the impact of this variant in *S. viridis*.

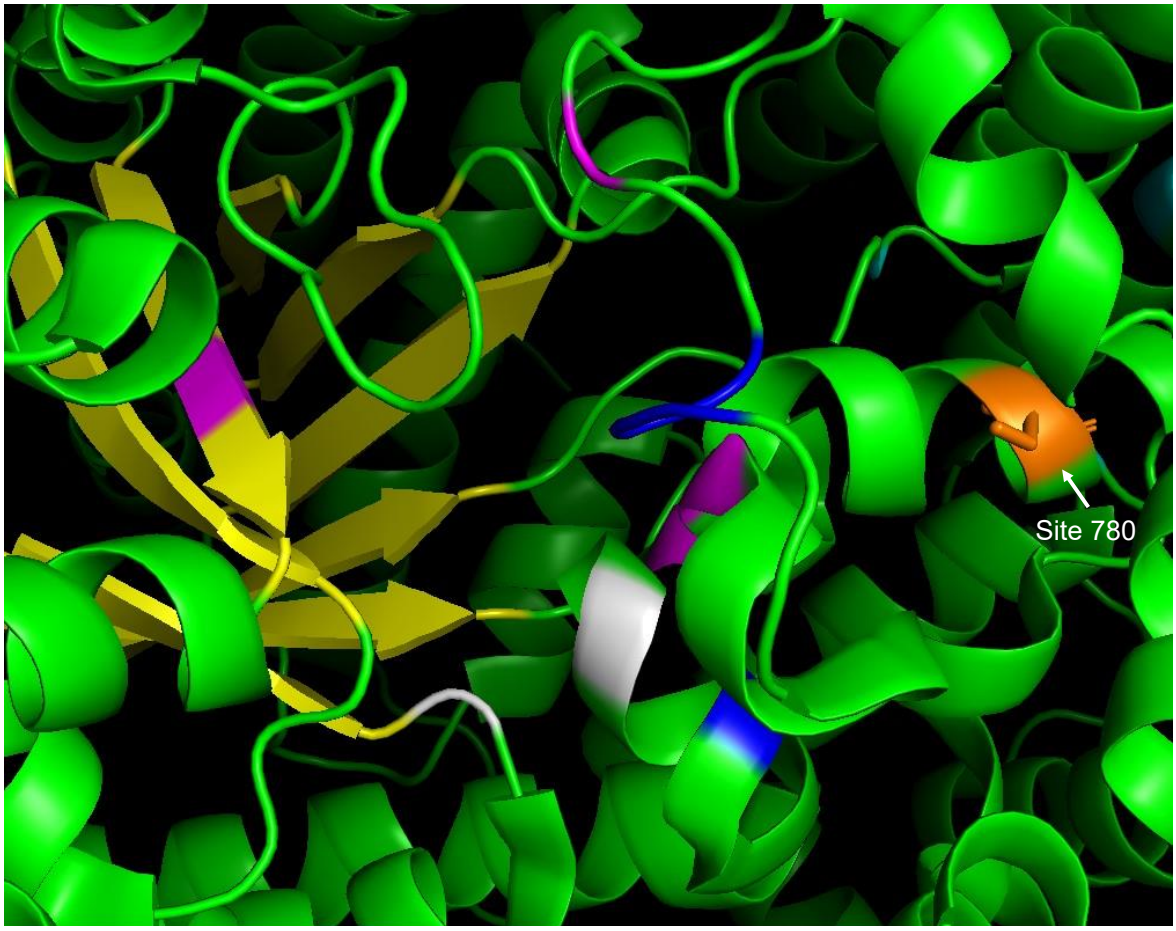


Figure 5.10: Site 780 is close to the C-terminal region of the β barrel and a mobile loop involved in catalysis. β barrel represented in yellow, PEP binding residues coloured in pink, magnesium binding residues in grey, bicarbonate binding residues in dark blue, aspartate/malate binding residues in light blue and site 780 in orange. PDB: 5VYJ.

Despite being common in bacteria, only one PEPC that contains glycine at position 780 has been characterised. This enzyme comes from *Thiobacillus thioparus* and has a $K_mHCO_3^-$ of 860 µM (Hoban and Lyric, 1975). It is hard to assess this value in the context of plant PEPCs, as substitution of serine for alanine in *F. trinervia* C_4 PEPC resulted in a change in $K_mHCO_3^-$ from 26.6 µM to 38.6 µM (DiMario and Cousins, 2019). Given the large difference between the values observed for the plant and bacterial enzymes, it is impossible to predict what impact glycine may have on the enzyme kinetics, so a single site mutation must be made to SvPEPC and the kinetics compared to the wild-type enzyme.

The data was clear that these three mutant variants performed well across all three replicates while all other variants disappeared. Thus, the next step was to evaluate how the kinetics of the enzyme had been altered by the change to glycine at position 780. If the enzyme kinetics were improved, then this would be an important mutation to try and reproduce in *Setaria* and crop plants. Furthermore, it would be interesting to determine how this glycine variant performs when competed against a larger pool of enzyme variants when all sites under selection are screened at once.

5.2.3 Identification of best-performing 6His::PEPC variants mutated at sites under selection in C_4 PEPC

5.2.3.1 A screen of 6His::PEPC variants for improved enzymatic activity in *E. coli*

Variant libraries generated at every site under selection in C_4 PEPC (Chapter 4.2.2) were cloned into the placeholder vector (Chapter 4.2.4) and transformed into *E. coli* Δppc. The population of bacteria was held in the logarithmic phase of growth for 72 hours in M9 glucose where 6His::PEPC activity was growth rate-limiting (Chapter 3.2.4). In total, 347 6His::PEPC variants were competed against each other over the course of the 72 hour screen (Figure 5.11). This set represents 82.6% of the 420 possible variants that could have been captured through mutagenesis of 21 sites independently. In total, 323 variants were analysed in the first biological replicate across all four technical replicates, and 325 variants were analysed in the second biological replicate across three technical replicates (Figure 5.11). Of the full set of 347 variants, 301 were present in one or more technical replicates of both biological replicates, of which all 20 possible amino acids were tested at sites 502, 596, 637, 665, 794, 807, 813 and 863 across both replicates.

Each culture was held in the log phase for at least 60 hours over the 72 hour experiment, with each replicate taking ~12 hours to reach the cell density at which the cell culture was maintained (Figure 5.12). The culture doubled increasingly quickly over time in all seven bioreactors, reflected in the increasing speed at which the cells returned to an OD of 0.43 once samples were taken and also the larger volume required to maintain the cells in log phase; during the second biological replicate the bioreactors used up fresh media at an average rate of 11.1 mL/hour/bioreactor between 26 and 50 hours which increased to 13.64

mL/hour/bioreactor between 50 and 72 hours. Faster growth of the bacterial culture over time is concordant with 6His::PEPC variants that confer a faster rate of *E. coli* growth becoming more highly represented in the culture over time.

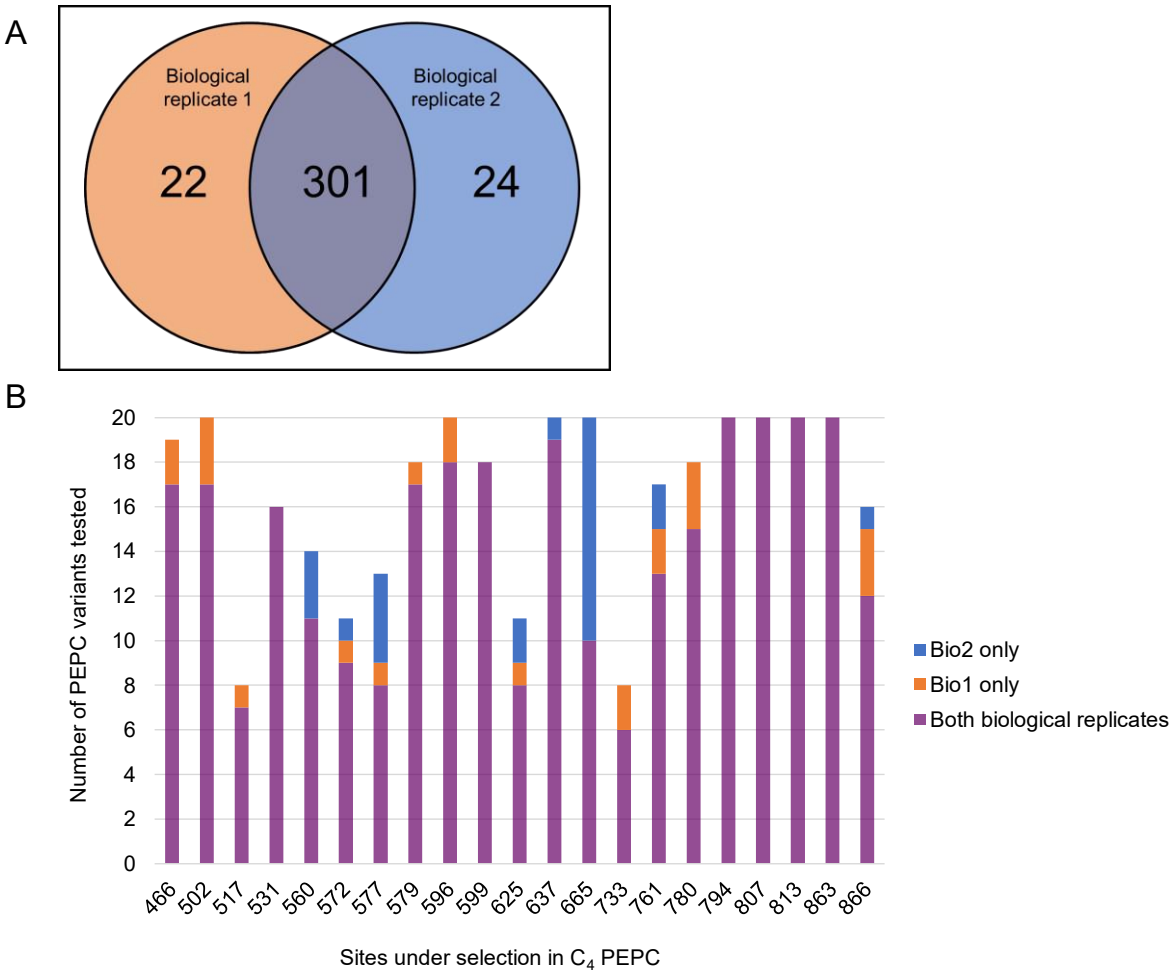


Figure 5.11: The number of 6His::PEPC variants competed in the screen shown by biological replicate (Bio) 1 and 2. (A) The total number of 6His::PEPC variants in a biological replicate. (B) The number of different variants generated at every site under selection (*Z. mays* residue numbers) across both biological replicates. For information on which 6His::PEPC variants were subject to co-culture, see Appendix 2.

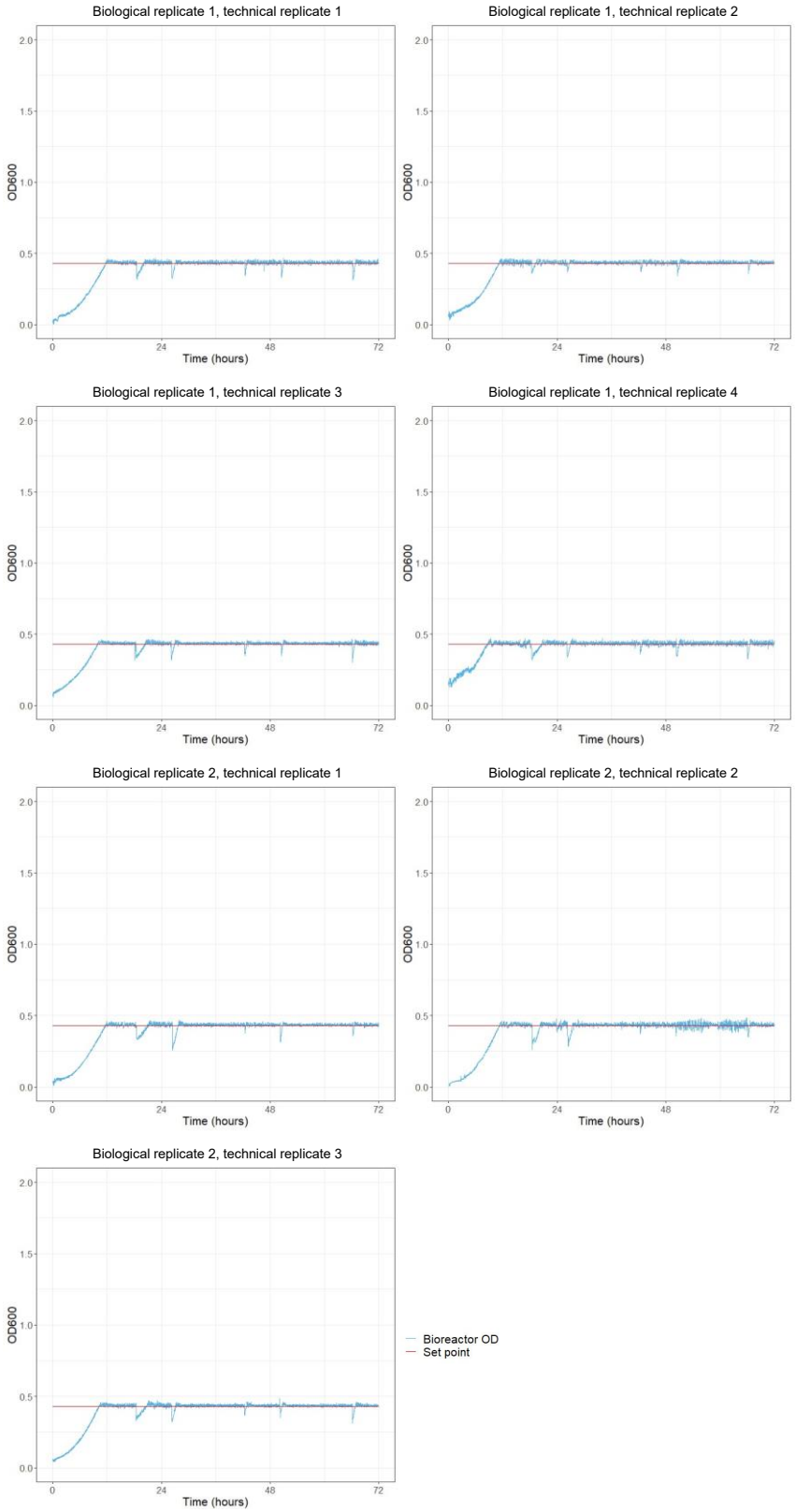


Figure 5.12: *E. coli* cultures in all seven replicates of the 6His::PEPC variant screen were held in the logarithmic phase of growth over the course of the screen.

5.2.3.2 P531A, F794V and N863K mutations confer a competitive advantage to *E. coli* dependent on 6His::PEPC for growth in M9 glucose

The top-performing variants as assessed by final abundance across all biological and technical replicates were P531A, F794V and N863K (numbered according to the equivalent residue in *Z. mays*) (Table 5.1, Figures 5.13 and 5.14). In addition, an A761N variant and the S780G variant discussed in the previous section comprised 10.4% and 12.6% final abundance in a single replicate, respectively (Table 5.1, Figures 5.13 and 5.15). However, the isolated incidence of this result suggested that these variants did not confer as substantial a growth advantage as variants P531A, F794V and N863K. It is worth noting here that successful variants contained wild-type amino acids at all other sites along the length of the *6His::PEPC* gene. Thus, the success of these three major variants will cause other less advantageous variants to appear less competitive compared to the wild-type sequence.

Table 5.1: All variants that finished with an abundance that comprised >5% of the population across seven replicates of a screen of 6His::PEPC variants.

Biological replicate	Technical replicate	6His::PEPC variant	Abundance as a proportion of the population after 72 hours (%)
1	0	P531A	57.11
1	0	F794V	16.12
1	0	N863K	11.34
1	1	N863K	27.09
1	1	P531A	26.90
1	1	F794V	6.09
1	2	P531A	80.24
1	2	F794V	11.62
1	3	P531A	59.05
1	3	S780G	12.56
1	3	A761N	10.47
1	3	F794V	7.18
2	0	P531A	58.72
2	0	P531V	20.20
2	0	N863K	5.57
2	1	P531A	30.52
2	1	N863K	19.48
2	1	P531V	15.95
2	2	P531A	79.36
2	2	P531V	7.20

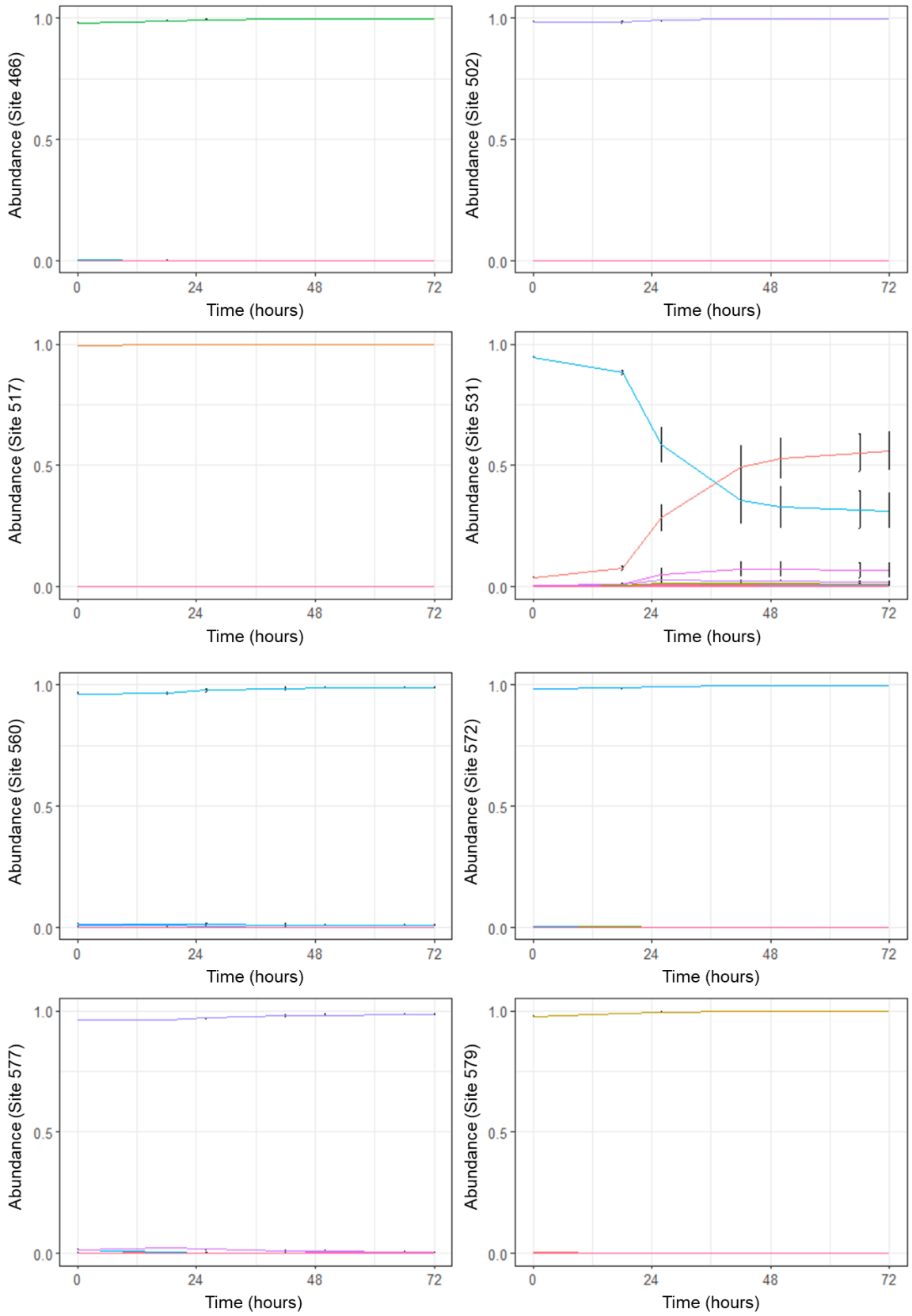


Figure 5.13 (I) Variation at sites 466 to 579 (1-8 out of 21 sites under selection in C_4 PEPC) (figure continued on next page)

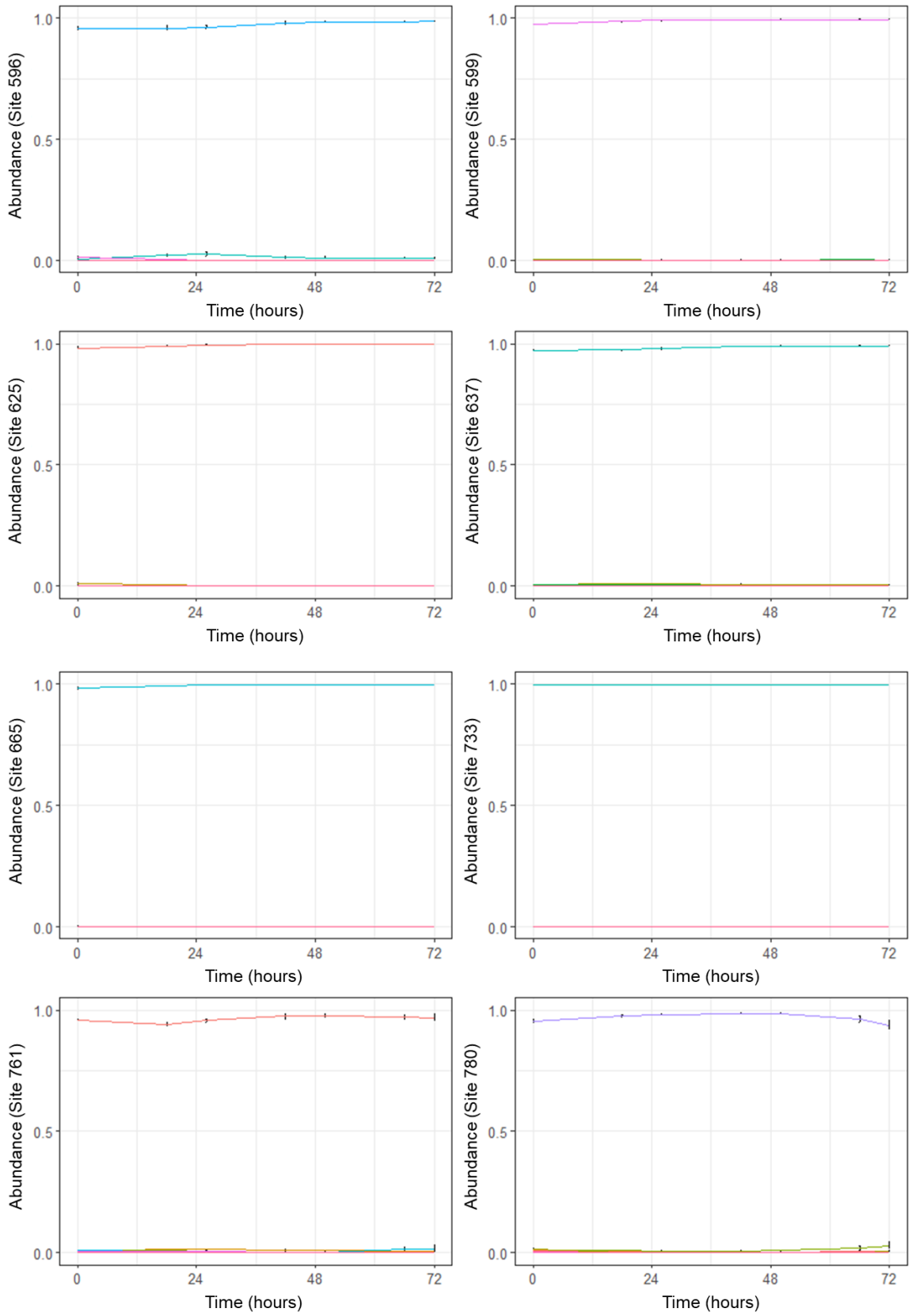


Figure 5.13 (II) Variation at sites 596 to 780 (9-16 out of 21 sites under selection in C_4 PEPC) (figure continued on next page)

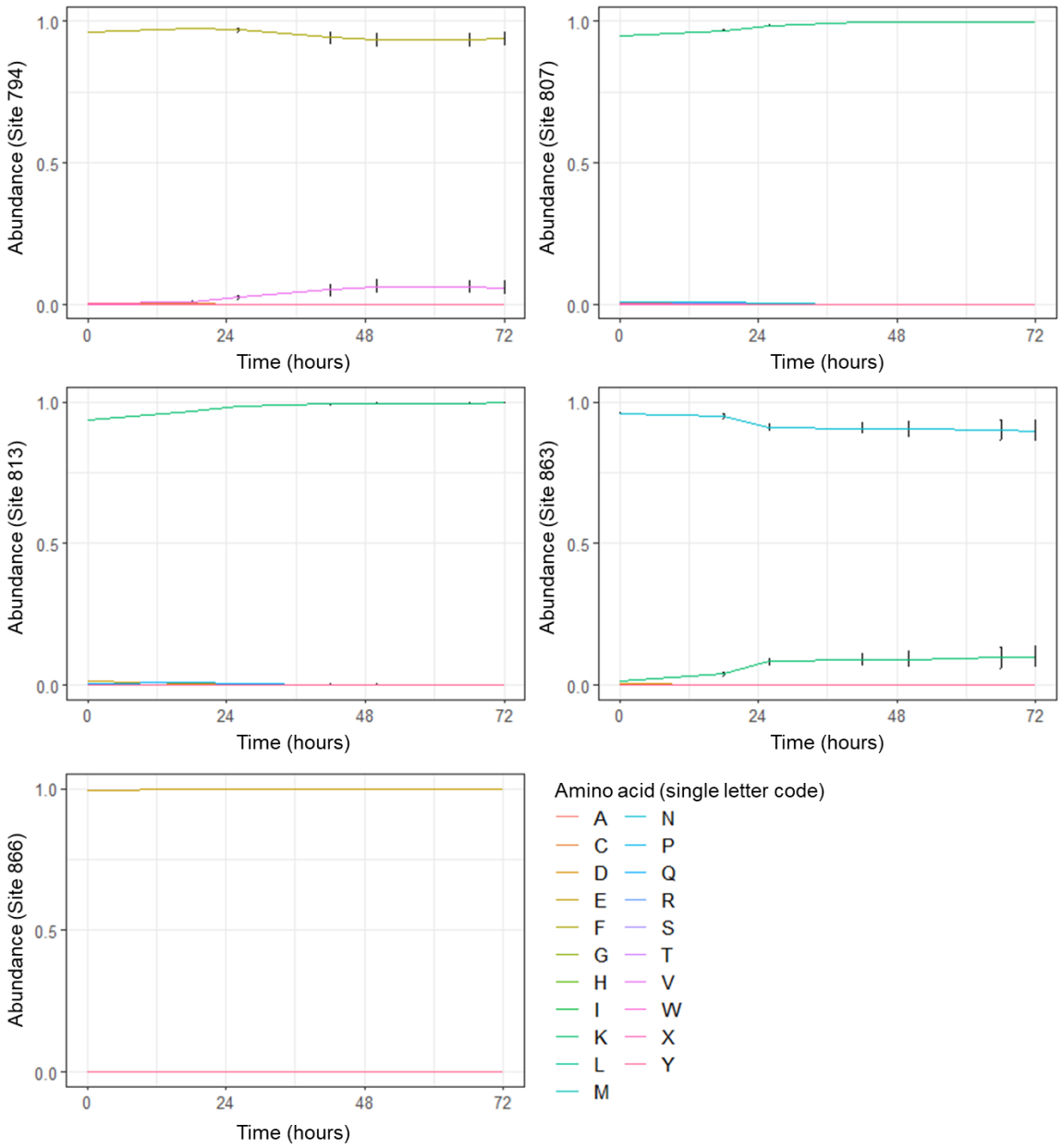


Figure 5.13 (III) Variation at sites 794 to 866 (17-21 out of 21 sites under selection in C_4 PEPC). Relative abundance of each variant is shown by site over the course of 72 hours and averaged across seven replicates. X represents a stop codon.

