Investigating Metabolic Burden and Stress Responses in Prokaryotic Model Organisms through an Integrated Computational and Synthetic Biology Approach

David Markham

Abstract

Expression of heterologous proteins in bacteria is commonplace in research and industry across many disciplines including medicine, agriculture, and civil engineering. Overexpression of proteins may result in greater yields; however, it can induce a stress response from metabolic burden which can reduce yields by mechanisms like cell death or sporulation. This thesis investigated the stress responses in prokaryotic model organisms and employed a combined approach of computational methods and synthetic biology to develop systems to detect these stresses. Furthermore, this project aimed to work toward portable systems - the ability to transfer a system from one species to another with minimal refactoring.

This work comprised of three major components:

1. A condition specific biomarker selection algorithm that takes an input of regulation data under different conditions and generates sets of biomarkers that best classify the selected condition versus all other conditions. This thesis thoroughly examines this algorithm through testing on diverse datasets including tiling array and RNA-seq regulation data.
2. A genetic logic gate was designed to use stress-specific biomarkers as inputs to produce a measurable output to detect the stress response. A testing system was developed to work between two species and early characterisation of this system was completed.
3. *in vivo* testing of a cross-species codon optimisation algorithm. A framework for testing coding sequences generated by a cross-species codon optimisation algorithm has been developed for this study.

In summary, this research explored approaches to detect stress responses in prokaryotic model organisms and worked toward making systems portable without the need to refactor between species. The approaches developed formed a foundation for combined computational and synthetic biology approaches.

Declaration

I confirm that all work presented within this thesis is my own work, unless explicitly stated. The work presented herein has not been submitted for the degree of PhD in Computing Science at Newcastle University and for no other degree at any other institution.

David Charles Markham

November 2023

COVID-19 Impact Statement

The work conducted in this thesis was significantly impacted by the COVID-19 pandemic. Halfway through the first year of my PhD, during early stages of my lab work, the lockdown was imposed, and the university was closed for a prolonged period. When I was able to return to the lab, there were severe restrictions imposed removing my access to our usual lab and office spaces and moving to an alternative lab space with different equipment. The disruption during this extended period of several months was extremely difficult to adapt to within the first year of my studies. Supervision was difficult to achieve due to the inability to interact with my supervisors and other colleagues face-to-face, dramatically slowing the pace of learning new skills and thus the progress of my PhD.

Purchasing of new equipment, including the equipment for the automation lab essential for my planned work, was delayed by the lockdown. In addition, the set-up and training on new equipment was delayed. These delays on essential equipment resulted in various changes to the planned work of my thesis which made the completion of aspects of work to be delayed, not completed, or changed to an alternative strategy. For example, the cloning of constructs for Chapter 5 was attempted to be automated by the new equipment available but had to be changed to manual cloning due to these setbacks.

COVID-19 also had a direct impact on my health as I had to isolate for symptomatic COVID during my studies. As a knock-on effect on my health, the stresses induced by the pandemic caused a prolonged bout of anxiety which frequently left me too physically fatigued or ill to continue my studies effectively.

The combination of all the effects from the COVID-19 pandemic resulted in me running out of time to complete the worked that I had intended for this thesis. Briefly, the work incomplete included: testing the AND gate *in vivo*; testing the Chimera Evolve coding sequences in *B. subtilis*; testing ROTC on a compiled dataset of *E. coli*. Additionally, it would have been preferable to have more time to continue the work on the different projects within the thesis. For example, testing the Chimera Evolve coding sequences with the addition of a burden generating device to limit the shared resources in the cell.

Acknowledgements

I want to express my sincere gratitude to my formal supervisors, Anil Wipat and Katherine James, for their unwavering guidance and support throughout this research journey. Their expertise has been a game-changer for me. A special shoutout to my informal supervisors, Wendy Smith, and Bradle Brown, for being exceedingly helpful with their support and mentoring throughout my studies and especially while I was trying to find my feet in the group. I owe a huge thanks to James Skelton, for consistently bailing me out of computational crises and bioinformatics puzzles, and to Polly Noble, for always helping me troubleshoot in the lab and being a supportive friend when I needed it most.

I would like to acknowledge everyone in the ICOS group and close friends who have contributed over the years. To name but a few: Brown Bradley, Chris Atallah, Dan (frog murderer) Herring, Jasmine Bird, Nadia Rostami, Ming Li, Yiming Huang, Zhen Ou, Silvia Navarro, and everyone else. A special acknowledgment goes to Dan (not frog murderer) and Josie Todd for their support during my thesis write-up and the exciting opportunity to join their new venture.

Lastly, heartfelt thanks to my family and friends for being my pillars of strength over these four years. Your encouragement has meant the world to me. This thesis is a culmination of the collective efforts and support of these amazing individuals, and I'm truly grateful for their contributions to my academic and personal growth.

Table of Contents

Table of Contents

Table of Contents

Table of Contents

Table of Contents

Table of Contents

Table of Figures

Table of Figures

Table of Tables

Acronyms

SMM Spizizen's Minimal Media

BMM Belitsky's Minimal Media

LB Luria–Bertani (media)

M9 M9 minimal media

IPTG Isopropyl ß-D-1-thiogalactopyranoside

CDS Coding Sequence

TU Transcription Unit

RBS Ribosome Binding Site

DNA Deoxyribonucleic Acid

cDNA Copy Deoxyribonucleic Acid

RNA Ribonucleic Acid

mRNA Messenger Ribonucleic Acid

tRNA Transfer Ribonucleic Acid

ncRNA Non-coding Ribonucleic Acid

asRNA Anti-sense Ribonucleic Acid

GFP Green Fluorescent Protein

RFP Red Fluorescent Protein

BFP Blue Fluorescent Protein

OD Optical Density

PCR Polymerase Chain Reaction

qPCR Quantitative Polymerase Chain Reaction

RT-qPCR Reverse Transcriptase Qualitative Polymerase Chain Reaction

RNA-seq Ribonucleic Acid Sequencing

NGS Next Generation Sequencing

QC Quality Control

TPM Transcripts Per Million

MSA Multiple Sequence Alignment

SH Shorthand

PCA Principal Component Analysis

AND logical and

RNAP Ribonucleic Acid Polymerase

CHAPTER 1

1 Introduction

1 Introduction

1.1 Motivation

Bacteria are exposed to various environmental conditions that may induce stress in the organism (Petersohn *et al.*, 2001; Beales, 2004; Bonilla, 2020; Scott *et al.*, 2010; Nicolas *et al.*, 2012; Smith *et al.*, 2023). These conditions may be easily monitored and controlled such as temperature, or they may be more complex such as intracellular or extracellular chemical or metabolic imbalances (Cabeen *et al.*, 2017; Guo and Gross, 2014; Vandana, Priyadarshanee and Das, 2023). Exposure to these conditions may trigger a stress response within the cell which can lead to negative impacts such as decreased metabolism or even cell death (Guan *et al.*, 2017; Kurland and Dong, 1996; Rath and Das, 2023). Understanding specific stress responses within bacteria is essential to mitigate the negative impacts of bacterial stress.

The research conducted in this thesis is centred around metabolic load-stress or metabolic burden, which is the metabolic exertion imposed when the organism is producing large amounts of a given product (Lynch and Marinov, 2015; Glick, 1995; Ceroni *et al.*, 2015; Snoeck, Guidi and De Mey, 2024). This stress is common in industrial settings where bacteria are grown to extract a desired product (Kurland and Dong, 1996; Jiang *et al.*, 2020).

Stress response in bacteria can be determined at a genetic level. When bacteria undergo stress, they will alter the regulation of certain genes and therefore change the direction of related metabolic pathways (De Nadal, Ammerer and Posas, 2011; Njenga *et al.*, 2023). For example, a common stress response among gram positive bacteria is sporulation. Sporulation is a state where the cell down regulates many metabolic pathways to reduce its nutritional requirement whilst strengthening its cell structure allowing it to survive in harsh conditions (Errington, 1993; Reder *et al.*, 2012; Freire *et al.*, 2023).

With an understanding of the genetic basis of a specific stress response, it is possible to select biomarkers that are specific to this stress and, thus determine when bacteria are experiencing this stress (den Besten *et al.*, 2010; Peña-Montenegro *et al.*, 2023). A biomarker in this instance is a gene that is regulated differently during the effects of the stress. It is important that the biomarkers are specific to the stress under investigation so that they may appropriately respond to

the stress. If it is known that the bacteria are undergoing load-stress, then it is possible to down regulate the pathway responsible for the stress, thus, mitigating it (Ceroni *et al.*, 2018). However, if the biomarkers are not specific, the bacteria may be undergoing a different type of stress therefore, down regulating the pathway may not alleviate the stress whatsoever.

The discipline of synthetic biology provides tools that can be applied to mitigating the stress response (Ceroni *et al.*, 2018; Pasini *et al.*, 2016; Darlington *et al.*, 2017; Wu *et al.*, 2016; Lo *et al.*, 2016). Synthetic biology allows for the introduction of genetic devices into organisms that may provide a new metabolic pathway, thus creating a desirable chemical product (Chotani *et al.*, 2000; Ingram *et al.*, 2010; Mugwanda *et al.*, 2023). As discussed previously, this can introduce a burden on the organism's metabolism (Ceroni *et al.*, 2015; Snoeck, Guidi and De Mey, 2024). Given knowledge of the stress response at a genetic level, the stress response can be detected and then countered with a complementary synthetic device that down-regulates the original pathway that introduced the stress. This will create a negative feedback loop that helps control the stress, therefore, increasing the efficiency of the pathway, in theory (Boo, Ellis and Stan, 2019).

Synthetic biology introduces the concept of logic gates *in vivo* (Moser *et al.*, 2012; Kim *et al.*, 2018; Bose *et al.*, 2023). This allows for further developing the synthetic devices introduced into the organisms and giving them increased specificity or functionality. For example, it may be necessary to use two biomarkers as an input to our synthetic device to allow it to respond specifically to load-stress and not to other stresses. This would feed in as two inputs of an AND gate where both biomarkers need to be present to activate the synthetic device (Shis and Bennett, 2013; Vishweshwaraiah *et al.*, 2021).

Synthetic devices can be further optimised by utilising different parts that may work better for different scenarios. The typical building blocks of a genetic device include: promoters that regulate transcription of the gene under the presence of specific small molecules or proteins; ribosome binding sites (RBS) that allow the ribosome to bind and initiate transcription; coding sequence (CDS) that carries the genetic information encoding the protein itself; a terminator which stops transcription. Each of these parts can be optimised to better suit the functionality of the device (Ceroni

et al., 2015; Stock and Gorochowski, 2024). This leads to a series of combinations of parts that need to be tested to determine which is more successful, a process known as combinatorial design (Naseri and Koffas, 2020).

Selection of a host species as a chassis for synthetic biology projects is an important part of the design stage (Freemont, 2019; Ma *et al.*, 2024; Xu *et al.*, 2023). However, moving from one species to another is not trivial and often requires significant refactoring before systems will work in a different species. Refactoring, in this context, refers to the practice of modification of an existing synthetic biological system (such as a genetic circuit) so that it may operate within a new chassis. Different host species, or even strains, have significant differences in their genes and metabolism which may result in synthetic systems not functioning the same way as they would in another host (Hitchcock, Hunter and Canniffe, 2020). As such, refactoring is often necessary: parts such as promoters and RBSs are often replaced; CDS components may need to be optimised for the host's codon usage bias (CUB); different vectors (e.g., plasmids) may have to be used with appropriate origins of replication for the host; etc. (de Lorenzo, Krasnogor and Schmidt, 2021).

The multidisciplinary Portabolomics project was initiated at Newcastle University with the aim to develop systems for synthetic biology that are portable between species (Krasnogor *et al.*, 2023). The Portabolomics project aimed towards the development of a "bio-adaptor" which can provide a standardised interface between a genetic circuit and the host organism to mitigate the dependence on refactoring when porting systems from one species to another.

The work in this thesis was conducted as part of the Portabolomics group, therefore, systems developed in this project were designed to be portable between organisms. The organisms chosen for this are *Escherichia coli* and *Bacillus subtilis* as both are well studied, industrially relevant and are each considered to be the model organism for gram-negative and gram-positive bacteria, respectively. This work was attributed to one arm of the Portabolomics project which was concerned with developing a genetic circuit which can identify when a host organism is undergoing a specific stress state and coupled with a computational and analytical approach that can select specific biomarkers for any given species. The project aimed to integrate this work with other systems that operate as the "bio-adaptor", aiding portability.

1.2 Research Question

“Can biological and computational systems be developed toward the mitigation of burden and stress responses that are portable between different prokaryotic model organisms?”

1.3 Aims and Objectives

The following research aims were addressed in this thesis:

- Develop systems that aid in the detection, mitigation, and understanding of stress and metabolic burden in model prokaryotes.
- Develop systems designed to operate portably between species, using the model organisms: *Escherichia coli* and *Bacillus subtilis.*

The research aims were approached by meeting these objectives:

- Describe and demonstrate the capability of the biomarker selection algorithm, ROTC, to produce sets of biomarkers of a given stress state.
- Design, build, and test a genetic dual-input AND gate toward the detection of a specific stress state.
- Attain *in vivo* experimental evidence to validate the effectiveness of the codon optimisation algorithm, Chimera Evolve.

1.4 Contributions and Outcomes

- An algorithm for biomarker selection, ROTC, has been introduced and investigated in detail to make it suitable for public use by researchers
    - ROTC was interrogated to determine an accurate model description to accompany the algorithm's release
    - ROTC was applied to two data sets of the transcriptome of *Bacillus subtilis* 168 under different stress conditions to investigate the effectiveness of generated sets of biomarkers
- A test system, genetic AND gate, was designed, built, and tested that was intended to be portable between *E. coli* and *B. subtilis*
    - A test system was developed which was designed to operate in both *B. subtilis* and *E. coli* whereas the inspiration for the design was only tested in *E. coli*
    - Each component of the AND gate was tested *in vivo* in *E. coli* and/or *B. subtilis* and demonstrated to function as expected albeit with room for optimisation
    - The AND gate was tested *in vitro* in an *E. coli* S30 extract cell free expression system and was not observed to function as expected
- Experimental evidence was gathered, *in vivo*, for the codon optimisation algorithm, Chimera Evolve
    - Five codon variants of the coding sequence for a green fluorescent protein, mGreenLantern, were generated by Chimera Evolve with a gradient of bias toward *E. coli* and *B. subtilis*
    - The codon variants were tested, alongside the unaltered sequence *in vivo* in *E. coli* DH5α but did not show the expected linear trend of difference in expression

1.5 Thesis Structure

The rest of this thesis is comprised of seven additional chapters. Chapter 2 contains background research, introducing key themes within this thesis and reviewing important papers related to this work.

Chapter 3 is the first research chapter and describes the biomarker selection algorithm ROTC (Recursive Orthogonal Threshold Classifier). Within chapter 3, ROTC is tested on two transcriptomic data sets of *Bacillus subtilis* 168.

Chapter 4 is a research chapter regarding the design, build and test of a genetic dual-input AND gate. Components of the AND gate were tested with simple inducible promoters in *Escherichia coli* DH5α, *Escherichia coli* BL21(DE3), *Bacillus subtilis* 168, and an *Escherichia coli* S30 cell extract system.

Chapter 5 is a research chapter describing the experimental validation of the codon optimisation algorithm, Chimera Evolve. In this chapter, many CDS variants of fluorescent proteins are generated by the algorithm and tested *in vivo* in *Escherichia coli* DH5α.

Chapter 6 rounds out the research covered in the previous chapters and formulates conclusions based on the research undertaken. Chapter 7 is a description of the methods used throughout the experiments of this work.

1.5.1 Acknowledgements for contribution

The ROTC algorithm, as introduced in Chapter 3, was developed before the work conducted in this thesis by members of Anil Wipat's research group. The version of ROTC worked upon within this thesis was developed by James Knight with significant rewrites conducted by myself, Anil Wipat, and Yiming Huang.

The datasets used in Chapter 3 for ROTC were sourced from the work of the BaSysBio consortium in their publication "Condition-dependent transcriptome reveals high-level regulatory architecture in *Bacillus subtilis*" (Nicolas *et al.*, 2012). Further RNA-seq experiments were conducted by Wendy Smith which contributed to an additional dataset.

The Chimera Evolve codon optimisation algorithm, as described in Chapter 5, was developed by James Skelton (Skelton *et al.*, 2020).

CHAPTER 2

2 Background

2 Background

2.1 Introduction

Bacteria undergo a variety of different stresses in their natural environments against which they have adapted (Bonilla, 2020; Beales, 2004; Moon *et al.*, 2023). When bacteria are subjected to stressors, they may exhibit a stress response to counteract the stress and merit their survival (Guan *et al.*, 2017; Angelini *et al.*, 2023). Despite the stress response being adapted to counteract the negative effects of a stress, it comes with a trade-off, often metabolic, which inhibits the growth of the population (Kurland and Dong, 1996; Guan *et al.*, 2017; Vasilakou, Van Loosdrecht and Wahl, 2020).

When bacteria are taken out of their natural environment and moved into a controlled environment, such as a batch fermenter vessel, it comes with a new slew of potential stresses (Moser *et al.*, 2012; Geissler *et al.*, 2022). Though physical and chemical stresses, such as temperature and inflow of nutrients, can be easier to maintain in a controlled environment, there are often problems in deliberately growing bacteria for production or research purposes. The major contributor of stress in a controlled environment is the phenomena known as metabolic load stress or burden which occurs when an organism is made to over-exert its metabolism (Glick, 1995; Cordell *et al.*, 2023). Bacteria are often used to produce heterologous proteins which can induce this metabolic burden (Wu *et al.*, 2016; Snoeck, Guidi and De Mey, 2024). Mitigating the impact of metabolic burden is essential to optimise the yields of engineered bacteria in research and industry (De Nadal, Ammerer and Posas, 2011; Naseri and Koffas, 2020).

Stress responses are regulated at a genetic level and typically divert the cells metabolism to processes that may directly mitigate the stress (Petersohn *et al.*, 2001; Tollerson and Ibba, 2020). As such, a stress response will negatively impact the growth of a cell population. Sporulation is an example of a stress response which frequently occurs in gram-positive bacteria which typically occurs when a cell runs out of nutrients required to survive, so the cell turns into a dormant spore (Reder *et al.*, 2012; Errington, 1993; Mutlu *et al.*, 2020). Clearly, sporulation is a behaviour best avoided in an industrial setting where production needs to continue. Additionally, some stress responses may be insufficient to aid the organism's survival and result in cell death, however, if the signals of a stress response can be

detected early then it may be possible to act against the stress and curb the problems before they start (Dahl *et al.*, 2013; Zhang *et al.*, 2022).

2.2 Gene regulation in prokaryotes

2.2.1 Biomarkers

The term “biomarker” is short for biological marker and is typically used in a medical setting to refer to a natural indicator of a disease (Strimbu and Tavel, 2010; Hansson, 2021). For example, a biomarker could be gene that is commonly associated with a given disease. In broader terms, the term biomarker may be used for many purposes referring to any measurable indicator of an interaction within or between living systems. In the context of this thesis, a biomarker is a gene whose expression can be measured as an indicator of a given cell state.

It is important to have a proper understanding of how gene regulation functions in prokaryotes when determining biomarkers. Unlike eukaryotes, prokaryotes typically have a more straight-forward approach to gene regulation so certain mechanisms like alternative splicing do not complicate things. However, prokaryotes still have interlinked regulatory networks that can be complex and there are many elements of certain species (e.g., functions of genes) that are still unknown.

Computational methods can be used to select biomarkers given sufficiently well annotated gene expression data (den Besten *et al.*, 2010; Huang *et al.*, 2023). If a biomarker gene needs to be exploited for autonomic regulation of a cell state, much needs to be known about the regulation of that gene. In this research, two model organisms have been selected for which there is an abundance of information about the gene function of these organisms: *Escherichia coli* and *Bacillus subtilis*.

2.2.2 The structure of a gene

The term gene is ambiguous; in general terms a gene is a single hereditary unit, but the definitions is confusing when considering what a gene is at a molecular level (Pearson, 2006; Fujiyoshi *et al.*, 2021). Often, a gene is defined as the sequence of nucleotides that is transcribed into an RNA sequence during transcription. Yet, there are differences between the nucleotides that get transcribed and those that are coding sequences for a given protein. In prokaryotes, several “genes” form an operon which are regulated by the same promoter thus are all contained within the

same mRNA molecule but are translated into several separate proteins (Kraikivski, 2021). A coding sequence (CDS) are the nucleotides which get translated into a protein within a given reading frame. Usually, a protein will be named and the CDS given the same name and called a “gene”.

For the sake of clarity with terminology, this thesis will typically refer to CDS when referring to a sequence that encodes a protein and the term gene or feature is used when referring to biomarkers or candidate biomarkers. The genes or features selected as biomarkers do not necessarily encode proteins, therefore, they are not referred to as a CDS but are referred to as genes because the sequence has been given a gene name.

An operon is a collection of genes that are all under the control of the same promoter, and a promoter is the name given to the sequence of nucleotides responsible for enabling or inhibiting the transcription of a gene (Foley, Cockburn and Koropatkin, 2016). Operons often have a transcription factor in common and typically share a nomenclature (Kraikivski, 2021). A transcription factor is a molecule, usually a protein, that assists in the regulation of a promoter. Typically, a transcription factor will bind upstream or downstream of a promoter sequence which changes its confirmation to either enable or repress transcription (Lewis, Doherty and Clarke, 2008; Rodriguez Ayala, Bartolini and Grau, 2020).

2.3 Transcriptomics

Transcriptomics is the study of the total RNA within a cell or cells, i.e., the transcriptome. To measure the RNA within cells for transcriptomic studies, the RNA is usually reverse transcribed into cDNA for quantification or sequencing. Up until the advent of NGS, microarrays used to be the primary method of transcriptomic analysis. RNA-seq was a game-changer in the field of transcriptomics, thus has taken over as the preferred method for transcriptomics (Lowe *et al.*, 2017; Thind *et al.*, 2021).

Microarrays use a solid surface called a chip to which many DNA probes are fixed (Lucchini, Thompson and Hinton, 2001; Israr *et al.*, 2024). The target DNA sequences are added with fluorescent probes which will attach to complementary probes on the chip’s surface. After washing steps, only the DNA that bound remains,

thus allowing for detection and quantification of specific regions of a genome. Microarrays are used for various purposes as the chips can be designed to target various sequences depending on the intent. Tiling arrays are a subset of microarrays designed for quantifying the transcriptome across the entire genome of an organism, including unknown regions. Tiling arrays can conduct entire genome, transcriptomic analysis by specialising the fixed probes to bind to contiguous regions of DNA (Mockler and Ecker, 2005; Jiang *et al.*, 2021).

RNA-seq leverages NGS technologies to fully the sequence the total RNA within a which is achieved by fragmentation and converting to cDNA (Corchete *et al.*, 2020). Unlike microarrays, RNA-seq is not reliant on probes that bind to specific known sequences of a genome and instead it sequences the entire transcriptome with no prior knowledge necessary. Downstream of sequencing, the reads are then assembled into contigs (if necessary) and mapped to a reference genome for quantification. Advantages of RNA-seq over microarrays include fewer background noise and a large dynamic range. RNA-seq does require library preparation in which poly-A sequences and adapters may be added and amplification may be performed depending on the method, which can introduce potential errors downstream (Wang, Gerstein and Snyder, 2009; Shi *et al.*, 2021).

2.3.1 Transcriptome of Bacillus subtilis 168

In 2012, the BaSysBio project published a large data set of the condition-dependent transcriptome of *B. subtilis* 168 (Nicolas *et al.*, 2012). The data were generated by a series of experiments subjecting *B. subtilis* 168 to conditions that the bacteria would experience in its natural environment as well as common research settings. The transcriptomes were measured by means of tiling array which gathered data for all known genes, small features such as small RNAs, and non-coding regions. The extent of this dataset has allowed for many studies to be completed using it as a basis for analysis, including Subtiwiki which is an online repository of genetic information for *B. subtilis* (Zhu and Stülke, 2018; Faria *et al.*, 2016).

Huang *et al*. conducted a study, utilising the BaSysBio tiling array data set, to identify a biomarker panel capable of distinguishing between different stress states (Huang *et al.*, 2021). The data were filtered and processed using Uniform Manifold Approximation and Projection (UMAP) for dimension reduction and then clustered

using a Leidan clustering algorithm to generate the clusters shown in figure 2.3.1 (Traag, Waltman and Van Eck, 2019; McInnes, Healy and Melville, 2020). The clusters suggest that certain cell states share commonalities in their transcriptome, thus are grouped together. For example, starvation and stationary phase were grouped together in cluster six which also makes sense from a biological perspective as they are both nutrient limited conditions.

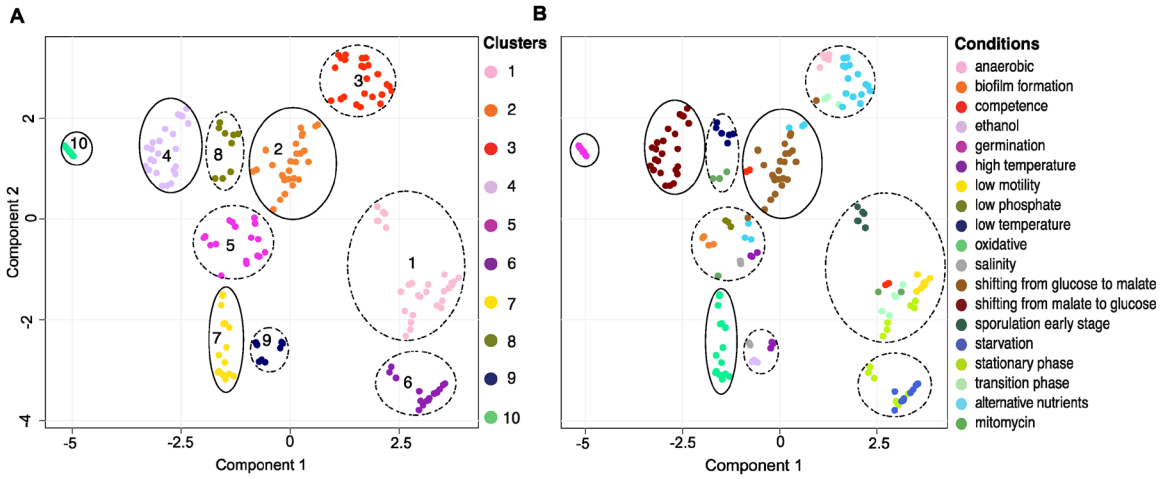

***Figure 2.3.1 Transcriptional landscape of Bacillus subtilis* 168**
Clusters identified by analysis of the BaSysBio tiling array data set of *B. subtilis* 168 from 'Computational Strategies for the Identification of a Transcriptional Biomarker Panel to Sense Cellular Growth States in Bacillus subtilis' (Huang *et al.*, 2021). Plot A shows the identified clusters grouped by colour and encircled by ellipses. Plot B shows the identified clusters encircled by ellipses and the growth conditions grouped by colour. Component 1 and Component 2 were generated by dimension reduction via UMAP.

Huang *et al.* went on to identify a minimal panel of ten biomarkers that could be used to differentiate between different cell states using a recursive feature elimination algorithm (Lazzarini and Bacardit, 2017). Later, a biomarker recommendation system was developed to generate stress sensing panels of biomarkers for different datasets and stress states (Huang *et al.*, 2023).

2.4 Synthetic biology

Synthetic biology is the name assigned to a discipline that applies the principles of engineering to biology. Typically, synthetic biology studies involve the construction of genetic components from the bottom up by combining standardised parts. The application of synthetic biology is seen in various areas including environmental

remediation, chemical synthesis, and medicine (Katz *et al.*, 2018; Tang *et al.*, 2020; Cubillos-Ruiz *et al.*, 2021).

Important concepts within synthetic biology are abstraction, modularisation, and standardisation which, when applied to genetics, allow the essential aspects of an organism's genetic code to be separated, tested in isolation, and combined with other modular elements like building blocks. Synthetic biology also leverages principles within computer science such as modelling, simulation, and machine learning (Endy, 2005; Van Brempt *et al.*, 2020).

2.4.1 Modular DNA Assembly

Modular DNA assembly is the method of combining standardised DNA parts in a specific order (Bird, Marles-Wright and Giachino, 2022). DNA assembly is a more general term given to ligating together fragments of DNA with complementary ends. Common methods of DNA assembly include Gibson assembly and BioBrick assembly (Gibson *et al.*, 2009; Røkke *et al.*, 2014). Modular assembly requires the use of standardised ends for each type of DNA part so that components can be exchanged without altering the design of the individual parts.

Standard DNA parts are one of the following types: promoter, RBS, CDS, and terminator. Modular cloning assigns specific ends to each of these types of parts allowing for the exchange and re-use of different standard parts within different constructs. There are many assembly methods and standards for modular cloning; for this work, Loop assembly was used with the Phytobricks standard (Pollak *et al.*, 2019; Cai, Carrasco Lopez and Patron, 2020).

Standard DNA parts are assembled at different levels within acceptor plasmids. Level 0 parts are a single DNA part (e.g., a promoter part) contained within an acceptor plasmid. Level 1 parts are a full transcription unit (promoter + RBS + CDS + terminator) contained within an acceptor plasmid. Level 2 parts are multiple transcription units contained within an acceptor plasmid. Each acceptor plasmid contains an antibiotic resistance gene so that the plasmid is upheld by the host.

In Loop assembly with Phytobricks standard, each part is given standard ends encoded by a letter. A promoter part contains an A 5' end and a B 3' end, a RBS part contains a B 5' end and a C 3' end, and so on (Figure 2.4.1). The acceptor

plasmid contains A and F sticky ends when digested by a specific restriction enzyme (SapI for standard Level 0 acceptor plasmids, BsaI for standard Level 1 acceptor plasmids).

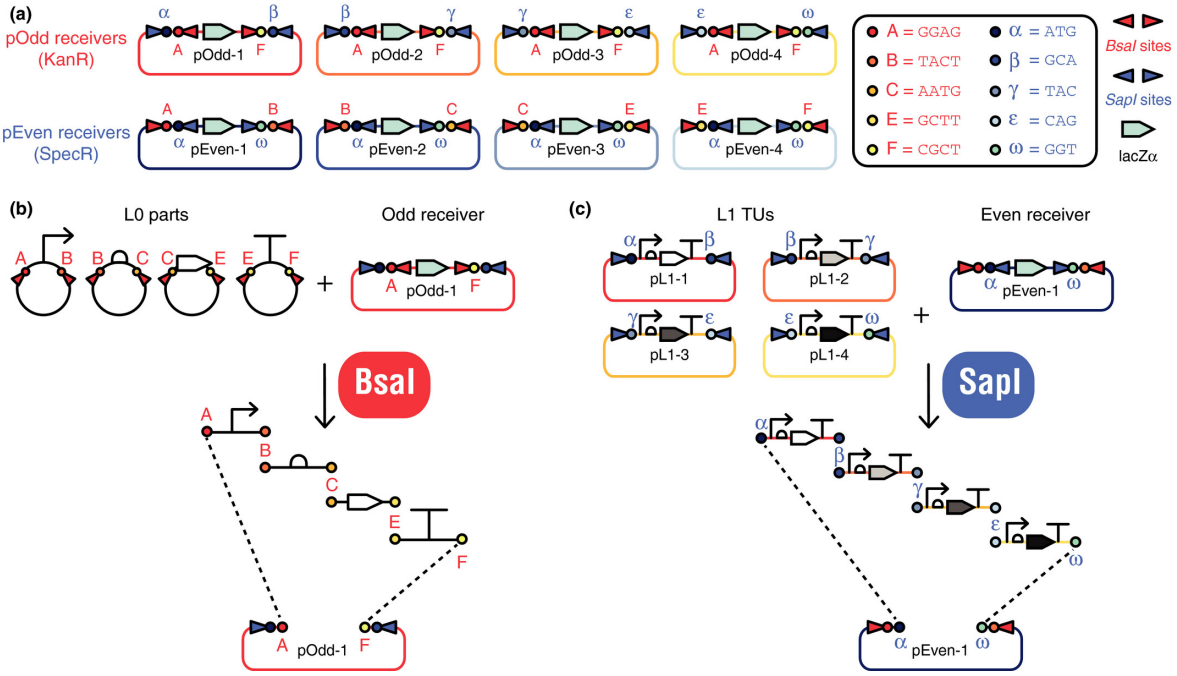

Figure 2.4.1 Diagrams describing the mechanisms of Loop assembly
Diagrams of Loop assembly mechanisms from 'Loop assembly: a simple and open system for recursive fabrication of DNA circuits' (Pollak *et al.*, 2019). Diagram a shows the standard acceptor plasmids, their restriction sites, and standard ends. Diagram b demonstrates how Level 0 parts combine to form a Level 1 part. Diagram c demonstrates how Level 1 parts combine to form a Level 2 part.

2.4.2 The Design, Build, Test, Learn Cycle

Synthetic biology has adopted the principle of the Design, Build, Test, Learn (DBTL) Cycle from engineering (Freemont, 2019). The design stage in synthetic biology relates to the conception of a genetic construct; the selection of parts for a genetic construct are usually derived from the natural genome of an organism though it may be heterologous, thus, much of the design stage involves researching existing genes and their operation within different hosts.

The build stage is the physical assembly of the genetic constructs and test stage is performing an experiment that can measure whether the construct works as expected. Screening, the validation that a construct has successfully been ported into a host, is usually part of the build stage and is simply a verification that the build

has worked. Nonetheless, screening is an essential step before continuing to the test stage.

The learn stage takes information obtained in the test stage to inform the design stage of the next iteration of the DBTL cycle. Computational methods assist with this stage such as Design of Experiments (DoE) (Singleton *et al.*, 2019). DoE is a method of multi-objective hypothesis testing where several variables are explored at the same time. DoE leverages machine learning to determine optimal values for each variable without having to exhaustively test every variable.

2.4.3 Genetic circuits

Synthetic biology allows for the construction of genetic devices analogous to components of computers (Goñi-Moreno and Nikel, 2019). Genetic constructs have been designed to act as basic switches and then combined to make logic gates (Bradley, Buck and Wang, 2016; Goñi-Moreno and Amos, 2012; Vishweshwaraiah *et al.*, 2021). A switch is possible due to the natural behaviour of promoters; some promoters are constitutive which means they are always switched on, whereas inducible promoters can be switched on and off (Gardner, Cantor and Collins, 2000; Cazier and Blazeck, 2021). Some inducible promoters can be switched on or off by the presence of a simple molecule. For example, the expression of the lac promoter (P_{lac}) is switched on by the presence of lactose or its analogue Isopropyl β- d-1-thiogalactopyranoside (IPTG) (Browning *et al.*, 2019).

To determine the state of a genetic switch, an output must be selected that can be measured. Fluorescent proteins are very common in research for achieving a measurable signal from genetic constructs (Nasu *et al.*, 2021). An example of a simple toggle switch would be P_{lac} linked to the expression of GFP; a robust RBS and terminator would be included either side of the CDS for GFP (Salis, Mirsky and Voigt, 2009; Bandiera *et al.*, 2020). For translational efficiency, the biophysical constraints of the nucleic acid sequence are important considerations for design (Reeve *et al.*, 2014; Rodnina, 2016). There must be sufficient spacing between the promoter and ribosome binding site to increase the efficiency of translation, which is often the subject of study for optimising synthetic biology designs (Lafleur, Hossain and Salis, 2022).

Combining two toggle switches in a system inherently creates a logical OR gate. If either switch is turned on, a signal is produced. However, different more complex logical operations are possible. Logical AND gates are very useful as they allow for the inclusion of two outputs simultaneously which is more robust that a single signal. To construct an AND gate, the outputs of both input switches must be combined in some way to form a signal that can be measured (Brophy and Voigt, 2014; Hicks, Bachmann and Wang, 2020).

Anderson *et al*. designed a dual input genetic AND gate for *E. coli* by utilising T7 RNA polymerase and the T7 promoter (P_{T7}) (Anderson, Voigt and Arkin, 2007). The second input of the AND gate encoded T7 RNA polymerase, however, the CDS was mutated to contain two amber stop codons (TAG) so that translation would not result in a functional polymerase. The first input encoded an amber suppressor tRNA (*supD*) which would replace a TAG codon with a Serine (Hoffman and Wilhelm, 1970). Thus, input two would only result in a functional T7 RNA polymerase if SupD is present in the system resulting in a logical AND gate.

2.5 Metabolic burden

Expression of heterologues, especially proteins, in bacteria is commonplace in research and industry particularly in fields like synthetic biology (Ingram *et al.*, 2010; Liu *et al.*, 2022). Overexertion of bacterial metabolism elicits a stress response, named metabolic burden, which is more commonly seen in engineered bacteria compared to the wildtype strains. Burden can be generated by various mechanisms (Figure 2.5.1), including resource competition and toxicity, which inhibit the metabolism of cells and lead to decreased cell growth and product yields (Boo, Ellis and Stan, 2019).

Resources within a cell are shared between the genomic DNA of the cell and any additional engineered DNA. When plasmid DNA is introduced into the cell, resources are redirected to its production, particularly when under selective pressure such as antibiotic resistance. The burden generated by the redirection of shared resources can result in a stress response and have deleterious effects on cell growth (Snoeck, Guidi and De Mey, 2024).

2 Background

Toxicity is another common factor resulting from the production of heterologues in engineered cells. Although the product intended for production in the cell may not be toxic, intermediates and byproducts may be unintentionally toxic. If the toxic products are permitted to build up within the cell, it can lead to decreased cell production. Resource competition within a cell may allow for toxic products to build up within a cell if the mechanisms required to break them down are being used elsewhere. Toxic products can also act as inhibitors to vital cell processes, such as respiration, leading to cell death (Jiang *et al.*, 2020). Other modes of toxicity include metabolic interference, degradation of cell structures, and creating an unviable environment within cells such as increased pH levels by acid byproducts (Lund *et al.*, 2020; Zhang and Voigt, 2018; Boo, Ellis and Stan, 2019).

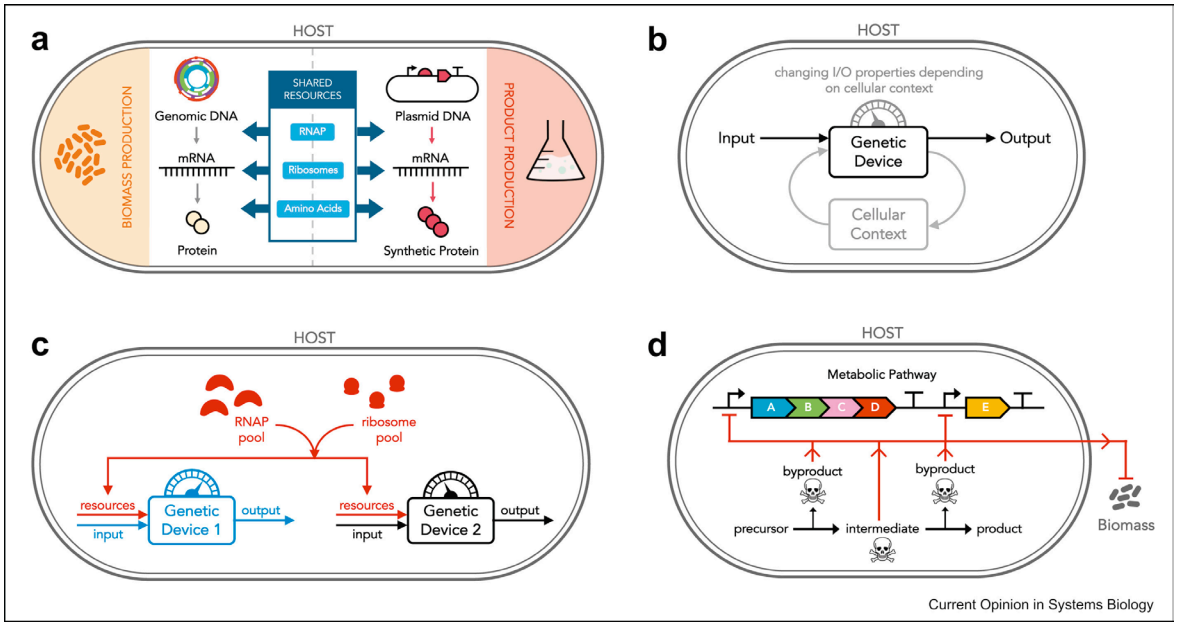

Figure 2.5.1 Diagram of mechanisms that generate metabolic burden
Four diagrams describing the intracellular mechanisms involved in metabolic burden from 'Host aware synthetic biology' (Boo, Ellis and Stan, 2019). Diagram a describes the sharing of resources between host genomic DNA and introduced plasmid DNA. Diagram b represents how the situation of a cell influences the behaviour of introduced DNA. Diagram c describes the mechanism of gene coupling. Diagram d describes how toxicity builds up in cells with a generalised metabolic pathway.

2.5.1 Codon optimisation

Codon optimisation is the process of ensuring that the codon composition of a given nucleotide sequence is tailored to the codon bias of the host organism (Bahiri-Elitzur and Tuller, 2021). There are various known codons used that encode for the same amino acid, however, each organism uses the codons with a different bias. Every organism has different compositions of tRNA molecules available; therefore, the codon composition of their genomes has adapted to match the anti-codons of their available tRNA pool (Hanson and Coller, 2018).

Ensuring that a sequence is codon optimised reduces the effect of metabolic burden by resource competition (Lipinszki *et al.*, 2018; Zhang *et al.*, 2022). Some tRNA molecules in an organism are only available in small amounts, thus the corresponding codons are known as rare codons. Lessening the frequency of rare codons in a heterologous DNA sequence is likely to increase the expression of proteins from the sequence as the tRNAs will be more available (Parvathy, Udayasuriyan and Bhadana, 2022).

Numerous algorithms have been developed for the purpose of codon optimisation (Sharp and Li, 1987; Zur and Tuller, 2014; Jain *et al.*, 2023). Simplistic codon optimisation algorithms utilise the codon frequency tables of a host organism. Codon frequency tables are determined by counting the use of every codon within all the coding sequences within a host genome; the overall codon composition of an organism is known as their codon usage bias (CUB). More advanced methods have been defined for codon optimisation which utilise larger sequences of codons within the host genome rather than relying on the individual codons (Zur and Tuller, 2014; Fox and Erill, 2010).

2.5.2 Regulatory feedback circuits to reduce metabolic burden

Systems to reduce metabolic burden have been created that utilise negative feedback to inhibit the production of the burden generating product when its production gets high enough. Examples of negative feedback circuits to reduce metabolic burden have been reviewed in ‘Host aware synthetic biology’ (Boo, Ellis and Stan, 2019).

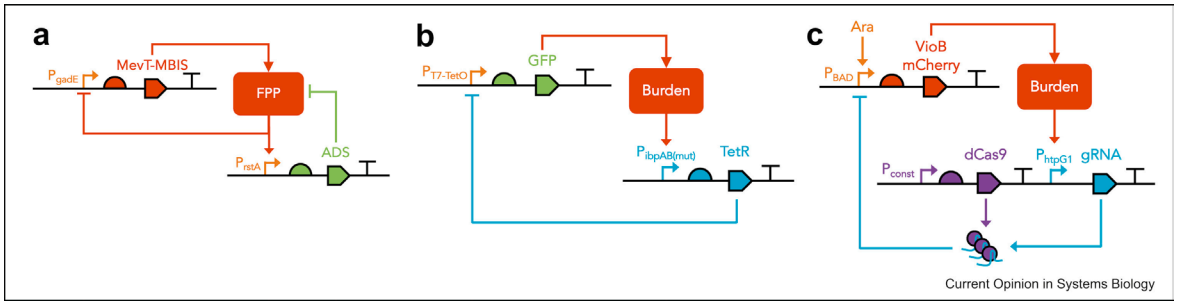

Figure 2.5.2 Burden controlled by negative feedback
Examples of three circuits designed to reduce burden in the host system, taken from Host aware synthetic biology (Boo, Ellis and Stan, 2019). Diagram a shows a system to control toxicity of the isoprenoid pathway (Dahl *et al.*, 2013). Diagram b shows a burden responsive feedback system from a global burden stress response (Dragosits, Nicklas and Tagkopoulos, 2012). Diagram c shows a specific host stress response feedback system utilising dCas9 (Ceroni *et al.*, 2018).

The first example (figure 2.5.2, a) regulates the production of the isoprenoid pathway which generates the toxic products farnesyl pyrophosphate (FPP) and HMG-CoA (Dahl *et al.*, 2013). The elegant negative feedback system utilises promoters that are regulated by FPP itself. P_{gadE} regulates the isoprenoid pathway and is inhibited in the presence of FPP, thus downregulating itself by negative feedback. Additionally, there is a second construct regulated by P_{rstA} which produces the enzyme amorphadiene synthase (ADS) which breaks down the toxic FPP product. P_{rstA} is upregulated by the presence of FPP, thus creating more ADS when there is too much FPP in the cell.

The next example (figure 2.5.2, b) addresses the stress response generated during overexpression of proteins (Dragosits, Nicklas and Tagkopoulos, 2012). The *IbpAB* operon is known to be differentially expressed under the expression of several recombinant proteins, so it has been identified as a general biomarker for burden in *E. coli*. The natural promoter for the *IbpAB* operon, P_{ibpAB}, was utilised in a second construct to produce the transcription factor, TetR. TetR inhibits the expression of the first construct which produces the recombinant protein that generated the stress response, thus completing the negative feedback loop.

The final example (figure 2.5.3, c) uses a dCas9 system to regulate the stress response generated by the specific recombinant protein produced, a fusion protein VioB-mCherry (Ceroni *et al.*, 2018). VioB-mCherry was regulated by the strong

inducible promoter, P_{BAD}, to generate burden in the cell. P_{htpG1} was determined to be upregulated specifically during the expression of the VioB-mCherry fusion protein so was selected to be the specific burden promoter for this system. P_{htpG1} was linked up to a guide RNA which targets the P_{BAD} promoter when linked up with dCas9, preventing expression from that promoter. So, when the production of VioB-mCherry triggers a stress response, according to P_{htpG1}, the continued production of VioB-mCherry will be inhibited due to being targeted by the dCas9 system.

