THE UNIVERSITY
of EDINBURGH

An integrated platform for the accelerated engineering of microorganisms: Application to industrial bioprocessing

Thesis presented for the degree of

Doctor of Philosophy,

University of Edinburgh

Stephen McColm

2019

Supervisors: Dr Colin J. Campbell and Dr Franck Escalettes

Declaration of Authorship

I, Stephen McColm, hereby certify that this thesis has been composed by me and that it is a record of my work, and that it has not been accepted in partial or complete fulfilment of any other degree or professional qualification.

Stephen McColm

The University of Edinburgh

2019

Signed:

Date:

Abstract

Due to climate change and uncertainties in global fuel prices, there is a need to adopt biomass derived feed stocks for sustainable manufacturing of fuels, chemicals and pharmaceuticals. As a result, many major industrial manufacturers are now seeking routes to their products that are sustainable, more efficient, and less waste or energy intensive. While bioprocesses to produce compounds ranging from therapeutic drugs to fuels have already been widely implemented, the current microbes being employed are often relatively inefficient and limited in the feedstocks they can utilise. Inherent to successful bioprocess development is the ability to rapidly and predictably engineer microbes for the efficient flux of simple biomass towards compounds of industrial significance. Current iterative and empirical processes for microbial strain improvement are limited and therefore improved enabling technologies to accelerate these processes are required.

To address these issues, this thesis describes the development of a platform for the rapid and predictable engineering of microbes for industrial bioprocesses. This has been achieved through complementing an accelerated DNA assembly technique for biosynthetic pathway construction with quantitative proteomics to identify pathway bottlenecks and guide subsequent rounds of pathway optimisation. Only through the ability to rapidly construct biosynthetic pathways and then assess the failure or success of the introduced pathways can microbes be engineered in an intuitive and predictable manner.

A central theme of this thesis is the optimisation and implementation of a DNA assembly technique for the construction of multicomponent pathways. Despite being a fundamental aspect of strain engineering, DNA assembly is often unreliable and time consuming. One limitation of this technique is the reduced efficiency observed in the assembly of multiple DNA fragments (as is often the case when constructing a heterologous pathway). To overcome this issue a ‘nested’ DNA assembly methodology has been developed for the predictable construction of combinatorial vector libraries and complex vectors resulting in the successful assembly of up to 10

fragments. Appropriate shake flask and microtiter plate assays were additionally developed to characterise these constructs. A parallel strand of this work has been the optimisation of the methodology to maximise throughput and efficiency whilst also ensuring the method is amenable to process automation.

To exemplify the power of proteomics in guiding strain engineering the reverse glyoxylate shunt was selected as a simple benchmark heterologous pathway in the commonly used host, *Escherichia coli*. This pathway allows for the conversion of tricarboxylic acid cycle intermediates malate and succinate to oxaloacetate and two molecules of acetyl-CoA. Strains have been engineered to overexpress the pathway genes and tryptic digestions of cell lysates carried out. Liquid chromatography, mass spectrometry and data analysis methods have been developed for the identification of over 100 proteins from these lysates. Work was then focused on developing quantitative acquisitions which will allow for the identification of pathway bottlenecks. The coupling of techniques for pathway engineering and pathway analysis will create a step change in the speed and predictability with which microbes can be engineered for industrial application.

Acknowledgements

I would firstly like to thank my supervisors, Dr Colin J. Campbell and Dr Franck Escalettes, for their valued input, guidance and support throughout his work. I greatly appreciate both the time and effort you have dedicated. I'm grateful for all I have learnt and the experience I have gained your supervision.

I am also very grateful for the mass spectrometry group in the School of Chemistry. Thank you to Dr Logan Mackay and Dr David Clarke for welcoming me into the lab and opening up the world of protein mass spectrometry. Your patience, training and advice has been invaluable.

I would like to thank the Royal Commission for Exhibition of 1851 for awarding me an industrial fellowship to carry out this work. The opportunity to work towards a PhD at the interface of industry and academia is something I am truly grateful for.

Special thanks must go to Ingenza for firstly nurturing my interest in biotechnology, training me in techniques spanning molecular biology and biochemistry and for supporting me through this work. Every member of staff I have worked has contributed to my learning and understanding and for that I extremely thankful.

Finally, I would like to thank Lois and my parents for your support throughout this PhD, without you this wouldn't have been possible.

Lay summary

Industrial biotechnology provides an opportunity to decrease the current reliance on fossil fuels through the development of bioprocesses to convert renewable carbon sources into industrially useful compounds. These products include, but are not limited to, fine and bulk chemicals, polymer intermediates, therapeutics, consumer products or renewable fuels.

This approach generally involves the engineering of a microbial host to efficiently catalyse the conversion of a sustainable feedstocks into the product of interest. This can be achieved – for example - through manipulating the microbial host to over-produce native and heterologous (foreign) enzymes through the introduction of DNA and reducing or eliminating competing reactions through the removal of native DNA.

A key bottleneck in the implementation of this approach, however, is the speed and predictability with which microbes can be engineered to produce the product of interest. This results in the development of a bioprocess becoming a costly and time consuming endeavour for the end user.

In this thesis I will develop suite of techniques which aim to bring increasing predictability and overcome persistent limitations associated with today's iterative and empirical processes for microbial strain improvement.

Glossary

Mass Spectrometry

ACN – Acetonitrile

CID – Collision Induced Dissociation

Da – Dalton

DDA – Data dependent acquisition

DIA – Data independent acquisition

ESI – Electrospray Ionisation

FA – Formic acid

ICAT - Isotope-coded affinity tagging

iTRAQ - Isobaric tags for relative and absolute quantification

kDa – Kilodalton

LC – Liquid chromatography

m/z – Mass to charge ratio

MALDI – Matrix assisted laser desorption ionisation

MS – Mass spectrometry

MS/MS – Tandem mass spectrometry

MS^E - Data independent acquisition which alternates between high energy and low energy scans

RT – Retention time

SILAC- Stable isotope labelling by amino acids in cell culture

Q-TOF – Quadrupole Time of Flight

Molecular Biology

ABTS – 2,2'-azino-bis(3-ethylbenzothiazoline-6-sulphonic acid)

bp – base pair

CFU – Colony forming unit

CRISPR - Clustered regularly interspaced short palindromic repeats

DNA – Deoxyribonucleic acid

dsDNA – Double stranded DNA

dNTPs - Deoxyribonucleotides

ELISA - Enzyme-linked immunosorbent assay

eGFP – Enhanced green fluorescent protein

ExoIII – Exonuclease III

LOl – Linker oligo long

LOs – Linker oligo short

LOA – Linker oligo annealed

MAGE – Multiplex automated genomic engineering

ORF – Open reading frame

Ori – Origin of replication

PCR – Polymerase chain reaction

PAGE - Polyacrylamide gel electrophoresis

PEG – Polyethylene glycol

POl – Part oligo long

POs – Part oligo short

POA – Part oligo annealed

PS – Phosphorothioate

RBS- - Ribosome binding site

mRNA – Messenger ribonucleic acid

TP – Truncated part

UTR – Untranslated region

General

°C – Degrees Celsius

µg – Microgram

µl – Microlitre

µM - Micromolar

ADH – Alcohol dehydrogenase

ATP – Adenosine triphosphate

BSA – Bovine serum albumin

CFE – Cell free extract

CoA – Coenzyme A

Da – Dalton

ddH_2O – Double distilled H_2O

DTT - dithiothreitol

E. coli – Escherichia coli

HRP – Horseradish peroxidase

g – Gram

IPTG - Isopropyl β-D-1-thiogalactopyranoside

Kb –Kilobase pair

L – Litre

LB – Luria-Bertani

mg – Milligram

mM – Millimolar

mRNA – Messenger ribonucleic acid

ng – Nanogram

nm – Nanometre

OD_{600} - Optical density at 600 nm

pg – Picogram

Reverse osmosis - RO

rGS – Reverse glyoxylate shunt

Room temperature – RT

RPM – Revolutions per minute

S. cerevisiae – *Saccharomyces cerevisiae*

SOC – Super optimal culture

SD – Standard deviation

SDS-PAGE - Sodium dodecyl sulphate polyacrylamide gel electrophoresis

U – Enzyme unit

WT – Wild type

Contents

1. Introduction

1.1 Production of chemicals through bioprocesses

The development of sustainable bioprocesses for the conversion of renewable feedstocks into fuels, chemicals and pharmaceuticals offers an attractive solution to the current reliance on petrochemical based routes[1, 2]. The current linear utilisation of carbon relies on taking petrochemical feedstocks from the earth for use, after which they end up as carbon dioxide in the atmosphere. In contrast a bioprocess can be defined as one which harnesses the power and diversity of nature via a micro-organism to convert a biological feedstock into an industrially relevant product. In this process carbon dioxide and water are converted to sugar by plants which can then be converted to the chemical of interest by microbial fermentation. These chemicals can then be used in the desired application and if they cannot be re-used, are converted back to CO_2 completing a closed loop cycle. Concerns over climate change and global fuel prices are driving the move towards a bio-based economy.

Despite the relatively recent initiative to move away from petrochemical feedstocks the industrial fermentation of microbes for chemical production predates the utilisation of petroleum based feedstock. Lactic acid fermentation was the first industrial fermentation processes for the production of a pure chemical. Initially described as a component of sour milk by Carl Wilhelm Scheele in 1780 it was not until 1858 when Louis Pasteur discovered that it was not a component of the milk but rather a metabolite generated through the fermentation of certain microorganisms, which he called lactic yeast[3]. By 1883 the process had been industrialised by Charles Avery who opened a lactate fermentation plant in Littleton, USA. Around the same time the German chemist Carl Wehmer was developing an industrial process for the production of citric acid through fermentation. Despite this process not progressing beyond pilot scale due to a combination of technical and sterility issues it laid the

ground work for the first commercial citric acid plant using the microorganism *Aspergillus niger* by Pfizer Inc. in 1923[4].

A further fermentation process with a long history is the use of *Clostridium acetobutylicum* for the production of acetone, butanol and ethanol (referred to as ABE fermentation)[5]. This process first rose to prominence during the 1st world war when acetone was a much sought after chemical due to its use as a solvent for the manufacturing of the smokeless explosive cordite. This resulted in Chaim Weizmann patenting his process utilising *Clostridium acetobutylicum* for industrialisation of the process. However, during the 1950's the process was superseded by cheaper petrochemical based production processes. The recent drive for renewable chemical production with a reduced carbon footprint has resulted in a renewed interest in the process from a number of companies[6, 7].

The following decades saw a focus on the development of a number of processes utilising microbes to produce natural products of pharmaceutical interest including, antibiotics, cholesterol lowering agents, anti-cancer drugs and immunosuppressants[8]. The improvement of these processes relied primarily on identification of a natural producer followed by strain mutagenesis and the screening of the resulting library for improved producers. This approach was particularly successful for antibiotics where a number of powerful screens were available with penicillin production using *Penicillin chrysogenum* increased by over 1,000-fold[9]. Today the development of strains for production of recombinant biopharmaceuticals proteins is a market with sales exceeding $100 billion[10].

1.2 Metabolic engineering

The advent of recombinant DNA technology during the 1970s and 1980s provided the tools for researchers to engineer microbial strains specifically for industrial bioprocesses[11, 12, 13]. This allowed researchers to use recombinant DNA techniques to either increase the production of native metabolites or to re-route metabolism for the production of new industrial compounds not natively produced by the chosen

host organism. This initially focussed on the production of proteins through the introduction of a single gene. Early success such as the engineering of *Escherichia coli* to produce human insulin[14], an idea which became feasible following the engineering of *E. coli* to produce human growth hormone[15], highlighted the ability to produce a chemically complex molecule through the overexpression of a single gene. This resulted in the conception of the first "molecular biology based" biotechnology company, Genentech, who licensed the technology to produce human insulin from a gene cloned in *E. coli* to the insulin producer Eli Lilly.

Since then the engineering of microbes through the introduction and deletion of genes to enable and maximise the production of a plethora of fuels, chemicals and pharmaceuticals from plant based renewable feedstocks such as starch, cellulose and lignocellulose as opposed to petrochemical feedstocks has been a primary goal of the fields of metabolic engineering and biotechnology.

Product	Host	Status	Companies	Reference
1,3 Propanediol	*E. coli*	Commercialised	DuPont, Tate & Lyle	www.duponttateandlyle.com
1,3 Butanediol	*E. coli*	Demonstration	Genomatica and Versalis	www.genomatica.com
1,4 Butanediol	*E. coli*	Commercialised	Genomatica and Dupont	www.genomatica.com
Isoprene	*S. cerevisiae*	Preparing commercialisation	Amyris, Braskem, Michelin	www.amyris.com
Isobutene	*E. coli*	Demonstration	Global Bioenergies	www.global-bioenergies.com
Squalene	*S. cerevisiae*	Commercialised	Amyris	www.amyris.com
PHA	*E. coli*	Commercialised	Metabolix	www.metabolix.com
Farnesene	*S. cerevisiae*	Commercialised	Amyris	www.amyris.com
Isobutanol	*S. cerevisiae*	Commercialised	Gevo	www.gevo.com

Table 1: Production of platform chemicals through the utilisation of engineered microbes in industrial bioprocesses (PHA – polyhydroxyalkanoate).

Despite there being many published reports of microbial strains being engineered for the heterologous production of fuels[16], chemicals[17] and pharmaceuticals[18, 19] relatively few have been progressed to an industrial scale (Table 1), whilst those that have, have proved to be extremely cost, time and labour intensive. These difficulties are directly linked to the efficiency and predictability with which microorganisms can be adapted for use in industrial bio-processes, a key issue this work aims to address.

1.3 Prominent examples of engineering microbes for industrial bioprocesses

1.3.1 Isobutanol

Gevo have commercialised a bioprocess utilising a genetically engineered yeast strain to produce isobutanol from renewable feedstocks. The branched chain higher alcohol isobutanol is commercially attractive both as a next generation biofuel due to its high octane value and also as a building block for commodity chemical production[20].

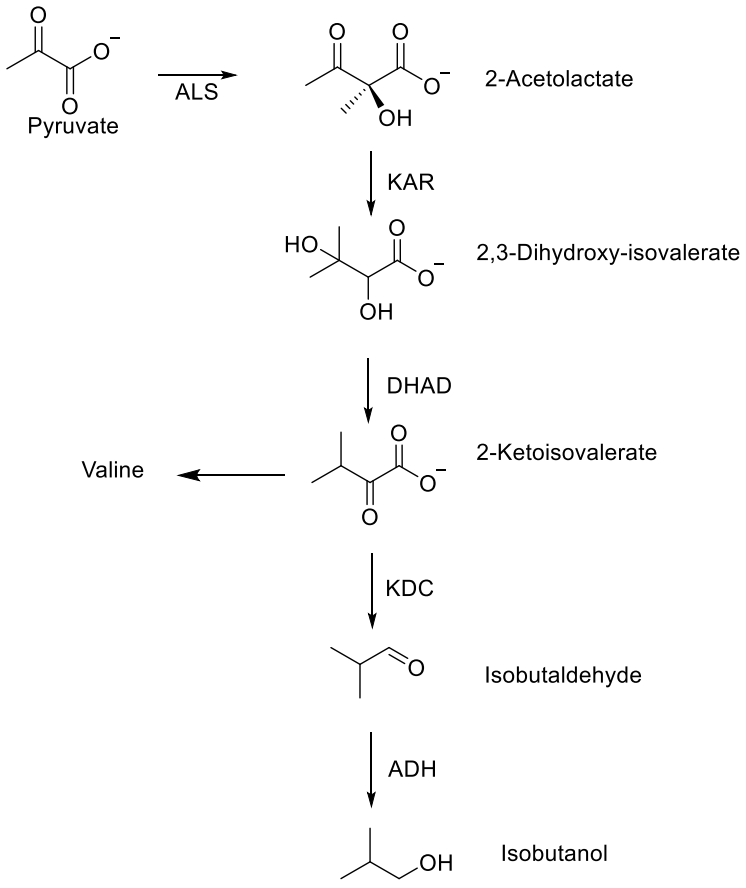


Figure 1: The metabolic pathway utilised for the production of isobutanol from pyruvate: ADH, alcohol dehydrogenase; ALS, acetolactate synthase; DHAD, dihydroxy-acid dehydratase; KAR, ketol-acid reductoisomerase; KDC, keto-acid decarboxylase.

Engineering of *S. cerevisiae* to produce isobutanol at high yields is achieved through diversion of carbon from the branched chain amino acid biosynthesis intermediate ketoisovalerate (KIV). The engineering of the strain to express a 2-keto acid decarboxylase[21] and alcohol dehydrogenase activity[22] results in isobutanol production from KIV (Figure 1). Further engineering of the strain to overproduce KIV and limit the formation of ethanol and acetic acid increased strain productivity under anaerobic conditions. Gevo has coupled this engineered strain to a process involving continuous stripping of isobutanol from the culture broth through flash evaporation resulting in the commercialisation of a bio-based process approaching 90% of theoretical yields[23].

1.3.2 1,3-Propanediol

Production of chemical building block 1,3-propanediol through engineering of *E. coli* by DuPont, in collaboration with Genencor and Tate and Lyle was industrialised with reported titers of up to 135 g l^{-1} [24]. The production of 1,3-propanediol was achieved via the glycolytic triose dihydroxyacetone phosphate (DHAP) which is reduced and dephosphorylated to glycerol. As *E. coli* is intrinsically a poor glycerol producer the team first introduced a glycerol production pathway from *S. cerevisiae*. This was followed by two genes from *Klebsiella pneumoniae* to dehydrate and reduce glycerol to 1,3-propanediol (Figure 2). Further optimisation of the strain was achieved through the elimination of *E. coli* pathways competing for consumption of glycerol, optimisation of glucose import and the use of a previously uncharacterised endogenous *E. coli* enzyme for the final reduction of 3-hydroxypropanal to 1,3-propanediol. This process continues to be used by DuPont for the bio-based production of 1,3-propanediol.

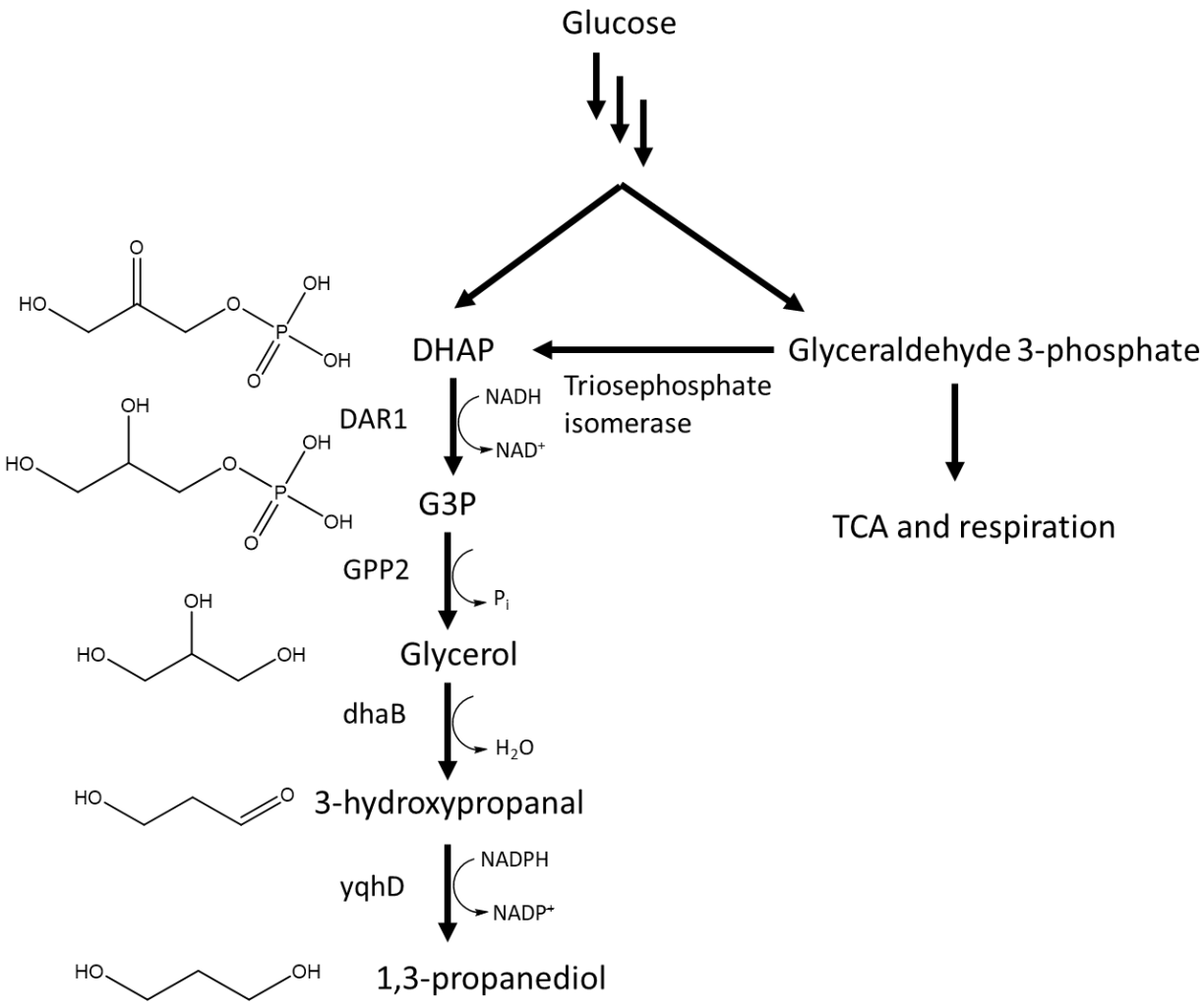


Figure 2: Metabolic engineering of E. coli for the production of 1,3-propanediol. Engineering focussed on the conversion of dihydroxyacetone phosphate (DHAP) to 1,3-propanediol. Glycerol production via glyceraldehyde 3-phosphate (G3P) was achieved through the introduction of glycerol 3-phosphate dehydrogenase (DAR1) and glycerol 3-phosphate phosphatase (GPP2) genes from Saccharomyces cerevisiae. Introduction of glycerol dehydratase (dhaB) from K. pneumoniae enabled the conversion of glycerol to 3-hydroxyproponal with the final reduction to 1,3-propanediol catalysed by an endogenous E. coli oxioreductase (yqhD).

1.3.3 Artemisinic acid

Jay Keasling's team at the University of California, Berkley were able to introduce in *Saccharomyces cerevisiae* a metabolic pathway responsible for the production of a precursor for the anti-malaria drug artemisinin, artemisinic acid[25]. Malaria affects over 200 million people annually[26] and disease control is hampered by the

occurrence of drug resistant strains of the malaria parasite *Plasmodium falciparum*. Whilst synthetic antimalarial drugs and vaccines are in development they are awaiting clinical trials[27]. Artemisinin is a sesquiterpene lactone endoperoxide extracted from the plant *Artemisia annua* which is a highly effective treatment however it can be in short supply[28] due to relatively low extraction yields and subjection to year to year variations in crop performance. Whilst the full chemical synthesis of artemisinin was achieved in 1983 it has proven too complex and costly for commercialisation[29]. Therefore, the semi-synthesis from microbially sourced artemisinic acid was identified as a potentially industrially attractive option (Figure 3).

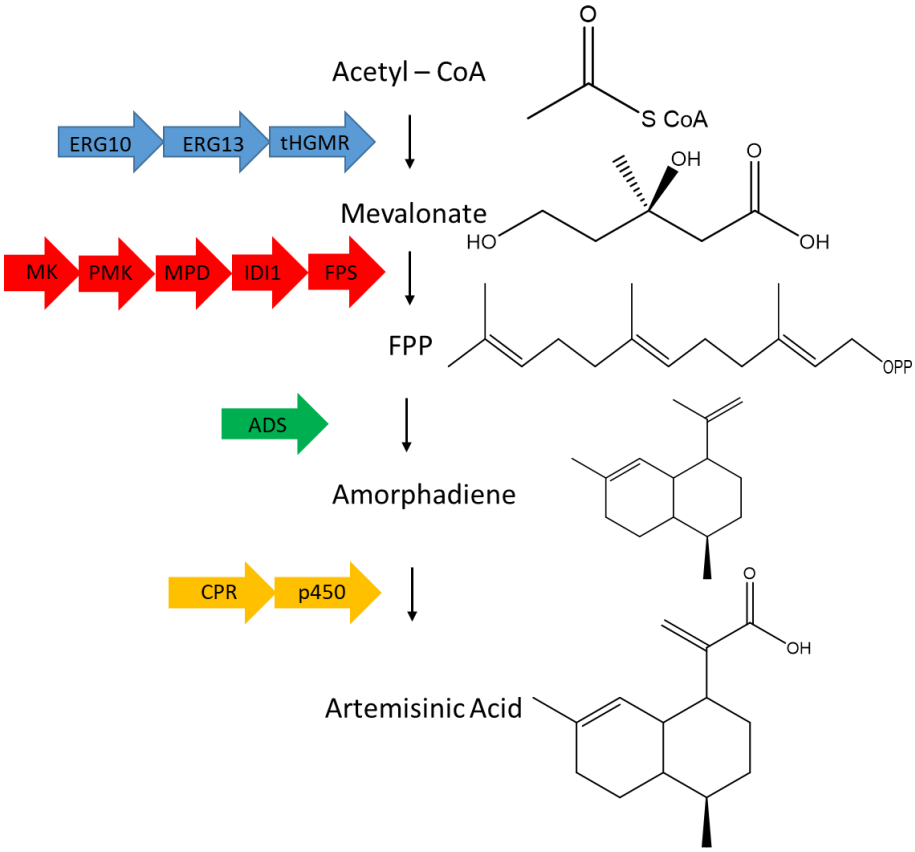


Figure 3: The biosynthetic production of artemisinic acid from renewable feedstocks overview. Artemisinic acid is produced from the mevalonate pathway intermediate farnesyl-pyrophosphate (FPP). In blue are the genes of the upper mevalonate pathway which converts acetyl-CoA to mevalonate. In red are the genes of the lower mevalonate pathway for the production of FPP. These genes were overexpressed in the production strain. In yellow is the amorphadiene synthase gene from A. annua which was heterologously expressed in the production strain. In yellow are the P450 enzyme (CYP71AV1) and its redox partner (CPR) from A. annua which catalyse the three step oxidation of amorphadiene to artmesinic acid.

During the development of this process both *E. coli* and *S. cerevisiae* were explored as well understood and genetically tractable potential hosts for artemisinic acid production. Strain engineering was initially focused on the accumulation of the precursor amorphadiene. To achieve this the group focussed on the native isoprenoid production pathways in both *E. coli* and *S. cerevisiae*, the 1-deoxy-D-xylulose 5-phosphate (DXP) and mevalonate pathways respectively. Initial efforts to overexpress rate limiting enzymes within the DXP pathway in *E. coli* only had limited success with terpenes produced in the low mg per L range[30]. Production in *E. coli* was improved through the heterologous expression of the *S. cerevisiae* mevalonate pathway and the introduction of a synthetic codon optimised amorphadiene synthase (ADS) from *A. annua*[31] which is responsible for the conversion of farnesyl pyrophosphate from the mevalonate pathway to amorphadiene, however the levels following optimisation of the fermentation process were still 50-fold lower than the estimated concentration required for an economically viable process[32].

Analysis of the limitations of this strain highlighted that when the upper part of the mevalonate pathway was overexpressed growth was inhibited. It was hypothesised that this inhibition in growth was the result of the accumulation of a pathway intermediate. A combination of operon optimisation through a combinatorial approach using tuneable intergenic regions[33] and targeted transcriptome and metabolite analysis showed that the accumulation of hydroxymethylglutaryl (HMG)-CoA resulted in the observed growth inhibition. The optimisation of this node to alleviate pathway bottlenecking coupled to process optimisation resulted in strain capable of producing 25 g per L amorphadiene[34].

The conversion of amorphadiene to artemisinic acid was found to be catalysed in the native host by a cytochrome P450 enzyme through the analysis of cell free extracts[35]. *E. coli* has typically been a poor expression host for eukaryotic P450 enzymes[32] and therefore the expression of the P450 and its cognate reductase was first tested in an amorphadiene producing *S. cerevisiae* strain resulting in 100 mg per L artemisinic acid[25]. This resulted in a change in host organism, despite the lower amorphadiene

yields produced in *S. cerevisiae* - 150 mg per L compared to *E. coli* 25 g per L - the inability of *E. coli* to efficiently express the P450 enzymes resulted in *S. cerevisiae* being identified as the optimal host. Further pathway and fermentation optimisation using the engineered *S. cerevisiae* strain resulted in a strain producing the 25 g per L artemisinic acid targeted at the beginning of the project[35]. This project is one of the first examples of engineering a strain for the use in a bioprocess using tools such as combinatorial DNA assembly and codon optimised gene synthesis for pathway construction coupled to transcriptomics and metabolomics analysis for heterologous pathway optimisation.

Despite the successes described above it took 10 years to take the proof of concept *Saccharomyces cerevisiae* artemisinin production strain through to one that was scaled to industrially relevant levels at an estimated cost of 50 M USD. It is believed to have taken 15 years at a cost of 130M USD for the development of the 1,3-propanediol process[8]. This highlights the difficulty and extensive time required in the optimisation of a proof of principle strain into an industrially applicable strain.

Using traditional approaches each stage of engineering is carried out sequentially adding to time and cost. The implementation of new technologies has the potential to decease this time and accelerate the implementation of the industrial bioprocess (Figure 4). This PhD focusses on addressing limitations in both the construction and characterisation of the microbes engineered for industrial applications.

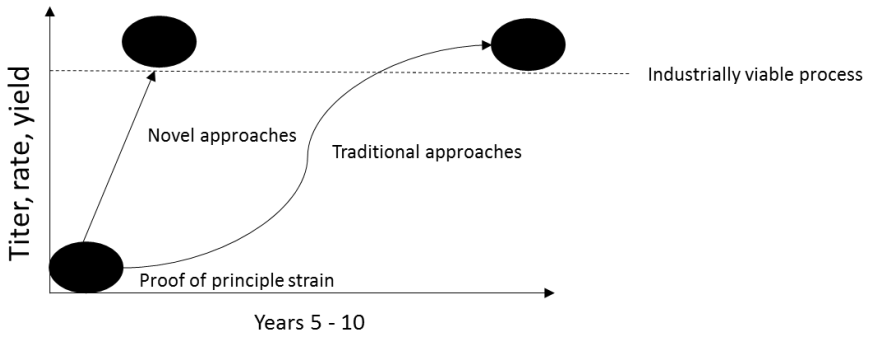


Figure 4: Development of an industrial strain from proof of principle can be a long and costly processes. As a result, the implementation of new technologies to accelerate this process is required (adapted from[8]).

1.4 Synthetic Biology

The field of synthetic biology has grown rapidly from 2004 when the first synthetic biology conference, Synthetic Biology 1.0, was held at Massachusetts Institute of Technology (MIT). A year later, Drew Endy published a review entitled *Foundations for engineering biology*[36] in which the application of three engineering principles (standardisation, decoupling and abstraction) to biotechnology were outlined: the use of DNA in the form of standardised biological parts which are well defined, described and characterised; reducing the complexity of engineering a system through splitting it into a number of smaller tasks which are eventually combined; and reduction of complexity through the use of parts which interact with each other in a known fashion, thus making them easier to model and combine. This review was one of the first to suggest the application of engineering principles to genetic engineering and has shaped the evolution of a rapidly growing field since[37,38,39,40]. Despite the field of synthetic biology spanning a broad range of topics including ones which are not directly related to the engineering of organisms for biotechnology, key synergies exist. Critically, this includes the predictable engineering of biological hosts.

The field has been driven by a number of technological breakthroughs particularly relevant to strain engineering, including: (i) Advance in the development of robust genetic parts such as those that provide tight control over gene transcription through synthetic promoter design[41] and the ability to rapidly combine multiple genetic fragments[42]. (ii) Reduction in the cost of DNA sequencing[43] – allowing for whole microbial genomes to be sequenced relatively cheaply - and the continued reduction in DNA synthesis cost[44,45] have not only allowed researchers to develop synthetic pathways which go beyond merely redirecting carbon flux through the host's natural pathways but also to select and introduce enzymatic functions through the identification of the appropriate genes from the abundance of sequenced organisms. (iii) The development and implementation of transcriptomics[46,47] , metabolomics[48, 49] and proteomics[50] techniques, known collectively as "omics", which provide a level of understanding of limitations in cell engineering not previously achievable. In

addition to these techniques efforts have been focused on the construction of genome scale metabolic models, commonly generated from omics data, which describe the inter-conversion of metabolites through enzyme catalysed chemical transformations[51]. These models have been used to characterise biological production systems and identify non-intuitive engineering strategies to optimise strain productivity[52] .

The tools developed in the field of synthetic biology can be combined to accelerate not only the engineering of microbes for industrial bioprocesses but also enhance the level of understanding of the engineered organism's limitations and guide subsequent rounds of strain improvement.

1.5 The design build test learn cycle - accelerating engineering of cellular metabolism

With the advent of synthetic biology, the engineering of microbes for use in a bioprocess can be considered less of a one off ad hoc process and instead can be thought of as a repeating cycle split into four primary components, design (D) of the biological system, building (B) of the necessary genetic constructs, test (T) of the generated cell system and learn (L) to generate information on the successes or failures of the cycle to feed into the next turn of the cycle. Through the utilisation of enabling tools developed in the field of synthetic biology each stage of this process - known as the design-test-build-learn cycle - can be optimised to accelerate microbe engineering (Figure 5).

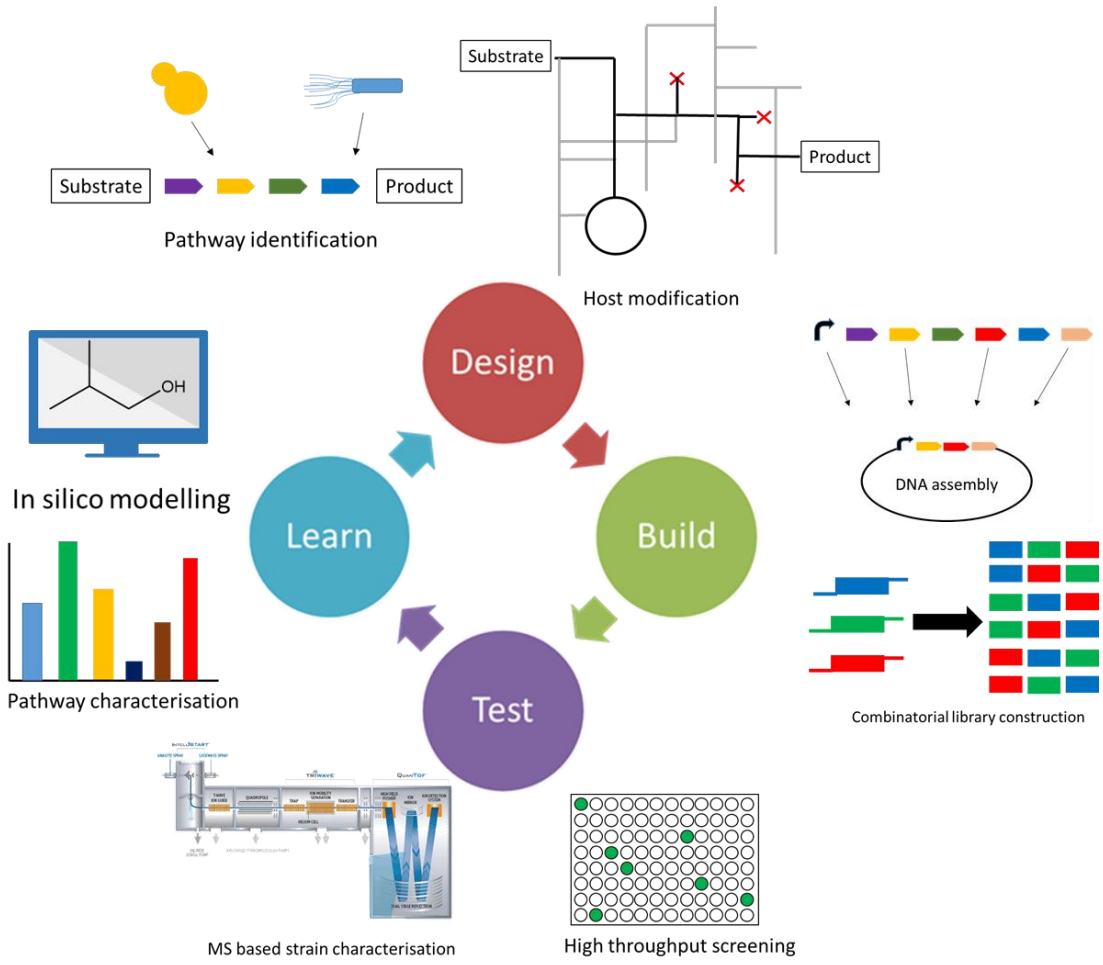


Figure 5: The design, build, test, learn cycle for the engineering of microbes for industrial bioprocesses.

1.5.1 Design

The engineering of a microbe standardly begins with the design stage during which a chemical of industrial interest is targeted. This is followed by the identification either of metabolic pathways present in nature to produce this target or the identification of genes from throughout nature that can be stitched together into a pathway to produce the chemical from an intracellular metabolite.

The simplest strategy typically relies on a classical strain engineering approach in which a natural producer of the target chemical is identified and exploited for further improvement through mutagenesis coupled to high throughput screening. This approach has proven to be been extremely efficient and has resulted in the bio-based

production of multiple biopharmaceuticals as mentioned previously. Implementation of a more rational engineering approach from a native producer for the overproduction of the compound of interest relies on several parameters including, ease of handling and culturing of the organism, in-depth knowledge of host metabolism and regulation, efficient procedures in genetic material transfer, availability of molecular biology tools and of genomic sequence data. Each of these parameters would be considered before engineering the native producer to enhance production of the desired compound.

In absence of a characterised natural producer or in the case of a genetically intractable host, the genes involved in the generation of the product of interest require to be transferred into a heterologous host. This approach usually utilises a well characterised host for which standard molecular biology tools are available such as *E. coli*[53], *S. cerevisiae*[54, 55] , *Bacillus subtilis*[56], *Corynebacterium glutamicum*[57] or *Pseudomonas putida*[58, 59] . Each of those micro-organisms has specific advantages which would guide the choice of host for specific applications. *P. putida* for example is extremely solvent tolerant[60] making it an attractive host for the production of certain chemicals whilst *B. subtilis* has very efficient protein secretion machinery[61] making it a promising choice for the industrial production of enzymes due to the savings associated with reduced downstream processing.

This stage also involves the engineering of the host strain to maximise target production, which can involve: (1) modulation of substrate uptake (2) down regulation of cellular pathways competing with the desired heterologous pathway; (3) enhancing cellular tolerance to the chemical of interest[62, 63] , through for example transporter and membrane engineering; (4) balancing of reducing power to drive flux towards the compound of interest[64].

1.5.2 Build

Following the choice of host and the definition of the required reactions to produce the compound of interest the required biosynthetic pathway is standardly constructed

in vitro. This involves the synthesis (and potentially host specific codon optimisation) or amplification of the genes required to encode the enzymes involved in the metabolic pathway and the combination with associated regulatory regions such as promoters, terminators and ribosome binding sites which can be used to modulate the amount of enzyme production at each stage of the pathway. The goal of this modulation is to balance pathway enzyme production levels, ensuring flux through the pathway avoiding any build up potentially toxic pathway intermediates[65,66] without putting an unnecessary burden on cellular metabolism[67]. The ability to rapidly construct these biosynthetic pathways in a combinatorial manner is an area which has seen significant growth recently and is reviewed later in this chapter (Section 1.6.1).

1.5.3 Test

The test stage of engineering an industrial microbe involves the utilisation of analytical techniques which allow for determination of the success of the design and build stages. This can include, confirmation of the success of the build stage (including successful biosynthetic pathway construction and implementation of gene knockouts), growth and physiological characterisation of the engineered strains and analysis of product formation.

The analysis of product formation often relies on techniques such as gas or liquid chromatography coupled to either UV absorbance or mass spectrometry (MS) based detection. Whilst these approaches combine reliable target identifications with accurate and precise target compound quantification, they lack the required throughput for the screening of large strain variant libraries. Consequently, a higher throughput screen or selection could be implemented as a preliminary step. Examples of high throughput screening methods include solid-phase assays in which product formation can be coupled to growth or a colorometric response and liquid-phase assays which are based on spectroscopic measurements such as colorimetric, UV absorbance or fluorescence in microtiter plates[68, 69]. As specific assays for target molecules are often lacking there has been a recent focus on the development of

biosensors[70, 71, 72]. A biosensor standardly functions through protein or transcript based sensing the of target molecule coupled to the expression of a reporter such as green fluorescent protein (GFP). The combination of this approach with fluorescent-activated cell sorting (FACS) can result in an extremely high-throughput screening platform.

Recent advances in the analysis of transcript, metabolite or protein abundance using mass spectrometry based omics' approaches for the optimisation of industrial microbe are described in detail later in this chapter (Section 1.7).

1.5.4 Learn

The learn stage is arguably the least systematic in metabolic engineering strategies and often relies on ad hoc observations and the intuition of individual researchers who define the next stages of engineering. With the increasing amount of data generated during the test cycle, techniques such as principal component analysis (PCA) have been used to analyse small proteomics dataset from engineered organisms and deduce patterns or trends to guide subsequent rounds of engineering[73]. In the future sophisticated data analysis tools such as machine learning[74] will likely be necessary to process the large diverse datasets collected during the learn cycle and deduce strain limitations to guide subsequent rounds of engineering

1.6 Accelerating the build stage

As metabolic engineers looked to accelerate the engineering of microbes for bioprocesses, it quickly became apparent that the inability to assemble multiple DNA fragments was one bottleneck which was holding the field back. Despite advances in the ability to construct DNA vectors, from the 1970s when the construction of a bacterial plasmid would be worthy of a high profile publication[75], standard techniques involving the digestion of DNA with a restriction enzyme followed by

subsequent ligation still only allow for the combination of two or three DNA fragments. When multiple pathway genes and their respective regulatory regions – promoters, ribosome binding site (RBS), terminators - required to be combined this involved laborious multi-stage cloning experiments which proved to be time consuming, unpredictable and expensive. This realisation coupled to decrease in the cost of synthetic DNA resulted in the development of techniques for the rapid and combinatorial assembly of genetic constructs.

1.6.1 DNA assembly methodologies

1.6.1.1 Traditional digestion and ligation methodology

Molecular cloning utilising DNA ligase and type II restriction enzymes has been the foundation of metabolic engineering for the past 40 years[76]. This "cut and paste" strategy relies on cleaving of DNA using restriction endonucleases to generate compatible sticky-ends. The DNA fragments are then co-incubated in a buffer which allows the compatible ends to anneal and in the presence of a DNA ligase which covalently seals nicks in the constructed vector. This approach is limited to assembly of a maximum three fragments of DNA, with each fragment requiring to be void of recognition sites for the restriction enzymes being used. This strategy is also difficult to standardise and automate. The limitations of this approach resulted in the development of several DNA assembly technologies, each with the goal of allowing the assembly of multiple DNA fragments in one reaction thus accelerating the building of genetic constructs.

1.6.1.2 BioBricks

BioBricks[77] was the first standardised assembly technique to gain traction amongst the scientific community. BioBricks are standardised fragments of DNA (known in DNA assembly nomenclature as parts) which can be joined together in a Lego-like

fashion. The BioBrick approach allows for the combination of up to three BioBricks with the resulting construct becoming a new BioBrick, thus allowing subsequent rounds of assembly. Both the assembly method and parts are standardised in the sense that the enzymes used are constant and therefore the 5' and 3' terminal sequences are constant. Large databases of these parts have been built up including the BioBricks foundation and MIT's Registry of Standardised Biological Parts. BioBricks require to be flanked by two specific restriction sites which cannot be present within the part itself (Figure 6). Through digestion and ligation two BioBricks can be combined - separated by a 6 base pair (bp) scar – with the resulting construct flanked by the original recognition sites meaning that this construct becomes a BioBrick in its own right. The main limitation of this approach is the time consuming construction of complex vectors due to the stepwise nature of the assembly process. A further potential limitation is that the introduction of 6 bp scars could be problematic if they occur within certain regions, for example within coding sequences in the case of protein domain assembly or between the ribosome binding site and the start codon in the case regulatory region optimisation.

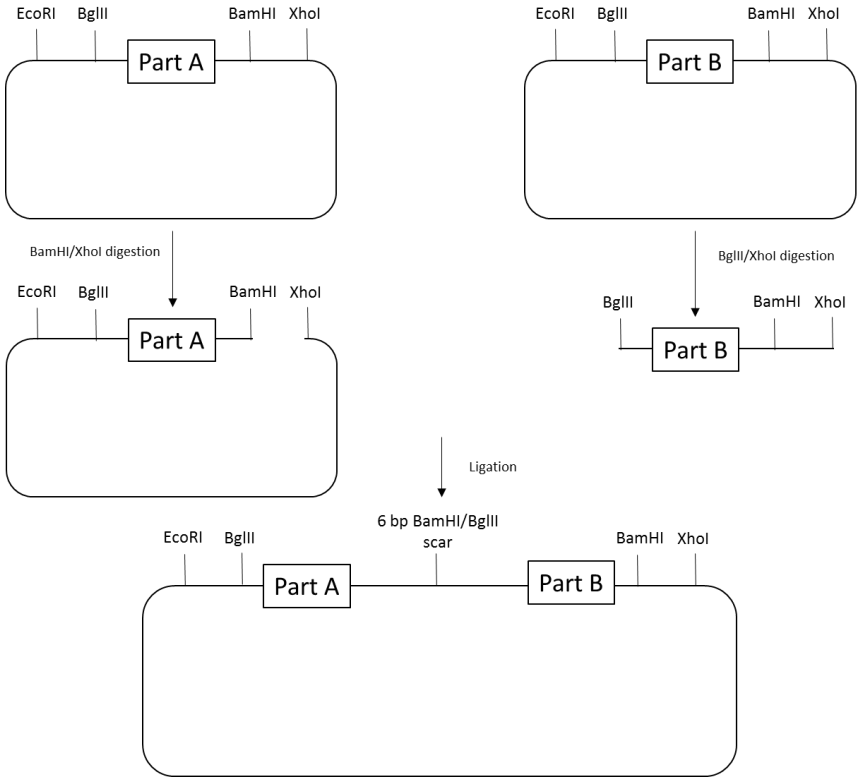


Figure 6: Assembly of two DNA fragments using the BioBricks technology. In this example, DNA assembly is achieved through digestion of Part A and B with BamHI/XhoI and BglII/XhoI respectively. This generates compatible overhangs allowing ligation of the two BioBricks. As the product of the BioBricks assembly is flanked by the original restriction sites further BioBricks can be added in a stepwise manner.

1.6.1.3 Gibson Assembly

Building on the success of BioBricks as a standardised system for DNA assembly, Gibson *et al*[78] described a method for the enzymatic assembly of over ten fragments of DNA. This technique falls into the category of long overlap DNA assembly methods. It differs from techniques which rely on digestion and ligation because the complementary overhangs between fragments are usually 15 nucleotides or more in comparison to the standard 4 nucleotide generated through restriction endonuclease digestion.

Gibson DNA assembly relies on a one pot reaction in which linear DNA parts with approximately 25 bp homology, a 5' exonuclease, a DNA polymerase and a DNA ligase are combined (Figure 7). Initially the exonuclease degrades the linear parts in

a 5’ to 3’ direction resulting in the generation of single stranded overhangs enabling complementary regions to anneal to each other. Next a high fidelity DNA polymerase fills in any gaps before the nicks in the final construct are sealed by a DNA ligase. The advantages of this technique in comparison to BioBricks are the increased number of parts which can be combined, the fact it is largely sequence independent in the sense that no specific restriction sites require to be omitted from part sequences and that it results in a scarless assembly. It is also extremely simple and amenable to automation as all three enzymes and the DNA to be assembled are mixed in one reaction, which is incubated at 50 °C for one hour after which the resultant constructs are directly transformed. All enzymes act simultaneously during the incubation, however the T5 exonuclease is the only temperature sensitive enzyme in the reaction resulting in slow heat inactivation through the reaction. Consequently, only the polymerase and ligase remain active towards the end of the reaction ensuring complete repair of the DNA construct.

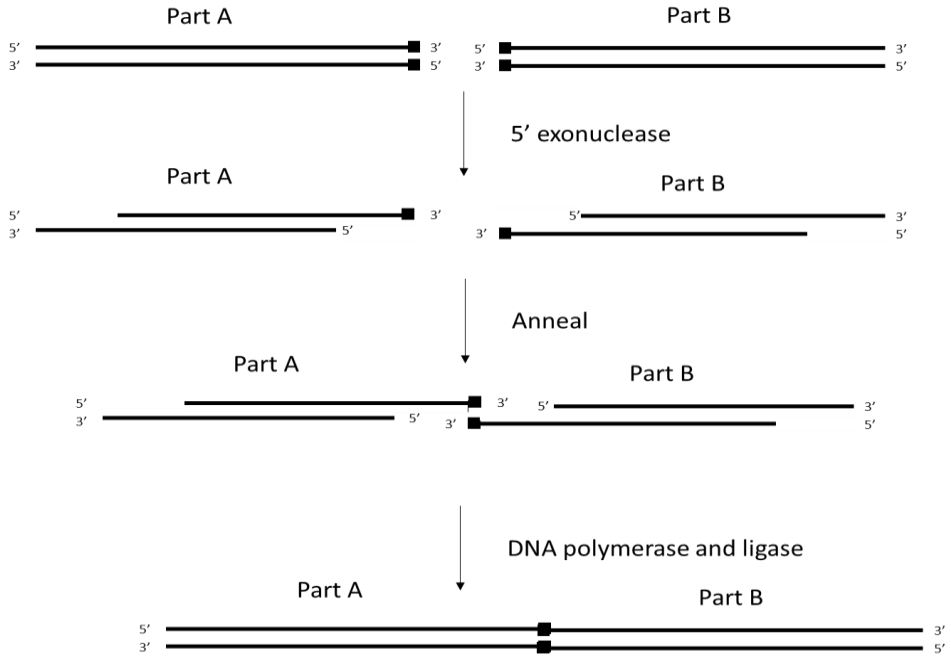


Figure 7: Assembly of two DNA fragments by Gibson assembly. Sequence overlaps depicted by black boxes. DNA assembly is achieved through the action of three enzymes. First a DNA exonuclease degrades the DNA fragments in a 5’ to 3’ direction generating single stranded overhangs which are complementary between the parts to be assembled. A DNA polymerase then fills in any gaps before a DNA ligase seals nicks in the final construct.

The part length, which is restricted to at least 250 bp and the incompatibility of the technique with repeated sequences represent the main limitations associated to the Gibson assembly technology. A minimum part length of 250 bp is required due to the potential degradation of the full length DNA fragment by the exonuclease prior to overhang annealing and polymerase-catalysed strand extension. Repeated sequences, for instance terminator elements located downstream of each gene in a pathway, are problematic as Gibson assembly relies on sequence homology which could result in aberrant constructs with missing parts. In order to circumvent this issue, sequential assemblies can be performed so that the part junctions including repeated sequences are not involved in the same reaction. However, this procedure adds time and complexity to the construction process. Torella *et al* [79] described a technique in which DNA parts were flanked by unique 40 bp nucleotide sequences thus limiting the potential for misassembly due to repeated sequences. However, this technique requires either an initial cloning or PCR step to attach these sequences to the parts.

1.6.1.4 Golden Gate assembly

A number of the issues encountered with Gibson assembly were addressed in the development of the Golden Gate assembly methodology and its variants [80, 81,] . This technique relies on type IIs endonucleases which are restriction enzymes that recognize asymmetric DNA sequences and cleave outside of their recognition sequence. Although several type IIs endonucleases are available BsaI is commonly used for Golden Gate assembly and will be used as the endonuclease for this description (Figure 8). Parts to be assembled using Golden Gate assembly are amplified, flanked by BsaI recognition sites (GGTCTCN^NNNN_). The PCR products are mixed with BsaI and ligase in a one pot reaction. An initial stage of digestion results in the part being liberated from the flanking BsaI sites resulting in 4 bp overhang sites allowing ligation between the parts to take place in a second stage. The initial digestion from the donor amplified fragment is reversible in that it is possible to ligate the BsaI containing sequences back onto the part, however once the parts are ligated to each other there are no longer any BsaI recognition sites present, this meaning that through cycles of digestion and ligation you can enrich towards the

formation of the desired product. In comparison to Gibson assembly this technique is less sequence independent as it requires the parts to be void of BsaI recognition sites. A second limitation is that with overhangs of only four base pairs in length and a desire to have at least two bp difference between each set of overhangs there is the potential that in complex assemblies it may not be possible to find specific overhangs for assembly.

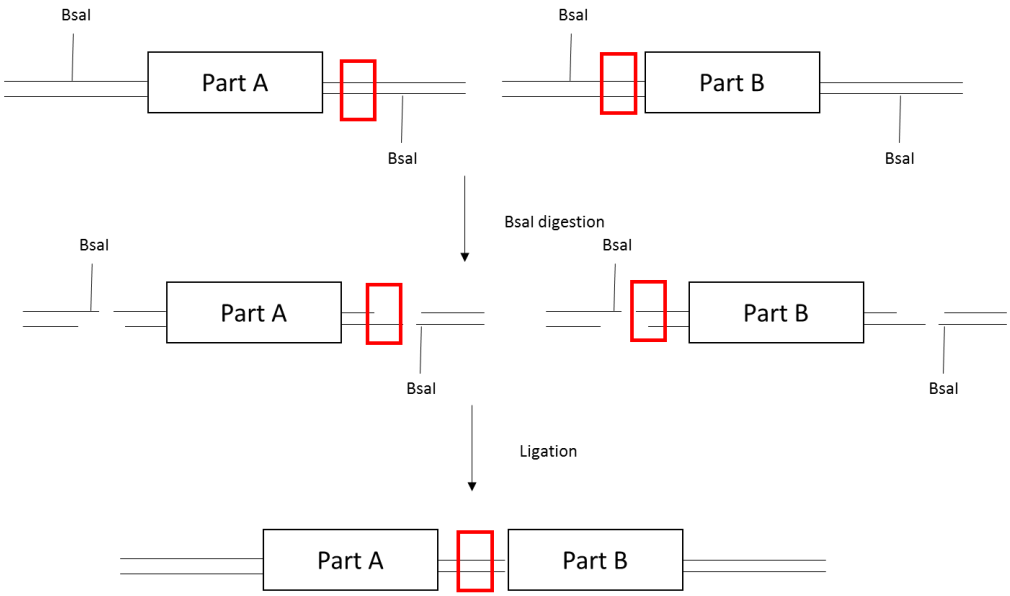


Figure 8: Assembly of two DNA fragments by Golden Gate assembly. Homology between parts highlight in boxes. DNA assembly is achieved through cycles of digestion and ligation. First, a type IIs restriction enzyme is used to digest the target DNA generating compatible four nucleotide overhangs which allow target parts to be ligated. To enrich the product of interest the digestion and ligation reactions are repeated multiple times in a cyclic manner.