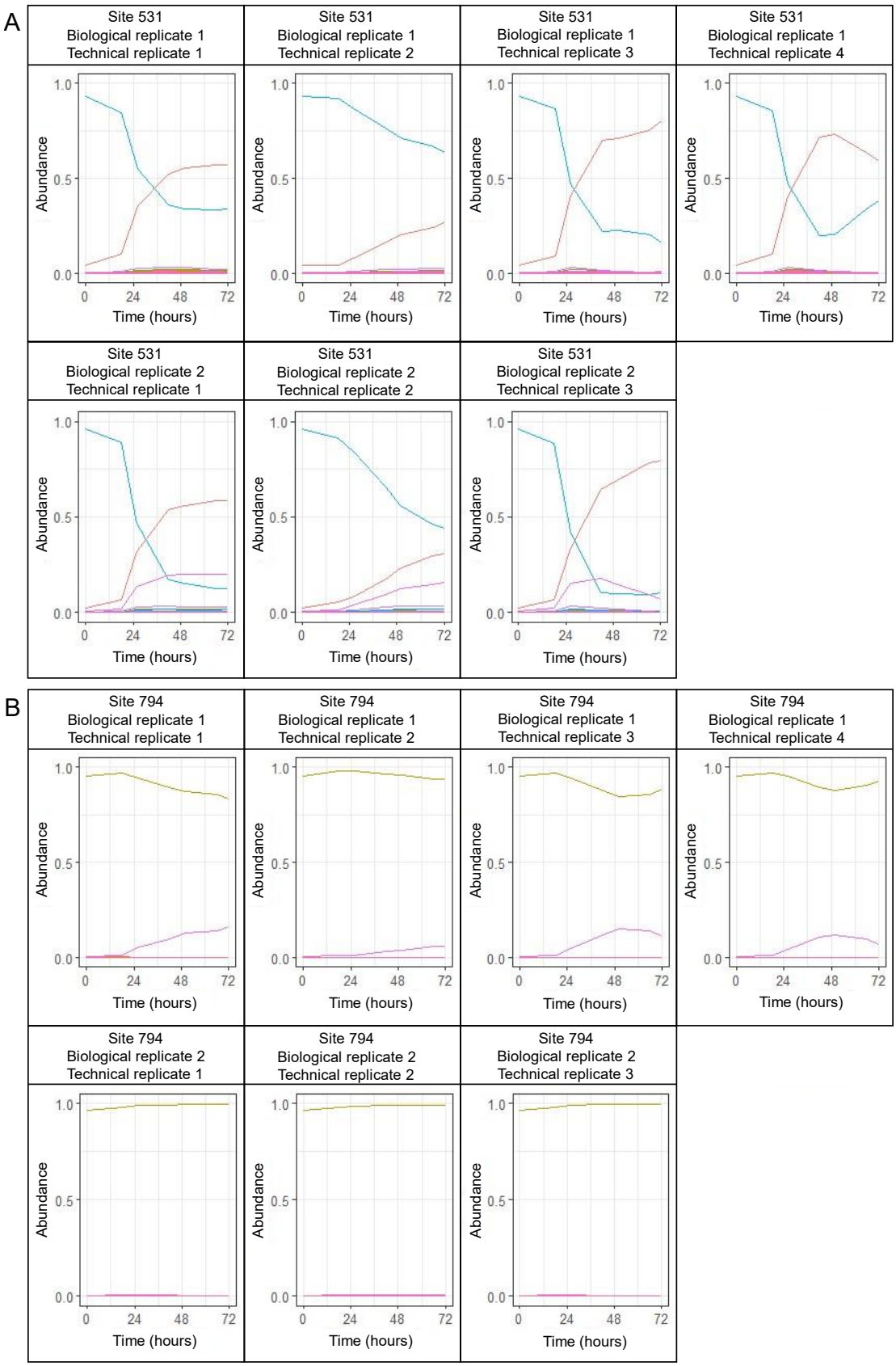


Figure 5.14 (I) Variation at sites 531 (A) and 794 (B) displayed by individual replicate (figure continued on next page)

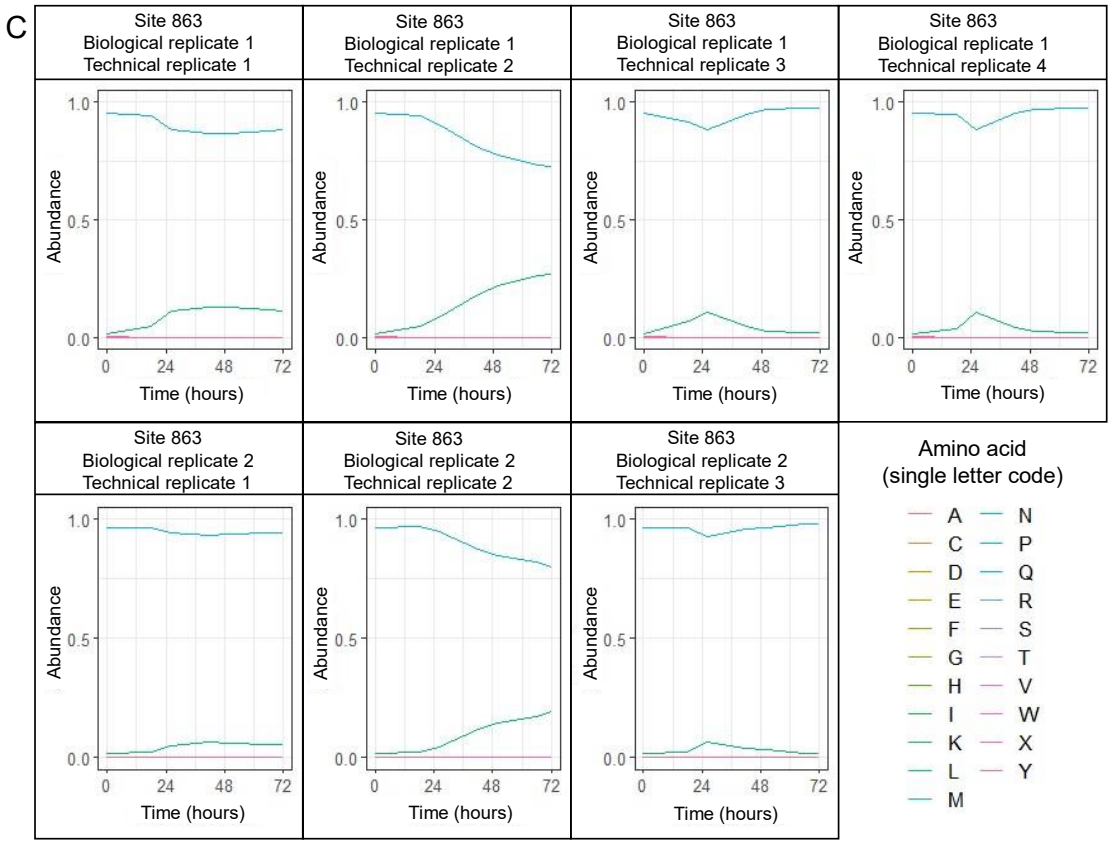


Figure 5.14 (II) Variation at site 863 displayed by individual replicate. X represents a stop codon.

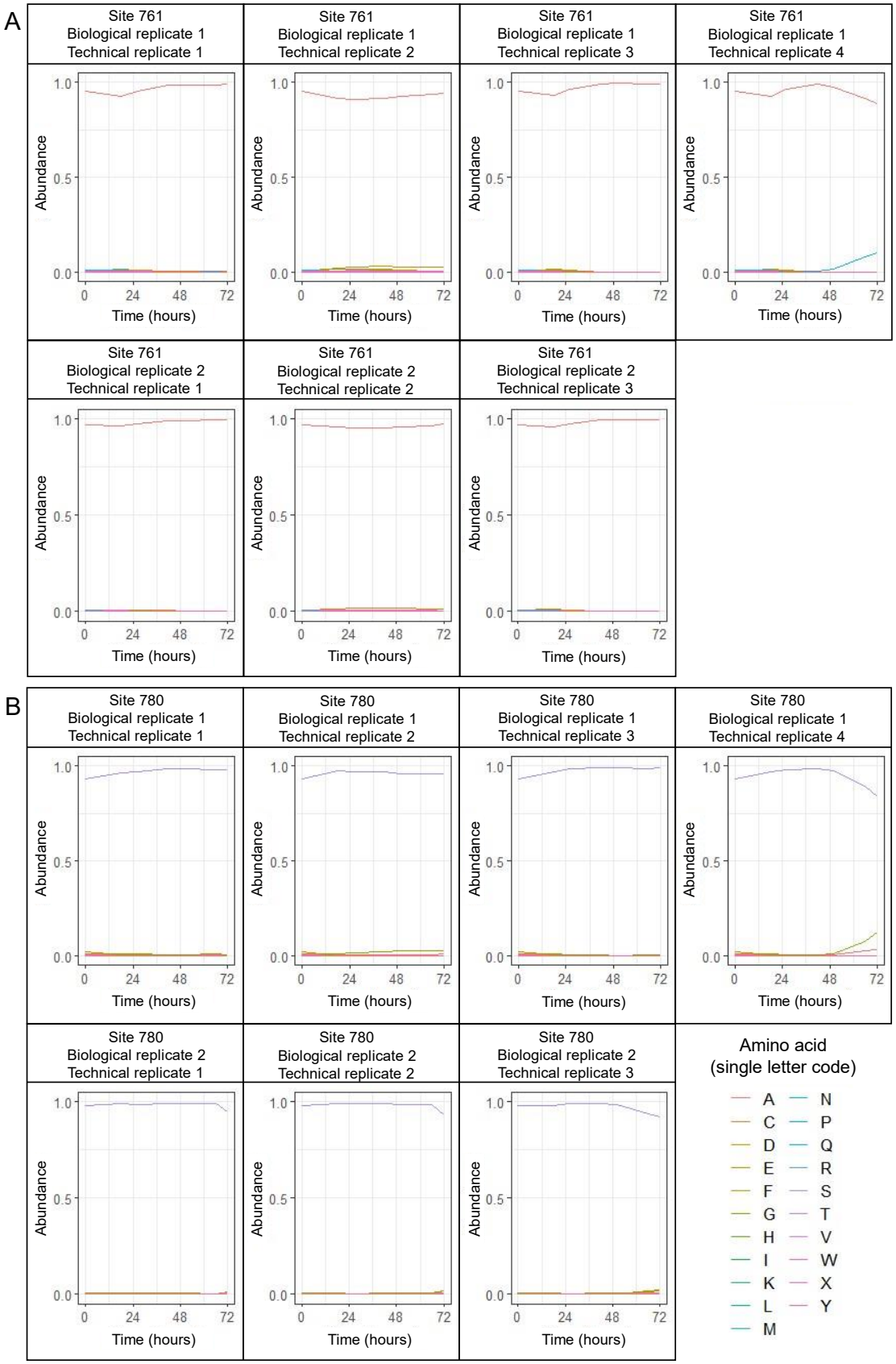


Figure 5.15: Variation at sites 761 (A) and 780 (B) increased but only in biological replicate 1, technical replicate 4. X represents a stop codon.

5.2.3.2.1 A P531A 6His::PEPC variant is most successful in the screen and has increased catalytic capacity

A 6His::PEPC P531A variant was consistently either the most or second-most abundant after 72 hours of continuous co-culture, and was observed at a relative abundance of ≥26.9% in every flask (Table 5.1, Figure 5.14 A). A P531V variant was also somewhat successful in the second biological replicate of the screen, ending at a relative abundance of >5% of the population in all three technical replicates (Table 5.1, Figure 5.14 A). Residue 531 is on a β sheet that comprises part of the eight-stranded β barrel (Figure 5.16). It is rare for a β sheet to contain a proline residue because the structure of proline is disruptive to the structure of a β sheet and the amino acid cannot complete the hydrogen bonding network (Li *et al.*, 1996; Bhattacharjee and Biswas, 2010). Despite this, proline has been selected for at site 531 in C_4 PEPC whereas alanine is typical at this site in non-C_4 PEPC (Christin *et al.*, 2007). Indeed, *Z. mays* C_4 PEPC and *E. coli* PEPC have proline and alanine at this position, respectively. The crystal structures of both enzymes show that the structure of the β sheet that contains residue 531 is similar, despite the proline residue at this position in the C_4 PEPC structure (Figure 5.17). Therefore, it appears that instead of a structural change, the greater flexibility and/or additional hydrogen bonding that alanine conferred at this residue in the P531A variant was beneficial for the catalytic capacity of the enzyme in *E. coli*.

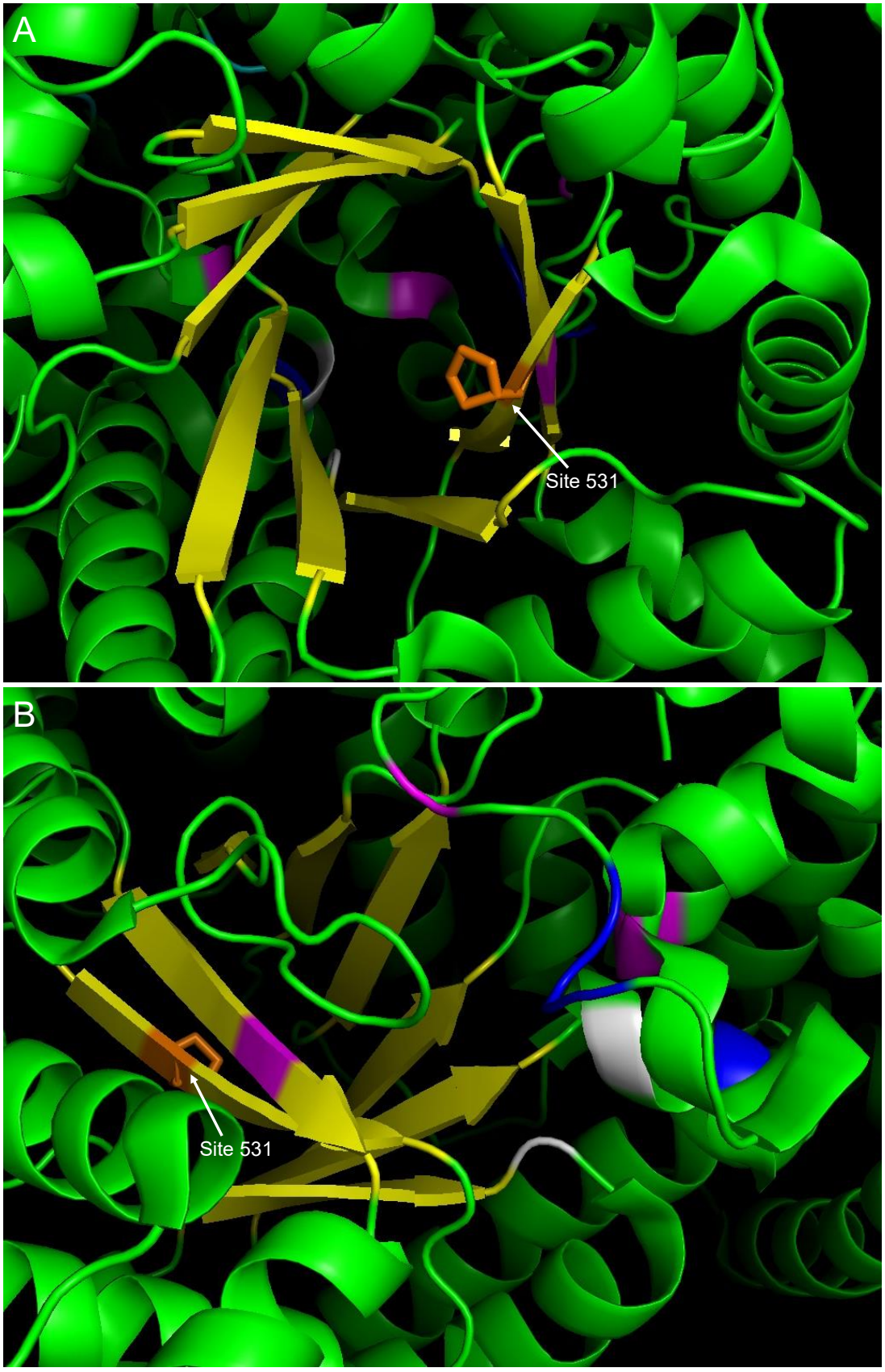


Figure 5.16: Site 531 is located in the β barrel, viewed from the N-terminal region (A) and C-terminal region (B). Figure legend continued on the next page.

Figure 5.16 (continued) The active site of the enzyme is at the C-terminal region of the β barrel. β barrel represented in yellow, PEP binding residues coloured in pink, magnesium binding residues in grey, bicarbonate binding residues in dark blue, aspartate/malate binding residues in light blue and site 531 in orange. PDB: 5VYJ.

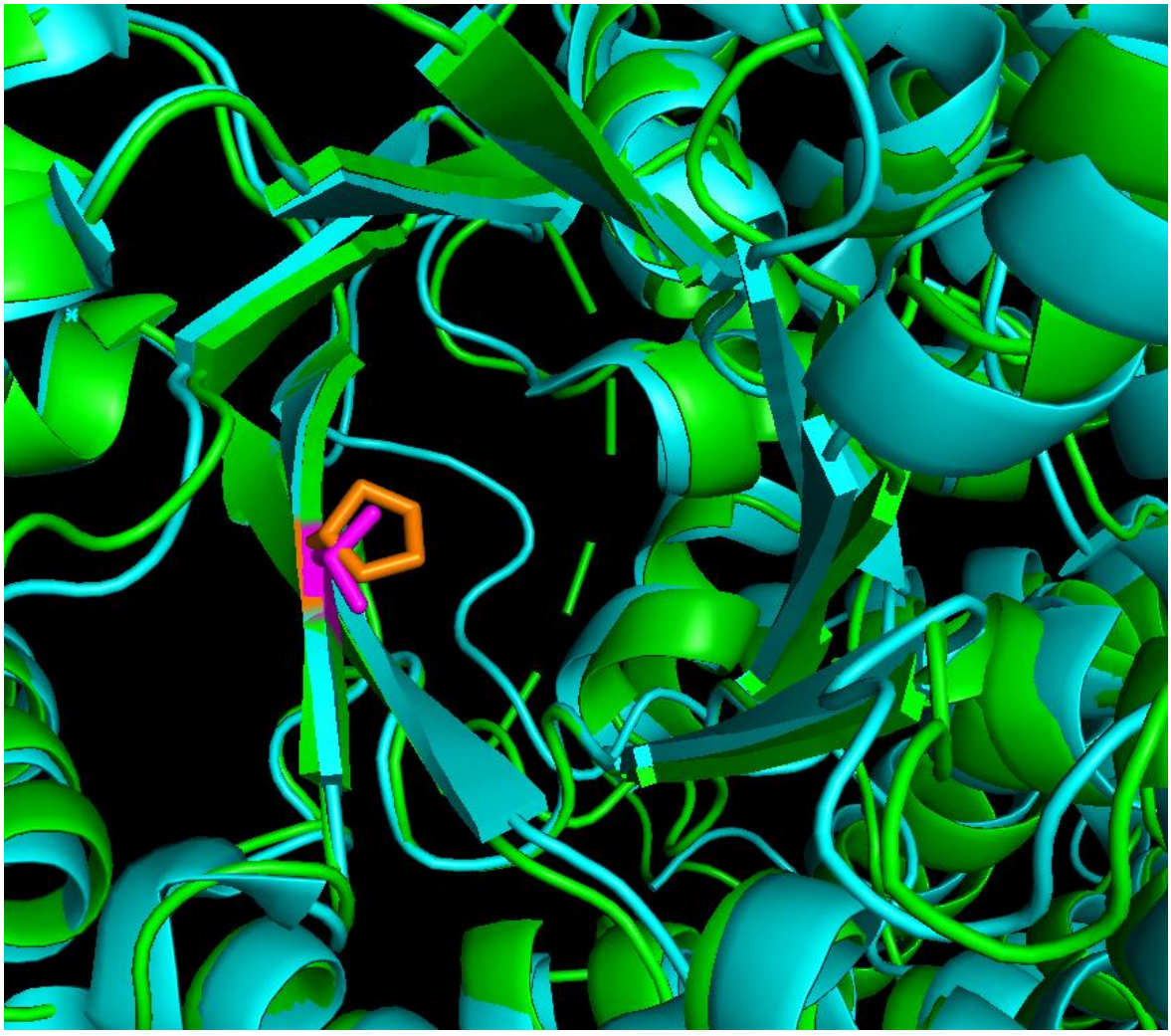


Figure 5.17: The β sheet containing site 531 is structurally similar despite alanine in *E. coli* PEPC (green) and proline in *Z. mays* C_4 PEPC (cyan) at site 531. There is an alanine residue at site 531 in *E. coli* PEPC (pink) and a proline residue at site 531 in *Z. mays* PEPC (orange). Enzymes are viewed from the N-terminal region of the β barrel. *E. coli* PEPC: 1FIY and *Z. mays* C_4 PEPC: 6U2T (PDB).

The P531A 6His::PEPC variant was present at the highest abundance at T_0 (Table 5.2). *E. coli* Δppc were cultured overnight in LB media with antibiotic selection following electroporation with 6His::PEPC variants to increase the relative abundance of transformants compared to untransformed *E. coli* Δppc. This facilitated inoculation of M9 glucose to an OD of 0.1 the following day with bacteria that expressed a 6His::PEPC variant, as opposed to untransformed bacteria which would nevertheless contribute to the OD value. It appeared that 6His::PEPC variants with a high rate of enzymatic activity were selected for during the overnight culture in LB, likely because this led to greater flux through the TCA cycle, and a faster bacterial central metabolism resulted in an increased bacterial growth rate. N863K, which also performed well in the screen, also started at a relatively high abundance (Table 5.2).

Despite the relatively high abundance of P531A and K863K at T_0 in both biological replicates, both T_0 cultures were also comprised of variants at an abundance similar to or higher than variants that were most successful in the screen (Table 5.2). For example, the A625E variant was represented at a similar abundance to the P531A variant in the second replicate, and F794V was not at a high abundance at the start of either biological replicate but became enriched in the population over the course of the screen. Therefore, the competitive advantage of P531A, F794V and N863K in M9 glucose, where bacterial growth required carbon flux through 6His::PEPC, is observed as a large increase in abundance during the screen and was not pre-determined by the initial abundance of the variant.

Table 5.2: The top 10 variants by T_0 abundance in two biological replicates of the screen of 6His::PEPC variants in *E. coli*.

Biological replicate	6His::PEPC variant	Abundance as a proportion of the site at the start of the screen (%)
1	P531A	4.65
1	S780F	2.28
1	Q596W	2.14
1	S577P	1.93
1	N863K	1.61
1	K813E	1.55
1	S577T	1.48
1	S780C	1.24
1	A761P	1.11
1	Q596M	1.03
2	P531A	2.26
2	A625E	2.19
2	P560R	2.17
2	N863K	1.73
2	K813E	1.59
2	P560Q	1.41
2	S577T	1.24
2	S577P	1.08
2	K807N	0.97
2	Q572H	0.78

The P531A, F794V and N863K variants became enriched in every replicate of the screen. Thus, in order to evaluate whether P531A, F794V or N863K mutations confer altered enzyme kinetics, these mutations were made to the SvPEPC sequence, the enzyme variants were purified from *E. coli* and preliminary kinetic characterisation was undertaken. This revealed that enrichment of the P531A variant appears to be due to a significantly higher K_{cat} compared to SvPEPC (Figure 5.18 A). This is despite the P531A mutation conferring a lower affinity of the enzyme for bicarbonate (reflected in a higher $K_mHCO_3^-$) (Figure 5.18 B). The mutation did not significantly change the affinity of the enzyme for PEP (K_mPEP) (Figure 5.18 C) and the specificity of the P531A variant for bicarbonate ($K_{cat}/K_mHCO_3^-$) was not significantly different to SvPEPC because the lower affinity of the P531A variant for bicarbonate occurs concurrent with an increased K_{cat} (Figure 5.18 D). Together, these results indicate that a proline to alanine substitution of the P531A variant conferred a higher maximum catalytic capacity but that this occurred at the expense of a lower affinity of the enzyme for the substrate bicarbonate.

Site 531 does not appear to be close enough (6 Å or less) to be interacting directly with any residues involved in substrate or effector binding; the nearest residue involved in substrate or effector binding is R456 which binds PEP and is 8.1 Å away from the site in the activated enzyme crystal structure (Figure 5.19). However, catalysis does occur at the C-terminal region of the β barrel which is comprised of β-strands including a strand containing residue 531 (Figure 5.19). Therefore, although it is unlikely that the residue interacts with substrate binding residues directly, it is possible that greater flexibility of the alanine residue alters the available conformations of the enzyme near the active site during catalysis in the P531A variant. Alternatively, the ability of alanine to form an additional hydrogen bond

could stabilise conformational changes around the active site during catalysis. Either or both of these features of an alanine residue at site 531 could facilitate more rapid catalytic turnover of the enzyme.

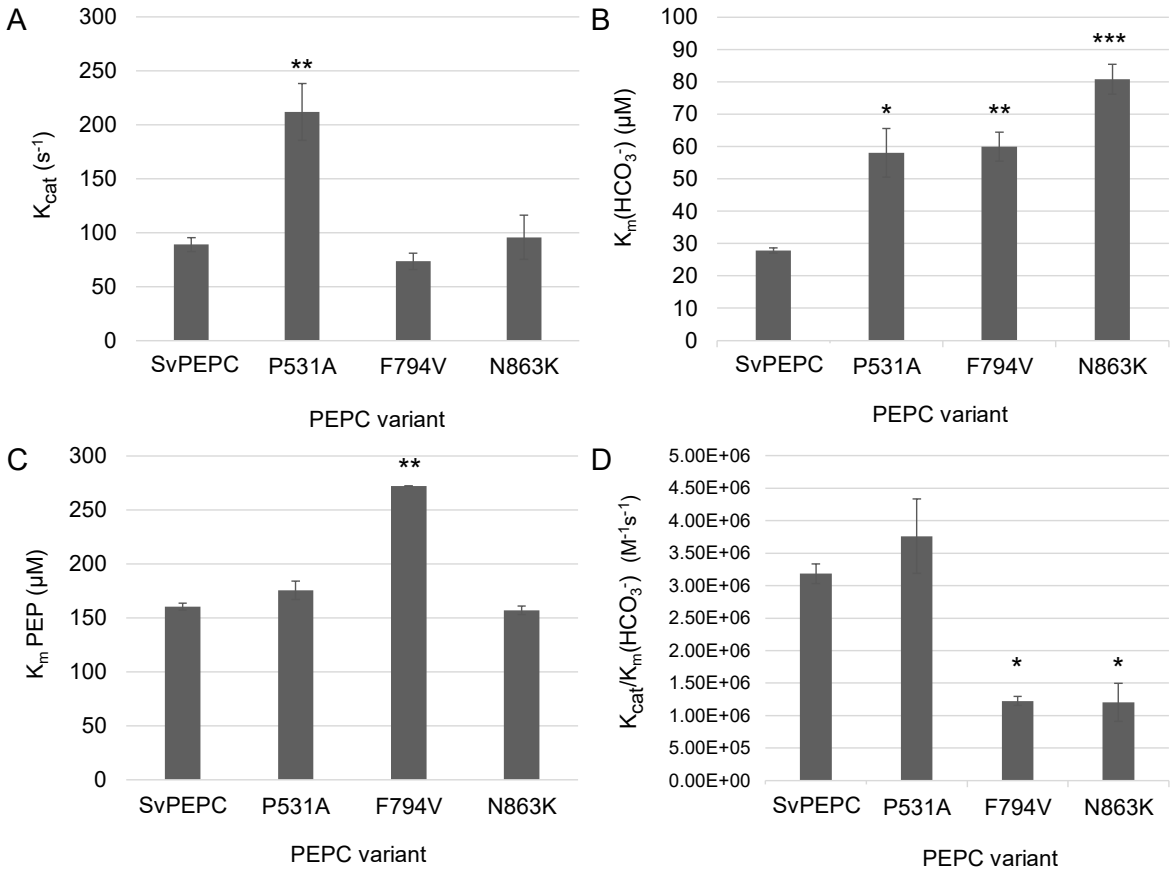


Figure 5.18: K_{cat} and affinity for substrates measured for the three most abundant variants after 72 hours selection in *E. coli* Δppc and compared to equivalent SvPEPC kinetics. Bars represent the mean ± SE. The enzyme assays were undertaken on SvPEPC or PEPC variants that do not have a His-tag. Statistically significant difference between SvPEPC and the three SvPEPC variants is indicated above the plot as determined by a one-way ANOVA and Tukey's HSD Test for multiple comparisons, n = 3 for all kinetics except K_mPEP measured for P531A, F794V and N863K variants where n=2. *** $p<0.001$, ** $p<0.01$, * $p<0.05$. Data was collected by Ryan Wessendorf, Washington State University. Statistical analysis undertaken by Philippa Sinclair.