2.6 Towards portable systems within engineering biology

Engineered systems are difficult to port from one species to another for many reasons (Brooks and Alper, 2021). A key reason being that different organisms will exhibit stress responses to different things as discussed in previous section; namely in codon optimisation where the codon bias varies between species. Additionally, different mechanisms work differently within different species such as promoters and ribosome binding sites. For example, promoters engineered to work in *E. coli* have lower performance in other species such as *B. subtilis* or simply won't work at all (Nyerges *et al.*, 2016; Ye *et al.*, 2022).

The work within this thesis forms part of a collaborative project named Synthetic Portabolomics, which aims to make systems portable between organisms with minimal refactoring requirements (Krasnogor *et al.*, 2023). The Portabolomics project crosses several disciplines from microbiology to data science all working toward the shared aim of making systems within synthetic biology more amenable to portability.

CHAPTER 3

3 Using ROTC for the selection of biomarkers that identify a given stress

3 Using ROTC for the selection of biomarkers that identify a given stress

3.1 Introduction

This chapter introduces the biomarker selection algorithm Recursive Orthogonal Threshold Classifier (ROTC)[1]. ROTC is a maximal margin classifier designed to take a matrix of transcriptomic data as input and find biomarkers, or sets of biomarkers, with an expression threshold that maximally separates treatment samples from control samples. Stress responses in prokaryotes can have complex transcriptional fingerprints and frequently share regulatory pathways and transcription factors (Buescher *et al.*, 2012; Brauer *et al.*, 2023; Njenga *et al.*, 2023); thus, making them difficult to distinguish with a singular biomarker (Huang *et al.*, 2021). ROTC was built with the intent of producing sets of biomarkers (individuals, pairs, or triples) to allow better distinguishing power (Van Der Kloet *et al.*, 2020). Additionally, ROTC was designed to find sets of biomarkers to operate as inputs for treatment-induced synthetic genetic logic gates. As such, it was intended to find biomarkers with a high expression threshold that maximally separate the intended condition from all other conditions provided.

Using data from a tiling array study on the transcriptome of *Bacillus subtilis* 168, pairs of biomarkers were generated specific to varied stress states such as growth at high temperature and growth in the presence of the antibiotic, mitomycin (Nicolas *et al.*, 2012). A new RNA-seq data set was also used in this study to contrast biomarkers generated using NGS data compared to original tiling array set. Biomarker sets generated by ROTC could be effective at distinguishing stress states with complex transcriptomic fingerprints according to the data used within this study.

3.1.1 Contributions

The ROTC algorithm was an existing algorithm developed by members of Anil Wipat's team, namely James Knight, with further development conducted by myself, Anil Wipat, and Yiming Huang. The tiling array data set used in this chapter was created by the BaSysBio consortium (Nicolas *et al.*, 2012). The RNA-seq data set was created by Anil Wipat's team: Wendy Smith conducted the experiments and prepared the samples, the sequencing was conducted by Azenta Life Sciences, the

[1] Source code for ROTC is available at https://github.com/intbio-ncl/ROTc

data were processed by myself using the nf-core/rnaseq pipeline (Ewels *et al.*, 2020).

3.1.2 Motivation

The motivation behind the development of ROTC was to implement an algorithm that selects sets of biomarkers that effectively distinguish cell states with complex transcriptional signatures. Historically, biomarker identification studies have been concerned with single biomarkers often for the identification of human diseases, especially cancer (Wehrens *et al.*, 2011; Dessì, Pascariello and Pes, 2013; Srivastava *et al.*, 2024). Studies have suggested that sets of biomarkers or biomarker panels are better for distinguishing cell states, as they allow for more combinations to attempt as the number of biomarkers increase (Van Der Kloet *et al.*, 2020). Most biomarker selection algorithms focus on the generation of single biomarkers and vary on whether they are exhaustive (O'Hara *et al.*, 2013; Dessì, Pascariello and Pes, 2013; Mandair, Reis-Filho and Ashworth, 2023). Approaches exist that generate panels of biomarkers, often integrating a machine learning approach for optimisation, but such approaches are still in their infancy and have not been widely adopted (Lazzarini and Bacardit, 2017; Wang *et al.*, 2022). ROTC uses an exhaustive search to generate sets of biomarkers, or individual biomarkers, the number of which is user-specified. Pairs of biomarkers allow for linkage to a simple dual-input genetic AND gate, therefore, biomarker pairs were the primary focus of this study.

ROTC was built with the aim to generate solutions of biomarker sets, when trained on the input data provided, that maximally separate a selected condition from all other conditions in the input data. Maximising the margins of the solutions ensures that the difference in expression of the chosen biomarkers is maximised to result in greater output from a genetic expression device such as a logic gate. As such, ROTC was tailored to generate biomarkers that can be linked up to inputs in a genetic logic gate.

Biological inference is required to convert biomarker genes into inputs of a logic gate. The regulatory mechanisms that underpin expression of a gene can be very complex involving concepts such as transcription factors, regulons, and operons (see chapter 2). The ideal scenario for a model produced by ROTC, is that the

biomarker genes are controlled by a simple natural promoter which can be used in an input device of a genetic logic gate. However, the regulation of a biomarker gene may not always be so simple, and therefore it is important that ROTC provide a list of potential solutions within which there should be a variety of options with different regulatory complexity.

While ROTC's primary use is for generating pairs of biomarkers, it is possible to generate individual biomarkers and triplets of biomarkers. Theoretically, ROTC could be applied to make sets of biomarkers of size *n*. However, in practice, this is computationally non-trivial as ROTC performs an exhaustive search on combinations of biomarkers; large amounts of computer memory are required to achieve a result where $n > 2$. For the reasons discussed above, it was decided to focus this study on searching for pairs of biomarkers to distinguish between specific stress states in *Bacillus subtilis* 168.

In theory, ROTC can be applied to any binary classification problem. ROTC is data agnostic in principle as the source of the data is inconsequential, which opens up the possibility for ROTC to be used in studies outside of transcriptional data. However, in its current state, the data must be sorted into a two-dimensional matrix with an accompanying one-dimensional binary factor used for classification of positive and negative samples.

3.2 Data Preparation

3.2.1 Tiling array data set

In 2012, the BaSysBio project produced an extensive set of tiling array data to cover the transcriptome of *Bacillus subtilis* 168 under different treatments (Nicolas *et al.*, 2012). The tiling array data covers 109 different treatments that were selected to represent environmental conditions that *B. subtilis* would undergo in its natural environment of the soil, as well as some common conditions that it may be subjected to in normal laboratory or industrial growth conditions. Treatments included different growth media such as LB, SMM, and M9, as well as supplementation of various nutrients such as fructose or glucose. Different growth phases were selected including exponential, transient, stationary, confluent, swarming, sporulation, and germination. Physical and chemical stresses were applied as treatments such as addition of mitomycin (an antibiotic), ethanol, diamide, or the application of high or low temperatures. A full list of the conditions including the experimental methods were included as supplementary information in the original publication of this data and are discussed in brief in section 3.4 of this chapter (Nicolas *et al.*, 2012). The data are available as a matrix of expression values in $\log_2$ normalised form.

To prepare the data into a form amenable to ROTC, the conditions were grouped and separated appropriately. Many experiments in the tiling array data set include an explicit control. For example, with diamide as a treatment, there were control samples without the addition of diamide. For this study, the explicit controls were separated from the treatment samples by the addition of a label; where appropriate, the tag of ‘Ctl’ was added to a sample name to indicate an explicit control for a treatment experiment and the addition of ‘Trt’ was added to the samples that included the treatment. In addition, many experiments included samples undergoing the same treatment but after different quantities of time (e.g., five minutes post induction by diamide). It was decided to group these samples together depending on the stress applied; even though the transcriptome does change with time, it was intended to attain biomarkers for a given treatment rather than a treatment at a given time interval.

The tiling array data set is very extensive both in the number of conditions tested, but also in the coverage of the genetic features in *Bacillus subtilis* 168. Not only

were known genes included, but also small regulatory features like asRNAs and regions that were defined as non-coding at the time. It was decided to include only the named genes, for the sake of computational efficiency and for ease of determining the natural promoter of the selected biomarkers.

ROTC requires a binary factor to label the position of the stress samples that need to be classified against all other samples, for ease of reference this factor will be referred to as the stress factor. The stress factor is generated after the treatment of interest is selected. Each sample in samples is iterated over and a 1 is put in the stress factor if it has been tagged as ‘Trt’ and the experiment name corresponds to the treatment of interest, all other values will have a 0 in the corresponding index of the stress factor. ROTC uses the stress factor to separate positive samples (those with a 1) from negative samples (those with a 0). See figure 3.2.1 for pseudocode describing the labelling of positive and negative samples.

```
# Pseudo code for creating labels vector
labels = [0] * length(data)

# Define user-defined 'treatment' as a string variable
treatment = "HiTm"

# Define the specific substring to search the sample name for
substring = concatenate(treatment, "_Trt")

# Iterate over each index i (1 to the length of samples)
for i in range from 1 to length(samples):
    # Check if the current sample name contains the substring
    if samples[i] contains substring:
        # Set the corresponding label to 1
        labels[i] = 1
```

Figure 3.2.1 pseudocode for labelling of samples in input data
The 'samples' vector represents a list of sample names in the imported data set, ‘data’. Each sample name follows a standard nomenclature: condition_Lbl.x, where 'condition' is the treatment or experiment, 'Lbl' is either ‘Trt’ or ‘Ctl’, and 'x' is the replicate number (an integer). “HiTm” is used as an example of a grouped treatment name.

3.2.2 RNA-seq data set

The tiling array data from the BaSysBio project is extensive and has been a widely used resource for transcriptomics and for those working with *Bacillus subtilis* as a model organism (Zhu and Stülke, 2018). However, tiling array data is no longer used widely and has mostly been replaced by data generated from NGS studies which, in the case of transcriptomic studies, is commonly RNA-seq (Negi *et al.*, 2022).

As part of this study and the wider Portabolomics project, an RNA-seq study was undertaken to supplement the tiling array data set generated in 2012. The experiments to exert stress conditions upon *Bacillus subtilis* 168 were conducted by Wendy Smith, replicating the methods outlined by the BaSysBio project (Nicolas *et al.*, 2012). The RNA samples were sent to Azenta Life Sciences[2] for sequencing, yielding sets of paired-end reads for each sample.

The RNA-seq study was on a smaller scale than the BaSysBio project, including fewer conditions. A minimal set of conditions was determined that most widely covers the transcriptional landscape uncovered in the original 2012 study. The subset of conditions were chosen based on the clusters identified by Huang *et al.*, to determine whether a single treatment from each cluster provides the diversity required to select biomarkers from a reduced data set (Huang *et al.*, 2021). The subset of conditions selected were as follows: anaerobic, competence, diamide, ethanol, nutrient shift from glucose to malate, nutrient shift from malate to glucose, low temperature, high temperature, and stationary phase in SMM.[3]

The RNA-seq data were processed using the nf-core/rnaseq pipeline (Ewels *et al.*, 2020), see methods 7.7. The nf-core/rnaseq pipeline involves pre-processing, alignment, and post-processing steps that result in the final read count tables per transcript which were used for downstream analysis. Additionally, the pipeline includes many quality control steps and reporting; some reports as summarised in the MultiQC report are shown in figures 3.2.1 and 3.2.3 (Ewels *et al.*, 2016). The transcripts per million (TPM) values were normalised with RUVg (remove unwanted variation) analysis using ERCC spike-in controls as a standard control (Risso *et al.*,

[2] Azenta, Inc. Corporate Headquarters, 200 Summit Drive, Burlington, MA 01803 USA
[3] Data are available on GEO (Gene Expression Omnibus) under accession number: GSE226559.

2014). Following this normalisation, the values were $\log_2$ transformed to form the basis of the new RNA-seq data set which is effectively a large matrix of feature count data per sample.

As before with the tiling array data, a stress factor for each condition was generated using the same method. ‘Ctl’ and ‘Trt’ labels were applied to the samples for each experiment appropriately. The stress factor will then assign ones and zeroes depending on the selected treatment of interest.

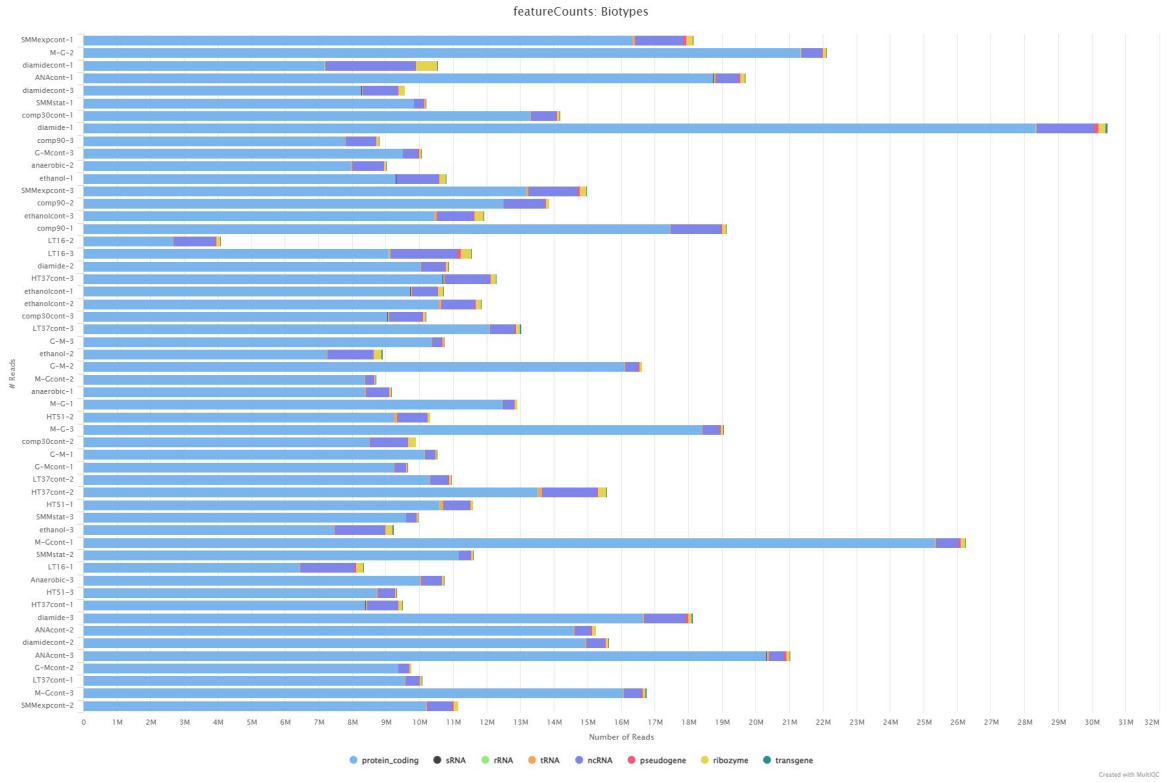

Figure 3.2.2 featureCounts Biotypes Plot
A plot generated in MultiQC (Ewels *et al.*, 2016) showing reads mapped to features of different biotypes as identified by featureCounts (Liao, Smyth and Shi, 2014). Each sample in the RNA-seq data set is represented as a separate bar on the plot.

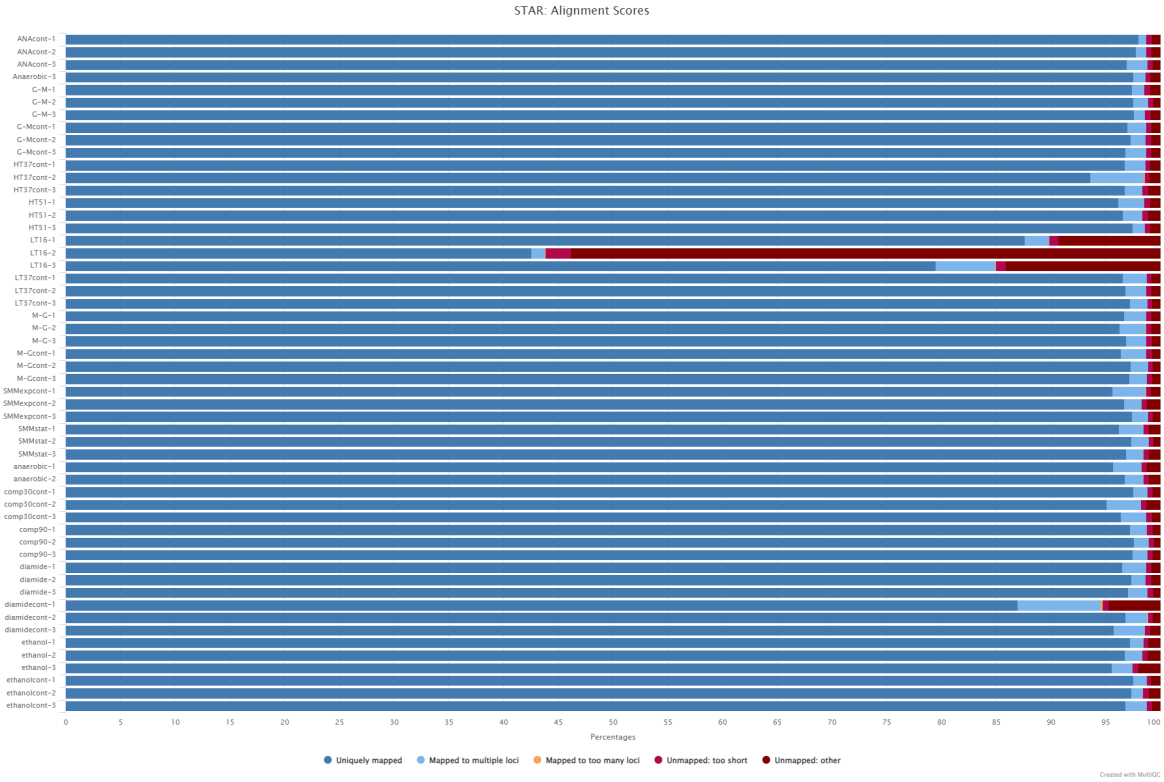

Figure 3.2.3 STAR Alignment Scores Plot
A plot generated in MultiQC (Ewels *et al.*, 2016) showing the percentage of reads for each sample that were correctly mapped to loci in the reference genome EB2 by STAR align (Dobin *et al.*, 2013). Each sample in the RNA-seq data set is represented as a separate bar on the plot.

3.3 Model Description

ROTC produces a model consisting of a set of biomarkers whose expression values under the stress being interrogated is maximally separated from the expression values in all other conditions. The data are split into positive samples (treatment) and negative samples (control) which the optimal model will maximally and completely separate. The positive and negative samples are separated by a boundary in the dimension of each feature in the model and a threshold is calculated as the mid-point between these boundaries. The threshold is then used as the classifier in the model for any new samples.

Each feature adds a dimension to this model space. Theoretically, this could be achieved for any number of dimensions but for the purpose of biomarker selection this has been done for one, two and three dimensions. Graphical representations of the biomarker selections models have been plotted for one, two, and three dimensions in figures 3.3.1, 3.3.2, and 3.3.3, respectively. Figures 3.3.1, 3.3.2, and 3.3.3 all use actual results from the top ranked biomarker solutions, generated by ROTC with the tiling array data set as input data, using growth at high temperature as the stress conditions being inspected.

ROTC calculates the upper bound, U, that separates the positive samples, S^+, as the minimum points in each dimension, d:

$$U_d = min(s_d^+)$$

If any negative sample is greater than the upper bound in every dimension, then it lies within the boundary of the positive samples and can't be separated i.e., it is misclassified. If a model contains misclassified samples, it is considered inseparable and is discarded.

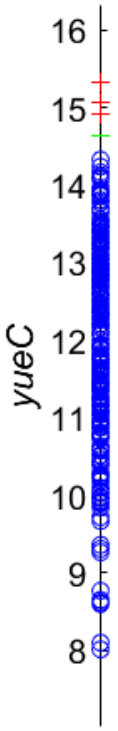

Figure 3.3.1 One-Dimensional Representation of a ROTC generated Model
Linear scatter plot of the top ranked individual biomarker solution for growth at high temperature as treatment. Plotted using data from the tiling array data set, values are $\log_2$ expression values of each sample with respect to the biomarker, *yueC*. Positive values (samples labelled as HiTm - growth at high temperature) are represented as red crosses; negative values (all other samples) are represented as blue circles. The threshold is represented as a green line mark on the axis. The upper and lower boundaries are not plotted but can be determined as the maximum negative value (lower bound) and the minimum positive value (upper bound) in one dimension.

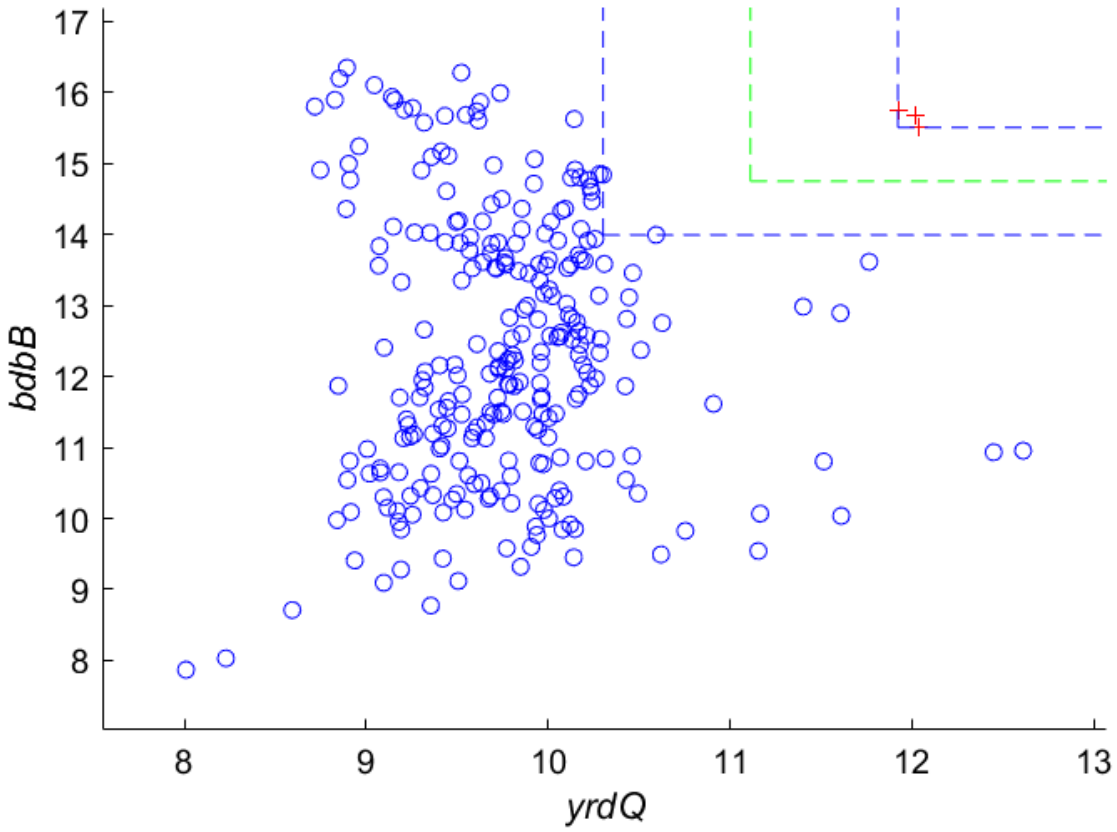

Figure 3.3.2 Two-Dimensional Representation of a ROTC generated Model
2D scatter plot of the top ranked biomarker pair solution for growth at high temperature as treatment. Plotted using data from the tiling array data set, values are $\log_2$ expression values for each sample, with co-ordinates in respect of the biomarker genes, *bdbB* and *yrdQ*, in each dimension. Positive values (samples labelled as HiTm - growth at high temperature) are represented as red crosses; negative values (all other samples) are represented as blue circles. The upper and lower bounds are represented as blue dashed lines, and the thresholds are represented as green, dashed lines.

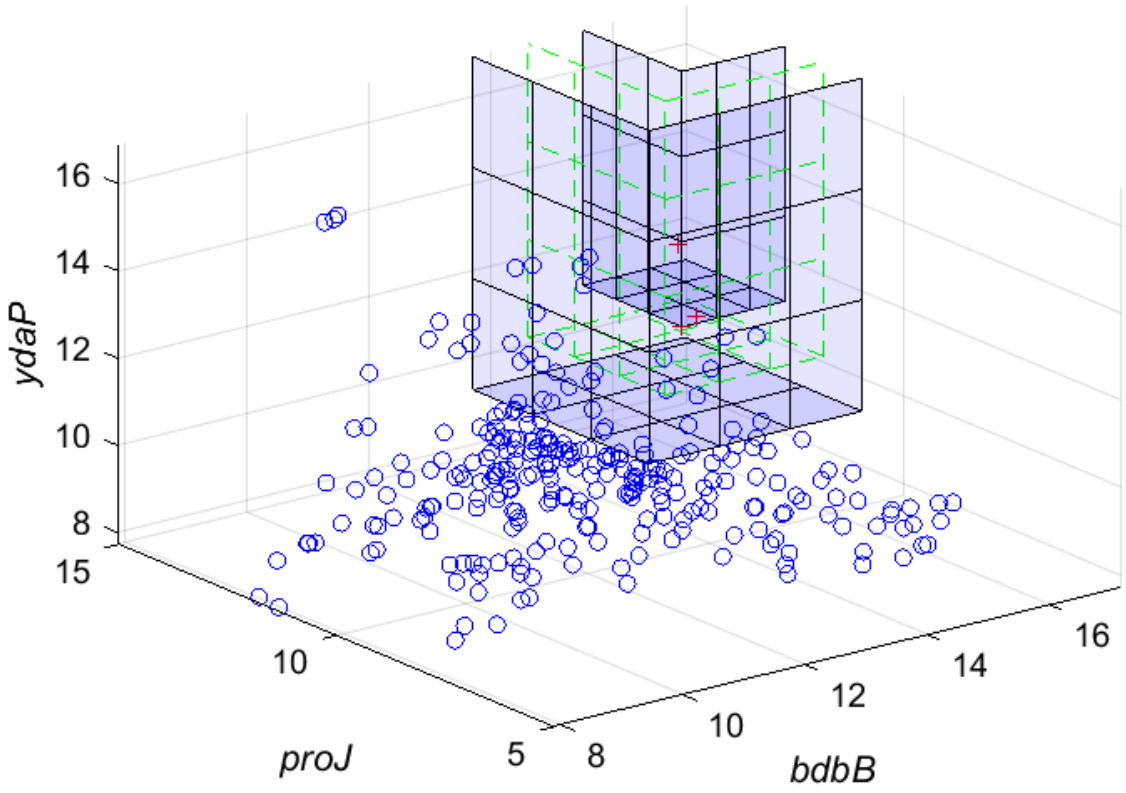

Figure 3.3.3 Three-Dimensional Representation of a ROTC generated Model
3D scatter plot of the top ranked biomarker triplet solution for growth at high temperature as treatment. Plotted using data from the tiling array data set, values are $\log_2$ expression values of each sample, with co-ordinates in respect to each biomarker gene (*bdbB*, *proJ*, *ydaP*). Positive values (samples labelled as HiTm - growth at high temperature) are represented as red crosses; negative values (all other samples) are represented as blue circles. The upper and lower bounds are represented as blue (surf) planes, and the thresholds are represented as green, dashed (mesh) planes.

To calculate the lower bound, L, that separates the negative samples, S^-, the distance between the upper bound and each negative point must be calculated i.e., the margin, M. The margin is calculated for every negative sample and in each dimension:

$$M_d^s = U_d - s_d$$

The lower bound for each dimension is then determined by the negative point closest to the upper bound where its margin is greater than the margin for the same data point in every other dimension where D is the total number of dimensions:

$$L_d = max(s_d^-) \; where \bigwedge_{i=D+1}^{i=0} M_d^s \geq M_{D-i}^s$$

The threshold, T, can then be calculated as the mid-point between the upper bound and the lower bound for each dimension:

$$T_d = \frac{U_d + L_d}{2}$$

Figure 3.3.4 contains an annotated example to illustrate how the lower bounds are determined. The points that define the lower bounds (negative point closest to the upper bound where its margin is greater than the margin for the same data point in every other dimension) have been encircled and their distances from the upper bound in each dimension have been labelled by arrows to highlight why these are the bound defining points. A diagonal line has also been plotted where the distances in both dimensions are equal; this diagonal separates the points so that those above the diagonal are where the distance in the dimension of Gene1 is greater than the distance in the dimension of Gene2, and the reverse is true for points below the diagonal. The point with the greatest value in its dimension above or below the diagonal will be the defining point for the lower bound of that dimension.

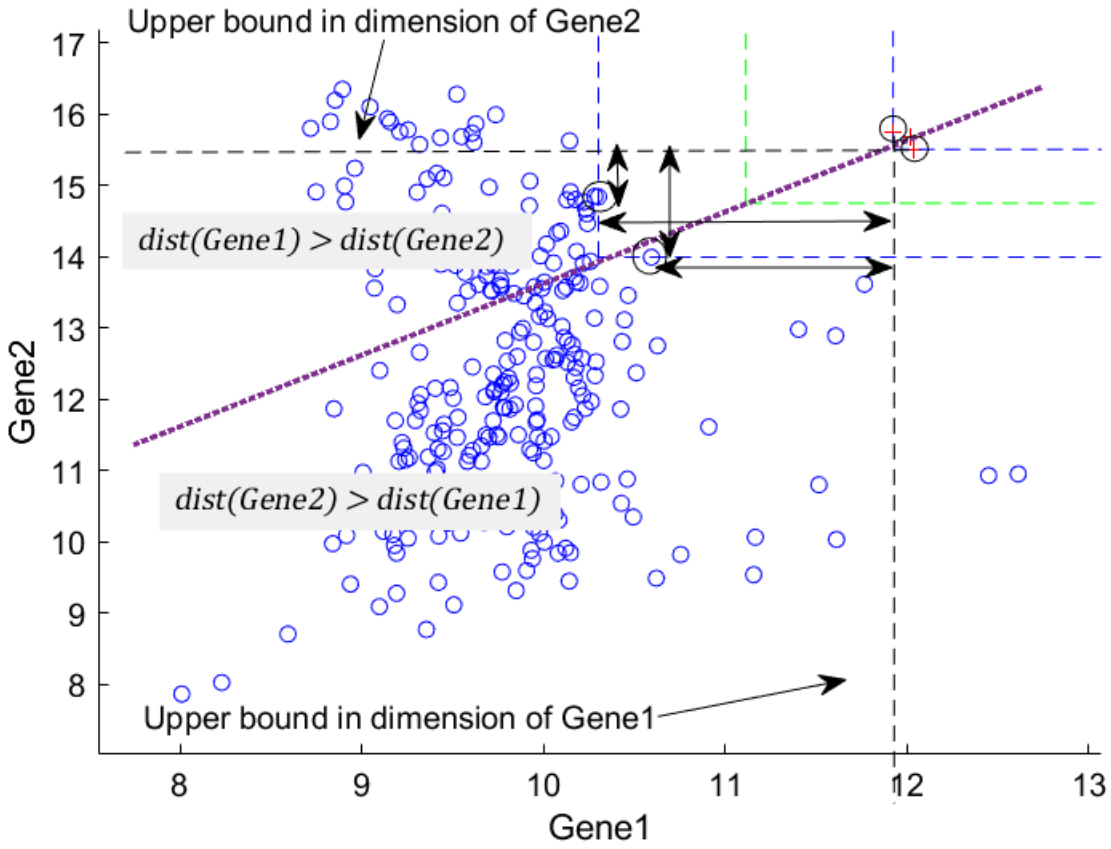

Figure 3.3.4 Annotated plot to illustrate how bounds are determined
Annotated 2D scatter plot of a biomarker gene pair solution. The diagonal, purple, dotted line represents the boundary where all points above the line in the dimension of Gene1 (x-axis) are where *dist(Gene2) > dist(Gene1)* and vice versa for points above the line in the dimension of Gene2 (y-axis). The upper and lower bounds are represented by blue dashed lines; the upper bounds are extended on this diagram so that the distance between the defining points (circled) is more apparent. The thresholds are represented by dashed green lines; positive values are represented as red crosses; negative values are represented as blue circles. The distances between boundary defining points (circled) are marked by double arrows.

In summary, the margin is the distance between the upper and lower bounds of the separated samples. The upper bounds are related to the minimum values of the positive samples; the lower bounds are related to the maximum values of the negative samples, but these will vary depending on the dimension of the lower bound. The threshold is the mid-point between the upper and lower bounds and is effectively the classifier component of the model. The margins are maximised during the exhaustive search function of the algorithm. The best solution is the model with the greatest margins with no misclassified samples, i.e., the solution must not be inseparable.

3.3.1 Results output of ROTC

ROTC returns all solutions, ranked in descending order of their margins, where a solution is a model (set of biomarkers) that fully separates negative samples from positive samples. Those with the largest margins have the greatest separation between positive and negative samples, and therefore are chosen as the optimal solutions. Inversely, those with the lowest margins have the least separation between positive and negative samples, and therefore are the least fit to operate as a binary classifier. Therefore, when the solutions are compared, the one with the lowest margin is ranked lower. The top ranked solution may not always be fit for purpose, so ROTC returns a list of viable solutions. A biomarker solution may not be viable for a number of reasons; for example, the natural regulation of the gene may have a sequence which is too complex for synthesis or may interfere with the metabolic pathway being introduced to the cell.

The top solution(s) will frequently be referred to throughout this chapter which refers to the top x solutions according to their rank in descending order of the lowest margin. For ease of reference, the term “minimum margin score” (MMS) has been defined as the lowest margin within a solution. MMS was used to determine how effective a biomarker is at separating positive and negative samples as it is the separation distance between these bounds in the dimension with the least separation.

It is worth noting that the MMS is only comparable within the dataset that it was taken from. In this study, those data sets are the tiling array data set and RNA-seq data set for *Bacillus subtilis* 168. Both data sets consist of values in $\log_2$ expression,

however, they are generated from differing technologies and therefore are not directly comparable. It is important to keep in mind that the value of a margin (and therefore, MMS) is on a $\log_2$ scale. For example, a difference of 1.5 which may be generated from "$15 - 13.5$" is large in terms of raw gene counts and is also different from a difference of 1.5 that is generated from "$8.5 - 6$". Therefore, it must be considered that MMS is not an absolute indicator of how good a solution is, rather that it is a convenient approach for ranking solutions.

3.4 Biomarker selection using tiling array data

The results presented in this section were generated using the tiling array data set, prepared as outlined in section 3.2.1. Each stress condition was selected as the treatment of interest, one at a time, and fed into ROTC to generate biomarkers. For reference, a summary of each stress condition is outlined here. A full summary of methods can be found in the supplementary materials of the original publication (Nicolas *et al.*, 2012). Please note that the shorthand abbreviations for each condition are, to some extent, an artefact of the sample names extracted from the original data set which has been maintained for ease of reference to the source of that data.

Addition of chemical stresses: diamide, hydrogen peroxide, paraquat, mitomycin, ethanol and salt (shorthand are diamide, H2O2, paraquat, mitomycin, Etha, and salt, respectively). Anaerobic growth conditions: aerobic growth was the implicit control for these experiments, anaerobic growth was measured with the addition of potassium nitrate and without for fermentative growth (shorthand: aero, nit, and ferm, respectively). Temperature induced stress: growth at low temperature, growth at high temperature, heat shock, and cold shock (shorthand: LoTm, HiTm, Heat, and Cold, respectively). Specific growth conditions were induced to represent: competence (the ability to uptake exogenous DNA), confluent growth, swarming, sporulation, germination (late sporulation), and biofilm production (shorthand: competence, confluent, swarming, sporulation, germination, and biofilm, respectively). Growth phases were observed in various media: exponential, stationary, and transient growth in LB, LB with addition of glucose, and M9 media (shorthand: LBexp, LBstat, LBtran, Gluexp, Glustat, Glutran, M9exp, M9stat, M9tran). Growth in minimal media: BMM, SMM and SMM in stationary phase

(shorthand: BMM, SMM, SMMPr). Growth in high phosphate media, low phosphate media, and low phosphate media post stationary phase (shorthand: HPh, LPh, LPhT). Growth with the supplementation of the following nutrients: pyruvate, glucose, fructose, gluconate, malate + glucose, malate, glycerol, and glucamate + succinate (shorthand: Pyr, Glu, Fru, Glucon, MG, Mal, Gly, GS). Nutrient shift from malate to glucose and glucose to malate. Finally, growth with a lack of nutrients to simulate starvation. ROTC was unable to find solutions for sporulation, glucogen + succinate, and for nutrient shift (SH: sporulation, GS, M2G, G2M).

3.4.1 Biomarker selection for diamide induced stress

All the grouped conditions were run through ROTC as a stress condition to generate biomarkers. In this section, diamide was selected as an example to demonstrate the kind of results that ROTC generates.

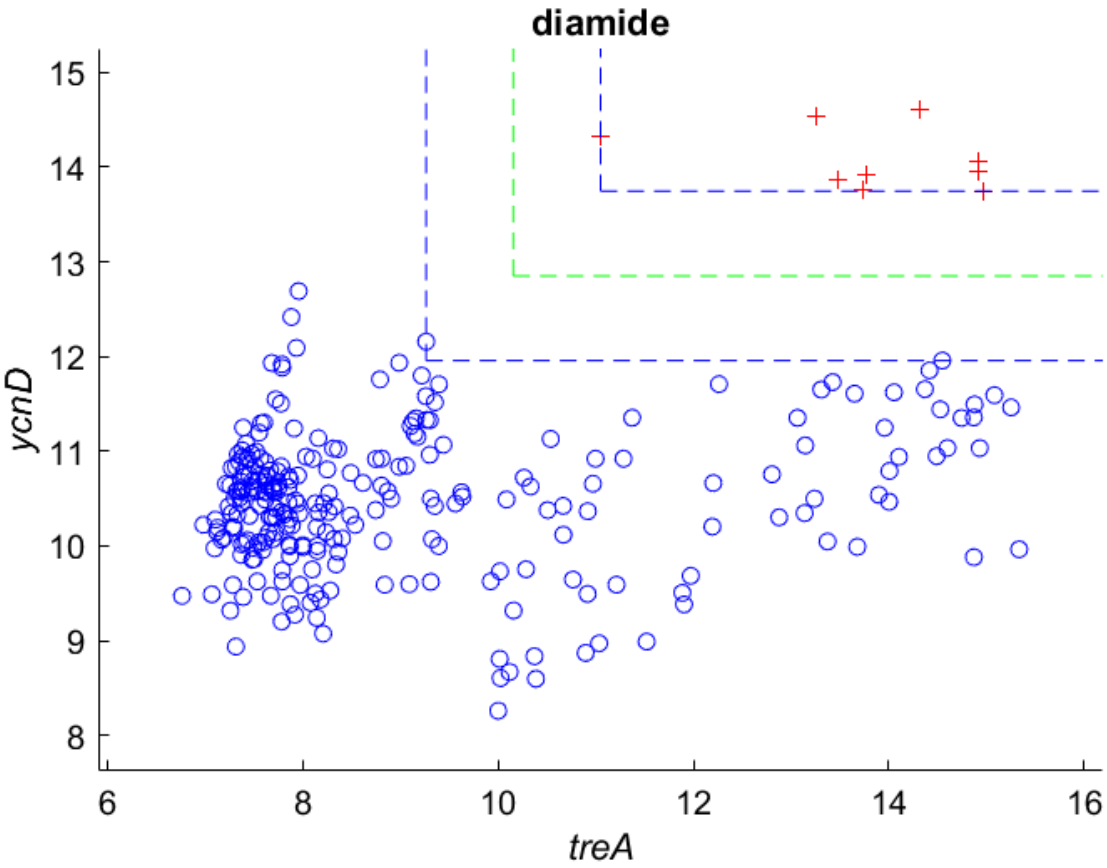

Figure 3.4.1 top biomarker pairs solution for diamide induced stress
Scatter plot for the top ranked biomarker pair solution for diamide induced stress. The x-axis shows the $\log_2$ expression values for Gene1 (treA), and the y-axis shows the $\log_2$ expression values for Gene2 (*ycnD*). Positive samples (those labelled as diamide) are represented as red crosses; negative samples are represented as blue circles. Blue dashed lines represent the upper and lower bounds; green dashed lines represent the thresholds. The title of the plot contains the shorthand for the condition selected as a stress treatment for ROTC.

Gene1	Gene2	Margin1	Margin2	Threshold1	Threshold2
treA	*ycnD*	1.787	1.788	10.153	12.852
ywnA	*yqiG*	1.752	1.773	13.006	12.274
fabHB	*treA*	1.773	1.739	12.506	10.177
ywnA	*yvrD*	1.752	1.735	13.006	13.451
ywnA	*yceJ*	1.752	1.727	13.006	8.512
mccA	*fabHB*	1.724	1.777	11.581	12.504
ywnA	*yrbC*	1.746	1.722	13.009	11.026
ywnA	*ywnF*	1.752	1.717	13.006	12.269
ywnA	*mccB*	1.686	1.780	13.039	11.737
arsB	*ydzF*	1.681	1.785	13.076	10.573
yugJ	*arsB*	1.760	1.681	13.854	13.076
arsB	*treR*	1.681	1.757	13.076	12.055

Table 3.4.1 top twelve biomarker pairs generated for diamide induced stress
The margin and threshold values are rounded to three decimal places. Margin1 and Margin2 correspond to the margins for Gene1 and Gene2, respectively. Threshold1 and Threshold2 correspond to the thresholds for Gene1 and Gene2, respectively.

Often in the top solutions for the same condition, the same biomarker gene often appears multiple times in different pairs which is expected as there is no filtering to prevent the same biomarker gene from appearing more than once. However, it is useful for analysis to examine the range of different solutions that use unique genes i.e., where no single biomarker gene appears more than once. This filtering has been conducted downstream and the results for pairs of unique biomarkers is plotted in figure 3.4.2 and the results are tabulated in table 3.4.2.

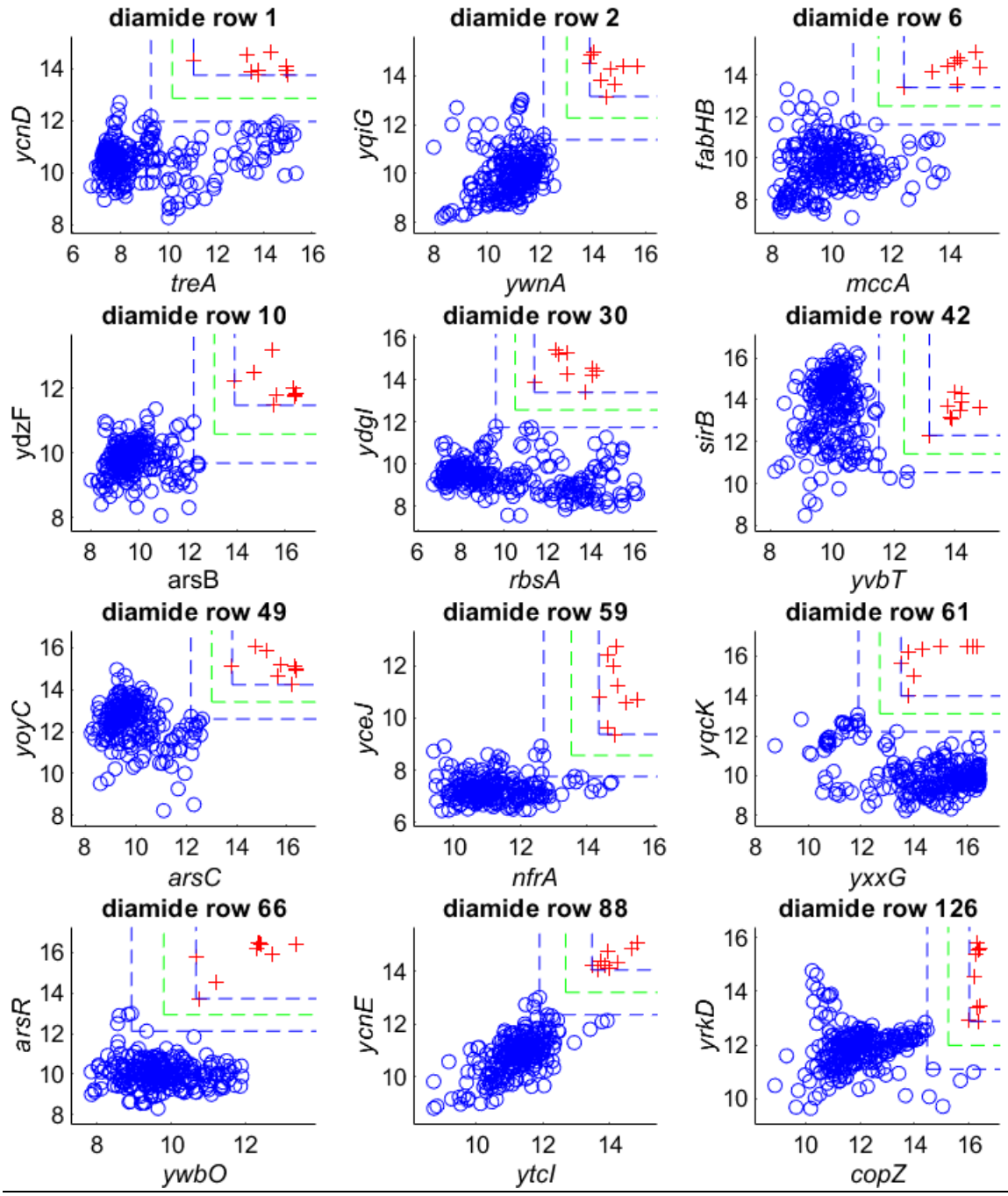

Figure 3.4.2 top twelve solutions of unique biomarker pairs for diamide induced stress

Unique solutions are where the same biomarker gene may only appear once in a list of solutions. The row number represents the solution's original position in the solution list output by ROTC as ranked in descending order of MMS. Positive samples are represented as red crosses; negative samples are represented as blue circles. Blue dashed lines represent the upper and lower bounds; green dashed lines represent the thresholds. The title of each subplot contains the shorthand name of the stress condition, followed by the row number of the solution plotted.