1.6.1.5 Alternative strategies

Despite BioBricks, Golden Gate, Gibson assembly and their variants being the most widely implemented DNA assembly strategies several other approaches are available, each with associated advantages and disadvantages. *In vivo* homologous recombination based approaches to assemble DNA have been implemented in a number of organisms but most widely in *S. cerevisiae*. The strategy is based on transformation of *S. cerevisiae* with DNA fragments flanked by regions of homology

defining the order the parts assemble. The efficient homologous recombination machinery of *S. cerevisiae* is then responsible for the assembly of the genetic construct. This approach has been shown to be extremely efficient and used for the construction of eight gene biosynthetic pathways[82] and even full synthetic genomes[83]. The main limitations associated with this strategy are the difficulty in *S. cerevisiae* competent cell preparation, the relatively low competency that can be achieved in comparison to *E. coli* (around 10^5 compared to 10^9 cfu/µg DNA) and the slower growth rate of *S. cerevisiae* in comparison to *E. coli* requiring transformation plates to incubated for up to four days compared to just one for *E. coli*.

Site specific recombinases, which are enzymes catalysing the introduction, removal, swapping or inversion of DNA fragments through specific recognition sites have also been implemented for DNA assembly[84]. Unlike the previously described approaches which rely on homology between DNA fragments, this strategy is reliant on the recombinase enzyme and its associated recognition sites which mediate the DNA assembly. A major advantage of such an approach is that fragments can be removed or swapped from previously built constructs. For example, when building a biosynthetic pathway, a suboptimal gene could be swapped for a more suitable variant without having to completely rebuild the genetic construct. Much study in this area has been based on the PhiC31 recombinase[85], with the focus on increasing the number of orthogonal recognition sites available and thus the number of parts which can be assembled. One limitation of recombinase based strategies is the length of the recombinase recognition sites, which are usually approximately 50bp inverted repeats which flank each assembled fragment as scar sequence and may negatively impact part performance.

1.6.1.6 inABLE®

Ingenza Ltd. currently uses a combinatorial genetic assembly technology (known commercially as inABLE®) for the efficient and selective assembly of complex DNA expression vectors encoding, for example, the enzymes responsible for multiple biosynthetic transformations within a pathway to produce chemical targets.

The inABLE® technology can be split into three stages, part design and preparation, part linker fusion and the assembly reaction (Figure 9).

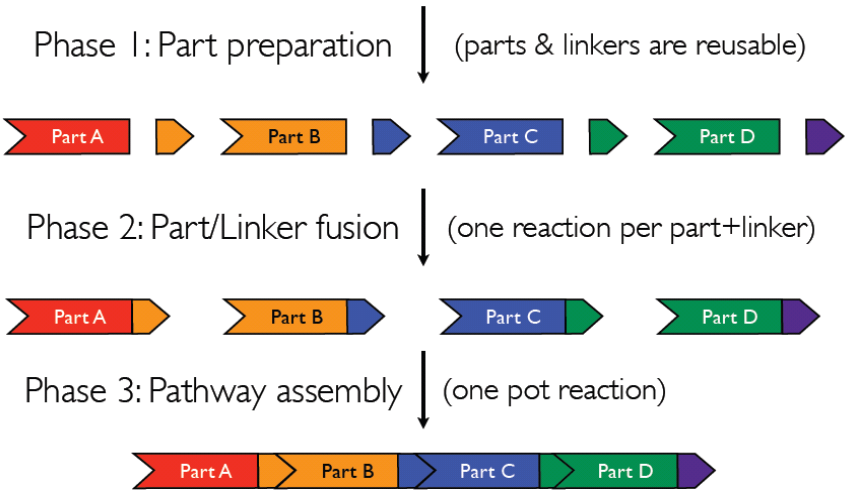


Figure 9: Overview of the assembly of DNA fragments using the inABLE technology.

The part is defined by the researcher and the only sequence constraints are that it is over 100bp in length and that it does not contain any EarI or SapI recognition sites. Once the part sequence has been defined computer aided design is implemented to define the most appropriate region to split the part into the part oligo, the linker oligo and the truncated part (Figure 10).

The splits are calculated to ensure that linker oligonucleotides have a melting temperature between 35 °C and 45 °C. A scar sequence of GCC is predefined as a high GC content is desirable during the ligation stage of the part linker fusion reaction. GCC also codes for alanine (or glycine in reverse) so if the introduction of the scar sequence within a coding region cannot be avoided a relatively innocuous amino acid will be introduced.

The truncated part is cloned into a compatible backbone flanked by recognition sites for the type IIs endonucleases EarI and SapI, allowing excision of truncated parts devoid of enzyme recognition sites. The part and linker oligonucleotides are split into part oligo short (POs), part oligo long (POl), linker oligo short (LOs) and linker oligo

long (LOl) and ordered as oligonucleotides. POl and POs are then annealed and 5' phosphorylated to create the "part oligo annealed" (POA). This is repeated for the linker oligonucleotides to produce the "linker oligo annealed" (LOA). The melting temperature of the overhang between the POA and LOA is prioritised for the split design. The designer software selects overhangs with a melting temperature over 70 °C which consequently leads to an average overhang length of 16 bp depending on the GC content. All truncated parts, POAs and LOAs can be recycled for any assemblies requiring those parts.

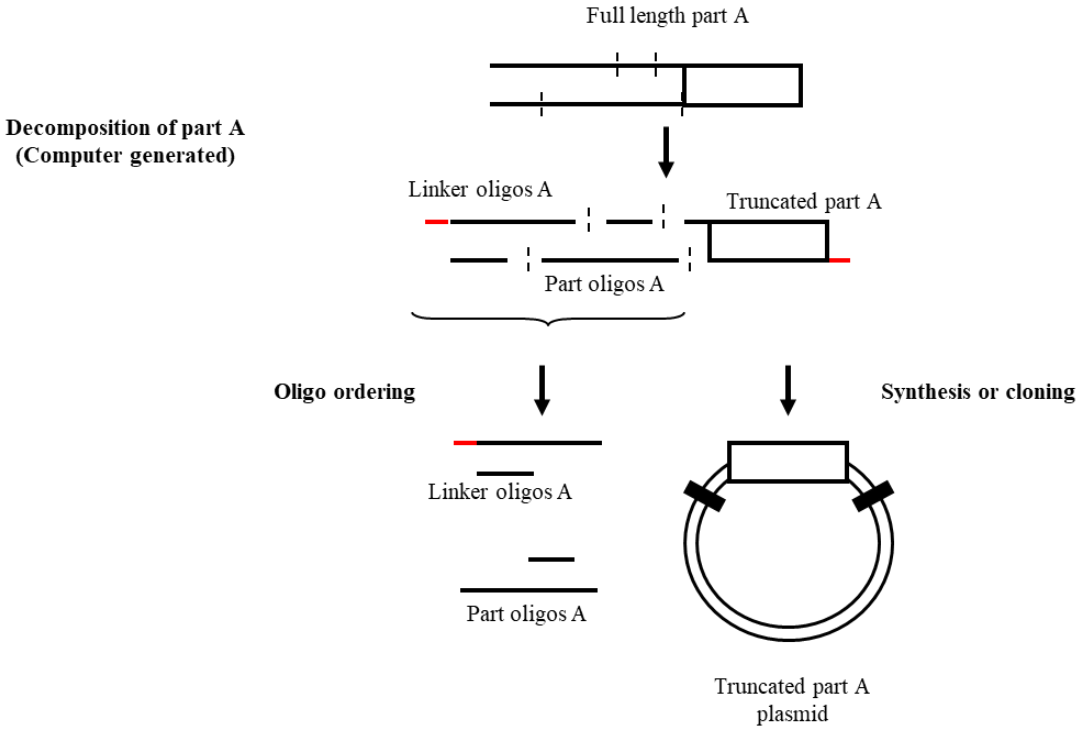


Figure 10: Phase1 – Primer and part design during which the full length DNA part is split into a truncated part and part and linker oligonucleotides.

A part linker fusion reaction is performed for each required part junction. These reactions are performed in parallel and therefore the time taken to perform the assembly is roughly independent from the number of parts necessary for the constructs. The part linker fusion reaction is performed through cycles of digestion and ligation. Initially an EarI digestion is performed resulting in the truncated part being cleaved

from the backbone. Secondly a ligation step is performed in which the corresponding POA and the LOA from the subsequent part are ligated to the 5' and 3' end of the truncated part respectively (Figure 11). In the same ligation step it is also possible for the truncated part to be ligated back to its original backbone. To enrich towards the desired product (truncated part with ligated 5' POA and 3' LOA) samples are incubated in a thermocycler which alternates between 37°C and 16°C corresponding to the optimum temperatures for digestion and ligation respectively. This enrichment through cycling is possible due to the nature of EarI (type IIS endonuclease) which cuts outside its own recognition site. As a result, the part linker fusion formed lacks EarI sites and therefore is removed from the digestion/ligation cycling. Only the residual truncated part inserted into its original EarI-harbouring backbone is able to enter the subsequent cycles of digestion. Part linker fusions are then purified from the backbone fragments via gel electrophoresis followed by DNA extraction.

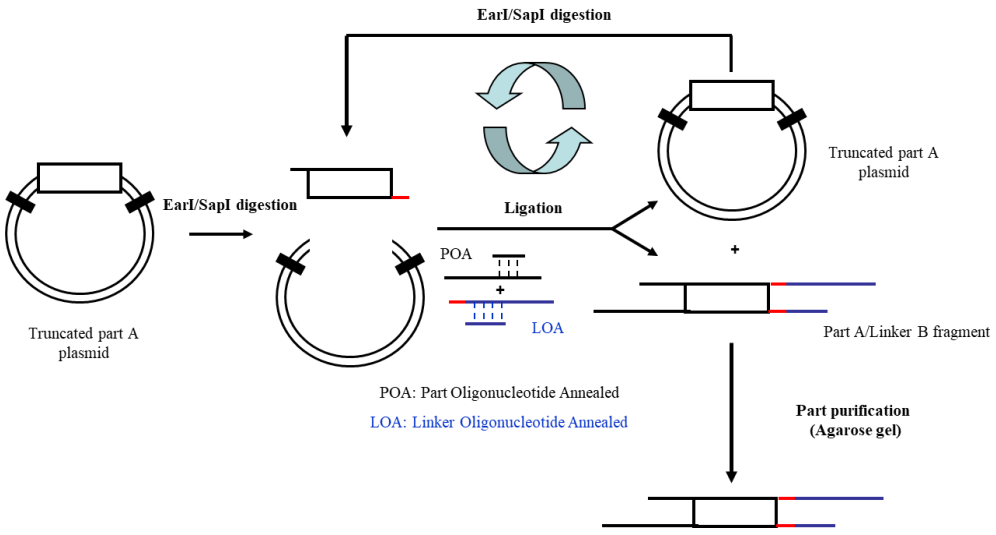


Figure 11: Phase 2 - The part/linker fusion reaction. During this reactions are prepared through cycles of EarI digestion/ligation.

The assembly reaction is a simple one pot reaction of constant time independent of the number of parts being assembled. Equimolar amounts of each part linker fusion are mixed and incubated at room temperature for 30 minutes. Parts anneal through the complementary 16 bp overhangs generated between the LOA and POA (Figure

12). The assembly reaction is then used to transform *Escherichia coli* and transformants screened for correctly assembled vectors.

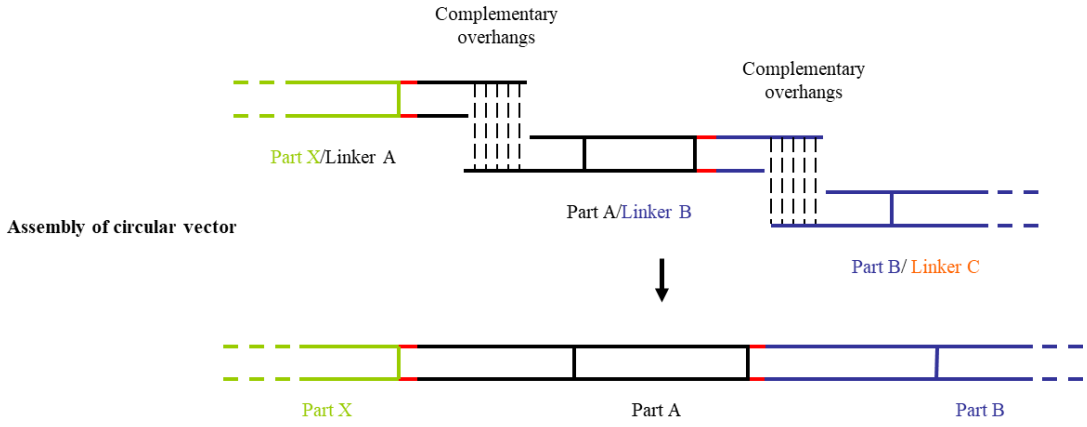


Figure 12: Phase 3 - The assembly reaction. During this reaction purified part/linker fusion products assemble as a result of complementary 16 nucleotide single stranded overhangs to generate final vector construct.

1.6.2 Further advances in accelerating the build cycle

Biosynthetic pathway construction capabilities can be further enhanced through the utilisation of techniques which are complementary to DNA assembly. A number of these can be directly tied to the decreasing cost of DNA synthesis[86] allowing large libraries of constructs to be built and characterised. A significant amount of research has been carried out particularly into the characterisation and construction of libraries of promoters[87,88] and terminators[89, 90], prediction of RBS strengths[91] and the effect of genome location on gene expression levels[92,93, 94] . Each of these approaches can be used in parallel with combinatorial DNA assembly strategies to fine tune gene expression levels and optimise pathway flux.

In parallel, techniques such as CRISPR/Cas9[95] and MAGE[96] have facilitated rapid and precise modifications of the host organisms genome.. The CRISPR/Cas system is an adaptive bacterial and archaeal immune system that can be utilised for rapid genome engineering. The system utilises a guide RNA (sgRNA) to guide the

endonuclease Cas9 to the targeted genome location where it catalyses a double strand break. During the repair of this break (usually through homologous recombination although non-homologous end joining is also possible), DNA introduction or deletion can be achieved. This approach has been used to engineer a number of organisms including *E. coli*, *P. putida*, *B. subtilis* and *S. cerevisiae* [97]. The system has also been adapted for multiplexed gene deletions and insertions [98, 99] as well as gene downregulation using a Cas9 protein which has been engineered to maintain DNA binding capacity without endonuclease activity (dCas9). The modified dCas9 protein can be targeted to the gene of interest to block transcription (CRISPR interference - CRISPRi)[100]. Such a strategy was recently used in *C. glutamicum* to downregulate three genes in order to maximise L-lysine and L-glutamate production[101]. It was found that such an approach resulted in yields comparable to those achieved through gene deletion, however the above strategy only took 3 days in comparison to a potentially time consuming and cumbersome gene deletion strategy. This approach has been further expanded to fuse dCas9 to an activator allowing for gene upregulation (CRISPR activation - CRISPRa)[102] facilitating the rapid up and downregulation of cellular pathways to optimise production formation.

Multiplexed automated genome engineering (MAGE) is a popular method for the generation of focused libraries at the genome level. This is achieved through the simultaneous introduction and incorporation of synthetic single stranded oligonucleotides at the DNA replication fork to introduce desired mutations[96]. To demonstrate the power of the technique enhancement of lycopene production in *E. coli* was previously targeted. The RBS of twenty genes previously shown to involved in lycopene production[103] were manipulated to modulate translational efficiency and four genes known to negatively impact lycopene production were targeted for knockout through the introduction of nonsense mutations. This approach resulted in a five-fold increase in lycopene production in 3 days. Whilst the production levels were lower than previously reported studies in which *E. coli* was rationally engineered[103,104], the time savings were significant. The MAGE technique has recently been adapted for use in *S. cerevisiae*[105] further expanding its potential utilisation in microbial engineering.

1.7 Enhancing the test stage

1.7.1 Complementing DNA assembly with omics approaches

The utilisation of synthetic biology tools for the engineering of industrial microbes is accelerating the ability to construct large numbers of engineered microbes during the build stage of strain engineering. However, the ability to screen these strains and vitally generate the information required to guide subsequent rounds of engineering is not progressing at the same rate. Whilst the success or failure of the engineering of a host will always be judged through the titer, rate and yield of product achieved by the engineered organism, high throughput screens which are directly linked to target product formation are unlikely to provide the required information to efficiently identify pathway bottlenecks and thus define targets for strain improvement.

A number of techniques to quantify cellular metabolite, transcript and protein abundances, known collectively as omics, have been developed which could be exploited to generate the amount of data required to this issue. The standard approaches however, are relatively low throughput (and utilise expensive instrumentation) so identification of the correct strategy to complement the build stage of engineering an industrial microbe is vital.

Metabolomics analysis provides the opportunity to explore pathway intermediates and potential bottlenecks however if competing pathways are directing intermediates away from the engineered pathway then bottleneck identification is not straight forward. Analysis of mRNA has rapidly developed with methodology in place allowing for the analysis of an organisms entire transcriptome in a single experiment[106, 107]. This is often implemented following the introduction of a pathway to ensure transcription of the pathway genes. mRNA levels however do not necessarily indicate the amount of protein being produced, with factors such as RBS strength, codon usage and mRNA secondary structures influencing how efficiently the mRNA is translated into protein.

Efficient pathway flux is often dependent on fine tuning of the amount of protein produced at each stage in the pathway since over production of each pathway protein results in significant metabolic burden being placed on the cell[108, 109]. The utilisation of accurate and high throughput approaches amenable to multiplexing to monitor protein expression can therefore dramatically accelerate strain engineering. Proteomics is however more challenging than transcriptomics as proteins cannot be amplified nor are they trivial to separate, both factors that can limit the throughput of such an approach. Western blot and ELISA are currently the most applied techniques for the specific analysis of proteins in unpurified extracts. Western blot analysis requires antibodies to be available or produced for each target protein. If quality antibodies are available Western Blot is highly specific and sensitive however it is limited in quantitation and the ability to examine multiple proteins in parallel. ELISA[110] in comparison is a quantitative technology but also limited in that a single target can only be examined in each assay. Although both techniques are used in day to day research due to ease of use and rapid processing, they do not provide the specificity or coverage that mass spectrometry based approaches can offer.

The utilisation of omics based platforms to characterise engineered microbes with the goal of identifying targets for strain improvement is still relatively rare. This is despite such an approach having proven successful on a number of occasions. For example, Pitera *et al*[111] demonstrated the use of a metabolomics based approach for the optimisation of a heterologous mevalonate pathway. Metabolomics analysis revealed a pathway bottleneck in which a pathway intermediate was being accumulated resulting in inhibition of cell growth. This informed the next stages of engineering which resulted in debottlenecking of the heterologous pathway and optimised flux towards the compound of interest. Transcriptomics analysis coupled to *in silico* modelling meanwhile was used to optimise L-valine production in *E. coli*, resulting in a 126.7% increase in product formation[112]. Multi-omics strategies have also been utilised, for example Brunk *et al* attempted to combine metabolomics, proteomics and predictions from *in silico* genome scale models to optimise carbon flux through the mevalonate pathway for terpenoid biosynthesis in *E. coli* [113]. Omics' approaches have also been utilised not only for the analysis of metabolically engineered strains but also for the identification of suitable host strains for further

engineering. In one such study a butyric acid producing *Clostridium* strain was identified (*Clostridium tyrobutyricum*) as a potential host for optimised butyric acid production, however limited knowledge of the genetic and metabolic characteristics of this strain had hampered metabolic engineering attempts. To address these issues genomic and proteome characterisation of the strain were undertaken. This analysis elucidated an alternative butyric acid production pathway in this strain in comparison to other clostridia species opening the door to the potential engineering of this strain to overproduce butyric acid[114]. A genomics and transcriptomics approach was also utilised in conjunction with *in silico* genome scale modelling to assess seven *E. coli* strains widely used in academic research and the biotechnology industry for their abilities to produce 40 chemicals under both aerobic and anaerobic conditions identifying certain strains more suited to produce certain chemicals[115].

An early example of using proteome analysis to maximise heterologous production in an engineered *E. coli* strain was performed by Han *et al* using two-dimensional gel electrophoresis[116]. It was found that the heterologous overproduction of serine rich (11.6% compared to an *E. coli* average of 5.6% serine) human leptin resulted in a significant decrease in abundance of the enzymes involved in serine amino acid biosynthetic pathways. They were able to identify one of the downregulated proteins, cysteine synthase A (encoded by *cysK*) and target it for over production. This strategy resulted in a two fold increase in cell growth and a four-fold increase in leptin production. To further exemplify the approach the cysteine synthase overproduction strain was shown to be able to produce another serine rich protein, interleukin-12 β chain (11.1% serine) at an enhanced level in comparison to the WT host. In one of the first examples of using mass spectrometry based proteomics to characterise engineered *E. coli*, Redding-Johanson *et al*[117] used targeted proteomics to optimise a heterologous pathway for production of the sesquiterpene amorpha-4,11-diene. Proteomics analysis revealed that two of the pathway proteins were only present at extremely low levels. Through codon optimisation of the gene sequences for expression in *E. coli* and the introduction of a strong promoter upstream of the coding sequence they were able to improve production three fold. These two studies identify the potential of using an omics approach to pathway optimisation but uptake

within the field has been limited as slow throughput and high data-complexity are still prohibitive.

Beyond characterisation of pathway proteins comparative proteomics can also be utilised to identify changes in the host proteome due to expression of the heterologous pathway. Batth *et al*[118]identified the need for proteomics analysis to have increased throughput as our ability to rapidly engineer microbes increased. They describe a platform for optimised sample throughput which was used to quantify 400 proteins from major metabolic *E. coli* pathways. Nowroozi et al[119] used targeted proteomics coupled to RBS modulation and a combinatorial assembly platform to optimise the mevalonate pathway for isoprenoid production in *E. coli*. Proteomics based approaches have also been used in order to characterise and quantify post translational modifications. This technique has gained significant traction in the field of polyketide synthase engineering. Through monitoring of the peptide which carries the phosphopantetheine modification bottlenecks in engineered polyketide function can be identified and optimised[120].

1.7.2 Mass spectrometry based protein analysis

Fundamentally all mass spectrometers measure the mass to charge ratio (*m/z*) of ions in the gas phase. From this, information on protein sequence, protein amounts and post-translation modifications can be derived. Simplistically, all instruments comprise of a sample inlet, which is often preceded by a form of chromatography to simplify the sample being analysed, a source in which gas phase ions are generated, a mass analyser and a detector. The detector will convert the number of ions reaching it - separated based on mass vs charge - into a mass spectrum.

The introduction of soft ionisation techniques led to highly sensitive mass spectrometry based protein and peptide analysis becoming the method of choice for analysing biological macromolecules. These techniques allow for the ionization of intact proteins and peptides which was not feasible with harsher chemical and electron ionisation approaches which had been used previously. John Fenn and

Koichi Tanaka developed Electrospray ionisation (ESI)[121] and Matrix-Assisted Laser Desorption (MALDI)[122] respectively for which they were awarded a share of the Nobel prize for Chemistry in 2002 and since this development MS has been a vital tool in protein research.

ESI (described in detail below) ionises the sample from solution and is therefore easily coupled to separation techniques such as liquid chromatography whilst MALDI sublimates and ionises the samples from a crystalline matrix through laser pulsing. MALDI has therefore been preferred for the analysis of relatively simple mixtures whilst in-line LC-ESI-MS systems have been preferred for the analysis of complex mixtures.

Mass analysers which can be utilised for protein analysis can be split into four primary categories, Time of Flight (TOF) (described below), ion trap, quadrupole (Q) and Fourier transform ion cyclotron resonance (FTICR). These analysers can be deployed alone or, in some cases, in tandem to exploit the advantages of the various strategies for analysis of proteins and peptides. In this study an electrospray quadrupole time of flight (Q-TOF) mass spectrometer was the instrument used for all protein analysis.

The development of these ionisation techniques and high resolution mass analyser coupled to advancements in protein/peptide separation, genome sequencing and the availability of protein database search tools have accelerated the advancement in the field of mass spectrometry based proteomics.

1.7.2.1 Proteins to peptides

Workflows for the identification and quantification of peptides from a complex mixture can generally be split into two main categories: Gel based and gel free techniques. In gel based strategies the proteins are first separated by 2D electrophoresis (2DE), stained and the proteins quantified by spot intensity. The desired spot is then excised, enzymatically digested into peptides and analysed via MS in a single step. The approach has been reported to suffer from limited dynamic

range with a systematic study of *Saccharomyces cerevisiae* highlighting that usually only the most abundant proteins can be observed using this approach[123]. In an attempt to address this issue, strategies such as more sensitive gel staining procedures have been developed[124].

Gel free mass spectrometry based protein quantification methodologies for the analysis of cellular extracts, such as those generated from an engineered industrial microbe, standardly utilise the 'shotgun' or bottom up technique[125] in which proteins are firstly digested into peptides typically between 5 and 14 amino acids in length and then fractionated prior to MS analysis (Figure 13).

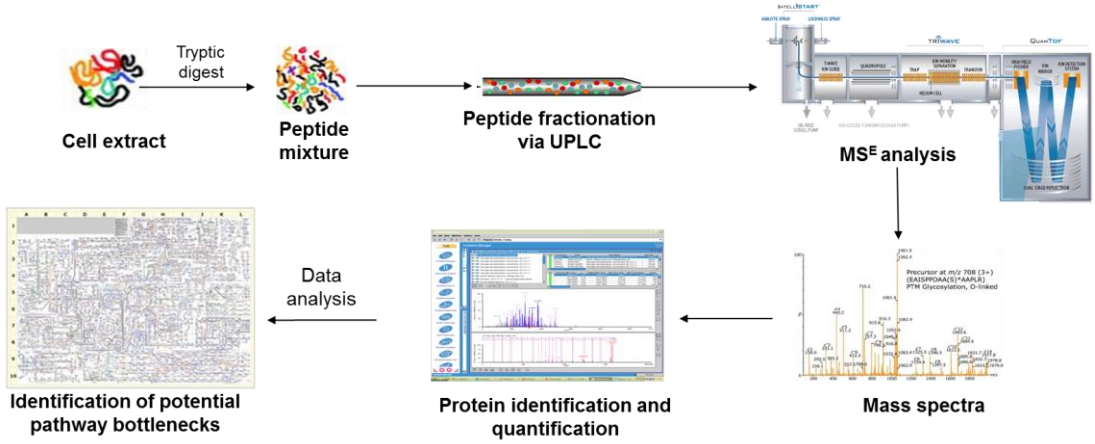


Figure 13: Proposed workflow for the mass spectrometry based quantification of target proteins from the cell extract of a strain engineered for utilisation in an industrial bioprocess.

The enzyme used for proteolysis in most approaches is trypsin which hydrolyses the peptide bond at the carboxyl side of the basic amino acids lysine and arginine, except when either are followed by proline. Efficiency and reproducibility of this initial step is vital if meaningful proteomics data is to be obtained[126]. For trypsin to efficiently cleave the protein into peptides, samples are denatured, reduced and alkylated prior to digestion. Denaturation and reduction are often performed in parallel through a combination of heat and treatment with a reducing agent such as 1, 4-dithiothreitol (DTT). To further reduce the potential for protein renaturation prior to digestion cysteine residues are alkylated, standardly through the addition of iodoacetamide. Commercial preparations of trypsin used for proteomics acquisitions

are standardly modified through reductive methylation of lysine residues to prevent autolysis which can result in the carry through of trypsin peptides into the analysis[127].

Due to the complexity of samples following tryptic digestion (thousands of peptides from a proteome sample at varying abundances) in line liquid chromatography, either high pressure liquid chromatography or as is the case in this study, ultra-high pressure liquid chromatography (UPLC), is standardly used to fractionate the sample and reduce the complexity of the sample entering the mass spectrometer. Reverse phased (RP) chromatography is generally the separation method of choice for proteomics based analysis. This approach employs a non-polar stationary phase such as C18 analytical column chemistry and a polar mobile phase comprised of aqueous solution and organic solvent. It allows peptide separation based on hydrophobicity and therefore the most water soluble peptides pass through the column quickest.

1.7.2.2 Electrospray ionisation

Electrospray ionisation (ESI) is a soft ionisation technique in which the analyte of interest is solubilised in a volatile buffer or solvent system. It is a technique which has been used to ionise a number of analytes from small organic and inorganic molecules to large polymers, nucleic acids and proteins. It is particularly suited to coupling to separation techniques such as liquid chromatography and has been one of the most popular ionisation methods for protein and peptide analysis.

There are three main steps in the production of gas phase ions from solution through electrospray ionisation: (1) Production of charged droplets at the capillary tip; (2) Droplet shrinkage through solvent evaporation and charge induced droplet fission resulting in small highly charged droplets capable of producing gas phase ions, and (3) the production of gas phase ions from these droplets.

The solution containing the analyte is passed through a conductive capillary under the influence of an electric field which is generated through having a large potential difference (~5 kV) between the capillary and the entrance to the mass spectrometer

which is standardly 0.3 – 2 cm from the capillary. This results in the charged analyte solution forming a Taylor cone, from which highly charged droplets are omitted.

The resultant droplets reduce in size due to solvent evaporation resulting in increased charge density. The increased charge density will result in the generation of a Coulombic repulsion which at a certain droplet radius exceeds the cohesive surface tension resulting in the splitting of the droplet. This process of repeated droplet fission results in very small charged droplets which are the precursor to gas phase ions.

Two mechanisms have then been proposed for the formation of gas phase ions. The ion evaporation model suggests the process of droplet fission continues until the droplet size has decreased to ≤10 nm and charged analytes can be directly desorbed into the gas phase at a point in which repulsion between ions at the surface of the droplet overcomes the cohesive force of the surface tension. It is believed for molecular species carrying a low number of charges that this mechanism is dominant. For molecular species carrying multiple charges it is believed that the charge residual model is dominant, in this model droplet generation would result in droplets containing one analyte molecule. The ion is not directly desorbed in this model, instead solvent evaporation continues until all the solvent is lost and the ion enters the gas phase[128].

1.7.2.3 Time of flight mass analysers

Time of flight (TOF) mass analysers are one of the oldest ion separation techniques. However, despite first being described in 1946[129] it was not until 1995 that they gained considerable popularity[130]. The approach relies on the separation of ions through the time taken for them to reach a detector in order of increasing mass to charge ratio. In its simplest form the approach relies on the acceleration, by an electric field, of a set of ions through a field free drift region within a linear vacuum tube to towards the detector. Ions that have differing *m/z* but acquire the same kinetic

energy will achieve a different velocity and therefore take differing amounts of time to reach the detector.

This can be shown through firstly using the equation to calculate kinetic energy (Ek) for any mass (Eqn. 1).

$$Ek = \frac{1}{2}mv^2$$

Equation 1: Calculation of kinetic energy

Where Ek is the kinetic energy after acceleration, m is the mass of ion and v is the velocity. Following the initial acceleration, the velocity of the ion remains constant as it travels through the field free region and therefore velocity can be calculated from:

$$v = \frac{d}{t}$$

Equation 2: Calculation of velocity

Where d is the distance travelled and t is the time taken for the ion to traverse the field free region and reach the detector. Introduction of v into Eqn. 1 results in:

$$Ek = \frac{1}{2}m(d/t)^2$$

Equation 3: Introduction of velocity calculation into kinetic energy equation

Which if solved for mass, yields:

$$m = \frac{2Ekt^2}{d^2}$$

Equation 4: Equation solved for mass

This equation is the basic time of flight relationship, for a given kinetic energy and distance the time of flight is proportional to the square of the mass [131].

Significant improvements in the resolution of TOF mass analysers has been achieved through the introduction of a "reflectron" within the flight tube[132]. This is a device which uses an electric field to first reduce the velocity of the ions to zero before reversing the ions back into the field free region. This approach not only increases

the length of the drift region but also corrects for any variations in the kinetic energies of the ions during the initial acceleration. This is achieved as ions with the same *m/z* but higher kinetic energies penetrate further into the reflectron and therefor taking longer to be reflected.

1.7.2.4 Peptide identification

Having measured the *m/z* and the intensity of the analysed peptides, primary sequence information is also standardly generated for the peptides analysed. This is achieved through tandem mass spectrometry (MS/MS and MS^2) in which two stages of MS are utilised. Initially the mass spectrometer scans all eluting precursor peptides recording mass and intensity. Peptides are selected for fragmentation through collision with an inert gas (nitrogen, argon or helium standardly) generating product ions in either a data dependent (fragmentation of specific peptides) or data independent (fragmentation of all peptides) manner. Following this fragments are associated to peptides which are assembled into proteins and the identifications statistically validated[133].

Data dependent acquisitions (DDA) rely on a serial process of peptide selection followed by fragmentation. A fixed number of peptides (usually 3-8) from a survey MS scan are selected for fragmentation. The selected peptides are fragmented, and data collected on the product ions over a period of time or until a defined ion current is breached. During this time all information on co-eluting peptides which were not selected for fragmentation is lost. These cycles of precursor and product ion analysis continue throughout the runtime. In an attempt to produce the highest quality fragmentation data standardly the most abundant peptides will be selected and therefore a key limitation of this approach is the potential loss of information on less abundant or less well ionised peptides.

An alternative to data dependent acquisitions, known as MS^E, was developed by Silva and co-workers and commercialised by Waters Corporation[134, 135]. In this data independent approach, no precursor identification is required. During this type of

acquisition, the quadrupole guides all ions into the collision cell which the instrument alternates between high energy and low energy over the course of the run. During the low energy scan data is generated on the *m/z* and retention time of the eluting peptide whilst during the high energy scan fragmentation data is collected. As this method is completely data independent there is no bias towards highly abundant peptides meaning data is collected for all peptides potentially maximising the number of proteins quantifiable.

Post-acquisition MS^E data processing is then employed in order to link fragments to their precursors, defining peptide sequences and associating them to proteins for identification. In this study all identifications were carried out using the identify algorithm in Protein Lynx Global Server (PLGS) software version 3.03 developed by Waters Corp. This software utilises peptide retention time, precursor and fragment ion intensity, charge states and precursor and fragment ion accurate masses for protein identification[136].

Data from an MS^E experiment is collected as three functions. A low energy function (Function 1), a high energy function (Function 2) and lockmass function (Function 3) which comprises of a reference compound infused throughout the run for accurate mass correction. An initial ion detection is performed from Function 1 and 2. First an algorithm, Apex 3D, is used to subtract noise and integrate ion current signals across the chromatographic elution. The output of this algorithm is a list of ions with intensities above a user defined threshold[137]. The second algorithm, Pep3D, collapses isotopes and charge states of both precursor and product ions into an Exact Mass and Retention Time (EMRT). The EMRT is a peptide of unknown sequence which is characterised by its mass, retention time and intensity. Fragment ions are then tentatively associated with their precursors based on elution profiles. It is likely that multiple precursors will co-elute from the column and therefore product ions will be associated to all potential precursors until later in the process.

Associated product and precursor ions are then filtered, with low molecular weight peptides - standardly below 750 Da under low energy and 350 Da under high energy - side lined to ensure highly specific initial protein identifications. Although most proteins will generate tryptic peptides smaller than 750 Da these have high sequence

similarity and therefore low specificity. Product ions with higher intensities than their predicted precursor are also filtered at this stage.

Prior to data searching a decoy database of *in silico* generated peptides is generated to allow for calculation of the false discovery rate (FDR). PLGS automatically constructs this database through generating one random sequence for each protein in the database to be searched. The random sequence contains the same number of amino acids as the original protein and would generate the same number of tryptic peptides however the order of the amino acids is randomised. Identical search parameters are then used to search both databases and the FDR calculated as a percentage by dividing the number of protein identifications from the decoy database by the total number of protein identifications (identifications for the decoy plus identifications from the protein database).

The software then begins database searching, which is performed in three stages, known as passes. In the first pass the software cycles through the data removing EMRT's which it considers to be peptides derived from the most confident protein identifications. This process continues until the rate of protein identification from the decoy database exceeds the user defined FDR. During pass two a subset database containing only proteins which were identified in pass one is created, however additional search parameters such as missed tryptic cleavages, variable peptide modifications and in source fragmentation are assigned to previously unassigned peptides that could have come from the proteins identified in the first pass. During the final pass all remaining precursor information is used to search the full database, however now total product ion intensity is allowed to exceed precursor ion intensity which can be a signature of in source fragmentation of highly labile peptides[136].

It is vital that a comprehensive protein database is constructed for the strain being characterised. This database should contain all proteins produced by the parental strain, heterologous proteins expressed as a result of strain engineering, trypsin if used for digestion and any proteins introduced during the sample preparation such as lysozyme for cell lysis. In this study the UniprotKB database (http://www.uniprot.org/) and the *E. coli* K12 proteome database from Uniprot were

downloaded for interrogation with MS^E data. The *E. coli* database had heterologous expressed protein sequences manually added when necessary.

1.7.2.5 Relative and absolute protein quantification

Quantification of proteins from complex mixtures can be achieved in a relative or absolute manner. Absolute quantitation techniques require the addition of internal standards such as synthetic peptides at known concentration which are designed to mimic the peptide of interest. In this approach, the peptide is synthesised with a stable isotope present within one amino acid of the peptide. This should result in the synthetic peptide and the target peptide having similar properties in terms of chromatography, ionisation and fragmentation but remain distinguishable due to the difference in mass.

Relative quantitation approaches can be divided into two main categories, isotope based and label free. Isotopic labelling can be performed through a number of techniques with the most popular including Stable Isotope Labelling by Amino Acids in Cell Culture (SILAC)[138], Isotope-coded Affinity Tag (iCAT)[139] or Isobaric Tags for Relative and Absolute quantification (iTRAQ)[140]. These techniques rely on either the incorporation of labelled amino acids into the protein during cell growth (SILAC), chemical labelling of proteins prior to digestion (iCAT) or enzymatic labelling of the peptides following digestion (iTRAQ). A primary advantage of these techniques in comparison to label free methodologies is that the samples being analysed are mixed, either prior to digestion (SILAC and iCAT) or prior to LC-MS analysis (Figure 14). This means that variations in sample preparation and chromatography that may exist sample to sample will not have as a profound effect on the data collected. Despite the success of these techniques label-free quantitation is becoming a more popular alternative due to limitations in the scope of labelled approaches, the increased time and complexity of sample preparation, the requirement for increased sample concentration or simply the cost[141].

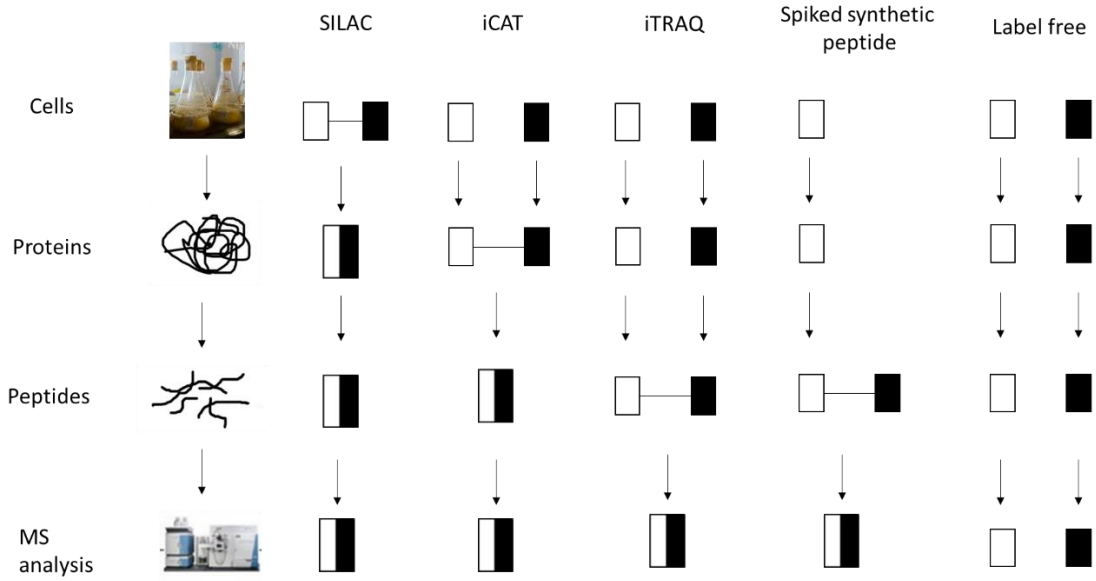


Figure 14: Comparison of typical quantitative proteomics techniques. Boxes in black and white represent two experimental conditions which require to be compared. Horizontal line represents sample mixing. At each stage where samples are processed in parallel – i.e. not mixed – quantification errors which are not compensated for may occur.

Label free methodologies in comparison are likely to be quicker and more cost effective due to the avoidance of additional sample processing steps and the cost of labelling reagents however as each sample is analysed in isolation robust digestion and LC methods require to be in place prior to analysis of complex mixtures. Label free methodologies rely on quantification through either spectral counting or signal intensity of the detected peptides.

Relative quantification through spectral counting is achieved through the comparison of the number of MS/MS peptide spectral matches from the same protein. This is based on the assumption that the number of peptide spectral matches is a correlation of total protein amount as more abundant proteins will generate more proteolytic peptides. Therefore, an increase in protein abundance results in an increase in protein sequence coverage, the number of identified unique peptides and the total number of identified MS/MS spectra. Liu *et al.* explored the correlation between sequence coverage, peptide number and spectral count (number of total identified MS/MS spectra) and protein abundance. This was achieved through spiking a yeast soluble

protein extract with specific quantities of six proteins, the protein mixes were then trypsin digested and the peptides analysed. It was found that only spectral count gave a strong correlation to relative protein abundance ($r^2 = 0.9997$)[142]. The accuracy of the approach was further improved through the definition of a normalised spectral abundance factor (NSAF) which also considers protein length. As larger proteins will tend to contribute more peptide/spectra than shorter ones the NSAF for a given protein is calculated as the number of spectral counts (Sc) identifying that protein divided by its length (L) divided by the sum of Sc/L for all proteins in the experiment[143].

When an ion is detected with a particular *m/z* information on intensity and retention time is also recorded. It was found that signal intensity from ESI correlates to ion concentration[144]. The potential to use this signal intensity based approach for protein/peptide quantification was initially explored through the analysis of 10 fmol – 100 pmol myoglobin digest via nano-LC coupled to LC/MS/MS[145]. The chromatographic peak area was extracted and calculated for five of the identified peptides and shown to increase with increasing concentration. Combination of peak areas and plotting against protein concentration resulted in a linear correlation ($r^2 = 0.991$) showing that the peak areas and concentration strongly correlates. This approach was further validated - by the same authors - in complex mixtures through the addition of horse myoglobin to human serum.

A number of factors require to be considered to ensure reliable protein quantification through peak intensity. These include highly reproducible LC retention times and the collection of sufficient data points across the curve to calculate peak area which can be an issue in MS/MS data dependent acquisitions as the MS scans used to collect peak area information are not collected during periods in which MS/MS information is being collected for protein identification. Experimental variation caused by sample preparation or injection variation and drifts in retention time and *m/z* can be reduced through data normalisation using spiked standards, abundant housekeeping proteins or the analysis of total ion area. Computational methods which align retention times across multiple samples, consider background noise and normalise peak abundance have been used to address these concerns. The ability to quantify multiple proteins within a cellular lysate extract following genetic engineering is a promising strategy

which allows for not only the determination of potential bottlenecks within the biosynthetic pathway but also the potential to examine unexpected effects engineering may have had on the host organism. The optimisation of tools such as the inABLE DNA assembly platform will allow for the rapid construction and introduction of synthetic pathways into host organisms and technologies such as proteomics are required to characterise the success or failure of these pathways and therefore direct the next round of engineering in a predictable manner. Through coupling combinatorial pathway construction, determination of product formation and a label free proteomics based strain interrogation approach the optimum abundance of each pathway protein can be understood. This understanding leads to more predictable pathway optimisation which would not be possible without the combination of these approaches.

1.8 Aims and Motivations

In this PhD, I will aim to address the primary issues cited by end users as to why bioprocesses are not more readily taken up – their development is too slow, too unpredictable and too expensive. This will be achieved through the development of a DNA assembly technique optimised for combinatorial pathway construction whilst making it feasible to rapidly construct and test many modified organisms. Although assays or screens provide information on the complete system they often do not provide the level of information required to guide the next stages of engineering. I will therefore focus on the development of a robust strain analysis tool which provides the level of information required to guide subsequent stages of strain optimisation. Proteomics gives the ability to analyse the engineered organism to a level which will result in the predictability that can only come from having an in depth understanding of the dynamic biological system. The combination of a DNA assembly technique optimised for rapid and efficient pathway construction and a proteomics platform for analysis of the engineered microbes will result in the development of an integrated platform for the accelerated engineering of industrial microbes.

2. Materials and Methods

2.1 Materials and reagents

All chemicals and solvents were purchased from Sigma Aldrich or VWR unless otherwise stated. *Escherichia coli* strains Top10 (Thermo Fischer) and BW25113 (Invitrogen) were used for plasmid construction and vector screening. All DNA modifying enzymes were purchased from New England Biolabs unless otherwise stated. Mass spectrometry grade trypsin was purchased from Thermo Fischer and prepared as per the manufacturer's instructions.

2.2 *E. coli* strains and plasmids

2.2.1 *E. coli* strains

Strain (Reference/supplier)	**Genotype**	**Application**
Top10 (Invitrogen)	mcrA, Δ(mrr-hsdRMS-mcrBC), Phi80lacZ(del)M15, ΔlacX74, deoR, recA1, araD139, Δ(ara-leu)7697, galU, galK, rps (SmR), endA1, nupG	Cloning strain
NEB 10-beta (NEB)	araD139 Δ(ara-leu)7697 fhuA lacX74 galK (f80 Δ(lacZ)M15) mcrA galU recA1 endA1 nupG rpsL (StrR) Δ(mrr-hsdRMS-mcrBC)	Cloning strain
BL21 (DE3) (Studier *et al.*[146])	F– ompT gal dcm lon hsdSB(rB–mB–) λ(DE3 [lacI lacUV5-T7p07 ind1 sam7 nin5]) [malB+]K-12(λS)	Protein expression
BW25113 (Datsenko *et al.*[147])	lacI+rrnBT14 ΔlacZWJ16 hsdR514 ΔaraBADAH33 ΔrhaBADLD78 rph-1 Δ(araB–D)567 Δ(rhaD–B)568 ΔlacZ4787(::rrnB-3) hsdR514 rph-1	Parental strain for reverse glyoxylate shunt screening host

Table 2: E. coli strains used in this study.

2.2.1 Plasmids

Plasmid name	Description	Reference
Assembly T1	pMB1 ori, AmpR, CampR, KanR, TetR	This work
3 bp overhang assembly product	pMB1 ori, TetR	This work
LOA secondary structure assembly product	pMB1, KanR	This work
Assembly Cellulase 1	pMB1 ori, AmpR, truncated ADH1p, alpha factor secretion signal, β-glucosidase a	This work
Assembly Cellulase 2	pMB1 ori, AmpR, truncated ADH1p, alpha factor secretion signal, β-glucosidase b	This work
Assembly Cellulase 3	pMB1 ori, AmpR, truncated PGK1p, alpha factor secretion signal, Endo/exoglucanase, TPS1t	This work
Assembly Cellulase 4	pMB1 ori, CampR, truncated CYC1t, ZeoR, AmpR, Cen4/ARSH4	This work
Assembly Cellulase 5	pMB1 ori, AmpR, PGK1p, alpha factor secretion signal, Endo/exoglucanase, TPS1t, ADH1p, alpha factor secretion signal, β-glucosidase a, CYC1t, ZeoR	This work
Assembly Cellulase 6	pMB1 ori, AmpR, PGK1p, alpha factor secretion signal, Endo/exoglucanase, TPS1t, ADH1p, alpha factor secretion signal, β-glucosidase b, CYC1t, ZeoR	This work
Nested Pr-Ss vector	pMB1 ori, AmpR, CampR, β-glucosidase, CYC1t, 2-micron ori, KanMX	This work
Assembly efficiency vector 1	pBR322 ori, AmpR, CampR	This work
Assembly efficiency vector 12	pBR322 ori, AmpR, KanR	This work
eGFP expression vector	pMB1ori, CampR, ProC, eGFP	This work
PN vector 1	pMB1 ori, KanR (SapI flanked), AmpR	This work
PN vector 2	pMB1 ori, KanR, ProC, eGFP, Syn ter (SapI flanked)	This work
pET21 MTKa/b	pBR322 ori KanR, LacI, T7pr, MTKa, MTKb	This work
pET21 MCL	pBR322 ori KanR, LacI, T7pr, MCL	This work
pET21 ICL	pBR322 ori KanR, LacI, T7pr, ICL	This work
pRGS_3.1	pBR322 ori KanR, pRecA, MTKa/b, MCL, ICL	This work
pRGS_3.2	pBR322 ori KanR, pA, MTKa/b, MCL, ICL	This work
pRGS_3.3	pBR322 ori KanR, pOSMY, MTKa/b, MCL, ICL	This work
pRGS_3.4	pBR322 ori KanR, pRecA, MTKa/b, MCL, ICL	This work
pRGS_4.1	pBR322 ori KanR, pRecA, MTKa/b, ICL, MCL	This work
pRGS_4.4	pBR322 ori KanR, pOSMY, MTKa/b, ICL, MCL	This work
pRGS_4.5	pBR322 ori KanR, pA, MTKa/b, ICL, MCL	This work
pRGS_5.1	pBR322 ori KanR, pL, MCL, MTKa/b, ICL	This work
pRGS_5.3	pBR322 ori KanR, pRecA, MCL, MTKa/b, ICL	This work
pRGS_5.5	pBR322 ori KanR, pOSMY, MCL, MTKa/b, ICL	This work
pRGS_5.7	pBR322 ori KanR, pA, MCL, MTKa/b, ICL	This work
pRGS_6.1	pBR322 ori KanR, pOSMY, MCL, ICL, MTKa/b	This work
pRGS_7.1	pBR322 ori KanR, pA, ICL, MTKa/b, MCL	This work
pRGS_7.2	pBR322 ori KanR, pOSMY, ICL, MTKa/b, MCL	This work

Table 3: Plasmids used in this study

2.3 Microbiology techniques

2.3.1 Culture media preparation

2.3.1.1 LB liquid media

37 g L^{-1} LB (Merck) prepared in RO H_2O.

2.3.1.2 LB agar

37 g L^{-1} LB agar (Merck) prepared in RO H_2O.

2.3.1.3 SOC liquid media

20 g L^{-1} Tryptone, 5 g L^{-1} yeast extract, 0.344 g L^{-1} NaCl, 0.185 g L^{-1} KCl, 2.4 g L^{-1} $MgSO_4$, 3.6 g L^{-1} glucose.

2.3.1.4 Ingenza minimal media

Stock Salts Solution (5X)

10 g L^{-1} $(NH_4)_2SO_4$,

73 g L^{-1} K_2HPO_4

10 g L^{-1} NaH2PO4 · 2H2O

10 g L^{-1} $(NH_4)_2$H-citrate

Trace Elements (500X)

0.5 g L^{-1} $CaCl_2.2H_2O$

10.03 g L^{-1} $FeCl_3$

0.18 g L^{-1} ZnSO4.7H2O

0.16 g L^{-1} $CuSO_4.5H_2O$

0.15 g L^{-1} $MnSO_4.H_2O$

0.18 g L^{-1} $CoCl_2.6H_2O$

22.3 g L^{-1} $Na_2EDTA.2H_2O$

To prepare 1 L minimal media

200 mL 5X Stock salt solution

20 mL Glucose 50% (w/v)

2 mL 1M 500X Trace elements

Optional: 15 g L^{-1} Select agar (Sigma)

2.3.1.5 Yeast nitrogen base

10x stock preparation

0.67 g L^{-1} Yeast nitrogen base (Sigma)

0.5 g L^{-1} Glucose

Working stock prepared through dilution in sterile H_2O prior to use

2.3.2 Antibiotic preparation

Ampicillin, chloramphenicol, kanamycin and tetracycline were purchased from Sigma. Ampicillin and kanamycin were prepared in H_2O, whilst chloramphenicol and tetracycline were prepared in 100 % [v/v] and 70 % [v/v] ethanol respectively. Stocks were prepared at 1000 x concentration and stored at – 20 °C. Working concentrations of each antibiotic found in Table 4.

Antibiotic	Working concentration (µg/mL)
Ampicillin	200
Chloramphenicol	34
Kanamycin	100
Tetracycline	10

Table 4: Antibiotic working concentrations.

2.4 DNA manipulation

2.4.1 Purification, isolation and quantification

Plasmid DNA was isolated using either plasmid mini or midi kits (Qiagen). DNA fragments were purified using Monarch DNA Gel Extraction Kit or PCR & DNA clean up kit (NEB). All DNA was quantified using the Qubit fluorometer HS DNA quantification kit (Thermo Fischer). Manufacturer's guidelines followed for each kit.

2.4.2 Digestion and Ligation

All digestions were performed using restriction endonucleases purchased from New England Biolabs (NEB). Reaction conditions were defined by the manufacturer's recommendations.

DNA ligations were performed using NEB Quick Ligation kit. As standard a 5:1 molar excess of insert to backbone (total 100ng DNA) was used in a final reaction volume of 20 µl with 1 µl Quick Ligase and 1x Quick Ligase buffer. Reactions were incubated at room temperature for 15 minutes prior to transformation.