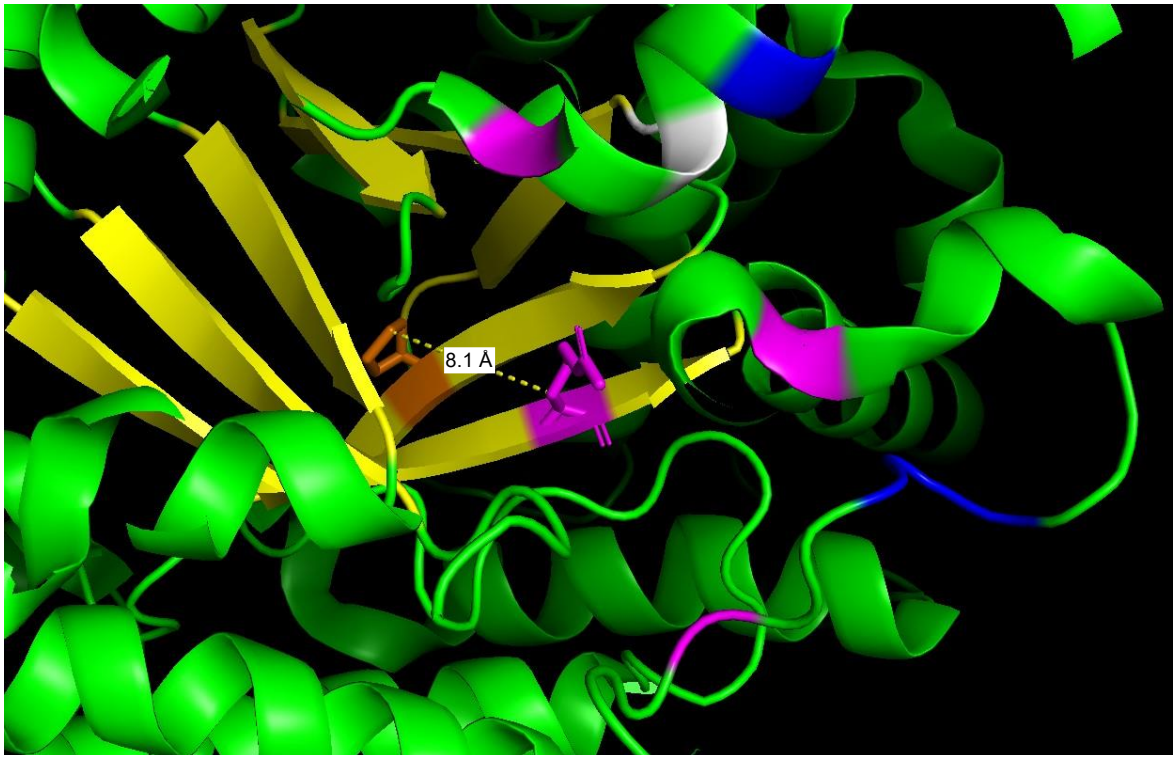


Figure 5.19: The nearest substrate binding site to 531 is PEP binding residue R456 located 8.1 Å away. β barrel represented in yellow, PEP binding residues coloured in pink, magnesium binding residues in grey, bicarbonate binding residues in dark blue and site 531 in orange. PDB: 5VYJ.

The selective pressures that influence the evolution of enzyme kinetics, for example catalytic capacity (K_{cat}) and affinity of an enzyme for its substrates, are complex and often remain unclear for a given enzyme (Bar-Even *et al.*, 2011). Theoretical optima are tailored to intracellular concentrations of substrate and products (Klipp and Heinrich, 1994), correlations observed between kinetic parameters imply possible trade-offs or phylogenetic constraint (Iñiguez *et al.*, 2020; Bouvier *et al.*, 2021), and the optimisation of specific kinetic parameters appears to be linked to the class of metabolism that the enzyme is involved in (Bar-Even *et al.*, 2011). For example, it is expected that an increased K_{cat} is selected for at the expense of K_m for an enzyme under a constant K_{cat}/K_m if an enzyme is involved in central metabolism because these enzymes are under strong selection to support high flux *in vivo* (Bar-Even *et al.*, 2011). Here, a strong artificial selection pressure for enzymatic rate has been imposed by limiting PEPC expression. Accordingly, the P531A variant has an increased K_{cat} at the expense of affinity for bicarbonate under a constant specificity of the enzyme for bicarbonate (Figure 5.18). In addition, the increase in K_{cat} at the expense of a lower affinity for bicarbonate for variant P531A suggests that SvPEPC could be subject to a kinetic trade-off between K_{cat} and affinity for bicarbonate (Figure 5.18).

Increased catalytic capacity is also thought to be selected for when substrates are in excess, without an equivalent selection pressure on the affinity of the enzyme for the substrate (Klipp and Heinrich, 1994; Garcia-Viloca *et al.*, 2004). The affinity of *E. coli* PEPC for PEP is 190 µM, similar to the SvPEPC, P531A and F863V variants (Figure 5.18 C), whereas the affinity of *E. coli* PEPC for bicarbonate (100 µM) is lower than the affinity of SvPEPC for bicarbonate (60.2 µM) (Kai *et al.*, 1999; Boyd, Gandin and Cousins, 2015). The higher affinity of SvPEPC for bicarbonate

compared to *E. coli* PEPC suggests that bicarbonate availability is less of a limitation to enzymatic rate of PEPC in *E. coli* compared to the selection pressure in a photosynthetic mesophyll cell. In addition, imposing reduced expression of PEPC during the screen could result in a build-up of substrates PEP and bicarbonate in the bacterial cytosol. Where bacterial growth is limited by the catalytic capacity of PEPC, and there is a lower selection pressure for a high affinity for bicarbonate compared to in a mesophyll cell, K_{cat} would be selected for regardless of a concomitant reduction in affinity of the enzyme for bicarbonate in *E. coli* Δppc.

In the model of C_4 photosynthesis, at low CO_2 partial pressure (C_m) in the mesophyll when PEPC is rate-limiting for photosynthesis, PEP is assumed to be saturating and together C_m, V_{max} and the affinity of PEPC for bicarbonate determine PEP carboxylation rate (von Caemmerer, 2021). PEPC with a high V_{max} and high affinity for bicarbonate should have the fastest carboxylation rate at a given low C_m. To evaluate the impact of the P531A mutation on carbon assimilation rate in *S. viridis*, the model of C_4 photosynthesis was adapted using the enzyme kinetics values in Figure 5.18. The rate of photosynthetic carbon fixation at low CO_2 concentrations, when PEPC is rate limiting, was predicted to be higher when a higher K_{cat} and lower affinity for bicarbonate characteristic of the P531A variant were incorporated into the model, compared to the model for wild-type *S. viridis*, assuming an equivalent enzyme expression level (Figure 5.20 A, B).

The predicted increase in carbon assimilation conferred by the P531A mutation in the model of C_4 photosynthesis is interesting given that proline has been selected for repeatedly at site 531 across independent origins of C_4 photosynthesis and is encoded in SvPEPC (Christin *et al.*, 2007). The repeated selection for proline from

alanine at site 531 in C_4 PEPC could suggest that the P531A mutation may not be beneficial *in planta* despite the associated increase in K_{cat} and higher modelled initial carbon assimilation rate in *S. viridis*. Despite being associated with lower K_{cat}, the proline residue at site 531 could be important in the regulation of enzyme expression or it could influence regulation of the enzyme by allosteric effectors, which was not measured in the enzyme kinetics characterisation in Figure 5.18 and was not accounted for in the model.

An alternative possibility is that the A531P mutation has occurred multiple times independently in lineages that have evolved C_4 photosynthesis in order to facilitate a change in enzyme activity that was advantageous during the initial stages of C_4 photosynthesis evolution but is no longer beneficial. Indeed, given that *E. coli* Δppc containing the P531A variant were consistently highly abundant at the end of each replicate of the bacterial screen, the P531A SvPEPC variant has a higher K_{cat} than the wild-type enzyme (Figure 5.18 A) and the mutation confers an increased modelled rate of photosynthetic carbon assimilation under low CO_2 conditions in *S. viridis* (Figure 5.20 B), the P531A mutation is an exciting target for plant mutagenesis. The *S. viridis* genome could be edited to introduce the P531A mutation to SvPEPC, then carbon assimilation rate (A_{net}) under different CO_2 concentrations in leaf mesophyll cells (C_m) could be measured to determine whether there are any changes to photosynthetic carbon assimilation rate at low CO_2 conditions. In addition, the plant phenotype could be characterised to distinguish any effect (or not) of the P531A variant on plant growth and physiology. Thus, screening 6His::PEPC variants in high-throughput in *E. coli* Δppc has led to the identification of a novel target that would be interesting to explore in plants.

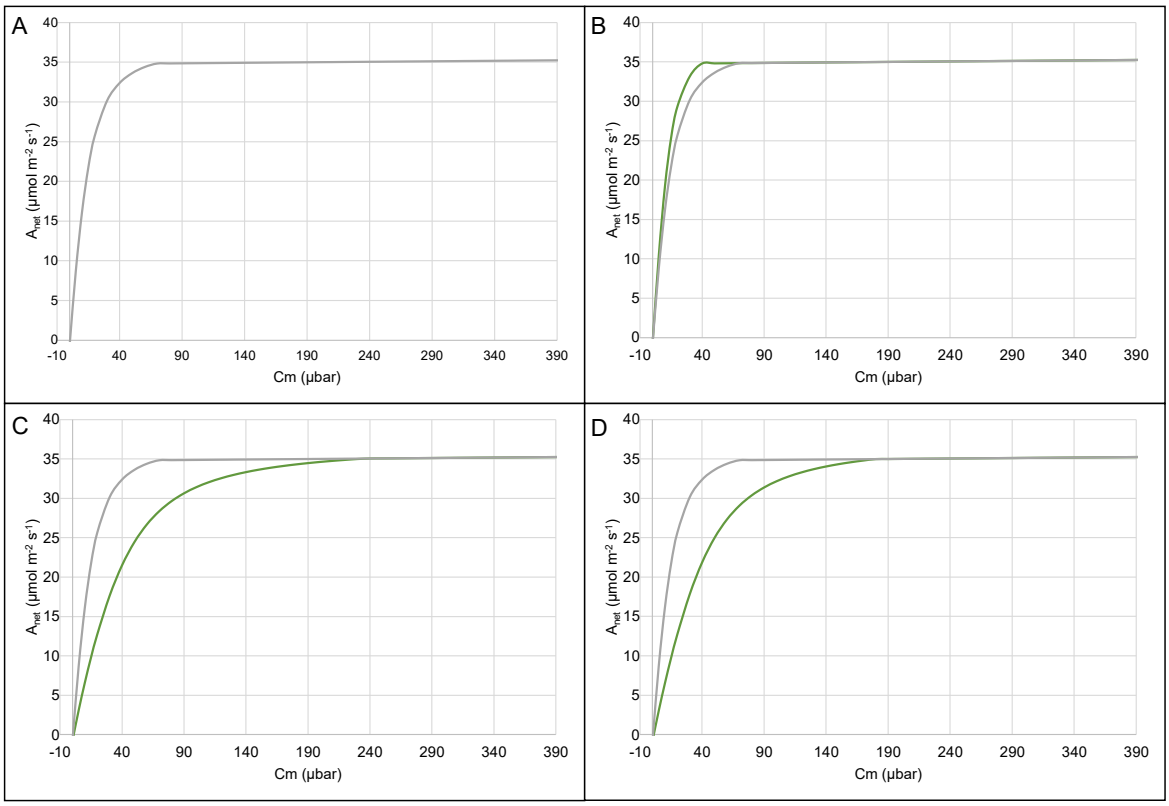


Figure 5.20: Predicted carbon assimilation rates (A_{net}) at different concentrations of intracellular CO_2 (C_m) in the mesophyll of *Setaria viridis*. Wild-type C_4 PEPC (A) is compared to the PEPC variants P531A (B), F794V (C) and N863K (D). The predicted assimilation rate graphs were calculated from the kinetics measured for the PEPC variants, whereby the average kinetics value for K_{cat} or $K_mHCO_3^-$ from all three technical replicates was used. The assimilation rate controlled by PEPC is represented in green and the wild-type *S. viridis* graph has been superimposed on the variant graphs in grey to better illustrate changes in rate. Model adapted from von Caemmerer, 2021, method detailed in Chapter 2.

C_4 PEPC has repeatedly evolved a higher affinity and specificity for bicarbonate compared to C_3 PEPC across independent origins of C_4 photosynthesis (Bauwe, 1986; DiMario and Cousins, 2019; Moody, Christin and Reid, 2020), apart from in the genus *Alloteropsis* (Phansopa *et al.*, 2020). Therefore, it is most likely that variants that have both a higher affinity and specificity for bicarbonate would be beneficial in a C_4 context. To select for a high affinity for bicarbonate while maintaining a high K_{cat} (and thereby selecting for a high specificity for bicarbonate), the screen in *E. coli* Δppc could be altered such that bicarbonate availability is reduced. This could be achieved by reducing the flow rate of CO_2 delivery to a bioreactor that has gas control capability or by knockdown of carbonic anhydrase in *E. coli* Δppc. *E. coli* requires carbonic anhydrase to grow on glucose media in air (Merlin *et al.*, 2003), so carbonic anhydrase knockout would be lethal but tightly controlled inducible expression should limit intracellular bicarbonate concentration. A screen using *E. coli* has selected for a P531A PEPC variant that has a higher K_{cat} than SvPEPC, at the expense of affinity for bicarbonate (Table 5.1, Figure 5.18). Therefore, the conditions of the screen could be altered to create a low intracellular bicarbonate concentration to select for 6His::PEPC variants that have both a high catalytic turnover (high K_{cat}) and high affinity for bicarbonate.

Site 531 exhibits a very different amino acid distribution across plant PEPCs (encompassing both C_3 and C_4 plant PEPCs) and PEPCs across the tree of life (Figure 5.21). Alanine is the most abundant amino acid (95.9%) at position 531 in plant PEPCs whereas proline is represented in only 2.6% plant PEPC sequences. Using the number of plant PEPC sequences with a serine residue at site 780 as a proxy for the number of C_4 species in the plant PEPC data set because serine is thought to be required for C_4 PEPC kinetics (Bläsing, Westhoff and Svensson,

2000), approximately 4.1% of the sequences encode C_4 PEPC. Thus, approximately two thirds of putative C_4 PEPC sequences have evolved proline at site 531. This provides further support to the hypothesis that adaptive mutations to C_4 cycle enzymes have occurred in some lineages but not others (Niklaus and Kelly, 2018).

In a broader sampling of taxa (which are mostly bacteria) there is a more diverse mix of amino acids at site 531, including amino acids with different properties. Alanine, serine, glutamine and asparagine are found at this site in over 10% of sequences in 39,581 unique PEPC sequences obtained from the NCBI dataset (Figure 5.21). These amino acids differ by size, polarity and charge. The relationship between the amino acid present at site 531 with catalytic capacity and affinity of C_4 *S. viridis* PEPC for bicarbonate, concurrent with the large variety of amino acids at this site across bacterial sequences, suggests that the residue is influential in enzyme kinetics and that different amino acids have been selected for in different environments. Further mutagenesis studies are required to identify the effect of the different amino acids at site 531 on PEPC kinetics.



Figure 5.21: Frequency of amino acids at sites 531, 794 and 863 (*Z. mays* numbering) from PEPC sequences across the tree of life and from plant PEPC sequences. Plant PEPC sequences are a subset of all the sequences. Polar amino acids and glycine are represented in green, basic amino acids in blue, neutral amino acids in purple, acidic amino acids in red and hydrophobic amino acids in black.

5.2.3.2.2 F794V and N863K 6His::PEPC variants have substitutions that have occurred repeatedly in C_4 PEPC

Two of the three top performing variants have amino acid changes that commonly occur when species evolve C_4 photosynthesis but have not occurred yet in *S. viridis*. Specifically, the phenylalanine found at position 794 in C_3 species is commonly mutated to valine in C_4 species, and the asparagine found at position 863 in C_3 species is commonly mutated to lysine in C_4 species (Christin *et al.*, 2007). However, in SvPEPC these positions are still phenylalanine and asparagine, respectively. In the screen, valine out-competed phenylalanine at residue 794 and lysine out-competed asparagine at residue 863 and thus it is likely that introduction of these substitutions would confer an advantage to SvPEPC function. Further, the repeated evolution of both these mutations independently across C_4 species suggests that it could be beneficial to introduce both together in the plant. The functionality of an enzyme in which both sites are simultaneously mutated could be established before plant engineering is undertaken by expressing the enzyme in *E. coli* Δppc to discover if growth of the bacteria in M9 glucose is rescued prior to enzyme kinetics analysis.

F794V and N863K variants do not have a significantly different K_{cat} compared to SvPEPC, unlike the P531A variant (Figure 5.18 A). Both variants also display a lower affinity and specificity for the substrate bicarbonate and either a similar or lower affinity for PEP compared to SvPEPC and were predicted to confer a lower initial photosynthetic assimilation rate *in planta* (Figure 5.18, Figure 5.20 C, D). The success of F794V and N863K variants, despite both variants having an equivalent rate of catalytic turnover and a lower affinity for bicarbonate compared to SvPEPC, supports the hypothesis that the selection pressure on the affinity and specificity of

6His::PEPC for bicarbonate in *E. coli* Δppc is lower than photosynthetic mesophyll cell (discussed above). In addition to changes to affinity for substrates bicarbonate and PEP, the co-option of PEPC into C_4 photosynthesis has also resulted in a higher tolerance of the enzyme to inhibitors malate and aspartate and increased activation by G6P (Bläsing, Westhoff and Svensson, 2000; Bläsing *et al.*, 2002; Jacobs *et al.*, 2008). These changes to enzyme regulation have occurred concomitant with amino acid substitutions at the sites under positive selection (Christin *et al.*, 2007). Moreover, enzyme kinetics characterisation was undertaken on C_3 and C_4 *Flaveria* PEPCs purified from *E. coli* which were not phosphorylated, so the different enzyme kinetics measured for the C_3 and C_4 PEPC orthologues must be underpinned by differences between the amino acid sequences, independent of other regulatory mechanisms such as phosphorylation (Bläsing, Westhoff and Svensson, 2000; Bläsing *et al.*, 2002; Jacobs *et al.*, 2008). Thus, given that mutations at sites 794 and 863 are conferring a competitive advantage *in vivo* during the screen but do not confer an increased K_{cat} or increased affinity for substrates, it is possible that these mutations influence allosteric regulation of PEPC.

Sites 794 and 863 do not appear to be close enough (6 Å or less) to be interacting directly with any known substrate or regulator binding residues (Figure 5.22). However, a mutation at site 794 or 863 could induce structural changes that alter effector regulation even if the interaction is not direct. Indeed, at site 863 a positively charged lysine residue in place of an uncharged asparagine residue appears to induce a salt bridge with glutamic acid at position 727 in *Z. mays* C_4 PEPC, which is not observed in the *E. coli* PEPC (Figure 5.23). The N863K mutation to SvPEPC could facilitate salt-bridge formation with a glutamic acid residue at the equivalent position in SvPEPC to the aforementioned residue in *Z. mays* C_4 PEPC (721 in

SvPEPC). Point mutations that are not local to the active site can also affect substrate specificity when a change to one residue causes sequential interaction changes that lead to and include the active site (Mendonça and Marana, 2011). A similar phenomenon could be occurring here where F794V and N863K mutations result in structural changes that have an allosteric effect on effector binding elsewhere in the enzyme.

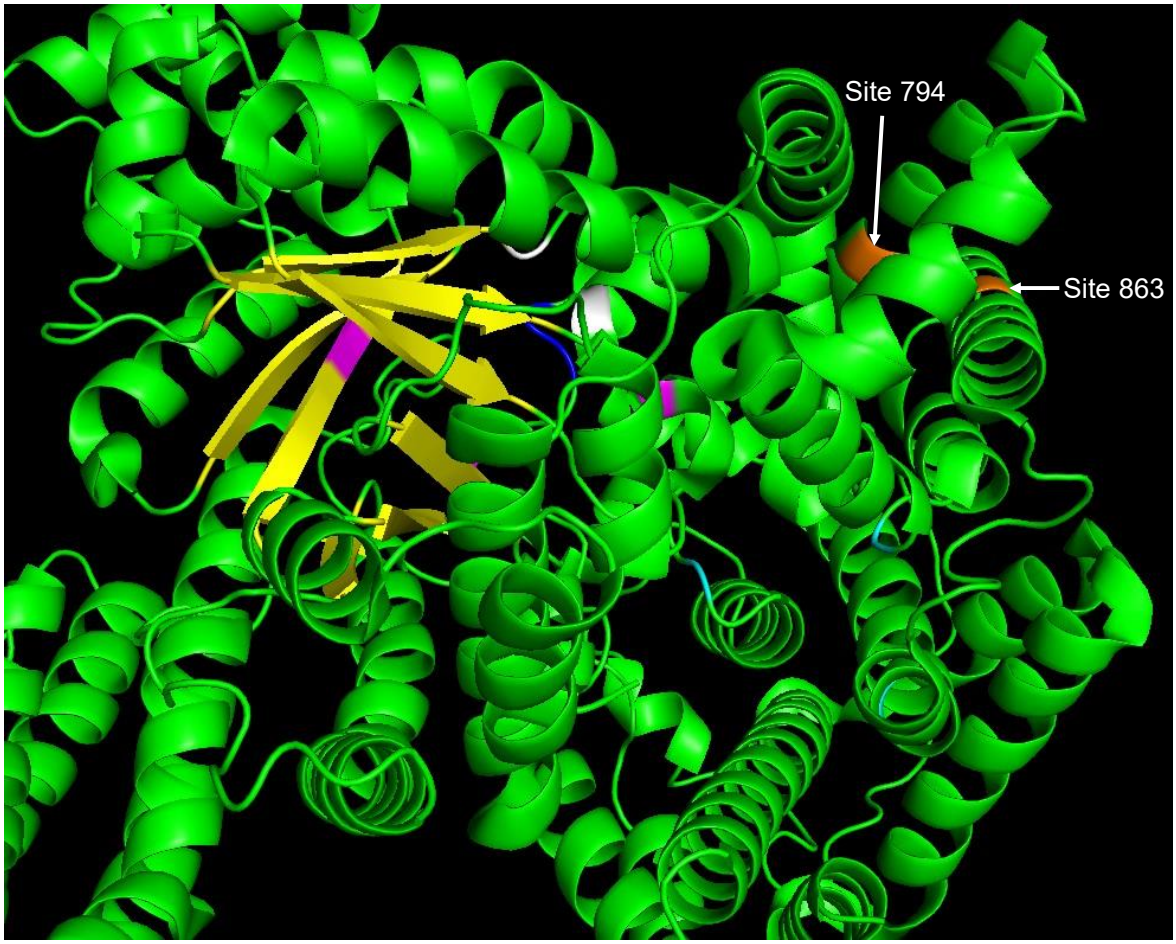


Figure 5.22: Sites 794 and 863 are not near to known substrate or effector binding sites. β barrel represented in yellow, PEP binding residues coloured in pink, magnesium binding residues in grey, bicarbonate binding residues in dark blue, aspartate/malate binding residues in light blue and sites 794 and 863 in orange. PDB: 5VYJ.

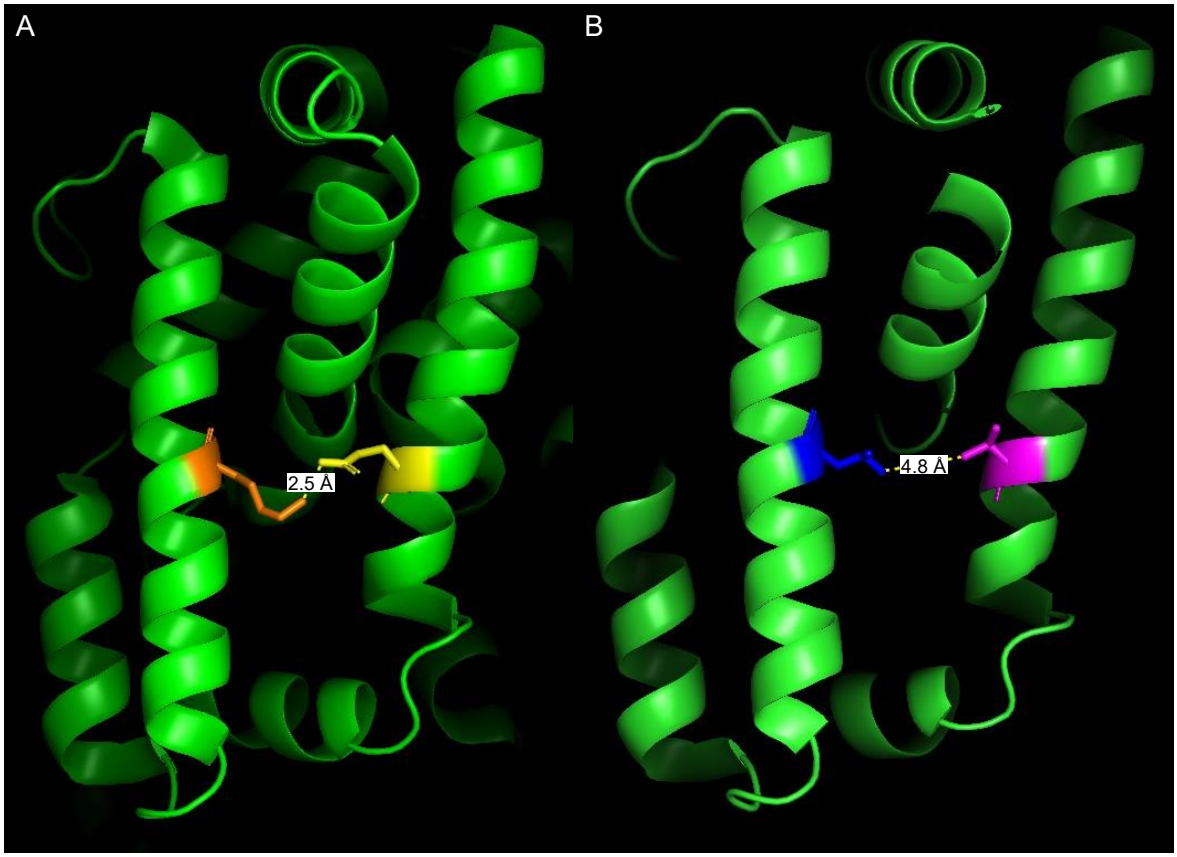


Figure 5.23: Salt bridge formation can occur in *Z. mays* C_4 PEPC between E727 (yellow) and K863 (orange) (A) but not in *E. coli* C_3 PEPC (B). *E. coli* residues are leucine (blue) and valine (pink). *E. coli* PEPC: 1FIY and *Z. mays* C_4 PEPC: 5VYJ (PDB).

C_4 PEPC characteristically has a lower affinity for inhibitors malate and aspartate and a higher level of activation by G6P compared to C_3 PEPC (Dong *et al.*, 1998; Bläsing *et al.*, 2002; Svensson, Bläsing and Westhoff, 2003). When *E. coli* growth rate is limited by the enzymatic rate of 6His::PEPC, demand for TCA cycle intermediates is likely to increase, so inhibitors including malate and aspartate could be present at a lower intracellular concentration during the screen in *E. coli* compared to under steady-state conditions in a photosynthetic mesophyll cell. Conversely, given that glucose is continually available to *E. coli* in the growth medium, and glucose is phosphorylated to form G6P on import into the bacterial cytoplasm (Fuentes *et al.*, 2013), if the competitive advantage of F794V and N863K variants is due to altered affinity for effectors then increased activation by G6P could be selected for during the screen as opposed to decreased inhibition by aspartate or malate.

The enzyme kinetics in Figure 5.18 were measured with 5 mM G6P in the enzyme assay mixture in keeping with other measurements in the literature (Boyd, Gandin and Cousins, 2015; DiMario and Cousins, 2019), but hexose 6-phosphates are present at a marginally higher concentration of 8.75 mM in *E. coli* during aerobic growth in M9 glucose (Bennett *et al.*, 2009). In *Z. mays* C_4 PEPC, increased enzyme activity conferred through G6P activation is almost fully realised by 5 mM and additional G6P does not confer further increases (Takahashi-Terada *et al.*, 2005; Yuan *et al.*, 2006), so G6P activation may not be contributing to increased enzymatic rate during the screen. Nonetheless, it would be interesting to characterise and compare the activator and inhibitor kinetics for the F794V and N863K variants to SvPEPC to discover whether or not these mutations alter

regulation by effectors and consequently if variants are likely to be beneficial to engineer *in planta*.

It is also possible that F794V and N863K variants are selected for during the screen because these mutations confer greater solubility, higher stability, improved protein folding or a longer lifespan compared to 6His::PEPC. The faster speed of translation in *E. coli* compared to eukaryotic cells can lead to misfolded or insoluble proteins during heterologous gene expression (Francis and Page, 2010). Further, localised misfolding induced by different speeds of translation can impact *E. coli* enzyme kinetics, even between identical enzymes (Jiang *et al.*, 2023). Sufficient 6His::PEPC must be soluble and functional to rescue the growth phenotype of *E. coli* Δppc in M9 glucose, however, if either F794V or N863K mutations improved protein solubility, folding or stability at the 37 °C culture temperature then this could give *E. coli* Δppc a competitive advantage in the screen. Alternatively, enzymes that undergo fewer catalytic cycles until replacement (CCR) are associated with higher metabolic burden and hence enzyme CCRs have been suggested as a target parameter to engineer higher crop yields (Tivendale *et al.*, 2020). If either a F794V or N863K mutation conferred higher CCR than 6His::PEPC, then these variants could be selected for because the bacterial cell can redirect more energy into other cellular processes. These hypotheses need to be tested, but screening plant enzymes in *E. coli* could accelerate the discovery of mutations that improve enzyme CCRs, stability or folding that can be tested *in planta*.

It is important to note that F794V and N863K variants may not have become enriched in the *E. coli* screen for the same reason. For example, the N863K mutation may confer changes to effector binding, increase enzyme stability or improve protein folding given the ability of lysine to form a salt bridge with glutamic

acid where asparagine cannot (Figure 5.23). However, the F794V mutation is on the inside of the protein and does not appear to be conferring any obvious structural changes, so the F794 variant could have been selected for because it maintains a functional enzyme but is less energy intensive for the bacteria to express; valine requires only 29 ATP to produce compared to 61 ATP for phenylalanine (Wagner, 2005). If these mutations do confer equivalent advantages to enzyme fitness in plants as they do in *E. coli*, it is surprising that neither have been selected for in *S. viridis*, as both could have evolved from the SvPEPC sequence through a single nucleotide polymorphism (SNP). Nevertheless, the fact that 794V and 863K have been repeatedly selected for across C_4 species and were abundant at the end of the *E. coli* screen makes them interesting candidates for engineering into *S. viridis*. Moreover, understanding why the F794V and N863K mutations give 6His::PEPC a competitive advantage in *E. coli* could help to elucidate why these mutations have been repeatedly selected for across C_4 PEPCs.