Row	Gene1	Gene2	Margin1	Margin2	Threshold1	Threshold2
1	*treA*	*ycnD*	1.787	1.788	10.153	12.852
2	*ywnA*	*yqiG*	1.752	1.773	13.006	12.274
6	*mccA*	*fabHB*	1.724	1.777	11.581	12.504
10	*arsB*	*ydzF*	1.681	1.785	13.076	10.573
30	*rbsA*	*ydgI*	1.791	1.655	10.499	12.574
42	*yvbT*	*sirB*	1.639	1.766	12.338	11.418
49	*arsC*	*yoyC*	1.633	1.635	13.014	13.414
59	*nfrA*	*yceJ*	1.650	1.614	13.523	8.569
61	*yxxG*	*yqcK*	1.609	1.791	12.701	13.105
66	*ywbO*	*arsR*	1.749	1.602	9.798	12.931
88	*ytcI*	*ycnE*	1.576	1.698	12.690	13.196
126	*copZ*	*yrkD*	1.539	1.784	15.239	11.992

Table 3.4.2 top twelve solutions of unique biomarker pairs for diamide induced stress

Margin and threshold values are rounded to three decimal places. Row represents the solution's original position in the solution list output by ROTC as ranked in descending order of minimum margin score. Margin1 and Margin2 correspond to the margins for Gene1 and Gene2, respectively. Threshold1 and Threshold2 correspond to the thresholds for Gene1 and Gene2, respectively.

For diamide as the selected treatment condition, the top solutions all have a high MMS relative to other conditions of the same data set. Despite the row numbers (which are representative of rank by MMS) increasing by an order of magnitude in table 3.4.2, the MMS is still large relative to other results from different conditions within the same data set. The results listed in table 3.4.2 indicate that ROTC is capable of generating hundreds of good solutions, which may be useful in certain use cases if there were many restraints that limited the number of genes that could be utilised in a solution.

3.4.2 Biomarker selection for all treatments in tiling array data set

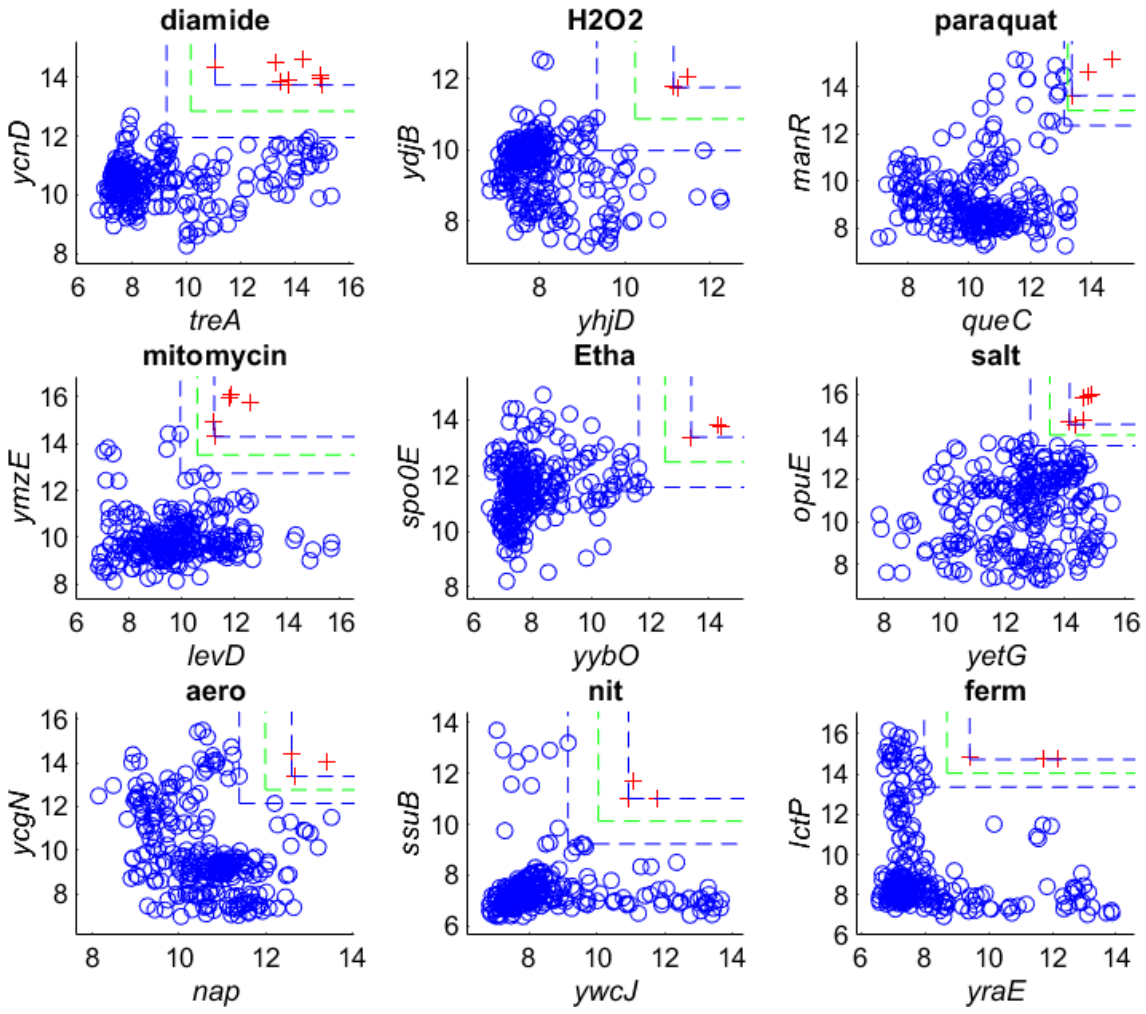

Figure 3.4.3 top biomarker pairs generated per treatment in tiling array data
Page 1 of 4
Scatter plots of the top ranked biomarker pair solution for every treatment condition in the tiling array data set. The x-axis shows the $\log_2$ expression values for Gene1 of the top ranked solution generated by ROTC. The y-axis shows the $\log_2$ expression values for Gene2 of the top ranked solution generated by ROTC. Positive samples are plotted using red crosses. Negative samples are plotted using blue circles. The blue dotted lines represent the upper and lower bounds of the solution. The green dotted lines represent the thresholds of the solution. Title of each plot denotes the shorthand of the condition that each graph plots the solution data for.

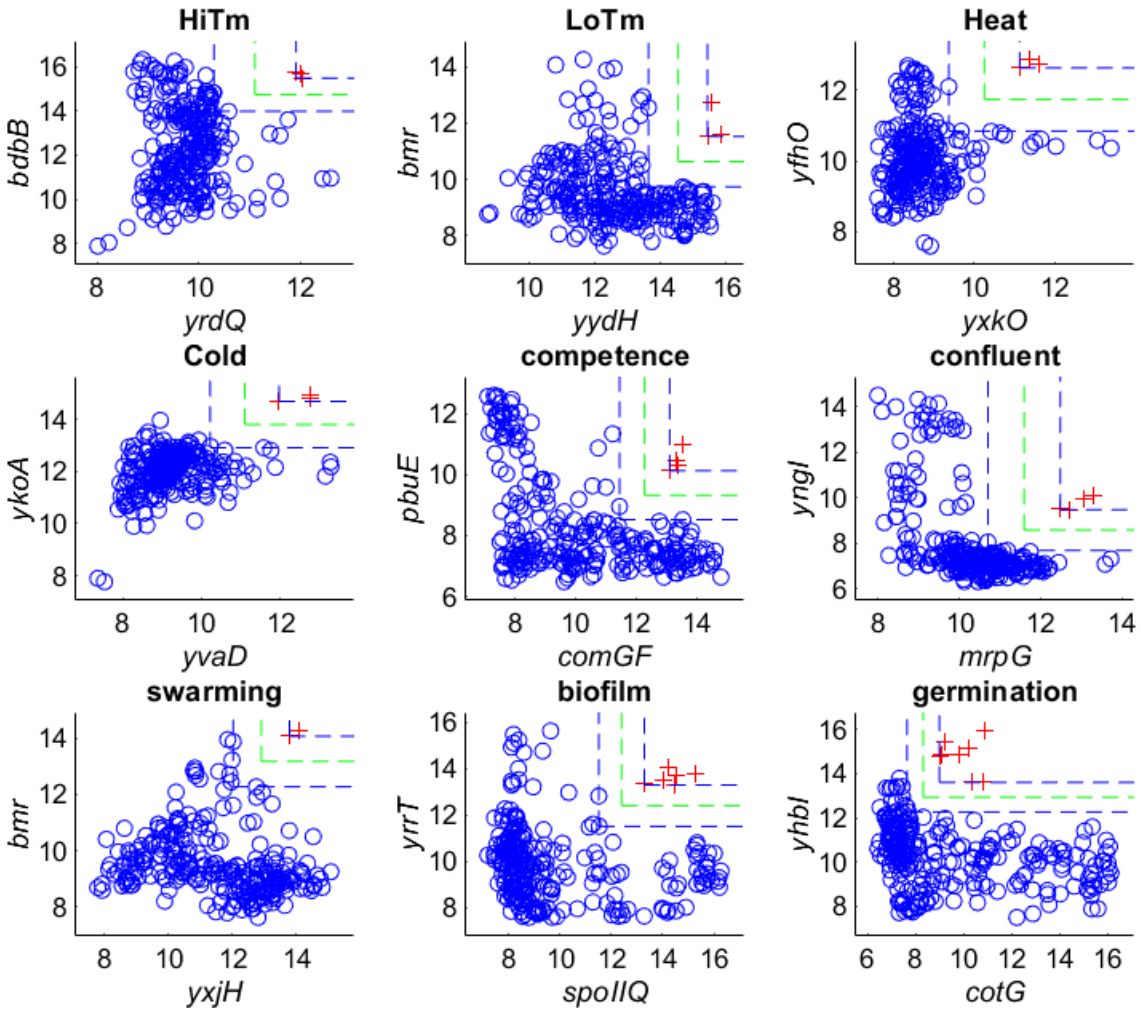

Figure 3.4.3 top biomarker pairs generated per treatment in tiling array data
Page 2 of 4
Scatter plots of the top ranked biomarker pair solution for every treatment condition in the tiling array data set. The x-axis shows the $\log_2$ expression values for Gene1 of the top ranked solution generated by ROTC. The y-axis shows the $\log_2$ expression values for Gene2 of the top ranked solution generated by ROTC. Positive samples are plotted using red crosses. Negative samples are plotted using blue circles. The blue dotted lines represent the upper and lower bounds of the solution. The green dotted lines represent the thresholds of the solution. Title of each plot denotes the shorthand of the condition that each graph plots the solution data for.

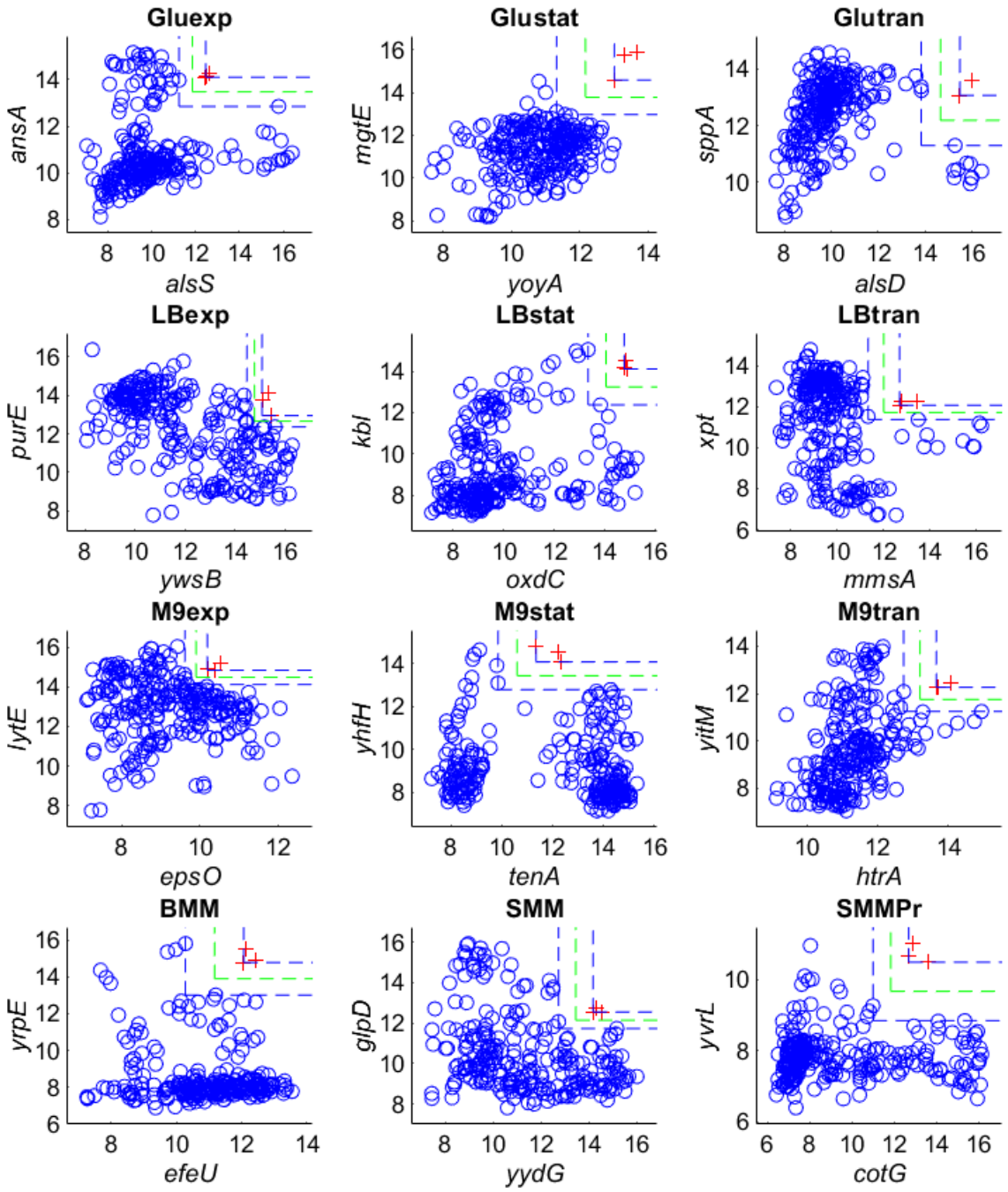

Figure 3.4.3 top biomarker pairs generated per treatment in tiling array data
Page 3 of 4
Scatter plots of the top ranked biomarker pair solution for every treatment condition in the tiling array data set. The x-axis shows the $\log_2$ expression values for Gene1 of the top ranked solution generated by ROTC. The y-axis shows the $\log_2$ expression values for Gene2 of the top ranked solution generated by ROTC. Positive samples are plotted using red crosses. Negative samples are plotted using blue circles. The blue dotted lines represent the upper and lower bounds of the solution. The green dotted lines represent the thresholds of the solution. Title of each plot denotes the shorthand of the condition that each graph plots the solution data for.

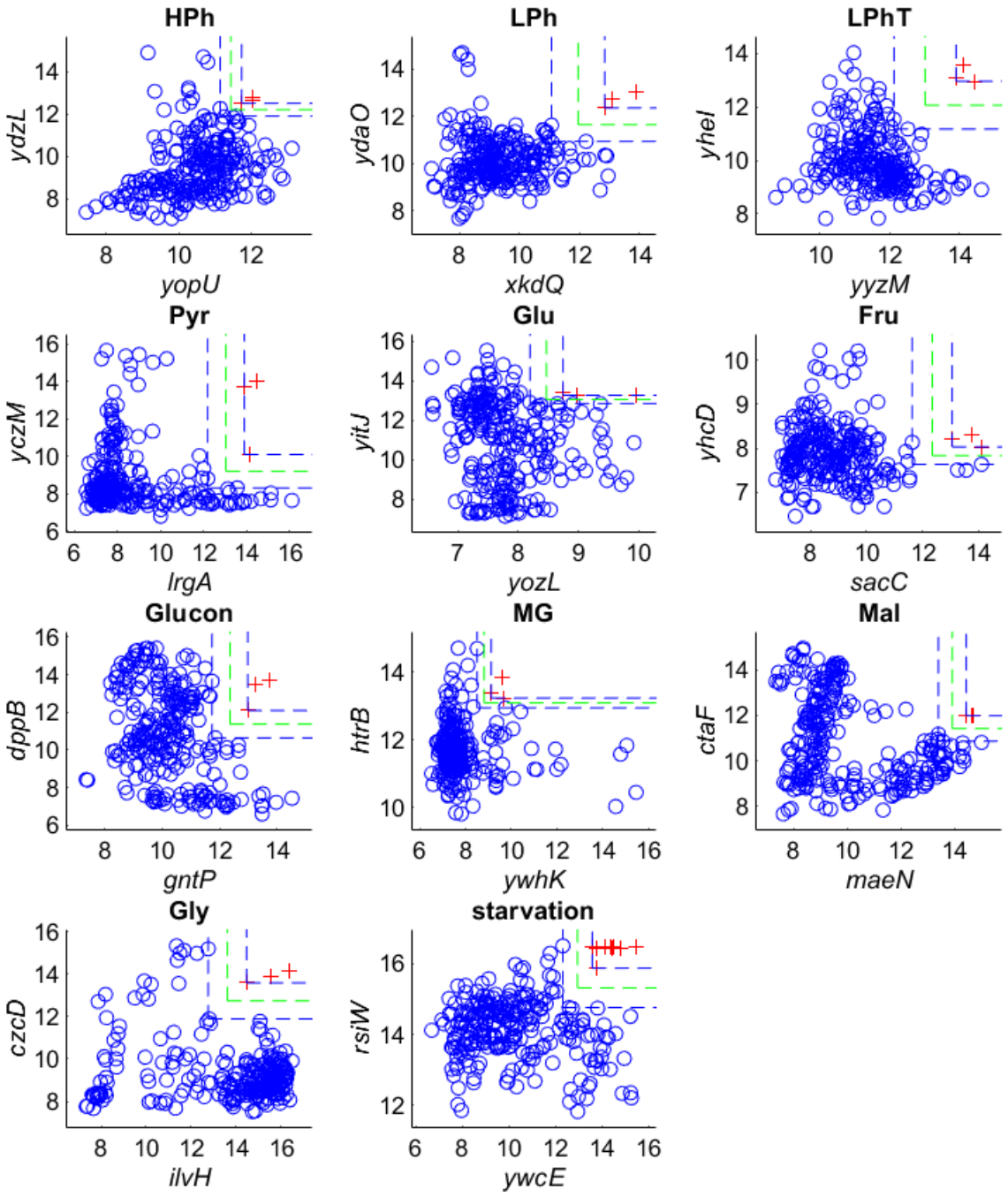

Figure 3.4.3 top biomarker pairs generated per treatment in tiling array data
Page 4 of 4
Scatter plots of the top ranked biomarker pair solution for every treatment condition in the tiling array data set. The x-axis shows the $\log_2$ expression values for Gene1 of the top ranked solution generated by ROTC. The y-axis shows the $\log_2$ expression values for Gene2 of the top ranked solution generated by ROTC. Positive samples are plotted using red crosses. Negative samples are plotted using blue circles. The blue dotted lines represent the upper and lower bounds of the solution. The green dotted lines represent the thresholds of the solution. Title of each plot denotes the shorthand of the condition that each graph plots the solution data for.

Condition	Gene1	Gene2	Margin1	Margin2	Threshold1	Threshold2
diamide	*treA*	*ycnD*	1.787	1.788	10.153	12.852
H2O2	*yhjD*	*ydjB*	1.796	1.767	10.239	10.866
paraquat	*queC*	*manR*	0.259	1.272	13.240	12.978
mitomycin	*levD*	*ymzE*	1.288	1.560	10.591	13.521
Etha	*yybO*	*spo0E*	1.796	1.794	12.520	12.473
salt	*yetG*	*opuE*	1.304	1.010	13.502	14.081
aero	*nap*	*ycgN*	1.210	1.247	11.996	12.766
nit	*ywcJ*	*ssuB*	1.793	1.779	10.040	10.120
ferm	*yraE*	*lctP*	1.431	1.363	8.688	14.038
HiTm	*yrdQ*	*bdbB*	1.622	1.510	11.113	14.747
LoTm	*yydH*	*bmr*	1.786	1.792	14.528	10.633
Heat	*yxkO*	*yfhO*	1.763	1.786	10.246	11.736
Cold	*yvaD*	*ykoA*	1.744	1.782	11.092	13.796
competence	*comGF*	*pbuE*	1.655	1.612	12.284	9.327
confluent	*mrpG*	*yngI*	1.784	1.788	11.596	8.577
swarming	*yxjH*	*bmr*	1.763	1.797	12.910	13.174
biofilm	*spoIIQ*	*yrrT*	1.774	1.791	12.397	12.408
germination	*cotG*	*yhbI*	1.356	1.340	8.318	12.935
HPh	*yopU*	*ydzL*	0.594	0.600	11.445	12.219
LPh	*xkdQ*	*ydaO*	1.776	1.420	11.959	11.634
LPhT	*yyzM*	*yheI*	1.776	1.786	13.007	12.064
Gluexp	*alsS*	*ansA*	1.215	1.248	11.858	13.477
Glustat	*yoyA*	*mgtE*	1.693	1.617	12.165	13.778
Glutran	*alsD*	*sppA*	1.641	1.773	14.654	12.184
LBexp	*ywsB*	*purE*	0.611	0.579	14.777	12.651
LBstat	*oxdC*	*kbl*	1.431	1.738	14.062	13.239
LBtran	*mmsA*	*xpt*	1.365	0.697	12.019	11.715
M9exp	*epsO*	*lytE*	0.572	0.719	9.908	14.490
M9stat	*tenA*	*yhfH*	1.499	1.285	10.608	13.446
M9tran	*htrA*	*yitM*	0.924	1.023	13.202	11.760
BMM	*efeU*	*yrpE*	1.789	1.765	11.174	13.897
SMM	*yydG*	*glpD*	1.443	0.811	13.457	12.132
SMMPr	*cotG*	*yvrL*	1.697	1.638	11.847	9.660
Pyr	*lrgA*	*yczM*	1.707	1.784	13.051	9.206
Glu	*yozL*	*yitJ*	0.547	0.420	8.470	13.058
Fru	*sacC*	*yhcD*	1.430	0.400	12.345	7.829
Glucon	*gntP*	*dppB*	1.246	1.452	12.362	11.370
MG	*ywhK*	*htrB*	0.615	0.292	8.815	13.082
Mal	*maeN*	*ctaF*	1.043	1.122	13.932	11.420
Gly	*ilvH*	*czcD*	1.702	1.701	13.615	12.740
starvation	*ywcE*	*rsiW*	1.260	1.111	12.944	15.317

Table 3.4.3 top biomarker pair solution for each treatment in the tiling array data set

Table containing the data corresponding to the top ranked solution for each condition in the tiling array data set. Margin and threshold values are rounded to three decimal places. Values under the “Condition” variable correspond to the shorthand definition of each treatment. The value in Margin1 corresponds to the margin of Gene1, likewise for Margin2 and Gene2, and likewise for Threshold1 and Threshold2 corresponding to the threshold for Gene1 and Gene2. The values for thresholds are $\log_2$ expression values, the values for margins are the difference between the upper and lower bounds $\log_2$ expression values.

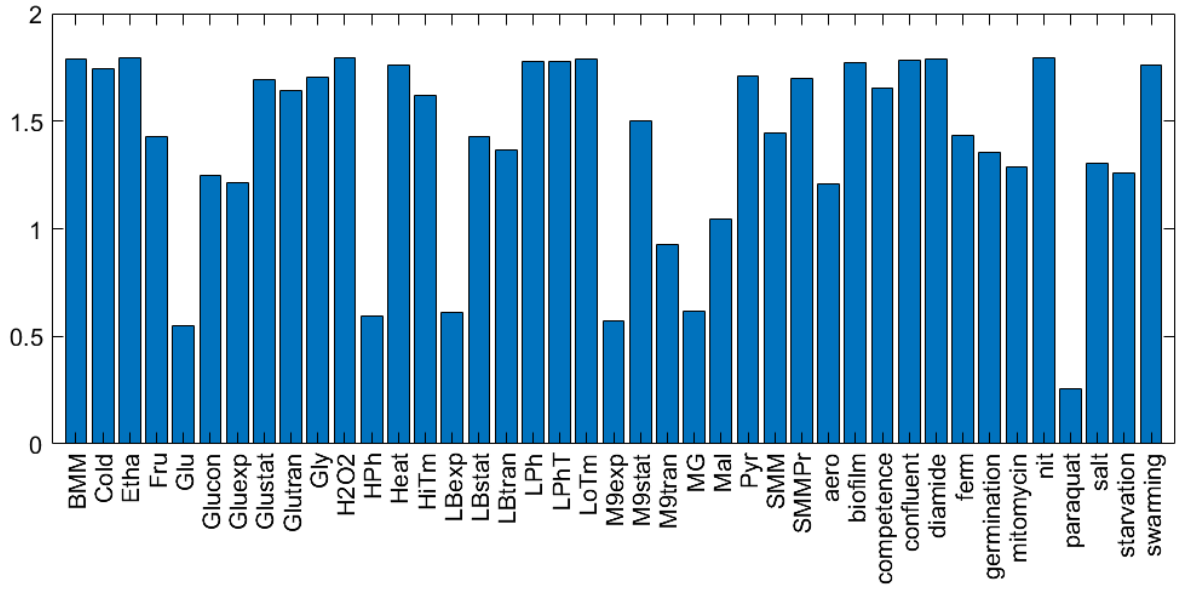

Figure 3.4.4 bar chart of the top biomarker pair solutions minimum margin score for each treatment in the tiling array data set
The lower of the margins in the top solution represents the MMS of the solution. The values for margins are the difference between the upper and lower bounds $\log_2$ expression values. The conditions are labelled according to the shorthand abbreviations. The gene names of the biomarker pair solutions can be cross referenced in Table 3.4.3.

Some conditions in the grouped tiling array data set have relatively small MMSs so stand out as being weaker solutions than the majority of the top ranked solutions (table 3.4.3, figure 3.4.4). It could be stated that the lower the MMS relative to its data set, the lower the confidence of the solution. Therefore, condition with the least confidence in figure 3.4.4, is paraquat with an MMS of 0.2589. Other conditions that rank in the lower quartile with a minimum margin score below 0.9671 are: M9tran, SMM, LBtran, HPh, LBexp, M9exp, Glu, Fru, and MG. Sporulation, GS, and nutrient shift are the conditions which rank the least as ROTC did not find any solutions for biomarker pairs that separate stress state from all other conditions.

With reference to section 3.4.1 where diamide induced stress was used as an example, it is useful to note the placement of diamide in the bar chart in figure 3.4.4. Diamide is not an exceptionally good example in terms of MMS and there are many conditions that exceed its MMS. The suggestion that all solutions from conditions with a similar or even greater MMS than diamide have equal or better robustness and quantity of solutions could be extrapolated from this, however, it can't be known for certain without conducting the same analysis for each condition.

3.5 Biomarker selection using RNA-seq data

As described in section 3.2.2, the RNA-seq data were generated using the same methods as used by the BaSysBio project. The conditions selection for the RNA-seq experiment are: growth in anaerobic conditions (SH: anaerobic); competence (SH: comp); addition of diamide (SH: diamide); addition of ethanol (SH: ethanol); nutrient shift from glucose to malate (SH: Glu2Mal); nutrient shift from malate to glucose (SH: Mal2Glu); growth at low temperature (SH: LoTemp); growth at high temperature (SH: HiTemp); stationary growth phase in SMM (SH: SMMstat). These conditions correspond (in shorthand) to the tiling array experiments as follows: anaerobic with nit; comp with competence; diamide with diamide; ethanol with Etha; Glu2Mal with G2M (no solutions found for tiling array); Mal2Glu with M2G (no solutions found for tiling array); LoTemp with LoTm; HiTemp with HiTm; SMMstat with SMMPr. Again, the shorthand names for these conditions are artefacts of the sample names used in the original data.

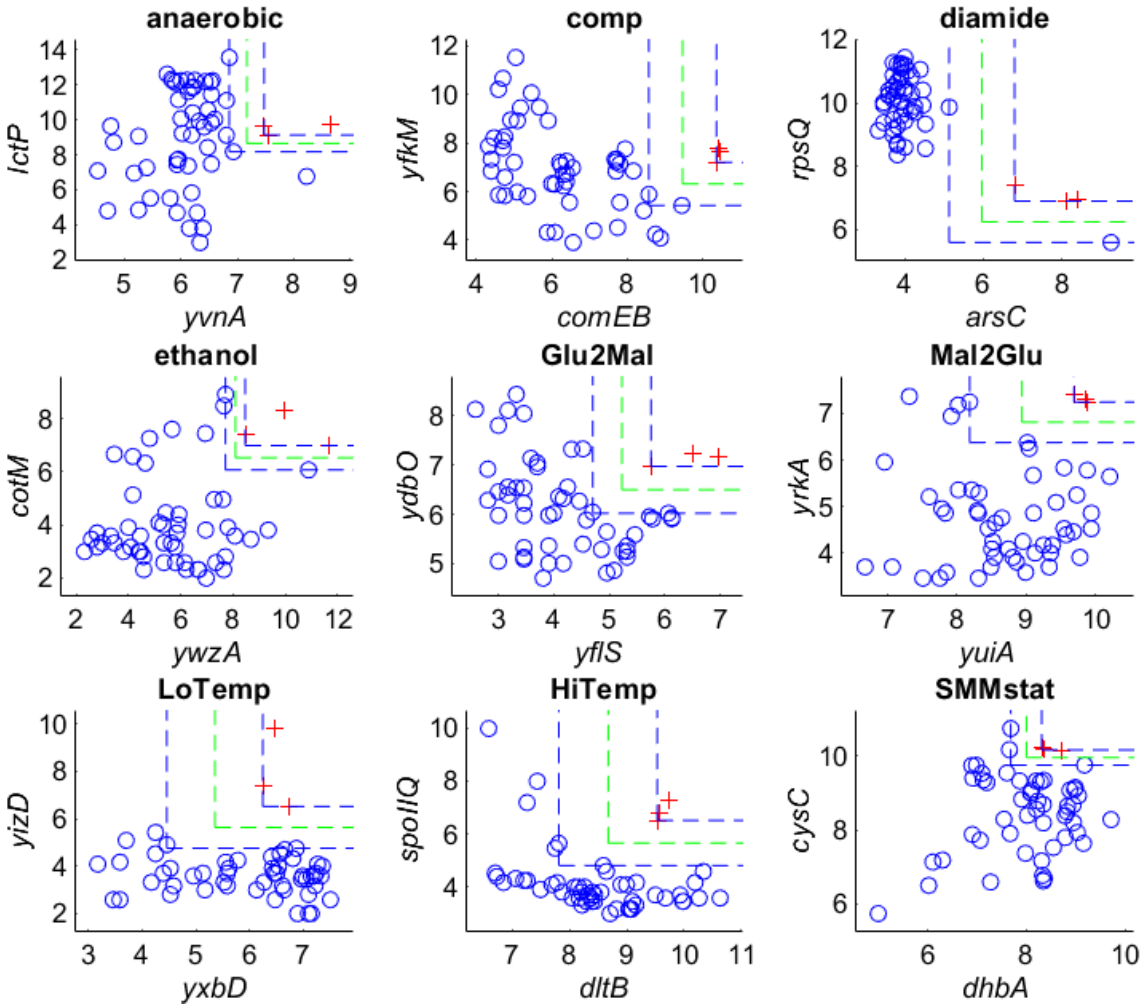

Figure 3.5.1 top ranked biomarker pair solution for each treatment in the RNA-seq data set

The x-axis shows the $\log_2$ expression values for Gene1 of the top ranked solution generated by ROTC. The y-axis shows the $\log_2$ expression values for Gene2 of the top ranked solution generated by ROTC. Positive samples are plotted using red crosses. Negative samples are plotted using blue circles. The blue dotted lines represent the upper and lower bounds of the solution. The green dotted lines represent the thresholds of the solution. Title of each plot denotes the shorthand of the condition that each graph plots the solution data for.

Condition	Gene1	Gene2	Margin1	Margin2	Threshold1	Threshold2
anaerobic	*yvnA*	*lctP*	0.618	0.954	7.167	8.652
comp	*comEB*	*yfkM*	1.798	1.773	9.484	6.313
diamide	*arsC*	*rpsQ*	1.665	1.310	5.962	6.240
ethanol	*ywzA*	*cotM*	0.753	0.911	8.091	6.522
Glu2Mal	*yflS*	*ydbO*	1.054	0.955	5.228	6.500
Mal2Glu	*yuiA*	*yrkA*	1.509	0.873	8.939	6.811
LoTemp	*yxbD*	*yizD*	1.788	1.769	5.354	5.639
HiTemp	*dltB*	*spoIIQ*	1.718	1.700	8.673	5.658
SMMstat	*dhbA*	*cysC*	0.626	0.410	8.000	9.966

Table 3.5.1 Top biomarker pairs for RNA-seq data
Table containing the data corresponding to the top ranked solution for each condition in the RNA-seq data set. Margin and threshold values are rounded to three decimal places. Values under the “Condition” variable correspond to the shorthand definition of each treatment. The value in Margin1 corresponds to the margin of Gene1, likewise for Margin2 and Gene2, and likewise for Threshold1 and Threshold2 corresponding to the threshold for Gene1 and Gene2. The values for thresholds are $\log_2$ expression values, the values for margins are the difference between the upper and lower bounds $\log_2$ expression values.

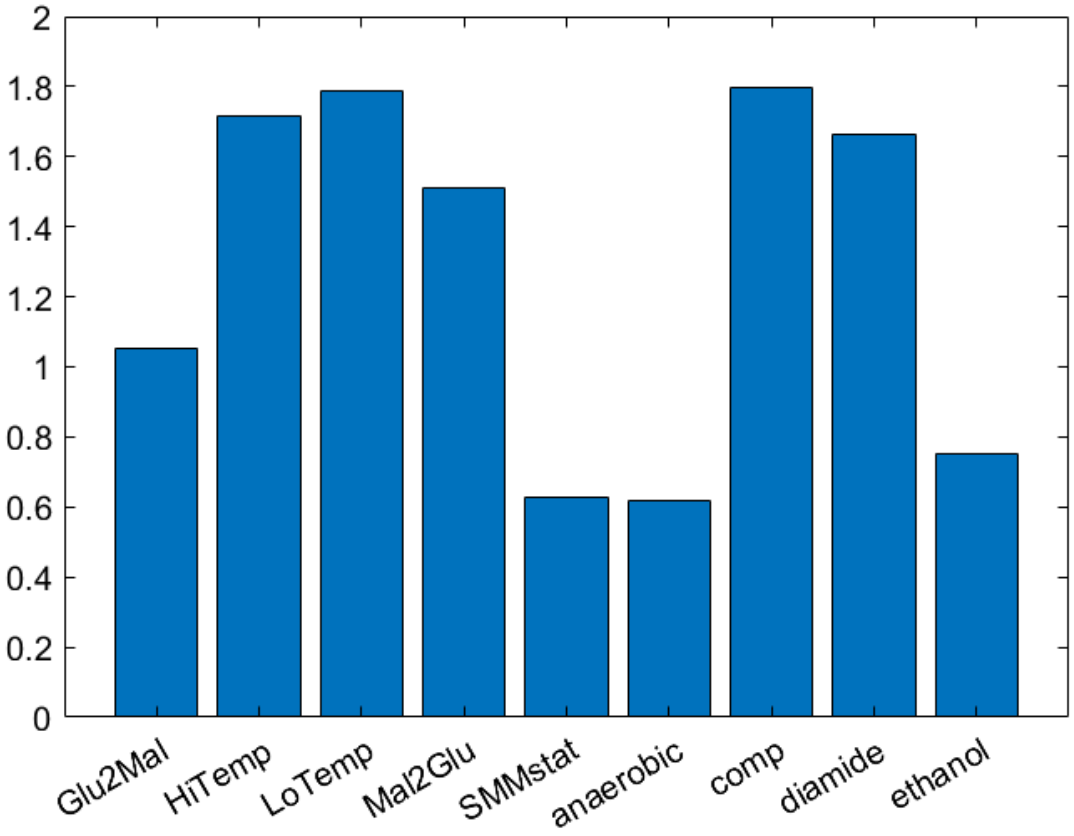

Figure 3.5.2 bar chart of the top biomarker pair solutions minimum margin score for each treatment in the RNA-seq data set
The lower of the margins in the top solution represents the MMS of the solution. The values for margins are the difference between the upper and lower bounds $\log_2$ expression values. The conditions are labelled according to the shorthand abbreviations. The gene names of the biomarker pair solutions can be cross referenced in Table 3.5.1.

Like in the tiling array data set, there are some conditions for which the MMS was low relative to its data set, including SMMstat, anaerobic, and ethanol. The solutions provided for SMMstat, anaerobic, and ethanol could be considered to have lower confidence than the other conditions in the data set based on their MMS.

3.6 Combining the results from both sets of data

3.6.1 Cross validation of biomarker models generated by ROTC

The results from the tiling array data set cannot be directly compared to those of the RNA-seq data set as they each operate within vastly different expression ranges. Instead of comparing the margins and thresholds from the top solutions generated by each data set, the top biomarker pair solution for each condition in the RNA-seq data set has been applied as a model on the samples from the tiling array data set and plotted. Vice versa, the top biomarker pair solution for each condition in the tiling array data set was applied as a model on the samples from the RNA-seq data set for the conditions that were shared between both data sets.

Figure 3.6.1 plots the solutions generated by ROTC using RNA-seq data set as input data against the $\log_2$ expression values for each labelled sample from the tiling array data set. Figure 3.6.2 plots the solutions generated by ROTC using tiling array data set as input data against the $\log_2$ expression values for each labelled sample from the RNA-seq data set. Using diamide stress as an example, the biomarker pair (*arsC*, *rpsQ*) was the top solution generated by ROTC using the RNA-seq data set as the input data, but the samples plotted in figure 3.6.1 originate from the tiling array data set – the positive samples are those labelled as diamide stress, and the negative samples are all other samples. The corresponding plot for diamide stress on figure 3.6.2 plots the biomarker pair (*treA, ycnD*), which was the top solution using the tiling array data set, but the samples on figure 3.6.2 originate from the RNA-seq data set where the positive samples are those labelled as diamide stress, and the negative samples are all other samples in the data set.

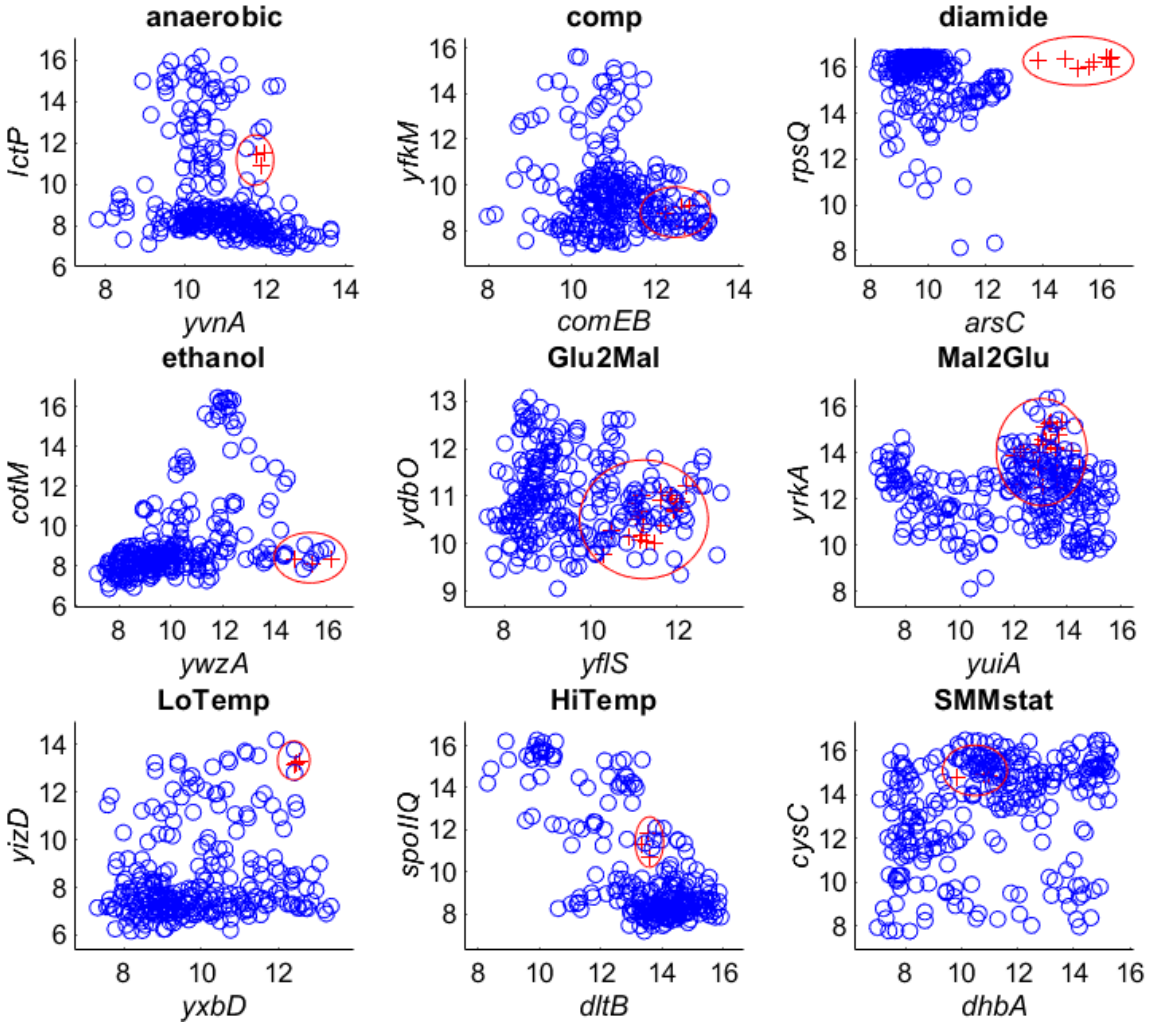

Figure 3.6.1 top biomarker pair solution for each condition in the RNA-seq data set plotted against the samples from the tiling array data set
The x-axis shows the $\log_2$ expression values (from tiling array data set) for Gene1 of the top ranked solution generated by ROTC (for the RNA-seq data set). The y-axis shows the $\log_2$ expression values for Gene2 of the top ranked solution generated by ROTC. Positive samples are plotted using red crosses and are encircled by red ellipses. Negative samples are plotted using blue circles. Title of each plot denotes the shorthand of the condition that each graph plots the solution data for.

The plots in figure 3.6.1 demonstrate that many negative samples are misclassified when the top solution from the RNA-seq data is applied to the tiling array data set. The results could be explained by the tiling array data set containing more treatments than those included in the RNA-seq data set, and the biomarker pair solutions generated from the RNA-seq data set were overfitted to their data set.

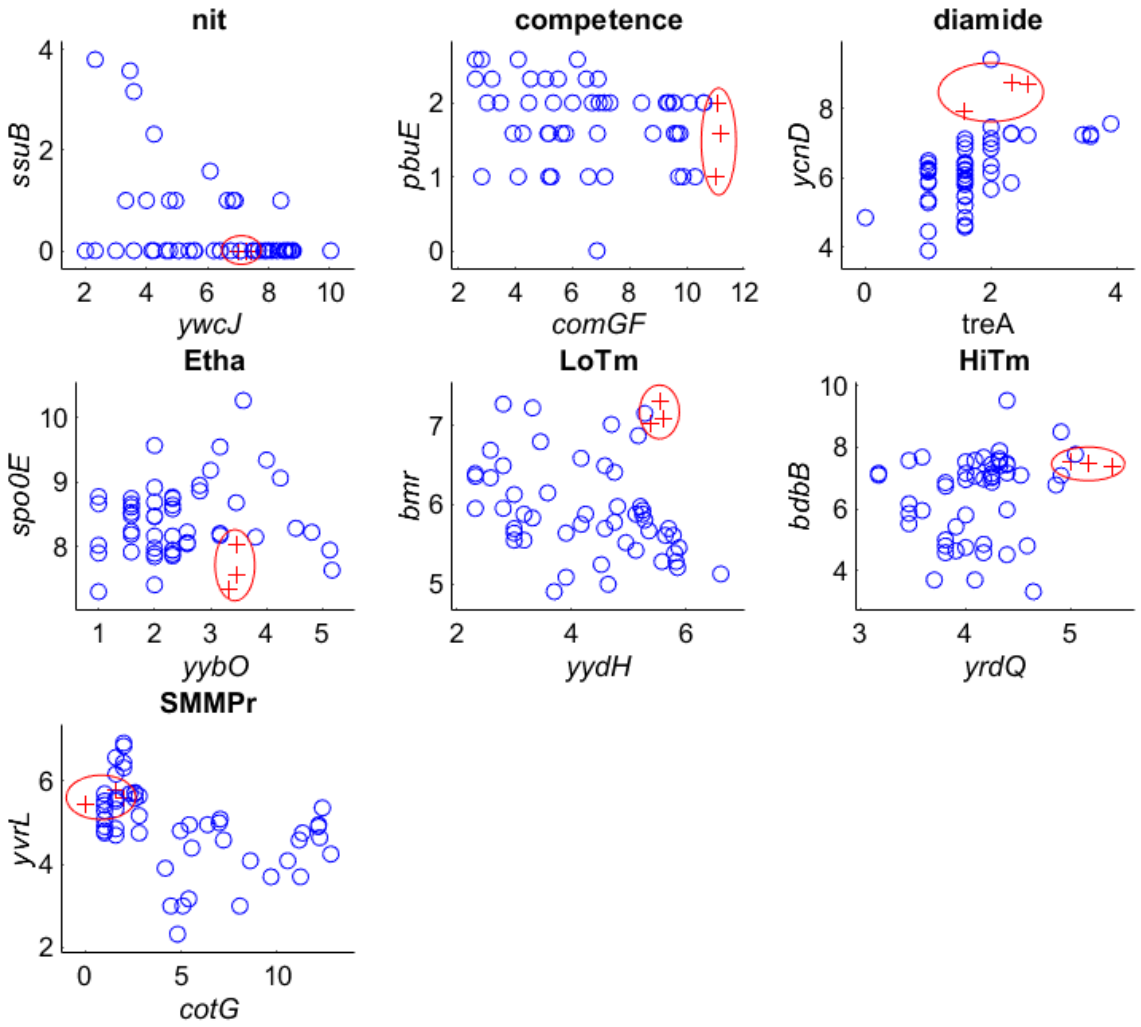

Figure 3.6.2 top biomarker pair solution for each shared condition of the tiling array data set plotted against the samples of the RNA-seq data set
The x-axis shows the $\log_2$ expression values (from RNA-seq data set) for Gene1 of the top ranked solution generated by ROTC (for the tiling array data set). y-axis shows the $\log_2$ expression values for Gene2 of the top ranked solution generated by ROTC. Positive samples are plotted using red crosses and are encircled by red ellipses. Negative samples are plotted using blue circles. Title of each plot denotes the shorthand of the condition that each graph plots the solution data for.