2.4.3 PCR

All oligonucleotides for PCR were synthesised by Sigma-Genosys and reconstituted to a concentration of 100 µM in molecular biology grade H_2O. Working stocks of PCR primers were generated through a further dilution in molecular biology grade H_2O to 10 µM.

PCR was routinely performed using Phusion high fidelity polymerase (NEB) using the following reaction conditions: 1x Phusion HF buffer, 200 µM dNTPs, 0.5 µM forward and reverse primer and 10ng template DNA in a total reaction volume of 50µl. Thermocycling conditions: Initial denaturation – 98 °C for 30 seconds, followed by 30 cycles of 98 °C for 5 seconds, 45 °C – 72 °C for 30 seconds, 72 °C for 15 seconds per kb and a final extension of 5 minutes.

2.4.4 Competent cell preparation

E. coli strains were made electrocompetent through streaking out cells from a master cell bank onto a LB agar plate supplemented with antibiotics as appropriate. The plate was incubated overnight at 37 °C to obtain single colonies.

A single colony was picked into 100 mL LB minus NaCl and incubated overnight at 37 °C 250 rpm. The following morning the culture was back diluted into 800 mL LB minus NaCl to an OD_{600} of 0.1. The culture was grown at 37 °C until an OD_{600} of 0.5 – 0.6 was reached at which point cells were chilled on ice for two hours. Following this stage all manipulations were performed at 4 °C or lower using ice chilled reagents.

Cells were pelleted through centrifugation (4000 rpm, 4 °C, 10 minutes) and the supernatant discarded. The cells were washed twice through resuspension in 200 mL ice cold RO H_2O followed by centrifugation (4000 rpm, 4 °C, 10 minutes) and removal of the supernatant. Following the second H_2O wash the pellet was resuspended in 40 mL ice cold 20 % (v/v) glycerol and centrifuged (5000 rpm, 4 °C,

15 minutes). The pellet was then resuspended in 8 mL ice cold 20 % (v/v) glycerol and 50 µl aliquots prepared. The aliquots were stored at – 80 °C prior to testing.

All chemically competent cells were purchased from NEB.

2.4.5 Electroporation

Electrocompetent cells (50 µL) were mixed with 1 µL of DNA in a 1 mm gap electroporation cuvette (Molecular BioProducts) and electroporated using a BioRad GenePulser™. Following a single pulse of 1.7 k, at a capacitance of 25 µFD and resistance of 200 Ω, the cells were recovered in 1 mL SOC. This was incubated at 37 °C, 250 rpm for 1 hour before an appropriate dilution of the cells were plated onto LB-agar supplemented with the appropriate antibiotic.

2.4.6 Heat shock transformation

Chemically competent cells (50 µl, NEB 10β) were mixed with 1 – 5 µl DNA and placed on ice for 30 minutes. Cells were heat shocked at 42 °C for 30 seconds and then placed on ice for 5 minutes. 950 µl SOC media was added to cells and 1 hour recovery performed at 37 °C, 250 rpm. An appropriate dilution of the cells was plated onto LB-agar supplemented with the appropriate antibiotic.

2.4.7 Calculation of transformation efficiency

The transformation efficiency of electrocompetent cells was determined through transformation using 50 pg of pUC18 (NEB) following the standard protocol. Following the recovery, a percentage (10%, 1%, 0.1% and 0.01%) was plated onto LB-Kan. Following overnight incubation (37 °C) the number of colonies were

counted, and the transformation efficiency calculated using Equation 5. A transformation efficiency of > 10^8 cfu/µg DNA was targeted.

$$Transformation\ efficiency = \frac{Number\ of\ colonies}{\mu g\ of\ pUC19\ \times\ \%\ dilution}$$

Equation 5 – Calculation of transformation efficiency.

2.4.8 Site directed mutagenesis

Site directed mutagenesis to remove incompatible restriction sites (standardly SapI and EarI) from DNA targeted for assembly via inABLE DNA assembly was performed following the protocol provided with the QuickChange II Site-Directed Mutagenesis Kit from Agilent. In brief two primers containing the desired mutation, flanked by unmodified nucleotide sequences were designed and used in a PCR reaction as indicated below.

5 µl 10X reaction buffer
X µl (50 ng) of dsDNA template
1 µl (10µM) primer #1
1 µl (10µM) primer #2
1 µl of dNTP mix (50x, 200 µM final concentration of each)
1 µl of PfuUltra HF DNA polymerase (2.5 U/µl)
ddH2O to a final volume of 50 µl

Thermocycling conditions: Initial denaturation – 94 °C for 30 seconds, followed by 12 cycles of 94 °C for 5 seconds, 50 °C for 30 seconds, 72 °C for 1 minute per kb of plasmid length. The PCR product was then digested at 37 °C for 1 hour using 1 µl Dpn I restriction enzyme (10 U/µl) to remove methylated plasmid starting material. 2 µl of the digested material was used to transform Top10 cells via electroporation and the resulting transformants screened for the expected mutation by restriction digestion and Sanger sequencing.

2.4.9 Capillary electrophoresis

DNA was analysed through electrophoresis using an Agilent 2200 Tapestation instrument following the manufacturer's instructions. In brief, samples were prepared by adding 1 µL of DNA to 10 µL of the appropriate sample buffer (either D1000, D5000 or Genomic DNA depending on the size of the DNA being analysed). The samples were vortexed for 1 min using an IKA vortexer with PCR tube adaptor at 2000 rpm prior to loading onto the Tapestation instrument. Results were analysed using the Agilent 2200 Tapestation Analysis software

2.5 inABLE DNA assembly

2.5.1 Truncated part and primer design

Part sequences were entered into the inABLE® Bio Python part designer program which splits the gene into a TP, POA, and LOA. PCR primers for amplification of a SapI flanked truncated part are also generated during this stage. The POA is composed of annealed part oligo short (POs) and part oligo long (POl) sequences whilst annealing of the linker oligo short (LOs) and linker oligo long (LOl) oligonucleotides generate the LOA. The oligonucleotides were synthesised by Sigma-Genosys and TP's were either synthesised by GenScript or PCR amplified.

2.5.2 Preparation of oligonucleotides for part linker fusion reaction

Phosphorylation of the 5'-end of each oligo previously designed (LOl, LOs, POl and POs) is required to allow efficient ligation with the truncated part during the part/linker fusion preparation. Annealing of the respective LOl, LOs, POl and POs was carried out to generate partially double stranded linkers harbouring specific 16 bp overhangs required for the assembly reaction between part/linker fusions.

Stock solutions (100 µM) of each oligonucleotide were first prepared by resuspending the lyophilized samples into water. Each primer pair (long and short) was then phosphorylated and annealed using T4 polynucleotide kinase (NEB) as detailed in Table 5.

Reagent	**Volume (µl)**	**Final Concentration**
Oligo long (100 µM)	10	10 µM
Oligo short (100 µM)	10	10 µM
PNK Buffer (10x)	10	1x
ATP (10 mM)	10	1 mM
T4 PNK (10 U/µl)	2	0.2 U
DTT (1 M)	0.5	5 mM
PEG 8000 (50% w/v)	10	5% w/v
Molecular grade water	47.5	-
Total Volume	**100**	-

Table 5: Reaction mixture for the phosphorylation of the inABLE oligonucleotides.

The reaction was incubated at 37°C for 30 minutes, followed by inactivation of the enzyme at 65°C for 20 minutes in a thermocycler. The reaction was then heated at 90°C and oligonucleotide annealing was achieved by gradually lowering the temperature down to 20°C (Table 6).

Temperature (°C)	Time (min)
37	30
65	20
90	5
85	1
80	1
75	1
70	1
65	1
60	0.5
55	0.5
50	0.5
45	0.5
40	0.5
35	0.5
30	0.5
25	0.5
20	0.5
Hold 4 °C	

Table 6: PCR program for oligonucleotide annealing to generate POA and LOA fragments

The resulting annealed part oligonucleotides (POA) and annealed linker oligonucleotides (LOA) were then stored at -20°C prior to use.

2.5.3 Truncated part amplification

When required, truncated parts were amplified using Phusion high fidelity polymerase (New England Biolabs) and the PCR primers previously designed (Section 2.5.1). Each PCR reaction was performed with multiple replicates in order to generate sufficient material for the part linker fusion reactions. Following PCR reactions samples were analysed by agarose gel electrophoresis. The expected bands were excised, and the DNA purified via gel extraction kit (Section 2.4.1).

2.5.4 Part linker fusion reactions

Part linker fusions were prepared by ligating the 5'-end of the truncated part TP with its corresponding part POA and the 3'-end of the truncated part with the LOA from the following part through cycles of EarI or SapI mediated digestion and T4 DNA ligase catalysed ligation.

The reactions were performed in a total volume of 50 µL and were incubated in a thermocycler. Each reaction comprised of 50 nM truncated part, 0.2 µM POA, 0.2 µM LOA, 1x NEB Buffer 4, 1 mM ATP, 12.5 U SapI/EarI and 500 U T4 DNA ligase. A ten-fold molar excess of LOA and POA to the truncated part was used to promote ligation between the truncated part and linkers during each ligation cycle.

Cycles of SapI/EarI digestion and ligation were achieved by alternating the temperature between 37°C and 16°C, which correspond to the optimum temperatures for the EarI digestion and for the ligation using the T4 DNA ligase respectively (Table 7).

Lid Temperature : 42°C	
Temperature	**Time (minutes)**
37	90
16	30
37	30
16	15
37	15
16	15
37	15
16	10
37	15
16	10
37	60
16	30
65	20
Hold at 4°C	

Table 7: PCR program for the preparation of part/linker fusions through cycles of digestion and ligation.

Upon completion part linker fusions were standardly run on a 1% agarose gel, the expected bands excised, and DNA extracted (Section 2.4.1).

2.5.5 Assembly reactions

Assembly reactions were performed through mixing 0.1 pmol of each purified part linker fusions together in the presence of 1 x NEB buffer 2 and 5% (w/v) PEG 8000. Reactions were incubated at 37 °C for 30 min prior to transformation.

2.5.6 Exonuclease III treatment

Following completion of the part linker fusion reaction cycling, 0.5 µl Exonuclease III (50 U), 6 µl NEB buffer 1 (10x) and 3.5 µl H_2O was added to each reaction (50 µl, 50 nM truncated part). The digestion reactions were incubated for 30 minutes at 37 °C before the enzyme was inactivated through incubating at 70 °C for 20 minutes.

2.6 Protein techniques

2.6.1 Protein expression

Overnight LB-Kan cultures of BL21 (DE3) harbouring expression vectors were grown at 37 °C and 250 rpm to an optical density at 600 nm (OD_{600}) of 0.6. Protein expression controlled by the T7 promoter of the plasmid pET21 was induced by addition of 1 mM IPTG (final concentration) and cultures were further incubated at 37°C and 250 rpm. Samples (1 mL) were taken prior to induction and after 4 and 24 hour's incubation post induction.

The time-point samples from each strain were then processed for SDS-PAGE analysis. The cells were lysed using the BugBuster protein extraction (Merck) reagent supplemented with 15 mg/mL Lysozyme (Sigma) and 3.4 U/µL Benzonase Nuclease (Sigma) following the manufacturer's instructions.

Following lysis, the fractions were separated through centrifugation (13, 000 rpm, 10 minutes). The soluble fraction was transferred to a new tube and the insoluble fraction re-suspended in the same volume of water as the volume of lysis buffer previously used.

2.6.2 SDS PAGE analysis

Samples were prepared for SDS-PAGE analysis through mixing 20 µl of each fraction to be analysed with 80 µl SDS-Sample buffer (25 µl 4x Bolt LDS Sample Buffer (Thermo Fisher), 10 µl 9 % (w/v) DTT and 45 µl H_2O). The mixture was heated for 5 minutes at 95°C prior to analysis.

20 µl of each SDS-sample preparation was then loaded on to SDS-PAGE 4-12% Bis-Tris gels to analyse the protein content of the soluble and insoluble fractions

Gel electrophoresis was performed in a dual Bolt mini (Life Technologies) tank using Bolt 4 – 12% Bis-Tris gels in a 1x MES running buffer, Gels were ran at 165 V for 38 minutes with a ladder run on each gel (Novex Sharp Pre-stained Protein Standard, Thermo). Separated proteins were visualised through Coomassie blue staining (Instant Blue, Expedeon).

2.6.3 Protein purification

Protein purification was performed using Ni-NTA spin columns (Qiagen). One gram of cell pellet was resuspended in 3 mL of 50mM HEPES buffer pH 7.5 and cells lysed via sonication (15 seconds on/45 seconds off, 5 minutes total, 45% amplitude). The cell free extract clarified by centrifugation (13, 000 rpm, 20 minutes 4 °C) and

600µl of the soluble fraction loaded onto a Ni-NTA column equilibrated with 600 µl 50 mM HEPES pH 7.5, 300 mM NaCl, 20 mM Imidazole. The column was centrifuged for 5 minutes at 1600 rpm and the flow through collected and stored on ice. The column was washed twice with 600 µl 50 mM HEPES pH 7.5, 300 mM NaCl, 50 mM Imidazole with centrifugation for 2 minutes at 2900 rpm after each wash. The purified protein was eluted from the column in two elution steps using 50 mM HEPES pH 7.5, 300 mM NaCl, 600 mM Imidazole. The process was repeated using a re-equilibrated column loaded with the flow through from step 1.

The four elution fractions were pooled in 12 mL HEPES pH 7.5 before being concentrated to 300 µl using a 15 mL 10 kDa molecular cut off filter (Merck), centrifuged at 3900 rpm, 30 minutes, 4°C. The purified protein was stored as a 10% (v/v) glycerol stock at – 80 °C prior to analysis.

2.6.4 In solution protein digest

For LC-MS analysis cells were pelleted through centrifugation (5, 000 rpm, 15 minutes, 4 °C), the supernatant was discarded and the cell pellet frozen at - 20 °C prior to analysis. Cells were resuspended in Tris-HCl buffer (0.4 M, pH 7.8) and lysed via sonication as previously described.

The cell lysate was clarified through centrifugation (13, 000 rpm, 30 minutes, 4 °C) and the supernatant used for protein precipitation. To a starting volume of 200 µl, 400 µl methanol was added and the sample mixed via vortexing. To this solution 100 µl chloroform was added and the sample mixed through vortexing. The samples were then centrifuged at 13, 000 rpm for two minutes and the top aqueous layer removed through pipetting. 400 µl of methanol was then added and the sample vortexed. The sample was centrifuged (13, 00 rpm, 3 minutes) and as the methanol removed through pipetting without disturbing the protein pellet. The pellet was then air dried for 15 minutes.

Protein pellets were re-suspended in 100 µl 6M urea through vortexing and sonication (Sonic bath 2 minutes). Protein concentrations were determined using the

Qubit fluorometer protein assay kit (Thermo) and normalised (~1 mg total protein in 100 µl). 5 µl of DTT (200 mM) was added to each sample to reduce disulphide bonds and samples were incubated at room temperature for 60 minutes. Alkylation was performed through the addition of 20µl iodoacetamide (200 mM), samples were mixed through vortexing and incubated in the dark at room temperature for 1 hour. The urea concentration was then diluted through the addition of 775 µl LC-MS grade H_2O. Sequencing grade trypsin (Pierce) prepared in ice cold LC-MS grade H2O at a final concentration of 0.2 µg/mL) was added at a ratio of 1:50 and samples incubated overnight at 37 °C. The digestion reaction was quenched through the addition of acetic acid to lower the pH below 6.

Prior to analysis peptides were purified using c18 SEP-Pak columns (Waters). Columns were flushed with 5 mL solution B (35% acetonitrile, 65% H_2O, 0.1 formic acid) followed by 10 mL solution A (2% acetonitrile, 98% H_2O, 0.1% formic acid). Peptide was then added and allowed to settle into the column. 10 mL of solution A used to wash peptides prior to elution using 2 x 1mL solution B. Purified peptides concentrated 20 x in a speed-vac.

2.7 Reverse glyoxylate shunt library construction and screening

E. coli Top10 was used as the host for all vector construction and BW25113 (ΔgltA ΔprpC) was used as a host to screen the vector library. All plasmids used in this study were prepared using the inABLE™ DNA assembly methodology. The MCL and MTK genes from *Rhodobacter sphaeroides* (B9KLE8) and *Methylococcus capsulatus str. Bath* (Q607L8/9) were codon optimised for *E. coli* expression whilst the WT *E. coli* ICL coding sequence was utilised. Genes were synthesised and sub-cloned as fragments compatible with the inABLE™ platform by DNA2.0.

E. coli strains used for vector construction and inABLE optimisation were grown in LB media supplemented with the appropriate antibiotic. Strains harbouring rGS constructs were screened in Ingenza minimal media supplemented with glucose (10 g L^{-1}), succinate (10 mM), kanamycin 50 µg/mL and select agar if required. For liquid

phase assays starter cultures were grown in LB media overnight at 37 °C in rotary shakers at 250 rpm. Fresh cultures containing Ingenza minimal media supplemented with glucose (10 g L^{-1}), succinate (10 mM) and kanamycin 50 µg/mL were inoculated to an OD of 0.1 using washed cells. OD measurements were conducted using a HACH spectrometer operating at 600 nm with measurements taken over eighty hours. Samples for proteomics were collected during mid log phase. For solid phase screening cells from starter cultures were washed using RO H_2O and diluted to allow plating of ~500 cells. Plates were incubated at 37°C in a static incubator.

2.8 Glucose oxidase peroxidase assay

Strains to be analysed were grown for 24 hours in YNB-Glucose media at 30°C, 250 rpm. OD_{600} of overnight cultures was measured and cells pelleted through centrifugation, 3000 rpm for 5 minutes. The resulting supernatants (40 µl) were incubated for 1 hour with 40 µl of substrate (1 mM substrate prepared in 100 mM NaOAc pH 4.5) in a 96-well microtiter plate at 37 °C. The reaction was the stopped through the addition of 20 µl 1M hydrochloric acid. 20 µl of 1M Tris solution was then added to each well followed by the addition of 80 µl assay mixture to detect glucose release. The assay mixture was comprised of 0.5 mM ABTS, 20 U HRP and 40 U glucose oxidase. The microtiter plates were sealed and incubated at 37 °C for twenty minutes before the absorbance at 420 nm was measured. A background level of glucose was determined for each supernatant following the same protocol but replacing the substrate with 100 mM NaOAc pH 4.5. Activity normalised based on OD_{600} readings of overnight cultures.

2.9 LC-MS based proteomics analysis

2.9.1 Reagents

Tryptically digested MassPrep digestion standard 1 and Leucine Enkephalin purchased from Waters. Trypsin-digested BSA MS Standard (CAM-modified) purchased from New England Biolabs. Mass spectrometry grade solvents were purchased from Fischer Chemical.

2.9.2 Sample preparation

Lyophilised MassPrep Standard mix 1 is an equimolar mix of four tryptically digested proteins (50 pmol of each). To prepare the standard for analysis the mix was reconstituted through the addition of 100 µl 0.1% formic acid, and vortexing until all material was completely re-suspended.

Lyophilised trypsin-digested BSA standard from NEB was reconstituted in 250 µl 0.1% formic acid through vortexing until all material was re-suspended giving a final concentration of 2 µM.

Leucine Enkephalin was used for lockmass correction at a concentration of 800 pg/µL in 75:25 H2O:ACN + 0.1% FA.

2.9.3 Liquid chromatography

All UPLC separations were performed on an Acquity UPLC system using a BEH C18 Column, 130Å, 1.7 µm, 1 mm X 100 mm analytical column (Waters Corporation). Solvent A was composed of 0.1% formic acid in water and Solvent B 0.1% formic acid in acetonitrile. Sample elution was standardly performed at a flow rate of 90 µl/minute by increasing the organic solvent from 5% to 40% over 25 minutes. For analysis of cell lysates the length of LC gradient was increased to 60 and 120 minutes. The Lockspray was infused throughout the run at a standard rate of 5 µL/min and sampled for 0.3 seconds every 30 seconds.

2.9.4 Mass spectrometer configuration

A Synapt G2 Q-TOF mass spectrometer controlled through MassLynx Version 4.1 (Waters Corporation) was used for all analysis in this study. The instrument was calibrated using a MS spectrum generated from sodium formate from *m/z* 50 - 1500

All subsequent acquisitions were lockmass corrected post acquisition using single point Leucine Enkephalin reference compound (*m/z* 556.2771). The reference compound was delivered into the mass spectrometer through the LockSpray interface. Acquisition occurred every 30 seconds for 0.3 seconds.

MS^E data was collected over the *m/z* range of using a method created the MS method editor in Mass Lynx. In the low energy scan a constant trap collision energy of 4 V was applied whilst in elevated energy mode the trap collision energy was ramped from 15 V to 40 V whilst the transfer energy was maintained at 2 V for low and high energy scans.

2.9.5 Processing of MS^E data and database interrogation

MS^E data for protein identification was processed using PLGS version 3.03. Data processing was performed using the parameters found in Table 8 and 9.

2.9.5.1 Processing parameters

Attribute	Value
Chromatographic Peak Width	Automatic
MS TOF Resolution	Automatic
Lock Mass for Charge 1	556.2771 Da/e
Lock Mass for Charge 2	N/A
Lock Mass Window	0.25 Da
Low Energy Threshold	135.0 counts
High Energy Threshold	30.0 counts
Elution Start Time	Not specified
Elution End Time	Not specified

Table 8: PLGS processing parameters.

2.9.5.2 Workflow parameters

Attribute	Value
Search Engine Type	PLGS
Databank	Variable
Taxonomy	N/A
Peptide Tolerance	10
Fragment Tolerance	10
Min Fragment Ion Matches per Peptide	3
Min Fragment Ion Matches per Protein	7
Min Peptide Matches per Protein	1
Maximum Protein Mass	250000
Primary Digest Reagent	Trypsin
Secondary Digest Reagent	None
Missed Cleavages	1
Fixed Modifier Reagents	Carbamidomethyl C
Fixed Modifier Reagent Groups	N/A
Variable Modifier Reagents	Oxidation M
False Discovery Rate	4

Table 9: PLGS Workflow parameters.

Processed data was used to interrogate either the UniProt KB database or an *E. coli* database downloaded from UniProt and supplemented with rGS protein sequences.

3. Adaptation of inABLE™ DNA assembly for biosynthetic pathway optimisation

3.1 Identification of inABLE™ DNA assembly limitations

3.1.1 Introduction

The engineering of microbes for industrial bioprocesses often relies on the introduction of potentially complex multi-step heterologous biosynthetic pathways in the microbe of choice as a first step. Traditionally the construction of the DNA vectors to achieve this would be through multiple rounds of digestion and ligation based molecular cloning however in recent years a rise in available DNA assembly techniques has accelerated the construction of these pathways and opened the door to the combinatorial construction of such pathways. The ability to mix and match coding sequences, promoters, terminators, ribosome binding sites and vector sequences allows a metabolic engineer to explore a much larger design space in a single experiment than was previously feasible. This increases the likelihood of identifying an optimal gene combination and expression balance to maximise flux through the heterologous pathway and has the potential to greatly accelerate the engineering of microbes for bio-based processes.

Ingenza currently uses a combinatorial genetic assembly technology (known commercially as inABLE®) for the efficient and selective assembly of complex DNA expression vectors encoding, for example, the enzymes responsible for multiple biosynthetic transformations within a pathway to produce chemical targets. Notwithstanding the advantages of this technique, it is also clear that with suitable innovation, this technology could be even more powerful than it is in its current form. This chapter focusses on the adaptation and optimisation of this technology into a robust platform for the combinatorial assembly of multicomponent pathways. In order to achieve this a systematic review of the current technology has been

undertaken and a number of areas which can limit the efficiency of the process identified.

The inABLE® technology can be split into three stages, part design and preparation, part linker fusion and the assembly reaction (described in detail in Section 1.6.1.6). The first step of the procedure involves the identification of the optimum splits within each part generating a truncated part and the linker and part oligonucleotides. Part linker fusions are then prepared firstly digesting the truncated part with SapI generating 3 nucleotide overhangs followed by ligating each truncated part with its part oligonucleotide at the 5'end and the linker oligonucleotide of the following truncated part at the 3'end, a process which generates sixteen nucleotide complementary overhangs between parts. The parts are then combined and assembled through these overhangs in a one pot reaction, without the addition of ligase. Each stage within the process is amenable to optimisation to improve the overall assembly efficiency. Three key areas contributing to assembly efficiency have been explored in this chapter, the effect of part number, the effect of homology with 3 nucleotide overhangs and the effect of secondary structure in 16 nucleotide overhangs.

3.1.2 Effect of part number on assembly efficiency

3.1.2.1 Introduction

With increasing pathway complexity, the number of fragments of DNA that require to be assembled also increases. Whilst standard digestion/ligation techniques which utilise overhangs of - on average - 3 nucleotides allow the annealing of up to three fragments of DNA, techniques such as inABLE DNA assembly utilise 16 nucleotide overhangs, increasing the number of fragments which can be assembled. Despite the ability to assemble a larger number of parts the efficiency of the assembly reaction will decrease as part number increases. For example, a recent study characterising Gibson assembly found that - in the authors hands - when an assembly of more than

four DNA parts was attempted to generate constructs totalling 4.8 kb below 50% of the isolates screened were correct[148]. To test the effect part number has on assembly efficiency, a study was designed to test assemblies containing between 2 and 10 parts, which is thought to be the upper limit of inABLE assembly process. The output of this experiment will guide the choice between the standard and alternative inABLE based approaches in future vector construction programs. This may be due to the complexity of the vector to be constructed or when a high assembly efficiency is required such as when constructing a combinatorial library.

3.1.2.2 Results

To explore the effect of part number on assembly efficiency an experiment was designed in which assembly efficiency could be directly calculated through counting of colonies on a selection plate without the requirement for PCR or restriction digestion based characterisation. An initial vector was constructed (Figure 15) which consisted of Ampicillin (Amp), Kanamycin (Kan), Chloramphenicol (Camp) and Tetracycline (Tet) resistance markers along with a pMB1 origin of replication. This vector provided the template DNA for subsequent truncated part design and amplification and also acted as a positive control for all subsequent transformations. Through selection of transformants on LB plates supplemented with Kan, Amp, Camp and Tet it is possible to directly screen for isolated containing the expected assembly product and therefore directly calculate assembly efficiency.

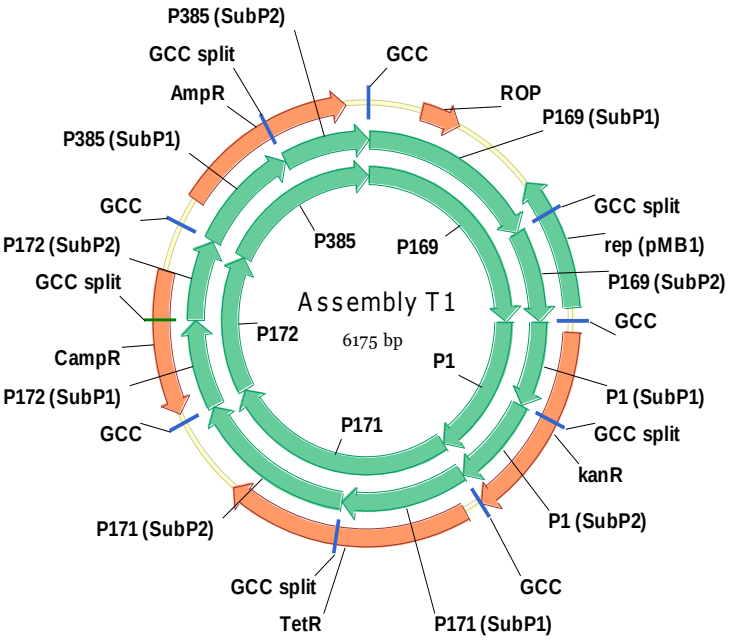


Figure 15: Map of assembly T1 which was used to determine the effect of part number on assembly efficiency. The construct is comprised of four genes which confer resistance to four different antibiotics (kanamycin, tetracycline, chloramphenicol and ampicillin) and a pMB1 E. coli origin of replication. Strains which contain correctly assembled construct can be selected for on agar containing the four antibiotics GCC sequences annotated on the map identify the introduced scar sequence and define the boundary of the parts assembled.

The parts required to construct Assembly T1 had previously been designed and truncated parts, POA and LOA samples were available from the inABLE inventory. In order to test assemblies containing between six and ten parts, additional truncated parts were designed by splitting the initial five parts used in Assembly T1 in two sub-parts. This ensured that regardless of the number of parts being assembled that the size of the final construct remained constant, which is important as increasing construct size is known to have a negative effect on transformation efficiency[149]. As described previously (Section 1.6.1.6) the inABLE procedure results in the introduction of a 3 bp scar (gcc) between parts. The introduction of this scar sequence is not feasible in the middle of one of the antibiotic resistance marker coding sequences as required for the tests involving 6 to 10 parts. To avoid the introduction of a scar within the markers the junction between sub-fragments within each part was defined as a codon encoding an alanine residue which could be substituted by the gcc scar resulting in a scarless assembly.

The ability to utilise the inABLE technique in a scarless manner potentially expands the scope of the approach to include techniques where splicing within a coding sequence is required such as combinatorial polyketide synthase construction[150]. A specific gcc fragment was selected within the pMB1 origin of replication (P169) for the sub-fragment split. Once the split points had been defined the sequences were used as input sequences for the part designer software to design the corresponding truncated parts, the linker oligonucleotides, the part oligonucleotides and PCR primers for the amplification of the truncated parts flanked by EarI/SapI sites (Section 2.5.1).

PCR reaction	Truncated part	Forward primer	Reverse primer	Expected product (bp)	Description
1	169	PCRf169	PCRr169	1487	pMB1 origin of replication
2	1	PCRf1	PCRr1	920	Kanamycin resistance marker
3	171	PCRf171	PCRr171	1592	Tetracycline resistance marker
4	172	PCRf172	PCRr172	876	Chloramphenicol resistance marker (CampR)
5	385	PCRf385	PCRr385	1045	Ampicillin resistance marker
6	A (1/171/172/385)	PCRf1	PCRr385	4588	KanR/TetR/CampR/AmpR
7	B (171/172/385)	PCRf171	PCRr385	3617	TetR/CampR/AmpR
8	C (172/385)	PCRf172	PCRr385	1975	CampR/AmpR
9	P169 (SubP1)	PCRf169	PCRr169 (SubP1)	982	5' section of pMB1 ori
10	P169 (SubP2)	PCRf169 (SubP2)	PCRr169	455	3' section of pMB1 ori
11	P1 (SubP1)	PCRf1	PCRr1 (SubP1)	416	5' section of KanR
12	P1 (SubP2)	PCRf1 (SubP2)	PCRr1	455	3' section of KanR
13	P171 (SubP1)	PCRf171	PCRr171 (SubP1)	673	5' section of TetR
14	P171 (SubP2)	PCRf171 (SubP2)	PCRr171	873	3' section of TetR
15	P172 (SubP1)	PCRf172	PCRr172 (SubP1)	435	5' section of CampR
16	P172 (SubP2)	PCRf172 (SubP2)	PCRr172	392	3' section of CampR
17	P385 (SubP1)	PCRf385	PCRr385 (SubP1)	557	5' section of AmpR
18	P385 (SubP2)	PCRf385 (SubP2)	PCRr385	440	3' section of AmpR

Table 10: PCR reactions to amplify the 18 truncated parts required for exploration of the effect of part number on assembly efficiency.

The eighteen truncated parts (Table 10) were amplified via PCR (Figures 16 and 17) and the corresponding part and linker oligonucleotides phosphorylated and annealed (Section 2.5.2).

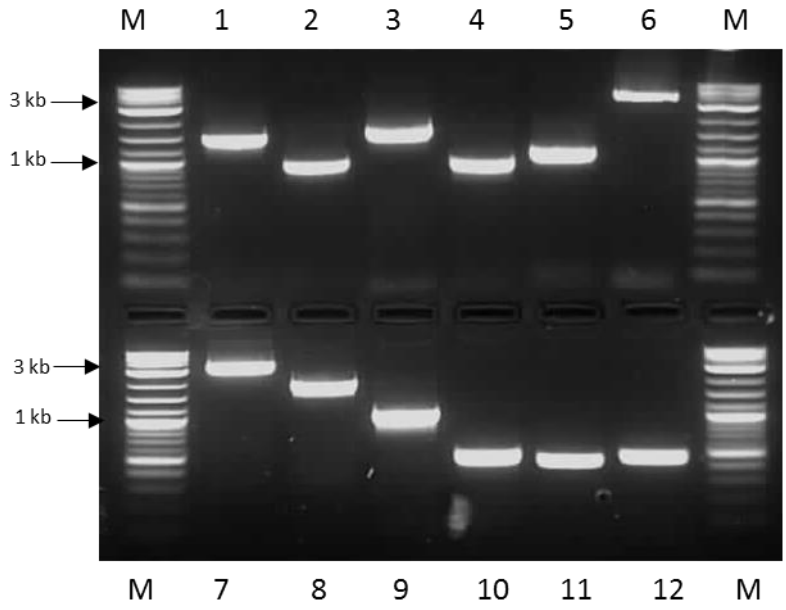


Figure 16: PCR amplification of truncated parts 1-12. Numbering and expected products outlined in Table 10.

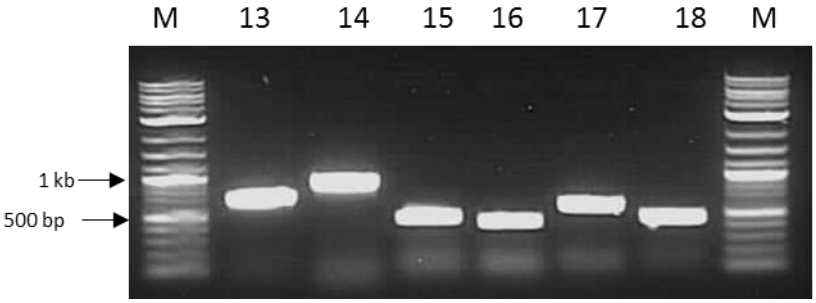


Figure 17: PCR amplification of truncated parts 13-18. Numbering and expected products outlined in Table 10.

The generated PCR products were used to generate the corresponding part linker fusions through ligation of the appropriate POA and LOA following the standard procedure (Section 2.5.4). Following purification of the part/linker fusion assembly reactions were performed in triplicate with the concentration of each truncated part in the reaction maintained at 2 nM. Following the completion of the assembly reaction

the product was used to transform *E. coli* and isolates harbouring the expected assembly product were selected for on LB-Kan, Amp, Camp, Tet plates. In parallel to each transformation reaction the assembly product (Assembly T1) was transformed at the final molar concentration (2 nM) expected for completion of the assembly reaction. Colonies were counted, and assembly efficiency calculated as a percentage of the number of transformants obtained using the positive control assembly, T1 (Table 11 and Figure 18).

	Average colony count	SD	Average assembly efficiency	SD
2 Part Assembly	1001	38.947	0.450	0.018
3 Part Assembly	324	69.415	0.146	0.031
4 Part Assembly	131	8.717	0.059	0.004
5 Part Assembly	40	1.571	0.018	0.001
6 Part Assembly	9	1.700	0.009	0.002
7 Part Assembly	3	1.247	0.003	0.001
8 Part Assembly	6	1.247	0.002	0.000
9 Part Assembly	0	0.000	0.000	0.000
10 Part Assembly	0	0.000	0.000	0.000

Table 11: Colony counts and assembly efficiencies for assembly reactions containing between 2 and 10 parts using the standard inABLE procedure. The average colony count and assembly efficiency presented are the result of experiments performed in triplicate. The assembly efficiency was calculated as a percentage of the number of transformants obtained using the positive control assembly.

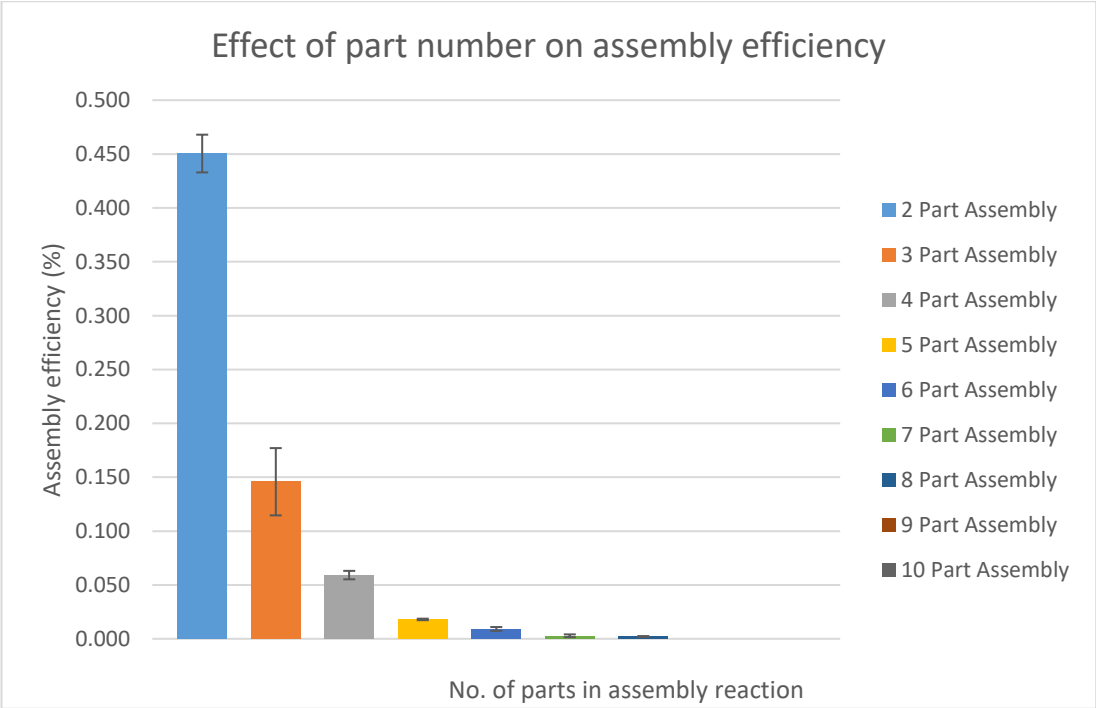


Figure 18: The effect of the number of DNA parts (between 2 and 10) being assembled on assembly efficiency.

3.1.2.3 Conclusions

A key advantage of the number of DNA assembly techniques which have been described in the past 10 years is the ability to assemble a higher number of parts than standard digestion/ligation approaches which are limited to 3 parts. However, this study highlights the considerable decrease in assembly efficiency as part number increases. This reinforces the requirement to move towards an alternative approach when vectors comprised of 5 or more parts or combinatorial libraries require to be constructed. One disadvantage of the strategy employed in this study is that the number of incorrect assemblies generated, and thus assembly accuracy, cannot be determined (as any colonies containing incorrect assemblies will not be viable on the selection plates). To explore the effect of part number on assembly accuracy the construction of a vector which results in a colorimetric response, such as the violacein pathway[151] or GFP expression could be utilised.

3.1.3 Effect of three base pair homology on assembly efficiency

3.1.3.1 Introduction

The second stage (part/linker fusion) of the inABLE process relies on efficient ligation of part and linker oligonucleotides to truncated parts during the part/linker fusion reaction. It is this reaction that generates the complementary 16 nucleotide overhangs between parts which are vital in the assembly reaction and as such any reaction which inhibits this ligation will impact the overall assembly efficiency. One such reaction is the potential for self-ligation of the truncated part through three nucleotide overhangs present at the 5' and 3' end of the truncated part after SapI/EarI digestion (Figure 19).

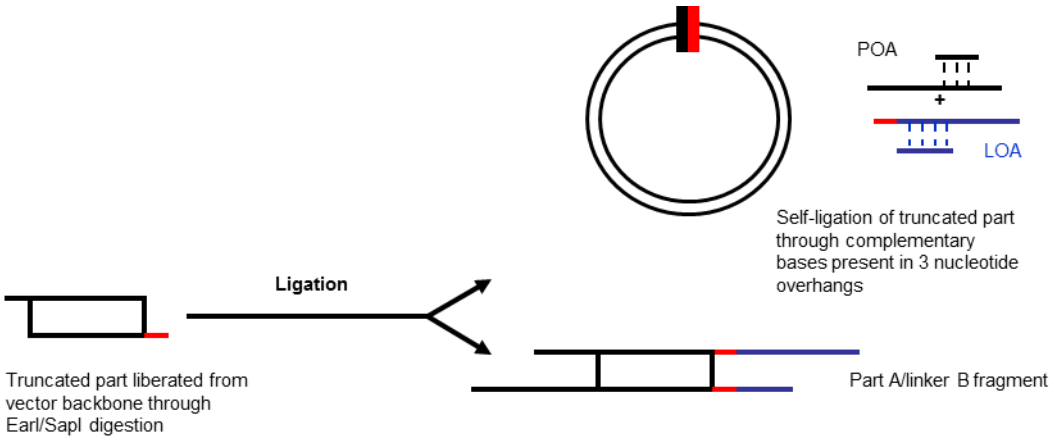


Figure 19: Self ligation of a truncated part through 3 nucleotide overhangs in the part/linker fusion reaction The formation of this product has the potential to decrease assembly efficiency as the self-ligated product cannot be re-digested by EarI/SapI and cannot ligate the POA or LOA fragments which are required for DNA assembly.

The possibility for this reaction is due to the nature of SapI and EarI as type IIs restriction enzymes. Type IIs restriction enzymes recognise asymmetric DNA sequences and cleave out with their recognition sites. SapI and EarI cleave in an n + 1, n + 4 manner which results in a 3 nucleotide overhang (Figure 20). This characteristic is vital as it firstly generates three nucleotide overhangs to which the

POA and LOA are ligated and secondly means the product of the part/linker fusion lacks recognition sites allowing for product enrichment through cycles of digestion and ligation.

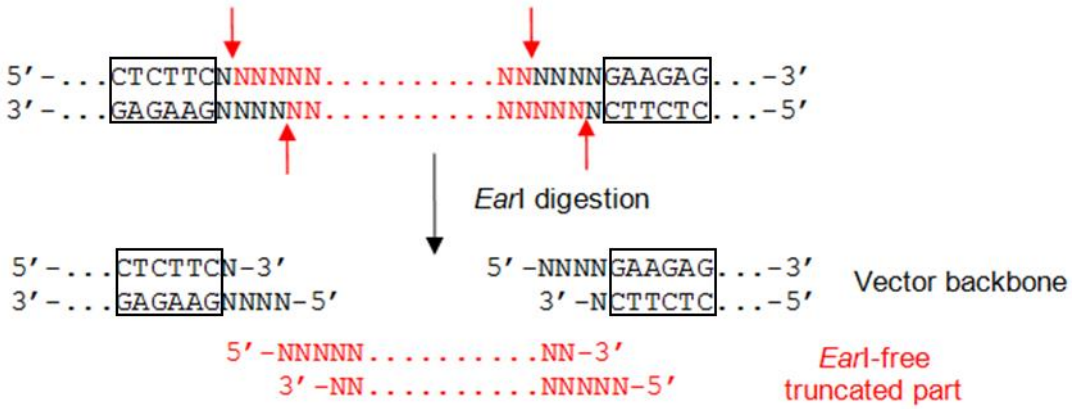


Figure 20: Type IIs restriction endonuclease cleave out with their recognition sites. EarI (and SapI) cuts n + 1, n + 4 resulting in a 3 nucleotide overhang. The EarI recognition site is boxed and the cleavage site indicated by red arrows.

The overhang at the 3' end of the truncated part is specified as being complementary to gcc. The maintenance of this sequence as an overhang means that any linker, which are standardised to contain a 5' gcc overhang, can be annealed to any truncated part providing the flexibility that makes the technique suitable for DNA assembly. The 5' overhang however is defined by the sequence of the full length part and as the part design software does not take complementary bases within the 3 nucleotide overhangs into account when defining the split point between the truncated part and the linkers, complementary bases between the upstream and downstream overhang can be generated following EarI/SapI digestion. The effect of complementary base pairs within these overhangs was explored to determine if this is a parameter which requires to be considered when looking to maximise assembly efficiency.

3.1.3.2 Results

To explore the effect of self-ligation of the truncated part during part linker fusion on assembly efficiency a two part assembly was utilised (Figure 21). The combination of one part containing a pMB1 origin of replication and second part containing a Tetracycline resistance marker allows for the simple calculation of assembly efficiency through selection of transformants on tetracycline containing media as only clones containing the correctly assembled product will be able to propagate.

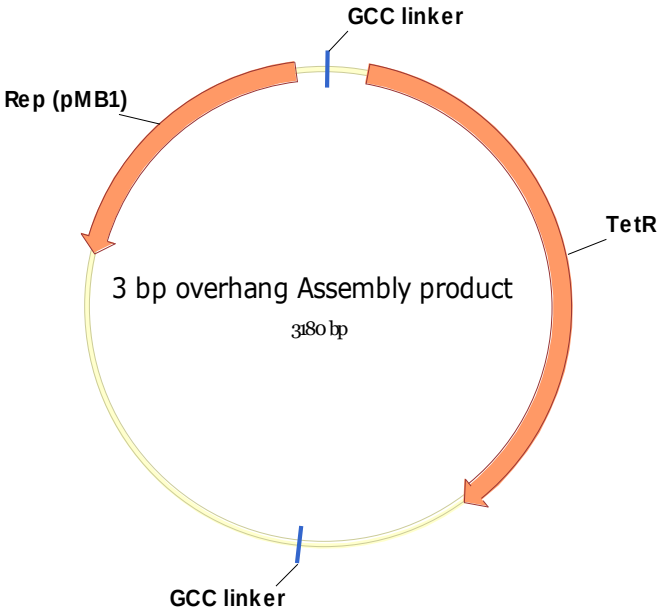


Figure 21: Expected assembly product used to explore the effect of complementary nucleotides within 3 base overhang on assembly efficiency. The construct comprises of a gene which confers tetracycline resistance (TetR) and an E. coli origin of replication (pMB1).

Primers were designed for amplification of the TetR truncated part generating varying 3 bp overhangs following SapI digestion. As the inABLE workflow specifies a consistent 3' overhang this remained constant and the 5' overhang was modified. Six overhangs generating between three and zero bases of compatibility and also the position of the complementary bases within the overhang were explored (Table 12).

	xxx	Overhang compatibility	Matches
A	gtg	gcc xx gtg	1 external match
B	cgg	gcc ''' cgg	3 matches
C	ggg	gcc x'' ggg	2 matches (1 external, 1 internal)
D	gta	gcc xxx gta	0 matches
E	tga	gcc x'x gga	1 internal match
F	ctg	gcc 'x' ctg	2 external matches

Table 12: Position and number of complementary bases flanking truncated part following SapI digestion.

Truncated parts were amplified through PCR, using alternative forward primers to introduce the desired 5' overhang and compatible appropriate POA and LOA prepared for each reaction. The amount of each part was maintained at 2 nM in each assembly reaction. A positive control reaction was performed in which the expected assembly product was diluted to 2 nM. All assembly reactions and subsequent transformations were prepared in triplicate as previously described and assembly efficiencies were calculated as a percentage of the number of transformants obtained from the positive control reaction (Table 13 and Figure 22).

Assembly	No. of CFU	Average	Assembly efficiency (%)	Average assembly efficiency (%)	Error
Assembly B (3 matches)	39	41	0.021	0.022	0.005
	52		0.028		
	31		0.016		
Assembly C (2 matches 1 internal, 1 external)	1440	1448	0.765	0.769	0.068
	1608		0.854		
	1296		0.688		
Assembly F (2 external matches)	2006	1642	1.066	0.872	0.151
	1608		0.854		
	1312		0.697		
Assembly A (1 external match)	3912	3819	2.078	2.028	0.115
	4024		2.137		
	3520		1.870		
Assembly E (1 internal match)	2968	2635	1.576	1.399	0.142
	2312		1.228		
	2624		1.394		
Assembly D (0 matches)	13600	12460	7.224	6.618	0.613
	12900		6.852		
	10880		5.779		

Table 13: The effect of complementary bases within three base overhangs following EarI/SapI digestion. Assembly efficiencies were calculated as a percentage of the number of transformants obtained from the positive control reaction.

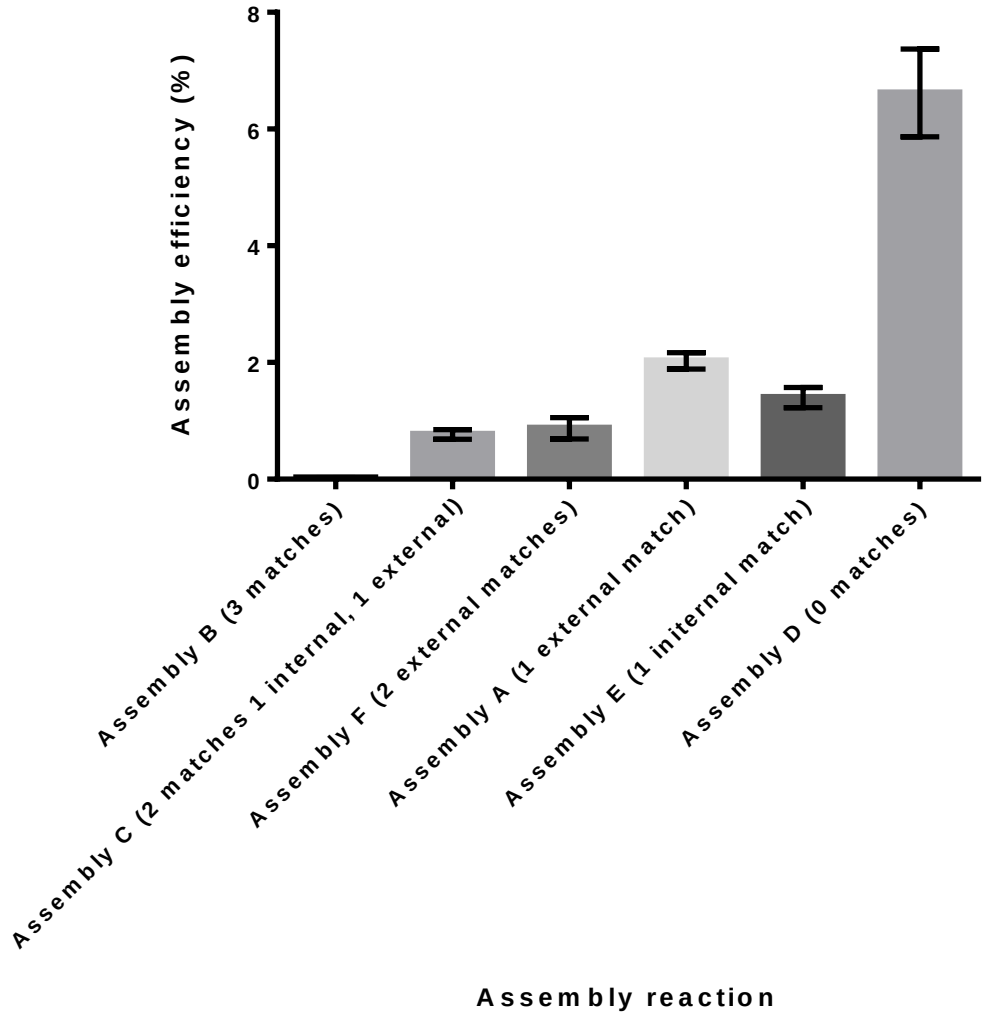


Figure 22: The effect of complementary bases within three base overhangs following EarI/SapI digestion. The assembly efficiency of the six reactions explored was calculated from triplicate experiments.

3.1.3.3 Conclusions

This study highlights the pronounced effect the presence of complementary bases within the 3 nucleotide overhangs has on the assembly efficiency. It was found that the presence of only one complementary base within these overhangs resulted in a decrease in assembly efficiency of 70% whilst the presence of three homologous bases resulted in a 99% decrease in the efficiency. It is therefore vital for robust and efficient assembly that complementary bases within these overhangs is avoided. To achieve this

the number of splits defined by the part design software can be increased allowing for identification of a split point that results in zero complementary bases following SapI digest – either via manual analysis or software modification - maximising assembly efficiency.

Interestingly the assembly efficiency in this experiment is considerably higher than seen for two parts in the previous part number experiment (0.45% versus 7% in this experiment with no complementary bases). As the assembly product was transformed in parallel in each experiment then this difference in efficiency cannot be attributed to the increased vector size or factors affecting *E. coli* transformation such as cell competency or transformation protocol. Analysis of the overhangs in the two part assembly revealed one complementary base in the overhangs generated following digestion of TPA (Table 10) with SapI. This is likely to be partly responsible for the lower efficiency observed in the part number experiment and further highlights the requirement to avoid complementary bases when looking to maximise assembly efficiency.

3.1.4 Effect of linker secondary structure on assembly efficiency

3.1.4.1 Introduction

During the final stage of the inABLE procedure (the assembly reaction) purified part linker fusions are mixed together and assembled through complementary 16 nucleotide single stranded overhangs (Figure 23). These overhangs are generated due to the ligation of appropriate part and linker oligonucleotides to adjoining truncated parts. One parameter which has not previously been considered is the presence of secondary DNA structure within the 16 nucleotide overhangs. The potential for the formation of secondary structures within these overhangs would likely prevent the efficient assembly of fragments. This is not currently taken into consideration when defining the split point between the truncated part and linkers. The effect of secondary structure within these overhangs was explored through alteration of the

overhang DNA sequence to manipulate the potential for secondary structure formation.

Figure 23: DNA assembly through POA and LOA annealing. Area in yellow highlights the 16 nucleotide complementary overhangs. Bases comprising the LOA in red and POA in blue.

3.1.4.2 Results

Experimental design allowed for assembly efficiency to be directly calculated through counting the number of colonies on antibiotic containing agar plates. On this occasion a KanR marker cassette and pMBI origin of replication were assembled. Since the linker sequence requires to be modified in order to modulate the strength of predicted secondary structure the KanR fragment available through the inABLE database could not be used due to the promoter being included in the linker. Modification of the linker therefore could have an impact on the transcription of the aminoglycoside 3'-phosphotransferase gene which confers kanamycin resistance. It is therefore possible that the modulation of the linker sequence may impair the promoter function and as a result kanamycin resistance in clones containing the expected vector. To circumvent this potential issue a new KanR fragment was cloned in which forty seven base pairs were added to the 5' end of the KanR cassette to allow for LOA modification without altering the promoter or 5' UTR and affecting transcription/translation.