The substitution of phenylalanine for valine or isoleucine at site 794 and of asparagine for lysine at site 863 has occurred across C_4 PEPCs (Christin *et al.*, 2007). However, analysis of a larger data set of plant PEPC sequences (C_3 and C_4 together) shows that at site 794, 65.6% sequences have phenylalanine and 29.1% have valine (Figure 5.21). At site 863, 50.3% sequences have asparagine, followed by glutamic acid (26.0%) and lysine (13.2%) (Figure 5.21). If, as hypothesised above, a maximum 4.1% sequences are C_4 PEPC, it appears that some C_3 plant PEPC sequences have valine at site 794 and lysine at site 863. Thus, perhaps these mutations are not required for C_4 PEPC-specific kinetics.

Across all the sequences deposited in NCBI for PEPC, comprised mostly of bacterial PEPC sequences, valine is most highly represented (31.1%) at site 794,

and alanine, phenylalanine and leucine are all represented at over 10% frequency, so it appears that there is more flexibility at the amino acid occurring at this site in bacterial PEPC compared to plant PEPCs (Figure 5.21). At site 863, glutamic acid is most common (65.3%), followed by leucine (11.2%), glutamine (7.4%), arginine (7.2%) and asparagine (2.5%), all amino acids with very different biochemical properties (Figure 5.21). Lysine only comprises 2.1% of total variation at site 863, so it is more highly abundant amongst plant PEPCs than PEPCs across the tree of life. The biological significance of the diverse range of amino acid biochemistries at site 863, different abundances of amino acids between plant and bacterial PEPC sequences at sites 794 and 863, and of selection for valine and lysine at sites 794 and 863 in C_4 PEPC, respectively, are yet to be fully understood.

5.3 Chapter summary

Single amino acid substitutions can greatly influence the affinity of C_4 PEPC for substrates, inhibitors and effectors (Bläsing, Westhoff and Svensson, 2000; Paulus, Schlieper and Groth, 2013; DiMario and Cousins, 2019). However, it is hard to predict which residues, out of a possible ~1000 that comprise the PEPC peptide sequence, will have the greatest impact on enzyme kinetics. It is even more challenging to predict which mutations might confer changes that would be beneficial in the context of C_4 photosynthesis. Concentrating on mutagenesis of sites under selection in PEPC and developing a high-throughput approach in *E. coli* has identified three variants – P531A, F794V and N863K – that have become enriched in the *E. coli* population. None of the three residues identified as conferring a competitive advantage in an *E. coli* Δppc screen are close to substrate or inhibitor binding residues, and as such are unlikely candidates for mutagenesis based on observation of the crystal structure. Preliminary enzyme kinetics characterisation indicated that these variants have altered enzyme kinetics compared to SvPEPC. Thus, focussing on mutating 6His::PEPC at sites under selection has simultaneously reduced the possible sequence space to be explored and led to the identification of the P531A variant, which has a higher catalytic capacity (K_{cat}) than SvPEPC.

The higher K_{cat} of the P531A variant was predicted to confer an increased rate of carbon fixation in *S. viridis* at low CO_2 concentrations. However, in C_4 PEPC, the reverse mutation appears to have been selected for, perhaps because of the way C_4 PEPC is regulated *in planta* or because proline at site 531 increases the affinity of the enzyme for bicarbonate despite a concomitant decrease in K_{cat}. The two second most abundant variants that were enriched in the screen have an F794V or

N863K mutation that introduce an amino acid that is typical of C_4 PEPC at these sites, including in maize, sugarcane and sorghum C_4 PEPCs. Therefore, P531A, F794V and N863K variants could all be taken forward for characterisation in *S. viridis*.

5.4 Authorship statement

The enzyme kinetics measurements that comprise Figure 5.18 were completed in collaboration with Ryan Wessendorf and Asaph Cousins at Washington State University. The kinetics measurements were taken by R. Wessendorf under the supervision of A. Cousins. All other material in this chapter is my own.

Chapter 6: General discussion

6.1 Overview

Crops that undertake C_4 photosynthesis, including maize, sorghum and sugarcane, are amongst the most highly cultivated in the world for food, feed and biofuel (Byrt, Grof and Furbank, 2011; van der Weijde *et al.*, 2013; Wang, Chan and Long, 2021). Therefore, strategies to enhance C_4 photosynthesis have the potential to deliver a large impact through increased crop yields (von Caemmerer, 2021; Wang, Chan and Long, 2021). Moreover, technologies that can be delivered *via* genome editing approaches have the potential to be readily implemented around the world (Metje-Sprink *et al.*, 2019; Sprink, Wilhelm and Hartung, 2022). Thus, there is a need to be able to identify changes that can be made to C_4 crops by genome editing that could increase their yield.

The enzymatic rate of PEPC, which catalyses the first irreversible carbon fixation step in C_4 photosynthesis, is rate limiting for C_4 photosynthesis under low CO_2 conditions in the mesophyll cell, such as under fluctuating light conditions or under drought stress (Pignon and Long, 2020; von Caemmerer, 2021; Wang, Chan and Long, 2021). PEPC variants that have an increased enzymatic rate in the mesophyll cell and can be delivered by genome editing are promising targets for improving yield in C_4 plants. Therefore, the work in this thesis has encompassed the development of a high-throughput screen of 6His::PEPC variants using *E. coli* as a host organism, focussing on 6His::PEPC variants that can be engineered into C_4 plants through genome editing. In this thesis, I demonstrated that growth of *E. coli* Δppc in M9 glucose can be made to be dependent on the expression level of 6His::PEPC (Chapter 3). I then developed and optimised a site saturating

mutagenesis technique to generate libraries of *6His::PEPC* variants (Chapter 4). Finally, I screened 347 6His::PEPC variants in *E. coli* Δppc and identified the three best-performing variants which were then subject to kinetic characterisation (Chapter 5). From this work, putative improved PEPC variants that are deliverable by genome editing have been identified that can now be tested *in planta*.

6.2 Exploiting natural variation between C_4 plants to identify PEPC variants with increased catalytic capacity

Of the 347 6His::PEPC variants mutated at sites under selection in C_4 PEPC, variants with P531A, F794V and N863K mutations repeatedly became most highly abundant during a screen for catalytic capacity in *E. coli* Δppc. Strikingly, the P531A variant that has a K_{cat} twice as high as wild-type SvPEPC became highly abundant across all replicates of the screen. When the kinetics were assayed for the P531A variant, a higher K_{cat} was coincident with a lower affinity for bicarbonate. This is notable because the opposite mutation has occurred repeatedly in nature; an alanine residue has been replaced repeatedly by a proline residue at site 531 across C_4 PEPCs (Christin *et al.*, 2007). Thus, these findings identify that a previously uncharacterised site in C_4 PEPC - site 531 - influences both K_{cat} and bicarbonate affinity in C_4 PEPC.

The fact that P531A is associated with an increase in K_{cat} and a decrease in bicarbonate affinity suggests that there could be a trade-off between these two kinetic parameters in PEPC. Fittingly, a correlation between K_{cat} and the affinity of PEPC for substrates PEP and bicarbonate is evident in the monocot genus *Alloteropsis*, which contains both C_3 and C_4 species (Phansopa *et al.*, 2020). In *Alloteropsis*, PEPCs in C_3 species and recently evolved C_4 species have a lower

K_{cat} and higher affinity for bicarbonate (Phansopa *et al.*, 2020). In contrast, PEPCs in older C_4 species have a higher K_{cat} and a lower affinity for bicarbonate (Phansopa *et al.*, 2020). The lower affinity for bicarbonate of *Alloteropsis* C_4 PEPCs that evolved the photosynthetic function earlier is in contrast with previous studies which have found that C_4 PEPCs evolve a higher affinity for bicarbonate compared to C_3 PEPCs (Bauwe, 1986; DiMario and Cousins, 2019; Moody, Christin and Reid, 2020). It is therefore suggested that the selection pressure on the affinity of PEPC for bicarbonate is highly context dependent (Phansopa *et al.*, 2020). However, C_4 PEPCs from *Panicum queenslandicum* and *Flaveria pringlei* have a higher rather than lower affinity for bicarbonate coincident with a relatively lower K_{cat} compared to C_3 PEPCs from closely related species (Moody, Christin and Reid, 2020). Further, *E. coli* PEPC has the highest K_{cat} of all PEPCs on BRENDA and a lower affinity for bicarbonate compared to most C_4 PEPCs (Bauwe, 1986; Kai *et al.*, 1999; DiMario and Cousins, 2019; Moody, Christin and Reid, 2020; DiMario *et al.*, 2021). The repeated evolution of proline at site 531 suggests that an increased affinity of PEPC for bicarbonate has been selected for at the expense of K_{cat} in C_4 PEPCs in species in the Poaceae family, including *S. viridis*, and that species within the genus *Alloteropsis* are an exception to this. In combination, the altered kinetics of the P531A mutation presented in this thesis and the PEPC kinetics previously characterised in both plant and bacterial PEPCs support the idea of a trade-off between K_{cat} and the affinity of the enzyme for bicarbonate.

6.3 F794V and N863K variants are candidates for characterisation *in planta*

In contrast to the P531A mutation, the F794V and N863K mutations identified in this screen were not associated with a significant increase in K_{cat} compared to SvPEPC.

The similar K_{cat} but lower affinity of the F794V and N863K variants for bicarbonate, resulting in a lower specificity of the two variants for bicarbonate, suggest that these variants are becoming enriched during the screen for a reason that has not been measured in the kinetics analysis. For example, it could be that the 6His::PEPC variants have altered regulation by effectors, for example greater activation by G6P or reduced inhibition by malate. Alternatively, it could be that these variants have similar kinetics to 6His::PEPC but are more effectively expressed in *E. coli* Δppc, for example if the variants are more soluble, less energetically expensive to produce or have more catalytic cycles until failure compared to 6His::PEPC. Therefore, it would be interesting to characterise the activator and inhibitor kinetics of variants F794V and N863K because these substitutions are repeatedly selected for as PEPC is co-opted into the C_4 photosynthesis pathway (Christin *et al.*, 2007). An understanding of why these mutations confer an advantage in *E. coli* Δppc might elucidate why these mutations are also selected for in C_4 PEPC.

The success of F794V and N863K variants within the 6His::PEPC background are particularly interesting because they have both arisen in key crop plants maize, sugarcane and sorghum C_4 PEPC and have evolved repeatedly in other C_4 species (Christin *et al.*, 2007). The F794V and N863K mutations were selected for in this screen and thus are assumed to increase flux through PEPC, however, the measured enzyme kinetics did not reveal an obvious mechanism underpinning this. Interestingly, both F794V and N863K mutations could have evolved in *S. viridis* by a single nucleotide change and confer a lower affinity for bicarbonate which results in a lower rate of initial carbon fixation using the model of C_4 photosynthesis. Thus, it is possible that these mutations have not evolved yet in *S. viridis* because they do not confer an advantage in the context of the SvPEPC. Interestingly, the

Andropogoneae origin of C_4 photosynthesis from which maize, sorghum and sugarcane have all evolved and the MPC grasses origin of C_4 photosynthesis from which *S. viridis* has evolved are closely related, occurring from plants within the same family (Poaceae), and both occurred around 20 million years ago (Sage, 2016). Therefore, despite the phylogenetic proximity between the two separate origins of C_4 in the Andropogoneae and MPC grasses suggesting similar standing variation and a similar time frame for the mutations F794V and N863K to have evolved, these mutations have not been selected for in SvPEPC. Diversity between C_4 photosynthesis cycles is well-documented (Rao and Dixon, 2016), and could lead to mutations being selected for in some lineages and not others. Despite these reservations, given that both F794V and N863K variants were selected for in the screen and the repeated selection for these mutations in C_4 species, it would be interesting to characterise the effect of both mutations in *S. viridis*.

6.4 Sites under selection as targets for 6His::PEPC mutagenesis

Variation in enzyme kinetics between C_4 PEPCs from lineages that have evolved C_4 photosynthesis at different times and between isoforms from the same genus or family suggest that C_4 PEPC is not equally well-adapted for its role in C_4 photosynthesis across different species (Niklaus and Kelly, 2019; Moody, Christin and Reid, 2020; Phansopa *et al.*, 2020; DiMario *et al.*, 2021). Sites under selection in C_4 PEPC were mutated because amino acid substitutions at these sites are likely to be conferring changes to enzyme kinetics as the enzyme is co-opted into the photosynthetic pathway given that these residues are evolving concurrent with kinetic changes (Christin *et al.*, 2007). A site saturating mutagenesis approach was undertaken because C_4 photosynthesis evolved relatively recently, so the optimal amino acid may not yet be mutated at a given site (Niklaus and Kelly, 2019). Single

sites have been shown to confer large differences in enzyme kinetics, including towards substrates PEP and bicarbonate (Bläsing, Westhoff and Svensson, 2000; Paulus, Schlieper and Groth, 2013; González-Segura *et al.*, 2018; DiMario and Cousins, 2019). However, of the sites under selection in C_4 PEPC only site 780, which is known to have a large influence over C_4 PEPC kinetics, has been interrogated experimentally (Bläsing, Westhoff and Svensson, 2000; DiMario and Cousins, 2019). In addition, apart from site 761 which is on a catalytic loop near the active site of the enzyme, and site 733 that has been suggested to be involved in enzyme kinetics but not experimentally characterised, none of the other 18 sites under selection in C_4 PEPC are near to any known effector binding pockets or the active site of the enzyme and are not obvious targets for mutagenesis. The substantial enrichment of the P531A variant during the screen and changes to SvPEPC kinetics conferred by the P531A mutation demonstrate that targeting sites under selection is an effective way of identifying residues that are influential in enzyme kinetics, namely that site 531 influences K_{cat} and bicarbonate affinity in C_4 PEPC.

Identification of sites under selection undertaken by Christin et al. (2007) was limited to residues 455 to 907 (*Z. mays* PEPC numbering) because this region of the enzyme is known to control enzyme kinetics (Bläsing, Westhoff and Svensson, 2000; Christin *et al.*, 2007). However, this domain was identified because it altered the affinity of C_4 PEPC for PEP rather than bicarbonate (Bläsing, Westhoff and Svensson, 2000). Moreover, domain swapping experiments show that residues 528 – 880, which have been shown to influence the affinity of PEPC for bicarbonate, do not fully explain all the variation between C_4 PEPCs with regards to bicarbonate kinetics, and residues 455 – 907 do not explain all the variation with regards to PEP

(Bläsing, Westhoff and Svensson, 2000; DiMario *et al.*, 2021). Thus, given that mutagenesis of sites under selection identified in C_4 PEPC between residues 455 and 907 has resulted in 6His::PEPC variants with altered kinetics, that sites 531, 794 and 863 are far from the active site but appear to be influential in PEPC activity, and residues outside of 455 – 907 must influence the affinity of C_4 PEPC for substrates, an equivalent bioinformatic approach for sites under selection could be extended to the entire C_4 PEPC gene. A large number of sites could be identified in a whole-gene analysis, so sites with differences in biochemistry which are likely to be having the largest effect on enzyme kinetics, for example N863K, could be prioritised for mutagenesis.

6.5 The application of directed evolution in bacteria to plant PEPC

"You get what you screen for" is the golden rule of directed evolution studies (You and Arnold, 1996). The screen of 6His::PEPC in *E. coli* Δppc was designed such that *6His::PEPC* expression level was limited and bacterial growth was solely dependent on 6His::PEPC activity. Thus, the enrichment of the P531A variant appears to be the result of a higher K_{cat} compared to SvPEPC. However, the lower affinity of the top three performing variants P531A, F794V and N863K for bicarbonate, and either equivalent or lower affinity for PEP, suggest that the selection pressures in the *E. coli* Δppc screen are not equivalent to the selection pressures in the mesophyll cell, namely for the higher affinity and specificity for bicarbonate that has repeatedly evolved across C_4 PEPCs (Bauwe, 1986; DiMario and Cousins, 2019; Moody, Christin and Reid, 2020). Conversely, optimisation of K_{cat} at the expense of affinity for substrates is theoretically optimal when substrates are in excess (Klipp and Heinrich, 1994), indicative of likely high abundance of PEP and bicarbonate in *E. coli* Δppc grown aerobically in M9 glucose. Moreover, K_{cat} is

likely to be selected for when flux through the central metabolic enzymes, which includes PEPC, is under strong selection, provided that an equivalent specificity for bicarbonate is maintained (Bar-Even *et al.*, 2011). Therefore, K_{cat} appears to have been selected for during the screen, but the conditions of the screen may need to be altered to better reflect the selection pressures in the C_4 photosynthesis mesophyll cell, for example for a higher affinity of PEPC for bicarbonate.

Modelled photosynthetic rate is dependent on both V_{max} and the affinity of the enzyme for bicarbonate (DiMario *et al.*, 2021; von Caemmerer, 2021; Wang, Chan and Long, 2021). Thus, PEPC variants that have a high affinity for bicarbonate and maintain a high K_{cat} should result in a higher rate of photosynthesis in plants when PEPC activity is rate limiting (von Caemmerer, 2021; Wang, Chan and Long, 2021). Therefore, development of a screen that is undertaken at a low bicarbonate concentration in *E. coli* Δppc should facilitate identification of variants with a high K_{cat} concomitant with a high affinity, and high specificity, of the enzyme for bicarbonate that has repeatedly evolved in C_4 PEPCs (Bauwe, 1986; DiMario and Cousins, 2019; Moody, Christin and Reid, 2020). In addition, not all of the residues that control bicarbonate affinity in C_4 PEPC are known (DiMario *et al.*, 2021). Thus, screening for PEPC variants with a high affinity for bicarbonate should both enable identification of variants that can be engineered into C_4 plants and improve our understanding of residues involved in bicarbonate affinity. Further, screening for both catalytic capacity and affinity for bicarbonate simultaneously would contribute to our understanding of the physiological constraints between K_{cat} and affinity for bicarbonate in PEPC, discussed above.

The scope of the screen for increased enzymatic rate of PEPC in *E. coli* was dependent primarily on the extent to which PEPC expression can be limiting to *E.*

coli Δppc growth rate. For example, if the growth rate of *E. coli* Δppc is reduced by 50% because of limited expression level of PEPC, enzymatic rate *in vivo* can be increased by 100% and this would be observed *via* an increase in bacterial growth rate. The extent to which 6His::PEPC was limiting to growth rate of *E. coli* Δppc in M9 glucose was demonstrated to be ~50% when comparing the growth rate of bacteria transformed with the highest and lowest strength of 6His::PEPC expression vector. This demonstrated that a hypothetical increase in growth rate of 100% was possible relative to the growth rate of the slowest growing strain that had the lowest expression of 6His::PEPC. When PEPC is overexpressed in *E. coli* Δppc grown in M9 glucose, bacterial growth rate remained unchanged or decreased (Chao and Liao, 1993). Thus, any increase in enzymatic rate beyond 100% could reduce the growth rate of the host bacterial cell. Accordingly, the K_{cat} of P531A variant demonstrated a 100% increase in K_{cat} compared to SvPEPC, which was predicted to be the maximal rate of increase that would be selected for in the bacterial screen.

6.6 Differences in PEPC regulation between bacteria and plants

Expression of SvPEPC effectively restores the growth phenotype of *E. coli* Δppc in M9 glucose. However, there are differences in gene expression between eukaryotes and prokaryotes that could introduce different selection pressures in *E. coli* Δppc compared to in the mesophyll cell of a C_4 plant. Most notably, C_4 PEPC activity is tightly regulated by reversible phosphorylation by PEPC kinase in plants, which confers reduced sensitivity of the enzyme to malate inhibition, increased sensitivity to activation by G6P and increases flux through the enzyme (Nimmo, 2003). However, PEPC is not phosphorylated in *E. coli* as the bacteria lack PEPC kinase and bacterial PEPC lacks the required N-terminal phosphorylation domain typical of plant PEPC (Macek *et al.*, 2008; Keseler *et al.*, 2021). To account for the

kinetic differences between phosphorylated and non-phosphorylated PEPC, an alternative screen of PEPC in *E. coli* could be undertaken whereby aspartic acid, which is commonly used as a phosphomimetic for phospho-serine, could be substituted at S15 (*Z. mays* numbering) to mimic covalent bonding of a phosphate group at this residue (Guerra-Castellano *et al.*, 2016). The S15D mutation would likely confer similar conformational and biochemical changes that are observed in a phosphorylated C_4 PEPC. The first step towards this alteration to the screen would be a requirement to confirm that the S15D variant of 6His::PEPC rescues the growth phenotype in M9 glucose. Indeed, it is possible that increased flux through the enzyme observed when PEPC is phosphorylated could help explain how a high bicarbonate affinity can be selected for across C_4 species concurrent with sufficient enzymatic rate during photosynthesis given the putative trade-off observed between K_{cat} and bicarbonate affinity, discussed above and observed in the P531A variant (Kai *et al.*, 1999; Moody, Christin and Reid, 2020; Phansopa *et al.*, 2020).

It remains unclear to what extent phosphorylation influences C_4 PEPC enzyme kinetics *in planta* and it does not appear to always be essential for C_4 PEPC activity; RNAi knockdown of PEPC kinase in the C_4 plant *Flaveria bidentis* does not confer a perturbed phenotype (Furumoto *et al.*, 2007). Thus, if phosphorylation is inessential or only fine-tunes PEPC activity in *S. viridis*, the effect on the transfer of results from the *E. coli* screen into *S. viridis* could be negligible. Furthermore, 5-enolpyruvylshikimate-3-phosphate synthase (EPSPS) and Rubisco variants identified by directed evolution in *E. coli* and subsequently conferred improved plant glyphosate resistance or increased plant growth, respectively (Wilson, Alonso and Whitney, 2016; Mao *et al.*, 2017; Wilson *et al.*, 2018; Zhou and Whitney, 2019). However, these enzymes are not regulated by phosphorylation. Despite differences

between the way genes are expressed in *E. coli* and plants, screening 6His::PEPC variants in *E. coli* is a good way to screen in high-throughput which would not be possible through individual plant transformations. Given that *E. coli* and plant PEPCs are regulated differently *in vivo*, the only way to optimise the screen in *E. coli* is to test the top performing variants in plants, and subsequently refine the parameters of the screen.

6.7 The application of directed evolution to crop plant PEPCs

The Darwin Assembly mutagenesis approach optimised in this study can be tailored to any enzyme, so could easily be applied to any crop plant C_4 PEPC. Sugarcane C_4 PEPC is a key target for engineering an increased K_{cat} because the enzyme V_{max} is low and limits photosynthesis under normal photosynthetic conditions, despite sugarcane being considered the most efficient biomass producer of all cultivated crops (Wang, Chan and Long, 2021). One way of increasing flux through C_4 PEPC in the plant would be to overexpress the enzyme to increase the amount of CO_2 that is fixed. However, constitutive overexpression of C_4 PEPC would place a higher metabolic burden on the plant which could negate any advantages from increased carbon fixation, so increasing the enzymatic rate *via* a small number of site changes would be a preferable alternative. Sugarcane C_4 PEPC is therefore a key target for the high-throughput screen of PEPC in *E. coli* developed here.

The low V_{max} of sugarcane PEPC compared to maize and sorghum is also present despite V794 and K863 residues in both maize and sugarcane. Therefore, despite having evolved from the same origin of C_4 photosynthesis, both enzymes have undergone different changes and currently occupy different places on the PEPC fitness landscape. Indeed, sugarcane PEPC could occupy a local fitness maximum that requires multi-site mutagenesis to escape. The pipeline developed here could

be adapted to accommodate multiple site saturation mutagenesis by Darwin Assembly. However, a maximum of two or three sites must be mutated simultaneously, to be delivered by sequential genome editing, in order for an edited plant to fall under GE regulations. To reduce the combinatorial complexity of mutating two or three sites under selection at once, coevolving sites could be prioritised as these sites could be functionally coupled and the optimal combinations may not have evolved yet in nature (Niklaus and Kelly, 2019). To accommodate the large number of possible variant libraries of enzymes that are co-mutated at two or three sites, many of the steps in the process could be automated in the future, for example the cloning, transformation and PCR steps, to increase throughput of mutagenesis and screening. Alternatively, random mutagenesis, or random mutagenesis and screening coupled to machine-learning insight could be undertaken, representative of a more traditional directed evolutionary approach and an improvement thereof, respectively (Wu *et al.*, 2019; Wang *et al.*, 2021). However, mutation of many independent residues simultaneously could not be delivered by a genome editing approach and these approaches would therefore be less suitable for rapid crop enhancement and commercialisation.

The P531A, F794V and N863K variants were so successful in the assay that almost all other variants were outcompeted. This demonstrates an advantage of screening variants in high-throughput in *E. coli*: the most advantageous variants will become enriched in the population over time and therefore the number of variants that are candidates to be tested in plants is reduced. Sites under selection where mutations have not conferred a fitness advantage in the screen could be under positive selection in C_4 PEPC for reasons other than control of enzyme kinetics, for example involvement in regulation of gene expression. Alternatively, it should be noted that

other sites under selection in C_4 PEPC could also control enzyme kinetics, but these might not be identified in a given screen for three reasons. First, if the selection pressure in the screen is not relevant to the kinetic parameter controlled by the residue. Second, if other single site changes confer a greater advantage to enzyme kinetics and rapidly become more highly enriched. Third, if the changes are only beneficial in conjunction with mutations at other sites under selection. Running multiple screens under different conditions and multi-site saturating mutagenesis approaches could both be rapid ways of identifying additional residues under selection that influence different parameters of C_4 PEPC kinetics either individually or in concert.

The sites under selection analysis was undertaken on grass species, with particularly dense sampling of species within the Panicoideae subfamily (Christin *et al.*, 2007). Agriculturally important crop plants maize, sugarcane and sorghum belong to the Panicoideae, and as such mutagenesis of sites under selection in these species could identify sites that can be engineered to better adapt C_4 PEPC in these crop plants. However, choosing this subset of C_4 species for analysis could bias the result because the *PEPC* genes are likely to have evolved from similar standing variation. Therefore, if Panicoideae PEPCs are increasing their adaptive fitness it may be towards a local rather than global maximum on the enzyme fitness landscape. Analysis to identify sites under selection could be repeated on species across greater phylogenetic diversity, and the greater number of C_4 PEPC sequences that have been made publicly available since 2007 included. Sites that are identified as under selection using a more in-depth and phylogenetically robust analysis could be mutated in crop plant C_4 PEPCs and variants subject to the screen

in *E. coli*. This would improve our understanding of residues that are important for C_4 PEPC kinetics.

6.8 The potential of directed evolution to engineer improved C_4 PEPC variants

The pipeline established in this thesis has demonstrated that hundreds of 6His::PEPC mutant variants can be screened simultaneously in *E. coli* Δppc. The Darwin Assembly mutagenesis approach was highly effective, resulting in a screen of 347 6His::PEPC variants out of a possible 420. Targeting sites under selection for mutagenesis has resulted in the identification of the role residue 531 has in determining K_{cat} and affinity for bicarbonate, a site in C_4 PEPC that has not been previously characterised. Two other variants, F794V and N863K, also became highly abundant during the screen and, given the repeated evolution of residues V794 and 863K in C_4 PEPCs, these are candidate variants for further characterisation in *S. viridis*. Selection for C_4 PEPC that have a higher K_{cat} compared to the wild-type enzyme is therefore a valuable tool for identifying residues that control PEPC kinetics and identifying single amino acid substitutions that can be engineered by genome editing *in planta*.

There are multiple ways in which the screen could be adapted in the future in order to change the parameters of the screen. A benefit of screening C_4 PEPC variants using a directed evolution approach in *E. coli* is that the pipeline is flexible and, once optimised, quick to undertake such that novel fitness solutions that may not have been explored during evolution can be rapidly identified. Once the parameters of the screen have been optimised, when the best-performing PEPC variants from a screen have been demonstrated to confer sustained high levels of C_4

photosynthesis under low stomatal conductance in *S. viridis*, an equivalent approach could be undertaken on C_4 PEPCs from key crop plants. This would facilitate identification of single site changes that confer improved PEPC activity in crop plant C_4 PEPCs that can be rapidly introduced without altering any gene expression or metabolic networks. Here, a proof of concept is presented for a pipeline developed to screen PEPC variants in high throughput in *E. coli* which could become a valuable tool for the design of higher yielding C_4 crop plants.

Appendix

A.1 Golden Gate cloning parts

Plasmid	Part	Antibiotic resistance	Reference
Setaria viridis PEPC* in pUC57	6His::PEPC CDS	Ampicillin	This study
pOGG121	Promoter	Spectinomycin	(Grant, 2018)
pOGG143	High strength RBS	Spectinomycin	(Grant, 2018)
pOGG145	Medium strength RBS	Spectinomycin	(Grant, 2018)
pOGG146	Low strength RBS	Spectinomycin	(Grant, 2018)
pOGG156	Terminator	Spectinomycin	(Grant, 2018)
pOGG024	Vector backbone	Gentamicin	(Geddes, Mendoza-Suárez and Poole, 2019)

*domesticated for *Escherichia coli* expression. RBS = ribosome binding site.