The plots in figure 3.6.2 highlight a different issue related to misclassification, in that the RNA-seq data appears unreliable for many of the selected biomarker genes. For example, the biomarker gene *ssuB* which was selected to identify anaerobic stress (shorthand: nit) has $\log_2$ expression values equalling zero for many of the samples (including all positive samples). Linear distributions are observed from lower values such as zero, one, or two whereas a more scattered distribution was expected, and these lower values were the result of lower amounts of reads mapped against these genes in the RNA-seq data. Many of these genes were filtered out

from the RNA-seq data set prior to being inputted to ROTC if their expression values were below a threshold of one TPM.

3.6.2 Finding solutions in common between both data sets

Solutions in common were found by searching both results for pairs of biomarkers with the same gene names. Each solution may have markedly different margins and thresholds depending on the data set so a method to rank these common solutions was defined. For every solution, the MMS was calculated relative to each data set which was denoted as MMS1 for the tiling array data set and MMS2 for the RNA-seq data set. Then, the minimum was taken between MMS1 and MMS2 and denoted as MMMS for Minimum-MMS. MMMS was used to rank the solutions in descending order.

As an example, the top solution in common for diamide has been plotted in figure 3.6.3 against the tiling array data set and the RNA-seq data set separately. The top solutions in common have their properties listed in table 3.6.1. Within these top solutions the same biomarker genes appear many times such as *copA* and *copz,* as well as biomarker genes from the same operon. The MMMS for the top solution was 1.166 which is relatively high compared to the MMMS for common solutions of different treatments.

For contrast, another example was selected where the solutions were less well separated when plotted on the tiling array data set; ethanol was selected as there are many misclassified samples on figure 3.6.1 for this treatment. The top common solution is plotted on figure 3.6.4 and the properties of the top solutions are listed in table 3.6.2. The MMMS for the top solution in common for ethanol treatment is 0.502, which is relatively much lower than that for diamide treatment. It must be noted that the value of 0.502 is the MMS for the RNA-seq data set which has much smaller margins regardless as the $\log_2$ expression values have a lower range.

The top solutions in common of both data sets for every treatment in common are listed in table 3.6.3 and ranked in descending order of their MMMS. The treatment with the lowest MMMS is stationary phase in SMM (shorthand = SMMstat/SMMPr). The treatment with the greatest MMMS is for growth in low temperature (shorthand = LoTm/LoTemp) and is followed up by growth in high temperature (shorthand = HiTm/HiTemp).

Combining the results from both data sets yielded some unexpected conclusions. In most cases, when the top biomarker pair solution from one data set was applied as a classifier to the samples of the other data set, the model would no longer separate the positive and negative samples. However, solutions in common could be found that separate positive and negative samples in both data sets. The solutions in common varied in confidence and margin size depending on the condition.

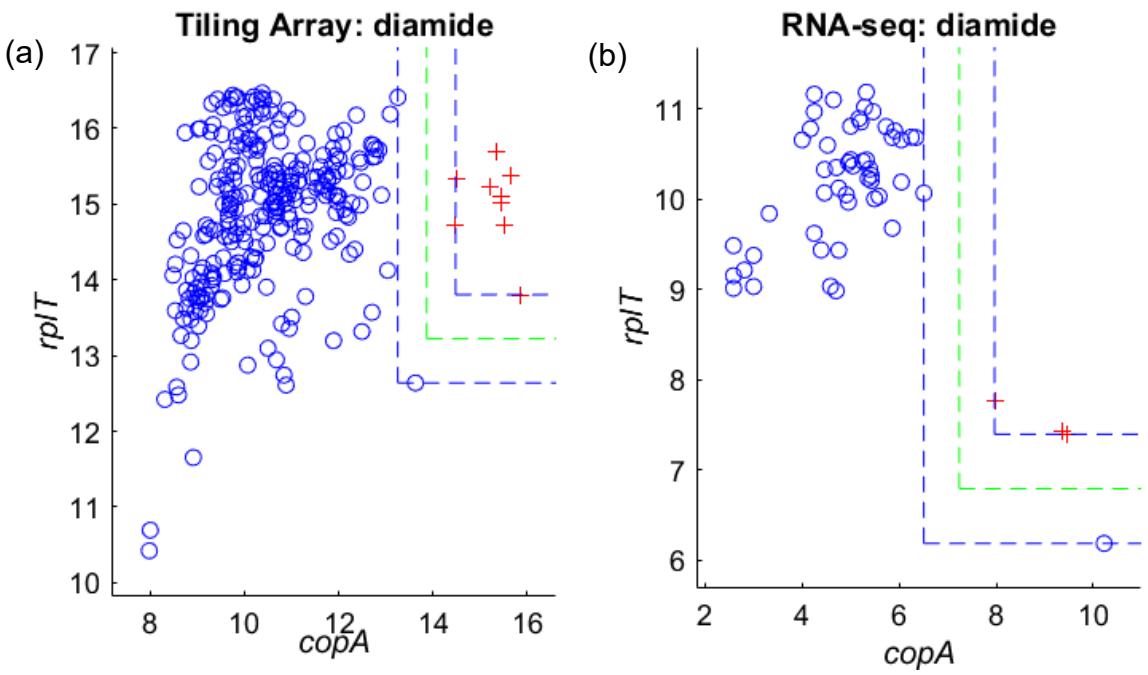

Figure 3.6.3 top biomarker pair solution in common for diamide in both data sets

The x-axis shows the $\log_2$ expression values for Gene1 (*copA*); the y-axis shows the $\log_2$ expression values for Gene2 (*rplT*). Positive samples are represented as red crosses; negative samples are represented as blue circles. Blue dashed lines represent the upper and lower bounds; green dashed lines represent the thresholds. Title of the plot contains the shorthand for the condition selected as a stress treatment for ROTC. Plot (a) plots the top solution in common against the tiling array data set ($\log_2$ expression values). Plot (b) plots the top solution in common against the RNA-seq data set ($\log_2$ expression values).

Solutions		Tiling Array Data Set Properties				RNA-seq Data Set Properties				Scores		
Gene 1	Gene 2	Margin 1	Margin 2	Threshold 1	Threshold 2	Margin 1	Margin 2	Threshold 1	Threshold 2	MMS 1	MMS 2	MMMS
copA	*rplT*	1.223	1.166	13.864	13.221	1.458	1.211	7.237	6.795	1.166	1.211	**1.166**
copA	*rpmI*	1.223	0.940	13.864	14.351	1.458	0.972	7.237	5.186	0.940	0.972	**0.940**
copA	*yugI*	1.223	1.082	13.864	13.489	1.458	0.887	7.237	6.983	1.082	0.887	**0.887**
copZ	*yugI*	1.329	1.279	15.344	13.391	1.335	0.887	7.668	6.983	1.279	0.887	**0.887**
copA	*thiN*	1.223	1.655	13.864	9.359	1.458	0.807	7.237	2.404	1.223	0.807	**0.807**
rpmH	*copA*	1.036	1.223	13.019	13.864	0.752	1.458	5.624	7.237	1.036	0.752	**0.752**
copZ	*cotD*	1.470	0.740	15.273	12.355	1.335	0.766	7.668	6.027	0.740	0.766	**0.740**
copA	*frr*	1.223	1.709	13.864	12.775	1.458	0.657	7.237	5.029	1.223	0.657	**0.657**
copA	*rpsNA*	1.223	1.576	13.864	14.709	1.458	0.654	7.237	7.621	1.223	0.654	**0.654**
copZ	*ackA*	1.539	1.286	15.239	12.703	1.335	0.637	7.668	4.489	1.286	0.637	**0.637**

Table 3.6.1 top ten biomarker pair solutions in common for diamide in both data sets
Margin, threshold, and MMS values are rounded to three decimal places. Margin1 and Margin2 correspond to the margins for Gene1 and Gene2, respectively. Threshold1 and Threshold2 correspond to the thresholds for Gene1 and Gene2, respectively. MMS1 and MMS2 show the minimum margin score for the tiling array data set solution and the RNA-seq data set solution, respectively. MMMS is the minimum of both minimum margin scores (Minimum-MMS).

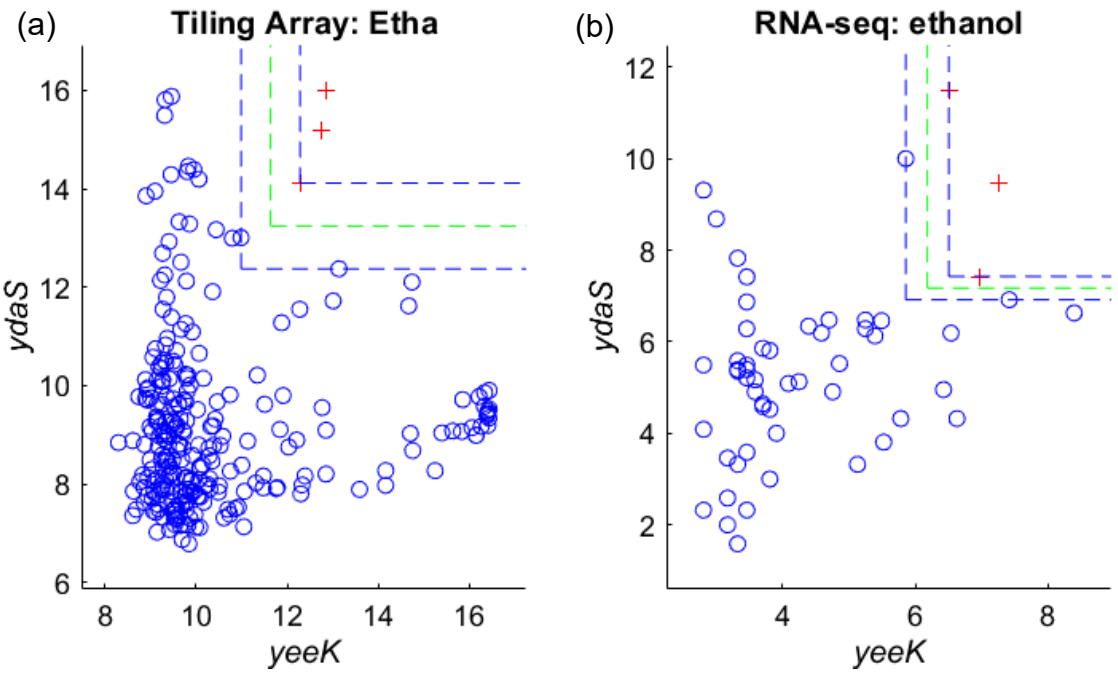

Figure 3.6.4 top biomarker pair solution in common for ethanol in both data sets

The x-axis shows the $\log_2$ expression values for Gene1 (*yeeK*); the y-axis shows the $\log_2$ expression values for Gene2 (*ydaS*). Positive samples are represented as red crosses; negative samples are represented as blue circles. Blue dashed lines represent the upper and lower bounds; green dashed lines represent the thresholds. Title of the plot contains the shorthand for the condition selected as a stress treatment for ROTC. Plot (a) plots the top solution in common against the tiling array data set ($\log_2$ expression values). Plot (b) plots the top solution in common against the RNA-seq data set ($\log_2$ expression values).

Solutions		Tiling Array Data Set Properties				RNA-seq Data Set Properties				Scores		
Gene1	Gene2	Margin 1	Margin 2	Threshold 1	Threshold 2	Margin 1	Margin 2	Threshold 1	Threshold 2	MMS 1	MMS 2	MMMS
yeeK	*ydaS*	1.296	1.741	11.632	13.246	0.650	0.507	6.183	7.173	1.296	0.507	**0.507**
ywzA	*yonD*	1.205	1.688	14.170	10.633	0.396	0.363	8.270	2.989	1.205	0.363	**0.363**
rpmEB	*ydbS*	0.363	0.367	15.363	12.306	0.348	0.447	6.955	8.043	0.363	0.348	**0.348**
rpmEB	*ydbT*	0.363	0.600	15.363	12.325	0.348	0.410	6.955	7.238	0.363	0.348	**0.348**
ywrE	*rpmEB*	0.568	0.363	11.448	15.363	0.564	0.348	4.806	6.955	0.363	0.348	**0.348**
yxnA	*yjoB*	0.669	0.750	12.707	11.198	0.341	0.355	5.077	5.604	0.669	0.341	**0.341**
yxjI	*rpmEB*	0.626	0.400	11.668	15.344	0.328	0.348	5.836	6.955	0.400	0.328	**0.328**
bofC	*yjoB*	0.459	0.522	13.086	11.312	0.379	0.322	5.096	5.620	0.459	0.322	**0.322**
ykgA	*yjoB*	1.725	1.555	13.736	10.796	0.655	0.322	5.882	5.620	1.555	0.322	**0.322**
ywzA	*yonE*	1.205	1.073	14.170	8.888	0.396	0.322	8.270	2.161	1.073	0.322	**0.322**

Table 3.6.2 top ten biomarker pair solutions in common for ethanol in both data sets
Margin, threshold, and MMS values are rounded to three decimal places. Margin1 and Margin2 correspond to the margins for Gene1 and Gene2, respectively. Threshold1 and Threshold2 correspond to the thresholds for Gene1 and Gene2, respectively. MMS1 and MMS2 show the minimum margin score for the tiling array data set solution and the RNA-seq data set solution, respectively. MMMS is the minimum of both minimum margin scores (Minimum-MMS).

			Tiling Array Data Set Properties				RNA-seq Data Set Properties				
Condition	Gene1	Gene2	Margin1	Margin2	Thresh1	Thresh2	Margin1	Margin2	Thresh1	Thresh2	MMMS
LoTemp	*yktD*	*yizD*	1.549	1.621	11.216	12.351	1.659	1.617	5.077	5.715	**1.549**
HiTemp	*yxlC*	*mreBH*	1.329	1.506	11.607	11.187	1.433	1.561	5.623	8.460	**1.329**
Diamide	*copA*	*rplT*	1.223	1.166	13.864	13.221	1.458	1.211	7.237	6.795	**1.166**
Ethanol	*yeeK*	*ydaS*	1.296	1.741	11.632	13.246	0.650	0.507	6.183	7.173	**0.507**
Competence	*yrhG*	*comEC*	0.421	0.776	9.072	10.712	0.531	0.643	3.435	8.198	**0.421**
Anaerobic	*yxeP*	*yjbB*	0.340	0.271	11.180	9.276	0.322	0.280	4.161	3.947	**0.271**
SMMstat	*bacB*	*ytrC*	0.309	0.200	11.973	11.947	0.241	0.126	7.902	5.522	**0.126**

Table 3.6.3 top biomarker pair solutions in common between both data sets for all conditions in common
Margin, threshold, and MMMS values are rounded to three decimal places. Conditions are ranked in descending order of MMMS. Thresh is short for Threshold. values are rounded to three decimal places. Margin1 and Margin2 correspond to the margins for Gene1 and Gene2, respectively. Threshold1 and Threshold2 correspond to the thresholds for Gene1 and Gene2, respectively. MMMS is the minimum of both minimum margin scores (Minimum-MMS).

3.7 Discussion

The work presented in this chapter shows that ROTC is a capable maximal margin classifier for one, two, and three dimensions. For the data sets tested in this study, ROTC generates candidate solutions of biomarkers for the given conditions with a few exceptions where no separable solutions were found.

The results for biomarkers of specific conditions are intended to function as biomarkers for that given condition alone and under no other conditions, however, this can't be stated definitively based on the results presented in this chapter. When the biomarkers generated from the RNA-seq data set were plotted on the tiling array data set, many negative samples were misclassified as positive samples; this behaviour could imply that ROTC is only capable of generating biomarker solutions suitable for the conditions provided as input data – a problem akin with overfitting in machine learning (El Naqa and Murphy, 2015). However, if the extensive list of conditions provided in the tiling array data set can be assumed as exhaustive of all natural stresses that *B. subtilis* may undergo, the biomarkers generated from this data set ought to be robust enough to distinguish against practically all other stresses. Based on the results generated from the RNA-seq data set, it seems that the subset of conditions is insufficient for an exhaustive search to select uniquely specific biomarkers.

When the biomarkers generated using the tiling array set were applied to the RNA-seq data set, there were other issues resulting in the misclassification of samples. The $\log_2$ TPM values were very low (frequently one or zero) which could have several contributing factors. The quality of the RNA-seq data may have been insufficient to pick up smaller reads or lower concentrations, which may be attributed to the method of library preparation (Shi *et al.*, 2021). There could also be an issue in the tiling array data set as these are a technology known for having noise in the data causing unwanted variation (Negi *et al.*, 2022).

ROTC works effectively and speedily in lower dimensions such as searching for individual biomarkers and pairs of biomarkers. However, there is a significant reduction in computational performance when searching for triples of biomarkers. To scale up ROTC for higher dimensions, it would be necessary to either significantly improve the memory efficiency and performance of the algorithm, or to

direct away from an exhaustive search in favour of machine learning techniques for optimisation, such as genetic algorithms or random forest models.

MMS is a convenient approach for ranking solutions generated by ROTC. However, as discussed in section 3.3.1, it is not an absolute indicator of how good a solution is because it depends on the nature of the input data. In the case of this study which utilised transcriptomic data in the form of $\log_2$ expression values, the magnitude of MMS depends on the values of the upper and lower bounds. It would be beneficial to work toward a standardised measure for how good a solution is to make the task of determining the best biomarker(s) less manual.

The biomarkers generated by ROTC show effective separation according to the data sets in this study. To determine how effective the biomarkers are for their intended application, as inputs to an AND gate, *in vivo* testing would be necessary. A general outline for *in vivo* testing as inputs to an AND gate would first require the identification of the natural promoters that the selected biomarker genes are regulated by. The natural promoters would be linked to a simple dual-input AND gate and tested under a range of conditions to test the hypothesis that the AND gate will only generate an input under the selected treatment condition only. It is critical that the margin be as large as possible for an effective AND gate to operate with any biomarker solution. Additionally, there is biological nuance pertaining to the regulation of the promoter sequences; any given promoter may require larger or smaller expression levels than expected which may result with indistinct ON/OFF states (Hicks, Bachmann and Wang, 2020). Ideally, the OFF state associated with negative samples should have as close to no expression as possible and the ON state associated with positive samples should have as much expression as possible to be measurable when the AND gate is ON. Therefore, it is essential that solutions generated by ROTC are studied carefully and tested *in vivo*.

In this study, ROTC was tested on two sources of data: tiling array data and RNA-seq data from *B. subtilis* 168 under an extensive range of treatments. ROTC can be tested on other data sets also to attain further evidence of its capability to maximally separate treatments. Of particular interest, testing ROTC on a similarly extensive set of data for *Escherichia coli* would be beneficial for this study where portability between *B. subtilis* and *E. coli* is the subject of research. Additionally, the application

of ROTC to more data from *B. subtilis* 168 would be useful to compare the biomarkers generated there with those generated within this study to see how many solutions remain in common and how many misclassifications occur for the solutions of each data set. Integration of NGS data from various studies on the same strain would be beneficial to account for noise and variation between experiments, similar to how the BaSysBio data set was integrated from several experiments in different institutions (Nicolas *et al.*, 2012).

It is important to tailor the input data for its intended purpose. The results presented in this chapter from the two *B. subtilis* 168 data sets show that ROTC provides a list of viable solutions regarding the presented data set. Furthermore, ROTC will provide the solutions with the greatest separation between the user defined control and treatment samples. Much of the potential error from generated solutions results from the quality of the data itself and how it has been presented to the algorithm, such as the labelling of the data and the quality of the sequencing data. The recommended preparation of the data would be to remove any unnecessary conditions from the control samples and keep only those that are required to generate the biomarkers needed which would result in stronger, less specific biomarkers. Generation of highly specific biomarkers is one of the benefits of ROTC as it will not provide any solutions that do not separate within the provided data set, but it comes at the cost of biomarker strength; as such, it is recommended to filter the data set appropriately depending on how specific the biomarkers need to be.

Overall, ROTC has demonstrated its capability to generate sets of biomarkers that distinguish highly specific stress signatures. ROTC was tested on two sets of data in this study, tiling array and RNA-seq of *Bacillus subtilis* 168. Biomarkers generated from a singular data set did not show immediate cross-compatibility with the other data set but combined, they could find solutions that were effective in both, albeit of lower strength. ROTC has room for further optimisation and has generated results that are worth testing *in vivo* to determine their performance in living systems. ROTC has the potential to be a useful tool for researchers doing transcriptomic analysis.

CHAPTER 4

4 A synthetic genetic AND gate for detecting a given stress response

4 A synthetic genetic AND gate for detecting a given stress response

4.1 Introduction

In this study, an AND gate was built and characterised based on a design by Anderson *et al.*, and adapted to operate in two species with minimal refactoring to enhance the portability of the system (Anderson, Voigt and Arkin, 2007). The AND gate was built for the purpose of detecting a specific stress response in a cell by using stress-specific biomarkers as inputs to the AND gate; thus, working as an *in vivo* demonstrator for the effectiveness of solutions produced by ROTC (see chapter 3). In this work, the AND gate was built and tested using simple inducible promoters to demonstrate its functionality as an AND gate.

Results are presented from assaying the individual components of the AND gate using *E. coli* DH5α, *B. subtilis* 168, and then testing the whole system in an *E. coli* S30 cell-free system. The results show that the individual components work as expected in isolation within the strains tested, but not when combined in the cell-free system. This study uncovers multiple factors that need additional testing to optimise the AND gate and ensure that it can be functional in both species. The results generated by testing the AND gate constructs in isolation supplies further evidence to the modularity of the T7 based system, by using different parts compared with the original study by Anderson *et al.*, and demonstrate the potential for portability between the two model organisms (*E. coli* DH5α and *B. subtilis* 168) tested within this study.

4.1.1 Contributions

The design of the AND gate in this chapter was based on the work of Anderson, Voigt and Arkin (Anderson, Voigt and Arkin, 2007). Gene synthesis was conducted by Integrated DNA Technologies (IDT) and Twist Bioscience. RNA extraction, library preparation and sequencing were outsourced to Azenta Life Sciences. RNA-seq raw reads were processed using the nf-core/rnaseq pipeline (Ewels *et al.*, 2020).

4.1.2 Motivation

Previous studies have shown that it is possible to utilise an AND gate to detect a specific stress response by using promoters as inputs that regulate the biomarkers detected by the stress response (Ceroni *et al.*, 2018). For this study, it was intended

to build a simple dual input AND gate with the goal of detecting a specific stress response using biomarkers generated by the ROTC (Recursive Orthogonal Threshold classifier) algorithm (Figure 4.1.1).

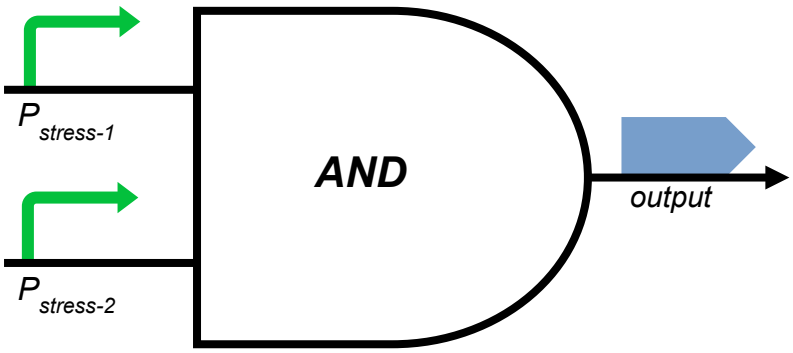

Figure 4.1.1 Diagram of a genetic stress detection genetic AND gate
Diagram of stress detection AND gate. $P_{stress\text{-}1}$ and $P_{stress\text{-}2}$ represent promoters based on stress-specific biomarkers generated by ROTC, and *output* represents a generic detectable product such as GFP.

ROTC is a biomarker selection algorithm with the purpose of generating condition specific biomarkers with properties tailored to work as inputs in a genetic AND gate. In chapter 3, ROTC was discussed in detail and used to generate many sets of biomarkers for different stress conditions using various data sets. In this chapter, a simple dual input AND gate was built to enable future testing of ROTC's ability to generate biomarkers that optimise AND gate performance. It was important for this study that the AND gate met the aims of portability between species. The species of interest are *B. subtilis* and *E. coli*, so the AND gate was designed with a test strategy to operate and be characterised in both species.

4.1.3 Genetic logic gates

Genetic logic gates have been developed in synthetic biology for a multitude of applications (Moser *et al.*, 2012; Sayut, Niu and Sun, 2009; Vishweshwaraiah *et al.*, 2021; Lebovich, Zeng and Andrews, 2023; Bose *et al.*, 2023). Genetic logic gates are inspired by electronic logic gates used in computing used to perform logical operations such as AND, OR, and NOT. Logical operations work by transforming inputs into an output; the outputs can be determined using truth tables. For example, the AND operation will only output true if every input is true. Examples of logic gates are given in figure 4.1.2 with associated truth tables.

Instead of electronic devices, genetic logic gates employ biochemical materials and mechanisms to transform input signals into an output. Typically, each input will be controlled by an inducible promoter allowing the input the turn to ON state when the promoter is activated. The output must be regulated by the activity of the inputs depending on the logical operation. For example, the output of an OR gate can be activated when either input is ON; this means that each input may encode the same product the activates the output. Genetic logic gates can be used individually as a simple actuator or combined for more complex purposes such as making biological circuits (Goñi-Moreno and Amos, 2012; Moser *et al.*, 2012; Anderson, Voigt and Arkin, 2007; Wu *et al.*, 2023).

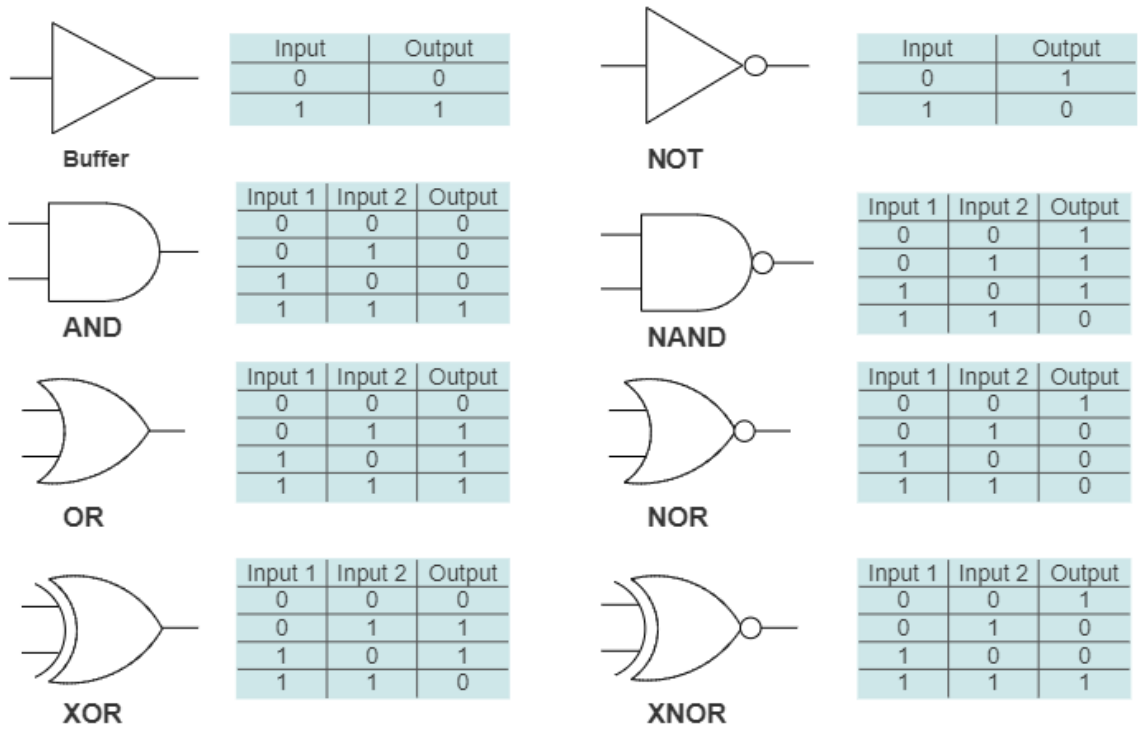

Figure 4.1.2 Logic gates and associated truth tables.
Symbols representing logic gates (Buffer, NOT, AND, NAND, OR, NOR, XOR, & XNOR). Labels for each symbol are located beneath each symbol, associated truth table is to the left of each symbol. Some logic gates use shorthand which is defined as follows: NAND = NOT AND; NOR = NOT OR; XOR = Exclusive OR; XNOR = Exclusive NOT OR. Dual input representations are shown for simplicity, but the logic extends to multiple inputs, except for Buffer and NOT which are single input. 0 is used in the truth tables to represent a value of false or OFF, 1 is used to represent a value of true or ON.

A genetic AND gate can only produce an output when both inputs are in the ON state. Both inputs can be controlled by inducible promoters, but the output must be driven by a culmination of both input products to accurately emulate an AND operation. AND gates have been designed exploiting various mechanisms within cells such as transcription factors, riboswitches and CRISPR (Ceroni *et al.*, 2018; Sharma, Nomura and Yokobayashi, 2008; Sayut, Niu and Sun, 2009; Vishweshwaraiah *et al.*, 2021).

4.1.4 T7 based genetic AND gate

Anderson, Voigt and Arkin described an AND gate based on the amber suppressor tRNA SupD (Anderson, Voigt and Arkin, 2007; Hoffman and Wilhelm, 1970). SupD binds to the amber stop codon, TAG, and encodes a serine where translation would normally terminate. Modifying a coding sequence (CDS) to include TAG amber stop codons allows it to be functionally translated only when SupD is present. Anderson, Voigt and Arkin modified T7 RNA polymerase (T7 RNAP) with two TAG amber stop codons for their AND gate. The modified CDS for T7 RNAP was named *T7pTag* and was coupled with the SupD tRNA as inputs to a modular AND gate. When both inputs are ON, the resulting product is a functional T7 RNAP which activates the P_{T7} promoter of an output device (Figure 4.1.3).

In their 2007 study, Anderson *et al.* demonstrated the modularity of the AND gate by testing it with different promoters linked to the input devices. First, the inputs were linked up to P_{sal} and P_{BAD} promoters which are activated by salicylate and arabinose, respectively. The output device, activated by the P_{T7} promoter, produces a green fluorescent protein which was used to characterise the system by fluorescence measurements.

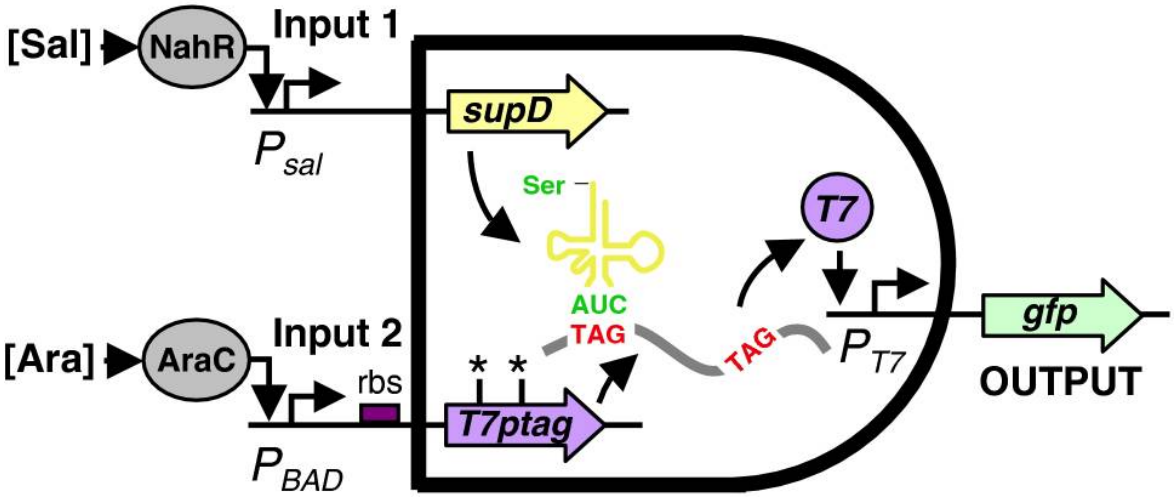

Figure 4.1.3 T7 based AND gate from 'Environmental signal integration by a modular AND gate'
Diagram of the T7 based genetic AND gate from 'Environmental signal integration by a modular AND gate' (Anderson, Voigt and Arkin, 2007). Depiction of the mechanism driving T7 RNAP production is shown within the gate. Promoter formation shown is with P_{sal} for Input 1 and P_{BAD} for Input 2.

After the initial characterisation of the gate, the input promoters and the output product were swapped with alternatives to demonstrate the gate's modularity. P_{sal} and P_{BAD} were changed to P_{mgrB} and P_{lux} which respond to magnesium limitation and quorum signal AI-1, respectively. The output was changed to the invasin gene, allowing the bacteria to invade mammalian cells, and can be detected by invasion assays. Both interpretations of the design led to positive results showing that the T7 based AND gate is a functional AND gate and it is tuneable in terms of exchanging promoters and output (Anderson, Voigt and Arkin, 2007).

4.2 Design

The modular T7 based AND gate described by Anderson *et al.* is a good candidate to be adapted for the purpose of detecting a stress response. As Anderson *et al.* demonstrated the modularity of the gate, it is possible to swap out various components to better suit the requirements for stress detection. Additionally, the stress detection AND gate is intended to operate in two species (*B. subtilis* and *E. coli*), so the parts selected for this design were chosen based on their ability to work in both species (Anderson, Voigt and Arkin, 2007).

4.2.1 Modification to design toward a portable AND gate

A testing system (see Figure 4.2.1) has been designed which produces a fluorescent protein as output to allow characterisation by fluorescence assays. A blue fluorescent protein, mTagBFP, was chosen for the output signal of the AND gate system. A red fluorescent protein, mCherry2, was selected for initial characterisation of the second input to the AND gate which produces the mutant T7 RNAP, T7pTag.

The following parts were selected for the test system: P_{grac} and P_{xyl} were chosen as inducible promoters; the consensus ribosome binding site (RBS) for *B. subtilis* was chosen for RBS; licBCAH and rrn0 were chosen as terminators. P_{grac} and P_{xyl} are both induced by easily sourced small molecules which make them good candidates for the AND gate: IPTG (isopropyl ß-D-1-thiogalactopyranoside) for P_{grac} and (D)-xylose for P_{xyl}. P_{grac} is a synthetic promoter created by the combination of the *gros* and *lac* operons from *B. subtilis* and *E. coli*, respectively (Phan, Nguyen and Schumann, 2006). P_{grac} was designed with the intention of creating an IPTG inducible expression vector for *B. subtilis* that also operates in *E. coli*. Though P_{xyl} originates from *B. subtilis*, it shares homology with the xylose operon in *E. coli* and has been shown to function in both species (Wilhelm and Hollenberg, 1985; Atanassov *et al.*, 2013).

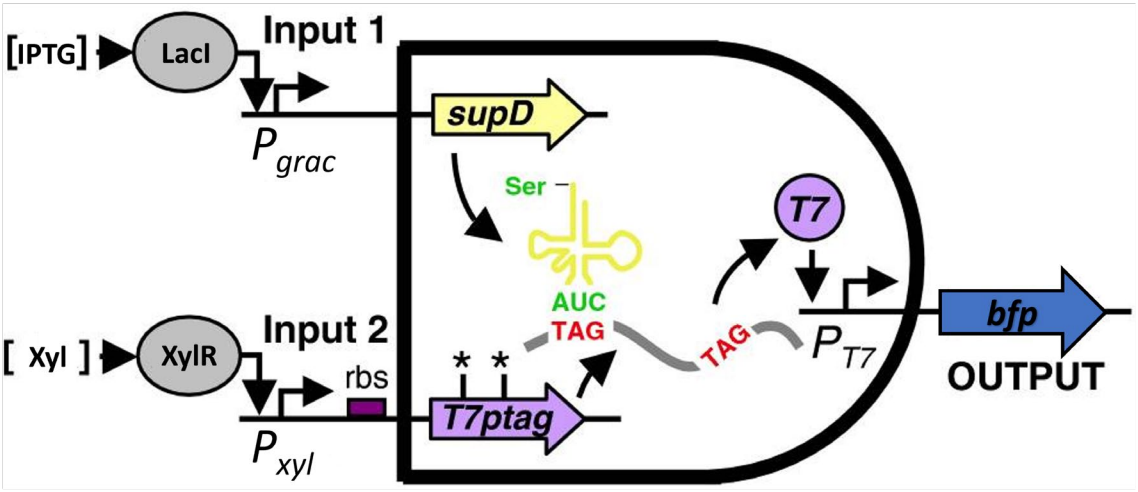

Figure 4.2.1 Modified T7-based AND gate design
Diagram of the altered T7 based genetic AND gate, inspired by 'Environmental signal integration by a modular AND gate' (Anderson, Voigt and Arkin, 2007). Output has been altered from GFP to BFP. Input 1 and Input 2 have changed to P_{grac} and P_{xyl}.

P_{grac} and P_{xyl} are regulated by repressors encoded by the *lacI* and *xylR* genes, respectively. Repressor proteins, LacI and XylR, block the binding of RNA polymerase (RNAP) thus preventing transcription. When an inducer molecule (IPTG or (D)-xylose) is added, the inducer binds to the repressor altering its conformity and unbinding from the operator sequence. RNAP is then able to bind to the operator and produce mRNA from the gene template (Figure 4.2.2). *E. coli* naturally has the *lacI* gene and *B. subtilis* naturally has *xylR,* but the repressor genes should be included on the plasmids conveying the AND gate to ensure that the promoters are regulated properly; insufficient repression leads to leaky expression which is not intended AND gate behaviour (Kim, Mogk and Schumann, 1996; Koreeda *et al.*, 2023).

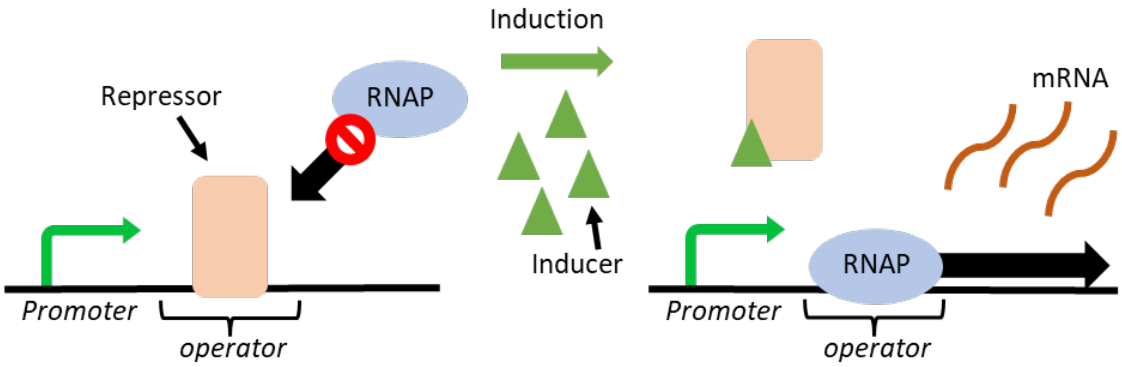

Figure 4.2.2 Diagram of regulation of inducible promoters
Generalised diagram modelling the interaction between regulatory molecules associated with the P_{grac} and P_{xyl} promoters. The mode of regulation for both promoters is the same but the molecules involved are different: LacI is the repressor for $P_{grac,}$ and its inducer is allolactose or analogues (IPTG); XylR is the repressor for $P_{xyl,}$ and its inducer is (D)-xylose.

pHT01 was selected as the backbone for the AND gate constructs as it operates in both *B. subtilis* and *E. coli* (Nguyen et al., 2005; Nguyen, Phan and Schumann, 2007; Phan, Nguyen and Schumann, 2006). pHT01, and derivative vectors, are shuttle vectors for *B. subtilis* for IPTG inducible expression via P_{grac}. For *E. coli*, the *ampR* gene conveys resistance to ampicillin; in *B. subtilis*, the *cmR* gene conveys resistance to chloramphenicol. An annotated plasmid map for pHT01 is shown in figure 4.2.3.

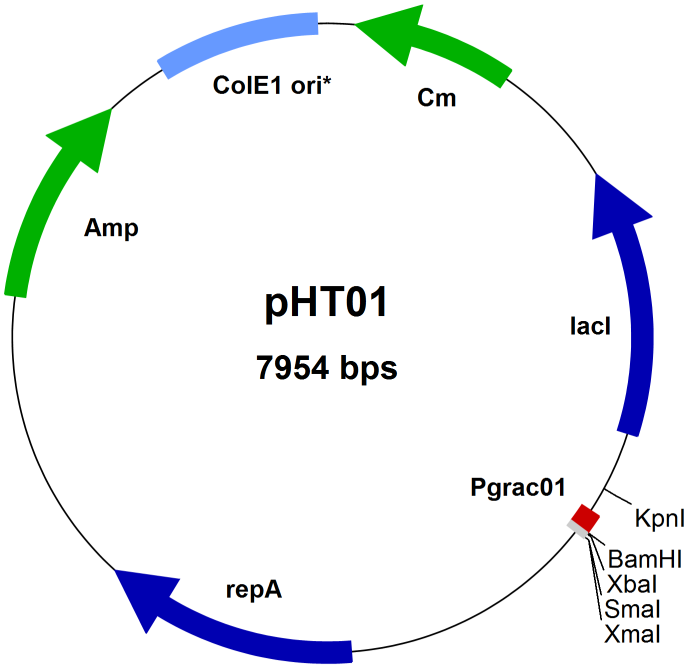

Figure 4.2.3 Plasmid map of pHT01

Plasmid map of pHT01 (Source: MoBiTec GmbH[4])

4.2.2 Constructs for isolated testing

The input and output devices of the AND gate are individual transcription units (TU) that work together to make an overall system producing an output, expressing mTagBFP. Before assembling the system, each TU was characterised to ensure that it operated as expected. Characterisation allows for optimising the TU before continuing, for example by trying different RBS or promoters if regulation needs improvement.

To allow for individual characterisation of each TU, a testing approach was defined during the design stage. Input 1 produces the amber suppressor tRNA, SupD, via transcription which can be identified by various techniques such as Northern blotting, RT-qPCR, or RNA-seq. Input 2 encodes *T7pTag*; because this produces a protein, the read-through of the promoter can be exploited to produce an additional

[4] MoBiTec GmbH, Lotzestraße 22a, D-37083 Goettingen, Germany

protein which is easier to detect. mCherry2 was added to the test construct for Input 2 so that this device can be characterised by measuring red fluorescence.

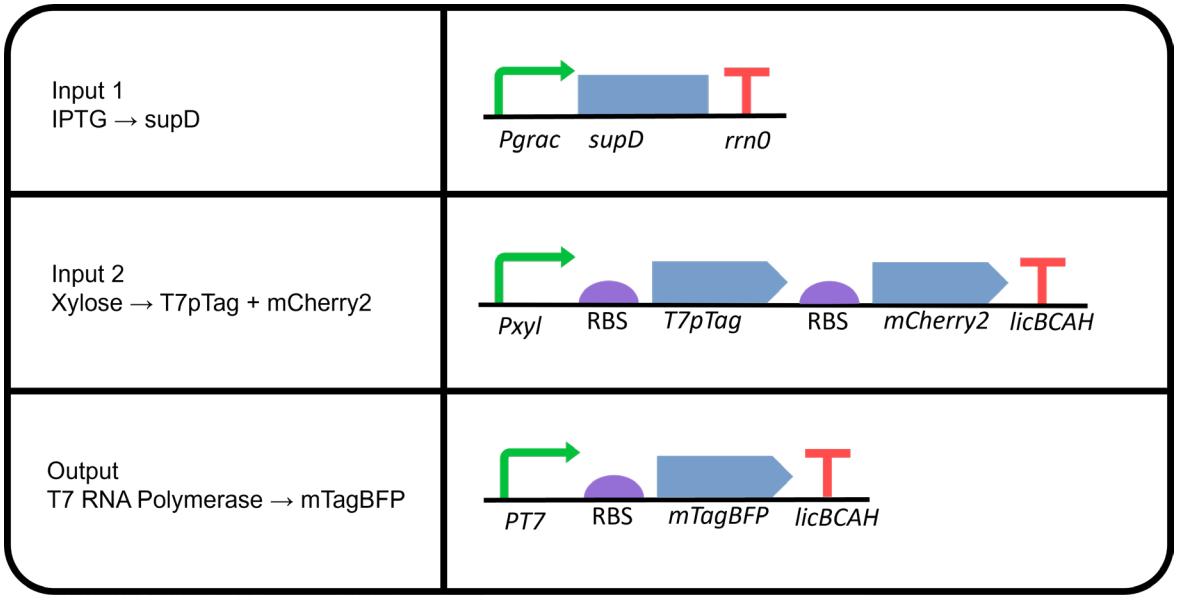

Figure 4.2.4 Visual representations of AND gate test constructs
Promoter parts are represented as green arrows. RBS parts are represented as purple semi-circles. Coding sequences are represented as blue polygons. Terminator parts are represented as red "T" shapes.