Eight LOl and the complementary POl sequences were designed to generate secondary structures within the 16 nucleotide single stranded overhang ranging in predicted strength from very strong ΔG (-7.47) to no predicted secondary structure and a ΔG value of 2.04 (Table 14).

LOA overhang sequence	GC %	Predicted secondary structure	ΔG (kcal. mole^{-1})
GAACGGTCTGCGTTGT	56.2	Moderate	-3.2
CAACGGTCTGCGTTGT	56.2	Moderate	-4.22
GAACGCTCTGCGTTGT	56.2	Strong	-4.05
CAACGCTCTGCGTTGT	56.2	Strong	-5.06
GAAGGGTCTGCGTTGT	56.2	None	1.02
GAAGGGTGTGCGTTGT	56.2	None	2.04
CAGCGCTCTGCGCTGT	68.8	Very strong	-6.15
CCGCGCTCTGCGCGGT	81.2	Very strong	-7.47

Table 14: Modification of 16 nucleotide overhang sequence to modulate predicted secondary structure strength.

Truncated parts were amplified through PCR and the appropriate POA and LOA prepared for each reaction. The amount of each part was maintained as 2 nM in each assembly reaction. A positive control reaction was performed in which the expected assembly product was diluted to 2 nM. All assembly reactions and subsequent transformations were prepared in triplicate as previously described. Assembly efficiencies were calculated as a percentage of the number of transformants obtained from the positive control reaction (Table 15 and Figure 24).

Assembly	No. of CFU	Average	Assembly efficiency (%)	Average assembly efficiency (%)	Error
Assembly H Very Strong (ΔG = -7.47)	200	230	0.136	0.157	0.024
	210		0.143		
	280		0.191		
Assembly G Very strong (ΔG = -6.15)	670	520	0.457	0.355	0.073
	430		0.293		
	460		0.314		
Assembly D Strong (ΔG = -5.06)	2660	2500	1.814	1.705	0.184
	2120		1.446		
	2720		1.855		
Assembly C Strong (ΔG = -4.05)	9790	9707	6.678	6.621	0.500
	10560		7.203		
	8770		5.982		
Assembly B Moderate (ΔG = -4.22)	12800	12120	8.731	8.267	0.350
	11560		7.885		
	12000		8.186		
Assembly A Moderate (ΔG = -3.2)	13120	11613	8.950	7.922	0.909
	9880		6.739		
	11840		8.076		
Assembly E None (ΔG = 1.02)	13040	12937	8.895	8.824	0.188
	12560		8.568		
	13210		9.011		
Assembly F None (ΔG = 2.04)	14600	15773	9.959	10.759	0.681
	15680		10.696		
	17040		11.623		

Table 15: The effect of the presence of secondary structures within 16 nucleotide overhangs on assembly efficiency. Number of colonies present on selective plates and the resulting assembly efficiencies. Assembly efficiencies were calculated as a percentage of the number of transformants obtained from the positive control reaction

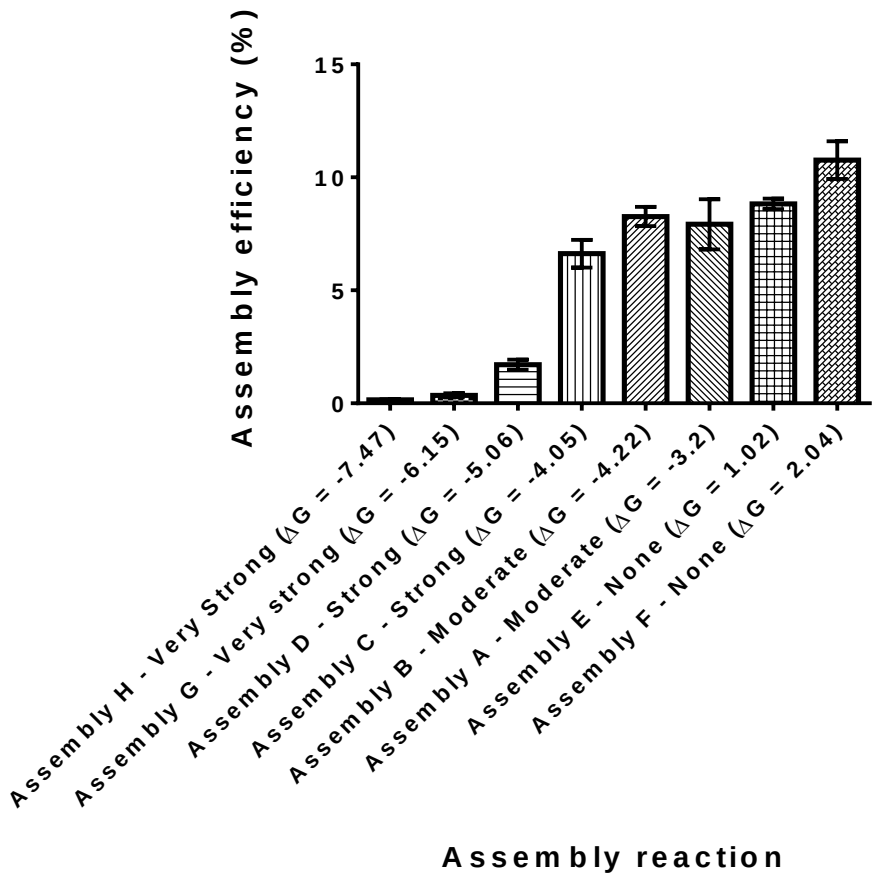


Figure 24: The effect of the presence of secondary structures within 16 nucleotide overhangs on assembly efficiency.

3.1.4.2 Conclusions

The presence of strong secondary structures within linker sequences results in a significant decrease in assembly efficiency. An analysis however of one hundred 16 nucleotide overhangs generated from the LOA sequences currently deposited within the Ingenza inABLE database predicted ΔG values between -2.81 and 2.55 kcal. mole^{-1} (Figure 25).

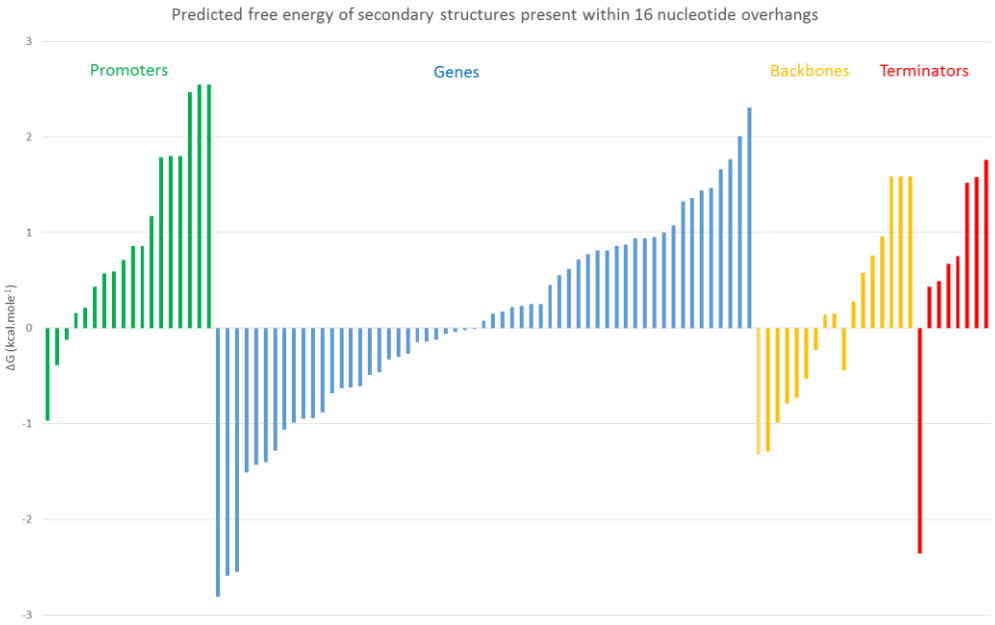


Figure 25: Analysis of predicted secondary structure within 16 nucleotide overhangs from 100 LOA sequences in the inABLE parts database.

This suggests that the decrease in assembly efficiency observed here will not be experienced for any of the linkers characterised. The presence of secondary structure however does have a strong impact on assembly efficiency in more extreme cases ($\Delta G < -4$ kcal. mole^{-1}) and should be explored as a possible cause, when difficult to assemble parts are encountered.

3.1.5 Conclusions

In this study the effect of the number of parts being assembled, the presence of complementary bases within truncated part overhangs following SapI digestion and the presence of secondary structure within 16 nucleotide overhangs on assembly efficiency were explored. On each occasion combinations of fragments comprising origins of replication and antibiotic resistance marker cassettes were utilised to allow for the rapid determination of assembly efficiency through selection of transformants on plates containing appropriate antibiotics. The results highlight that each of the

factors explored can have a pronounced impact on assembly efficiency and should be considered when looking to maximise reaction efficiency. Secondary structure within overhangs and complementary bases flanking the truncated part can both be addressed through careful DNA fragment design however the number of parts being assembled is dependent on the construct being assembled. In order to address this, the development of an alternative inABLE workflow was initiated.

3.2 Development and implementation of nested inABLE™

3.2.1 Introduction

The key advantage gained from using the inABLE® platform is the selectivity it offers in assembling DNA parts, permitting assembly of multiple parts in a single reaction. Traditional methods allow for the efficient ligation of 2 (or at most 3) parts – using inABLE® up to 9 DNA parts in a single reaction have previously been assembled. However, in its current form the technology has limitations. Firstly, to assemble a full biosynthetic pathway, combinations of a larger number of genomic regions (>10 parts) are desirable. Since the current procedure leads to a construct which is no longer compatible with the technology (lacking EarI sites), complex assemblies harbouring more than ten parts cannot be generated using the technique. Secondly a powerful use of this technology is the ability to perform combinatorial assemblies, for example the combination of a gene with multiple regulatory regions (promoters, terminators, ribosome binding sites), however as described in Section 3.1.2, as part number increases assembly efficiency decreases making the combinatorial construction of a vector comprised of multiple fragments unpractical.

To address these limitations the development of a “nested” inABLE technology was implemented. The essential innovation in this approach is that an initial DNA vector is prepared in pieces from inABLE parts in a form that allows it to be used in subsequent assemblies – i.e. a convergent rather than a linear synthesis. The key piece of novel enabling technology which forms the basis of this work is the design of DNA

parts suitable for “nested” inABLE. In theory such an approach will be able to yield assemblies of up to 100 parts or ensure that when constructing vector libraries, the final assembly step is as efficient as possible, maximizing library coverage.

As outlined in Section 3.1.2, a primary limitation of the inABLE technology is that as part number increases the assembly efficiency rapidly decreases. In response to this issue a step-wise approach in which complex assemblies could be split into smaller sub-assemblies followed by a final two or three part assembly was envisaged. In its current format the product of an inABLE® assembly cannot be used in subsequent assemblies. As described in Section 2.5.4, the part linker fusion reactions involve cycles of digestion and ligation using EarI (recognition site - CTCTTC) and T4 DNA ligase. Although EarI is standardly used truncated parts are designed so that they are flanked by a SapI/EarI recognition site (GCTCTTC) (Figure 26).

Figure 26: Comparison of EarI and SapI recognition and cleavage sites. The ability to alternate between the utilisation of these enzymes is key when adopting a nested assembly approach.

This gives the possibility to use SapI for a first round of reactions and fuse linkers containing the EarI recognition sites to the outmost parts. This would result in the product of the first round of assembly being flanked by EarI recognition sites which can then be used in a subsequent round of assembly (Figure 27).

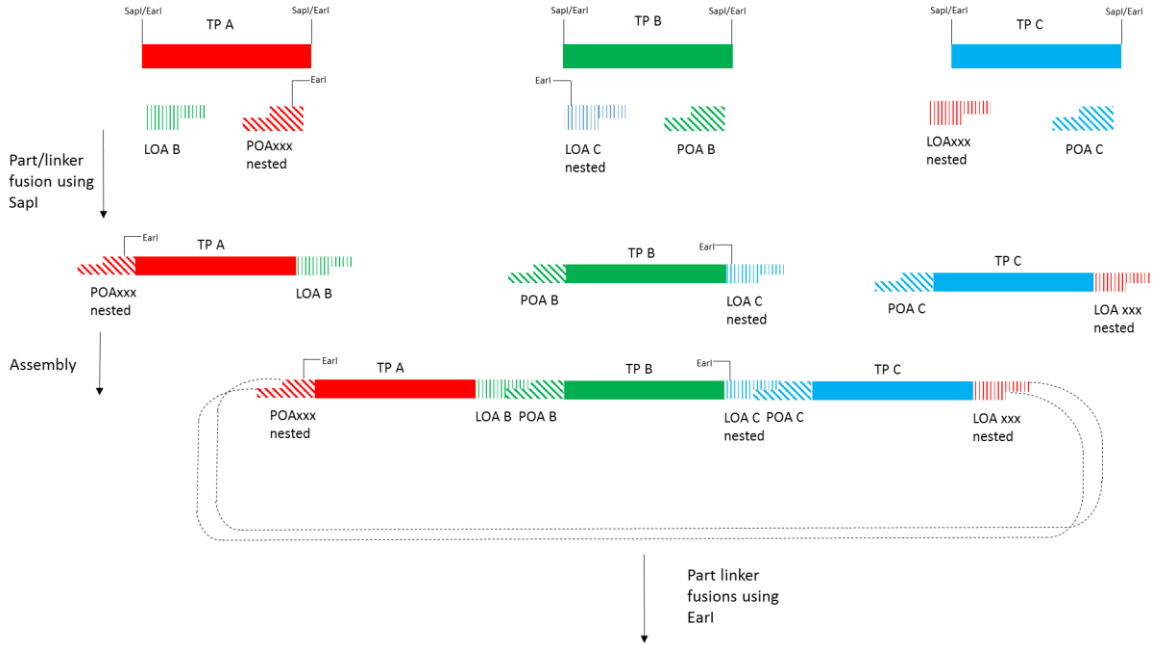


Figure 27: Nested assembly of three DNA parts. Following the first assembly Part A and Part B are flanked by EarI recognition sites allowing for combination with other parts in a subsequent inABLE assembly provided EarI is used in the part linker fusion reaction.

The ability to switch between EarI and SapI for subsequent rounds of part/linker fusions gives the possibility to perform complex assemblies in two steps. The primary advantages of this new approach are the ability to build increasingly complex vectors as assemblies would no longer be limited to 10 parts and the efficient construction of multipart vector libraries through the use of a final step two or three part assembly. To explore these potential benefits, the construction of a combinatorial library and a complex vector (>7 parts) were initiated.

3.2.2 Construction of a multi-part vector using a nested inABLE approach

3.2.2.1 Introduction

To explore the ability to use the nested inABLE approach in order to construct complex vectors the construction of *Saccharomyces cerevisiae* vectors for the

liberation of glucose from cellulose was initiated. Cellulose is a structural polysaccharide found in plants which is the most abundant organic polymer in the biosphere[152]. It is comprised of β-1,4 linked D-glucose monomers organised into ordered crystalline domains and disordered amorphous domains. Efficient degradation of this feedstock is thought to require the synergistic action of a minimum of three enzyme classes, endoglucanases which cleave β-1,4-glucosidic bonds at internal sites within the cellulose fibre, exoglucanases which cleave cellobiose from the exposed ends in a progressive manner and β-glucosidases which liberate glucose from cellobiose[153, 154] (Figure 28). The addition of blends of these enzymes to cellulosic feedstocks is not cost efficient for large scale biomass processing. Ideally a consolidated bioprocessing approach in which the production strain produces its own biomass degrading enzymes, assimilates the liberated sugars and produces the chemical of interest would be of industrial interest[155].

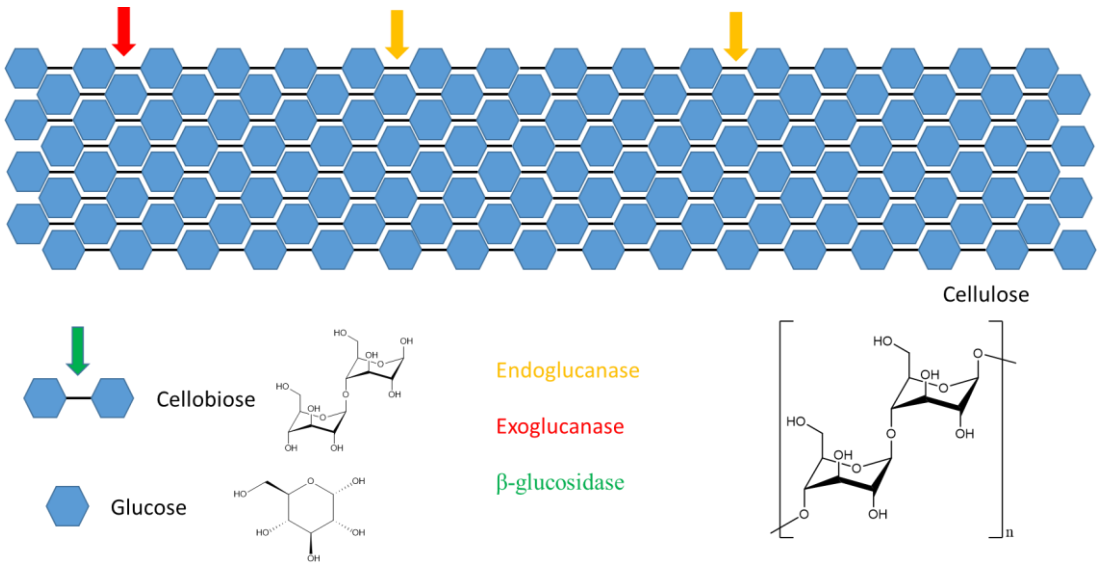


Figure 28: Degradation of cellulose to glucose through the concerted activities of endoglucanase, exoglucanase and β-glucosidase.

3.2.2.2 Results

To engineer *S. cerevisiae* to access glucose from cellulosic material the construction of two expression vectors for cellulose degradation was initiated. An endoglucanase/exoglucanase fusion protein[156] alongside two β-glucosidases previously identified from rumen metagenomic data and previously characterised *in vitro* by Ingenza were targeted for concerted expression. Previously, these enzymes had been characterised individually but expression in *S. cerevisiae* or the concerted activity of the enzymes had not been explored. In order to offer the best chance of successful expression and secretion, the native secretion signals were identified in each enzyme using Signal P version 4.0[157]. The signal peptides were removed and replaced by the well characterised *S. cerevisiae* alpha factor secretion signal[158], the genes were codon optimised for expression in *S. cerevisiae* using the algorithm provided by Genscript. Following optimisation, the sequences were then adapted for the inABLE® DNA assembly process and truncated parts designed, synthesised and cloned into compatible vector backbones (Figure 29). Construction of this vector using a standard inABLE based approach would involve an eight part assembly, to de-risk this construction a nested strategy was initiated.

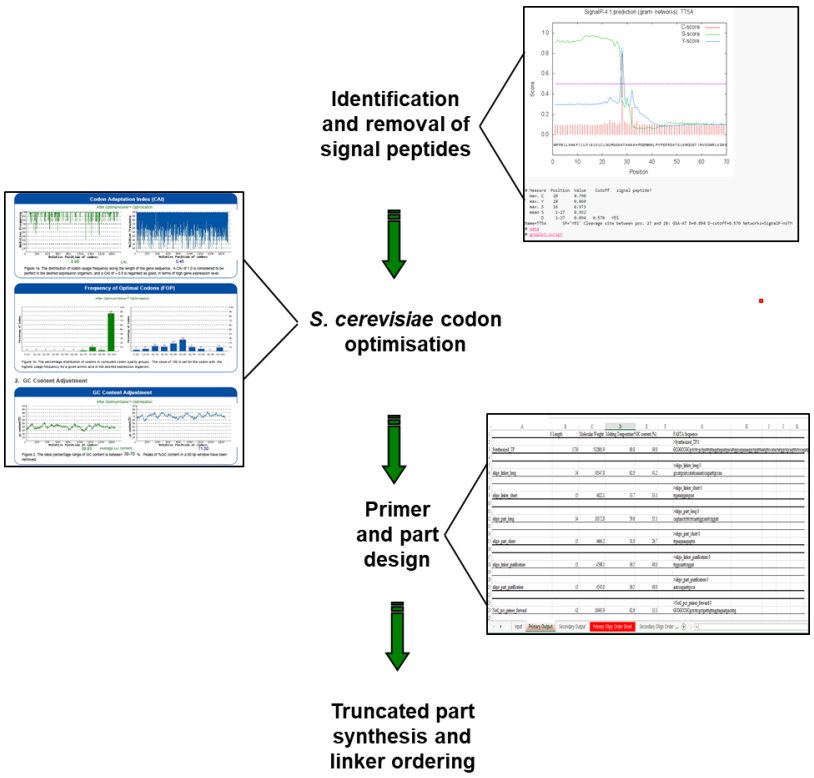


Figure 29: Design of cellulose degradation gene targets for inABLE assembly. Native signal peptides were removed and the genes codon-optimised prior to truncated part and linker design.

To facilitate using the product of one round of assembly in a subsequent part linker fusion reaction several specific oligonucleotides required to be designed. Appropriate POs, POl, LOl and LOs oligonucleotides were designed to introduce EarI recognition sites flanking the initially assembled truncated parts (Table 16). The oligonucleotides designed are standardised so that they are compatible with any nested approach. EarI sites present within the *S. cerevisiae*/*E. coli* shuttle backbone which would be tolerated in the standard inABLE procedure were disrupted through site directed mutagenesis (Section 2.4.8).

Oligonucleotide	Sequence	Description
POsxxx nested	gtgctggtCTCTTCg	POA fused to 5' end of first TP in the nested assembly, introducing EarI site. POA is specific to the 3 nucleotide overhang formed following SapI
POlxxx (nnn) nested	nnncGAAGAGaccagcaccaacaatgcagatatc	
LOsxxx nested	ggaagaatgtttcat	LOA fused to backbone part. Provides overhang complementary to POAxxx nested.
LOlxxx nested	gccatgaaacattcttccgatatctgcattgttg	
LOs172 nested	cgcctggtgggatcctCTCTTCg	LOA fused to the 3' end of the final TP in the nested assembly, introducing EarI site. Linker is specific to the backbone used to construct the nested vector.
LOl172 nested	gcccGAAGAGaggatcccaccaggcgtttaagggcaccaata	

Table 16: Nested primers used for construction of initial nested backbone. The EarI recognition sites which are used in the second round of assembly are highlighted in capital letters.

In a first round of assembly intermediate vectors comprising a *S. cerevisiae*/*E. coli* shuttle backbone, an endo/exoglucanase expression cassette and two β-glucosidase expression cassettes were constructed. These were then combined in a second round of assembly resulting in two dual expression vectors (Table 17).

Intermediate nested vectors	Final Expression vectors
β-glucosidase A cassette (Assembly Cellulase 1)	Endo/Exoglucanase/β-glucosidase A (Assembly Cellulase 5)
β-glucosidase B cassette (Assembly Cellulase 2)	
Endo/Exoglucanase cassette (Assembly Cellulase 3)	Endo/Exoglucanase/β-glucosidase B (Assembly Cellulase 6)
S. cerevisiae/*E. coli* backbone (Assembly Cellulase 4)	

Table 17: Nested strategy for construction of S. cerevisiae dual expression vectors.

Four initial nested vectors were constructed (Figures 30 and 31) following the part/linker fusion and assembly methods previously outlined. Critically in this first round of part linker fusion reactions, the enzyme SapI was used.

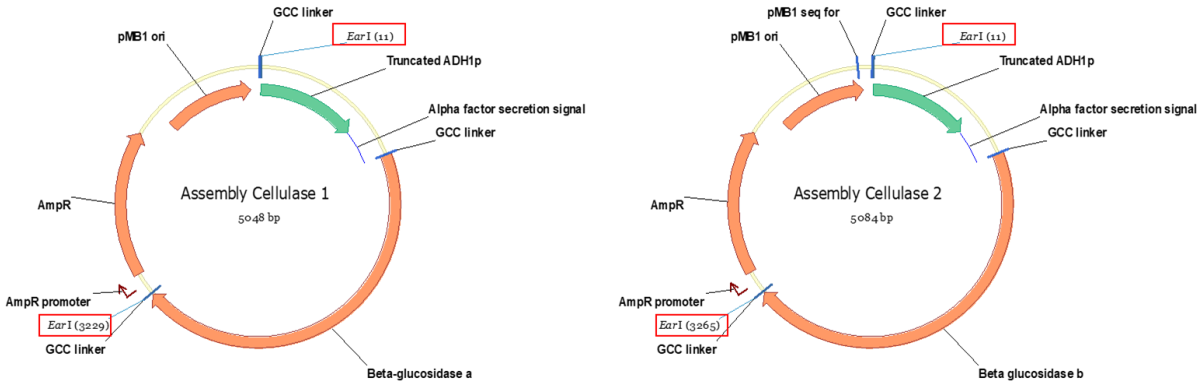


Figure 30: Nested intermediate β-glucosidase vectors. Introduced EarI recognition sites highlighted

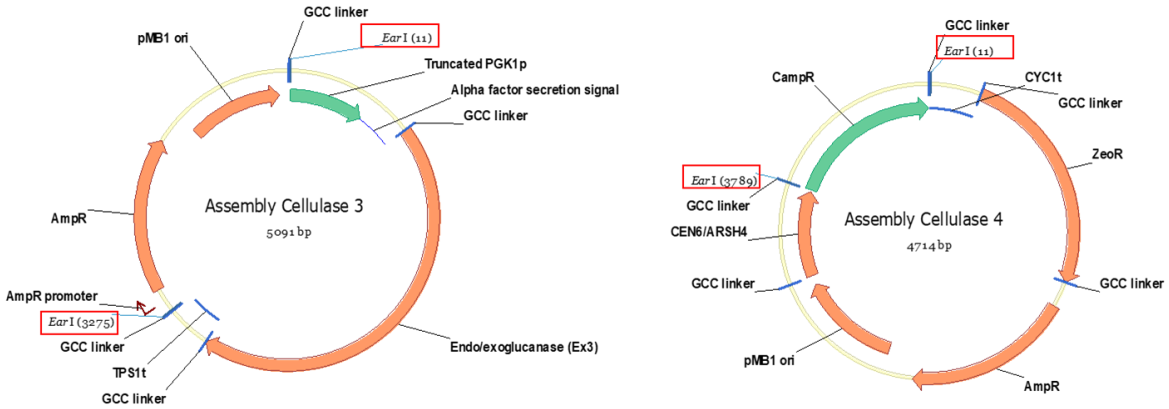


Figure 31: Nested intermediate Endo/exoglucanase and E. coli/S. cerevisiae shuttle backbone vectors. Introduced EarI recognition sites highlighted.

Construction of intermediate vectors was confirmed through EarI digestion (Figures 32, 33, 34 and 35) and sequencing of the junction regions between assembled parts. The construction of these vectors was achieved with relatively low efficiency (average of 23% of the clones characterised harboured the expected construct). This highlights the value of a nested approach as it has previously been shown that as the number of parts being assembled increases the efficiency further decreases (Section

3.1.2), potentially making the isolation of a correctly assembled product a lengthy process.

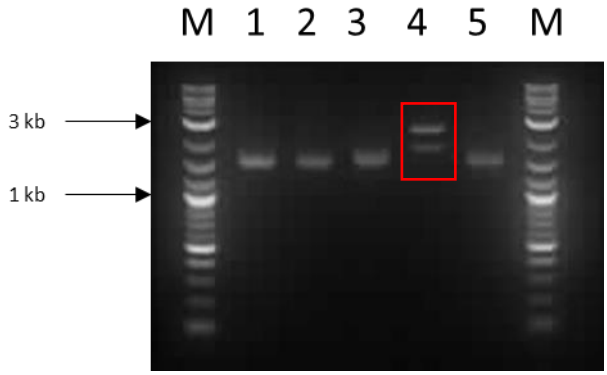


Figure 32: Screening of nested assembly Cellulase 1 vectors via EarI digestion. Screening of five vectors (lanes 1 - 5) resulted in one vector (lane 4) which yielded DNA fragments of the expected sizes (3.2 kb and 1.8 kb).

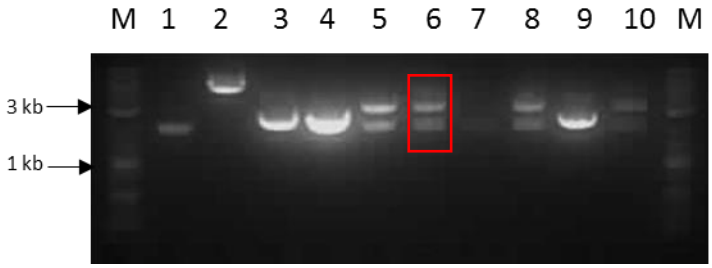


Figure 33: Screening of nested assembly Cellulase 2 vectors via EarI digestion. Screening of ten vectors (lanes 1 - 10) resulted in four vectors (lanes 5, 6, 7 ,8 and 10) which yielded DNA fragments of the expected sizes (3.2 kb and 1.8 kb).

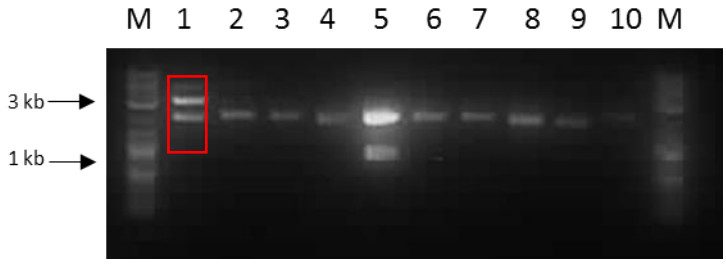


Figure 34: Screening of nested assembly Cellulase 3 via EarI digestion. Screening of ten vectors (lanes 1 - 10) resulted in one vector (lanes 1) which yielded DNA fragments of the expected sizes (3.2 kb and 1.8 kb).

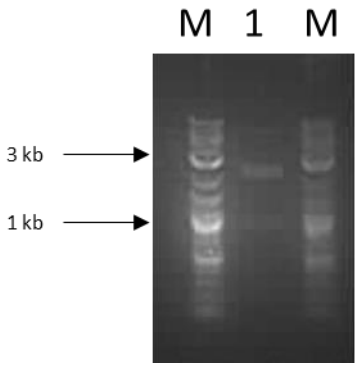


Figure 35: Screening of nested assembly Cellulase 4 via EarI digestion. Screening of one vector (lane 1) resulted in one vector (lane 1) which yielded DNA fragments of the expected sizes (3.7 kb and 0.9 kb).

Two final *S. cerevisiae* dual expression vectors were constructed through combination of the nested endo/exoglucanase, β-glucosidase and backbone parts (Figure 36).

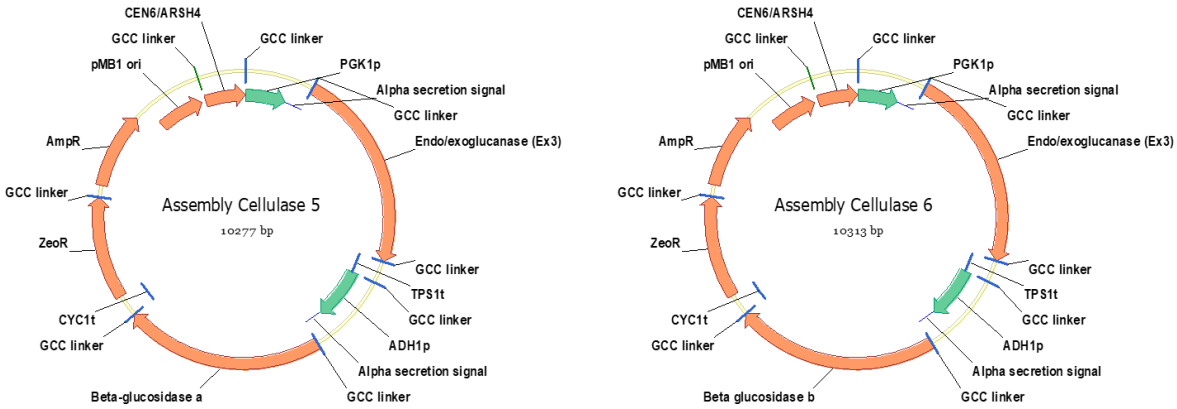


Figure 36: Final S. cerevisiae vectors constructed using a nested approach for concerted expression and secretion of an endo/exoglucanase fusion and a β-glucosidase.

Part/linker fusions and assemblies were performed as previously but this time using EarI in the part linker fusion reaction and the resulting assembly reactions used to transform *E. coli* as previously described. Ten random clones were picked, the corresponding vectors were isolated and the constructs screened via PvuI restriction

digestion (Figures 37 and 38) with successful assemblies confirmed via DNA sequencing.

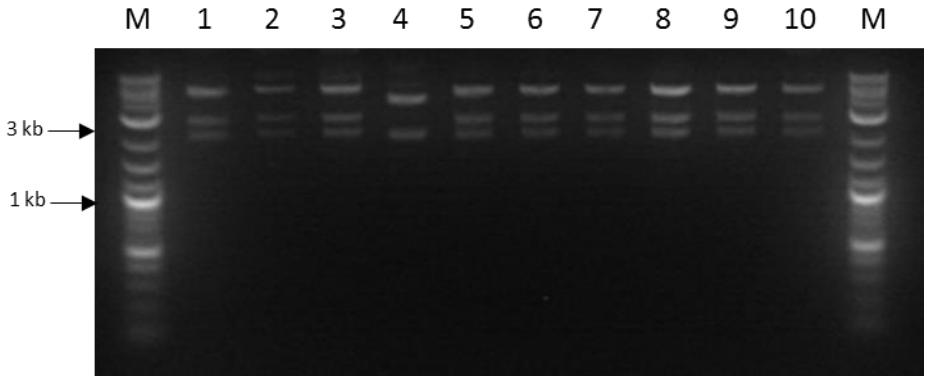


Figure 37: Screening of assembly Cellulase 5 vectors via PvuI digestion. Screening of ten vectors (lanes 1 - 10) resulted in nine vectors (lane 1 – 3 and 5 - 10) which yielded DNA fragments of the expected sizes (5.4 kb, 2.8 kb and 2.0 kb).

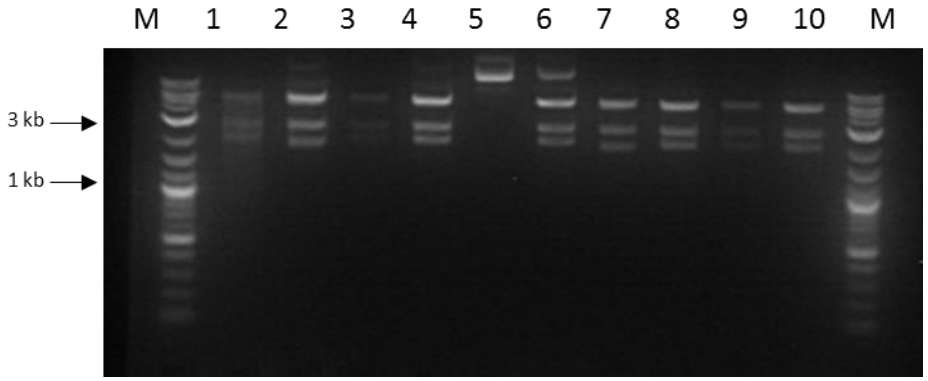


Figure 38: Screening of assembly Cellulase 5 vectors via PvuI digestion. Screening of ten vectors (lanes 1 - 10) resulted in nine vectors (lane 1 – 4 and 6 - 10) which yielded DNA fragments of the expected sizes (5.4 kb, 2.8 kb and 2.0 kb).

Screening of isolates following the final nested assembly reaction highlighted that 90% of the clones characterised contained the expected DNA construct, with >1000 transformants obtained. It is extremely unlikely that such an efficiency would have been achieved for the construction of a vector of this complexity using the standard inABLE procedure. This result highlights the application and benefits of a nested

approach to DNA assembly when looking to construct vectors comprised of a large number (5+) parts. Despite the increased time investment required to construct intermediate vectors these can be re-used in subsequent assemblies and this investment is further offset by the time saved in identifying positive isolates due to increased reaction efficiency.

The constructed vectors were introduced into an industrial ethanol producing *S. cerevisiae* strain (Ethanol Red, Fermentis)[159] following the lithium acetate transformation method described by Gietz et al[160] generating strains Cellulase 68 and 69. To characterise rate, intermediates and products of cellulose breakdown, a shake flask assay and complementary analytical methodology that had previously been developed by Ingenza were utilised.

The two *S. cerevisiae* strains engineered for concerted expression of the exonuclease/endonuclease fusion and either beta-glucosidase (Cellulase 68 and 69) were initially screened at shake flask level using commercial corn mash as a substrate. Corn mash is currently used as a primary feedstock for the bioethanol industry in the United States. In this process the corn is treated with amylases (glucoamylase and α-amylase) to degrade the starch present into glucose. *S. cerevisiae* is then added to assimilate the liberated glucose and produce ethanol[161]. Currently in this process however, the cellulose is not targeted. Analysis of the dried distillers grains and solubles (DDGS) which constitutes the remaining feedstock at the end of a corn ethanol fermentation contains between 9 and 16 % cellulose by weight[162].

The engineered strains and a negative control (the non-engineered parental strain) were incubated with and without commercial glucoamylase which is standardly added to the process to catalyse the saccharification of starch to glucose. Ethanol production throughout the experiment was monitored by CO_2 loss. Strains Cellulase 68 and 69 displayed slightly enhanced ethanol production rate in comparison to the parental control in the presence of commercial glucoamylase. A HPLC based analytical method was used for quantification of DP4+ (saccharides with four or more glucose units), DP3 (maltotriose), maltose, glucose, glycerol and ethanol at the assay end point (65 hours). Whilst Cellulase 68 and 69 did not display increased final

ethanol yield, the slightly faster rate of ethanol production rate displayed in this experiment is potentially an advantage to end users (Figure 39).

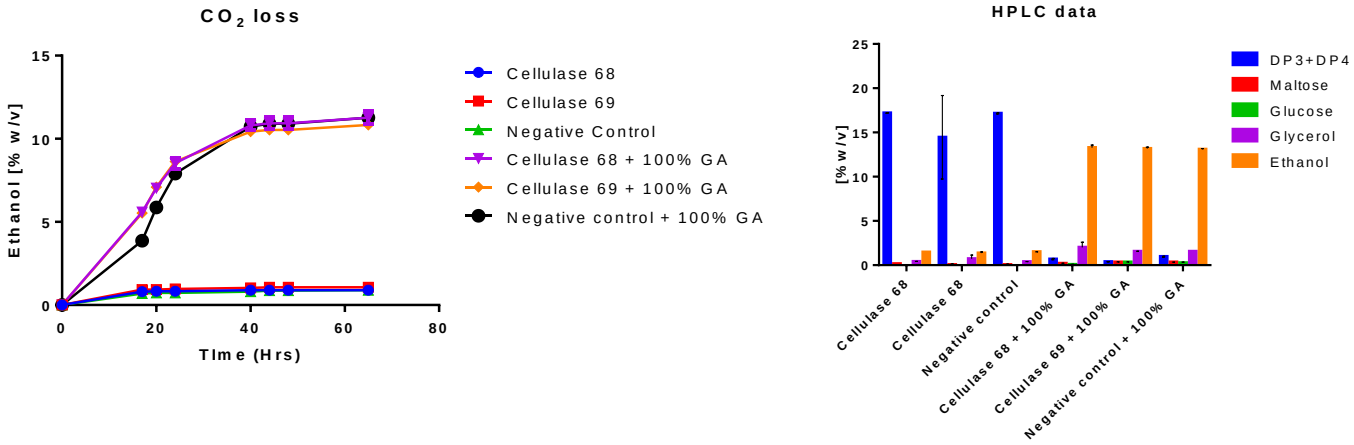


Figure 39: Shake flask assay of Cellulase 68 and 69 (CO2 loss) (Left). Shake flask assay of Cellulase 68 and 69 (End point HPLC analysis) (Right).

3.2.2.3 Conclusions

This proof of concept study demonstrates the ability to use a nested inABLE strategy in which the product of one assembly reaction is compatible with a second round of assembly. The technique is feasible due to the introduction of EarI recognition sites in linker sequences during the construction of initial vectors which can then be utilised in a second round of part/linker fusion. In this study a vector comprised of eight parts, which is on the limit of the standard technology, was built in a nested manner with 90% of the clones screened following the final assembly reaction containing the expected construct. The utilisation of a nested strategy when constructing vectors with over five parts has numerous advantages. Firstly, through de-risking the vector construction by splitting it into stages and minimising the number of parts in the final assembly reaction assembly efficiency was increased, limiting the time taken to identify clones containing the expected isolates. Secondly each of the intermediate parts are re-usable, meaning that if, for example, one of the

genes in the construct is determined to be sub-optimal it can be replaced with a more suitable candidate without having to reconstruct the multi-part construct afresh.

3.2.3 Overcoming the detrimental effect of part number on assembly efficiency

3.2.3.1 Introduction

Protein secretion can be a powerful of tool for multiple biotechnology applications, for example it may provide a simple and cost efficient downstream purification platform avoiding the requirement to lyse cells at scale when looking to develop a biologics production strain. Alternatively, when looking to engineer a microbe to access polymers such as starch, cellulose or lignin as feedstocks and avoid the purchasing of purified enzyme cocktails the engineered organism requires to be able to efficiently secrete hydrolytic enzymes to degrade these substrates into accessible sugars.

An N-terminal secretion signal is required to direct the targeted protein into the cells secretory pathway. The protein is first transferred into the endoplasmic reticulum (ER) either during or directly following translation. In the ER the protein is post-translationally modified before it is transferred to the Golgi apparatus for further modification, sorting and finally secretion. This complex process involves multiple steps and can quickly become overwhelmed if the targeted protein is highly overexpressed, resulting in misfolded inactive protein. It is difficult to predict the optimal secretion signal and promoter strength to maximise the level of protein secreted for a given gene, with protein specific optimisation often required[163]. In order to attempt to optimise the abundance of extracellular target protein, the combinatorial construction of a library of *S. cerevisiae* expression vectors containing 32 combinations of promoters and secretion signals coupled to a cellulase gene and a vector backbone comprised of the required origins of replication and selection markers was embarked upon.

3.2.3.2 Results

The construction of the considered library would likely not be feasible through a six part assembly due to the assembly efficiency being too low when utilising the standard technology. To maximise the final assembly reaction efficiency a nested strategy was implemented. To enable library construction a cellulase coding sequence, terminator, *E. coli* backbone and *S. cerevisiae* backbones were first assembled using the nested inABLE methodology (section 3.2.1). The cellulase gene from *Cytophaga hutchinsonii* had previously been cloned in *E. coli* and identified as being an interesting candidate due to its potential dual exoglucanase/β-glucosidase characteristics[164]. The native secretion signal was identified and removed, and the gene codon optimised for *S. cerevisiae* expression prior to truncated part design (described in Section 3.2.2.2).

Four constitutive *S. cerevisiae* promoters; Alcohol dehydrogenase 1 (ADH1p); Phosphoglycerate Kinase 1 (PGK1p); Glyceraldehyde-3-phosphate dehydrogenase 1 (GAPDH1p) and Profilin 1 (PFY1p) of varying strength were selected from the Ingenza inABLE database. Eight secretion signal sequences were identified from the *Pichia pastoris* PichiaPink™ secretion signal collection (Invitrogen). As these secretion signal sequences had previously been codon optimised for *P. pastoris,* the wild type sequence was identified from the organism's genomic DNA and codon optimised for *S. cerevisiae* using the algorithm provided by DNA 2.0. The identified secretion signal sequences were added to the promoter specific reverse PCR primer (PCRr) (Table 18) allowing for the fusion of the secretion signal sequence to the promoter truncated parts. This resulted in the generation of 32 re-usable promoter secretion signal truncated parts.

Truncated part number	Promoter	Secretion signal	Modified PCRr primer sequence
406	ADH1p	*S. cerevisiae* α factor	ttagatggatccgctcttcgGGCTAATGCGGAGGATGCTGCGAATAAAACTGCAGTAAAAA TTGAAGGAAATCTCATtgtatatgagatagttgattgtatgc
407		*A. niger* α amylase	ttagatggatccgctcttcgGGCCAAAGCAGGTGCAGCGACCTGCAGACCGTACAGAAACA AAGACCACCAAGCGACCATtgtatatgagatagttgattgtatgc
408		*A. awamori* GLA1 glucoamylase	ttagatggatccgctcttcgGGCCAAACCAGAACAAACCAAACCAGACAAAGCCAACAAAG ATCTAAAAGACATtgtatatgagatagttgattgtatgc
409		Human serum albumin	ttagatggatccgctcttcgggcAGAGTAAGCAGAAGAAAAAAGAAACAACAAGCTTATAA AGGTAACCCACTTCATtgtatatgagatagttgattgtatgc
410		*K. marxianus* INU1 inulase	ttagatggatccgctcttcgGGCACTGACTCCTGCCAATGGAAGCAACAAGGAGTATGCTA ACTTCATtgtatatgagatagttgattgtatgc
411		*S. cerevisiae* SUC2 invertase	ttagatggatccgctcttcgGGCAGATATTTTGGCTGCAAAACCAGCCAAAAGGAAAAGGA AAGCTTGCAAAAGCATtgtatatgagatagttgattgtatgc
412		*S. cerevisiae* M1 killer toxin	ttagatggatccgctcttcgGGCTACGACTAGATGTAGTAATGTGATGAAAAATAATATACT GACAGATCTAACTAATACTTGGGTTGGCTTAGTCATtgtatatgagatagttgattgtatgc
413		*G. gallus* lysozyme	ttagatggatccgctcttcgggcACCTTGACAGATACCCAACAAAGCAGTCAATCCCAACAA GACCAAAACAAGACACATTGGGTCGTTCTTACCCAGCATtgtatatgagatagttgattgtatgc

Table 18: Design of modified PCRr primers for addition of secretion signal to ADH1 promoter truncated part. Introduced secretion signals in capitals. Underlined 3'-end fragment of the selected promoters. Primer design carried out following the same strategy for PGK1, GAPDH1 and PFY1 promoters.

The recipient nested vector was constructed comprised of a cellulose coding sequence, *S. cerevisiae* terminator and *E. coli*/*S. cerevisiae* backbone parts flanked by EarI recognition sites (Figure 40), providing a platform to allow for the combination of this construct with the multiple promoter/secretion signal parts in a second round of assembly to generate the required library of expression vectors.

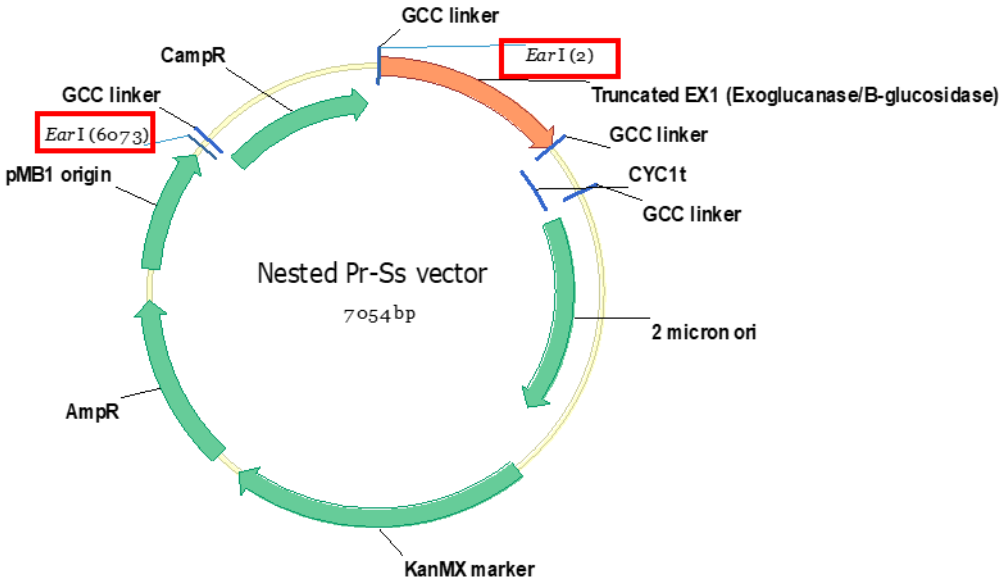


Figure 40: Initial nested vector for promoter secretion signal library construction. Introduced EarI sites highlighted in red boxes In a second stage of inABLE assembly promoter/secretion signal combinations can be introduced upstream of the Ex1 β-glucosidase.

To build this vector a four part assembly was performed as described in Section 2.5. Around fifty colonies were obtained following *E. coli* transformation of which 10 were picked at random and screened for successful assemblies via EarI restriction digest (Figure 41). The construction of this initial vector was achieved at low efficiency with 10% of the clones characterised containing the expected construct.

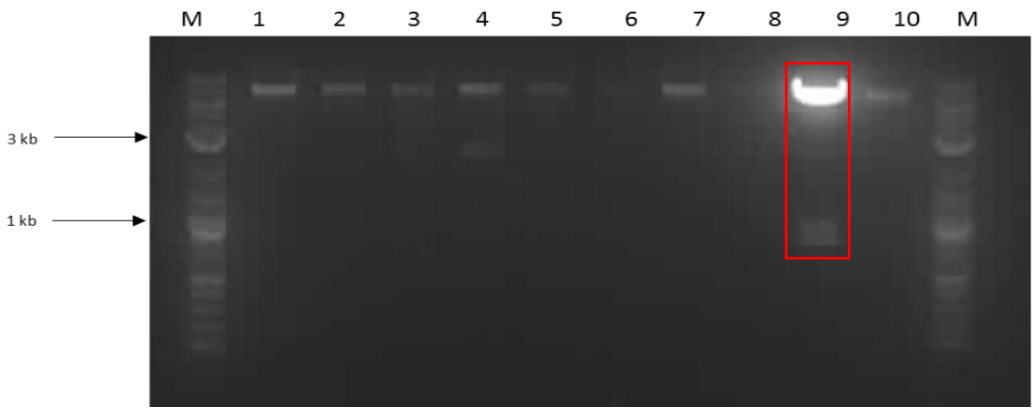


Figure 41: Screening of nested backbone assembly via an EarI restriction digest. Screening of ten vectors (lanes 1 - 10) resulted in one vector (lane 9) which yielded DNA fragments of the expected sizes (6.1 kb and 1.0 kb).

The low efficiency of this four part assembly highlights the requirement to implement a nested approach. Sanger sequencing analysis of the vector confirmed the expected assembly. The backbone was tested in a second round of inABLE for library construction through combination with the 32 promoter secretion signal parts, vitally this part linker fusion reaction was performed using EarI.

To characterise the library 60 colonies were picked and the promoter and secretion signal combination present in the construct identified through colony PCR. From the sixty clones analysed, nine failed to generate a PCR product either due the isolates harbouring a mis-assembled construct or due to the PCR reaction failing. From the library of 32 at least one positive was isolated for 28 of the potential constructs (Figure 42).

Construct	Promoter	Secretion signal	Number identified
1	ADH1p	*A. niger* alpha Amylase	3
2		*A. awarmori* Glucoamylase	1
3		*H. sapien* serum albumin	2
4		*K. maxianus* Inulinase	3
5		*S. cerevisiae* Invertase	2
6		*S. cerevisiae* M1 killer toxin	2
7		*G. gallus* Lysozyme	2
8		*S. cerevisiae* α factor	1
9	PGK1p	*A. niger* alpha Amylase	2
10		*A. awarmori* Glucoamylase	1
11		*H. sapien* serum albumin	
12		*K. maxianus* Inulinase	2
13		*S. cerevisiae* Invertase	2
14		*S. cerevisiae* M1 killer toxin	2
15		*G. gallus* Lysozyme	2
16		*S. cerevisiae* α factor	1
17	GAPDH1p	*A. niger* alpha Amylase	
18		*A. awarmori* Glucoamylase	4
19		*H. sapien* serum albumin	
20		*K. maxianus* Inulinase	1
21		*S. cerevisiae* Invertase	2
22		*S. cerevisiae* M1 killer toxin	1
23		*G. gallus* Lysozyme	
24		*S. cerevisiae* α factor	1
25	PFY1p	*A. niger* alpha Amylase	2
26		*A. awarmori* Glucoamylase	1
27		*H. sapien* serum albumin	4
28		*K. maxianus* Inulinase	2
29		*S. cerevisiae* Invertase	2
30		*S. cerevisiae* M1 killer toxin	2
31		*G. gallus* Lysozyme	
32		*S. cerevisiae* α factor	1
		Total	51

Figure 42: Identification of promoter secretion signal combinations following library construction through colony PCR. From the library of 32 combinations at least one positive was isolated for 28 of the potential constructs

The above library was used to transform *S. cerevisiae* following the lithium acetate/single stranded carrier DNA method outlined by Gietz et al[113], with 1.5 µg of the library used to transform 10^8 freshly prepared competent cells. Transformants were selected for on YPD-G418 and 100 colonies were initially picked for screening.

Screening of the library was envisioned through a liquid phase glucose oxidase/peroxidase liquid phase assay. This assay relies on the incubation of the

culture supernatant - which should contain the secreted enzyme – with cellulose model substrates and the detection of liberated glucose. Glucose is detected through the oxidation of glucose by glucose oxidase resulting in the formation of and hydrogen peroxidase and gluconic acid. Hydrogen peroxidase is in turn used by horseradish peroxidase for the oxidation of ABTS which results in a colorometric response that can be monitored at 420 nm[165] (Figure 43).