A.1.1 *Setaria viridis* C_4 PEPC (Sevir.4G143500) codon optimised for expression in *Escherichia coli* (SC module), written 5' to 3'

CACTCTGTGGTCTCAAATGCATCATCACCATCACCACGGAGGCGCCTCGAAG
CCGGTAGAGAAGCACCACAGCATTGATGCTCAATTACGTTTATTGGCTCCGG
GTAAGGTGTCGGAAGATGACAAATTGGTCGAGTACGATGCTTTACTTATCGA
CCGTTTCTTGGATATCTTCCAAGACCTTCACGGTCCGAACATTCGTGAATTCG
TCCAGGAATGTTACGAAGTCGCTGCGGAATATGAGCGTGACCGCGATGCCG
CAAAATTATCCGAGCTGGGATCACGCTTGACAAAGCTGTCCCCAAACGACGC
AATCGTGGTAGCTTCAAGCTTCAGTAATATGCTTAACTTAACAAACCTTGCCG
AGGAGGTACAACTGGCTCATCTTCGTCGTAACAAATTAAAGCGTGGGGACTT
CGCTGACGAGGGGTTTGCCGCCACGGAAAGTGACATCGAAGAAACCTTAAA
ACGCCTGGTGACTGAATTGGGCAAGTCGAAAGAAGAAGTCTTTGATGCCCTG
AAAAACCAAACTGTTGACTTAGTATTCACAGCTCATCCCACACAATCCGTCCG
CCGTTCCCTGTTGCAAAAGCATGCGCGCATCCGTAATTGCCTTACACAATTAT

ACGCAAAAGACATTACTGAGGACGACAAACAAGAATTGGATGAGGCTCTTCA
GCGCGAAATTCAGGCGTGTTTTCGTACTGATGAGATTCGTCGCGCGCAGCC
GACCCCCCAAGACGAAATGCGTTATGGTATGTCCTACTTTCACGAAACGATC
TGGAAGGGCGTTCCAAAGTTCCTTCGCCGTGTAGATACGGCCCTGAAAAACA
TTGGAATTGACGAACGTTTGCCGTACAACGCGCCGCTTATCCAGTTTAGCAG
TTGGATGGGAGGCGACCGTGACGGGAACCCCCGCGTAACCCCCGAAGTGA
CACGCGACGTTTGCCTTTTAGCGCGTATGATGGCTAGTAACTTGTACTTTTCT
CAGATCGAAGAGCTTATGTTCGAATTAAGTATGTGGCGCTGCAATGACGAGC
TGCGCGCGAGAGCAGAGGAGCTGCATGCCGCTTCGCAGAAAATCTCTAAAC
ATTATATCGAGTTTTGGCGCCAGTTACCCGCGAGTGAACCGTATCGCGTGGT
CCTGGGTTATGTCCGCGACAAGTTATATAGCACCCGCGAACGCTCGCGTCAC
TTACTTACAAGTGGCTTTAGCGATATCCCAGAAGATAGCGCTTTTAAAAATGT
TGAAGAGTTCCTGGAGCCGCTGGAATTATGCTATAAGTCTTTATGCGATTGTG
GCGATAAGACCATTGCCGATGGCTCGTTACTTGACTTTATGCGTCAAGTAAG
CACCTTCGGTCTTTCTATGGTCAAACTGGATATTCGTCAGGAATCTGAACGTC
ATACTGATGTGATTGACGCTATCACAACCCATCTTGGCATTGGGTCCTATCGC
GAGTGGTCGGAAGAAAAACGTCAGGAGTGGTTATTATCCGAACTGCGTGGAA
AGCGTCCCTTGTTATCGAAAGACATGCCTCAGACAGAGGAAATCGCTGACGT
GCTTGGATGTTTTCACGTGCTGGCAGAATTACCTCGCGATTCTTTTGGGCCG
TATATTATCTCGATGGCGACGGCTCCGTCCGACGTTTTAGCAGTCGAGCTGT
TACAGCGTGAGTGTCACGTCAAGCAGCCCTTACCGGTTGTTCCACTTTTTGA
GAAATTAGCTGACTTGCAGTCAGCACCTGCCTCAATCGAGCGTTTATTCTCAC
TGGACTGGTATATGAATCGTATCGGCGGGAAGCAACAGGTGATGGTAGGATA
TAGTGATTCTGGTAAGGATGCCGGACGCCTGTCGGCGGCATGGCATCTGTAT
AAAGCTCAAGAGGAAATGGCAAAGGTTGCAAAGCGTTATGGCGTAAAATTGA
CTATGTTCCACGGCCGCGGGGGGACTGTCGGACGTGGTGGGGGTCCTACTC
ATCTGGCTATTCTGTCCCAACCCCCTGACACGATCAATGGCTCACTGCGTGT
GACTGTTCAAGGGGAAGTAATCGAGACTAGCTTCGGGGAGGAACATCTTTGT
TTCCGCACTTTACAACGTTTCACAGCTGCCACGCTTGAACATGGAATGCATCC
TCCAGTATCGCCAAAACCTGAGTGGCGTGCGTTAATGGATGAAATCGCCGCT
GTGGCCACTGACGAGTATCGCTCTGTCGTCATGCGCGAGCCGCGCTTCGTG
GAGTACTTCCGCAGCGCGACTCCCGAAACTGAGTACGGCCGCCTTAATATTG
GATCGCGTCCTGCCAAACGCAAGCCTAAGGGAGGGATCGAAAGCCTGCGTG
CCATTCCGTGGATTTTCTCTTGGACGCAAACCCGCTTTCACCTTCCTGTCTGG
CTGGGCTTTGGCGCAGCGTTTAAGCATGCGATGAAGAAGGATATTAAGAACT
TTCAGATGCTTAAAGAAATGTATAATGAGTGGCCGTTTTTCCGTGTTACTCTT
GACTTACTTGAAATGGTCTTAGCAAAGGGTGATCCGACCATCGCTGGATTGT
ATGACCAACTGCTTGTAGCTGACGAGCTTAAGCCTTTTGGCGAACAATTACG
CAACAATTACGTAGAAACCGAAAAGCTTATCCTTCAGGTCGCCGGACATAAA
GAGATTCTTGAGAGTGATCCTGGGCTTAAGCAGCAGCTGCGTTTACGTGATC
CATATATCACTATCCTGAATGTTTGGCAAGCTTATACCTTAAAGCGCATCCGT
GACCCTAATTTCAAAGTAACTCCGCAACCGCCTTTGAGCAAGGAGTTTGCAG
ATGAGAATCAACCGCGCGGTATTGTGAAATTGAATCCTGCCAGCGAATATGG
ACCGGGACTAGAGGATACCCTTATTTTAACCATGAAGGGTATCGCTGCAGGA
ATGCAGAACACTGGCTAACCTCAGCGCTTTGAGACCACGAAGTG

A.1.2 Golden Gate parts, written 5’ to 3’

pOGG121 (P module):

GGAGTTTACGGCTAGCTCAGTCCTAGGTATAGTGCTAGC

pOGG143 (U module): TACTCTAGAATTAAAGAGGAGAAATTAACC

pOGG145 (U module): TACTCTAGAGCTAATCTTCGCGTACTAAAAAGGCAAGT

pOGG146 (U module): TACTCTAGAGCTAATCTTCGCGTACTAAGTAGGCAAGT

pOGG156 (T module):

GCTTAACGCATGAGAAAGCCCCCGGAAGATCACCTTCCGGGGGCTTTTTTAT

TGCGCTCCTT

All gifted by Poole Lab (Oxford).

A.1.3 Placeholder fragment, written 5’ to 3’

LacZ fragment delivered from Twist with two cloning sites: BsaI and BpiI

CACTCTGTGGTCTCAAATGATGTCTTCTCTAGAATTAAAGAGGAGAAATTAAC
CATGACCATGATTACGCCAAGCTTGCATGCCTGCAGGTCGACTCTAGAGGAT
CCCCGGGTACCGAGCTCGAATTCACTGGCCGTCGTTTTACAACGTCGTGACT
GGGAAAACCCTGGCGTTACCCAACTTAATCGCCTTGCAGCACATCCCCCTTT
CGCCAGCTGGCGTAATAGCGAAGAGGCCCGCACCGATCGCCCTTCCCAACA
GTTGCGCAGCCTGAATGGCGAATGGCGCCTGATGCGGTATTTTCTCCTTACG
CATCTGTGCGGTATTTCACACCGCATATGGTGCACTCTCAGTACAATCTGCTC
TGATGCCGCATAGGAAGACATGCTTTGAGACCACGAAGTG

BsaI sites in yellow, BpiI sites in blue, lacZ alpha fragment in grey.

A.1.4 Placeholder vector, written 5’ to 3’

CACGTAGGCCGACACGGTGCCGGGGCCGTTGAGCTTCACGACGGAGATCCA
GCGCTCGGCCACCAAGTCCTTGACTGCGTATTGGACCGTCCGCAAAGAACG
TCCGATGAGCTTGGAAAGAGTCTTTTGGCTGACCACCACGGCGTTCTGGTGG
CCCATCTGCGCGACGAGGTGATGCAGCAGCATTGCCGCCGTGGGTTTCCTC
GCAATAAGCCCGGCCCACGCCTCATGCGCTTTGCGTTCCGTTTGGACCCAGT
GACCGGGCTTGTTCTTGGCTTGAATGCCGATTTCTCTGGACTGCGTGGCCAT
GCTTATCTCCATGCGGTAGGGGTGCCGCACGGTTGCGGCACCATGCGCAAT
CAGCTGCAACTTTTCGGCAGCGCGACAACAATTATGCGTTGCGTAAAAGTGG

CAGTCAATTACAGATTTTCTTTAACCTACGCAATGAGCTATTGCGGGGGGTGC
CGCAATGAGCTGTTGCGTACCCCCCTTTTTTAAGTTGTTGATTTTTAAGTCTTT
CGCATTTCGCCCTATATCTAGTTCTTTGGTGCCCAAAGAAGGGCACCCCTGC
GGGGTTCCCCCACGCCTTCGGCGCGGCTCCCCCTCCGGCAAAAAGTGGCC
CCTCCGGGGCTTGTTGATCGACTGCGCGGCCTTCGGCCTTGCCCAAGGTGG
CGCTGCCCCCTTGGAACCCCCGCACTCGCCGCCGTGAGGCTCGGGGGGCA
GGCGGGCGGGCTTCGCCCTTCGACTGCCCCCACTCGCATAGGCTTGGGTCG
TTCCAGGCGCGTCAAGGCCAAGCCGCTGCGCGGTCGCTGCGCGAGCCTTG
ACCCGCCTTCCTCTTGGTGTCCAACCGGCAAGCGAAGCGCGCAGGCCGCAG
GCCGGAGGCTTTTCCCCAGAGAAAATTAAAAAAATTGATGGGGCAAGGCCGC
AGGCCGCGCAGTTGGAGCCGGTGGGTATGTGGTCGAAGGCTGGGTAGCCG
GTGGGCAATCCCTGTGGTCAAGCTCGTGGGCAGGCGCAGCCTGTCCATCAG
CTTGTCCAGCAGGGTTGTCCACGGGCCGAGCGAAGCGAGCCAGCCGGTGG
CCGCTCGCGGCCATCGTCCACATATCCACGGGCTGGCAAGGGAGCGCAGC
GACCGCGCAGGGCGAAGCCCGGAGAGCAAGCCCGTAGGGGGTTACTCTAG
GGCGGCGGATTTGTCCTACTCAGGAGAGCGTTCACCGACAAACAACAGATAA
AACGAAAGGCCCAGTCTTTCGACTGAGCCTTTCGTTTTATTTGATGCCTGCCG
AATTCGGATCCGGAGtttacggctagctcagtcctaggtatagtgctagctactctagagctaatcttcgcg
tactaagtaggcaagtaatgatgtcttctctagaattaaagaggagaaattaaccatgaccatgattacgccaagc
ttgcatgcctgcaggtcgactctagaggatccccgggtaccgagctcgaattcactggccgtcgttttacaacgtcg
tgactgggaaaaccctggcgttacccaacttaatcgccttgcagcacatccccctttcgccagctggcgtaatagc
gaagaggcccgcaccgatcgcccttcccaacagttgcgcagcctgaatggcgaatggcgcctgatgcggtatttt
ctccttacgcatctgtgcggtatttcacaccgcatatggtgcactctcagtacaatctgctctgatgccgcataggaa
gacatgcttaacgcatgagaaagcccccggaagatcaccttccgggggctttttttattgcgctccttcgctCTTGG
ACTCCTGTTGATAGATCCAGTAATGACCTCAGAACTCCATCTGGATTTGTTCA
GAACGCTCGGTTGCCGCCGGGCGTTTTTTATTGGTGAGAATCCAGGCAATTG
ACATAAGCCTGTTCGGTTCGTAAACTGTAATGCAAGTAGCGTATGCGCTCAC
GCAACTGGTCCAGAACCTTGACCGAACGCAGCGGTGGTAACGGCGCAGTGG
CGGTTTTCATGGCTTGTTATGACTGTTTTTTTGTACAGCCTATGCCTCGGGCA
TCCAAGCAGCAAGCGCGTTACGCCGTGGGTCGATGTTTGATGTTATGGAGCA
GCAACGATGTTACGCAGCAGCAACGATGTTACGCAGCAGGGCAGTCGCCCT
AAAACAAAGTTAGGTGGCTCAAGTATGGGCATCATTCGCACATGTAGGCTCG
GCCCTGACCAAGTCAAATCCATGCGGGCTGCTCTTGATCTTTTCGGTCGTGA
GTTCGGCGACGTAGCCACCTACTCCCAACATCAGCCGGACTCCGATTACCTC
GGGAACTTGCTCCGTAGTAAGACATTCATCGCGCTTGCTGCCTTCGACCAAG
AAGCGGTTGTTGGCGCTCTCGCGGCTTACGTTCTGCCCAAGTTTGAGCAGCC
GCGTAGTGAGATCTATATCTATGATCTCGCAGTCTCCGGAGAGCACCGGAGG
CAGGGCATTGCCACCGCGCTCATCAATCTCCTCAAGCATGAGGCCAACGCG
CTTGGTGCTTATGTGATCTACGTGCAAGCAGATTACGGTGACGATCCCGCAG
TGGCTCTCTATACAAAGTTGGGCATACGGGAAGAAGTGATGCACTTTGATAT
CGACCCAAGTACCGCCACCTAACAATTCGTTCAAGCCGAGATCGGCTTCCCG
GCCGCGGAGTTGTTCGGTAAATTGACTACTTTTCCGCTGCATAACCCTGCTT
CGGGGTCATTATAGCGATTTTTTCGGTATATCCATCCTTTTTCGCACGATATA
CAGGATTTTGCCAAAGGGTTCGTGTAGACTTTCCTTGGTGTATCCAACGGCG
TCAGCCGGGCAGGATAGGTGAAGTAGGCCCACCCGCGAGCGGGTGTTCCTT
CTTCACTGTCCCTTATTCGCACCTGGCGGTGCTCAACGGGAATCCTGCTCTG
CGAGGCTGGCCGTAGGCCGGCCCTACCGGCGCGGCAGCGTTACCCGTGTC
GGCGGCTCCAACGGCTCGCCATCGTCCAGAAAACACGGCTCATCGGGCATC
GGCAGGCGCTGCTGCCCGCGCCGTTCCCATTCCTCCGTTTCGGTCAAGGCT
GGCAGGTCTGGTTCCATGCCCGGAATGCCGGGCTGGCTGGGCGGCTCCTC

GCCGGGGCCGGTCGGTAGTTGCTGCTCGCCCGGATACAGGGTCGGGATGC
GGCGCAGGTCGCCATGCCCCAACAGCGATTCGTCCTGGTCGTCGTGATCAA
CCACCACGGCGGCACTGAACACCGACAGGCGCAACTGGTCGCGGGGCTGG
CCCCACGCCACGCGGTCATTGAC

A.1.5 PEPC PCR product fragment, written 5’ to 3’

GGACTGGAGAAGACCAAATGCATCATCACCATCACCACGGAGGCGCCTCGA
AGCCGGTAGAGAAGCACCACAGCATTGATGCTCAATTACGTTTATTGGCTCC
GGGTAAGGTGTCGGAAGATGACAAATTGGTCGAGTACGATGCTTTACTTATC
GACCGTTTCTTGGATATCTTCCAAGACCTTCACGGTCCGAACATTCGTGAATT
CGTCCAGGAATGTTACGAAGTCGCTGCGGAATATGAGCGTGACCGCGATGC
CGCAAAATTATCCGAGCTGGGATCACGCTTGACAAAGCTGTCCCCAAACGAC
GCAATCGTGGTAGCTTCAAGCTTCAGTAATATGCTTAACTTAACAAACCTTGC
CGAGGAGGTACAACTGGCTCATCTTCGTCGTAACAAATTAAAGCGTGGGGAC
TTCGCTGACGAGGGGTTTGCCGCCACGGAAAGTGACATCGAAGAAACCTTAA
AACGCCTGGTGACTGAATTGGGCAAGTCGAAAGAAGAAGTCTTTGATGCCCT
GAAAAACCAAACTGTTGACTTAGTATTCACAGCTCATCCCACACAATCCGTCC
GCCGTTCCCTGTTGCAAAAGCATGCGCGCATCCGTAATTGCCTTACACAATT
ATACGCAAAAGACATTACTGAGGACGACAAACAAGAATTGGATGAGGCTCTT
CAGCGCGAAATTCAGGCGTGTTTTCGTACTGATGAGATTCGTCGCGCGCAGC
CGACCCCCCAAGACGAAATGCGTTATGGTATGTCCTACTTTCACGAAACGAT
CTGGAAGGGCGTTCCAAAGTTCCTTCGCCGTGTAGATACGGCCCTGAAAAAC
ATTGGAATTGACGAACGTTTGCCGTACAACGCGCCGCTTATCCAGTTTAGCA
GTTGGATGGGAGGCGACCGTGACGGGAACCCCCGCGTAACCCCCGAAGTG
ACACGCGACGTTTGCCTTTTAGCGCGTATGATGGCTAGTAACTTGTACTTTTC
TCAGATCGAAGAGCTTATGTTCGAATTAAGTATGTGGCGCTGCAATGACGAG
CTGCGCGCGAGAGCAGAGGAGCTGCATGCCGCTTCGCAGAAAATCTCTAAA
CATTATATCGAGTTTTGGCGCCAGTTACCCGCGAGTGAACCGTATCGCGTGG
TCCTGGGTTATGTCCGCGACAAGTTATATAGCACCCGCGAACGCTCGCGTCA
CTTACTTACAAGTGGCTTTAGCGATATCCCAGAAGATAGCGCTTTTAAAAATG
TTGAAGAGTTCCTGGAGCCGCTGGAATTATGCTATAAGTCTTTATGCGATTGT
GGCGATAAGACCATTGCCGATGGCTCGTTACTTGACTTTATGCGTCAAGTAA
GCACCTTCGGTCTTTCTATGGTCAAACTGGATATTCGTCAGGAATCTGAACGT
CATACTGATGTGATTGACGCTATCACAACCCATCTTGGCATTGGGTCCTATCG
CGAGTGGTCGGAAGAAAAACGTCAGGAGTGGTTATTATCCGAACTGCGTGGA
AAGCGTCCCTTGTTATCGAAAGACATGCCTCAGACAGAGGAAATCGCTGACG
TGCTTGGATGTTTTCACGTGCTGGCAGAATTACCTCGCGATTCTTTTGGGCC
GTATATTATCTCGATGGCGACGGCTCCGTCCGACGTTTTAGCAGTCGAGCTG
TTACAGCGTGAGTGTCACGTCAAGCAGCCCTTACCGGTTGTTCCACTTTTTGA
GAAATTAGCTGACTTGCAGTCAGCACCTGCCTCAATCGAGCGTTTATTCTCAC
TGGACTGGTATATGAATCGTATCGGCGGGAAGCAACAGGTGATGGTAGGATA
TAGTGATTCTGGTAAGGATGCCGGACGCCTGTCGGCGGCATGGCATCTGTAT
AAAGCTCAAGAGGAAATGGCAAAGGTTGCAAAGCGTTATGGCGTAAAATTGA
CTATGTTCCACGGCCGCGGGGGGACTGTCGGACGTGGTGGGGGTCCTACTC
ATCTGGCTATTCTGTCCCAACCCCCTGACACGATCAATGGCTCACTGCGTGT
GACTGTTCAAGGGGAAGTAATCGAGACTAGCTTCGGGGAGGAACATCTTTGT
TTCCGCACTTTACAACGTTTCACAGCTGCCACGCTTGAACATGGAATGCATCC
TCCAGTATCGCCAAAACCTGAGTGGCGTGCGTTAATGGATGAAATCGCCGCT
GTGGCCACTGACGAGTATCGCTCTGTCGTCATGCGCGAGCCGCGCTTCGTG

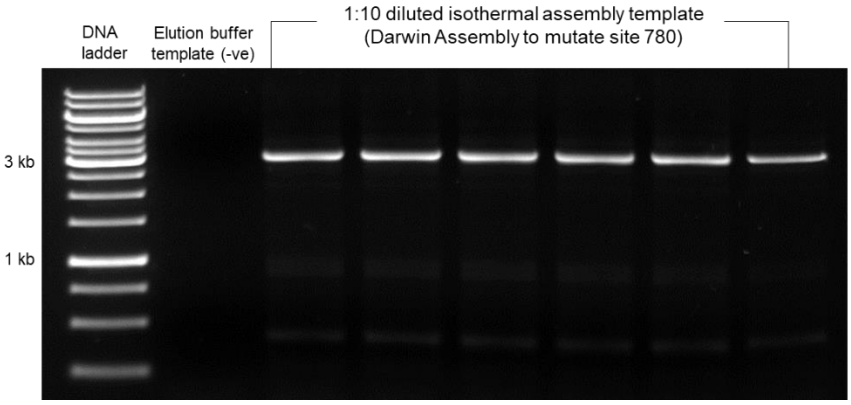
GAGTACTTCCGCAGCGCGACTCCCGAAACTGAGTACGGCCGCCTTAATATTG
GATCGCGTCCTGCCAAACGCAAGCCTAAGGGAGGGATCGAAAGCCTGCGTG
CCATTCCGTGGATTTTCTCTTGGACGCAAACCCGCTTTCACCTTCCTGTCTGG
CTGGGCTTTGGCGCAGCGTTTAAGCATGCGATGAAGAAGGATATTAAGAACT
TTCAGATGCTTAAAGAAATGTATAATGAGTGGCCGTTTTTCCGTGTTACTCTT
GACTTACTTGAAATGGTCTTAGCAAAGGGTGATCCGACCATCGCTGGATTGT
ATGACCAACTGCTTGTAGCTGACGAGCTTAAGCCTTTTGGCGAACAATTACG
CAACAATTACGTAGAAACCGAAAAGCTTATCCTTCAGGTCGCCGGACATAAA
GAGATTCTTGAGAGTGATCCTGGGCTTAAGCAGCAGCTGCGTTTACGTGATC
CATATATCACTATCCTGAATGTTTGGCAAGCTTATACCTTAAAGCGCATCCGT
GACCCTAATTTCAAAGTAACTCCGCAACCGCCTTTGAGCAAGGAGTTTGCAG
ATGAGAATCAACCGCGCGGTATTGTGAAATTGAATCCTGCCAGCGAATATGG
ACCGGGACTAGAGGATACCCTTATTTTAACCATGAAGGGTATCGCTGCAGGA
ATGCAGAACACTGGCTAACCTCAACGCTTTAGTCTTCGGT

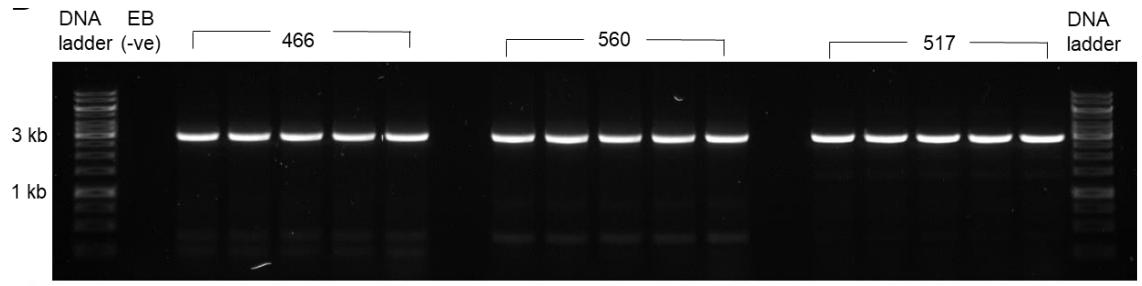
A.2 Additional Darwin Assembly results

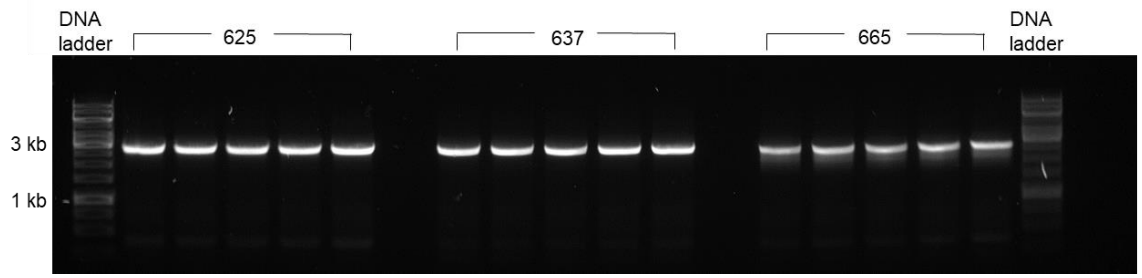
A.2.1 Gels of large-scale PCR for libraries

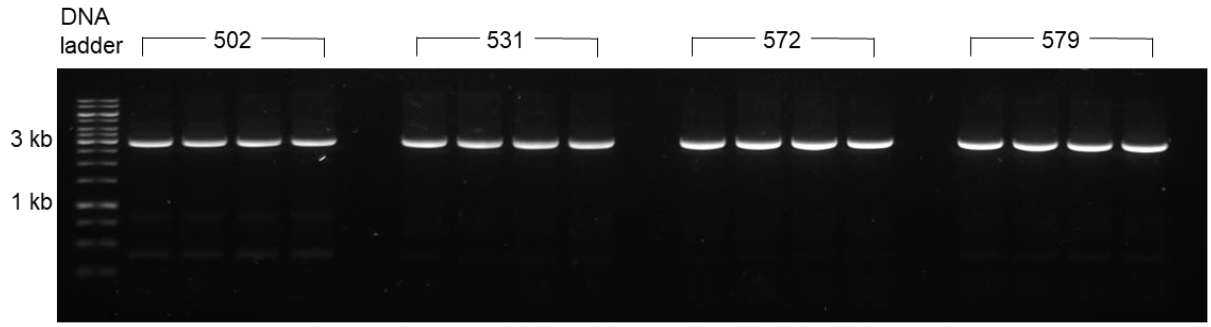
N.B. 780, 466, 560, 517, 625, 637, 665 scaled up 5X.

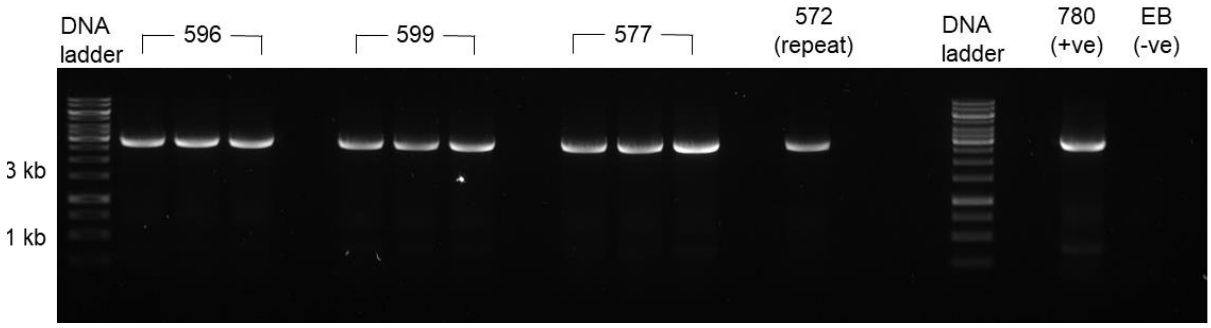
502, 531, 572, 579, 596, 599, 577, 733, 761, 794, 807, 813, 863, 866 all scaled up 3X.