4.2.3 AND gate cloning strategy

Input 2 was built using two fragment Gibson assembly (see Chapter 7.2.2 – Methods) at the *Eco*RI insertion site of pHT01 (Gibson *et al.*, 2009). DNA fragments were synthesised by Integrated DNA Technologies[5]. The insert constructs of Input 1 and Output were considered too complex for standard DNA synthesis, so the cloning of these constructs was outsourced to Twist Bioscience[6]. The resultant plasmids were named pInput1, pInput2 and pOutput (Table 4.2.1).

Additional expression devices for the repressors LacI and XylR were synthesised for pInput1 and pInput2, respectively. The repressor expression devices were designed to be inserted on the reverse strand (relative to the input construct) to prevent read-through and were the repressor genes, *lacI* or *xylR,* constitutively expressed by P_{veg}. Addition of the LacI expression device was successful for Input

[5] Integrated DNA Technologies, Inc., 1710 Commercial Park, Coralville, Iowa 52241, USA
[6] Twist Bioscience HQ, 681 Gateway Blvd, South San Francisco, CA 94080, USA

1, however, addition of XylR expression to Input 2 could not be achieved. It was envisaged that a lack of XylR may result in leaky expression for Input 2, especially in *E. coli* which does not naturally contain the same *xyl* operon as found in *B. subtilis*.

Name	Vector	Restriction Site	Insertions	Deletions
pInput1	pHT01	EcoRI	P_{grac} → *supD* P_{veg} → *lacI*	P_{grac} → *lacI*
pInput2	pHT01	EcoRI	P_{xyl} → *T7pTag + mCherry2*	P_{grac} → *lacI*
pOutput	pHT01	EcoRI	P_{T7} → *mTagBFP*	P_{grac} → *lacI*

Table 4.2.1 Plasmids created for AND gate
Table listing the modifications made to produce the plasmids used in the AND gate study. The EcoRI restriction site was used to remove the P_{grac} → *lacI* sequence in pHT01 and replace it with the insert needed for each construct.

Methods relating to cloning are found in the following sections of Chapter 8: Gibson assembly (7.2.2); transformation of chemically competent *E. coli* cells (7.3.2); transformation of *B. subtilis* (7.3.3); preparation of overnight culture (7.1.2); extraction of plasmid from *E. coli* (7.4.1). Screening of constructs was conducted by restriction digest (7.5.1) and checking the size and number of fragments by DNA gel electrophoresis (7.5.2). Constructs were sequence validated using Sanger sequencing at Eurofins Genomics[7]. Sequence validated constructs were stored as glycerol stocks (7.1.4) at -80ºC and streaked onto LB agar plates for cultures prior to experiments.

[7] Eurofins Genomics Europe Shared Services GmbH, Anzinger Str. 7a, 85560 Ebersberg, Germany

Name	Genetic Parts (including plasmid backbone)	Host	Experiments
Input 1 + LacI	(P_{grac}, *supD*, *rrn0* term), (P_{veg}, *lacI*), pHT01: *CmR, AmpR, ColE1* ori	*E. coli* DH5α	RNA-seq. Section 4.3.3
		E. coli S30 extract	Cell free fluorescence assay. Section 4.6.3
Input 1	(P_{grac}, *supD*, *rrn0* term), pHT01: *CmR, AmpR, ColE1* ori	*E. coli* DH5α	N/A
Input 2 + XylR	(P_{xyl}, rbs, *T7pTag, mCherry2,* rbs, *licBCAH* term), (P_{veg}, *xylR*), pHT01: *CmR, AmpR, ColE1* ori	N/A	Cloning unsuccessful
Input 2	(P_{xyl}, rbs, *T7pTag, mCherry2,* rbs, *licBCAH* term), pHT01: *CmR, AmpR, ColE1* ori	*E. coli* DH5α	RFP assay. Section 4.4.5
		B. subtilis 168	RFP assay. Section 4.4.4
		E. coli S30 extract	Cell free fluorescence assay. Section 4.6.3
Output	(P_{T7}, rbs, *mTagBFP, licBCAH* term), pHT01: *CmR, AmpR, ColE1* ori	*E. coli* DH5α	BFP assay. Section 4.5.3
		E. coli BL21(DE3)	BFP assay. Section 4.5.3
		E. coli S30 extract	Cell free fluorescence assay. Section 4.6.3

Table 4.2.2 Summary of constructs and strains created for AND gate
Table summarising all constructs designed for this chapter, which host strains they were cloned into, and for which experiments those strains were used.

4.3 Confirming the presence of SupD tRNA in induced cells

4.3.1 Introduction

The test construct for input 1 is designed to produce SupD tRNA following induction by IPTG. RNA-seq was employed to confirm the presence of SupD tRNA in an induced sample of the input 1 strain. As a preliminary check, only one biological replicate for each sample was used for this study. In future work, to fully characterise the input 1 construct, at least three biological replicates would be required for statistical significance and a gradient of inducer concentrations should be tested.

The input 1 test construct could have been characterised in alternative ways to RNA-seq, such as with fluorescent RNA aptamers or qPCR methodologies (Ruijter *et al.*, 2013; Corchete *et al.*, 2020). RT-qPCR would be an effective approach at characterising the presence of SupD tRNA and was considered as an approach, however, RNA-seq was eventually chosen due to familiarity with the methods and the richness of the data. Data from RT-qPCR requires normalisation against a set of house-keeping genes which are known to be expressed at stable levels regardless of environmental conditions, such as 16S rRNA genes (Johnson *et al.*, 2019; Casas *et al.*, 2022). As such, RT-qPCR experiments require many samples with different probes for each sequence to be analysed, whereas, RNA-seq can quantify the total RNA within the entire sample (Grätz *et al.*, 2022).

RNA-seq reads were mapped against the *E. coli* reference genome (EB1) which contains the following genes of interest: *serU* and *lacI*. *serU* is the gene found in *E. coli* that encodes SupD tRNA. If SupD is present in the sample, it is expected that transcript counts will be mapped against the *serU* gene. *lacI* is expressed on pInput1 by constitutive expression, so it is expected to find transcript counts in abundance mapped against the *lacI* gene.

E. coli DH5α was used as a control sample and two samples of *E. coli* DH5α with pInput1 were used as test samples, one induced with IPTG and one without IPTG. It is expected that the *serU* gene will see the most expression in the IPTG induced test sample; there may be some leaky expression of *serU* observed in the non-induced test sample and some expression in the control sample as *serU* is located on the chromosome. Transcript counts for *lacI* are expected to be high in the test

samples as this is constitutively expressed on pInput1, and lower in the control sample as this is located on the chromosome only.

4.3.2 Results

Methods for this experiment are documented in chapter 7, section 7.8.1. The results of these experiments are outlined in this section.

Gene	Transcript Counts			Fold Change		
	Control	Non-induced	Induced	Non-induced / Control	Induced / Control	Induced / Non-induced
lacI	1148	556296	433149	484.58	377.31	0.78
serU	7	27	160.939	3.86	22.99	5.96
rrfF	0.143	3	2.5	20.98	17.48	0.83
rrfH	0.143	0	2.5	0.00	17.48	--
yihU	123	700	1805	5.69	14.67	2.58
frmR	626	2573	7294	4.11	11.65	2.83

Table 4.3.1 Most differentially expressed genes between induced and control samples
Transcript counts generated by STAR 2.6.1d (Dobin *et al.*, 2013) and Salmon 1.10.1 (Patro *et al.*, 2017) through the nf-core/rnaseq pipeline (Ewels *et al.*, 2020). Control represents the sample of *E. coli* DH5α without a plasmid; non-induced represents the sample of *E. coli* DH5α with pInput1; induced represents the sample of *E. coli* DH5α with pInput1 and 1 mM IPTG added during the exponential growth phase. Fold change is calculated by dividing the transcript of the leading sample by the transcript counts of the following sample. Genes with a fold change of at least ten-fold between Induced and Control are displayed in the table and ranked from highest to lowest.

Due to the lack of biological replicates, statistical significance can't be drawn from the observed results (Table 4.3.1), however, the RNA-seq analysis does show the expected results. *lacI* shows a several hundred-fold increase between the samples with pInput1 and the control group with no plasmid. *serU* shows a ~23-fold increase in the induced sample, a ~six-fold increase between non-induced and induced input 1 samples, and a ~four-fold increase between the non-induced input 1 sample and the control sample.

Genes *rrfF* and *rrfH* show a greater than ten-fold increase between the induced input 1 and the control group, however, the transcript counts are extremely low (≤3) so are not likely to be significant. Genes *yihU* and *frmR* show fold changes of ~15 and ~12, respectively.

Despite the lack of biological replicates, it seems apparent that pInput1 is present in the samples provided for sequencing. *lacI* is successfully being expressed in large amounts by the P_{veg} promoter. The *serU* gene is also being expressed by the P_{grac} promoter and there seems to be tight expression due to the much lower fold change between non-induced and control than that between induced and control. Expression of the *serU* gene suggest that SupD tRNA is present after the construct is induced.

4.4 Characterization of Input 2: (D)-xylose → mCherry2

4.4.1 Introduction

The test construct for input 2 was designed to co-express mCherry2 alongside the *T7pTag* CDS under induction by (D)-xylose[8]. As the *xylR* repressor gene is not found in *E. coli* DH5α, poorly regulated expression in *E. coli* DH5α was expected. Leaky expression in *B. subtilis* 168 is likely due to the scarcity of XylR expressed from the chromosome relative to the copy number of pInput2. The input 2 test construct was characterised in *B. subtilis* 168 and *E. coli* DH5α by measuring red fluorescence under induction by (D)-xylose. It can be inferred that if RFP fluorescence is greater, more *T7pTag* is being expressed because of its co-expression with mCherry2.

4.4.2 Results: Bacillus subtilis 168

Methods for this experiment are documented in chapter 7, section 7.8.2. The results of these experiments are outlined in this section.

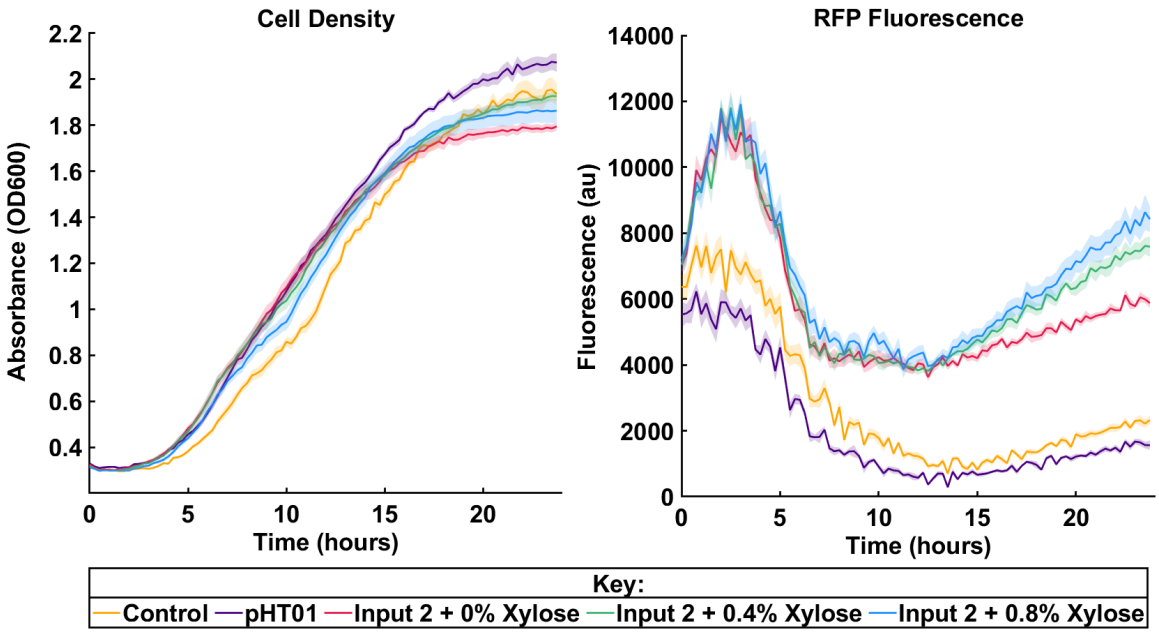

Figure 4.4.1 RFP fluorescence from Input 2 in Bacillus subtilis 168
Plot (a) shows blank corrected absorbance at 600 nm for each strain at each concentration of (D)-xylose. Plot (b) shows fluorescence measured at emission 561 nm (band 20) and excitation 610 nm (band 20); values were blank corrected, and

[8] All (D)-xylose solutions were prepared as w/v

growth corrected. Solid lines represent the mean average of three replicates; shaded translucent regions represent the standard error.

All groups of *B. subtilis* 168 exhibited similar patterns in growth curve (shown in plot (a) of figure 4.4.1) suggesting that the presence of a plasmid, antibiotic, nor the concentration of (D)-xylose had a significant impact on cell growth. Fluorescence measurements in plot (b) of figure 4.4.1 start off with substantial noise which lessens at approximately hour ten. After hour ten, a clear trend can be seen that the strains containing pInput2 are exhibiting red fluorescence and the control strains are not. There is an increase in fluorescence between the non-induced input 2 group and the induced input 2 groups of approximately 30% toward the end of the experiment at hour 24.

4.4.3 Results: Escherichia coli DH5α

Methods for this experiment are documented in chapter 7, section 7.8.3. The results of these experiments are outlined in this section.

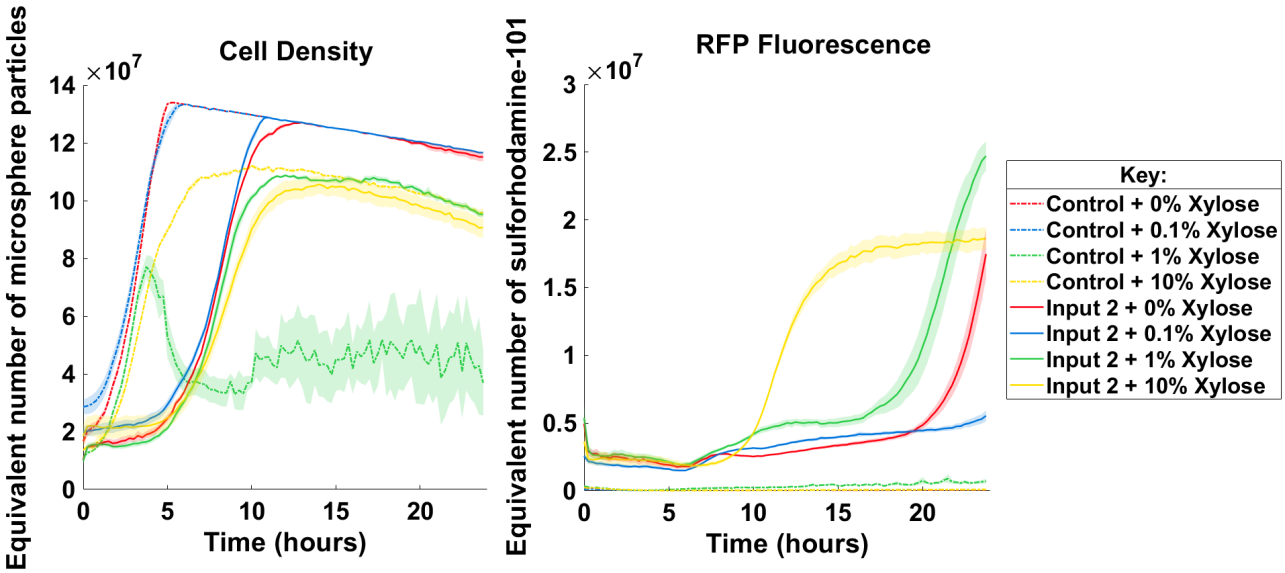

Figure 4.4.2 RFP fluorescence from Input 2 in Escherichia coli DH5α
Plot (a) shows blank corrected absorbance at 600 nm for each strain at each concentration of (D)-xylose, calibrated to equivalent number of microsphere particles. Plot (b) shows fluorescence measured at emission 561 nm (band 20) and excitation 610 nm (band 20); values were blank corrected, growth corrected and calibrated to equivalent molecules of calibrant: sulforhodamine-101. Solid lines represent the mean average of 5 replicates; shaded translucent regions represent the standard error.

The growth curves show in plot (a) of figure 4.4.2 show some differences between groups. There is a delay in exponential growth between the control groups and the input 2 groups; the control groups started exponential growth between hour zero and hour five, whereas the input 2 groups started exponential growth between hour four and hour 12. Groups at 1% or 10% (D)-xylose started stationary phase with a lower cell density than groups at 0% and 0.1% (D)-xylose. The group “Control + 1% Xylose” exhibited very atypical behaviour after hour four of the experiment; this group declined in cell density rapidly and then levelled out with very noisy measurements for the remainder of the experiment. Replicates within the same group neighbour one another in the plate, so it is possible that something happened in this region of the plate that caused this behaviour. Possible causes could be some condensation forming on the plate seal at this location that obfuscated the readings, or some antibiotic was unintentionally added to these wells causing erratic cell growth and death patterns.

The fluorescence measurements shown in plot (b) of figure 4.4.2 show a significant difference in fluorescence between the control groups and the input 2 groups starting at hour zero. The atypical behaviour in “Control + 1% Xylose” can be observed in the fluorescence measurements also; this further supports the hypothesis that something on the plate was obfuscating the measurements in this locality. “Input 2 + 10% Xylose” shows a sigmoidal curve in red fluorescence which plateaus after hour fifteen. “Input 2 + 1% Xylose” begins to exponentially increase fluorescence after hour seventeen, and “Input 2 + 0% Xylose” begins to exponentially increase fluorescence after hour twenty. “Input 2 + 0.1% Xylose” does not increase exponentially within the 24 hours of this experiment. If this experiment was repeated for 48 hours or longer, the trend for all these groups could be observed in full; it is possible that all of them exhibit a sigmoidal curve.

4.5 Characterization of the Output construct in *E. coli* BL21(DE3)

4.5.1 Introduction

pOutput was designed to produce mTagBFP in the presence of T7 RNAP. T7 RNAP is designed to be the product when both inputs are turned on; however, for testing output in isolation, T7 RNAP can be introduced by using an engineered strain. *E. coli* BL21(DE3) is a frequently used protein expression strain which produces T7 RNAP under induction by IPTG, thus, works ideally as a testing strain for pOutput.

T7 expression systems do not occur naturally in *B. subtilis,* though many efforts have been made to engineer a T7 expression system in *B. subtilis* (Conrad et al., 1996; Ji et al., 2021; Ye et al., 2022). These efforts have proved difficult in previous studies due to the interference of the native RNAP of *B. subtilis* (Ye et al., 2022; Conrad et al., 1996). Although many researchers have engineered *B. subtilis* to produce T7 RNAP, there are no commercially available strains. Due to difficulties in acquiring a T7 RNAP expression system for *B. subtilis*, it was decided to conduct this work in *E. coli* alone. Once the expression system of the input devices has been shown to work in *E. coli*, this could be ported to *B. subtilis* 168 where it could behave both as an AND gate and a T7 expression system.

Blue fluorescence was measured under IPTG induction of *E. coli* strains with pOutput. *E. coli* DH5α was used as a control group with no T7 RNAP production, and *E. coli* BL21(DE3) was used for testing with T7 RNAP present. For each strain (DH5α and BL21(DE3)), three treatment groups were tested: no plasmid as a negative BFP control, pOutput with IPTG as the induced test group, and pOutput without IPTG as the non-induced test group.

For the negative control groups without pOutput, no blue fluorescence should be observed as there is no CDS present encoding a blue fluorophore. In *E. coli* DH5α with pOutput, there may be some blue fluorescence due to leaky expression, but IPTG induction should not make a difference. In *E. coli* BL21(DE3), there should be blue fluorescence observed in both non-induced and induced groups, but fluorescence should grow more rapidly following IPTG induction. *E. coli* BL21(DE3) produces T7 RNAP under induction by IPTG, but it has been shown that delayed expression of T7 RNAP occurs with omission of IPTG (Zhang *et al.*, 2015).

4.5.2 Results

Methods for this experiment are documented in chapter 7, section 7.8.4. The results of these experiments are outlined in this section.

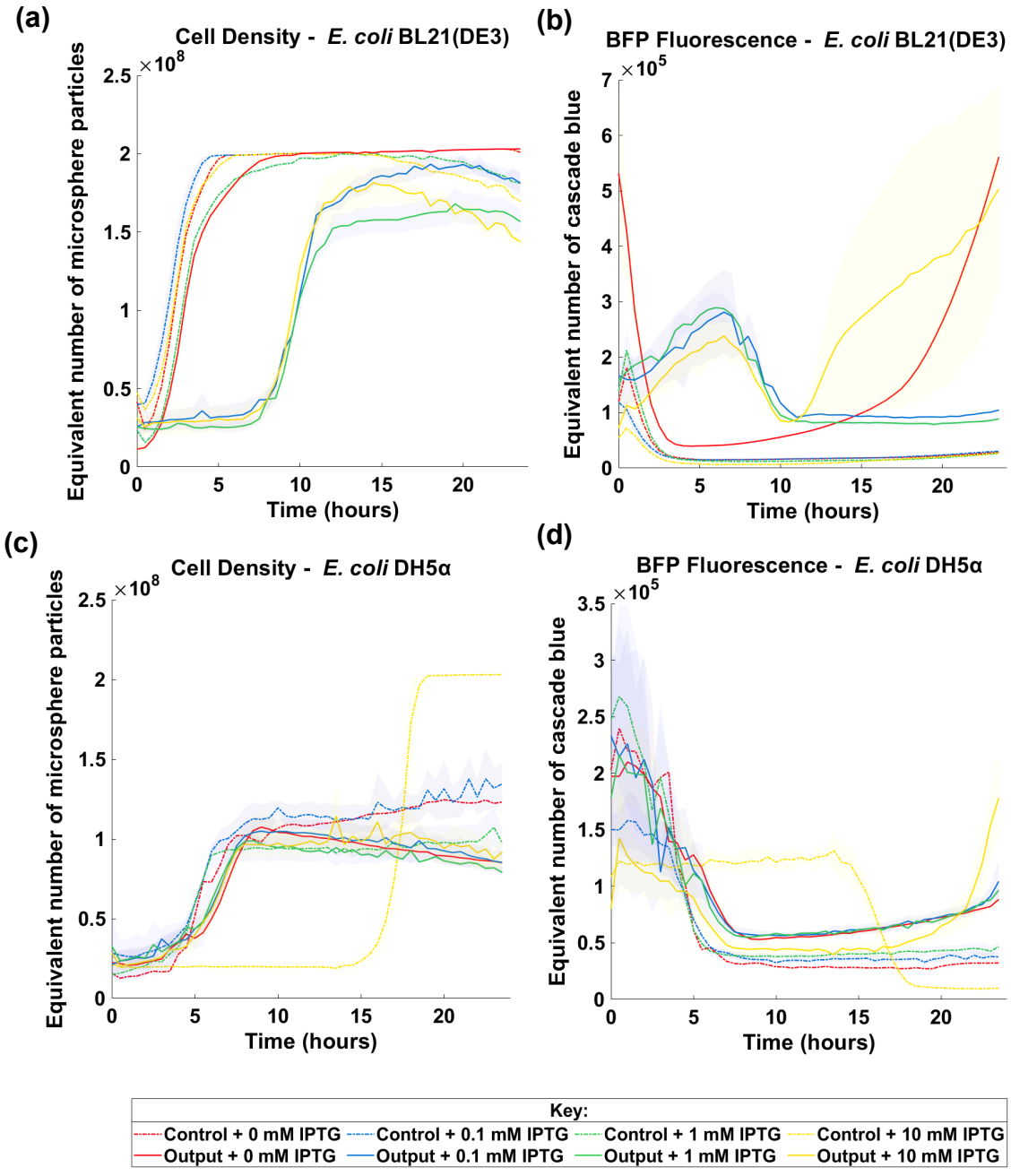

Figure 4.5.1 BFP fluorescence of Output – T7 → mTagBFP

Plots (a) and (c) shows blank corrected absorbance at 600 nm for each strain at each concentration of IPTG, calibrated to equivalent number of microsphere particles. Plots (b) and (d) shows fluorescence measured at emission 402 nm (band 20) and excitation 458 nm (band 30); values were blank corrected, growth corrected and calibrated to equivalent molecules of calibrant: cascade blue. Solid lines

represent the mean average of 3 replicates; shaded translucent regions represent the standard error.

The growth curves shown in plot a of figure 4.5.1 split into two groups of trends: the first group contains the controls and output without IPTG, and exhibit exponential growth within the first five hours of the experiment; the second group contains the output with IPTG induction, and exhibit exponential growth after the first five hours of the experiment. Plot (c) of figure 4.5.1 shows the growth curves for *E. coli* DH5α which all display the same trend with the exception of "Control + 10 mM IPTG". "Control + 10 mM IPTG" does not reach exponential phase until hour fifteen of the experiment, after which it reaches a cell density above 2×10^8 equivalent number of microsphere particles. The inverse trend is shown in the growth corrected fluorescence data in plot (d) of figure 4.5.1. The remaining groups in plot (c) of figure 4.5.1 start exponential growth earlier at approximately hour four of the experiment but reach a much lower cell density circa 1×10^8 equivalent number of microsphere particles.

The fluorescence measurements shown in plots (b) and (d) of figure 4.5.1 become less noisy after hour ten of the experiment. In plot (d) for *E. coli* DH5α, all control groups (except for "Control + 10 mM IPTG") show very low fluorescence values which can be attributed to background fluorescence. The output groups very gradually increase in fluorescence to low values, a trend most apparent in "Output + 10 mM IPTG" where the fluorescence sharply increases to approximately 1.8×10^5 equivalent molecules of cascade blue. This could suggest slight leaky expression in the T7 promoter. In plot (b) for *E. coli* BL21(DE3), all control groups show very low fluorescence, and all output groups show some increased blue fluorescence. "Output + 10 mM IPTG" steadily rises from ~1×10^5 to ~5×10^5 equivalent molecules of cascade blue, with a wide standard error, between hours ten and 24. "Output + 0 mM IPTG" shows a rapid increase in fluorescence at approximately hour 17 of the experiment when it increases from ~1.5×10^5 to ~5.5×10^5 equivalent molecules of cascade blue. Groups with 0.1 mM IPTG and 1 mM IPTG do not increase in fluorescence.

4.6 Testing the AND gate using *E. coli* S30 extract, cell free expression system

4.6.1 Introduction

To enable testing of the AND gate system without the need for additional cloning, the circular DNA was added to a cell free expression system (CFS). The key advantage of using a cell free system for this scenario is that the plasmids can be directly added to the solution rather than following a transformation process. It would not be possible to transform all three plasmids of the AND gate into a single organism as they utilise the same selection pressure of ampicillin resistance, so only a single plasmid would be maintained by the strain and the others discarded.

The "*E. coli* S30 Extract System for Circular DNA" from Promega was used for these experiments and the "*E. coli* T7 S30 Extract System for Circular DNA" from Promega was used in these experiments as a positive control system containing T7 RNA polymerase (Pratt *et al.*, 2004; Suzuki *et al.*, 2002).

The AND gate system was tested with all combinations of inducers to determine its functionality as an AND gate: IPTG + (D)-xylose (ON), IPTG only (OFF), (D)-xylose only (OFF), and neither inducer (OFF). If the system is working correctly, blue fluorescence would only be seen when the AND gate is in the ON position. However, it is likely that blue fluorescence would be seen in the other states as well but in low amounts which could then be reduced by optimisation of the system.

The control system (using T7 RNAP) was tested with the pOutput in isolation and compared with the *E. coli* S30 extract system without T7RNAP as a baseline. Additionally, pInput2 was tested in isolation with (D)-xylose to determine whether mCherry2 could be expressed within the *E. coli* S30 extract system.

4.6.2 Results

Methods for this experiment are documented in chapter 7, section 7.8.5. The results of these experiments are outlined in this section.

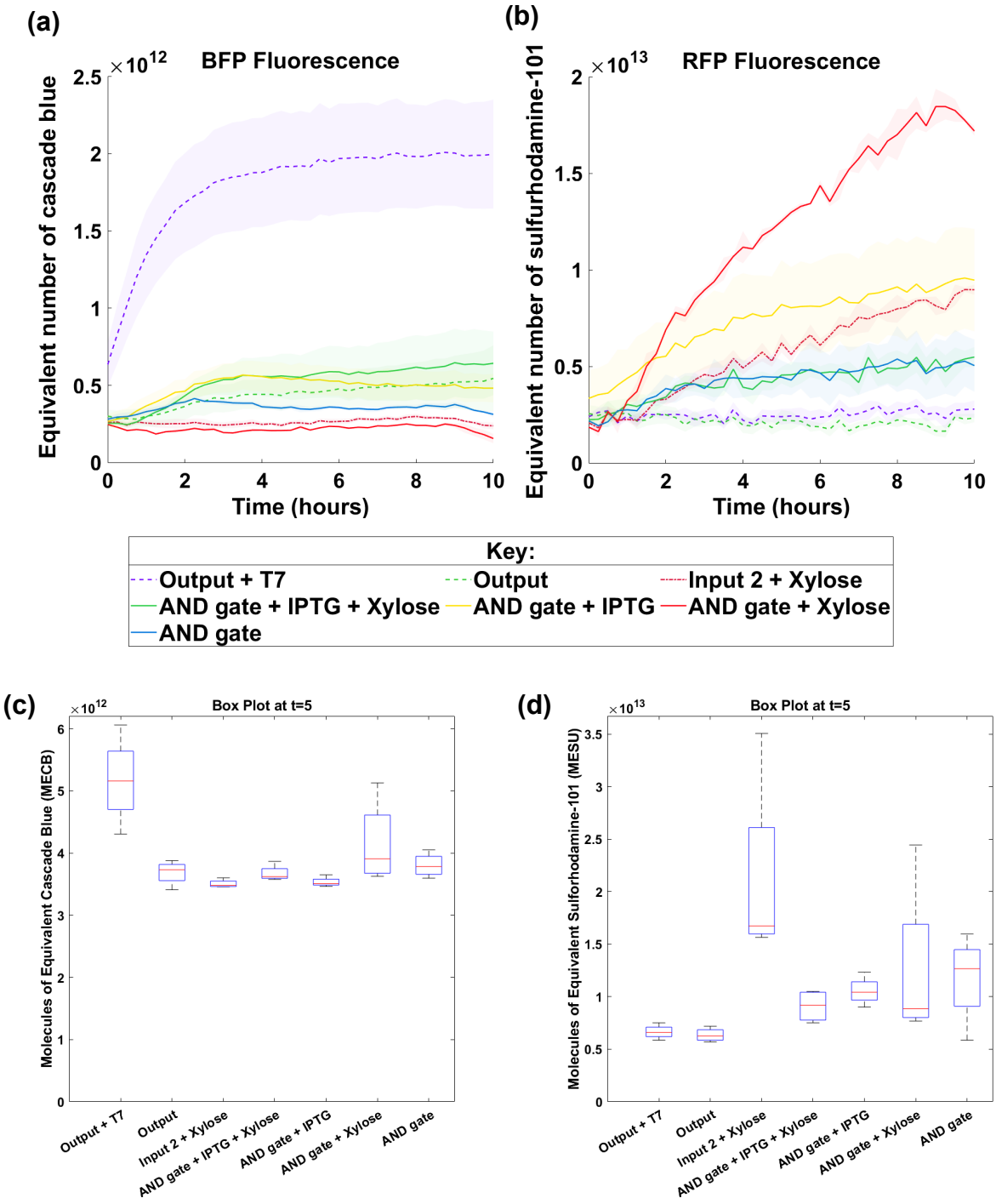

Figure 4.6.1 Fluorescence over time of AND gate in E. coli S30 CFS
Fluorescence measurements at: emission 402 nm (band 20) and excitation 458 nm (band 30) for BFP; emission 561 nm (band 20) and excitation 610 nm (band 20) for RFP. 10 µL *E. coli* S30 extract CFS was grown at 37°C for 10 hours with

measurements taken every 30 minutes, for each reaction. Circular DNA was added so that each plasmid had an amount of 1 pmol in each reaction. The experiment names are described as follows: “Output + T7” is pOutput in the T7 *E. coli* S30 extract CFS; “Output” is pOutput in the *E. coli* S30 extract CFS (without T7); “Input 2 + Xylose” is pInput2 with 1% (D)-xylose; “AND gate + IPTG + Xylose” is pInput1, pInput2 & pOutput with 1 mM IPTG and 1% (D)-xylose; “AND gate + IPTG” is the same as previous without (D)-xylose; “AND gate + Xylose” is the same as previous without IPTG but with (D)-xylose; “AND gate” is the same as previous without either inducer. All fluorescence data was blank corrected and calibrated to equivalent number of calibrant molecules: cascade blue for BFP and sulforhodamine-101 for RFP. Plots (a) and (b) show the mean average fluorescence for each reaction from four replicates. Standard error is represented by the shaded translucent regions for each line. Plots (c) and (d) are box plots showing the distribution of the data for each reaction at t=5 (the measurement at the fifth hour in the experiment).

The BFP fluorescence chart shown in plot (a) of figure 4.6.1 demonstrates that the group “Output + T7” fluoresces considerably more than the other groups. This suggests that the T7 RNA polymerase in the *E. coli* T7 S30 extract is activating P_{T7} to express *mTagBFP* in this system. The other groups have low fluorescence which can be attributed to background fluorescence and variation in the samples.

The RFP fluorescence chart shown in plot (b) of figure 4.6.1 demonstrates that the two Output groups have very low fluorescence values which is expected as they do not contain a CDS for a red fluorophore. The other groups all have varying levels of red fluorescence. “AND gate + Xylose” displays the greatest fluorescence, followed by “AND gate + IPTG” which has much variation between replicates which can be seen by the standard error or the distribution in the box plot in plot (d) of figure 4.6.1. “Input 2 + Xylose” was intended as a control group for positive RFP production; while it does show red fluorescence, it is in modest quantities relative to all the groups. “AND gate” and “AND gate + IPTG + Xylose”, which represent fully OFF and fully ON AND gate respectively, show very close measurements of red fluorescence and rank lowest out of all the groups containing a CDS for *mCherry2*.

4.6.3 Discussion

The results from assays in *E. coli* S30 extracts do not indicate that the AND gate is functional *in vitro* in *E. coli* S30 extract. There was no increased BFP production in the AND gate when in the “on” state, nor was there increased BFP production in any of the AND gate groups. Similarly, for RFP production, it was not displayed that RFP was produced in increased quantities when 1% (D)-xylose was present.

The positive control group, with T7 RNA polymerase present in the system, demonstrates that P_{T7} was functional in the system. Based on the measurements taken in this assay, it could not be determined whether P_{grac} works in the system as no measurements for the presence of SupD tRNA were made due to the cost and time restrictions that would be required for quantitative analysis of SupD. Based on the RFP measurements made in the assay, it seems possible that P_{xyl} does not work as it was expected under these conditions as the RFP measurements did not follow the expected behaviour of induction – presence of (D)-xylose equating to more RFP detected and absence of (D)-xylose equating to less RFP detected.

Before making conclusions that P_{xyl} does not work in *E. coli* S30 extract, more concentrations of (D)-xylose would need to be tested and for a greater duration than the ten hours used for the experiments of this work. The system was only measured for ten hours because after this time, the system begins to evaporate which obfuscates the measurements. Larger reaction volumes would allow for running this experiment for longer; ideally, reaction volumes should be 100 µL in a 96-well plate as this is known to work well for *in vivo* reactions presented in this chapter.

To postulate some reasons that P_{xyl} may not work in *E. coli* S30 extract, it is necessary to point out that cell-free systems do not contain the same materials as a living *E. coli* strain. *E. coli* S30 extract by Promega only supports commonly used promoters for *E. coli*. Both P_{xyl} and P_{grac} originate from *B. subtilis* so these may not function normally in a cell-free system.

The regulator for P_{xyl} is XylR, which was not present on pInput2. It was expected that this would lead to leaky expression of *T7pTag* and *mCherry2*. Additionally, as identified by Fages-Lartaud *et al*., *mCherry2* contains an internal RBS which produces a truncated RFP leading to unregulated RFP expression where this CDS is included (Fages-Lartaud *et al.*, 2022). Either reason could be the cause of the basal RFP expression seen in this assay which was seemingly unaffected by 1% (D)-xylose. Again, it is important to note that 1% (D)-xylose is not necessarily optimal for the system used in the assay and more concentrations need to be tested.

4.7 Discussion

Overall, the results presented in this chapter provide some initial characterisation data for the AND gate design. Input 2 demonstrates portability in its ability to produce mCherry2 in both *E. coli* DH5α and *B. subtilis* 168 without refactoring. Another RNA-seq experiment in *B. subtilis* 168 would provide data to show whether Input 1 is equally portable. Likewise, for Output which need testing in *B. subtilis* in the presence of T7 RNA polymerase. Although, the AND gate system did not work as expected in the assay within an *E. coli* S30 extract CFS, there are some clear directions for future work to follow up and get the system working.

The results from the RNA-seq experiment, presented in section 4.3, suggests that the input 1 construct is working as expected and the regulation of P_{grac} seems to be tight in *E. coli* DH5α. A more robust quantitative approach would be preferable for characterisation of the input 1 construct. More replicates should be included, as well as a gradient of concentrations of the inducer, IPTG. Understanding how IPTG concentration affects the input 1 device *in vivo* is important for optimising the whole AND gate downstream. Finally, the input 1 construct needs the same approach applied in *B. subtilis* 168 to confirm that it operates in both species.

Input 2 was characterised using coupled transcription/translation of *mCherry2*, assuming that the production of *T7pTag* is closely linked as they are on the same mRNA transcript. The results presented in section 4.4 suggest that input 2 operates in both *B. subtilis* 168 and *E. coli* DH5α, although only a small selection of inducer concentrations was selected. More experimentation could be done to optimise this system and the addition of XylR expression on the plasmid would be beneficial to tighten up the leaky expression that was observed.

Output exhibited blue fluorescence when subjected to T7 RNAP in *E. coli* BL21(DE3) strain (section 4.5) and in the *E. coli* T7 S30 extract CFS (section 4.6). Output needs to be tested in *B. subtilis* to determine whether the AND gate design will work in that species. As stated in section 4.5, developing a T7 RNAP expression system for *B. subtilis* is no easy task (Ye *et al.*, 2022). Therefore, testing of the output construct in *B. subtilis* might best be substituted with testing of the full AND gate system once it has been shown to work in *E. coli*.

Further testing in cell-free systems would be beneficial, however, further characterisation of the test constructs in isolation should be conducted prior to this to better understand the interaction between the promoters, inducers, and other regulatory elements (e.g., XylR). The cell-free systems may benefit from higher concentrations of DNA and inducer molecules in the system. Determining optimal conditions for these systems would require a lot of experimentation which could be assisted by machine learning techniques such as Design of Experiments (DoE) (Singleton *et al.*, 2019). DoE would be an interesting approach; however, the direction of testing may be better invested in cloning the system together as whole for *in vivo* testing. The benefit of *in vitro* testing is that the test constructs can still be used, thus, are easily changed before committing to constructing a new plasmid.

As shown by Anderson *et al*., the AND gate design is modular, so it is intended to swap the promoters with stress-specific promoters identified by ROTC. An issue with this design is that it restricts portability: *E. coli* and *B. subtilis* have different stress responses so will have different sets of biomarkers for a given stress response. However, minus the input promoters, the rest of the AND gate should be portable between both species and the work so far in this thesis has not suggested anything to the contrary.

The AND gate system did not function as expected within the experiments summarised in section 4.6. The reasons for which were discussed in 4.6.4 including the lack of regulation machinery, such as transcription factors, within a cell-free system. To fully explore the functionality of the AND gate system, the constructs within pInput1, pInput2, and pOutput should be combined within a single chassis so that the system can be tested in full as intended.

4.7.1 Summary and Conclusion

The findings of the experiments conducted throughout this chapter are briefly summarised below:

- pInput1 showed increased expression under induction in *E. coli* DH5α (Table 4.3.1)
- pInput2 showed increased expression under induction in *B. subtilis* 168 (Figure 4.4.1)
- pInput2 showed increased expression under induction in *E. coli* DH5α (Figure 4.4.2)
- pOutput showed increased expression under induction in *E. coli* BL21(DE3), but not in *E. coli* DH5α (Figure 4.5.1)
- pInput2 did not show increased expression under induction within *E. coli* S30 cell extract (Figure 4.6.1)
- pOutput showed increased expression in the presence of T7 RNAP within *E. coli* S30 cell extract (Figure 4.6.1)
- pInput1 + pInput2 + pOutput did not show increased expression within *E. coli* S30 cell extract (Figure 4.6.1)

The characterisation data collected within this chapter supports the functionality of the individual test constructs (pInput1, pInput2, pOutput) independently. Further work is required to fully characterise these constructs in both *E. coli* and *B. subtilis*. Furthermore, optimisations can be made to tighten up leaky expression, as observed with pInput2. The experiments conducted within *E. coli* S30 cell extract did not demonstrate functionality of the AND gate system. The constructs contained on pInput1, pInput2, and pOutput should be combined within a single chassis to test the AND gate system as a whole without the restrictions of a cell-free system.

CHAPTER 5

5 *in vivo* experimental validation of a cross-species codon optimisation algorithm, Chimera Evolve

5 *in vivo* experimental validation of a cross-species codon optimisation algorithm, Chimera Evolve

5.1 Introduction

Chimera Evolve is an evolutionary algorithm for codon optimisation designed to produce sequences optimised toward multiple target organisms. The algorithm can be run in one of two modes, minimum or weighted. Minimum mode optimises a sequence for each target organism and selects the fittest solution, whereas weighted mode allows the user to provide a set of weights to the algorithm which informs the algorithm how much bias to use toward each target organism. Results have shown that Chimera Evolve performs well *in silico*, however, it lacks testing to show how well the optimised sequences perform *in vivo* (Skelton *et al.*, 2020).

In this study, the coding sequences (CDS) of two fluorescent proteins were codon optimised using Chimera Evolve with a range of weights toward *E. coli* MG1655 and *B. subtilis* 168 as target organisms. The resultant CDSs were cloned into expression vectors and tested in *E. coli* DH5α. The codon optimised CDSs did not show any significant difference in fluorescence between themselves nor with the non-optimised CDS. The results do not support the hypothesis that CDSs optimised with a greater weight toward one species perform best in that species. However, further experimentation is required to satisfy that conclusion.

5.1.1 Contributions

The Chimera Evolve algorithm was developed by James Skelton (Skelton *et al.*, 2020). DNA synthesis was completed by Twist Bioscience.

5.1.2 Motivation

Allocation of translational resources is an important aspect of metabolic load stress (Boo, Ellis and Stan, 2019; Zhang *et al.*, 2022). When introducing heterologous DNA into an organism, it will create metabolic load on the cell from transcription, translation and other factors as discussed in Chapter 2. Translational resources within a cell are finite and need to be managed appropriately to prevent strain on the cell. If the usage of these resources is optimised, it allows for more resources to be available for other processes in the cell including further expression of heterologous proteins (Zhou *et al.*, 2016; Zegarra *et al.*, 2023).

5 *in vivo* experimental validation of a cross-species codon optimisation algorithm, Chimera Evolve

Codon usage bias (CUB) is the term used to describe the different frequencies of codons in the coding sequences of different species. Codon optimisation is the process of altering a CDS in-line with an organism's CUB. Codon optimised CDSs of heterologous proteins free up translational resources in the host organism and reduce the impact of metabolic load (Bahiri-Elitzur and Tuller, 2021; Terpe, 2006). As such, codon optimisation is a commonly used approach for maximising the yield of heterologous proteins in genetic engineering and related fields such as synthetic biology (Ghavim *et al.*, 2017; Lipinszki *et al.*, 2018; Parvathy, Udayasuriyan and Bhadana, 2022).

In synthetic biology and related fields, systems often need to be refactored when porting from one species to another. For example, expression devices for heterologous proteins may have the CDS optimised for each host species and contain different promoters for regulation. Systems with increased portability reduce the need to refactor when porting between species. The Portabolomics project aims to produce portable systems between multiple hosts (Krasnogor *et al.*, 2023). Chimera Evolve meets the aims of portability, as the user can optimise a single CDS for multiple organisms at the same time.

Chimera Evolve has proven effective at generating candidate sequences as shown by the comparison against the Chimera UGEM algorithm in the initial study by Skelton et al. Three coding sequences were selected for comparison in ARS score and the scores were like those generated by Chimera UGEM; in the case of CDS P42212, sequences generated by Chimera Evolve even had ARS scores surpassing those generated by Chimera UGEM despite Chimera Evolve optimising for two species at once. The authors suggest that Chimera Evolve explores a design space that is not considered by Chimera UGEM, thus is able to surpass what was initially thought to be a ceiling in ARS score (Skelton *et al.*, 2020). However, there is not currently experimental *in vivo* evidence to support the algorithm. As part of the investigation into metabolic load stress and portable systems between microbial systems, Chimera Evolve was tested *in vivo* to determine whether the resulting CDSs show better expression in the species they are optimised for.

5.1.3 Codon Optimisation

Early approaches to codon optimisation assumed that one amino acid is most optimally encoded by a single codon in each species. This led to the development of metrics such as the codon adaptation index (CAI) (Sharp and Li, 1987), the frequency of optimal codons (FOP) (Ikemura, 1981) and the relative codon adaptation (RCA) (Fox and Erill, 2010).

CAI is one of the most frequently used metrics in codon optimisation (Bahiri-Elitzur and Tuller, 2021). To calculate CAI, first the codon frequencies for all coding sequences in the target organism's genome are determined. The collated coding sequences from the genome of a given organism are referred to as the reference set. Each codon in the CDS to be optimised is given a frequency score based on how frequently that codon appears in the reference set. The CAI of a CDS is calculated by taking the mean average of all the frequency scores assigned to each codon in the sequence. Methods such as CAI have been shown to perform effectively but do not consider greater factors such as the secondary structure of the resulting mRNA molecule (Bahiri-Elitzur and Tuller, 2021; Zur and Tuller, 2014).