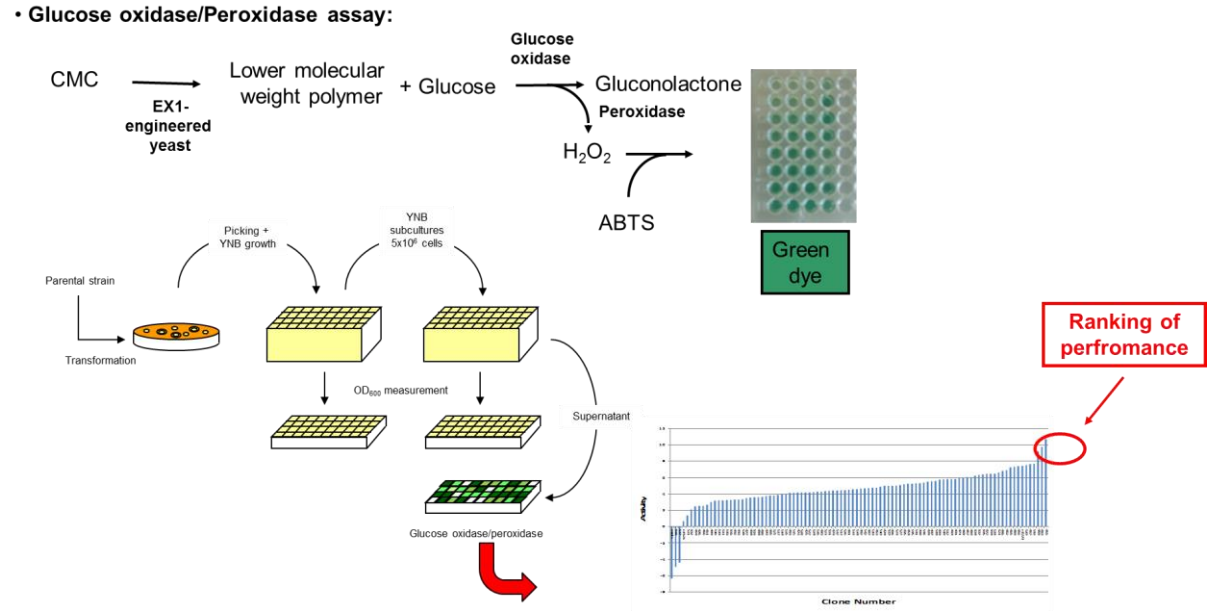


Figure 43: Screening of cellulase expressing strains through glucose oxidase/peroxidase coupled assay. In this approach, culture supernatant is incubated with cellobiose resulting in the formation of glucose through the action of the secreted β-glucosidase.

The glucose oxidase assay was performed using both carboxymethyl cellulose and cellobiose which are model substrates for both exoglucanase and β-glucosidase activities. Despite attempts to optimise both the protein expression and assay conditions, reproducible glucose release could not be detected for any of the library strains characterised.

3.2.3.2 Conclusions

In order to optimise extracellular protein production, the combinatorial construction of a library of vectors containing 32 combinations of promoters and secretion signals was designed. The construction of this library however would likely not be feasible through a six part assembly due to the assembly efficiency being too low. To enable the library construction a recipient vector comprised of the cellulase coding sequence, a *S. cerevisiae* terminator, and *E. coli*/*S. cerevisiae* shuttle backbone was first assembled using the nested inABLE methodology. Eight secretions signals were fused to the four promoter truncated parts via PCR allowing for a final two part assembly, maximising assembly efficiency.

This final assembly resulted in thousands of transformants with the library diversity confirmed through a colony PCR based screen. A high throughput liquid phase assay was designed to characterise the library following the introduction of the library into *S. cerevisiae*. Unfortunately, no activity could be detected for any of the constructs. Further work to determine if this result is due to the low activity of the enzyme on the substrates explored or inefficient protein expression/secretion is required.

3.2.4 Conclusions

In this chapter multiple parameters which can impede efficient multi-component pathway construction have been identified. It is clear to maximise assembly efficiency, as is often the case when constructing combinatorial libraries, it is required to limit the potential for secondary structure formation within linkers and avoid the possibility of truncated part self-ligation during part linker fusion. The effect of part number on assembly efficiency has also been defined, highlighting the requirement for the development of the nested approach when multiple DNA fragments require to be assembled at high efficiency.

The expansion of the standard inABLE technology through a nested approach was investigated through the construction of a complex vector (>7 parts) and a combinatorial library. Nested linkers which allow for the stepwise assembly of parts were designed and tested. The two primary advantages of this approach are the ability to construct increasingly complex vectors which may be required when building a multi-gene pathway, and the efficient construction of combinatorial libraries which is powerful tool in pathway optimisation. These applications have been explored and the relevant plasmids and libraries were successfully constructed using the nested approach. The efficiency in which these constructs were prepared would not have been achieved using the standard technology. Identification of key parameters which limit the efficiency of the inABLE DNA assembly platform and the implementation a nested assembly workflow are key developments for acceleration of the optimisation of biosynthetic pathways.

4. The development of an inABLE 2.0 platform

4.1 Gel free inABLE

4.1.1 Introduction

Despite the advantages described in the previous chapter the throughput and efficiency of the inABLE DNA assembly platform is still amenable to optimisation. DNA assembly approaches are standardly ranked based on throughput, accuracy, flexibility and their amenability to process automation. The further development of an inABLE 2.0 platform was envisioned through optimisation of each of these parameters.

A major bottleneck within the current inABLE DNA workflow is the requirement for part/linker fusion purification through gel electrophoresis or biotinylated primers and streptavidin beads[166]. These purification approaches have limitations and present a bottleneck in the current workflow. The running of agarose gels and excision of the DNA of interest is the currently preferred method for isolating the DNA of interest. However, this is a cumbersome, low throughput approach with the added disadvantage that the DNA extracted from the gel is likely a combination of the desired part/linker fusion and fragments lacking one or both linkers which are vital in the assembly stage.

The utilisation of biotinylated primers and streptavidin beads is a potential solution to these issues. However, this approach requires an additional stage of processing increasing the length of time taken to construct the vector of interest by up to six hours. This section describes the development of a method to address this bottleneck in the current workflow. It outlines a high throughput approach, easily automated through the use of a liquid handling robot which results in the isolation of only the DNA of interest whilst degrading the remaining contaminating DNA in the reaction.

4.1.1.1 Current methods

At present, gel electrophoresis is the standard method to purify desired DNA from a mixture[167]. This approach utilizes an electric field for DNA migration through a porous agarose gel. The fragments of DNA travel through the gel at a speed inversely proportional to their length. Therefore, smaller fragments travel further on the gel and DNA of the desired size can be identified through comparison with a co-electrophoresed set of DNA fragments of a known size. The desired DNA is subsequently isolated from the gel through manual excision of the gel fragment and extraction of DNA from the gel matrix using commercially available kits (i.e., QIAquick gel extraction – Qiagen). This process however is low throughput and provides a bottleneck in the inABLE DNA assembly process. This is also a manual process not amenable to automation.

An additional limitation to this approach is the resolution that can be achieved through gel electrophoresis. The DNA of interest contains 5' and 3' linkers (as described in Section 1.6.1.6) which constitute an increase in size of ~ 50 base pairs, over a DNA part which lacks either the 5' and/or or the 3' linker. The part itself can often be ≥ 1000 base pairs in length so this change in size may be only ≤ 5% and thereby indistinguishable by gel electrophoresis. As a result, DNA lacking one (or both) of the linkers can readily be carried through into the assembly stage where its presence is likely to be detrimental to the efficiency of the assembly reaction.

To date attempts to accelerate this process have largely focused on the utilisation of biotinylated linkers and streptavidin beads[168]. This approach improves throughput and is in theory specific for purification of DNA which contains both 5' and 3' linkers. The technique relies on the extremely high affinity with which biotin binds to streptavidin. Biotinylated primers are available from all major primer manufacturers whilst small scale spin columns packed with streptavidin beads are produced by multiple companies. Biotinylated oligonucleotides (referred to as purification oligonucleotides) complementary to the 16 nucleotide single stranded overhangs of inABLE parts can be ordered. However, it is not feasible to confirm whether both 5' and 3' biotinylated linkers have been attached to a part. For

instance, presence of only one linker would still result in the DNA part-linker binding to the streptavidin. In order to purify the desired DNA fragment, two rounds of purification are required. In the first round a biotinylated oligo specific for the 5' linker is utilised before the process is repeated using a purification oligo specific for the 3' linker. This adds additional steps to the process increasing the time to assemble the required vector, an extra cost to purchase the additional purification oligonucleotides, the need for two stages of purification, the unavoidable loss of product at each stage and the potential for contamination with non-specifically bound DNA. This results in a less efficient, albeit somewhat higher throughput process than what is achievable with gel extraction based purification.

4.1.1.2 Exonucleases and phosphorothioate bonds

Exonucleases are enzymes which degrade DNA through hydrolysis of phosphodiester bonds in a stepwise manner from the end of the polynucleotide chain, as opposed to endonucleases which cleave internally in the chain. Exonucleases cleave nucleotides one at a time from the end of polynucleotide chain in either a 5' → 3' or 3' → 5' direction. To date, 17 exonucleases have been identified in *E. coli*[169] which play key cellular roles in DNA repair and recombination, genome stability and mutation avoidance.

Exonucleases cleave phosphodiester bonds from a free DNA end and therefore do not degrade supercoiled circular plasmid starting material (although a number are able to initiate degradation from nicks in plasmid DNA[169]). The use of an exonuclease treatment instead of the standard gel-based strategy to remove contaminating linearised backbone can therefore be envisioned. However, an alternative strategy such as counter selection - as detailed in section 4.1.2.1 - will be required to select against contamination of the assembly reaction with supercoiled plasmid starting material.

Phosphorothioate bonds – which have been found naturally in species of Streptomyces[170, 171] are generated through replacing a non-bridging oxygen within

the phosphate backbone with a sulfur atom (Figure 44). The introduction of the sulfur atom renders the internucleotide linkage resistant to endo- and exonuclease cleavage[172, 173].

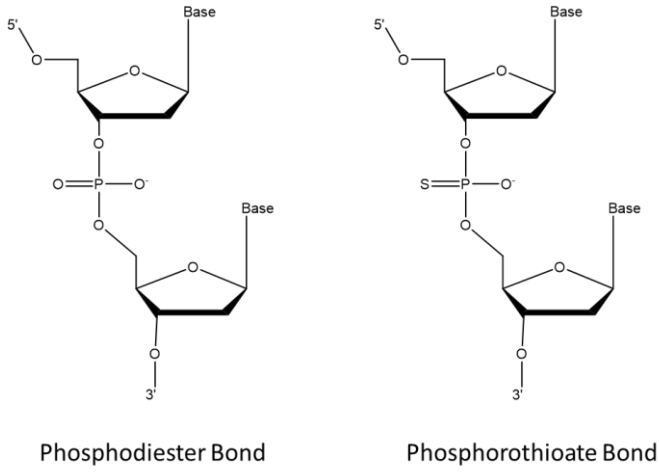


Figure 44: Generation of a phosphorothioate bond through the introduction of a sulphur atom within the DNA backbone.

This property can be used to protect DNA of interest from endo- and exonuclease attack, and has been utilised to protect DNA in a number of approaches including next generation sequencing[174] and single stranded DNA recombineering[175, 176] although not for DNA assembly. Through the introduction of phosphorothioate bonds within the 5' and 3' linkers utilised in the inABLE procedure, only DNA which has linkers annealed at both ends is protected, while the remaining DNA is unprotected from exonuclease degradation.

Such an approach can offer specificity which cannot be rivalled by gel electrophoresis (due to the resolution which can be achieved) and biotin/streptavidin purification (a separation with potential for contamination by non-specifically bound DNA or a part with a linker attached at only one end). The process does not involve additional steps during ligation of linkers to parts and replaces the gel electrophoresis stage with the addition of an exonuclease and incubation for a maximum of 30 minutes. This enzyme addition and incubation has the potential to be extremely high throughout and is amenable to automation, unlike gel electrophoresis.

4.1.2 Results

4.1.2.1 Backbone counter selection

The simplest method to accelerate the assembly process is to simply avoid purification of the part/linker fusion. A concern for such an approach is that standardly the part including the backbone will carry the same *E. coli* origin of replication and antibiotic resistance marker as the final assembly plasmid. If the corresponding part linker fusion reaction does not reach completion, plasmid starting material will not be removed through gel extraction and will be carried into the assembly reaction without an agarose gel based purification. The presence of the unwanted cloned backbone part in the assembly reaction will result in a high percentage of transformants containing this vector starting material rather than the assembly product.

One potential solution to this issue is the use of a double antibiotic selection approach. In such a strategy an additional antibiotic marker fragment would be included in each assembly reaction and selection on agar containing both this antibiotic and the one on the vector backbone would select against carry through of starting material. However, this approach requires the inclusion of the additional part in each assembly reaction, reducing assembly efficiency whilst the maintenance of two antibiotic resistance markers puts an unnecessary metabolic burden on the host.

An alternative strategy to address the issue of backbone contamination is through the introduction of a counter-selectable marker into the truncated part carrying vector as this fragment of DNA will not be present in the desired final assembly product. The SacB gene from *Bacillus subtilis* which encodes a levansucrase[177] can be used to confer sucrose sensitivity in *E. coli*[178]. To explore the potential to use the *sacB* gene as counter selection against TP vector starting material contamination the *sacB* gene was cloned into the vector TP59 (KanR, pBR322 fragment) generating vector TP184 (Figure 45). Importantly the *sacB* gene is cloned into the carrying vector fragment and therefore is not present in final assembly product. Contamination of the assembly reaction with starting material (TP184 plasmid) therefore, in the case of an

incomplete part linker fusion reaction, can be counter selected against through the addition of sucrose (10% [w/v]) in transformation selection media.

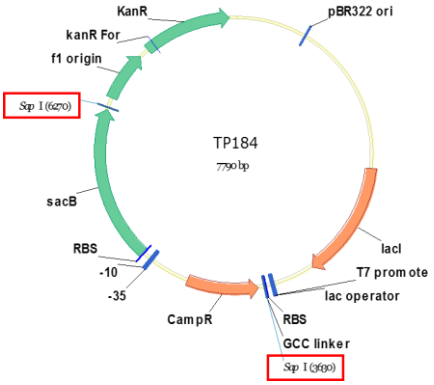


Figure 45: TP184 vector containing counter selectable SacB marker cloned adjacent to the chloramphenicol marker in the carrying vector.

Activity of the *sacB* gene in vector TP184 was confirmed by *E. coli* transformation and the absence of growth on LB-Kan supplemented with 10% [w/v] sucrose (Suc10). As expected, dilutions of the reaction selected on LB-Kan resulted in cell growth (Figure 46).

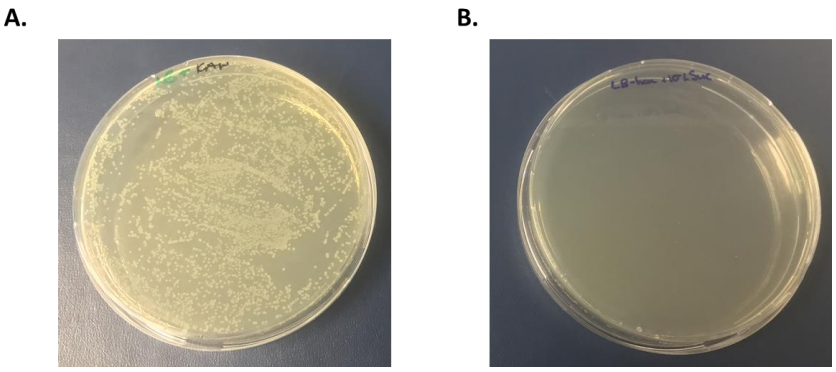


Figure 46: Selection of E. coli cells harbouring the SacB containing TP184 on LB-Kan and LB-Kan with 10% [w/v] sucrose. Selection of cells on sucrose results in cell death due the expression of the sacB gene from plasmid TP184.

To explore gel purification free inABLE DNA assembly coupled to counter selection of the potentially contaminating backbone a two part assembly was performed, combining TP184 and TP171 which results in the construction of a vector which confers resistance to kanamycin and tetracycline. Three purification approaches were tested to determine assembly efficiency when an agarose gel based purification step is omitted. As well as the standard purification, a silica-membrane-based purification (PCR purification) and direct use of fragments without any purification were also explored. Transformants were selected for on LB-Tet + Kan and LB-Kan supplemented with Suc_{10} and the number of colonies counted (Figure 47).

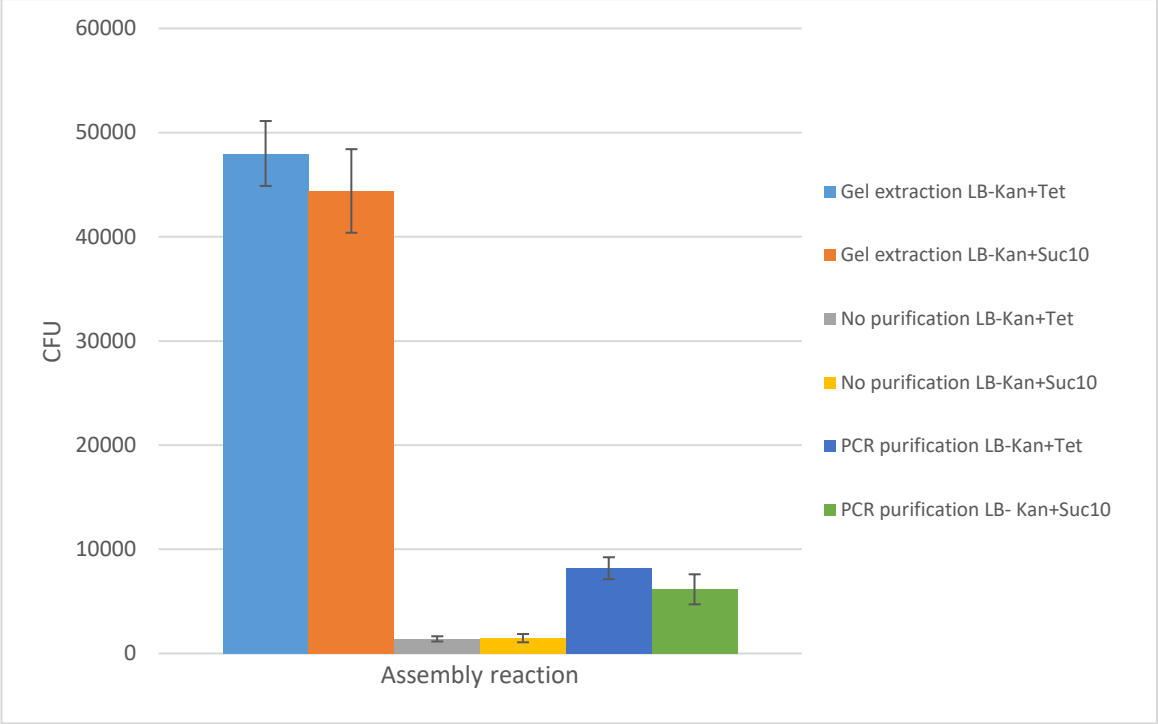


Figure 47: Characterisation of the utilisation of the SacB counter selection in a Gel free inABLE approach. Three purification approaches were compared: a silica-membrane-based purification (PCR purification), direct use of fragments without any purification and the standard gel extraction based purification. Average colony counts are the result of triplicate experiments.

Selection of transformants on LB-Kan + Tet and LB-Kan + Suc_{10} results in a similar number of transformants confirming that that SacB counter selection is a suitable approach in the inABLE workflow to limit carry through of starting material into the

assembly reaction. However, using a silica membrane spin column to purify part linker fusions rather than gel extraction results in a decrease in the number of transformants by 83% whilst performing no purification following the part linker fusion reaction results in a drop by 97%. The avoidance of an agarose gel based purification will enhance the throughput of the procedure, however, the dramatic drop in efficiency means that such an approach is not compatible with a multi-part assembly strategy as is often required when constructing biosynthetic pathways.

Despite the observed decrease in efficiency such a strategy may be a promising tool when looking to construct relatively simple constructs (2-3 parts) in a high throughput manner as many as thousands of isolates harbouring the expected construct were obtained without the requirement to purify the DNA fragments via gel electrophoresis which offers significant time savings and allows for the automation of the process.

To increase the number of parts that can be assembled in a gel-free assembly a novel approach was devised which combines an exonuclease treatment to degrade contaminating DNA coupled to phosphorothioate based protection of part linker fusions as means to purify DNA within the current DNA assembly workflow.

4.1.2.2 Exonuclease identification

Identification of a suitable exonuclease depends on its substrate specificity, directionality and sensitivity to phosphorothioate bonds. SapI digestion of a truncated part results in the generation of 5’ overhangs comprising of three nucleotides to which part and linker oligonucleotides are annealed generating 16 nucleotide 3’ overhangs (Section 2.5.2). The backbone from which the truncated part is cleaved has 3’ three nucleotide recessions (Figure 48a).

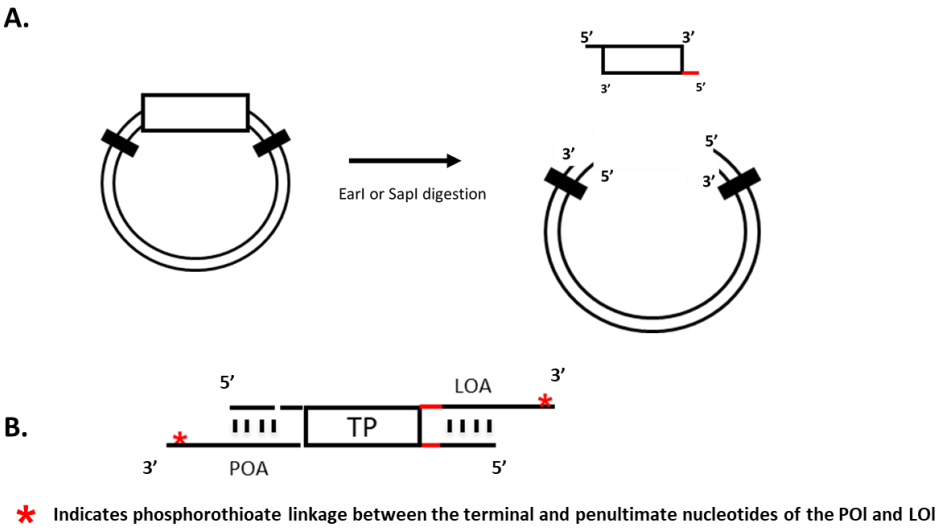


Figure 48: A. Overhangs generated following SapI digestion with 5’ and 3’ ends annotated. B. Overhangs generated following POA and LOA annealing with 5’ and 3’ ends annotated and introduced phosphorothioate bond highlighted.

Since the logical point to introduce phosphorothioate bonds is at the end of the 16 nucleotide 3’ overhangs (Figure 48b) the exonuclease of choice should act in a 3’ to 5’ manner and be unable to cleave phosphorothioate bonds allowing for the protection of correctly ligated part linker fusions. Through inversion of the POA and LOA it is also feasible to generate 5’ 16 nucleotide extensions (Figure 49) and therefore 5’ to 3’ exonuclease unable to cleave phosphorothioate bonds were included in the search.

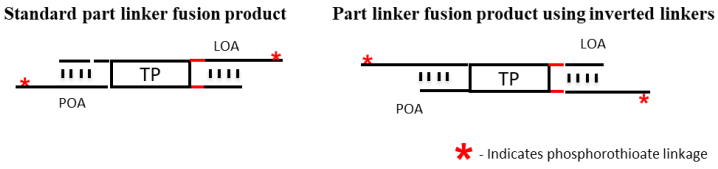


Figure 49: Inversion of POA and LOA sequences to generate 5' overhangs. The ability to generate 5’ overhangs expands the number of exonuclease that can be explored.

A review of available exonuclease characteristics (Table 19) immediately identified Exonuclease III from *E. coli* as a particularly interesting candidate as it prefers substrates with blunt or recessed 3' termini with the 3' extensions over 4 bases or longer essentially being resistant to cleavage.

Exonuclease	Polarity	Initiates on DNA with				Phosphorothioate cleavage
		5' ext	3' ext	Blunt	Nick	
Exonuclease I	3'-5'	No	+/-	+/-	NR	-
Exonuclease III		Yes	+/-	Yes	Yes	-
Exonuclease T		No	Yes	+/-	NR	-
BAL-31		Yes	Yes	Yes	Yes	NA
Exonuclease V	Both	Yes	Yes	Yes	No	NA
Exonuclease VII		+/-	+/-	No	No	+
Exonuclease Lambda	5'-3'	+/-	Yes	Yes	No	-
Rec-JF		+/-	No	+/-	No	-

Table 19: Characteristics of Exonucleases available from New England Biolabs. +/-: activity greatly reduced relative to preferred substrate.

An initial test using Exonuclease III was carried out to determine if this enzyme would efficiently degrade a SapI digested truncated part (Figure 50a) and secondly if the ligation of part and linker oligonucleotides (at this stage without phosphorothioate modification) would render the fragment of DNA resistant to exonuclease degradation (Figure 50b)

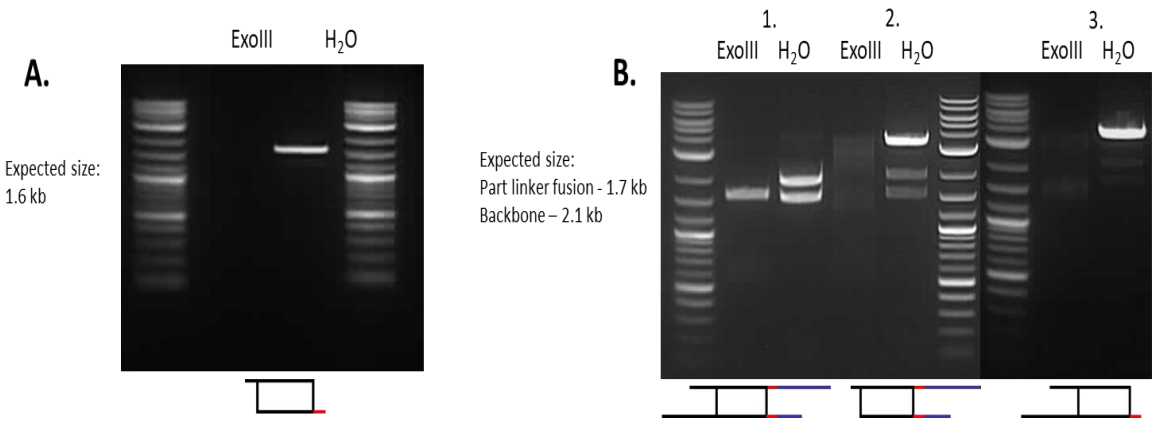


Figure 50: Treatment of the possible outcomes of a part linker fusion with Exonuclease III. A. Truncated part digested from backbone, gel extracted and subjected to exonuclease treatment (No POA or LOA attached). B. Part/linker fusion reaction performed with TP/POA/LOA (1), TP/LOA (2) or TP/POA (3).

The results of this initial study show that as expected exonuclease III has the capability to degrade truncated parts completely lacking both POA and LOA (Figure 50a), annealed to only the LOA (Figure 50b-2) or the POA (Figure 50b-2) whilst the full length part linker fusion (without phosphorothioate bonds) appears to be resistant to degradation (Figure 50b-1). This is of particular interest as with the current gel electrophoresis based protocol it is not possible to distinguish between part linker fusions and fragments of DNA lacking either POA, LOA or both. In the current approach these contaminating fragments are carried through into the assembly reaction potentially decreasing assembly efficiency.

The remaining exonucleases identified as having attractive characteristics were screened through the digestion of part linker fusion reactions performed using POA and LOA fragments with phosphorothioate bonds, In the case of 3'- 5' exonucleases POA and LOA sequences were only modified to introduce the phosphorothioate bonds. For 5' – 3' the POA and LOA sequences were inverted generating 16 nucleotide 5' extensions rather than the standard 3' extensions, with a phosphorothioate bond again introduced between the final two bases of the 16 nucleotide overhang. The products of the exonuclease treatment were analysed via gel electrophoresis to determine if the exonucleases tested were able to remove the backbone fragment whilst not digesting the protected part linker fusion (Figure 51).

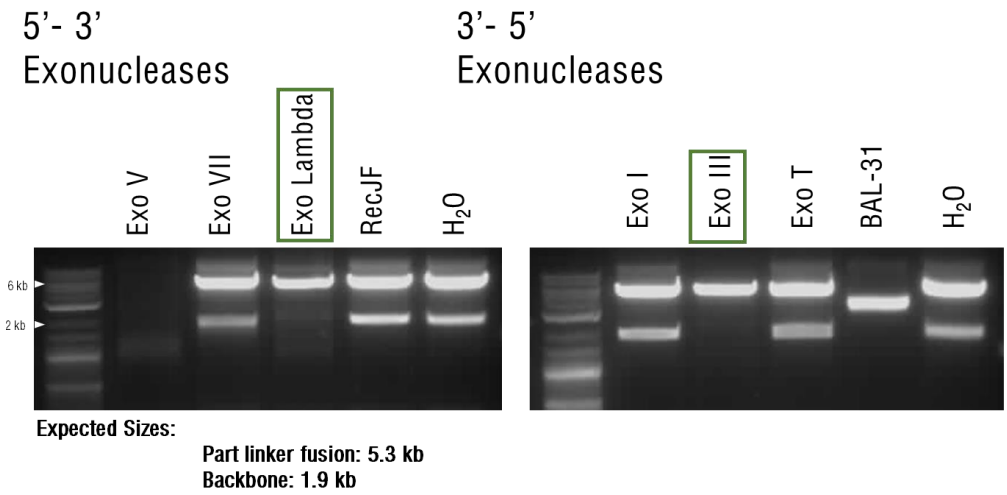


Figure 51: Digestion of phosphorothioate protected part/linker fusion reactions with candidate 5' - 3' and 3' - 5' exonucleases. The desired exonuclease reaction would result in the removal of the backbone fragment (1.9 kb) whilst the protected part linker fusion (5.3 kb) remains undigested.

Only Exonuclease lambda and Exo III displayed the required characteristics. As Exonuclease lambda operates in a 5' to 3' direction - and to utilise this enzyme the entire LOA and POA inABLE back catalogue would require to be reconfigured - the most suitable candidate identified from this initial characterisation all further tests were performed using *E. coli* Exonuclease III.

4.1.2.3 Exonuclease treatment

The drop in efficiency observed when part linker fusions are not gel extracted is likely due to the carryover of contaminating DNA fragments present in the part/linker fusion reaction into the assembly reaction. This DNA includes the vector backbone and unbound POA and LOA fragments both of which are removed through gel electrophoresis. These fragments will compete in the assembly reaction with the construction of the desired vector.

Exonuclease III treatment without the use of phosphorothioate bonds was explored to determine if this alone is sufficient to remove contaminating DNA and maintain or even enhance assembly efficiency. As described previously, a review of the characteristics of Exonuclease III suggest that it should be suitable to remove contaminating DNA from the reaction without digesting the part linker fusion (removing the requirement for phosphorothioate bonds). This is due to its preferred substrates being blunt or recessed 3' termini with 3' extensions over 4 bases or longer reported as being resistant to cleavage[179]. These characteristics should prevent the degradation of the part/linker fusion which contains 16 nucleotide 3' extensions whilst contaminating vector backbone (recessed 3' termini) would be removed from the reaction.

To explore the potential to use an exonuclease treatment alone to purify part linker fusions a two part assembly comprised of TP171 and TP184 was utilised which confer resistance to tetracycline and kanamycin respectively. Part linker fusions were treated with Exonuclease III (Section 2.5.6) and either purified by agarose gel, purified using a PCR spin column or not purified following exonuclease treatment. In parallel the inABLE procedure was performed as standard using agarose gel based purification. The resulting assembly reactions were used to transform *E. coli*, the transformants selected for on media supplemented with tetracycline and kanamycin and the number of colonies counted (Figure 52).

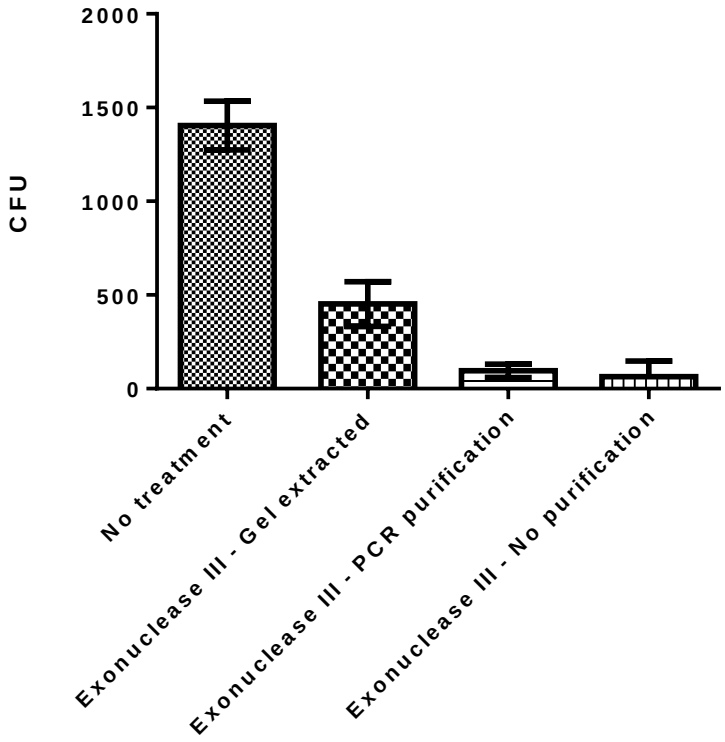


Figure 52: The effect of exonuclease treatment on assembly efficiency. Following exonuclease treatment, three purification approaches were compared: a silica-membrane-based purification (PCR purification), direct use of fragments without any purification and the standard gel extraction based purification. The average number of colonies present on LB Kan + Tet selection plate are the result of triplicate experiments.

Treatment of part linker fusion reactions with Exonuclease III under the conditions tested above resulted in a pronounced decrease in assembly efficiency. In this experiment the standard inABLE procedure yielded ~ 1400 colonies carrying the expected assembly. When an exonuclease treatment was used followed by purification via agarose gel the number of transformants dropped to ~ 450 colonies, a decrease in assembly efficiency of 68 %. The efficiency further decreases when the exonuclease treatment is followed by a PCR clean up and no purification (up to a 93% decrease).

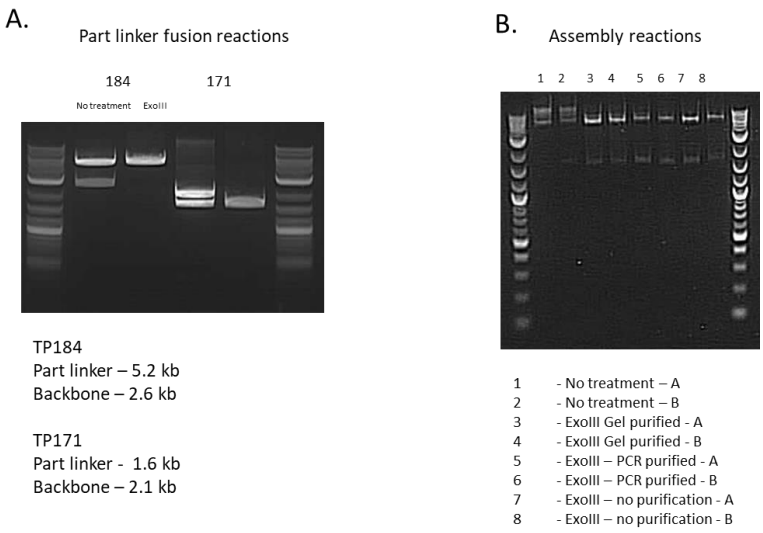


Figure 53: Analysis of part linker fusions and assembly products following Exonuclease treatment. Exonuclease treatment has successfully removed the backbone fragments (A) however the resulting part/linker fusions appear to be unable to assembly. This can be observed through the lack of higher molecular weight bands present in lanes 3 – 8 in image B.

Analysis of part linker fusion reactions by gel electrophoresis (Figure 53a) following exonuclease treatment gel analysis confirmed that, as expected, the contaminating backbone fragment was completely digested whilst it appeared that the part/linker fusion remained undigested. However, analysis of the assembly reaction products via agarose gel suggested that when treated with Exonuclease III the two DNA fragments do not assemble as seen for the standard inABLE workflow (Figure 53b). It is hypothesised that this is due to degradation of the 16 base overhang by Exonuclease III. Whilst 16 base overhangs are not the preferred substrate it is likely to occur as a secondary reaction at a reduced rate compared to backbone degradation. To avoid the issue, the utilisation of phosphorothioate bonds to protect these 16 base overhangs was explored.

4.1.2.4 Coupling phosphorothioate containing linkers with an Exonuclease III treatment

The previously attempted two part assembly utilising three purification approaches was repeated utilising modified phosphorothioate containing linkers. A single phosphorothioate bond was introduced at the 3' of each LOl and POl used in the assembly reaction. Linker preparation and part linker fusion reactions were performed as standard. Following completion of the cycling reaction, part linker fusions reactions were divided, and half treated with Exonuclease III whilst the remaining samples were stored on ice prior to purification. As previously, three part purification approaches were explored and transformants were selected for on LB + Kan + Tet and LB + Kan supplemented with Suc_{10} (Figure 54).

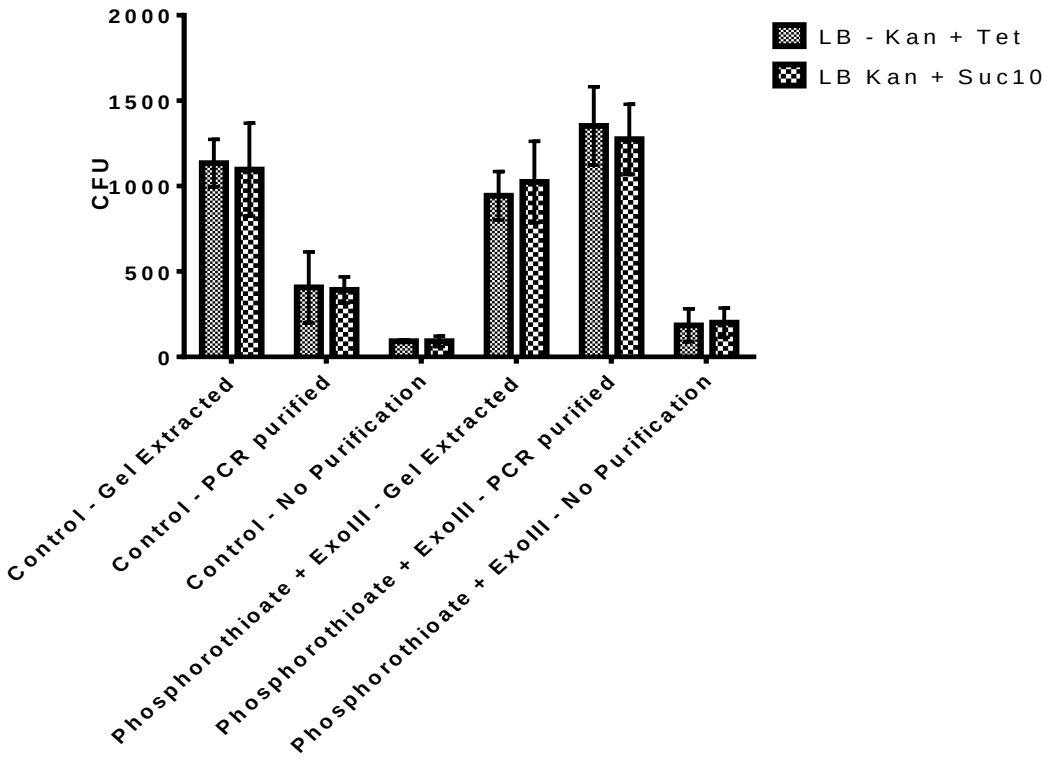


Figure 54: The effect of coupled phosphorothioate protection and exonuclease treatment on assembly efficiency. As previously, three purification approaches were compared: a silica-membrane-based purification (PCR purification), direct use of fragments without any purification and the standard gel extraction based purification. Average CFU are the result of experiments performed in triplicate.

Comparable numbers of transformants were observed when transformants were selected on LB- Kan + Tet in comparison to Kan and Suc_{10} containing media whilst no transformants were observed on negative control plates as previously observed.

In the absence of an exonuclease treatment, a decrease in assembly efficiency when a gel extraction was not performed was observed (64% decrease in efficiency when a spin column purification was performed and a 92% decrease in efficiency when no purification was performed). However, when an exonuclease treatment was performed, coupled to phosphorothioate treatment and the fragments gel purified the assembly efficiency did not decrease highlighting that the presence of the phosphorothioate bonds prevents overhang degradation. Critically, when an exonuclease treatment and phosphorothioate linkers was coupled to spin column purification the assembly efficiency was maintained. Therefore, through the removal of the backbone DNA fragment and/or aberrant part linker fusions the assembly efficiency is maintained without the requirement to run an agarose gel. This is a key result as it opens the door to process automation whilst not having to sacrifice assembly efficiency.

To determine if this strategy is scalable beyond two parts the effect on assembly efficiency of Exonuclease III treatment in conjunction with PCR purification (the most promising gel free work flow from the two-part assembly) was then explored to combine five DNA part-linker fusions to confirm the approach was suitable for the assembly of more complex vectors. A five-part assembly was then conducted in triplicate using the protocol as described for the two-part approach. The standard inABLE procedure was also performed in parallel for the five-part assembly.

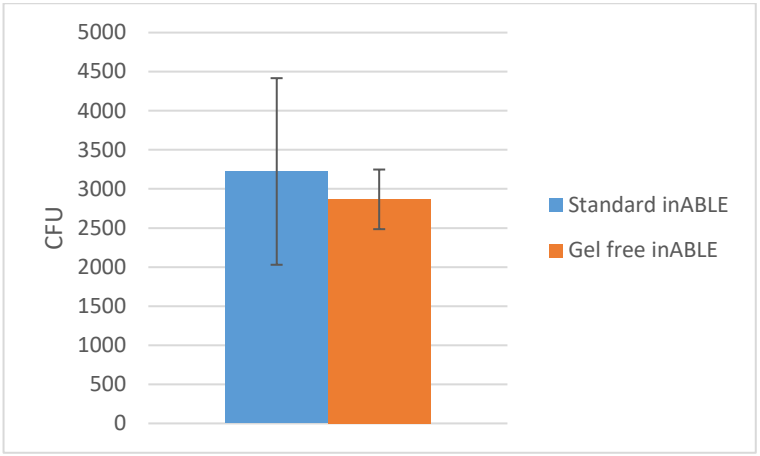


Figure 55: Comparison of standard inABLE and Gel free inABLE assembly for five DNA fragments. Four DNA fragments containing antibiotic resistance markers and one containing an E. coli of replication were assembled. The average number of transformants was calculated from experiments performed in triplicate.

Analysis of the number of transformants carrying the expected assembly product (Figure 55) and the part linker fusion and assembly products via capillary electrophoresis (Figure 56) highlights that the gel free approach is suitable for the assembly of larger numbers of DNA fragments without sacrificing DNA assembly efficiency. The ability to automate this process offers the potential to significantly enhance process efficiency.

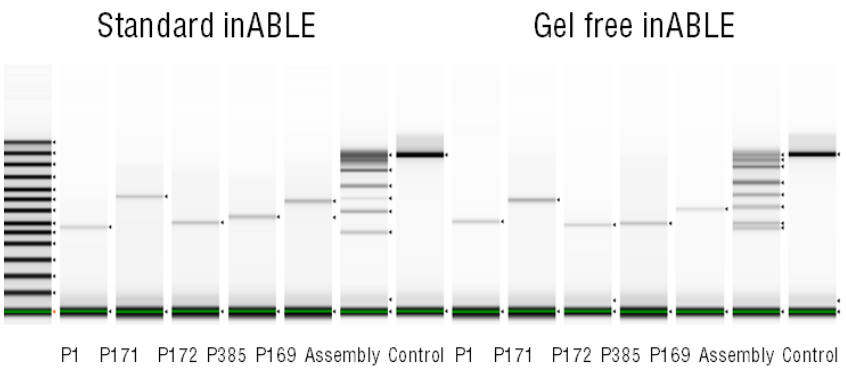


Figure 56: Analysis of part linker fusions and assembly products from a standard and gel free five part construct via capillary electrophoresis. The five fragments assembled P1, P171, P172, P385 and P169 can be observed a single band for both the standard and gel free inABLE. Analysis of the assembly products highlights the generation of higher molecular weight products for both approaches suggesting the successful generation of assembly products.

4.1.2.5 The effect of phosphorothioate bonds alone on assembly efficiency

It was hypothesised that the introduction of a phosphorothioate bond alone may result in an increase in assembly efficiency. The present workflow does not require the addition of ligase during the assembly reaction and attempts to include ligase in the assembly reaction results in a modest increase in assembly efficiency, but a pronounced increase in the number of incorrect assemblies. Therefore, the current process relies on *in vivo* ligation of the assembly product to repair nicks present in the assembled construct. These nicks are also a potential target for cellular exonucleases. *E. coli* Exonuclease III - for example - is able to initiate degradation at plasmid nicks in a 3' → 5' manner[169]. The introduction of phosphorothioate bonds at the 3' end of the POl and LOl has the potential to confer to resistance to cellular exonucleases at nicks present in the assembled vector *in vivo.*

An experiment was therefore performed in which five DNA fragments were assembled, four of which confer resistance to an antibiotic and one an *E. coli* origin of replication allowing selection of cells containing the expected assembly on media supplemented with four antibiotics. Part linker fusion reactions were performed using either standard linker or phosphorothioate modified linkers. All fragments were purified via the running of an agarose gel and assembled in triplicate as standard.

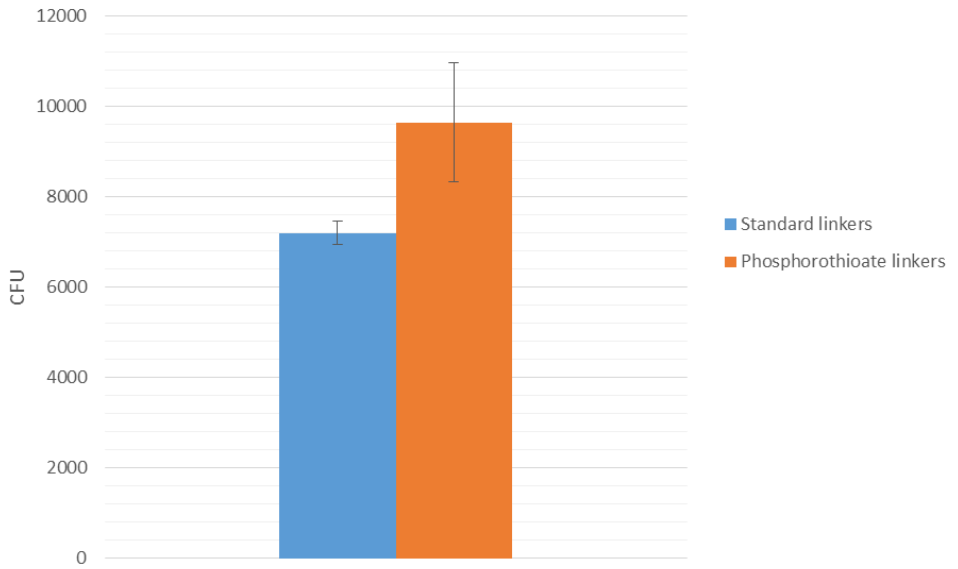


Figure 57: Comparison of the effect of phosphorothioate bonds alone on assembly efficiency. Part linker fusion reactions were performed using either standard linker or phosphorothioate modified linkers and all fragments were purified via the running of an agarose gel followed by gel extraction. The average number of transformants was calculated from experiments performed in triplicate.

No significant difference in assembly efficiency were observed when comparing standard to phosphorothioate containing linkers (Figure 57). This result suggests that the utilisation of phosphorothioate linkers alone does not provide a noticeable advantage when assembling five fragments of DNA, however such an effect may be noticeable when a higher number of fragments are assembled as the number of nicks present in the assembly product is directly correlated to the number of parts being assembled.

4.1.2.6 The effect of Exonuclease treatment and phosphorothioate bonds on assembly accuracy

In each of the previous examples the assembly of DNA fragments conferring antibiotic resistance or containing origins of replication allowed for a direct comparison of assembly efficiency through the number of colonies present on selection plates containing the appropriate antibiotics. However, this approach does

not allow for the study of assembly accuracy, i.e. the ratio of correctly assembled products to misassemblies, as cells carrying misassemblies will not be viable on the selection plate.

To explore assembly accuracy a four-part assembly (one fragment containing an antibiotic resistance marker and origin of replication and three comprised of a gene and associated regulatory regions) which had previously proven difficult to construct using the standard inABLE approach was explored. Three workflows were compared, the standard inABLE procedure, phosphorothioate bonds only and coupled exonuclease treatment and phosphorothioate bonds. All fragments were purified via gel electrophoresis with assembly reactions and *E. coli* transformation performed as described previously. Transformants were screened via colony PCR and the ratio of correctly assembled fragments to misassemblies presented as assembly accuracy (Figure 58).

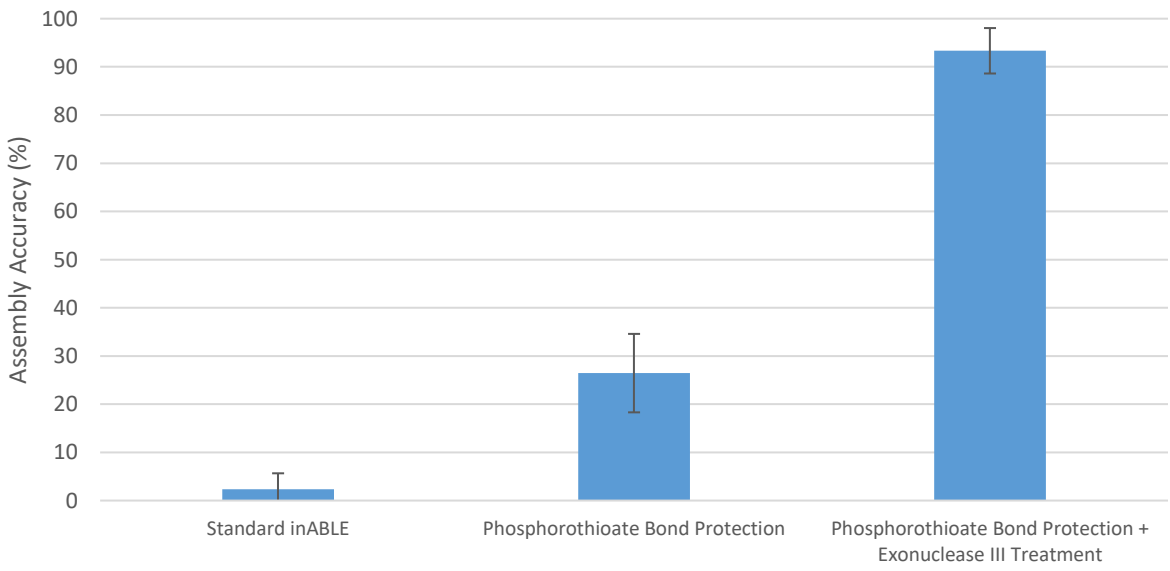


Figure 58: The effect of phosphorothioate bonds and coupled phosphorothioate/exonuclease treatment on assembly efficiency. The standard inABLE procedure was compared to phosphorothioate bonds only and coupled exonuclease treatment and phosphorothioate bonds. All fragments were purified via gel electrophoresis and transformants screened by cPCR to identify correctly assembled vectors.

In the absence of both phosphorothioate bonds and exonuclease treatment an average assembly efficiency of 2.3% was achieved which is comparable to previous results when trying to build this construct. The introduction of phosphorothioate bonds into the linker sequences resulted in a slight increase in assembly efficiency to 26.4%, however when phosphorothioate bonds were coupled to exonuclease treatment the assembly accuracy rose to 93.3% (Table 20).

	Conditions		**Experiment 1**	**Experiment 2**	**Standard Deviation**	**Average Assembly Accuracy (%)**
	Phosphorothioate Bond Protection	Exonuclease III Treatment	Assembly Accuracy (%)	Assembly Accuracy (%)		
Standard inABLE	(-)	(-)	0	4.7	3.3	2.3
Phosphorothioate Bond	(+)	(-)	32.2	20.7	8.1	26.4
Coupled Phosphorothioate Bond + Exonuclease III Treatment	(+)	(+)	90	96.7	4.7	93.3

Table 20: The effect of phosphorothioate bonds and coupled phosphorothioate/exonuclease treatment on assembly efficiency - Colony counts.

4.1.3 Conclusions

The utilisation of phosphorothioate linkers and exonuclease treatment in this study has shown the ability to maintain assembly efficiency whilst removing the requirement to run an agarose gel, greatly reducing the length of time to build the required constructs whilst also increasing the predictability and robustness of the procedure and permitting automation of the process. The new approach is completely compatible with the current inABLE workflow, not requiring changes to the part and primer design, the previously cloned truncated parts or the part linker fusion or assembly reactions. Only the linkers are modified to include phosphorothioate bonds. As described in the above work, the presence of these bonds can then be utilised to purify the part-linker DNA using an exonuclease treatment. Specifically, this provides a DNA purification method that results in

greater process efficiency through higher throughput and specificity in reduced time while permitting automation of the entire process by eliminating the need for gel electrophoresis. This is a key breakthrough in the development of a second generation inABLE procedure and is subject to US patent application number – 15901431 (METHOD FOR ASSEMBLY OF POLYNUCLEIC ACID SEQUENCES USING PHOSPHOROTHIOATE BONDS WITHIN LINKER OLIGOS, 2018)

4.2 Accelerating the part/linker fusion reaction

4.2.1 Introduction

One of the most time consuming stages of the inABLE DNA assembly process is the cycling of digestion and ligations reactions that are utilised to generate a part linker fusion, which takes a total of 335 minutes. The cyclic nature of this reaction is feasible, as described previously, due to the characteristics of SapI which as a type IIs restriction enzyme cuts out with its recognition site. This means that the part/linker fusion is void of SapI recognition sites and essentially removed from the reaction. Through cycling between rounds of digestion and ligation it is possible to enrich towards the product of interest. This is a similar strategy to one employed in golden gate based DNA assembly strategies to enrich towards the product of interest.