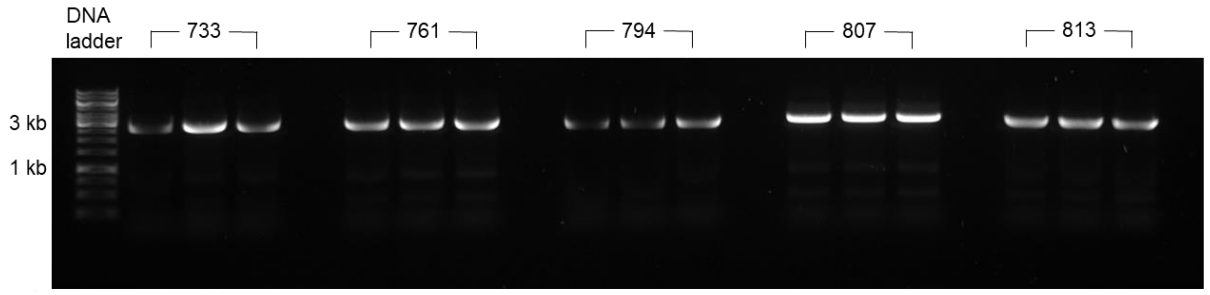


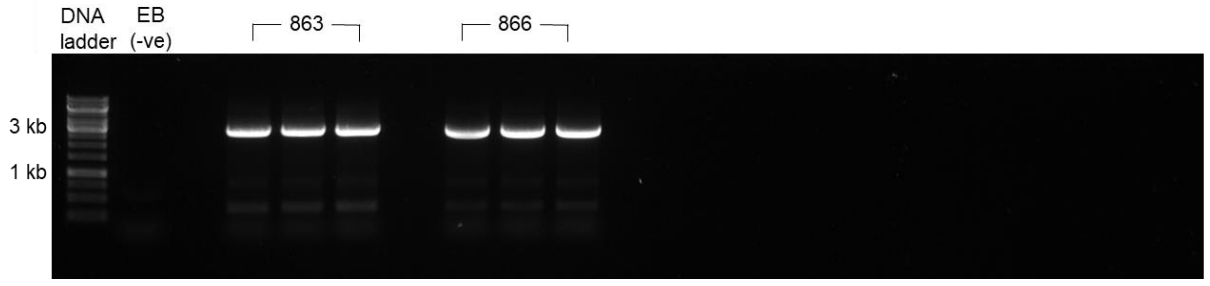





DNA ladder
502
531
572
579
3 kb
1 kb


DNA ladder
596
599
577
572 (repeat)
DNA ladder
780 (+ve)
EB (-ve)
3 kb
1 kb


DNA ladder
733
761
794
807
813
3 kb
1 kb


DNA ladder
EB (-ve)
863
866
3 kb
1 kb

A.2.1 Variants tested/not tested in the large screen of 347 6His::PEPC variants together (across both biological replicates)

Site under selection	Residue number (*Z. mays* numbering)	Amino acids **not** tested
1	466	A
2	502	None
3	517	A, E, H, I, K, L, M, N, P, Q, T, V
4	531	D, I, K, Y
5	560	C, D, E, F, V, Y
6	572	A, C, D, F, G, I, V, W, Y
7	577	C, F, H, I, L, R, W
8	579	I, N
9	596	None
10	599	H, Y
11	625	C, F, H, I, L, M, N, R, Y
12	637	None
13	665	None
14	733	A, C, D, E, F, G, H, P, Q, S, W, Y
15	761	F, M, W
16	780	K, Q
17	794	None
18	807	None
19	813	None
20	863	None
21	866	F, H, L, Y

A.3 Primers

A.3.1 Darwin Assembly boundary and outnest primers

Name	Sequence
PS 5' boundary MUT 18.01.21	GGAGAAGACCAAATGCATCATCACCATCACC with a 5’ Biotin-TEG label
PS 5' outnest MUT 18.01.21	GGACTGGAGAAGACCAAATGCATC

5’ boundary and outnest primers

Name	Sequence
3' boundary MUT PROTECT 19.01	GAATGCAGAACACTGGCTAACCTCAACGCTTTAGTCTTCGG*T 3’ modification is an inverted 3’ base dT
3' outnest MUT 19.01 protect	ACCGAAGACTAAAGCGTTGAGGTTAG

3’ boundary and outnest oligos

5' BandO for PEPC PCR	GGACTGGAGAAGACCAAATGCATCATCACCATCACC
3' BandO for PEPC PCR PROTECT	ACCGAAGACTAAAGCGTTGAGGTTAGCCAGTGTTCTGCATTC

‘Boundary and Outnest’ primers for simulating Darwin Assembly product (A.1.5).

A.3.2 Darwin Assembly mutagenic inner primers

Primer name	Primer sequence
L466I NDT	CATACTGATGTGNDTGACGCTATC
L466I VMA	CATACTGATGTGVMAGACGCTATC
L466I ATG	CATACTGATGTGATGGACGCTATC
L466I TGG	CATACTGATGTGTGGGACGCTATC
G502X NDT	GTCCCTTGTTANDTAAAGACATGC
G502X VMA	GTCCCTTGTTAVMAAAAGACATGC
G502X ATG	GTCCCTTGTTAATGAAAGACATGC
G502X TGG	GTCCCTTGTTATGGAAAGACATGC
IN_6_NDT (site 517)	GTGCTTGGANDTTTTCACG
IN_6_VMA (site 517)	GTGCTTGGAVMATTTCACG
IN_6_ATG (site 517)	GTGCTTGGAATGTTTCACG
IN_6_TGG (site 517)	GTGCTTGGATGGTTTCACG
A531P NDT	CGATTCTTTTGGGNDTTATATTATCTCG
A531P VMA	CGATTCTTTTGGGVMATATATTATCTCG

A531P ATG	CGATTCTTTTGGGATGTATATTATCTCG
A531P TGG	CGATTCTTTTGGGTGGTATATTATCTCG
R560P NDT	CCCTTANDTGTTGTTCCAC
R560P VMA	CCCTTAVMAGTTGTTCCAC
R560P ATG	CCCTTAATGGTTGTTCCAC
R560P TGG	CCCTTATGGGTTGTTCCAC
E572Q NDT	GACTTGNDTTCAGCACCTG
E572Q VMA	GACTTGVMATCAGCACCTG
E572Q ATG	GACTTGATGTCAGCACCTG
E572Q TGG	GACTTGTGGTCAGCACCTG
A577S NDT	ACCTGCCNDTATCGAGC
A577S VMA	ACCTGCCVMAATCGAGC
A577S ATG	ACCTGCCATGATCGAGC
A577S TGG	ACCTGCCTGGATCGAGC
A579E NDT	CCTCAATCNDTCGTTTATTCTC
A579E VMA	CCTCAATCVMACGTTTATTCTC
A579E ATG	CCTCAATCATGCGTTTATTCTC
A579E TGG	CCTCAATCTGGCGTTTATTCTC
E596Q NDT	GAAGCAANDTGTGATGGTAGG
E596Q VMA	GAAGCAAVMAGTGATGGTAGG
E596Q ATG	GAAGCAAATGGTGATGGTAGG
E596Q TGG	GAAGCAATGGGTGATGGTAGG
I599V NDT	CAGGTGATGNDTGGATATAGTG
I599V VMA	CAGGTGATGVMAGGATATAGTG
I599V ATG	CAGGTGATGATGGGATATAGTG
I599V TGG	CAGGTGATGTGGGGATATAGTG
V625A NDT	GAGGAAATGNDTAAGGTTGC
V625A VMA	GAGGAAATGVMAAAGGTTGC
V625A ATG	GAGGAAATGATGAAGGTTGC
V625A TGG	GAGGAAATGTGGAAGGTTGC
M637L NDT	CGTAAAATTGACTNDTTTCCACGG
M637L VMA	CGTAAAATTGACTVMATTCCACGG
M637L ATG	CGTAAAATTGACTATGTTCCACGG
M637L TGG	CGTAAAATTGACTTGGTTCCACGG
H665N NDT	GACACGATCNDTGGCTCACTG
H665N VMA	GACACGATCVMAGGCTCACTG
H665N ATG	GACACGATCATGGGCTCACTG
H665N TGG	GACACGATCTGGGGCTCACTG
F733V NDT	CTGTCGTCNDTCGCGAG
F733V VMA	CTGTCGTCVMACGCGAG
F733V ATG	CTGTCGTCATGCGCGAG
F733V TGG	CTGTCGTCTGGCGCGAG
S761A NDT	CGTCCTNDTAAACGCAAG
S761A VMA	CGTCCTVMAAAACGCAAG
S761A ATG	CGTCCTATGAAACGCAAG
S761A TGG	CGTCCTTGGAAACGCAAG
A780S NDT	GTGGATTTTCNDTTGGACGC
A780S VMA	GTGGATTTTCVMATGGACGC

A780S ATG	GTGGATTTTCATGTGGACGC
A780S TGG	GTGGATTTTCTGGTGGACGC
F794V NDT	TGGCTGGGCNDTGGCGCAGCGTTT
F794V VMA	TGGCTGGGCVMAGGCGCAGCGTTT
F794V ATG	TGGCTGGGCATGGGCGCAGCGTTT
F794V TGG	TGGCTGGGCTGGGGCGCAGCGTTT
R807K NDT	GAAGGATATTNDTAACTTTCAGATGC
R807K VMA	GAAGGATATTVMAAACTTTCAGATGC
R807K ATG	GAAGGATATTATGAACTTTCAGATGC
R807K TGG	GAAGGATATTTGGAACTTTCAGATGC
R813K NDT	CAGATGCTTNDTGAAATGTATAATGAG
R813K VMA	CAGATGCTTVMAGAAATGTATAATGAG
R813K ATG	CAGATGCTTATGGAAATGTATAATGAG
R813K TGG	CAGATGCTTTGGGAAATGTATAATGAG
N863K NDT	CAATTACGCAACNDTTACGTAGAAAC
N863K VMA	CAATTACGCAACVMATACGTAGAAAC
N863K ATG	CAATTACGCAACATGTACGTAGAAAC
N863K TGG	CAATTACGCAACTGGTACGTAGAAAC
D866E NDT	CAATTACGTANDTACCGAAAAGC
D866E VMA	CAATTACGTAVMAACCGAAAAGC
D866E ATG	CAATTACGTAATGACCGAAAAGC
D866E TGG	CAATTACGTATGGACCGAAAAGC

A.3.3 Colony PCR primers

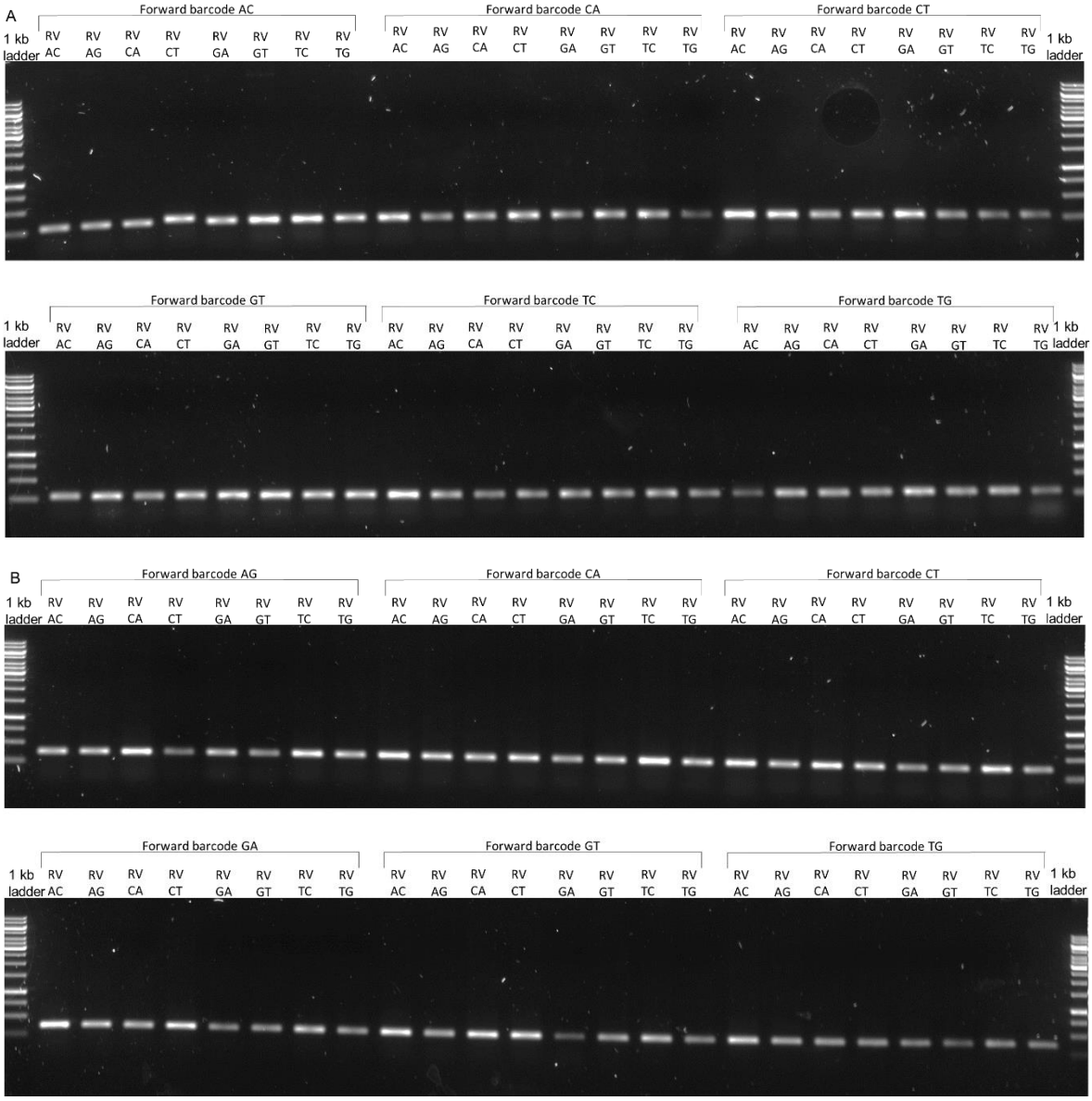
FW (0284)	GGCGGATTTGTCCTA
RV (0283)	AGCGTTCTGAACAAA
PEPC_RV_2395_2414	AGATGAGTAGGACCCCCACC

A.3.4 NGS barcoded primers

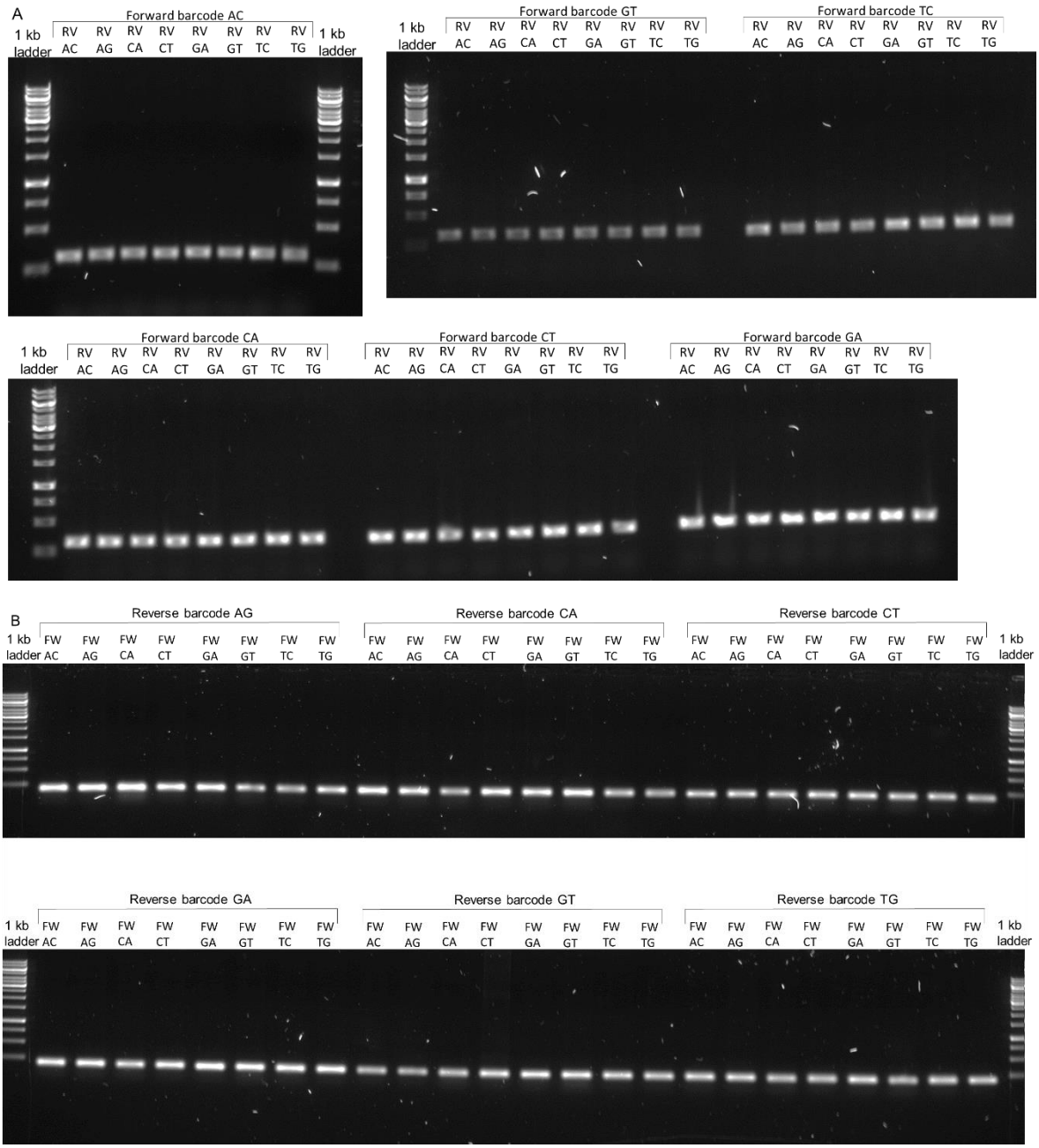
Name	Primer sequence
Frag1_NGS_FW_AC	ATCGACATTCGTCAGGAATCTGAACGTC
Frag1_NGS_FW_CA	ATCGCAATTCGTCAGGAATCTGAACGTC
Frag1_NGS_FW_CT	ATCGCTATTCGTCAGGAATCTGAACGTC
Frag1_NGS_FW_GT	ATCGGTATTCGTCAGGAATCTGAACGTC
Frag1_NGS_FW_TC	ATCGTCATTCGTCAGGAATCTGAACGTC
Frag1_NGS_FW_TG	ATCGTGATTCGTCAGGAATCTGAACGTC
Frag1_NGS_RV_AC	TGACACGCTAAAACGTCGGACGGAGC
Frag1_NGS_RV_AG	TGACAGGCTAAAACGTCGGACGGAGC
Frag1_NGS_RV_CA	TGACCAGCTAAAACGTCGGACGGAGC
Frag1_NGS_RV_CT	TGACCTGCTAAAACGTCGGACGGAGC

Frag1_NGS_RV_TC	TGACTCGCTAAAACGTCGGACGGAGC
Frag1_NGS_RV_TG	TGACTGGCTAAAACGTCGGACGGAGC
Frag1_NGS_RV_GA	GACTGAGCTAAAACGTCGGACGGAGC
Frag1_NGS_RV_GT	GACTGTGCTAAAACGTCGGACGGAGC
Frag2_NGS_FW_CA	GCATCAGTGTCACGTCAAGCAGCC
Frag2_NGS_FW_CT	GCATCTGTGTCACGTCAAGCAGCC
Frag2_NGS_FW_GA	GCATGAGTGTCACGTCAAGCAGCC
Frag2_NGS_FW_GT	GCATGTGTGTCACGTCAAGCAGCC
Frag2_NGS_FW_TG	GCATTGGTGTCACGTCAAGCAGCC
Frag2_NGS_RV_AC	TCAGACAACAGTCACACGCAGTGAG
Frag2_NGS_RV_AG	TCAGAGAACAGTCACACGCAGTGAG
Frag2_NGS_RV_CA	TCAGCAAACAGTCACACGCAGTGAG
Frag2_NGS_RV_CT	TCAGCTAACAGTCACACGCAGTGAG
Frag2_NGS_RV_GA	TCAGGAAACAGTCACACGCAGTGAG
Frag2_NGS_RV_GT	TCAGGTAACAGTCACACGCAGTGAG
Frag2_NGS_RV_TC	TCAGTCAACAGTCACACGCAGTGAG
Frag2_NGS_RV_TG	TCAGTGAACAGTCACACGCAGTGAG
Frag2_NGS_FW_AG	GCATAGGTGTCACGTCAAGCAGCC
Frag3_NGS_FW_AC	AGCTACACGAGTATCGCTCTGTCGTC
Frag3_NGS_FW_CA	AGCTCAACGAGTATCGCTCTGTCGTC
Frag3_NGS_FW_CT	AGCTCTACGAGTATCGCTCTGTCGTC
Frag3_NGS_FW_GA	AGCTGAACGAGTATCGCTCTGTCGTC
Frag3_NGS_FW_GT	AGCTGTACGAGTATCGCTCTGTCGTC
Frag3_NGS_FW_TC	AGCTTCACGAGTATCGCTCTGTCGTC
Frag3_NGS_FW_AG	AGCTAGACGAGTATCGCTCTGTCGTC
Frag3_NGS_RV_AC	TGACACGTCAAGAGTAACACGGAAAAACG
Frag3_NGS_RV_AG	TGACAGGTCAAGAGTAACACGGAAAAACG
Frag3_NGS_RV_CT	TGACCTGTCAAGAGTAACACGGAAAAACG
Frag3_NGS_RV_TG	TGACTGGTCAAGAGTAACACGGAAAAACG
Frag3_NGS_RV_GT	GCTAGTGTCAAGAGTAACACGGAAAAACG
Frag3_NGS_RV_CA	TGACCAGTCAAGAGTAACACGGAAAAACG
Frag3_NGS_RV_TC	GCTATCGTCAAGAGTAACACGGAAAAACG
Frag3_NGS_RV_GA	GCTAGAGTCAAGAGTAACACGGAAAAACG
Frag4_NGS_FW_AC	CAGTACCCATCGCTGGATTGTATGACC
Frag4_NGS_FW_AG	CAGTAGCCATCGCTGGATTGTATGACC
Frag4_NGS_FW_CT	CAGTCTCCATCGCTGGATTGTATGACC
Frag4_NGS_FW_GA	CAGTGACCATCGCTGGATTGTATGACC
Frag4_NGS_FW_GT	CAGTGTCCATCGCTGGATTGTATGACC
Frag4_NGS_FW_TC	CAGTTCCCATCGCTGGATTGTATGACC
Frag4_NGS_FW_TG	CAGTTGCCATCGCTGGATTGTATGACC
Frag4_NGS_FW_CA	TAGCCACCATCGCTGGATTGTATGACC
Frag4_NGS_RV_AG	GTCAAGGTGATATATGGATCACGTAAACGC
Frag4_NGS_RV_CA	GTCACAGTGATATATGGATCACGTAAACGC
Frag4_NGS_RV_GA	GTCAGAGTGATATATGGATCACGTAAACGC
Frag4_NGS_RV_GT	GTCAGTGTGATATATGGATCACGTAAACGC
Frag4_NGS_RV_CT	CATGCTGTGATATATGGATCACGTAAACGC
Frag4_NGS_RV_TG	CATGTGGTGATATATGGATCACGTAAACGC

A.4 Single amplicon of barcoded primer products



A.4 (I) All combinations of barcoded primers for fragments 1 (A) and 2 (B) result in a single PCR product. Figure is continued on the next page.



A.4 (II) All combinations of barcoded primers for fragments 3 (A) and 4 (B) result in a single PCR product. The DNA barcode is at the 5' end of each primer following four 'buffer' bases.

Bibliography

Aldous, S. H. *et al.* (2014) ‘Evolution of the phospho*enol*pyruvate carboxylase protein kinase family in C_3 and C_4 *Flaveria* spp.’, *Plant Physiology*, 165(3), pp. 1076–1091.

Alvarez, C. E. *et al.* (2019) ‘Molecular adaptations of NADP-malic enzyme for its function in C_4 photosynthesis in grasses’, *Nature Plants*, 5, pp. 755–765.

Asano, Y. *et al.* (2011) ‘Functional expression of a plant hydroxynitrile lyase in *Escherichia coli* by directed evolution: creation and characterization of highly *in vivo* soluble mutants’, *Protein Engineering, Design and Selection*, 24(8), pp. 607–616.

Ashworth, J. M. and Kornberg, H. L. (1966) ‘The anaplerotic fixation of carbon dioxide by *Escherichia coli*’, *Proceedings of the Royal Society of London. Series B. Biological Sciences*, 165(999), pp. 179–188.

Aune, T. E. V. and Aachmann, F. L. (2010) ‘Methodologies to increase the transformation efficiencies and the range of bacteria that can be transformed’, *Applied Microbiology and Biotechnology*, 85, pp. 1301–1313.

Baba, T. *et al.* (2006) ‘Construction of *Escherichia coli* K-12 in-frame, single-gene knockout mutants: the Keio collection.’, *Molecular Systems Biology*, 2006.0008.

Bakrim, N. *et al.* (1993) ‘Regulatory phosphorylation of C_4 phospho*enol*pyruvate carboxylase (a cardinal event influencing the photosynthesis rate in sorghum and maize)’, *Plant Physiology*, 101(3), pp. 891–897.

Bar-Even, A. *et al.* (2011) ‘The moderately efficient enzyme: evolutionary and physicochemical trends shaping enzyme parameters’, *Biochemistry*. American Chemical Society, 50(21), pp. 4402–4410.

Bassham, J. A. *et al.* (1956) ‘Intermediates in the photosynthetic cycle’, *Biochimica et Biophysica Acta*, 21, pp. 376–377.

Bauwe, H. (1986) ‘An efficient method for the determination of K_m values for HCO_3^- of phospho*enol*pyruvate carboxylase’, *Planta*, 169, pp. 356–360.

Bauwe, H. and Chollet, R. (1986) ‘Kinetic properties of phospho*enol*pyruvate carboxylase from C_3, C_4, and C_3-C_4 intermediate species of *Flaveria* (Asteraceae)’, *Plant Physiology*, 82(3), pp. 695–699.

Bennett, B. D. *et al.* (2009) ‘Absolute metabolite concentrations and implied enzyme cti ve site occupancy in *Escherichia coli*’, *Nature Chemical Biology*, 5, pp. 593–599.

Berry, J. A., Downton, W. J. S. and Tregunna, E. B. (1970) ‘ The photosynthetic carbon metabolism of *Zea mays* and *Gomphrena globosa* : the location of the CO_2 fixation and the carboxyl transfer reactions ’, *Canadian Journal of Botany*, 48(4), pp. 777–786.

Bhattacharjee, N. and Biswas, P. (2010) ‘Position-specific propensities of amino acids in the -strand’, *BMC Structural Biology*, 10, pp 1-10.

Bhatwa, A. *et al.* (2021) ‘Challenges associated with the formation of recombinant protein inclusion bodies in *Escherichia coli* and strategies to address them for industrial applications’, *Frontiers in Bioengineering and Biotechnology*, 9, 630551.

Blanke, M. M. and Lenz, F. (1989) ‘Fruit photosynthesis’, *Plant, Cell and Environment*, 12(1), pp. 31–46.

Bläsing, O. E. *et al.* (2002) ‘The non-photosynthetic phospho*enol*pyruvate carboxylases of the C_4 dicot *Flaveria trinervia* - Implications for the evolution of C_4 photosynthesis’, *Planta*, 215, pp. 448–456.

Bläsing, O. E., Westhoff, P. and Svensson, P. (2000) ‘Evolution of C_4 phospho*enol*pyruvate 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(36), pp. 27917–27923.

Bouvier, J. W. *et al.* (2021) ‘Rubisco adaptation is more limited by phylogenetic constraint than by catalytic trade-off’, *Molecular Biology and Evolution*, 38(7). pp. 2880–2896.

Bowes, G., Ogren, W. L. and Hageman, R. H. (1971) ‘Phosphoglycolate production catalyzed by ribulose diphosphate carboxylase’, *Biochemical and Biophysical Research Communications*, 45(3), pp. 716–722.

Bowes, G. and Salvucci, M. E. (1989) ‘Plasticity in the photosynthetic carbon metabolism of submersed aquatic macrophytes’, *Aquatic Botany*, 34(1–3), pp. 233–266.

Boxall, S. F. *et al.* (2017) ‘Phosphorylation of phospho*enol*pyruvate carboxylase is essential for maximal and sustained dark CO_2 fixation and core circadian clock operation in the obligate Crassulacean Acid Metabolism species *Kalanchoë fedtschenkoi*’, *The Plant Cell*, 29(10), pp. 2519–2536.

Boyd, R. A., Gandin, A. and Cousins, A. B. (2015) ‘Temperature responses of C_4 photosynthesis: biochemical analysis of Rubisco, phosphoenolpyruvate carboxylase, and carbonic anhydrase in *Setaria viridis* ’, *Plant Physiology*, 169(3), pp. 1850–1861.

Bräutigam, A. and Gowik, U. (2016) ‘Photorespiration connects C_3 and C_4 photosynthesis’, *Journal of Experimental Botany*, 67(10), pp. 2953–2962.

Brown, W. V. (1975) ‘Variations in anatomy, associations, and origins of Kranz tissue’, *American Journal of Botany*, 62(4), p. 395–402.

Brutnell, T. P. *et al.* (2010) ‘*Setaria viridis*: a model for C_4 photosynthesis ’, *The Plant Cell*, 22(8), pp. 2537–2544.

BTX (2022) *Making Electrocompetent Bacteria*. Available at: https://www.btxonline.com/media/wysiwyg/protocol_db/pdfs/MAKING ELECTROCOMPETENT BACTERIA.pdf (Accessed: 20th April 2023).