The Chimera Average Repetitive Substring (ARS) measure (Zur and Tuller, 2014) is calculated by taking substrings from the CDS and comparing those to a reference set of the target organism. Each codon in a CDS (S) is given a score (S_i^j), where j is the codon and i is the index of the starting position. Each codon score is the length of the longest substring that also appears in the reference set. ARS score is calculated as the sum of all codon scores divided by the length of the CDS.

$$\boldsymbol{ARS} = \frac{\sum S_i^j}{|\boldsymbol{S}|}$$

The Chimera ARS approach has been shown to predict gene expression more accurately than CAI for heterologous and endogenous proteins based on existing data from *E. coli* (Zur and Tuller, 2014). ARS score is employed by Chimera Evolve to produce optimal candidates for codon optimised sequences.

5.1.4 The Chimera Evolve algorithm

Chimera Evolve implements an evolutionary algorithm (EA) to explore the solution landscape in codon optimisation; a method which, to the author's knowledge, previous codon optimisation algorithms have not attempted. Chimera Evolve can optimise a CDS for multiple species at the same time by one of two modes, minimum or weighted mode. For either mode, a CDS is supplied to the algorithm alongside the reference sets of CDSs for each target organism. Minimum mode optimises the sequences with equal favour to each provided organism. In weighted mode, a weighting toward each of these species is provided and the algorithm optimises the given CDS relative to these provided weights.

ARS is used to assess a fitness score for use in the EA. Minimum mode calculates the ARS score relative to each reference set and uses the lowest of these scores as the fitness. Weighted mode calculates the ARS score in the same way but then each score is multiplied by its corresponding weight; the fitness is the normalised sum of the weighted ARS scores.

In Skelton *et al*., Chimera Evolve is presented alongside data used to parameterise the algorithm with default values for crossover, mutations, generations, and generation start size (Skelton *et al.*, 2020). Three proteins were chosen for codon optimisation with *E. coli* MG1655 and *B. subtilis* 168 as the target organisms. The results were compared against a similar algorithm employing ARS score, Chimera UGEM (Diament *et al.*, 2019). The scores of successful candidates were comparable and, in some cases, better than those generated by Chimera UGEM.

5.2 Experimental Design

5.2.1 Selection of genetic parts to build transcription units

For experimental validation of Chimera Evolve *in vivo*, two heterologous proteins were selected to be optimised by the algorithm. The resulting candidate CDSs were synthesised and cloned into a transcriptional unit (TU) with a constitutive promoter, ribosome binding site (RBS) and terminator. The TUs were then introduced into the target organisms that they were optimised for, and the resulting protein expression measured over time. Fluorescent proteins were selected for testing as the expression can be quantified by measuring fluorescence. For these experiments,

two CDSs encoding fluorescent proteins were optimised over a gradient of weights toward two model organisms, *E. coli* MG1655 and *B. subtilis* 168. The hypothesis is that the weight that is most in favour of one organism should be expressed most in that organism and vice versa.

The key limiting factor for testing the algorithm is the cost for DNA synthesis, so, it was decided to keep a modest number of tests. Five sets of weights were picked for Chimera Evolve weighted mode: (0, 1); (0.25, 0.75); (0.5, 0.5); (0.75, 0.25); (1, 0). Tuples for the weight arguments are provided in the fomrat (*a*, *b*), where *a* is the weight toward *B. subtilis* 168 and *b* is the weight toward *E. coli* MG1655. The wildtype sequence for each input CDS were used as controls.

Two fluorescent proteins were chosen for testing, mGreenLantern and mCherryM10L. mGreenLantern is a green fluorescent protein (GFP) based on EGFP which has been shown to have greater fluorescence intensity in eukaryotes (Campbell *et al.*, 2020). mCherryM10L is a modification of the red fluorescent protein (RFP), mCherry, with an amino acid substitution at position 10. The substitution is from methionine to leucine to remove an internal RBS that was identified by Fages-Lartaud *et al*., and they showed that the substitution had no significant impact on the protein (Fages-Lartaud *et al.*, 2022). Green and red fluorescence are easily measured and differentiated spectra, plus standards exist for calibration to equivalent number of calibrant molecules per cell (See methods 7.6). Well-characterised proteins were included in the tests as positive controls for fluorescence, superfolder GFP and mRFP (Pédelacq *et al.*, 2006; Campbell *et al.*, 2002).

Originally, combinatorial design was chosen as an approach for the testing process and to use Design of Experiments to explore the optimal pairing of promoter and RBS. Combinatorial design was removed from the scope of this project, but the promoters and RBSs were selected with this in mind. The promoters chosen are from the Anderson collection with a variety of strengths. RBSs with a range of strengths were selected from Salis *et al*.'s study into optimal spacers for ribosome binding sites (Salis, Mirsky and Voigt, 2009). For testing in *B. subtilis* 168, three additional constitutive promoters were selected from those best characterised on

the iGEM parts registry. The selection of promoters and RBS for this study are shown in Table 5.2.1 and Table 5.2.2, respectively.

As Chimera Evolve is a stochastic algorithm, it was run several times to obtain three CDS variants of each protein at each weighted threshold. This way, three replicates are obtained for each weight and CDS to be tested.

In total, the number of parts synthesised was forty-five: seven promoters - four from the Anderson collection, three for *B. subtilis* 168; four RBS parts with spacers identified in 'Automated design of synthetic ribosome binding sites to control protein expression' (Salis, Mirsky and Voigt, 2009); thirty-three CDS parts – three wildtype sequences for mGreenLantern, mCherryM10L & superfolder GFP, fifteen variants for mGreenLantern and fifteen variants for mCherryM10L; one terminator part – B0015.

Promoter Name	Strength	Intended Host	Sequence
J23101	Strong	*E. coli*	tttacagctagctcagtcctaggtattatgctagc
J23106*	Medium	*E. coli*	tttacggctagctcagtcctaggtatagtgctagc
J23117	Weak	*E. coli*	ttgacagctagctcagtcctagggattgtgctagc
J23119	Strong	*E. coli*	ttgacagctagctcagtcctaggtataatgctagc
P_{veg}*	Strong	*B. subtilis*	ggagttctgagaattggtatgccttataagtccaattaacagttgaaaa cctgcataggagagctatgcgggtttttttattttacataatgatacataatt taccgaaacttgcggaacataattgaggaatcatagaattttgtcaaa ataattttattgacaacgtcttattaacgttgatataatttaaattttatttgac aaaaatgggctcgtgttgtacaataaatgtagt
P_{liaG}	Weak	*B. subtilis*	agtcaatgtatgaatggatacgggatatgaatcaataagtacgtgaaa gagaaaagcaacccagatatgatagggaacttttctctttcttgttttaca ttgaatctttacaatcctattgatataatctaagctagtgtattttgcgtttaa tagt
P_{lepA}	Medium	*B. subtilis*	caaaaatcagaccagacaaaagcggcaaatgaataagcggaacg gggaaggatttgcggtcaagtccttcccttccgcacgtatcaattcgca agcttttcctttataatagaatgaatga

Table 5.2.1 Promoters selected for the expression devices
Promoters accompanied by an asterisk were sequence validated and have been used to generate data in Figure 5.3.1 or Figure 5.3.2.

RBS Name	Strength	Sequence
B0030	Strong	attaaagaggagaaa**ttaagc**
B0031	Weak	tcacacaggaaacc**ggttcg**
B0032*	Medium	tcacacaggaaag**gcctcg**
B0033	Weak	tcacacaggac**ggccgg**

Table 5.2.2 Ribosome binding sites and spacers selected for the expression devices
Spacer sequence of 6 base pairs is emphasised in bold. RBSs accompanied by an asterisk were sequence validated and have been used to generate data in Figure 5.3.1 and Figure 5.3.2.

5.2.2 CDS variants generated by Chimera Evolve

Chimera Evolve was run using the default parameter set (300 mutations, 100 crossover events, 1000 generations, and an initial generation size of 200) which was shown to result in the highest mean ARS scores out of all sets tested for either mode, minimum or weighted (Skelton *et al.*, 2020). The CDS for mGreenLantern and mCherryM10L were provided as the target CDS and reference sets were provided for *B. subtilis* 168 and *E. coli* MG1655. Each CDS was ran through Chimera Evolve for each weight tuple [(0,1), (0.25,0.75), (0.5,0.5), (0.75,0.25), (1,0)] at least thrice.

Each output sequence was uploaded to Benchling to check whether the sequence is legal, i.e., does not have restriction sites that will be used for cloning. If the sequence contained illegal restriction sites, the algorithm was run again until a legal sequence was produced. The process was repeated until three candidate sequences for each CDS at each weighting were produced.

The candidate sequences were then aligned against each other using Clustal Omega (Sievers and Higgins, 2021) to produce a percent identity matrix. If any of the sequences were found to be identical with another, the original CDS and weight tuples would be run through Chimera Evolve again until all sequences were discrete. The percent identity matrices are shown in Table 5.2.5 and Table 5.2.6.

A shorthand naming convention was created for the CDSs which follows the format of *Xw-r* where *X* denotes the CDS name (G = mGreenLantern, C = mCherryM10L, sf = superfolder GFP), *w* denotes the weight tuple and *r* denotes the replicate number. The weight tuples, *w*, are assigned digits as follows 1 = (0, 1), 2 = (0.25, 0.75), 3 = (0.5, 0.5), 4 = (0.75, 0.25), and 5 = (1, 0). Tuples for the weight arguments are provided in the format (*a*, *b*), where *a* is the weight toward *B. subtilis* 168 and *b* is the weight toward *E. coli* MG1655. Metadata relating to the CDSs including shorthand names are presented in Table 5.2.3 and 5.2.4.

Shorthand Name	CDS Name	Weight (*w*)	*B. subtilis* weight	*E. coli* weight	Weight Tuple	Replicate	Fitness Score	ARS score *E. coli*	ARS score *B. subtilis*
sf0*	superfolder GFP	0	--	--	--	--	3.176	3.303	3.176
G0*	mGreenLantern	0	--	--	--	--	3.038	3.163	3.038
G1*	mGreenLantern	1	0	1	(0,1)	1	3.874	3.874	3.159
G2*	mGreenLantern	2	0.25	0.75	(0.25,0.75)	1	3.729	3.862	3.331
G3*	mGreenLantern	3	0.5	0.5	(0.5,0.5)	1	3.626	3.657	3.594
G4*	mGreenLantern	4	0.75	0.25	(0.75,0.25)	1	3.712	3.326	3.841
G5*	mGreenLantern	5	1	0	(1,0)	1	3.782	3.113	3.782
C0	mCherry-M10L	0	--	--	--	--	3.199	3.352	3.199
C1	mCherry-M10L	1	0	1	(0,1)	1	3.843	3.843	3.186
C2	mCherry-M10L	2	0.25	0.75	(0.25,0.75)	1	3.714	3.818	3.403
C3	mCherry-M10L	3	0.5	0.5	(0.5,0.5)	1	3.636	3.691	3.581
C4	mCherry-M10L	4	0.75	0.25	(0.75,0.25)	1	3.725	3.318	3.860
C5	mCherry-M10L	5	1	0	(1,0)	1	3.831	3.165	3.831

Table 5.2.3 CDS variants – Replicate 1 + wildtypes
CDS variants and their attributes: Table 5.2.3 contains information relating to the wildtype sequences and the first stochastic replicate, Table 5.2.4 contains information relating to the other two stochastic replicates (none of which were sequence validated nor used for assays). CDSs accompanied by an asterisk were sequence validated and have been used to generate data in Figure 5.3.1 and Figure 5.3.2. Wildtype sequences (sf0, G0, C0) are the original unchanged CDS so the weight and replicate fields do not apply, and the fitness score was calculated as the minimum ARS score for either target organism. Fitness score for the CDS weighted variants is the value calculated by Chimera Evolve; the mean average of ARS scores multiplied by their weight. Chimera ARS Score has been calculated for each sequence with respect to *E. coli* MG1655 and *B. subtilis* 168 separately.

Shorthand Name	CDS Name	Weight (*w*)	*B. subtilis* weight	*E. coli* weight	Weight Tuple	Replicate	Fitness Score	ARS score *E. coli*	ARS score *B. subtilis*
G1-2	mGreenLantern	1	0	1	(0,1)	2	3.866	3.866	3.138
G2-2	mGreenLantern	2	0.25	0.75	(0.25,0.75)	2	3.715	3.845	3.326
G3-2	mGreenLantern	3	0.5	0.5	(0.5,0.5)	2	3.638	3.657	3.619
G4-2	mGreenLantern	4	0.75	0.25	(0.75,0.25)	2	3.714	3.322	3.845
G5-2	mGreenLantern	5	1	0	(1,0)	2	3.820	3.117	3.820
G1-3	mGreenLantern	1	0	1	(0,1)	3	3.879	3.879	3.201
G2-3	mGreenLantern	2	0.25	0.75	(0.25,0.75)	3	3.708	3.841	3.310
G3-3	mGreenLantern	3	0.5	0.5	(0.5,0.5)	3	3.628	3.669	3.586
G4-3	mGreenLantern	4	0.75	0.25	(0.75,0.25)	3	3.694	3.314	3.820
G5-3	mGreenLantern	5	1	0	(1,0)	3	3.824	3.176	3.824
C1-2	mCherry-M10L	1	0	1	(0,1)	2	3.881	3.881	3.233
C2-2	mCherry-M10L	2	0.25	0.75	(0.25,0.75)	2	3.727	3.835	3.403
C3-2	mCherry-M10L	3	0.5	0.5	(0.5,0.5)	2	3.636	3.699	3.572
C4-2	mCherry-M10L	4	0.75	0.25	(0.75,0.25)	2	3.713	3.347	3.835
C5-2	mCherry-M10L	5	1	0	(1,0)	2	3.826	3.089	3.826
C1-3	mCherry-M10L	1	0	1	(0,1)	3	3.881	3.881	3.186
C2-3	mCherry-M10L	2	0.25	0.75	(0.25,0.75)	3	3.725	3.843	3.369
C3-3	mCherry-M10L	3	0.5	0.5	(0.5,0.5)	3	3.638	3.648	3.627
C4-3	mCherry-M10L	4	0.75	0.25	(0.75,0.25)	3	3.724	3.314	3.860
C5-3	mCherry-M10L	5	1	0	(1,0)	3	3.877	3.161	3.877

Table 5.2.4 CDS variants – Replicate 2 + Replicate 3
CDS variants and their attributes: Table 5.2.3 contains information relating to the wildtype sequences and the first stochastic replicate, Table 5.2.4 contains information relating to the other two stochastic replicates (none of which were sequence validated nor used for assays).

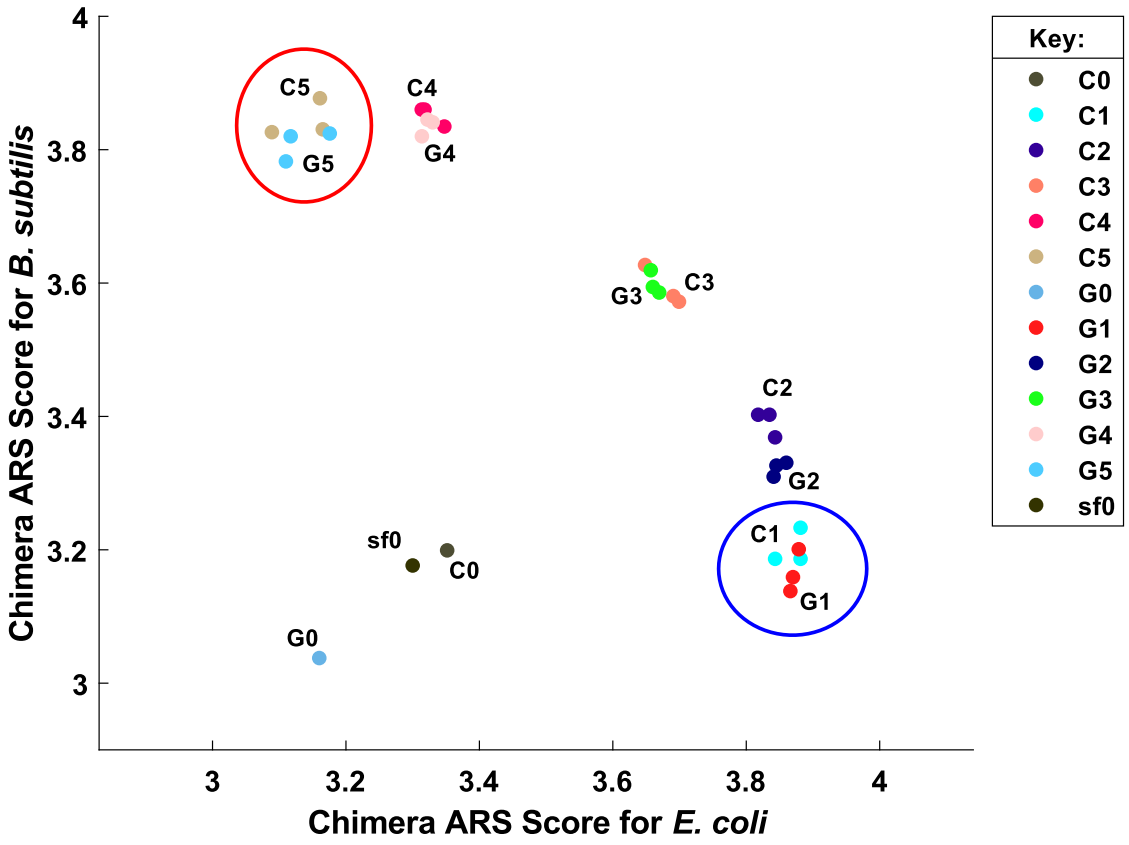

Figure 5.2.1 Distribution of Chimera ARS scores
Chimera ARS scores for each CDS variant with *E. coli* MG1655 and *B. subtilis* 168 as target organisms, separately. C0, G0 and sf0 are the wildtype sequences for mCherryM10L, mGreenLantern and superfolder GFP, respectively. G1-G5 and C1-C5 are Chimera Evolve generated variants with the weightings G1/C1 = (0,1), G2/C2 = (0.25,0.75), G3/C3 = (0.5,0.5), G4/C4 = (0.75,0.25), G5/C5 = (1,0). The blue ellipse encapsulates the values of G1 & C1; the red ellipse encapsulates the values of G5 & C5.

Chimera ARS scores were calculated for each CDS variant with *E. coli* MG1655 and *B. subtilis* 168 as target organisms, separately. Clusters can be seen for the variants with the same weight parameters applied i.e., C1 and G1 cluster together with the greatest bias toward *E. coli* MG1655 as shown by the blue ellipse. Based on the ARS scores alone, it is expected that C1 and G1 would be most expressed in *E. coli* MG1655 and C5 and G5 would be expressed most in *B. subtilis* 168.

Percent Identity

75 80 85 90 95 100

	G0	G3-2	G3	G1-3	G2-3	G1-2	G3-3	G1	G2	G2-2	G5	G4-3	G5-2	G4	G4-2	G5-3
G0	100	79.05	79.45	81.69	80.9	81.82	79.05	82.21	81.42	79.05	76.94	76.42	75.89	77.73	76.15	77.47
G3-2	79.05	100	87.88	87.62	86.69	86.17	89.46	86.56	88.14	87.75	82.21	85.64	83.4	86.3	87.62	84.32
G3	79.45	87.88	100	86.3	89.46	85.38	88.41	85.64	87.88	85.9	83.79	85.77	84.19	85.9	86.69	85.11
G1-3	81.69	87.62	86.3	100	86.43	88.41	85.24	89.06	89.06	86.69	79.71	82.08	81.82	81.95	82.48	80.5
G2-3	80.9	86.69	89.46	86.43	100	85.77	87.09	86.82	87.62	89.46	82.21	83.4	81.16	84.45	84.58	82.74
G1-2	81.82	86.17	85.38	88.41	85.77	100	84.06	89.46	88.67	86.96	79.97	80.76	81.03	82.61	82.48	81.16
G3-3	79.05	89.46	88.41	85.24	87.09	84.06	100	85.64	89.72	88.27	83.4	85.24	83.93	86.17	89.2	84.45
G1	82.21	86.56	85.64	89.06	86.82	89.46	85.64	100	89.99	87.88	80.11	81.95	79.84	82.08	82.08	81.55
G2	81.42	88.14	87.88	89.06	87.62	88.67	89.72	89.99	100	90.78	81.29	83.4	81.55	84.98	83.93	82.48
G2-2	79.05	87.75	85.9	86.69	89.46	86.96	88.27	87.88	90.78	100	82.08	85.38	82.87	85.51	86.03	83.4
G5	76.94	82.21	83.79	79.71	82.21	79.97	83.4	80.11	81.29	82.08	100	84.32	86.17	87.09	87.35	87.35
G4-3	76.42	85.64	85.77	82.08	83.4	80.76	85.24	81.95	83.4	85.38	84.32	100	87.35	85.64	87.35	86.82
G5-2	75.89	83.4	84.19	81.82	81.16	81.03	83.93	79.84	81.55	82.87	86.17	87.35	100	86.56	86.96	87.75
G4	77.73	86.3	85.9	81.95	84.45	82.61	86.17	82.08	84.98	85.51	87.09	85.64	86.56	100	89.99	88.01
G4-2	76.15	87.62	86.69	82.48	84.58	82.48	89.2	82.08	83.93	86.03	87.35	87.35	86.96	89.99	100	89.86
G5-3	77.47	84.32	85.11	80.5	82.74	81.16	84.45	81.55	82.48	83.4	87.35	86.82	87.75	88.01	89.86	100

Table 5.2.5 DNA sequence Percent Identity Matrix for mGreenLantern GFP codon variants
Percent Identity Matrix created by Clustal2.1, of all mGreenLantern CDS variants generated by Chimera Evolve. A value of 100% means that the sequences are identical. The shorthand names for CDS variants can be cross-referenced using Table 5.2.3 and 5.2.4; the format follows the nomenclature of *Gw-r* where *G* denotes the CDS name (mGreenLantern), *w* denotes the weight tuple and *r* denotes the replicate number.

Percent Identity: 75 80 85 90 95 100

	C5-2	C4	C5-3	C5	C4-2	C4-3	C0	C3	C3-3	C1	C3-2	C2-3	C2	C2-2		C1-2	C1-3
C5-2	100	82.8	85.87	84.4	82.93	85.47	78.8	82.67	80.13	81.73	80.53	80.8	80.13	79.87		79.2	80.8
C4	82.8	100	87.87	84.4	85.73	84.93	82.67	84.93	85.07	81.6	84.53	84	84.8	82.13		82.4	81.6
C5-3	85.87	87.87	100	85.47	85.47	88.13	81.2	83.87	83.33	80.8	82.8	82.8	82	80.13		81.6	81.07
C5	84.4	84.4	85.47	100	87.73	85.73	81.47	83.6	83.87	81.07	83.47	80.93	82	81.47		81.73	80.53
C4-2	82.93	85.73	85.47	87.73	100	89.87	81.73	85.33	86.53	82.8	86.4	85.87	84.8	83.87		83.2	80.93
C4-3	85.47	84.93	88.13	85.73	89.87	100	80.4	84.8	85.07	82.53	85.2	84.93	83.07	83.73		81.6	81.07
C0	78.8	82.67	81.2	81.47	81.73	80.4	100	83.33	83.6	83.2	84.27	82.8	82.8	81.73		84.53	84
C3	82.67	84.93	83.87	83.6	85.33	84.8	83.33	100	87.33	86.53	89.07	87.47	88.13	87.33		86.93	86.4
C3-3	80.13	85.07	83.33	83.87	86.53	85.07	83.6	87.33	100	85.07	89.73	87.87	86.27	86.4		86.67	84.27
C1	81.73	81.6	80.8	81.07	82.8	82.53	83.2	86.53	85.07	100	85.2	89.73	88.27	88		88.67	89.33
C3-2	80.53	84.53	82.8	83.47	86.4	85.2	84.27	89.07	89.73	85.2	100	89.87	88.53	88.67		87.73	86.8
C2-3	80.8	84	82.8	80.93	85.87	84.93	82.8	87.47	87.87	89.73	89.87	100	92.67	91.2		90.93	89.87
C2	80.13	84.8	82	82	84.8	83.07	82.8	88.13	86.27	88.27	88.53	92.67	100	92.8		89.47	88.27
C2-2	79.87	82.13	80.13	81.47	83.87	83.73	81.73	87.33	86.4	88	88.67	91.2	92.8	100		90.13	90.27
C1-2	79.2	82.4	81.6	81.73	83.2	81.6	84.53	86.93	86.67	88.67	87.73	90.93	89.47	90.13		100	91.33
C1-3	80.8	81.6	81.07	80.53	80.93	81.07	84	86.4	84.27	89.33	86.8	89.87	88.27	90.27		91.33	100

Table 5.2.6 DNA sequence Percent Identity Matrix for mCherryM10L RFP codon variants
Percent Identity Matrix created by Clustal2.1, of all mCherryM10L CDS variants generated by Chimera Evolve. A value of 100% means that the sequences are identical. The shorthand names for CDS variants can be cross-referenced using Table 5.2.3 and 5.2.4; the format follows the nomenclature of *Cw-r* where *C* denotes the CDS name (C = mCherryM10L), *w* denotes the weight tuple and *r* denotes the replicate number.

5.2.3 Cloning strategy

Loop Assembly was chosen as the method to create the constructs for this experiment, which is a type of Golden Gate Assembly that utilises standardised acceptor plasmids to scale up assemblies into larger transcription units (TU) (Pollak *et al.*, 2019). The assembly starts by using linear DNA direct from synthesis with compatible ends for the assembly standard, which in this case was Phytobricks (Cai, Carrasco Lopez and Patron, 2020).

The general methodology behind Loop Assembly is outlined here, and the overall motivation behind Loop Assembly was discussed in chapter 2. The linear parts are cut using restriction enzyme, SapI, and are cloned into the universal loop acceptor plasmid – pSB1C00. The universal acceptor plasmid contains a mRFP TU which is cut out by SapI and replaced with the new Level 0 part by T4 Ligase. The white colonies are selected for, and the red ones ignored as background, in a process known as red/white screening. The result is a Level 0 part in an acceptor plasmid that can be used in the next step. The next step involves putting the Level 0 parts together in a TU, Level 1 part. Because each part has specific ends, they will ligate together in a set order. Again, red/white screening is used but this time the acceptor plasmid is an Odd level acceptor plasmid (pOdd-3) and the restriction enzyme used is BsaI. The result is a Level 1 part in an Odd level acceptor plasmid that can be used in the next stage. The next stage would be to assemble multiple (up to four) TUs in an Even level acceptor plasmid, however, that is not required for this experiment, so the cloning ends with Level 1 parts as TUs in Odd level acceptor plasmids. The advantage of modular assembly here is that each part only needs to be synthesised once but still allows flexibility when building up TUs, i.e., many different promoters can be tested in TUs with the same RBS, CDS & terminator parts.

pOdd plasmids operate in *E. coli* DH5α but do not have an origin of replication to allow operation in *B. subtilis* 168. TUs need to be cloned into a suitable expression vector, pHT01, to work with both organisms. The TUs can be extracted from the pOdd vector by PCR and ligated into pHT01 so that it can operate in both *E. coli* DH5α and *B. subtilis* 168. This experiment had to be scaled back due to setbacks in the automated cloning cycle; setup of essential equipment was delayed, and optimisation of protocols required more time than anticipated which resulted in an incomplete automated cloning cycle. Combinatorial design was removed from scope as it involved the most work and was least required to obtain experimental validation for Chimera Evolve. As the automated workflow was not completed in time for the experiment, much of the cloning had to be completed manually. To reduce the manual labour and resources needed for this process, the number of variants and replicates was reduced. Only one fluorescent protein, mGreenLantern was chosen to be tested and all the stochastic replicates were removed. This reduced the amount of Level 0 cloning required down to two promoter parts (one for each species), one RBS part, seven CDS parts – superfolder GFP as a positive control, mGreenLantern and its five codon variants, and one terminator part. The CDS parts are named in Table 5.3 (J23106, P_{veg}, B0032+spacer, sf0, G0, G1, G2, G3, G4, G5 & B0015). Level 1 parts sum to fourteen which is the seven CDS parts combined with J23106 promoter and the seven CDSs with P_{veg}.

5.3 *in vivo* fluorescence measurement of Chimera Evolve generated CDS variants in *Escherichia coli*

The hypothesis was that the variant optimised most in favour of *E. coli* MG1655 (G1) will be expressed most in *E. coli* DH5α and least in *B. subtilis* 168. This hypothesis also infers the inverse behaviour for the variant optimised most in favour of *B. subtilis* 168 (G5), and for the other variants to form a linear trend in between these two extremes.

The strains of *E. coli* DH5α containing the Level 1 constructs were grown on LB Agar + 100 µg/mL Kanamycin. The Level 1 parts were expressed in *E. coli* DH5α with pOdd-3 as the plasmid backbone. For testing in *B. subtilis* 168, the TUs would have required cloning into a separate vector which was not completed due to time constraints.

5.3.1 GFP Assay with J23106 promoter

Methods for this experiment are documented in chapter 7, section 7.8.6. The results of these experiments are outlined in this section.

Figure 5.3.1 shows the cell density and green fluorescence over time during the testing of the mGreenLantern constructs with J23106 promoter in *E. coli* DH5α. The strains show similar growth patterns and trends with fluorescence. No difference in fluorescence was observed between the control group (*E. coli* DH5α with no plasmid), and the test groups.

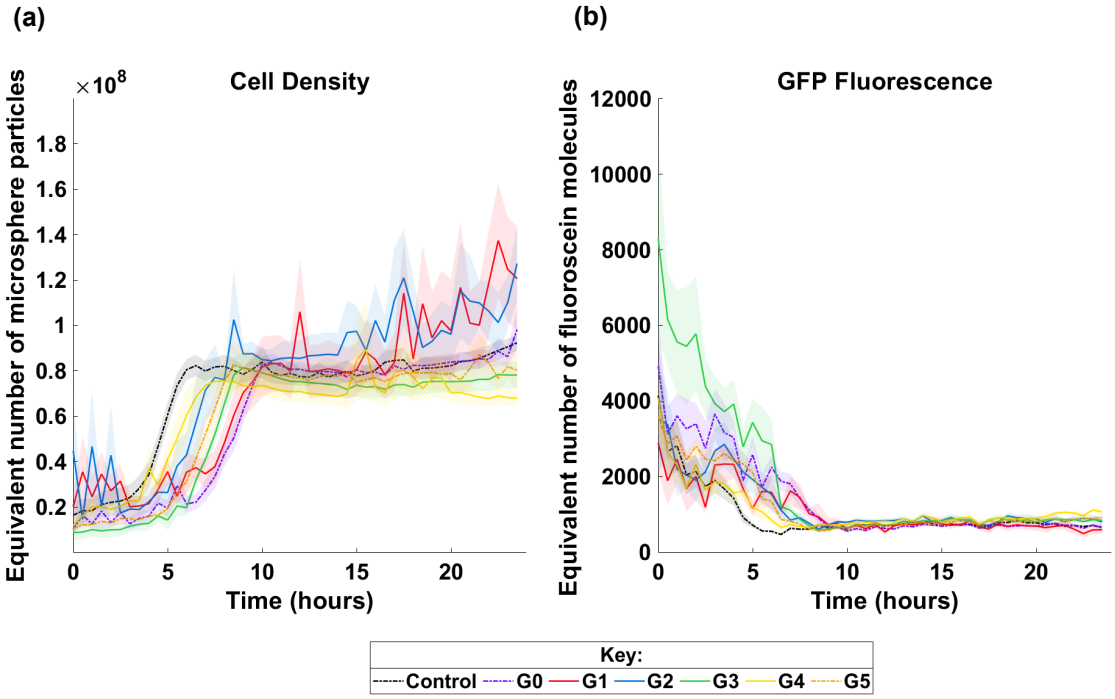

Figure 5.3.1 Green Fluorescence of CDS variants in E. coli DH5α with J23106 promoter

Control is *E. coli* DH5α with no plasmid. G0 is the wildtype sequence for mGreenLantern. G1, G2, G3, G4, G5 are Chimera Evolve generated variants with the weightings (0,1), (0.25,0.75), (0.5,0.5), (0.75,0.25). (1,0) for G1, G2, G3, G4 and G5 respectively, where the left digit represents the weight toward *B. subtilis* 168 and the right digit represents the weight toward *E. coli* MG1655. Cells were grown in 100 µL of LB media with 100 µg/mL of Kanamycin for strains with a plasmid, for 24 hours at 37°C in a CLARIOstar Plus plate reader with readings taken every 30 minutes. (a) shows the cell density of *E. coli* DH5α over time. Values are blank corrected before calibration to equivalent number of microspheres. The mean average of 5 replicates is represented by solid lines, standard error is represented by the translucent regions. (b) shows fluorescence at excitation 470 nm (band 15 nm) and emission 515 nm (band 20 nm). Values are blank corrected before calibration to equivalent number of fluorescein molecules. The mean average of 5 replicates is represented by solid lines, standard error is represented by the translucent regions.

5.3.2 GFP Assay with P_{veg} promoter

Methods for this experiment are documented in chapter 7, section 7.8.6. The results of these experiments are outlined in this section.

Figure 5.3.2 shows the cell density and green fluorescence over time during the testing of the mGreenLantern constructs with P_{veg} promoter in *E. coli* DH5α. The strains show similar growth patterns, however, G5 reached exponential growth phase sooner than the others.

All strains except the control group (*E. coli* DH5α with no plasmid), began fluorescing at similar times. The sf0 group showed greater fluorescence than all other groups. G1, G2, G3, G5 & G0 show similar fluorescence measurements over time. The G4 group had lower fluorescence than all other groups with plasmids.

The distribution of fluorescence for each group is shown in Figure 5.3.3. The pattern of Chimera ARS scores for each CDS does not match the distribution of fluorescence data observed. The sf0 group shows higher maximum fluorescence than the other groups and the G4 group shows lower fluorescence than the other groups. The difference in fluorescence between groups is most apparent in the final four hours of the experiment, as shown in plot (b) of Figure 5.3.3.

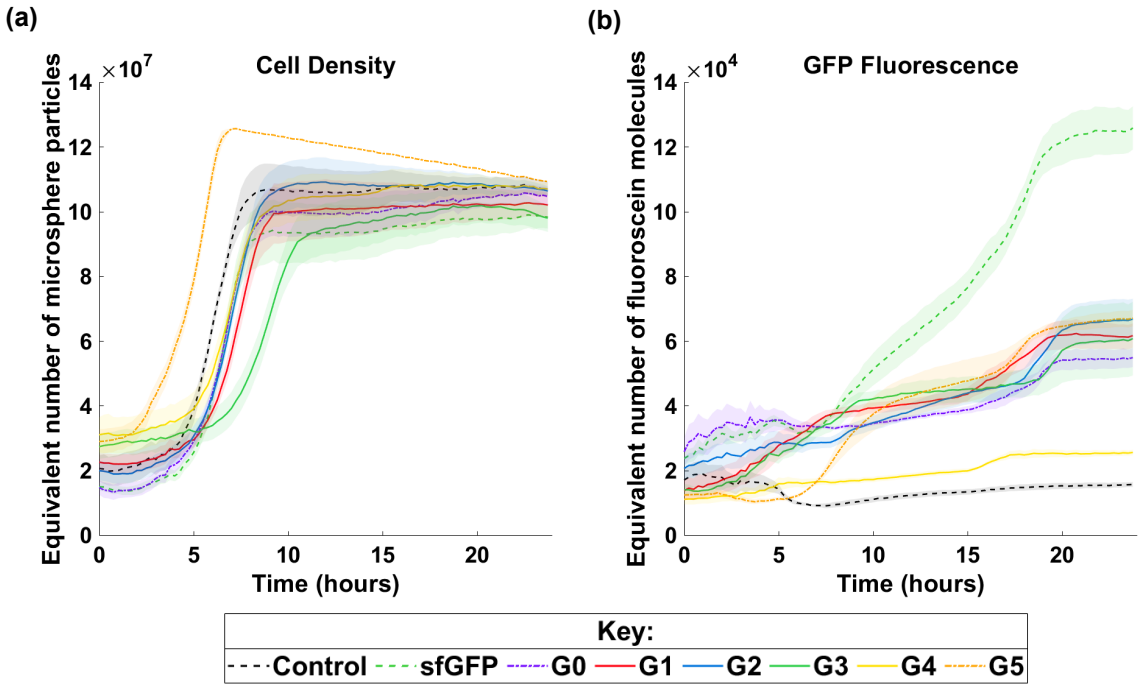

Figure 5.3.2 Green Fluorescence of CDS variants in E. coli DH5α with P_{veg} promoter

Control is *E. coli* DH5α with no plasmid. G0 and sf0 are the wildtype sequence for mGreenLantern and superfolder GFP, respectively. G1, G2, G3, G4, G5 are Chimera Evolve generated variants with the weightings (0,1), (0.25,0.75), (0.5,0.5), (0.75,0.25). (1,0) for G1, G2, G3, G4 and G5 respectively, where the left digit represents the weight toward *B. subtilis* 168 and the right digit represents the weight toward *E. coli* MG1655. Cells were grown in 100 µL of LB media with 100 µg/mL of Kanamycin for strains with a plasmid, for 24 hours at 37°C in a CLARIOstar Plus plate reader with readings taken every 15 minutes. (a) shows the cell density of *E. coli* DH5α over time. Values are blank corrected before calibration to equivalent number of microspheres. The mean average of 5 replicates is represented by solid lines, standard error is represented by the translucent regions. (b) shows fluorescence at excitation 470 nm (band 15 nm) and emission 515 nm (band 20 nm). Values are blank corrected before calibration to equivalent number of fluorescein molecules. The mean average of 5 replicates is represented by solid lines, standard error is represented by the translucent regions.

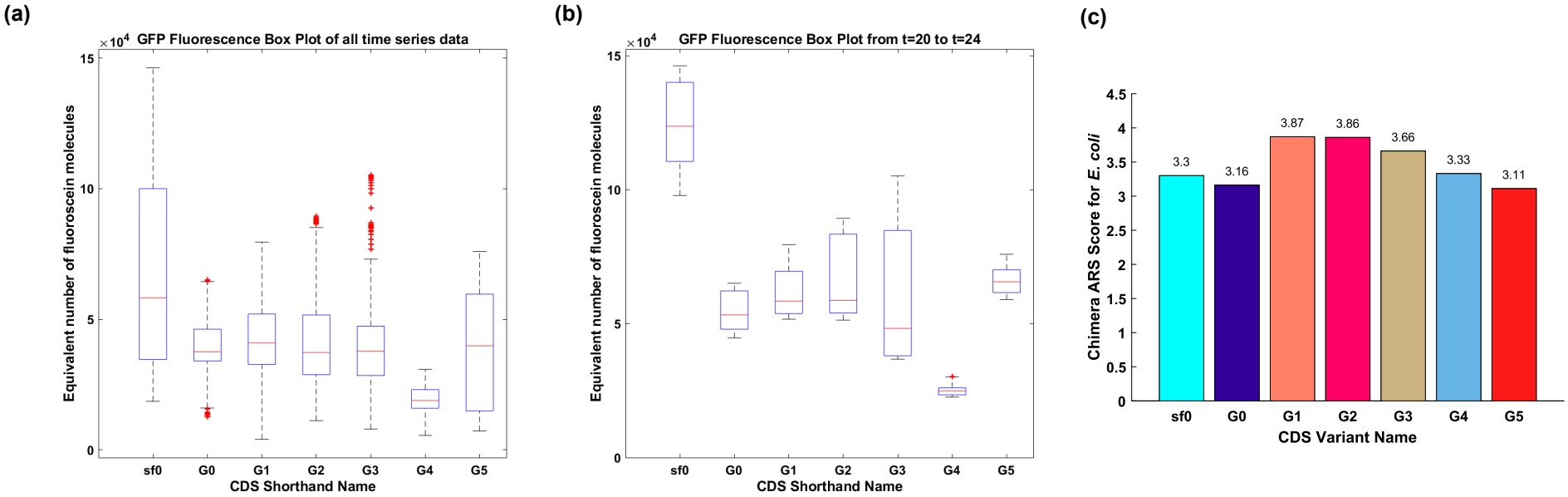

Figure 5.3.3 Box plots for green fluorescence of CDS variants in E. coli DH5α with P_{veg} promoter
Box plots of the data from the experiment shown in Figure 5.3. (a) Each box plot represents all data for each CDS variant across the entire series (b) Each box plot represents a subset of the data from 20 hours to 24 hours for each CDS variant (c) Bar chart showing the Chimera ARS scores of each CDS variant with *E. coli* MG1655 as the target organism.

5.4 Discussion

5.4.1 Absence of expression using the J23106 promoter

In the assays shown in 5.3.1, no fluorescence was observed from the mGreenLantern constructs using J23106 as a promoter. As the Level 1 cloning for the P_{veg} promoter was done in tandem with J23106, it was noted that the strains with the P_{veg} promoter were fluorescing in *E. coli* DH5α*,* so it was decided to conduct the assay using the P_{veg} constructs.

To the author's knowledge, it has not been observed previously that J23106 has expression issues in *E. coli* DH5α. J23106 is typically characterised as a robust, medium strength promoter (Yang *et al.*, 2023; Tan, Hsiang and Ng, 2021; Jia *et al.*, 2014). However, the results in 5.3.1 provide some evidence that J23106 may not always work as expected in *E. coli* DH5α.

The constructs used in these assays were validated by Sanger sequencing, so there is not an issue with the sequence being incorrect. Additionally, as these constructs work with the P_{veg} promoter, it does not appear that there is a problem with the other genetic parts in the TU or the vector in isolation.

5.4.2 Observations from in vivo testing of mGreenLantern CDS variants

The final assay conducted was to examine GFP expression over time with these different codon variants. The expected result to support the hypothesis for this section of research was a linear trend of increased expression (measured by fluorescence intensity) as the codon variants increase their bias toward *E. coli* MG1655. However, the codon variants did not display the linear trend that was expected so the original hypothesis could not be supported based on the results of these experiments.

G5, with the weight (1,0), was optimised with the most bias towards *B. subtilis* 168 and had the lowest Chimera ARS score with *E. coli* MG1655 as the target organism. As such, G5 was expected to fluoresce the least out of all mGreenLantern variants. Contradictorily, it exhibited the greatest fluorescence out of all of them, albeit without statistical significance. G4, with the weight (0.75,0.25), fluoresced the least out of the five variants. It is expected that G4 would perform poorly compared to the other

variants, however, not to perform worse than G5. Testing of the stochastic variants would provide additional evidence to support these results and would provide robustness to the data provided. Currently, there is insufficient evidence to confirm nor deny whether the ARS score correlates with protein expression for these CDSs.

The wildtype sequence, G0, did not show statistically significant difference in expression levels between itself and the other codon variants of the same CDS (except for G4 which showed weaker fluorescence than all other variants). It is worth noting that superfolder GFP did show a significantly higher fluorescence than any of the mGreenLantern variants, contradictory to Campbell *et al.*'s findings in eukaryotic cells (Campbell *et al.*, 2020). Superfolder GFP may be preferential in *E. coli* DH5α and the results presented here provide evidence to support that (Pédelacq *et al.*, 2006). However, it is more likely that the instrument settings in fluorescence measurement were not optimised for spectra of mGreenLantern. Figure 5.4.1 demonstrates that the emission spectra of superfolder GFP is higher than that of mGreenLantern when excited at 470 nm, which is what the plate reader was set to excite at (Lambert, 2019). Further configuration of the plate reader settings would help for characterisation of mGreenLantern in *E. coli* DH5α.

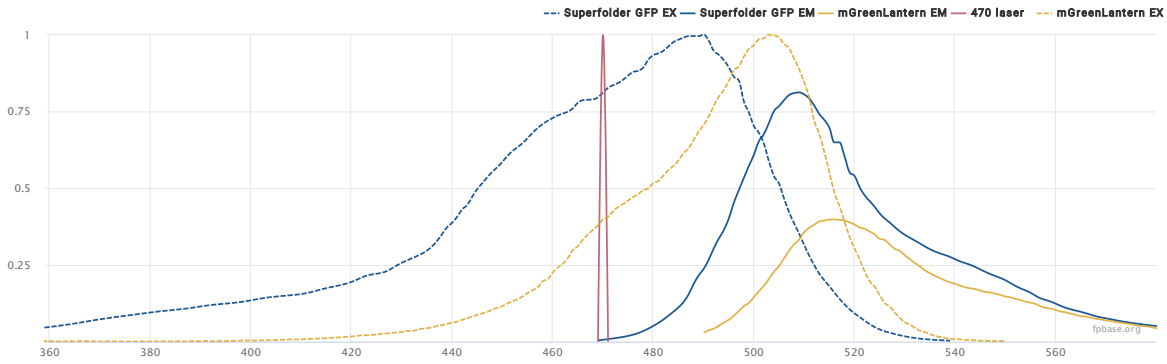

Figure 5.4.1 Spectra of Superfolder GFP and mGreenLantern
Excitation (EX) and emission (EM) ranges of Superfolder GFP (shown in blue) and mGreenLantern (shown in yellow) generated using FPbase. The emission curves are normalised based on the light source from a laser at 470 nm (shown in red).

There are many factors that may contribute to the results not showing the expected pattern. It could be related to the secondary structure of the mRNA; features such as hairpin loops can prevent the binding of ribosomes, block the movement of tRNAs, or inhibit extension of the polypeptide chain (Boo, Ellis and Stan, 2019; Bahiri-Elitzur and Tuller, 2021). In theory, usage of the Chimera ARS score for optimisation should negate this effect to an extent (Bahiri-Elitzur and Tuller, 2021; Skelton *et al.*, 2020; Diament *et al.*, 2019; Zur and Tuller, 2014). However, single substrings within a large CDS may not be sufficient to compensate for the entire molecule.

It may be the case that the cells in this experiment were not experiencing enough metabolic stress for a difference in expression to be observed. The cells may have their translational resources in enough abundance that the expression of heterologous protein is not affected by CUB. The P_{veg} promoter is strong, but it is a promoter for *B. subtilis* 168 and is not optimised for expression in *E. coli* DH5α, so the transcription rate may have been low enough for CUB to have had no observable impact in this experiment.

Without further experimentation, conclusions cannot be drawn as to the effectiveness of Chimera Evolve *in vivo*. However, the results provided in this chapter lean toward there being no significant difference between the codon variants.