This process currently involves six rounds of digestion and six rounds of ligations taking a total of close to six hours to run (Table 21).

Stage	Lid Temperature: 42°C	
	Temperature (°C)	Time
1	37	90
2	16	30
3	37	30
4	16	15
5	37	15
6	16	15
7	37	15
8	16	10
9	37	15
10	16	10
11	37	60
12	16	30
Total	335 minutes	

Table 21: The cycles of digestion and ligation in a standard part linker fusion reaction. Digestion is performed at 37 °C which is the optimal temperature for EarI and SapI activity whilst ligation is performed at 16°C, the optimal temperature for T4 DNA ligase.

Due to its stability, SapI is characterised by NEB as an enzyme which is not recommended for digestions of longer than one hour[180]. Therefore, in order to determine if any advantage is gained from the cycling of digestion and ligation a time course experiment was implemented.

4.2.2 Results

4.2.2.1 Assembly efficiency

To explore the effect of digestion/ligation cycling on assembly efficiency, the construction of two vectors was explored in parallel, one of which conferred resistance to Amp and Camp and a second which conferred resistance to Amp and Kan.

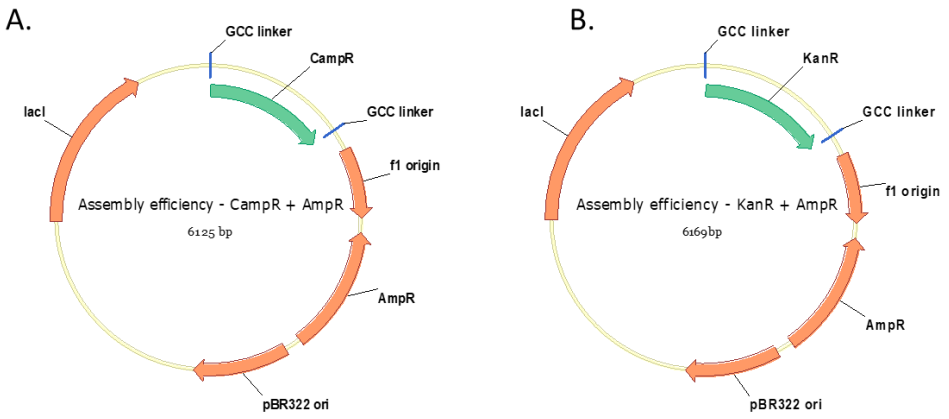


Figure 59: Vectors constructed to determine the effect of the number of digestion and ligation cycles on assembly efficiency. Two vectors were constructed to explore this, the first vector (A.) was constructed through a two part assembly and confers resistance to Camp and Amp whilst the second (B.) confers resistance to Kan and Amp.

Following each ligation reaction cycle, samples were removed from the thermocycler and the SapI and T4 DNA ligase present in the reactions heat inactivated through incubation at 65 °C for 20 minutes. Part linker fusions were first visualised on an agarose gel to monitor the presence of starting material in the reaction (Figure 60). The presence of visible staring material would highlight an incomplete part/linker fusion reaction suggesting additional rounds of digestion and ligation are required.

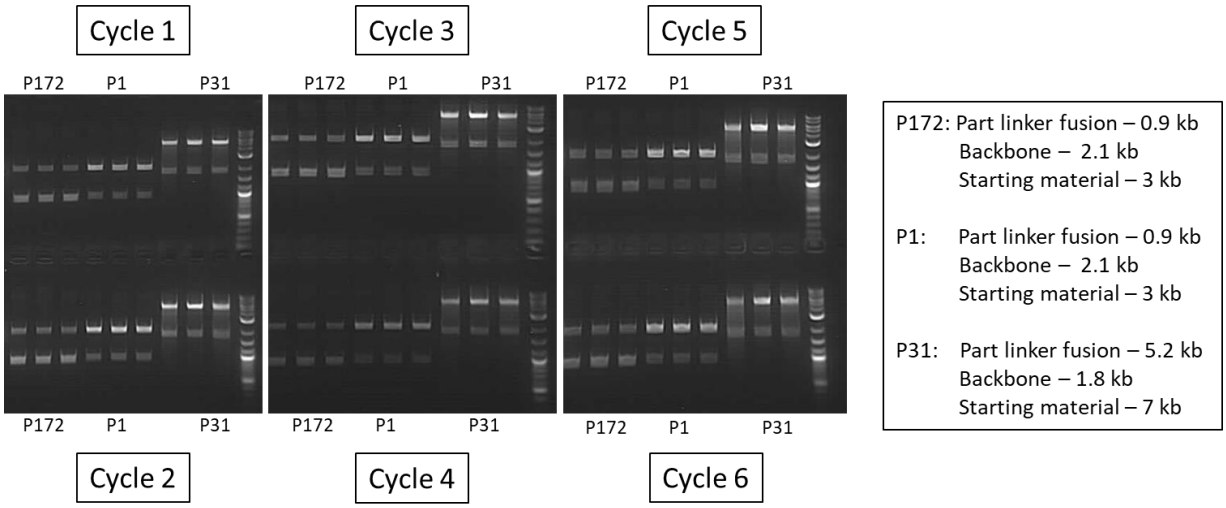


Figure 60: Visualisation of part linker fusion reaction products following each ligation cycle.

The appropriate DNA fragments were gel extracted and assembled following the standard procedure. Through the construction of vectors harbouring double antibiotic resistance markers assembly efficiency was calculated through the number of colonies present following selection on media containing the appropriate antibiotics (Table 22).

		Expt 1 - CampR				Expt 2 - KanR		
Cycle	CFU	Average	S.D	Relative Efficiency (%)	CFU	Average	S.D	Relative Efficiency (%)
1	384	446	44.21	112	312	413	42	117
	484				348			
	470				386			
2	368	415	34.27	104	380	457	36	129
	448				388			
	430				472			
3	352	371	35.41	93	412	457	36	129
	341				459			
	421				501			
4	232	255	24.07	64	251	268	65	76
	288				198			
	244				354			
5	254	362	96.81	91	232	323	64	91
	344				376			
	489				360			
6	344	399	41.35	100	340	354	69	100
	444				444			
	408				277			

Table 22: The effect of the number of digestion/ligation cycles on assembly efficiency. Colony counts were determined following transformation of assembly product and selection on plates containing the appropriate antibiotics to confirm successful vector assembly

The number of colonies present on selection plates at each time does not increase significantly throughout the time course experiment (Figure 61). This result and the visual analysis (Figure 60) of the part linker fusions suggest that there is no

enhancement in assembly efficiency achieved through cycling between ligation and digestion. This is likely due to the reported low stability of SapI in the reaction resulting in it being inactive following the first cycle of incubation at 37 °C. This reduction in the number of cycles can significantly enhance the throughput of the inABLE DNA assembly platform.

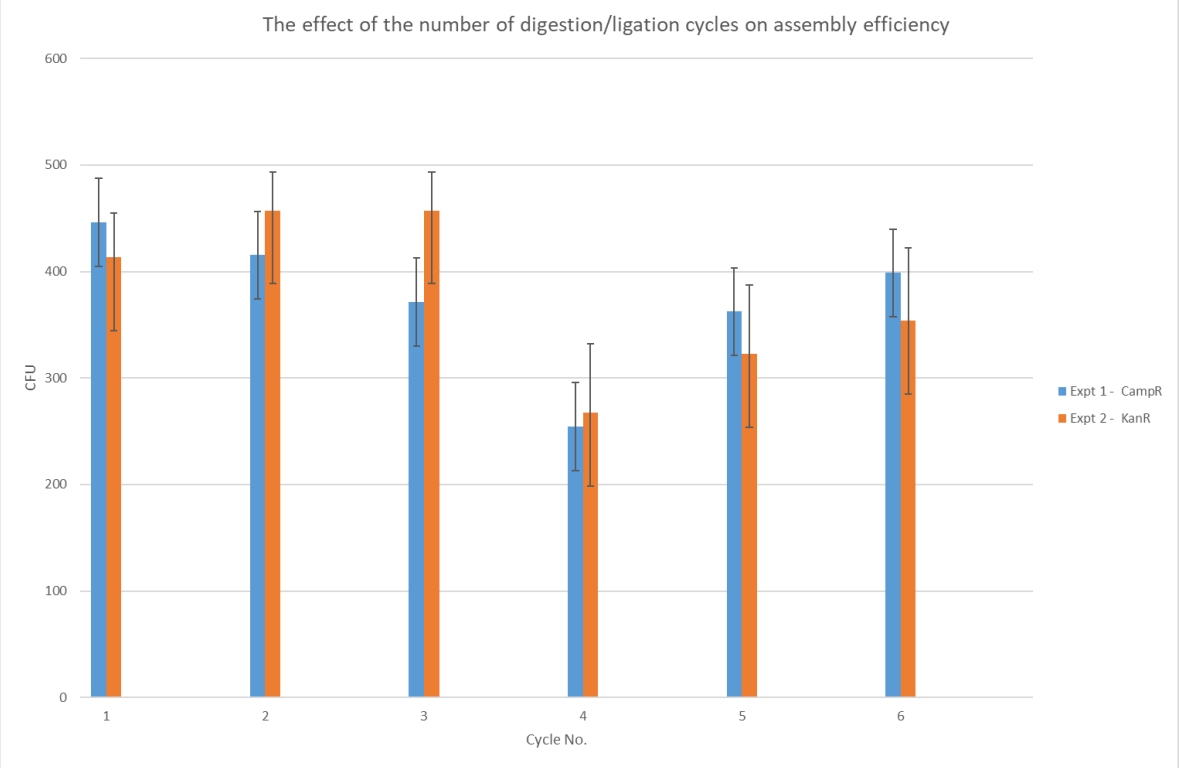


Figure 61: Average number of colony forming units from triplicate experiments on selection plates following each ligation stage within the part linker fusion reaction

4.2.2.2 Assembly accuracy

A limitation with the above strategy is that due to the double antibiotic selection approach only cells transformed with the expected construct will be viable on the selection media. This approach allows for calculation of assembly efficiency but not accuracy, which is a key aspect of any DNA assembly approach. To explore the ratio of positive assemblies to mis-assemblies a screen based on the expression of

enhanced green fluorescent protein (eGFP) was implemented. In this study a constitutive *E. coli* promoter (ProC)[181], an enhanced GFP variant from the pSEVA plasmid collection[182] and an *E. coli* backbone comprised of a CampR marker and a pMB1 origin of replication were assembled (Figure 62).

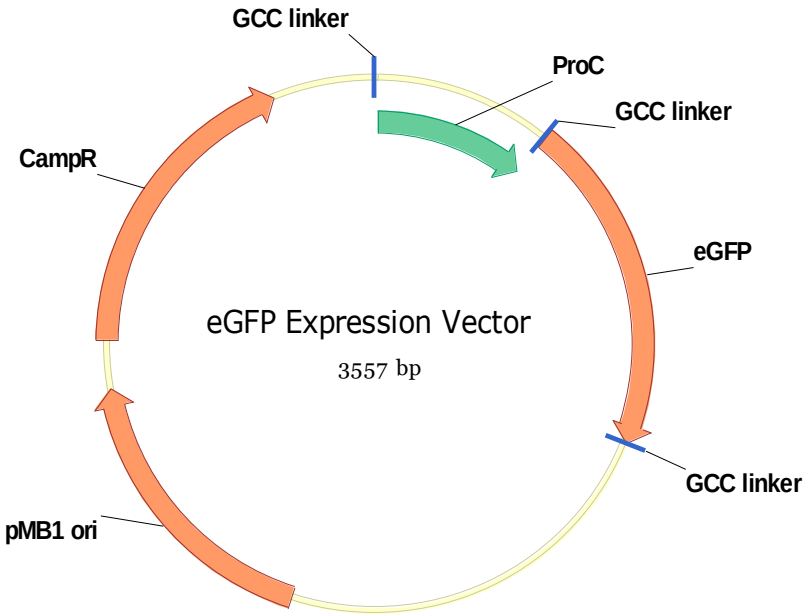


Figure 62: Target vector construct comprised of three DNA fragments. Only cells containing the expected construct will be viable on chloramphenicol containing media and produce eGFP.

The minimal conditions identified in the assembly efficiency test (1 cycle of digestion and 1 cycle of ligation) were directly compared to the standard part linker fusion reaction. Following the part linker fusion and assembly reactions, transformants were selected for on chloramphenicol containing plates. Only transformants containing the expected construct would be viable on this media and fluoresce when the plates are visualised using a Safe Imager 2.0™ blue light transilluminator (Invitrogen). In parallel, a negative control construct was prepared in which the promoter was purposely omitted. This was performed to ensure that read through from a promoter in the *E. coli* backbone (such as the one controlling transcription of the CampR gene) could not be responsible for the observed

fluorescence. When this negative control construct was introduced into *E. coli* and the resultant plate visualised no fluorescence was observed.

Cycling	Total colonies	Fluorescent colonies	Average colony number	S.D	Average Fluorescent colonies	S.D	Assembly accuracy (%)
Standard	123	117	107.7	18.3	99.0	16.6	92.0
	118	103					
	82	77					
Minimal	112	106	94.3	12.6	84.3	15.5	89.4
	87	76					
	84	71					

Table 23: Colony counts following transformation of assembly product, selection on plates containing the appropriate antibiotics and visualisation of GFP output to confirm successful vector assembly.

As observed in the assembly efficiency experiment the number of colonies present on selection plates between a minimal and standard cycling regime does not increase significantly. In this experiment however, the presence of transformants on the selection plate is not 100% indicative of a fully assembled construct. Crucially, a similar number of fluorescing colonies is observed in both standard and minimal cycling regimes, highlighting that the implementation of a minimal cycling approach does not negatively impact assembly accuracy. (Table 23; Figure 63 A and B).

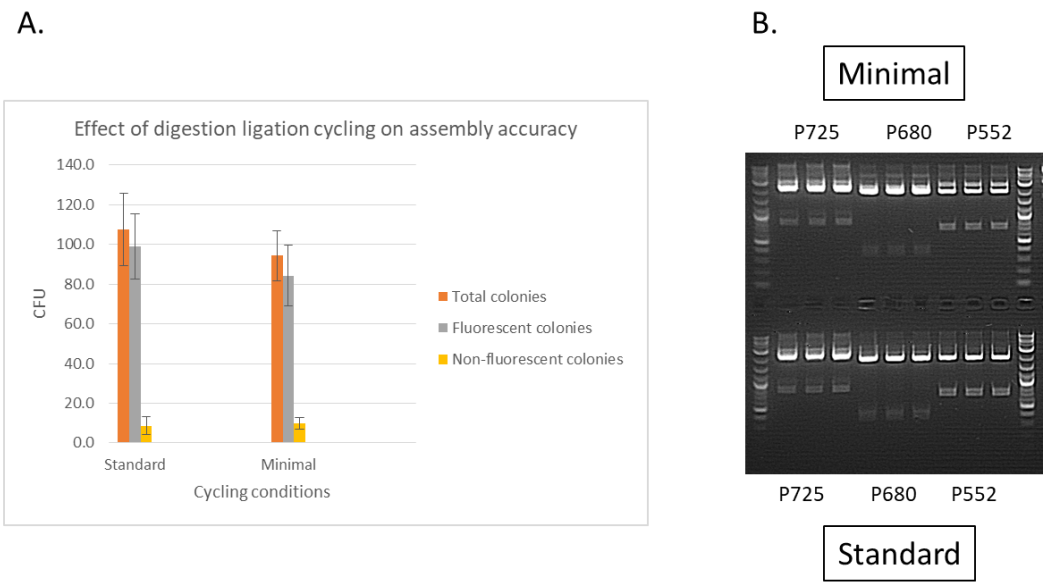


Figure 63: The effect of part linker fusion cycling on assembly accuracy from triplicate reactions.

4.2.3 Conclusions

This study highlights that reducing the number of cycles in the part linker fusion reaction does not have a significant effect on either the efficiency or the accuracy of the inABLE DNA procedure however the reduced cycling time can dramatically increase the throughput of the procedure. As previously described SapI is not recommended for digestions of longer than one hour, meaning that following the first round of incubation at 37 °C the SapI is unlikely to be active in subsequent rounds of digestion. Secondly, in the part linker fusion reaction a 10 to 1 molar excess of the linkers to the truncated part is present. Therefore, during the ligation reaction, ligation of the linkers to the truncated part should be favoured over re-ligation of the truncated part into the vector backbone. It is important to note that EarI is expected to be more stable in the part linker fusion reaction. However, since EarI has a recognition site 1 bp shorter than SapI (CTCTTC and GCTCTTC respectively) its use places more stringent sequence requirements on fragments which can be assembled using the inABLE approach increasing the likelihood for site-directed mutagenesis requirement.

The ability to reduce the part linker fusion reaction from almost six to two hours can have a significant effect on the timeline with which vectors can be constructed using the inABLE platform and also further open the door to automation. One could assume an additional acceleration in vector construction by coupling this approach with a faster growing host than *E. coli* such as *Vibrio natriegens* which has recently been characterised as a molecular biology host with a doubling time of <10 minutes[183].

4.3 Multiple rounds of SapI mediated assembly through protection of SapI sites using phosphorothioate bonds (Phospho-nested DNA assembly)

4.3.1 Introduction

One limitation of the current inABLE technology is that the product of one assembly reaction cannot be directly used in a second round of assembly. This limits the potential to, for example, add additional genes to a previously assembled heterologous pathway. This is a result of the utilisation of type IIs restriction enzymes in the part linker fusion reaction the nature of which result in no recognition sites being present in the final assembly product. The BioBricks platform (reviewed in Section 1.7.1.2) for example addresses the issue through the use of a four restriction endonuclease approach in which pairs of enzymes cleave different sites to produce compatible cohesive ends. Following ligation, a non-cleavable scar is formed while active sites flank the assembled DNA and can be used to add further fragments. This four enzyme strategy however places significant sequence constraints on the parts being assembled meaning fragments will often require to be modified to mutate non-compatible sequences prior to assembly. The Golden Gate assembly has also been adapted for iterative DNA assembly as described in the MoClo and Golden braid methodologies[184, 185] both of which rely on alternative use between two type IIs enzymes in a multi-level system that allows the products of one assembly reaction to be used in further reactions.

One possible approach within the inABLE workflow (as described in section 3.2) is to alternate between two type IIs enzymes for each round of assembly (i.e. 1st round using SapI and append linkers containing EarI recognition sites for a second round of assembly). This approach has previously been exemplified and is compatible with the inABLE DNA assembly framework as each truncated part is flanked by a SapI site and by default also an EarI site. Whilst this approach allows for the addition of parts to previously constructed plasmid without resorting to PCR you are limited to two rounds of assembly as it is not possible to use SapI containing linkers during the EarI mediated second round of part linker fusion since they would be cleaved. The part sequence also requires to be void of both EarI and SapI recognition sites decreasing the flexibility of the approach.

The possibility to mediate sequential rounds of SapI mediated DNA assembly through the introduction of SapI sites protected by phosphorothioate bonds in a first round of assembly on linker fragments was therefore explored as a potential solution. Following replication of the assembly product in *E. coli* the phosphorothioate bonds will be lost making SapI sites available for a second round of assembly. Such an approach can be repeated for unlimited rounds of assembly. A similar approach has recently been implemented for DNA assembly approaches which use the Dcm methylation sensitive type IIs restriction endonuclease BsaI such as in the Golden Gate strategy and its derivatives[186]. This approach utilises the introduction of BsaI sites on a synthetic fragment of DNA. It has been shown that the methylation of either of the cytosines the BsaI recognition site GGTCTC confers resistance to BsaI cleavage[187], in the same way that restriction enzyme-producing organisms protect their own DNA through the expression of site specific methylases. This can be leveraged to prevent cleavage of BsaI recognition sites contained on linker sequences in an initial round of assembly. The protection of the BsaI sites is then lost following introduction of the assembly product into an *E. coli* host which does not maintain the methylation pattern freeing BsaI sites for subsequent rounds of assembly. As SapI or EarI are not reported to be sensitive to DNA methylation the potential to utilise the

presence of phosphorothioate bonds to protect SapI sites on linker sequences will be explored.

4.3.2 Results

4.3.2.1 Protection of SapI recognition sites via phosphorothioate bonds

Recently, Zhao et al [188] explored the effect of phosphorothioate or 2'-O-Methyl nucleotide substitutions on the cleavage efficiency of six type II endonucleases. It was found that introduction of nucleotide substitutions can modulate the efficiency of enzymatic cleavage in a position dependent manner. Only moderate inhibition from phosphorothioate bond introduction was observed regardless of the position of the substitution in the case of XhoI and XbaI enzymes. Significant reduction in enzymatic cleavage was observed when the phosphodiester bond cleaved by the restriction endonucleases SpeI and EcoRI was substituted with a PS bond. Finally, complete inhibition of cleavage bond for PstI and SphI was observed once again when the bond cleaved was modified to a PS bond. In general, 2'-O-Methyl nucleotide substitutions appeared a more robust method for restriction endonuclease site protection and this modification could be explored further for use in SapI protection in the future. No similar study has been performed using SapI or other type IIs restriction endonucleases.

To explore if SapI recognition sites can be protected through the replacement of phosphodiester bonds with phosphorothioate bonds a total of eight, 100 bp DNA fragments were synthesised containing a central SapI site and each bond within the recognition site and the cleavage sites comprised of either a phosphorothioate or phosphodiester bond (Table 24).

Name	Sequence
PS_1	5'-GGCAGTTATTGGTGCCCTTAAACGCCTGGTGGGATCC**G*CTCTTC**GGGCTATATCTCCTTCTTAAAGTTAAACAAAATTATTTCTAGAGGG -3'
PS_2	5' - GGCAGTTATTGGTGCCCTTAAACGCCTGGTGGGATCC**GC*TCTTC**GGGCTATATCTCCTTCTTAAAGTTAAACAAAATTATTTCTAGAGGG – 3'
PS_3	5' - GGCAGTTATTGGTGCCCTTAAACGCCTGGTGGGATCC**GCT*CTTC**GGGCTATATCTCCTTCTTAAAGTTAAACAAAATTATTTCTAGAGGG – 3'
PS_4	5' - GGCAGTTATTGGTGCCCTTAAACGCCTGGTGGGATCC**GCTC*TTC**GGGCTATATCTCCTTCTTAAAGTTAAACAAAATTATTTCTAGAGGG – 3'
PS_5	5' - GGCAGTTATTGGTGCCCTTAAACGCCTGGTGGGATCC**GCTCT*TC**GGGCTATATCTCCTTCTTAAAGTTAAACAAAATTATTTCTAGAGGG – 3'
PS_6	5' - GGCAGTTATTGGTGCCCTTAAACGCCTGGTGGGATCC**GCTCTT*C**GGGCTATATCTCCTTCTTAAAGTTAAACAAAATTATTTCTAGAGGG – 3'
PS_7	5' - GGCAGTTATTGGTGCCCTTAAACGCCTGGTGGGATCC**GCTCTTC***GGGCTATATCTCCTTCTTAAAGTTAAACAAAATTATTTCTAGAGGG – 3'
PS_8	5' – GGCAGTTATTGGTGCCCTTAAACGCCTGGTGGGATCC**GCTCTTC**GGGC*TATATCTCCTTCTTAAAGTTAAACAAAATTATTTCTAGAGGG – 3'

*Table 24: DNA fragments used to explore protection of EarI/SapI recognition sites (in bold) through the introduction of phosphorothioate bonds. Recognition site in bold and phosphorothioate bonds denoted by *.*

Analysis of digestion products by capillary electrophoresis revealed that EarI/SapI sites are only successfully protected through the introduction of phosphorothioate bonds at the cleavage site (Figure 64, lanes 9 and 18), introduction of the bond at any site within the recognition sequence has no noticeable effect on SapI or EarI activity.

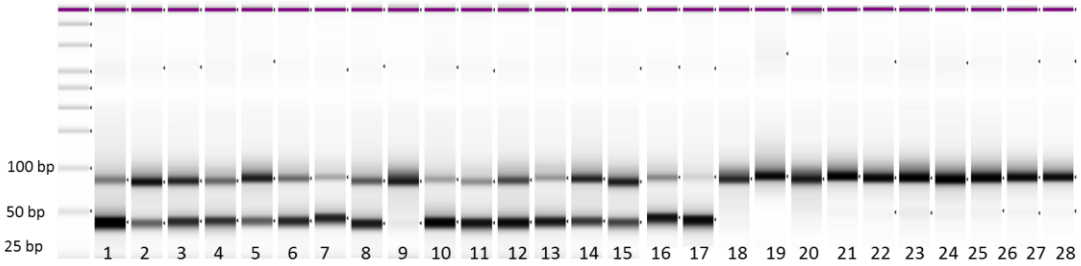


Figure 64: Capillary electrophoresis analysis of products following SapI and EarI digestion. Lanes 1-9 show, 1 – No phosphorothioate bond; 2 – PS_1, PS_2, PS_3, PS_4, PS_5, PS_6, PS_7, PS_8; digested with SapI. 10-18 digested with EarI and 19-28 H_2O controls

This initial result opens the door to potentially implementing a "phospho-nested" DNA assembly strategy. Further study of alternative modifications to inhibit

SapI/EarI modification may further expand the flexibility of this approach for iterative rounds of DNA assembly.

4.3.2.2 Design of initial phospho-nested approach.

An initial two part assembly was implemented to explore a potential "phospho-nested strategy". In previous nested approaches specific POA and LOA sequences were ordered to introduce EarI recognition sites to be utilised in a second round of DNA assembly (Section 3.2). These sequences were modified to now introduce phosphorothioate protected SapI recognition sites rather than EarI sites (Table 25).

Linker	**Sequence**	**Description**
LOA204 - nested	Gcc*gcccGAAGAGCacaggtggcacttttcggggaaatgtgcgcggaa Cgg*gCTTCTCGtgtccaccgtgaaaagcc	SapI recognition sites in bold.
POAxxx - nested	gtgctggtGCTCTTCg*ctg ctatagacgtaacaaccacgaccaCGAGTTGcgac*gac	Phosphorothioate bonds denoted by *

*Table 25: DNA fragments used to explore protection of EarI/SapI recognition sites through the introduction of phosphorothioate bonds. SapI recognition site in bold and bonds denoted by **

As seen above, one limitation of this approach is the requirement to add additional scar sequences with each round assembly. The reason for the further scar sequence introduction is that phosphorothioate bonds require to be introduced during DNA synthesis. The bonds cleaved in a scarless system would be not generated during DNA synthesis rather through the ligation of the linker to the truncated part during the part linker fusion reaction at which point it is not possible to introduce a phosphorothioate bond.

This is not an issue with the annealing of the LOA in which a GCC scar is added using the standard technology and therefore parts are designed so this does not cause issues (for example it is standardly introduced directly following a stop codon in a gene part). However, the introduction a scar during the ligation of the POA has the

potential to be problematic. The annealing of a POA to a truncated part is generally scarless. This is because the addition of any scar at this point is within the sequence of the full length part (i.e. within the coding sequence in a gene part) and therefore requires to be avoided. This is a limitation of the proposed "phospho-nested" approach and requires careful fragment design to ensure any scar sequences introduced do not have a negative effect on construct performance.

To alleviate the issue during this proof of principle study the DNA fragment previously constructed for the exploration of the effect of 16 bp overhang secondary structure on assembly efficiency (Section 3.1.4) was utilised in this experiment. This part comprises of a KanR marker cassette with a 5' extension of 49 bp, allowing for the introduction of scar sequences during POA ligation without the disruption of the 5' UTR which may impact cassette performance.

Part linker fusion reactions were performed as described previously using "phospho-nested" linkers (Table 25) with and without phosphorothioate bonds. Selection of transformants on plates containing Amp + Kan allowed for the direct comparison of assembly efficiency in this two part approach.

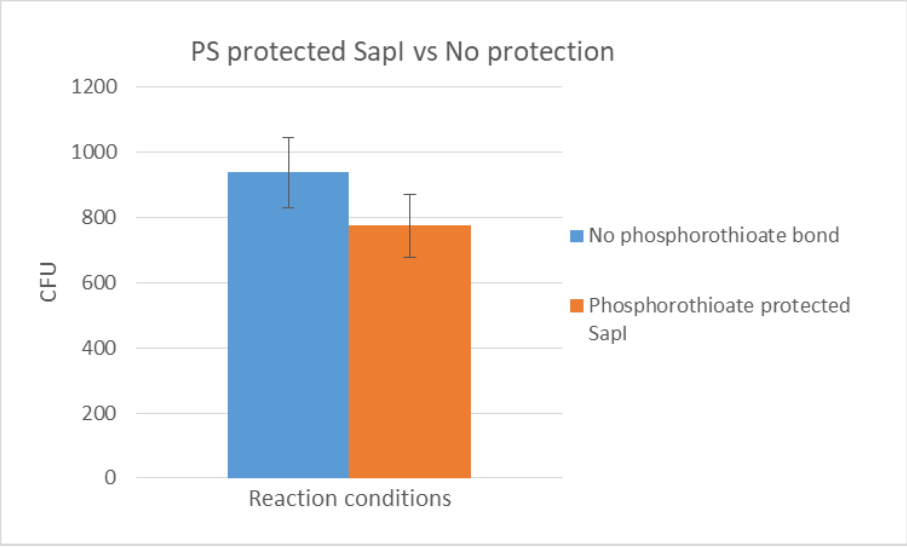


Figure 65: Comparison of the utilisation of nested linkers containing SapI sites plus or minus phosphorothioate protection. Average colony counts following selection of transformants on LB media supplemented with Amp and Kan calculated from triplicate experiments.

It was expected that without the phosphorothioate mediated protection of SapI sites within the linkers that the assembly efficiency would dramatically decrease due to the cleavage of SapI sites contained within the linker sequences resulting in less efficient DNA assembly. However, assembly efficiency was largely maintained with or without phosphorothioate introduction (Figure 65). Subsequent treatment of assembly reactions with SapI resulted in a complete loss of transformants in the assembly products lacking phosphorothioate protection whilst assembly products with protection resulted in a similar number of transformants as previously. This confirms that the phosphorothioate bonds are conferring resistance to SapI in the assembled product (Figure 66).

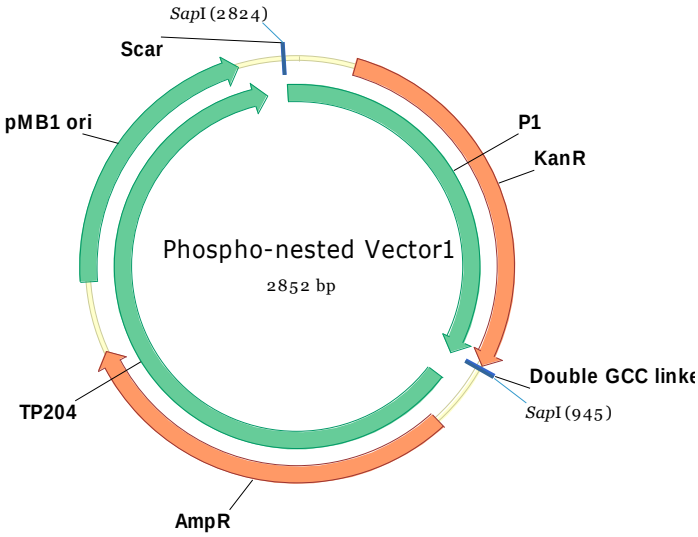


Figure 66: Vector constructed using a "phospho-nested" strategy. This strategy results in SapI recognition sites which are protected by phosphorothioate bonds being introduced during the assembly process. Following replication of the assembly product in E. coli the phosphorothioate bonds are lost allowing the SapI sites to be utilised for a second round of inABLE assembly.

It is likely that a difference in assembly efficiency between the protected and standard linker was not observed due to the stability of SapI within the part linker fusion reaction. SapI is characterised as a restriction enzyme which is not recommended for incubation times of longer than 1 hour[180]. Therefore, towards the end of the part linker fusion cycling program only ligation reactions would be

occurring allowing for repair of linkers and formation of the required part linker fusions. Analysis of assembly products from both protected and non-protected experiments through SapI digestion and DNA sequencing confirmed that even without protection the expected construct was being formed and the introduced recognitions sites can be utilised for additional rounds of inABLE based vector construction.

4.3.2.3 Utilisation for bioengineering

In order to make the phospho-nested approach compatible with the construction of biosynthetic pathways whilst maintaining its compatibility with previously cloned inABLE truncated parts a revised construction strategy was implemented.

To utilise this approach for biosynthetic pathway generation a fully functional vector would require to be constructed to which additional parts can be added as required. Ideally, this should be achieved without the introduction of the associated scar sequences in potentially problematic regions. If a similar approach to the previously exemplified nested strategy (Section 3.2) was adopted this would not feasible as the 1st part of the assembly reaction product remains truncated and likely non-functional. In addition, scar sequences would be introduced during POA ligation meaning they would also impact part performance.

To avoid this a Phospho-nested LOA comprised of a short synthetic *E. coli* terminator[89] flanked by two phosphorothioate protected SapI recognition sites specific for the expression vector backbone (CampR/pMB1 ori) was constructed (Figure 67). The design of this fragment ensures that each of the parts assembled are full length and as such will result in a functional expression vector to be constructed and a terminator flanked by SapI sites introduced after the final gene of the construct.

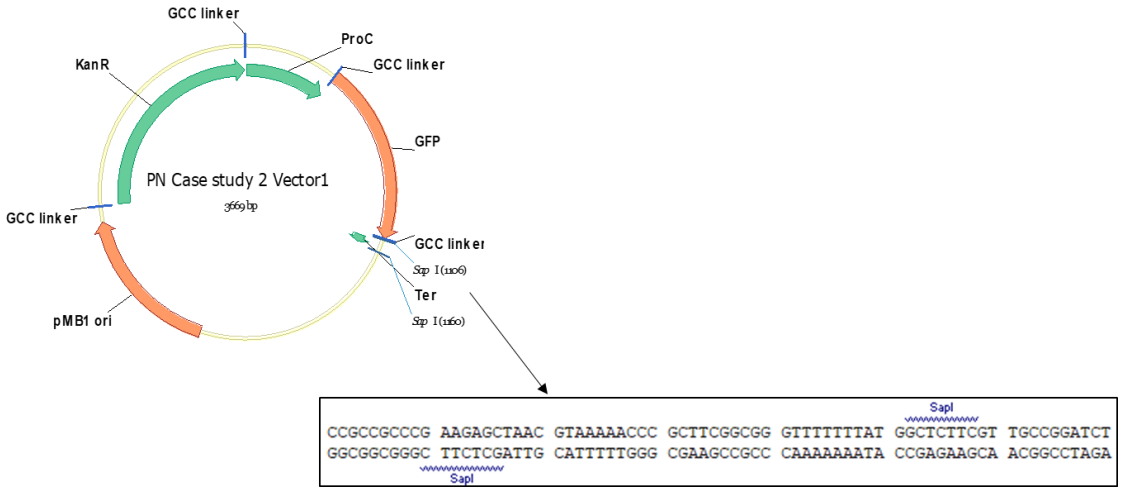


Figure 67: Product of phospho-nested assembly reaction. E. coli terminator flanked by SapI recognition sites which can be utilised in subsequent rounds of inABLE assembly.

If additional genes require to be added to the construct, then a part linker fusion reaction using SapI can be performed. The T7 terminator flanked by SapI sites can be relocated to follow the additional genes added in this second round of assembly. This approach also ensures that the additional scar sequences which are a result of the phospho-nested strategy are introduced out with gene coding sequences or key regulatory regions.

Linker	Sequence	Description
LOA PN1	GCC*GCCCGAAGAGCTAACGTAAAAACCCGCTTCGGCGGGTTTTTTTATGGCTCTTCG*TTGCCGGATCTGCATCGCAGGATGCTGCTGGCTA CGG*GCTTCTCGATTGCATTTTTGGGCGAAGCCGCCCAAAAAAATACCGAGAAGCAAC*GGCCTAGACGTAGCG	LOA containing terminator flanked by protected SapI sites.

Table 26: Phospho-nested linker sequences. DNA contains a short E. coli terminator flanked by SapI sites which are protected with phosphorothioate bonds (*).

To explore this strategy four DNA fragments were assembled, a KanR marker fragment, pMB1 origin of replication, the ProC *E. coli* constitutive promoter and an eGFP coding sequence along for selection of transformants containing the expected assembly through resistance to Kanamycin and production of eGFP. The fragments were assembled as standard except LOA PN (Table 26) was annealed to the GFP CDS part.

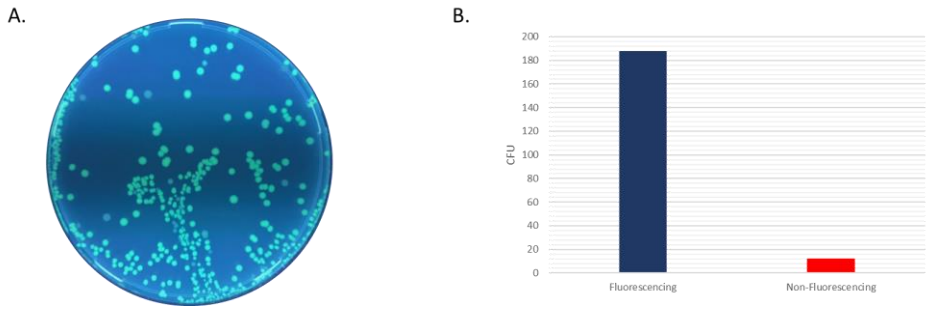


Figure 68: Analysis of transformants from Phospho-nested assembly reaction: A. Visualisation of fluorescent vs non-fluorescent colonies using blue light transilluminator. B. Fluorescent vs non-fluorescent colonies.

Colonies following transformation were visualised using a Safe Imager 2.0 blue light transilluminator and the number of fluorescent colonies to non-fluorescent colonies compared. 94% of the colonies characterised were fluorescent suggesting correctly assembled constructs (Figure 68). Sanger sequencing confirmed the successful introduction of the synthetic terminator flanked by SapI sites after the GFP gene confirming the construction of a functional expression vector which is compatible with additional rounds of inABLE assembly.

When additional parts require to be added to this vector TP SapI is again used in the part linker fusion reaction. As the previously assembled DNA has replicated in *E. coli* the phosphorothioate modifications will be lost and the introduced SapI sites unprotected

4.3.3 Conclusions

In this section the use of phosphorothioate bonds for the protection of SapI recognition sites with the goal of recursive DNA assembly was explored. Such an approach builds on the previously exemplified nested inABLE strategy (Section 3.2). This approach has the potential to allow for unlimited rounds of inABLE assembly using a single restriction enzyme provided carefully designed DNA fragments are employed.

The full protection of SapI recognition sites was defined through using short synthetic fragments of DNA with phosphorothioate bonds introduced throughout the recognition site and also at the cleavage site. It was found that only the introduction of a phosphorothioate bond within the cleavage site inhibits SapI digestion. The use of protected SapI recognition sites within the inABLE DNA assembly workflow was then validated in initial proof of principle experiments confirming the ability to construct functional vectors which are compatible with additional rounds of inABLE assembly. The development of the phospho-nested strategy allows for recursive DNA assembly and is a valuable addition to the inABLE toolbox.

5. Proteomics for metabolic pathway optimisation

5.1 Label free protein identification – Proof of principle

5.1.1 Introduction

The utilisation of mass spectrometry based proteomics approaches for the assessment of heterologous pathway implementation has the potential to accelerate the engineering of microbes for industrial bioprocesses. Many factors contribute to pathway performance which would not be identified when only monitoring product titers. For example, a vital consideration when introducing a multi gene pathway is the fine tuning of protein production at each stage of the pathway[189]. Over production of each pathway protein can result in significant metabolic burden placed on the cell thus diminishing productivity whilst the underproduction of pathway enzymes may result in the formation of pathway bottlenecks. The optimal balance of pathway proteins is difficult to predict and is therefore often addressed through the construction of libraries of genetic elements. The level of expression is manipulated through varying factors including promoter strength, copy number, RBS strength and gene codon usage. The ability to directly monitor and quantify multiple proteins in the engineered system therefore can greatly benefit attempts to optimise heterologous metabolic pathways.

One challenge associated with characterising an engineered strain at the protein level is the lack of high throughout techniques which are readily available. The separation of a specific protein of interest from the cellular proteome can be time consuming and labour intensive as proteins have similar physical characteristics – such as molecular weight (mW) and isoelectric point (pI) – and are produced at significantly varying concentrations. Immunoblot assays have traditionally been used for the selective identification of proteins. These approaches are fast, convenient and for the

same protein can be multiplexed, however can prove challenging when multiple proteins require to be detected.

Advances in LC-MS and LC-MS/MS based approaches have allowed for the identification and quantification of multiple proteins from cell lysates. Multiple techniques for both relative and absolute quantification including the inclusion of an internal synthetic peptide and isotopic labelling through the incorporation of labelled amino acids into the protein during cell growth (SILAC)[138], chemical labelling of proteins prior to digestion (iCAT)[139] or enzymatic labelling of the peptides following digestion (iTRAQ)[140]. Despite the success of these techniques, label-free approaches are often favoured as quick and cost effective options requiring no additional stages of sample preparation making it extremely compatible with previously developed strain engineering platforms. These approaches also avoid the limitations of labelling approaches including the cost of reagents, the efficiency of the labelling reaction and limitation of the number of samples that can be analysed.

In this chapter the implementation of a MS-based platform for the label free quantification of multiple pathway proteins will be explored. The implementation will initially focus on the robust identification of target proteins using the MS^E platform before attempting relative quantification of target proteins from engineered microbes. The successful implementation of such a strategy has the potential to complement the previously exemplified pathway construction tools and accelerate the engineering of strains for industrial bioprocesses.

5.1.2 Results

5.1.2.1 High and Low Energy scans

The MS^E protein analysis approach developed by Waters Corporation, is a data independent approach in which there is no precursor peptide selection for fragmentation and instead the instrument alternates between high energy and low energy scans. During the low energy scan the *m/z* and retention time of the intact

eluting tryptic peptide is collected and as such the collision energy is set to minimise fragmentation within the collision cell. It should be noted that fragmentation of a peptide can also occur within the source and therefore parameters such as cone voltage should also be considered during optimisation. During the elevated energy scan the goal is to collect information on fragment ions and therefore the collision energy should be such that the intact peptides are efficiently fragmented. However unlike in a data dependent acquisition (DDA) when the collision energy for any given precursor is defined by the *m/z* and the charge in an MS^E experiment multiple precursors may be co-eluting from the column over a range of *m/z* and charges. A collision energy ramp is therefore implemented which is suitable to fragment peptides spanning the range of *m/z* and charges expected from a tryptic digest. As there is no mass selection, bias towards the identification of highly abundant peptides is reduced and fragmentation data is generated for all eluting peptides.

During an MS^E acquisition three functions are generated within the raw data file by the Mass Lynx software. Function one comprises of low energy information, function two contains high energy data whilst function three contains lockmass data which is a reference compound infused throughout the run and used for data correction post acquisition.

To perform an initial study of generating MS^E data using a single peptide a sample of luteinizing hormone-releasing hormone (LHRH) was analysed through LC-MS^E on a Synapt G2 instrument. Low and high energy data generated showing the parental ion and the generated fragment ions can be found below (Figure 69).

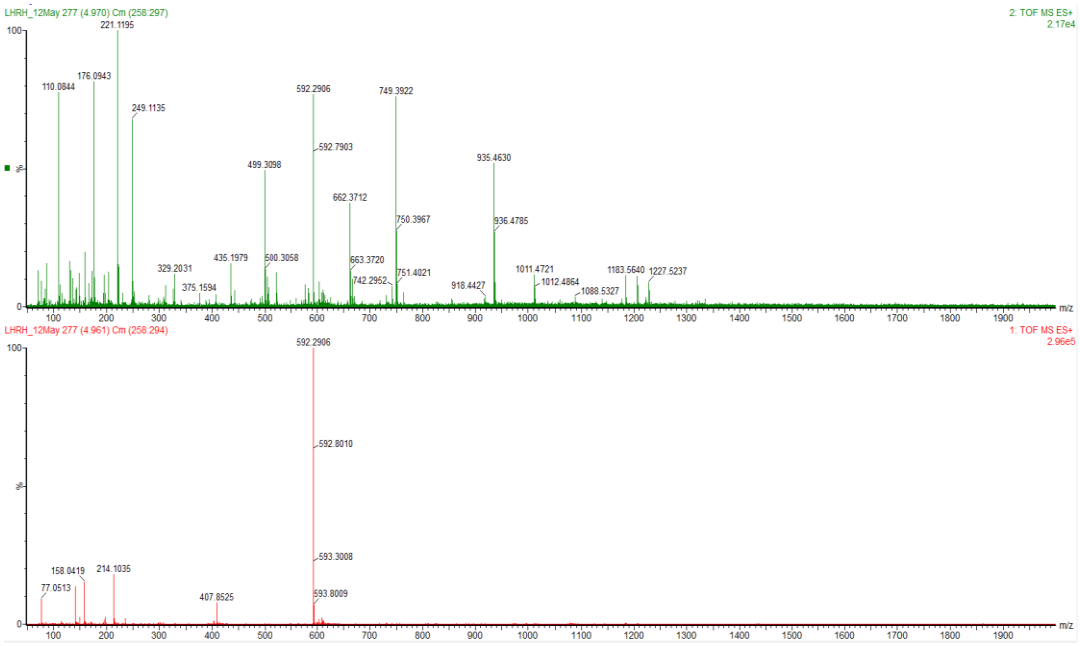


Figure 69: High (upper) and low (lower) energy scans of LHRH on a Synapt G2 instrument mimicking a MS^E analysis which utilises high and low energy scans to generate intact and fragmentation data on analysed peptides.

5.1.2.2 MS^E based protein identification – Single protein

To further explore MS^E data acquisition, tryptically digested BSA was purchased (NEB) and analysed using LC-MS^E on a Synapt G2 instrument and the resultant MS^E data analysed for protein identification in PLGS (Reviewed in Section 1.7.2.4) with the data searched against the SwissProt protein database. A lockmass of Leucine Enkephalin (Leu-Enk) was infused throughout the run and used for post-acquisition mass correction with chromatographic separation of the peptides achieved over 30 minutes. Over triplicate experiments the average sequence coverage achieved was 31%. An example of the PLGS protein identification output for this analysis is found below (Figure 70).

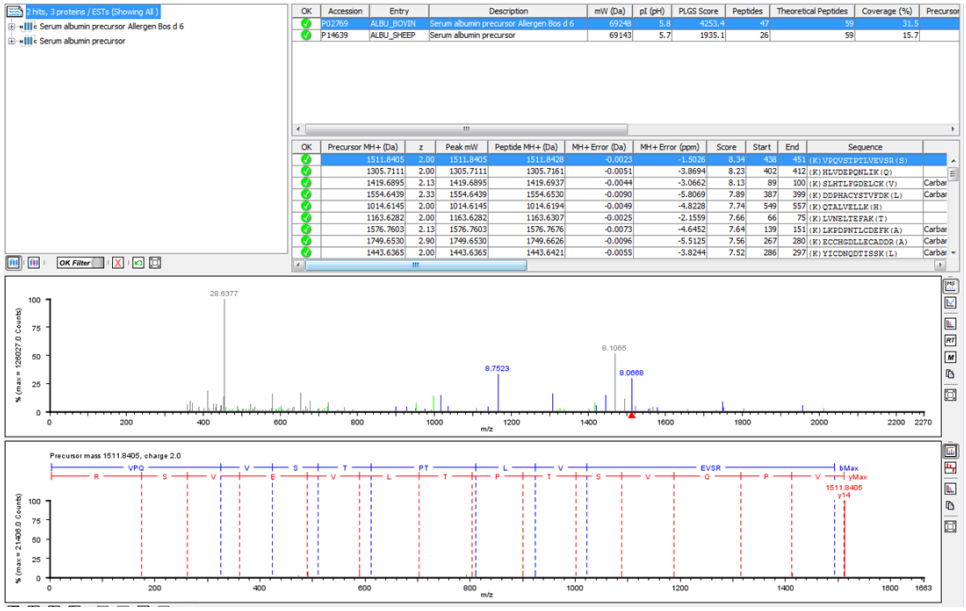


Figure 70: Example output of PLGS processing of MS^E data generated from the analysis of a commercially available tryptic digest of BSA.

5.1.2.3 MS^E based protein identification – Multiple proteins

To further expand on this initial MS^E based protein identification proof of principle a MassPrep standard was purchased form Waters Corporation. This is an equimolar mix of four tryptically digested proteins; alcohol dehydrogenase from *S. cerevisiae* (UniProt P00330 and P00331), phosphorylase b from rabbit muscle (UniProt P00489), bovine serum albumin (UniProt P02769), and enolase from *S. cerevisiae* [UniProt P00924 and P00295]). The mixture was analysed using the same workflow that was utilised for BSA analysis.

Accession	Entry	mW (Da)	pI (pH)	Peptides	Theoretical Peptides	Coverage (%)	Precursor RMS Mass Error (ppm)	Products	Products RMS Mass Error (ppm)
P00330	ADH1_YEAST	36668	6.2739	20	25	44.38	3.05	133	7.6165
P00489	PHS2_RABIT	97096	6.7954	39	82	44.8	3.92	287	8.0808
P00924	ENO1_YEAST	46642	6.1538	16	36	29.5	4.08	90	7.5545
P02769	ALBU_BOVIN	69248	5.7583	23	64	32.7	4.45	110	9.3348

Table 27: Output of PLGS analysis of data generated through MS^E characterisation of MassPrep protein standard which comprises of peptides generated from the tryptic digestion of S. cerevisiae alcohol dehydrogenase and enolase, rabbit phosphorylase B and BSA.

The data was searched against the SwissProt protein database and returned identifications for the four target proteins. The proteins characterised ranged in size from between 37 kDa and 97 kDa with an average protein coverage of 38% achieved (Table 27). This was achieved in a single analysis using as little as 180 ng of protein with limited chromatographic optimisation.

5.1.3 Conclusions

In this section a data independent protein identification workflow has been implemented and proof of concept protein identifications performed for a single protein and a mix of four proteins. The MS^E acquisition approach was implemented on a Synapt G2 instrument with the appropriate chromatography and the data analysis steps to analyse MS^E data for protein analysis also implemented. This initial work provides the framework for the analysis of increasingly complex mixtures using this approach in the subsequent sections of this chapter.

5.2 Label free protein identification – The reverse glyoxylate shunt

5.2.1 Introduction

To explore the use of the data independent proteomics workflow to guide strain engineering approaches the implementation of a reverse glyoxylate shunt in *E. coli* was identified as test case. The glyoxylate shunt avoids the decarboxylation steps of the TCA cycle and therefore allows the production of TCA cycle intermediates from acetyl-coA without carbon loss. The operation of this pathway in an acetyl-CoA producing direction however had not been described until Mainguet et al in 2013[190]. The introduction of a rGS from malate to isocitrate into *E. coli* requires the expression of two heterologous enzymes, *Methylococcus capsulatus* malate thiokinase (MTK) and *Rhodobacter sphaeroides* malyl-CoA lyase (MCL), along with the overexpression of the native *E. coli* isocitrate lyase (ICL) (Uniprot numbers Q607L8/9, B9KLE8 and P0A9G6 respectively). The introduction of this synthetic pathway into *E. coli* was demonstrated through the conversion of four carbon TCA intermediates such as malate and succinate into oxaloacetate plus two molecules of acetyl-CoA without carbon loss (Figure 71). Common sugars are standardly metabolised to acetyl Co-A through the decarboxylation of pyruvate and therefore, the ability to produce acetyl-CoA without the loss of carbon associated with the standard route through pyruvate is of interest for the production of a number of industrially useful compounds including fuels, fatty acids and isoprenoids. In this study the pathway from malate to isocitrate was explored.

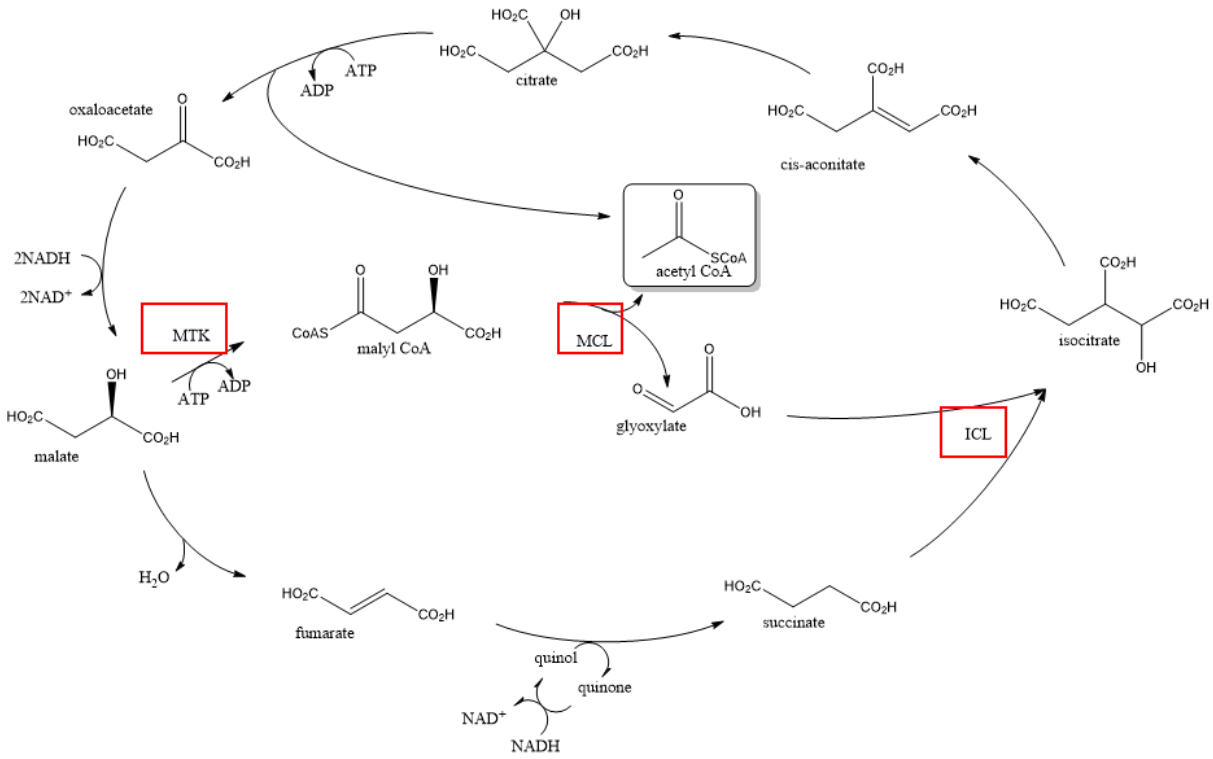


Figure 71: Conversion of four carbon TCA intermediates such as malate into oxaloacetate plus two molecules of acetyl-CoA through the reverse glyoxylate shunt. MTK, MCL and ICL highlighted by red boxes.