Burgess, A. J. *et al.* (2022) ‘Improving crop yield potential: underlying biological processes and future prospects’, *Food and Energy Security*, 12(1), p. e435.

Byrt, C. S., Grof, C. P. L. and Furbank, R. T. (2011) ‘C_4 plants as biofuel feedstocks: optimising biomass production and feedstock quality from a lignocellulosic perspective’, *Journal of Integrative Plant Biology*, 53(2), pp. 120–135.

von Caemmerer, S. *et al.* (1997) ‘Reduction of ribulose-1,5-bisphosphate carboxylase/oxygenase by antisense RNA in the C_4 plant *Flaveria bidentis* leads to reduced assimilation rates and increased carbon isotope discrimination’, *Plant Physiology*, 113(2), pp. 469–477.

von Caemmerer, S. *et al.* (2004) ‘Carbonic anhydrase and C_4 photosynthesis: a transgenic analysis’, *Plant, Cell and Environment*, 27(6), pp. 697–703.

von Caemmerer, S. *et al.* (2014) ‘Single cell C_4 photosynthesis in aquatic and terrestrial plants: a gas exchange perspective’, *Aquatic Botany*, 118, pp. 71–80.

von Caemmerer, S. (2021) ‘Updating the steady-state model of C_4 photosynthesis.’, *Journal of Experimental Botany*, 72(17), pp. 6003–6017.

von Caemmerer, S. and Furbank, R. T. (2016) ‘Strategies for improving C_4 photosynthesis’, *Current Opinion in Plant Biology*, 31, pp. 125-134.

Carter, P. J. *et al.* (1993) ‘Circadian rhythms in the activity of a plant protein kinase’, *The EMBO Journal*, 10(8), pp. 2063–2068.

Chao, Y. P. and Liao, J. C. (1993) ‘Alteration of growth yield by overexpression of phosphoenolpyruvate carboxylase and phospho*enol*pyruvate carboxykinase in *Escherichia coli*’, *Applied and Environmental Microbiology*, 59(12), pp. 4261–4265.

Chastain, C. J. *et al.* (2011) ‘Functional evolution of C_4 pyruvate, orthophosphate dikinase’, *Journal of Experimental Botany*, 62(9), pp. 3083–3091.

Chen, K. and Arnold, F. H. (1993) ‘Tuning the activity of an enzyme for unusual environments: sequential random mutagenesis of subtilisin E for catalysis in dimethylformamide.’, *Proceedings of the National Academy of Sciences of the United States of America*, 90(12), pp. 5618–5622.

Chen, K. and Arnold, F. H. (2020) ‘Engineering new catalytic activities in enzymes’, *Nature Catalysis*, 3(3), pp. 203–213.

Cheng, Y. N. *et al.* (2020) ‘Enhancing catalytic efficiency of an *Actinoplanes utahensis* Echinocandin B deacylase through random mutagenesis and site-directed mutagenesis’, *Applied Biochemistry and Biotechnology*, 190(4), pp. 1257–1270.

Chica, R. A., Doucet, N. and Pelletier, J. N. (2005) ‘Semi-rational approaches to engineering enzyme activity: combining the benefits of directed evolution and rational design’, *Current Opinion in Biotechnology*, 16(4), pp. 378–384.

Chockalingam, K. *et al.* (2005) ‘Directed evolution of specific receptor–ligand pairs for use in the creation of gene switches’, *Proceedings of the National Academy of Sciences of the United States of America*, 102(16), pp. 5691–5696.

Chollet, R., Vidal, J. and O’Leary, M. H. (1996) ‘Phosphoenolpyruvate carboxylase: a ubiquitous, highly regulated enzyme in plants’, *Annual Review of Plant Physiology and Plant Molecular Biology*, 47(1), pp. 273–298.

Christin, P.-A. *et al.* (2007) ‘C_4 photosynthesis evolved in grasses *via* parallel adaptive genetic changes’, *Current Biology* , 17(14), pp. 1241–1247.

Christin, P.-A. *et al.* (2011) ‘C_4 eudicots are not younger than C_4 monocots’, *Journal of Experimental Botany*, 62(9), pp. 3171–3181.

Chronopoulou, E. G. *et al.* (2018) ‘Expanding the plant GSTome through directed evolution: DNA shuffling for the generation of new synthetic enzymes with engineered catalytic and binding properties’, *Frontiers in Plant Science*, 9, 1737.

Cobb, R. E., Chao, R. and Zhao, H. (2013) ‘Directed evolution: past, present, and future’, *AIChE Journal*, 59(5), pp. 1432–1440.

Colombo, S. L., Pairoba, C. F. and Andreo, C. S. (1996) ‘Inhibitory effect of shikimic acid on PEP carboxylase activity’, *Plant and Cell Physiology*, 37(6), pp. 870–872.

Confédération paysanne *et al.* (2018) 'Case C-528/16: Confédération paysanne and Others v. Premier ministre and Ministre de l'Agriculture, de l'Agroalimentaire et de la Forêt'. Available at https://curia.europa.eu/juris/document/document.jsf;jsessionid=1E83EA59BCA815F1204ABB5BA393A015?text=&docid=207002&pageIndex=0&doclang=en&mode=req&dir=&occ=first&part=1&cid=5202041 (Accessed: 20th April 2023).

Cousins, A. B. *et al.* (2007) 'The role of phospho*enol*pyruvate carboxylase during C_4 photosynthetic isotope exchange and stomatal conductance', *Plant Physiology*, 145(3), pp. 1006–1017.

Cozens, C. and Pinheiro, V. B. (2018) 'Darwin Assembly: fast, efficient, multi-site bespoke mutagenesis', *Nucleic Acids Research*, 46(8), pp. e51.

Cozzone, A. J. (1998) 'Regulation of acetate metabolism by protein phosphorylation in enteric bacteria', *Annual Review of Microbiology*, 52(1), pp. 127–164.

Crétin, C. (1991) 'Production in *Escherichia coli* of active *Sorghum* phospho*enol*pyruvate carboxylase which can be phosphorylated', *Plant Molecular Biology*, 17, pp. 83–88.

DEFRA (2021) *The regulation of genetic technologies*. Available at: https://consult.defra.gov.uk/agri-food-chain-directorate/the-regulation-of-genetic-technologies (Accessed: 20th April 2023).

Dennig, A. *et al.* (2011) 'OmniChange: the sequence independent method for simultaneous site-saturation of five codons', *PLOS ONE*, 6(10). e26222.

Deroche, M. -E and Carrayol, E. (1988) 'Nodule phospho*enol*pyruvate carboxylase: a review', *Physiologia Plantarum*, 74(4), pp. 775–782.

Dharmarajan, L. *et al.* (2009) ‘Expression, purification and crystallization of an archaeal-type phospho*enol*pyruvate carboxylase.’, *Acta Crystallographica Section F, Structural Biology and Crystallization Communications*, 65, pp. 1193–1196.

DiMario, R. J. *et al.* (2021) ‘Kinetic variation in grass phospho*enol*pyruvate carboxylases provides opportunity to enhance C_4 photosynthetic efficiency’, *The Plant Journal*, 105(6), pp. 1677–1688.

DiMario, R. J. and Cousins, A. B. (2019) ‘A single serine to alanine substitution decreases bicarbonate affinity of phospho*enol*pyruvate carboxylase in C_4 *Flaveria trinervia*’, *Journal of Experimental Botany*, 70(3), pp. 995–1004.

Doncaster, H. D. and Leegood, R. C. (1987) ‘Regulation of phospho*enol*pyruvate carboxylase activity in maize leaves’, *Plant Physiology*, 84(1), pp. 82–87.

Dong, L.-Y., Hata, S. and Izui, K. (1997) ‘High-level expression of maize C_4-type phospho*enol*pyruvate carboxylase in *Escherichia coli* and its rapid purification’, *Bioscience, Biotechnology, and Biochemistry*, 61(3), pp. 545–546.

Dong, L. Y. *et al.* (1998) ‘Cloning, expression, and characterization of a root-form phospho*enol*pyruvate carboxylase from *Zea mays*: comparison with the C_4-form enzyme’, *Plant and Cell Physiology*, 39(8), pp. 865–873.

Dower, W. J., Miller, J. F. and Ragsdale, C. W. (1988) ‘High efficiency transformation of *E. coli* by high voltage electroporation.’, *Nucleic Acids Research*, 16(13), pp. 6127–6145.

Driever, S. M. *et al.* (2017) ‘Increased SBPase activity improves photosynthesis and grain yield in wheat grown in greenhouse conditions’, *Philosophical Transactions of the Royal Society B: Biological Sciences*, 372(1730), 20160384.

Duff, S. M. G. *et al.* (1993) ‘An engineered change in the L-malate sensitivity of a site-directed mutant of sorghum phospho*enol*pyruvate carboxylase: the effect of sequential mutagenesis and S-carboxymethylation at position 8’, *Archives of Biochemistry and Biophysics*, 306(1), pp. 272–276.

Edwards, E. J. *et al.* (2010) ‘The origins of C_4 grasslands: integrating evolutionary and ecosystem science’, *Science*, 328(5978), pp. 587–591.

Edwards, G. E. and Voznesenskaya, E. V (2011) ‘C_4 photosynthesis: Kranz forms and single-cell C_4 in terrestrial plants', in Raghavendra, A. S. and Sage, R. F. (eds) ‘C_4 Photosynthesis and Related CO_2 Concentrating Mechanisms’ pp. 29–61. Springer, Dordrecht.

Ehleringer, J. and Pearcy, R. W. (1983) ‘Variation in quantum yield for CO_2 uptake among C_3 and C_4 plants’, *Plant Physiology*, 73(3), pp. 555–559.

Ehleringer, J. R. *et al.* (1991) ‘Climate change and the evolution of C_4 photosynthesis’, *Trends in Ecology and Evolution*, 6(3), pp. 95–99.

Ehleringer, J. R. and Monson, R. K. (1993) ‘Evolutionary and ecological aspects of photosynthetic pathway variation’, *Annual Review of Ecology and Systematics*, 24, pp. 411–439.

Endo, T. *et al.* (2008) ‘Maize C_4-form phospho*enol*pyruvate carboxylase engineered to be functional in C_3 plants: mutations for diminished sensitivity to feedback inhibitors and for increased substrate affinity', *Journal of Experimental Botany*, 59(7), pp. 1811–1818.

Engelmann, S. *et al.* (2003) ‘Molecular evolution of C_4 phospho*enol*pyruvate carboxylase in the genus *Flaveria* - a gradual increase from C_3 to C_4 characteristics’, *Planta*, 217, pp. 717–725.

Engelmann, S., Bläsing, O. E. and Westhoff, P. (2002) ‘Serine 774 and amino acids 296 to 437 comprise the major C_4 determinants of the C_4 phospho*enol*pyruvate carboxylase of *Flaveria trinervia*’, *FEBS Letters*, 524(1–3), pp. 11–14.

Engler, C., Kandzia, R. and Marillonnet, S. (2008) ‘A one pot, one step, precision cloning method with high throughput capability’, *PLOS ONE*, 3(11), e3647.

Ermakova, M. *et al.* (2019) ‘Overexpression of the Rieske FeS protein of the Cytochrome b_6f complex increases C_4 photosynthesis in *Setaria viridis*’, *Communications Biology*, 2(314).

Ermakova, M. *et al.* (2020) ‘On the road to C_4 rice: advances and perspectives’, *The Plant Journal*, 101(4), pp. 940–950.

Ettema, T. J. G. *et al.* (2004) ‘Identification and functional verification of archaeal-type phospho*enol*pyruvate carboxylase, a missing link in archaeal central carbohydrate metabolism’, *Journal of Bacteriology*, 186(22), pp. 7754–7762.

European Commission (2001) *Directive 2001/18/EC of the European Parliament and of the Council*. Available at: https://www.legislation.gov.uk/eudr/2001/18/contents (Accessed: 20th April 2023).

European Commission (2003) *Regulation 1829/2003/EC*. Available at: https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32003R1829&from=EN (Accessed: 7th March 2023).

European Commission (2021) *Legislation for plants produced by certain new genomic techniques*. Available at: https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/13119-Legislation-for-plants-produced-by-certain-new-genomic-techniques_en (Accessed: 20th April 2023).

European Commission (2023) *EU Register of authorised GMOs*. Available at: https://webgate.ec.europa.eu/dyna2/gm-register/ (Accessed: 23rd January 2023).

FAO (2022) *FAOSTAT* (Food and Agriculture Organization). Available at: https://www.fao.org/faostat/en/#data (Accessed: 21st November 2022).

Fasan, R. *et al.* (2011) 'Improved product-per-glucose yields in P450-dependent propane biotransformations using engineered *Escherichia coli*', *Biotechnology and Bioengineering*, 108(3), pp. 500–510.

Feodorova, T. A. *et al.* (2010) 'Biogeographic patterns of diversification and the origins of C_4 in *Cleome* (Cleomaceae)', *Systematic Botany*, 35(4), pp. 811–826.

Francis, D. M. and Page, R. (2010) 'Strategies to optimize protein expression in *E. coli*', *Current Protocols in Protein Science*, 61(1), pp. 5.24.1-5.24.29.

Freitag, H. and Stichler, W. (2000) 'A remarkable new leaf type with unusual photosynthetic tissue in a central Asiatic genus of Chenopodiaceae', *Plant Biology*, 2(2), pp. 154–160.

Fuentes, L. G. *et al.* (2013) 'Modification of glucose import capacity in *Escherichia coli*: physiologic consequences and utility for improving DNA vaccine production', *Microbial cell factories*, 12, 42.

Furbank, R. T. (2011) ‘Evolution of the C_4 photosynthetic mechanism: are there really three C_4 acid decarboxylation types?’, *Journal of Experimental Botany*, 62(9), pp. 3103–3108.

Furumoto, T. *et al.* (2007) ‘Phosphorylation of phospho*enol*pyruvate carboxylase is not essential for high photosynthetic rates in the C_4 species *Flaveria bidentis*’, *Plant Physiology*, 144(4), pp. 1936–1945.

Gao, Y. and Woo, K. C. (1995) ‘Site-directed mutagenesis of Lys600 in phospho*enol*pyruvate carboxylase of *Flaveria trinervia*: its roles in catalytic and regulatory functions’, *FEBS Letters*, 375(1–2), pp. 95–98.

Garcia-Viloca, M. *et al.* (2004) ‘How enzymes work: analysis by modern rate theory and computer simulations’, *Science*, 303(5655), pp. 186–195.

Gatica-Arias, A. (2020) ‘The regulatory current status of plant breeding technologies in some Latin American and the Caribbean countries’, *Plant Cell, Tissue and Organ Culture*, 141(2), pp. 229–242.

Geddes, B. A., Mendoza-Suárez, M. A. and Poole, P. S. (2019) ‘A bacterial expression vector archive (BEVA) for flexible modular assembly of Golden Gate-compatible vectors’, *Frontiers in Microbiology*, 9, 3345.

Geerlof, A. (2010) *M9 mineral medium*. Available at: https://static.igem.org/mediawiki/2019/2/20/T--Tuebingen--M9_recipe.pdf (Accessed: 20th April 2023).

Ghannoum, O., Evans, J. R. and von Caemmerer, S. (2010) 'Nitrogen and water use efficiency of C_4 plants' in Raghavendra, A. S. and Sage, R. F. (eds) 'C_4 Photosynthesis and Related CO_2 Concentrating Mechanisms' pp. 129–146. Springer, Dordrecht.

Gionfriddo, M., De Gara, L. and Loreto, F. (2019) 'Directed evolution of plant processes: towards a green (r)evolution?', *Trends in Plant Science*, 24(11), pp. 999–1007.

Gokarn, R. R. *et al.* (2001) 'The physiological effects and metabolic alterations caused by the expression of *Rhizobium etli* pyruvate carboxylase in *Escherichia coli*.', *Applied Microbiology and Biotechnology*. 56(1–2), pp. 188–195.

González-Segura, L. *et al.* (2018) 'Identification of the allosteric site for neutral amino acids in the maize C_4 isozyme of phospho*enol*pyruvate carboxylase: the critical role of Ser-100', *Journal of Biological Chemistry*, 293(26), pp. 9945–9957.

Gonzalez, J. E., Long, C. P. and Antoniewicz, M. R. (2017) 'Comprehensive analysis of glucose and xylose metabolism in *Escherichia coli* under aerobic and anaerobic conditions by ^{13}C metabolic flux analysis.', *Metabolic Engineering*, 39, pp. 9–18.

Gowik, U. *et al.* (2004) '*cis*-Regulatory elements for mesophyll-specific gene expression in the C_4 plant *Flaveria trinervia*, the promoter of the C_4 phospho*enol*pyruvate carboxylase gene', *The Plant Cell*, 16(5), pp. 1077–1090.

Gowik, U. *et al.* (2006) 'Evolution of C_4 phospho*enol*pyruvate carboxylase in the genus *Alternanthera*: gene families and the enzymatic characteristics of the C_4 isozyme and its orthologues in C_3 and C_3/C_4 *Alternantheras'*, *Planta*, 223, pp. 359–368.

Grant, K. (2018) *Engineering Rhizobacteria as Synthetic Biology Chassis*. University of Oxford.

Green, M. R. and Sambrook, J. (2020) 'Transformation of *Escherichia coli* by Electroporation', *Cold Spring Harbor Protocols*. Cold Spring Harbor Laboratory Press, 2020(6), p. pdb-prot101220.

Gu, X., Liu, L. and Zhang, H. (2021) 'Transgene-free Genome Editing in Plants', *Frontiers in Genome Editing*, 3, pp. 805317.

Güémez-Toro, R., Mújica-Jiménez, C. and Muñoz-Clares, R. A. (2012) 'Allosteric regulation of the photosynthetic C_4 isoenzyme of phospho*enol*pyruvate carboxylase: a comparative study between enzymes from monocot and dicot plants', *Journal of the Mexican Chemical Society*, 56(1), pp. 58–66.

Guerra-Castellano, A. *et al.* (2016) 'Structural and functional characterization of phosphomimetic mutants of cytochrome *c* at threonine 28 and serine 47', *Biochimica et Biophysica Acta - Bioenergetics*, 1857(4), pp. 387–395.

Haberlandt, G. (1896) 'Physiologische Pflanzenanatomie', Leipzig, Germany: Wilhelm Engelman.

Halford, N. G. (2019) 'Legislation governing genetically modified and genome-edited crops in Europe: the need for change', *Journal of the Science of Food and Agriculture*, 99(1), pp. 8–12.

Hanahan, D. (1983) 'Studies on transformation of *Escherichia coli* with plasmids', *Journal of Molecular Biology*, 166(4), pp. 557–580.

Hanahan, D., Jessee, J. and Bloom, F. R. (1991) 'Plasmid transformation of *Escherichia coli* and other bacteria', *Methods in Enzymology*, 204, pp. 63–113.

Hartwell, J. *et al.* (1996) ‘Higher plant phospho*enol*pyruvate carboxylase kinase is regulated at the level of translatable mRNA in response to light or a circadian rhythm’, *The Plant Journal*, 10(6), pp. 1071–1078.

Hartwell, J. *et al.* (1999) ‘Phospho*enol*pyruvate carboxylase kinase is a novel protein kinase regulated at the level of expression’, *The Plant Journal*, 20(3), pp. 333–342.

Haskett, T. L. *et al.* (2022) ‘Control of nitrogen fixation and ammonia excretion in *Azorhizobium caulinodans*’, *PLOS Genetics*, 18(6), e1010276.

Hatch, M. D. (1987) ‘C_4 photosynthesis: a unique elend of modified biochemistry, anatomy and ultrastructure’, *Biochimica et Biophysica Acta - Reviews on Bioenergetics*, 895(2), pp. 81–106.

Hatch, M. D. and Osmond, C. B. (1976) ‘Compartmentation and transport in C_4 photosynthesis’, in Stocking C. R. and Heber U. (eds) ‘Transport in Plants III: Intracellular Interactions and Transport Processes’ pp. 144–184. Springer, Berlin Heidelberg.

Hatch, M. D. and Slack, C. R. (1966) ‘Photosynthesis by sugar-cane leaves. A new carboxylation reaction and the pathway of sugar formation’, *The Biochemical Journal*, 101(1), pp. 103–111.

Hattersley, P. W. and Browning, A. J. (1981) ‘Occurrence of the suberized lamella in leaves of grasses of different photosynthetic types. I. In parenchymatous bundle sheaths and PCR (“Kranz”) sheaths’, *Protoplasma*, 109, pp. 371–401.

Hibberd, J. M. and Covshoff, S. (2010) ‘The regulation of gene expression required for C_4 photosynthesis’, *Annual Review of Plant Biology*, 61, pp. 181–207.

Hoban, D. J. and Lyric, R. M. (1975) 'Phospho*enol*pyruvate carboxylase of *Thiobacillus thioparus*. I. General properties', *Canadian Journal of Biochemistry*, 53(8), pp. 875–880.

Hogrefe, H. H. *et al.* (2002) 'Creating randomized amino acid libraries with the Quikchange® multi site-directed mutagenesis kit', *BioTechniques*, 33(5), pp. 1158–1165.

Holme, I. B., Gregersen, P. L. and Brinch-Pedersen, H. (2019) 'Induced genetic variation in crop plants by random or targeted mutagenesis: convergence and differences', *Frontiers in Plant Science*, 10, 1468.

IDT (2023) *IDT primer design guidelines*. Available at: https://eu.idtdna.com/pages/education/decoded/article/designing-pcr-primers-and-probes (Accessed: 25th January 2023).

Iñiguez, C. *et al.* (2020) 'Evolutionary trends in RuBisCO kinetics and their co-evolution with CO_2 concentrating mechanisms', *The Plant Journal*, 101(4), pp. 897–918.

ISAAA (2019) *ISAAA Brief 55-2019: Executive Summary*. Available at: https://www.isaaa.org/resources/publications/briefs/55/executivesummary/default.asp#:~:text=supporting self-sufficiency on a,%2C health%2C and social benefits. (Accessed: 16th March 2023).

Izui, K. *et al.* (1981) 'Regulation of *Escherichia coli* phospho*enol*pyruvate carboxylase by multiple effectors *in vivo*. II. Kinetic studies with a reaction system containing physiological concentrations of ligands', *Journal of Biochemistry*, 90(5), pp. 1321–1331.

Izui, K. *et al.* (2004) 'Phospho*enol*pyruvate carboxylase: a new era of structural biology', *Annual Review of Plant Biology*, 55, pp. 69–84.

Jacobs, B. *et al.* (2008) 'Evolution of C_4 phospho*enol*pyruvate carboxylase in *Flaveria*: determinants for high tolerance towards the inhibitor L-malate', *Plant, Cell and Environment*, 31(6), pp. 793–803.

Jahn, M. *et al.* (2016) 'Copy number variability of expression plasmids determined by cell sorting and Droplet Digital PCR.', *Microbial Cell Factories*, 15, 211.

Jain, R. and Srivastava, R. (2009) 'Metabolic investigation of host/pathogen interaction using MS2-infected *Escherichia coli*', *BMC Systems Biology*, 3, 121.

Janc, J. W., O'Leary, M. H. and Cleland, W. W. (1992) 'A kinetic investigation of phospho*enol*pyruvate carboxylase from *Zea mays*', *Biochemistry*, 31(28), pp. 6421–6426.

Jiang, Y. *et al.* (2023) 'How synonymous mutations alter enzyme structure and function over long timescales', *Nature Chemistry*, 15(3), pp. 308–318.

Jiao, J.-A.. *et al.* (1991) '*In vivo* regulatory phosphorylation site in C_4-leaf phospho*enol*pyruvate carboxylase from maize and sorghum', *Plant Physiology*, 96(1), pp. 297–301.

Jiménez-Osés, G. *et al.* (2014) 'The role of distant mutations and allosteric regulation on LovD active site dynamics', *Nature Chemical Biology*, 10(6), pp. 431–436.

Juers, D. H., Matthews, B. W. and Huber, R. E. (2012) '*LacZ* β-galactosidase: structure and function of an enzyme of historical and molecular biological importance', *Protein Science*, 21(12), pp. 1792–1807.

Kai, Y. *et al.* (1999) ‘Three-dimensional structure of phospho*enol*pyruvate carboxylase: a proposed mechanism for allosteric inhibition’, *Proceedings of the National Academy of Sciences of the United States of America*, 96(3), pp. 823–828.

Kai, Y., Matsumura, H. and Izui, K. (2003) ‘Phosphoenolpyruvate carboxylase: three-dimensional structure and molecular mechanisms’, *Archives of Biochemistry and Biophysics*, 414(2), pp. 170–179.

Kanai, R. and Edwards, G. E. (1999) ‘The biochemistry of C_4 photosynthesis’ in Sage, R. F. and Monson, R. K. (eds) ‘C_4 Plant Biology’ pp. 49–87. Academic Press.

Kaur, Jashandeep, Kumar, A. and Kaur, Jagdeep (2018) ‘Strategies for optimization of heterologous protein expression in *E. coli*: roadblocks and reinforcements’, *International Journal of Biological Macromolecules*, 106, pp. 803–822.

Kellogg, E. A. (1999) ‘Phylogenetic aespects of the evolution of C_4 photosynthesis’, in Sage, R. F. and Monson, R. K. (eds) ‘C_4 Plant Biology’ pp. 411–444. Academic Press.

Keseler, I. M. *et al.* (2021) ‘The EcoCyc Database in 2021’, *Frontiers in Microbiology*, 12, 711077.

Kiirats, O. *et al.* (2002) ‘Bundle sheath diffusive resistance to CO_2 and effectiveness of C_4 photosynthesis and refixation of photorespired CO_2 in a C_4 cycle mutant and wild-type *Amaranthus edulis*’, *Plant Physiology*, 130(2), pp. 964–976.

Klipp, E. and Heinrich, R. (1994) ‘Evolutionary optimization of enzyme kinetic parameters; effect of constraints’, *Journal of Theoretical Biology*, 171(3), pp. 309–323.

Kobayashi, A. *et al.* (2006) 'Growth phase-dependent expression of drug exporters in *Escherichia coli* and its contribution to drug tolerance', *Journal of Bacteriology*, 188(16), pp. 5693–5703.

Kornberg, H. L. and Krebs, H. A. (1957) 'Synthesis of cell constituents from C_2-units by a modified tricarboxylic acid cycle', *Nature*, 179(4568), pp. 988–991.

Kovach, M. E. *et al.* (1994) 'pBBR1MCS: a broad-host-range cloning vector', *BioTechniques*, 16(5), pp.800-802.

Kromdijk, J. *et al.* (2016) 'Improving photosynthesis and crop productivity by accelerating recovery from photoprotection', *Science*, 354(6314), pp. 857–861.

Ku, M. S. *et al.* (1999) 'High-level expression of maize phospho*enol*pyruvate carboxylase in transgenic rice plants', *Nature Biotechnology*, 17, pp. 76–80.

Kucera, R. and Cantor, E. (2018) *Breaking through the Limitations of Golden Gate Assembly– The Co-Evolution of Test Systems, Engineered Enzymes and Understanding Ligase Fidelity*, *New England Biolabs Technical Note*. Available at: https://international.neb.com/-/media/nebus/files/application-notes/technote_breaking_through_the_limits_of_golden_gate_assembly.pdf?rev=d8ac2ce3ef664c6c875e126d86153ec1&hash=5558D350DF89CEE31187713BEF575D80 (Accessed: 27th January 2023).

Langdale, J. A. (2011) 'C_4 cycles: past, present, and future research on C_4 photosynthesis', *The Plant Cell*, 23(11), pp. 3879–3892.

Langdale, J. A. (2022) 'Electro-transformationn of *E. coli*'. Available at: https://langdalelab.files.wordpress.com/2015/07/electro_ecoli.pdf (Accessed: 14th December 2022).

Lara, M. V. *et al.* (2006) ‘Species having C_4 single-cell-type photosynthesis in the Chenopodiaceae family evolved a photosynthetic phospho*enol*pyruvate carboxylase like that of Kranz-type C_4 species’, *Plant Physiology*, 142(2), pp. 673–684.

Lassner, M. and Bedbrook, J. (2001) ‘Directed molecular evolution in plant improvement’, *Current Opinion in Plant Biology*, 4(2), pp. 152–156.

Lema, M. A. (2019) ‘Regulatory aspects of gene editing in Argentina’, *Transgenic Research*, 28, pp. 147–150.

Leong, B. J. and Hanson, A. D. (2023) ‘Continuous directed evolution of a feedback-resistant *Arabidopsis* arogenate dehydratase in plantized *Escherichia coli*’, *ACS Synthetic Biology*, 12(1), pp. 43–50.

Lepiniec, L. *et al.* (1994) ‘Phospho*enol*pyruvate carboxylase: structure, regulation and evolution’, *Plant Science*, 99(2), pp. 111–124.

Li, G., Wang, J. bo and Reetz, M. T. (2018) ‘Biocatalysts for the pharmaceutical industry created by structure-guided directed evolution of stereoselective enzymes’, *Bioorganic and Medicinal Chemistry*, 26(7), pp. 1241–1251.

Li, S.-C. *et al.* (1996) ‘Alpha-helical, but not beta-sheet, propensity of proline is determined by peptide environment’, *Proceedings of the National Academy of Sciences of the United States of America*, 93(13), pp. 6676–6681.