5.4.3 Future Work

There are multiple avenues to explore for continuation of this work and to provide further experimental evidence for Chimera Evolve. The experiments conducted for this chapter could be repeated with different methods of measurement. Measuring a range of different settings (e.g., emission and excitation spectra, gain, focal length, number of flashes, etc.) on the CLARIOstar Plus plate reader could provide useful data for characterising the fluorescent proteins and repeating the experiment with the optimal settings. Flow cytometry would be a useful technique to employ as this would provide measurements per cell. This would remove much of the noise relating to the growth of the population and the following data correction steps would be removed.

The assays need to be tested in *B. subtilis* 168 to determine what trend is seen in this organism, allowing comparison between both species. Both CDSs, mGreenLantern and mCherryM10L, should be tested to give a context on more than one CDS. It has already been observed that Chimera Evolve performs differently depending on the CDS (Skelton *et al.*, 2020), so testing more CDSs with different attributes (e.g., protein size, protein function, species of origin) would give a wider coverage of results to support any found conclusions. Testing the stochastic variants would help add more replicates behind each weight provided, thus increasing the robustness of the data gathered.

Combinatorial design would be a useful approach in these experiments as it can determine the optimal promoter and RBS combination to maximise the expression in each organism. It would also aid in negating factors relating to regulation and transcription of the genes as this experiment is primarily concerned with translation and codon usage bias. Using a very strong constitutive promoter and RBS may induce sufficient metabolic load stress in the cell to reduce the availability of translational resources. A separate source of metabolic load stress (for example, the vioB_mCherry fusion protein) could be introduced into the cell in attempt to ensure that the cell's resources are limited.

Finally, these results should be compared with other codon optimisation algorithms *in vivo*. Experiments could either be completed in separate labs with standardised

approaches or this could be done on a smaller scale by selecting a subset of algorithms to benchmark against. The clear candidate algorithm for comparison would by Chimera UGEM as this was used in Skelton *et al*.'s presentation of Chimera Evolve, however, algorithms using different measures (such as CAI) would allow for comparison of different metrics.

5.4.4 Summary

Chimera Evolve is effective at producing candidate CDSs with ARS scores falling under local optima. The experimental results presented in this chapter do not support the hypothesis that higher ARS scores correlate with greater protein expression. However, there is more experimentation to be completed before it can be concluded that the candidate sequences generated by Chimera Evolve are not best optimised for the species targeted. The results presented in this chapter provide a foundation for future experiments to work upon, have highlighted key areas to explore, and has introduced a methodology that can be developed further.

CHAPTER 6

6 Discussion and Conclusions

6.1 Summary

The aim of this project was to work toward biological and computational systems to aid in the detection, mitigation, and understanding of stress and metabolic burden in model prokaryotes. Additionally, there was a goal to develop systems that work portably between model organisms: *E. coli* and *B. subtilis*. The aims of this research were addressed by the exploration of three objectives which are listed as follows. To describe and demonstrate the capability of the biomarker selection algorithm, ROTC, to produce sets of biomarkers of a given stress state; which was presented in chapter three of this thesis and is discussed in 6.1.1. The design, build, and test of a genetic dual-input AND gate toward the detection of a specific stress state; which was presented in chapter four of this thesis and is discussed in 6.1.2. To attain *in vivo* experimental evidence to validate the effectiveness of the codon optimisation algorithm, Chimera Evolve; which was presented in chapter five of this thesis and is discussed in 6.1.3.

6.1.1 Using ROTC for the selection of biomarkers that identify a given stress

Chapter three introduced the biomarker selection algorithm, ROTC. ROTC was used on two data sets, one from tiling array data and one from RNA-seq data, to generate sets of stress specific biomarkers. ROTC demonstrated its capability to produce sets of biomarkers that accurately classify a given stress state versus all other conditions provided. Interestingly, the models produced from the tiling array data set did not necessarily work for the RNA-seq data set and vice versa. However, many solutions in common could be found between both data sets which performed well.

Within the landscape of existing biomarker selection algorithms, ROTC can be added as an additional tool for researchers to apply in their studies. Fundamental techniques such as recursive feature elimination (RFE) (Guyon *et al.*, 2002) and support vector machine (SVM) (Sun *et al.*, 2016) based classifiers remain useful tools for biomarker selection, but ROTC may work better for given contexts such as the intended purpose here: generating inputs to a genetic logic gate. Other algorithms are very capable at generating larger sets of biomarkers and biomarker panels, such as RGIFE, BIM, or PanelomiX (Lazzarini and Bacardit, 2017; Huang

et al., 2023; Robin *et al.*, 2013; Huang *et al.*, 2021). In instances where discriminatory panels of biomarkers are beneficial, these techniques clearly excel. Whereas ROTC is more suitable for sets of biomarkers of fewer than three, where the number of biomarkers can be specified by the user. Continued work into ROTC will aid the useability of the algorithm and its effectiveness.

6.1.2 A synthetic genetic AND gate for detecting a given stress response

Chapter four presented the design, build, and test of a genetic AND gate. The AND gate developed was based upon the work in 'Environmental signal integration by a modular AND gate' as it is a relatively simplistic example of a dual input AND gate that had already been demonstrated to be modular (Anderson, Voigt and Arkin, 2007). While previous studies have generated genetic circuits for regulating against stress and metabolic burden, this study was intended to highlight how capable ROTC is at distinguishing stress states with complex transcriptional fingerprints (Ceroni *et al.*, 2018; Dragosits, Nicklas and Tagkopoulos, 2012). To realise the aim of demonstrating that ROTC can effectively distinguish cell states *in vivo*, an investigation would involve subjecting bacteria to multiple different stress conditions and ensuring that only the targeted stress state activated the AND gate; a study which would likely take part of the future work discussed in 6.2.

In this thesis, the design was adapted to operate in *B. subtilis* as well as *E. coli* with the intention of integrating the gate with stress responsive promoters as generated by ROTC. Each of the three constructs were redesigned as testing constructs for proof of principle and adapted to use promoter and RBS combinations that work in both *B. subtilis* and *E. coli*. The first input generated the tRNA molecule, SupD, whose presence was checked using RNA-seq in *E. coli* DH5α. The second input generated the modified T7 RNAP and a fluorescent protein, mCherry2, for validation by fluorescence assays. The presence of mCherry2 was determined by fluorescence assays in both *E. coli* DH5α and *B. subtilis* 168, though the regulation appeared leaky so needs future optimisation. The final construct was the output which was activated by a T7 promoter and produced a fluorescent protein, mTagBFP, for validation by fluorescence assays which was validated in *E. coli* BL21(DE3).

The AND gate was tested in a cell free expression system, *E. coli* S30 extract for circular DNA, but the results did not demonstrate the circuit functioning as an AND gate. Though cell free systems have many advantages such as the reduced impact of confounders in experimental results, there are also many drawbacks to using these systems for research (Shin and Noireaux, 2012; Maharjan and Park, 2023). Not only is there a significant decrease in yield from a cell free system and an increase in cost, but there are also limitations in the regulatory features available in the extract (Brookwell, Oza and Caschera, 2021). The growth point at which the cell lysate is extracted for production of a cell free system directly influences what is available within the system for regulation; studies have been conducted to develop cell free systems at different growth stages of *E. coli*, thus, providing different regulatory elements in the system (Failmezger *et al.*, 2017). However, with the commercially available S30 extract kits, the regulation available is limited which is a factor when considering the functionality of the T7 based AND gate tested in this research.

Development of genetic logic gates remains on ongoing field of study in synthetic biology and different mechanisms are frequently being applied to the concept of logic gates. Additionally, new toolkits are being released with useful standardised parts such as joint universal modular plasmids (JUMP) (Valenzuela-Ortega and French, 2021). The T7 based AND gate built for this work is still a good candidate for a modular and portable AND gate as shown in the original publication and supported by the findings from the testing of individual components in this research (Anderson, Voigt and Arkin, 2007). Recent studies do provide some insight and useful tools for testing the AND gate used in this work as an entire system; for example, the standardised *Bacillus* SANDBOX plasmids utilised in 'Chimeric MerR-Family Regulators and Logic Elements for the Design of Metal Sensitive Genetic Circuits in *Bacillus subtilis*' (Ghataora, Gebhard and Reeksting, 2023) or JUMP (Valenzuela-Ortega and French, 2021) would aid in the cloning of the AND gate onto a single plasmid for testing. Additionally, introducing proper standards in areas such as cloning sites make optimisation of the circuit easier when parts need swapping out for optimisation (Bird, Marles-Wright and Giachino, 2022). The AND

gate was not built using a modular cloning strategy like Loop Assembly, but it would be beneficial for future work to implement (Pollak *et al.*, 2019).

6.1.3 in vivo experimental validation of a cross-species codon optimisation algorithm, Chimera Evolve

In chapter five, codon variants with a gradient of biases toward two organisms, *E. coli* and *B. subtilis* were generated by Chimera Evolve, synthesised, built, and tested in *E. coli* DH5α (Skelton *et al.*, 2020). The codon variants were successfully expressed in *E. coli* DH5α but the findings did not suggest that the bias toward *E. coli* increases fluorescence. Chimera Evolve is unique in its ability to optimise a coding sequence with a bias toward more than one organism, thus it is important to provide experimental evidence in its support. Although the results from the experiments conducted in this thesis did not show a linear trend in difference in expression as the bias from one species to another changed, the experiment did highlight important directions for future work (see 6.2).

The insight that codon variants do not immediately result in an observable difference in expression is an important observation which can be applied to other studies looking at codon optimisation *in vivo*. There have been many studies conducted that have observed increased expression of heterologous proteins after codon optimisation (Boël *et al.*, 2016; Fu *et al.*, 2020; Zhou *et al.*, 2016). Typically, the codon optimisation studies are focussed on matching the codon usage bias (CUB) of the host organism, but some work has been done to experimentally validate the usage of the Chimera average repetitive substring score (cARS). The Chimera UGEM algorithm optimises using ARS and was experimentally validated using heterologous proteins in the green alga, *Chlamydomonas reinhardtii* (Diament et al., 2019). However, to the author's knowledge, Chimera Evolve remains the only algorithm to optimise sequences with a weight toward more than one organism and as such will require experimental validation to determine the performance in both organisms.

6.2 Further work

The work completed in this thesis provides insight for future work to continue the study. Each chapter focused on a different section of research which can be combined as part of a larger project investigating the stress response and metabolic burden. The vision for the study would be to construct a genetic circuit that can detect and/or regulate a given stress response. The circuit would be modular and portable to allow for the adaptation to any given stress response in any given organism.

The conceptualised circuit to bring the work in this thesis together would involve two key components. The first component would be a burden generating device which would operate by overexpressing a recombinant protein, such as the VioB-mCherry fusion protein used in 'Burden-driven feedback control of gene expression' (Ceroni *et al.*, 2018). The second component would be a dual input AND gate, as described in chapter four. The inputs to the AND gate would be natural promoters determined by ROTC, using data relevant to the transcriptome under stress by overexpression of the recombinant protein (e.g., VioB-mCherry) generated by the first component. Once the conceptualised circuit is functional, the recombinant protein generating the metabolic burden would be codon optimised by Chimera Evolve to alleviate some of the metabolic burden associated with shared resources.

Chimera Evolve itself needs additional experiments to be conducted to determine its effectiveness *in vivo*. In chapter five, it was discussed how the lack of a linear trend of expression between the codon variants could be due to various factors and ideas were presented about how this could be investigated further. Primarily, the codon variants need to be tested in both organisms for which they were optimised for and more than one CDS should be tested to give a range of results. Additionally, CDS variants generated by alternative algorithms should be tested alongside to provide a comparison. Another algorithm utilising ARS score would be beneficial to test and to create a range of ARS scores that emulate the bias generated by Chimera Evolve which would help to determine whether ARS score makes a difference to observable *in vivo* expression. However, the most pressing theory to test would be to measure the CDS variants while the bacteria are undergoing stress

from metabolic burden as this would help ensure that the shared translational resources are limited enough to make an observable difference.

6.3 Conclusion

In summary, the work conducted in this thesis has resulted in some useful insights regarding the conceptualised method to alleviate metabolic load stress in prokaryotic model organisms. Not all results were positive, but they do provide useful information that can be applied to the continuation of this research. ROTC was presented and shown to be effective at generating candidate sets of biomarkers for a given cell state. An AND gate was built and tested, which worked in isolation but not when pulled together in a cell free expression system. Chimera Evolve was given evidence for *in vivo* validation, however, the results did not give the expected result. Overall, the findings presented within this thesis present evidence that supports future endeavours in this area; whether that be the description of ROTC, the results from testing the AND gate in multiple strains, or the lack of variation found between codon variants generated by Chimera Evolve. Continuation of the work in this thesis would be beneficial for the study of metabolic load stress and portability in bacteria.

CHAPTER 7

7 Methods

7.1 Microbiology Methods

Unless otherwise stated, all bacterial cultures were grown in LB Broth at 37 ºC, shaking at ~200 rpm, in an Eppendorf Innova S44i shaking orbital incubator for the duration specified in the given method.

7.1.1 Preparation of growth media

LB Broth: For 1 L of LB, the following chemicals were mixed in sterile, pure water: 10 g tryptone, 10 g sodium chloride, 5 g yeast extract. The solutions were sterilised by autoclaving at 121 ºC and allowing to cool.

LB Agar: For 1 L of LB, the following chemicals were mixed in sterile, pure water: 15 g agar, 10 g tryptone, 10 g sodium chloride, 5 g yeast extract. The solutions were sterilised by autoclaving at 121 ºC and allowing to cool until solidified.

SMM: For 1 L of SMM, the following chemicals were mixed in sterile, pure water: 2 g ammonium sulphate, 14 g dipotassium hydrogen phosphate, 6 g potassium dihydrogen phosphate, 1 g sodium citrate dehydrate, 0.2 g magnesium sulphate. The solutions were sterilised by autoclaving at 121 ºC and allowing to cool.

Preparation of antibiotics

Antibiotic solutions were prepared at the following concentrations for long-term storage at -20 ºC: ampicillin at 50 mg/mL in water; carbenicillin at 100 mg/mL in water; chloramphenicol at 100 mg/mL in ethanol; kanamycin at 25 mg/mL in water.

Unless otherwise specified, the concentration of antibiotic required for cell culturing were as follows: 100 µg/mL ampicillin for *E. coli*; 100 µg/mL carbenicillin for *E. coli*; 5 µg/mL chloramphenicol for *B. subtilis*; 100 µg/mL chloramphenicol for *E. coli*; 50 µg/mL kanamycin for *E. coli*.

LB Agar Plates

Sterile LB agar was melted using a water bath at a constant temperature of ~70 ºC, then allowed to cool before pouring into sterile petri dishes. If antibiotic was required, it was added to the liquid LB agar once the agar had cooled to below 50 ºC, before pouring into sterile petri dishes. The LB agar plates were left in a sterile environment

to solidify, then were sealed, and left at room temperature overnight before being refrigerated at 4 ºC for short-term storage.

7.1.2 Preparation of overnight culture

Overnight cultures were prepared by inoculating 5 mL of LB broth in a sterile 50 mL Falcon tube, with cells (grown on an agar plate) using a sterile loop. Antibiotic was added to the broth as required and at the necessary concentration. Cultures were incubated “overnight” for ~16 hours at 37 ºC, shaking at 250 rpm in an orbital incubator (Eppendorf Innova S44i).

7.1.3 Cell growth on agar plates

Cell cultures were grown on LB agar plates, with antibiotic at the necessary concentration as required, by either streaking or spreading and allowed to grow in a static incubator at 37 ºC for ~16 hours until colonies form.

Streaking: cells were transferred to a sterile LB agar plate using a sterile loop in straight lines in a small area of the plate, followed by several streaks across each section to gradually decrease the density of cells across the plate.

Spreading: cells were dispensed onto a sterile LB agar plate using a Gilson pipette and spread across the plate using sterile glass beads (no more than 200 µL were dispensed onto any one plate).

7.1.4 Preparation of glycerol stocks

Following growth in overnight culture (see 7.1.2), 500 µL of liquid culture were dispensed into a sterile 1.8 mL cryotube. 500 µL of 50% (v/v) glycerol was added to the culture and pipetted up and down until mixed. The glycerol stock was labelled and stored at -80 ºC for long-term storage.

7.2 DNA Assembly

7.2.1 Design and DNA synthesis

Design of DNA constructs used throughout this study were aided by Benchling. Benchling was used to create a repository of parts used in the study, to analyse them in detail, search for restriction sites, design primers, simulate assemblies, run virtual restriction digests and to visualise alignments of sequencing results.

IDT were used for DNA synthesis of gBlocks for Gibson Assembly and the synthesis of oligonucleotide primers. Twist Bioscience were used for the synthesis of gene fragments (such as those used in Chapter 5) and for the outsourcing of cloning, i.e., integration of a synthesised gene fragment into a plasmid backbone at a specified locus.

7.2.2 Gibson Assembly

Reagents for Gibson Assembly were sourced from New England Biolabs. Prior to assembly, the backbone was digested using the procedure in 7.5.1 and run on an agarose gel to confirm the digestion (7.5.2). Gibson assembly was conducted using three ratios on insert to backbone (1:1, 2:1, 3:1) plus controls (no insert as a negative control, HiFi Assembly Positive Control Mix as a positive control). 50 ng of the digested backbone was added to PCR tubes and appropriate ratio of moles of part were added to each of the reactions and made up to 10 µL with nuclease-free water. 10 µL of 2X Gibson Assembly Master Mix was added to each reaction and mixed by pipetting up and down. Each reaction was incubated for 20 minutes at 50 ºC, then held at 4 ºC prior to downstream transformation (see 7.3.2).

7.2.3 Loop Assembly

NEBridge Ligase Master Mix was used for Loop Assembly. The protocol used was a modified version of the manufacturer's recommendations. DNA parts were stored in nuclease-free water at concentration of 10 fmol/µL at -20 ºC.

For level 0 parts: 2 µL of DNA backbone and 2 µL of the DNA part insert were added to a PCR tube. 1.5 µL of NEBridge Ligase Master Mix was added to the reaction mix, then 0.5 µL of restriction enzyme SapI was added to the reaction mix. The reaction mix was made up to 10 µL with nuclease-free water and mixed until

homogenous. The reaction was incubated in a thermocycler for the following program: repeat for 100 cycles [37 ºC for 60 seconds, 16 ºC for 90 seconds], 50 ºC for 10 minutes, hold at 4 ºC.

For level 1 parts: 2 µL of DNA backbone and 1 µL of each DNA part (1.5 µL of AB part, 1.5 µL of BC part, 1.5 µL of CE part, 1.5 µL of EF part) insert were added to a PCR tube. 1.5 µL of NEBridge Ligase Master Mix was added to the reaction mix, then 0.5 µL of restriction enzyme BsaI was added to the reaction mix. The reaction mix was made up to 10 µL with nuclease-free water and mixed until homogenous. The reaction was incubated in a thermocycler for the following program: repeat for 100 cycles [37 ºC for 60 seconds, 16 ºC for 90 seconds], 50 ºC for 10 minutes, hold at 4 ºC.

Level 0 parts used a pSB1C00 acceptor plasmid as a backbone which conveys chloramphenicol resistance in *E. coli*. Level 0 assembly mixes were transformed into *E. coli* (see 7.3.2) and grown on LB agar plates with 100 µg/mL chloramphenicol.

Level 1 parts used a pOdd-3 acceptor plasmid as a backbone which conveys kanamycin resistance in *E. coli*. Level 1 assembly mixes were transformed into *E. coli* (see 7.3.2) and grown on LB agar plates with 50 µg/mL kanamycin.

7.3 Transformation

7.3.1 Preparation of chemically competent E. coli cells

Two buffers were required for this protocol named TF-1 and TF-2. TF-1 was made by mixing 1.48 g potassium chloride, 0.59 g potassium acetate, 0.3 g calcium chloride dihydrate, and 30 g glycerol in 190 mL of sterilised, pure water. TF-1 was then adjusted to pH 6.4 with acetic acid. TF-2 was made by mixing 0.148 g, 2.2 g calcium chloride dihydrate, and 30 g glycerol in 196 mL of sterilised, pure water. Both solutions were sterilised by autoclaving at 121 ºC for 20 minutes and allowed to cool to 4 ºC. Once cooled, 10 mL of 1 M manganese chloride tetrahydrate (prepared in sterile, pure water and filter sterilised) solution was added to TF-1. 4 mL of 0.5 M MOPS buffer (prepared in sterile, pure water; adjusted to pH 6.8 with potassium hydroxide; filter sterilised) was added to TF-2.

A streak plate was prepared (see 7.1.3) of the required *E. coli* strain for preparation of chemically competent cells. From the streak plate, an overnight culture was prepared (see 7.1.2). In a sterile 250 mL conical flask, 40 mL of LB broth was inoculated with 400 µL of overnight culture and allowed to grow at 37 ºC, shaking at 250 rpm until OD_{600} was between 0.4 and 0.5. The culture was then transferred to a sterile 50 mL Falcon tube to pellet by centrifugation at 4 ºC, the supernatant was discarded. The pellet was re-suspended in 8 mL of chilled TF-1 buffer and placed on ice for 15 minutes before pelleting by centrifugation as before. The supernatant was discarded, and the pellet resuspended in 4 mL of TF-2 buffer. The resuspended mixture was transferred into 1.5 mL centrifuge tubes as 50 µL aliquots. The aliquots were snap frozen in liquid nitrogen and stored at -80 ºC.

7.3.2 Transformation of chemically competent E. coli cells

An aliquot of chemically competent *E. coli* cells from -80 ºC storage was placed on ice to thaw. ~5 ng of circular DNA in nuclease-free water were dispensed into the thawed competent cells and mixed by flicking the tube 4-5 times. The mixture was then placed on ice for 30 minutes. Heat shock was applied to the mixture by submerging the tube in a 42 ºC water bath for 30 seconds before being placed back on ice for 2 minutes. 350 µL of sterile SOC media was added to the mixture and pipetted up and down until mixed. The culture was incubated to recover for 1 hour

at 37 ºC, shaking at 250 rpm. 100 µL of the culture was spread onto an LB Agar plate with the necessary antibiotic and allowed to grow overnight (~16 hours) until colonies form.

7.3.3 Transformation of B. subtilis 168

Two specialist growth media were required for this protocol based on SMM: Spizizen-starvation media (SSM) and Spizizen-plus media (SMM+). For SMM+, the following chemical solutions are mixed in 10 mL of SMM: 125 µL of 40% (w/v) glucose, 100 µL of tryptophan, 60 µL of magnesium sulphate, 10 µL of cas-amino acids, 5 µL of iron ammonium citrate. For SSM, the following chemical solutions are mixed in 10 mL of SMM: 125 µL of 40% (w/v) glucose, 60 µL of magnesium sulphate. All listed chemicals were filter sterilised before addition to sterile SMM.

An overnight culture was prepared in 5 mL of SMM (see 7.1.2). In a 50 mL Falcon tube, 5 mL SMM+ was inoculated with 300 µL of overnight culture and incubated for 3 hours at 37 ºC, shaking at 180 rpm, until OD_{600} was between 0.4 and 0.5. 5 mL of SSM (adjusted to 37 ºC) was added to the culture and incubated for a further 2 hours. Aliquots of 400 µL were prepared in 1.5 mL microcentrifuge tubes.

~100 ng circular DNA was added to the aliquot and mixed by pipetting up and down. The mixture was incubated for recovery for 1 hour at 37 ºC, shaking at 180 rpm. 200 µL of the culture was spread onto LB agar plates (see 7.1.3) with the appropriate antibiotic and incubated overnight (~16 hours) until colonies formed.

7.4 Extraction of plasmid DNA from *E. coli*

7.4.1 via Monarch® Plasmid DNA Miniprep Kit

All solutions were acquired from the Monarch® Plasmid DNA Miniprep Kit for this protocol and stored according to the instructions of the manufacturer.

An overnight culture of *E. coli* (see 7.1.2) was pelleted by centrifugation at 16000 xg and the supernatant discarded. The pellet was resuspended in 200 µL of Resuspension Buffer by pipetting up and down until there were no visible clumps and transferred to a 1.5 mL microcentrifuge tube. 200 µL of Lysis Buffer was added to the mixture and the tube inverted ~5 times until the solution changed colour to dark pink and became transparent. The solution was left to incubate for 1 minute. 400 µL of Neutralisation Buffer was added and the tube inverted gently until a white precipitate was formed. The solution was left to incubate for 2 minutes. The solution was spun down for 5-6 minutes in a centrifuge at 16000g to clarify the lysate. The supernatant was transferred to a clean spin column and spun for 1 minute at 16000 xg, the flow-through was discarded. 200 µL of Wash Buffer 1 was added to the spin column and then spun for 1 minute at 16000 xg, the flow-through was discarded. 400 µL of Wash Buffer 2 was added to the spin column and then spun for 1 minute at 16000 xg, the flow-through was discarded. The spin column was transferred to a clean 1.5 mL microcentrifuge tube. 50 µL of nuclease-free water (warmed to 50 ºC) was added to the centre of the spin column and incubated for 1 minute. The spin column within the clean tube was then spun for 1 minute at 16000 xg and the flow-through kept as the purified plasmid sample.

1-2 µL of the sample was measured on a Nanodrop to obtain a concentration of dsDNA. If the concentration was too low or contained impurities, the process was repeated from a new overnight culture.

7.4.2 via Genopure Plasmid Midi Kit

All solutions were acquired from the Genopure Plasmid Midi Kit by Sigma Aldrich for this protocol and stored according to the instructions of the manufacturer.

A 100 mL culture of *E. coli* was grown in LB broth (with antibiotic at the appropriate concentration) at 37 ºC, shaking at 250 rpm overnight. The overnight culture was

divided and transferred into two 50 mL Falcon tubes and was pelleted by centrifugation at 16000 xg at 4 ºC, and the supernatant discarded. Each pellet was resuspended in 4 mL of Suspension Buffer by pipetting up and down until there were no visible clumps. 4 mL of Lysis Buffer was added to the mixture and the tube inverted ~6 times and incubated for 2-3 minutes at ambient temperature (~15-22 ºC). 4 mL of Neutralisation Buffer (cooled to 4 ºC) was added and the tube inverted gently until a homogenous suspension was formed. The solution was left to incubate for 5 minutes at 4 ºC. The lysate was then separated from the solution by filtration using a sterile filter paper in a funnel inserted into a collection tube; the solution was filtered three times to ensure full separation of lysate from solution. A filter column was prepared by insertion into a collection tube, loading with 2.5 mL of Equilibration Buffer, allowing the liquid to leave the column by gravity flow, and discarding the flow through. The cleared lysate was loaded into the filter column and allowed to empty by gravity flow; this step was conducted twice and then the flow through was discarded. 4 mL of Wash Buffer was added to the column, allowed to empty by gravity flow, and the flow through was discarded; this step was conducted three times. In a new collection tube, 2.5 mL of Elution Buffer (warmed to 50 ºC) was added to the column, allowed to empty by gravity flow, and the flow through was collected; this step was conducted twice. The eluates were combined, precipitated with 3.6 mL of isopropanol (equilibrated to ambient temperature), and aliquoted out into 2 mL microcentrifuge tubes. The DNA was pelleted by high-speed centrifugation at 4 ºC, 15000 xg for 30 minutes; the supernatant was carefully discarded. 3 mL of 70 % (v/v) ethanol (cooled to 4 ºC) was added to the DNA pellets, and centrifuged at 4 ºC, 15000 xg for 10 minutes. The ethanol was carefully removed from the pellets, and the pellets allowed to dry in a sterile environment. The DNA pellets were resuspended in sterile, pure water (warmed to 50 ºC) with volumes of approximately 5 µL per pellet, then the pellets were combined.

1-2 µL of the sample was measured on a Nanodrop to obtain a concentration of dsDNA. If the concentration was too low or contained impurities, the process was repeated from a new overnight culture.

7.5 DNA Analysis

7.5.1 Restriction digest

Restriction enzymes and their recommended buffers were sourced from New England Biolabs. Depending on the enzyme used, the protocol may vary, and this was modified according to the manufacturer's recommendations. Below is the typical procedure followed throughout this thesis.

In a PCR tube, 1 µg of DNA suspended in nuclease-free water was added with 5 µL of buffer (usually 10X CutSmart Buffer) and made up to 49 µL with nuclease-free water. 1 µL of enzyme was added and mixed into the solution by pipetting up and down. The PCR tube was added to a thermocycler under the following program: 37 ºC digest for 1 hour, 65 ºC deactivation for 20 minutes, hold indefinitely at 4 ºC.

7.5.2 Agarose gel DNA electrophoresis

Agarose gel DNA electrophoresis would be used to verify the size and number of fragments from a restriction digest as well as the efficiency of the digest. The agarose gel was prepared by mixing 10 g/L of agarose within 1X TAE buffer. The mixture was heated and stirred until the agarose is fully dissolved. When the agarose solution was cooled to below 50 ºC, 0.5 µL of Nancy Cybersafe Dye was added to 50 mL of the solution. The solution was then stirred and poured into a clamped gel cast with the appropriate size comb and allowed to cool until solid. The clamps were removed, and the gel was submerged in an electrophoresis tank containing 1X TAE buffer, then the comb was carefully removed. 10 µL of DNA sample was mixed with 2 µL of 10X Purple Loading Dye and pipetted into a well of the gel. Once all samples and ladders were loaded into the gel, the electrodes were attached, and the gel was left to run at 100 V for 45 minutes. After this time, the gel was removed from the tank and imaged under UV light to attain an image for analysis using DNA ladders to determine the size of DNA fragments in the samples.

7.5.3 Sanger sequencing

Sanger sequencing was utilised for sequence verification of various constructs, especially assemblies, throughout this thesis. Eurofins Genomics and the services formerly GATC, were the service of choice for all sequence verification. Returned

results were aligned using MAFFT v7 (Katoh and Standley, 2013) and examined using Benchling.

Stock solutions of primers were kept at 100 µM for long-term storage at -20 ºC and at 10 µM as a working stock at 4 ºC. To prepare samples for sequencing 2.5 µL of forward primer was added to a 1.5 mL snap-cap microcentrifuge tube with 250-500 ng of DNA and then made up to a total volume of 10 µL using nuclease-free water. The same procedure was carried out for the reverse primer and for any other number of additional primers.

7.6 Plate Reader Calibration

CLARIOstar Plus plate reader was calibrated using the following calibrants: 10 µM fluorescein, 10 µM cascade blue, 2 µM sulforhodamine 101, 3×10^9 microspheres/mL. Reagents were sourced from the iGEM annual distribution kits.

Serial dilutions of each calibrant were prepared in a black 96-well plate with clear, flat bottoms. Rows A and B contained fluorescein for calibration of GFP; rows C and D contained sulforhodamine-101 for calibration of RFP; rows E and F contained Cascade Blue for calibration of BFP; rows G and H contained microspheres for calibration of cell density. 200 µL of the following calibrants were pipetted into column 1 of the plate: 10 µM fluorescein, 10 µM cascade blue, 2 µM sulforhodamine 101, 3×10^9 microspheres/mL. Following the initial dispense step, the rest of the serial dilution was carried out across the columns of the plate by aspirating 100 µL from the previous column and dispensing into the next column, then mixing in 100 µL of solvent and continuing the steps with the next column. The dilution process continued as before until the penultimate column, where 100 µL is aspirated out of the mixed wells and dispensed into a waste receptacle. Finally, the last column had 100 µL of solvent dispensed into it which serves as a blank.

End-point readings were taken from the plate for the following measurements: green fluorescence at 488 nm excitation, 530 nm emission (band 30 nm); red fluorescence at 561 nm excitation, 610 nm emission (band 20 nm); blue fluorescence at 405 nm excitation, 450 nm emission (band 50 nm); absorbance at 600 nM. The calibration

factors for each calibrant were calculated by plotting a standard curve against the dilutions, assisted with a spreadsheet model provided by the iGEM engineering hub.

7.7 RNA-seq

7.7.1 Computational methods for the processing of RNA-seq reads

RNA-seq reads were run through the nf-core/rnaseq pipeline (Ewels *et al.*, 2020) using a samplesheet to match the paired-end read files (with file extension: ".fastq.gz"). The nf-core/rnaseq pipeline was executed using Illumina iGenomes for the reference genome (EB1 for *E. coli* DH5α, EB2 for *B. subtilis* 168). An additional fasta file was provided for ERCC spike-in controls where relevant. Other than those values specified above, all values and parameters were kept as default for the nf-core/rnaseq workflow, version 3.14 (Ewels *et al.*, 2020).

The nf-core/rnaseq workflow, version 3.14, consists of sixteen steps in five stages. The first stage is pre-processing and quality control (QC): fastq files are merged with cat; sub-sampled and strandedness inferred with fq and Salmon (Patro *et al.*, 2017); QC reports generated with FastQC (Andrews, 2010); Unique Molecular Identifiers (UMIs) extracted by UMI-tools (Smith, Heger and Sudbery, 2017); adapters removed and quality trimming with Trim Galore! (Krueger, 2015); BBSplit was used to remove contaminants; rRNA was removed by SortMeRNA (Kopylova, Noé and Touzet, 2012). Following the pre-processing and QC stage, outputs were produced including QC reports. The second stage is full genome alignment and quantification; by default, STAR (Dobin *et al.*, 2013) was used for alignment against iGenomes reference genomes, and quantification of transcripts was performed by Salmon (Patro *et al.*, 2017). The third stage is optional pseudo-alignment by Salmon or Kallisto; for this work, full alignment was conducted instead of this stage. The fourth stage is post-processing which handles sorting, indexing and de-duplication, and the fifth stage conducts final QC checks.

The gene count files outputted by Salmon were used for downstream analysis. Normalisation was performed by the R package: RUVseq v1.36 (Risso *et al.*, 2014). Where ERCC spike-in controls were utilised, RUVg was the selected method of normalisation.

7.8 Experimental Design and Methods

7.8.1 Detection of SupD tRNA from pInput1 via RNA-seq

Methods are outlined below for the results recorded in chapter 4, section 4.3.2.

E. coli DH5α with pInput1 was grown on LB agar plates with 100 µg/mL of carbenicillin, and *E. coli* DH5α without a plasmid was grown on LB agar plates as a control group. The strains on plates were used to prepare overnight cultures (7.1.2).

One in ten dilutions were prepared in 50 mL of LB broth (with 100 µg/mL of carbenicillin for strains with plasmids) in sterile 250 mL conical flasks with paper bungs and allowed to grow at 37ºC, shaking at 250 rpm in a shaking orbital incubator until absorbance measurements reached an OD_{600} of 0.3. The flasks were then removed from incubation for induction by IPTG. The cultures nominated for induction had IPTG added to a concentration of 1 mM, then all cultures were returned to incubation until an OD_{600} of 0.6. *E. coli* DH5α grows markedly faster than *E. coli* DH5α with pInput1 (figure 7.8.1), so cultures growing with pInput1 were set to grow earlier and allowed more time to reach the correct OD_{600} measurement.

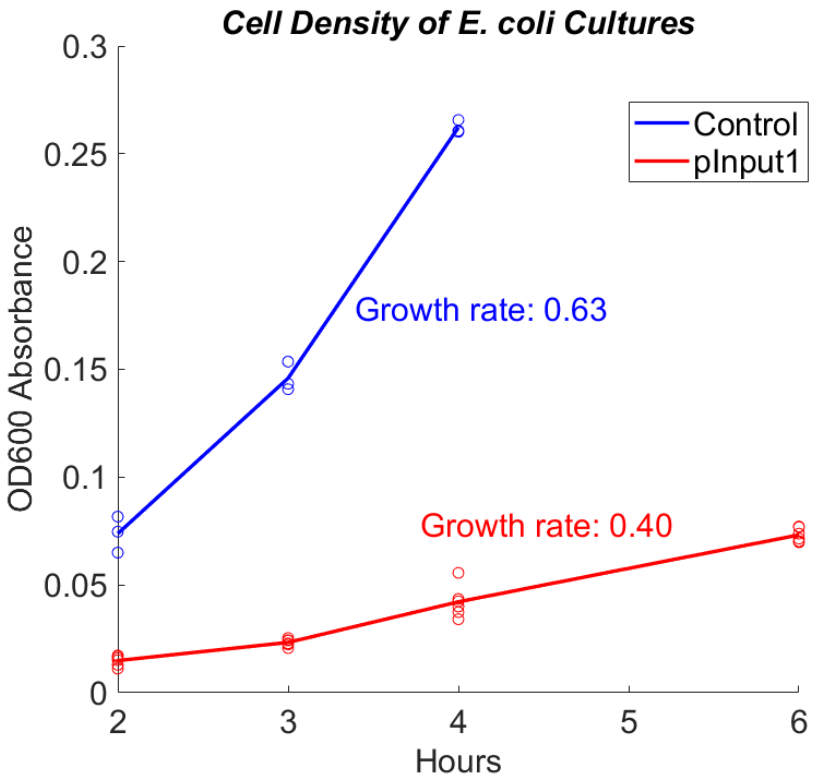

Figure 7.8.1 Growth Rates of E. coli strains used for RNA-seq
OD600 readings taken from 1 mL subsamples of cultures using a spectrophotometer. Control group were cultures of *E. coli* DH5α without plasmid, pInput1 group were cultures of *E. coli* DH5α with pInput1. Approximate growth rates calculated using the log difference in OD600 between the final reading and the initial reading, as plotted.

Aliquots of 1 mL were prepared for each culture and spun down at 16602 x g for one minute to produce a cell pellet. Supernatant was removed from the cell pellet and the pellet was flash frozen using liquid nitrogen. The frozen cell pellets were stored at -80ºC.

RNA extraction, library preparation and sequencing were outsourced to Azenta Life Sciences[9]. Frozen cell pellets were sent to Azenta Life Sciences[9] on dry ice for

[9] Azenta, Inc. Corporate Headquarters, 200 Summit Drive, Burlington, MA 01803 USA

these procedures. RNA-seq raw reads were processed using the nf-core/rnaseq pipeline (see section 7.7 for methods) (Ewels *et al.*, 2020).

7.8.2 RFP Assay for pInput2 in Bacillus subtilis 168

Methods are outlined below for the results recorded in chapter 4, section 4.4.2.

pInput2 and the original vector, pHT01, were transformed according to the protocol in section 7.3.3. *B. subtilis* 168 was grown on LB agar plates or LB agar with 5 µg/mL of chloramphenicol for strains containing plasmids. Overnight cultures were prepared from these plates in LB broth, containing 5 µg/mL of chloramphenicol for strains containing plasmids (7.1.2).

1 µL of overnight culture was inoculated into 99 µL of media in wells in a black 96-well plate with clear flat bottoms. Media used was LB broth, containing 5 µg/mL of chloramphenicol for strains containing plasmids and a concentration of (D)-xylose (0%, 0.4%, or 0.8% w/v) for pInput2. The plate was incubated in a CLARIOstar Plus plate reader at 25ºC, 500 rpm for 24 hours. Measurements were taken every 15 minutes for: absorbance at 600 nm; fluorescence at emission 561 nm (band 20) and excitation 610 nm (band 20), gain 1500.

7.8.3 RFP Assay for pInput2 in Escherichia coli DH5α

Methods are outlined below for the results recorded in chapter 4, section 4.4.3.

pInput2 was transformed into chemically competent cells of *E. coli* DH5α (7.3.2), then prepared as glycerol stocks stored at -80ºC. Streak plates were prepared for *E. coli* DH5α on LB agar plates and for *E. coli* DH5α with pInput2 on LB agar plates with 100 µg/mL of carbenicillin (7.1.3). Overnight cultures were prepared from these plates in LB broth, with 100 µg/mL of carbenicillin for strains containing plasmids (7.1.2).

1 µL of overnight culture was inoculated into 99 µL of media in wells in a black 96-well plate with clear flat bottoms. Media used was LB broth, containing 100 µg/mL of carbenicillin for pInput2, and a given concentration of (D)-xylose: 0%, 0.1%, 1%, or 10% (w/v). The plate was incubated in a CLARIOstar Plus plate reader at 37ºC, 500 rpm for 24 hours. Measurements were taken every 15 minutes for: absorbance

at 600 nm; fluorescence at emission 561 nm (band 20) and excitation 610 nm (band 20), gain 1500.

7.8.4 BFP Assay for pOutput in E. coli BL21(DE3) *and E. coli* DH5α

Methods are outlined below for the results recorded in chapter 4, section 4.5.2.

pOutput was transformed into chemically competent cells of *E. coli* BL21(DE3) and *E. coli* DH5α (7.3.2), then prepared as glycerol stocks stored at -80ºC. Streak plates were prepared for *E. coli* BL21(DE3) and *E. coli* DH5α on LB agar plates and for the same strains with the pOutput on LB agar plates with 100 µg/mL of carbenicillin (7.1.3). Overnight cultures were prepared from these plates in LB broth, with 100 µg/mL of carbenicillin for strains containing plasmids (7.1.2).

1 µL of overnight culture was inoculated into 99 µL of media in wells in a black 96-well plate with clear flat bottoms. Media used was LB broth, containing 100 µg/mL of carbenicillin for pOutput, and a given concentration of IPTG: 0 mM, 0.1 mM, 1 mM, or 10 mM. The plate was incubated in a CLARIOstar Plus plate reader at 37ºC, 500 rpm for 24 hours. Measurements were taken every 15 minutes for: absorbance at 600 nm; fluorescence at emission 402 nm (band 20) and excitation 458 nm (band 30), gain 1000.

7.8.5 Fluorescence assay in E. coli S30 *cell extract expression system*

Methods are outlined below for the results recorded in chapter 4, section 4.6.2.

E. coli S30 extract systems were prepared according to the manufacturer's guidelines but with the total volume reduced from 50 µL to 10 µL per reaction to allow for more reactions to be tested. Each reaction contains: 1 µL of complete amino acid mixture, 4 µL of S30 premix, 3 µL of S30 extract and 2 µL of remaining volume for circular DNA and inducer in nuclease-free water. Each plasmid was prepared so that the total amount per reaction was 1 pmol. Final concentrations for IPTG and (D)-xylose were 1 mM and 1% respectively.

The following reactions were prepared with 4 replicates in a black 384-well plate with clear flat bottoms: *E. coli* S30 extract system without DNA added; *E. coli* T7 S30 extract system without DNA added; *E. coli* S30 extract system with pOutput; *E. coli* T7 S30 extract system with pOutput; *E. coli* S30 extract system with Input 2 +

(D)-xylose; *E. coli* S30 extract system with pInput1, pInput2 & pOutput (AND gate) + IPTG + (D)-xylose; *E. coli* S30 extract system with AND gate + IPTG; *E. coli* S30 extract system with AND gate + (D)-xylose; *E. coli* S30 extract system with AND gate without inducers.

The plate was agitated using a BioShake at 1000 rpm before being incubated at 37ºC, 500 rpm for 10 hours in a CLARIOstar Plus plate reader. Measurements were taken at emission 402 nm (band 20) and excitation 458 nm (band 30), gain 1000 for BFP and excitation 610 nm (band 20), gain 1500 for RFP every 15 minutes. Fluorescence readings were blank corrected using the systems without DNA as a blank before calibration to equivalent number of calibrant molecules (cascade blue for BFP, sulforhodamine-101 for RFP).

7.8.6 GFP Assays for CDS variants generated by Chimera Evolve

Methods are outlined below for the results recorded in chapter 5, section 5.3.

For each strain tested in the experiment, a single colony was selected and grown overnight (7.1.2). 1 µL of overnight culture was diluted with 99 µL LB media (with 100 µg/mL Kanamycin for strains containing plasmids) for a total volume of 100 µL in each well of a black 96-well plate with clear flat-bottomed wells. Plates were sealed using BreatheEasy clear, gas-permeable membranes. Plates were incubated in a CLARIOstar Plus plate reader at 37°C for 24 hours, shaking at 500 rpm in between measurements. Measurements were taken every 15 - 30 minutes for absorbance at 600 nm and fluorescence at excitation 470 nm (band 15 nm) and emission 515 nm (band 20 nm), Gain 1000.

Bibliography

8 Bibliography

Anderson, J. C., Voigt, C. A. and Arkin, A. P. (2007) 'Environmental signal integration by a modular AND gate', *Molecular systems biology,* 3, pp. 133-133.

Andrews, S. 2010. FastQC: a quality control tool for high throughput sequence data. Babraham Bioinformatics, Babraham Institute, Cambridge, United Kingdom.

Angelini, L. L., Dos Santos, R. A. C., Fox, G., Paruthiyil, S., Gozzi, K., Shemesh, M. and Chai, Y. (2023) 'Pulcherrimin protects *Bacillus subtilis* against oxidative stress during biofilm development', *npj Biofilms and Microbiomes,* 9(1).

Atanassov, I., Stefanova, K., Tomova, I. and Kamburova, M. (2013) 'Seamless GFP and GFP-Amylase Cloning in Gateway Shuttle Vector, Expression of the Recombinant Proteins in *E. Coli* and *Bacillus Megaterium* and Assessment of the GFP-Amylase Thermostability', *Biotechnology & Biotechnological Equipment,* 27(5), pp. 4172-4180.

Bahiri-Elitzur, S. and Tuller, T. (2021) 'Codon-based indices for modeling gene expression and transcript evolution', *Computational and Structural Biotechnology Journal,* 19, pp. 2646-2663.

Bandiera, L., Gomez-Cabeza, D., Gilman, J., Balsa-Canto, E. and Menolascina, F. (2020) 'Optimally Designed Model Selection for Synthetic Biology', *ACS synthetic biology.,* 9(11), pp. 3134-3144.

Beales, N. (2004) 'Adaptation of microorganisms to cold temperatures, weak acid preservatives, low pH, and osmotic stress: A review', *Comprehensive reviews in food science and food safety.,* 3(1), pp. 1-20.

Bird, J. E., Marles-Wright, J. and Giachino, A. (2022) 'A User's Guide to Golden Gate Cloning Methods and Standards', *ACS Synthetic Biology,* 11(11), pp. 3551-3563.

Bibliography

Bonilla, C. Y. (2020) 'Generally stressed out bacteria: Environmental stress response mechanisms in gram-positive bacteria', *Integrative and Comparative Biology,* 60(1), pp. 126-133.

Boo, A., Ellis, T. and Stan, G.-B. (2019) 'Host-aware synthetic biology', *Current Opinion in Systems Biology,* 14, pp. 66-72.