5.2.2 Results

5.2.2.1 Generation of purified rGS protein extracts

To validate the ability to detect the enzymes involved in the rGS through an MS^E approach each protein was independently purified. To allow for purification of the three target proteins, vectors were assembled using the standard inABLE procedure to incorporate an N-terminus histidine tag comprised of ten histidine residues (His_{10}). The His_{10} tag was introduced through the use of modified linker oligos. Design of modified linker oligos to incorporate His tag outlined in Table 28.

Gene	Modified linker oligo long	Modified linker oligo short
MTK	gccatgGGCCATCATCATCATCATCA TCATCATCATCACAGCAGCGGCCA TATCGACGACGACGACAAGaacatcc atgagtatcaggcaaaagaac	ctcatggatgttcatCTTGTCGTCGTCGTCGATA TGGCCGCTGCTGTGATGATGATGATGA TGATGATGATGATGGCCCAT
MCL	gccatgGGCCATCATCATCATCATCA TCATCATCATCACAGCAGCGGCCA TATCGACGACGACGACAAGagctttcg tctgcagcctgcaccgcctgc	gcagacgaaagctcatCTTGTCGTCGTCGTCGA TATGGCCGCTGCTGTGATGATGATGAT GATGATGATGATGATGGCCCAT
ICL	gccatgGGCCATCATCATCATCATCA TCATCATCATCACAGCAGCGGCCA TATCGACGACGACGACAAGaaaaccc gtacacaacaaattgaagaat	tgtacgggttttcatCTTGTCGTCGTCGTCGATA TGGCCGCTGCTGTGATGATGATGATGA TGATGATGATGATGGCCCAT

Table 28: Design of linker oligos for incorporation of N-terminal His tag. His tag highlighted in capitals. 3' end of gene underlined.

E. coli expression vectors were constructed through a standard two part inABLE assembly using the His_{10} sequence containing linkers which are specific to each gene truncated part (Figure 72).

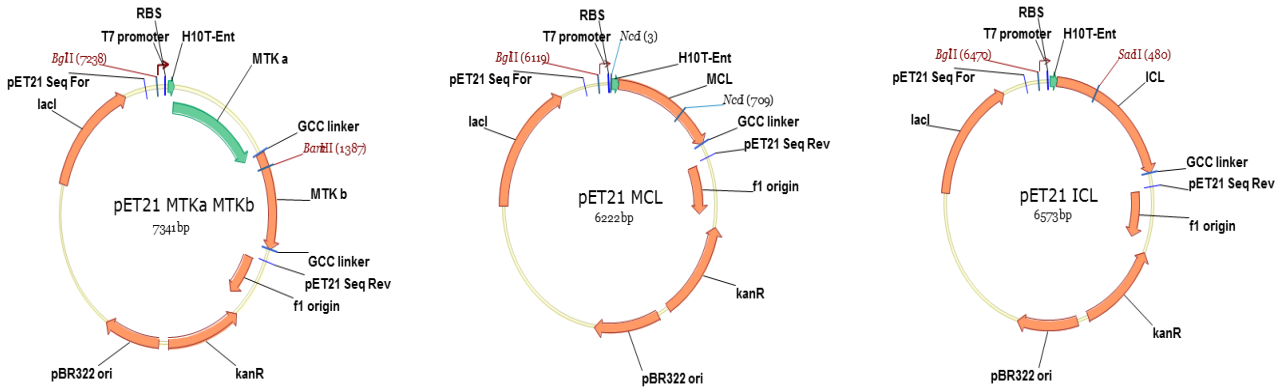


Figure 72: E. coli vectors constructed to introduce N-terminal His tag to reverse glyoxylate shunt enzymes. Restriction sites used for screening annotated on maps.

Part linker fusions were performed, and gene parts assembled with a pET21 backbone part in a two part assembly. Constructs were screened via restriction analysis and positives confirmed via sequencing (Figure 73).

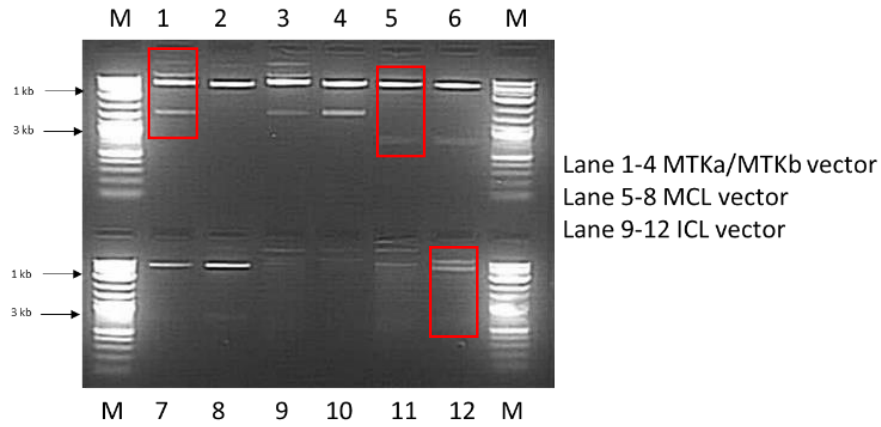
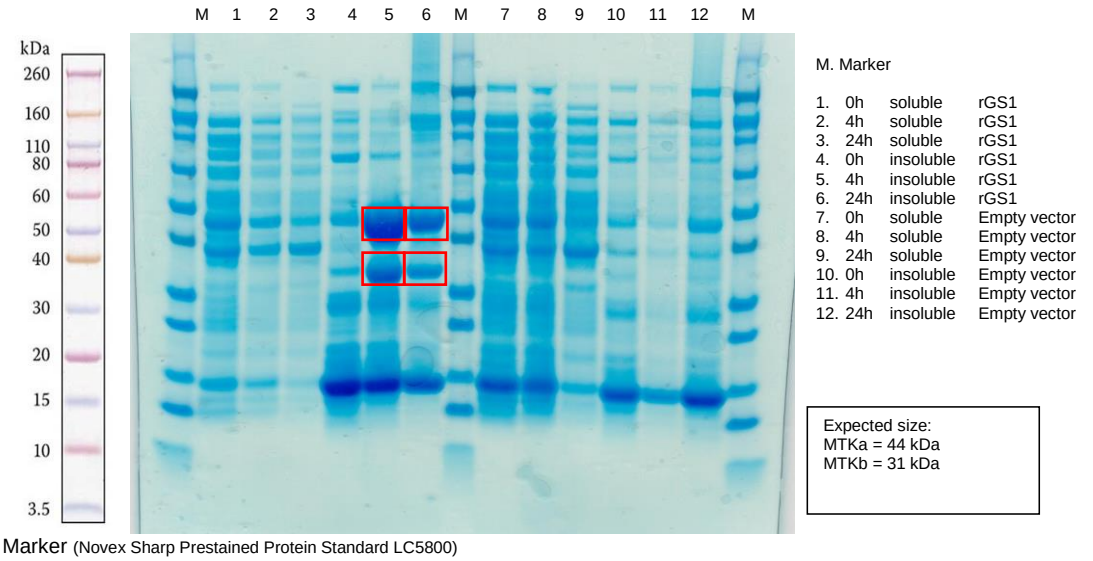


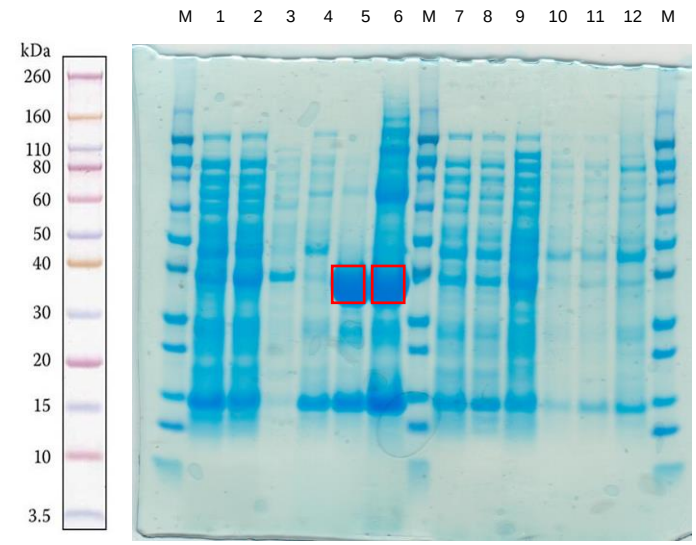
Figure 73: Agarose gel screening of rGS vectors through restriction digest. Expected bands (kb): MTKa/MTKb – 5.9, 1.5; MCL – 5.4, 0.7, 0.1; ICL – 6.0, 0.5.

The *E. coli* expression strain BL21 (DE3) was transformed with the three vectors and a single colony picked from each transformation plate to test for protein expression (Figures 74, 75 and 76).



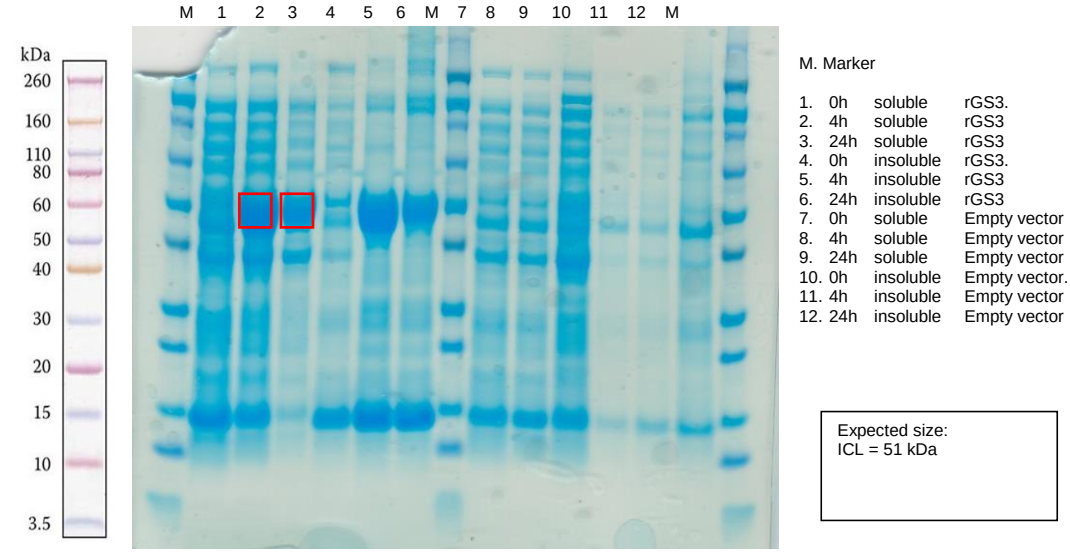
Marker (Novex Sharp Prestained Protein Standard LC5800)

Figure 74: SDS-PAGE analysis of soluble and insoluble fractions from expression of MTKa and MTKb. Target proteins highlighted by red boxes. Protein expression was induced with 1 mM IPTG and cells incubated at 37 °C post induction.



Marker (Novex Sharp Prestained Protein Standard LC5800)

Figure 75: SDS-PAGE analysis of soluble and insoluble fractions from expression of MCL. Target protein highlighted by red box. Protein expression was induced with 1 mM IPTG and cells incubated at 37 °C post induction.



Marker (Novex Sharp Prestained Protein Standard LC5800)

Figure 76: SDS-PAGE analysis of soluble and insoluble fractions from expression of ICL. Target protein highlighted by red box. Protein expression was induced with 1 mM IPTG and cells incubated at 37 °C post induction.

Since MTKa/MTKb and MCL were primarily expressed in an insoluble form expression work was repeated at a lower post induction temperature. The reduced temperature may result in an increase in protein solubility due to slower protein synthesis resulting in enhanced protein folding[191]. The expression protocol and sample preparation were the same as described earlier in this section with the exception that once cultures reached OD_{600} 0.6 they were immediately placed on ice prior to IPTG addition and incubated at 16 °C post induction (Figures 76 and 77).

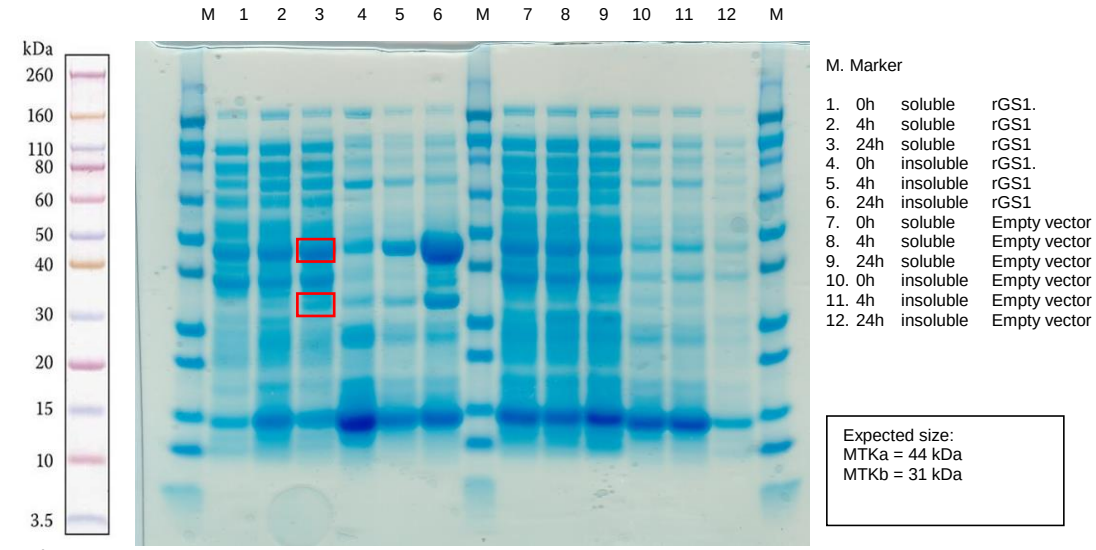


Figure 77: SDS-PAGE analysis of soluble and insoluble fractions from optimised expression of MTK a/b. Target proteins highlighted by red boxes. Protein expression was induced with 1 mM IPTG and cells incubated at 16 °C post induction.

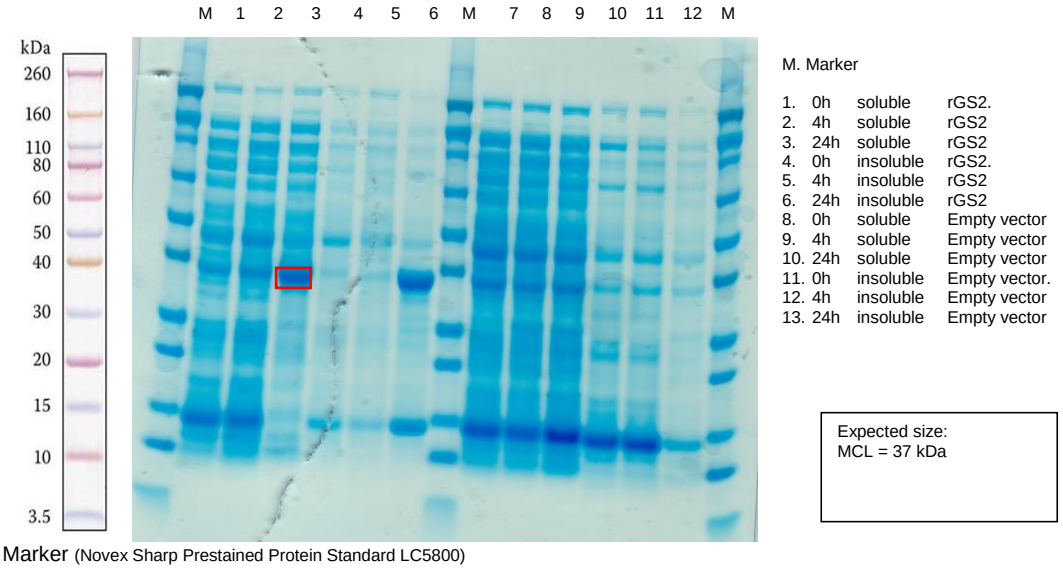
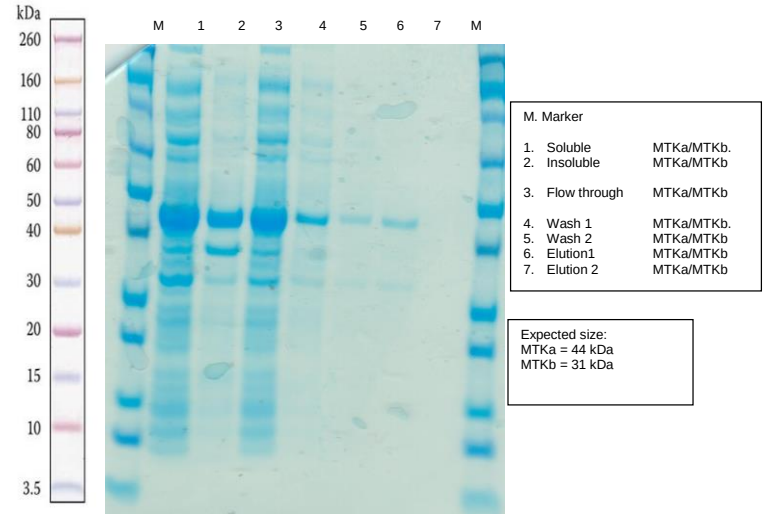


Figure 78: SDS-PAGE analysis of soluble and insoluble fractions from optimised expression of MCL. Target protein highlighted by red box. Protein expression was induced with 1 mM IPTG and cells incubated at 16 °C post induction.

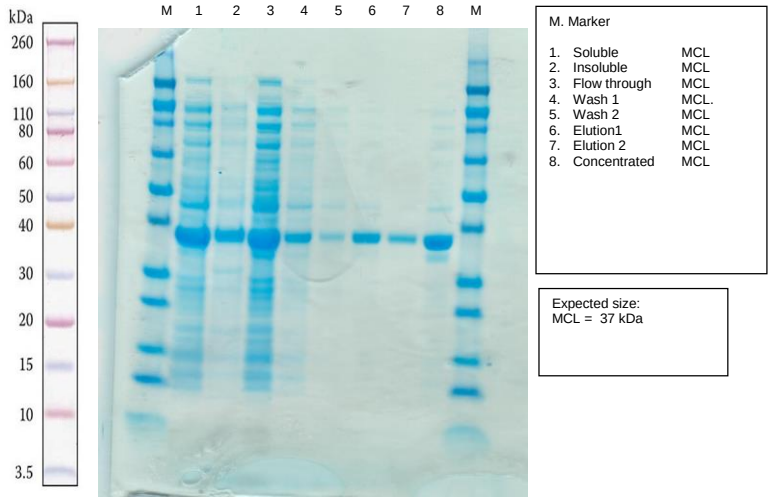
Following expression condition optimisation soluble protein was detected for the three targets by SDS-PAGE. Optimised conditions were used for culture scale up to provide biomass for protein purification. Protein purification was performed using NiNTA spin columns (Section 2.6.3) and SDS-PAGE analysis of the resulting fractions was performed (Figures 79, 80 and 81).



Marker (Novex Sharp Prestained Protein Standard LC5800)

Figure 79: SDS-PAGE analysis of fractions from MTKa and MTKb purification. Purified protein can be observed in the elution fraction 1.

.....



Marker (Novex Sharp Prestained Protein Standard LC5800)

Figure 80: SDS-PAGE analysis of fractions from MCL purification. Purified protein can be observed in elution fractions 1 and 2.

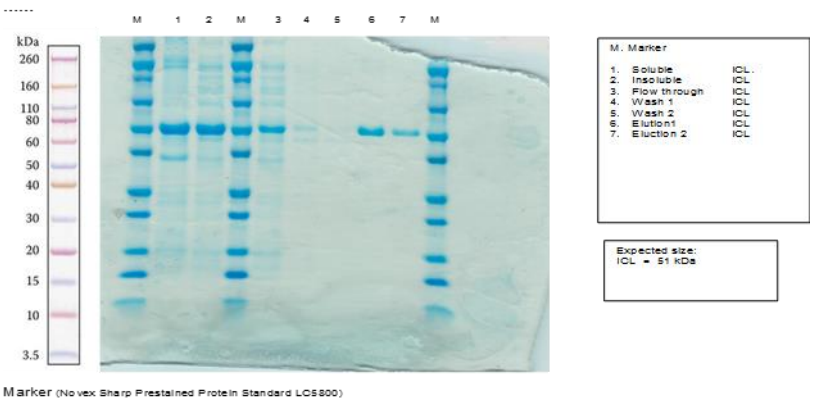


Figure 81: SDS-PAGE analysis of fractions from ICL purification. Purified protein can be observed in elution fractions 1 and 2.

SDS-PAGE analysis highlighted a considerable amount of target protein present in the column flow through fraction (Lane 3). This is likely due to overloading of the columns which are limited to a binding capacity of 300 µg. Purified protein concentration was measured using the BCA assay[192] and protein extracts stored at -80 °C as a 10% (v/v) glycerol stock prior to in solution tryptic digestion.

5.2.2.2 MS^E characterisation of purified rGS proteins

Equimolar amounts of the four purified rGS proteins (MTKa/b, MCL and ICL) were mixed and then reduced, alkylated and digested with trypsin as described in section 2.6.4. The resulting peptides were separated using a chromatographic gradient of 35 minutes (Figure 82) as previously described and analysed via MS^E. A Lockspray of Leucine enkephalin was infused throughout the run as previously described (Section 2.9.4).

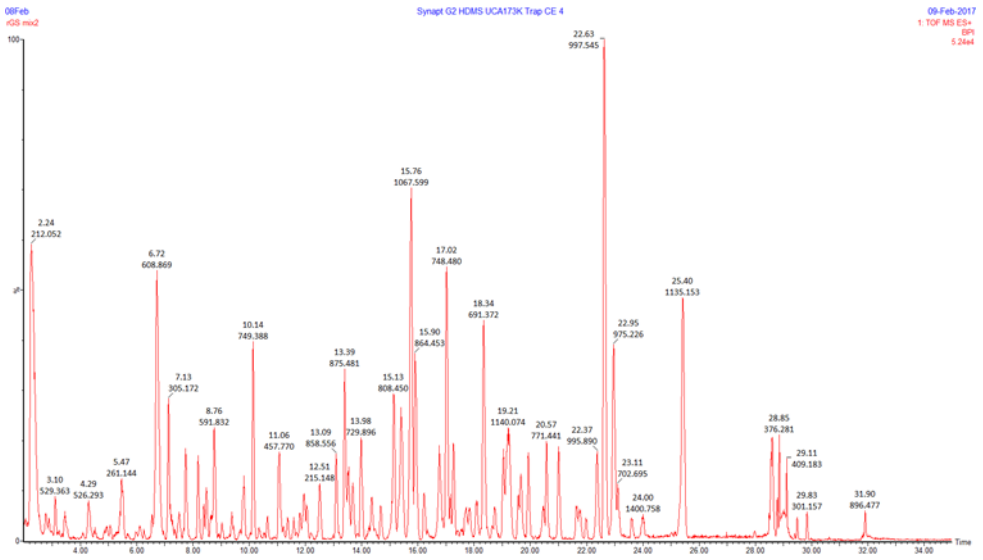


Figure 82: Chromatographic separation of tryptic peptides generated from an equimolar mix of purified target proteins - MTKa/b, MCL and ICL.

The resulting data was analysed using PLGS with data searched against a custom *E. coli* K-12 Uniprot database with His-tagged versions of the four reverse glyoxylate shunt proteins manually added. Interrogation of the database identified each of the proteins with an average sequence coverage of between 78% and 91% achieved from three technical replicates (Figure 83). An example protein sequence coverage for each target is shown in Figure 84.

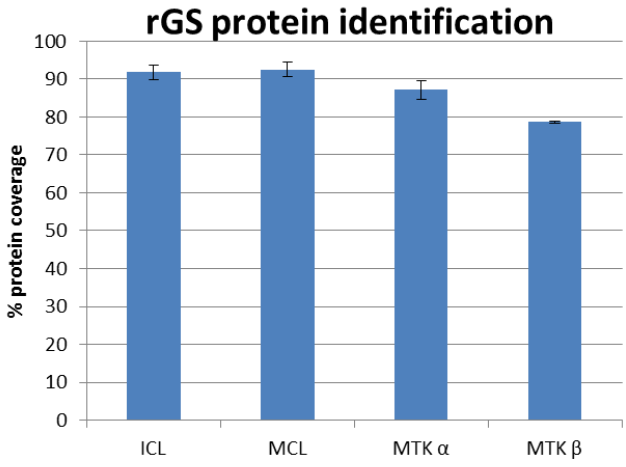


Figure 83: Average sequence coverage of the four target proteins (MTKa/b, MCL and ICL) generated from triplicate data.

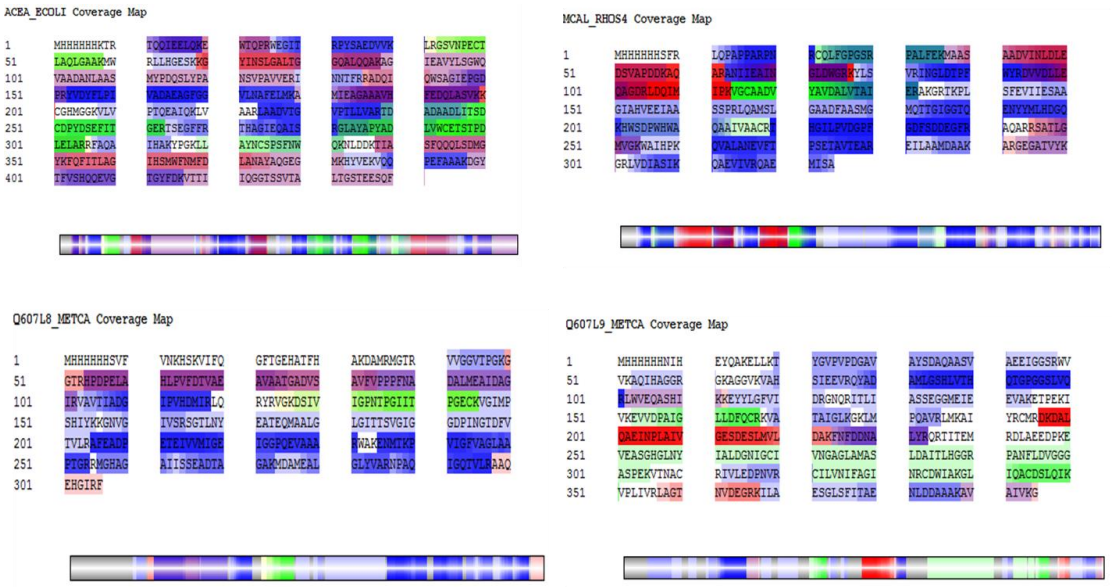


Figure 84: Peptide coverage maps for each of the target proteins. Generated from MSE characterisation of an equimolar mix of the four purified targets. Blue indicates a peptide match, red a partial peptide match, green a modified peptide match and yellow a partial modified peptide match.

5.2.2.3 MS^E characterisation of rGS1 overexpression strain

To expand on this successful identification of the purified targets the ability to identify one of the target proteins against an *E. coli* proteome background was explored. To study this an *E. coli* strain engineered to overexpress ICL through IPTG induced T7 based expression was utilised. On this occasion the cells were lysed, and the entire cellular proteome digested, following the in solution tryptic digestion protocol described in Section 2.6.4. The resulting peptides were separated over a chromatographic gradient of 120 minutes and analysed using Synapt G2 mass spectrometer configured for MS^E acquisition (Figure 85). The resulting data was analysed using PLGS and searched against the custom database generated in Section 5.2.2.2 to confirm that identification of a single target proteins from an *E. coli* cell lysate background was achievable.

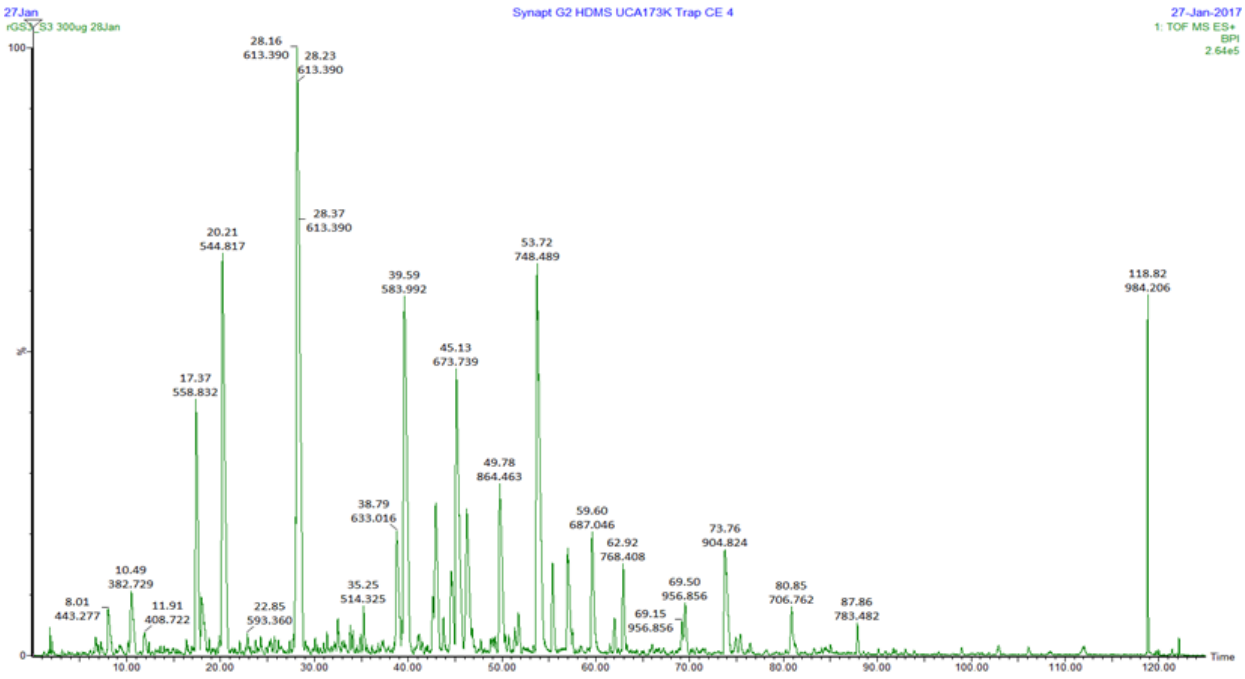


Figure 85: Chromatographic separation of E. coli proteome tryptic peptides over 120 minutes. The strain characterised is engineered for the overexpression of E. coli ICL.

The analysis of this data via PLGS resulted in the identification of 173 *E. coli* cellular proteins. From these identified proteins, ICL had the highest sequence coverage with 93.5% (Figure 86).

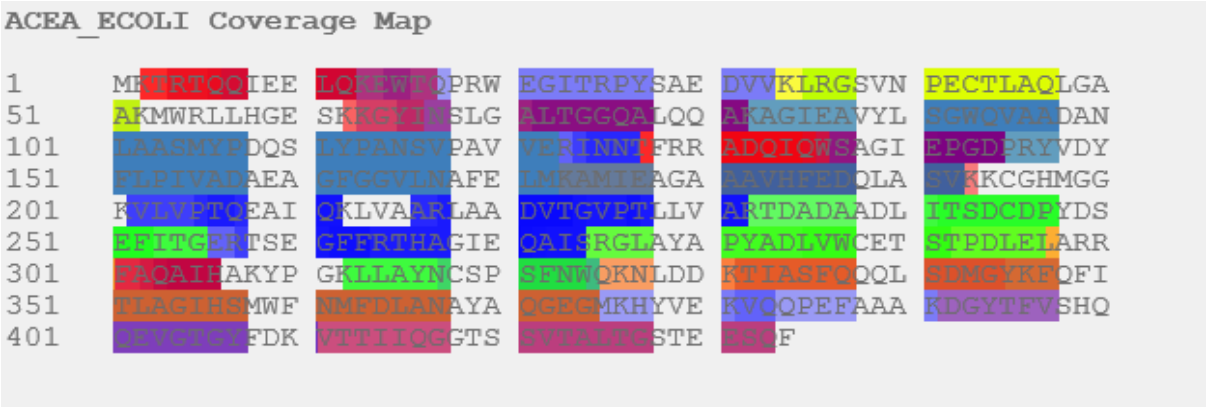


Figure 86: Isocitrate lyase protein coverage. Blue indicates a peptide match, red a partial peptide match, green a modified peptide match and yellow a partial modified peptide match.

In this study proteins were identified which ranged in size from 8 to 155 kDa and isoelectric points from 3.75 to 11.8 (Figure 87). The average number of peptides per protein identification in this dataset was 12.6 and no protein was identified from a single peptide whilst only 11% were identified from four or less peptides. The average protein coverage of the 173 proteins identified was 44.1%.

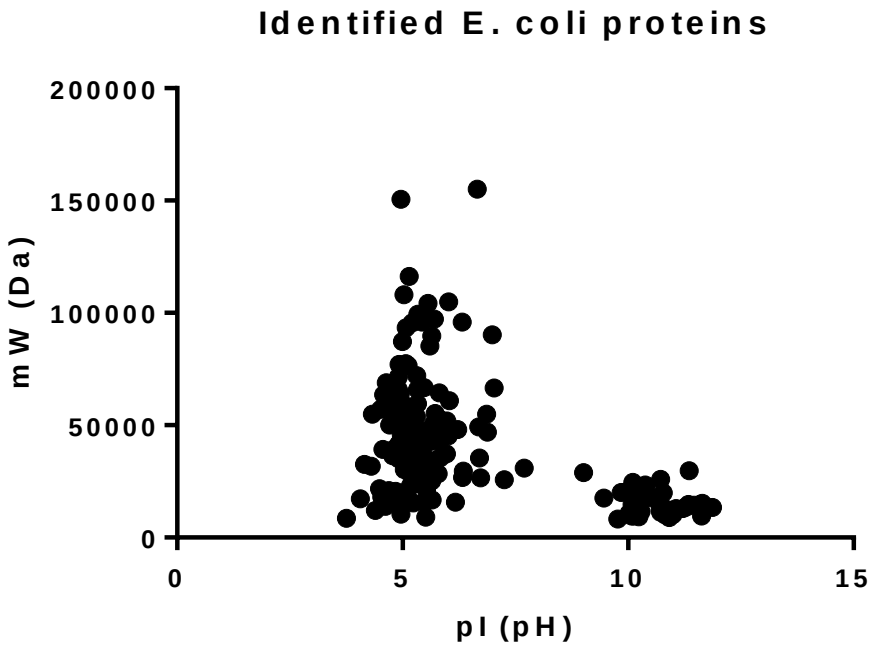


Figure 87: Molecular weight and isoelectric point of proteins identified from E. coli lysate by MS^E analysis.

Analysis of the identified proteins revealed 20 which are involved in *E. coli* central carbon metabolism (Table 29). These proteins are likely to be of particular interest when engineering cellular metabolism as they are involved in the oxidation of main carbon sources whilst all natural metabolites produced by *E. coli* are derived from 12 precursors which are either intermediates of central carbon metabolism or co-factors including NAD(P)H and ATP[193].

Accession	Entry	Description	mW (Da)	Coverage (%)	Pathway
P06994	MDH_ECOLI	Malate dehydrogenase	32317	89.4231	TCA
P08324	ENO_ECOLI	Enolase	45495	71.2297	Glycolysis
P00891	CISY_ECOLI	Citrate synthase	48031	66.5105	TCA
P06958	ODP1_ECOLI	Pyruvate dehydrogenase E1 component	99474	66.3657	Glycolysis
P06977	G3P1_ECOLI	Glyceraldehyde 3-phosphate dehydrogenase A	35379	63.9394	Glycolysis
P08200	IDH_ECOLI	Isocitrate dehydrogenase [NADP]	45727	61.0577	Glyoxylate cycle
P11665	PGK_ECOLI	Phosphoglycerate kinase	40961	58.0311	Glycolysis
P31217	PMG1_ECOLI	Phosphoglycerate mutase 1	28407	53.4137	Glycolysis
P07459	SUCD_ECOLI	Succinyl-CoA synthetase alpha chain	29627	51.7361	TCA
P30148	TALB_ECOLI	Transaldolase B	35066	48.7342	PPP
P11604	ALF_ECOLI	Fructose-bisphosphate aldolase class II	38991	48.6034	Glycolysis
P27302	TKT1_ECOLI	Transketolase 1	72156	45.5505	PPP
P00350	6PGD_ECOLI	6-phosphogluconate dehydrogenase_ decarboxylating	51449	43.1624	PPP
P07460	SUCC_ECOLI	Succinyl-CoA synthetase beta chain	41366	43.0412	TCA
P10444	DHSA_ECOLI	Succinate dehydrogenase flavoprotein subunit	64381	38.7755	TCA
P07014	DHSB_ECOLI	Succinate dehydrogenase iron-sulfur protein	26752	34.0336	TCA
P00864	CAPP_ECOLI	Phosphoenolpyruvate carboxylase	99000	27.9728	TCA
P22259	PPCK_ECOLI	Phosphoenolpyruvate carboxykinase [ATP]	59606	27.2222	Gluconeogenesis
P78258	TALA_ECOLI	Transaldolase A	35636	18.3544	PPP
P14178	KPY1_ECOLI	Pyruvate kinase I	50697	17.0213	Glycolysis

Table 29: Enzymes involved in E. coli central carbon metabolism identified via MS^E analysis of an E. coli cellular extract.

5.2.3 Conclusions

The aim of this section was to establish that the MS^E workflow was suitable for the identification of heterologously expressed proteins. To achieve this a purification strategy for the 4 target proteins was successfully designed through the introduction of a N-terminal His tag and Ni-NTA affinity chromatography. The purified targets were mixed in equimolar amounts and in-solution tryptic digestion performed. The peptides were analysed via the previously exemplified MS^E workflow and the data used to interrogate a custom *E. coli* protein database that was supplemented with the heterologous proteins. This resulted in the successful identification of each of the target protein with up to 91 % protein coverage.

An *E. coli* cellular lysate from a strain engineered to over express once of the target proteins (ICL) was then lysed. Protein expression was induced, the cells lysed and the cellular proteins tryptically digested in-solution. The analysis of this lysate resulted in identifying the target protein with a protein coverage of 93.5%. In addition to this a further 172 *E. coli* cellular proteins were identified spanning a range of molecular weights, isoelectric points and cellular functions in a single experiment.

5.3 Proteomics guided strain engineering: Application to the reverse glyoxylate shunt

5.3.1 Introduction

Inherent to the successful re-engineering of metabolic pathways to make high value chemicals is the development and utilisation of technologies that aid both their construction and the optimisation. Earlier chapters of this PhD have described methodologies to rapidly assemble biosynthetic pathway. This study will aim to complement this approach through applying MS^E proteomics directly to the optimisation of a metabolic engineering approach. In the previous section of this chapter a workflow was developed for the identification of target proteins from cellular extracts. This will now be expanded to attempt label free quantification of pathway proteins using strains engineered to express the previously described reverse glyoxylate shunt. A library of genetic configurations designed to modulate the amount of each pathway protein will be screened for flux through the pathway using both solid and liquid phase analysis. Strains will then be characterised via MS^E with the goal of identifying pathway bottlenecks and the identification of an optimal balance of protein levels.

5.3.2 Results

5.3.2.1 Construction of library to modulate gene expression

The fine tuning of heterologous protein production levels is a key aspect to be considered when looking to optimise biosynthetic pathway implementation. This approach can be used to avoid the accumulation of pathway intermediates and limit the burden on the host cell[67,194]. In the case of the reverse glyoxylate shunt the expression of four heterologous enzymes requires to be balanced (MTK α/β, MCL and ICL).

Multiple strategies can be implemented in order to modulate protein production including, strength of transcriptional promoter, RBS sequence, transcriptional terminator sequence, plasmid copy number and codon optimisation. In this study the three genes were designed to be expressed from a single plasmid in a synthetic operon. Four constitutive promoters, standardly used for heterologous protein production, of varying strength were implemented (lambda phage pL, *E. coli* AspC, *E. coli* RecA and *E. coli* OSMY). To further modulate the amount of protein being produced the order of the three genes within the synthetic operon was shuffled (Figure 88). Previous studies on the order of genes within a synthetic operon have suggested that as the distance between the start of a gene and the end of the operon increases ("transcriptional distance") so does the level of expression. This is due to genes at the start of the operon having more time to be translated during transcription of the remainder of the operon [195].

Promoter	Strength (1 – Strong 4- Weak)
pL	1
pAspC	2
pRecA	3
pOSMY	4

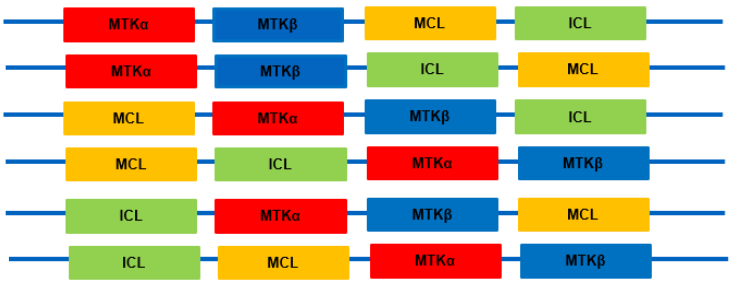


Figure 88: Promoter strength and order of genes within the synthetic operon explored. Each potential combination was constructed resulting in 24 combinations.

A high throughput inABLE based combinatorial DNA assembly methodology was implemented to construct the 24 possible vectors. In this study plasmid copy number and ribosome binding site (RBS) upstream of each gene were kept constant.

5.3.2.2 Solid and liquid phase screening of library

To assess the effect of gene order and promoter strength on the reverse glyoxylate shunt activity *in vivo* an *E. coli* screening strain was obtained. The construction of this strain was achieved through the deletion of citrate synthase (*gltA*) and 2-methyl citrate synthase (*prpC*) which has previously been reported to have minor citrate synthase activity[196] generating a glutamate auxotroph as previously described[190]. To confirm the generated strain displays the expected phenotype, glutamate auxotrophy was assessed through plating the Glu$^-$ strain (Δ*gltA* Δ*prpC)* onto LB rich media and minimal media +/- glutamate. The auxotroph was unable to grow on minimal media lacking the glutamate addition confirming the expected auxotrophy (Figure 89).

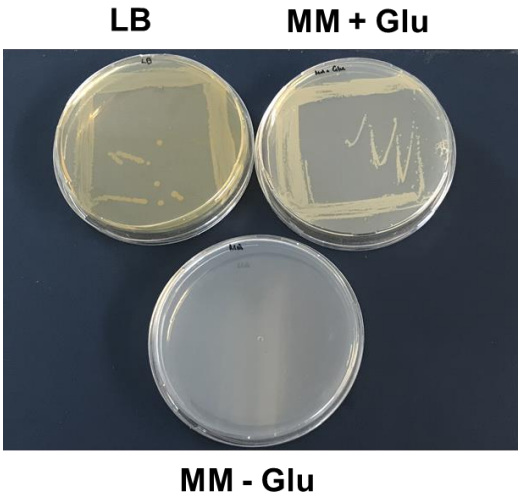


Figure 89: Confirmation of glutamate auxotrophy in Glu$^-$ (ΔgltA ΔprpC) strain through growth LB rich media and minimal media +/- glutamate agar plates.

An initial vector for expression of the three genes comprising the rGS under the control of the pRecA promoter was introduced into the resulting strain (ΔgltA ΔprpC) and the complementation of glutamate auxotrophy through channelling carbon across the rGS was confirmed by the ability of the resulting strain to grow on glucose minimal media supplemented with succinate (Figure 90).

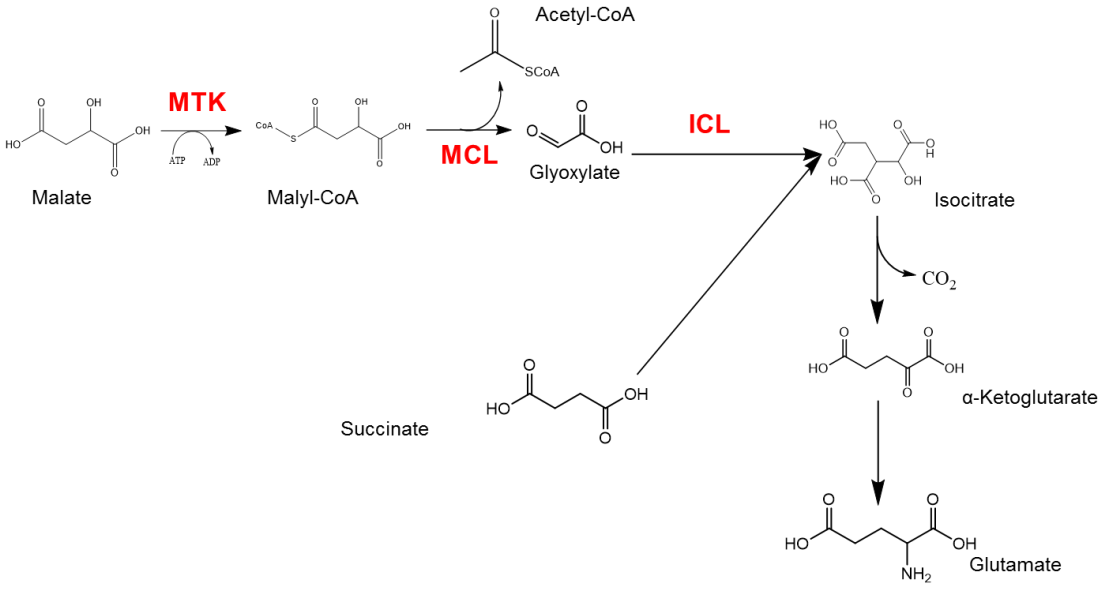


Figure 90: Metabolic pathway for complementing glutamate auxotrophy through the reverse glyoxylate shunt.

The introduction of an empty vector failed to rescue the auxotrophy and growth was not detected. As the cell is unable to produce glutamate through any other route than the reverse glyoxylate shunt the growth rate on this media is a direct consequence of the engineered strains ability to channel carbon through the heterologous pathway (Figure 91).

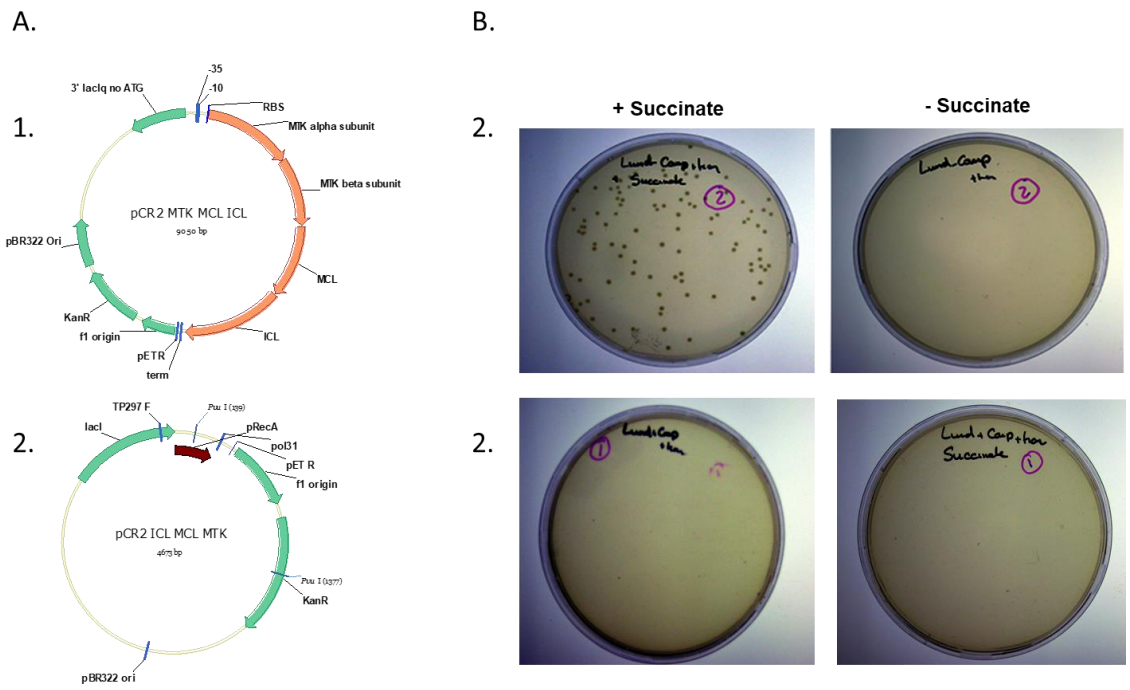


Figure 91: Complementation of succinate auxotrophy via the rGS. A1: Initial rGS expression construct. B1: Plating of Glu- strain harbouring rGS expression vector +/- succinate. A2/B2 Repeated with negative control.

This was exploited as an initial screening platform to characterise the vector library in a solid phase screen. Following selection on glucose minimal media plates supplemented with succinate, forty of the largest colonies were isolated and the operon gene order and transcriptional promoter defined through sequencing (Table 30).

Strain ID	Promoter	Gene Order
3.1	pRecA	MTK-MCL-ICL
3.2	pA	
3.3	pOSMY	
3.4	pRecA	
4.1	pRecA	MTK-ICL-MCL
4.4	pOSMY	
4.5	pA	
5.1	pL	MCL-MTK-ICL
5.3	pRecA	
5.5	pOSMY	
5.7	pA	
6.1	pOSMY	MCL-ICL-MTK
7.1	pA	ICL-MTK-MCL
7.2	pOSMY	

Table 30: Genetic characterisation of fastest growing libraries members identified from solid phase screen

This initial approach identified 14 strains with unique genetic configurations which were able to grow on minimal media supplemented with succinate and therefore channel carbon through the reverse glyoxylate shunt. A negative control strain containing an empty vector was again unable to proliferate on this media. Whilst the results of this initial study suggested that multiple genetic configurations result in an active rGS the optimal gene order or promoter for rGS activity was difficult to define. In fact, each of the four promoters implemented was observed and only the operon order ICL::MCL::MTK was not detected.

To further characterise the 14 strains identified in the first round of screening the growth profile of each strain in minimal media supplemented with glucose and succinate was observed (Figure 92). This approach allowed ranking of isolates however once again no pattern of optimal genetic configuration and therefore optimal balance of pathway proteins was easily identifiable. In fact, the three strains which displayed the quickest growth rates (4.1, 4.4 and 5.1) differed gene order and promoter (Table 30). Strain 5.1 carries a strong promoter and the MCL gene in operon position one suggesting that high expression of this protein may be required for optimised flux. However, in strains 4.1 and 4.4 the MCL gene is in position three with weaker promoters controlling transcription of the operon.

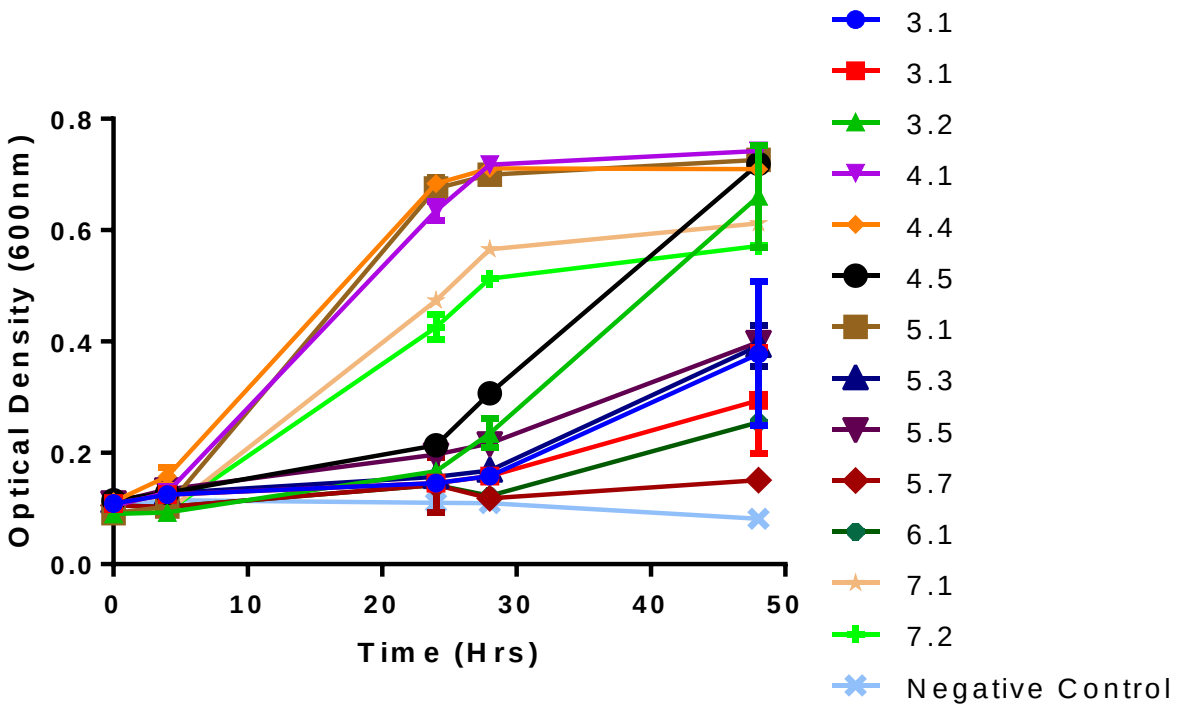


Figure 92: Growth curves of 14 library members identified in solid phase screen. Growth curves the result of triplicate experiments generated in minimal media supplemented with glucose and succinate. Strain ID found in Table 30.