Lian, J. *et al.* (2014) ‘Directed evolution of a cellodextrin transporter for improved biofuel production under anaerobic conditions in *Saccharomyces cerevisiae*’, *Biotechnology and Bioengineering*, 111(8), pp. 1521–1531.

Long, S. P., Marshall-Colon, A. and Zhu, X. G. (2015) ‘Meeting the global food demand of the future by engineering crop photosynthesis and yield potential’, *Cell*, 161(1), pp. 56–66.

López-Llano, J., Campos, L.A. and Sancho, J., (2006) ‘α-helix stabilization by alanine relative to glycine: roles of polar and apolar solvent exposures and of backbone entropy’, *Proteins: Structure, Function and Genetics*, 64(3), pp. 769–778.

Ludwig, M. (2016) ‘Evolution of carbonic anhydrase in C_4 plants’, *Current Opinion in Plant Biology*, 31, pp. 16–22.

Macek, B. *et al.* (2008) ‘Phosphoproteome analysis of *E. coli* reveals evolutionary conservation of bacterial Ser/Thr/Tyr phosphorylation’, *Molecular and Cellular Proteomics*, 7(2), pp. 299–307.

Mao, C. *et al.* (2017) ‘Error-prone PCR mutation of *Ls-EPSPS* gene from *Liriope spicata* conferring to its enhanced glyphosate-resistance’, *Pesticide Biochemistry and Physiology*, 141, pp. 90–95.

Marillonnet, S. and Grützner, R. (2020) ‘Synthetic DNA assembly using Golden Gate cloning and the hierarchical modular cloning pipeline’, *Current Protocols in Molecular Biology*, 130(1), e115.

Masumoto, C. *et al.* (2010) ‘Phospho*enol*pyruvate carboxylase intrinsically located in the chloroplast of rice plays a crucial role in ammonium assimilation’, *Proceedings of the National Academy of Sciences of the United States of America*, 107(11), pp. 5226–5231.

Matsumura, H. *et al.* (2002) ‘Crystal structures of C_4 form maize and quaternary complex of *E. coli* phospho*enol*pyruvate carboxylases’, *Structure*, 10(12), pp. 1721–1730.

Matsumura, H., Izui, K. and Mizuguchi, K. (2006) ‘A novel mechanism of allosteric regulation of archaeal phospho*enol*pyruvate carboxylase: a combined approach to structure-based alignment and model assessment’, *Protein Engineering, Design and Selection*, 19(9), pp. 409–419.

McAlister, L. E., Evans, E. L. and Smith, T. E. (1981) ‘Properties of a mutant *Escherichia coli* phosphoenolpyruvate carboxylase deficient in coregulation by intermediary metabolites’, *Journal of Bacteriology*, 146(1), pp. 200–208.

Mendonça, L. M. F. and Marana, S. R. (2011) ‘Single mutations outside the active site affect the substrate specificity in a β-glycosidase’, *Biochimica et Biophysica Acta - Proteins and Proteomics*, 1814(12), pp. 1616–1623.

Merlin, C. *et al.* (2003) ‘Why Is carbonic anhydrase essential to *Escherichia coli*?’, *Journal of Bacteriology*, 185(21), pp. 6415–6424.

Metje-Sprink, J. *et al.* (2019) ‘DNA-free genome editing: past, present and future’, *Frontiers in Plant Science*, 9, 1957.

Miller, E. M. and Nickoloff, J. A. (1995) ‘*Escherichia coli* electrotransformation’, in Nickoloff, J.A. (eds) ‘Electroporation Protocols for Microorganisms. Methods in Molecular Biology’ vol 47, pp. 105–114. Humana Press, New Jersey.

Moody, N. R., Christin, P. A. and Reid, J. D. (2020) ‘Kinetic modifications of C_4 PEPC are qualitatively convergent, but larger in *Panicum* than in *Flaveria’*, *Frontiers in Plant Science*, 11, 1014.

Morikawa, M. *et al.* (1980) ‘Regulation of *Escherichia coli* phospho*enol*pyruvate carboxylase by multiple effectors in vivo I. Estimation of the activities in the cells grown on various compounds’, *Journal of Biochemistry*, 87(2), pp. 441–449.

Mueller-Cajar, O. and Whitney, S. M. (2008) ‘Directing the evolution of Rubisco and Rubisco activase: first impressions of a new tool for photosynthesis research’, *Photosynthesis Research*, 98, pp. 667–675.

Muhaidat, R., Sage, R. F. and Dengler, N. G. (2007) ‘Diversity of Kranz anatomy and biochemistry in C_4 eudicots’, *American Journal of Botany*, 94(3), pp. 362–381.

Muñoz-Clares, R. A. *et al.* (2020) ‘Structural and biochemical evidence of the glucose 6-phosphate-allosteric site of maize C_4-phosphoenolpyruvate carboxylase: its importance in the overall enzyme kinetics’, *The Biochemical Journal*, 477(11), pp. 2095–2114.

New England BioLabs (2023) *BbsI*. Available at: https://international.neb.com/products/r0539-bbsi#Product Information (Accessed: 27th January 2023).

Niklaus, M. and Kelly, S. (2019) ‘The molecular evolution of C_4 photosynthesis: opportunities for understanding and improving the world’s most productive plants’, *Journal of Experimental Botany*, 70(3), pp. 795–804.

Nimmo, H. G. (2003) ‘Control of the phosphorylation of phospho*enol*pyruvate carboxylase in higher plants’, *Archives of Biochemistry and Biophysics*, 414(2), pp. 189–196.

Nishikido, T. *et al.* (1965) ‘Inhibition of the carbon dioxide fixation in *E. coli* by the compounds related to TCA cycle’, *Biochemical and Biophysical Research Communications*, 21(2), pp. 94–99.

Nishikido, T. and Takanashi, H. (1973) ‘Glycine activation of PEP carboxylase from monocotyledoneous C_4 plants’, *Biochemical and Biophysical Research Communications*, 53(1), pp. 126–133.

Norrander, J., Kempe, T. and Messing, J. (1983) ‘Construction of improved M13 vectors using oligodeoxynucleotide-directed mutagenesis’, *Gene*, 26(1), pp. 101–106.

O’Leary, B., Park, J. and Plaxton, W. C. (2011) ‘The remarkable diversity of plant PEPC (phospho*enol*pyruvate carboxylase): recent insights into the physiological functions and post-translational controls of non-photosynthetic PEPCs’, *Biochemical Journal*, 436(1), pp. 15–34.

O’Leary, M. H. (1982) ‘Phosphoenolpyruvate carboxylase: an enzymologist’s view’, *Annual Review of Plant Physiology*, 33(1), pp. 297–315.

Ölçer, H., Lloyd, J. C. and Raines, C. A. (2001) ‘Photosynthetic capacity is differentially affected by reductions in sedoheptulose-1,7-bisphosphatase activity during leaf development in transgenic tobacco plants’, *Plant Physiology*, 125(2), pp. 982–989.

Ort, D. R. *et al.* (2015) ‘Redesigning photosynthesis to sustainably meet global food and bioenergy demand’, *Proceedings of the National Academy of Sciences of the United States of America*, 112(28), pp. 8529–8536.

Outlaw, Jr., W. H. (1990) ‘Kinetic properties of guard-cell phospho*enol*pyruvate carboxylase’, *Biochemie und Physiologie der Pflanzen*, 186(5–6), pp. 317–325.

Pace, N. C. and Scholtz, M. J. (1998) ‘A helix propensity scale based on experimental studies of peptides and proteins’, *Biophysical Journal*, 75(1), 422–427.

Pairoba, C. F., Colombo, S. L. and Andreo, C. S. (1996) ‘Flavonoids as inhibitors of NADP-malic enzyme and PEP carboxylase from C_4 plants’, *Bioscience, Biotechnology, and Biochemistry*, 60(5), pp. 779–783.

Parry, M. A. J. *et al.* (2013) ‘Rubisco activity and regulation as targets for crop improvement’, *Journal of Experimental Botany*, 64(3), pp. 717–730.

Patel, H. M., Kraszewski, J. L. and Mukhopadhyay, B. (2004) ‘The phospho*enol*pyruvate carboxylase from *Methanothermobacter thermautotrophicus* has a novel structure’, *Journal of Bacteriology*, 186(15), pp. 5129–5137.

Paulus, J. K., Schlieper, D. and Groth, G. (2013) ‘Greater efficiency of photosynthetic carbon fixation due to single amino-acid substitution’, *Nature Communications*, 4, 1518.

Phansopa, C. *et al.* (2020) ‘Lateral gene transfer acts as an evolutionary shortcut to efficient C_4 biochemistry’, *Molecular Biology and Evolution*, 37(11), pp. 3094–3104.

Pignon, C. P. and Long, S. P. (2020) ‘Retrospective analysis of biochemical limitations to photosynthesis in 49 species: C_4 crops appear still adapted to pre-industrial atmospheric [CO_2]’, *Plant Cell and Environment*, 43(11), pp. 2606–2622.

Plaxton, W. C. and Podestá, F. E. (2006) ‘The functional organization and control of plant respiration’, *Critical Reviews in Plant Sciences*, 25(2), pp. 159–198.

Pryor, J. M. *et al.* (2020) ‘Enabling one-pot Golden Gate assemblies of unprecedented complexity using data-optimized assembly design’, *PLOS ONE*, 15(9), e0238592.

Puigbò, P. *et al.* (2007) ‘OPTIMIZER: a web server for optimizing the codon usage of DNA sequences’, *Nucleic Acids Research*, 35, pp. W126–W131.

Püllmann, P. *et al.* (2019) ‘Golden Mutagenesis: an efficient multi-site-saturation mutagenesis approach by Golden Gate cloning with automated primer design’, *Scientific Reports*, 9, 10932.

Rajagopalan, A. V., Devi, M. T. and Raghavendra, A. S. (1994) ‘Molecular biology of C_4 phospho*enol*pyruvate carboxylase: structure, regulation and genetic engineering’, *Photosynthesis Research*, 39, pp. 115–135.

Ranson, S. L. and Thomas, M. (1960) ‘Crassulacean Acid Metabolism’, *Annual Review of Plant Physiology*, 11, pp. 81–110.

Rao, X. and Dixon, R. A. (2016) ‘The differences between NAD-ME and NADP-ME subtypes of C_4 photosynthesis: more than decarboxylating enzymes’, *Frontiers in Plant Science*, 7, 1525.

Ray, D. K. *et al.* (2013) ‘Yield trends are insufficient to double global crop production by 2050’, *PLOS ONE*, 8(6), e66428.

Reshes, G. *et al.* (2008) ‘Timing the start of division in *E. coli*: a single-cell study’, *Physical Biology*, 5(4), 46001.

Rhoades, M. M. and Carvalho, A. (1944) ‘The function and structure of the parenchyma sheath plastids of the maize leaf’, *Bulletin of the Torrey Botanical Club*, 71(4), pp. 335–346.

Romero, P. A. and Arnold, F. H. (2009) ‘Exploring protein fitness landscapes by directed evolution’, *Nature Reviews Molecular Cell Biology*, 10(12), pp. 866–876.

Rosnow, J. J., Edwards, G. E. and Roalson, E. H. (2014) ‘Positive selection of Kranz and non-Kranz C_4 phospho*enol*pyruvate carboxylase amino acids in Suaedoideae (Chenopodiaceae)’, *Journal of Experimental Botany*, 65(13), pp. 3595–3607.

Sabe, H. *et al.* (1984) ‘Molecular cloning of the phospho*enol*pyruvate carboxylase gene, *ppc*, of *Escherichia coli*’, *Gene*, 31(1-3), pp. 279–283.

Sage, R. F. (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(14), pp. 4039–4056.

Sage, R. F., Sage, T. L. and Kocacinar, F. (2012) ‘Photorespiration and the evolution of C_4 photosynthesis’, *Annual Review of Plant Biology*, 63, pp. 19–47.

Sako, Y. *et al.* (1996) ‘Purification and characterization of phospho*enol*pyruvate carboxylase from the hyperthermophilic archaeon *Methanothermus sociabilis*’, *FEBS Letters*, 392(2), pp. 148–152.

Sako, Y. *et al.* (1997) ‘Biochemical relationship of phospho*enol*pyruvate carboxylases (PEPCs) from thermophilic archaea’, *FEMS Microbiology Letters*, 153(1), pp. 159–165.

Sakoda, K. *et al.* (2022) ‘Towards improved dynamic photosynthesis in C_3 crops by utilizing natural genetic variation’, *Journal of Experimental Botany*, 73(10), pp. 3109–3121.

Sangwan, R. S., Singh, N. and Plaxton, W. C. (1992) 'Phospho*enol*pyruvate carboxylase activity and concentration in the endosperm of developing and germinating castor oil seeds', *Plant Physiology*, 99(2), pp. 445–449.

Sauer, U. and Eikmanns, B. J. (2005) 'The PEP-pyruvate-oxaloacetate node as the switch point for carbon flux distribution in bacteria', *FEMS Microbiology Reviews*, 29(4), pp. 765–794.

Savile, C. K. *et al.* (2010) 'Biocatalytic asymmetric synthesis of chiral amines from ketones applied to sitagliptin manufacture', *Science*, 329(5989), pp. 303–305.

Sayers, J. R., Schmidt, W. and Eckstein, F. (1988) '5′–3′ Exonucleases in phosphorothioate-based oligonucleotide-directed mutagenesis', *Nucleic Acids Research*, 16(3), pp. 791–802.

Saze, H. *et al.* (2001) 'Thioredoxin-mediated reductive activation of a protein kinase for the regulatory phosphorylation of C_4-form phospho*enol*pyruvate carboxylase from maize', *Plant and Cell Physiology*, 42(12), pp. 1295–1302.

Sedelnikova, O. V., Hughes, T. E. and Langdale, J. A. (2018) 'Understanding the genetic basis of C_4 Kranz anatomy with a view to engineering C_3 crops', *Annual Review of Genetics*, 52, pp. 249–270.

Sharkey, T.D. (1986). 'Theoretical and experimental observations on O_2 sensitivity of C_3 photosynthesis' in Marcelle, R., Clijsters, H., van Poucke, M. (eds) 'Biological Control of Photosynthesis. Advances in Agricultural Biotechnology', vol 19, pp. 115–116. Springer, Dordrecht.

Sheng, Y., Mancino, V. and Birren, B. (1995) ‘Transformation of *Escherichia coli* with large DNA molecules by electroporation’, *Nucleic Acids Research*, 23(11), pp. 1990–1996.

Silverstein, R. and Willis, M. S. (1973) ‘Concerted regulation *in vitro* of phosphoenolpyruvate carboxylase from *Escherichia coli*’, *Journal of Biological Chemistry*, 248(24), pp. 8402–8407.

Smith, L. H. *et al.* (1996) ‘Light regulation of phospho*enol*pyruvate carboxylase in barley mesophyll protoplasts is modulated by protein synthesis and calcium, and not necessarily correlated with phosphoenolpyruvate carboxylase kinase activity’, *Planta*, 200, pp. 174–180.

Smith, T. E. (1971) ‘*Escherichia coli* phospho*enol*pyruvate carboxylase. Physical and chemical properties’, *Journal of Biological Chemistry*, 246(13), pp. 4234–4241.

Smith, T. E. (1977) ‘*Escherichia coli* phosphoenolpyruvate carboxylase: studies on the mechanism of multiple allosteric interactions’, *Archives of Biochemistry and Biophysics*, 183(2), pp. 538–552.

Sowa, S. *et al.* (2021) ‘Legal and practical challenges to authorization of gene edited plants in the EU’, *New Biotechnology*, 60, pp. 183–188.

Sprink, T., Hartung, F., Metje-Sprink, J. (2022) ‘Transgene-Free Genome Editing in Plants’ in Wani, S. H., Hensel, G. (eds) Genome Editing pp. 171–186. Springer, Cham.

Sprink, T., Wilhelm, R. and Hartung, F. (2022) ‘Genome editing around the globe: an update on policies and perceptions’, *Plant Physiology*, 190(3), pp. 1579–1587.

Steel, H. *et al.* (2020) ‘In situ characterisation and manipulation of biological systems with Chi.Bio’, *PLOS Biology*, 18(7), p. e3000794.

Steel, H. (2023) *Chi.Bio*. Available at: chi.bio (Accessed: 14th April 2023).

Stemmer, W. P. C. and Morris, S. K. (1992) ‘Enzymatic inverse PCR: a restriction site independent, single-fragment method for high-efficiency, site-directed mutagenesis’, *BioTechniques*, 13(2), pp. 214–220.

Still, C. J. *et al.* (2003) ‘Global distribution of C_3 and C_4 vegetation: carbon cycle implications’, *Global Biogeochemical Cycles*, 17(1), pp. 6-1–6-14.

Studer, A. J. *et al.* (2014) ‘A limited role for carbonic anhydrase in C_4 photosynthesis as revealed by a *ca1ca2* double mutant in maize’, *Plant Physiology*, 165(2), pp. 608–617.

Svensson, P., Bläsing, O. E. and Westhoff, P. (1997) ‘Evolution of the enzymatic characteristics of C_4 phospho*enol*pyruvate carboxylase. A comparison of the orthologous PPCA phospho*enol*pyruvate carboxylases of *Flaveria trinervia* (C_4) and *Flaveria pringlei* (C_3)’, *European Journal of Biochemistry*, 246(2), pp. 452–460.

Svensson, P., Bläsing, O. E. and Westhoff, P. (2003) ‘Evolution of C_4 phospho*enol*pyruvate carboxylase’, *Archives of Biochemistry and Biophysics*, 414(2), pp. 180–188.

Takahashi-Terada, A. *et al.* (2005) ‘Maize phospho*enol*pyruvate carboxylase: mutations at the putative binding site for glucose 6-phosphate caused desensitization and abolished responsiveness to regulatory phosphorylation’, *Journal of Biological Chemistry*, 280(12), pp. 11798–11806.

Tang, L. *et al.* (2012) ‘Construction of “small-intelligent” focused mutagenesis libraries using well-designed combinatorial degenerate primers.’, *BioTechniques*, 52(3), pp. 149–158.

Tausta, S. L. *et al.* (2002) ‘Maize C_4 and non-C_4 NADP-dependent malic enzymes are encoded by distinct genes derived from a plastid-localized ancestor’, *Plant Molecular Biology*, 50, pp. 635–652.

Terada, K. and Izui, K. (1991) ‘Site-directed mutagenesis of the conserved histidine residue of phospho*enol*pyruvate carboxylase’, *European Journal of Biochemistry*, 202(3), pp. 797–803.

Terada, K., Murata, T. and Izui, K. (1991) ‘Site-directed mutagenesis of phospho*enol*pyruvate carboxylase from *E. coli*: the role of His579 in the catalytic and regulatory functions’, *Journal of Biochemistry*, 109(1), pp. 49–54.

Thakur, C. S. *et al.* (2010) ‘Growth of wildtype and mutant *E. coli* strains in minimal media for optimal production of nucleic acids for preparing labeled nucleotides’, *Applied Microbiology and Biotechnology*, 88, pp. 771–779.

ThermoFisher Scientific (2023) *BpiI (BbsI) (10U/uL)*. Available at: https://www.thermofisher.com/order/catalog/product/ER1011 (Accessed: 27th January 2023).

Tian, X. *et al.* (2021) ‘Will reaching the maximum achievable yield potential meet future global food demand?’, *Journal of Cleaner Production*, 294, 126285.

Tilman, D. *et al.* (2011) ‘Global food demand and the sustainable intensification of agriculture’, *Proceedings of the National Academy of Sciences of the United States of America*, 108(50), pp. 20260–20264.

Ting, I. P. and Osmond, C. B. (1973) ‘Photosynthetic phospho*enol*pyruvate carboxylases characteristics of alloenzymes from leaves of C_3 and C_4 plants’, *Plant Physiology*, 51(3), pp. 439–447.

Tivendale, N. D. *et al.* (2020) ‘Enzymes as parts in need of replacement – and how to extend their working life’, *Trends in Plant Science*, 25(7), pp. 661–669.

Tolbert, N. E. and Yamazaki, R. K. (1969) ‘Leaf peroxisomes and their relation to photorespiration and photosynthesis’, *Annals of the New York Academy of Sciences*, 168(2), pp. 325–341.

Tomoiaga, D. *et al.* (2022) ‘High rates of plasmid cotransformation in *E. coli* overturn the clonality myth and reveal colony development’, *Scientific Reports*, 12, 11515.

Tong, M. *et al.* (2020) ‘Gene dispensability in *Escherichia coli* grown in thirty different carbon environments’, *mBio*, 11(5).

Tripathi, N. K. and Shrivastava, A. (2019) ‘Recent developments in bioprocessing of recombinant proteins: expression hosts and process development’, *Frontiers in Bioengineering and Biotechnology*, 7, 420.

Trudeau, D. L. *et al.* (2018) ‘Design and in vitro realization of carbon-conserving photorespiration’, *Proceedings of the National Academy of Sciences of the United States of America*, 115(49), pp. e11455–e11464.

Tsai, C. S. (1967) ‘Spontaneous decarboxylation of oxalacetic acid’, *Canadian Journal of Chemistry*, 45(8), pp. 873–880.

Turnbull, C., Lillemo, M. and Hvoslef-Eide, T. A. K. (2021) ‘Global Regulation of Genetically Modified Crops Amid the Gene Edited Crop Boom – A Review’, *Frontiers in Plant Science*, 12, 630396.

Vance, C. P. *et al.* (1994) ‘Primary assimilation of nitrogen in alfalfa nodules: molecular features of the enzymes involved’, *Plant Science*, 101(1), pp. 51–64.

Vidal, J. and Chollet, R. (1997) ‘Regulatory phosphorylation of C_4 PEP carboxylase’, *Trends in Plant Science*, 2(6), pp. 230–237.

Villarreal, J. C. and Renner, S. S. (2012) ‘Hornwort pyrenoids, carbon-concentrating structures, evolved and were lost at least five times during the last 100 million years’, *Proceedings of the National Academy of Sciences of the United States of America*, 109(46), pp. 18873–18878.

Voznesenskaya, E. V. *et al.* (2001) ‘Kranz anatomy is not essential for terrestrial C_4 plant photosynthesis’, *Nature*, 414, pp. 543–546.

Wagner, A. (2005) ‘Energy constraints on the evolution of gene expression’, *Molecular Biology and Evolution*, 22(6), pp.1365-1374.

Wang, Y. *et al.* (2021) ‘Directed evolution: methodologies and applications’, *Chemical Reviews*, 121(20), pp. 12384–12444.

Wang, Y. H. *et al.* (1992) ‘Site-directed mutagenesis of the phosphorylatable serine (Ser8) in C_4 phospho*enol*pyruvate carboxylase from sorghum: the effect of negative charge at position 8’, *Journal of Biological Chemistry*, 267(24), pp. 16759– 16762.

Wang, Y., Chan, K. X. and Long, S. P. (2021) ‘Towards a dynamic photosynthesis model to guide yield improvement in C_4 crops’, *The Plant Journal*, 107(2), pp. 343–359.

Wang, Y., Long, S. P. and Zhu, X. G. (2014) ‘Elements required for an efficient NADP-malic enzyme type C_4 photosynthesis’, *Plant Physiology*, 164(4), pp. 2231–2246.

Weber, E. *et al.* (2011) ‘A modular cloning system for standardized assembly of multigene constructs’, *PLOS ONE*, 6(2), e16765.

van der Weijde, T. *et al.* (2013) ‘The potential of C_4 grasses for cellulosic biofuel production’, *Frontiers in Plant Science*, 4, 107.

Weiss, T. *et al.* (2020) ‘Optimization of multiplexed CRISPR/Cas9 system for highly efficient genome editing in *Setaria viridis*’, *The Plant Journal*, 104(3), pp. 828–838.

Wells, J. A., Vasser, M. and Powers, D. B. (1985) ‘Cassette mutagenesis: an efficient method for generation of multiple mutations at defined sites’, *Gene*, 34(2–3), pp. 315–323.

Westhoff, P. and Gowik, U. (2004) ‘Evolution of C_4 phospho*enol*pyruvate carboxylase. Genes and proteins: a case study with the genus *Flaveria’*, *Annals of Botany*, 93(1), pp. 13–23.

Whelan, A. I., Gutti, P. and Lema, M. A. (2020) ‘Gene Editing Regulation and Innovation Economics’, *Frontiers in Bioengineering and Biotechnology*, 8, 303.

Williams, B. P., Aubry, S. and Hibberd, J. M. (2012) ‘Molecular evolution of genes recruited into C_4 photosynthesis’, *Trends in Plant Science*, 17(4), pp. 213–220.

Wilson, R. H. *et al.* (2018) ‘An improved *Escherichia coli* screen for Rubisco identifies a protein-protein interface that can enhance CO_2-fixation kinetics’, *Journal of Biological Chemistry*, 293(1), pp. 18–27.

Wilson, R. H., Alonso, H. and Whitney, S. M. (2016) ‘Evolving *Methanococcoides burtonii* archaeal Rubisco for improved photosynthesis and plant growth’, *Scientific Reports*, 6(1), 22284.

Wolfenden, R., Lewis, C. A. and Yuan, Y. (2011) ‘Kinetic challenges facing oxalate, malonate, acetoacetate, and oxaloacetate decarboxylases’, *Journal of the American Chemical Society*, 133(15), pp. 5683–5685.

Wolt, J. D. and Wolf, C. (2018) ‘Policy and governance perspectives for regulation of genome edited crops in the United States’, *Frontiers in Plant Science*, 9, 1606.

Wu, Y. *et al.* (2012) ‘Glyceraldehyde-3-phosphate dehydrogenase: a universal internal control for Western blots in prokaryotic and eukaryotic cells’, *Analytical Biochemistry*, 423(1), pp. 15–22.

Wu, Z. *et al.* (2019) ‘Machine learning-assisted directed protein evolution with combinatorial libraries’, *Proceedings of the National Academy of Sciences of the United States of America*, 116(18), pp. 8852–8858.

Xiong, A. S. *et al.* (2007) ‘High efficiency and throughput system in directed evolution *in vitro* of reporter gene’, *Applied Microbiology and Biotechnology*, 74, pp. 160–168.

Xiong, A. S. *et al.* (2011) ‘A thermostable β-glucuronidase obtained by directed evolution as a reporter gene in transgenic plants’, *PLOS ONE*, 6(11), e26773.

Xu, D. *et al.* (2018) ‘Removal of feedback inhibition of *Corynebacterium glutamicum* phospho*enol*pyruvate carboxylase by addition of a short terminal peptide’, *Biotechnology and Bioprocess Engineering*, 23, pp. 72–78.

Yang, H. *et al.* (2018) ‘Evolving artificial metalloenzymes *via* random mutagenesis’, *Nature Chemistry*, 10(3), pp. 318–324.

Yano, M. *et al.* (1995) ‘Catalytic role of an arginine residue in the highly conserved and unique sequence of phospho*enol*pyruvate carboxylase’, *Journal of Biochemistry*, 117(6), pp. 1196–1200.

Yano, M. and Izui, K. (1997) ‘The replacement of Lys620 by serine desensitizes *Escherichia coli* phospho*enol*pyruvate carboxylase to the effects of the feedback inhibitors L- aspartate and L-malate’, *European Journal of Biochemistry*, 247(1), pp. 74–81.

Ying, S. *et al.* (2017) ‘Regulatory phosphorylation of bacterial-type PEP carboxylase by the Ca^{2+}-dependent protein kinase RcCDPK1 in developing castor oil seeds’, *Plant Physiology*, 174(2), pp. 1012–1027.

Ymer, S. (1991) ‘Head inactivation of DNA ligase prior to electroporation increases transformation efficiency’, *Nucleic Acids Research*, 19(24), 6960.

You, L. and Arnold, F. H. (1996) ‘Directed evolution of subtilisin E in *Bacillus subtilis* to enhance total activity in aqueous dimethylformamide’, *Protein Engineering*, 9(1), pp. 77–83.

Yuan, J. *et al.* (2006) ‘The regulatory role of residues 226-232 in phospho*enol*pyruvate carboxylase from maize’, *Photosynthesis Research*, 88, pp. 73–81.

Zhao, H. and Arnold, F. H. (1997) ‘Combinatorial protein design: strategies for screening protein libraries’, *Current Opinion in Structural Biology*, 7(4), pp. 480–485.

Zhao, H., Chockalingam, K. and Chen, Z. (2002) ‘Directed evolution of enzymes and pathways for industrial biocatalysis’, *Current Opinion in Biotechnology*, 13(2), pp. 104–110.

Zhou, H., Akçay, E. and Helliker, B. R. (2019) ‘Estimating C_4 photosynthesis parameters by fitting intensive A/C_i curves’, *Photosynthesis Research*, 141, pp. 181–194.

Zhou, Y. and Whitney, S. (2019) ‘Directed evolution of an improved Rubisco; *in vitro* analyses to decipher fact from fiction’, *International Journal of Molecular Sciences*, 20(20), 5019.

Zhu, X. G., Long, S. P. and Ort, D. R. (2008) ‘What is the maximum efficiency with which photosynthesis can convert solar energy into biomass?’, *Current Opinion in Biotechnology*, 19(2), pp. 153–159.

Zhu, X. G., Long, S. P. and Ort, D. R. (2010) ‘Improving photosynthetic efficiency for greater yield’, *Annual Review of Plant Biology*, 61, pp. 235–261.