Bose, D., Roy, A., Roy, L. and Chatterjee, S. (2023) 'Nucleic Acid Sensors and Logic Gates', *Nucleic Acid Biology and its Application in Human Diseases*: Springer Nature Singapore, pp. 271-319.

Boël, G., Letso, R., Neely, H., Price, W. N., Wong, K.-H., Su, M., Luff, J. D., Valecha, M., Everett, J. K., Acton, T. B., Xiao, R., Montelione, G. T., Aalberts, D. P. and Hunt, J. F. (2016) 'Codon influence on protein expression in *E. coli* correlates with mRNA levels', *Nature,* 529(7586), pp. 358-363.

Bradley, R. W., Buck, M. and Wang, B. (2016) 'Recognizing and engineering digital-like logic gates and switches in gene regulatory networks', *Current opinion in microbiology.,* 33, pp. 74-82.

Brauer, A. M., Shi, H., Levin, P. A. and Huang, K. C. (2023) 'Physiological and regulatory convergence between osmotic and nutrient stress responses in microbes', *Current opinion in cell biology.,* 81, pp. 102170.

Brooks, S. M. and Alper, H. S. (2021) 'Applications, challenges, and needs for employing synthetic biology beyond the lab', *Nature Communications,* 12(1).

Brookwell, A., Oza, J. P. and Caschera, F. (2021) 'Biotechnology Applications of Cell-Free Expression Systems', *Life,* 11(12), pp. 1367.

Brophy, J. A. N. and Voigt, C. A. (2014) 'Principles of genetic circuit design', *Nature Methods,* 11(5), pp. 508-520.

Browning, D. F., Godfrey, R. E., Richards, K. L., Robinson, C. and Busby, S. J. W. (2019) 'Exploitation of the *Escherichia coli lac* operon promoter for controlled

recombinant protein production', *Biochemical Society Transactions,* 47(2), pp. 755-763.

Buescher, J. M., Liebermeister, W., Jules, M., Uhr, M., Muntel, J., Botella, E., Hessling, B., Kleijn, R. J., Le Chat, L., Lecointe, F., Mäder, U., Nicolas, P., Piersma, S., Rügheimer, F., Becher, D., Bessieres, P., Bidnenko, E., Denham, E. L., Dervyn, E., Devine, K. M., Doherty, G., Drulhe, S., Felicori, L., Fogg, M. J., Goelzer, A., Hansen, A., Harwood, C. R., Hecker, M., Hubner, S., Hultschig, C., Jarmer, H., Klipp, E., Leduc, A., Lewis, P., Molina, F., Noirot, P., Peres, S., Pigeonneau, N., Pohl, S., Rasmussen, S., Rinn, B., Schaffer, M., Schnidder, J., Schwikowski, B., Van Dijl, J. M., Veiga, P., Walsh, S., Wilkinson, A. J., Stelling, J., Aymerich, S. and Sauer, U. (2012) 'Global Network Reorganization During Dynamic Adaptations of *Bacillus subtilis* Metabolism', *Science,* 335(6072), pp. 1099-1103.

Cabeen, M. T., Russell, J. R., Paulsson, J. and Losick, R. (2017) 'Use of a microfluidic platform to uncover basic features of energy and environmental stress responses in individual cells of *Bacillus subtilis*', *PLoS Genetics,* 13(7).

Cai, Y.-M., Carrasco Lopez, J. A. and Patron, N. J. (2020) 'Phytobricks: Manual and Automated Assembly of Constructs for Engineering Plants', *Methods in Molecular Biology*: Springer US, pp. 179-199.

Campbell, B. C., Nabel, E. M., Murdock, M. H., Lao-Peregrin, C., Tsoulfas, P., Blackmore, M. G., Lee, F. S., Liston, C., Morishita, H. and Petsko, G. A. (2020) 'mGreenLantern: a bright monomeric fluorescent protein with rapid expression and cell filling properties for neuronal imaging', *Proceedings of the National Academy of Sciences,* 117(48), pp. 30710-30721.

Campbell, R. E., Tour, O., Palmer, A. E., Steinbach, P. A., Baird, G. S., Zacharias, D. A. and Tsien, R. Y. (2002) 'A monomeric red fluorescent protein', *Proceedings of the National Academy of Sciences,* 99(12), pp. 7877-7882.

Casas, A. I., Hassan, A. A., Manz, Q., Wiwie, C., Kleikers, P., Egea, J., López, M. G., List, M., Baumbach, J. and Schmidt, H. H. H. W. (2022) 'Un-biased

housekeeping gene panel selection for high-validity gene expression analysis', *Scientific Reports,* 12(1).

Cazier, A. P. and Blazeck, J. (2021) 'Advances in promoter engineering: Novel applications and predefined transcriptional control', *Biotechnology journal.,* 16(10), pp. 2100239.

Ceroni, F., Algar, R., Stan, G. B. and Ellis, T. (2015) 'Quantifying cellular capacity identifies gene expression designs with reduced burden', *Nature Methods,* 12(5), pp. 415-418.

Ceroni, F., Boo, A., Furini, S., Gorochowski, T. E., Borkowski, O., Ladak, Y. N., Awan, A. R., Gilbert, C., Stan, G. B. and Ellis, T. (2018) 'Burden-driven feedback control of gene expression', *Nature Methods,* 15(5), pp. 387-393.

Chotani, G., Dodge, T., Hsu, A., Kumar, M., LaDuca, R., Trimbur, D., Weyler, W. and Sanford, K. (2000) 'The commercial production of chemicals using pathway engineering', *Biochimica et Biophysica Acta - Protein Structure and Molecular Enzymology,* 1543(2), pp. 434-455.

Conrad, B., Savchenko, R. S., Breves, R. and Hofemeister, J. (1996) 'A T7 promoter-specific, inducible protein expression system for *Bacillus subtilis*', *Molecular and General Genetics MGG,* 250(2), pp. 230-236.

Corchete, L. A., Rojas, E. A., Alonso-López, D., De Las Rivas, J., Gutiérrez, N. C. and Burguillo, F. J. (2020) 'Systematic comparison and assessment of RNA-seq procedures for gene expression quantitative analysis', *Scientific Reports,* 10(1).

Cordell, W. T., Avolio, G., Takors, R. and Pfleger, B. F. (2023) 'Milligrams to kilograms: making microbes work at scale', *Trends in biotechnology.,* 41(11), pp. 1442-1457.

Cubillos-Ruiz, A., Guo, T., Sokolovska, A., Miller, P. F., Collins, J. J., Lu, T. K. and Lora, J. M. (2021) 'Engineering living therapeutics with synthetic biology', *Nature Reviews Drug Discovery,* 20(12), pp. 941-960.

Bibliography

Dahl, R. H., Zhang, F., Alonso-Gutierrez, J., Baidoo, E., Batth, T. S., Redding-Johanson, A. M., Petzold, C. J., Mukhopadhyay, A., Lee, T. S., Adams, P. D. and Keasling, J. D. (2013) 'Engineering dynamic pathway regulation using stress-response promoters', *Nature Biotechnology,* 31(11), pp. 1039-1046.

Darlington, A. P. S., Kim, J., Jiménez, J. I. and Bates, D. G. (2017) 'Design of a translation resource allocation controller to manage cellular resource limitations', *IFAC-PapersOnLine,* 50(1), pp. 12653-12660.

de Lorenzo, V., Krasnogor, N. and Schmidt, M. (2021) 'For the sake of the Bioeconomy: define what a Synthetic Biology Chassis is!', *New biotechnology.,* 60, pp. 44-51.

De Nadal, E., Ammerer, G. and Posas, F. (2011) 'Controlling gene expression in response to stress', *Nature Reviews Genetics,* 12(12), pp. 833-845.

den Besten, H. M. W., Arvind, A., Gaballo, H. M. S., Moezelaar, R., Zwietering, M. H. and Abee, T. (2010) 'Short- and long-term biomarkers for bacterial robustness: A framework for quantifying correlations between cellular indicators and adaptive behavior', *PLoS ONE,* 5(10).

Dessì, N., Pascariello, E. and Pes, B. (2013) 'A Comparative Analysis of Biomarker Selection Techniques', *BioMed Research International,* 2013, pp. 1-10.

Diament, A., Weiner, I., Shahar, N., Landman, S., Feldman, Y., Atar, S., Avitan, M., Schweitzer, S., Yacoby, I., Tuller, T. and Hancock, J. (2019) 'ChimeraUGEM: unsupervised gene expression modeling in any given organism', *Bioinformatics.,* 35(18), pp. 3365-3371.

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M. and Gingeras, T. R. (2013) 'STAR: ultrafast universal RNA-seq aligner', *Bioinformatics.,* 29(1), pp. 15-21.

Dragosits, M., Nicklas, D. and Tagkopoulos, I. (2012) 'A synthetic biology approach to self-regulatory recombinant protein production in *Escherichia coli*', *Journal of Biological Engineering,* 6(1), pp. 2.

Bibliography

El Naqa, I. and Murphy, M. J. (2015) 'What Is Machine Learning?', *Machine Learning in Radiation Oncology*: Springer International Publishing, pp. 3-11.

Endy, D. (2005) 'Foundations for engineering biology', *Nature,* 438(7067), pp. 449-453.

Errington, J. (1993) '*Bacillus subtilis* sporulation: Regulation of gene expression and control of morphogenesis', *Microbiological Reviews,* 57(1), pp. 1-33.

Ewels, P., Magnusson, M., Lundin, S. and Käller, M. (2016) 'MultiQC: summarize analysis results for multiple tools and samples in a single report', *Bioinformatics,* 32(19), pp. 3047-3048.

Ewels, P. A., Peltzer, A., Fillinger, S., Patel, H., Alneberg, J., Wilm, A., Garcia, M. U., Di Tommaso, P. and Nahnsen, S. (2020) 'The nf-core framework for community-curated bioinformatics pipelines', *Nature Biotechnology,* 38(3), pp. 276-278.

Fages-Lartaud, M., Tietze, L., Elie, F., Lale, R. and Hohmann-Marriott, M. F. (2022) 'mCherry contains a fluorescent protein isoform that interferes with its reporter function', *Front Bioeng Biotechnol,* 10, pp. 892138.

Failmezger, J., Rauter, M., Nitschel, R., Kraml, M. and Siemann-Herzberg, M. (2017) 'Cell-free protein synthesis from non-growing, stressed *Escherichia coli*', *Scientific Reports,* 7(1).

Faria, J. P., Overbeek, R., Taylor, R. C., Conrad, N., Vonstein, V., Goelzer, A., Fromion, V., Rocha, M., Rocha, I. and Henry, C. S. (2016) 'Reconstruction of the Regulatory Network for *Bacillus subtilis* and Reconciliation with Gene Expression Data', *Frontiers in Microbiology,* 7.

Foley, M. H., Cockburn, D. W. and Koropatkin, N. M. (2016) 'The Sus operon: a model system for starch uptake by the human gut Bacteroidetes', *Cellular and Molecular Life Sciences,* 73(14), pp. 2603-2617.

Fox, J. M. and Erill, I. (2010) 'Relative Codon Adaptation: A Generic Codon Bias Index for Prediction of Gene Expression', *DNA research :,* 17(3), pp. 185-196.

Bibliography

Freemont, P. S. (2019) 'Synthetic biology industry: data-driven design is creating new opportunities in biotechnology', *Emerging Topics in Life Sciences,* 3(5), pp. 651-657.

Freire, V., del Río, J., Gómara, P., Salvador, M., Condón, S. and Gayán, E. (2023) 'Comparative study on the impact of equally stressful environmental sporulation conditions on thermal inactivation kinetics of *B. subtilis* spores', *International journal of food microbiology.,* 405, pp. 110349.

Fu, H., Liang, Y., Zhong, X., Pan, Z., Huang, L., Zhang, H., Xu, Y., Zhou, W. and Liu, Z. (2020) 'Codon optimization with deep learning to enhance protein expression', *Scientific Reports,* 10(1).

Fujiyoshi, K., Bruford, E. A., Mroz, P., Sims, C. L., O'Leary, T. J., Lo, A. W. I., Chen, N., Patel, N. R., Patel, K. P., Seliger, B., Song, M., Monzon, F. A., Carter, A. B., Gulley, M. L., Mockus, S. M., Phung, T. L., Feilotter, H., Williams, H. E. and Ogino, S. (2021) 'Standardizing gene product nomenclature—a call to action', *Proceedings of the National Academy of Sciences,* 118(3), pp. e2025207118.

Gardner, T. S., Cantor, C. R. and Collins, J. J. (2000) 'Construction of a genetic toggle switch in *Escherichia coli*', *Nature,* 403(6767), pp. 339-342.

Geissler, A. S., Poulsen, L. D., Doncheva, N. T., Anthon, C., Seemann, S. E., González-Tortuero, E., Breüner, A., Jensen, L. J., Hjort, C., Vinther, J. and Gorodkin, J. (2022) 'The impact of PrsA over-expression on the *Bacillus subtilis* transcriptome during fed-batch fermentation of alpha-amylase production', *Frontiers in Microbiology,* 13.

Ghataora, J. S., Gebhard, S. and Reeksting, B. J. (2023) 'Chimeric MerR-Family Regulators and Logic Elements for the Design of Metal Sensitive Genetic Circuits in *Bacillus subtilis*', *ACS synthetic biology.,* 12(3), pp. 735-749.

Ghavim, M., Abnous, K., Arasteh, F., Taghavi, S., Nabavinia, M., Alibolandi, M. and Ramezani, M. (2017) 'High level expression of recombinant human growth hormone

in *Escherichia coli*: crucial role of translation initiation region', *Research in pharmaceutical sciences.,* 12(2), pp. 168.

Gibson, D. G., Young, L., Chuang, R.-Y., Venter, J. C., Hutchison, C. A. and Smith, H. O. (2009) 'Enzymatic assembly of DNA molecules up to several hundred kilobases', *Nature Methods,* 6(5), pp. 343-345.

Glick, B. R. (1995) 'Metabolic load and heterologous gene expression', *Biotechnology Advances,* 13(2), pp. 247-261.

Goñi-Moreno, A. and Amos, M. (2012) 'A reconfigurable NAND/NOR genetic logic gate', *BMC Systems Biology,* 6(1), pp. 126.

Goñi-Moreno, A. and Nikel, P. I. (2019) 'High-Performance Biocomputing in Synthetic Biology–Integrated Transcriptional and Metabolic Circuits', *Frontiers in Bioengineering and Biotechnology,* 7.

Grätz, C., Bui, M. L. U., Thaqi, G., Kirchner, B., Loewe, R. P. and Pfaffl, M. W. (2022) 'Obtaining Reliable RT-qPCR Results in Molecular Diagnostics—MIQE Goals and Pitfalls for Transcriptional Biomarker Discovery', *Life,* 12(3), pp. 386.

Guan, N., Li, J., Shin, H. D., Du, G., Chen, J. and Liu, L. (2017) 'Microbial response to environmental stresses: from fundamental mechanisms to practical applications', *Applied Microbiology and Biotechnology,* 101(10), pp. 3991-4008.

Guo, M. S. and Gross, C. A. (2014) 'Stress-induced remodeling of the bacterial proteome', *Current Biology,* 24(10), pp. R424-R434.

Guyon, I., Weston, J., Barnhill, S. and Vapnik, V. (2002) *Machine Learning,* 46(1/3), pp. 389-422.

Hanson, G. and Coller, J. (2018) 'Codon optimality, bias and usage in translation and mRNA decay', *Nature Reviews Molecular Cell Biology,* 19(1), pp. 20-30.

Hansson, O. (2021) 'Biomarkers for neurodegenerative diseases', *Nature Medicine,* 27(6), pp. 954-963.

Bibliography

Hicks, M., Bachmann, T. T. and Wang, B. (2020) 'Synthetic Biology Enables Programmable Cell-Based Biosensors', *ChemPhysChem,* 21(2), pp. 132-144.

Hitchcock, A., Hunter, C. N. and Canniffe, D. P. (2020) 'Progress and challenges in engineering cyanobacteria as chassis for light-driven biotechnology', *Microbial Biotechnology,* 13(2), pp. 363-367.

Hoffman, E. P. and Wilhelm, R. C. (1970) 'Genetic Mapping and Dominance of the Amber Suppressor, *Su1* (*supD*), in *Escherichia coli* K-12', *Journal of Bacteriology,* 103(1), pp. 32-36.

Huang, Y., Sinha, N., Wipat, A. and Bacardit, J. (2023) 'A knowledge integration strategy for the selection of a robust multi-stress biomarkers panel for *Bacillus subtilis*', *Synthetic and systems biotechnology,* 8(1), pp. 97-106.

Huang, Y., Smith, W., Harwood, C., Wipat, A. and Bacardit, J. (2021) 'Computational Strategies for the Identification of a Transcriptional Biomarker Panel to Sense Cellular Growth States in *Bacillus subtilis*', *Sensors (Basel, Switzerland),* 21(7), pp. 2436.

Ikemura, T. (1981) 'Correlation between the abundance of *Escherichia coli* transfer RNAs and the occurrence of the respective codons in its protein genes: A proposal for a synonymous codon choice that is optimal for the *E. coli* translational system', *Journal of molecular biology.,* 151(3), pp. 389-409.

Ingram, L. O., Jarboe, L. R., Zhang, X., Wang, X., Moore, J. C. and Shanmugam, K. T. (2010) 'Metabolic engineering for production of biorenewable fuels and chemicals: Contributions of synthetic biology', *Journal of Biomedicine and Biotechnology,* 2010.

Israr, J., Alam, S., Siddiqui, S., Misra, S., Gupta, D. and Kumar, A. (2024) 'Recent Progress in Microarray and its Role in Genomics', *Advances in Genomics*: Springer Nature Singapore, pp. 199-212.

Bibliography

Jain, R., Jain, A., Mauro, E., Leshane, K. and Densmore, D. (2023) 'ICOR: improving codon optimization with recurrent neural networks', *BMC Bioinformatics,* 24(1).

Ji, M., Li, S., Chen, A., Liu, Y., Xie, Y., Duan, H., Shi, J. and Sun, J. (2021) 'A wheat bran inducible expression system for the efficient production of α-L-arabinofuranosidase in *Bacillus subtilis*', *Enzyme and microbial technology.,* 144, pp. 109726.

Jia, H., Liang, T., Wang, Z., He, Z., Liu, Y., Yang, L., Zeng, Y., Liu, S., Tang, L., Wang, J., Chen, Y. and Xie, Z. (2014) 'Multistage Regulator Based on Tandem Promoters and CRISPR/Cas', *ACS Synthetic Biology,* 3(12), pp. 1007-1010.

Jiang, L., Guo, Y., Yu, H., Hoff, K., Ding, X., Zhou, W. and Edwards, J. (2021) 'Detecting SARS-CoV-2 and its variant strains with a full genome tiling array', *Briefings in Bioinformatics,* 22(6).

Jiang, T., Li, C., Teng, Y., Zhang, R. and Yan, Y. (2020) 'Recent advances in improving metabolic robustness of microbial cell factories', *Current opinion in biotechnology.,* 66, pp. 69-77.

Johnson, J. S., Spakowicz, D. J., Hong, B.-Y., Petersen, L. M., Demkowicz, P., Chen, L., Leopold, S. R., Hanson, B. M., Agresta, H. O., Gerstein, M., Sodergren, E. and Weinstock, G. M. (2019) 'Evaluation of 16S rRNA gene sequencing for species and strain-level microbiome analysis', *Nature Communications,* 10(1).

Katoh, K. and Standley, D. M. (2013) 'MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability', *Molecular Biology and Evolution,* 30(4), pp. 772-780.

Katz, L., Chen, Y. Y., Gonzalez, R., Peterson, T. C., Zhao, H. and Baltz, R. H. (2018) 'Synthetic biology advances and applications in the biotechnology industry: a perspective', *Journal of Industrial Microbiology and Biotechnology,* 45(7), pp. 449-461.

Bibliography

Kim, L., Mogk, A. and Schumann, W. (1996) 'A xylose-inducible *Bacillus subtilis* integration vector and its application', *Gene.,* 181(1-2), pp. 71-76.

Kim, S. G., Noh, M. H., Lim, H. G., Jang, S., Jang, S., Koffas, M. A. G. and Jung, G. Y. (2018) 'Molecular parts and genetic circuits for metabolic engineering of microorganisms', *FEMS Microbiology Letters,* 365(17).

Kopylova, E., Noé, L. and Touzet, H. (2012) 'SortMeRNA: fast and accurate filtering of ribosomal RNAs in metatranscriptomic data', *Bioinformatics,* 28(24), pp. 3211-3217.

Koreeda, A., Taguchi, R., Miyamoto, K., Kuwahara, Y. and Hirooka, K. (2023) 'Protein expression systems combining a flavonoid-inducible promoter and T7 RNA polymerase in *Bacillus subtilis*', *Bioscience, Biotechnology, and Biochemistry,* 87(9), pp. 1017-1028.

Kraikivski, P. (2021) 'The lac Operon', *Case Studies in Systems Biology*: Springer International Publishing, pp. 137-147.

Krasnogor, N., Zuliani, P., Bacardit, J., Zenkin, N., Daniel, R. A., Lord, P., Yuzenkova, Y., Murray, H., Woods, S., Kaiser, M. and Wipat, A. (2023) *Synthetic Portabolomics: Leading the way at the crossroads of the Digital and the Bio Economies*. Available at: http://portabolomics.ico2s.org 2023).

Krueger, F. (2015) 'Trim Galore!: A wrapper around Cutadapt and FastQC to consistently apply adapter and quality trimming to FastQ files, with extra functionality for RRBS data', *Babraham Institute*.

Kurland, C. G. and Dong, H. (1996) 'Bacterial growth inhibition by overproduction of protein', *Molecular Microbiology,* 21(1), pp. 1-4.

Lafleur, T. L., Hossain, A. and Salis, H. M. (2022) 'Automated model-predictive design of synthetic promoters to control transcriptional profiles in bacteria', *Nature Communications,* 13(1).

Bibliography

Lambert, T. J. (2019) 'FPbase: a community-editable fluorescent protein database', *Nature Methods,* 16(4), pp. 277-278.

Lazzarini, N. and Bacardit, J. (2017) 'RGIFE: a ranked guided iterative feature elimination heuristic for the identification of biomarkers', *BMC Bioinformatics,* 18(1).

Lebovich, M., Zeng, M. and Andrews, L. B. (2023) 'Algorithmic Programming of Sequential Logic and Genetic Circuits for Recording Biochemical Concentration in a Probiotic Bacterium', *ACS synthetic biology.,* 12(9), pp. 2632-2649.

Lewis, P. J., Doherty, G. P. and Clarke, J. (2008) 'Transcription factor dynamics', *Microbiology.,* 154(7), pp. 1837-1844.

Liao, Y., Smyth, G. K. and Shi, W. (2014) 'featureCounts: an efficient general purpose program for assigning sequence reads to genomic features', *Bioinformatics,* 30(7), pp. 923-930.

Lipinszki, Z., Vernyik, V., Farago, N., Sari, T., Puskas, L. G., Blattner, F. R., Posfai, G. and Gyorfy, Z. (2018) 'Enhancing the Translational Capacity of *E. coli* by Resolving the Codon Bias', *ACS Synthetic Biology,* 7(11), pp. 2656-2664.

Liu, J., Wang, X., Dai, G., Zhang, Y. and Bian, X. (2022) 'Microbial chassis engineering drives heterologous production of complex secondary metabolites', *Biotechnology Advances,* 59, pp. 107966.

Lo, T. M., Chng, S. H., Teo, W. S., Cho, H. S. and Chang, M. W. (2016) 'A Two-Layer Gene Circuit for Decoupling Cell Growth from Metabolite Production', *Cell Systems,* 3(2), pp. 133-143.

Lowe, R., Shirley, N., Bleackley, M., Dolan, S. and Shafee, T. (2017) 'Transcriptomics technologies', *PLOS Computational Biology,* 13(5), pp. e1005457.

Lucchini, S., Thompson, A. and Hinton, J. C. D. (2001) 'Microarrays for microbiologists', *Microbiology.,* 147(6), pp. 1403-1414.

Lund, P. A., De Biase, D., Liran, O., Scheler, O., Mira, N. P., Cetecioglu, Z., Fernández, E. N., Bover-Cid, S., Hall, R., Sauer, M. and O'Byrne, C. (2020)

Bibliography

'Understanding How Microorganisms Respond to Acid pH Is Central to Their Control and Successful Exploitation', *Frontiers in Microbiology,* 11.

Lynch, M. and Marinov, G. K. (2015) 'The bioenergetic costs of a gene', *Proceedings of the National Academy of Sciences of the United States of America,* 112(51), pp. 15690-15695.

Ma, S., Su, T., Lu, X. and Qi, Q. (2024) 'Bacterial genome reduction for optimal chassis of synthetic biology: a review', *Critical reviews in biotechnology.,* 44(4), pp. 660-673.

Maharjan, A. and Park, J. H. (2023) 'Cell-free protein synthesis system: A new frontier for sustainable biotechnology-based products', *Biotechnology and Applied Biochemistry,* 70(6), pp. 2136-2149.

Mandair, D., Reis-Filho, J. S. and Ashworth, A. (2023) 'Biological insights and novel biomarker discovery through deep learning approaches in breast cancer histopathology', *npj Breast Cancer,* 9(1).

McInnes, L., Healy, J. and Melville, J. (2020) 'UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction', Available at: arXiv.

Mockler, T. C. and Ecker, J. R. (2005) 'Applications of DNA tiling arrays for whole-genome analysis', *Genomics.,* 85(1), pp. 1-15.

Moon, S., Ham, S., Jeong, J., Ku, H., Kim, H. and Lee, C. (2023) 'Temperature Matters: Bacterial Response to Temperature Change', *Journal of Microbiology,* 61(3), pp. 343-357.

Moser, F., Broers, N. J., Hartmans, S., Tamsir, A., Kerkman, R., Roubos, J. A., Bovenberg, R. and Voigt, C. A. (2012) 'Genetic circuit performance under conditions relevant for industrial bioreactors', *ACS Synthetic Biology,* 1(11), pp. 555-564.

Mugwanda, K., Hamese, S., Van Zyl, F., Winschau, Prinsloo, E., Du Plessis, M., Dicks, M. T., Leon and Thimiri Govinda Raj, B., Deepak (2023) 'Recent advances

in genetic tools for engineering probiotic lactic acid bacteria', *Bioscience Reports,* 43(1).

Mutlu, A., Kaspar, C., Becker, N. and Bischofs, I. B. (2020) 'A spore quality–quantity tradeoff favors diverse sporulation strategies in *Bacillus subtilis*', *The ISME Journal,* 14(11), pp. 2703-2714.

Naseri, G. and Koffas, M. A. G. (2020) 'Application of combinatorial optimization strategies in synthetic biology', *Nature Communications,* 11(1).

Nasu, Y., Shen, Y., Kramer, L. and Campbell, R. E. (2021) 'Structure- and mechanism-guided design of single fluorescent protein-based biosensors', *Nature Chemical Biology,* 17(5), pp. 509-518.

Negi, A., Shukla, A., Jaiswar, A., Shrinet, J. and Jasrotia, R. S. (2022) 'Applications and challenges of microarray and RNA-sequencing', pp. 91-103.

Nguyen, H. D., Nguyen, Q. A., Ferreira, R. C., Ferreira, L. C. S., Tran, L. T. and Schumann, W. (2005) 'Construction of plasmid-based expression vectors for *Bacillus subtilis* exhibiting full structural stability', *Plasmid :,* 54(3), pp. 241-248.

Nguyen, H. D., Phan, T. T. P. and Schumann, W. (2007) 'Expression Vectors for the Rapid Purification of Recombinant Proteins in *Bacillus subtilis*', *Current Microbiology,* 55(2), pp. 89-93.

Nicolas, P., Mäder, U., Dervyn, E., Rochat, T., Leduc, A., Pigeonneau, N., Bidnenko, E., Marchadier, E., Hoebeke, M., Aymerich, S., Becher, D., Bisicchia, P., Botella, E., Delumeau, O., Doherty, G., Denham, E. L., Fogg, M. J., Fromion, V., Goelzer, A., Hansen, A., Härtig, E., Harwood, C. R., Homuth, G., Jarmer, H., Jules, M., Klipp, E., Le Chat, L., Lecointe, F., Lewis, P., Liebermeister, W., March, A., Mars, R. A. T., Nannapaneni, P., Noone, D., Pohl, S., Rinn, B., Rügheimer, F., Sappa, P. K., Samson, F., Schaffer, M., Schwikowski, B., Steil, L., Stülke, J., Wiegert, T., Devine, K. M., Wilkinson, A. J., Van Dijl, J. M., Hecker, M., Völker, U., Bessières, P. and Noirot, P. (2012) 'Condition-dependent transcriptome reveals high-level regulatory architecture in *Bacillus subtilis*', *Science,* 335(6072), pp. 1103-1106.

Bibliography

Njenga, R., Boele, J., Öztürk, Y. and Koch, H.-G. (2023) 'Coping with stress: How bacteria fine-tune protein synthesis and protein transport', *Journal of biological chemistry.,* 299(9), pp. 105163.

Nyerges, Á., Csörgő, B., Nagy, I., Bálint, B., Bihari, P., Lázár, V., Apjok, G., Umenhoffer, K., Bogos, B., Pósfai, G. and Pál, C. (2016) 'A highly precise and portable genome engineering method allows comparison of mutational effects across bacterial species', *Proceedings of the National Academy of Sciences,* 113(9), pp. 2502-2507.

O'Hara, S., Wang, K., Slayden, R. A., Schenkel, A. R., Huber, G., O'Hern, C. S., Shattuck, M. D. and Kirby, M. (2013) 'Iterative feature removal yields highly discriminative pathways', *BMC Genomics,* 14(1), pp. 832.

Parvathy, S. T., Udayasuriyan, V. and Bhadana, V. (2022) 'Codon usage bias', *Molecular Biology Reports,* 49(1), pp. 539-565.

Pasini, M., Fernández-Castané, A., Jaramillo, A., de Mas, C., Caminal, G. and Ferrer, P. (2016) 'Using promoter libraries to reduce metabolic burden due to plasmid-encoded proteins in recombinant *Escherichia coli*', *New Biotechnology,* 33(1), pp. 78-90.

Patro, R., Duggal, G., Love, M. I., Irizarry, R. A. and Kingsford, C. (2017) 'Salmon provides fast and bias-aware quantification of transcript expression', *Nature Methods,* 14(4), pp. 417-419.

Pearson, H. (2006) 'What is a gene?', *Nature,* 441(7092), pp. 398-401.

Petersohn, A., Brigulla, M., Haas, S., Hoheisel, J. D., Völker, U. and Heckler, M. (2001) 'Global analysis of the general stress response of *Bacillus subtilis*', *Journal of Bacteriology,* 183(19), pp. 5617-5631.

Peña-Montenegro, T. D., Kleindienst, S., Allen, A. E., Eren, A. M., McCrow, J. P., Sánchez-Calderón, J. D., Arnold, J. and Joye, S. B. (2023) 'Species-specific responses of marine bacteria to environmental perturbation', *ISME Communications,* 3(1).

Bibliography

Phan, T. T. P., Nguyen, H. D. and Schumann, W. (2006) 'Novel plasmid-based expression vectors for intra- and extracellular production of recombinant proteins in *Bacillus subtilis*', *Protein expression and purification.,* 46(2), pp. 189-195.

Pollak, B., Cerda, A., Delmans, M., Álamos, S., Moyano, T., West, A., Gutiérrez, R. A., Patron, N. J., Federici, F. and Haseloff, J. (2019) 'Loop assembly: a simple and open system for recursive fabrication of DNA circuits', *New Phytologist,* 222(1), pp. 628-640.

Pratt, S. D., David, C. A., Black-Schaefer, C., Dandliker, P. J., Xuei, X., Warrior, U., Burns, D. J., Zhong, P., Cao, Z., Saiki, A. Y., Lerner, C. G., Chovan, L. E., Soni, N. B., Nilius, A. M., Wagenaar, F. L., Merta, P. J., Traphagen, L. M. and Beutel, B. A. (2004) 'A strategy for discovery of novel broad-spectrum antibacterials using a high-throughput *Streptococcus pneumoniae* transcription/translation screen', *J Biomol Screen,* 9(1), pp. 3-11.

Pédelacq, J.-D., Cabantous, S., Tran, T., Terwilliger, T. C. and Waldo, G. S. (2006) 'Engineering and characterization of a superfolder green fluorescent protein', *Nature Biotechnology,* 24(1), pp. 79-88.

Rath, S. and Das, S. (2023) 'Oxidative stress-induced DNA damage and DNA repair mechanisms in mangrove bacteria exposed to climatic and heavy metal stressors', *Environmental pollution.,* 339, pp. 122722.

Reder, A., Albrecht, D., Gerth, U. and Hecker, M. (2012) 'Cross-talk between the general stress response and sporulation initiation in *Bacillus subtilis* - the σB promoter of *spo0E* represents an AND-gate', *Environmental Microbiology,* 14(10), pp. 2741-2756.

Reeve, B., Hargest, T., Gilbert, C. and Ellis, T. (2014) 'Predicting Translation Initiation Rates for Designing Synthetic Biology', *Frontiers in Bioengineering and Biotechnology,* 2.

Bibliography

Risso, D., Ngai, J., Speed, T. P. and Dudoit, S. (2014) 'Normalization of RNA-seq data using factor analysis of control genes or samples', *Nature Biotechnology,* 32(9), pp. 896-902.

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C. and Müller, M. (2013) 'PanelomiX: A threshold-based algorithm to create panels of biomarkers', 1(1), pp. 57-64.

Rodnina, M. V. (2016) 'The ribosome in action: Tuning of translational efficiency and protein folding', *Protein Science,* 25(8), pp. 1390-1406.

Rodriguez Ayala, F., Bartolini, M. and Grau, R. (2020) 'The Stress-Responsive Alternative Sigma Factor SigB of *Bacillus subtilis* and Its Relatives: An Old Friend With New Functions', *Frontiers in Microbiology,* 11.

Ruijter, J. M., Pfaffl, M. W., Zhao, S., Spiess, A. N., Boggy, G., Blom, J., Rutledge, R. G., Sisti, D., Lievens, A., De Preter, K., Derveaux, S., Hellemans, J. and Vandesompele, J. (2013) 'Evaluation of qPCR curve analysis methods for reliable biomarker discovery: Bias, resolution, precision, and implications', *Methods.,* 59(1), pp. 32-46.

Røkke, G., Korvald, E., Pahr, J., Øyås, O. and Lale, R. (2014) 'BioBrick Assembly Standards and Techniques and Associated Software Tools', *DNA Cloning and Assembly Methods*: Humana Press, pp. 1-24.

Salis, H. M., Mirsky, E. A. and Voigt, C. A. (2009) 'Automated design of synthetic ribosome binding sites to control protein expression', *Nature Biotechnology,* 27(10), pp. 946-950.

Sayut, D. J., Niu, Y. and Sun, L. (2009) 'Construction and Enhancement of a Minimal Genetic AND Logic Gate', *Applied and Environmental Microbiology,* 75(3), pp. 637-642.

Scott, M., Gunderson, C. W., Mateescu, E. M., Zhang, Z. and Hwa, T. (2010) 'Interdependence of cell growth and gene expression: Origins and consequences', *Science,* 330(6007), pp. 1099-1102.

Bibliography

Sharma, V., Nomura, Y. and Yokobayashi, Y. (2008) 'Engineering Complex Riboswitch Regulation by Dual Genetic Selection', *Journal of the American Chemical Society,* 130(48), pp. 16310-16315.

Sharp, P. M. and Li, W.-H. (1987) 'The codon adaptation index-a measure of directional synonymous codon usage bias, and its potential applications', *Nucleic Acids Research,* 15(3), pp. 1281-1295.

Shi, H., Zhou, Y., Jia, E., Pan, M., Bai, Y. and Ge, Q. (2021) 'Bias in RNA-seq Library Preparation: Current Challenges and Solutions', *BioMed Research International,* 2021, pp. 1-11.

Shin, J. and Noireaux, V. (2012) 'An *E. coli* Cell-Free Expression Toolbox: Application to Synthetic Gene Circuits and Artificial Cells', *ACS synthetic biology.,* 1(1), pp. 29-41.

Shis, D. L. and Bennett, M. R. (2013) 'Library of synthetic transcriptional AND gates built with split T7 RNA polymerase mutants', *Proceedings of the National Academy of Sciences of the United States of America,* 110(13), pp. 5028-5033.

Sievers, F. and Higgins, D. G. (2021) 'The Clustal Omega Multiple Alignment Package', *Methods Mol Biol,* 2231, pp. 3-16.

Singleton, C., Gilman, J., Rollit, J., Zhang, K., Parker, D. A. and Love, J. (2019) 'A design of experiments approach for the rapid formulation of a chemically defined medium for metabolic profiling of industrially important microbes', *PLoS ONE,* 14(6).

Skelton, D. J., Eland, L. E., Sim, M., White, M. A., Davenport, R. J. and Wipat, A. (2020) 'Codon optimisation for maximising gene expression in multiple species and microbial consortia', *bioRxiv*, pp. 2020.06.30.177766.

Smith, T., Heger, A. and Sudbery, I. (2017) 'UMI-tools: modeling sequencing errors in Unique Molecular Identifiers to improve quantification accuracy', *Genome research,* 27(3), pp. 491-499.

Bibliography

Smith, W. P. J., Wucher, B. R., Nadell, C. D. and Foster, K. R. (2023) 'Bacterial defences: mechanisms, evolution and antimicrobial resistance', *Nature Reviews Microbiology,* 21(8), pp. 519-534.

Snoeck, S., Guidi, C. and De Mey, M. (2024) '"Metabolic burden" explained: stress symptoms and its related responses induced by (over)expression of (heterologous) proteins in Escherichia coli', *Microbial Cell Factories,* 23(1).

Srivastava, S., Jayaswal, N., Kumar, S., Sharma, P. K., Behl, T., Khalid, A., Mohan, S., Najmi, A., Zoghebi, K. and Alhazmi, H. A. (2024) 'Unveiling the potential of proteomic and genetic signatures for precision therapeutics in lung cancer management', *Cellular signalling.,* 113, pp. 110932.

Stock, M. and Gorochowski, T. E. (2024) 'Open-endedness in synthetic biology: A route to continual innovation for biological design', *Science Advances,* 10(3).

Strimbu, K. and Tavel, J. A. (2010) 'What are biomarkers?', *Current Opinion in HIV and AIDS,* 5(6), pp. 463-466.

Sun, C.-Y., Su, T.-F., Li, N., Zhou, B., Guo, E.-S., Yang, Z.-Y., Liao, J., Ding, D., Xu, Q., Lu, H., Meng, L., Wang, S.-X., Zhou, J.-F., Xing, H., Weng, D.-H., Ma, D. and Chen, G. (2016) 'A chemotherapy response classifier based on support vector machines for high-grade serous ovarian carcinoma', *Oncotarget,* 7(3), pp. 3245-3254.

Suzuki, Y., Suzuki, A., Tamaru, A., Katsukawa, C. and Oda, H. (2002) 'Rapid Detection of Pyrazinamide-Resistant *Mycobacterium tuberculosis* by a PCR-Based In Vitro System', *Journal of clinical microbiology.,* 40(2), pp. 501-507.

Tan, S.-I., Hsiang, C.-C. and Ng, I. S. (2021) 'Tailoring Genetic Elements of the Plasmid-Driven T7 System for Stable and Robust One-Step Cloning and Protein Expression in Broad *Escherichia coli*', *ACS Synthetic Biology,* 10(10), pp. 2753-2762.

Bibliography

Tang, T.-C., An, B., Huang, Y., Vasikaran, S., Wang, Y., Jiang, X., Lu, T. K. and Zhong, C. (2020) 'Materials design by synthetic biology', *Nature Reviews Materials,* 6(4), pp. 332-350.

Terpe, K. (2006) 'Overview of bacterial expression systems for heterologous protein production: from molecular and biochemical fundamentals to commercial systems', *Applied Microbiology and Biotechnology,* 72(2), pp. 211-222.

Thind, A. S., Monga, I., Thakur, P. K., Kumari, P., Dindhoria, K., Krzak, M., Ranson, M. and Ashford, B. (2021) 'Demystifying emerging bulk RNA-Seq applications: the application and utility of bioinformatic methodology', *Briefings in Bioinformatics,* 22(6).

Tollerson, R., II and Ibba, M. (2020) 'Translational regulation of environmental adaptation in bacteria', *Journal of Biological Chemistry,* 295(30), pp. 10434-10445.

Traag, V. A., Waltman, L. and Van Eck, N. J. (2019) 'From Louvain to Leiden: guaranteeing well-connected communities', *Scientific Reports,* 9(1).

Valenzuela-Ortega, M. and French, C. (2021) 'Joint universal modular plasmids (JUMP): a flexible vector platform for synthetic biology', *Synthetic Biology,* 6(1).

Van Brempt, M., Clauwaert, J., Mey, F., Stock, M., Maertens, J., Waegeman, W. and De Mey, M. (2020) 'Predictive design of sigma factor-specific promoters', *Nature Communications,* 11(1).

Van Der Kloet, F. M., Buurmans, J., Jonker, M. J., Smilde, A. K. and Westerhuis, J. A. (2020) 'Increased comparability between RNA-Seq and microarray data by utilization of gene sets', *PLOS Computational Biology,* 16(9), pp. e1008295.

Vandana, Priyadarshanee, M. and Das, S. (2023) 'Bacterial extracellular polymeric substances: Biosynthesis and interaction with environmental pollutants', *Chemosphere.,* 332, pp. 138876.

Bibliography

Vasilakou, E., Van Loosdrecht, M. C. M. and Wahl, S. A. (2020) '*Escherichia coli* metabolism under short-term repetitive substrate dynamics: adaptation and trade-offs', *Microbial Cell Factories,* 19(1).

Vishweshwaraiah, Y. L., Chen, J., Chirasani, V. R., Tabdanov, E. D. and Dokholyan, N. V. (2021) 'Two-input protein logic gate for computation in living cells', *Nature Communications,* 12(1).

Wang, A., Liu, H., Yang, J. and Chen, G. (2022) 'Ensemble feature selection for stable biomarker identification and cancer classification from microarray expression data', *Computers in biology and medicine.,* 142, pp. 105208.

Wang, Z., Gerstein, M. and Snyder, M. (2009) 'RNA-Seq: a revolutionary tool for transcriptomics', *Nature Reviews Genetics,* 10(1), pp. 57-63.

Wehrens, R., Franceschi, P., Vrhovsek, U. and Mattivi, F. (2011) 'Stability-based biomarker selection', *Analytica chimica acta.,* 705(1-2), pp. 15-23.

Wilhelm, M. and Hollenberg, C. P. (1985) 'Nucleotide sequence of the *Bacillus subtilis* xylose isomerase gene: extensive homology between the *Bacillus* and *Escherichia coli* enzyme', *Nucleic acids research.,* 13(15), pp. 5717-5722.

Wu, G., Yan, Q., Jones, J. A., Tang, Y. J., Fong, S. S. and Koffas, M. A. G. (2016) 'Metabolic Burden: Cornerstones in Synthetic Biology and Metabolic Engineering Applications', *Trends in Biotechnology,* 34(8), pp. 652-664.

Wu, Y., Li, Y., Jin, K., Zhang, L., Li, J., Liu, Y., Du, G., Lv, X., Chen, J., Ledesma-Amaro, R. and Liu, L. (2023) 'CRISPR–dCas12a-mediated genetic circuit cascades for multiplexed pathway optimization', *Nature Chemical Biology,* 19(3), pp. 367-377.

Xu, X., Meier, F., Blount, B. A., Pretorius, I. S., Ellis, T., Paulsen, I. T. and Williams, T. C. (2023) 'Trimming the genomic fat: minimising and re-functionalising genomes using synthetic biology', *Nature Communications,* 14(1).

Yang, Q., Guo, Y., Xiang, Y., Chen, L., Liu, G., Liu, Y., Shi, J., Hu, L., Liang, Y., Yin, Y., Cai, Y. and Jiang, G. (2023) 'Toward efficient bioremediation of methylmercury

in sediment using *merB* overexpressed *Escherichia coli*', *Water research.,* 229, pp. 119502.

Ye, J., Li, Y., Bai, Y., Zhang, T., Jiang, W., Shi, T., Wu, Z. and Zhang, Y.-H. P. J. (2022) 'A facile and robust T7-promoter-based high-expression of heterologous proteins in *Bacillus subtilis*', *Bioresources and Bioprocessing,* 9(1).

Zegarra, V., Bedrunka, P., Bange, G. and Czech, L. (2023) 'How to save a bacterial ribosome in times of stress', *Seminars in cell & developmental biology.,* 136, pp. 3-12.

Zhang, S. and Voigt, C. A. (2018) 'Engineered dCas9 with reduced toxicity in bacteria: implications for genetic circuit design', *Nucleic acids research.*

Zhang, Z., Kuipers, G., Niemiec, Ł., Baumgarten, T., Slotboom, D. J., De Gier, J.-W. and Hjelm, A. (2015) 'High-level production of membrane proteins in *E. coli* BL21(DE3) by omitting the inducer IPTG', *Microbial Cell Factories,* 14(1).

Zhang, Z.-X., Nong, F.-T., Wang, Y.-Z., Yan, C.-X., Gu, Y., Song, P. and Sun, X.-M. (2022) 'Strategies for efficient production of recombinant proteins in *Escherichia coli*: alleviating the host burden and enhancing protein activity', *Microbial Cell Factories,* 21(1).

Zhou, Z., Dang, Y., Zhou, M., Li, L., Yu, C.-H., Fu, J., Chen, S. and Liu, Y. (2016) 'Codon usage is an important determinant of gene expression levels largely through its effects on transcription', *Proceedings of the National Academy of Sciences,* 113(41), pp. E6117-E6125.

Zhu, B. and Stülke, J. (2018) 'SubtiWiki in 2018: from genes and proteins to functional network annotation of the model organism *Bacillus subtilis*', *Nucleic Acids Research,* 46(D1), pp. D743-D748.

Zur, H. and Tuller, T. (2014) 'Exploiting hidden information interleaved in the redundancy of the genetic code without prior knowledge', *Bioinformatics,* 31(8), pp. 1161-1168.

Bibliography