Three strains (5.1, 5.7 and 6.1) were selected for further analysis of heterologous protein abundance. One slow grower, one intermediate and one fast grower were taken forward for initial MS^E based quantification of rGS protein levels to determine if an optimal protein abundance at each stage in the pathway could be identified and secondly how the genetic configuration affected protein output.

5.3.2.3 Heterologous pathway protein quantification

The three strains selected (rGS 5.1, rGS 5.7 and rGS 6.1) displayed differing growth rates in the liquid phase assay (rGS5.1 > rGS 6.1 > rGS5.7) (Figure 92) and contain differing genetic configurations (Table 30), however no clear correlation between growth rate, promoter strength and operon order could be deciphered. Analysis of the

soluble fractions generated from the three strains prior to tryptic digestion through SDS-PAGE was complicated by the target proteins having similar sizes, for example Malyl–CoA lyase (37 kDa), Malate thiokinase β-subunit (31 kDa) and Malate α-subunit (44 kDa). A tentative hypothesis would be that based on the SDS-PAGE, strain 5.1 – the fastest grower – is producing higher amounts of MCL than the other strains characterised by the band present above 32 kDa in lanes 1 to 3 in the SDS-PAGE which is not present in the other lanes (Figure 93). However, without target specific antibodies this is difficult to confirm.

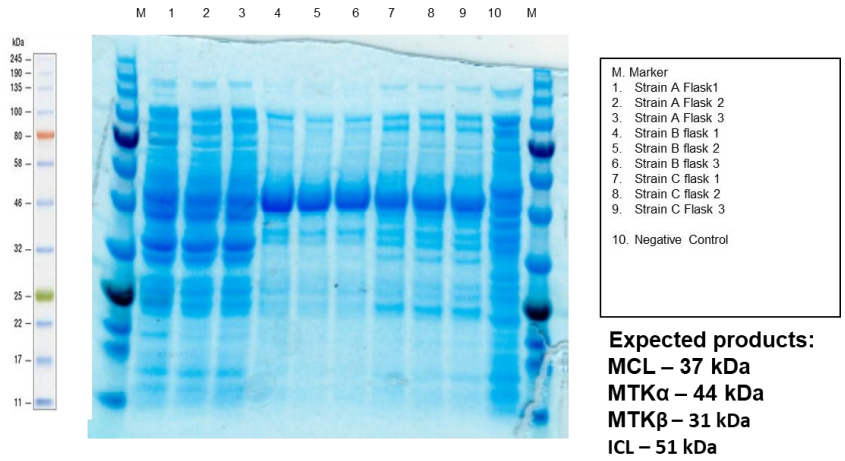


Figure 93: SDS-PAGE analysis of the three rGS expressing strains proteomes prior to tryptic digestion. Strain A, B and C relate to rGS 5.1, rGS 5.7 and rGS 6.1 respectively.

Tryptically digested cellular proteomes from biological triplicates of the three strains were prepared for MS^E analysis as previously described. The data was collected as previously however in order to allow for relative protein quantification the resulting data was processed using the Progenesis QI for Proteomics platform (Nonlinear Dynamics) for label free relative protein quantification.

Relative protein quantification was performed using the Hi-3 methodology described by Silva et al[135]. Quantification was carried out following peptide and protein identification and was achieved through averaging the abundance of the three most

abundant peptides per protein giving a value for the protein signal. This averaged value allowed for relative quantification of the same protein across runs. It is possible - with the addition of a calibrant - to convert this number into an absolute value.

In this study relative quantification between strains was performed. Following protein quantification proteins with a P value > 0.05 and a max fold change < 2 were omitted. This resulted in 32 quantified proteins including the four reverse glyoxylate shunt proteins for analysis. To act as a technical replicate a pool prepared from all the samples analysed was run in triplicate throughout the experiment to monitor for column condition, sample degradation or other factors that may impact the analysis. The max fold change in the four rGS proteins from this pool analysis was 1.8 fold. As an additional control the chloramphenicol acetyltransferase which is expressed from the plasmid acting as a selection marker was also quantified giving a max fold change of 1.7 between samples.

Results from each run were grouped per strain, the abundance normalised across runs and relative average abundances calculated for each of the target proteins (Figure 94). These results suggest that the least active strain (rGS 5.7), produces high levels of isocitrate lyase but low levels of the other rGS proteins whilst the most active strain (5.1) appears not to overproduce any one of the pathway proteins and instead displays a more balanced pathway protein output.

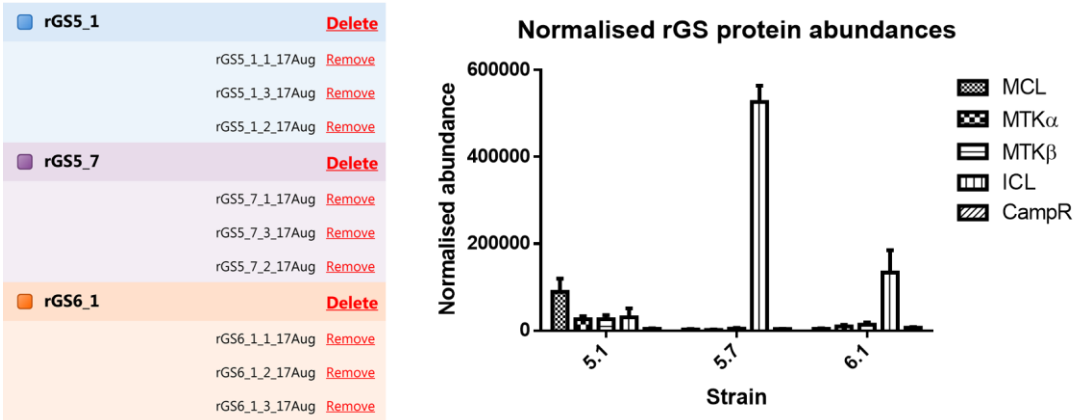


Figure 94: Average normalised relative abundances generated from the three reverse glyoxylate shunt expressing strains explored in this study.

To further characterise the differences in protein abundance between the strains analysed, the fold difference in each protein between the strains was calculated. It was found that the highest fold change observed for any of the proteins quantified was a 33.76 fold increase in malyl-CoA lyase abundance between the most active strain (5.1) and the least active strain (5.7) (Table 31). This result suggests that the successful implementation of the reverse glyoxylate shunt relies on the abundance of this enzyme.

Target Protein	Max fold change in abundance	Strain with highest abundance	Strain with lowest abundance
ICL	17.04	rGS5.7	rGS5.1
MTKα	15	rGS5.1	rGS5.7
MTKβ	5.74	rGS5.1	rGS5.7
MCL	33.76	rGS5.1	rGS5.7

Table 31: Maximum relative fold changes in the target rGS proteins observed between engineered strains. The largest fold change observed was a 33.76 fold increase in MCL abundance in strain rGS5.1 (the best performing strain) in comparison to rGS5.7 (the worst performing strain)

5.3.3.4 Identification of pathway bottleneck through individual gene downregulation

To further explore this hypothesis, a series of vectors were constructed to simulate the under expression of each pathway enzyme and the effect this has on carbon flux through the reverse glyoxylate shunt. The pathway was split between two plasmids one of which included a high copy origin of replication (pBR322 ~15-20 copies[197]) and strong promoter whilst the other included a lower copy ori (p15a ~ 10 copies[198]) and a weak promoter. Each of the genes were introduced independently to the high copy promoter vector to determine the effect of individual gene up regulation on carbon flux. A control construct in which the full operon was expressed from the high copy plasmid was constructed in parallel. The resultant vectors were introduced into the Glu^- strain and, as previously, screened on minimal media supplemented with glucose and succinate (Figure 95).

Strain			Relative expression		
			Mtk	*Mcl*	*Icl*
1	Negative control				
2	Positive control		High	High	High
3	HLL	*Mtk* high	High	Low	Low
4	LHL	*Mcl* high	Low	High	Low
5	LLH	*Icl* high	Low	Low	High

High = expression from **high** copy number (pBR322) **strong** promoter plasmid
Low = expression from **low** copy number (p15a) **weak** promoter plasmid

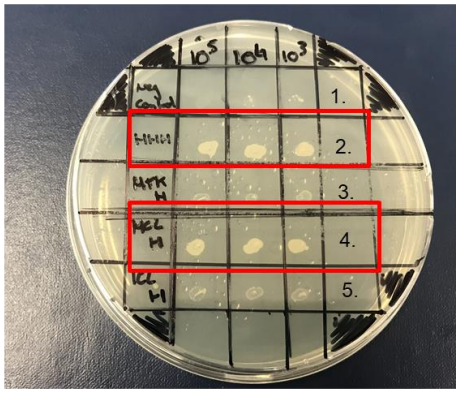


Figure 95: The effect of individual gene downregulation on the ability of strains to channel carbon through the rGS. Growth on a minimal media agar plate supplemented with glucose and succinate was only achieved when MCL was expressed using a strong promoter from a high copy number plasmid.

These results show that growth on this media and therefore flux through the reverse glyoxylate shunt was only achieved for the positive control strain and when the MCL gene was expressed from the high copy/strong promoter construct. This result supports the finding from the MS^E based analysis, highlighting that high levels of expression of MCL is key for successful carbon channelling through the reverse glyoxylate shunt.

5.3.3 Conclusions

A key consideration when embarking on a metabolic pathway engineering project is the fine tuning of expression of heterologous pathway proteins to optimise strain performance. Through modulation of protein levels, unnecessary metabolic burden can be avoided whilst also ensuring pathway intermediates do not build up. In this study the reverse glyoxylate shunt was used as a test case to exemplify the utilisation of proteomics coupled to high throughput screening and synthetic biology approaches to accelerate metabolic engineering programs. This study highlights the

greater depth of knowledge that can be garnered from techniques such as LC-MS based proteomics and how this knowledge, which would not be generated from traditional product formation assays, can be used to guide subsequent rounds of metabolic engineering.

In this section a library of vectors designed to modulate the levels of expression of the enzymes involved in the introduction of reverse glyoxylate shunt into *E. coli* were characterised. This characterisation combined with the use of solid and liquid phase analysis with label free protein quantification to identify pathway protein abundances and correlate these abundances to pathway performance. It was found through label free protein quantitation that the best performing strain had a 33-fold increase in malyl-CoA lyase abundance in comparison to the worst performing isolate suggesting the abundance of this enzyme is key for pathway performance. This finding was confirmed through the simulated under-expression of each pathway protein, which confirmed that MCL required to be highly expressed to facilitate flux through the engineered pathway.

6. Conclusions

The replacement of conventional processes to manufacture valuable industrial products and the selection of optimal biosynthetic routes requires the construction, and in most cases subsequent context-dependent evaluation, and optimization of multicomponent biosynthetic pathways to generate commercial end products. This PhD has focused on the development of an industrial platform for microbial strain improvement which addresses the persistent limitations associated with today's iterative and empirical approaches. To achieve this, I have developed a DNA assembly platform which is optimised for both accuracy and efficiency and coupled this approach to mass spectrometry based proteomics techniques and direct screening methods to identify strains with improved process efficiency.

The DNA assembly technique is state of the art, enabling the optimised combinatorial assembly of DNA fragments for large scale gene/pathway assembly and optimisation. To achieve this a systematic overview of the previous inABLE technology was undertaken and key factors limiting DNA assembly identified and addressed. Despite this, the process still required the purification of DNA fragments through gel electrophoresis, which limits the potential to automate the process. To address this key technical hurdle a novel purification platform based on phosphorothioate bond mediated DNA protection and exonuclease degradation of contaminating DNA was developed. It was found that the DNA of interest could be purified through exonuclease addition, removing the requirement to run agarose gels, greatly enhancing process efficiency and ensuring the technique is amenable to process automation. The introduction of this step did not reduce assembly efficiency in comparison to gel electrophoresis and in fact was found to increase assembly accuracy in previously difficult to assemble constructs.

This innovation and optimisation has resulted in a strategy which can be considered alongside the Gibson[78] and Golden Gate[80] approaches - two of the most widely used DNA assembly approaches[199] – as the method of choice for combinatorial assembly of DNA. As the cost of DNA synthesis continues to decrease[86], the cloning of a single

gene will become a process which is no longer carried out by the researcher and instead is outsourced to DNA synthesis providers. This highlights the fact that it is increasingly important that DNA assembly techniques are optimised for the large scale production of combinatorial libraries as the complete chemical synthesis of these libraries will likely remain cost prohibitive for the foreseeable future. The ability to mix and match target genes with regulatory regions to identify levels of expression which results in optimal flux through the required pathway without placing unnecessary burden on the host organism has proven a successful strain engineering strategy[33, 200]. The platform developed in this thesis provides the ability to build combinatorial DNA libraries rapidly and accurately in an automated manner and has the potential to greatly accelerate strain engineering programs. A patent has been applied for to protect this invention and work has now been initiated on the automation of the process through transferring the optimised protocol onto a liquid handling robot.

Obvious synergy exists between this approach and label free quantitative proteomics coupled to versatile, solid phase screening and selection. The generation of information on pathway protein abundances in combination with target production and cellular growth rate has been used previously to identify clear and focused targets for the next stages of pathway optimsiation[73,117]. In this thesis, methods were developed to identify engineered cells of interest, pathway bottlenecks and guide subsequent rounds of pathway optimisation. The label free MS^E platform, initially described by Silva *et al*[134,135], was identified as a suitable approach for the flexible analysis of protein abundances in engineered microbes. Robust liquid chromatography, MS and data analysis methods were developed to identify and quantify multiple key cellular proteins from engineered strain lysates. The approach was applied to the implementation of a reverse glyoxylate shunt in *E. coli* and this strategy coupled to liquid and solid phase screening approaches allowed for the ranking of strain performance and the identification of pathway bottlenecks limiting flux through the pathway.

MS^E analysis highlighted a 33-fold increase in abundance of one pathway protein – MCL from *R. sphaeroides* - when comparing a well performing strain against a poor performing strain. The findings from the MS^E analysis were then confirmed through

simulated under expression of each pathway enzyme *in vivo* where it was shown that only when the MCL was overexpressed was flux through the pathway observed.

This study highlights the greater depth of knowledge that can be garnered from techniques such as LC-MS based proteomics and how this knowledge, which would not be generated from traditional product formation assays, can be used to guide subsequent rounds of metabolic engineering. Future work will focus on expanding the scope of this approach to study not only heterologous proteins but also the effect the introduction of the biosynthetic pathway has had on the host proteome. This data would then be used to inform the next stages of strain optimisation.

This thesis aimed to address the primary issues cited by end users as to why bioprocesses are not more readily taken up – their development is too slow, too unpredictable and too expensive. The integration of the methodologies described in this study for pathway engineering and analysis have the potential to address these concerns through the creation of a step change in the speed and predictability with which microbes can be engineered for industrial applications. Ingenza is currently applying this platform to synthetic biology programs for the bio-production of plastic, nylon and rubber, amongst other products.

7. Bibliography

1. Keasling, J. Manufacturing Molecules through Metabolic Engineering. *Science (80-.).* **330**, 1355–1358 (2010).

2. Keasling, J. D. Synthetic biology for synthetic chemistry. *ACS Chem. Biol.* **3**, 64–76 (2008).

3. Wee, Y., Kim, J. & Ryu, H. Biotechnological Production of Lactic Acid and Its Recent Applications. *Food Technol. Biotechnol.* **44**, 163–172 (2006).

4. Demain, A. L. History of Industrial Biotechnology. in *Industrial Biotechnology: Sustainable Growth and Economic Success* 17–77 (Wiley-VCH Verlag GmbH & Co. KGaA, 2010). doi:10.1002/9783527630233.ch1

5. Dürre, P. New insights and novel developments in clostridial acetone/butanol/isopropanol fermentation. *Appl. Microbiol. Biotechnol.* **49**, 639–648 (1998).

6. Davies, E. T. Green Biologics Ltd.: Commercialising bio-n-butanol. *Green Process. Synth.* **2**, 2013 (2013).

7. Green, E. M. Fermentative production of butanol-the industrial perspective. *Curr. Opin. Biotechnol.* **22**, 337–343 (2011).

8. Nielsen, J. & Keasling, J. D. Engineering Cellular Metabolism. *Cell* **164**, 1185–1197 (2016).

9. van den Berg, M. A. Functional characterisation of penicillin production strains. *Fungal Biol. Rev.* **24**, 73–78 (2010).

10. Walsh, G. Biopharmaceutical benchmarks 2014. *Nat. Biotechnol.* **32**, 992–1000 (2014).

11. Stephanopoulos, G. & Vallino, J. J. Metabolite Overproduction. *Science (80-.).* **252**, 1675 (1991).

12. Bailey, J. E. Toward a Science of Metabolic Engineering Published by : American Association for the Advancement of Science Stable. *Science (80-.).* **252**, 1668–1675 (1991).

13. Woolston, B. M., Edgar, S. & Stephanopoulos, G. Metabolic Engineering: Past and Future. *Annu. Rev. Chem. Biomol. Eng.* **4**, 259–288 (2013).

14. Goeddel, D. V. *et al.* Expression in *Escherichia coli* of chemically synthesized genes for human insulin. *Proc. Natl. Acad. Sci.* **76**, 106–110 (1979).

15. Goeddel, D. *et al.* Direct expression in *Escherichia coli* of a DNA sequence coding for human growth hormone. *Nature.* **281**, 544–548 (1979).

16. Atsumi, S. & Liao, J. C. Metabolic engineering for advanced biofuels production from *Escherichia coli*. *Curr. Opin. Biotechnol.* **19**, 414–419 (2008).

17. Webb, J. P. *et al.* Efficient bio-production of citramalate using an engineered *Escherichia coli* strain. *Microbiology* 1–9 (2017). doi:10.1099/mic.0.000581

18. DeJong, J. M. *et al.* Genetic engineering of taxol biosynthetic genes in *Saccharomyces cerevisiae*. *Biotechnol. Bioeng.* **93**, 212–224 (2006).

19. Galanie, S., Thodey, K., Trenchard, I. J., Interrante, M. F. & Smolke, C. D. Complete biosynthesis of opiods in yeast. *Science (80-.).* **349**, 1095–1100 (2015).

20. Connor, M. R. & Liao, J. C. Microbial production of advanced transportation fuels in non-natural hosts. *Curr. Opin. Biotechnol.* **20**, 307–315 (2009).

21. Atsumi, S., Hanai, T. & Liao, J. C. Non-fermentative pathways for synthesis of branched-chain higher alcohols as biofuels. *Nature* **451**, 86–89 (2008).

22. Atsumi, S. *et al.* Engineering the isobutanol biosynthetic pathway in *Escherichia coli* by comparison of three aldehyde reductase/alcohol dehydrogenase genes. *Appl. Microbiol. Biotechnol.* **85**, 651–657 (2010).

23. Peralta-Yahya, P. P., Zhang, F., Del Cardayre, S. B. & Keasling, J. D. Microbial engineering for the production of advanced biofuels. *Nature* **488**, 320–328 (2012).

24. Nakamura, C. E. & Whited, G. M. Metabolic engineering for the microbial production of 1,3-propanediol. *Curr. Opin. Biotechnol.* **14**, 454–459 (2003).

25. Ro, D. K. *et al.* Production of the antimalarial drug precursor artemisinic acid in engineered yeast. *Nature* **440**, 940–943 (2006).

26. World Health Organization. *World Malaria Report 2017*.

27. Vennerstrom, J. L. *et al.* Identification of an antimalarial synthetic trioxolane drug development candidate. *Nature* **430**, 900–904 (2004).

28. Enserink, M. Source of new hope against malaria is in short supply. *Science (80-.).* **307**, 33 (2005).

29. Schmid, G. & Hofheinz, W. Total Synthesis of Qinghaosu. *J. Am. Chem. Soc.* **105**, 624–625 (1983).

30. Reiling, K. K. *et al.* Mono and diterpene production in *Escherichia coli*. *Biotechnol. Bioeng.* **87**, 200–212 (2004).

31. Martin, V. J. J., Piteral, D. J., Withers, S. T., Newman, J. D. & Keasling, J. D. Engineering a mevalonate pathway in *Escherichia coli* for production of terpenoids. *Nat. Biotechnol.* **21**, 796–802 (2003).

32. Paddon, C. J. & Keasling, J. D. Semi-synthetic artemisinin: A model for the use of synthetic biology in pharmaceutical development. *Nat. Rev. Microbiol.* **12**, 355–367 (2014).

33. Pfleger, B. F., Pitera, D. J., Smolke, C. D. & Keasling, J. D. Combinatorial engineering of intergenic regions in operons tunes expression of multiple genes. *Nat. Biotechnol.* **24**, 1027–1032 (2006).

34. Tsuruta, H. *et al.* High-level production of amorpha-4, 11-diene, a precursor of the antimalarial agent artemisinin, in *Escherichia coli*. *PLoS One* **4**, (2009).

35. Paddon, C. J. *et al.* High-level semi-synthetic production of the potent antimalarial artemisinin. *Nature* **496**, 528–532 (2013).

36. Endy, D. Foundations for engineering biology. *Nature* **438**, 449–453 (2005).

37. Nielsen, J. *et al.* Engineering synergy in biotechnology. *Nat. Chem. Biol.* **10**, 319–322 (2014).

38. Chubukov, V., Mukhopadhyay, A., Petzold, C. J., Keasling, J. D. & Martín, H. G.

Synthetic and systems biology for microbial production of commodity chemicals. *npj Syst. Biol. Appl.* **2**, 16009 (2016).

39. Keasling, J. D. Synthetic biology for synthetic fuels. *ACS Chem. Biol.* **3**, 64–76 (2011).

40. Keasling, J. D. Synthetic biology and the development of tools for metabolic engineering. *Metab. Eng.* **14**, 189–195 (2012).

41. Blazeck, J. & Alper, H. S. Promoter engineering: Recent advances in controlling transcription at the most fundamental level. *Biotechnol. J.* **8**, 46–58 (2013).

42. Ellis, T., Adie, T. & Baldwin, G. S. DNA assembly for synthetic biology: from parts to pathways and beyond. *Integr. Biol.* **3**, 109 (2011).

43. Lu, H., Giordano, F. & Ning, Z. Oxford Nanopore MinION Sequencing and Genome Assembly. *Genomics, Proteomics Bioinforma.* **14**, 265–279 (2016).

44. Hughes, R. A. & Ellington, A. D. Synthetic DNA Synthesis and Assembly : Putting the Synthetic in Synthetic Biology. (2017). doi:10.1101/CSHPERSPECT.A023812

45. Lausted, C. *et al.* POSaM: a fast, flexible, open-source, inkjet oligonucleotide synthesizer and microarrayer. *Genome Biol.* **5**, R58 (2004).

46. Kizer, L., Pitera, D. J., Pfleger, B. F. & Keasling, J. D. Application of functional genomics to pathway optimization for increased isoprenoid production. *Appl. Environ. Microbiol.* **74**, 3229–3241 (2008).

47. Oh, M. K. & Liao, J. C. DNA microarray detection of metabolic responses to protein overproduction in Escherichia coli. *Metab. Eng.* **2**, 201–209 (2000).

48. Bajad, S. U. *et al.* Separation and quantitation of water soluble cellular metabolites by hydrophilic interaction chromatography-tandem mass spectrometry. *J. Chromatogr. A* **1125**, 76–88 (2006).

49. Reaves, M. L. & Rabinowitz, J. D. Metabolomics in systems microbiology. *Curr. Opin. Biotechnol.* **22**, 17–25 (2011).

50. Han, M. J., Lee, J. W. & Lee, S. Y. Understanding and engineering of microbial cells based on proteomics and its conjunction with other omics studies. *Proteomics* **11**, 721–743 (2011).

51. Saha, R., Chowdhury, A. & Maranas, C. D. Recent advances in the reconstruction of metabolic models and integration of omics data. *Curr. Opin. Biotechnol.* **29**, 39–45 (2014).

52. Ranganathan, S., Suthers, P. F. & Maranas, C. D. OptForce: An optimization procedure for identifying all genetic manipulations leading to targeted overproductions. *PLoS Comput. Biol.* **6**, (2010).

53. Nordberg, E., Karlsson, L., Johansson, O. H. & Liden, G. *Escherichia coli* as a Well-Developed Host for Metabolic Engineering. *Metab. Pathw. Eng. Handb. Fundam.* (2010). doi:10.1201/9781439802977.ch21

54. Jensen, M. K. & Keasling, J. D. Recent applications of synthetic biology tools for yeast metabolic engineering. *FEMS Yeast Res.* **15**, 1–10 (2015).

55. Besada-Lombana, P. B., McTaggart, T. L. & Da Silva, N. A. Molecular tools for pathway engineering in *Saccharomyces cerevisiae*. *Curr. Opin. Biotechnol.* **53**, 39–49 (2018).

56. Ila, S. *Bacillus subtilis* and its relatives: molecular biological and industrial workhorses. *Trends Biotechnol.* **10**, 247–256 (1992).

57. Vertès, A. A., Inui, M., Yukawa, H. & Verte, A. A. Manipulating *Corynebacteria*, from Individual Genes to Chromosomes- Minireview. *Appl. Environ. Microbiol.* **71**, 7633–7642 (2005).

58. Elmore, J. R., Furches, A., Wolff, G. N., Gorday, K. & Guss, A. M. Development of a high efficiency integration system and promoter library for rapid modification of *Pseudomonas putida* KT2440. *Metab. Eng. Commun.* **5**, 1–8 (2017).

59. Martínez-García, E. & de Lorenzo, V. Molecular tools and emerging strategies for deep genetic/genomic refactoring of *Pseudomonas*. *Curr. Opin. Biotechnol.* **47**, 120–132 (2017).

60. Wierckx, N. J. P., Ballerstedt, H., Bont, J. a M. De & Wery, J. Engineering of Solvent-Tolerant *Pseudomonas putida* S12 for Bioproduction of Phenol from Glucose. *Appl. Environ. Microbiol.* **71**, 8221–8227 (2005).

61. Brockmeier, U. *et al.* Systematic Screening of All Signal Peptides from *Bacillus*

subtilis: A Powerful Strategy in Optimizing Heterologous Protein Secretion in Gram-positive Bacteria. *J. Mol. Biol.* **362**, 393–402 (2006).

62. Dunlop, M. J. Engineering microbes for tolerance to next- generation biofuels. 1–9 (2011).

63. Tomko, T. A. & Dunlop, M. J. Expression of Heterologous Sigma Factor Expands the Searchable Space for Biofuel Tolerance Mechanisms. *ACS Synth. Biol.* **6**, 1343–1350 (2017).

64. Shen, C. R. *et al.* Driving forces enable high-titer anaerobic 1-butanol synthesis in *Escherichia coli*. *Appl. Environ. Microbiol.* **77**, 2905–2915 (2011).

65. Sankaranarayanan, M. *et al.* Production of 3-hydroxypropionic acid by balancing the pathway enzymes using synthetic cassette architecture. *J. Biotechnol.* **259**, 140–147 (2017).

66. Anthony, J. R. *et al.* Optimization of the mevalonate-based isoprenoid biosynthetic pathway in Escherichia coli for production of the anti-malarial drug precursor amorpha-4,11-diene. *Metab. Eng.* **11**, 13–19 (2009).

67. Ceroni, F., Algar, R., Stan, G. B. & Ellis, T. Quantifying cellular capacity identifies gene expression designs with reduced burden. *Nat. Methods* **12**, 415–418 (2015).

68. Quaglia, D., Ebert, M. C. C. J. C., Mugford, P. F. & Pelletier, J. N. Enzyme engineering: A synthetic biology approach for more effective library generation and automated high-throughput screening. *PLoS One* **12**, 1–14 (2017).

69. Dietrich, J. A., McKee, A. E. & Keasling, J. D. *High-Throughput Metabolic Engineering: Advances in Small-Molecule Screening and Selection*. *Annual Review of Biochemistry* **79**, (2010).

70. Van Rossum, T., Kengen, S. W. M. & Van Der Oost, J. Reporter-based screening and selection of enzymes. *FEBS J.* **280**, 2979–2996 (2013).

71. Dietrich, J. A., Shis, D. L., Alikhani, A. & Keasling, J. D. Transcription factor-based screens and synthetic selections for microbial small-molecule biosynthesis. *ACS Synth. Biol.* **2**, 47–58 (2013).

72. Zhang, J., Jensen, M. K. & Keasling, J. D. Development of biosensors and their

application in metabolic engineering. *Curr. Opin. Chem. Biol.* **28**, 1–8 (2015).

73. Alonso-Gutierrez, J. *et al.* Principal component analysis of proteomics (PCAP) as a tool to direct metabolic engineering. *Metab. Eng.* **28**, 123–133 (2015).

74. Tarca, A. L., Carey, V. J., Chen, X., Romero, R. & Drăghici, S. Machine Learning and Its Applications to Biology. *PLoS Comput. Biol.* **3**, e116 (2007).

75. Cohen, S. N., Chang, A. C. Y., Boyer, H. W. & Helling, R. B. Construction of Biologically Functional Bacterial Plasmids In Vitro. *Proc. Natl. Acad. Sci.* **70**, 3240–3244 (1973).

76. Jackson, D. A., Symons, R. H. & Berg, P. Biochemical Method for Inserting New Genetic Information into DNA of Simian Virus 40: Circular SV40 DNA Molecules Containing Lambda Phage Genes and the Galactose Operon of *Escherichia coli*. *Proc. Natl. Acad. Sci.* **69**, 2904–2909 (1972).

77. Shetty, R. P., Endy, D. & Knight, T. F. Engineering BioBrick vectors from BioBrick parts. *J. Biol. Eng.* **2**, 1–12 (2008).

78. Gibson, D. G. *et al.* Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nat. Methods* **6**, 343–345 (2009).

79. Torella, J. P. *et al.* Unique nucleotide sequence-guided assembly of repetitive DNA parts for synthetic biology applications. *Nat. Protoc.* **9**, 2075–2089 (2014).

80. Engler, C., Kandzia, R. & Marillonnet, S. A one pot, one step, precision cloning method with high throughput capability. *PLoS One* **3**, (2008).

81. Engler, C., Gruetzner, R., Kandzia, R. & Marillonnet, S. Golden gate shuffling: A one-pot DNA shuffling method based on type ils restriction enzymes. *PLoS One* **4**, (2009).

82. Shao, Z., Zhao, H. & Zhao, H. DNA assembler, an in vivo genetic method for rapid construction of biochemical pathways. *Nucleic Acids Res.* **37**, 1–10 (2009).

83. Gibson, D. G. *et al.* Complete Checmical Synthesis, Assembly, and Cloning of a Mycoplasma genitalium Genome. *Science (80-.).* **319**, 1215–1221 (2008).

84. Olorunniji, F. J. *et al.* Multipart DNA Assembly Using Site-Specific Recombinases from the Large Serine Integrase Family. in *Site-Specific Recombinases: Methods and Protocols* (ed. Eroshenko, N.) 303–323 (Springer New York, 2017). doi:10.1007/978-

1-4939-7169-5_19

85. Colloms, S. D. *et al.* Rapid metabolic pathway assembly and modification using serine integrase site-specific recombination. *Nucleic Acids Res.* **42**, (2014).

86. Siying, Ma; Nicholas, Tang; and Jingdong, T. DNA Synthesis, Assembly and Applications in Synthetic Biology. *Curr. Opin. Chem. Biol.* **16**, 260–267 (2013).

87. Mutalik, V. K. *et al.* Precise and reliable gene expression via standard transcription and translation initiation elements. *Nat. Methods* **10**, 354–360 (2013).

88. Redden, H. & Alper, H. S. The development and characterization of synthetic minimal yeast promoters. *Nat. Commun.* **6**, 1–9 (2015).

89. Chen, Y. J. *et al.* Characterization of 582 natural and synthetic terminators and quantification of their design constraints. *Nat. Methods* **10**, 659–664 (2013).

90. Curran, K. A. *et al.* Short Synthetic Terminators for Improved Heterologous Gene Expression in Yeast. *ACS Synth. Biol.* **4**, 824–832 (2015).

91. Salis, H. M., Mirsky, E. A. & Voigt, C. A. Automated design of synthetic ribosome binding sites to control protein expression. *Nat. Biotechnol.* **27**, 946–950 (2009).

92. Bryant, J. A., Sellars, L. E., Busby, S. J. W. & Lee, D. J. Chromosome position effects on gene expression in *Escherichia coli* K-12. *Nucleic Acids Res.* **42**, 11383–11392 (2014).

93. Flagfeldt, D. B., Siewers, V., Huang, L. & Nielsen, J. Characterization of chromosomal integration sites for heterologous gene expression in *Saccharomyces cerevisiae*. *Yeast* **26**, 545–551 (2009).

94. Apel, A. R. *et al.* A Cas9-based toolkit to program gene expression in *Saccharomyces cerevisiae*. *Nucleic Acids Res.* **45**, 496–508 (2017).

95. Jinek, M. *et al.* A Programmable Dual-RNA – Guided. **337**, 816–822 (2012).

96. Wang, H. H. *et al.* Programming cells by multiplex genome engineering and accelerated evolution. *Nature* **460**, 894–898 (2009).

97. Choi, K. R. & Lee, S. Y. CRISPR technologies for bacterial systems: Current achievements and future directions. *Biotechnol. Adv.* **34**, 1180–1209 (2016).

98. Horwitz, A. A. *et al.* Efficient Multiplexed Integration of Synergistic Alleles and

Metabolic Pathways in Yeasts via CRISPR-Cas. *Cell Syst.* **1**, 88–96 (2015).

99. Bao, Z. *et al.* Homology-Integrated CRISPR-Cas (HI-CRISPR) System for One-Step Multigene Disruption in *Saccharomyces cerevisiae*. *ACS Synth. Biol.* **4**, 585–594 (2015).

100. Larson, M. H. *et al.* CRISPR interference (CRISPRi) for sequence-specific control of gene expression. *Nat. Protoc.* **8**, 2180–2196 (2013).

101. Cleto, S., Jensen, J. V. K., Wendisch, V. F. & Lu, T. K. *Corynebacterium glutamicum* Metabolic Engineering with CRISPR Interference (CRISPRi). *ACS Synth. Biol.* **5**, 375–385 (2016).

102. Bikard, D. *et al.* Programmable repression and activation of bacterial gene expression using an engineered CRISPR-Cas system. *Nucleic Acids Res.* **41**, 7429–7437 (2013).

103. Golden, U. & Shuffling, G. cDNA Libraries. **729**, 167–181 (2011).

104. Farmer, W. R. & Liao, J. C. Improving lycopene production in *Escherichia coli* by engineering metabolic control. *Nat. Biotechnol.* **18**, 533–537 (2000).

105. Barbieri, E. M., Muir, P., Akhuetie-Oni, B. O., Yellman, C. M. & Isaacs, F. J. Precise Editing at DNA Replication Forks Enables Multiplex Genome Engineering in Eukaryotes. *Cell* **171**, 1453-1467.e13 (2017).

106. Wang, z. et al. RNA-seq: A revolutionary tool for transcriptomics. *Bioinforma. A Pract. Guid. to Anal. Genes Proteins* **10**, 57–63 (2001).

107. Tjaden, B. *et al.* Transcriptome analysis of *Escherichia coli* using high-density oligonucleotide probe arrays. *Nucleic Acids Res.* **30**, 3732–8 (2002).

108. Stoebel, D. M., Dean, A. M. & Dykhuizen, D. E. The cost of expression of *Escherichia coli* lac operon proteins is in the process, not in the products. *Genetics* **178**, 1653–1660 (2008).

109. Pasini, M. *et al.* Using promoter libraries to reduce metabolic burden due to plasmid-encoded proteins in recombinant *Escherichia coli*. *N. Biotechnol.* **33**, 78–90 (2016).

110. Van Weemen, B. K. & Schuurs, A. H. W. M. Immunoassay using antigen-enzyme conjugates. *FEBS Lett.* **15**, 232–236 (1971).

111. Pitera, D. J., Paddon, C. J., Newman, J. D. & Keasling, J. D. Balancing a heterologous mevalonate pathway for improved isoprenoid production in *Escherichia coli*. *Metab. Eng.* **9**, 193–207 (2007).

112. Lee, S. J. *et al.* Metabolic Engineering of *Escherichia coli* for Enhanced Production of l-valine, Based on Genome Comparison and In Silico Gene Knockout Simulation. *Appl. Environ. Microbiol.* **71**, 7880–7887 (2005).

113. Brunk, E. *et al.* Characterizing Strain Variation in Engineered *E. coli* Using a Multi-Omics-Based Workflow. *Cell Syst.* **2**, 335–346 (2016).

114. Lee, J., Jang, Y. S., Han, M. J., Kim, J. Y. & Lee, S. Y. Deciphering *Clostridium tyrobutyricum* metabolism based on the whole-genome sequence and proteome analyses. *MBio* **7**, 1–12 (2016).

115. Monk, J. M. *et al.* Multi-omics Quantification of Species Variation of *Escherichia coli* Links Molecular Features with Strain Phenotypes. *Cell Syst.* **3**, 238-251.e12 (2016).

116. Han, M., Jeong, K. J., Yoo, J. & Lee, S. Y. Engineering Escherichia coli for increased productivity of serine -rich proteins based on proteome profiling. *Appl. Environ. Microbiol.* **69**, 5772–5781 (2003).

117. Redding-Johanson, A. M. *et al.* Targeted proteomics for metabolic pathway optimization: Application to terpene production. *Metab. Eng.* **13**, 194–203 (2011).

118. Batth, T. S. *et al.* A targeted proteomics toolkit for high-throughput absolute quantification of *Escherichia coli* proteins. *Metab. Eng.* **26**, 48–56 (2014).

119. Nowroozi, F. F. *et al.* Metabolic pathway optimization using ribosome binding site variants and combinatorial gene assembly. *Appl. Microbiol. Biotechnol.* **98**, 1567–1581 (2014).

120. Hagen, A. *et al.* Engineering a Polyketide Synthase for in Vitro Production of Adipic Acid. *ACS Synth. Biol.* **5**, 21–27 (2016).

121. Fenn JB, Mann M, Meng CK, Wong SF, W. C. Electrospray ionization for mass spectrometry of large biomolecules. *Science (80-.).* **246**, 64–71 (1989).

122. Tanaka, K. *et al.* Protein and polymer analyses up to m/z 100 000 by laser ionization time-of-flight mass spectrometry. *Rapid Commun. Mass Spectrom.* **2**, 151–153 (1988).

123. Gygi, S. P., Corthals, G. L., Zhang, Y., Rochon, Y. & Aebersold, R. Evaluation of two-dimensional gel electrophoresis-based proteome analysis technology. *Proc. Natl. Acad. Sci.* **97**, 9390–9395 (2000).

124. Unlu, M., Morgan, M. E. & Minden, J. S. Difference gel electrophoresis: a single gel method for detecting changes in protein extracts. *Electrophoresis* **18**, 2071–2077 (1997).

125. Yates, J. R. The revolution and evolution of shotgun proteomics for large-scale proteome analysis. *J. Am. Chem. Soc.* **135**, 1629–1640 (2013).

126. Klammer, A. A. & MacCoss, M. J. Effects of modified digestion schemes on the identification of proteins from complex mixtures. *J. Proteome Res.* **5**, 695–700 (2006).

127. Tm, R. H. R., Means, G. E. & Brown, W. D. Stabilization of bovine trypsin by reductive methylation. *Biochim. Biophys. Acta* **492**, 316–321 (1977).

128. Kebarle, P. A brief overview of the present status of the mechanisms involved in electrospray mass spectrometry. *J. Mass Spectrom.* **35**, 804–817 (2000).

129. Wolff, M. M. & Stephens, W. E. A pulsed mass spectrometer with time dispersion. *Rev. Sci. Instrum.* **24**, 616–617 (1953).

130. Brown, R. S. & Lennon, J. J. Mass Resolution Improvement by Incorporation of Pulsed Ion Extraction in a Matrix-Assisted Laser Desorption/Ionization Linear Time-of-Flight Mass Spectrometer. *Anal. Chem.* **67**, 1998–2003 (1995).

131. Hoffmann, E. De & Stroobant, V. *Mass Spectrometry - Priniples and Applications. Mass spectrometry reviews* **29**, (2007).

132. El-Aneed, A., Cohen, A. & Banoub, J. Mass spectrometry, review of the basics: Electrospray, MALDI, and commonly used mass analyzers. *Appl. Spectrosc. Rev.* **44**, 210–230 (2009).

133. Altelaar, A. F. M., Munoz, J. & Heck, A. J. R. Next-generation proteomics: Towards an

integrative view of proteome dynamics. *Nat. Rev. Genet.* **14**, 35–48 (2013).

134. Silva, J. C. *et al.* Quantitative proteomic analysis by accurate mass retention time pairs. *Anal. Chem.* **77**, 2187–2200 (2005).

135. Silva, J. C., Gorenstein, M. V., Li, G.-Z., Vissers, J. P. C. & Geromanos, S. J. Absolute Quantification of Proteins by LCMS ^{E}. *Mol. Cell. Proteomics* **5**, 144–156 (2006).

136. Li, G. Z. *et al.* Database searching and accounting of multiplexed precursor and product ion spectra from the data independent analysis of simple and complex peptide mixtures. *Proteomics* **9**, 1696–1719 (2009).

137. Bond, N. J., Shliaha, P. V., Lilley, K. S. & Gatto, L. Improving Qualitative and Quantitative Performance for MS ^{E} -based Label-free Proteomics. *J. Proteome Res.* **12**, 2340–2353 (2013).

138. Ong, S.-E. *et al.* Stable Isotope Labeling by Amino Acids in Cell Culture, SILAC, as a Simple and Accurate Approach to Expression Proteomics. *Mol. Cell. Proteomics* **1**, 376–386 (2002).

139. Gygi, S. P. *et al.* Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. *Nat. Biotechnol.* **17**, 994–9 (1999).

140. Ross, P. L. *et al.* Multiplexed Protein Quantitation in *Saccharomyces cerevisiae* Using Amine-reactive Isobaric Tagging Reagents. *Mol. Cell. Proteomics* **3**, 1154–1169 (2004).

141. Zhu, W., Smith, J. W. & Huang, C. M. Mass spectrometry-based label-free quantitative proteomics. *J. Biomed. Biotechnol.* **2010**, (2010).

142. Liu, H., Sadygov, R. G. & Yates, J. R. A model for random sampling and estimation of relative protein abundance in shotgun proteomics. *Anal. Chem.* **76**, 4193–4201 (2004).

143. Florens, L. *et al.* Analyzing chromatin remodeling complexes using shotgun proteomics and normalized spectral abundance factors. *Methods* **40**, 303–311 (2006).

144. Voyksner, R. D. & Lee, H. Investigating the use of an octupole ion guide for ion storage and high-pass mass filtering to improve the quantitative performance of

electrospray ion trap mass spectrometry. *Rapid Commun. Mass Spectrom.* **13**, 1427–1437 (1999).

145. Chelius, D. & Bondarenko, P. V. Quantitative profiling of proteins in complex mixtures using liquid chromatography and mass spectrometry. *J. Proteome Res.* **1**, 317–323 (2002).

146. Studier, F. W. & Moffattf, B. A. Use of Bacteriophage T7 RNA Polymerase to Direct Selective High-level Expression of Cloned Genes. 113–130 (1986).

147. Datsenko, K. A. & Wanner, B. L. One-step inactivation of chromosomal genes in *Escherichia coli* K-12 using PCR products. *Proc. Natl. Acad..Sci. USA* **97**, 6640–6645 (2000).

148. Kok, S. De *et al.* Rapid and reliable DNA assembly via ligase cycling reaction. *ACS Synth. Biol.* **3**, 97–106 (2014).

149. Walker, S. a & Klaenhammer, T. R. The effect of increasing plasmid size on transformation efficiency in Escherichia coli. *J. Exp. Microbiol. Immunol.* **2**, 207–223 (2002).

150. Yuzawa, S. *et al.* Comprehensive in Vitro Analysis of Acyltransferase Domain Exchanges in Modular Polyketide Synthases and Its Application for Short-Chain Ketone Production. *ACS Synth. Biol.* **6**, 139–147 (2017).

151. Pemberton, J. M., Vincent, K. M. & Penfold, R. J. Cloning and heterologous expression of the violacein biosynthesis gene cluster from *Chromobacterium violaceum*. *Curr. Microbiol.* **22**, 355–358 (1991).

152. Klemm, D., Heublein, B., Fink, H. & Bohn, A. Polymer Science Cellulose : Fascinating Biopolymer and Sustainable Raw Material Angewandte. 3358–3393 (2005). doi:10.1002/anie.200460587

153. Béguin, P. & Aubert, J.-P. The biological degradation of cellulose. *FEMS Microbiol Rev* **13**, 25–58 (1994).

154. Teeri, T. T. Crystalline cellulose degradation: New insight into the function of cellobiohydrolases. *Trends Biotechnol.* **15**, 160–167 (1997).

155. Lynd, L. R., Zyl, W. H. Van, Mcbride, J. E. & Laser, M. Consolidated bioprocessing of

cellulosic biomass : an update. 577–583 (2005). doi:10.1016/j.copbio.2005.08.009

156. Duedu, K. O. & French, C. E. Characterization of a *Cellulomonas fimi* exoglucanase/xylanase-endoglucanase gene fusion which improves microbial degradation of cellulosic biomass. *Enzyme Microb. Technol.* **93**–**94**, 113–121 (2016).

157. Petersen, T. N., Brunak, S., Von Heijne, G. & Nielsen, H. SignalP 4.0: Discriminating signal peptides from transmembrane regions. *Nat. Methods* **8**, 785–786 (2011).

158. Bitter, G. a, Chen, K. K., Banks, a R. & Lai, P. H. Secretion of foreign proteins from *Saccharomyces cerevisiae* directed by alpha-factor gene fusions. *Proc. Natl. Acad. Sci. U. S. A.* **81**, 5330–5334 (1984).

159. Snoek, T. *et al.* Large-scale robot-assisted genome shuffling yields industrial *Saccharomyces cerevisiae* yeasts with increased ethanol tolerance. *Biotechnol. Biofuels* **8**, 1–19 (2015).

160. Gietz, R. D. & Woods, R. A. Transformation of yeast by lithium acetate/single-stranded carrier DNA/polyethylene glycol method. *Methods Enzymol.* **350**, 87–96 (2002).

161. Baxter, S. Directed evolution of an industrial N -acetyl-amino acid racemase. (2011).

162. Xu, W., Reddy, N. & Yang, Y. Extraction, characterization and potential applications of cellulose in corn kernels and Distillers' dried grains with solubles (DDGS). *Carbohydr. Polym.* **76**, 521–527 (2009).

163. Mori, A. *et al.* Signal peptide optimization tool for the secretion of recombinant protein from *Saccharomyces cerevisiae*. *J. Biosci. Bioeng.* **120**, 518–525 (2015).

164. Liu, C.-K. The cellulose degradation system of Cytophaga hutchinsonii. *PhD Thesis, Univ. Edinburgh* (2012).

165. Werner, W., City, A., Gawehn, K., Gmbh, B. M. & Tutzing, B. W. On the Properties of a New Chromogen /or the Determination of Glucose in Blood According to the GOD/POD Method. *Anal. Bioanal. Chem.* 224–228 (1970).

166. Hultman, T., Stahl, S., Homes, E. & Uhlén, M. Direct solid phase sequencing of genomic and plasmid DNA using magnetic beads as solid support. *Nucleic Acids Res.* **17**, 4937–4946 (1989).

167. Aaij, C. & Borst, P. The gel electrophoresis of DNA. *BBA Sect. Nucleic Acids Protein Synth.* **269**, 192–200 (1972).

168. Che, A., Knight, T., Canton, B., Kelly, J. & Shetty, R. Method for Assembly of Polynucleic Acid Sequences. (2010).

169. Lovett, S. T. The DNA exonucleases of *Escherichia coli*. *EcoSal Plus* **4**, (2014).

170. Liu, G. *et al.* Structural basis for the recognition of sulfur in phosphorothioated DNA. *Nat. Commun.* doi:10.1038/s41467-018-07093-1

171. Zhou, X. *et al.* A novel DNA modification by sulphur. **57**, 1428–1438 (2005).

172. Putney, S. D., Benkovic, S. J. & Schimmel, P. R. A DNA fragment with an alpha-phosphorothioate nucleotide at one end is asymmetrically blocked from digestion by exonuclease III and can be replicated in vivo. *Proc. Natl. Acad. Sci.* **78**, 7350–7354 (1981).

173. Vosberg, H. P. & Eckstein, F. Effect of deoxynucleoside phosphorothioates incorporated in DNA on cleavage by restriction enzymes. *J. Biol. Chem.* **257**, 6595–6599 (1982).

174. Schöfl, G. *et al.* 2.7 million samples genotyped for HLA by next generation sequencing: Lessons learned. *BMC Genomics* **18**, 1–16 (2017).

175. Van Pijkeren, J. P., Neoh, K. M., Sirias, D., Findley, A. S. & Britton, R. A. Exploring optimization parameters to increase ssDNA recombineering in *Lactococcus lactis* and *Lactobacillus reuteri*. *Bioengineered* **3**, 209–217 (2012).

176. Mosberg, J. A., Lajoie, M. J. & Church, G. M. Lambda red recombineering in *Escherichia coli* occurs through a fully single-stranded intermediate. *Genetics* **186**, 791–799 (2010).

177. Gay, P., Le Coq, D., Steinmetz, M., Ferrari, E. & Hoch, J. A. Cloning structural gene sacB, which codes for exoenzyme levansucrase of *Bacillus subtilis*: Expression of the gene in *Escherichia coli*. *J. Bacteriol.* **153**, 1424–1431 (1983).

178. Gay, P., Le Coq, D., Steinmetz, M., Berkelman, T. & Kado, C. I. Positive selection procedure for entrapment of insertion sequence elements in gram-negative bacteria. *J Bacteriol* **164**, 918–921 (1985).

179. Henikoff, S. Unidirectional digestion with exonuclease III creates targeted breakpoints for DNA sequencing. *Gene* **28**, 351–359 (1984).

180. Restriction Endonucleases - Survival in a Reaction. Available at: https://www.neb.com/tools-and-resources/usage-guidelines/restriction-endonucleases-survival-in-a-reaction.

181. Davis, J. H., Rubin, A. J. & Sauer, R. T. Design, construction and characterization of a set of insulated bacterial promoters. *Nucleic Acids Res.* **39**, 1131–1141 (2011).

182. Silva-Rocha, R. *et al.* The Standard European Vector Architecture (SEVA): A coherent platform for the analysis and deployment of complex prokaryotic phenotypes. *Nucleic Acids Res.* **41**, 666–675 (2013).

183. Weinstock, M. T., Hesek, E. D., Wilson, C. M. & Gibson, D. G. *Vibrio natriegens* as a fast-growing host for molecular biology. *Nat. Methods* **13**, 849–851 (2016).

184. Weber, E., Engler, C., Gruetzner, R., Werner, S. & Marillonnet, S. A modular cloning system for standardized assembly of multigene constructs. *PLoS One* **6**, (2011).

185. Sarrion-Perdigones, A. *et al.* GoldenBraid: An iterative cloning system for standardized assembly of reusable genetic modules. *PLoS One* **6**, (2011).

186. Van Hove, B., Guidi, C., De Wannemaeker, L., Maertens, J. & De Mey, M. Recursive DNA Assembly Using Protected Oligonucleotide Duplex Assisted Cloning (PODAC). *ACS Synth. Biol.* **6**, 943–949 (2017).

187. Storch, M. *et al.* BASIC: A New Biopart Assembly Standard for Idempotent Cloning Provides Accurate, Single-Tier DNA Assembly for Synthetic Biology. *ACS Synth. Biol.* **4**, 781–787 (2015).

188. Zhao, G. *et al.* Enzymatic cleavage of type II restriction endonucleases on the 2'-O-methyl nucleotide and phosphorothioate substituted DNA. *PLoS One* **8**, 1–13 (2013).

189. Alper, H. *et al.* Tuning genetic control through promoter engineering. *Pnas* **103**, 3006–3007 (2006).

190. Mainguet, S. E., Gronenberg, L. S., Wong, S. S. & Liao, J. C. A reverse glyoxylate shunt to build a non-native route from C4 to C2 in *Escherichia coli*. *Metab. Eng.* **19**, 116–127 (2013).

191. Jhamb, K. & Sahoo, D. K. Production of soluble recombinant proteins in *Escherichia coli*: Effects of process conditions and chaperone co-expression on cell growth and production of xylanase. *Bioresour. Technol.* **123**, 135–143 (2012).

192. Smith, P. K. *et al.* Measurement of protein using bicinchoninic acid. *Anal. Biochem.* **150**, 76–85 (1985).

193. Noor, E., Eden, E., Milo, R. & Alon, U. Central Carbon Metabolism as a Minimal Biochemical Walk between Precursors for Biomass and Energy. *Mol. Cell* **39**, 809–820 (2010).

194. Glick, B. R. Metabolic load and heterogenous gene expression. *Biotechnol. Adv.* **13**, 247–261 (1995).

195. Lim, H. N., Lee, Y. & Hussein, R. Fundamental relationship between operon organization and gene expression. *Proc. Natl. Acad. Sci.* **108**, 10626–10631 (2011).

196. Maloy, S. R. & Nunn, W. D. Genetic regulation of the glyoxylate shunt in *Escherichia coli* K-12. *J. Bacteriol.* **149**, 173–180 (1982).

197. Lee, C., Kim, J., Shin, S. G. & Hwang, S. Absolute and relative QPCR quantification of plasmid copy number in *Escherichia coli*. **123**, 273–280 (2006).

198. Green, MR and Sambrook, J and Sambrook, J. *Molecular cloning: a laboratory manual*. (Cold Spring Harbor Laboratory Press, 2012).

199. Kahl, L. J. & Endy, D. A survey of enabling technologies in synthetic biology A survey of enabling technologies in synthetic biology. *J. Biol. Eng.* **7**, 1 (2013).

200. Borkowski, O., Ceroni, F., Stan, G. B. & Ellis, T. Overloaded and stressed: whole-cell considerations for bacterial synthetic biology. *Curr. Opin. Microbiol.* **33**, 123–130 (2016).